1 Övningar

1.1 Övningar på kapitel 1, konditionstal, stabilitet, flyttalsaritmetik

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

2. På föreläsningen härledde vi en uppskattning 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?

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

4. När man parallellkopplar motstånd med resistanserna R_1, $R_2, \ldots, R_n$ så ges den totala resistansen, R, av:

$$\frac{1}{R} = \frac{1}{R_1} + \frac{1}{R_2} + \cdots + \frac{1}{R_n}$$

   Antag att osäkerheten i R_k är $\pm 0.1 R_k$. Härled en begränsning av felet i R. Vad gäller vid seriekoppling?

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

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

7. Vi vill mäta höjden på en flaggstång genom att stega ut en sträcka, s, från stången och mäta vinkeln, α, mellan markplanet och flaggstångens topp. Antag att vi har begränsningar på relativa osäkerheter i s och α, så vi har begränsningar av $|\Delta s/s|$ och $|\Delta\alpha/\alpha|$, hur stor relativ osäkerhet har vi i höjden h? För vilka s och α är problemet att mäta h speciellt illa konditionerat? Verkar dina slutsatser rimliga? Om inte, kanske du får göra andra antaganden om felen i s och α. Du kan anta att vi har små osäkerheter så att det går att använda partiella derivator för att uppskatta osäkerheterna.

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

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

   Vi kan tänka oss detta som en funktion också, nämligen den som avbildar α på $\mathbf{x}$, 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 $\mathbf{x}$ som funktion av α. För vilka α är $\mathbf{x}$ känslig för förändringar i α?

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

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

10. Studera hur känslig lösningen till begynnelsevärdesproblemet, $y'(t) = cy(t)$, $y(0) = y_0$ är för ändringar i y_0. c och y_0 är givna reella tal och y är den sökta funktionen (som beror av t).

11. Låt talföljden a_k ges av $a_{k+2} = a_{k+1} + a_k, k = 0, 1, 2, \ldots$ där a_0 och a_1 är givna. Studera hur a_n (givet något stort n) beror av förändringar i a_0 och a_1. Då $a_0 = a_1 = 1$ har vi Fibonaccis talföljd. (Lösningen kräver kunskap om hur man löser linjära differensekvationer.)

12. Under föreläsningen studerade vi hur känsliga nollställena, till ett andragradspolynom, är för störningar i polynomets koefficienter. Försök att generalisera resultatet till ett polynom av grad n (inte en så lätt uppgift). Ledning: betrakta ett nollställe, r, som en funktion av koefficienterna, $a_0, a_1, \ldots, a_n$ och använd implicit derivering. Antag att polynomet har distinkta nollställen.

13. Kontrollera ovanstående uppskattning med hjälp av Matlab. Ledning: `roots` och `poly`.

14. Antag att vi arbetar med fyrsiffrig decimal aritmetik. Beräkna följande summor samt de absoluta och relativa felen. `6.278 + 4.039`, `6.278e10 + 4.039e10` och. `6.278e-10 + 4.039e-10`

15. Visa att addition enligt IEEE är en stabil algoritm.

16. Visa, med hjälp av föregående övning, att matrisaddition enligt IEEE är stabil.

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

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

18. Vi har ett flyttalsystem med basen 10, t=4, 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?

19. Vad blir resultatet av följande Matlab-beräkning? `sum(100.^[1 5 10 15 20])`

20. Vi har ett flyttalsystem med basen 10. 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?

21. Jämför uttrycken $\log x - \log y$ och $\log(x/y)$ ur beräkningssynpunkt. För vilka x och y förväntar vi oss problem?

22. Vilket av uttrycken $x^2 - y^2$ och $(x - y)(x + y)$ ger oss mindre avrundningsfel? För vilka x, y är detta speciellt tydligt?

23. Vi vill lösa ekvationen $x^2 + ax + b = 0$ då vi vet att a och b båda är positiva och där a är mycket större än b. På föreläsningen sa vi att 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 föreläsningen (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.

24. Längden av en vektor definieras som $[\sum_{k=1}^{n} x_k^2]^{1/2}$. Är det lämpligt att implementera en datorprogram direkt efter formeln? Vad skall man göra istället (svårt)?

25. Låt $f(x) = (e^x - 1)/x$. Vi vet att $f(x) \to 1$ då $x \to 0$.
a) Troliggör detta genom att beräkna $f(10^{-k}), k = 1, \ldots, 16$. b) Ge kommandot `man expm1` .

26. Vi vill approximera $f'(x)$ med differenskvoten, $(f(x + h) - f(x))/h$. Vad är ett lämpligt värde för h? Vad gäller om vi använder approximationen $(f(x + h) - f(x - h))/(2h)$?

27. Den harmoniska serien $\sum_{n=1}^{\infty} 1/n$ är divergent. När vi använder flyttalsaritmetik gäller inte detta, varför? Efter ett antal termer ändras inte summan, uppskatta detta antal.

28. Vi kan beräkna standardavvikelsen, σ, på två olika sätt:

$$\sigma = \left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{1/2} = \left[\frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right]\right]^{1/2}$$

$\bar{x}$ är medelvärdet av x-värdena. Vilken formel är att föredra? Varför?

1.2 Övningar på kapitel 2, linjära ekvationssystem

1. $\mathbf{A}$ är en kvadratisk matris vars alla radsummor är noll. Visa att $\mathbf{A}$ är singulär.

2. En magisk kvadrat, av ordning n, är en $n \times n$-matris, $\mathbf{M}_n$, vars alla radsummor, kolonnsummor och två diagonalsummor är lika. Dessutom skall elementen i matrisen vara $1, 2, \ldots, n^2$. Matlabkommandot `magic(n)` returnerar en sådan matris av ordning `n` (större än två).
   Visa ett $\mathbf{M}_n$ har ett egenvärde $n(n^2+1)/2$.

3. a) Visa att matrisen $\mathbf{A}$ är singulär.
$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 2 & 1 \\ 1 & 3 & 2 \end{bmatrix}$$
   b) Hur många lösningar har systemet $\mathbf{Ax} = [2, 4, 6]^T$?

4. Beräkna $\mathbf{A}^{-1}$ då
$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & -1 & 0 \\ 1 & -2 & 1 \end{bmatrix}$$

5. $\mathbf{A}$ är kvadratisk med $\mathbf{A}^2 = \mathbf{0}$. Visa att $\mathbf{A}$ är singulär.

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

7. $\mathbf{A}$ är ickesingulär. Visa att $(\mathbf{A}^T)^{-1} = (\mathbf{A}^{-1})^T$. Vi skriver därför normalt $\mathbf{A}^{-T}$.

8. Beskriv, i punktform, hur man lämpligen löser systemet:
$$\begin{bmatrix} \mathbf{L}_1 & \mathbf{0} \\ \mathbf{B} & \mathbf{L}_2 \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} \mathbf{b} \\ \mathbf{c} \end{bmatrix}$$
   $\mathbf{L}_1$ och $\mathbf{L}_2$ är ickesingulära undertriangulära matriser. Vektorerna har partitionerats så att de passar ihop med blocken i matrisen.

9. Låt $\mathbf{L}_k$ beteckna en enhetsmatris men där elementen i kolonn k under diagonalen får vara skilda från noll (vi använde sådana matriser vid härledningen av LU-faktoriseringen). Visa att a) $\mathbf{L}_k$ är ickesingulär. b) $\mathbf{L}_k = \mathbf{I} - \mathbf{m}_k\mathbf{e}_k^T$ där $\mathbf{m}_k$ är en vektor vars första k element är noll. c) $\mathbf{L}^{-1} = \mathbf{I} + \mathbf{m}_k\mathbf{e}_k^T$. d) $\mathbf{L}_k\mathbf{L}_j = \mathbf{I} - \mathbf{m}_k\mathbf{e}_k^T - \mathbf{m}_j\mathbf{e}_k^T$ om $j > k$.

10. a) Beräkna LU-faktoriseringen av matrisen nedan. b) När är matrisen singulär?
$$\begin{bmatrix} 1 & a \\ c & b \end{bmatrix}$$

11. Visa att en symmetrisk och positivt definit matris $\mathbf{A}$ 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 $\mathbf{D}$, i $\mathbf{LDL}^T$-faktoriseringen (Du får anta att den existerar).

12. Beräkna LU-faktoriseringen av matrisen nedan. Beräkna sedan Choleskyfaktoriseringen.
$$\begin{bmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 5 \end{bmatrix}$$

13. Visa att matrisen nedan saknar LU-faktoriseringen:
$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

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

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

Gör om samma sak men använd definitionen av positivt definit matris. Gör det slutligen med egenvärden.

15. Låt $\mathbf{A} = \mathbf{C}\mathbf{C}^T$ vara Choleskyfaktoriseringen av $\mathbf{A}$. Visa att elementen i $\mathbf{C}$ inte kan bli godtyckligt stora fastän vi inte pivoterar. Man kan visa mycket mer, försök att knyta ihop elementen i $\mathbf{C}$ med de i $\mathbf{A}$.

16. $\mathbf{A}, \mathbf{B}, \mathbf{C} \in \mathbb{R}^{n\times n}$ där $\mathbf{B}$ och $\mathbf{C}$ är ickesingulära. Visa, i punktform, hur man beräknar $\mathbf{x}$ på ett bra sätt (bra vad avser beräkningsfel, cpu-tid och minnesbehov) då $\mathbf{x} = \mathbf{B}^{-1}(2\mathbf{A} + \mathbf{I})(\mathbf{C}^{-1} + \mathbf{A})\mathbf{b}$.

17. $\mathbf{0} \neq \mathbf{u}, \mathbf{v} \in \mathbb{R}^n$. a) Visa att $\text{rang}(\mathbf{u}\mathbf{v}^T) = 1$. Visa att om $\text{rang}(\mathbf{A}) = 1$ så är $\mathbf{A}$ av formen $\mathbf{u}\mathbf{v}^T$.

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

19. Vi generaliserar ovanstående och vill bevisa att $(\mathbf{A} - \mathbf{U}\mathbf{V}^T)^{-1} = \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{U}(\mathbf{I} - \mathbf{V}^T\mathbf{A}^{-1}\mathbf{U})^{-1}\mathbf{V}^T\mathbf{A}^{-1}$. $\mathbf{A}$ är ickesingulär, men $\mathbf{U}, \mathbf{V}$ behöver inte vara kvadratiska matriser (de kan vara vektorer till exempel).

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

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

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

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

24. Låt $\mathbf{D} = \text{diag}(d_1, \ldots, d_n)$ med alla $d_j \neq 0$. Beräkna $\kappa(\mathbf{D})$.

25. Beräkna $\kappa_1(\mathbf{A})$ som funktion av α då:

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

26. Visa att en positivt definit matris är ickesingulär och att inversen är positivt definit.

27. Antag att $\mathbf{A} = \mathbf{B}\mathbf{B}^T$ där $\mathbf{B}$ är ickesingulär. Visa att $\mathbf{A}$ är symmetrisk och positivt definit.

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

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

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

29. Visa att antalet additioner (subtraktioner) och multiplikationer för att beräkna Choleskyfaktorisering är $n^3/6 + \cdots$.

30. NA-Net, `http://www.netlib.org/na-net/`, är en web/e-post-baserad sammanslutning av personer som är intresserade av numerisk analys. Vem som helst kan vara med, från professionella numeriker till allmänt intresserade. Följande fråga skickades in till NA-Net av en professor i datalogi;

> Let S be a symmetric positive definite matrix, and D a diagonal matrix whose entries are in the interval]0,1]. Is the product DS positive definite? In my numerical experiments it appears to be true. Apparently the proof must be easy, but I was enable to find one.

Numerisk analys

Hade Du kunnat hjälpa honom? Man får använda den allmänna definitionen av positivt definit (ingen symmetri krävs).

31. Vi vill lösa $\mathbf{Cz} = \mathbf{d}$ där $\mathbf{C}$ är en **komplex** och kvadratisk matris. $\mathbf{z}$ och $\mathbf{d}$ är komplexa vektorer. Låt oss införa real- och imaginärdelar, $\mathbf{C} = \mathbf{A} + i\mathbf{B}$, $\mathbf{z} = \mathbf{x} + i\mathbf{y}$ och $\mathbf{d} = \mathbf{b} + i\mathbf{c}$. Visa att lösningen ges av problemet:
$$\begin{bmatrix} \mathbf{A} & -\mathbf{B} \\ \mathbf{B} & \mathbf{A} \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} \mathbf{b} \\ \mathbf{c} \end{bmatrix}$$
Är detta en bra metodik (jämfört med att angripa det komplexa problemet direkt)?

32. Antag att $\mathbf{T}$ är en **tridiagonal** matris. a) Diskutera hur man kan implementera en effektiv algoritm för att lösa $\mathbf{Tx} = \mathbf{b}$. Är det bra att använda $\mathbf{T}^{-1}$? b) Implementera Choleskyfaktorisering i det fall då $\mathbf{T}$ dessutom är symmetrisk och positivt definit.

33. Antag att $\mathbf{T}$ är en tridiagonal och ickesingulär matris. Visa att $\mathbf{T}^{-1}$ i allmänhet är full (en anledning till att inte använda inverser). Detta är en **svår** övning.

1.3 Övningar på kapitel 3, minstakvadratproblem

1. Visa med ett enkelt exempel att det är skillnad på linjära ekvationssystem, $\mathbf{Ax} = \mathbf{b}$, och linjära minstakvadratproblem. Är båda problemtyperna alltid lösbara?

2. Använd Matlab för att beräkna den andragradskurva som bäst (i minstakvadratmetodens mening) anpassar följande data:

t	b
1	2.5
2	7.0
3	2.9
4	6.4
5	-3.0
6	-5.8
7	-17.4
8	-31.2
9	-36.7
10	-51.2

($b_k \approx \alpha + \beta t_k + \gamma t_k^2$.)
Hur stor är residualen i minimum?

3. Vi vill lösa minstakvadratproblemet
$$\min_{\mathbf{x}} ||\mathbf{Ax} - \mathbf{b}||_2$$
då $\mathbf{A}$ har ortogonala kolumner ($\mathbf{a}_j^T \mathbf{a}_k = 0$ då $j \neq k$). Hur förenklar denna egenskap hos $\mathbf{A}$ lösandet av problemet?

4. Vi vill anpassa mätpunkter (t_k, N_k) (alla $N_k > 0$) till funktionen
$$N(t) = N_0 e^{-\lambda t}$$
Gör en lämplig omskriving av problemet så att parametrarna, N_0 och λ, i modellen kan bestämmas med hjälp av ett (linjärt) minstakvadratproblem.

5. Givet mätpunkterna

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

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

6. Nedan har vi matrisen, $\mathbf{A}$, i ett minstakvadraproblem. De tre vektorerna är förslag på residualvektorer, $\mathbf{r} = \mathbf{b} - \mathbf{A}\mathbf{x}$, där $\mathbf{x}$ är lösningen. Vilken av vektorerna är en tänkbar residualvektor?

$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}, \quad \mathbf{r}_a = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \quad \mathbf{r}_b = \begin{bmatrix} -1 \\ -1 \\ 1 \\ 1 \end{bmatrix}, \quad \mathbf{r}_c = \begin{bmatrix} -1 \\ 1 \\ 1 \\ -1 \end{bmatrix}$$

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

8. Antag att $\mathbf{A} \in \mathbb{R}^{n \times n}$ är både ortogonal och triangulär. a) Visa att $\mathbf{A}$ är diagonal.
b) Vilka diagonalelement har $\mathbf{A}$?

9. Antag att den partitionerade matrisen nedan är ortogonal ($\mathbf{A}$ och $\mathbf{C}$ är kvadratiska). Visa att $\mathbf{A}$ och $\mathbf{C}$ måste vara ortogonala och att $\mathbf{B} = \mathbf{0}$.

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

10. $\mathbf{A}$ är en kvadratisk matris. a) Visa att två godtyckligt valda villkor nedan medför det tredje.

$$\mathbf{A}^T = \mathbf{A}, \quad \mathbf{A}^T\mathbf{A} = \mathbf{I}, \quad \mathbf{A}^2 = \mathbf{I}$$

b) Hitta ett exempel (ej en permutationsmatris) på en $\mathbf{A}$ av ordning tre som uppfyller villkoren ovan.

11. Visa att matrisen $\mathbf{H} = \mathbf{I} - (2/\mathbf{v}^T\mathbf{v})\mathbf{v}\mathbf{v}^T$ både är ortogonal och symmetrisk.

12. Studera minstakvadratproblemet ur störningssynpunkt då

$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & \sigma \\ 0 & 0 \end{bmatrix}, \quad 0 < \sigma \ll 1, \quad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} \quad \text{med} \quad b_3 \neq 0$$

och då vi stör $\mathbf{A}$ med

$$\mathbf{E} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & \epsilon \end{bmatrix}, \quad 0 < \epsilon \ll \sigma$$

Svårare: Vi definierar $\kappa_2(\mathbf{A})$, för en matris med full kolonnrang, på följande sätt:
$\kappa_2(\mathbf{A}) = ||\mathbf{A}||_2/||\mathbf{F}||_2$ där $\mathbf{F}$ är den matris med minsta $||\mathbf{F}||_2$ sådan att $\text{rang}(\mathbf{A} + \mathbf{F}) < \text{rang}(\mathbf{A})$. Dvs. $\mathbf{F}$ är den minsta matrisen som minskar $\mathbf{A}$:s rang (med ett).

Visa att $\kappa(\mathbf{A}) = 1/\sigma$ genom att ta fram den minimala störningen som gör $\mathbf{A}$ rangdefekt.

13. 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.)

14. Låt $0 < \epsilon < \sqrt{\epsilon_{mach}}$ (så att $fl(1 + \epsilon^2) = 1$) och definiera

$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ \epsilon & 0 \\ 0 & \epsilon \end{bmatrix}$$

Visa att $fl(\mathbf{A}^T\mathbf{A})$ är singulär.

1.4 Övningar på kapitel 5, 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.

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

3. Detta är ett något akademiskt exempel, eftersom vi kan beräkna rötterna exakt. Det visar dock på de problem som uppstår när vi inte kan finna något användbart M i feluppskattningen. 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**.

4. Sätt upp Newtons metod för problemet $x^2 = 1$ och visa att metoden aldrig konvergerar om $x_0 = \alpha i$, $0 \neq \alpha \in \mathbb{R}$. (Vi studerar komplexa x_k med andra ord.) Visa även att det finns cykler (genom att bestämma en med hjälp av Mathematica till exempel). Dvs. hitta $p > 1$ så att $x_0 = x_p = x_{2p} = x_{3p} = \cdots$. Visa slutligen att metoden är konvergerar för **alla reella** $x_0 \neq 0$ (svårare). Avsikten med denna övning är att Du skall se hur komplicerade konvergensegenskaperna hos Newtons metod är.

5. Sätt upp Newtons metod för följande problem: a) $x^3 - 2x - 5 = 0$. b) $e^{-x} = x$. c) $x \sin x = 1$.
   b) Kör de tre metoderna med $x_0 = 1.5$ och skriv ut felen. Kommentar?

6. Newtons metod används ibland för att implementera kvadratrotsfunktionen.
   a) Vi vill beräkna $\sqrt{y}$, sätt upp Newtons metod för problemet.
   b) Antag att x_0 erhålles ur en tabell. Hur många iterationer krävs för att få 24 respektive 53 bitars noggrannhet givet att vi har fyra bitar i x_0?

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

8. a) Visa att $x_{k+1} = x_k - f(x_k)(x_k - x_{k-1})/(f(x_k) - f(x_{k-1}))$ är ett alternativt sätt att formulera sekantmetoden. b) Varför är denna formulering sämre än standardformuleringen?

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

10. Formulera Newtons metod för följande två problem.

$$\begin{cases} x_1^2 + x_2^2 - 1 = 0 \\ x_1^2 - x_2 = 0 \end{cases}, \quad \begin{cases} x_1^2 + x_1 x_2^3 - 9 = 0 \\ 3x_1^2 x_2 - x_2^3 - 4 = 0 \end{cases}$$

11. Tag ett steg av Newtons metod för problemet:

$$\begin{cases} x_1^2 - x_2^2 = 0 \\ 2x_1 x_2 = 1 \end{cases}, \quad \mathbf{x}^{(0)} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

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

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?

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

15. Försök att hitta så många rötter som möjligt till systemet:

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

16. **Svårare.** Låt f och δ vara reella funktioner av en reell variabel (dvs. $f : \mathbb{R} \to \mathbb{R}$, analogt för δ). Vi antag att x^* är ett nollställe till f, $f(x^*) = 0$, och vi vill studera hur x^* påverkas när vi stör funktionen f med funktionen $\epsilon\delta$, för små ϵ. Låt oss vidare anta att de ingående funktionerna är tillräckligt snälla så att den störda roten, $x^*(\epsilon)$, kan utvecklas i en potensserie i ϵ. Dvs. $x^*(\epsilon) = x^* + \sum_{k=1}^{\infty} c_k \epsilon^k$ (för konstanter $c_1, c_2, \ldots$). För tillräckligt små ϵ kommer alltså roten att flytta sig ungefär $c_1\epsilon$. Vi får då följande identitet:

$$f\left(x^* + \sum_{k=1}^{\infty} c_k \epsilon^k\right) + \epsilon\, \delta\left(x^* + \sum_{k=1}^{\infty} c_k \epsilon^k\right) \equiv 0$$

Använd Taylorutveckling för att få ett uttryck för c_1. Du kan anta att alla derivator, som Du måste dividera med, är skilda från noll.

Låt oss nu vidare anta att $f(x^*) = f(x^* + \gamma) = 0$ där gapet, $\gamma \neq 0$. Använd Taylorutveckling för att uppskatta $f'(x^*)$ (i uttrycket för c_1) i termer av γ (och en högre derivata av f) samt ge en tolkning av Ditt resultat.

Applicera slutligen Din formel på fallet när $f(x) = x^2 + ax + b$ och $\delta(x) = x$ respektive $\delta(x) = 1$ och jämför med resultatet som vi fick fram i början av kursen.

1.5 Övningar på kapitel 7, interpolation

1. Först lite träning på polynom. Antag att polynomet p har distinkta nollställen $r_1, \ldots, r_n$. Antag nu att $q(r_k) = 0, k = 1, \ldots, n$. Innebär det att $q = p$? Vad gäller om vi dessutom vet att q har gradtalet n? Vad gäller om vi dessutom vet att $q(\rho) = p(\rho)$ där $\rho \neq r_k, k = 1, \ldots, n$?

2. Antag att polynomet p har ett nollställe r av multiplicitet m. Visa att $p^{(k)}(r) = 0, k = 0, \ldots, m-1$. Kan det inträffa att $p^{(m)}(r) = 0$?

3. Antag att p är ett polynom som inte är konstant. Visa att $|p(t)| \to \infty$ då $|t| \to \infty$.

4. Visa att ett polynom, p, inte kan vara konstant, c, på ett intervall utan att $p(t) = c$ för alla t. Dvs. visa att om $p(t) = c$ för $t \in \{\xi \mid |\xi - t_0| \leq \epsilon\}$, givet ett fixt t_0 och där $\epsilon > 0$, så är $p(t) = c$ för alla t.

5. Normalt gäller att ett polynom p, av grad $n-1$, bestäms entydigt av n villkor $p(t_k) = y_k, k = 1, \ldots, n$. Villkoren måste dock ha viss beskaffenhet för att detta skall gälla. Om vi t.ex. kräver att $p(t_k) = 0, k = 1, \ldots n$ så duger nollpolynomet oavsett vad n är. Visa att det krävs ett polynom av grad minst $n-1$ för att satisfiera följande villkor: Vi kräver att $p(t_k) = y_k, k = 1, \ldots, n$, med $n > 2$ och där $t_k < t_{k+1}, k = 1, \ldots, n-1$. Vidare gäller att $\text{sign}(y_{k+1} - y_k) = (-1)^k, k = 1, \ldots, n-1$, där sign är "tecknet av" ($\text{sign}(a) = -1$ om $a < 0$, $\text{sign}(0) = 0$, $\text{sign}(a) = 1$ om $a > 0$). Annorlunda uttryckt, om vi förbinder punkterna med räta linjesegment så får vi en sicksack-linje.
Observera att det inte räcker att kräva att y_k inte är konstant. (t_k, y_k) med $y_k = t_k^2$ bestämmer ju inte ett polynom av grad $n-1$ ($p(t) = t^2$ duger ju bra).

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

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

8. Antag att $t_1, \ldots, t_n$ är distinkta. Visa att att Vandermondes matris, $\mathbf{A}$, med elementen, $a_{i,j} = t_i^{j-1}$ är ickesingulär.

9. Vi vill interpolera $(t_k, y_k), k = 1, \ldots, n$ 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?

10. Denna övning visar att ekvidistant interpolation kan upphov till väldiga svängningar hos interpolanten. Låt p vara det polynom av grad $n-1$ som satisfierar $p(k) = (-1)^k\epsilon, k = 1, \ldots, n$, $\epsilon > 0$. Enligt övningen ovan så krävs verkligen ett polynom av grad $n-1$ för detta. Visa att $p(0) = -(2^n - 1)\epsilon$ och att $p(n+1) = (-1)^n(2^n - 1)\epsilon$. Polynomet kan svänga tämligen kraftigt i intervallet $[1, n]$ också ($p(3/2)$ blir tämligen stort) men det är svårare att räkna på.

11. Skriv ett Matlabprogram som givet en uppsättning punkter, $(t_k, y_k), k = 1, \ldots, n$, skapar en uppsättning styckvisa andragradspolynom, som interpolerar punkterna. Vidare skall gälla att interpolanten, q, har kontinuerlig förstaderivata. Detta ger oss ett villkor för lite. Låt oss lägga på villkoret $\alpha q'(t_1) + \beta q'(t_n) = \gamma$ (där inte alla α, β, γ är noll).

12. Bestäm (utan att använda Chebyshev-satsen) t_1 och t_2 så att

$$\max_{-1 \le t \le 1} \mid (t - t_1)(t - t_2) \mid$$

minimeras. Det skall gälla att $-1 \le t_1 < t_2 \le 1$.

13. Vi bestämmer interpolationspolynomet, p_n, på $[0, 1]$ som interpolerar e^t i punkterna $0 = t_1 < t_2 < \ldots < t_n = 1$. Visa att oavsett hur vi väljer t_k-punkterna (i övrigt) så gäller:

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

Visa att om vi väljer Chenbyshevpunkterna så gäller att:

$$\max_{0 \le t \le 1} \left|e^t - p_n(t)\right| \le \frac{e}{n!\ 2^{2n-1}}$$

Varför gäller även resultaten för t.ex. $\cos t$?

1.6 Övningar på kapitel 8, kvadratur

1. Skriv en integrationsrutin som använder andragradspolynomen från sista övningen på övningslappen för kapitel 7.

2. Vi har följande kvadraturformel:

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

där vi vet att det polynomiella gradtalet är minst ett. Visa att $\sum_{k=1}^{n} w_k = 1$.

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

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

4. Använd Taylorutveckling för att härleda en ensidig, andra ordningens differensapproximation av $f'(x)$. Differensapproximationen skall utnyttja funktionsvärdena $f(x)$, $f(x+h)$ och $f(x+2h)$.

5. Så här ska man inte göra. Varför?

```
>> quad('1./x', -1, 1.1)
Warning: Minimum step size reached; singularity possible.
ans = 6.036795665603418e-01
```

Är talet av något intresse? Vad hade hänt med en vanlig enkel trapetsmetod?

6.
```
>> quadl('exp(x)./sqrt(x)', 0, 1)
Warning: Divide by zero.
ans = 2.925303907172242e+00
```
Problemet är att $f(0) = \infty$. Dock existerar integralen. Den listiga `quadl` klarar ändå detta problem, det exakta värdet är `2.92530349181436...` (enligt Maple). Använd substitutionen $x = t^2$ och visa att vi då får en snäll integral (där integranden inte har någon singularitet). Ett annat alternativ är partiell integration (integrera $1/\sqrt{x}$). Det kan vara farligt att ändra integrationsintervallet för att bli av med singulariteten. $\int_0^1 dx/x$ är divergent men $\int_a^1 dx/x$ är ändlig för varje $a > 0$.

7. I följande problem är integrationsområdet inte ändligt.
$$\int_0^\infty \frac{dx}{(1+x^2)^{4/3}}$$
Ett tidsödande och lite farligt sätt är att testa `quadl` med olika intervallgränser:

```
>> quadl('1./(1+x.^2).^(4/3)', 0, 10)
ans = 1.107402448680100e+00
>> quadl('1./(1+x.^2).^(4/3)', 0, 100)
ans = 1.119972821978428e+00
>> quadl('1./(1+x.^2).^(4/3)', 0, 1000)
ans = 1.120245308919566e+00
>> quadl('1./(1+x.^2).^(4/3)', 0, 10000)
ans = 1.120251181152987e+00
>> quadl('1./(1+x.^2).^(4/3)', 0, 100000)
ans = 1.120251297910802e+00
```

(Man borde givetvis integrera över delintervallen $[0, 10], [10, 100]$ etc. men ovanstående är lättare att tolka.)
Farligt, eftersom integranden kanske växer utanför integrationsområdet. I detta fall avtar dock integranden och vi torde kanske kunna lita på de första siffrorna.
I vissa fall kan man skriva problemet $\int_a^\infty f(x)dx = \int_a^b f(x)dx + \int_b^\infty f(x)dx$ där den andra integralen enkelt kan begränsas (och den första har **ändligt** integrationsintervall). Gör det!
Ibland kan man göra en substitution, testa $t = 1/(1+x)$. Ett problem med denna substitution är att integrandens derivata är oändlig för $t = 0$ (integranden uppför sig som $t^{2/3}$ för t nära noll). Detta går att åtgärda med ytterligare en substitution, $t = u^3$.

1.7 Övningar på kapitel 9, ordinära differentialekvationer

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

2. Gör som i föregående övning men då $t_0 = 3$.

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

4. Studera lokala felets utseende för Heuns metod och problemet $y' = \lambda y$, $y(0) = 1$. Metoden brukar skrivas

$$\begin{aligned} s_1 &= hf(t_{k-1}, y_{k-1}) \\ s_2 &= hf(t_{k-1} + h, y_{k-1} + s_1) \\ y_k &= y_{k-1} + (s_1 + s_2)/2 \end{aligned}$$

5. En övning på stabilitet: Vi vill undersöka hur störningar, ϵ och δ, påverkar lösningen till följande problem: $y' = \lambda y$, $y(0) = y_0 \neq 0$. För att förenkla problemställningen nöjer vi oss med att studera störda problem på formen: $z' = (1 + \epsilon)\lambda z$, $z(0) = (1 + \delta)y_0$. (z är alltså lösningen till det störda problemet). För vilka $\lambda \neq 0$ gäller att $z(t) - y(t) \to 0$ då $t \to \infty$ då $0 < |\epsilon| < 1$ och $0 < |\delta| < 1$? Existerar några $\lambda \neq 0$ för vilket gäller att det relativa felet, $(z(t) - y(t))/y(t) \to 0$ då $t \to \infty$?

6. Låt $s(t)$ vara en störning i följande ode-problem: $y' = \lambda y + s(t)$, $y(0) = y_0$. Vi vill studera hur lösningens störningskänslighet beror av tecknet på $\lambda \in \mathbb{R}$, $\lambda \neq 0$. $s(t)$ är en kontinuerlig funktion som satisfierar:

$$s(t) = \begin{cases} 0, & t \leq 1 - \delta \\ > 0, & 1 - \delta < t < 1 \\ 0, & 1 \leq t \end{cases}, \quad 0 < \delta < 1$$

   Låt $z(t)$ vara lösning till det ostörda problemet ($z' = \lambda z, z(0) = y_0$) och visa att för $t \geq 1$:

$$|y(t) - z(t)| = \left| s(\xi) \frac{1 - e^{\lambda\delta}}{\lambda} \right| \; e^{\lambda(t-1)}, \quad 1 - \delta < \xi < 1$$

   Vad händer med felet om $\lambda < 0$ respektive $\lambda > 0$?
   Ledning: använd integrerande faktor för att få ett uttryck för $y(t)$ och applicera sedan integralkalkylens medelvärdessats på integralen.

7. Skriv om följande ekvationer som första ordningens system: a) $y'' = t + y + y'$, b) $y''' = y'' + ty$, c) $y''' = y'' - 2y' + y - t + 1$. Använd följande begynnelsvillkor, $y(0) = 1$, $y'(0) = -1$ för a) och dessutom $y''(0) = 3$ för b) och c).

8. Skriv om följande system ekvationer (rörelse-ekvationer för ett tvåkropparsproblem) som ett första ordningens system:

$$\begin{cases} y_1'' = -GMy_1/(y_1^2 + y_2^2)^{3/2} \\ y_2'' = -GMy_2/(y_1^2 + y_2^2)^{3/2} \end{cases}$$

   Använd följande begynnelsvillkor, $y_1(0) = 1$, $y_1'(0) = -1$ $y_2(0) = 0$, $y_2'(0) = 4$.

9. 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.

10. Vilka lösningar har följande problem?

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

11. Jag löste problemet nedan med hjälp av Matlabs `ode45`, och fick då utskriften:

```
Warning: Failure at t=1.999919e+00. Unable to meet
integration tolerances without reducing the step
size below the smallest value allowed (7.105140e-15)
at time t.
```

    Förklara!

$$y' = 2y^{3/2}, \quad y(0) = 1/4$$

12. Eulers metod kan härledas på följande sätt:

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

vilket ger metoden $y_{k+1} = y_k + hf(t_k, y_k)$. Härled en högre ordningens metod genom att ta med nästa term i Taylorutvecklingen (Du får approximera y'' på något bra sätt).

2 Några ord om $fl()$, för IEEE-övningar

$fl(x)$ är det flyttal som ligger närmast x. IEEE kräver att `+, -, *` och `/` skall beräknas korrekt avrundade (om resultatet existerar), vilket medför att:

$$fl(a \otimes b) = (1+\epsilon)(a \otimes b), \quad \text{med} \quad |\epsilon| \leq \epsilon_{mach}$$

där $\otimes$ är en av `+, -, *` och `/`. Detta förutsätter att a och b är flyttal (redan avrundade). ϵ är det aktuella avrundningsfelet som skall vara begränsat av ϵ_{mach}. ϵ kan vara mycket mindre än ϵ_{mach}, t.ex. torde $fl(1.0+3.0) = 4.0$ exakt, dvs. med $\epsilon = 0$. Om man skall begränsa det totala avrundningsfelet får man samla på sig alla avrundningsfel som kan uppstå och sedan på slutet göra en begränsning (denna process kan vara tämligen besvärlig). Det är vanligt att man indexerar de olika avrundningsfelen enligt detta exempel (jag skriver inte ut `*` och vi antar att a, b och c är flyttal):

$$fl(a+bc) = fl(a+fl(bc)) = fl(a+bc(1+\epsilon_1)) = (a+bc(1+\epsilon_1))(1+\epsilon_2)$$

där $|\epsilon_k| \leq \epsilon_{mach}$, $k = 1, 2$. Så, om vi multiplicerar ihop faktorerna

$$fl(a+bc) = a+bc+bc\epsilon_1+(a+bc)\epsilon_2+bc\epsilon_1\epsilon_2 \quad \text{eller} \quad |fl(a+bc)-(a+bc)| = |bc\epsilon_1+(a+bc)\epsilon_2+bc\epsilon_1\epsilon_2|$$

Vi kan nu ge en övre begränsning av det absoluta felet:

$$|fl(a+bc)-(a+bc)| \leq |bc|\epsilon_{mach}+|a+bc|\epsilon_{mach}+|bc|\epsilon_{mach}^2 = ((1+\epsilon_{mach})|bc|+|a+bc|)\epsilon_{mach}$$

För det relativa felet observerar vi (om $a+bc \neq 0$):

$$\frac{|fl(a+bc)-(a+bc)|}{|a+bc|} \leq \frac{(1+\epsilon_{mach})|bc|+|a+bc|}{|a+bc|}\epsilon_{mach} = \left[\frac{(1+\epsilon_{mach})|bc|}{|a+bc|}+1\right]\epsilon_{mach}$$

Så det relativa felet är litet om $|bc/(a+bc)|$ inte är för stort. Om däremot $bc \approx 1$, $a+bc \approx 0$, till exempel, kan vi få ett stort relativt fel.

Observera att ett uttryck som $|a\epsilon_1 - b\epsilon_2| \leq (|a|+|b|)\epsilon_{mach}$ **även** om a och b har **samma** tecken (ϵ_1 och ϵ_2 kan ju ha olika tecken). Att uppskatta $|(a+b)\epsilon_1| \leq (|a|+|b|)\epsilon_{mach}$ är dock onödigt pessimistiskt, ty a och b kan ju ha olika tecken (notera att det är **samma** ϵ_1 för båda termerna).

Om man inte kräver en **strikt** gräns utan endast en uppskattning kan man tillåta sig att slänga t.ex. $\epsilon_1\epsilon_2$-termer (produkter av termer) ty $\epsilon_{mach}^2 \ll \epsilon_{mach}$.

Det är ett elände att få $1+\epsilon_1$-termer i nämnaren (det blir då ännu krångligare). Om vi tillåter oss att ändra ϵ_{mach} lite kan vi flytta upp ϵ-termen i täljaren:

$$\frac{1}{1+\epsilon} = 1+\epsilon', \quad \text{med} \quad |\epsilon'| \leq \tilde{\epsilon}_{mach}$$

Där $\tilde{\epsilon}_{mach}$ är **lite** större än ϵ_{mach}. Det är enkelt att inse, ty: $1/(1+\epsilon) = 1-\epsilon+\epsilon^2-\ldots$.

En annan förenkling är följande

$$\prod_{k=1}^{n}(1+\epsilon_k) = 1+n\epsilon \quad \text{med} \quad |\epsilon| \leq \hat{\epsilon}_{mach}$$

om n inte är alltför stort och där $\hat{\epsilon}_{mach}$ är **lite** större än ϵ_{mach}.
Vi antar från och med nu att ϵ_{mach} är detta lite större värde.

Det är enkelt att inse varför ovanstående förenkling fungerar. Tag t.ex. $n = 3$, då ser vi att:

$$|(1+\epsilon_1)(1+\epsilon_2)(1+\epsilon_3)| = |1+\epsilon_1+\epsilon_2+\epsilon_3+\epsilon_1\epsilon_2+\epsilon_1\epsilon_3+\epsilon_2\epsilon_3+\epsilon_1\epsilon_2\epsilon_3| \leq 1+3\epsilon_{mach}+3\epsilon_{mach}^2+\epsilon_{mach}^3$$

Om $\epsilon_{mach} \approx 10^{-16}$ så är $\epsilon_{mach}^2 \approx 10^{-32}$ och $\epsilon_{mach}^3 \approx 10^{-48}$ och man inser att n måste vara väldigt stort för att $n\epsilon_{mach}^2$ skall komma i närheten av ϵ_{mach}.