

Numerisk analys

Föreläsningsanteckningar

Robert Nyqvist

Förord

Dessa föreläsningsanteckningar är endast ett komplement till kurslitteraturen.¹ Ändringar, tillägg och korrigeringar sker kontinuerligt under aktuell kurs, oftast efter det att varje föreläsning ägt rum. Hur innehållet är fördelat per föreläsning markeras med datum i marginalen och där nästa datum markerar slutet av föreläsningen. Avsnitt som inte direkt ingår i kursen men som bedöms som intressanta att ta med i framställningen markeras i marginalen med trafikmärket *varning för avsmalnaden väg*. Ett sådant avsnitt kan läsas kursivt.

Även en del övningsuppgifter ges i slutet av varje kapitel. Några av uppgifterna är är hämtade från tidigare skrivningar. Det datum då tentamen ägde rum ges i slutet av dessa uppgifter. Om det finns ett svar till en uppgift är uppgiftsnumret en hyperlänk till svaret i facit. Om det till en uppgift finns ett lösningsförslag markeras med ett L i vänstermarginalen. Klicka på L:et för att hoppa till lösningsförslaget. Även korsreferenser till bla formler, figurer och litteraturförteckningen är hyperlänkar. Med Alt och vänsterpil går du tillbaka till den sida där du klickade på en hyperlänk.



[Alt] + [←]

¹Versionsnumret konvergerar mot Feigenbaums konstant $\delta = 4.66921166091029906\dots$

Innehåll

Förord	iii
1 Felanalys och datoraritmetik	1
1.1 Grundläggande definitioner i felanalys	1
1.2 Felfortplantning	2
1.3 Konditionstal	4
1.4 Bakåtfel	5
1.5 Positionssystem	5
1.6 Flyttalssystem	8
1.7 Framåt- och bakåtanalys	12
1.8 Fördjupning	13
1.9 Övningsuppgifter	15
2 Icke-linjära ekvationer	17
2.1 Fixpunktmetoden	17
2.2 Intervallhalveringsmetoden	22
2.3 Newton[-Raphson]s metod	24
2.4 Sekantmetoden	26
2.5 Konvergenshastighet och feluppskattning	27
2.6 Iterativa metoder för ekvationssystem	28
2.7 Övningsuppgifter	39
3 Interpolation	45
3.1 Funktionsapproximation	45
3.2 Interpolation av en funktion	51
3.3 Lagranges interpolation	54
3.4 Newtons interpolationsformel	56
3.5 Potensberäkning	61
3.6 Linjära splinefunktioner	62
3.7 Kubiska splinefunktioner	65
3.8 Bézierkurvor	73
3.9 Övningsuppgifter	87
4 Numerisk integration	93
4.1 Trapetsmetoden	94
4.2 Newton-Cotes kvadraturformler	96
4.3 Rombergs metod	99
4.4 Monte Carlo-metoden	100
4.5 Övningsuppgifter	102

5 Numerisk linjär algebra	105
5.1 Linjära ekvationssystem	105
5.2 Pivotering	106
5.3 LU-faktorisering	109
5.4 Matrismodell och konditionstal	120
5.5 Minsta kvadratmetoden	123
5.6 QR-faktorisering	133
5.7 Singulärvärdesuppdelning	142
5.8 Egenvärdesproblem	147
5.9 Övningsuppgifter	153
6 Ordinära differentialekvationer	159
6.1 Numerisk derivering	159
6.2 Begynnelsevärdesproblem	162
6.3 Runge-Kuttas metod	164
6.4 System av differentialekvationer	166
6.5 Randvärdesproblem	171
6.6 Övningsuppgifter	174
Facit	179
Lösningsförslag	187
Litteraturförteckning	239
Erkännande	241
That's all Folks!	243

Kapitel 1

Felanalys och datoraritmetik

1.1 Grundläggande definitioner i felanalys

20160329

Vid numeriska beräkningar kan det uppkomma flera olika typer av fel.

- Fel i indata, tex från mätinstrument.
- Avrundningsfel, tex att vi representerar π med endast fem siffror.
- Trunkeringsfel, tex serien

$$\sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} + \dots = \frac{\pi^2}{6} \approx 1.644934067$$

som utgörs av oändligt många termer väljer vi att approximera med

$$\sum_{k=1}^{1000} \frac{1}{k^2} = \frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} + \dots + \frac{1}{1000^2} \approx 1.643934567$$

som består av ett ändligt antal termer.

- Överspill eller underspill, tex är resultatet för stort för att kunna lagras på datorn. Notera att "överspill" inte betyder att vi nått oändligheten.
- Division med noll; $1/0$ är inte definierad i \mathbb{R} eller \mathbb{C} , men ändå i $\mathbb{C}_{\infty} = \mathbb{C} \cup \{\infty\}$.
- Programmeringsfel, tex en oändlig loop.
- Fel i design av maskinvaran eller ett hårdvarufel.

I de flesta fall har vi inte kännedom om vilka fel som eventuellt har inträffat – använder vi närmedvärdet vid datorberäkningar så kan vi utgå från att flera avrundningar har gjorts under beräkningarna.

Låt x vara ett tal, som här motsvarar det exakta värdet, dvs är givet med oändlig precision, och låt \hat{x} vara en approximation av x . Det *absoluta felet i \hat{x}* definieras enligt

$$\delta x = \hat{x} - x.$$

En uppskattning Δx av beloppet av det absoluta felet, dvs $|\delta x| \leq \Delta x$, kallas för en *felgräns*. Vi skriver

$$x = \hat{x} \pm \Delta x$$

vilket är ekvivalent med

$$\hat{x} - \Delta x \leq x \leq \hat{x} + \Delta x.$$

Det *relativa felet i \hat{x}* definieras enligt

$$\frac{\delta x}{x} = \frac{\hat{x} - x}{x},$$

då $x \neq 0$. Det största heltalet t som uppfyller

$$|\delta x| \leq 0.5 \cdot 10^{-t}$$

anegs *antal korrekta siffror i \hat{x}* . De siffror, exklusive inledande nollor, i den del av \hat{x} som är större än eller lika med 10^{-t} kallas för *signifikanta siffror*.

Exempel 1.1. Låt $x = \pi$ och $\hat{x} = 3.1416$. Då är

$$\delta x = 3.1416 - \pi = 0.00000734641\dots \leq 7.4 \cdot 10^{-6}$$

det absoluta felet och speciellt är $\Delta x = 7.4 \cdot 10^{-6}$ en främst för approximationen \hat{x} av x . Det relativa felet ges av

$$\frac{3.1416 - \pi}{\pi} = 0.00000233843\dots \approx 0.00023\%.$$

Från $\delta x \leq 0.074 \cdot 10^{-4}$ följer att

$$\pi = 3.1416 \pm 0.074 \cdot 10^{-4} \quad \text{och} \quad 3.1415926 \leq \pi \leq 3.14161074.$$

Vidare ger

$$|\Delta x| \leq 0.0074 \cdot 10^{-4} \leq 0.4 \cdot 10^{-4}$$

att vi har fyra korrekta decimaler i \hat{x} . Notera att vanligtvis vet vi inte det exakta värdet x , och kan därför bara uppskatta storleken hos det absoluta respektive relativa felet. ◊

1.2 Felfortplantning

Med en liten approximation introduceras ett fel, som sedan kommer följa med under beräkningarna och som i värsta fall kan eskalera och generera ett mycket större fel. Hur mycket felet växer beror på vilken metod som man använder.

Exempel 1.2. Vi ska studera det reella talet

$$\alpha = \frac{1}{(\sqrt{122} - 11)^2} = 485.9979424\dots$$

Låt oss approximera $x = \sqrt{122} = 11.04536102\dots$ med

$$\hat{x}_1 = 11.0454 \quad \text{och} \quad \hat{x}_2 = 11.05.$$

Då är

$$\hat{\alpha}_1 = \frac{1}{(\hat{x}_1 - 11)^2} = 485.164 \quad \text{och} \quad \hat{\alpha}_2 = \frac{1}{(\hat{x}_2 - 11)^2} = 400.$$

Vi kan skriva om uttrycket för α till

$$\frac{1}{243 - 22\sqrt{122}} \quad \text{och} \quad 243 + 22\sqrt{122}.$$

Det första uttrycket ger att

$$\frac{1}{243 - 22\hat{x}_1} \approx 833.333 \quad \text{och} \quad \frac{1}{243 - 22\hat{x}_2} = -10.$$

Det andra uttrycket för α ger oss ett bättre resultat, nämligen

$$243 + 22\hat{x}_1 \approx 485.999 \quad \text{och} \quad 243 + 22\hat{x}_2 = 486.1.$$

Exemplet illustrerar hur små fel i nämnaren kan resultera i stora fel i slutändan. \diamond

Sats 1.1 (Medelvärdessatsen). *Låt $a, b \in \mathbb{R}$. Antag att $f: \mathbb{R} \rightarrow \mathbb{R}$ är kontinuerlig på $[a, b]$ och deriverbar på (a, b) . Då är*

$$f(b) - f(a) = f'(\xi)(b - a)$$

för något $\xi \in (a, b)$.

Med hjälp av medelvärdessatsen, se sats 1.1, är det möjligt att uppskatta hur felet i utdata beror på felet i indata till en metod. Vi har nämligen att

$$\delta f(x) = f(\hat{x}) - f(x) = f'(\xi)(\hat{x} - x) = f'(\xi)\delta x,$$

där ξ ligger mellan \hat{x} och x . I praktiken approximerar man $f'(\xi)$ med $f'(\hat{x})$, dvs

$$\delta f(x) \approx f'(\hat{x})\delta x,$$

vilket kallas *felfortplantningsformeln*. Se kursboken för formler för uppskattningsfelet vid aritmetik

Exempel 1.3. Låt $f(x) = \log x$ och $x = 1.25 \pm 0.03$. Då är $\delta x = 0.03$ och

$$\delta f(x) = f'(\xi)\delta x = \frac{1}{\xi} \cdot \delta x.$$

Felfortplantningsformeln ger att

$$\delta f(x) \approx \frac{1}{1.25} \cdot 0.03 \approx 0.024.$$

Vi kan också uppskattar felet enligt

$$\delta f(x) \leq \frac{1}{1.22} \cdot 0.03 \leq 0.025.$$

Eftersom $f'(x)$ är en avtagande funktion då $x > 0$, väljer vi $\xi = 1.22$. Antag att vi funnit att $\log(1.25) \approx 0.22$. Vi kan då skriva

$$f(1.25) = \log(1.25) = 0.22 \pm 0.025$$

med en korrekt decimal, ty $0.025 \leq 0.5 \cdot 10^{-1}$. \diamond

1.3 Konditionstal

20160330

Låt funktionsvärdet $f(x)$ beteckna lösningen till ett givet problem, tex "beräkna ett funktionsvärde" eller "lös en ekvation". Problemets *konditiontal* ges av

$$\kappa = \left| \frac{\delta f(x)}{f(x)} \right| / \left| \frac{\delta x}{x} \right|$$

och är ett mått på hur känsligt problemet är. Om κ är litet medför liten skillnad i indata liten skilnad i utdata – man säger att problemet är *välkonditionerat*. Är deromot κ stort kan även små skillnader i indata ge stora skillnader i utdata och då säger man att problemet är *illkonditionerat*.

Exempel 1.4. Låt $f(x) = 1/x^2 = x^{-2}$. Felfortplantningsformeln ger att

$$\delta f(x) \approx f'(\hat{x})\delta x = -\frac{2}{x^3}\delta x = -2x^{-3}\delta x.$$

Alltså är

$$\kappa \approx \left| \frac{-2x^{-3}\delta x}{x^{-2}} \right| / \left| \frac{\delta x}{x} \right| = \left| \frac{-2x^{-3}\delta x}{x^{-2}} \cdot \frac{x}{\delta x} \right| = 2$$

konditionstalet för att beräkna x^{-2} för ett givet x . \diamond

Anmärkning. Då f är en deriverbar funktion har vi

$$\kappa = \left| \frac{\delta f(x)}{f(x)} \right| / \left| \frac{\delta x}{x} \right| \approx \left| \frac{f'(x)\delta x}{f(x)} \right| \cdot \left| \frac{x}{\delta x} \right| = \left| x \frac{f'(x)}{f(x)} \right|,$$

vilket vi kan se i föregående exempel.

Exempel 1.5. Låt $f(x) = (x - 11)^{-2}$. I exempel 1.2 studerade vi $f(\sqrt{122})$ och upptäckte att f kan vara mycket känslig för små förändringar i indata. Konditionstalet för f är

$$\kappa(x) \approx \left| x \frac{-2(x - 11)^{-3}}{(x - 11)^{-2}} \right| = \left| \frac{2x}{x - 11} \right|,$$

dvs $\kappa(x) \rightarrow \infty$ då $x \rightarrow 11$. Om tex $x = 10^{-1} \sqrt{122} = 1.10454\dots$, så är

$$\kappa(x) = \left| \frac{2 \sqrt{122}}{\sqrt{122} - 110} \right| \approx 0.22,$$

vilket är ett litet konditiontal. Med approximationerna $\hat{x}_1 = 1.10454$ och $\hat{x}_2 = 1.105$ av x får vi

$$f(\hat{x}_1) \approx 0.010212 \quad \text{och} \quad f(\hat{x}_2) \approx 0.010213.$$

Som väntat ger en avvikelse i indata inte så stor avvikelse i utdata. Det relativafelet för $f(\hat{x}_2)$ är

$$\frac{\delta f(x)}{f(x)} \approx \frac{f(\hat{x}_2) - f(x)}{f(x)} \approx 0.000094 = 0.0094\%.$$

Jämför med det relativafelet för $f(\hat{x}_2)$ i exempel 1.2, som är $-0.176951 \approx -17.7\%$. \diamond

1.4 Bakåtfel

Antag att f är inverterbar i en omgivning till x , som också innehåller \hat{x} . Då kan vi uppskatta det sk *bakåtfel* i indata i förhållande till felet i utdata med

$$\delta x = \hat{x} - x \approx f^{-1}(\hat{f}(x)) - x.$$

Med bakåtfel kan man ge ett mått på hur stabilt en algoritm är. Man säger att en algoritm är *stabil* om små skillnande i indata inte påverkar utdata.

Exempel 1.6. Låt $f(x) = \sqrt{1+x}$. Vi approximrar f med

$$\hat{f}(x) = 1 + \frac{x}{2} - \frac{x^2}{8},$$

dvs ett Taylorpolynom kring $x = 0$. Då $0 < x < 1$ är f inverterbar och

$$y = \sqrt{1+x} \Leftrightarrow x = y^2 - 1,$$

dvs $f^{-1}(y) = y^2 - 1$. Låt $x = 0.4$. Då är $\hat{f}(x) = 1.18$. Bakåtfellet är

$$f^{-1}(\hat{f}(x)) - x = f^{-1}(1.18) - 0.4 = (1.18^2 - 1) - 0.4 = -0.0076$$

och framåtfellet är

$$\hat{f}(x) - f(x) \approx -0.0032.$$

Det betyder att med felgränsen $\Delta x = 0.0076$ i indata till approximationen \hat{f} av f har vi i utdata en felgräns på $\Delta f(x) = 0.0032$. \diamond

1.5 Positionssystem

Sats 1.2 (Divisionsalgoritmen). *Låt a och b vara heltal där $b > 0$. Då existerar det entydigt bestämda heltalen q och r sådana att*

$$a = bq + r \quad \text{och} \quad 0 \leq r < b.$$

Heltalen q och r kallas kvot respektive rest.

Sats 1.3. *Låt b vara ett heltal större än 1. För varje positivt heltal a existerar det entydigt bestämda heltalen n, d_0, d_1, \dots, d_n sådana att*

$$a = d_n b^n + d_{n-1} b^{n-1} + \dots + d_2 b^2 + d_1 b + d_0, \tag{1.1}$$

där varje d_i satisfierar $0 \leq d_i < b$ och $d_n \neq 0$.

Eftersom utvecklingen (1.1) är entydig inför vi skrivsättet $a = (d_n d_{n-1} \dots d_2 d_1 d_0)_b$. Heltalet b kallas *basen* för talsystemet och heltalen d_0, d_1, \dots, d_n kallas för *siffrorna för a* . Sats 1.3 slår alltså fast existensen av *positionssystemet*, dvs betydelsen av varje siffra beror på dess position i representationen av heltalet. Om $b = 10$, så säger man att basen är *decimal*. Vi kommer att utlämna b i $(d_n \dots d_1 d_0)_b$ om $b = 10$. Om $b = 2$, så säges basen vara *binär*.

Exempel 1.7. Om $b = 10$, så är $157 = 1 \cdot 100 + 5 \cdot 10 + 7 = 1 \cdot 10^2 + 5 \cdot 10^1 + 7 \cdot 10^0$. \diamond

Man brukar räkna positionssystemet som ett av de stora framstegen under mänsklighetens utveckling – i paritet med uppfinnandet av hjulet samt pilbågen och bemästrandet av elden. Positionssystemet styrka ligger i att aritmetik är enkelt, dvs vi har smidiga algoritmer för att addera, subtrahera, multiplicera och dividera heltal när de representeras på formen (1.1). Andra talsystem som tex det romerska är betydligt mer omständligare.

Exempel 1.8. Att tex beräkna $13 + 28 = 41$ och $13 \cdot 28 = 364$ då de ingående talen är givna med romerska siffror ger oss uttrycken

$$\text{XIII} + \text{XXVIII} = \text{XLI} \quad \text{respektive} \quad \text{XIII} \cdot \text{XXVIII} = \text{CCCLXIV}.$$

Kan du se något mönster? ◊

Exempel 1.9. Utvecklingen av heltal på formen (1.1) ger oss tex en enkel metod för att beräkna produkten av två heltal. Studera följande uppställning för att beräkna produkten $25\,412 \cdot 325 = 8\,258\,900$.

$$\begin{array}{r} 25412 \\ \times \quad 325 \\ \hline 127060 \\ 50824 \\ + \quad 76236 \\ \hline 8258900 \end{array}$$

Vi har att

$$\begin{aligned} 25\,412 \cdot 325 &= 25\,412 \cdot (3 \cdot 10^2 + 2 \cdot 10 + 5) \\ &= 25\,412 \cdot 3 \cdot 10^2 + 25\,412 \cdot 2 \cdot 10 + 25\,412 \cdot 5. \end{aligned}$$

Vilket förklarar varför vi skjuter resultatet ett och två steg åt vänster när man multiplicerar 25 412 med tiotalssiffran 2 respektive med hundratatalssiffran 3. Med andra ord, vi utnyttjar positionssystemet. Vidare är tex

$$\begin{aligned} 25\,412 \cdot 5 &= (2 \cdot 10^4 + 5 \cdot 10^3 + 4 \cdot 10^2 + 1 \cdot 10 + 2) \cdot 5 \\ &= (2 \cdot 10^4 + 5 \cdot 10^3 + 4 \cdot 10^2 + 1 \cdot 10) \cdot 5 + 10 \\ &= (2 \cdot 10^4 + 5 \cdot 10^3 + 4 \cdot 10^2) \cdot 5 + 5 \cdot 10 + (1 \cdot 10 + 0) \\ &= (2 \cdot 10^4 + 5 \cdot 10^3) \cdot 5 + (2 \cdot 10 + 0) \cdot 10^2 + 6 \cdot 10 + 0 \\ &= 2 \cdot 10^4 \cdot 5 + (2 \cdot 10 + 5) \cdot 10^3 + 2 \cdot 10^3 + 0 \cdot 10^2 + 6 \cdot 10 + 0 \\ &= (1 \cdot 10 + 0) \cdot 10^4 + 2 \cdot 10^4 + 7 \cdot 10^3 + 0 \cdot 10^2 + 6 \cdot 10 + 0 \\ &= 1 \cdot 10^4 + 2 \cdot 10^4 + 7 \cdot 10^3 + 0 \cdot 10^2 + 6 \cdot 10 + 0 \\ &= 127060. \end{aligned}$$

Notera hur kvoten q i $q \cdot 10 + r$ följer med som minne till nästa siffra i 25 412 från höger till vänster multipliceras med 5. Givetvis bestämmer man inte $25\,412 \cdot 5$ på ett så tillkrånglat sätt som ovan (ovanstående långa uträkning visar bara med ett exempel hur man bevisar att metoden verkligen fungerar i det allmänna fallet). Istället multiplicerar man i tur och ordning från höger till vänster varje siffra i 25 412 med 5, addera eventuellt minne, skriver ned entalssiffran och behåller eventuell tiotalssiffran som minne (eftersom $9 \cdot 9 = 81$ kan vi aldrig få en hundratatalssiffran). På samma sätt följer den avslutande additionen i uppställningen av $25\,412 \cdot 325$. ◊

Exempel 1.10. Med hjälp av divisionsalgoritmen, sats 1.2, kan vi bestämma siffrorna för ett heltal. Om $b = 10$ och $a = 157$, så får vi följande uppställning.

$$\begin{aligned} 157 &= 15 \cdot 10 + 7 \\ 15 &= 1 \cdot 10 + 5 \\ 1 &= 0 \cdot 10 + 1 \end{aligned}$$

Proceduren är enkel: bestäm kvot och rest vid division med b , dela kvoten med b för att erhålla en ny kvot och avbryt när kvoten blir 0. I vårt fall har vi att

$$157 = (1 \cdot 10 + 5) \cdot 10 + 7 = 1 \cdot 10^2 + 5 \cdot 10 + 7.$$

Notera att vi läser resterna nedifrån och upp i uppställningen. Om $b = 2$, så får vi istället följande resultat.

$$\begin{aligned} 157 &= 78 \cdot 2 + 1 \\ 78 &= 39 \cdot 2 + 0 \\ 39 &= 19 \cdot 2 + 1 \\ 19 &= 9 \cdot 2 + 1 \\ 9 &= 4 \cdot 2 + 1 \\ 4 &= 2 \cdot 2 + 0 \\ 2 &= 1 \cdot 2 + 0 \\ 1 &= 0 \cdot 2 + 1 \end{aligned}$$

Uppifrån och ned har vi att

$$\begin{aligned} 157 &= 78 \cdot 2 + 1 \\ &= (39 \cdot 2 + 0) \cdot 2 + 1 \\ &= ((19 \cdot 2 + 1) \cdot 2 + 0) \cdot 2 + 1 \\ &= (((9 \cdot 2 + 1) \cdot 2 + 1) \cdot 2 + 0) \cdot 2 + 1 \\ &= ((((4 \cdot 2 + 1) \cdot 2 + 1) \cdot 2 + 1) \cdot 2 + 0) \cdot 2 + 1 \\ &= (((((2 \cdot 2 + 0) \cdot 2 + 1) \cdot 2 + 1) \cdot 2 + 1) \cdot 2 + 0) \cdot 2 + 1 \\ &= ((((((1 \cdot 2 + 0) \cdot 2 + 0) \cdot 2 + 1) \cdot 2 + 1) \cdot 2 + 1) \cdot 2 + 0) \cdot 2 + 1. \end{aligned}$$

Jämför med Horners metod, se avsnitt 1.8.1. Multiplicerar vi in samtliga 2:or får vi att

$$157 = 1 \cdot 2^7 + 0 \cdot 2^6 + 0 \cdot 2^5 + 1 \cdot 2^4 + 1 \cdot 2^3 + 1 \cdot 2^2 + 0 \cdot 2 + 1.$$

Alltså är $157 = 10011101_{\text{två}}$. ◊

Sats 1.4. Låt α vara ett reellt tal sådant att $0 \leq \alpha < 1$ och låt b vara ett heltal större än 1. Då kan α uttryckas på ett entydigt sätt på formen

$$\alpha = d_1 b^{-1} + d_2 b^{-2} + d_3 b^{-3} + \dots,$$

där siffrorna d_i för α i basen b är heltalet sådana att $0 \leq d_i < b$ för alla positiva heltalet i , med restriktionen att för varje heltalet N existerar det ett heltalet n sådant att $n \geq N$ och $d_n \neq b - 1$.

Exempel 1.11. Om $\alpha = 1/16$, så är

$$\alpha = 0.0625 = 0 \cdot 10^{-1} + 6 \cdot 10^{-2} + 2 \cdot 10^{-3} + 5 \cdot 10^{-4}$$

och från

$$2^{-4} = 0 \cdot 2^{-1} + 0 \cdot 2^{-2} + 0 \cdot 2^{-3} + 1 \cdot 2^{-4}$$

följer att $\alpha = 0.0001_{\text{två}}$. ◊

Ett reellt tal r har en *heltalsdel* a och en *bråktalsdel* α , dvs $r = a + \alpha$, där a är ett heltal och $0 \leq \alpha < 1$. Låt $a = \lfloor r \rfloor$ och $\alpha = \{r\}$. Andra vanliga beteckningar för bråktalsdelen är $\text{frac}(r)$ och $r \bmod 1$. En bråktalsdels utveckling kan vara oändlig, som tex

$$\frac{1}{13} = 0.076923076923076923076923076\dots$$

$$\sqrt{2} = 1.414213562373095048801688724\dots$$

och

$$\pi = 3.141592653589793238462643383\dots$$

Man kan visa att ett reellt tal är ett rationellt tal om och endast om dess bråktalsdelen är ändlig eller upprepas från om med en siffra oavsett bas, jämför med första exemplet ovan. Varken $\sqrt{2}$ eller π är ett rationellt tal, så deras bråktalsdel är oändlig och saknar upprepning.

Exempel 1.12. Vi har att $0.999\dots = 1$. Att så är fallet följer från

$$0.999\dots = 9 \cdot 10^{-1} + 9 \cdot 10^{-2} + 9 \cdot 10^{-3} + \dots = \sum_{n=1}^{\infty} \frac{9}{10^n} = \frac{9}{10-1} = 1,$$

då den geometriska serien är konvergent eftersom $1/10 < 1$. Notera att $0.999\dots$ inte är en tillåten representation av 1 i sats 1.4 \diamond

Algoritm 1.1. Låt r vara ett reellt tal sådant att $0 \leq r < 1$. Algoritmen nedan bestämmer den binära representationen $(0.d_1d_2d_3\dots)_{\text{två}}$ av r .

1. [Initiera] Sätt $f \leftarrow \{r\}$.
2. [Iterera] För $n = 1, 2, 3, \dots$ sätt $d_n \leftarrow \lfloor 2f \rfloor$ och $f \leftarrow \{2f\}$

Exempel 1.13. Låt $r = 367/50 = 7.34$. Då är $f_0 = \{7.34\} = 0.34$. De två första iterationerna ger att

$$d_1 = \lfloor 2f_0 \rfloor = \lfloor 0.68 \rfloor = 0 \quad \text{och} \quad f_1 = \{2f_0\} = 0.68$$

och

$$d_2 = \lfloor 2f_1 \rfloor = \lfloor 1.36 \rfloor = 1 \quad \text{och} \quad f_2 = \{2f_1\} = 0.36.$$

Slutligen har vi att $r = (111.0101\dots)_{\text{två}}$. \diamond

1.6 Flyttalssystem

Låt x vara ett reellt tal och låt β vara ett heltal större än 1. En dator kan troligtvis inte representera x exakt och därför lagras en approximation av x på formen

$$\pm m \cdot \beta^e,$$

där *mantissan* uppfyller

$$m = d_0.d_1d_2\dots d_t = d_0 + d_1 \cdot \beta^{-1} + d_2 \cdot \beta^{-2} + \dots + d_{t-1} \cdot \beta^{-(t-1)}$$

där $d_k \in \{0, 1\}$ och $d_0 \neq 0$ för $k = 0, 1, \dots, t-1$. Alltså har vi att $1 \leq m < \beta$. Den sk *exponenten* e är ett heltal som uppfyller

$$L \leq e \leq U.$$

Talet 0 representeras genom att sätta $d_0 = d_1 = \dots = d_{t-1} = e = 0$. Antal element som vi kan lagra på denna form ges av

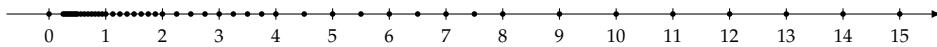
$$\frac{2 \cdot (\beta - 1) \cdot \beta^{t-1} \cdot (U - L + 1) + 1}{d_0, d_1, \dots, d_{t-1}, e, 0} < \infty.$$

Notera att alla dessa tal är rationella, då de har en ändlig bråktalsdel. Låt $\text{fl}(x)$ beteckna approximationen av x i aktuellt flyttalssystem.

Exempel 1.14. Låt $\beta = 2$, $t = 4$, $L = -2$ och $U = 3$. Notera att då måste $d_0 = 1$ när vi representerar ett reellt tal skilt från 0. Antal element på formen

$$\pm(1.d_1d_2d_3)_{\text{två}} \cdot 2^e, \quad -2 \leq e \leq 3$$

är 97. I figuren nedan ser vi hur de 48 positiva talen samt 0 är fördelade på tallinjen.



De resterande 48 talen är negativ och ger en spegelvänt bild. Vi kan tex inte representera talet 8.5. Låt $a = 1/3$ och $b = 4/5$. Då är

$$a = (0.0101010101\dots)_{\text{två}} \quad \text{och} \quad b = (0.1100110011\dots)_{\text{två}}.$$

Vidare är

$$c = a + b = \frac{17}{15} = (1.0010001000\dots)_{\text{två}}.$$

Med flyttalssystemet ovan kan vår dator lagra approximationer av a och b enligt

$$a \approx \text{fl}(a) = 1.010 \cdot 2^{-2} \quad \text{respektive} \quad b \approx \text{fl}(b) = 1.100 \cdot 2^{-1}.$$

Det ger oss att

$$\text{fl}(a + b) = 0.1010 \cdot 2^{-1} + 1.1000 \cdot 2^{-1} = 11.0010 \cdot 2^{-1} \approx 1.100 \cdot 2^0$$

dvs $a + b$ lagras i datorn som $1.100 \cdot 2^0 = 3/2$. Observera att med binär bas är $1+1=10$, vilket ger 1 i minne till nästa siffra (som också kallas *bit*), dvs additionen kan beskrivas med följande uppställning.

$$\begin{array}{r} & 1 & 1 \\ & 0.1010 \\ + & 1.1000 \\ \hline & 11.0010 \end{array}$$

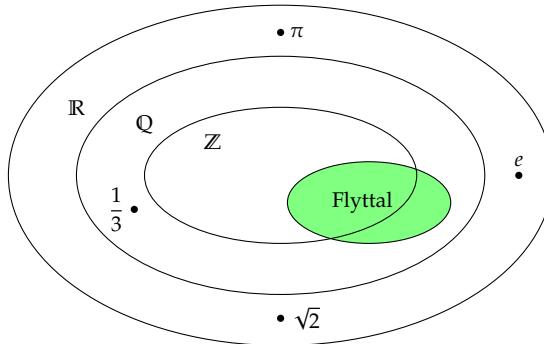
Det absoluta felet är $3/2 - 17/15 = 11/30 \approx 0.367$. ◊

De flesta problemen i denna kurs är formulaterade över \mathbb{R} , men numeriskt kan vi endast "lösa" dem i en ändlig delmängd av \mathbb{Q} , se figur 1.1.

Exempel 1.15. Till skillnad från Matlab använder Mathematica sk *arbiträr precision* vid bla flyttalsberäkningar, vilket innebär att programvaran representera reella tal med så hög precision som hårdvaran tillåter. Det är också möjligt att själv bestämma vilken precision man vill arbeta med. Låt oss mata in ett närmevärde till Ramanujans konstant

$$e^{\pi \sqrt{163}}$$

med 50 siffrors noggrannhet:



Figur 1.1. Katt bland hermelinerna

```
In[1] := a = N[e^π √163], 50]
Out[1] = 2.625374126407687439999999999925007259719818568888 × 1017
```

Därefter adderar vi 1.5 till denna approximation och bestämmer golvet av resultatet (med *golvet* [a] av a avses det största heltal som är mindre än eller lika med a):

```
In[2] := b = Floor[a + 1.5]
Out[2] = 262 537 412 640 768 736
```

Nu bör b vara lika med $\lfloor a \rfloor + 1$ eller $\lfloor a \rfloor + 2$ och speciellt har vi att $-1.5 \leq a - b \leq 1.5$. Fast det är klar, när vi utför beräkningarna i en dator kan vi inte vara säkra på att så är fallet och med den önskade precisionen får vi faktiskt följande resultat.

```
In[3] := a - b
Out[3] = 7.99999999999925007259719818568888
```

Inget vidare bra resultat. Det kanske blir bättre om vi struntar i golvfunktionen?

```
In[4] := a - (a + 1.5)
Out[4] = 0
```

Med andra ord kan inte *Mathematica* representera $a + 1.5$, utan avrundar till det närmsta tal, som i detta fall råkar vara a . Tänk på att vi med parentesen styr i vilken ordning operationerna ska utföras. Hur långt från sanningen är vi?

```
In[5] := e^π √163.0 - a
Out[5] = -480
```

Låter vi *Mathematica* själv styra över precisionen så får vi ett värde som avviker en hel del från vårt a . ◇

Exempel 1.16 (Avrundning och avhuggning). Låt k vara ett positivt heltal och $x \in \mathbb{R}$. Då finns det ett heltal n sådant att

$$x = \pm(d_0.d_1d_2 \dots d_{t-1}d_t d_{t+1} \dots) \cdot 10^n,$$

där d_i är heltal vilka uppfyller $1 \leq d_0 \leq 9$ och $0 \leq d_i \leq 9$ för alla $i \geq 1$. Med t siffrors avrundning av x avses talet

$$\text{fl}_{\text{round}}(x) = \pm(r_0.r_1r_2 \dots r_{k-2}r_{k-1}) \cdot 10^n,$$

där $r_0r_1 \dots r_{t-1}$ är det heltal som är närmast $d_0d_1 \dots d_{t-1}d_t \dots$ och $0 \leq r_i \leq 9$. I gränsfallet då $d_t = 5$ och $d_j = 0$ för alla $j > t$, så väljs siffrorna r_i så att r_{t-1} är jämn. Om $t = 4$, så

har vi bla att

$$\begin{array}{ll} \text{fl}_{\text{round}}(0.23456) = 0.2346 & \text{fl}_{\text{round}}(15.7802) = 15.78 \\ \text{fl}_{\text{round}}(0.0012078) = 0.001208 & \text{fl}_{\text{round}}(78.286) = 78.290 \\ \text{fl}_{\text{round}}(0.45015) = 0.4502 & \text{fl}_{\text{round}}(0.45025) = 0.4502 \\ \text{fl}_{\text{round}}(1.48991) = 1.490 & \text{fl}_{\text{round}}(1.49991) = 1.500. \end{array}$$

Med t siffrors avhuggning av x avses talet

$$\text{fl}_{\text{chop}}(x) = \pm(d_0.d_1d_2\dots d_{t-1}) \cdot 10^n,$$

tex har vi att $\text{fl}_{\text{chop}}(15.138) = 15.13$ och $\text{fl}_{\text{chop}}(78.286) = 78.280$, då $t = 4$. Om $t = 4$, så är

$$\text{fl}_{\text{round}}\left(\frac{140.5}{0.008513}\right) = \text{fl}_{\text{round}}(16504.170092\dots) = 16500$$

medan

$$\text{fl}_{\text{round}}\left(\frac{0.008513}{140.5}\right) = \text{fl}_{\text{round}}(0.00006059074\dots) = 0.00006059.$$

Det relativafelet är 0.000253 respektive 0.0000123. \diamond

Det minsta positiva tal vi kan representera i ett flyttalssystem är $1.00\dots 0 \cdot \beta^L = \beta^L$, som kallas *underspillsgräns*. Det största tal man kan representera i ett flyttalssystem ges av

$$\begin{aligned} & ((\beta - 1) + (\beta - 1)\beta^{-1} + (\beta - 1)\beta^{-2} + \dots + (\beta - 1)\beta^{-(t-1)})\beta^U \\ &= (\beta - 1 + 1 - \beta^{-1} + \beta^{-1} - \beta^{-2} + \dots + \beta^{-(t-1)+1} - \beta^{-(t-1)})\beta^U \\ &= (\beta - \beta^{-(t-1)})\beta^U = (\beta - \beta^{1-t})\beta^U = (1 - \beta^{-t})\beta^{U+1}, \end{aligned}$$

som är flyttalssystemets *överspillsgräns*. Låt $x \in \mathbb{R}$. Då existerar det ett reellt tal α och ett heltalet e sådana att $1 \leq \alpha < \beta$ och $L \leq e \leq U$ samt $x = \pm\alpha\beta^e$. Genom att addera 1 till d_{t-1} får man nästa möjliga mantissa i ett flyttalssystem. Med andra ord ges skillnaden mellan två tal på formen $m = d_0 + d_1\beta^{-1} + \dots + d_{t-1}\beta^{-(t-1)}$ av $\beta^{-(t-1)}$. Eftersom α måste ligga mellan två sådana tal gäller det att

$$-\frac{\beta^{-(t-1)}}{2} \leq m - \alpha \leq \frac{\beta^{-(t-1)}}{2}$$

för det m som är närmast α . Alltså är $\text{fl}(x) = \pm m\beta^e$. Alltså är

$$\left| \frac{\text{fl}(x) - x}{x} \right| = \left| \frac{(\pm m\beta^e) - (\pm \alpha\beta^e)}{\pm \alpha\beta^e} \right| = \frac{|m - \alpha|}{\alpha} \leq \frac{\beta^{-(t-1)}/2}{1} = \frac{\beta^{1-t}}{2}$$

det relativafelet vid flyttalsrepresentation av reella tal. Konstanten $\mu = \beta^{1-t}/2$ kallas för flyttalssystemets *avrundningsenhets* eller *maskintal*.

Exempel 1.17 (Utskiftning). Låt $(\beta, t, L, U) = (10, 3, -2, 2)$ samt låt

$$x = 1.54 \cdot 10^2, \quad y = 1.91 \cdot 10^{-1} \quad \text{och} \quad z = 4.27 \cdot 10^{-1}.$$

Vi kan addera de tre talen enligt bla $(x + y) + z$ och $x + (y + z)$. För att kunna addera skriver vi om y och z så att de har samma exponent som x , dvs

$$y = 0.00191 \cdot 10^2 \quad \text{och} \quad z = 0.00427 \cdot 10^2.$$

Vi får att

$$\begin{aligned}\text{fl}(\text{fl}(x + y) + z) &= \text{fl}(\text{fl}(1.54191 \cdot 10^2) + z) \\ &= \text{fl}(1.54 \cdot 10^2 + z) \\ &= \text{fl}(1.54427 \cdot 10^2) = 1.54 \cdot 10^2 = x\end{aligned}$$

och

$$\begin{aligned}\text{fl}(x + \text{fl}(y + z)) &= \text{fl}(x + \text{fl}(0.00618 \cdot 10^2)) \\ &= \text{fl}(x + 0.00618 \cdot 10^2) \\ &= \text{fl}(1.54618 \cdot 10^2) = 1.55 \cdot 10^2 \neq x.\end{aligned}$$

När vi adderar för flyttalssystemet små tal med stora tal kan det inträffa att den mindre termen är så litet att avrundningen efter additionen returnerar som summa den stora termen, dvs $\text{fl}(x + y) = x$ trots att $y \neq 0$. Vidare ser vi att den associativa lagen $(x + y) + z = x + (y + z)$ inte gäller i ett flyttalssystem. \diamond

1.7 Framåt- och bakåtanalys

20160406

Vi har att

$$\begin{aligned}\left| \frac{\text{fl}(x) - x}{x} \right| &\leq \mu \\ \Leftrightarrow \\ -\mu \leq \frac{\text{fl}(x) - x}{x} &\leq \mu \\ \Leftrightarrow \\ \mp \mu x &\leq \text{fl}(x) + x \leq \pm \mu x \\ \Leftrightarrow \\ x \mp \mu x &\leq \text{fl}(x) \leq x \pm \mu x \\ \Leftrightarrow \\ x(1 \mp \mu) &\leq \text{fl}(x) \leq x(1 \pm \mu),\end{aligned}$$

dvs

$$\text{fl}(x) = x(1 + \delta)$$

för något δ sådant att $|\delta| \leq \mu$.

Exempel 1.18. Studera summan $S = x + y + z$. Låt

$$S_1 = x_1, \quad S_2 = x + y \quad \text{och} \quad S_3 = x + y + z.$$

Alltså är $S = S_3$. I ett flyttalssystem såsom IEEE har vi att

$$\text{fl}(a \odot b) = (a \odot b)(1 + \delta),$$

för något δ sådant att $|\delta| \leq \mu$. I framåtanalys studerar man beräkningarna i den ordning de utförs för att uppskatta fel i slutresultatet. I vårt har vi att

$$S_1 = x$$

$$\begin{aligned}\hat{S}_2 &= \text{fl}(S_1 + y) = \text{fl}(x + y) = (x + y)(1 + \delta_1) \\ \hat{S}_3 &= \text{fl}(\hat{S}_2 + z) = (\hat{S}_3 + z)(1 + \delta_2) = ((x + y)(1 + \delta_1) + z)(1 + \delta_2) = \hat{x} + \hat{y} + \hat{z}\end{aligned}$$

där

$$\begin{aligned}\hat{x} &= x(1 + \delta_1)(1 + \delta_2) \\ \hat{y} &= y(1 + \delta_1)(1 + \delta_2) \\ \hat{z} &= z(1 + \delta_2)\end{aligned}$$

och $|\delta_1| \leq \mu$ och $|\delta_2| \leq \mu$. Det ger att

$$\begin{aligned}|\hat{S}_3 - S| &= |\hat{x} + \hat{y} + \hat{z} - x - y - z| \\ &= |x(\delta_1 + \delta_2 + \delta_1\delta_2) + y(\delta_1 + \delta_2 + \delta_1\delta_2) + z\delta_2| \\ &\leq |x| \cdot |\delta_1 + \delta_2 + \delta_1\delta_2| + |y| \cdot |\delta_1 + \delta_2 + \delta_1\delta_2| + |z| \cdot |\delta_2| \\ &\leq |x|(2\mu + O(\mu^2)) + |y|(2\mu + O(\mu^2)) + |z|\mu\end{aligned}$$

Vi ser att x och y har större inverkan på felet i \hat{S}_3 än vad z har. Alltså bör man vänta med att addera stora tal. Vid bakåtanalsys ställer vi oss frågan hur känslig funktionen eller algoritmen är för skillnader i indata och speciellt hur mycket indata kan variera och ändå ge samma utdata. Vi betraktar avrundningar som en del av algoritmen. Uppskattning av det relativa felet i indata ger att

$$\begin{aligned}\left| \frac{\hat{x} - x}{x} \right| &= \left| \frac{x(1 + \delta_1)(1 + \delta_2) - x}{x} \right| = |\delta_1 + \delta_2 + \delta_1\delta_2| \leq 2\mu + O(\mu^2), \\ \left| \frac{\hat{y} - y}{y} \right| &= \left| \frac{y(1 + \delta_1)(1 + \delta_2) - y}{y} \right| = |\delta_1 + \delta_2 + \delta_1\delta_2| \leq 2\mu + O(\mu^2)\end{aligned}$$

och

$$\left| \frac{\hat{z} - z}{z} \right| = \left| \frac{z(1 + \delta_2) - z}{z} \right| = |\delta_2| \leq \mu.$$

Med andra ord kan x och y variera litet mer än z



1.8 Fördjupning

1.8.1 Horners metod

Låt n vara ett positivt heltal. Studera polynomet



$$p(x) = c_n x^n + c_{n-1} x^{n-1} + \cdots + c_2 x^2 + c_1 x + c_0,$$

där koefficienterna c_0, c_1, \dots, c_n är reella tal. Antalet additioner är n och antalet multiplikationer är

$$n + (n - 1) + \cdots + 2 + 1 + 0 = \frac{n(n + 1)}{2},$$

dvs det totala antalet operationer för att beräkna $p(x)$ är således

$$n + \frac{n(n + 1)}{2} = \frac{n(n + 3)}{2}.$$

Då n är stort är det önskvärt att minimera antal operationer som krävs för att beräkna funktionsvärdet $p(x)$. Vi skriver om polynomet enligt

$$p(x) = (\cdots ((c_n x + c_{n-1}) \cdot x + c_{n-2}) \cdot x + \cdots + c_1) \cdot x + c_0.$$

Då får vi ett uttryck som består av endast $2n$ operationer.

Algoritm 1.2 (Horners metod). Låt $p(x) = c_n x^n + \cdots + c_1 x + c_0$ och $a \in \mathbb{R}$. Följande procedur beräknar $p(a)$.

1. [Initiera] Sätt $b \leftarrow c_n$.
2. [Iterera] För $k = n - 1, \dots, 1, 0$ sätt $b \leftarrow ba + c_k$.
3. [Klar] Returnera b .

Exempel 1.19. Låt $p(x) = 5x^4 - 8x^3 + x^2 + 7x + 3$. Då är

$$p(x) = (((5x - 8)x + 1)x + 7)x + 3.$$

Jämför med

$$p(x) = 5 \cdot x \cdot x \cdot x \cdot x - 8 \cdot x \cdot x \cdot x + 1 \cdot x \cdot x + 7 \cdot x + 3.$$

Här är $n = 4$ och

$$c_0 = 3, c_1 = 7, c_2 = 1, c_3 = -8 \quad \text{och} \quad c_4 = 5.$$

Antag att vi vill beräkna $p(0.25)$. Med Horners metod får vi

$$\begin{aligned} b &\leftarrow c_4 = 5 \\ b &\leftarrow ba + c_3 = 5 \cdot 0.25 - 8 = -6.75 \\ b &\leftarrow ba + c_2 = -6.75 \cdot 0.25 + 1 = -0.6875 \\ b &\leftarrow ba + c_1 = -0.6875 \cdot 0.25 + 7 = 6.82813 \\ b &\leftarrow ba + c_0 = 6.82813 \cdot 0.25 + 3 = 4.70703. \end{aligned}$$

Alltså är $p(0.25) = 4.70703$ eller $p(1/4) = 1205/256$ uttryckt med rationella tal. \diamond

Det är troligt att Horners metod var känd i Kina redan under 300-talet fKr och att den då nämndes i boken *Chiu Chang Suan Shu* (svensk överlättning: *Nio kapitel om den matematiska konsten*). Metoden återupptäcktes 1819 av den engelske skolläraren William George Horner (1786–1837).

1.8.2 Stort ordo



Låt $f: \mathbb{R} \rightarrow \mathbb{R}$. Om det existerar positiva konstanter c och C samt en funktion $g: \mathbb{R} \rightarrow \mathbb{R}$ sådana att

$$|f(x)| \leq C|g(x)| \quad \text{då} \quad |x| \leq c,$$

så säges f vara *stort ordo g nära noll*, vilket betecknas

$$f(x) = O(g(x)) \quad \text{då} \quad x \rightarrow 0.$$

Antag att funktionerna f_1 och f_2 uppfyller

$$|f_1(x) - f_2(x)| \leq C|g(x)| \quad \text{då} \quad |x| \leq c.$$

Då säges f_2 approximerar f_1 *stort ordo g*, vilket betecknas

$$f_1(x) = f_2(x) + O(g(x)) \quad \text{då} \quad x \rightarrow 0.$$

Förväxla inte denna definition med "stort ordo då $x \rightarrow \infty$ ", som tex används för att beskriva en komplexiteten hos en algoritms.

Exempel 1.20. Låt $f(x) = x^4 + 7x^3 - x^2$. Visa att $f(x) = O(x^2)$. ◊

Lösning. Låt $c = 2$ och antag att $|x| \leq c$. Vi får då att

$$\frac{|f(x)|}{|x^2|} = \left| \frac{x^4 + 7x^3 - x^2}{x^2} \right| = |x^2 + 7x - 1| \leq 17,$$

dvs $|f(x)| \leq 17|x^2|$ då $|x| \leq 2$. I detta fall är $C = 17$. Störst värde för $|x^2 + 7x - 1|$ över intervallet $[-2, 2]$ erhålls då $x = 2$. □

Anmärkning. Största och minsta värde för en funktion över ett interval kan vi finna antingen i de punkter där funktionens derivata är noll, i de punkter där derivatan inte existerar eller i intervallets ändpunkter. Om $g(x) = x^2 + 7x - 1$, så är $g'(x) = 2x + 7$. Derivatan av g existerar för alla punkter i intervallet $[-2, 2]$. Då är $g'(x) = 0$ ekvivalent med $x = -7/2$. Men $-7/2$ tillhör inte det aktuella intervallet. Då återsår endast intervallets ändpunkter. Från $g(-2) = -11$ och $g(2) = 17$ följer att

$$-11 \leq x^2 + 7x - 1 \leq 17$$

då $-2 \leq x \leq 2$.

Exempel 1.21. På samma sätt som i föregående exempel kan vi visa att

$$|f(x)| = |x^4 + 7x^3 - x^2| \leq 34|x|$$

då $|x| \leq 2$, dvs $f(x) = O(x)$ då $x \rightarrow 0$. ◊

Låt m och n vara positiva heltal. Ett mer korrekt skrivsätt är $f(x) \in O(x^n)$, vilket ska tolkas som att funktionen f tillhör klassen $O(x^n)$, som innehåller alla funktioner som konvergerar lika snabbt eller snabbare mot 0 som x^n gör då $x \rightarrow 0$. Ju större n desto snabbare konvergerar x^n mot 0. Alltså har vi att $O(x^m) \subseteq O(x^n)$ då $m \geq n$.

1.9 Övningsuppgifter

- L 1. Bestäm både binärt och decimalt samtliga tal som kan representeras i flyttalsystemet i exempel 1.14.
- L 2. Bestäm den binära representationen av följande tal. (20120603)
 - (a) 25
 - (b) $\frac{25}{4}$
 - (c) 2.75
- 3. Skriv om följande binära tal till decimalform. (20130109)
 - (a) $101.101_{\text{två}}$
 - (b) $0.11011_{\text{två}}$
 - (c) $0.00001_{\text{två}}$
- L 4. Skriv om följande binära tal till decimal form.
 - (a) $1.0110101_{\text{två}}$
 - (b) $11.0010010001_{\text{två}}$
- 5. Bestäm den binära representationen av följande tal. (20130823)
 - (a) 27_{tio}
 - (b) 1.75_{tio}
 - (c) $\left(\frac{127}{128}\right)_{\text{tio}}$
- L 6. Skriv om följande rationella tal på binär representation $(0.d_1d_2\dots d_n)_{\text{två}}$.
 - (a) $7/16$
 - (b) $15/16$
 - (c) $23/32$
 - (d) $75/128$

7. Bestäm de flyttal på formen $m \cdot 2^e$ som bäst approximerar följande tal i det flyttalssystemet som ges av $L = -3$, $U = 3$ och $t = 5$. (20150108)
 (a) 3 (b) 3/5 (c) 3/15
8. Representera $\sqrt{3}$ i ett binärt flyttalssystem där $t = 4$, $L = -2$ och $U = 2$. Bestäm sedan det absoluta felet. (20140110)
- L 9. Givet ett normaliserat binärt flyttalssystem där $t = 5$, $L = -4$ och $U = 4$.
- (a) Bestäm de tal $\text{fl}(a)$ och $\text{fl}(b)$ i flyttalssystemet som bäst approximerar de decimala talen $a = 13.4$ respektive $b = 0.226$.
 (b) Beräkna de absoluta felet δa och δb .
 (c) Utför beräkningen $\text{fl}(a) + \text{fl}(b)$ i det givna flyttalssystemet och bestäm därefter det absoluta felet för resultatet (20140603)
- L 10. Låt $\hat{x} = 1.0101 \cdot 2^2$ vara ett flyttal givet i ett normaliserat flyttalssystem. Antag att \hat{x} är ett närmevärde till x . Bestäm samtliga möjliga x då vi vet att antal signifikanta siffror är 2. (20150822)
- L 11. Låt $x = 0.d_1d_2 \dots d_n$ i basen b . Visa att x är ett rationellt tal, dvs att om x har en ändligt bråktalsdel, så är $x = \alpha/\beta$, där α och β är heltalet och $\beta \neq 0$.
- L 12. Visa att det finns talföljder $(x_n)_{n=0}^{\infty}$ av rationella tal sådana att de konvergerar mot ett irrationellt tal.
- L 13. Bestäm det absoluta felet och det relativta felet. Bestäm också antal signifikanta siffror i respektive approximation.
 (a) $x = 2.71828182$ (b) $y = 98.350$ (c) $z = 0.000068$
 $\hat{x} = 2.7182$ $\hat{y} = 98.000$ $\hat{z} = 0.00006$
14. Låt $f(x) = x^4 + 7x^3 - x^2$. Visa att $f(x) \neq O(x^3)$.
15. Låt m och n vara positiva heltalet. När gäller $x^m = O(x^n)$? Då $m \geq n$ eller då $m \leq n$?
16. Låt m och n vara positiva heltalet, där $m \geq n$. Antag att

$$f_1(x) = O(x^m) \quad \text{och} \quad f_2(x) = O(x^n).$$

Visa att

- (a) $f_1(x) + f_2(x) = O(x^n)$ (b) $f_1(x) - f_2(x) = O(x^n)$
 (c) $f_1(x)f_2(x) = O(x^{m+n})$ (d) $f_1(x)/f_2(x) = O(x^{m-n})$

Kapitel 2

Icke-linjära ekvationer

Vissa ekvationer som tex

$$3x^2 - 7x + 1 = 0 \quad \text{och} \quad \sin(x - 7) = \frac{1}{\sqrt{2}},$$

är möjliga att lösa för hand. Men hur gör vi om vi vill bestämma lösningarna till tex ekvationerna

$$x \cos x - e^{\sin x} = 0$$

och

$$x^7 + 2x^6 - 4x^5 + 18x^4 + 94x^3 - x^2 + 27x + 5 = 0$$

för vilka det saknas explicita analytiska metoder?

2.1 Fixpunktmetoden

Låt $g: \mathbb{R} \rightarrow \mathbb{R}$. Om $a \in \mathbb{R}$ uppfyller $g(a) = a$, så kallas a för en *fixpunkt till* g . Antag att vi vill lösa ekvationen $f(x) = 0$, där $f(x)$ är en funktion. Sätt $g(x) = f(x) + x$. Från

$$g(a) = a \Leftrightarrow f(a) + a = a \Leftrightarrow f(a) = 0$$

följer att a är en lösning till $f(x) = 0$ om och endast om a är en fixpunkt till g . Låt $x_0 \in \mathbb{R}$ och definiera

$$x_{n+1} = g(x_n)$$

för alla icke-negativa heltal n . Det ger oss rekursivt följen $(x_n)_{n=0}^\infty = (x_0, x_1, x_2, \dots)$.

Sats 2.1. *Antag att $g: \mathbb{R} \rightarrow \mathbb{R}$ är kontinuerlig. Om följen (x_n) konvergerar mot a , så är a en fixpunkt till g .*

Bevis. Eftersom $\lim_{n \rightarrow \infty} x_n = a$ och g är kontinuerlig har vi att

$$g(a) = g\left(\lim_{n \rightarrow \infty} x_n\right) = \lim_{n \rightarrow \infty} g(x_n) = \lim_{n \rightarrow \infty} x_{n+1} = a,$$

vilket skulle visas. \square

Sats 2.2 (Bolzanos sats eller satsen om mellanliggande värden). *Antag att f är en reellvärd kontinuerlig funktion på det slutna intervallet $[a, b]$. Om c är ett reellt tal som ligger mellan $f(a)$ och $f(b)$, så finns det ett reellt tal $\xi \in [a, b]$ sådant att $f(\xi) = c$.*

Sats 2.3 (Brouwers fixpunktssats). *Låt $g: \mathbb{R} \rightarrow \mathbb{R}$ vara kontinuerlig på $[a, b]$. Om g är surjektiv på $[a, b]$, så har g en fixpunkt i $[a, b]$.*

Bevis. Eftersom g är surjektiv gäller att $a \leq g(x) \leq b$ för alla $a \leq x \leq b$. Om $g(a) = a$ eller $g(b) = b$, så är vi klara. Annars gäller att

$$g(a) > a \quad \text{och} \quad g(b) < b.$$

Bilda funktionen $f(x) = g(x) - x$. Då är

$$f(a) > 0 \quad \text{och} \quad f(b) < 0.$$

Bolzanos sats, se sats 2.2, ger nu att det finns ett $c \in I$ sådant att $f(c) = 0$. Alltså gäller att $g(c) - c = 0$, eller ekvivalent $g(c) = c$, dvs c är en fixpunkt till g . \square

Sats 2.4. *Låt I vara ett begränsat och slutet intervall i \mathbb{R} . Antag att $g: I \rightarrow I$ är kontinuerlig och att $|g'(x)| < 1$ för alla $x \in I$. Då har g exakt en fixpunkt i intervallet I .*

Bevis. Antag att a och b är två olika fixpunkter till g i I , dvs $g(a) = a$ och $g(b) = b$. Medelvärdessatsen ger att det finns ett reellt tal c mellan a och b sådant att

$$g(a) - g(b) = g'(c)(a - b) \Leftrightarrow g'(c) = \frac{g(a) - g(b)}{a - b}.$$

Det ger att

$$g'(c) = \frac{a - b}{a - b} = 1.$$

Men det strider mot antagandet att $|g'(x)| < 1$ för alla $x \in I$. \square

Sats 2.5. *Antag att $g \in C^1$ och att a är en fixpunkt till g . Bilda rekursivt följen $(x_n)_{n=0}^\infty$ enligt $x_n = g(x_{n-1})$, där x_0 väljes enligt nedan.*

- (a) *Om $|g'(a)| < 1$, så existerar det ett öppet interval I kring a sådant att följen (x_n) konvergerar mot a för alla $x_0 \in I$.*
- (b) *Om $|g'(a)| > 1$, så finns det ett öppet interval I kring a sådant att för varje $x_0 \in I$ finns det ett k så att $x_k \notin I$.*

Anmärkning. I fall (a) säger man att a är en *attraherande fixpunkt*, och i fall (b) kallas a för en *repellerande fixpunkt*. Om $|g'(a)| = 1$, så säges a vara en *neutral fixpunkt*.

Bevis. (a) Eftersom g är kontinuerlig finns det ett $\varepsilon > 0$ sådant att $|g'(x)| < K < 1$ för alla $x \in I_\varepsilon = [a - \varepsilon, a + \varepsilon]$ och någon konstant K . Medelvärdessatsen ger nu att

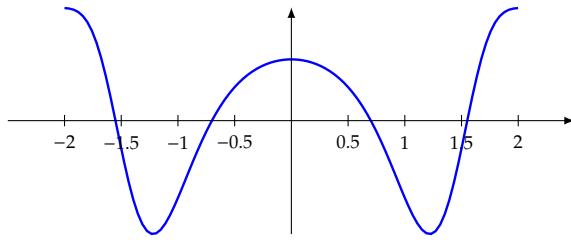
$$|g(x) - a| = |g(x) - g(a)| = |g'(c)| \cdot |x - a| < K|x - a| < |x - a| < \varepsilon,$$

för något $c \in I_\varepsilon$ och för alla $x \in I_\varepsilon$. Det visat att $g(x) \in I_\varepsilon$. Låt $x_0 \in I_\varepsilon$. Vi får då att

$$\begin{aligned} |x_n - a| &= |g(x_{n-1}) - g(a)| < K|x_{n-1} - a| \\ &= K|g(x_{n-2}) - g(a)| < K^2|x_{n-2} - a| = \dots < K^n|x_0 - a|. \end{aligned}$$

Alltså har vi att $x_n \rightarrow a$ då $n \rightarrow \infty$, eftersom $K^n \rightarrow 0$. (b) Eftersom g är kontinuerlig finns det ett $\varepsilon > 0$ sådant att $|g'(x)| > L > 1$ för alla $x \in I_\varepsilon = [a - \varepsilon, a + \varepsilon]$ och någon konstant L . Medelvärdessatsen ger nu att

$$|g(x) - a| = |g(x) - g(a)| = |g'(d)| \cdot |x - a| > L|x - a| > |x - a| > \varepsilon,$$

Figur 2.1. Funktionsgrafen $y = \cos e^{x \sin x}$.

för något $d \in I_\varepsilon$ och för alla $x \in I_\varepsilon$. Låt $x_0 \in I_\varepsilon$. Antag att $x_n \in I_\varepsilon$ för alla positiva heltalet n . Vi får då att

$$\begin{aligned} |x_n - a| &= |g(x_{n-1}) - g(a)| > L|x_{n-1} - a| \\ &= L|g(x_{n-2}) - g(a)| > L^2|x_{n-2} - a| = \dots > L^n|x_0 - a| > L^n\varepsilon. \end{aligned}$$

Men $L^n \rightarrow \infty$ då $n \rightarrow \infty$, vilket betyder att för tillräckligt stort heltalet k är $|x_k - a| > \varepsilon$. Det motsäger att $x_n \in I_\varepsilon$ för alla n . \square

Exempel 2.1. Antag att vi vill bestämma de två minsta positiva lösningarna till

$$\cos e^{x \sin x} = 0.$$

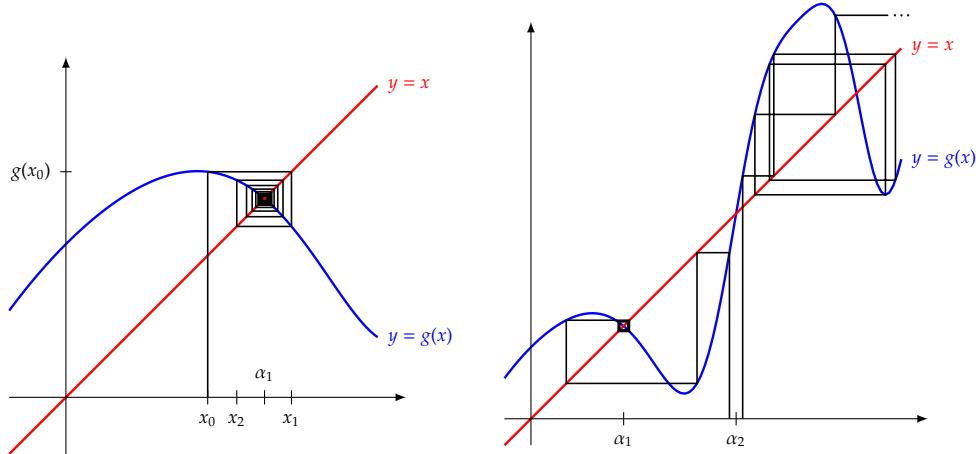
Låt $f(x) = \cos e^{x \sin x}$ och $g(x) = f(x) + x = \cos e^{x \sin x} + x$, se figur 2.1. Vi ser att de två sökta lösningarna ligger nära 0.5 respektive 1.6. Den första lösningen a_1 är en fixpunkt till g och med startvärdet $x_0 = 0.5$ får vi följande iteration:

$$\begin{aligned} x_1 &= g(x_0) \approx 0.795436 \\ x_2 &= g(x_1) \approx 0.602580 \\ x_3 &= g(x_2) \approx 0.765555 \\ x_4 &= g(x_3) \approx 0.636957 \\ x_5 &= g(x_4) \approx 0.746953 \\ x_6 &= g(x_5) \approx 0.656764 \\ x_7 &= g(x_6) \approx 0.734177 \\ x_8 &= g(x_7) \approx 0.669628 \\ x_9 &= g(x_8) \approx 0.725064 \\ &\vdots \\ x_{73} &= g(x_{72}) \approx 0.700535 \\ x_{74} &= g(x_{73}) \approx 0.700533 \\ x_{75} &= g(x_{74}) \approx 0.700535. \end{aligned}$$

Alltså är, $a_1 \approx 0.70053$. Notera att $|g'(a_1)| \leq 0.86 < 1$. Grafiskt kan processen illustreras på följande vis. Gå från punkten $(x_0, 0)$ till $(x_0, g(x_0))$. Sätt $x_1 = g(x_0)$ och gå horisontellt till (x_1, x_1) , som är en punkt på linjen $y = x$. Därefter går vi till $(x_1, g(x_1))$, osv, se vänstra bilden i figur 2.2. För den andra lösningen a_2 har vi att

$$|g'(a_2)| \geq 5.8 > 1.$$

För denna kommer således inte fixpunktmetoden att fungera. Om vi väljer $x_0 < a_2$, så kommer följdens konvergens mot a_1 istället. Om vi väljer $x_0 > a_2$, så kommer följdens divergen, se högra bilden i figur 2.2. \diamond



Figur 2.2. Fixpunktsiteration för första respektive andra lösningen.

Algoritm 2.1 (Fixpunktsmetoden). Låt g vara en funktion, $x_0 \in \mathbb{R}$, $\varepsilon > 0$ och n_{\max} ett positivt heltal. Denna algoritm bestämmer en approximation \hat{a} av en fixpunkt a till g sådan att $\delta a = \hat{a} - a \leq \varepsilon$ om det är möjligt inom n_{\max} iterationer.

1. [Initiera] Sätt $n \leftarrow 1$.
 2. [Iterera] Sätt $x_n \leftarrow g(x_{n-1})$.
 3. [Klar?] Om $|x_n - x_{n-1}| > \varepsilon$ och $n < n_{\max}$, sätt $n \leftarrow n + 1$ och gå till steg 2.
 4. [Utdata] Om $|x_n - x_{n-1}| \leq \varepsilon$, så returnera x_n annars returnera "Misslyckades".
- Stoppkriteriet $|x_n - x_{n-1}| > \varepsilon$ kan ersättas med $|f(x_n)| > \varepsilon$, där f är den funktion som härrör från ekvation $f(x) = 0$, dvs $g(x) = f(x) + x$. Det andra stoppkriteriet $n < n_{\max}$ är till för att undvika en eventuell oändlig loop.

Genom att välja ε kan vi styra önskad precision. Men tänk på att vi inte kan få bättre noggrannhet än vad vi uppnår vid beräkning av g . Även små fel i eventuella koefficienter i uttrycket för g kan försämmra möjligheterna att uppnå hög precision i approximationen av fixpunkten a .

Exempel 2.2 (Den logistiska avbildningen). Låt $f(x) = kx(1-x)$, där k är en positiv reell konstant. Vi ser att f har nollställena $x = 0$ och $x = 1$, dvs $f(0) = f(1) = 0$. Sätt

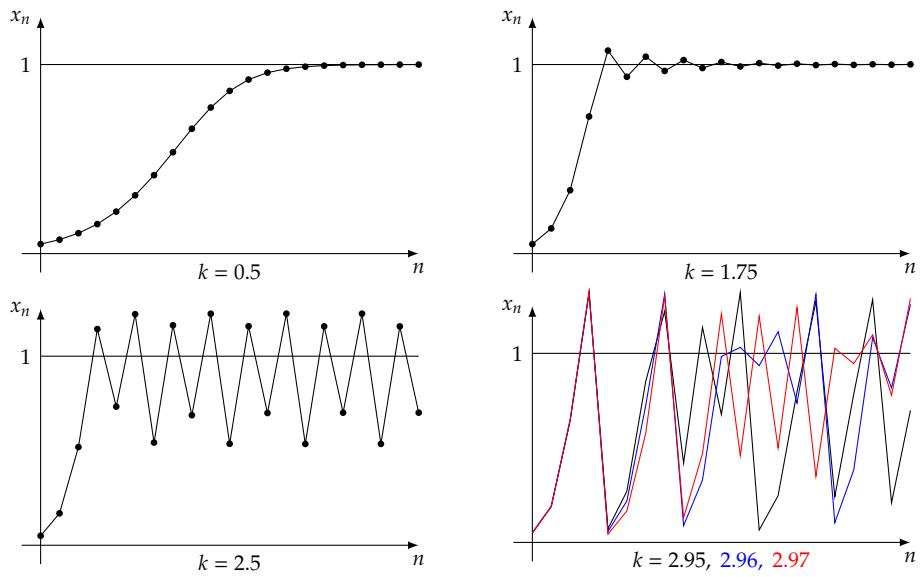
$$g(x) = f(x) + x = (1+k)x - kx^2.$$

Då är 0 och 1 fixpunktter till g . Låt $x_0 = 0.05$ och $x_{n+1} = g(x_n)$. Hur talföljden (x_n) beter sig beror mycket på konstanten k , se figur 2.3. Eftersom $g'(1) = 1-k$, så kommer inte fixpunktsmetoden att konvergera mot 1 om $k \geq 2$. Då $k = 5/2 = 2.5$, så kommer talföljden konvergera mot fyra värden och växelvis "hoppa" mellan dessa. Notera speciellt i fjärde bilden hur små ändringar av k kan resultera i stora skillnader. ◇

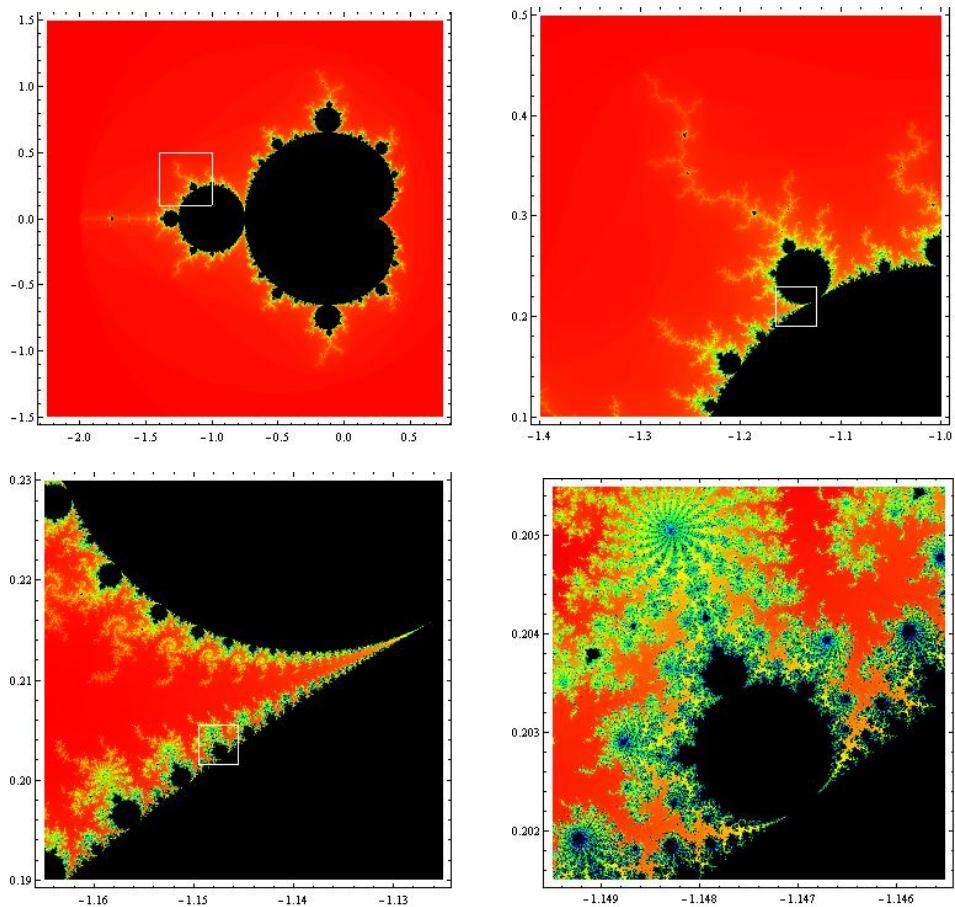
Exempel 2.3 (Komplexta dynamiska system). Låt

$$z_0 = 0 \quad \text{och} \quad z_{n+1} = g(z_n) = z_n^2 + c$$

Mängden av alla $c \in \mathbb{C}$ sådana att $z_n \not\rightarrow \infty$ då $n \rightarrow \infty$ kallas för *Mandelbrotsmängden*, se de svarta områderna i figur 2.4. För en animation se figur 2.5. ◇



Figur 2.3. Iterationer med den logistiska avbildningen.



Figur 2.4. Mandelbrotsmängden för $g(z) = z^2 + c$. Den vita ramen i en bild visar vilken del som nästa bild är en förstoring av.



Figur 2.5. Inzoomning i Mandelbrotsmängden (YouTube: Mandelbrot Zoom 10²²⁷.

2.2 Intervallhalveringsmetoden

Antag att vi vill lösa ekvationen

$$f(x) = 0,$$

där $f: \mathbb{R} \rightarrow \mathbb{R}$ är kontinuerlig. Antag vi funnit ett interval $[a, b]$ sådant att $f(a)$ och $f(b)$ har olika tecken, dvs en av $f(a)$ och $f(b)$ är positiv och den andre är negativ. Enligt sats 2.2 existerar det ett reellt tal c sådant att $a \leq c \leq b$ och $f(c) = 0$. Intervallhalveringsmetoden går ut på att jämför funktionsvärdena i intervallets ändpunkter samt mittpunkt. På det sättet man avgöra i vilket delintervall som lösningen c ligger. Det ger oss bättre lokalisering av lösningen. Med metoden kan man bestämma en grov uppskattning av lösningen, vilken sedan kan användas som startvärde till en iterativ metod.

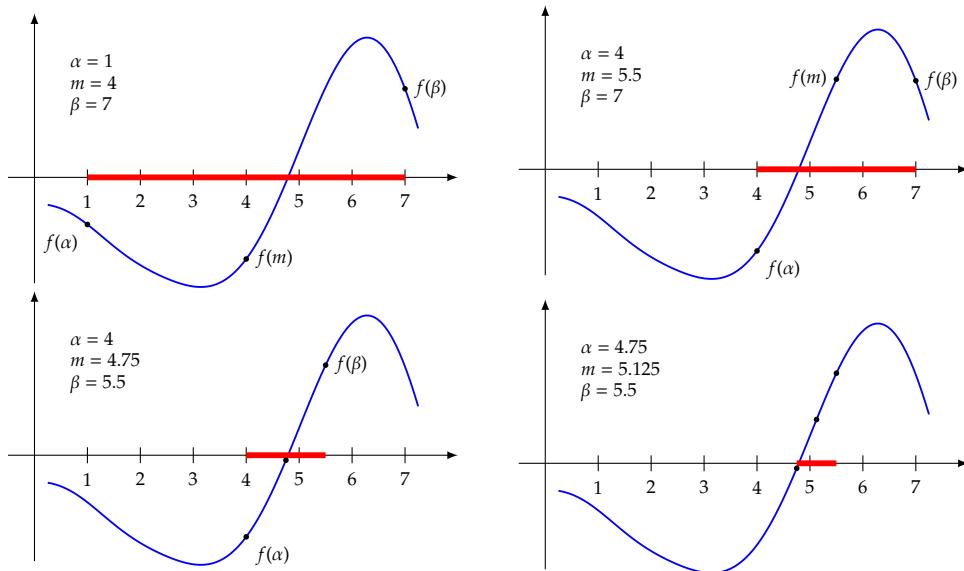
Algoritm 2.2 (Intervallhalveringsmetoden). Låt $f: \mathbb{R} \rightarrow \mathbb{R}$ vara en kontinuerlig på intervallet $[a, b]$. Antag att $f(a)f(b) < 0$. Denna algoritm returnerar ett interval (α, β) som innehåller en lösning till $f(x) = 0$ samt sådan att $\beta - \alpha < \varepsilon$, där $\varepsilon > 0$.

1. [Initiera] Sätt $\alpha \leftarrow a$ och $\beta \leftarrow b$.
2. [Klar?] Om $\beta - \alpha < \varepsilon$, gå till steg 7.
3. [Mittpunkt] Sätt $m \leftarrow (\alpha + \beta)/2$.
4. [Vänster delintervall?] Om $f(\alpha)f(m) < 0$, sätt $\beta \leftarrow m$ och gå till steg 2.
5. [Höger delintervall?] Om $f(\beta)f(m) < 0$, sätt $\alpha \leftarrow m$ och gå till steg 2.
6. [Lösning funnen] Sätt $\alpha \leftarrow m$ och $\beta \leftarrow m$. Alltså är $f(m) = 0$.
7. [Utdata] Returnera (α, β) .

Låt x vara ett reellt tal. Med $\text{sign}(x)$ menas *signum av x* som är lika med 1 om $x > 0$, lika -1 om $x < 0$ och lika med 0 om $x = 0$, dvs

$$\text{sign}(x) = \begin{cases} -1 & \text{om } x < 0 \\ 0 & \text{om } x = 0 \\ 1 & \text{om } x > 0. \end{cases}$$

Denna funktion kan användas för att enkelt avgöra vilket delintervall som den sökta lösningen tillhör.



Figur 2.6. Intervallhalveringsmetoden steg för steg.

Exempel 2.4. Låt

$$f(x) = x \cos x - e^{\sin x}$$

och $\varepsilon = 2^{-6}$. Då har tex $f(1)$ och $f(7)$ olika tecken. Startar vi med intervallet (α, β) , så ger intervallhalveringsmetoden följande resultat.

α	m	β	$\text{sign}(f(\alpha)f(m))$	$\beta - \alpha$
1.00000000	4.00000000	7.00000000	1	6.0000000
4.00000000	5.50000000	7.00000000	-1	3.0000000
4.00000000	4.75000000	5.50000000	1	1.5000000
4.75000000	5.12500000	5.50000000	-1	0.7500000
4.75000000	4.93750000	5.12500000	-1	0.3750000
4.75000000	4.84375000	4.93750000	-1	0.1875000
4.75000000	4.79687500	4.84375000	-1	0.0937500
4.75000000	4.77343750	4.79687500	1	0.0468750
4.77343750	4.78515625	4.79687500	1	0.0234375
4.78515625	4.78515625	4.79687500	1	0.0117188

Efter nio iterationer har vi funnit ett interval $(\alpha, \beta) = (4.78515625, 4.79687500)$ vars längd är mindre än $\varepsilon = 0.015625$. I figur 2.6 ser vi de fyra första intervallerna. \diamond

Sats 2.6. Låt $f \in C[a, b]$ och antag att det existerar ett $c \in (a, b)$ som uppfyller $f(c) = 0$. Låt talföljden (c_n) bestå av alla mittpunkter som erhålls rekursivt med intervallhalveringsmetoden. Då gäller att

$$|c - c_n| \leq \frac{b - a}{2^{n+1}} \quad \text{för alla } n = 0, 1, 2, \dots,$$

dvs $c_n \rightarrow c$ då $n \rightarrow \infty$.



Bevis

2.3 Newton[-Raphson]s metod

Låt $f: \mathbb{R} \rightarrow \mathbb{R}$ vara differentierbar och $x_0 \in \mathbb{R}$. Då definieras *Newtons metod*, även känd som *Newton-Raphsons metod*, för alla icke-negativa heltalet n enligt

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}. \quad (2.1)$$

Notera att med

$$g(x) = x - \frac{f(x)}{f'(x)}$$

är Newtons metod en fixpunktmetod enligt $x_{n+1} = g(x_n)$. Att härleda uttrycket för Newtons metod är enkelt. Låt a vara en lösning till $f(x) = 0$. Antag att x_n är given. Vi vill att nästa element x_{n+1} i talföljden är närmare till a , se figur 2.7. Tangenten till funktionsgrafen $y = f(x)$ i punkten $(x_n, f(x_n))$ ges av

$$y - f(x_n) = f'(x_n)(x - x_n).$$

Sätt x_n till det reella tal där tangenten skär x -axeln. Det betyder att punkten $(x_{n+1}, 0)$ tillhör tangenten, dvs

$$0 - f(x_n) = f'(x_n)(x_{n+1} - x_n),$$

vilket efter omskrivning ger oss formeln (2.1).

Sats 2.7. *Antag att $f \in C^2$ och att a är en lösning till $f(x) = 0$ för vilket gäller att $f'(a) \neq 0$. Sätt $g(x) = x - f(x)/f'(x)$. Då existerar det ett öppet interval I kring a sådant att för alla reella tal $x_0 \in I$ och följd (x_n) som ges av $x_n = g(x_{n-1})$ gäller att $x_n \rightarrow a$ då $n \rightarrow \infty$.*

Bevis. Vi ser att a är en fixpunkt till g , ty

$$g(a) = a - \frac{f(a)}{f'(a)} = a - \frac{0}{f'(a)} = a.$$

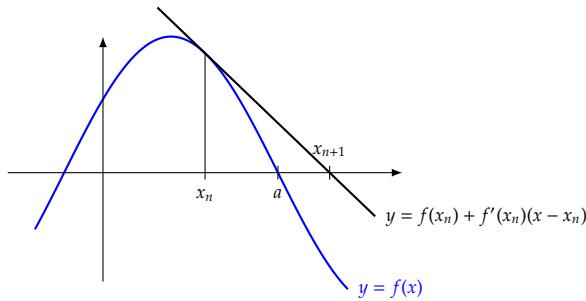
Vidare är

$$g'(x) = 1 - \frac{(f'(x))^2 - f(x)f''(x)}{(f'(x))^2} = \frac{f(x)f''(x)}{(f'(x))^2}.$$

Det ger att

$$|g'(a)| = \left| \frac{f(a)f''(a)}{(f'(a))^2} \right| = 0 < 1.$$

Enligt sats 2.5 existerar det ett öppet interval I kring fixpunkten a så att för alla $x_0 \in I$ konvergerar (x_n) mot a . \square



Figur 2.7. Motivering av Newtons metod.

Exempel 2.5. Låt $f(x) = \cos e^{x \sin x}$. Vi hade problem att bestämma den näst minsta positiva lösningen a_2 till $f(x) = 0$ med fixpunktmetoden, se exempel 2.1. Derivatan av funktionen f är

$$f'(x) = -\sin(e^{x \sin x})(x \cos x + \sin x)e^{x \sin x}.$$

Uttrycket (2.1) är något komplicerat i detta exempel. Newtons metod ger oss följande resulat.

$$\begin{aligned} x_0 &= 1.7 \\ x_1 &= x_0 - \frac{f(x_0)}{f'(x_0)} = 1.7 - \frac{f(1.7)}{f'(1.7)} \approx 1.50426 \\ x_2 &= x_1 - \frac{f(x_1)}{f'(x_1)} \approx 1.55106 \\ x_3 &= x_2 - \frac{f(x_2)}{f'(x_2)} \approx 1.55051 \\ x_4 &= x_3 - \frac{f(x_3)}{f'(x_3)} \approx 1.55051 \end{aligned}$$

Metoden fungerar och redan efter fyra iterationer har vi bra approximation av den sökta lösningen. \diamond

Exempel 2.6. Det kan dock hända att man även misslyckas med Newtons metod. Antag att vi vill lösa

$$x^2 - x + 1 = 0.$$

Sätt $f(x) = x^2 - x + 1$. Iterationsformeln (2.1) ges i detta fall av

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = x_n - \frac{x_n^2 - x_n + 1}{2x_n - 1} = \frac{x_n^2 - 1}{2x_n - 1}.$$

Det är alltid lämpligt att förenkla uttrycket så mycket som möjligt och även skriva om det för att undvika att stora beräkningsfel, se exempel 1.15. Eftersom ingen lösning är nära $1/2$ så behövs här ingen omskrivning. Med $x_0 = 0$ som startvärde så får vi följden $(0, 1, 0, 1, 0, \dots)$, dvs en följd som divergerar.

Faktum är att oavsett vilket reellt tal vi väljer som startvärde kommer Newtons metod att misslyckas. Anledningen är att $f(x) = 0$ har endast två komplexa lösningar, nämligen $(1 - i\sqrt{3})/2$ och $(1 + i\sqrt{3})/2$, där i är den imaginära enheten, dvs $i^2 = -1$. Om x_n är ett reellt tal, så är också x_{n+1} ett reellt tal. Låt $x_0 = 0.3 + 0.6i$. Då fås att

$$\begin{aligned} x_1 &= 0.5875 + 0.8625i, & x_2 &= 0.500091 + 0.861603i, & x_3 &= 0.5 + 0.866037i, \\ x_4 &= 0.5 + 0.866025i & \text{och} & & x_5 &= 0.5 + 0.866025i. \end{aligned}$$

Startar vi med ett komplext tal nära en av lösningarna så har vi större chans att bestämma en approximation av lösningen. \diamond

Exempel 2.7. Ekvationen

$$\sin x - \frac{x}{2} = 0$$

har tre reella lösningar, $a_1 \approx -1.89549$, $a_2 = 0$ och $a_3 \approx 1.89549$. Sätt $f(x) = \sin x - x/2$. Vi vill således lösa ekvationen $f(x) = 0$. Newtons iterationsformel ges av

$$x_{n+1} = x_n - \frac{\sin x_n - x_n/2}{\cos x_n - 1/2} = \frac{2(\sin x_n - x_n/2)}{2\cos x_n - 1}.$$

Med tex $x_0 = 2$ som startvärde får vi en talföljd som konvergerar mot $a_3 \approx 1.89549$. Men valet av startvärde är känsligt. Nedan redovisas vad som händer om man väljer olika x_0 i intervallet $[1.00, 1.20]$, dvs $x_n \rightarrow L$ då $n \rightarrow \infty$.

x_0	L	x_0	L	x_0	L	x_0	L
1.00	a_3	1.05	∞	1.10	a_3	1.15	a_2
1.01	a_1	1.06	a_3	1.11	a_1	1.16	a_2
1.02	∞	1.07	a_3	1.12	a_1	1.17	a_3
1.03	a_1	1.08	a_3	1.13	∞	1.18	a_3
1.04	a_3	1.09	a_3	1.14	a_2	1.19	a_3

En liten ändring av x_0 resulterar i stor skillnad i slutändan, där det är svårt att förutsäga om talföljden kommer konvergerar mot någon av lösningarna eller rent av divergera. Man kallas oftast detta för *kaos*. \diamond

Exempel 2.8. Låt

$$f(x) = \frac{x}{\sqrt{|x|}}$$

då $x \neq 0$ och $f(0) = 0$. Då är

$$f'(x) = \frac{2|x| - \text{sign}(x)x}{2\sqrt[3]{|x|}}$$

då $x \neq 0$. Därmed är

$$x - \frac{f(x)}{f'(x)} = x - \frac{x}{\sqrt{|x|}} \frac{2\sqrt[3]{|x|}}{2|x| - \text{sign}(x)x} = x - \frac{2x|x|}{2|x| - \text{sign}(x)x}.$$

Tag $x_0 = a > 0$. Det ger att

$$x_1 = a - \frac{2a|a|}{2|a| - \text{sign}(a)a} = a - \frac{2a^2}{2a - a} = a - 2a = -a$$

och

$$x_2 = -a - \frac{2(-a)|-a|}{2|-a| - \text{sign}(-a)(-a)} = -a + \frac{2a^2}{2a - a} = -a + 2a = a.$$

Det betyder att $x_3 = -a$, $x_4 = a$, $x_5 = -a$, osv. \diamond

2.4 Sekantmetoden

Om talföljden (x_n) konvergerar, så minskar avståndet mellan konsekutiva element i följen då n ökar, dvs $x_n - x_{n-1} \rightarrow 0$ då $n \rightarrow \infty$. Från definitionen av derivata av en funktion följer att

$$f'(x_n) \approx \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}.$$

Om $f'(x_n) \neq 0$, så har vi således att

$$\frac{1}{f'(x_n)} \approx \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})}.$$

Insättning i (2.1) ger oss formeln

$$x_{n+1} = x_n - f(x_n) \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})}. \quad (2.2)$$

Sekantmetoden går ut på att man väljer två startvärden x_0 och x_1 . Därefter beräknar man nästa tal i följen med

$$x_{n+1} = g(x_{n-1}, x_n),$$

där g är högerledet av (2.2). Notera att sekantmetoden inte är en fixpunktmetod eftersom nästa tal i följen (x_n) beror på de två föregående. En fördel med denna metod är att vi inte behöver beräkna $f'(x_n)$.

2.5 Konvergenshastighet och feluppskattning

20160412

Låt (x_n) vara en följd sådan att $x_n \rightarrow a$ då $n \rightarrow \infty$. Definiera $\delta x_n = x_n - a$. Om p är det största positiva tal sådant att

$$\lim_{n \rightarrow \infty} \frac{|x_{n+1} - a|}{|x_n - a|^p} = \lim_{n \rightarrow \infty} \frac{|\delta x_{n+1}|}{|\delta x_n|^p} = C < \infty,$$

så säger man att p är konvergensordningen för talföljden och C kallas för den asymptotiska felkonstanten. Om $p = 1$ eller $p = 2$, så säges konvergensen vara linjär respektive kvadratisk. Ju större p ju snabbare konvergens.

Sats 2.8. Låt a var en lösning till $f(x) = 0$, där $f \in C^2$, dvs både första- och andradervatan av f är kontinuerlig. Antag att a är ett enkelt nollställe till funktionen f , dvs $f'(a) \neq 0$. Om talföljden (x_n) är genererad med Newtons iterationsformel, så är

$$|\delta x_{n+1}| \approx \frac{|f''(a)|}{2|f'(a)|} |\delta x_n|^2$$

för tillräckligt stora n . Med andra ord är konvergenshastigheten vid enkelrötter kvadratisk, och den asymptotiska felkonstanten för Newtons metod för en enkelrot a är $C = |f''(a)/(2f'(a))|$.

Bevis. Vi har att

$$\delta x_{n+1} = x_{n+1} - a = x_n - \frac{f(x_n)}{f'(x_n)} - a.$$

Taylorutveckling av f kring x_n ges av

$$f(x) = f(x_n) + f'(x_n)(x - x_n) + \frac{1}{2}f''(c_n)(x - x_n)^2$$

för något c_n mellan x och x_n . Med $x = a$ så får vi att

$$0 = f(x_n) + f'(x_n)(a - x_n) + \frac{1}{2}f''(x_n)(a - x_n)^2$$

eller ekvivalent

$$x_n - \frac{f(x_n)}{f'(x_n)} - a = \frac{f''(c_n)}{2f'(x_n)}(x_n - a)^2$$

Alltså har vi att

$$\delta x_{n+1} = \frac{f''(c_n)}{2f'(x_n)} \delta x_n^2.$$

Notera att c_n ligger mellan a och x_n . Eftersom f' och f'' är kontinuerliga och $x_n \rightarrow a$ då $n \rightarrow \infty$, så följer det att $f'(x_n) \rightarrow f'(a)$ och $f''(c_n) \rightarrow f''(a)$ då $n \rightarrow \infty$. \square

Antag att α är en lösning till ekvationen $f(x) = 0$ och att $\hat{\alpha}$ är en approximation av α . Sätt $\delta\alpha = \hat{\alpha} - \alpha$. Vi har att

$$f(\hat{\alpha}) = f(\alpha) + f'(\alpha)(\hat{\alpha} - \alpha) + O((\hat{\alpha} - \alpha)^2) = 0 + f'(\alpha)\delta\alpha + O(\delta\alpha^2).$$

Om $\hat{\alpha}$ är nära α , dvs om $\delta\alpha$ är litet, så är

$$f(\hat{\alpha}) \approx f'(\alpha)\delta\alpha$$

Om $f'(\alpha) \neq 0$, dvs om α är en enkelrot, så är

$$\delta\alpha \approx \frac{f(\hat{\alpha})}{f'(\alpha)} \approx \frac{f(\hat{\alpha})}{f'(\hat{\alpha})}.$$

Notera att denna uppskattning av felet i $\hat{\alpha}$ är oberoende av vilken metod vi använt för att bestämma $\hat{\alpha}$.

Exempel 2.9. Låt $f(x) = \cos e^{x \sin x}$ och $\hat{\alpha} = 0.70053$. Då är $f(\hat{\alpha}) = 0 + \delta f$. Det ger att

$$|\delta\alpha| \lesssim \frac{|0 + \delta f|}{|f'(\hat{\alpha})|} \approx 0.5399 |\delta f|$$

som ett mått på osäkerheten i approximationen i relation till det absoluta felet vid funktionsberäkning. Den metodoberoende feluppskattningen är

$$|\delta\alpha| \lesssim \left| \frac{f(\hat{\alpha})}{f'(\hat{\alpha})} \right| \approx 4.22 \cdot 10^{-6}.$$

I exempel 2.1 behövde vi 75 iterationer med fixpunktmetoden för att bestämma $\hat{\alpha}$. \diamond

2.6 Iterativa metoder för ekvationssystem

Låt n vara ett heltal större än 1. Antag att funktionerna $f_k: \mathbb{R}^n \rightarrow \mathbb{R}$ är differentierbara, där $k = 1, 2, \dots, n$. Studera ekvationssystemet

$$\begin{cases} f_1(x_1, x_2, \dots, x_n) = 0 \\ f_2(x_1, x_2, \dots, x_n) = 0 \\ \vdots \\ f_n(x_1, x_2, \dots, x_n) = 0. \end{cases} \quad (2.3)$$

Sätt $x = (x_1, x_2, \dots, x_n)$ och $f(x) = (f_1(x), f_2(x), \dots, f_n(x))$. Då är (2.3) ekvivalent med

$$f(x) = \mathbf{0}.$$

Notera att $f: \mathbb{R}^n \rightarrow \mathbb{R}^n$.

Exempel 2.10. För systemet

$$\begin{cases} x + xy^3 = \log(z) \\ e^{x+y} \sin(yz) = x \\ xyz - \cos(x^y) = 0 \end{cases}$$

har vi att $x = (x, y, z)$ samt

$$f_1(x, y, z) = x + xy^3 - \log(z), \quad f_2(x, y, z) = e^{x+y} \sin(yz) - x$$

och $f_3(x, y, z) = xyz - \cos(x^y)$.

Den vektorvärda funktionen f har tre komponenter, dvs $f(x) = (f_1(x), f_2(x), f_3(x))$. \diamond

2.6.1 Vektornorm

Låt n vara ett positivt heltal. En funktion $\|\cdot\|: \mathbb{R}^n \rightarrow \mathbb{R}$, enligt $x \mapsto \|x\|$, som uppfyller

- (a) $\|x\| \geq 0$, med likhet om och endast om $x = \mathbf{0}$
- (b) $\|ax\| = |a| \cdot \|x\|$
- (c) $\|x + y\| \leq \|x\| + \|y\|$ (triangelolikheten)

för alla vektorer x och y i \mathbb{R}^n och alla reella tal a , kallas för en *vektornorm på \mathbb{R}^n* . Om axiomen ovan är uppfyllda utom att $\|x\| = 0$ för något $x \neq \mathbf{0}$, så säges funktionen $\|\cdot\|$ vara en *seminorm*.

Exempel 2.11. En mycket viktig familj av vektornormer är de sk ℓ^p -normerna:

$$\|x\|_p = (\|x_1\|^p + \|x_2\|^p + \cdots + \|x_n\|^p)^{1/p}, \quad p \geq 1,$$

där $x = (x_1, x_2, \dots, x_n) \in \mathbb{R}^n$. I denna kurs är det speciellt tre normer som är av intresse, nämligen *absolutnormen*

$$\|x\|_1 = |x_1| + |x_2| + \cdots + |x_n|,$$

den *Euklidiska normen*

$$\|x\|_2 = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$$

och *maximumnormen*

$$\|x\|_\infty = \max_{1 \leq i \leq n} |x_i|.$$

Det krävs ett bevis att de tre axiomen i definitionen av vektornorm är samtliga uppfyllda för en given funktion, tex ℓ^p -normen. \diamond

Exempel 2.12. Låt $x = (4, -5, 3)$. Beräkna $\|x\|_p$ för $p = 1, 2$ och ∞ . \diamond

Lösning. Vi får att

$$\begin{aligned} \|x\|_1 &= |4| + |-5| + |3| = 12, & \|x\|_2 &= \sqrt{4^2 + (-5)^2 + 3^2} = 5\sqrt{2} \\ \text{och } \|x\|_\infty &= \max(|4|, |-5|, |3|) = 5. \end{aligned}$$

Normen av en och samma vektor beror vilken vektornorm vi använder. \square

Vi vill tolka $\|x - y\|$ som *avståndet mellan x och y* . Med andra ord vill vi kunna tolka normen $\|x\| = \|x - 0\|$ som avståndet från nollvektorn till x , dvs vektorlängd. Men är det en rimlig tolkning? Överensstämmer definitionen med vad vi vanligtvis förknippar med ett avstånd? För alla normer gäller följande.

- (a) $\|x - y\| \geq 0$, med likhet om och endast om $x - y = \mathbf{0}$, dvs $x = y$

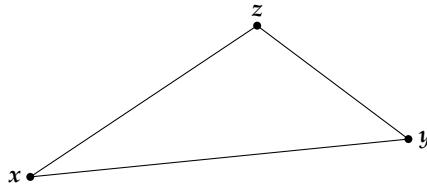
Tolkning: avståndet mellan två vektorer är alltid positivt och lika med noll endast då vi mäter från en vektor till samma vektor.

- (b) $\|x - y\| = \|(-1)(y - x)\| = |-1| \cdot \|y - x\| = \|y - x\|$

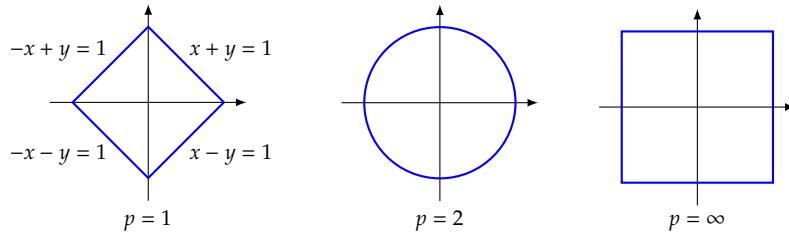
Tolkning: det spelar ingen roll om vi mäter från x till y eller tvärtom, avståndet är detsamma.

- (c) $\|x - y\| = \|x - z + z - y\| \leq \|x - z\| + \|z - y\|$

Tolkning: enligt triangelolikheten är $\|x - y\|$ det kortaste avståndet mellan x och y , för mäter vi först mellan x och en tredje vektor z och sedan mellan y och z så blir det längre, om nu inte z råkar "ligga mellan" x och y , se figur 2.8.



Figur 2.8

Figur 2.9. Enhetsfären i \mathbb{R}^2 med avseende på olika vektornormer.

Alltså överensstämmer $\|x - y\|$ med vad vi normalt sett uppfattar i begreppet avstånd.

Exempel 2.13. Mängden av alla vektorer x i planet \mathbb{R}^2 sådana att $\|x\| = 1$ kallas för *enhetsfären* (begreppet kan generaliseras till valfri dimension). Rita enhetsfären i \mathbb{R}^2 med avseende på absolutnormen, den Euklidiska normen och maximumnormen. ◇

Lösning. Låt $x = (x, y)$. Då är

$$\|x\|_1 = 1 \Leftrightarrow |x| + |y| = 1 \quad (2.4)$$

I tex andra kvadranten är $|x| = -x$ och $|y| = y$, eftersom $x \leq 0$ och $y \geq 0$. Det ger att (2.4) är ekvivalent med $-x + y = 1$, dvs en linje. På samma sätt finner man en linje i respektive kvadrant. Dessa fyra linjesegment bildar tillsammans den sökta enhetsfären, se vänstra bilden i figur 2.9. För den Euklidiska normen gäller att

$$\|x\|_2 = 1 \Leftrightarrow \sqrt{x^2 + y^2} = 1 \Leftrightarrow x^2 + y^2 = 1, \quad x, y \in \mathbb{R}.$$

Vilket vi känner igen som cirkelns ekvation, se andra bilden i figur 2.9. Slutligen har vi för maximumnormen att

$$\|x\|_\infty = 1 \Leftrightarrow \max(|x|, |y|) = 1$$

vilket är ekvivalent med att

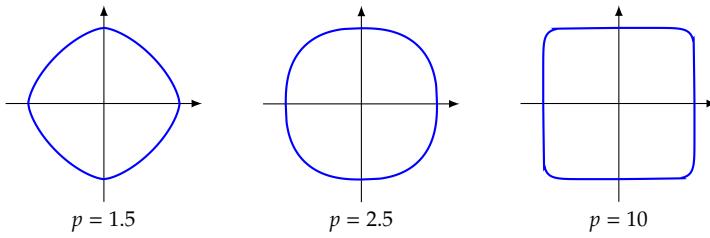
$$(|x| = 1 \text{ och } -1 \leq y \leq 1) \text{ eller } (|y| = 1 \text{ och } -1 \leq x \leq 1).$$

Vidare är $|x| = 1$ ekvivalent med $x = 1$ eller $x = -1$, dvs två vertikala linjer. Analogt för likheten $|y| = 1$, som är ekvivalent med $y = 1$ eller $y = -1$, dvs två horisontella linjer, se högra bilden i figur 2.9. Genom att ändra p kommer enhetsfären med avseende på ℓ^p -normen ändra form, se figur 2.10 ◇

2.6.2 Partiell derivata

Vi kommer att behöva derivera funktioner som beror på flera variabler. Den *partiella derivatan* av $f(x_1, x_2, \dots, x_n)$ med avseende på x_i betecknas

$$f'_{x_i}(x_1, x_2, \dots, x_n) = \frac{\partial f}{\partial x_i}, \quad i = 1, 2, \dots, n,$$



Figur 2.10. Enhetsfären för ytterligare några värden på p .

och bestäms genom att man betraktar (för en stund) de övriga variablerna som konstanter och deriverar f som om det är en funktion i en variabel. Resultatet är en funktion i lika många variabler som f . För en mer formell definition av partiell derivata hänvisas läsaren till en kurs i flervariabelanalys.

Exempel 2.14. Låt $f(x, y, z) = xy^3 + x \sin(yz) - x^2z$. Då är

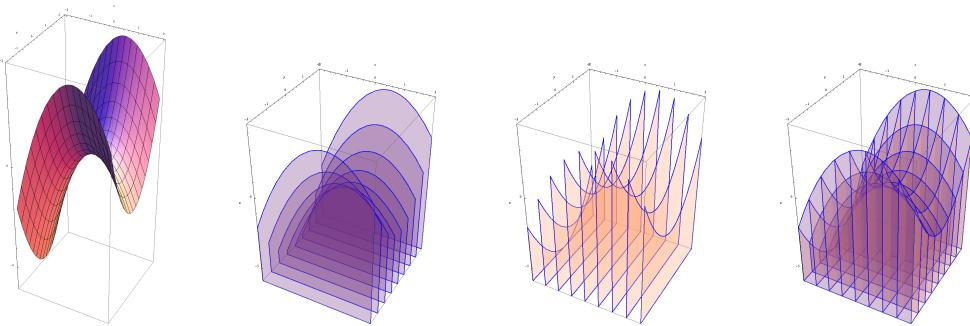
$$\begin{aligned} f'_x(x, y, z) &= \frac{\partial f}{\partial x} = y^3 + \sin(yz) - 2xz, \\ f'_y(x, y, z) &= \frac{\partial f}{\partial y} = 3xy^2 + xz \cos(yz) \end{aligned}$$

och

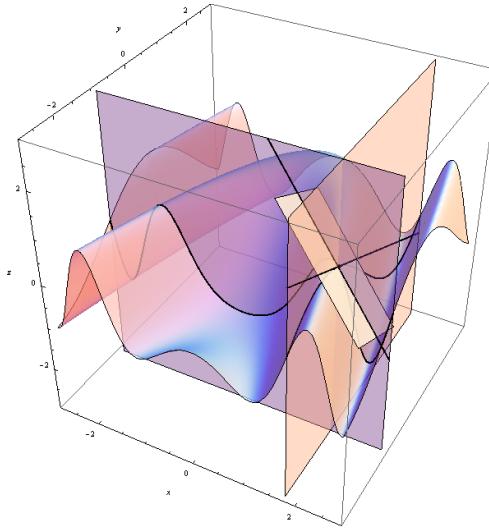
$$f'_z(x, y, z) = \frac{\partial f}{\partial z} = xy \cos(yz) - x^2.$$

Antal förstaderivator är lika med antal variabler. Vi kan givetvis även derivera en gång till. I detta exempel skulle det ge oss nio stycken andraderivator (vi kan derivera varje förstaderivata med avseende på var och en av de tre variabler). Några av andraderivatorna är lika, men det är en annan historia. ◇

Exempel 2.15. Låt $f(x, y) = x - x^2 + y^2$. De punkter $(x, y, z) \in \mathbb{R}^3$ för vilka $z = f(x, y)$ beskriver en yta, se första bilden i figur 2.11. Andra och tredje bilden i figur 2.11 illustrerar de "spant som håller upp ytan". I den andra bilden varierar vi x och håller y konstant. I den tredje bilden håller vi x konstant och varierar y . När vi deriverar f med avseende på en av variablerna och betraktar den andra som konstant, så bestämmer vi de funktioner som beskriver hur mycket spanten kant mot ytan förändras i respektive led. I den fjärde bilden är de två olika spanten tillsammans, jämför med näset på ytan i bilden ovan. ◇



Figur 2.11



Figur 2.12. $f(x, y) = \sin(x^2 + y)$

Exempel 2.16. Låt $f: \mathbb{R}^2 \rightarrow \mathbb{R}$. Då beskriver $f'_x(x, b)$ och $f'_y(a, y)$ hur funktionsytan förändras i x - respektive y -led, dvs förändringen av den kurva som ges av skärningen mellan funktionsytan $y = f(x, y)$ och de vertikala planen $y = b$ respektive $x = a$. I dessa plan beskriver $f'_x(x, b)$ och $f'_y(a, y)$ lutningen hos tangenten till motsvarande kurva. Dessa två linjer definierar ett plan, nämligen tangentplanet till funktionsytan i punkten $(a, b, f(a, b))$, se figur 2.12. \diamond

2.6.3 Newtons metod

Iterationsformeln i Newton-Raphsons metod för för en obekant ges som bekant av

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

Talföljden (x_n) konvergerar mot en lösning till ekvationen $f(x) = 0$. Vi vill kunna använda en liknande metod för att lösa ett ekvationssystem

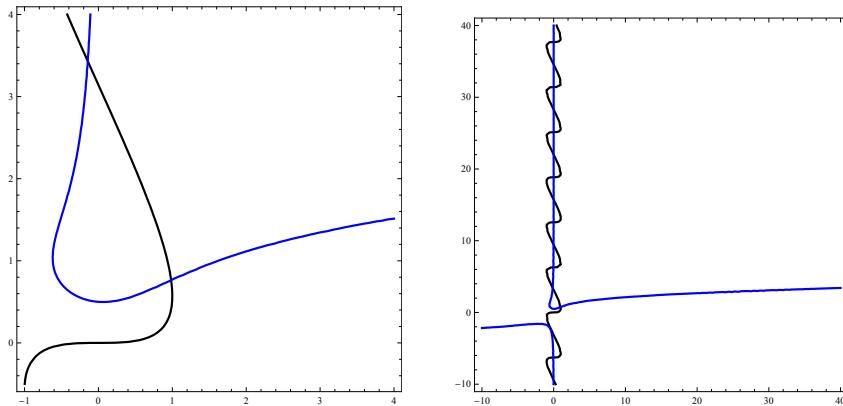
$$f(x) = 0,$$

där $f(x) = (f_1(x), f_2(x), \dots, f_n(x))$. Vi behöver en vettig tolkning av "1/ $f'(x)$ ". Eftersom funktionen f beror på n variabler och är vektorvärd, får vi en vektorvärd funktion vid derivering med avseende på respektive variabel,

$$\frac{\partial f}{\partial x_i} = \left(\frac{\partial f_1}{\partial x_i}, \frac{\partial f_2}{\partial x_i}, \dots, \frac{\partial f_n}{\partial x_i} \right), \quad i = 1, 2, \dots, n.$$

Då definieras Jacobis funktionalmatris av f som $n \times n$ -matrisen

$$J = J_f = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \cdots & \frac{\partial f_n}{\partial x_n} \end{pmatrix}.$$



Figur 2.13. Svart kurva är de punkter (x, y) sådana att $f_1(x, y) = 0$ och blå kurva motsvarar $f_2(x, y) = 0$, där f_1 och f_2 är funktionerna i exempel 2.17.

Med andra ord samtliga förstaderivator av f . Vi ska tolka denna matris som $f'(x)$. Notera att matriselementen i J är funktioner på formen $\mathbb{R}^n \rightarrow \mathbb{R}$. Vi kan nu tolka uttrycket " $1/f'(x)$ " som $J(x)^{-1}$, dvs bestäm Jacobis funktionalmatris, sätt in x i samtliga förstaderivator och bestäm sedan inversen till matrisen $J(x)$. Matriselementen i $J(x)$ är reella tal, för en given vektor x .

Algoritm 2.3 (Newtons metod). Låt $f: \mathbb{R}^n \rightarrow \mathbb{R}^n$ vara differentierbar, $x_0 \in \mathbb{R}^n$, $\varepsilon > 0$ och k_{\max} ett positivt heltal. Denna algoritm bestämmer en följd (x_n) som konvergerar mot en lösning till ekvationssystemet $f(x) = 0$, sådan att $\|x_k - x_{k-1}\| < \varepsilon$ eller avbryter om $k > k_{\max}$.

1. [Initiera] Sätt $k \leftarrow 0$.
2. [Iterera] Sätt $k \leftarrow k + 1$ och $x_k \leftarrow x_{k-1} - J(x_{k-1})^{-1}f(x_{k-1})$.
3. [Avbryt?] Om $\|x_k - x_{k-1}\| < \varepsilon$, avbryt och returnera x_k . Om $k > k_{\max}$, avbryt och returnera "Misslyckades!". Annars gå till steg 2. Metoden kan härledas med hjälp av Taylors formel av andra ordningen i flera variabler, se övningsuppgift 25.

Anmärkning. Det är inte nödvändigt att beräkna inversen $J(x_{k-1})^{-1}$ till matrisen $J(x_{k-1})$. Istället kan vi lösa det linjära ekvationssystemet

$$J(x_{k-1})y_{k-1} = -f(x_{k-1}),$$

och sedan i steg 2 sätta $x_k \leftarrow x_{k-1} + y_{k-1}$. I så fall är första stopkriteriet i steg 3 ekvivalent med $\|y_{k-1}\| < \varepsilon$.

Exempel 2.17. Låt

$$f_1(x, y) = \sin(x + y) - x \quad \text{och} \quad f_2(x, y) = x^2 - xy^3 - 2y + 1.$$

Studera ekvationssystemet

$$\begin{cases} f_1(x, y) = 0 \\ f_2(x, y) = 0 \end{cases} \Leftrightarrow \begin{cases} \sin(x + y) - x = 0 \\ x^2 - xy^3 - 2y + 1 = 0. \end{cases}$$

Punkterna (x, y) som satisfiera en av ekvationerna bildar en kurva i planet \mathbb{R}^2 , se figur 2.13. En lösning till ekvationssystemet motsvarar en skärningspunkt mellan

de två kurvorna. Vi ser att en lösning av de oändligt många lösningarna ligger nära punkten $(1, 1)$. En approximation av lösningen är $(0.9842, 0.7645)$. Sätt

$$f(x, y) = (f_1(x, y), f_2(x, y)).$$

Då är

$$J(x, y) = \begin{pmatrix} \frac{\partial f_1}{\partial x} & \frac{\partial f_1}{\partial y} \\ \frac{\partial f_2}{\partial x} & \frac{\partial f_2}{\partial y} \end{pmatrix} = \begin{pmatrix} \cos(x+y) - 1 & \cos(x+y) \\ 2x - y^3 & -3xy^2 - 2 \end{pmatrix}.$$

Som startvektor väljer vi $x_0 = (1, 1)$. Då är

$$f(x_0) = \begin{pmatrix} f_1(1, 1) \\ f_2(1, 1) \end{pmatrix} = \begin{pmatrix} -0.0907 \\ -1.0000 \end{pmatrix}$$

och

$$J(x_0) = \begin{pmatrix} -1.41615 & -0.41615 \\ 1.00000 & -5.00000 \end{pmatrix}.$$

Vektorn y_0 är lösningen till $J(x_0)y = -f(x_0)$, dvs

$$\begin{pmatrix} -1.41615 & -0.41615 \\ 1.00000 & -5.00000 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 0.0907 \\ 1.0000 \end{pmatrix}.$$

Vi får att $y_0 = (-0.00498421, -0.200997)$ och därmed är

$$x_1 = x_0 + y_0 = (0.995016, 0.799003).$$

Det ger i sin tur att

$$\|x_1 - x_0\|_2 = \|y_0\|_2 = \sqrt{(-0.00498421)^2 + (-0.200997)^2} \approx 0.201059.$$

Upprepas vi dessa steg får vi att

$$\begin{aligned} x_2 &= (0.984841, 0.765576) & \|y_1\|_2 &= 0.0349409 \\ x_3 &= (0.984219, 0.764479) & \|y_2\|_2 &= 0.00126181 \\ x_4 &= (0.984218, 0.764478) & \|y_3\|_2 &= 1.7145 \cdot 10^{-6} \\ x_5 &= (0.984218, 0.764478) & \|y_4\|_2 &= 3.12688 \cdot 10^{-12}. \end{aligned}$$

Redan efter fem iterationer har vi en god approximation av lösningen. \diamond

2.6.4 Fixpunktmetoder

Linjära ekvationssystem

20160413

Vi börjar med den enklaste typen av ekvationssystem, nämligen de linjära. Respektive metod kommer att presenteras med hjälp av exempel.

Exempel 2.18 (Jacobis metod). Studera det linjära ekvationssystemet

$$\begin{cases} 5x + y - 2z = 3 \\ -x + 12y + 3z = 5 \\ x - 5y + 3z = 1. \end{cases} \quad (2.5)$$

k	v_k	$\ v_{k-1} - v_k\ _1$	$\ v_{k-1} - v_k\ _2$
0	(1.0, 0.5, 0.7)		
1	(0.78, 0.325, 0.833333)	0.528333	0.311131
2	(0.868333, 0.273333, 0.615)	0.358333	0.241126
3	(0.791333, 0.335278, 0.499444)	0.2545	0.15205
4	(0.732722, 0.35775, 0.628352)	0.209991	0.143378
5	(0.779791, 0.320639, 0.685343)	0.14117	0.0827081
6	(0.810009, 0.310314, 0.607801)	0.118085	0.0838596
7	(0.781058, 0.332217, 0.58052)	0.0781367	0.0454119
8	(0.765764, 0.336625, 0.626676)	0.0658576	0.0488235
9	(0.783345, 0.323811, 0.63912)	0.0428387	0.0250626
10	(0.790886, 0.322165, 0.611904)	0.0364026	0.0282894
11	(0.780328, 0.329598, 0.606647)	0.0232464	0.0139403
12	(0.776739, 0.330032, 0.622554)	0.01993	0.0163122
13	(0.783015, 0.325757, 0.624474)	0.0124718	0.00783289
14	(0.784638, 0.325799, 0.615256)	0.0108843	0.00936004
15	(0.780942, 0.328239, 0.614786)	0.00660525	0.00445334

Tabell 2.1. Iteration med Jacobis metod.

Ekvationssystemet kan skrivas om till

$$\begin{cases} x = \frac{3-y+2z}{5} \\ y = \frac{5+x-3z}{12} \\ z = \frac{1-x+5y}{3} \end{cases} \quad \text{och rekursivt enligt} \quad \begin{cases} x_{k+1} = \frac{3-y_k+2z_k}{5} \\ y_{k+1} = \frac{5+x_k-3z_k}{12} \\ z_{k+1} = \frac{1-x_k+5y_k}{3}. \end{cases}$$

Jacobis metod är en fixpunktmetod och därmed iterativ — vi använder föregående värde på x , y och z för att beräkna nästa. Låt oss starta med

$$v_0 = (x_0, y_0, z_0) = (1.0, 0.5, 0.7).$$

Då är

$$\begin{cases} x_1 = \frac{3-0.5+2 \cdot 0.7}{5} = 0.78 \\ y_1 = \frac{5+1.0-3 \cdot 0.7}{12} = 0.325 \\ z_1 = \frac{1-1.0+5 \cdot 0.5}{3} \approx 0.833. \end{cases}$$

Alltså är $v_1 = (x_1, y_1, z_1) \approx (0.78, 0.325, 0.833)$, vilket ger oss att

$$\|v_0 - v_1\|_1 = |1.0 - 0.78| + |0.5 - 0.325| + |0.7 - 0.833| \approx 0.528,$$

och

$$\|v_0 - v_1\|_2 = \sqrt{(1.0 - 0.78)^2 + (0.5 - 0.325)^2 + (0.7 - 0.833)^2} \approx 0.311$$

Vi beräknar på samma sätt v_k tills avståndet mellan två konsekutiva vektorer är tillräckligt liten, se tabell 2.1. Jämför med den exakta lösningen som är

$$(x, y, z) = \frac{1}{55}(43, 18, 34) \approx (0.781818, 0.327273, 0.618182)$$

Vi ser att metoden ger oss en följd av vektorer som konvergerar mot lösningen, dock mycket långsmat i detta exempel. \diamond

Exempel 2.19 (Gauss-Seidels metod). Låt oss studera ekvationsystem (2.5) igen. Efter det att vi beräknat x_{k+1} skulle vi kunna använda det värdet istället för x_k när vi beräknar y_{k+1} och z_{k+1} . Av samma skäl kan vi använda y_{k+1} när vi beräknar z_{k+1} . Alltså kan vi skriva om det rekursiva systemet enligt

$$\begin{cases} x_{k+1} = \frac{3 - y_k + 2z_k}{5} \\ y_{k+1} = \frac{5 + x_{k+1} - 3z_k}{12} \\ z_{k+1} = \frac{1 - x_{k+1} + 5y_{k+1}}{3}. \end{cases}$$

Med $v_0 = (x_0, y_0, z_0) = (1.0, 0.5, 0.7)$ får vi efter en iteration att

$$\begin{cases} x_1 = \frac{3 - 0.5 + 2 \cdot 0.7}{5} = 0.78 \\ y_1 = \frac{5 + 0.78 - 3 \cdot 0.7}{12} \approx 0.3067 \\ z_1 = \frac{1 - 0.78 + 5 \cdot 0.3067}{3} \approx 0.5844. \end{cases}$$

Redan efter sex iterationer uppnår vi ett bättre resultat än med Jacobis metod, se tabell 2.2. Vanligtvis konvergerar denna metod snabbar mot en lösning än vad Jacobis metod gör. \diamond

Konvergens



En kvadratisk matris $A = (a_{i,j})$ av ordning n säges vara *strikt diagonaldominant* om

$$|a_{kk}| > |a_{k,1}| + \cdots + |a_{k,k-1}| + |a_{k,k+1}| + \cdots + |a_{k,n}|,$$

för alla $k = 1, 2, \dots, n$. Man kan visa att ett linjärt ekvationssystem vars koefficientmatris är strikt diagonaldominant har exakt en lösning och både Jacobis och Gauss-Seidels metod konvergerar mot lösningen oavsett val av startvektor (beviset ligger utanför ramen för kursen).

k	v_k	$\ v_{k-1} - v_k\ _1$	$\ v_{k-1} - v_k\ _2$
0	(1.0, 0.5, 0.7)		
1	(0.78, 0.306667, 0.584444)	0.528889	0.314851
2	(0.772444, 0.334926, 0.634062)	0.0854321	0.0575982
3	(0.78664, 0.323705, 0.610628)	0.0488505	0.029607
4	(0.77951, 0.328969, 0.621778)	0.0235442	0.0142434
5	(0.782917, 0.326465, 0.61647)	0.0112195	0.00678665
6	(0.781295, 0.327657, 0.618997)	0.00534201	0.00323134

Tabell 2.2. Iteration med Gauss-Seidels metod.

Exempel 2.20. Åven om en matris inte är strikt diagonaldominant kan metoderna ändå fungera. Vi har nämligen att koeficientmatrisen till ekvationssystemet (2.5), dvs

$$\mathbf{A} = \begin{pmatrix} 5 & 1 & -2 \\ -1 & 12 & 3 \\ 1 & -5 & 3 \end{pmatrix}$$

är inte strikt diagonaldominant eftersom

$$|3| < |1| + |-5|$$

på den tredje raden ($k = 3$). \diamond

Icke-linjära ekvationssystem

Skriv om ekvationssystemet $f(\mathbf{x}) = \mathbf{0}$, dvs (2.3), enligt

$$\begin{cases} f_1(x_1, x_2, \dots, x_n) = 0 \\ f_2(x_1, x_2, \dots, x_n) = 0 \\ \vdots \\ f_n(x_1, x_2, \dots, x_n) = 0 \end{cases} \Leftrightarrow \begin{cases} g_1(x_1, x_2, \dots, x_n) = x_1 \\ g_2(x_1, x_2, \dots, x_n) = x_2 \\ \vdots \\ g_n(x_1, x_2, \dots, x_n) = x_n. \end{cases}$$

Notera att det finns flera olika sätt att göra detta på. Låt $\mathbf{g}: \mathbb{R}^n \rightarrow \mathbb{R}^n$ enligt

$$\mathbf{g}(\mathbf{x}) = (g_1(\mathbf{x}), g_2(\mathbf{x}), \dots, g_n(\mathbf{x})).$$

Då är ekvationssystemet $f(\mathbf{x}) = \mathbf{0}$ ekvivalent med $\mathbf{g}(\mathbf{x}) = \mathbf{x}$. Med andra ord, är en sökt lösning \mathbf{x} till $f(\mathbf{x}) = \mathbf{0}$ en fixpunkt till \mathbf{g} . Vad är då lämpliga funktioner g_1, g_2, \dots, g_n att härleda från ett givet ekvationssystem $f(\mathbf{x}) = \mathbf{0}$ så att vi har en chans att lyckas iterera oss fram till en lösning? Antag att \mathbf{p} är en fixpunkt till \mathbf{g} , dvs $\mathbf{g}(\mathbf{p}) = \mathbf{p}$. Om

$$\left| \frac{\partial g_i}{\partial x_1}(\mathbf{p}) \right| + \left| \frac{\partial g_i}{\partial x_2}(\mathbf{p}) \right| + \cdots + \left| \frac{\partial g_i}{\partial x_n}(\mathbf{p}) \right| < 1$$

för alla $i = 1, 2, \dots, n$, så kommer iterationen

$$\mathbf{x}_{k+1} = \mathbf{g}(\mathbf{x}_k), \quad k = 0, 1, \dots$$

att konvergera mot \mathbf{p} , om \mathbf{x}_0 väljs tillräckligt nära \mathbf{p} .

Exempel 2.21. Låt $f_1, f_2: \mathbb{R}^2 \rightarrow \mathbb{R}$ funktionerna i exempel 2.17. För att bestämma g_1 och g_2 kan vi addera $\mathbf{x} = (x, y)$ till båda led i $f(\mathbf{x}) = \mathbf{0}$, dvs

$$\begin{cases} \sin(x+y) - x + x = x \\ x^2 - xy^3 - 2y + 1 + y = y \end{cases} \Leftrightarrow \begin{cases} \sin(x+y) = x \\ x^2 - xy^3 - y + 1 = y. \end{cases}$$

Alltså är $g_1(x, y) = \sin(x+y)$ och $g_2(x, y) = x^2 - xy^3 - y + 1$. Det ger att

$$\begin{aligned} \frac{\partial g_1}{\partial x} &= \cos(x+y) & \frac{\partial g_1}{\partial y} &= \cos(x+y) \\ \frac{\partial g_2}{\partial x} &= 2x - y^3 & \frac{\partial g_2}{\partial y} &= -3xy^2 - 1. \end{aligned}$$



Bevis?

k	\mathbf{x}_k	$\ \mathbf{x}_{k-1} - \mathbf{x}_k\ _2$
0	(1.000000, 1.000000)	
1	(0.909297, 0.666667)	0.345453
2	(0.999987, 0.741356)	0.117486
3	(0.985492, 0.777960)	0.0393693
4	(0.981499, 0.761716)	0.0167274
5	(0.985173, 0.763759)	0.00420341
6	(0.984176, 0.765136)	0.00169997
7	(0.984108, 0.764297)	0.000842016
8	(0.984269, 0.764466)	0.000233198
9	(0.984211, 0.764506)	0.0000710447
10	(0.984214, 0.764467)	0.0000395171

Tabell 2.3. Iteration med fixpunktmetoden.

Med $\mathbf{p} = (0.9842, 0.7645)$ är

$$\left| \frac{\partial g_1}{\partial x}(\mathbf{p}) \right| + \left| \frac{\partial g_1}{\partial y}(\mathbf{p}) \right| \approx 0.3539 < 1,$$

men

$$\left| \frac{\partial g_2}{\partial x}(\mathbf{p}) \right| + \left| \frac{\partial g_2}{\partial y}(\mathbf{p}) \right| \approx 4.2473 > 1.$$

Istället för att addera y till $f_2(x, y) = 0$ kan vi addera $3y$, vilket ger oss att

$$g_2(x, y) = \frac{x^2 - xy^3 + y + 1}{3}.$$

Därmed är

$$\left| \frac{\partial g_2}{\partial x}(\mathbf{p}) \right| + \left| \frac{\partial g_2}{\partial y}(\mathbf{p}) \right| \approx 0.7491 < 1.$$

Låt $\mathbf{x}_0 = (x_0, y_0) = (1, 1)$. Första iterationen ger oss

$$\begin{cases} x_1 = g_1(x_0, y_0) = \sin(1+1) \approx 0.909297 \\ y_1 = g_2(x_0, y_0) = \frac{1^2 - 1 \cdot 1^3 + 1 + 1}{3} \approx 0.666667 \end{cases}$$

Alltså är $\mathbf{x}_1 = (0.909297, 0.666667)$ och

$$\|\mathbf{x}_0 - \mathbf{x}_1\|_2 = \sqrt{(1 - 0.909297)^2 + (1 - 0.666667)^2} \approx 0.3455.$$

Efter tio iterationer har vi ett resultat nära den sökta lösningen, se tabell 2.3. \diamond

Exempel 2.22 (Seidels metod). Denna metod är ett försök till en förbättring av fixpunktmetoden från föregående exempel på samma sätt som Gauss-Seidels metod för linjära ekvationssystem med avseende på Jacobis metod. Iterationsformeln för ekvationssystemet i exempel 2.21 kanske skrivas om enligt

$$\begin{cases} x_{k+1} = g_1(x_k, y_k) = \sin(x_k + y_k) \\ y_{k+1} = g_2(x_{k+1}, y_k) = \frac{x_{k+1}^2 - x_{k+1}y_k^3 + y_k + 1}{3}, \end{cases}$$

för icke-negativa heltal k . Detaljerna lämnas som övning, se övningsuppgift 32. \diamond

2.6.5 Metodoberoende feluppskattning för ekvationssystem

Låt $f(x) = (f_1(x), f_2(x), \dots, f_n(x))$ och låt \hat{x} vara en approximation av lösningen α till ekvationen $f(x) = 0$. Sätt $\delta\alpha = \hat{x} - \alpha = (\delta\alpha_1, \delta\alpha_2, \dots, \delta\alpha_n)$. Genom att Taylorutveckla varje f_i kring α , se övningsuppgif 25, får vi att

$$f_i(\hat{x}) = f_i(\alpha + \delta\alpha) = f_i(\alpha) + \sum_{j=1}^n \frac{\partial f_i}{\partial x_j}(\alpha) \delta\alpha_j + O(\|\delta\alpha\|^2)$$

för $i = 1, 2, \dots, n$, vilket kan skrivas

$$f(\hat{x}) = J(\alpha)\delta\alpha + O(\|\delta\alpha\|^2) \approx J(\hat{x})\delta\alpha$$

eftersom $f(\alpha) = 0$. Löser vi ut $\delta\alpha$ får vi att $\delta\alpha \approx J(\hat{x})^{-1}f(\hat{x})$ och därmed

$$\|\delta\alpha\| \approx \|J(\hat{x})^{-1}f(\hat{x})\|$$

för valfri vektornorm.

Exempel 2.23. I exempel 2.17 fann vi $x_5 = (0.984218, 0.764478)$. Då är

$$f(x_5) \approx (-4.56928 \cdot 10^{-7}, -1.11405 \cdot 10^{-6})$$

samt

$$J(x_5) \approx \begin{pmatrix} -1.17696 & -0.176963 \\ 1.52165 & -3.72561 \end{pmatrix} \quad \text{och} \quad J(x_5)^{-1} \approx \begin{pmatrix} -0.800487 & 0.0380223 \\ -0.326944 & -0.252883 \end{pmatrix}$$

Det ger att $J(x_5)^{-1}f(x_5) \approx (3.23406 \cdot 10^{-7}, 4.31114 \cdot 10^{-7})$ och

$$\|J(x_5)^{-1}f(x_5)\|_2 \approx 5.38935 \cdot 10^{-7}.$$

Jämför med $\|y_4\|_2 = 3.12688 \cdot 10^{-12}$ i exempel 2.17. ◊

2.7 Övningsuppgifter

- L 1. Låt $g(x) = x^2 + x - 4$. Kan man använda fixpunktsiteration för att finna lösningarna till ekvationen $g(x) = x$? Motivera ditt svar.
- L 2. Låt $g(x) = x - 0.0001x^2$ och $p_0 = 1$. Sätt $p_n = g(x_{n-1})$, där $n \geq 1$.
 - (a) Visa att $p_0 > p_1 > p_2 > \dots$.
 - (b) Visa att $p_n > 0$ för alla n .
 - (c) Eftersom följen $(p_n)_{n=0}^\infty$ är avtagande och nedåt begränsad har den ett gränsvärde. Bestäm detta gränsvärde.
- L 3. Låt $g(x) = 0.5x + 1.5$ och $p_0 = 4$. Sätt $p_{n+1} = g(p_n)$ för $n \in \mathbb{N}$.
 - (a) Visa att $p = 3$ är en fixpunkt till g .
 - (b) Visa att $|p - p_n| = |p - p_{n-1}|/2$ för alla $n \in \mathbb{N}$.
 - (c) Visa att $|p - p_n| = |p - p_0|/2^n$ för alla $n \in \mathbb{N}$.
- 4. Låt f vara en reellvärd funktion. Förklara varför de två påståenderna

" $f(a)$ och $f(b)$ har olika tecken" och " $f(a)f(b) < 0$ "

är ekvivalenta.

L 5. Låt

$$f(x) = \frac{2}{3}x^2 - 2x - \frac{2}{3}.$$

och studera ekvationen $f(x) = 0$.

- (a) Bestäm den funktion $g(x)$ som hör till fixpunktmetoden vid lösning av ekvationen $f(x)$.
- (b) Låt $x_0 = 1$ och bestäm de fyra elementen x_1, x_2, x_3 och x_4 i talföljden som ges av fixpunktmetoden med x_0 som initialvärde.
- (c) Bestäm fixpunkterna till $g(x)$.
- (d) Visa att om man försöker lösa ekvationen $f(x) = 0$ med fixpunktmetoden kommer man att misslyckas. (20140603)

6. Låt $f(x) = 0.08x^3 - 0.16x^2 - 0.4x + 0.48$ och $g(x) = f(x) + x$.

- (a) Visa att $-2, 1$ och 3 är fixpunkter till g .
- (b) För vilken eller vilka av fixpunkterna till g konvergerar fixpunktmetoden? Motivering krävs. (20150108)

L 7. Låt $f: \mathbb{R} \rightarrow \mathbb{R}$ och sätt $g(x) = f(x) + x$. Antag att

$$g(x_1) = x_2, g(x_2) = x_3, \dots, g(x_{n-1}) = x_n, g(x_n) = x_1.$$

där $x_1, x_2, \dots, x_n \in \mathbb{R}$. Visa att

$$f(x_1) + f(x_2) + \dots + f(x_n) = 0. \quad (20150822)$$

L 8. Vad händer om man använder intervallhalveringsmetoden på funktionen

$$f(x) = \frac{1}{x-2}$$

över följande intervall?

- (a) $[3, 7]$
- (b) $[1, 7]$

L 9. Antag att intervallhalveringsmetoden används för att bestämma ett nollställe till funktionen $f(x)$ i intervallet $[2, 7]$. Hur många intervallhalveringar måste vi göra för att uppnå en noggrannhet på $5 \cdot 10^{-9}$?

10. Låt f vara en reellvärd kontinuerlig funktion definierad på intervallet $[a, b]$, sådan att $f(a)f(b) < 0$. Sätt $c = (a+b)/2$ samt $I_1 = [a, c]$ och $I_2 = [c, b]$. Antag att $f(x) = 0$ för exakt tre olika $x \in (a, b)$ och att $f(c) \neq 0$. Visa att man med intervallhalveringsmetoden alltid finner det nollställe till f som tillhör ett av delintervallen I_1 och I_2 och de två andra tillhör det andra delintervallet. (20120821)

L 11. Låt $f(x) = x^3 - 2x^2 - 11x + 12$.

- (a) Visa att $-3, 1$ och 4 är lösningar till ekvationen $f(x) = 0$.
- (b) Vilken lösning finner man med hjälp av intervallhalveringsmetoden om man startar med intervallet $[-4, 7]$? Motivering krävs.

- (c) Varför är intervallet $[-4, 3]$ ett dåligt val att utgå från? (20140822)
- 12.** Man kan lätt lura sig själv om man bara titta på en bild. I härledningen av formeln för Newtons metod utgick vi från figur 2.7 och där ser vi tydligt att x_{n+1} är närmare till a än vad x_n är. Är det alltid så?
- 13.** Låt $f(x) = x^2 - x + 1$, se exempel 2.6. Genererar de 100 första talen i talföljden (x_m) för nedanstående startvärdet. Ser du något mönster?
- (a) $x_0 = 0.1$ (b) $x_0 = 0.9$ (c) $x_0 = 1.9$
- L 14.** Låt $f(x) = (x - 2)^4$.
- (a) Härled Newtons formel $x_{k+1} = g(x_k)$ för f .
 (b) Sätt $x_0 = 2.1$ och bestäm x_1, x_2, x_3 och x_4 .
 (c) Konvergerar talföljden kvadratiskt eller linjärt?
- L 15.** Låt $f(x) = \cos(x)$.
- (a) Härled Newtons formel $x_{k+1} = g(x_k)$ för f .
 (b) Kan vi använda $x_0 = 3$ för att finna nollstället $x = 3\pi/2$?
 (c) Kan vi använda $x_0 = 5$ för att finna nollstället $x = 3\pi/2$?
- L 16.** Bestäm de fem första elementen i den talföljd (x_0, x_1, x_2, \dots) som man erhåller då man löser ekvationen $e^{-x} = x$ med hjälp av Newton-Raphsons metod och startvärdet $x_0 = 0.25$. (20120603)
- 17.** Approximera den reella lösningen till ekvationen $x = e^{-x^2}$ med hjälp av Newtons metod och fem iterationer. (20130109)
- L 18.** Studera ekvationen
- $$xe^x = 1.$$
- Bestäm först Newton-Raphsons iterationsformel $x_{n+1} = g(x_n)$ med vilken man kan finna en approximativ lösning till ekvationen ovan. Låt $x_0 = 1$. Bestäm sedan talen x_1, x_2, x_3, x_4 och x_5 med hjälp av g . (20130607)
- 19.** Bestäm ett nollställe till $f(x) = x^3 - 3x^2 - 7x + 21$ med Newtons metod. Som initiativvärde ska du använda $x_0 = 2.5$ och avbryta först när du uppnår en noggrannhet om minst $\varepsilon = 10^{-5}$. (20130823)
- 20.** Finn en approximation av den positiva lösningen till ekvationen
- $$x^2 = \cos x$$
- med hjälp av Newtons metod. Använd $x_0 = 0.75$ som startvärde och bestäm de fyra första talen i talföljden (x_1, x_2, \dots) . (20140110)
- 21.** I definitionen av ℓ^p -norm, se exempel 2.11, gäller att $p \geq 1$. Men vad händer om man tillåter att $0 < p < 1$? Rita "enhetssfären" då
- (a) $p = 0.7$ (b) $p = 0.5$ (c) $p = 0.3$.
- 22.** Visa att om $0 < p < 1$ och $n > 1$, så uppfyller inte "normen" i exempel 2.11 triangololikheten och är därmed inte en vektornorm på \mathbb{R}^n .

- L 23. Låt n vara ett positivt heltal och $\|\cdot\|_1: \mathbb{R}^n \rightarrow \mathbb{R}$ enligt

$$\|\mathbf{x}\|_1 = \sum_{k=1}^n |x_k| = |x_1| + |x_2| + \cdots + |x_n|,$$

där $\mathbf{x} = (x_1, x_2, \dots, x_n) \in \mathbb{R}^n$. Visa att $\|\cdot\|_1$ uppfyller

- (a) $\|\mathbf{x}\|_1 \geq 0$
- (b) $\|\mathbf{x}\|_1 = 0$ om och endast om $\mathbf{x} = \mathbf{0}$
- (c) $\|c\mathbf{x}\|_1 = |c| \cdot \|\mathbf{x}\|_1$

för alla $\mathbf{x} \in \mathbb{R}^n$ och alla reella tal c .

24. Bestäm samtliga partiella förstaderivator av följande funktioner.

(a) $f(x, y) = xy^2 + e^{xy}$	(b) $f(x, y) = xy^2 - ye^{xy}$
(c) $f(x, y) = x^{\sin xy}$	(d) $g(x, y, z) = \cos\left(xz \ln \frac{x}{y}\right)$

25. (Taylors formel av andra ordningen i flera variabler) Låt $f: \mathbb{R}^n \rightarrow \mathbb{R}$ och $\mathbf{a}, \mathbf{h} \in \mathbb{R}^n$.

Visa att

$$f(\mathbf{a} + \mathbf{h}) = f(\mathbf{a}) + \sum_{i=1}^n f'_i(\mathbf{a})h_i + \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n f''_{i,j}(\mathbf{a})h_i h_j + O(|\mathbf{h}|^3),$$

där $\mathbf{h} = (h_1, h_2, \dots, h_n)$ samt där f'_i betecknar den partiella derivatan av f med avseende på den i :te variabeln och där $f''_{i,j}$ betecknar den partiella andraderivatan av f med avseende på först den i :te och sedan den j :te variabeln.

26. Bestäm iterationsformeln

$$\mathbf{x}_k = \mathbf{x}_{k-1} - J(\mathbf{x}_{k-1})f(\mathbf{x}_{k-1})$$

för Newtons metod, då $f(\mathbf{x}) = \mathbf{0}$ är ett linjärt ekvationssystem.

- L 27. Låt f_1 och f_2 vara funktioner $\mathbb{R}^2 \rightarrow \mathbb{R}$ vars samtliga partiella derivator är kontinuerliga. Visa att Newtons metod för ekvationssystemet

$$\begin{cases} f_1(x, y) = 0 \\ f_2(x, y) = 0 \end{cases} \quad \text{kan skrivas} \quad \begin{cases} x = g_1(x, y) \\ y = g_2(x, y), \end{cases}$$

dvs som en fixpunktmetod, där

$$g_1(x, y) = x - \frac{f_1(x, y) \frac{\partial f_2}{\partial y}(x, y) - f_2(x, y) \frac{\partial f_1}{\partial y}(x, y)}{\det(J(x, y))}$$

$$g_2(x, y) = y - \frac{f_2(x, y) \frac{\partial f_1}{\partial x}(x, y) - f_1(x, y) \frac{\partial f_2}{\partial x}(x, y)}{\det(J(x, y))}.$$

28. För vilka reella tal a och b är matrisen

$$\begin{pmatrix} a & 0 & -2 \\ 3 & 7 & b \\ a & -1 & b \end{pmatrix}$$

strikt diagonaldominant?

(20140822)

29. Studera ekvationssystemet

$$\begin{cases} 2x - y = 4 \\ 3x + 5y = 1. \end{cases}$$

Låt $p_0 = \mathbf{0}$ och bestäm p_k för $k = 1, 2, 3$ med Gauss-Seidels metod. Konvergerar metoden? Motivera ditt svar. (20120821)

L 30. Studera ekvationssystemet

$$\begin{cases} -x + 3y = 1 \\ 6x - 2y = 2. \end{cases}$$

Låt $p_0 = \mathbf{0}$ och bestäm p_k för $k = 1, 2, 3$ med

- (a) Jacobis metod (b) Gauss-Seidels metod.

Avgör om iterationen konvergerar eller divergerar i respektive deluppgift.

31. Använd Gauss-Seidels metod för att finna en lösning till ekvationssystemet

$$\begin{cases} 2x + y = 1 \\ -x + 3y = 2. \end{cases}$$

Redovisa de tre första iterationerna, v_1 , v_2 och v_3 , med $v_0 = (0.5, 0.5)$ som startvektorn. Bestäm också det absoluta felet för v_3 . (20130823)

L 32. Slutför exempel 2.22 genom att iterera tio gånger med $x_0 = (1, 1)$ som startvektor. Konvergensen kommer inte att vara bättre än fixpunktmetoden.

L 33. Bestäm analytiskt fixpunkterna till $g(x) = (g_1(x), g_2(x), \dots, g_n(x))$. Med andra ord lös ekvationen $x = g(x)$ för hand.

(a) $\begin{cases} g_1(x, y) = x - y^2 \\ g_2(x, y) = -x + 6y \end{cases}$ (b) $\begin{cases} g_1(x, y) = (x^2 - y^2 - x - 3)/3 \\ g_2(x, y) = (-x + y - 1)/3 \end{cases}$

(c) $\begin{cases} g_1(x, y) = \sin(y) \\ g_2(x, y) = -6x + y \end{cases}$ (d) $\begin{cases} g_1(x, y, z) = 9 - 3y - 2z \\ g_2(x, y, z) = 2 - x + z \\ g_3(x, y, z) = -9 + 3x + 4y - z \end{cases}$

L 34. Låt $x_0 = (x_0, y_0) = (-0.3, -1.3)$ och

$$\begin{cases} g_1(x, y) = \frac{y - x^3 + 3x^2 + 3x}{7} \\ g_2(x, y) = \frac{y^2 + 2y - x - 2}{2}. \end{cases}$$

Bestäm de tre första punkterna som genereras med x_0 som startpunkt och med var och en av följande metoder för att finna en approximation av lösningen till ekvationen $x = g(x)$, där $g(x) = g(x, y) = (g_1(x, y), g_2(x, y))$.

- (a) Fixpunktmetoden (b) Seidels metod

L 35. Approximera en lösning till ekvationssystemet

$$\begin{cases} 3xy = 1 \\ \sin(xy) = x \end{cases}$$

med Seidels metod. Redovisa de fyra första iterationerna där du som startvektor använder $x_0 = (0.7, 0.5)$. (20130109)

Kapitel 3

Interpolation

3.1 Funktionsapproximation

3.1.1 Serier

En serie

$$S = \sum_{n=1}^{\infty} a_n = a_1 + a_2 + a_3 + a_4 + \dots$$

säges *konvergera* om det finns ett gränsvärde till de partiella summorna

$$S_N = \sum_{n=1}^N a_n = a_1 + a_2 + \dots + a_N$$

dvs om

$$\lim_{N \rightarrow \infty} S_N < \infty$$

och i så fall definieras *summan av serien* som detta gränsvärde. Om S konvergerar, så gäller att $a_n \rightarrow 0$, då $n \rightarrow \infty$. Omvändningen gäller dock inte, tex har vi att $1/n \rightarrow 0$ då $n \rightarrow \infty$, men

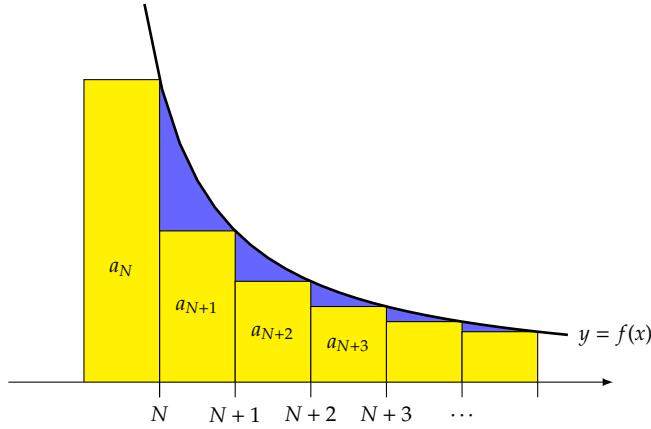
$$\sum_{n=1}^{\infty} \frac{1}{n} = \infty.$$

Trots det växer de partiella summorna S_N mycket långsamt, tex är

$$S_{100\,000} = \sum_{k=0}^{10^5} \frac{1}{k} \approx 12.0901 \quad \text{och} \quad S_{1\,000\,000} = \sum_{k=0}^{10^6} \frac{1}{k} \approx 14.3927.$$

Att enbart förlita sig på numeriska beräkningar kan alltså vara vansktigt. Vid numeriska beräkningar approximerar vi en (förhoppningsvis) konvergent serie genom att beräkna en av de partiella summorna samt uppskattar felet, nämligen den sk
resttermen $R_N = S - S_N$. Antag att $f: \mathbb{R} \rightarrow \mathbb{R}$ är positiv och avtagande. Om $a_n = f(n)$ är termerna i en konvergent serie, så är

$$R_N = S - S_N = \sum_{n=1}^{\infty} a_n - \sum_{n=1}^N a_n = \sum_{n=N+1}^{\infty} a_n \leq \int_N^{\infty} f(x) dx.$$



Figur 3.1

I figur 3.1 motsvarar arean av staplarna a_{N+1}, a_{N+2}, \dots resttermen R_N . Arean under kurvan $y = f(x)$ då $x \geq N$, dvs integralen ovan, är större än R_N . Notera att denna uppskattning av felet är i många fall grov.

Exempel 3.1. Låt

$$S = \sum_{n=1}^{\infty} \frac{1}{n^2} \quad \text{och} \quad S_N = \sum_{n=1}^N \frac{1}{n^2}.$$

Vi har att

$$R_N = S - S_N = \sum_{n=N+1}^{\infty} \frac{1}{n^2} \leq \int_N^{\infty} \frac{dx}{x^2} = \left[-\frac{1}{x} \right]_N^{\infty} = \frac{1}{N}$$

Vi vet att $S = \pi^2/6 \approx 1.6449340668$. Med 11 signifikanta siffror och olika värden på N får vi resultatet i tabell 3.1. \diamond

N	S_N	R_N	$1/N$
10	1.5497677312	0.0951663357	0.1
100	1.6349839002	0.0099501667	0.01
1000	1.6439345667	0.0009995002	0.001
10 000	1.6448340718	0.0000999950	0.0001
100 000	1.6449240669	0.0000100000	0.00001
1 000 000	1.6449330668	0.0000010000	0.000001

Tabell 3.1

3.1.2 Taylorutveckling

Låt $f: \mathbb{R} \rightarrow \mathbb{R}$ vara en funktion vars $n + 1$ första derivatore är kontinuerliga i en omgivning I till a . Då gäller Taylors formel, dvs

$$\begin{aligned} f(x) &= f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f^{(3)}(a)}{3!}(x-a)^3 \\ &\quad + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n + \frac{f^{(n+1)}(\xi)}{(n+1)!}(x-a)^{n+1} \end{aligned}$$

för alla $x \in I$, där ξ ligger mellan a och x . Notera att ξ beror på x och n , dvs $\xi = \xi(x, n)$. Den n :te Taylorpolynomet p_n till f i punkten a ges av

$$p_n(x) = f(a) + \frac{f'(a)}{1!}(x - a) + \cdots + \frac{f^{(n)}(a)}{n!}(x - a)^n,$$

och motsvarande felterm

$$E_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}(x - a)^{n+1}$$

kallas *Lagranges rest*. Alltså är

$$f(x) = p_n(x) + E_n(x).$$

Vi kan använda p_n för att approximera f och uppskatta felet med hjälp av E_n .

Exempel 3.2. Låt $f(x) = x \cos x$ och $a = 0$. För att bestämma Taylorutvecklingen av funktionen f kring $a = 0$, kan vi först bestämma Taylorutvecklingen av $g(x) = \cos x$ och sedan utnyttja att $f(x) = xg(x)$. Vi får att

$$g'(x) = -\sin x, \quad g''(x) = -\cos x, \quad g^{(3)}(x) = \sin x \quad \text{och} \quad g^{(4)}(x) = \cos x,$$

där mönstret ovan upprepar sig för högre derivator. Från

$$g(0) = 1, \quad g'(0) = 0, \quad g''(0) = -1, \quad g^{(3)}(0) = 0 \quad \text{och} \quad g^{(4)}(0) = 1$$

följer att

$$\begin{aligned} \cos x &= g(x) = 1 - \frac{1}{2!}(x - 0)^2 + \frac{1}{4!}(x - 0)^4 - \frac{1}{6!}(x - 0)^6 + \cdots + \frac{g^{(2n+1)}(\xi)}{(2n+1)!}(x - 0)^{2n+1} \\ &= 1 - \frac{x^2}{2} + \frac{x^4}{24} - \frac{x^6}{720} + \cdots + (-1)^n \frac{x^{2n}}{(2n)!} + (-1)^{n+1} \frac{x^{2n+1}}{(2n+1)!} \sin \xi. \end{aligned}$$

Alltså är

$$f(x) = x \cos x = x - \frac{x^3}{2} + \frac{x^5}{24} - \frac{x^7}{720} + \cdots + (-1)^n \frac{x^{2n+1}}{(2n)!} + (-1)^{n+1} \frac{x^{2n+2}}{(2n+1)!} \sin \xi. \quad (3.1)$$

Notera att om vi istället direkt Taylorutvecklar f måste vi derivera f . Det ger samma resultat, förutom Lagranges rest, som då ges av

$$\frac{f^{(m)}(\xi)}{m!}(x - 0)^m = (-1)^{m/2} \frac{x^m}{m!} (\xi \cos \xi + m \sin \xi), \quad (3.2)$$

där m är ett jämnt heltal. Det lämnas som övning att bevisa (3.2), se övningsuppgift 2. Notera att det rör sig om olika ξ i (3.1) respektive (3.2). Låt $n = 2k - 1$, där k är ett positivt heltal. Då är n ett udda heltal. Sätt

$$p_n(x) = x - \frac{x^3}{2} + \frac{x^5}{24} \pm \cdots + (-1)^{k+1} \frac{x^n}{(n-1)!},$$

där vi utnyttjar att $n! = n \cdot (n-1)!$ och $f^{(n)}(0) = (-1)^{k+1}n$, se övningsuppgift 2. Utgår vi från (3.2) ges Lagranges rest av

$$E_n(x) = (-1)^k \frac{x^{n+1}}{(n+1)!} (\xi \cos \xi + (n+1) \sin \xi),$$

där $m = n + 1 = 2k$ och därmed är $m/2 = k$.

Låt $\alpha = 0.48$ och $\beta = 2.17$. Vi vill approximera $f(\alpha) \approx 0.425758$ och $f(\beta) \approx -1.22385$ med hjälp av Taylorutvecklingen av f som vi bestämde ovan. Vi får att

$$\begin{aligned} p_1(x) &= x, \\ p_3(x) &= x - \frac{x^3}{2}, \\ p_5(x) &= x - \frac{x^3}{2} + \frac{x^5}{24} \\ p_7(x) &= x - \frac{x^3}{2} + \frac{x^5}{24} - \frac{x^7}{720} \\ p_9(x) &= x - \frac{x^3}{2} + \frac{x^5}{24} - \frac{x^7}{720} + \frac{x^9}{40320} \end{aligned}$$

och specifikt är

$$\begin{array}{ll} p_1(0.48) = 0.48 & p_1(2.17) = 2.17 \\ p_3(0.48) = 0.424704 & p_3(2.17) = -2.93916 \\ p_5(0.48) = 0.425766 & p_5(2.17) = -0.934281 \\ p_7(0.48) = 0.425758 & p_7(2.17) = -1.24897 \\ p_9(0.48) = 0.425758 & p_9(2.17) = -1.22251. \end{array}$$

Om vi Taylorutvecklar f kring $a = 0$ så kommer approximationen av $f(x)$ troligtvis bli bättre ju närmare x är a . Vi ser i vårt exempel att redan för $n = 3$ ger motsvarande Taylorpolynom en bra approximation av $f(\alpha)$, medan vi måste använda ett polynom av större grad för att få en bra approximation av $f(\beta)$, se även figur 3.2. Vi har tex att

$$f(0.48) = p_5(0.48) + E_5(0.48) = 0.425766 + E_5(0.48)$$

och

$$f(2.17) = p_9(2.17) + E_9(2.17) = -1.2225 + E_9(2.17).$$

En grov uppskattning av Lagranges rest ges av

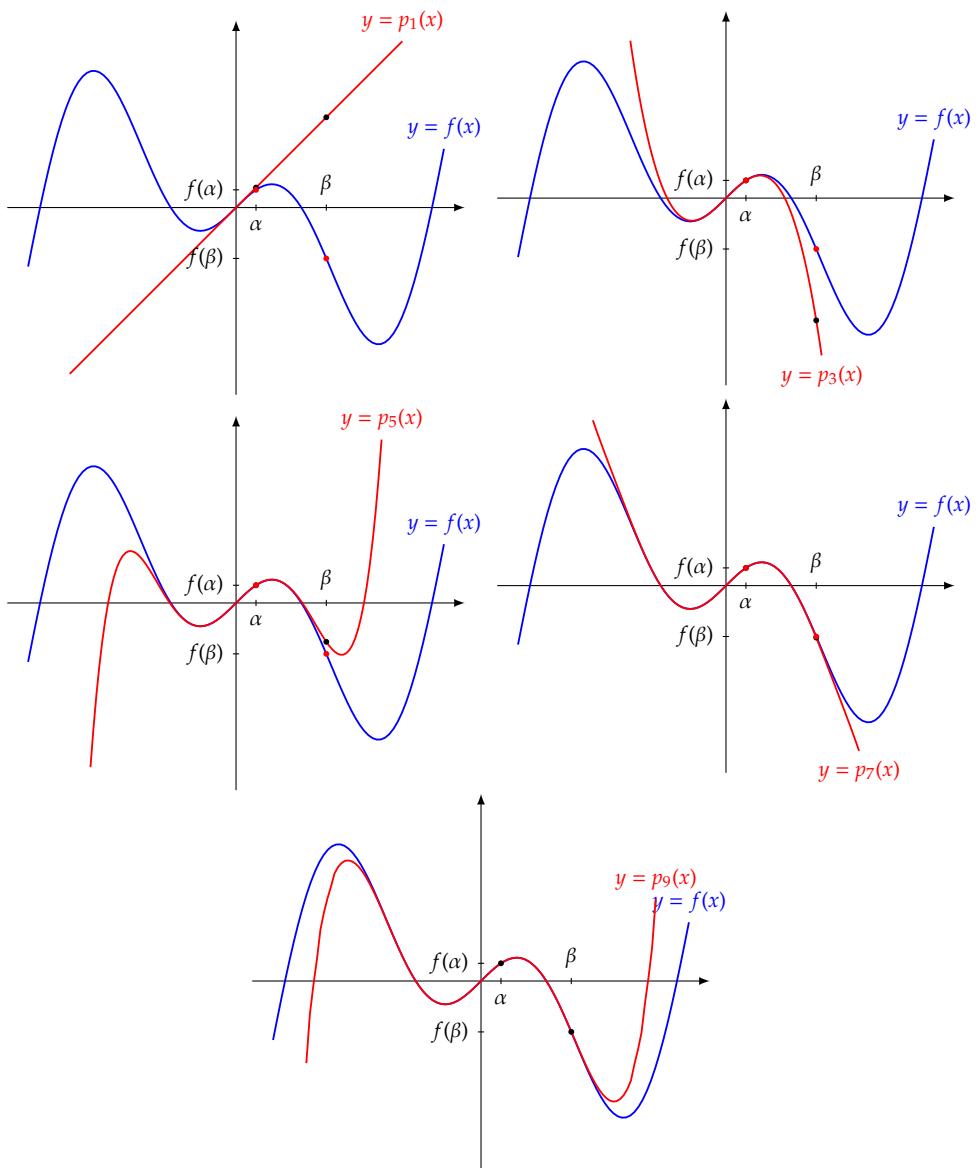
$$\begin{aligned} |E_n(x)| &= \left| (-1)^k \frac{x^{n+1}}{(n+1)!} (\xi \cos \xi + (n+1) \sin \xi) \right| \\ &= |-1|^k \frac{|x^{n+1}|}{|(n+1)!|} |\xi \cos \xi + (n+1) \sin \xi| \\ &\leq \frac{|x|^{n+1}}{(n+1)!} (|\xi| \cdot |\cos \xi| + |n+1| \cdot |\sin \xi|) \\ &\leq \frac{|x|^{n+1}}{(n+1)!} (|\xi| + n+1), \end{aligned}$$

där den första olikheten följer från triangelolikheten och den andra olikheten följer från det faktum att $|\cos \xi| \leq 1$ och $|\sin \xi| \leq 1$. Notera att ξ är ett element mellan $a = 0$ och x . Alltså är

$$|E_5(0.48)| \leq \frac{0.48^6}{6!} (0.48 + 6) \approx 0.00011$$

och

$$|E_9(2.17)| \leq \frac{2.17^{10}}{10!} (2.17 + 10) \approx 0.00776.$$



Figur 3.2. Illustration av hur approximationen av f med p_n förbättras då vi ökar n .

Resultatet är här bättre om x är nära det a som vi Taylorutvecklade f kring. Genom att skriva om uttrycket $f(x)$ kan vi komma närmare $a = 0$. Låt y vara det reella tal sådant att $x = y + \pi/2$, tex ger $2.17 = y + \pi/2$ att $y \approx 0.599204$. Från den trigonometriska likheten $\cos(x + \pi/2) = -\sin x$ följer att

$$f(x) = x \cos x = x \cos\left(y + \frac{\pi}{2}\right) = -x \sin y.$$

Alltså är $f(2.17) \approx -2.17 \sin(0.599204)$ och med ett Taylorpolynom av låg grad till sinus kring $a = 0$ uppnår vi förhoppningsvis en god approximation av $f(2.17)$. \diamond

3.1.3 Horners metod för derivata

Låt b_n, \dots, b_1, b_0 vara de tal som bestäms i tur ordning vid beräkning av $p(a)$ i algoritmen 1.2, dvs

$$b_n = c_n \quad \text{och} \quad b_k = b_{k+1}a + c_k,$$

där $k = n-1, \dots, 1, 0$ och $p(x) = c_n x^n + \dots + c_1 x + c_0$. Enligt divisionsalgoritmen för polynom existerar det ett polynom $q(x)$ och en konstant r sådana att

$$p(x) = (x - a)q(x) + r.$$

För att bestämma q och r studerar vi produkten

$$\begin{aligned} & (x - a)(b_n x^{n-1} + b_{n-1} x^{n-2} + \dots + b_2 x + b_1) \\ &= b_n x^n + b_{n-1} x^{n-1} + \dots + b_2 x^2 + b_1 x \\ &\quad - ab_n x^{n-1} - ab_{n-1} x^{n-2} - \dots - ab_2 x - ab_1 \\ &= b_n x^n + (b_{n-1} - ab_n) x^{n-1} + \dots + (b_1 - ab_2) x - ab_1 \\ &= c_n x^n + c_{n-1} x^{n-1} + \dots + c_1 x + (b_0 - ab_1) - b_0 \\ &= c_n x^n + c_{n-1} x^{n-1} + \dots + c_1 x + c_0 - b_0 \\ &= p(x) - b_0, \end{aligned}$$

där vi utnyttjat att $c_n = b_n$ och $c_k = b_{k+1} - ab_k$ för alla $k = 0, 1, \dots, n-1$. Sätt

$$q(x) = b_n x^{n-1} + b_{n-1} x^{n-2} + \dots + b_2 x + b_1.$$

Vi har visat att

$$(x - a)q(x) = p(x) - b_0$$

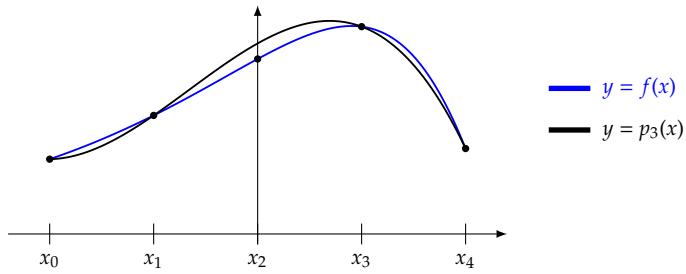
eller ekvivalent

$$p(x) = (x - a)q(x) + b_0.$$

Eftersom kvoten och resten i divisionsalgoritmen är entydigt bestämd och därmed har vi att $r = b_0$. Det ger att

$$p'(x) = (x - a)q'(x) + q(x)$$

och vi kan konstatera att $p'(a) = q(a)$. Det ger oss en naturlig utvidning av Horners metod för att beräkna $p'(a)$, dvs vi ska beräkna funktionsvärdet av ett polynom.



Figur 3.3

3.2 Interpolation av en funktion

Låt $f: \mathbb{R} \rightarrow \mathbb{R}$. Antag att det är svårt att beräkna $f(a)$ för ett givet tal a . Två anledningar kan vara uttrycket för $f(x)$ är komplicerat eller helt enkelt okänt. Antag nu att vi vet att $y_i = f(x_i)$ i några punkter x_i för vilka

$$x_0 < x_1 < \dots < x_n.$$

Vi vill bestämma en funktion $p: \mathbb{R} \rightarrow \mathbb{R}$ för vilken $p(a)$ är enkel att beräkna och som uppfyller

$$p(x_i) = y_i, \quad \text{for } i = 0, 1, \dots, n.$$

Om $x_0 \leq a \leq x_n$, så kallas $p(a)$ för en *interpolation av $f(a)$* och vi säger att funktionen p *interpolerar f i intervallet $[x_0, x_n]$* . Om $a \neq x_i$, så gäller troligtvis att $p(a) \neq f(a)$. När a ligger utanför det minsta intervallet som innehåller samtliga x_i , dvs om $a \notin [x_0, x_n]$, så kallas $p(a)$ för en *extrapolation av $f(a)$* .

Exempel 3.3. Vi ska intrpolera funktionen

$$f(x) = \sin e^x$$

över intervallet $[-1, 1]$, se blå kurva i figur 3.3. Antag att vi endast har kännedom fem punkter (x, y) sådan att $y = f(x)$, nämligen

$$(x_0, y_0) = (-1.0, 0.359638), (x_1, y_1) = (-0.5, 0.570020), (x_2, y_2) = (0.0, 0.841471), \\ (x_3, y_3) = (0.5, 0.996965) \text{ och } (x_4, y_4) = (1.0, 0.410781),$$

vilka är markerade i figur 3.3. Till att börja med intrpolera vi f med ett polynom av grad 3, där vi använder punkterna x_0, x_1, x_3 och x_4 . Alltså ska vi bestämma koeficienterna till

$$p_3(x) = a_3x^3 + a_2x^2 + a_1x + a_0,$$

sådan att $p_3(x_i) = y_i$, for $i = 0, 1, 3, 4$, dvs

$$\begin{cases} p_3(x_0) = y_0 \\ p_3(x_1) = y_1 \\ p_3(x_3) = y_3 \\ p_3(x_4) = y_4 \end{cases}$$

eller ekvivalent

$$\begin{cases} -a_3 + a_2 - a_1 + a_0 = 0.359638 \\ -0.125a_3 + 0.25a_2 - 0.5a_1 + a_0 = 0.570020 \\ 0.125a_3 + 0.25a_2 + 0.5a_1 + a_0 = 0.996965 \\ a_3 + a_2 + a_1 + a_0 = 0.410781. \end{cases}$$

Hade vi även använt punkten (x_2, y_2) , så skulle vi erhållit ett överbestämt ekvationsystem. Lösningen ger oss polynomet

$$p_3(x) \approx -0.535164x^3 - 0.531045x^2 + 0.560736x + 0.916254, \quad (3.3)$$

se svart kurva i figur 3.3. Notera att kurvan som ges polynomet går genom de punkter vi använde, men missar (x_2, y_2) . Man behöver inte ställa upp ekvationsystemet ovan och lösa det med Gausselimination. Man kan istället gå tillväga på följande sätt. Skriv polynomet på formen

$$p_3(x) = c_0 + c_1(x - x_0) + c_2(x - x_0)(x - x_1) + c_3(x - x_0)(x - x_1)(x - x_3). \quad (3.4)$$

Genom att i tur och ordning använda punkterna (x_i, y_i) kan härleda en linjär ekvation vars lösning är koefficienten c_i . Från $y_0 = p_3(x_0) = c_0$ följer det att $c_0 = 0.359638$. Det ger i sin tur att

$$\begin{aligned} y_1 &= p_3(x_1) = c_0 + c_1(x_1 - x_0) \\ &\Leftrightarrow \\ 0.570020 &= 0.359638 + c_1(-0.5 - (-1)) \\ &\Leftrightarrow \\ c_1 &= 0.420766. \end{aligned}$$

Härnäst får vi att

$$\begin{aligned} y_3 &= p_3(x_3) = c_0 + c_1(x_3 - x_0) + c_2(x_3 - x_0)(x_3 - x_1) \\ &\Leftrightarrow \\ 0.996965 &= 0.359638 + 0.420766(0.5 - (-1)) + c_2(0.5 - (-1))(0.5 - (-0.5)) \\ &\Leftrightarrow \\ c_2 &= 0.004120 \end{aligned}$$

och till slut att

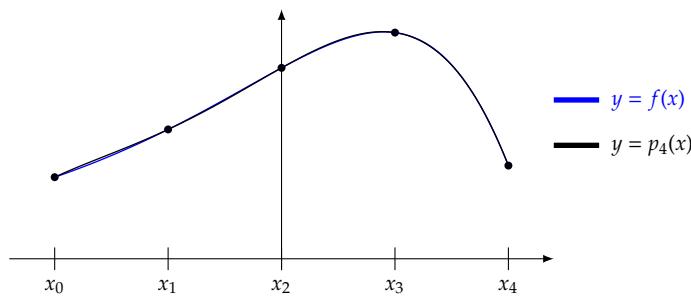
$$\begin{aligned} y_4 &= p_3(x_4) = c_0 + c_1(x_4 - x_0) + c_2(x_4 - x_0)(x_4 - x_1) + c_3(x_4 - x_0)(x_4 - x_1)(x_4 - x_3) \\ &\Leftrightarrow \\ c_4 &= \frac{y_4 - c_0 - c_1(x_4 - x_0) - c_2(x_4 - x_0)(x_4 - x_1)}{(x_4 - x_0)(x_4 - x_1)(x_4 - x_3)} = -0.535164. \end{aligned}$$

Insättning i (3.4) ger oss polynomet

$$\begin{aligned} p_3(x) &= 0.359638 + 0.420766(x - (-1.0)) + 0.004120(x - (-1.0))(x - (-0.5)) \\ &\quad - 0.535164(x - (-1.0))(x - (-0.5))(x - 0.5), \end{aligned}$$

som är samma polynom som (3.3), vilket inses tex efter förenkling (lämnas som övning). Vi har utnyttjat likheterna

$$\begin{aligned} y_0 &= p_3(x_0) = c_0 \\ y_1 &= p_3(x_1) = c_0 + c_1(x_1 - x_0) \\ y_3 &= p_3(x_3) = c_0 + c_1(x_3 - x_0) + c_2(x_3 - x_0)(x_3 - x_1) \\ y_4 &= p_3(x_4) = c_0 + c_1(x_4 - x_0) + c_2(x_4 - x_0)(x_4 - x_1) + c_3(x_4 - x_0)(x_4 - x_1)(x_4 - x_3). \end{aligned}$$



Figur 3.4

För att istället bestämma ett fjärdegradspolynom behöver vi även (x_2, y_2) . Oavsett vilken av metoderna ovan får vi att

$$p_4(x) = -0.299132x^4 - 0.535164x^3 - 0.157129x^2 + 0.560736x + 0.841471 \quad (3.5)$$

interpolerar f , se figur 3.4. Detaljerna lämnas som övning. Slutligen, antag att vi vill bestämma $f(0.75)$, som är ungefär 0.854503. De två polynomen ger approximationen

$$p_3(0.75) \approx 0.812321 \text{ respektive } p_4(0.75) \approx 0.853218.$$

Känner vi inte till själva uttrycket för f är dessa värden acceptabla. \diamond

Sats 3.1. Låt x_0, x_1, \dots, x_n vara olika reella tal och y_0, y_1, \dots, y_n reella tal. Då existerar det ett entydigt bestämt polynom p_n av grad mindre än eller lika med n sådan att

$$p_n(x_i) = y_i,$$

för alla $i = 0, 1, \dots, n$.

Bevis. Låt $P_n(\mathbb{R})$ beteckna mängden av alla polynom med reella koefficienter och av grad högst n . Definiera funktionen $F: P_n(\mathbb{R}) \rightarrow \mathbb{R}^{n+1}$ enligt

$$p(x) \mapsto (p(x_0), p(x_1), \dots, p(x_n)).$$

För alla $a \in \mathbb{R}$ och alla $p(x), q(x) \in P_n(\mathbb{R})$ gäller att

$$F(ap(x)) = (ap(x_0), ap(x_1), \dots, ap(x_n)) = a(p(x_0), p(x_1), \dots, p(x_n)) = aF(p(x))$$

och

$$\begin{aligned} F(p(x) + q(x)) &= (p(x_0)q(x_0), p(x_1) + q(x_1), \dots, p(x_n) + q(x_n)) \\ &= (p(x_0), p(x_1), \dots, p(x_n)) + (q(x_0), q(x_1), \dots, q(x_n)) \\ &= F(p(x)) + F(q(x)). \end{aligned}$$

Alltså är F en linjär avbildning. Om $p(x) \in P_n(\mathbb{R})$ uppfyller $F(p(x)) = \mathbf{0}$, dvs om $p(x_i) = 0$ för alla $i = 0, 1, \dots, n$, så betyder det att $p(x)$ har $n+1$ olika nollställen. Men eftersom ett polynom av grad $k \leq n$ har exakt k nollställen måste $p(x)$ vara nollpolynomet 0. Det visar att $\text{null } F = \{0\}$ eller ekvivalent att F injektiv. Dimensionsatsen ger nu att

$$n+1 = \dim P_n(\mathbb{R}) = \dim \text{null } F + \dim \text{range } F = 0 + \dim \text{range } F,$$

dvs $\dim \text{range } F = n+1$. Det visar att $\text{range } F = \mathbb{R}^{n+1}$ eller ekvivalent att F är surjektiv. Det visar att för $(y_0, y_1, \dots, y_n) \in \mathbb{R}^{n+1}$ existerar det exakt ett polynom $p(x) \in P_n(\mathbb{R})$ sådant att $p(x_i) = y_i$ då $i = 0, 1, \dots, n$. \square



Kräver linjär algebra

Alternativt bevis. Existensen av polynomet $p(x)$ följer direkt från tex konstruktionen med hjälp av Lagranges koefficientpolynom, se nästa avsnitt och speciellt (3.6). Det återstår att visa att $p(x)$ är entydigt bestämt. Antag att även polynomet $q(x)$ är av grad högst n samt uppfyller $q(x_i) = y_i$ för alla $i = 0, 1, \dots, n$. Sätt $f(x) = p(x) - q(x)$. Då är

$$f(x_i) = p(x_i) - q(x_i) = y_i - y_i = 0,$$

för alla $i = 0, 1, \dots, n$. Alltså har f minst $n + 1$ olika nollställen. Men eftersom $f(x)$ är ett polynom av grad högst n så måste $f(x)$ vara nollpolynomet, dvs $p(x) = q(x)$. \square

Sats 3.2. *Antag att de $n + 1$ första derivatorna av funktionen f existerar och är kontinuerliga över det minsta intervallet I som innehåller x, x_0, x_1, \dots, x_n . Om p_n är det entydigt bestämda polynom av grad mindre än eller lika med n som uppfyller*

$$p_n(x_i) = f(x_i) = y_i$$

för alla $i = 0, 1, \dots, n$, så ges felet i $x \in I$ av

$$E_n(x) = f(x) - p_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} (x - x_0)(x - x_1) \cdots (x - x_n),$$

där $\xi = \xi(x) \in I$.

Bevis. Om $x = x_i$ för något $i = 0, 1, \dots, n$, så är $E_n(x) = 0$ eftersom $f(x_i) = p(x_i)$. Antag i fortsättningen att $x \neq x_i$, där $i = 0, 1, \dots, n$. Låt $g: \mathbb{R} \rightarrow \mathbb{R}$ enligt

$$g(t) = f(t) - p_n(t) - a(t - x_0)(t - x_1) \cdots (t - x_n)$$

där

$$a = \frac{f(x) - p_n(x)}{(x - x_0)(x - x_1) \cdots (x - x_n)}.$$

Vi ser att

$$g(x) = g(x_0) = g(x_1) = \cdots = g(x_n) = 0,$$

dvs funktionen g har minst $n + 2$ olika nollställen. Enligt Rolles sats finns det ett $\xi \in I$ sådant att $g^{(n+1)}(\xi) = 0$. Från

$$g^{(n+1)}(t) = f^{(n+1)}(t) - a(n+1)!$$

följer nu satsen. \square

3.3 Lagranges interpolation



Om $x_0 < x < x_1$, så ges den linjära interpolationen av $f(x)$ av

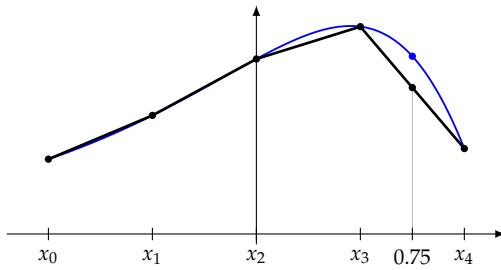
$$p_1(x) = y_0 + \frac{x - x_0}{x_1 - x_0} (y_1 - y_0).$$

dvs ekvationen för den linje som går genom (x_0, y_0) och (x_1, y_1) .

Exempel 3.4. Låt $f(x) = \sin e^x$. Med de punkter (x_i, y_i) som vi utgick från i exempel 3.3 har vi fyra delintervall. För var och en av dessa kan vi interpolera f med ett linjärt polynom, dvs vi ska skapa ett polygontåg, se figur 3.5. Vid approximation av $f(0.75)$ använder vi x_3 och x_4 , eftersom $x_3 \leq 0.75 \leq x_4$. Det ger att

$$p_1(0.75) = y_3 + \frac{x - x_3}{x_4 - x_3} (y_4 - y_3) \approx 0.997 + \frac{0.75 - 0.5}{1.0 - 0.5} (0.411 - 0.997) \approx 0.704.$$

Jämför med $f(0.75) = \sin e^{0.75} \approx 0.854503$. \diamond



Figur 3.5

Vi ska studera en annan metod att bestämma det entydigt bestämda polynom p_n av grad n eller lägre som interpolerar en funktion f i $n + 1$ givna punkter x_i , dvs vi har att $p_n(x_i) = f(x_i)$. Antag att $x_0 < x_1$. Sätt

$$L_{1,0}(x) = \frac{x - x_1}{x_0 - x_1} \quad \text{och} \quad L_{1,1}(x) = \frac{x - x_0}{x_1 - x_0}.$$

Då är

$$\begin{aligned} p_1(x) &= y_0 + \frac{x - x_0}{x_1 - x_0}(y_1 - y_0) \\ &= \frac{y_0(x_1 - x_0) + (x - x_0)(y_1 - y_0)}{x_1 - x_0} \\ &= -y_0 \frac{x - x_1}{x_1 - x_0} + y_1 \frac{x - x_0}{x_1 - x_0} \\ &= y_0 L_{1,0}(x) + y_1 L_{1,1}(x). \end{aligned}$$

Notera att

$$L_{1,0}(x_0) = 1 \quad \text{och} \quad L_{1,1}(x_0) = 0$$

samt

$$L_{1,0}(x_1) = 0 \quad \text{och} \quad L_{1,1}(x_1) = 1.$$

Alltså är $p_1(x_0) = y_0 = f(x_0)$ och $p_1(x_1) = y_1 = f(x_1)$. Generellt har vi att

$$p_n(x) = \sum_{k=0}^n y_k L_{n,k}(x)$$

där *Lagranges koefficientpolynom* har formen

$$L_{n,k}(x) = \frac{(x - x_0) \cdots (x - x_{k-1})(x - x_{k+1}) \cdots (x - x_n)}{(x_k - x_0) \cdots (x_k - x_{k-1})(x_k - x_{k+1}) \cdots (x_k - x_n)}.$$



Bevis

där faktorerna $x - x_k$ och $x_k - x_k$ saknas i täljaden respektive nämnaren. Vi ser att

$$L_{n,k}(x_i) = \begin{cases} 1 & \text{if } i = k \\ 0 & \text{if } i \neq k. \end{cases}$$

Med andra ord är

$$p_n(x_i) = y_0 L_{n,0}(x_i) + \cdots + y_i L_{n,i}(x_i) + \cdots + y_n L_{n,n}(x_i) = y_i \cdot 1 = y_i, \quad (3.6)$$

dvs med hjälp av Lagranges koefficientpolynom erhåller vi ett polynom som interpolerar samtliga givna punkter.

Exempel 3.5. Tag ett djupt andetag! Om $n = 4$, så är

$$\begin{aligned}L_{4,0}(x) &= \frac{(x - x_1)(x - x_2)(x - x_3)(x - x_4)}{(x_0 - x_1)(x_0 - x_2)(x_0 - x_3)(x_0 - x_4)} \\L_{4,1}(x) &= \frac{(x - x_0)(x - x_2)(x - x_3)(x - x_4)}{(x_1 - x_0)(x_1 - x_2)(x_1 - x_3)(x_1 - x_4)} \\L_{4,2}(x) &= \frac{(x - x_0)(x - x_1)(x - x_3)(x - x_4)}{(x_2 - x_0)(x_2 - x_1)(x_2 - x_3)(x_2 - x_4)} \\L_{4,3}(x) &= \frac{(x - x_0)(x - x_1)(x - x_2)(x - x_4)}{(x_3 - x_0)(x_3 - x_1)(x_3 - x_2)(x_3 - x_4)} \\L_{4,4}(x) &= \frac{(x - x_0)(x - x_1)(x - x_2)(x - x_3)}{(x_4 - x_0)(x_4 - x_1)(x_4 - x_2)(x_4 - x_3)}.\end{aligned}$$

Pust! Med samma förutsättningar som i exempel 3.3 får vi tex att

$$\begin{aligned}L_{4,3}(x) &= \frac{(x + 1)(x + 0.5)(x - 0)(x - 1)}{(0.5 + 1)(0.5 + 0.5)(0.5 - 0)(0.5 - 1)} \\&= -\frac{x(x + 1)(x + 0.5)(x - 1)}{0.375} \\&\approx -2.66667x(x + 1)(x + 0.5)(x - 1).\end{aligned}$$

De fyra övriga fås på samma sätt, vilket lämnas som övning. Till slut får vi att

$$\begin{aligned}L_{4,0}(x) &= 0.666667x(x + 0.5)(x - 0.5)(x - 1) \\L_{4,1}(x) &= -2.66667x(x + 1)(x - 0.5)(x - 1) \\L_{4,2}(x) &= 4(x + 1)(x + 0.5)(x - 0.5)(x - 1) \\L_{4,3}(x) &= -2.66667x(x + 1)(x + 0.5)(x - 1) \\L_{4,4}(x) &= 0.666667x(x + 1)(x + 0.5)(x - 0.5).\end{aligned}$$

Därmed är

$$\begin{aligned}p_4(x) &= \sum_{k=0}^4 y_k L_{4,k}(x) \\&= y_0 L_{4,0}(x) + y_1 L_{4,1}(x) + y_2 L_{4,2}(x) + y_3 L_{4,3}(x) + y_4 L_{4,4}(x) \\&= -0.299132x^4 - 0.535164x^3 - 0.157129x^2 + 0.560736x + 0.841471.\end{aligned}$$

Jämför med (3.5). Andas ut. ◊

3.4 Newtons interpolationsformel



Låt $f: \mathbb{R} \rightarrow \mathbb{R}$ vara en funktion och x_0, x_1, \dots, x_n reella tal, där $x_0 < x_1 < \dots < x_n$. De sk *dividerade differenserna* för f med avseende på noderna x_0, x_1, \dots, x_k ges rekursivt av

$$f[x_i] = f(x_i) = y_i$$

och

$$f[x_0, x_1, \dots, x_k] = \frac{f[x_1, x_2, \dots, x_k] - f[x_0, x_1, \dots, x_{k-1}]}{x_k - x_0}.$$

Vi har tex att

$$f[x_0, x_1] = \frac{f[x_1] - f[x_0]}{x_1 - x_0} = \frac{y_1 - y_0}{x_1 - x_0}$$

och

$$f[x_1, x_2] = \frac{f[x_2] - f[x_1]}{x_2 - x_1} = \frac{y_2 - y_1}{x_2 - x_1}.$$

Därmed är

$$f[x_0, x_1, x_2] = \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0} = \frac{\frac{y_2 - y_1}{x_2 - x_1} - \frac{y_1 - y_0}{x_1 - x_0}}{x_2 - x_0}.$$

För att underlätta beräkningarna av de dividerade differenserna arbeta man med en uppställning i tabellform.

Sats 3.3. Det entydigt besämda polynomet p_n som uppfyller $p_n(x_i) = f(x_i) = y_i$ ges av

$$\begin{aligned} p_n(x) &= f[x_0] + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_1) \\ &\quad + \cdots + f[x_0, x_1, \dots, x_n](x - x_0)(x - x_1) \cdots (x - x_{n-1}). \end{aligned}$$



Bevis

Exempel 3.6. Låt $f(x) = \sin e^x$ och låt (x_i, y_i) vara samma punkter som i exempel 3.3. Vi ska skapa en uppställning på följande form.

x_0	y_0	$f[x_0, x_1]$			
x_1	y_1	$f[x_1, x_2]$	$f[x_0, x_1, x_2]$		
x_2	y_2	$f[x_2, x_3]$	$f[x_0, x_1, x_2, x_3]$	$f[x_0, x_1, x_2, x_3, x_4]$	
x_3	y_3	$f[x_3, x_4]$	$f[x_1, x_2, x_3, x_4]$		
x_4	y_4				

Notera att $y_i = f[x_i]$. I detalj har vi tex att

$$f[x_0, x_1] = \frac{y_1 - y_0}{x_1 - x_0} = \frac{0.570020 - 0.359638}{-0.5 - (-1.0)} \approx 0.420766,$$

$$f[x_1, x_2] = \frac{y_2 - y_1}{x_2 - x_1} = \frac{0.841471 - 0.570020}{0 - (-0.5)} \approx 0.542901$$

och

$$f[x_0, x_1, x_2] = \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0} = \frac{0.542901 - 0.420766}{0 - (-1.0)} \approx 0.122136.$$

Kolumn för kolumn får vi till slut följande tabell.

x_i	$y_i = f[x_i]$	$f[\cdot, \cdot]$	$f[\cdot, \cdot, \cdot]$	$f[\cdot, \cdot, \cdot, \cdot]$	$f[\cdot, \cdot, \cdot, \cdot, \cdot]$
-1.0	0.359638				
		0.420766			
-0.5	0.570020		0.122136		
			0.542901	-0.236032	
0.0	0.841471		-0.231912		-0.299132
			0.310989	-0.834296	
0.5	0.996965		-1.48336		
			-1.17237		
1.0	0.410781				

Notera hur varje dividerad differens utgör ett hörn i en "liggande likbent triangel" och där vi finner x -värdena genom att följa "sidorna". Vi har tex att

$$\begin{aligned} f[x_1, x_2, x_3, x_4] &= \frac{f[x_2, x_3, x_4] - f[x_1, x_2, x_3]}{x_4 - x_1} \\ &= \frac{-1.48336 - (-0.231912)}{1.0 - (-0.5)} \approx -0.834296. \end{aligned}$$

Alltså är

$$\begin{aligned} p_4(x) &= f[x_0] + f[x_0, x_1](x+1) + f[x_0, x_1, x_2](x+1)(x+0.5) \\ &\quad + f[x_0, x_1, x_2, x_3](x+1)(x+0.5)(x-0) \\ &\quad + f[x_0, x_1, x_2, x_3, x_4](x+1)(x+0.5)(x-0)(x-0.5) \\ &= 0.359638 + 0.420766(x+1) + 0.122136(x+1)(x+0.5) \\ &\quad - 0.236032(x+1)(x+0.5)(x-0) \\ &\quad - 0.299132(x+1)(x+0.5)(x-0)(x-0.5) \\ &= -0.299132x^4 - 0.535164x^3 - 0.157129x^2 + 0.560736x + 0.841471. \end{aligned}$$

Jämför med (3.5). ◊

Exempel 3.7 (Tre olika metoder). Låt punkterna (x_k, y_k) , där $k = 0, 1, 2, 3$, ges av

$$\left(-1, \frac{15}{4}\right), \left(0, \frac{1}{4}\right), \left(\frac{3}{2}, -\frac{5}{4}\right) \text{ respektive } \left(2, -\frac{3}{4}\right).$$

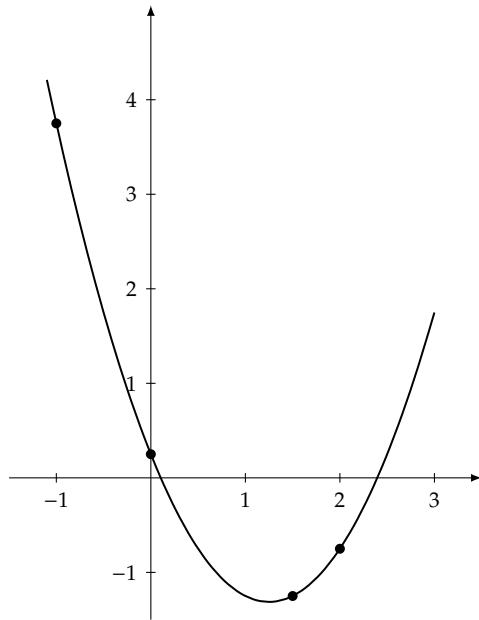
Låt oss studera hur man med Lagranges interpolationsformel och Newtons interpolationsformel bestämmer det entydiga tredjegradspolynom p_3 vars kurva $y = p_3(x)$ går genom samtliga fyra givna punkter. Lagranges koefficientpolynom ges av

$$\begin{aligned} L_{3,0}(x) &= \frac{(x-0)(x-3/2)(x-2)}{(-1-0)(-1-3/2)(-1-2)} = \frac{2}{15}x(x-3/2)(x-2) \\ L_{3,1}(x) &= \frac{(x-(-1))(x-3/2)(x-2)}{(0-(-1))(0-3/2)(0-2)} = -\frac{1}{3}(x+1)(x-3/2)(x-2) \\ L_{3,2}(x) &= \frac{(x-(-1))(x-0)(x-2)}{(3/2-(-1))(3/2-0)(3/2-2)} = -\frac{8}{15}x(x+1)(x-2) \\ L_{3,3}(x) &= \frac{(x-(-1))(x-0)(x-3/2)}{(2-(-1))(2-0)(2-3/2)} = \frac{1}{3}x(x+1)(x-3/2). \end{aligned}$$

Då är

$$\begin{aligned} p_3(x) &= \sum_{k=0}^3 y_k L_{3,k}(x) \\ &= \frac{15}{4} \left(\frac{2}{15}x(x-3/2)(x-2) \right) + \frac{1}{4} \left(-\frac{1}{3}(x+1)(x-3/2)(x-2) \right) \\ &\quad + \left(-\frac{5}{4} \right) \left(-\frac{8}{15}x(x+1)(x-2) \right) + \left(-\frac{3}{4} \right) \left(\frac{1}{3}x(x+1)(x-3/2) \right) \\ &= x^2 - \frac{5}{2}x + \frac{1}{4}. \end{aligned}$$

Notera att vi erhåller ett polynom av grad 2 och inte 3. Det beror på att punkterna råkar ligga på parabeln $y = p_3(x)$, se figur 3.6. Med Newtons dividerade differenser får vi



Figur 3.6

följande uppställning.

$$\begin{array}{ccccccc}
 -1 & \frac{15}{4} & & -\frac{7}{2} & & 1 & 0 \\
 & & & & & & \\
 0 & \frac{1}{4} & & -1 & & 1 & 0 \\
 & & & & & & \\
 \frac{3}{2} & -\frac{5}{4} & & & & 1 & \\
 & & & & & & \\
 2 & -\frac{3}{4} & & & & &
 \end{array}$$

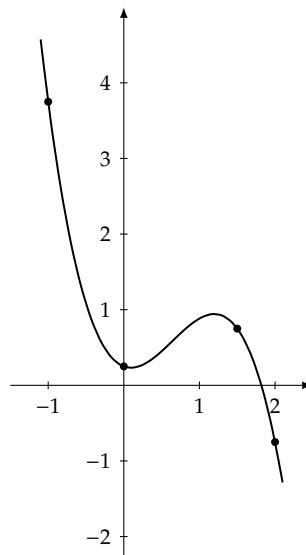
Alltså är

$$f[x_0] = \frac{15}{4}, \quad f[x_0, x_1] = -\frac{7}{2}, \quad f[x_0, x_1, x_2] = 1 \quad \text{och} \quad f[x_0, x_1, x_2, x_3] = 0.$$

Det ger att

$$\begin{aligned}
 p_3(x) &= f[x_0] + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_1) \\
 &\quad + f[x_0, x_1, x_2, x_3](x - x_0)(x - x_1)(x - x_2) \\
 &= \frac{15}{4} - \frac{7}{2}(x - (-1)) - 1 \cdot (x - (-1))(x - 0) \\
 &\quad + 0 \cdot (x - (-1))(x - 0)(x - 3/2) \\
 &= x^2 - \frac{5}{2}x + \frac{1}{4}.
 \end{aligned}$$

En tredje metod för att bestämma $p_3(x)$ är att ställa upp det ekvationssystem som ges



Figur 3.7

av $p_3(x_k) = y_k$, för alla $k = 0, 1, 2, 3$, där $p_3(x) = a_0 + a_1x + a_2x^2 + a_3x^3$, dvs

$$\begin{cases} a_0 - a_1 + a_2 - a_3 = \frac{15}{4} \\ a_0 = \frac{1}{4} \\ a_0 + \frac{3}{2}a_1 + \frac{9}{4}a_2 + \frac{27}{8}a_3 = -\frac{5}{4} \\ a_0 + 2a_1 + 4a_2 + 8a_3 = -\frac{3}{4}. \end{cases}$$

Gausselimination med pivotering (förstås) följt av bakåtsubstitution ger lösningen

$$a_0 = \frac{1}{4}, \quad a_1 = -\frac{5}{2}, \quad a_2 = 1 \quad \text{och} \quad a_3 = 0,$$

dvs koefficienterna till det polynom p_3 vi redan funnit tidigare med Lagranges respektive Newtons inrerpolarionsformel. ◇

Exempel 3.8. Låt punkterna (x_k, y_k) , där $k = 0, 1, 2, 3$, ges av

$$\left(-1, \frac{15}{4}\right), \left(0, \frac{1}{4}\right), \left(\frac{3}{2}, \frac{3}{4}\right) \quad \text{respektive} \quad \left(2, -\frac{3}{4}\right),$$

dvs nästan samma datamängd som i exempel 3.7. Det är endast $y_2 = 3/4$ som skiljer sig från $y_2 = -5/4$ i föregående exempel. Det polynom som av grad 3 som interpolera dessa fyra punkter ges av

$$p_3(x) = -\frac{16}{15}x^3 + \frac{31}{15}x^2 - \frac{11}{30}x + \frac{1}{4},$$

se figur 3.7. Detaljerna lämnas som övning. ◇



3.5 Potensberäkning

Låt a och b vara reella tal. Vi ska här studera hur man kan gå till väga för att beräkna a^b . Det gör vi genom att dela upp i olika fall med avseende på exponenten b . Vi antar att a och b är positiva.

3.5.1 Heltalsexponent

Antag att b är ett heltalet. Låt

$$b = (b_n \dots b_1 b_0)_{\text{två}} = b_n \cdot 2^n + \dots + b_1 \cdot 2 + b_0$$

vara den binära representationen av b . Potenslagarna ger att

$$a^b = a^{b_n \cdot 2^n + \dots + b_1 \cdot 2 + b_0} = a^{b_n \cdot 2^n} \cdots a^{b_1 \cdot 2} \cdot a^{b_0} = (a^{2^n})^{b_n} (a^{2^{n-1}})^{b_{n-1}} \cdots (a^2)^{b_1} \cdot a^{b_0}.$$

Notera att b_k är antingen 0 eller 1, för alla $k = 0, 1, \dots, n$. Det betyder att

$$(a^{2^k})^{b_k} = \begin{cases} 1 & \text{om } b_k = 0 \\ a^{2^k} & \text{om } b_k = 1. \end{cases}$$

Vidare är

$$a^{2^k} \cdot a^{2^k} = (a^{2^k})^2 = a^{2 \cdot 2^k} = a^{2^{k+1}}. \quad (3.7)$$

Det ger oss en möjlighet att reducera antalet multiplikationer som måste utföras. Låt tex $a = 1.04$ och $b = 57$. Potensen $a^b = 1.04^{57}$ består av 56 multiplikationer. Vi har att

$$b = 57 = (111001)_{\text{två}} = 1 \cdot 2^5 + 1 \cdot 2^4 + 1 \cdot 2^3 + 0 \cdot 2^2 + 0 \cdot 2 + 1.$$

Vidare får vi från (3.7) att

$$\begin{aligned} a^2 &= a \cdot a = 1.04 \cdot 1.04 = 1.0816 \\ a^{2^2} &= a^2 \cdot a^2 = 1.0816 \cdot 1.0816 = 1.16986 \\ a^{2^3} &= a^{2^2} \cdot a^{2^2} = 1.16986 \cdot 1.16986 = 1.36857 \\ a^{2^4} &= a^{2^3} \cdot a^{2^3} = 1.36857 \cdot 1.36857 = 1.87298 \\ a^{2^5} &= a^{2^4} \cdot a^{2^4} = 1.87298 \cdot 1.87298 = 3.50806. \end{aligned}$$

Så här långt har vi utfört fem multiplikationer. Nu utnyttjar vi den binära representationen av b och får att

$$\begin{aligned} a^b &= a^{1 \cdot 2^5 + 1 \cdot 2^4 + 1 \cdot 2^3 + 0 \cdot 2^2 + 0 \cdot 2 + 1} = a^{2^5} \cdot a^{2^4} \cdot a^{2^3} \cdot a \\ &= 3.50806 \cdot 1.87298 \cdot 1.36857 \cdot 1.04 = 9.35191. \end{aligned}$$

Med denna metod krävdes det endast nio multiplikationer!

3.5.2 Rationell exponent

Antag att $b = p/q$, där p och q är positiva heltalet. Då är

$$a^b = a^{p/q} = (a^p)^{1/q}.$$

Vi såg ovan hur vi beräknar potenser då exponenten är ett heltal. Därför räcker det att studera fallet $b = 1/q$. Vi söker det reella tal x som uppfyller

$$a^{1/q} = x \quad \Rightarrow \quad a = x^q.$$

Sätt $f(x) = x^q - a$ och lös ekvationen $f(x) = 0$ med tex Newton-Raphsons metod. Eftersom q är ett heltal så kan vi beräkna $f(x)$ och $f'(x)$ med den metod som vi studerade ovan.

3.5.3 Irrationell exponent

Vid flyttalsaritmetik kan vi inte hamna i den situation att b är irrationell – om b är givet som ett flyttal i en dator är b ett rationellt tal. Men även om vi kan bestämma p och q i $b = p/q$, så kan det vara omöjligt att beräkna a^p om p är stort, tex riskera vi överspill om $a > 1$ och underspill om $0 < a < 1$. Oavsett om b är rationellt eller inte kan vi utgå från

$$a^b = e^{\ln a^b} = \exp(\ln a^b) = \exp(b \ln a).$$

Vi kan beräkna logaritm- och exponentialfunktionen med hjälp av tex deras Taylorutvecklingar, dvs

$$\ln(1+x) = \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k} x^k \quad \text{respektive} \quad e^x = \sum_{k=0}^{\infty} \frac{1}{k!} x^k.$$

Motsvarande Taylorpolynom

$$p_n(x) = x - \frac{x^2}{2} + \frac{x^3}{3} \pm \cdots + (-1)^{n+1} \frac{x^n}{n}$$

respektive

$$Q_n(x) = 1 + x + \frac{x^2}{2} + \cdots + \frac{x^n}{n!}$$

av tillräcklig hög grad n ger en god approximation för alla x nära 0. Konvergensen är dock mycket långsam. Om $|a-1| < 1$, dvs om a är nära 1, så kan vi använda p_n och approximera $\ln a$ med $c = p_n(a-1)$. Från logaritmlagen $\ln(1/x) = -\ln x$ följer att vi kan approximera $\ln a$ med $c = -p_n(1/(a-1))$ för stora a . Därefter är det dags att beräkna $\exp(bc)$. Om bc är nära 0, så är $a^b \approx Q_n(bc)$. Om $|bc|$ är stort, så bestämmer vi det heltal k sådant att $0 < |bc|/2^k < 1$. Då är

$$\exp(bc) = e^{bc} = \left(e^{bc/2^k}\right)^{2^k} = \exp(bc/2^k)^{2^k}.$$

Notera att dividera med 2^k är detsamma som högerskift k steg vid binär representation. Således bestämmes vi $d = Q_n(bc/2^k)$ och därefter $a^b \approx d^{2^k}$. För att beräkna d^{2^k} krävs k multiplikationer, se diskussionen om heltalexponenter ovan.

3.6 Linjära splinefunktioner

Om vi för att förbättra interpolationen ökar antal punkter, så kommer det att öka graden av interpolationspolynomet. En lösning är introducera konceptet *splinefunktion*. Låt m och n vara positiva heltal och $[a, b]$ ett slutet interval uppdelat enligt

$$a = x_0 < x_1 < \cdots < x_n = b.$$

Då är en *spline av ordning* $2m + 1$ en funktion $S: [a, b] \rightarrow \mathbb{R}$ vilken interpolerar *noderna*

$$(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n),$$

och är styckvis definierad med polynom av grad $2m + 1$, med ett polynom för varje delintervall $[x_{i-1}, x_i]$. Vidare ska $S^{(k)}$ vara kontinuerlig för alla $k = 0, 1, \dots, 2m$.

För att bestämma en spline funktion S utgår man således från det man vet, nämligen noderna $(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n)$, och de önskade egenskaperna hos S , för att härleda ekvationer med avseende på koeficienterna i polynomen

$$\begin{aligned} S_1(x) &= c_{1,0} + c_{1,1}x + c_{1,2}x^2 + \dots + c_{1,2m+1}x^{2m+1} \\ S_2(x) &= c_{2,0} + c_{2,1}x + c_{2,2}x^2 + \dots + c_{2,2m+1}x^{2m+1} \\ &\vdots \\ S_n(x) &= c_{n,0} + c_{n,1}x + c_{n,2}x^2 + \dots + c_{n,2m+1}x^{2m+1}. \end{aligned}$$

Splinefunktionen S ges därför av

$$S(x) = \begin{cases} S_1(x) & \text{om } x_0 \leq x \leq x_1 \\ S_2(x) & \text{om } x_1 \leq x \leq x_2 \\ \vdots \\ S_n(x) & \text{om } x_{n-1} \leq x \leq x_n \end{cases}$$

Vi kommer endast studera *linjära* och *kubiska splinefunktioner*, dvs då $m = 0$ respektive då $m = 1$. Vi börjar med det linjära fallet och vänta med att studera de kubiska splinefunktionerna till nästa avsnitt. Låt

$$S(x) = S_i(x), \quad x_{i-1} \leq x \leq x_i \quad \text{och} \quad i = 1, 2, \dots, n$$

där varje S_i är ett linjärt polynom, dvs $S_i(x) = a_i + b_i(x - x_{i-1})$. Villkoret att $S^{(0)} = S$ ska vara kontinuerlig är ekvivalent med

$$S_i(x_i) = S_{i+1}(x_i) \Leftrightarrow a_i + b_i(x_i - x_{i-1}) = a_{i+1} + b_{i+1}(x_i - x_i),$$

för alla $i = 1, 2, \dots, n - 1$. Alltså är

$$a_i + b_i(x_i - x_{i-1}) = a_{i+1}.$$

Vidare ska S interpolera noderna, dvs

$$S(x_i) = y_i, \quad i = 0, 1, \dots, n,$$

eller ekvivalent

$$S_i(x_{i-1}) = a_i = y_{i-1} \quad \text{och} \quad S_n(x_n) = y_n, \quad i = 1, 2, \dots, n.$$

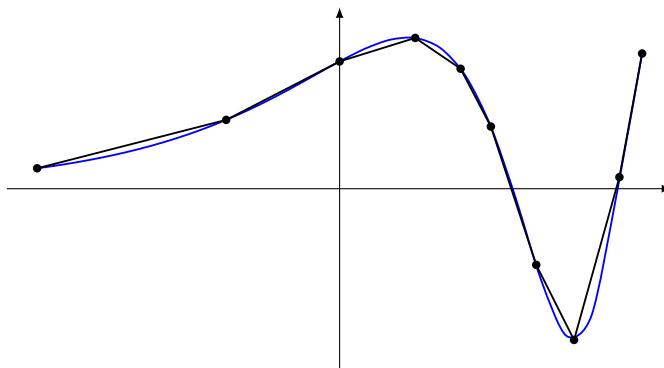
Vi har visat att

$$a_i = y_{i-1} \quad \text{och} \quad b_i = \frac{y_i - y_{i-1}}{x_i - x_{i-1}},$$

för alla $i = 1, 2, \dots, n$. Alltså är

$$S(x) = S_i(x) = y_{i-1} + \frac{y_i - y_{i-1}}{x_i - x_{i-1}}(x - x_{i-1}),$$

då $x_{i-1} \leq x \leq x_i$.



Figur 3.8

Exempel 3.9. Låt $f(x) = \sin e^x$ och $y_k = f(x_k)$, där

$$(x_k)_{k=0}^9 = (-2, -0.75, 0, 0.5, 0.8, 1, 1.3, 1.55, 1.85, 2).$$

Då är

$$(y_k)_{k=0}^9 \approx (0.134923, 0.454995, \dots, 0.893855).$$

Koefficienterna i tex $S_1(x) = a_1 + b_1(x - x_0)$ är således

$$a_1 = y_0 \approx 0.134923 \quad \text{och} \quad b_1 = \frac{0.454995 - 0.134923}{-0.75 - (-2)} \approx 0.256058.$$

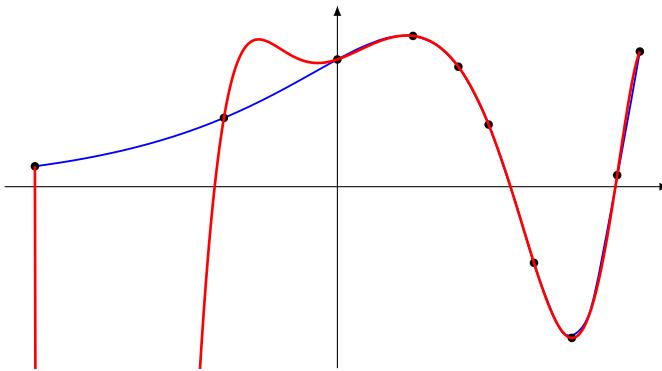
Alltså är $S_1(x) = 0.134923 + 0.256058(x - (-2))$. På sammasätt bestämmer vi koefficienterna för S_2, S_3, \dots, S_9 . Den sökta splinefunktionen ges av

$$S(x) = \begin{cases} S_1(x) = 0.134923 + 0.256058(x + 2) & \text{om } -2 \leq x \leq -0.75 \\ S_2(x) = 0.454995 + 0.515301(x + 0.75) & \text{om } -0.75 \leq x \leq 0 \\ S_3(x) = 0.841471 + 0.310989x & \text{om } 0 \leq x \leq 0.5 \\ S_4(x) = 0.996965 - 0.679206(x - 0.5) & \text{om } 0.5 \leq x \leq 0.8 \\ S_5(x) = 0.793203 - 1.91211(x - 0.8) & \text{om } 0.8 \leq x \leq 1 \\ S_6(x) = 0.410781 - 3.04777(x - 1) & \text{om } 1 \leq x \leq 1.3 \\ S_7(x) = -0.503551 - 1.98579(x - 1.3) & \text{om } 1.3 \leq x \leq 1.55 \\ S_8(x) = -1 + 3.58853(x - 1.55) & \text{om } 1.55 \leq x \leq 1.85 \\ S_9(x) = 0.0765592 + 5.44864(x - 1.85) & \text{om } 1.85 \leq x \leq 2, \end{cases}$$

se figur 3.8. Funktionen S beskriver ett polygontåg genom de givna noderna. Låt oss approximera $f(0.3) \approx 0.9757$ med S . Om $x = 0.3$, så är $x_2 < x < x_3$ och

$$S(0.3) = S_3(0.3) = 0.841471 + 0.310989 \cdot 0.3 = 0.934768.$$

Vid interpolation med endast ett polynom erhålls ett polynom p_9 av grad nio som visserligen går genom samtliga tio noder, men avviker mycket från $y = f(x)$ i intervallet $[-2, 0]$, se figur 3.9. Vi tex att $p_9(-1.74) \approx -44.1$. Det lämnas som övning att bestämma p_9 . \diamond



Figur 3.9

3.7 Kubiska splinefunktioner

Sats 3.4. Låt $(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n)$ vara noder, där $a = x_0 < x_1 < \dots < x_n = b$. Då existerar det en entydigt bestämd kubisk splinefunktion $S: [a, b] \rightarrow \mathbb{R}$ sådan att

$$S(x_i) = y_i, \quad i = 0, 1, \dots, n,$$

med

$$S_i^{(r)}(x_i) = S_{i+1}^{(r)}(x_i), \quad i = 1, 2, \dots, n-1 \quad \text{och} \quad r = 0, 1, 2,$$

där $S(x) = S_i(x)$, då $x_{i-1} \leq x \leq x_i$, och $\deg S_i \leq 3$, tillsammans med ett av följande villkor:

- (a) $S''_1(x_0) = S''_n(x_n) = 0$ naturlig splinefunktion
- (b) $S'_1(x_0) = \alpha$ och $S'_n(x_n) = \beta$ "clamped boundary"
- (c) $S_1^{(3)}(x_1) = S_2^{(3)}(x_1)$ och $S_{n-1}^{(3)}(x_{n-1}) = S_n^{(3)}(x_{n-1})$ "not-a-knot"

där α och β är konstanter.

Anmärkning. De extra randvillkoren krävs för att göra S entydigt definierad. Det finns fler andra villkor än de som nämns i satsen.

Bevisskiss och tillika konstruktion. Villkoren för S ger att $S^{(r)}(x) = S_i^{(r)}(x)$ då $x_{i-1} \leq x \leq x_i$ och $r = 0, 1, 2$. Låt $h_i = x_i - x_{i-1}$ för $i = 1, 2, \dots, n$ och $k_i = S'(x_i)$ för $i = 0, 1, \dots, n$. Så här långt känner vi endast värdet på varje x_i , h_i och y_i , medan samtliga k_i är för stunden okända. Ansätt

$$S_i(x) = a_i + b_i u_i + c_i u_i^2 + d_i u_i^3,$$

där

$$S_i(x) = S_i(u_i) = S_i(u_i(x)) \quad \text{och} \quad u_i = u_i(x) = \frac{x - x_{i-1}}{x_i - x_{i-1}} = \frac{x - x_{i-1}}{h_i}.$$

Observera att definitionen av u_i beror på i , dvs vi använder med andra ord olika variablersubstitutioner för olika $S_i(x)$. Anledningen är bla att varje polynom $S_i(x)$ endast är definierad på intervallet $[x_{i-1}, x_i]$ och att när vi studerar $S_i(x)$ då $x = x_{i-1}$ eller $x = x_i$ ger det enkla uttryck. Speciellt, om $x = x_{i-1}$ eller $x = x_i$, så är

$$u_i(x_{i-1}) = \frac{x_{i-1} - x_{i-1}}{h_i} = 0 \quad \text{respektive} \quad u_i(x_i) = \frac{x_i - x_{i-1}}{h_i} = \frac{h_i}{h_i} = 1, \quad (3.8)$$

eftersom $h_i = x_i - x_{i-1}$. Vi vill bestämma koefficienterna a_i, b_i, c_i och d_i . Villkoret att S ska vara kontinuerlig, dvs att

$$S_i^{(0)}(x_{i-1}) = S_{i+1}^{(0)}(x_{i-1}), \quad i = 1, 2, \dots, n-1,$$

ger tillsammans med villkoren $S(x) = S_i(x)$, då $x_{i-1} \leq x \leq x_i$, och $S(x_i) = y_i$ att

$$S_i(x_{i-1}) = S_{i-1}(x_{i-1}) = S(x_{i-1}) = y_{i-1} \quad (3.9)$$

för alla $i = 1, 2, \dots, n$. Vidare är

$$S_i(x_i) = S(x_i) = y_i \quad (3.10)$$

Från ansättningen av S_i och (3.8) följer att

$$S_i(x_{i-1}) = a_i + b_i \cdot 0 + c_i \cdot 0^2 + d_i \cdot 0^3 = a_i$$

och

$$S_i(x_i) = a_i + b_i \cdot 1 + c_i \cdot 1^2 + d_i \cdot 1^3 = a_i + b_i + c_i + d_i.$$

Därmed ger (3.9) och (3.10) att

$$a_i = y_{i-1} \quad \text{respektive} \quad a_i + b_i + c_i + d_i = y_i. \quad (3.11)$$

Villkoren för S med avseende på förstaderivatan ger att

$$S'_i(x_{i-1}) = S'_{i-1}(x_{i-1}) = S'(x_{i-1}) = k_{i-1} \quad \text{och} \quad S'_i(x_i) = S'(x_i) = k_i. \quad (3.12)$$

Vi betraktar u_i som en funktion av x och därmed ger kedjeregeln att

$$\frac{d}{dx} u_i^m = m u_i^{m-1} u'_i = \frac{m u_i^{m-1}}{h_i}, \quad m \in \mathbb{Z}_+,$$

där faktorn $1/h_i$ är den inre derivatan av u_i . Alltså är

$$S'_i(x) = \frac{b_i}{h_i} + \frac{2c_i}{h_i} u_i + \frac{3d_i}{h_i} u_i^2.$$

Från (3.8) följer det därmed att

$$S'_i(x_{i-1}) = \frac{b_i}{h_i} \quad \text{och} \quad S'_i(x_i) = \frac{b_i}{h_i} + \frac{2c_i}{h_i} + \frac{3d_i}{h_i}$$

och likheterna (3.12) är alltså ekvivalenta med

$$b_i = h_i k_{i-1} \quad \text{respektive} \quad b_i + 2c_i + 3d_i = h_i k_i. \quad (3.13)$$

Vi har härlett ett linjärt ekvationssystem med koefficienterna a_i, b_i, c_i och d_i som obekanta, nämligen enligt (3.11) och (3.13) är

$$\begin{cases} a_i &= y_{i-1} \\ a_i + b_i + c_i + d_i &= y_i \\ b_i &= h_i k_{i-1} \\ b_i + 2c_i + 3d_i &= h_i k_i. \end{cases}$$

Notera att för varje S_i hör ett sådant ekvationssystem och att vi inte vet värdet på k_i eller k_{i-1} . Ekvationssystemet har den entydiga lösningen

$$\begin{cases} a_i = y_{i-1} \\ b_i = h_i k_{i-1} \\ c_i = 3(y_i - y_{i-1}) - h_i(2k_{i-1} + k_i) \\ d_i = 2(y_{i-1} - y_i) + h_i(k_{i-1} + k_i). \end{cases} \quad (3.14)$$

Vi behöver bestämma k_i och k_{i-1} . Härnäst deriverar vi en gång till och får då att

$$S_i''(x) = \frac{2c_i}{h_i^2} + \frac{6d_i}{h_i^2} u_i.$$

Speciellt är

$$S_i''(x_i) = \frac{2c_i + 6d_i}{h_i^2} \quad \text{och} \quad S_{i+1}''(x_i) = \frac{2c_{i+1}}{h_{i+1}^2},$$

eftersom

$$u_{i+1} = \frac{x - x_i}{h_{i+1}} \quad \text{och} \quad h_{i+1} = x_{i+1} - x_i$$

för $S_{i+1}(x)$. Det betyder att ekvationen $S_i''(x_i) = S_{i+1}''(x_i)$ är ekvivalent med

$$\frac{2c_i + 6d_i}{h_i^2} = \frac{2c_{i+1}}{h_{i+1}^2},$$

för alla $i = 1, 2, \dots, n-1$. Från (3.14) följer att

$$\begin{aligned} \frac{1}{h_i^2} \left\{ 2(3(y_i - y_{i-1}) - h_i(2k_{i-1} + k_i)) - 6(2(y_i - y_{i-1}) + h_i(k_{i-1} + k_i)) \right\} \\ = \frac{1}{h_{i+1}^2} 2(3(y_{i+1} - y_i) - h_{i+1}(2k_i + k_{i+1})). \end{aligned}$$

eller ekvivalent

$$h_{i+1}k_{i-1} + 2(h_i + h_{i+1})k_i + h_i k_{i+1} = r_i$$

där

$$r_i = 3 \left(h_{i+1} \frac{y_i - y_{i-1}}{h_i} + h_i \frac{y_{i+1} - y_i}{h_{i+1}} \right),$$

då $i = 1, 2, \dots, n-1$. Vi har $n-1$ linjära ekvationer med k_0, k_1, \dots, k_n som obekanta. Det saknas två ekvationer för att för att vi ska ha ett kvadratiskt system. Dessa två erhålls med hjälp av en av randvillkoren. Låt oss studera dem en i taget. **Naturlig splinefunktion.** Från $S_1''(x_0) = 0$ och $S_n''(x_n) = 0$ tillsammans med

$$S_1''(x_0) = \frac{2c_1}{h_1^2} \quad \text{och} \quad S_n''(x_n) = \frac{2c_n + 6d_n}{h_n^2},$$

följer det att $c_1 = 0$ respektive $c_n + 3d_n = 0$. Från (3.14) erhåller vi

$$3(y_1 - y_0) - h_1(2k_0 + k_1) = 0$$

och

$$3(y_n - y_{n-1}) - h_n(2k_{n-1} + k_n) + 3(2(y_{n-1} - y_n) + h_n(k_{n-1} + k_n)) = 0,$$

eller ekvivalent

$$2k_0 + k_1 = r_0 \quad \text{respektive} \quad k_{n-1} + 2k_n = r_n$$

där

$$r_0 = 3 \frac{y_1 - y_0}{h_1} \quad \text{och} \quad r_n = 3 \frac{y_n - y_{n-1}}{h_n}.$$

Clamped boundary. Villkoren $S'_1(x_0) = \alpha$ och $S'_n(x_n) = \beta$ tillsammans med

$$S'_1(x_0) = \frac{b_1}{h_1} \quad \text{och} \quad S'_n(x_n) = \frac{b_n}{h_n} + \frac{2c_n}{h_n} + \frac{3d_n}{h_n},$$

ger att

$$b_1 = \alpha h_1 \quad \text{respektive} \quad b_n + 2c_n + 3d_n = \beta h_n.$$

Insättning från (3.14) i dessa två ekvationer medför att

$$h_1 k_0 = \alpha h_1$$

respektive

$$\begin{aligned} h_n k_{n-1} + 2(3(y_n - y_{n-1}) - h_n(2k_{n-1} + k_n)) \\ + 3(2(y_{n-1} - y_n) + h_n(k_{n-1} + k_n)) = \beta h_n. \end{aligned}$$

Det ger till slut att

$$k_0 = r_0 \quad \text{och} \quad k_n = r_n,$$

där

$$r_0 = \alpha \quad \text{och} \quad r_n = \beta.$$

Not-a-knot. Tredjederivatan av S_i är

$$S_i^{(3)}(x) = \frac{6d_i}{h_i^3},$$

dvs en konstant funktion. Villkoren $S_1^{(3)}(x_1) = S_2^{(3)}(x_1)$ och $S_{n-1}^{(3)}(x_{n-1}) = S_n^{(3)}(x_{n-1})$ är ekvivalenta med

$$\frac{6d_1}{h_1^3} = \frac{6d_2}{h_2^3} \quad \text{respektive} \quad \frac{6d_{n-1}}{h_{n-1}^3} = \frac{6d_n}{h_n^3}$$

Från (3.14) följer att första ekvationen kan skrivas om till

$$\begin{aligned} \frac{2(y_0 - y_1) + h_1(k_0 + k_1)}{h_1^3} &= \frac{2(y_1 - y_2) + h_2(k_1 + k_2)}{h_2^3} \\ &\Leftrightarrow \\ \frac{1}{h_1^2}(k_0 + k_1) - \frac{1}{h_2^2}(k_1 + k_2) &= \frac{2}{h_2^3}(y_1 - y_2) - \frac{2}{h_1^3}(y_0 - y_1) \\ &\Leftrightarrow \\ h_2^2 k_0 + (h_2^2 - h_1^2) k_1 - h_1^2 k_2 &= r_0 \end{aligned}$$

där

$$r_0 = 2 \left(\frac{h_1^2}{h_2} (y_1 - y_2) - \frac{h_2^2}{h_1} (y_0 - y_1) \right).$$

På samma sätt härledder man motsvarande formel för k_{n-2} , k_{n-1} och k_n .

Oavsett val av randvillkor erhåller vi två linjära ekvationer och därmed har vi ett linjärt ekvationssystem med $n + 1$ obekanta och lika många ekvationer. I tex fallet med naturlig splinefunktion har vi härlett det linjära ekvationssystemet

$$\mathbf{A}\mathbf{x} = \mathbf{b},$$

där \mathbf{A} är $(n + 1) \times (n + 1)$ -matrisen

$$\begin{pmatrix} 2 & 1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ h_2 & 2(h_1 + h_2) & h_1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & h_3 & 2(h_2 + h_3) & h_2 & \cdots & 0 & 0 & 0 \\ 0 & 0 & h_4 & 2(h_3 + h_4) & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 2(h_{n-2} + h_{n-1}) & h_{n-2} & 0 \\ 0 & 0 & 0 & 0 & \cdots & h_n & 2(h_{n-1} + h_n) & h_{n-1} \\ 0 & 0 & 0 & 0 & \cdots & 0 & 1 & 2 \end{pmatrix}$$

samt

$$\mathbf{x} = \begin{pmatrix} k_0 \\ k_1 \\ k_2 \\ \vdots \\ k_n \end{pmatrix} \quad \text{och} \quad \mathbf{b} = \begin{pmatrix} r_0 \\ r_1 \\ r_2 \\ \vdots \\ r_n \end{pmatrix}.$$

De två andra randvillkoren skiljer sig från ovanstående i första och sista raden. Genom att lösa detta ekvationssystemet bestämmer vi alla k_0, k_1, \dots, k_n , som i sin tur ger oss samtliga koefficienter a_i, b_i, c_i och d_i till de kubiska polynome $S_i(x)$, dvs

$$S_i(x) = a_i + b_i \frac{x - x_{i-1}}{h_i} + c_i \frac{(x - x_{i-1})^2}{h_i^2} + d_i \frac{(x - x_{i-1})^3}{h_i^3},$$

för $i = 1, 2, \dots, n$. Det ger oss splinefunktionen $S(x)$. Att denna är entydigt bestämd följer från det faktum att koefficienterna är lösningar till ett icke-singulärt linjärt ekvationssystem. \square

Exempel 3.10 (Naturlig splinefunktion). Låt

$$\begin{aligned} (x_0, y_0) &= (0.0, 0.5) & (x_1, y_1) &= (0.6, 1.3) \\ (x_2, y_2) &= (1.2, 0.1) & (x_3, y_3) &= (2.1, 0.8). \end{aligned}$$

Då är $n = 3$ och från $h_i = x_i - x_{i-1}$ följer det att

$$h_1 = 0.6, \quad h_2 = 0.6 \quad \text{och} \quad h_3 = 0.9.$$

Med randvillkoret för naturlig splinefunktion ges koefficientsmatrisen av

$$\mathbf{A} = \begin{pmatrix} 2 & 1 & 0 & 0 \\ h_2 & 2(h_1 + h_2) & h_1 & 0 \\ 0 & h_3 & 2(h_2 + h_3) & h_2 \\ 0 & 0 & 1 & 2 \end{pmatrix} = \begin{pmatrix} 2 & 1 & 0 & 0 \\ 0.6 & 2.4 & 0.6 & 0 \\ 0 & 0.9 & 3.0 & 0.6 \\ 0 & 0 & 1 & 2 \end{pmatrix}.$$

Låt $\mathbf{x} = (k_0, k_1, k_2, k_3)$. Vidare är

$$r_0 = 3 \frac{y_1 - y_0}{h_1} = 4.0$$

$$\begin{aligned}r_1 &= 3 \left(h_2 \frac{y_1 - y_0}{h_1} + h_1 \frac{y_0 - y_1}{h_2} \right) = 0.0 \\r_2 &= 3 \left(h_3 \frac{y_2 - y_1}{h_2} + h_2 \frac{y_3 - y_2}{h_3} \right) = -4.0 \\r_3 &= 3 \frac{y_3 - y_2}{h_1} = 3.5.\end{aligned}$$

Sätt $\mathbf{b} = (r_0, r_1, r_2, r_3) = (4.0, 0.0, -4, 0, 3.5)$. Systemet

$$\mathbf{Ax} = \mathbf{b} \Leftrightarrow \begin{pmatrix} 2 & 1 & 0 & 0 \\ 0.6 & 2.4 & 0.6 & 0 \\ 0 & 0.9 & 3.0 & 0.6 \\ 0 & 0 & 1 & 2 \end{pmatrix} \begin{pmatrix} k_0 \\ k_1 \\ k_2 \\ k_3 \end{pmatrix} = \begin{pmatrix} 4.0 \\ 0.0 \\ -4,0 \\ 3.5 \end{pmatrix}$$

har lösningen $x \approx (2.02, -0.04, -1.86, 2.68)$. Koefficienterna till S_1 är

$$\begin{cases} a_1 = y_0 = 0.5 \\ b_1 = h_1 k_0 \approx 1.21 \\ c_1 = 3(y_1 - y_0) - h_1(2k_0 + k_1) \approx 0.0 \\ d_1 = 2(y_0 - y_1) + h_1(k_0 + k_1) \approx -0.41. \end{cases}$$

Eftersom $x_0 = 0.0$ och $h_1 = 0.6$ har vi att

$$\begin{aligned}S_1(x) &= a_1 + b_1 \frac{x - x_0}{h_1} + c_1 \frac{(x - x_0)^2}{h_1^2} + d_1 \frac{(x - x_0)^3}{h_1^3} \\&\approx 0.5 + 2.020x - 1.909x^3, \quad 0 \leq x \leq 0.6.\end{aligned}$$

De andra två polynome bestäms på liknande sätt och ges av

$$\begin{aligned}S_2(x) &\approx -2.374 + 14.389x - 17.281x^2 + 5.840x^3 & 0.6 \leq x \leq 1.2 \\S_3(x) &\approx 9.283 - 14.753x + 7.005x^2 - 0.906x^3 & 1.2 \leq x \leq 2.1.\end{aligned}$$

Splinefunktionen i sin helhet ges av

$$S(x) = \begin{cases} S_1(x) & \text{om } 0.0 \leq x \leq 0.6 \\ S_2(x) & \text{om } 0.6 \leq x \leq 1.2 \\ S_3(x) & \text{om } 1.2 \leq x \leq 2.1, \end{cases}$$

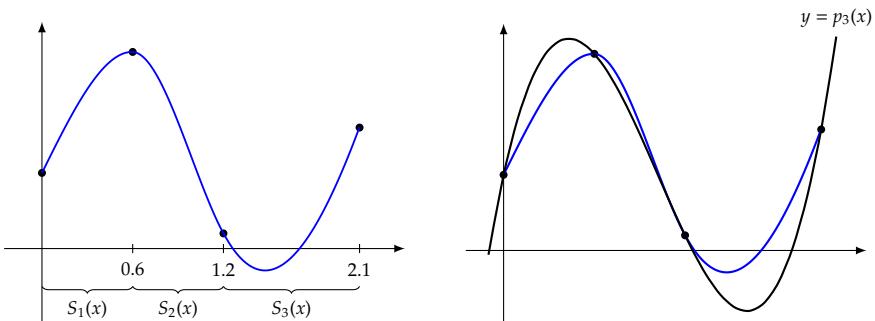
se vänstra bilden i figur 3.10. Se även figur 3.11 för en illustration av hur S är uppbyggd av de tre tredjegradspolynomen S_1 , S_2 och S_3 . Eftersom vi har fyra punkter kan vi istället bestämma ett interpolationspolynom p_3 av grad 3, nämligen

$$p_3(x) = 2.20459x^3 - 6.74603x^2 + 4.5873x + 0.5.$$

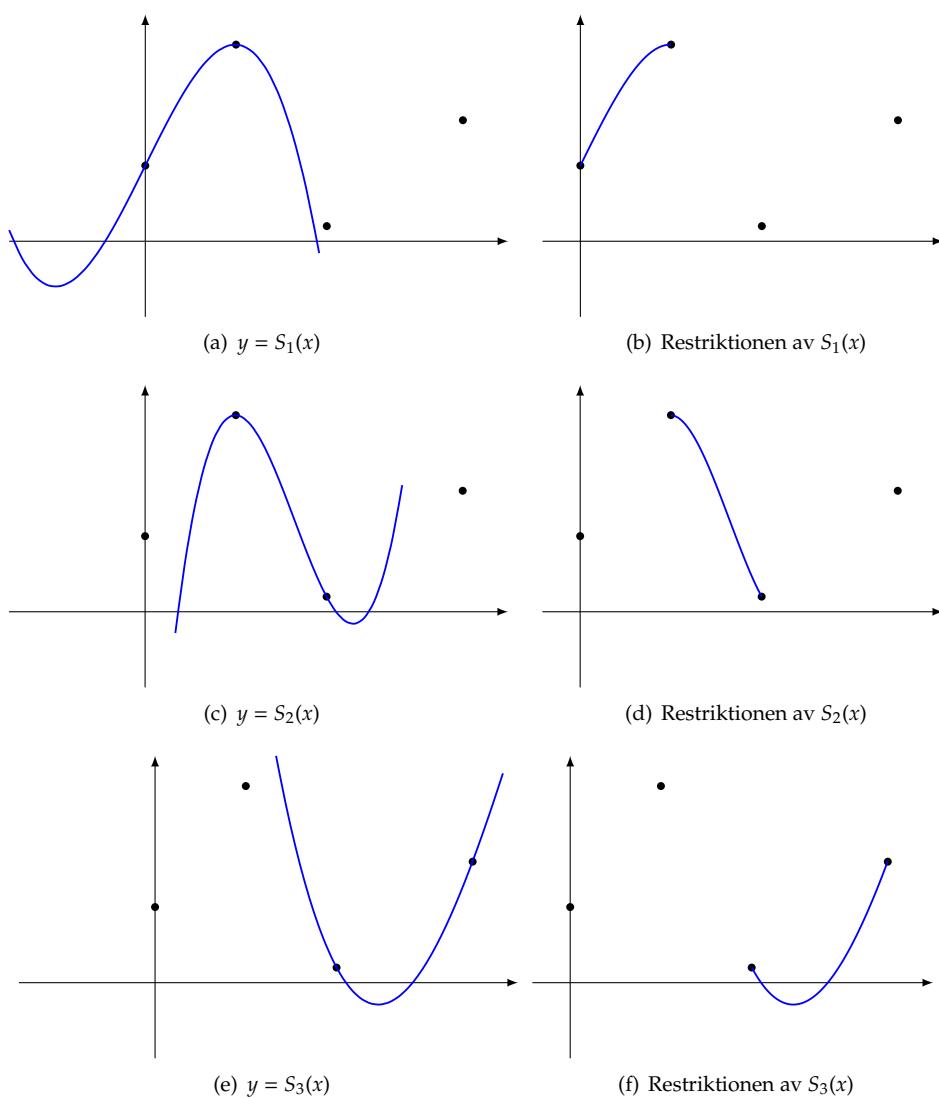
Detaljerna lämnas som övning. Splinefunktionen S ger en helt annan interpolation av noderna än p_3 , se högra bilden i figur 3.10. \diamond

Exempel 3.11 (Clamped boundary). Vi utgår från samma noder som i exempel 3.10. Koefficientsmatrisen ges av

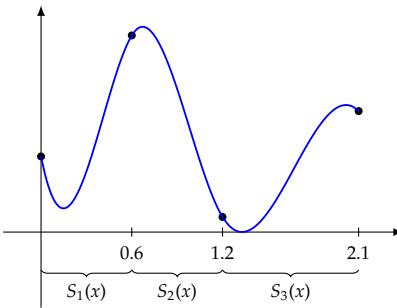
$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ h_2 & 2(h_1 + h_2) & h_1 & 0 \\ 0 & h_3 & 2(h_2 + h_3) & h_2 \\ 0 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0.6 & 2.4 & 0.6 & 0 \\ 0 & 0.9 & 3.0 & 0.6 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$



Figur 3.10



Figur 3.11



Figur 3.12

Antag att vi vill att $S'(x_0) = S'(0.0) = \alpha = -5$ och $S'(x_3) = S'(2.1) = \beta = -1$. Högerledet av ekvationssystemet då ges av vektor

$$\begin{aligned} r_0 &= \alpha = -5 \\ r_1 &= 3\left(h_2 \frac{y_1 - y_0}{h_1} + h_1 \frac{y_0 - y_1}{h_2}\right) = 0.0 \\ r_2 &= 3\left(h_3 \frac{y_2 - y_1}{h_2} + h_2 \frac{y_3 - y_2}{h_3}\right) = -4.0 \\ r_3 &= \beta = -1. \end{aligned}$$

Sätt $\mathbf{b} = (r_0, r_1, r_2, r_3) = (-5, 0.0, -4, 0, -1)$. Systemet

$$\mathbf{Ax} = \mathbf{b} \Leftrightarrow \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0.6 & 2.4 & 0.6 & 0 \\ 0 & 0.9 & 3.0 & 0.6 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} k_0 \\ k_1 \\ k_2 \\ k_3 \end{pmatrix} = \begin{pmatrix} -5.0 \\ 0.0 \\ -4.0 \\ -1.0 \end{pmatrix}$$

har lösningen $\mathbf{x} \approx (-5, 1.65766, -1.63063, -1)$. Koefficienterna till S_1 är

$$\begin{cases} a_1 = y_0 = 0.5 \\ b_1 = h_1 k_0 \approx -3 \\ c_1 = 3(y_1 - y_0) - h_1(2k_0 + k_1) \approx 7.40541 \\ d_1 = 2(y_0 - y_1) + h_1(k_0 + k_1) \approx -3.60541. \end{cases}$$

Eftersom $x_0 = 0.0$ och $h_1 = 0.6$ har vi att

$$\begin{aligned} S_1(x) &= a_1 + b_1 \frac{x - x_0}{h_1} + c_1 \frac{(x - x_0)^2}{h_1^2} + d_1 \frac{(x - x_0)^3}{h_1^3} \\ &\approx 0.5 - 5x + 20.5706x^2 - 16.6917x^3, \quad 0 \leq x \leq 0.6. \end{aligned}$$

De andra två polynome bestäms på liknande sätt och ges av

$$\begin{aligned} S_2(x) &\approx -6.72162 + 29.1081x - 32.9429x^2 + 11.1862x^3 & 0.6 \leq x \leq 1.2 \\ S_3(x) &\approx 21.5386 - 41.5425x + 25.9326x^2 - 5.16813x^3 & 1.2 \leq x \leq 2.1, \end{aligned}$$

se figur 3.12. Notera lutningen i ändpunktarna x_0 och x_3 . ◊

Exempel 3.12 (Not-a-knot). Vi utgår från samma noder som i exempel 3.10. Koefficientsmatrisen ges av

$$\mathbf{A} = \begin{pmatrix} h_2^2 & h_2^2 - h_1^2 & -h_1^2 & 0 \\ h_2 & 2(h_1 + h_2) & h_1 & 0 \\ 0 & h_3 & 2(h_2 + h_3) & h_2 \\ 0 & h_3^2 & h_3^2 - h_2^2 & -h_2^2 \end{pmatrix} = \begin{pmatrix} 0.36 & 0 & -0.36 & 0 \\ 0.6 & 2.4 & 0.6 & 0 \\ 0 & 0.9 & 3.0 & 0.6 \\ 0 & 0.81 & 0.45 & -0.36 \end{pmatrix}.$$

Låt $\mathbf{x} = (k_0, k_1, k_2, k_3)$. Vidare är

$$\begin{aligned} r_0 &= 2\left(h_1^2 \frac{y_1 - y_2}{h_2} - h_2^2 \frac{y_0 - y_1}{h_1}\right) = 2.4 \\ r_1 &= 3\left(h_2 \frac{y_1 - y_0}{h_1} + h_1 \frac{y_0 - y_1}{h_2}\right) = 0.0 \\ r_2 &= 3\left(h_3 \frac{y_2 - y_1}{h_2} + h_2 \frac{y_3 - y_2}{h_3}\right) = -4.0 \\ r_3 &= 2\left(h_2^2 \frac{y_2 - y_3}{h_3} - h_3^2 \frac{y_1 - y_2}{h_2}\right) = -3.8. \end{aligned}$$

Sätt $\mathbf{b} = (r_0, r_1, r_2, r_3) = (2.4, 0.0, -4.0, -3.8)$. Systemet

$$\mathbf{Ax} = \mathbf{b} \Leftrightarrow \begin{pmatrix} 0.36 & 0 & -0.36 & 0 \\ 0.6 & 2.4 & 0.6 & 0 \\ 0 & 0.9 & 3.0 & 0.6 \\ 0 & 0.81 & 0.45 & -0.36 \end{pmatrix} \begin{pmatrix} k_0 \\ k_1 \\ k_2 \\ k_3 \end{pmatrix} = \begin{pmatrix} 2.4 \\ 0.0 \\ -4.0 \\ -3.8 \end{pmatrix}$$

har lösningen $\mathbf{x} \approx (4.15873, -0.412698, -2.50794, 6.49206)$. Koefficienterna till S_1 är

$$\begin{cases} a_1 = y_0 = 0.5 \\ b_1 = h_1 k_0 \approx 2.49524 \\ c_1 = 3(y_1 - y_0) - h_1(2k_0 + k_1) \approx -2.34286 \\ d_1 = 2(y_0 - y_1) + h_1(k_0 + k_1) \approx 0.647619. \end{cases}$$

Eftersom $x_0 = 0.0$ och $h_1 = 0.6$ har vi att

$$\begin{aligned} S_1(x) &= a_1 + b_1 \frac{x - x_0}{h_1} + c_1 \frac{(x - x_0)^2}{h_1^2} + d_1 \frac{(x - x_0)^3}{h_1^3} \\ &\approx 0.5 + 4.15873x - 6.50794x^2 + 2.99824x^3, \quad 0 \leq x \leq 0.6. \end{aligned}$$

De andra två polynome bestäms på liknande sätt och ges av

$$\begin{aligned} S_2(x) &\approx -0.7 + 8.15873x - 9.84127x^2 + 2.99824x^3 & 0.6 \leq x \leq 1.2 \\ S_3(x) &\approx -0.7 + 8.15873x - 9.84127x^2 + 2.99824x^3 & 1.2 \leq x \leq 2.1. \end{aligned}$$

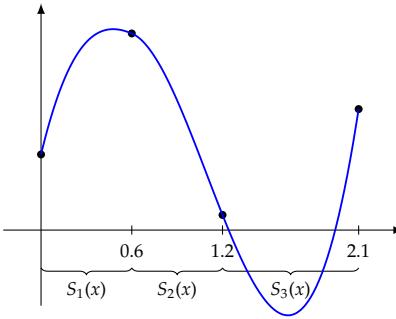
se figur 3.13. ◊

3.8 Bézierkurvor

20160419

Vi inleder med en kort introduktion till vektoranalys i planet. Antag att $x, y: \mathbb{R} \rightarrow \mathbb{R}$ är differentierbara funktioner. Låt $v: \mathbb{R} \rightarrow \mathbb{R}^2$ enligt

$$v(t) = (x(t), y(t)).$$



Figur 3.13

Då bildar punkterna (a, b) i mängden

$$\{v(t) : t \in \mathbb{R}\}$$

en kontinuerlig kurva C i planet \mathbb{R}^2 . Derivatan v' av v definieras som

$$v'(t) = (x'(t), y'(t)).$$

Låt $a \in \mathbb{R}$. Man kan tolka vektorn $v'(a)$ geometriskt som den vektor som är tangent till kurvan C i punkten $v(a)$ och vars längd är lika med storleken på förändringen av $v(t)$ då $t = a$. Riktningen hos $v'(a)$ motsvarar rörelsen då vi följer kurvan C för växande t .

Exempel 3.13. Låt $v(t) = (2t^3 e^{-t^2}, t e^{-t^2})$. Då är tex

$$v(-1) \approx (-0.7358, -0.3679) \quad \text{och} \quad v(1.2) \approx (0.8188, 0.2843).$$

För varje värde på t motsvarar $v(t)$ en vektor, se tex de fyra ortsvektorna i figur 3.14. Genom att variera t erhåller vi de punkter $(x(t), y(t))$ som utgör kurvan C . I figur 3.14 motsvarar den blå kurvan C för $v(t)$ då $-2 \leq t \leq 2$. Vi har att $v(t) \rightarrow (0, 0)$ då $t \rightarrow \pm\infty$. Derivering komponentvis ger att

$$v'(t) = e^{-t^2}(6t^2 - 4t^4, 1 - 2t^2).$$

Notera att även $v'(t)$ är en vektorvärd funktion, dvs för olika värden på t är $v'(t)$ en vektor. Denna vektor tangerar kurvan i punkten $v(t)$, se figur 3.14. Vidare pekar vektorn $v'(t)$ i den riktning som $v(t)$ följer kurvan för växande t och dess längd $\|v'(t)\|$ anger hur fort vi färdas just då (i många tillämpningar tolkas t som en tidsvariabel). ◇

Exempel 3.14. Grafen till funktionen

$$v(t) = (t \cos(3t), t \sin(3t))$$

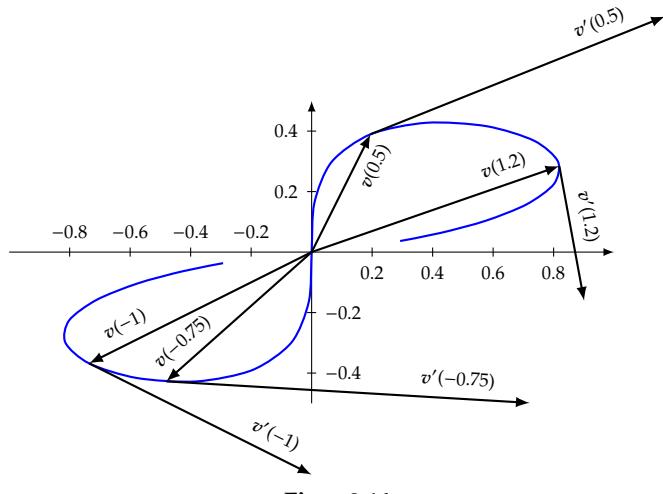
är en spiral. I den vänstra bilden i figur 3.15 illustrerar C då $0 \leq t \leq 15$. Låt $F: \mathbb{R}^2 \rightarrow \mathbb{R}^2$ vara den linjära avbildningen vars avbildningsmatris i standardbasen ges av

$$A = \begin{pmatrix} 2 & -3 \\ 4 & 1 \end{pmatrix}.$$

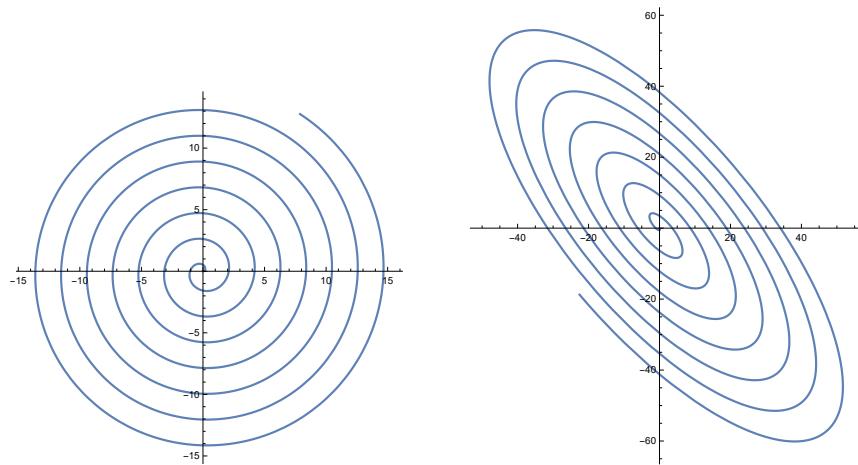
Då är

$$F(v(t)) = Av(t) = (2t \cos(3t) - 3t \sin(3t), -4t \cos(3t) + t \sin(3t))$$

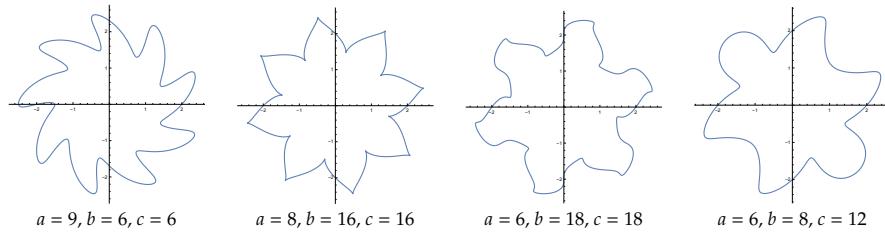
också en vektorvärd funktion, se den högra bilden i figur 3.15. ◇



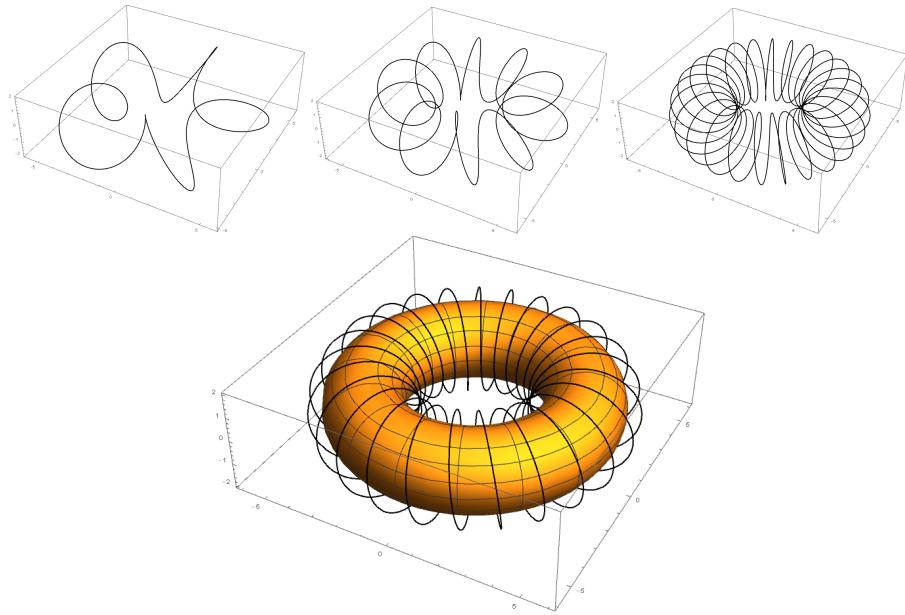
Figur 3.14



Figur 3.15



Figur 3.16



Figur 3.17

Exempel 3.15. Kurvan till en vektorvärd funktion kan vara sluten. Låt

$$v(t) = \left(2 + \frac{1}{2} \sin(at)\right) \left(\cos\left(t + \frac{1}{c} \sin(bt)\right), \sin\left(t + \frac{1}{c} \sin(bt)\right)\right).$$

Med olika val av värden på parametrarna a , b och c får kurvan olika form, se figur 3.16. Redan då $0 \leq t \leq 2\pi$ fås en sluten kurva. Utökar vi intervallet för t ger det ytterligare varv av samma kurva. \diamond

Exempel 3.16. Det går att definiera vektorvärda funktioner i tre dimensioner. I tex linjär algebra beskriver man linjer och plan på parameterform. Funktionen

$$S(t) = (a + b \sin(ct)) \cos(t), (a + b \sin(ct)) \sin(t), b \cos(ct)$$

får man spole runt en torus, se figur 3.17. Ytan till en torus kan i sin tur beskrivas med hjälp av funktionen

$$T(s, t) = ((a + b \sin(s)) \cos(t), (a + b \sin(s)) \sin(t), b \cos(s)).$$

Radien på torusen bestäms av a och radien på "röret" bestäms av b . Med c styr man hur många varv spolen ska göra kring "röret". \diamond

Låt n vara ett positivt heltal och P_0, P_1, \dots, P_n punkter i planet \mathbb{R}^2 , sk *kontrollpunkter*. Funktionen $b_n: [0, 1] \rightarrow \mathbb{R}^2$ som definieras enligt

$$b_n(t) = \sum_{i=0}^n B_{i,n}(t)P_i,$$

kallas för en *Bézierkurva* av grad n och där $B_{i,n}$ är det i :te Bernsteinpolynomet av grad n , som definieras som

$$B_{i,n}(t) = \binom{n}{i} t^i (1-t)^{n-i} = \frac{n!}{i!(n-i)!} t^i (1-t)^{n-i},$$

där $i = 0, 1, \dots, n$. Vi har tex att Bernsteinpolynomen av grad 1 är

$$B_{0,1}(t) = 1 - t \quad \text{och} \quad B_{1,1}(t) = t,$$

se första bilden i figur 3.18. Bernsteinpolynom av grad 2 ges av

$$B_{0,2}(t) = (1-t)^2, \quad B_{1,2}(t) = 2t(1-t) \quad \text{och} \quad B_{2,2}(t) = t^2$$

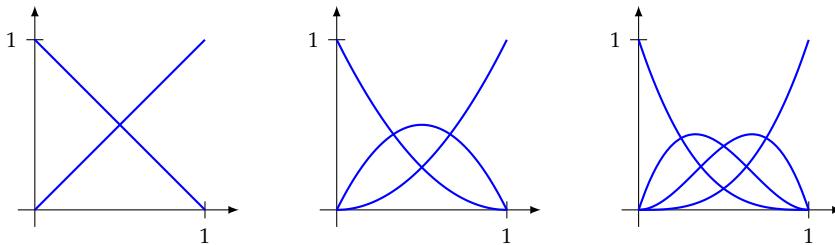
se andra bilden i figur 3.18. Bernsteinpolynom av grad 3 är

$$B_{0,3}(t) = (1-t)^3, \quad B_{1,3}(t) = 3t(1-t)^2, \quad B_{2,3}(t) = 3t^2(1-t) \quad \text{och} \quad B_{3,3}(t) = t^3,$$

se tredje bilden i figur 3.18. Om $n = 3$, så är

$$b_3(t) = P_0 B_{0,3}(t) + P_1 B_{1,3}(t) + P_2 B_{2,3}(t) + P_3 B_{3,3}(t).$$

Med andra ord bestämmer graden n hur många kontrollpunkter som behövs för att definiera b_n och $b_n(t)$ är en vektor med två polynom av grad n eller lägre.



Figur 3.18

Exempel 3.17. Givet kontrollpunkterna

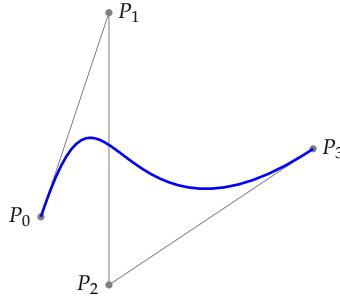
$$P_0 = (1, 0), \quad P_1 = (2, 3), \quad P_2 = (2, -1) \quad \text{och} \quad P_3 = (5, 1).$$

Bestäm $b_3(t)$.

◊

Lösning. Vi har att

$$\begin{aligned} b_3(t) &= \sum_{i=0}^3 P_i B_{i,3}(t) = P_0 B_{0,3}(t) + P_1 B_{1,3}(t) + P_2 B_{2,3}(t) + P_3 B_{3,3}(t) \\ &= (1, 0)(1-t)^3 + (2, 3)3t(1-t)^2 + (2, -1)3t^2(1-t) + (5, 1)t^3 \end{aligned}$$



Figur 3.19

$$\begin{aligned}
 &= (1 - 3t + 3t^2 - t^3, 0) + (6t - 12t^2 + 6t^3, 9t - 18t^2 + 9t^3) \\
 &\quad + (6t^2 - 6t^3, -3t^2 + 3t^3) + (5t^3, t^3) \\
 &= (1 + 3t - 3t^2 + 4t^3, 9t - 21t^2 + 13t^3).
 \end{aligned}$$

I figur 3.19 ser vi kurvan $(x, y) = b_3(t)$, där $0 \leq t \leq 1$. \square

Exempel 3.18. I figur 3.20 ser vi flera olika exempel på Bézierkurvor. Det lämnas som övning att bestämma respektive funktion b_3 , se övningsuppgift 18. \diamond

Lemma 3.5. Låt n vara ett positivt heltal och k ett heltal sådant att $0 \leq k \leq n$.

$$(a) \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

$$(b) (a+b)^n = \sum_{k=0}^n \binom{n}{k} a^k b^{n-k}$$

$$(c) \binom{n}{0} = \binom{n}{n} = 1$$

$$(d) \binom{n}{k} = \binom{n}{n-k}$$

$$(e) \binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$$

$$(f) k \binom{n}{k} = n \binom{n-1}{k-1} \text{ och } (n-k) \binom{n}{k} = n \binom{n-1}{k}.$$

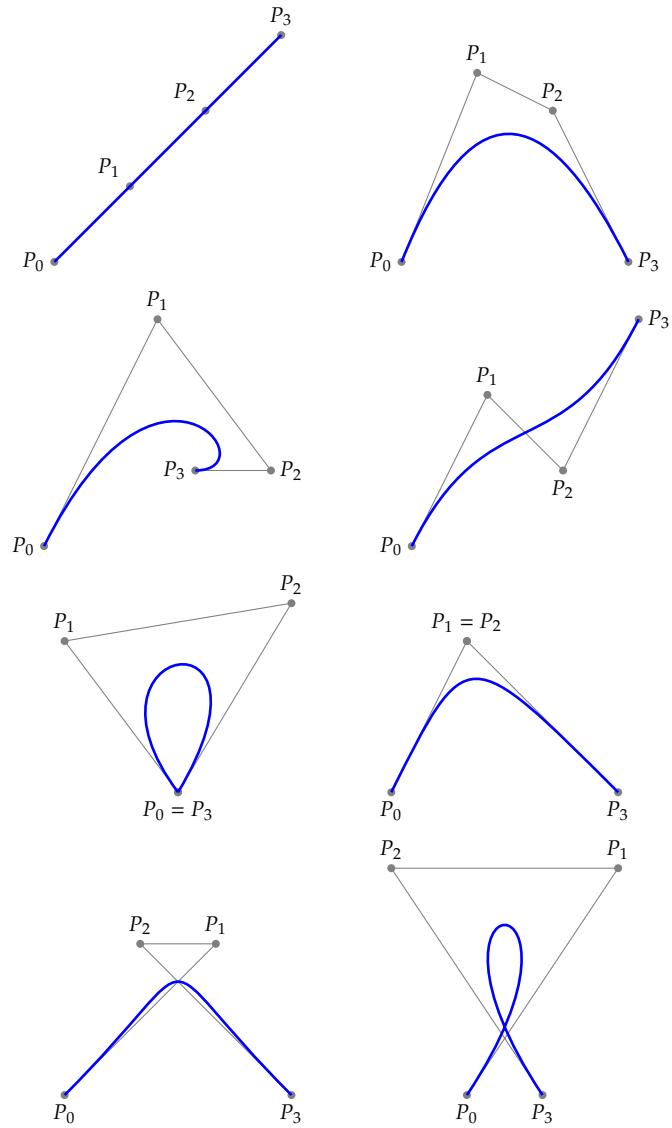
Bevis. Se godtycklig bok om diskret matematik. \square

Sats 3.6. Bernsteinpolynomen har följande egenskaper.

$$(a) B_{0,n}(t) = (1-t)^n \text{ och } B_{n,n}(t) = t^n.$$

$$(b) Sätt B_{0,0}(t) = 1 \text{ och } B_{i,n}(t) = 0 \text{ för } i < 0 \text{ och } i > n. Då är$$

$$B_{i,n}(t) = (1-t)B_{i,n-1}(t) + tB_{i-1,n-1}(t), \quad i = 1, 2, \dots, n-1.$$



Figur 3.20

- (c) Om $0 \leq t \leq 1$, så är $B_{i,n}(t) \geq 0$.
- (d) $B_{0,n}(0) = 1$ och $B_{i,n}(0) = 0$ om $i \neq 0$.
- (e) $B_{n,n}(1) = 1$ och $B_{i,n}(1) = 0$ om $i \neq n$.
- (f) $B_{0,n}(t) + B_{1,n}(t) + \cdots + B_{n,n}(t) = 1$ för alla t .
- (g) $B_{i,n}(t) = B_{n-i,n}(1-t)$ för alla $i = 0, 1, \dots, n$.
- (h) $B'_{i,n}(t) = n(B_{i-1,n-1}(t) - B_{i,n-1}(t))$.
- (i) Bernsteins polynomen av grad n bildar en bas för det vektorrum som består av alla polynom av grad mindre än eller lika med n .

Bevis. (a) Det följer direkt från definitionen av $B_{i,n}$, nämligen

$$B_{0,n}(t) = \binom{n}{0} t^0 (1-t)^{n-0} = (1-t)^n$$

och

$$B_{n,n}(t) = \binom{n}{n} t^n (1-t)^{n-n} = t^n.$$

(b) Om $i = 0$, så är $B_{i-1,n}(t) = B_{-1,n}(t) = 0$ och

$$(1-t)B_{0,n-1}(t) + tB_{-1,n-1}(t) = (1-t)(1-t)^{n-1} = (1-t)^n = B_{0,n}(t),$$

och om $i = n$, så är $B_{n,n-1}(t) = 0$ och

$$(1-t)B_{n,n-1}(t) + tB_{n-1,n-1}(t) = t(1-t)^{n-1} = t^n = B_{n,n}(t).$$

Låt nu $i = 1, 2, \dots, n-1$. Då är

$$\begin{aligned} (1-t)B_{i,n-1}(t) + tB_{i-1,n-1}(t) &= (1-t)\binom{n-1}{i} t^i (1-t)^{n-1-i} + t\binom{n-1}{i-1} t^{i-1} (1-t)^{n-1-(i-1)} \\ &= \left\{ \binom{n-1}{i} + \binom{n-1}{i-1} \right\} t^i (1-t)^{n-i} \\ &= \binom{n}{i} t^i (1-t)^{n-i} = B_{i,n}(t), \end{aligned}$$

enligt lemma 3.5. (c) Påståendet följer eftersom binomialkoefficienterna är positiva heltal och att både t^i och $(1-t)^{n-i}$ är icke-negativa då $0 \leq t \leq 1$. (d) Från (a) följer det att $B_{0,n}(0) = (1-0)^n = 1$. Om $i \neq 0$, så är

$$B_{i,n}(0) = \binom{n}{0} 0^i (1-t)^{n-i} = 0.$$

(e) Från (a) har vi att $B_{n,n}(1) = 1^n = 1$. Om $i \neq n$, så är

$$B_{i,n}(1) = \binom{n}{n} 1^i (1-1)^{n-i} = 0.$$

(f) Från binomialsatsen, se (b) i lemma 3.5, följer det att

$$\sum_{i=0}^n B_{i,n}(t) = \sum_{i=0}^n \binom{n}{i} t^i (1-t)^{n-i} = (t + (1-t))^n = 1^n = 1.$$

(g) Vi har att

$$B_{n-i,n}(t) = \binom{n}{n-i} t^{n-i} (1-t)^{n-(n-i)} = \binom{n}{n-i} t^{n-i} (1-t)^i.$$

Alltså är

$$B_{n-i,n}(1-t) = \binom{n}{n-i} (1-t)^{n-i} (1-(1-t))^i = \binom{n}{i} t^i (1-t)^{n-i} = B_{i,n}(t).$$

(h) Från (f) i lemma 3.5 följer det att

$$\begin{aligned} B'_{i,n}(t) &= \frac{d}{dt} \binom{n}{i} t^i (1-t)^{n-i} \\ &= \binom{n}{i} (it^{i-1}(1-t)^{n-i} - (n-i)t^i(1-t)^{n-i-1}) \\ &= n \binom{n-1}{i-1} t^{i-1} (1-t)^{(n-1)-(i-1)} - n \binom{n-1}{i} t^i (1-t)^{n-1-i} \\ &= n B_{i-1,n-1}(t) - n B_{i,n-1}(t). \end{aligned}$$

(i) Antag att de reella talen $\lambda_0, \lambda_1, \dots, \lambda_n$ uppfyller

$$\lambda_0 B_{0,n}(t) + \lambda_1 B_{1,n}(t) + \dots + \lambda_n B_{n,n}(t) = 0.$$

Genom att identifiera koefficienter till polynomet i vänsterledet med nollpolynomet erhåller vi ett linjärt ekvationssystem med avseende på $\lambda_0, \lambda_1, \dots, \lambda_n$. Ekvationsystemet är triangulärt eftersom termen av lägst grad i $B_{i,n}(t)$ är av grad i . Därmed följer det att $\lambda_0 = \lambda_1 = \dots = \lambda_n = 0$ är den enda lösningen och det bevisar att Bernsteinpolynomen av grad n är linjärt oberoende. Vektorrummet av alla polynom av grad mindre än eller lika med n har dimensionen $n+1$. Alltså måste Bernsteinpolynomen bilda en bas för detta vektorrum. \square

Sats 3.7. Låt v vara en vektor och låt $P_i = P_0 + s_i v$ för $s_i \in \mathbb{R}$, dvs kontrollpunktarna ligger på en rät linje. Då är Bézierkurvan $b_n(t)$ en del av denna linje.

Bevis. Vi har att

$$\begin{aligned} b_n(t) &= \sum_{i=0}^n B_{i,n}(t) P_i = \sum_{i=0}^n B_{i,n}(t) (P_0 + s_i v) \\ &= P_0 \sum_{i=0}^n B_{i,n}(t) + \sum_{i=0}^n s_i B_{i,n}(t) v. \end{aligned}$$

Låt

$$f(t) = \sum_{i=0}^n s_i B_{i,n}(t).$$

Då är

$$b_n(t) = P_0 + f(t)v,$$

eftersom $\sum_{i=0}^n B_{i,n}(t) = 1$. Det bevisar att $b_n(t)$ ligger på samma linje som kontrollpunkterna, för alla t . \square

Låt P_0, P_1, \dots, P_n vara kontrollpunkter för en Bézierkurva $\mathbf{b}_n(t)$. För $i = 0, 1, \dots, n$ lår $Q_i = P_{n-i}$ vara kontrollpunkterna till Bézierkurvan $c_n(t)$. Med andra ord använder vi samma kontrollpunkter, men in omvänt ordning. Motsvarar \mathbf{b}_n och c_n samma kurva? Ja! Vi har nämligen att om $0 \leq t \leq 1$, så är $\mathbf{b}_n(t) = c_n(1-t)$, vilket följer från

$$\begin{aligned}\mathbf{b}_n(t) &= \sum_{i=0}^n B_{i,n}(t)P_i = \sum_{i=0}^n B_{n-i,n}(1-t)P_i \\ &= \sum_{i=0}^n B_{i,n}(1-t)P_{n-i} = \sum_{i=0}^n B_{i,n}(1-t)Q_i = c_n(1-t).\end{aligned}$$

Däremot får vi olika kurvor om vi kastar om kontrollpunkterna i en annan ordning. Antag att P_a, P_b, P_c och P_d är kontrollpunkterna för en Bézierkurva av ordning 3. Det finns $4!/2 = 12$ olika Bézierkurvor för fyra valda punkter, eftersom vi visade ovan att uppsättningarna P_a, P_b, P_c, P_d och P_d, P_c, P_b, P_a motsvarar samma kurva, se figur 3.21.

Sats 3.8. *De två kontrollpunkterna P_0 och P_n är ändpunkter till motsvarande Bézierkurv \mathbf{b}_n , dvs $\mathbf{b}_n(0) = P_0$ och $\mathbf{b}_n(1) = P_n$.*

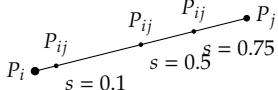
Bevis. Vi har att

$$\mathbf{b}_n(0) = \sum_{i=0}^n B_{i,n}(0)P_i = P_0,$$

eftersom $B_{0,n}(0) = 1$ och $B_{i,n}(0) = 0$ då $i \neq 0$. Den andra likheten visas analogt. \square

Låt s vara ett reellt tal sådant att $0 \leq s \leq 1$. Definiera P_{ij} som den mellanliggande punkt på linjesegmentet mellan P_i och P_j som uppfyller that

$$s = \frac{|P_{ij} - P_i|}{|P_j - P_i|},$$



dvs P_{ij} ligger med förhållandet $s:1$ från P_i på sträckan P_iP_j . Bilden i marginalen visar positionen hos P_{ij} för olika värden på s . För varje s erhåller vi tre punkter, P_{01} , P_{12} och P_{23} . Förbind punkterna P_{01} och P_{12} samt P_{12} och P_{23} med linjesegment. Bestäm på samma sätt punkterna P_{012} och P_{123} på respektive linjesegment, dvs P_{012} ligger med förhållandet $s:1$ från punkten P_{01} på sträckan $P_{01}P_{12}$, och P_{123} ligger med förhållandet $s:1$ från P_{12} på sträckan $P_{12}P_{23}$. Slutligen förbind punkterna P_{012} och P_{123} samt låt P_{0123} vara den punkt som ligger med förhållandet $s:1$ från P_{012} på sträckan $P_{012}P_{123}$, se figur 3.22. Då är

$$P_{0123} = \mathbf{b}_3(s).$$

För att bevisa detta konstaterar vi först att

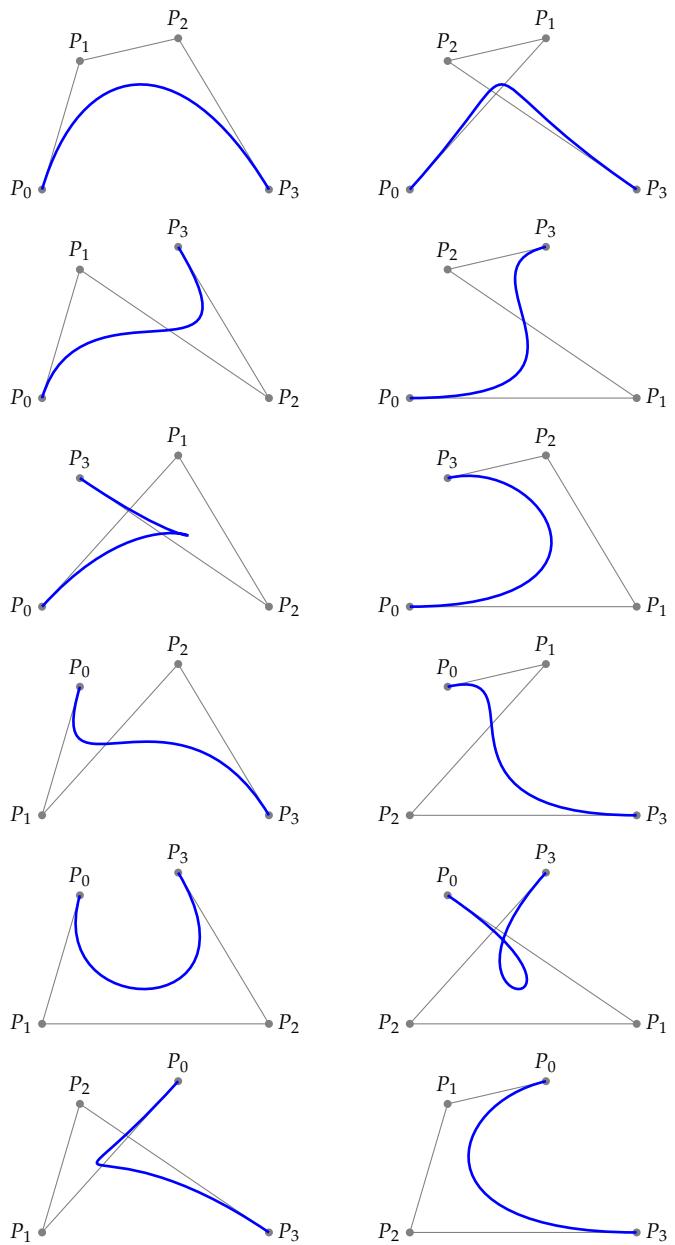
$$P_{ij} = s(P_j - P_i) + P_i \quad \text{och} \quad P_{ijk} = s(P_{jk} - P_{ij}) + P_{ij}.$$

Därmed följer det att

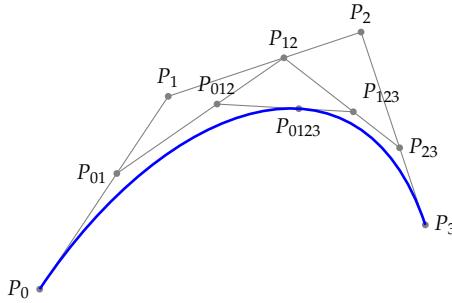
$$P_{ijk} = s^2 P_k - 2s^2 P_j + s^2 P_i + 2s P_j - 2s P_i + P_i,$$

och till slut att

$$\begin{aligned}P_{0123} &= s(P_{123} - P_{012}) + P_{012} \\ &= (1-s)^3 P_0 + 3(1-s)^2 s P_1 + 3(1-s) s^2 P_2 + s^3 P_3\end{aligned}$$



Figur 3.21



Figur 3.22

$$= \sum_{i=0}^3 B_{0,i}(s)P_i = \mathbf{b}_3(s),$$

vilket skulle visas.

Sats 3.9. Samtliga derivator av \mathbf{b}_n är kontinuerliga, dvs $\mathbf{b}_n \in C^\infty[0, 1]$.

Bevis. Låt $P_i = (x_i, y_i) \in \mathbb{R}^2$, där $i = 0, 1, \dots, n$, vara kontrollpunkterna till $\mathbf{b}_n(t)$. Då gäller att $\mathbf{b}_n(t) = (x(t), y(t))$ där

$$x(t) = \sum_{i=0}^n B_{i,n}(t)x_i \quad \text{och} \quad y(t) = \sum_{i=0}^n B_{i,n}(t)y_i,$$

dvs $\mathbf{b}_n(t)$ är en vektorvärd funktion med ett polynom av grad n eller läggre i varje komponent. Därför är \mathbf{b}_n oändligt deriverbar och \mathbf{b}_n samt alla dess derivator är kontinuerlig. \square

Sats 3.10. Låt \mathbf{b}_n vara en Bézierkurva. Då är $\mathbf{b}'_n(0) = n(P_1 - P_0)$ och $\mathbf{b}'_n(1) = n(P_n - P_{n-1})$.

Bevis. Förstaderivatan av $\mathbf{b}_n(t)$ ges av

$$\mathbf{b}'_n(t) = (x'(t), y'(t)),$$

vilken också kan skrivas på formen

$$\mathbf{b}'_n(t) = \frac{d}{dt} \sum_{i=0}^n B_{i,n}(t)P_i = \sum_{i=0}^n B'_{i,n}(t)P_i = n \sum_{i=0}^n (B_{i-1,n-1}(t) - B_{i,n-1}(t))P_i,$$

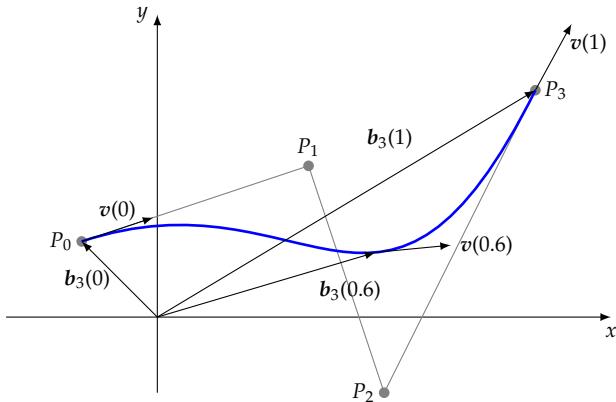
enligt sats 3.6. Speciellt är

$$\mathbf{b}'_n(0) = n \sum_{i=0}^n (B_{i-1,n-1}(0) - B_{i,n-1}(0))P_i = n(P_1 - P_0),$$

eftersin $B_{i,n}(0) = 1$ om $i = 0$ och annars är $B_{i,n}(0) = 0$. På samma sätt finner vi att

$$\mathbf{b}'_n(1) = n(P_n - P_{n-1}).$$

Alltså är vektorn $\mathbf{b}'_n(0)$ parallell med linjen som går genom punkterna P_0 och P_1 , och vektorn $\mathbf{b}'_n(1)$ är parallell med linjen som går genom punkterna P_{n-1} och P_n . \square



Figur 3.23

Anmärkning. Generellt gäller att vektorn $b'_n(t)$ är parallell med den linje som tangerar kurvan $(x, y) = b_n(t)$ i punkten $b_n(t)$. Det betyder att kontrollpunktterna P_0 och P_1 tillsammans bestämmer lutningen hos kurvan i den försat ändpunkten. Lutningen i den andra ändpunkten bestäms med hjälp av kontrollpunktterna P_{n-1} och P_n .

Exempel 3.19. Låt $P_0 = (-1, 1)$, $P_1 = (2, 2)$, $P_2 = (3, -1)$ och $P_3 = (5, 3)$. Då är

$$b_3(t) = (3t^3 - 6t^2 + 9t - 1, 11t^3 - 12t^2 + 3t + 1).$$

och

$$b'_3(t) = (9t^2 - 12t + 9, 33t^2 - 24t + 3).$$

Det ger tex att $b'_3(0) = (9, 3) = 3(3, 1) = 3(P_1 - P_0)$, se figur 3.23 Vanligtvis är tangentvektorn $b'_n(t)$ lång – dess längd motsvarar den fart som $b_n(t)$ rör sig utmed kurvan då t förändras med konstant fart. I figur 3.23 har vi istället ritat vektorn v , som är normeringen av b'_3 , dvs $v(t) = b'_3(t)/|b'_3(t)|$. \diamond

En delmängd av \mathbb{R}^n kallas för en *konvex mängd* om linjesegmentet mellan varje par av punkter i mängden helt tillhör mängden. Om det existerar två punkter i en delmängd sådan att någon del av linjesegmentet mellan punkterna ligger utanför delmängden, säges delmängden vara *icke-konvex* (ibland även *konkav mängd*, men det finns flera olika definitioner av konkav mängd). Låt $C \subset \mathbb{R}^n$. Då är C en konvex mängd om och endast om

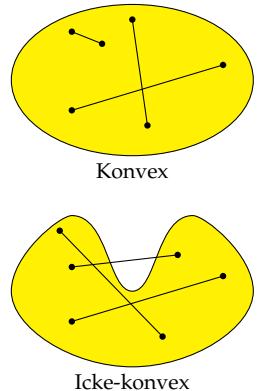
$$\lambda x + (1 - \lambda)y \in C$$

för alla $x, y \in C$ och alla $\lambda \in [0, 1]$. Med induktion följer det att om $x_1, x_2, \dots, x_k \in C$, så gäller att den *konvessa kombinationen*

$$\lambda_1 x_1 + \lambda_2 x_2 + \dots + \lambda_k x_k$$

tillhör den konveka mängden C , för alla icke-negativa reella tal $\lambda_1, \lambda_2, \dots, \lambda_k$ sådana att $\lambda_1 + \lambda_2 + \dots + \lambda_k = 1$. Låt S vara en mängd av punkter. Med det *konvessa höljet* menar vi snittet av alla konvessa mängder som innehåller S . Med andra ord är det konvessa höljet till S den minsta konvessa mängden som innehåller S .

Sats 3.11. Bézierkurvan $b_n(t)$ ligger i det konvessa höljet till (eng. convex hull) till mängden av dess kontrollpunkter.

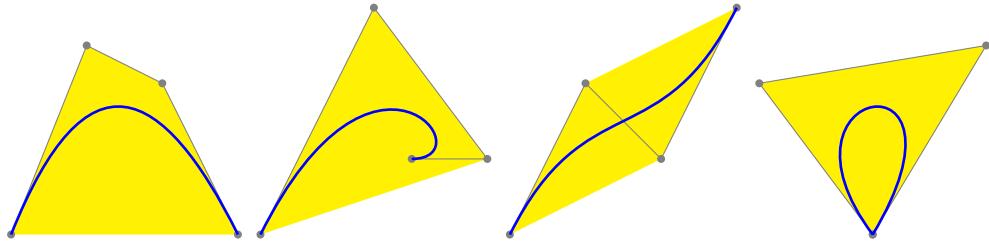


Bevis. Eftersom $B_{i,n}(t) \geq 0$ och

$$\sum_{i=0}^n B_{i,n}(t) = 1,$$

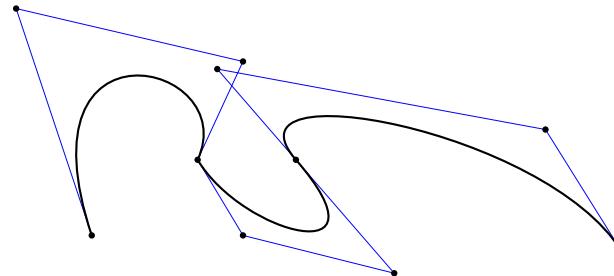
följer det att Bézierkurvan är en konvex kombination av dess kontrollpunkter. Alltså är Bézierkurvan innehållen i det konvexa hörnet till dess kontrollpunkter. \square

Exempel 3.20. I figur 3.24 ser vi några exempel på Bézierkurvor och motsvarande konvexa hörje. \diamond



Figur 3.24

Exempel 3.21. Vid tillämpning konstrueras ofta en kurva så att en följd av två konsekutiva Bézierkurvor har samma ändpunkt. Med andra ord, om $\{P_k\}_{k=0}^n$ och $\{Q_k\}_{k=0}^n$ är kontrollpunkter till två konsekutiva Bézierkurvor, som betecknas p_n respektive q_n , så väljs kontrollpunkterna så att $P_n = Q_0$. Vidare för att få till en "mjuk" övergång från en Bézierkurva till nästa krävs att vektorerna $p'_n(1)$ och $q'_n(0)$ är parallella. Vi åstadkommer det genom att välja kontrollpunkterna så att $P_{n-1}, P_n = Q_0$ och Q_1 är kolinjära, se figur 3.25. \diamond

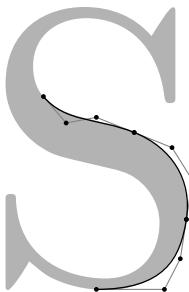


Figur 3.25. Tre kubiska Bézierkurvor.

Exempel 3.22. En vanlig tillämpning av kubiska Bézierkurvor är *typsnitt*, där konturen av varje bokstav och tecken beskrivs som en styckvis definierad Bézierkurva, se figur 3.26. För att specificera elektorniskt bokstavs kontur räcker det att lagra koordinaterna för kontrollpunkterna. All geometri i PostScript beskrivs med hjälp av kubiska Bézierkurvor. \diamond

Exempel 3.23. Låt $P_{i,j}$ vara punkter i \mathbb{R}^3 , där $i = 0, 1, \dots, m$ och $j = 0, 1, \dots, n$. Då definieras *Bézierten* av *ordning* (m, n) enligt

$$b_{n,m}(s, t) = \sum_{i=0}^m \sum_{j=0}^n B_{i,m}(s) B_{j,n}(t) P_{i,j},$$



Figur 3.26

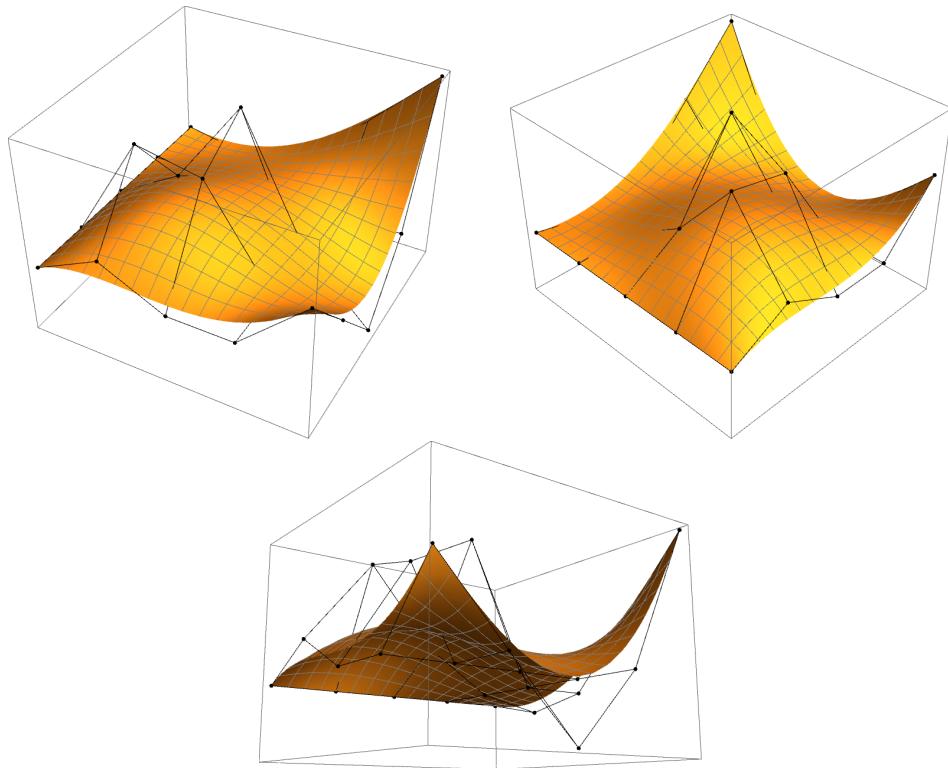
där $0 \leq s \leq 1$ och $0 \leq t \leq 1$. I figur 3.27 ser vi en Bézieryta av ordning $(5, 5)$, ur olika vinklar.

◊

3.9 Övningsuppgifter

- L 1. Bestäm Taylorpolynomet av grad 5 till $f(x) = 1/(1+x)$ kring punkten $a = 0$. Ange också motsvarande restterm.
 - L 2. Bevisa (3.2).
 - L 3. Polynomet

$$p(x) = -0.02x^3 + 0.1x^2 - 0.2x + 1.66$$
 interpolerar punkterna $(1, 1.54)$, $(2, 1.5)$, $(3, 1.42)$ och $(5, 0.66)$.
 - (a) Beräkna $p(4)$.
 - (b) Beräkna $p'(4)$.
 - (c) Beräkna integralen över intervallet $[1, 4]$ med $p(x)$ som integrand.
 - (d) Beräkna $p(5.5)$.
 - (e) Visa hur man bestämmer koefficienterna till $p(x)$.
 - L 4. Låt $f(x) = x^3$. Bestäm Lagranges interpolationspolynom $p_n(x)$
 - (a) då $n = 1$ för punkterna $x_0 = -1$ och $x_1 = 0$
 - (b) då $n = 2$ för punkterna $x_0 = -1$, $x_1 = 0$ och $x_2 = 1$
 - (c) då $n = 3$ för punkterna $x_0 = -1$, $x_1 = 0$, $x_2 = 1$ och $x_3 = 2$
 - (d) då $n = 1$ för punkterna $x_0 = 1$ och $x_1 = 2$
 - (e) då $n = 2$ för punkterna $x_0 = 0$, $x_1 = 1$ och $x_2 = 2$.
 - L 5. Låt $f(x) = x + 2/x$. Bestäm Lagranges interpolationspolynom $p_n(x)$
 - (a) då $n = 2$ för punkterna $x_0 = 1$, $x_1 = 2$ och $x_2 = 2.5$
 - (b) då $n = 3$ för punkterna $x_0 = 0.5$, $x_1 = 1$, $x_2 = 2$ och $x_3 = 2.5$.
- Approximera i båda fallen $f(1.2)$ och $f(1.5)$.



Figur 3.27

- L 6. Bestäm Newtons interpolationspolynom p_1, p_2, p_3 och p_4 för punkterna

$$x_0 = 1, \quad x_1 = 3, \quad x_2 = 4 \quad \text{och} \quad x_3 = 4.5$$

och med koefficienterna

$$a_0 = 4, \quad a_1 = -1, \quad a_2 = 0.4, \quad a_3 = 0.01 \quad \text{och} \quad a_4 = -0.002.$$

Låt $c = 2.5$ och beräkna $p_1(c), p_2(c), p_3(c)$ och $p_4(c)$.

- L 7. Bestäm Newtons interpolationspolynom p_1, p_2, p_3 och p_4 för punkterna

$$x_0 = -1, \quad x_1 = 0, \quad x_2 = 1 \quad \text{och} \quad x_3 = 4$$

och med koefficienterna

$$a_0 = 7, \quad a_1 = 3, \quad a_2 = 0.1, \quad a_3 = 0.05 \quad \text{och} \quad a_4 = -0.04.$$

Låt $c = 3$ och beräkna $p_1(c), p_2(c), p_3(c)$ och $p_4(c)$.

8. Låt $f(x) = \sin(x^2)$. Bestäm till $f(x)$ Lagranges interpolationspolynom $p_2(x)$ för punkterna $x_0 = 0, x_1 = 1$ och $x_2 = 2$. (20130109)

- L 9. Interpolera funktionen $f(x) = \cos(\sin x)$ med ett tredjegradsplynom med hjälp av Lagranges koefficientplynom $L_{3,k}$ i de givna punkterna $x_0 = 0.0, x_1 = 0.5, x_2 = 1.0$ och $x_3 = 1.5$. (20140110)

- L 10. Låt $f(x) = x^{1/2}$, $\alpha = 4.5$ och $\beta = 7.5$. Utgå från följande data.

k	x_k	$f(x_k)$
0	4.0	2.00000
1	5.0	2.23607
2	6.0	2.44949
3	7.0	2.64575
4	8.0	2.82843

- (a) Bestäm de dividerade differenserna.
- (b) Bestäm Newtons interpolationspolynom p_1, p_2, p_3 och p_4 .
- (c) Beräkna $p_n(\alpha)$ och $p_n(\beta)$ för $n = 1, 2, 3, 4$.
- (d) Jämför resultaten i föregående deluppgift med $f(\alpha)$ och $f(\beta)$.

11. Låt $f(x) = \ln(1 + x)$ och $x_k = 2 + 0.25k$ där $k = 0, 1, 2$. Bestäm först de dividerade differenserna för $f(x_k)$ och använd sedan resultatet för att bestämma Newtons interpolationspolynom $p_2(x)$. (20120821)

12. Låt $f(x) = x^3$. Bestäm de dividerade differenserna $f[x_0], f[x_0, x_1]$ och $f[x_0, x_1, x_2]$, där x_0, x_1 och x_2 är godtyckliga reella tal. (20130823)

Ledning: $a^3 - b^3 = (a - b)(a^2 + ab + b^2)$.

13. Låt punkterna (x_k, y_k) , där $k = 0, 1, \dots, 12$, ges av följande tabell.

k	0	1	2	3	4	5	6	7	8	9	10	11	12
x_k	-3	$-\frac{5}{2}$	-2	$-\frac{3}{2}$	-1	$-\frac{1}{2}$	0	$\frac{1}{2}$	1	$\frac{3}{2}$	2	$\frac{5}{2}$	3
y_k	1	$\frac{7}{2}$	-1	1	-3	$\frac{1}{2}$	0	3	-4	3	$-\frac{7}{2}$	-3	5

Eftersom $x_0 < x_1 < \dots < x_{12}$, så existerar det ett polynom p_{12} av grad 12 eller lägre som interpolerar punkterna.

- (a) Bestäm $L_{12,0}(x)$ och $L_{12,5}(x)$.
- (b) Ställ upp i sin helhet tabellen för Newtons dividerade differenser.
- (c) Bestäm $p_{12}(x)$ med valfri metod.

- L 14. Låt

$$S(x) = \begin{cases} S_1(x) & \text{om } 1 \leq x \leq 2 \\ S_2(x) & \text{om } 2 \leq x \leq 3. \end{cases}$$

Avgör vilka av följande definitioner av S_1 och S_2 som gör S till en kubisk spline.

- (a) $S_1(x) = \frac{19}{2} - \frac{81}{4}x + 15x^2 - \frac{13}{4}x^3$ $S_2(x) = -\frac{77}{2} + \frac{207}{4}x - 21x^2 + \frac{11}{4}x^3$
- (b) $S_1(x) = 11 - 24x + 18x^2 - 4x^3$ $S_2(x) = -54 + 72x - 30x^2 + 4x^3$
- (c) $S_1(x) = 18 - \frac{75}{2}x + 26x^2 - \frac{11}{2}x^3$ $S_2(x) = -70 + \frac{189}{2}x - 40x^2 + \frac{11}{2}x^3$
- (d) $S_1(x) = 13 - 31x + 23x^2 - 5x^3$ $S_2(x) = -35 + 51x - 22x^2 + 3x^3$

15. Låt

$$S(x) = \begin{cases} S_1(x) & \text{om } 0 \leq x \leq 1 \\ S_2(x) & \text{om } 1 \leq x \leq 2, \end{cases}$$

där

$$S_1(x) = x^3 + ax^2 + bx + c \quad \text{och} \quad S_2(x) = 2x^3 - x^2 + 7x + 11.$$

Bestäm a, b och c så att S blir en kubisk spline. (20120821)

L **16.** Låt S beteckna en kubisk splinefunktion som interpolerar noderna

$$(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n).$$

Antag att $S(x_0) = S(x_n)$, dvs $y_0 = y_n$. Härlled de två extra linjära ekvationer med avseende på k_0, k_1, \dots, k_n som utgår från

$$S'_1(x_0) = S'_n(x_n) \quad \text{och} \quad S''_1(x_0) = S''_n(x_n),$$

dvs de sk *periodiska randvillkoren*. (20130607)

17. Låt $S_1(x) = 1 - x + 2x^2 - x^3$ och $S_2(x) = a + bx + cx^2 + dx^3$ samt sätt

$$S(x) = \begin{cases} S_1(x) & \text{om } 0 \leq x \leq 1 \\ S_2(x) & \text{om } 1 \leq x \leq 2. \end{cases}$$

Bestäm koefficienterna a, b, c och d så att funktionen S är en kubisk splinefunktion för vilken $S(2) = 2$. (20140822)

Ledning: Aktuellt randvillkor är "clamped boundary", något man egentligen inte behövs veta för att lösa uppgiften.

18. Bestäm b_3 för var och en av följande uppsättning av kontrollpunkter.

- (a) $P_0 = (0, 0), P_1 = (1, 1), P_2 = (2, 2), P_3 = (3, 3)$
- (b) $P_0 = (0, 0), P_1 = (2, 5), P_2 = (2, 2), P_3 = (3, 0)$
- (c) $P_0 = (0, 0), P_1 = (1.5, 3), P_2 = (3, 1), P_3 = (2, 1)$
- (d) $P_0 = (0, 0), P_1 = (1, 2), P_2 = (2, 1), P_3 = (3, 3)$
- (e) $P_0 = (1.5, 0), P_1 = (0, 2), P_2 = (3, 2.5), P_3 = (1.5, 0)$
- (f) $P_0 = (0, 0), P_1 = (1, 2), P_2 = (1, 2), P_3 = (3, 0)$
- (g) $P_0 = (0, 0), P_1 = (2, 2), P_2 = (1, 2), P_3 = (3, 0)$
- (h) $P_0 = (1, 0), P_1 = (3, 3), P_2 = (0, 3), P_3 = (2, 0)$

Varje deluppgift motsvarar i tur och ordning en av kurvorna i figur 3.20.

L **19.** Utveckla följande Bernsteinpolynom.

$$\text{(a)} \quad B_{2,4}(t) \quad \text{(b)} \quad B_{3,5}(t) \quad \text{(c)} \quad B_{5,7}(t)$$

L **20.** Bestäm den Bézierkurva av grad 3 som definieras av kontrollpunkterna

$$P_0 = (-1, 2), P_1 = (1, 0), P_2 = (2, 3) \quad \text{och} \quad P_3 = (-1, 0). \quad \text{(20120603)}$$

L **21.** Bestäm den Bézierkurva av grad n som ges av följande kontrollpunkter.

$$\text{(a)} \quad n = 3, \{(1, 3), (3, -1), (2, 4), (3, 0)\}$$

- (b) $n = 4, \{(-2, 3), (-1, 3), (3, 5), (3, 4), (2, 3)\}$
(c) $n = 5, \{(1, 1), (2, 2), (3, 4), (4, 4), (5, 2), (6, 1)\}$

22. Den kvadratiska Bézierkurvan $\mathbf{b}_2(t) = (-1+4t-t^2, 1+2t-4t^2)$ har kontrollpunkterna

$$P_0 = (-1, 1), P_1 = (x, y) \text{ och } P_2 = (2, -1).$$

Bestäm samtliga möjliga punkter P_1 .

(20130823)

23. Låt n och i vara positiva heltal samt $j = n + i$. Med $\mathbf{b}_n^{[i, i+1, \dots, j]}(t)$ betecknar vi den Bézierkurva av ordning n som ges av kontrollpunkterna P_i, P_{i+1}, \dots, P_j . Visa att

$$\mathbf{b}_n^{[0, 1, \dots, n]}(t) = (1-t)\mathbf{b}_{n-1}^{[0, 1, \dots, n-1]}(t) + t\mathbf{b}_{n-1}^{[1, 2, \dots, n]}(t).$$

24. Låt $P_0 = (-1, -2), P_1 = (3, 0)$ och $P_2 = (-1, 3)$.

- (a) Bestäm den kvadratiska Bézierkurva $\mathbf{b}_2(t)$ vars kontrollpunkter ges i tur och ordning av P_0, P_1 och P_2 .
(b) Låt $\mathbf{b}_2(t)$ vara funktionen i föregående deluppgift. Bestäm samtliga giltiga t för vilka $\mathbf{b}_2(t) = (0, y)$ samt bestäm motsvarande y . (20150822)

L 25. Låt

$$\mathbf{b}_3(t) = (-2t^3 - 3t^2 + 6t - 1, -9t^3 + 12t^2 - 3t + 1)$$

vara en kubisk Bézierkurva. Bestäm motsvarande kontrollpunkter. (20150108)

L 26. Låt $\mathbf{b}_n(t)$ vara en Bézierkurva av grad n , där $n \geq 2$. Visa att

- (a) $\mathbf{b}_n''(0) = n(n-1)(P_2 - 2P_1 + P_0)$
(b) $\mathbf{b}_n''(1) = n(n-1)(P_n - 2P_{n-1} + P_{n-2})$.

Kapitel 4

Numerisk integration

20160420

Låt $f: \mathbb{R} \rightarrow \mathbb{R}$ vara integrerbar över intervallet $[a, b]$. Antag att vill beräkna integralen

$$\int_a^b f(x) dx.$$

Med andra ord, vill vi bestämma det tal I som integralen motsvarar och vi ska här studera metoder som ger oss en approximation av I . Notera att definitionen av en integral **inte** ges av

$$\int_a^b f(x) dx = F(b) - F(a), \quad (4.1)$$

där F är en primitiv funktion till integranden f . Även om vi inte har en explicit formel för F så kan mycket väl integralen existerar, tex är

$$\int e^{x^3} dx = \frac{x\Gamma(1/3, -x^3)}{3\sqrt[3]{-x^3}}$$

där Γ döljer i princip samma integral, nämligen

$$\Gamma(a, z) = \int_z^\infty t^{a-1} e^{-t} dt.$$

Det är även vanskt att använda (4.1) då vi kan bestämma en primitiv funktion till integranden – det kan inträffa att integralen likvänt inte existerar. Studera tex

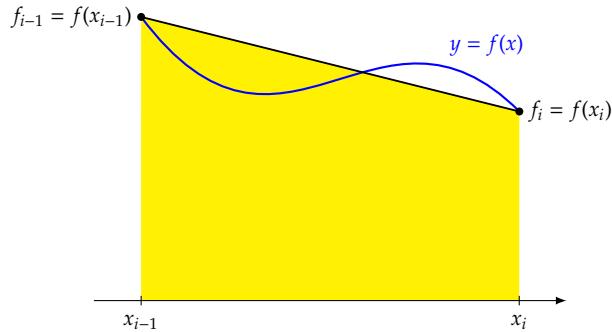
$$\int_{-1}^2 \frac{1}{x^3} dx = \left[-\frac{1}{2x^2} \right]_{-1}^2 = -\frac{1}{2 \cdot 2^2} + \frac{1}{2 \cdot (-1)^2} = -\frac{3}{8}.$$

Vi bestämmer en korrekt primitiv funktion men eftersom integranden inte är kontinuerlig i det aktuella intervallet kan vi inte använda (4.1). Integralen ovan existerar inte, dvs beräkningen är inkorrekt. Metoder för att symboliskt bestämma en primitiv funktion till en funktion ingår inte i kursen.

Sats 4.1 (Första medelvärde-satsen för integraler). *Låt f och g vara kontinuerliga reellvärda funktioner på intervallet $[a, b]$. Om g inte ändrar tecken i intervallet (a, b) , så är*

$$\int_a^b f(t)g(t) dt = f(\theta) \int_a^b g(t) dt$$

för något tal $\theta \in (a, b)$.



Figur 4.1

4.1 Trapetsmetoden

Antag att vi vill beräkna integralen

$$\int_a^b f(x) dx.$$

Dela upp intervallet $[a, b]$ i delintervall $[x_{i-1}, x_i]$, som alla är av samma längd. Vi interpolerar integranden f med en rät linje i respektive delintervall, se figur 4.1. Det ger oss en trapets, dvs en fyrhörning där två sidor är parallella. Låt $h = x_i - x_{i-1}$. Arean av trapetsen är

$$\frac{h}{2}(f_{i-1} + f_i).$$

Notera att uttrycket för arean är oberoende av inbördes förhållande mellan f_{i-1} och f_i samt deras tecken, eftersom

$$\int_{x_{i-1}}^{x_i} \left(\frac{f_i - f_{i-1}}{x_i - x_{i-1}}(x - x_{i-1}) + f_{i-1} \right) dx = \frac{h}{2}(f_{i-1} + f_i).$$

Tidigare såg vi att felet vid interpolation av en funktion med en rät linje är

$$\frac{f''(\xi_i(x))}{2}(x - x_{i-1})(x - x_i),$$

där $\xi_i(x)$ ligger i det minsta intervallet som innehåller x, x_{i-1} och x_i . Alltså är

$$\int_{x_{i-1}}^{x_i} f(x) dx = \frac{h}{2}(f_{i-1} + f_i) + \underbrace{\int_{x_{i-1}}^{x_i} \frac{f''(\xi_i(x))}{2}(x - x_{i-1})(x - x_i) dx}_{E(f)}.$$

Låt $x = x_{i-1} + th$. Med detta variabelbyte är $x = x_{i-1}$ då $t = 0$, och $x = x_i$ då $t = 1$, eftersom $h = x_i - x_{i-1}$. Vidare är

$$\frac{dx}{dt} = h.$$

Alltså ska vi ersätta dx med hdt . Slutligen har vi att $x - x_{i-1} = th$ och

$$x - x_i = x_{i-1} + th - x_i = th - h = (t - 1)h,$$

vilket ger oss att feltermen kan skrivas

$$E(f) = \frac{h^3}{2} \int_0^1 f''(\xi_i(t))t(t-1)dt$$

Medelvärdeessatsen för integraler, se sats 4.1 ger oss att

$$E(f) = \frac{h^3}{2} f''(\xi_i(\theta)) \int_0^1 t(t-1)dt = -\frac{h^3}{12} f''(\xi_i),$$

för något θ sådant att $0 < \theta < 1$, eller ekvivalent $x_i < \xi_i < x_{i+1}$. Låt m vara ett positivt heltal och $h = (b-a)/m$. Då delar punkterna

$$x_i = a + ih,$$

för $i = 0, 1, \dots, m$, intervallet $[a, b]$ i m delintervall av längd h . Då följer det att

$$\begin{aligned} \int_a^b f(x)dx &= \sum_{i=1}^m \int_{x_{i-1}}^{x_i} f(x)dx = \sum_{i=1}^m \left(\frac{h}{2}(f_{i-1} + f_i) - \frac{h^3}{12} f''(\xi_i) \right) \\ &= \underbrace{\frac{h}{2}(f_0 + 2f_1 + \dots + 2f_{m-1} + f_m)}_{T(h)} - \frac{h^3}{12} \sum_{i=1}^m f''(\xi_i). \end{aligned}$$

Antag att f'' är kontinuerlig. Om $a \leq \xi \leq b$, så existerar det tal c och d sådana att

$$c \leq f''(\xi) \leq d.$$

Alltså är

$$mc \leq f''(\xi_1) + f''(\xi_2) + \dots + f''(\xi_m) \leq md,$$

eftersom $a \leq \xi_i \leq b$ för alla $i = 1, 2, \dots, m$. Men vi har också att

$$ma \leq mf''(\xi) \leq mb.$$

Därmed existerar det ett tal ξ sådant att $a < \xi < b$ och

$$\sum_{i=1}^m f''(\xi_i) = mf''(\xi).$$

Eftersom $b-a = hm$ följer det att

$$-\frac{h^3}{12} \sum_{i=1}^m f''(\xi_i) = -\frac{b-a}{12} h^2 f''(\xi).$$

Alltså är det lokala trunkeringsfelet $O(h^3)$ och det totala trunkeringsfelet är $O(h^2)$.

Exempel 4.1. Vi vill beräkna integralen

$$\int_{-2}^3 e^{\sin x} dx \approx 7.0925863,$$

med hjälp av trapetsmetoden. Då tex $m = 2$ är $h = (3 - (-2))/2 = 2.5$. Alltså är

$$x_0 = -2, \quad x_1 = -2 + 2.5 = 0.5 \quad \text{och} \quad x_2 = -2 + 2.5 \cdot 2 = 3.$$

m	h	$T(h)$
1	5	3.88592
2	2.5	5.98083
4	1.25	6.96652
8	0.625	7.06077
16	0.3125	7.08466
32	0.15625	7.09061
100	0.05	7.09238
250	0.02	7.09255
500	0.01	7.09258

Tabell 4.1

Figur 4.2

Sätt $f(x) = e^{\sin x}$. Då är

$$f_0 = f(x_0) = e^{\sin(-2)}, \quad f_1 = f(x_1) = e^{\sin(0.5)} \quad \text{och} \quad f_2 = f(x_2) = e^{\sin(3)}$$

Det ger att

$$T(2.5) = \frac{1}{h}(f_0 + 2f_1 + f_2) \approx \frac{1}{2.5}(0.402807 + 2 \cdot 1.61515 + 1.15156) \approx 5.98083.$$

är approximationen av integralen med trapetsmetoden och steglängden $h = 2.5$. Ju större heltalet m destor fler delintervall och därmed erhåller vi bättre approximation av integralen, se tabell 4.1. Trapetsmetoden kan illustreras som summan av arean av trapetserna över respektive delintervall, se figur 4.2. \diamond

4.2 Newton-Cotes kvadraturformler

Låt p_n vara ett polynom, där $\deg p_n \leq n$, som interpolerar f i x_0, x_1, \dots, x_n . Då är

$$p_n(x) = \sum_{i=0}^n f(x_i)L_i(x) = \sum_{i=0}^n f_i L_i(x),$$

där L_i är Lagranges interpolationspolynom, dvs

$$L_i(x) = \frac{(x - x_0) \cdots (x - x_{i-1})(x - x_{i+1}) \cdots (x - x_n)}{(x_i - x_0) \cdots (x_i - x_{i-1})(x_i - x_{i+1}) \cdots (x_i - x_n)}.$$

Om vi integrerar p_n , så får vi att

$$\int_a^b p_n(x) dx = \sum_{i=0}^n f(x_i) \int_a^b L_i(x) dx = \sum_{i=0}^n A_i f_i.$$

Newton-Cotes' kvadraturformel ges av

$$\int_a^b f(x) dx = \sum_{i=0}^n A_i f_i + E(f).$$

Notera att integralen A_i enbart beror på x_0, x_1, \dots, x_n . Trunkeringsfelet ges av

$$E(f) = \int_a^b \frac{f^{(n+1)}(\xi(x))}{(n+1)!} (x - x_0)(x - x_1) \cdots (x - x_n) dx.$$

vilket följer från sats 3.2.

Exempel 4.2. Om $n = 1$, så är $x_0 = a$, $x_1 = b$ och $h = x_1 - x_0$. Alltså är,

$$A_0 = \int_{x_0}^{x_1} L_0(x) dx = \int_{x_0}^{x_1} \frac{x - x_1}{x_0 - x_1} dx = \frac{x_1 - x_0}{2} = \frac{h}{2}$$

och

$$A_1 = \int_{x_0}^{x_1} L_1(x) dx = \int_{x_0}^{x_1} \frac{x - x_0}{x_1 - x_0} dx = \frac{x_1 - x_0}{2} = \frac{h}{2}.$$

Med andra ord inget annat är trapetsmetoden: $h(f_0 + f_1)/2$. \diamond

Simpsons formel Låt $n = 2$, $h = (b - a)/2$ och $x_i = a + ih$, för $i = 0, 1, 2$. Då är

$$A_0 = \int_{x_0}^{x_2} \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)} dx = \frac{(x_2 - x_0)(3x_1 - 2x_0 - x_2)}{6(x_1 - x_0)} = \frac{h}{3},$$

$$A_1 = \int_{x_0}^{x_2} \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} dx = \frac{(x_2 - x_0)^3}{6(x_1 - x_0)(x_2 - x_1)} = \frac{4h}{3}$$

och

$$A_2 = \int_{x_0}^{x_2} \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)} dx = \frac{(x_2 - x_0)(x_0 - 3x_1 + 2x_2)}{6(x_2 - x_1)} = \frac{h}{3}.$$

Generellt ges Simpsons formel av

$$\int_a^b f(x) dx = S(h) + E(f),$$

där

$$S(h) = \frac{h}{3}(f_0 + 4f_1 + 2f_2 + 4f_3 + \cdots + 2f_{m-2} + 4f_{m-1} + f_m)$$

och där m är ett jämnt heltal samt $h = (b - a)/m$, med $x_i = a + ih$ och $f_i = f(x_i)$ för alla $i = 0, 1, \dots, m$. Om $f^{(4)}$ är kontinuerlig, så är

$$E(f) = -\frac{b-a}{180} h^4 f^{(4)}(\eta),$$

för något tal η sådant att $a < \eta < b$.

Exempel 4.3. Låt oss återigen studera vi integralen

$$\int_{-2}^3 e^{\sin x} dx \approx 7.0925863.$$

Med Simpsons formel interpolerar vi delar av integranden med andragradspolynom och erhåller på det sättet en bättre approximation än trapetsmetoden med samma steglängd, se tabell 4.2. \diamond

m	h	$S(h)$	$E(f)$
2	2.5	6.67912929034	-0.413457
4	1.25	7.29508484119	0.202499
8	0.625	7.29508484119	-0.000393633
16	0.3125	7.09262456698	0.0000383047
32	0.15625	7.09258891452	$2.65219 \cdot 10^{-6}$
64	0.078125	7.09258643170	$1.69374 \cdot 10^{-7}$
100	0.05	7.09258629086	$2.85327 \cdot 10^{-8}$
250	0.02	7.09258626306	$7.32808 \cdot 10^{-10}$
500	0.01	7.09258626237	$4.64508 \cdot 10^{-11}$

Tabell 4.2

Exempel 4.4. Låt x vara ett positivt reellt tal. I talteorin definieras $\pi(x)$ som antalet primtal mindre än eller lika med x . Tyvärr finns det inte något enkelt uttryck för funktionen π . Det vi däremot kan göra är att approximera $\pi(x)$ med en funktion som är givet av ett uttryck som är enklare att använda vid beräkning. Låt f och g vara funktioner på \mathbb{R} , dvs $f: \mathbb{R} \rightarrow \mathbb{R}$ och $g: \mathbb{R} \rightarrow \mathbb{R}$. Om

$$\lim_{x \rightarrow \infty} \frac{f(x)}{g(x)} = 1,$$

så säger man att f är asymptotisk med g , vilket betecknas $f \sim g$. Definitionen säger att funktionsgraferna $y = f(x)$ och $y = g(x)$ ligger nära varanadra – i varje fall när x går mot oändligheten. Låt

$$\text{Li}(x) = \int_2^x \frac{dt}{\log t}.$$

Primtalssatsen säger att $\pi(x) \sim \text{Li}(x)$. I figur 4.3 är grafen för $y = \pi(x)$ och $y = \text{Li}(x)$ svart respektive blå. Vi har tex att

$$\pi(100\,000) = 9592 \quad \text{och} \quad \text{Li}(100\,000) \approx 9628.76.$$

Låt $m = 100\,000$. Då är $h = (100\,000 - 2)/m = 0.99998$. Trapetsmetoden och Simpsons formel ger då att

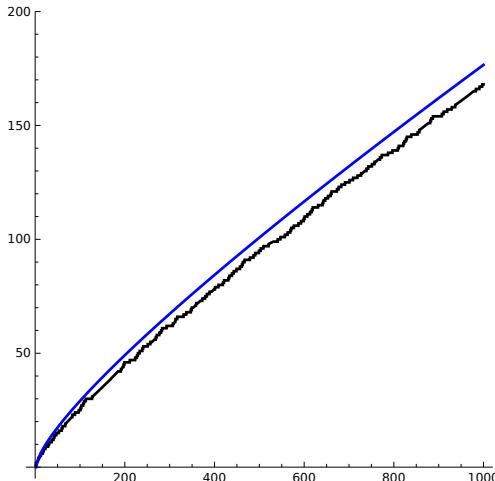
$$\text{Li}(100\,000) \approx T(h) \approx 9628.84 \quad \text{och} \quad \text{Li}(100\,000) \approx S(h) \approx 9628.78.$$

Med tanke på steglängden h är approximationen bra. \diamond

Simpsons 3/8-regel Låt $m = 3k$, där k är ett positivt heltal. Då $n = 3$ ger Newton-Cotes' kvadraturformel att

$$\int_a^b f(x) dx = \frac{3h}{8} \sum_{i=0}^k (f_{3i} + 3f_{3i+1} + 3f_{3i+2} + f_{3i+3}) + O(h^4).$$

Vid härledning delar man lämpligast först in intervallet $[a, b]$ i fyra delintervall.



Figur 4.3

Booles regel Låt $m = 4k$, där k är ett positivt heltal. Då $n = 4$ ger Newton-Cotes' kvadraturformel att

$$\int_a^b f(x) dx = \underbrace{\frac{2h}{45} \sum_{i=0}^k (7f_{4i} + 32f_{4i+1} + 12f_{4i+2} + 32f_{4i+3} + 7f_{4i+4})}_{B(h)} + O(h^6).$$

Det lämnas som övning att härleda denna formel.

4.3 Rombergs metod

Låt T vara approximationsfunktionen i trapetsmetoden, dvs

$$T(h) = h \left(\frac{f_0}{2} + f_1 + f_2 + \cdots + f_{m-1} + \frac{f_m}{2} \right).$$

Om vi låter h bli mindre och mindre, så får vi en bättre och bättre approximation av integralen. Men vi kan inte numeriskt beräkna $T(0)$ då det är ett gränsvärde på formen $0 \cdot \infty$. Tänk på att då h blir mindre och mindre ökar antalet termer f_i . Ett sätt att lösa detta är att med en metod som kallas *Richardsons extrapolation*. Låt

$$I(f) = \int_a^b f(x) dx.$$

Man kan visa att

$$T(h) = I(f) + a_1 h^2 + a_2 h^4 + \cdots + a_k h^{2k} + O(h^{2k+2})$$

där koefficienterna a_1, a_2, \dots, a_k är oberoende av h . Då är

$$T(2h) = I(f) + 4a_1 h^2 + 16a_2 h^4 + \cdots + 2^{2k} a_k h^{2k} + O(h^{2k+2}).$$

Det ger att

$$4T(h) - T(2h) = 3I(f) - 12a_2 h^4 - \cdots - (2^k - 4)a_k h^{2k} + O(h^{2k+2}) = 3I(f) + O(h^4).$$



Bevis

m	$T_1(h)$	$R_1 + \Delta_1/3$	$R_2 + \Delta_2/15$	$R_3 + \Delta_3/63$	$R_4 + \Delta_4/255$
2	5.980828194				
4	6.966520679	7.295084841			
8	7.060774642	7.092192629	7.078666482		
16	7.084662086	7.092624567	7.092653363	7.092875377	
32	7.090607207	7.092588915	7.092586538	7.092585477	7.092584340
64	7.092091626	7.092586432	7.092586266	7.092586262	7.092586265

Tabell 4.3

eller ekvivalent

$$I(f) = \frac{4T(h) - T(2h)}{3} + O(h^4).$$

Med andra ord har vi funnit en approximation av integralen $I(f)$ vars fel är i storleksordningen $O(h^4)$ jämfört felet i trapetsmetoden som är $O(h^2)$. Detta är första steget i Richardsons extrapolation av trapetsmetoden. Sätt

$$R_1(h) = T(h) \quad \text{och} \quad R_{k+1}(h) = R_k(h) + \frac{R_k(h) - R_k(2h)}{4^k - 1},$$

där k är ett positivt heltal. Man kan visa att

$$I(f) = R_k(h) + O(h^{2k+2}),$$

se övningsuppgift 11. Vi erhåller en triangulär tabell:

$$\begin{array}{cccc} R_0(2^i h) & & & \\ R_0(2^{i-1} h) & R_1(2^{i-1} h) & & \\ R_0(2^{i-2} h) & R_1(2^{i-2} h) & R_2(2^{i-2} h) & \\ R_0(2^{i-3} h) & R_1(2^{i-3} h) & R_2(2^{i-3} h) & R_3(2^{i-3} h) \\ \dots & & & \end{array}$$

Dessa värden beräknas tills

$$|R_k(2^i h) - R_k(2^{i+1} h)| < \varepsilon,$$

och då används $R_k(2^j h)$ som en approximation av $I(f)$.

Exempel 4.5. Ån en gång studerar vi integralen

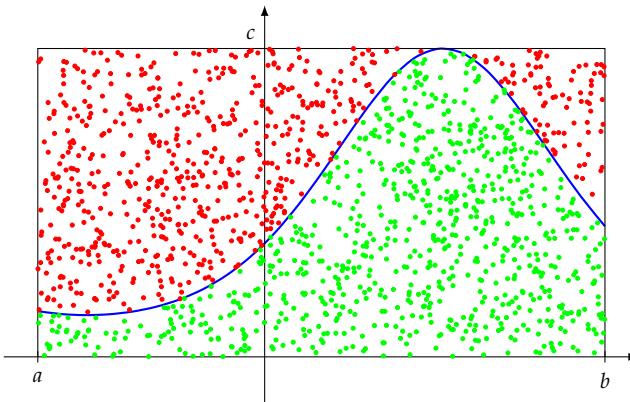
$$\int_{-2}^3 e^{\sin x} dx \approx 7.0925863.$$

Med Rombergs metod får vi resultatet i tabell 4.3. Med tex $\varepsilon = 5 \cdot 10^{-6}$ erhåller vi approximationen 7.092586266. \diamond

4.4 Monte Carlo-metoden

20160426

Detta är en probabilistisk metod – namnet anspelar på kasinot i Monte Carlo. Antag att f är positiv över intervallet $[a, b]$. Sätt $c = \max_{a \leq x \leq b} f(x)$. Låt A beteckna arean av den rektanglen som består av alla punkter (x, y) för vilka $a \leq x \leq b$ and $0 \leq y \leq c$.



Figur 4.4

Eftersom $f(x) \leq c$ då $a \leq x \leq b$ så ligger grafen $y = f(x)$ i denna rektangel. Därmed har vi att även området under $y = f(x)$ också är innehållan i rektangeln. Alltså är

$$I = \int_a^b f(x) dx \leq A = (b - a) \cdot c.$$

Välj slumpmässigt n punkter (α, β) i rektangeln. Låt m vara antal punkter som ligger under kurvan $y = f(x)$, dvs de för vilka $f(\alpha) \geq \beta$. Då är

$$I \approx \frac{Am}{n} = \frac{(b - a)c m}{n}.$$

För att uppnå en bra approximation av integralen krävs oftast man genererar många slumptal.

Exempel 4.6. Låt $n = 1500$ och studera (igen) integralen

$$\int_{-2}^3 e^{\sin x} dx \approx 7.0925863.$$

Då är $0 \leq e^{\sin x} \leq e$ för alla $x \in [-2, 3]$. Det betyder att de slumptal α och β som vi ska generera ska uppfylla

$$-2 \leq \alpha \leq 3 \quad \text{respektive} \quad 0 \leq \beta \leq e.$$

Antag att $m = 788$ av de $n = 1500$ slumpmässigt valda punkterna (α, β) visade sig ligga under kurvan $y = e^{\sin x}$, dvs i det område som motsvarar integralen, se de gröna punkterna i figur 4.4. Arean av rektangeln är $A = (3 - (-2))e = 5e$. Det ger oss att

$$\int_{-2}^3 e^{\sin x} dx \approx \frac{5e \cdot 788}{1500} \approx 7.14002.$$

Notera att kvoten m/n är en approximation av hur stor andel av rektangeln som integralen upptar. \diamond

4.5 Övningsuppgifter

- L 1. Använd trapetsmetoden för att beräkna integralen

$$\int_{-1}^4 (2 - |1 - x|) dx,$$

med $n = 5$ delintervall.

(20140603)

2. Beräkna integralen

$$\int_{0.2}^{1.7} e^{-x^2} dx$$

med trapetsmetoden då $n = 10$.

(20150822)

- L 3. Approximera integralen

$$I = \int_0^1 f(x) dx$$

med trapetsregeln, Simpsons formel, Simpsons 3/8-formel och Booles formel, för följande funktioner $f(x)$.

- (a) $\sin(\pi x)$ (b) $1 + e^{-x} \cos(4x)$ (c) $\sin(\sqrt{x})$

- L 4. Använd Simpsons formel för att approximera integralen

$$I = \int_{0.2}^{1.4} (x^2 + e^{2x}) dx$$

med $n = 4$ samt bestäm det relativa felet.

(20120603)

5. Approximera integralen

$$\int_1^2 \frac{\sin x}{x} dx,$$

med Simpsons formel, då $n = 3$.

(20130109)

- L 6. Beräkna integralen

$$\int_0^2 \cos(\sin x) dx$$

med hjälp av Simpsons formel, då $n = 5$.

(20140110)

- L 7. Approximera följande integraler med trapetsregeln ($n = 10$) och med Simpsons formel ($n = 5$).

(a) $\int_{-1}^1 \frac{dx}{1+x^2}$

(b) $\int_0^1 (2 + \sin(2\sqrt{x})) dx$

(c) $\int_{0.25}^4 \frac{dx}{\sqrt{x}}$

(d) $\int_0^4 x^2 e^{-x} dx$

(e) $\int_0^2 2x \cos(x) dx$

(f) $\int_0^\pi \sin(2x) e^{-x} dx$

- L 8. Ställ upp en tabell med de tre första raderna i enlighet med Rombergs metod för var och en av följande integraler.

(a) $\int_0^3 \frac{\sin(2x)}{1+x^2} dx$

(b) $\int_0^3 \sin(4x) e^{-2x} dx$

(c) $\int_{0.04}^1 \frac{dx}{\sqrt{x}}$

(d) $\int_0^2 \frac{dx}{x^2 + 0.1}$

(e) $\int_{1/(2\pi)}^2 \sin \frac{1}{x} dx$

(f) $\int_0^2 \sqrt{4-x^2} dx$

9. Låt $m = 2k$. Visa att

$$T(h) = \frac{T(2h)}{2} + h \sum_{i=1}^k f_{2i-1}$$

10. Låt k vara ett positivt heltal och $h_k = (b - a)/2^k$. Visa att

$$\begin{aligned} T(h_k) &= \frac{T(h_{k-1})}{2} + h \sum_{i=1}^{2^{k-1}} f_{2i-1}, \\ S(h_k) &= \frac{4T(h_k) - T(h_{k-1})}{3} \quad \text{och} \\ B(h_k) &= \frac{16S(h_k) - S(h_{k-1})}{15} \end{aligned}$$

Med andra ord är $R_2(h_k) = S(h_k)$. Är $R_3(h_k) = B(h_k)$?

L **11.** Visa att $R_k(h) = I(f) + O(h^{2k+2})$ för alla $k \in \mathbb{N}$.

Kapitel 5

Numerisk linjär algebra

5.1 Linjära ekvationssystem

Låt n vara ett positivt heltal. Vi ska studera metoder för lösning av linjära ekvationssystem på formen

$$\begin{cases} a_{1,1}x_1 + a_{1,2}x_2 + \cdots + a_{1,n}x_n = b_1 \\ a_{2,1}x_1 + a_{2,2}x_2 + \cdots + a_{2,n}x_n = b_2 \\ \vdots \\ a_{n,1}x_1 + a_{n,2}x_2 + \cdots + a_{n,n}x_n = b_n, \end{cases} \quad (5.1)$$

där $a_{i,j}$ och b_i är givna reella tal för alla $i = 1, 2, \dots, n$ och $j = 1, 2, \dots, n$. Bilda matrisen

$$\mathbf{A} = (a_{i,j})_{n \times n} = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{1,1} & a_{1,2} & \cdots & a_{1,n} \end{pmatrix}$$

samt vektorerna $\mathbf{x} = (x_1, x_2, \dots, x_n)$ och $\mathbf{b} = (b_1, b_2, \dots, b_n)$. Då kan det linjära ekvationssystemet (5.1) skrivas

$$\mathbf{Ax} = \mathbf{b},$$

där \mathbf{x} och \mathbf{b} i sin tur skrivs som kolonnmatriner. En vanlig metod för att lösa ett linjärt ekvationssystem är Gausselimination och de metoder vi ska studera baseras på denna. Men det är en långsam metod då den kräver många beräkningsoperationer och det är därför av intresse att förbättra den. Ett annat skäl är att flera av operationerna kan generera fel – stora sådana.

Om alla matriselement över (eller under) huvuddiagonalen i matrisen \mathbf{A} är noll, så kallas \mathbf{A} för en *övertriangulär matris* (respektive *undertriangulär matris*). Alternativt kallas \mathbf{A} för en *högertriangulär* respektive *vänstertriangulär* matris. En triangulär matris med enbart ettor i huvuddiagonalen kallas för en *enhetstriangulär matris*. Om \mathbf{A} är en övertriangulär, så ges motsvarande linjära ekvationssystem av

$$\begin{cases} a_{1,1}x_1 + a_{1,2}x_2 + \cdots + a_{1,n}x_n = b_1 \\ a_{2,2}x_2 + \cdots + a_{2,n}x_n = b_2 \\ \vdots \\ a_{n,n}x_n = b_n, \end{cases}$$

och som kan enkelt lösas med *bakåtsubstitution*, dvs

$$x_n = \frac{b_n}{a_{n,n}} \quad \text{och} \quad x_i = \frac{1}{a_{i,i}} \left(b_i - \sum_{j=i+1}^n a_{i,j} x_j \right),$$

där $i = n-1, n-2, \dots, 1$. För ett undertriangulära system

$$\begin{cases} a_{1,1}x_1 &= b_1 \\ a_{2,1}x_1 + a_{2,2}x_2 &= b_2 \\ &\vdots \\ a_{n,1}x_1 + a_{n,2}x_2 + \cdots + a_{n,n}x_n &= b_n \end{cases}$$

använts en liknande metod, som kallas *framåtsubstitution*, vilken ges av

$$x_1 = \frac{b_1}{a_{1,1}} \quad \text{och} \quad x_i = \frac{1}{a_{i,i}} \left(b_i - \sum_{j=1}^{i-1} a_{i,j} x_j \right),$$

där $i = 2, 3, \dots, n$. För att undvika stora beräkningsfel är det önskvärt att diagonalelementen $a_{i,i}$ inte ligger alltför nära 0.

Exempel 5.1. Diagonalmatriser är de enda matriser som är både över- och undertriangulära. Bland dessa finner två av de viktigaste matriserna, nämligen enhetsmatrisen och nollmatrisen. Två andra exemplen är

$$\mathbf{A}_1 = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & 5 & 6 & 7 \\ 0 & 0 & 8 & 9 \\ 0 & 0 & 0 & 10 \end{pmatrix} \quad \text{och} \quad \mathbf{A}_2 = \begin{pmatrix} 7 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 5 & 9 & 0 & 6 \end{pmatrix},$$

som är över- respektive undertriangulär. Om $\mathbf{b} = (11, 12, 13, 14)$, så finner vi lösningen till ekvationssystemet $\mathbf{A}_1 \mathbf{x} = \mathbf{b}$ med bakåtsubstitutionen enligt

$$\begin{aligned} x_4 &= \frac{b_4}{a_{4,4}} = \frac{14}{10} = \frac{7}{5} \\ x_3 &= \frac{1}{a_{3,3}}(b_3 - a_{3,4}x_4) = \frac{1}{8}(13 - 9 \cdot \frac{7}{5}) = \frac{1}{20} \\ x_2 &= \frac{1}{a_{2,2}}(b_2 - a_{2,3}x_3 - a_{2,4}x_4) = \frac{1}{4}(12 - 6 \cdot \frac{1}{20} - 7 \cdot \frac{7}{5}) = \frac{19}{50} \\ x_1 &= \frac{1}{a_{1,1}}(b_1 - a_{1,2}x_2 - a_{1,3}x_3 - a_{1,4}x_4) = \frac{1}{1}(11 - 2 \cdot \frac{19}{50} - 3 \cdot \frac{1}{20} - 4 \cdot \frac{7}{5}) = \frac{449}{100}. \end{aligned}$$

Eftersom $a_{3,3} = 0$ i \mathbf{A}_2 , så har $\mathbf{A}_2 \mathbf{x} = \mathbf{b}$ en entydig lösning och därför kommer inte framåtsubstitution att fungera i detta fall. \diamond

5.2 Pivoterings

Vi antar i fortsättningen att matrisen \mathbf{A} är icke-singulär, dvs att \mathbf{A} är inverterbar. Det betyder att det existerar en matris \mathbf{B} sådan att $\mathbf{AB} = \mathbf{BA} = \mathbf{I}$, där \mathbf{I} betecknar enhetsmatrisen. I fortsättningen betecknas \mathbf{B} med \mathbf{A}^{-1} . Målet med Gausselimination är att skriva om ett linjärt ekvationssystem till ett triangulärt system. Under processen

upprepar man en kombination av operationerna: byt plats på två rader eller två kolonner, multiplicera en rad med en konstant skilt från noll och addera en rad till en annan. Notera att om man skiftar plats på två kolonner, så innebär det att man ändrar om ordningen för de obekanta, dvs elementen i x byter platser – vilket ställer till det litet vid matrisrepresnetationen av ett linjärt ekvationssystem.

Under Gausseliminationen消除 man kolonn för kolonn alla element under huvuddiagonalen. Efter $k - 1$ steg ska man ha ett ekvationssystem på formen

$$\left\{ \begin{array}{l} a_{1,1}x_1 + a_{1,2}x_2 + \cdots + a_{1,k}x_k + a_{1,k+1}x_{k+1} + \cdots + a_{1,n}x_n = b_1 \\ a_{2,2}x_2 + \cdots + a_{2,k}x_k + a_{2,k+1}x_{k+1} + \cdots + a_{2,n}x_n = b_2 \\ \vdots \\ a_{k,k}x_k + a_{k,k+1}x_{k+1} + \cdots + a_{k,n}x_n = b_k \\ \vdots \\ a_{i,k}x_k + a_{i,k+1}x_{k+1} + \cdots + a_{i,n}x_n = b_i \\ \vdots \\ a_{n,k}x_k + a_{n,k+1}x_{k+1} + \cdots + a_{n,n}x_n = b_n. \end{array} \right.$$

För att till slut kunna utföra bakåtsubstitution måste $a_{i,i} \neq 0$, för alla $i = 1, 2, \dots, n$.

Algoritm 5.1 (Gausseliminering med partiell pivotering). Låt A vara en kvadratisk matris av ordning n som är icke-singulär. Vidare, låt $b \in \mathbb{R}^n$. Denna algoritm transformerar det linjära ekvationssystemet $Ax = b$ till ett triangulärt system. Upprepa nedanstående steg i tur och ordning för $k = 1, 2, \dots, n - 1$.

1. [Pivotelement] Finn det i sådant att $k \leq i \leq n$ och

$$|a_{i,k}| = \max_{k \leq j \leq n} |a_{j,k}|.$$

Element $a_{i,k}$ kallas *pivot* (uttalas *pivå*).

2. [Radbyte] Byt plats på rad k och i i A och motsvarande element i b .
3. [Multiplikator] Sätt

$$m_{j,k} \leftarrow \frac{a_{j,k}}{a_{k,k}},$$

för $j = k+1, k+2, \dots, n$. Notera att $|m_{j,k}| \leq 1$ och därmed troligtvis möjligt att få en god approximation med ett flyttal. Tänk på att $a_{j,k}$ och $a_{k,k}$ avser de "nya" koefficienterna efter radbytet i föregående steg.

4. [Eliminering] För alla $j = k+1, k+2, \dots, n$ sätt $a_{j,i} \leftarrow 0$ då $i = 1, 2, \dots, k$ samt

$$a_{j,i} \leftarrow a_{j,i} - m_{j,k}a_{k,i} \quad \text{och} \quad b_j \leftarrow b_j - m_{j,k}b_k$$

då $i = k+1, k+2, \dots, n$.

Anmärkning. Det finns ingen anledning att beräkna $a_{j,i} - m_{j,k}a_{k,i}$ för $j = 1, 2, \dots, k$ eftersom redan $a_{i,j}$ är lika med 0 för alla $i = 1, 2, \dots, k-1$ och från konstruktionen av $m_{j,k}$ följer att

$$a_{j,k} - m_{j,k}a_{k,k} = a_{j,k} - \frac{a_{j,k}}{a_{k,k}}a_{k,k} = 0.$$

Vi kan med andra ord sätta dessa matriselement till 0 och därmed undvika eventuella avrundningsfel. För att snabba upp algoritmen kan man låta dessa element behålla sina gamla värden. De kommer ändå inte användas mer, varken under Gausselimineringen eller den efterföljande bakåtsubstitutionen.

Exempel 5.2. Studera ekvationssystemet

$$\begin{cases} 4x - 10y + 30z = 4 \\ 3x + 20y + 60z = 1 \\ 17x + 5y - 8z = 2, \end{cases}$$

vilken representeras av totalmatrisen

$$\left(\begin{array}{ccc|c} 4.000 & -10.00 & 30.00 & 4.000 \\ 3.000 & 20.00 & 60.00 & 1.000 \\ 17.00 & 5.000 & -8.000 & 2.000 \end{array} \right),$$

där vi använder fyra siffrors avrundning, se exempel 1.16. Ekvationsystemet har en entydig lösning, nämligen

$$x = \frac{417}{2203} \approx 0.1893, \quad y = -\frac{1667}{11015} \approx -0.1513 \quad \text{och} \quad z = \frac{127}{2203} \approx 0.05765. \quad (5.2)$$

Vi löser ekvationsystemet med partiell pivotering. Största elementet i första kolonnen är $a_{3,1} = 17.00$. Därför byter vi plats på första och andra raden och får

$$\left(\begin{array}{ccc|c} 17.00 & 5.000 & -8.000 & 2.000 \\ 3.000 & 20.00 & 60.00 & 1.000 \\ 4.000 & -10.00 & 30.00 & 4.000 \end{array} \right).$$

Då är

$$m_{2,1} = \text{fl}_{\text{round}} \frac{a_{2,1}}{a_{1,1}} = \text{fl}_{\text{round}} \frac{3.000}{17.00} = 0.1765$$

och

$$m_{3,1} = \text{fl}_{\text{round}} \frac{a_{3,1}}{a_{1,1}} = \text{fl}_{\text{round}} \frac{4.000}{17.00} = 0.2353.$$

Genom att multiplicera första raden i totalmatrisen med $-m_{2,1}$ och $-m_{3,1}$ och adderar resultatet till andra respektive tredje raden erhåller vi matrisen

$$\left(\begin{array}{ccc|c} 17.00 & 5.000 & -8.000 & 2.000 \\ 0 & 19.12 & 61.41 & 0.647 \\ 0 & -11.18 & 31.88 & 3.529 \end{array} \right).$$

Vi nollställer elementen i första kolonnen på andra och tredje raden, eftersom avrundningsfel kan leda till att de inte blir noll. Varje beräkning har genomförts med fyra siffrors avrundning efter varje operation, som tex

$$a_{2,2} \leftarrow \text{fl}_{\text{round}}(a_{2,2} - \text{fl}_{\text{round}}(m_{2,1}a_{2,1})) = 19.12.$$

Eftersom $|19.12| \geq |-11.18|$ gör vi inget radbyte inför nästa eliminering. Då är

$$m_{3,2} = \text{fl}_{\text{round}} \frac{a_{3,2}}{a_{2,2}} = -\text{fl}_{\text{round}} \frac{11.18}{19.12} = -0.5847.$$

Multiplicerar vi andra raden med $-m_{3,2}$ och adderar resultatet till tredje raden erhåller vi matrisen

$$\left(\begin{array}{ccc|c} 17.00 & 5.000 & -8.000 & 2.000 \\ 0 & 19.12 & 61.41 & 0.647 \\ 0 & 0 & 67.79 & 3.907 \end{array} \right).$$

Bakåtsubstitution ger i tur ordning

$$z = 0.05763, \quad y = -0.1513 \quad \text{och} \quad x = 0.1892,$$

jämför med (5.2). Avrundningsfelen fortplantar sig och ger viss effekt, tex är

$$x = \text{fl}_{\text{round}} \left(\frac{\text{fl}_{\text{round}}(\text{fl}_{\text{round}}(b_1 - \text{fl}_{\text{round}}(a_{1,2}x_2)) - \text{fl}_{\text{round}}(a_{1,3}x_3))}{a_{1,1}} \right).$$

Utan pivotering riskerar vi få stora avrundningsfel vid beräkning av bla $m_{j,k}$, se exempel 1.16. \diamond

5.3 LU-faktorisering

Det finns flera metoder att skriva en matris som en produkt av två eller fler matriser, vilka har egenskaper som gör dem enklare att handska med än den ursprungliga matrisen. Vi ska i detta avsnit studera LU-faktorisering.

5.3.1 Permutationsmatriser och Gausstransformationer

En matris $P = (p_{ij})_{n \times n}$ kallas för en *permutationsmatris* om exakt ett element på varje rad och varje kolonn är lika med 1, och övriga element är 0.

Exempel 5.3. Låt

$$P = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix}.$$

Vi ser att P är en permutationsmatris. Namnet "permutationsmatris" inser man tex genom att studera produkten

$$\begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \\ d \\ e \end{pmatrix} = \begin{pmatrix} b \\ e \\ a \\ d \\ c \end{pmatrix},$$

dvs P permutterar elementen i vektorn (a, b, c, d, e) . \diamond

Låt $r, s \in \{1, 2, \dots, n\}$, där $r \neq s$, och låt $P_{r,s}$ vara en permutationsmatris som endast skiljer sig från enhetsmatrisen i fyra positioner enligt

$$p_{r,r} = 0, \quad p_{s,s} = 0, \quad p_{r,s} = 1 \quad \text{och} \quad p_{s,r} = 1. \quad (5.3)$$

Då kallas $P_{r,s}$ för en *elementär permutationsmatris*. Om $r = s$, så sätter vi $P_{r,r} = I$. Från definitionen följer det att $P_{r,s} = P_{s,r}$. Låt A^T beteckna transponatet av matrisen A .

Lemma 5.1. *Låt A vara en kvadratiska matris av ordning n . Då gäller att*

- (a) $P_{r,s}$ är symmetrisk, dvs $P_{r,s}^T = P_{r,s}$.
- (b) $P_{r,s}A$ skiftar rad r och s i A .

- (c) $A\mathbf{P}_{r,s}$ skiftar kolonn r och s i A .
- (d) $\mathbf{P}_{r,s}^{-1} = \mathbf{P}_{r,s}$.
- (e) $\mathbf{P}_{r,s}$ är ortogonal, dvs $\mathbf{P}_{r,s}^{-1} = \mathbf{P}_{r,s}^T$.
- (f) en produkt av två eller flera elementära permutationsmatriser är en permutationsmatris och varje permutationsmatris kan skrivas som en produkt av elementära permutationsmatriser (varje produkt av permutationsmatriser är således en permutationsmatris).
- (g) varje permutationsmatris är ortogonal,
- (h) om $r \neq s$, så är $\det \mathbf{P}_{r,s} = -1$.
- (i) determinanten av en permutationsmatris \mathbf{P} är -1 eller 1 , dvs antalet elementära permutationsmatriser i en faktorisering av \mathbf{P} är alltid udda respektive jämn.
- (j) antalet permutationsmatriser och elementära permutationsmatriser av ordning n är

$$n! \quad \text{respektive} \quad \frac{n(n-1)}{2} + 1.$$

Bevis. (a) Låt $\mathbf{P}_{r,s} = (p_{i,j})$ och $\mathbf{P}_{r,s}^T = (p'_{i,j})$. Då är $p'_{i,j} = p_{j,i}$. Vi behöver bara studera de fyra matriselement där $\mathbf{P}_{r,s}$ skiljer sig från enhetsmatrisen. Från (5.3) följer att

$$p'_{r,r} = p_{r,r} = 0, \quad p'_{s,s} = p_{s,s} = 0, \quad p'_{r,s} = p_{s,r} = 1 \quad \text{och} \quad p'_{s,r} = p_{r,s} = 1.$$

Det visar att transponatet av en elementär permutationsmatris är lika med matrisen, dvs elementära permutationsmatriser är symmetriska.

(b) Den i :te raden i $\mathbf{P}_{r,s}\mathbf{A}$ erhålls genom att beräkna skalärprodukten av den i :te raden i $\mathbf{P}_{r,s}$ med kolonnerna i \mathbf{A} , där rader och kolonner betraktas som vektorer. Den i :te raden i $\mathbf{P}_{r,s}$ består av nollor, förutom en etta i den j :te kolonnen. Om $i \neq r$ och $i \neq s$, så är $j = i$ och i så fall är den i :te raden i $\mathbf{P}_{r,s}\mathbf{A}$ den i :te raden i \mathbf{A} . Om $i = r$ (eller $i = s$), så är $j = s$ (eller $j = r$) och den i :te raden i $\mathbf{P}_{r,s}\mathbf{A}$ är den s :te (eller den r :te) raden i \mathbf{A} . Med andra ord har den r :te och s :te raden i \mathbf{A} bytt plats, medan de övriga är oförändrade.

(c) Från $(\mathbf{A}\mathbf{P}_{r,s})^T = \mathbf{P}_{r,s}^T\mathbf{A}^T = \mathbf{P}_{r,s}\mathbf{A}^T$ följer att $(\mathbf{A}\mathbf{P}_{r,s})^T$ erhålls genom att skiffta rad r och s i \mathbf{A}^T , enligt (b). Det visar påståendet eftersom transponatet av en matris transformera dess rader till kolonner, och vice versa.

(d) Det följer från (b) att $\mathbf{P}_{r,s}\mathbf{P}_{r,s} = (p''_{i,j})$ är lika med to $\mathbf{P}_{r,s}$, där den r :te och s :te raden har bytt plats. För de fyra elementen på de kritiska positionerna gäller att

$$p''_{r,r} = 1, \quad p''_{s,s} = 1, \quad p''_{r,s} = 0 \quad \text{och} \quad p''_{s,r} = 0,$$

vilket visar att alla ettor hittar vi i huvuddiagonalen. Allts är $\mathbf{P}_{r,s}\mathbf{P}_{r,s} = \mathbf{I}$. Ett alternativt bevis. Studera produkten $\mathbf{P}_{rs}\mathbf{P}_{rs}\mathbf{A}$. Till att börja med har rad r och rad s i $\mathbf{B} = \mathbf{P}_{rs}\mathbf{A}$ skifftat plats jämfört med \mathbf{A} och därefter skifftar $\mathbf{P}_{rs}\mathbf{B}$ samma rader in gång till, dvs tillbaka till deras ursprungliga positioner. Eftesom detta är sant för alla matriser \mathbf{A} , så måste $\mathbf{P}_{rs}\mathbf{P}_{rs} = \mathbf{I}$.

(e) Det följer från (a) och (d) att $\mathbf{P}_{r,s}^{-1} = \mathbf{P}_{r,s} = \mathbf{P}_{r,s}^T$.

(f) Låt k vara antal elementära permutationsmatriser som ingår i produkten. Vi bevisar påståendet med induktion över k . Om $k = 1$, så är det självklart att "produkten" är en permutationsmatris. Antag att varje produkt av $k - 1$ elementära permutationsmatriser är en permutationsmatris och låt \mathbf{P} beteckna en sådan produkt.

Notera att varje rad och varje kolonn i P innehåller exakt en 1. Låt $P_{r,s}$ vara en elementär permutationsmatris. Så är $P_{rs}P$ en produkt av k elementära permutationsmatriser. Eftersom $P_{rs}P$ erhålls genom att byta plats på två rader i P , kommer det finnas exakt en 1:a på varje rad och varje kolonn. Alltså är produkten $P_{rs}P$ en permutationsmatris.

För att bevisa det andra påståendet använder vi (b) och (d). Vi kan kasta om ordningen av raderna i en permutationsmatris P så att vi erhåller enhetsmatrisen. Det kan vu åstadkomma genom att multiplicera P från vänster med lämpligt valda elementära permutationsmatriser, dvs

$$P_{r_k,s_k} \cdots P_{r_2,s_2} P_{r_1,s_1} P = I.$$

Eftersom $P_{r,s}^{-1} = P_{r,s}$ så följer det att

$$P = P_{r_1,s_1} P_{r_2,s_2} \cdots P_{r_k,s_k}.$$

Det visar att varje permutationsmatris kan skrivas som en produkt av elementära permutationsmatriser. Notera att produkten inte är entydig.

(g) Från (f) vet vi att P kan skrivas som en produkt elementära permutationsmatriser. Därmed följer det att

$$\begin{aligned} PP^T &= (P_{r_1,s_1} P_{r_2,s_2} \cdots P_{r_k,s_k})(P_{r_1,s_1} P_{r_2,s_2} \cdots P_{r_k,s_k})^T \\ &= (P_{r_1,s_1} P_{r_2,s_2} \cdots P_{r_k,s_k})(P_{r_k,s_k}^T \cdots P_{r_2,s_2}^T P_{r_1,s_1}^T) = I, \end{aligned}$$

eftersom $P_{r,s}^{-1} = P_{r,s}^T$. Det bevisar att $P^{-1} = P^T$, dvs P är ortogonal.

(h) Om vi byter plats på rader i en matris ändrar determinanten tecknen. Då vi byter plats på rad r och rad s i $P_{r,s}$ så erhåller vi enhetsmatrisen. Eftersom det $I = 1$, följer det att $\det P_{r,s} = -\det I = -1$.

(i) Antag att vi behöver byta plats på rader k gånger för att transformera P till I . Då är $\det P = (-1)^k \det I = (-1)^k$. Eftersom determinanten av en matris är invariant, så måste k alltid vara av samma paritet oavsett hur P faktoriseras med elementära permutationsmatriser.

(j) Vid konstruktion av en permutationsmatris väljer vi de positioner där ettorna ska placeras. För den första raden har vi n möjligheter. För den andra raden har vi sedan $n-1$ möjligheter eftersom vi kan välja den kolonn som vi valt för ettan på första raden. För den tredje raden har vi $n-2$ möjligheter, osv. Från multiplikationsprincipen följer det att antal permutationsmatriser är

$$n \cdot (n-1) \cdot (n-2) \cdots 2 \cdot 1 = n!.$$

Det finns

$$\binom{n}{2} = \frac{n!}{2!(n-2)!} = \frac{n(n-1)}{2}$$

olika heltalspar (r,s) sådana att $1 \leq r < s \leq n$. Eftersom också enhetsmatrisen är en elementär permutationsmatris adderar vi 1 till ovanstående. \square

Låt $A = (a_{ij})$ vara en icke-singulär kvadratisk matris av ordning n och låt k vara ett heltal sådant att $1 \leq k < n$. Definiera kolonnmatriisen

$$m_k = (0 \quad \cdots \quad 0 \quad m_{k+1,k} \quad \cdots \quad m_{n,k})^T,$$

där de k första elementen är 0 och där $m_{i,k} = a_{i,k}/a_{k,k}$ för $i = k+1, k+2, \dots, n$. Låt M_k beteckan matrisen $(0 \quad \cdots 0 \quad m_k \quad 0 \quad \cdots \quad 0)$, där m_k är den k :te kolonnen i M_k och 0

är nollvektorn i \mathbb{R}^n . Sätt $L_k = I + M_k$, dvs den undertriangulära matrisen

$$L_k = L_k(A) = \begin{pmatrix} 1 & & & & & & 0 \\ & 1 & & & & & \\ & & \ddots & & & & \\ & & & 1 & & & \\ & & & 0 & 1 & & \\ & & & 0 & m_{k+1,k} & 1 & \\ & & & 0 & m_{k+2,k} & 0 & 1 \\ & & & \vdots & \vdots & \vdots & \ddots \\ 0 & & 0 & m_{n,k} & 0 & & 1 \end{pmatrix},$$

där alla andra element är 0 (allt utanför huvuddiagonalen och den del av kolonn k som ligger under huvuddiagonalen). Matrisen L_k kallas för den k :te *Gausstransformationen* av A . Notera att L_k beror på A , därav skrivsättet $L_k(A)$.

Lemma 5.2. Om $i \leq j$, så är $M_i M_j = O$.

Bevis. De enda element som inte är 0 i M_i finns i kolonn i . I produkten $M_i M_j$ multipli-
ceras dessa element med element på den i :te raden i M_j . Men alla element på denna
rad är lika med 0. Alltså är produkten lika med nollmatrisen. \square

Lemma 5.3. Inversen till L_k ges av $L_k^{-1} = I - M_k$.

Bevis. Från lemma 5.2 följer det att

$$L_k(I - M_k) = (I + M_k)(I - M_k) = I^2 + M_k I - I M_k - M_k^2 = I,$$

vilket bevisar lemmat. \square

Anmärkning. På matrisform har vi således att

$$L_k^{-1} = I - M_k = \begin{pmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & & -m_{k+1,k} & 1 & & \\ & & & \vdots & & \ddots & \\ & & & -m_{n,k} & & & 1 \end{pmatrix},$$

där alla övriga element är 0.

Lemma 5.4. Låt $a_k = (a_{1,k} \ a_{2,k} \ \cdots \ a_{n,k})^T$ vara den k :te kolonnen av en icke-singulär kvadratisk matris A av ordning n . Då är

$$L_k^{-1} a_k = (a_{1,k} \ a_{2,k} \ \cdots \ a_{k,k} \ 0 \ \cdots \ 0)^T.$$

Bevis. Först noterar vi att $L_k^{-1} a_k = a_k - M_k a_k$, enligt lemma 5.3. Det betyder att det i :te elementet i $M_k a_k$ är lika med 0 om $1 \leq i \leq k$, och lika med

$$\frac{a_{i,k}}{a_{k,k}} \cdot a_{k,k} = a_{i,k}$$

om $k < i \leq n$. Alltså elimineras de $n - k$ sista elementen i a_k . \square

Från beviset av lemma 5.4 följer det att produkten $L_k^{-1}A$ lämnar alla de k första raderna i A oförändrade och där de övriga raderna i $L_k^{-1}A$ fås genom att multiplicera den k :te raden i A med en konstant och subtrahera resultatet från den aktuella raden. I detalj kan det beskrivas på följande vis. Låt $a_j = (a_{1,j}, a_{2,j}, \dots, a_{n,j})$ vara den j :te kolonnen i A . Då är

$$\begin{aligned} L_k^{-1}a_j &= (I - M_k)a_j = a_k - M_k a_j \\ &= \begin{pmatrix} a_{1,j} \\ \vdots \\ a_{k,j} \\ a_{k+1,j} \\ \vdots \\ a_{n,j} \end{pmatrix} - \begin{pmatrix} 0 & & & & a_{1,j} \\ & \ddots & & & \vdots \\ & & 0 & & a_{k,j} \\ & & m_{k+1,k} & 0 & a_{k+1,j} \\ & & \vdots & & \vdots \\ & & m_{n,k} & & 0 \end{pmatrix} \begin{pmatrix} a_{1,j} \\ \vdots \\ a_{k,j} \\ a_{k+1,j} \\ \vdots \\ a_{n,j} \end{pmatrix} = \begin{pmatrix} a_{1,k} \\ \vdots \\ a_{k,k} \\ a_{k+1,j} - a_{k+1,k}a_{k,j}/a_{k,k} \\ \vdots \\ a_{n,j} - a_{n,k}a_{k,j}/a_{k,k} \end{pmatrix} \end{aligned}$$

Om $j = k$, så är $L_k^{-1}a_k = (a_{1,k}, \dots, a_{k,k}, 0, \dots, 0)$. Antag att A_{k-1} är matrisen som vi erhållit efter de $k-1$ första stegen under Gausselimineringen. Låt L_k vara den k :te Gausstransformationen av $P_{k,l}A_{k-1}$, för något heltalet l där $k \leq l \leq n$. Då motsvarar produkten $L_k^{-1}P_{k,l}A_{k-1}$ det k :te steget av Gausseliminationen av A .

Lemma 5.5. Om $i \leq j$, så är $L_i L_j = I + M_i + M_j$.

Bevis. Vi har att

$$L_i L_j = (I + M_i)(I + M_j) = I^2 + IM_i + M_j I + M_i M_j = I + M_i + M_j,$$

eftersom $M_i M_j = O$ då $i \leq j$. \square

Anmärkning. Alltså är $L_i L_j$ en matris på formen

$$\begin{pmatrix} & & & & \\ & \diagdown & & & \\ & & \diagdown & & \\ & & & \diagdown & \\ & & & & \end{pmatrix}$$

då $i \leq j$.

Lemma 5.6. Låt $\Lambda_k = L_1 L_2 \cdots L_k$, där $1 \leq k < n$. Då är

$$\Lambda_k = I + M_1 + M_2 + \cdots + M_k.$$

Med andra ord, Λ_k är en undertriangulär enhetsmatris, där de k första kolonnerna tas från respektive kolonn i M_k . Övriga element är lika med 0.

Bevis. Från definitionen av Gausstransformation följer att

$$\Lambda_k = (I + M_1)(I + M_2) \cdots (I + M_k) = I + M_1 + M_2 + \cdots + M_k,$$

eftersom alla termer med en produkt av minst två av matriserna M_i är lika med nollmatrisen, enligt lemma 5.2. \square

Anmärkning. Vi kan även grafiskt beskriva matrisen Λ_k som en matris på formen

$$\begin{pmatrix} & & & & \\ & \diagdown & & & \\ & & \diagdown & & \\ & & & \diagdown & \\ & & & & \end{pmatrix}.$$

Om vi startar med L_1 och multiplicera från höger med L_2, L_3, \dots, L_{n-1} , så fylls det triangelformade "hålet" under huvuddiagonalen med element skilda från 0. En undertriangulär enhetsmatris där alla element under huvuddiagonalen och till höger om kolonn k kallas i fortsättning för en Λ_k -matris

Exempel 5.4. Låt

$$L_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & 0 & 1 & 0 \\ 4 & 0 & 0 & 1 \end{pmatrix}, \quad L_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 5 & 1 & 0 \\ 0 & 6 & 0 & 1 \end{pmatrix} \quad \text{och} \quad L_3 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 7 & 1 \end{pmatrix}.$$

Då är

$$L_1 L_2 L_3 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & 5 & 1 & 0 \\ 4 & 6 & 7 & 1 \end{pmatrix}.$$

Detaljerna, dvs matrismultiplikationerna, lämnas som övning till läsaren. \diamond

5.3.2 Matrisfaktorisering

Lemma 5.7. Om $1 < k \leq l \leq n$ och $k \neq n$, så är $P_{k,l}\Lambda_{k-1}P_{k,l}L_k$ en Λ_k -matris.

Bevis. Först konstaterar vi att $M_i P_{k,l} = M_i$ för alla $i = 1, 2, \dots, k-1$, eftersom när vi multiplicerar med $P_{k,l}$ från höger skiftar vi plats på kolonn k och l th i M_i . Men elementen i båda dessa kolonner är alla lika med 0. Alltså är

$$\begin{aligned} P_{k,l}\Lambda_{k-1}P_{k,l}L_k &= P_{k,l}(I + M_1 + M_2 + \dots + M_{k-1})P_{k,l}(I + M_k) \\ &= (P_{k,l}P_{k,l} + P_{k,l}(M_1 P_{k,l} + M_2 P_{k,l} + \dots + M_{k-1} P_{k,l}))(I + M_k) \\ &= (I + P_{k,l}(M_1 + M_2 + \dots + M_{k-1}))(I + M_k) \\ &= I + P_{k,l}(M_1 + M_2 + \dots + M_{k-1}) + M_k, \end{aligned}$$

eftersom $P_{k,l}^{-1} = P_{k,l}$ och $M_i M_k = O$ då $i \leq k$. De element skilda från noll i M_i och som skifter position i $P_{k,l}M_i$, för $i = 1, 2, \dots, k-1$, kommer fortfarande vara under huvuddiagonalen eftersom $k \leq l$. Det bevisar lemmat. \square

Sats 5.8. Låt A vara en icke-singulär kvadratisk matris. Då finns det en permutationsmatris P sådan att vi kan skriva

$$PA = LU,$$

där L är en undertriangulär enhetsmatris och U är en övertriangulär matris. För denna permutationsmatris P är LU-faktorisering entydig.

Anmärkning. Påståendet i satsen är ekvivalent med *alla linjära ekvationssystem med en entydig lösning kan med Gausseminering skrivas som till ett triangulärt system*. Beviset är konstruktivt och ger oss en metod för att bestämma P , L och U .

Bevis. Sätt $A_0 = A$ och låt A_k beteckna matrisen som erhålls efter det k första stegen under Gausseminimationen med partiell pivotering, dvs

$$A_k = L_k^{-1}P_{k,l}A_{k-1},$$

för något heltal l sådant att $k \leq l \leq n$, och där L_k är det k :te Gausstransformationen av $P_{k,l}A_{k-1}$. Till slut får vi matrisen $U = A_{n-1}$, som är övertriangulär. Eftersom A

är icke-singulär är alla diagonalelement i U skilda från 0. Låt P_k beteckna den permutationsmatris som användes i det k :te steget. Då är

$$U = L_{n-1}^{-1} P_{n-1} \cdots L_2^{-1} P_2 L_1^{-1} P_1 A,$$

eller ekvivalent

$$P_1 A = L_1 P_2 L_2 \cdots P_{n-1} L_{n-1} U. \quad (5.4)$$

Sätt $B_1 = L_1$ och $B_k = P_k B_{k-1} P_k L_k$, där $1 \leq k < n$. Enligt lemma 5.7 är B_k en Λ_k -matris, för alla k . Om vi multiplicerar båda led i (5.4) med $P_{n-1} \cdots P_3 P_2$, så får vi

$$\begin{aligned} P_{n-1} \cdots P_2 P_1 A &= P_{n-1} \cdots P_3 P_2 L_1 P_2 L_2 \cdots P_{n-1} L_{n-1} U \\ &= P_{n-1} \cdots P_3 P_2 B_1 P_2 L_2 \cdots P_{n-1} L_{n-1} U \\ &= P_{n-1} \cdots P_3 B_2 P_3 L_3 \cdots P_{n-1} L_{n-1} U \\ &= P_{n-1} \cdots P_4 B_3 P_4 L_4 \cdots P_{n-1} L_{n-1} U \\ &= \cdots = B_{n-1} U, \end{aligned}$$

där $P = P_{n-1} \cdots P_2 P_1$ är en permutationsmatris och $L = B_{n-1}$ är en undertriangulär enhetsmatris. \square

Exempel 5.5. Låt

$$A = \begin{pmatrix} 0.2 & 0.1 & 3.0 & 1.2 \\ 5.1 & 0.5 & 0.4 & 2.1 \\ 1.1 & 4.7 & 2.8 & 0.7 \\ 0.4 & 0.7 & 1.9 & 3.6 \end{pmatrix}.$$

Vi väljer 5.1 i första kolonnen som pivotelement och byter därför plats på första och andra raden, dvs $P_1 = P_{1,2}$. Den första Gausstransformationen ges av is

$$L_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0.0392 & 1 & 0 & 0 \\ 0.2157 & 0 & 1 & 0 \\ 0.0784 & 0 & 0 & 1 \end{pmatrix}.$$

Elementen i första kolonnen i L_1 ges av

$$m_{2,1} = \frac{0.2}{5.1} = 0.0392, \quad m_{3,1} = \frac{1.1}{5.1} = 0.2157, \quad \text{och} \quad m_{4,1} = \frac{0.4}{5.1} = 0.0784.$$

Det ger att

$$A_1 = L_1^{-1} P_1 A = \begin{pmatrix} 5.1 & 0.5 & 0.4 & 2.1 \\ 0 & 0.080 & 2.984 & 1.118 \\ 0 & 4.592 & 2.714 & 0.247 \\ 0 & 0.661 & 1.869 & 3.435 \end{pmatrix}.$$

Som pivotelement i andra kolonnen väljer vi 4.592, dvs vi byter plats på andra och tredje raden med hjälp av $P_2 = P_{2,3}$. Den andra Gausstransformationen ges av

$$L_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0.018 & 1 & 0 \\ 0 & 0.144 & 0 & 1 \end{pmatrix}.$$

Det ger att

$$A_2 = L_2^{-1}P_2A_1 = \begin{pmatrix} 5.1 & 0.5 & 0.4 & 2.1 \\ 0 & 4.592 & 2.714 & 0.247 \\ 0 & 0 & 2.937 & 1.113 \\ 0 & 0 & 1.478 & 3.340 \end{pmatrix}.$$

I tredje kolonnen väljer vi 2.937 som pivotelement. Eftersom vi inte behöver byta plats på några rader ges perumtationsmatrisen denna gång av $P_3 = I$. Den tredje Gausstransformationen ges av

$$L_3 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0.503 & 1 \end{pmatrix}.$$

Det ger till slut att

$$U = L_3^{-1}P_3A_2 = \begin{pmatrix} 5.1 & 0.5 & 0.4 & 2.1 \\ 0 & 4.592 & 2.714 & 0.247 \\ 0 & 0 & 2.937 & 1.113 \\ 0 & 0 & 0 & 2.839 \end{pmatrix}$$

Vi kan summera beräkningarna ovan enligt

$$U = L_3^{-1}P_3A_2 = L_3^{-1}P_3(L_2^{-1}P_2A_1) = L_3^{-1}P_3(L_2^{-1}P_2(L_1^{-1}P_1A))$$

och eftersom $P_{i,j}^{-1} = P_{i,j}$ är detta ekvivalent med

$$P_3P_2P_1A = P_3P_2L_1P_2L_2P_3L_3U.$$

Sätt

$$P = P_3P_2P_1 = IP_{2,3}P_{1,2} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

och

$$L = P_3P_2L_1P_2L_2P_3L_3 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0.216 & 1 & 0 & 0 \\ 0.039 & 0.018 & 1 & 0 \\ 0.078 & 0.144 & 0.503 & 1 \end{pmatrix}.$$

Vi har bestämt matriser P , L och U sådana att $PA = LU$. ◊

Antag att $PA = LU$, där P är en permutationsmatris, L en undertriangulär enhetsmatris och U en övertriangulär matris. Då är

$$Ax = b \Leftrightarrow PAx = Pb \Leftrightarrow LUx = Pb.$$

Med andra ord, en lösning till $Ax = b$ fås genom att lösa de två triangulära systemen

$$Ly = Pb \quad \text{och} \quad Ux = y,$$

i tur och ordning, med bakåt- respektive framåtsubstitution.

Exempel 5.6 (LU-faktorisering för hand). Låt

$$\mathbf{A} = \begin{pmatrix} 10.0 & -5.0 & 7.0 \\ 6.0 & 2.0 & 5.0 \\ -3.0 & -1.0 & 1.0 \end{pmatrix}.$$

Vi ska faktorisera matisen \mathbf{A} utan pivotering, dvs utan att skifta plats på rader. Det ger oss en LU-faktorisering utan en permutationsmatris. Steg för steg får vi följande:

$$\begin{array}{lll} \text{Steg 0:} & \begin{pmatrix} 10.0 & -5.0 & 7.0 \\ 6.0 & 2.0 & 5.0 \\ -3.0 & -1.0 & 1.0 \end{pmatrix} & \mathbf{L} = \begin{pmatrix} 1 & 0 & 0 \\ * & 1 & 0 \\ * & * & 1 \end{pmatrix} \\ \text{Steg 1:} & \begin{pmatrix} 10.0 & -5.0 & 7.0 \\ 6.0 & 2.0 & 5.0 \\ -3.0 & -1.0 & 1.0 \end{pmatrix} \xrightarrow{-0.6} & \mathbf{L} = \begin{pmatrix} 1 & 0 & 0 \\ 0.6 & 1 & 0 \\ * & * & 1 \end{pmatrix} \\ \text{Steg 2:} & \begin{pmatrix} 10.0 & -5.0 & 7.0 \\ 0 & 5.0 & 0.8 \\ -3.0 & -1.0 & 1.0 \end{pmatrix} \xrightarrow{0.3} & \mathbf{L} = \begin{pmatrix} 1 & 0 & 0 \\ 0.6 & 1 & 0 \\ -0.3 & * & 1 \end{pmatrix} \\ \text{Steg 3:} & \begin{pmatrix} 10.0 & -5.0 & 7.0 \\ 0 & 5.0 & 0.8 \\ 0 & -2.5 & 3.1 \end{pmatrix} \xrightarrow{0.5} & \mathbf{L} = \begin{pmatrix} 1 & 0 & 0 \\ 0.6 & 1 & 0 \\ -0.3 & -0.5 & 1 \end{pmatrix} \end{array}$$

Slutligen har vi erhållit matriserna

$$\mathbf{U} = \begin{pmatrix} 10.0 & -5.0 & 7.0 \\ 0 & 5.0 & 0.8 \\ 0 & 0 & 3.5 \end{pmatrix} \quad \text{och} \quad \mathbf{L} = \begin{pmatrix} 1 & 0 & 0 \\ 0.6 & 1 & 0 \\ -0.3 & -0.5 & 1 \end{pmatrix}.$$

Det lämnas som övning att kontrolla att $\mathbf{LU} = \mathbf{A}$. ◊

Exempel 5.7 (Dolittles metod). Låt

$$\mathbf{A} = \begin{pmatrix} 2 & 2 & 1 & 4 \\ 1 & 2 & 1 & 0 \\ 2 & 0 & 2 & 1 \\ 1 & 2 & 0 & 2 \end{pmatrix}.$$

Studera

$$\begin{pmatrix} 2 & 2 & 1 & 4 \\ 1 & 2 & 1 & 0 \\ 2 & 0 & 2 & 1 \\ 1 & 2 & 0 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ * & 1 & 0 & 0 \\ * & * & 1 & 0 \\ * & * & * & 1 \end{pmatrix} \begin{pmatrix} * & * & * & * \\ 0 & * & * & * \\ 0 & 0 & * & * \\ 0 & 0 & 0 & * \end{pmatrix}.$$

Vid multiplikation av första raden i \mathbf{L} med de fyra kolonnerna i \mathbf{U} ger att

$$\begin{pmatrix} 2 & 2 & 1 & 4 \\ 1 & 2 & 1 & 0 \\ 2 & 0 & 2 & 1 \\ 1 & 2 & 0 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ * & 1 & 0 & 0 \\ * & * & 1 & 0 \\ * & * & * & 1 \end{pmatrix} \begin{pmatrix} 2 & 2 & 1 & 4 \\ 0 & * & * & * \\ 0 & 0 & * & * \\ 0 & 0 & 0 & * \end{pmatrix}.$$

Multiplicerar vi andra, tredje och fjärde raden i \mathbf{L} med första kolonnen i \mathbf{U} får vi att

$$\begin{pmatrix} 2 & 2 & 1 & 4 \\ 1 & 2 & 1 & 0 \\ 2 & 0 & 2 & 1 \\ 1 & 2 & 0 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1/2 & 1 & 0 & 0 \\ 1 & * & 1 & 0 \\ 1/2 & * & * & 1 \end{pmatrix} \begin{pmatrix} 2 & 2 & 1 & 4 \\ 0 & * & * & * \\ 0 & 0 & * & * \\ 0 & 0 & 0 & * \end{pmatrix}$$

Multiplikation av andra raden i L med andra, tredje och fjärde kolonnen i U ger

$$\begin{pmatrix} 2 & 2 & 1 & 4 \\ 1 & 2 & 1 & 0 \\ 2 & 0 & 2 & 1 \\ 1 & 2 & 0 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1/2 & 1 & 0 & 0 \\ 1 & * & 1 & 0 \\ 1/2 & * & * & 1 \end{pmatrix} \begin{pmatrix} 2 & 2 & 1 & 4 \\ 0 & 1 & 1/2 & -2 \\ 0 & 0 & * & * \\ 0 & 0 & 0 & * \end{pmatrix},$$

där tex det fjärde elementet på andra raden i U fås från

$$0 = \frac{1}{2} \cdot 4 + 1 \cdot u_{2,4} + 0 \cdot u_{3,4} + 0 \cdot u_{4,4} \Leftrightarrow u_{2,4} = -2.$$

Härnäst multiplicerar vi tredje och fjärde raden i L med andra kolonnen i U, vilket ger resultatet

$$\begin{pmatrix} 2 & 2 & 1 & 4 \\ 1 & 2 & 1 & 0 \\ 2 & 0 & 2 & 1 \\ 1 & 2 & 0 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1/2 & 1 & 0 & 0 \\ 1 & -2 & 1 & 0 \\ 1/2 & 1 & * & 1 \end{pmatrix} \begin{pmatrix} 2 & 2 & 1 & 4 \\ 0 & 1 & 1/2 & -2 \\ 0 & 0 & * & * \\ 0 & 0 & 0 & * \end{pmatrix}.$$

Multiplikation av tredje raden i L med tredje och fjärde kolonnen i U ger att

$$\begin{pmatrix} 2 & 2 & 1 & 4 \\ 1 & 2 & 1 & 0 \\ 2 & 0 & 2 & 1 \\ 1 & 2 & 0 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1/2 & 1 & 0 & 0 \\ 1 & -2 & 1 & 0 \\ 1/2 & 1 & * & 1 \end{pmatrix} \begin{pmatrix} 2 & 2 & 1 & 4 \\ 0 & 1 & 1/2 & -2 \\ 0 & 0 & 2 & -7 \\ 0 & 0 & 0 & * \end{pmatrix}.$$

Sista saknade elementet i L erhålls genom att multiplicera fjärde raden i L med tredje kolonnen i U, dvs

$$\begin{pmatrix} 2 & 2 & 1 & 4 \\ 1 & 2 & 1 & 0 \\ 2 & 0 & 2 & 1 \\ 1 & 2 & 0 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1/2 & 1 & 0 & 0 \\ 1 & -2 & 1 & 0 \\ 1/2 & 1 & -1/2 & 1 \end{pmatrix} \begin{pmatrix} 2 & 2 & 1 & 4 \\ 0 & 1 & 1/2 & -2 \\ 0 & 0 & 2 & -7 \\ 0 & 0 & 0 & * \end{pmatrix}.$$

Slutligen multiplicerar vi fjärde raden i L med fjärde kolonnen i U och får då

$$\begin{pmatrix} 2 & 2 & 1 & 4 \\ 1 & 2 & 1 & 0 \\ 2 & 0 & 2 & 1 \\ 1 & 2 & 0 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1/2 & 1 & 0 & 0 \\ 1 & -2 & 1 & 0 \\ 1/2 & 1 & -1/2 & 1 \end{pmatrix} \begin{pmatrix} 2 & 2 & 1 & 4 \\ 0 & 1 & 1/2 & -2 \\ 0 & 0 & 2 & -7 \\ 0 & 0 & 0 & -3/2 \end{pmatrix}.$$

En bra övning är att gå genom hela exemplet och reda ut alla detaljer för hand. ◇

5.3.3 Matrisinvers

Antag att A är icke-singulär matris av ordning n med LU faktoriseringen PA = LU, för någon permutationsmatris P, där L är en undertriangulär enhetsmatris och U är en övertriangulär matri. Då är

$$(PA)^{-1} = (LU)^{-1} \Leftrightarrow A^{-1}P^{-1} = U^{-1}L^{-1} \Leftrightarrow A^{-1} = U^{-1}L^{-1}P.$$

Låt

$$e_1 = (1, 0, \dots, 0), e_2 = (0, 1, \dots, 0), \dots, e_n = (0, 0, \dots, 1)$$

vara standardbasen i \mathbb{R}^n . Eftersom $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$, så är kolonn k i \mathbf{A}^{-1} lösningen till ekvationssystemet

$$\mathbf{A}\mathbf{x} = \mathbf{e}_k.$$

För varje $k = 1, 2, \dots, n$ löser vi de två ekvationssystemen

$$\mathbf{L}\mathbf{y}_k = \mathbf{P}\mathbf{e}_k \quad \text{och} \quad \mathbf{U}\mathbf{x}_k = \mathbf{y}_k.$$

Här ser vi nyttan av att först LU-faktorisera \mathbf{A} , då det besparas oss från att utföra Gausselimination om och om igen. Men det krävs likväld många operationer och därför bör man undvika att bestämma matrisinversen tills det är absolut nödvändigt.

LU-faktoriseringen kan också användas för att beräkna determinanten av \mathbf{A} . Från den multiplikativa egenskapen hos determinanter följer att

$$\det(\mathbf{PA}) = \det(\mathbf{P}) \det(\mathbf{A})$$

Eftersom $\det \mathbf{P} = \pm 1$ har vi att

$$\frac{1}{\det \mathbf{P}} = \det \mathbf{P}.$$

Därmed följer det från $\mathbf{PA} = \mathbf{LU}$ att

$$\det \mathbf{A} = \frac{1}{\det \mathbf{P}} \det(\mathbf{LU}) = \det(\mathbf{P}) \det(\mathbf{L}) \det(\mathbf{U}).$$

Determinanten av en triangulär matris är lika med produkten av dess diagonalelement. Det betyder att $\det \mathbf{L} = 1$.

Exempel 5.8. Låt

$$\mathbf{A} = \begin{pmatrix} 0.2 & 0.1 & 3.0 & 1.2 \\ 5.1 & 0.5 & 0.4 & 2.1 \\ 1.1 & 4.7 & 2.8 & 0.7 \\ 0.4 & 0.7 & 1.9 & 3.6 \end{pmatrix}.$$

Vi såg i exempel 5.5 att

$$\mathbf{PA} = \mathbf{LU}$$

där

$$\mathbf{L} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0.22 & 1 & 0 & 0 \\ 0.04 & 0.02 & 1 & 0 \\ 0.08 & 0.14 & 0.50 & 1 \end{pmatrix}, \quad \mathbf{U} = \begin{pmatrix} 5.1 & 0.5 & 0.4 & 2.1 \\ 0 & 4.59 & 2.71 & 0.25 \\ 0 & 0 & 2.94 & 1.11 \\ 0 & 0 & 0 & 2.84 \end{pmatrix}$$

och $\mathbf{P} = \mathbf{P}_{2,3}\mathbf{P}_{1,2}$. Låt $\mathbf{x}_k = (x_{1,k}, x_{2,k}, \dots, x_{n,k})$ vara kolonn k i \mathbf{A}^{-1} . Till atta börja med löser vi $\mathbf{Ly}_1 = \mathbf{Pe}_1$, dvs

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0.22 & 1 & 0 & 0 \\ 0.04 & 0.02 & 1 & 0 \\ 0.08 & 0.14 & 0.50 & 1 \end{pmatrix} \begin{pmatrix} y_{1,1} \\ y_{2,1} \\ y_{3,1} \\ y_{4,1} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}.$$

Därefter löser vi $\mathbf{Ux}_1 = \mathbf{y}_1$, dvs

$$\begin{pmatrix} 5.1 & 0.5 & 0.4 & 2.1 \\ 0 & 4.59 & 2.71 & 0.25 \\ 0 & 0 & 2.94 & 1.11 \\ 0 & 0 & 0 & 2.84 \end{pmatrix} \begin{pmatrix} x_{1,1} \\ x_{2,1} \\ x_{3,1} \\ x_{4,1} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \\ -0.50 \end{pmatrix}$$

Det ger oss att första kolonnen i A^{-1} är $x_1 = (0.06, -0.23, 0.41, -0.18)$. På samma sätt bestämmemr vi x_2, x_3 , och x_4 . Slutligen får vi att

$$A^{-1} = \begin{pmatrix} 0.0637 & 0.2051 & -0.0023 & -0.1404 \\ -0.2314 & -0.0416 & 0.2132 & 0.0600 \\ 0.4077 & -0.0081 & 0.0121 & -0.1335 \\ -0.1773 & -0.0104 & -0.0476 & 0.3522 \end{pmatrix}.$$

Vi avslutar exemplet med att beräkna determinanten av A , nämligen

$$\det A = \det(P) \det(L) \det(U) = 5.1 \cdot 4.58 \cdot 2.94 \cdot 2.84 = 195.293,$$

eftersom $\det L = 1$ och $\det P = \det(P_{2,3}) \det(P_{1,2}) = (-1) \cdot (-1) = 1$. \diamond

5.4 Matrisnorm och konditionstal

Låt $\text{Mat}_n(\mathbb{R})$ beteckna mängden av alla reella kvadratiska matriser av ordning n , där n är ett positivt heltal större än 1. Vidare, låt $\|\cdot\|$ vara en vekornorm på \mathbb{R}^n , se avsnitt 2.6.1. Då definieras *matrisnormen* $\text{Mat}_n(\mathbb{R}) \rightarrow \mathbb{R}$ som *induceras* av vektornormen $\|\cdot\|$ enligt

$$\|A\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|}, \quad A \in \text{Mat}_n(\mathbb{R}),$$

där "sup" avser *supremum*, dvs det minsta reella tal som är större än eller lika med alla positiva reella tal $\|Ax\|/\|x\|$ som erhålls då x "genomlöper" samtliga element i \mathbb{R}^n förrutom nollvektorn. Notera att Ax är en vektor, dvs uttrycket $\|Ax\|$ kräver endast vektornormen – ingen risk för "rundgång" i definitionen.

Lemma 5.9. *Den inducerade matrisnormen uppfyller*

$$\|A\| = \max_{\|x\|=1} \|Ax\|.$$

Bevis. Antag att $\|y\| \neq 0$. Sätt $a = \|y\|$ och $x = (1/a)y$. Observera att $a > 0$ och $\|x\| = 1$. Det följer då att

$$\|A\| = \sup_{y \neq 0} \frac{\|Ay\|}{\|y\|} = \sup_{y \neq 0} \frac{\|Ay\|}{a} = \sup_{y \neq 0} \left\| \frac{1}{a} Ay \right\| = \sup_{y \neq 0} \left\| A \left(\frac{1}{a} y \right) \right\| = \sup_{\|x\|=1} \|Ax\|. \quad (5.5)$$

Eftersom funktionen $x \mapsto \|Ax\|$ är kontinuerlig och mängden av alla enhetsvektorer är kompakt, dvs sluten och begränsad, så antar funktionen sitt maximum i nämnd mängd. Alltså kan supremum ersättas med maximum i sista ledet i (5.5). \square

Sats 5.10. *Den inducerade matrisnormen uppfyller*

- (a) $\|I\| = 1$
- (b) $\|A\| \geq 0$, med likhet om och endast om $A = \mathbf{O}$
- (c) $\|aA\| = |a| \cdot \|A\|$
- (d) $\|A + B\| \leq \|A\| + \|B\|$
- (e) $\|Ax\| \leq \|A\| \cdot \|x\|$

$$(f) \|AB\| \leq \|A\| \cdot \|B\|,$$

för alla matriser A och B , vektorer x och reella tal a .

Bevis. Lämnas som övning. □



Sats 5.11. De inducerade matrisnormerna med avseende på absolutnormen, Euklidiska normen och maximumnormen ges av

$$\|A\|_1 = \max_{1 \leq j \leq n} (|a_{1,j}| + |a_{2,j}| + \cdots + |a_{n,j}|)$$

$$\|A\|_2 = \max \{ \sqrt{\lambda} : \lambda \text{ är ett egenvärde till } A^T A \}$$

respektive

$$\|A\|_\infty = \max_{1 \leq i \leq n} (|a_{i,1}| + |a_{i,2}| + \cdots + |a_{i,n}|),$$

där A^T betecknar transponatet av A .

Bevis. Lämnas också som övning. □



Exempel 5.9. Låt

$$A = \begin{pmatrix} 4 & -14 & 6 \\ 0 & 25 & 1 \\ 2 & 5 & -19 \end{pmatrix}.$$

Då är

$$\begin{aligned} \|A\|_1 &= \max(|4| + |0| + |2|, |-14| + |25| + |5|, |6| + |1| + |-19|) \\ &= \max(6, 44, 26) = 44, \end{aligned}$$

dvs summan av absolutbeloppen av elementen i respektive kolonn, och

$$\begin{aligned} \|A\|_\infty &= \max(|4| + |-14| + |6|, |0| + |25| + |1|, |2| + |5| + |-19|) \\ &= \max(24, 26, 26) = 26, \end{aligned}$$

dvs summan av absolutbeloppen av elementen i respektive rad. Normen $\|A\|_2$ är mer komplicerad att beräkna. Egenvärdena till matrisen

$$A^T A = \begin{pmatrix} 20 & -46 & -14 \\ -46 & 846 & -154 \\ -14 & -154 & 398 \end{pmatrix}$$

är lösningarna till den karakteristiska ekvationen

$$\det(A^T A - \lambda I) = 0 \Leftrightarrow \lambda^3 - 1264\lambda^2 + 335\,560\lambda - 5\,053\,504 = 0,$$

vilken kan lösas tex med Newton-Raphson metod. Notera att det finns mer direkta metoder för att numeriskt bestämma egenvärdena till en matris. Vi finner att $A^T A$ har egenvärdena

$$\lambda_1 \approx 895.641, \lambda_2 \approx 352.346 \text{ och } \lambda_3 \approx 16.0136.$$

Alltså är

$$\|A\|_2 = \sqrt{\lambda_1} \approx 29.927.$$

Från sats 5.10 följer att en matrisnorm kan precis som vektornorm också användas för att mäta avstånd, dvs $\|A\|_p = \|A - O\|_p$ motsvarar "längden av matrisen". ◊

En matris A säges vara *illa-konditionerad* om det finns en matris B sådan att en liten ändring av något eller några element i A eller B ger stora förändringar i $A^{-1}B$.

Exempel 5.10. Låt

$$A_1 = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 1.997 & 4 \\ 1 & 2.002 & 5 \end{pmatrix} \quad \text{och} \quad A_2 = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 0 & 4 \\ 1 & 3 & 6 \end{pmatrix}.$$

Vi bestämmer lösningen x_i till $A_1x = b_i$ för

$$b_1 = (7, 8, 9), \quad b_2 = (7, 8.01, 9) \quad \text{och} \quad b_3 = (7, 7.99, 9),$$

vilka ges av

$$x_1 = A_1^{-1}b_1 = (4, 0, 1), \quad x_2 = A_1^{-1}b_2 = (8.992, -2.500, 1.002) \\ \text{respektive} \quad x_3 = A_1^{-1}b_3 = (-0.992, 2.500, 0.997).$$

Motsvarande lösningar till $A_2x = b_i$ för samma b_i som ovan ges av

$$x_1 = (5.143, -0.143, 0.714), \quad x_2 = (5.147, -0.147, 0.716) \\ \text{respektive} \quad x_3 = (5.139, -0.139, 0.713).$$

Vi ser att matrisen A_1 är illa-konditionerad. \diamond

Låt x vara lösningen till det linjära ekvationssystemet $Ax = b$ och låt \bar{x} vara en approximation av x , där A är icke-singulär och $b \neq 0$. Sätt $\delta x = \bar{x} - x$. Det absoluta felet och det relativa felet ges då av

$$\|\delta x\| \quad \text{respektive} \quad \frac{\|\delta x\|}{\|x\|}$$

Vi kan betrakta δx i \bar{x} som en störning av x . Eftersom \bar{x} är en approximation av lösningen, gäller att

$$A\bar{x} = \bar{b} \quad \Leftrightarrow \quad A(x + \delta x) = b + \delta b \quad \Leftrightarrow \quad \delta x = A^{-1}\delta b.$$

Egenskaperna för matris- och vektornorm ger oss att

$$\|\delta x\| = \|A^{-1}\delta b\| \leq \|A^{-1}\| \cdot \|\delta b\|.$$

Vidare är

$$\|b\| = \|Ax\| \leq \|A\| \cdot \|x\| \quad \Leftrightarrow \quad \frac{1}{\|x\|} \leq \frac{\|A\|}{\|b\|}.$$

Kombinerar vi ovanstående två uppskattningar erhåller vi

$$\frac{\|\delta x\|}{\|x\|} \leq \|A\| \cdot \|A^{-1}\| \cdot \frac{\|\delta b\|}{\|b\|}.$$

Konstanten $\kappa(A) = \|A\| \cdot \|A^{-1}\|$ kallas för *konditionstalet för matrisen A*. Om $\kappa(A)$ är stort, så kan det relativa felet i lösningen bli stort, även om det relativa felet i högerledet b är litet.

Exempel 5.11. Låt

$$\mathbf{A} = \begin{pmatrix} 0.5 & 2.8 & 1.1 \\ 0.1 & 0.7 & 4.3 \\ 3.7 & 0.3 & 0.2 \end{pmatrix}.$$

Då är $\kappa_1(\mathbf{A}) \approx 2.67$, $\kappa_2(\mathbf{A}) \approx 2.08$ och $\kappa_\infty(\mathbf{A}) \approx 2.72$. Detaljerna lämnas som övning. \diamond

Sats 5.12. Låt \mathbf{A} vara en icke-singulär matris och \mathbf{x} en lösning till $\mathbf{Ax} = \mathbf{b}$. Antag att

$$\|\mathbf{A}^{-1}\| \cdot \|\delta\mathbf{A}\| = \kappa(\mathbf{A}) \frac{\|\delta\mathbf{A}\|}{\|\mathbf{A}\|} = \tau < 1,$$

Då är också $\mathbf{A} + \delta\mathbf{A}$ icke-singulär samt lösningen \mathbf{y} till det störda systemet

$$(\mathbf{A} + \delta\mathbf{A})\mathbf{y} = \mathbf{b} + \delta\mathbf{b}$$

uppfyller

$$\frac{\|\mathbf{x} - \mathbf{y}\|}{\|\mathbf{x}\|} \leq \frac{\kappa(\mathbf{A})}{1 - \tau} \left(\frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|} + \frac{\|\delta\mathbf{A}\|}{\|\mathbf{A}\|} \right).$$



Bevis?

Från $1 = \|\mathbf{I}\| = \|\mathbf{AA}^{-1}\| \leq \|\mathbf{A}\| \cdot \|\mathbf{A}^{-1}\|$ följer att $\kappa(\mathbf{A}) \geq 1$, för alla icke-singulära matriser \mathbf{A} . Om $\kappa(\mathbf{A})$ är ett litet tal, så säges \mathbf{A} vara *väl-konditionerad*. Annars säges matrisen vara *illa-konditionerad*.

Exempel 5.12. Låt \mathbf{A}_1 och \mathbf{A}_2 vara matriserna som vi studerade i exempel 5.10. Då är

$$\kappa_2(\mathbf{A}_1) \approx 5490.48 \quad \text{och} \quad \kappa_2(\mathbf{A}_2) \approx 18.5382,$$

vilket indikerar att \mathbf{A}_1 är mer känslig för små ändringar än \mathbf{A}_2 . \diamond

5.5 Minsta kvadratmetoden

20160504

Exempel 5.13. Givet följande statistiska data på formen (x, y) ,

$$(0.1, 2.8), (0.4, 2.2), (1.1, 2.1), (1.8, 1.6), (2.3, 1.9), (3.1, 1.7).$$

Vi vill approximera ovanstående data med det linjära polynomet $f(x) = a + bx$. Likheterna $f(x_i) = y_i$, där $i = 1, 2, \dots, 6$, motsvarar ett överbestämt ekvationssystem

$$\begin{cases} a + 0.1b = 2.8 \\ a + 0.4b = 2.2 \\ a + 1.1b = 2.1 \\ a + 1.8b = 1.6 \\ a + 2.3b = 1.9 \\ a + 3.1b = 1.7 \end{cases} \Leftrightarrow \begin{pmatrix} 1 & 0.1 \\ 1 & 0.4 \\ 1 & 1.1 \\ 1 & 1.8 \\ 1 & 2.3 \\ 1 & 3.1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 2.8 \\ 2.2 \\ 2.1 \\ 1.6 \\ 1.9 \\ 1.7 \end{pmatrix},$$

som saknar lösning. \diamond

Låt $\mathbf{A} = (\mathbf{a}_1 \ \mathbf{a}_2 \ \dots \ \mathbf{a}_m)$ vara en $n \times m$ -matris, där $n \geq m$ och $\mathbf{a}_j \in \mathbb{R}^n$ är den j :te kolonnen i \mathbf{A} . Om $\mathbf{x} = (x_1, x_2, \dots, x_m)$, så kan \mathbf{Ax} skrivas som en linjärkombination av kolonnerna i \mathbf{A} enligt

$$\mathbf{Ax} = x_1 \mathbf{a}_1 + x_2 \mathbf{a}_2 + \dots + x_m \mathbf{a}_m.$$

Alltså spänner vektorerna $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_m$ upp ett underrum V till \mathbb{R}^n , dvs V är mängden av alla vektorer på formen \mathbf{Ax} . Notera att $\mathbf{Ax} \in \mathbb{R}^n$ och $\mathbf{x} \in \mathbb{R}^m$. Att mängden V är

ett *underrum* menas den är en delmängd till \mathbb{R}^n sådan att $a\mathbf{v} \in V$ och $\mathbf{u} + \mathbf{v} \in V$ för alla $a \in \mathbb{R}$ och alla $\mathbf{u}, \mathbf{v} \in V$, se övningsuppgift 18. Speciellt tillhör a_j mängden V .

Om $\mathbf{b} \in V$, så är det linjär ekvationssystemet $\mathbf{Ax} = \mathbf{b}$ lösbart. Om $\mathbf{b} \notin V$, så vill vi finna \mathbf{x} sådan att *residualen* $\mathbf{r} = \mathbf{b} - \mathbf{Ax}$ är så liten som möjligt, med avseende på på den Euklidiska vektornormen $\|\cdot\|_2$. Låt $\mathbf{y} = \mathbf{Ax}$. Då gäller att $\mathbf{y} \in V$. Det kan visas att varje vektor $\mathbf{b} \in \mathbb{R}^n$ kan skrivas entydigt på formen

$$\mathbf{b} = \mathbf{b}_1 + \mathbf{b}_2,$$

där \mathbf{b}_1 är ortogonal mot alla vektorer i V , och $\mathbf{b}_2 \in V$. Eftersom $\mathbf{b}_2 + \mathbf{y} \in V$ så är vektorerna \mathbf{b}_1 och $\mathbf{b}_2 + \mathbf{y}$ ortogonala. Alltså är

$$\|\mathbf{r}\|_2^2 = \|\mathbf{b} - \mathbf{y}\|_2^2 = \|\mathbf{b}_1 + \mathbf{b}_2 - \mathbf{y}\|_2^2 = \|\mathbf{b}_1\|_2^2 + \|\mathbf{b}_2 - \mathbf{y}\|_2^2,$$

enligt Pythagoras sats, se övningsuppgift 19. Eftersom \mathbf{b} är konstant är också \mathbf{b}_1 och \mathbf{b}_2 det. För att minimera $\|\mathbf{r}\|_2$ ska vi välja \mathbf{x} sådan att $\mathbf{b}_2 = \mathbf{y} = \mathbf{Ax}$, dvs residualen \mathbf{r} ska vara ortogonal mot V . Speciellt har vi att

$$\mathbf{a}_j^T \cdot \mathbf{r} = 0 \quad \text{for } j = 1, 2, \dots, m,$$

vilket kan i matrisform skrivas enligt

$$\mathbf{A}^T \mathbf{r} = \mathbf{0} \Leftrightarrow \mathbf{A}^T(\mathbf{b} - \mathbf{Ax}) = \mathbf{0} \Leftrightarrow \mathbf{A}^T \mathbf{Ax} = \mathbf{A}^T \mathbf{b}.$$

Den sista likheten kalla för *normalekvationerna* och om kolonnerna i \mathbf{A} är linjärt oberoende, så har normalekvationerna en entydig lösning. Notera att $\mathbf{A}^T \mathbf{A}$ är en symmetrisk matris av ordning m . Om $m = 2$, så är

$$\mathbf{A} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \dots & \dots \\ 1 & x_m \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} a \\ b \end{pmatrix} \quad \text{och} \quad \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}$$

att

$$\|\mathbf{r}\|_2^2 = \|\mathbf{b} - \mathbf{Ax}\|_2^2 = \sum_{k=1}^m (y_k - a - bx_k)^2.$$

Det vertikala avståndet från en punkt (x_k, y_k) till linjen $y = f(x)$ ges av just

$$|y_k - f(x_k)| = |y_k - a - bx_k|.$$

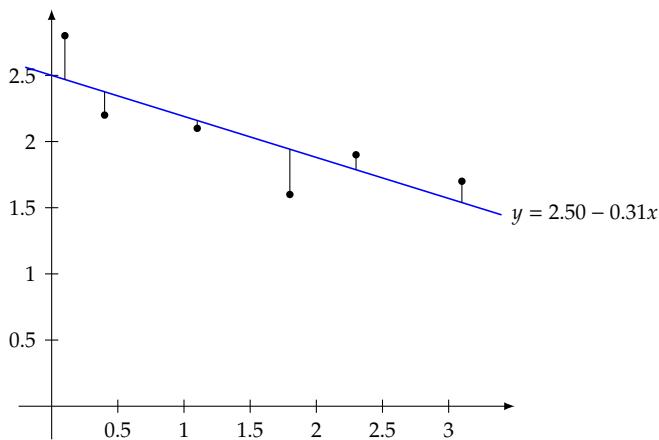
Att vi sedan kvadrerar samtliga termer gör att vi kan ta bort absolutbeloppet.

Exempel 5.14. Normalekvationerna för exempel 5.13 ges av

$$\mathbf{A}^T \mathbf{Ax} = \mathbf{A}^T \mathbf{b} \Leftrightarrow \begin{pmatrix} 6.0 & 8.8 \\ 8.8 & 19.52 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 12.3 \\ 15.99 \end{pmatrix}. \quad (5.6)$$

Lösningen är $(a, b) \approx (2.50, -0.31)$. Alltså är $f(x) = 2.50 - 0.31x$. Summan av kvadraterna av de vertikala avstånden från noderna till linjen, dvs residualerna, är så små som möjligt. En liten ändring av linje, lutning och/eller förskjutning, gör summan större, se figur 5.1. Residualerna $e_k = y_k - f(x_k)$ redovisas i tabell 5.1. I detta exempel ges *maximumfelet* E_∞ , *medelvärdesfelet* E_1 och *kvadratiska medelvärdesfelet* E_2 av

$$E_\infty(f) = \max_{1 \leq k \leq n} |e_k| \approx 0.346673,$$



Figur 5.1

k	x_k	y_k	$f(x_k)$	$ e_k $	$ e_k ^2$
1	0.1	2.8	2.47364	0.326361	0.1065110
2	0.4	2.2	2.38065	0.180645	0.0326327
3	1.1	2.1	2.16366	0.063659	0.0040525
4	1.8	1.6	1.94667	0.346673	0.1201820
5	2.3	1.9	1.79168	0.108317	0.0117325
6	3.1	1.7	1.54370	0.156300	0.0244298

Tabell 5.1

$$E_1(f) = \frac{1}{n} \sum_{k=1}^n |e_k| \approx 0.196993$$

och

$$E_2(f) = \left(\frac{1}{n} \sum_{k=1}^n |e_k|^2 \right)^{1/2} \approx 0.223436,$$

där $n = 6$ är antal givna punkter. \diamond

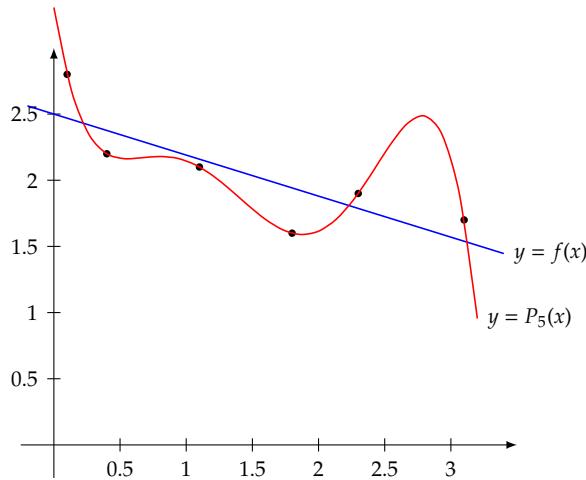
Anmärkning. Vi skulle istället kunna bestämma interpolationspolynomet P_n med någon av de metoder som vi studerade i kapitel 3. Eftersom vi har sex givna punkter så ska P_n vara av grad fem eller lägre. Oavsett metod får vi att

$$P_5(x) = 3.30545 - 6.16594x + 12.1319x^2 - 10.5696x^3 + 3.97178x^4 - 0.527442x^5.$$

Resultatet är ett polynom som oscillerar kraftigt, se figur 5.2.

Exempel 5.15. Det är möjligt att anpassa ett plan i minsta kvadratmetodens mening till en mängd punkter i rummet. Antag att vår data ges av

$$\begin{aligned} (-1.0, -1.3, 1.3) & \quad (-0.9, 0.3, 2.0) & (-1.1, 0.6, 2.8) \\ (-0.2, -0.6, 0.0) & \quad (0.4, -0.1, 1.4) & (-0.1, 0.6, 2.0) \\ (1.3, -0.8, -1.1) & \quad (0.9, -0.3, 0.3) & (0.8, 1., 0.6), \end{aligned}$$



Figur 5.2

dvs punkter på formen (x_k, y_k, z_k) . Låt $f(x, y) = a + bx + cy$. Då motsvarar $f(x_k, y_k) = z_k$, där $k = 1, 2, \dots, 9$ av ekvationssystemet

$$\begin{cases} a - 1.0b - 1.3b = 1.3 \\ a - 0.9b + 0.3b = 2.0 \\ a - 1.1b + 0.6b = 2.8 \\ a - 0.2b - 0.6b = 0.0 \\ a + 0.4b - 0.1b = 1.4 \\ a - 0.1b + 0.6b = 2.0 \\ a + 1.3b - 0.8b = -1.1 \\ a + 0.9b - 0.3b = 0.3 \\ a + 0.8b + 1.0b = 0.6. \end{cases}$$

Låt \mathbf{A} och \mathbf{b} vara koefficientmatrisen respektive högerledet i ekvationssystemet samt låt $\mathbf{x} = (a, b, c)$. Normalekvationerna ges då av

$$\mathbf{A}^T \mathbf{Ax} = \mathbf{A}^T \mathbf{b} \Leftrightarrow \begin{pmatrix} 9.00 & 0.10 & -0.60 \\ 0.10 & 6.37 & -0.12 \\ -0.60 & -0.12 & 4.60 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} 9.30 \\ -6.50 \\ 3.04 \end{pmatrix}.$$

Det ger att $a \approx 1.097$, $b \approx -1.023$ och $c \approx 0.777$. Det plan som bäst approximerar punkterna i minsta kvadratmetodens mening ges av $z = 1.097 - 1.023x + 0.777y$. \diamond

5.5.1 En alternativ metod att härleda normalekvationerna



Låt $f(x) = a + bx$ och

$$E(a, b) = \sum_{k=1}^n e_k^2 = \sum_{k=1}^n |f(x_k) - y_k|^2 = \sum_{k=1}^n (a + bx_k - y_k)^2.$$

Med andra ord är E en funktion som beror på a och b , som är summan av kvadraterna av de vertikala avstånden från punkterna (x_k, y_k) och linjen $y = f(x)$. Vi har att

$$\frac{\partial E}{\partial a} = \sum_{i=1}^n 2(a + bx_i - y_i) \quad \text{och} \quad \frac{\partial E}{\partial b} = \sum_{i=1}^n 2x_i(a + bx_i - y_i).$$

Vi minimera $E(a, b)$ genom att lösa ekvationssystemet

$$\begin{cases} \frac{\partial E}{\partial a} = 0 \\ \frac{\partial E}{\partial b} = 0, \end{cases} \Leftrightarrow \begin{cases} a \sum_{i=1}^n 1 + b \sum_{i=1}^n x_i - \sum_{i=1}^n y_i = 0 \\ a \sum_{i=1}^n x_i + b \sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i y_i = 0. \end{cases}$$

Notera att första summan $\sum_{i=1}^n 1$ är lika med antal givna punkter, dvs n , och det erhållna ekvationssystemet är linjärt.

Exempel 5.16. Med samma förutsättning som i exempel 5.13 får vi att

$$\begin{aligned} \sum_{i=1}^6 1 &= 6, & \sum_{i=1}^6 x_i &= 8.8, & \sum_{i=1}^6 y_i &= 12.3, \\ \sum_{i=1}^6 x_i^2 &= 19.52 & \text{och} & & \sum_{i=1}^6 x_i y_i &= 15.99. \end{aligned}$$

Alltså ges normalekvationerna av

$$\begin{cases} 6a + 8.8b - 12.3 = 0 \\ 8.8a + 19.52b - 15.99 = 0. \end{cases}$$

Jämför med ekvationssystemet (5.6). ◊

5.5.2 Kurvanpassning med polynom

Exempel 5.17. Låt



$$\begin{aligned} (x_1, y_1) &= (-1, 5), & (x_2, y_2) &= (0, -1), & (x_3, y_3) &= (1, 1), \\ (x_4, y_4) &= (2, 2) & \text{och} & & (x_5, y_5) &= (3, 7). \end{aligned}$$

Vi vill bestämma det andragradspolynomet $f(x) = a + bx + cx^2$ som bästa anpassa punkterna enligt minsta kvadratmetoden. Likheterna $f(x_k) = y_k$ för $k = 1, 2, 3, 4, 5$ ger ekvationssystemet

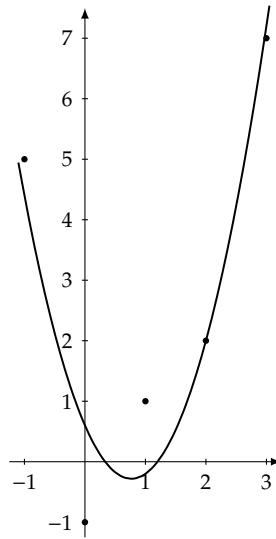
$$\begin{cases} f(-1) = 5 \\ f(0) = -1 \\ f(1) = 1 \\ f(2) = 2 \\ f(3) = 7 \end{cases} \Leftrightarrow \begin{cases} a - b + c = 5 \\ a + 0 + 0 = -1 \\ a + b + c = 1 \\ a + 2b + 4c = 2 \\ a + 3b + 9c = 7. \end{cases}$$

Låt A vara koefficientmatrisen för ekvationssystemet, $x = (a, b, c)$ och y högerledet. Då ges normalekvationerna av

$$A^T A x = A^T y \Leftrightarrow \begin{pmatrix} 5 & 5 & 15 \\ 5 & 15 & 35 \\ 15 & 35 & 99 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} 14 \\ 21 \\ 77 \end{pmatrix}.$$

Detta ekationsystem av lösningen

$$a = \frac{3}{5}, \quad b = -\frac{23}{10} \quad \text{och} \quad c = \frac{3}{2}.$$



Figur 5.3

Alltså är

$$f(x) = \frac{3}{5} - \frac{23}{10}x + \frac{3}{2}x^2$$

det andragradspolynom som bäst anpassar de givna punkterna i minsta kvadratmetodens mening, se figur 5.3. \diamond

Exempel 5.18. Givet följande statistiska data på formen (x_k, y_k) , där $k = 1, 2, \dots, 9$.

$$(0, 4.7), \quad (0.5, 1.5), \quad (1, 0.1), \quad (1.5, 0.5), \quad (2, 1.1), \\ (2.5, 0.6), \quad (3, 0.2), \quad (3.5, 0.8) \quad \text{och} \quad (4, 4.2)$$

Vi vill approximera data med ett polynom av grad fyra, dvs med ett polynom på formen

$$f(x) = c_0 f_0(x) + c_1 f_1(x) + c_2 f_2(x) + c_3 f_3(x) + c_4 f_4(x),$$

där $f_0(x) = 1$ och $f_i(x) = x^i$ för $i = 1, 2, 3, 4$. Alltså är

$$f(x) = c_0 \cdot 1 + c_1 \cdot x + c_2 \cdot x^2 + c_3 \cdot x^3 + c_4 \cdot x^4.$$

Låt $x = (c_0, c_1, c_2, c_3, c_4)$ och

$$y = (4.7, 1.5, 0.1, 0.5, 1.1, 0.6, 0.2, 0.8, 4.2).$$

Då ges koefficientmatrisen A i ekvationen $Ax = y$ av

$$A = \begin{pmatrix} f_0(x_1) & f_1(x_1) & f_2(x_1) & f_3(x_1) & f_4(x_1) \\ f_0(x_2) & f_1(x_2) & f_2(x_2) & f_3(x_2) & f_4(x_2) \\ f_0(x_3) & f_1(x_3) & f_2(x_3) & f_3(x_3) & f_4(x_3) \\ f_0(x_4) & f_1(x_4) & f_2(x_4) & f_3(x_4) & f_4(x_4) \\ f_0(x_5) & f_1(x_5) & f_2(x_5) & f_3(x_5) & f_4(x_5) \\ f_0(x_6) & f_1(x_6) & f_2(x_6) & f_3(x_6) & f_4(x_6) \\ f_0(x_7) & f_1(x_7) & f_2(x_7) & f_3(x_7) & f_4(x_7) \\ f_0(x_8) & f_1(x_8) & f_2(x_8) & f_3(x_8) & f_4(x_8) \\ f_0(x_9) & f_1(x_9) & f_2(x_9) & f_3(x_9) & f_4(x_9) \end{pmatrix}$$

$$\begin{aligned}
 &= \begin{pmatrix} f_0(0) & f_1(0) & f_2(0) & f_3(0) & f_4(0) \\ f_0(0.5) & f_1(0.5) & f_2(0.5) & f_3(0.5) & f_4(0.5) \\ f_0(1) & f_1(1) & f_2(1) & f_3(1) & f_4(1) \\ f_0(1.5) & f_1(1.5) & f_2(1.5) & f_3(1.5) & f_4(1.5) \\ f_0(2) & f_1(2) & f_2(2) & f_3(2) & f_4(2) \\ f_0(2.5) & f_1(2.5) & f_2(2.5) & f_3(2.5) & f_4(2.5) \\ f_0(3) & f_1(3) & f_2(3) & f_3(3) & f_4(3) \\ f_0(3.5) & f_1(3.5) & f_2(3.5) & f_3(3.5) & f_4(3.5) \\ f_0(4) & f_1(4) & f_2(4) & f_3(4) & f_4(4) \end{pmatrix} \\
 &= \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 0.5 & 0.25 & 0.125 & 0.0625 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1.5 & 2.25 & 3.375 & 5.0625 \\ 1 & 2 & 4 & 8 & 16 \\ 1 & 2.5 & 6.25 & 15.625 & 39.0625 \\ 1 & 3 & 9 & 27 & 81 \\ 1 & 3.5 & 12.25 & 42.875 & 150.063 \\ 1 & 4 & 16 & 64 & 256 \end{pmatrix}.
 \end{aligned}$$

Normalekvationerna $\mathbf{A}^T \mathbf{A} \mathbf{x} = \mathbf{A}^T \mathbf{y}$ är

$$\begin{pmatrix} 9 & 18 & 51 & 162 & 548.25 \\ 18 & 51 & 162 & 548.25 & 1930.5 \\ 51 & 162 & 548.25 & 1930.5 & 6983.81 \\ 162 & 548.25 & 1930.5 & 6983.81 & 25761.4 \\ 548.25 & 1930.5 & 6983.81 & 25761.4 & 96424.3 \end{pmatrix} \begin{pmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \\ c_4 \end{pmatrix} = \begin{pmatrix} 13.7 \\ 25.5 \\ 88.55 \\ 328.65 \\ 1255.21 \end{pmatrix}.$$

Lösningen ges av

$$c_0 \approx 4.799, c_1 \approx -11.69, c_2 \approx 10.66, c_3 \approx -3.859 \quad \text{och} \quad c_4 \approx 0.479.$$

Alltså är

$$f(x) = 4.799 - 11.69x + 10.66x^2 - 3.859x^3 + 0.479x^4.$$

Det kvadratiska-medelvärdesfelet är

$$E_2(f) = \left(\frac{1}{9} \sum_{i=1}^9 (f(x_i) - y_i)^2 \right)^{1/2} \approx 0.181526,$$

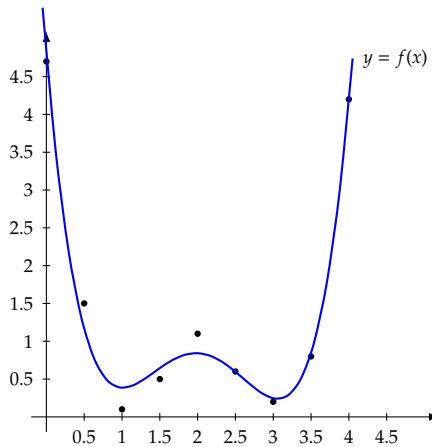
se även figur 5.4. ◊

5.5.3 Linjärisering

Exempel 5.19. De nio punkterna

$$\begin{aligned}
 &(-0.5, 0.6), (0.4, 0.75), (0.4, 1.2), (1.1, 1.5), (1.7, 3.2), \\
 &(2.4, 4.4), (2.8, 5.2), (3.3, 6.7) \quad \text{och} \quad (3.5, 9.1)
 \end{aligned}$$





Figur 5.4

verkar lika utmed en kurva ges av funktionen $f(x) = be^{ax}$, se figur 5.5. Om vi ställer upp likheterna $f(x_k) = y_k$ för $k = 1, 2, \dots, 9$ så får vi ekvationssystemet

$$\begin{cases} be^{-0.5a} = 0.6 \\ be^{0.4a} = 0.75 \\ be^{0.4a} = 1.2 \\ be^{1.1a} = 1.5 \\ be^{1.7a} = 3.2 \\ be^{2.4a} = 4.4 \\ be^{2.8a} = 5.2 \\ be^{3.3a} = 6.7 \\ be^{3.5a} = 9.1, \end{cases}$$

som inte är linjärt. Det betyder att vi inte kan skriva om det på matrisform och sedan bestämma normalekvationerna genom att multiplicera med transponatet av koefficientenmatrisen (den existerar inte). Genom att skriva om likheten $y = f(x)$ kan vi erhålla ett linjärt ekvationssystem. Processen kallas för *linjärisering*. I vårt exempel får vi att

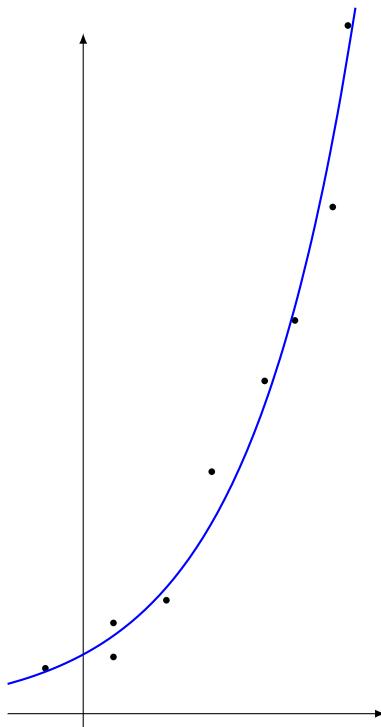
$$y = be^{ax} \Leftrightarrow \ln(y) = \ln(be^{ax}) = \ln(b) + ax.$$

Låt $A = a$, $B = \ln(b)$, $X = x$ och $Y = \ln(y)$. Då är

$$y = be^{ax} \Leftrightarrow Y = B + AX.$$

Istället för att approximera punkterna (x_i, y_i) med funktionen $f(x)$, approximerar vi punkterna (X_i, Y_i) med en linje. Vi måste beräkna varje Y_i , tex är

$$Y_2 = \ln(y_2) = \ln(0.75) \approx -0.287682.$$



Figur 5.5

Vi har transformerat likheterna $f(x_k) = y_k$ till $AX_k + B = Y_k$, vilka motsvarar ett linjärt ekvationssystem $\mathbf{Ax} = \mathbf{Y}$. Här är $x = (A, B)$, $X = (X_i)$,

$$\mathbf{A} = (\mathbf{X} \quad \mathbf{1}) = \begin{pmatrix} -0.5 & 1 \\ 0.4 & 1 \\ 0.4 & 1 \\ 1.1 & 1 \\ 1.7 & 1 \\ 2.4 & 1 \\ 2.8 & 1 \\ 3.3 & 1 \\ 3.5 & 1 \end{pmatrix} \quad \text{och} \quad \mathbf{Y} = (Y_i) = \begin{pmatrix} -0.510826 \\ -0.287682 \\ 0.182322 \\ 0.405465 \\ 1.16315 \\ 1.4816 \\ 1.64866 \\ 1.90211 \\ 2.20827 \end{pmatrix}.$$

Det ger oss normalekvationerna

$$\mathbf{A}^T \mathbf{Ax} = \mathbf{A}^T \mathbf{Y} \Leftrightarrow \begin{cases} 41.41A + 15.1B = 24.8146 \\ 15.1A + 9B = 8.19307. \end{cases}$$

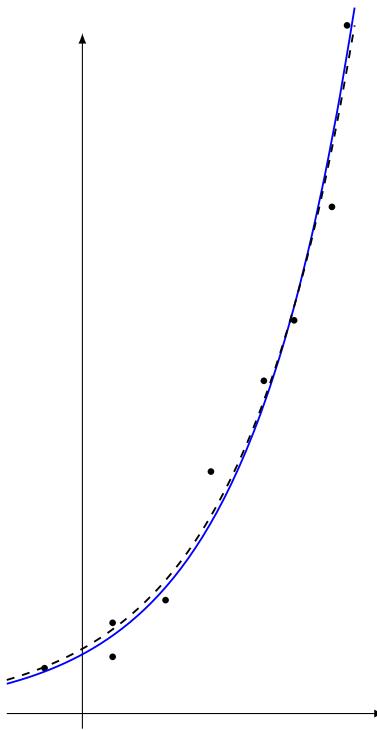
Lösningen är $A = 0.688529$ och $B = -0.244857$. Alltså är

$$a = A = 0.688529 \quad \text{och} \quad b = e^B \approx e^{-0.244857} \approx 0.782816.$$

Funktionen

$$f(x) = 0.782816e^{0.688529x}.$$

är en approximation av data i minsta kvadratmetodens mening, se figur 5.5. ◊



Figur 5.6. Heldragn kurvan motsvarar approximationen i exempel 5.19 och streckad kurvan motsvarar approximationen i exempel 5.20.

Exempel 5.20. Det är möjligt att approximera i minsta kvadratmetodens mening utan att linjärisera data. Med samma förutsättningar som i föregående exempel, låt

$$\begin{aligned} E(a, b) &= \sum_{i=1}^9 (f(x_i) - y_i)^2 = \sum_{i=1}^9 (be^{ax_i} - y_i)^2 \\ &= (be^{-0.5a} - 0.6)^2 + (be^{0.4a} - 1.2)^2 + (be^{0.4a} - 0.75)^2 \\ &\quad + (be^{1.1a} - 1.5)^2 + (be^{1.7a} - 3.2)^2 + (be^{2.4a} - 4.4)^2 \\ &\quad + (be^{2.8a} - 5.2)^2 + (be^{3.3a} - 6.7)^2 + (be^{3.5a} - 9.1)^2. \end{aligned}$$

Då är

$$\frac{\partial E}{\partial a} = 2 \sum_{i=1}^9 bx_i e^{ax_i} (be^{ax_i} - y_i) \quad \text{och} \quad \frac{\partial E}{\partial b} = 2 \sum_{i=1}^9 e^{ax_i} (be^{ax_i} - y_i).$$

Sätter vi de partiella derivatorna lika med noll erhåller vi ett icke-linjärt ekvations-system med a och b som obekanta. Vi kan lösa det med någon av det metoder som vi studerade tidigare (vilket lämnas som övning). Lösningen är $a \approx 0.656583$ och $b \approx 0.856082$. Det ger oss approximationen

$$f(x) = 0.856082e^{0.656583x},$$

se figur 5.6. ◊

Exempel 5.21. Antag att den enda information vi har om funktionen $f \in C[0.5, 3]$ är följande data

x	0.5	1	1.5	2	2.5	3
$f(x)$	0.2144	1.1625	1.9046	2.2403	2.3921	2.4250

och att

$$f(x) = \frac{x^3}{ax^3 + b}.$$

Vi vill bestämma a och b . Låt $X = 1/x^3$ och $Y = 1/y$. Om $x \neq 0$ och $y \neq 0$, så är

$$y = \frac{x^3}{ax^3 + b} \Leftrightarrow \frac{1}{y} = \frac{ax^3 + b}{x^3} \Leftrightarrow Y = A + BX,$$

där $A = a$ och $B = b$. Vi har linjäriserat problemet. Den transformerade data:

X	8	1	0.2963	0.125	0.064	0.0370
Y	4.6638	0.8602	0.5250	0.4464	0.4180	0.4124

Vi vill bestämma A och B så att $F(X) = A + BX$ approximerar data ovan enligt minsta kvadratmetoden. Det ger oss det linjära ekvationssystemet

$$\begin{cases} A + 8B = 4.6638 \\ A + B = 0.8602 \\ A + 0.2963B = 0.5250 \\ A + 0.125B = 0.4464 \\ A + 0.064B = 0.4180 \\ A + 0.0370B = 0.4124. \end{cases}$$

Normalekvationerna $\mathbf{A}^T \mathbf{A} \mathbf{x} = \mathbf{A}^T \mathbf{Y}$ ges därmed av

$$\begin{pmatrix} 6 & 9.5223 \\ 9.522 & 65.1089 \end{pmatrix} \begin{pmatrix} A \\ B \end{pmatrix} = \begin{pmatrix} 7.3258 \\ 38.4243 \end{pmatrix}.$$

Vi finner lösningen $A = 0.3793$ och $B = 0.5360$. Alltså är

$$f(x) = \frac{x^3}{0.3793x^3 + 0.5360},$$

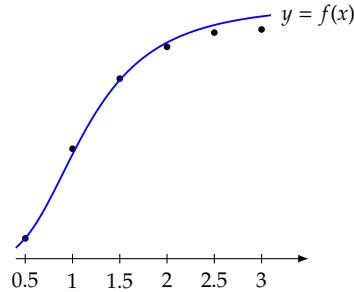
se figur 5.7. Felet är $\|f(x) - y\|_2 \approx 0.1766$. ◊

5.6 QR-faktorisering

En $m \times n$ -matris M säges vara *semiortogonal* om $M^T M = I_n$ eller $MM^T = I_m$. Notera att de två produkterna $M^T M$ och MM^T är en $n \times n$ -matris respektive en $m \times m$ -matris. Om $m = n$, så är en semiortogonal matris även *ortogonal*, dvs $M^T M = I_n = MM^T$ och därmed följer för ortogonala kvadratiska matriser att $M^T = M^{-1}$.

Låt $A = (a_{i,j})$ vara en $m \times n$ -matris, där $m \geq n$. Vi ska i detta avsnitt studera metoder att finna matriser Q och R sådana att

$$A = QR$$



Figur 5.7

och där Q är en kvadratisk matris av ordning m som är ortogonal, dvs $Q^T Q = Q Q^T = I$, och R är en övertriangulär $m \times n$ -matris. Om $m > n$, så innehåller de $m, -n$ sista raderna i R endast av 0:or. Lå R_1 vara den kvadratiska matrisen av ordning n som består av de n första raderna i R . Vidare delar vi upp Q i två delmatriser Q_1 och Q_2 , där den förstnämnda innehåller de n första kolonnerna i Q . Alltså har vi att

$$A = QR = (Q_1 \quad Q_2) \begin{pmatrix} R \\ O \end{pmatrix} = Q_1 R_1 + Q_2 O = Q_1 R_1,$$

där O är nollmatrisen med $m - n$ rader och n kolonner. Med andra ord har vi två olika faktoriseringar av A , vilka kallas för *full QR-faktoriseringar av A* respektive *kompakt QR-faktoriseringar av A*. Notera att Q_1 och Q_2 är båda semiortogonala.

5.6.1 Tillämpning av QR-faktorisering

Antag att vi vill lösa ekvationssystemet $Ax = b$ och har QR-faktoriserat matrisen A . Då kan vi skriva om ekvationssystemet enligt

$$Ax = b \Leftrightarrow QRx = b \Leftrightarrow Rx = Q^T b,$$

dvs till ett triangulärt ekvationssystem som vi enkelt löser med bakåtsubstitution, där det inte behövs någon Gausselimination.

Om vi betraktar vektorerna u och v som en kolonnmatriiser så följer det från definitionen av skalärprodukt att

$$(u, v) = u^T v.$$

Den Euklidiska vektornormen definieras som

$$\|v\|_2 = \sqrt{(v, v)}.$$

Om M är en ortogonal matris, så är

$$\|Mv\|_2^2 = (Mv, Mv) = (Mv)^T Mv = v^T M^T Mv = v^T v = (v, v) = \|v\|_2^2 \quad (5.7)$$

för alla vektorer v . För att finna den bästa approximationen i enlighet med minsta kvadratmetoden till det överbestämda ekvationssystemet $Ax = b$ vill minimera residualen r . Vi har att

$$\|r\|_2^2 = \|b - Ax\|_2^2 = \|b - QRx\|_2^2 = \|Q(Q^T b - Rx)\|_2^2 = \|Q^T b - Rx\|_2^2$$

$$= \left\| \begin{pmatrix} Q_1^T b \\ Q_2^T b \\ \vdots \\ 0 \end{pmatrix} - \begin{pmatrix} R_1 x \\ 0 \end{pmatrix} \right\|_2^2 = \|Q_1^T b - R_1 x\|_2^2 + \|Q_2^T b\|_2^2,$$

där den sista likheten följer från definitionen av den Euklidiska vektornormen. För att få r så liten som möjligt ska vi således välja x så att $R_1 x = Q_1^T b$, dvs vi ska lösa ett triangulärt ekvationssystem.

Exempel 5.22. QR-faktorisering av koefficientmatrisen A i exempel 5.13 ger resultatet

$$A = Q_1 R_1 = \begin{pmatrix} -0.408248 & -0.531438 \\ -0.408248 & -0.414781 \\ -0.408248 & -0.142581 \\ -0.408248 & 0.129619 \\ -0.408248 & 0.324047 \\ -0.408248 & 0.635133 \end{pmatrix} \begin{pmatrix} -2.44949 & -3.59258 \\ 0 & 2.57164 \end{pmatrix}.$$

Här lämnas detaljerna som övning (se följande två avsnitt för beskrivning av metoder att bestämma faktoriseringen). Det ger oss ekvationssystemet

$$R_1 x = Q_1^T b \Leftrightarrow \begin{pmatrix} -2.44949 & -3.59258 \\ 0 & 2.57164 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} -5.021450 \\ -0.797157 \end{pmatrix},$$

vilket har lösningen $a = 2.50464$ och $b = -0.30998$. \diamond

5.6.2 Gram-Schmidts ortogonaliseringprocess

Låt v_1, v_2, \dots, v_n vara vektorer i \mathbb{R}^m och antag att de är linjärt oberoende. Med *Gram-Schmidts ortogonalitetsprocess* kan man bestämma vektorer w_1, w_2, \dots, w_n som är normerade och parvist ortogonala, dvs

$$(w_i, w_i) = 1 \quad \text{och} \quad (w_i, w_j) = 0, \quad i \neq j,$$

samt sådana att

$$\begin{aligned} v_1 &= a_{1,1} w_1 \\ v_2 &= a_{1,2} w_1 + a_{2,2} w_2 \\ v_3 &= a_{1,3} w_1 + a_{2,3} w_2 + a_{3,3} w_3 \\ &\vdots \\ v_n &= a_{1,n} w_1 + a_{2,n} w_2 + a_{3,n} w_3 + \cdots + a_{n,n} w_n, \end{aligned}$$

där $a_{i,j} \in \mathbb{R}$. Speciellt är samtliga $a_{i,i} \neq 0$ eftersom de två uppsättningarna av vektorer är varför sig linjärt oberoende. För att bestämma w_1, w_2, \dots, w_n går man till väga på följande vis. Till att börja med sätt

$$w_1 = \frac{1}{\|v_1\|_2} v_1.$$

För $i = 2, 3, \dots, n$ sätt

$$u_i = v_i - (v_i, w_1) w_1 - (v_i, w_2) w_2 - \cdots - (v_i, w_{i-1}) w_{i-1}$$

och

$$w_i = \frac{1}{\|u_i\|_2} u_i.$$

För ett bevis att dessa w_1, w_2, \dots, w_n uppfyller ovanstående krav hänvisas läsaren till en kurs i linjär algebra.

Låt $A = (A_1 \ A_2 \ \dots \ A_n)$ och antag att kolonnerna A_1, A_2, \dots, A_n i A är linjärt oberoende. Vi kan med hjälp av Gram-Schmidts ortogonaliseringssprocess bestämma en uppsättning normerade och parvist ortogonala vektorer Q_1, Q_2, \dots, Q_n så att

$$\begin{aligned} A_1 &= r_{1,1}Q_1 \\ A_2 &= r_{1,2}Q_1 + r_{2,2}Q_2 \\ &\vdots \\ A_n &= r_{1,n}Q_1 + r_{2,n}Q_2 + \dots + r_{n,n}Q_n. \end{aligned} \tag{5.8}$$

där

$$r_{i,j} = (Q_i, A_j)$$

eftersom Q_1, Q_2, \dots, Q_n är en ortonormerad uppsättning av vektorer. Bilda matriserna

$$Q = (Q_1 \ Q_2 \ \dots \ Q_n) \quad \text{och} \quad R = \begin{pmatrix} r_{1,1} & r_{1,2} & r_{1,3} & \cdots & r_{1,n} \\ 0 & r_{2,2} & r_{2,3} & \cdots & r_{2,n} \\ 0 & 0 & r_{3,3} & \cdots & r_{3,n} \\ \dots & \dots & \dots & \ddots & \dots \\ 0 & 0 & 0 & \cdots & r_{n,n} \end{pmatrix}.$$

Då följer från (5.8) att

$$A = QR.$$

Elementet på rad i och kolonn j i $Q^T Q$ är lika med

$$Q_i^T Q_j = 1 \quad \text{eller} \quad Q_i^T Q_j = 0,$$

då $i = j$ respektive $i \neq j$, dvs ty $Q^T Q = I$. Vi har *QR-faktoriserat* matrisen A .

Exempel 5.23 (Kompakt QR-faktorisering). Låt

$$A = \begin{pmatrix} 0.4 & 1.3 & -1.1 \\ 4.1 & 5.2 & 8.6 \\ -0.4 & -0.1 & 6.2 \\ 3.7 & 4.8 & 8.3 \\ 5.2 & 0.3 & 6.4 \end{pmatrix}.$$

Då är

$$A_1 = \begin{pmatrix} 0.4 \\ 4.1 \\ -0.4 \\ 3.7 \\ 5.2 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 1.3 \\ 5.2 \\ -0.1 \\ 4.8 \\ 0.3 \end{pmatrix} \quad \text{och} \quad A_3 = \begin{pmatrix} -1.1 \\ 8.6 \\ 6.2 \\ 8.3 \\ 6.4 \end{pmatrix}.$$

Första kolumnen i Q ges av

$$Q_1 = \frac{1}{\|A_1\|_2} A_1 = \begin{pmatrix} 0.0525861 \\ 0.539007 \\ -0.0525861 \\ 0.486421 \\ 0.683619 \end{pmatrix}.$$

Härnäst bestämmer vi andra kolumnen, dvs

$$q_2 = A_2 - (A_2, Q_1)Q_1 = A_2 - (A_2^T Q_1)Q_1$$

och

$$Q_2 = \frac{1}{\|q_2\|_2} q_2 = \begin{pmatrix} 0.213861 \\ 0.480428 \\ 0.0389361 \\ 0.456165 \\ -0.716833 \end{pmatrix}.$$

Den tredje och sista kolonnen i Q ges av

$$q_3 = A_3 - (A_3, Q_1)Q_1 - (A_3, Q_2)Q_2$$

och

$$Q_3 = \frac{1}{\|q_3\|_2} q_3 = \begin{pmatrix} -0.343999 \\ 0.0237456 \\ 0.93456 \\ 0.0857486 \\ 0.0186148 \end{pmatrix}.$$

Alltså är

$$Q = (Q_1 \quad Q_2 \quad Q_3) = \begin{pmatrix} 0.0525861 & 0.213861 & -0.343999 \\ 0.539007 & 0.480428 & 0.0237456 \\ -0.0525861 & 0.0389361 & 0.93456 \\ 0.486421 & 0.456165 & 0.0857486 \\ 0.683619 & -0.716833 & 0.0186148 \end{pmatrix}$$

Det återstår att bestämma matrisen R. Matriselementen ges av skalärprodukterna

$$\begin{aligned} r_{1,1} &= Q_1^T A_1 = 7.60658 & r_{1,2} &= Q_1^T A_2 = 5.41637 & r_{1,3} &= Q_1^T A_3 = 12.664 \\ r_{2,2} &= Q_2^T A_2 = 4.74689 & r_{2,3} &= Q_2^T A_3 = 3.33627 & r_{3,3} &= Q_3^T A_3 = 7.20773. \end{aligned}$$

Således är

$$R = \begin{pmatrix} 7.60658 & 5.41637 & 12.664 \\ 0 & 4.74689 & 3.33627 \\ 0 & 0 & 7.20773 \end{pmatrix}$$

och $A = QR$ samt $Q^T Q = I$. ◊

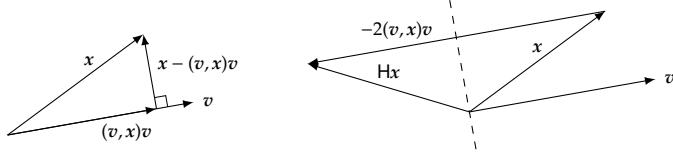
5.6.3 Householderspeglingar

Låt $v = (v_1 \quad v_2 \quad \cdots \quad v_m)^T \in \mathbb{R}^m$ vara en enhetsvektor, dvs $\|v\|_2 = 1$. Bilda matrisen

$$H = I - 2vv^T = \begin{pmatrix} 1 - 2v_1^2 & -2v_1v_2 & -2v_1v_3 & \cdots & -2v_1v_m \\ -2v_2v_1 & 1 - 2v_2^2 & -2v_2v_3 & \cdots & -2v_2v_m \\ -2v_3v_1 & -2v_3v_2 & 1 - 2v_3^2 & \cdots & -2v_3v_m \\ \dots & \dots & \dots & \dots & \dots \\ -2v_mv_1 & -2v_mv_2 & -2v_mv_3 & \cdots & 1 - 2v_m^2 \end{pmatrix}.$$

Vi ser att H är symmetrisk, vilket också följer direkt med hjälp av räknelagarna för transponat enligt

$$H^T = (I - 2vv^T)^T = I^T - 2(vv^T)^T = I - 2(v^T)^T v^T = I - 2vv^T = H.$$



Figur 5.8

Vidare är

$$H^T H = H^2 = (I - 2vv^T)(I - 2vv^T) = I^2 - 4vv^T + 4vv^T vv^T = I - 4vv^T + 4vv^T = I,$$

eftersom $v^T v = \|v\|_2^2 = 1$. Alltså är H semiortogonal. Men eftersom H är en kvadratisk matris, så är H därför ortogonal. Vi har således visat att $H^{-1} = H^T = H$. Eftersom H är ortogonal gäller att

$$\|Hx\|_2 = \|x\|_2$$

för alla $x \in \mathbb{R}^m$. Med andra ord, en linjär avbildning med H som avbildningsmatris är *isometrisk*, dvs den bevarar längd. Men vad för typ av avbildning motsvarar H ? Vi utgår här från att matrisen för avbildningen är iven med avseende på standardbasen. Låt u vara den ortogonala projektionen av $x \in \mathbb{R}^m$ i v . Enligt projekionsformeln är

$$u = \frac{(v, x)}{\|v\|_2^2} v = (v, x)v,$$

eftersom $\|v\|_2^2 = 1$, se även vänstra bilden i figur 5.8. För varje vektor $x \in \mathbb{R}^m$ är

$$Hx = (I - 2vv^T)x = Ix - 2vv^T x = x - 2(v^T x)v = x - 2(v, x)v,$$

dvs den ortogonala speglingen av x i de plan som är ortogonalt mot v , se högra bilden i figur 5.8 där den streckade linjen motsvarar planet som är ortogonalt mot v . Denna avbildning kallas för *Householderspegling*. Låt $x, y \in \mathbb{R}^m$, där $\|x\|_2 = \|y\|_2$, och sätt

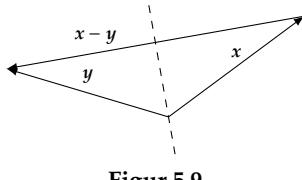
$$v = \frac{1}{\|x - y\|_2}(x - y).$$

Då är

$$x^T y = (x, y) = (y, x) = y^T x \quad \text{och} \quad x^T x = \|x\|_2^2 = \|y\|_2^2 = y^T y.$$

Det ger att

$$\begin{aligned} 2(x - y)(x - y)^T x &= 2(xx^T x - xy^T x - yx^T x + yy^T x) \\ &= xx^T x - xy^T x - yx^T x + yy^T x \\ &\quad + xx^T x - xy^T x - yx^T x + yy^T x \\ &= xx^T x - xy^T x - yx^T x + yy^T x \\ &\quad + xy^T y - xx^T y - yy^T y + yx^T y \\ &= (x^T x - x^T y - y^T x + y^T y)x \\ &\quad - (x^T x - x^T y - y^T x + y^T y)y \\ &= (x - y)^T (x - y)x - (x - y)^T (x - y)y \\ &= \|x - y\|_2^2 (x - y). \end{aligned}$$



Figur 5.9

Alltså är

$$Hx = (I - 2vv^T)x = x - \frac{\|x - y\|_2^2}{\|x - y\|_2^2}(x - y) = y,$$

dvs en Householderspeglings som avbildar x på y . Man kan även visa att så är fallet med ett resonemang enligt följande. Vektorn v är parallell med $x - y$ och eftersom x och y har samma längd är y speglingen av x i det plan som är ortogonalt mot v , se figur 5.9.

Låt A vara en $m \times n$ -matris. Man kan med Householderspeglings steg för steg bestämma matrisen R i QR-faktoriseringen av A enligt

$$H_n \cdots H_2 H_1 A = R.$$

Vi vill att när vi multiplicerar $H_{k-1} \cdots H_1 A$ med H_k ska det lämna de $k-1$ första kolonnerna i $H_{k-1} \cdots H_1 A$ oförändrade medan kolonn k ska bli motsvarande kolonn i den sökta matrisen R , dvs nollar under diagonalen. Låt I_{k-1} vara enhetsmatrisen av ordning $k-1$ och skriv matriserna H_k och $H_{k-1} \cdots H_1 A$ på blockform enligt

$$H_k = \begin{pmatrix} I_{k-1} & O \\ O & H' \end{pmatrix} \quad \text{respektive} \quad H_{k-1} \cdots H_1 A = \begin{pmatrix} A_{1,1} & A_{1,2} \\ O & A_{2,2} \end{pmatrix},$$

där

- H' är en $(m-k+1) \times (m-k+1)$ -matris,
- $A_{1,1}$ är en övertriangulär $(k-1) \times (k-1)$ -matris,
- $A_{1,2}$ är en $(k-1) \times (n-k+1)$ -matris,
- $A_{2,2}$ är en $(m-k+1) \times (n-k+1)$ -matris

och där O representerar nollmatriser av olika typ. Då är

$$H_k H_{k-1} \cdots H_1 A = \begin{pmatrix} A_{1,1} & A_{1,2} \\ O & H' A_{2,2} \end{pmatrix},$$

vilket är det vi önskade. Vidare vill vi att första kolonnen i $H' A_{2,2}$ ges av $(r_{k,k}, 0, \dots, 0)$. Med denna konstruktion av Householdermatriserna motsvarar $(A_{1,1} \ A_{2,2})$ de $k-1$ första raderna i R , dvs steg för steg erhåller vi rader och kolonner i R . Låt v_k beteckna den vektorn som används för att generera H_k . För att få önskad form på matrisen H_k måste de $k-1$ första elementen i vektorn v_k vara lika med 0. Låt

$$A_k = (r_{1,k} \ \dots \ r_{k-1,k} \ a_{k,k} \ \dots \ a_{m,k})^T$$

vara kolonn k i $H_{k-1} \cdots H_1 A$. Sätt

$$A'_k = (0 \ \dots \ 0 \ a_{k,k} \ \dots \ a_{m,k})^T.$$

Låt e_k vara kolonn k i enhetsmatrisen av ordning m . Vi vill avbilda A'_k på $r_{k,k}e_k$. För att kunna bestämma v_k måste vektorerna A'_k och $r_{k,k}e_k$ ha samma längd. Det får vi genom att sätta

$$r_{k,k} = \|A'_k\|_2.$$

Då är

$$v_k = \frac{1}{\|A'_k - r_{k,k}e_k\|_2} (A'_k - r_{k,k}e_k)$$

och därmed är H_k bestämd. Sätt $Q^T = H_n \cdots H_2 H_1$, dvs $Q^T A = R$. Då är

$$Q = (Q^T)^T = (H_n \cdots H_2 H_1)^T = H_1^T H_2^T \cdots H_n^T$$

och eftersom samtliga H_k är ortogonala följer det att

$$Q^T Q = H_n \cdots H_2 H_1 H_1^T H_2^T \cdots H_n^T = I,$$

dvs matrisen Q är ortogonal och därmed är $A = QR$.

Exempel 5.24 (Full QR-faktorisering). Låt

$$A = \begin{pmatrix} 0.4 & 1.3 & -1.1 \\ 4.1 & 5.2 & 8.6 \\ -0.4 & -0.1 & 6.2 \\ 3.7 & 4.8 & 8.3 \\ 5.2 & 0.3 & 6.4 \end{pmatrix},$$

se exempel 5.23. Vi berjar med att bestämma H_1 . Från

$$A'_1 = (0.4 \quad 4.1 \quad -0.4 \quad 3.7 \quad 5.2)^T \quad \text{och} \quad r_{1,1} = \|A'_1\|_2 = 7.60658$$

följer att

$$\begin{aligned} v_1 &= \frac{1}{\|A'_1 - r_{1,1}e_1\|_2} (A'_1 - r_{1,1}e_1) \\ &= (-0.688264 \quad 0.39157 \quad -0.038202 \quad 0.353368 \quad 0.496626)^T \end{aligned}$$

och

$$\begin{aligned} H_1 &= I - 2v_1 v_1^T \\ &= \begin{pmatrix} 0.0525861 & 0.539007 & -0.0525861 & 0.486421 & 0.683619 \\ 0.539007 & 0.693345 & 0.0299175 & -0.276737 & -0.388928 \\ -0.0525861 & 0.0299175 & 0.997081 & 0.0269987 & 0.0379442 \\ 0.486421 & -0.276737 & 0.0269987 & 0.750262 & -0.350984 \\ 0.683619 & -0.388928 & 0.0379442 & -0.350984 & 0.506726 \end{pmatrix}. \end{aligned}$$

Så här långt har vi att

$$H_1 A = \begin{pmatrix} 7.60658 & 5.41637 & 12.664 \\ 0 & 2.8581 & 0.769295 \\ 0 & 0.128478 & 6.96397 \\ 0 & 2.68658 & 1.23327 \\ 0 & -2.67022 & -3.53163 \end{pmatrix}$$

För att kunna bestämma H_2 behöver vi

$$A'_2 = \begin{pmatrix} 0 & 2.8581 & 0.128478 & 2.68658 & -2.67022 \end{pmatrix}^T$$

och

$$r_{2,2} = \|A'_2\|_2 = 4.74689.$$

Därmed är

$$\begin{aligned} v_2 &= \frac{1}{\|A'_2 - r_{2,2}e_2\|_2} (A'_2 - r_{2,2}e_2) \\ &= \begin{pmatrix} 0 & -0.446039 & 0.0303402 & 0.634435 & -0.630572 \end{pmatrix}^T \end{aligned}$$

och

$$H_2 = I - 2v_2v_2^T = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0.602099 & 0.0270658 & 0.565965 & -0.562519 \\ 0 & 0.0270658 & 0.998159 & -0.0384977 & 0.0382633 \\ 0 & 0.565965 & -0.0384977 & 0.194985 & 0.800114 \\ 0 & -0.562519 & 0.0382633 & 0.800114 & 0.204758 \end{pmatrix}.$$

Det andra steget av faktoriseringen slutförs genom att multiplicera med H_2 . Det ger

$$H_2 H_1 A = \begin{pmatrix} 7.60658 & 5.41637 & 12.664 \\ 0 & 4.74689 & 3.33627 \\ 0 & 0 & 6.78936 \\ 0 & 0 & -2.41794 \\ 0 & 0 & 0.0973473 \end{pmatrix}.$$

Det återstår att bestämma H_3 . Vi har att

$$A'_3 = \begin{pmatrix} 0 & 0 & 6.78936 & -2.41794 & 0.0973473 \end{pmatrix}^T$$

och

$$r_{3,3} = \|A'_3\|_2 = 7.20773.$$

Det ger i sin tur att

$$v_3 = \frac{1}{\|A'_3 - r_{3,3}e_3\|_2} (A'_3 - r_{3,3}e_3) = \begin{pmatrix} 0 & 0 & -0.170358 & -0.984585 & 0.0396398 \end{pmatrix}^T$$

och

$$H_3 = I - 2v_3v_3^T = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0.941956 & -0.335465 & 0.013506 \\ 0 & 0 & -0.335465 & -0.938813 & 0.0780575 \\ 0 & 0 & 0.013506 & 0.0780575 & 0.996857 \end{pmatrix}.$$

Alltså är

$$R = H_3 H_2 H_1 A = \begin{pmatrix} 7.60658 & 5.41637 & 12.664 \\ 0 & 4.74689 & 3.33627 \\ 0 & 0 & 7.20773 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

och

$$Q = H_1^T H_2^T H_3^T = \begin{pmatrix} 0.0525861 & 0.213861 & -0.343999 & -0.863139 & 0.296908 \\ 0.539007 & 0.480428 & 0.0237456 & -0.0932631 & -0.685126 \\ -0.0525861 & 0.0389361 & 0.93456 & -0.343627 & 0.065099 \\ 0.486421 & 0.456165 & 0.0857486 & 0.335468 & 0.659861 \\ 0.683619 & -0.716833 & 0.0186148 & -0.125202 & 0.0528479 \end{pmatrix}.$$

Det lämnas som övning att kontrollera att $A = QR$ och $Q^T Q = I$. \diamond

5.7 Singulärvärdesuppdelning

20160510

Låt $A = (a_{i,j})$ vara en matris av typen $m \times n$, där $m \geq n$. Om $m \leq n$, så studerar vi istället A^T . Vi vill finna matriser U, Σ och V sådana att

$$A = U\Sigma V^T$$

där U och V är kvadratiska och ortogonala matriser av ordningen m respektive n samt där Σ är en diagonalmatris av typen $m \times n$. Matrisfaktoriseringen kan illustreras enligt

$$\boxed{\quad} = \boxed{\quad} \boxed{\quad} \boxed{\quad} \boxed{\quad}.$$

Låt $\sigma_1, \sigma_2, \dots, \sigma_n$ vara diagonalelementen i Σ , och låt u_i och v_i beteckna den i :te kolonnen i U respektive V . De $m - n$ sista raderna i Σ består endast av 0:or och dessa multipliceras med elementen i kolonnerna $u_{n+1}, u_{n+2}, \dots, u_m$, dvs matrisprodukten beror inte på dessa kolonner. Alltså är

$$A = \sum_{k=1}^n \sigma_k u_k v_k^T.$$

Vi kan sortera diagonalelementen i Σ , de sk *singulära värdena till A*, enligt

$$\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_n.$$

Notera att σ_k är icke-negativa tal, se sats 5.13 och övningsuppgift 17.

Sats 5.13. *Antag att A kan singulärvärdesuppdelas, dvs $A = U\Sigma V^T$. Låt u_k och v_k beteckna den k :te kolonnen i U respektive V . Då gäller att*

- (a) $Av_k = \sigma_k u_k$
- (b) $A^T u_k = \sigma_k v_k$
- (c) σ_k^2 är ett egenvärde till $A^T A$ med v_k som tillhörande egenvektor
- (d) σ_k^2 är ett egenvärde till AA^T med u_k som tillhörande egenvektor.

Bevis. (a) Eftersom kolonnerna i V är ortonormala är $v_k^T v_k = 1$ och $v_i^T v_k = 0$, då $i \neq k$. Det ger att

$$Av_k = \left(\sum_{i=1}^n \sigma_i u_i v_i^T \right) v_k = \sum_{i=1}^n \sigma_i u_i v_i^T v_k = \sigma_k u_k.$$

(b) Följer på samma sätt som ovan eftersom

$$\mathbf{A}^T = \sum_{k=1}^n \sigma_k \mathbf{v}_k \mathbf{u}_k^T.$$

(c) Vi har att

$$\mathbf{A}^T \mathbf{A} \mathbf{v}_k = \mathbf{A}^T (\sigma_k \mathbf{u}_k) = \sigma_k \mathbf{A}^T \mathbf{u}_k = \sigma_k^2 \mathbf{v}_k,$$

vilket skulle visas. (d) Visas på samma sätt som ovan. \square

Antag att $\sigma_r \neq 0$ och $\sigma_{r+1} = \dots = \sigma_n = 0$ och bilda matrisen $\Sigma_1 = \text{diag}(\sigma_1, \sigma_2, \dots, \sigma_r)$. Låt \mathbf{U}_1 och \mathbf{V}_1 vara de matriser som består av de r första kolonnerna i \mathbf{U} respektive \mathbf{V} , och låt \mathbf{U}_2 och \mathbf{V}_2 vara de övriga kolonnerna i respektive matris. Då är

$$\mathbf{A} = (\mathbf{U}_1 \quad \mathbf{U}_2) \begin{pmatrix} \Sigma_1 & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{pmatrix} \begin{pmatrix} \mathbf{V}_1^T \\ \mathbf{V}_2^T \end{pmatrix} = \mathbf{U}_1 \Sigma_1 \mathbf{V}_1^T,$$

vilket kallas för *kompakt singulärvärdesuppdelning* av \mathbf{A} .

Sats 5.14. *För varje matris \mathbf{A} existerar det en singulärvärdesuppdelning $\mathbf{A} = \mathbf{U} \Sigma \mathbf{V}^T$.*

Bevis. Vi bevisar påståendet med induktion över dimensionen av \mathbf{A} . Antag till att börja med att $m \geq n = 1$. Sätt

$$\mathbf{U}_1 = \frac{1}{\|\mathbf{A}\|_2} \mathbf{A}, \quad \Sigma_1 = \|\mathbf{A}\|_2 \quad \text{och} \quad \mathbf{V} = 1.$$

Då är $\mathbf{A} = \mathbf{U}_1 \Sigma_1 \mathbf{V}^T$ en kompakt singulärvärdesuppdelning av \mathbf{A} . För att konstruera en full singulärvärdesuppdelning av \mathbf{A} väljer vi $m - 1$ vektorer i \mathbb{R}^m vilka tillsammans med kolonnen \mathbf{U}_1 bildar en bas för \mathbb{R}^m . Med Gram-Schmidts ortogonaliseringssprocess bestämmer vi en ortonormal bas för \mathbb{R}^m utifrån \mathbf{U}_1 och den valda vektorerna. Låt de nya vektorerna vara kolonnerna i matrisen \mathbf{U}_2 och bilda matrisen

$$\mathbf{U} = (\mathbf{U}_1 \quad \mathbf{U}_2).$$

Notera att \mathbf{U}_1 är en $m \times 1$ -matris och som vektor har normen 1. Då är \mathbf{U} en ortogonal kvadratisk matris av ordning m . Låt

$$\Sigma = (\Sigma_1 \quad 0 \quad \cdots \quad 0)^T$$

vara en $m \times 1$ -matris. Då är $\mathbf{A} = \mathbf{U} \Sigma \mathbf{V}^T$ en full singulärvärdesuppdelning av \mathbf{A} .

Antag att varje matris av typen $(m-1) \times (n-1)$ har en singulärvärdesuppdelning. Låt \mathbf{A} vara en $m \times n$ -matris, skild från nollmatrisen. Sätt $\sigma_1 = \|\mathbf{A}\|_2$, dvs σ_1 är kvadratroten av det största egenvärdet till $\mathbf{A}^T \mathbf{A}$. Eftersom mängden

$$S = \{x \in \mathbb{R}^n : \|x\|_2 = 1\}$$

är kompakt och funktionen $x \mapsto \|\mathbf{A}x\|_2$ är kontinuerlig och begränsad på S , så antar den sitt maximum i S , dvs det existerar ett $v_1 \in S \subset \mathbb{R}^n$ sådant att $\sigma_1 = \|\mathbf{A}v_1\|_2$. Sätt

$$\mathbf{u}_1 = \frac{1}{\|\mathbf{A}v_1\|_2} \mathbf{A}v_1 = \frac{1}{\sigma_1} \mathbf{A}v_1.$$

Vi kan med Gram-Schmidts ortogonaliseringssprocess och med utgångspunkt från vektorerna u_1 och v_1 bestämma en ortonormal bas för \mathbb{R}^m respektive \mathbb{R}^n samt bilda matriserna

$$U = \begin{pmatrix} u_1 & U_2 \end{pmatrix} \quad \text{och} \quad V = \begin{pmatrix} v_1 & V_2 \end{pmatrix}$$

vars kolonner är respektive uppsättning av basvektorer. Från konstruktionen följer att U och V är ortogonala kvadratiska matriser av ordning m respektive n . Då är

$$U^T AV = \begin{pmatrix} u_1^T \\ U_2^T \end{pmatrix} A \begin{pmatrix} v_1 & V_2 \end{pmatrix} = \begin{pmatrix} u_1^T A v_1 & u_1^T A V_2 \\ U_2^T A v_1 & U_2^T A V_2 \end{pmatrix}$$

Från definitionen av u_1 följer att

$$u_1 = \frac{1}{\sigma_1} Av_1 \quad \Leftrightarrow \quad \sigma_1 = u_1^T Av_1,$$

eftersom $u_1^T u_1 = (u_1, u_1) = \|u_1\|_2^2 = 1$. Första raden i $U^T AV$ ges sådels av

$$(\sigma_1 \quad u_1^T AV_2).$$

Vi har att

$$\left\| (\sigma_1 \quad u_1^T AV_2) \right\|_2 \geq \sigma_1.$$

Men eftersom U och V är ortogonala, så är

$$\begin{aligned} \|U^T AV\|_2 &= \sup_{x \neq 0} \frac{\|U^T AVx\|_2}{\|x\|_2} = \sup_{x \neq 0} \frac{\|AVx\|_2}{\|x\|_2} = \sup_{x \neq 0} \frac{\|AVx\|_2}{\|Vx\|_2} \cdot \frac{\|Vx\|_2}{\|x\|_2} \\ &= \sup_{x \neq 0} \frac{\|AVx\|_2}{\|Vx\|_2} \cdot \frac{\|x\|_2}{\|x\|_2} = \sup_{y \neq 0} \frac{\|Ay\|_2}{\|y\|_2} = \|A\|_2 = \sigma_1, \end{aligned}$$

enligt (5.7) och det faktum att V är inverterbar. Från

$$\|U^T AV\|_2 = \max_{\|x\|_2=1} \|U^T AVx\|_2$$

följer att vektornormen av en rad i $U^T AV$ är mindre än eller lika med matrisnormen av $U^T AV$. Alltså måste

$$\sigma_1 \leq \left\| (\sigma_1 \quad u_1^T AV_2) \right\|_2 \leq \|U^T AV\|_2 = \|A\|_2 = \sigma_1.$$

Således måste $u_1^T AV = \mathbf{0}^T$, ty annars är normen av första raden i $U^T AV$ större än σ_1 . Eftersom u_1 och Av_1 är parallella är varje kolonn i U_2 ortogonal mot Av_1 . Därmed följer det att $U_2^T Av_1 = \mathbf{0}$. Sätt $\tilde{A} = U_2^T AV_2$. Vi har funnit att

$$U^T AV = \begin{pmatrix} \sigma_1 & \mathbf{0}^T \\ \mathbf{0} & \tilde{A} \end{pmatrix},$$

där \tilde{A} är en $(m-1) \times (n-1)$ -matris. Enligt induktionsantagandet har \tilde{A} en singulärvärdesuppdelning, dvs $\tilde{A} = \tilde{U} \tilde{\Sigma} \tilde{V}^T$. Det ger att

$$U^T AV = \begin{pmatrix} \sigma_1 & \mathbf{0}^T \\ \mathbf{0} & \tilde{U} \tilde{\Sigma} \tilde{V}^T \end{pmatrix} = \begin{pmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & \tilde{U} \end{pmatrix} \begin{pmatrix} \sigma_1 & \mathbf{0}^T \\ \mathbf{0} & \tilde{\Sigma} \end{pmatrix} \begin{pmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & \tilde{V} \end{pmatrix}^T$$

eller ekvivalent

$$A = U \begin{pmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & \tilde{U} \end{pmatrix} \begin{pmatrix} \sigma_1 & \mathbf{0}^T \\ \mathbf{0} & \tilde{\Sigma} \end{pmatrix} \begin{pmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & \tilde{V} \end{pmatrix}^T V^T.$$

Det visar att A har en singulärvärdesuppdelning, ty

$$U \begin{pmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & \tilde{U} \end{pmatrix} \quad \text{och} \quad \begin{pmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & \tilde{V} \end{pmatrix}^T V^T$$

är ortogonala eftersom produkten av ortogonala matriser är i sig ortogonal. \square

Algoritm 5.2. Låt A vara en $m \times n$ -matris med full rang. Nedanstående algoritm bestämmer en singulärvärdesuppdelning av A .

1. Bestäm egenvärdena $\lambda_1, \lambda_2, \dots, \lambda_n$ och motsvarande egenvektorer v_1, v_2, \dots, v_n till matrisen $A^T A$, så att $\lambda_1 > \lambda_2 > \dots > \lambda_n$ och $\|v_1\|_2 = \|v_2\|_2 = \dots = \|v_n\|_2 = 1$.
2. Bilda matrisen $V = (v_1 \ v_2 \ \dots \ v_n)$.
3. Sätt $\sigma_k = \sqrt{\lambda_k}$, för $k = 1, 2, \dots, n$.
4. Bilda matrisen $\Sigma = \text{diag}(\sigma_1, \sigma_2, \dots, \sigma_n)$.
5. Sätt $u_k = \frac{1}{\sigma_k} A v_k$, för $k = 1, 2, \dots, n$.
6. Bilda matrisen $U = (u_1 \ u_2 \ \dots \ u_n)$.

Exempel 5.25. Låt

$$A = \begin{pmatrix} 0.4 & 1.3 & -1.1 \\ 4.1 & 5.2 & 8.6 \\ -0.4 & -0.1 & 6.2 \\ 3.7 & 4.8 & 8.3 \\ 5.2 & 0.3 & 6.4 \end{pmatrix}.$$

Då är

$$A^T A = \begin{pmatrix} 57.86 & 41.20 & 96.33 \\ 41.20 & 51.87 & 84.43 \\ 96.33 & 84.43 & 223.46 \end{pmatrix},$$

vilken har egenvärdena

$$\lambda_1 = 302.835, \lambda_2 = 17.7643 \quad \text{och} \quad \lambda_3 = 12.5904$$

med tillhörande normerade egenvektorer

$$\begin{aligned} v_1 &= (0.393198, 0.350515, 0.850021), \\ v_2 &= (-0.276824, -0.836459, 0.472974) \end{aligned}$$

respektive

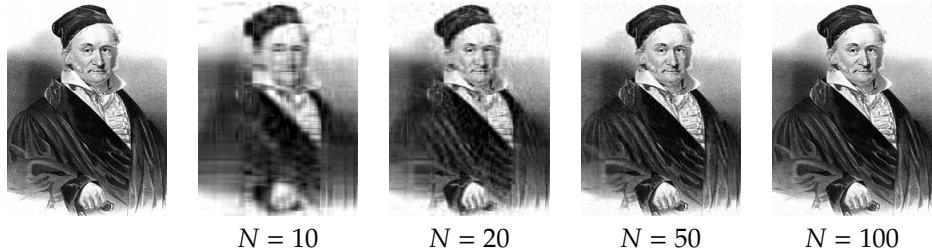
$$v_3 = (0.876792, -0.421278, -0.231863).$$

Alltså är

$$V = \begin{pmatrix} 0.393198 & -0.276824 & 0.876792 \\ 0.350515 & -0.836459 & -0.421278 \\ 0.850021 & 0.472974 & -0.231863 \end{pmatrix}$$

Från $\sigma_k = \sqrt{\lambda_k}$ följer att

$$\sigma_1 = 17.4022, \sigma_2 = 4.21477 \quad \text{och} \quad \sigma_3 = 3.5483.$$



Figur 5.10. Komprimering med singulärvärdesuppdelning

Därmed är

$$\Sigma = \begin{pmatrix} 17.4022 & 0 & 0 \\ 0 & 4.21477 & 0 \\ 0 & 0 & 3.5483 \end{pmatrix}.$$

Det återstår att bestämma matrisen U . Vi finner att

$$u_1 = \frac{1}{\sigma_1} Av_1 = (-0.0185077, 0.61745, 0.291791, 0.585701, 0.436148)$$

$$u_2 = \frac{1}{\sigma_2} Av_2 = (-0.407709, -0.336197, 0.741871, -0.264207, 0.317126)$$

och

$$u_3 = \frac{1}{\sigma_3} Av_3 = (0.0163754, -0.166227, -0.492107, -0.197973, 0.831107).$$

Således är

$$U = \begin{pmatrix} -0.0185077 & -0.407709 & 0.0163754 \\ 0.61745 & -0.336197 & -0.166227 \\ 0.291791 & 0.741871 & -0.492107 \\ 0.585701 & -0.264207 & -0.197973 \\ 0.436148 & 0.317126 & 0.831107 \end{pmatrix}$$

och $A = U\Sigma V^T$.

◊

Exempel 5.26 (Minsta kvadratmetoden). Vi vill minimera residualen

$$\begin{aligned} \|r\|_2^2 &= \|Ax - b\|_2^2 = \|U\Sigma V^T x - b\|_2^2 = \|U(\Sigma V^T x - U^T b)\|_2^2 \\ &= \|\Sigma V^T x - U^T b\|_2^2 = \|\Sigma_1 V_1^T x - U_1^T b - U_2^T b\|_2^2 \\ &= \|\Sigma_1 V_1^T x - U_1^T b\|_2^2 + \|U_2^T b\|_2^2. \end{aligned}$$

Således ska vi välja x så att

$$\Sigma_1 V_1^T x - U_1^T b = 0 \quad \Leftrightarrow \quad x = V_1 \Sigma_1^{-1} U_1^T b.$$

Om $\Sigma_1 = \text{diag}(\sigma_1, \sigma_2, \dots, \sigma_n)$, så är $\Sigma_1^{-1} = \text{diag}(1/\sigma_1, 1/\sigma_2, \dots, 1/\sigma_n)$. Matrisen $V_1 \Sigma_1^{-1} U_1^T$ kallas för *pseudoinversen till A*.

◊

Exempel 5.27 (Komprimering). Låt matrisen A representera den bild längst till vänster i figur 5.10, vilken består av 267×200 bildpunkter. En singulärvärdesuppdelning av A , dvs $A = U\Sigma V^T$, ger oss matriserna U , Σ och V , vilka är av typen 267×267 , 267×200 respektive 200×200 . De 200 elementen σ_k i diagonalen i Σ är

$$134.334, 24.5032, 20.6997, \dots, 0.131371, 0.118127, 0.109.$$

Låt u_k och v_k vara den k :te kolonnen i U respektive V . Då är

$$A = \sum_{k=1}^{200} \sigma_k u_k v_k^T.$$

Eftersom diagonalelementen i Σ är sorterade i fallande ordning har de försat singulära värdena σ_k mest betydelse för A . Vi kan trunkera summan enligt

$$A_N = \sum_{k=1}^N \sigma_k u_k v_k^T.$$

Resultatet för olika värden på N ser vi i andra till och med femte bilden i figur 5.10. Notera att A_N är precis som A en 267×200 -matris för varje positivt heltalet $N \leq 200$. Trunkeringen är detsamma som att behålla de N första kolonnerna i U och V samt de N första raderna och kolonnerna i Σ . \diamond

5.8 Egenvärdesproblem

Låt A vara en kvadratisk matris av ordning n . Antag att A har n olika egenvärden, vilka vi betecknar $\lambda_1, \lambda_2, \dots, \lambda_n$, och som uppfyller

$$|\lambda_1| > |\lambda_2| \geq |\lambda_3| \leq \dots \leq |\lambda_n|.$$

Vidare, låt v_1, v_2, \dots, v_n vara egenvektorer till respektive egenvärde, dvs

$$Av_k = \lambda_k v_k$$

för $k = 1, 2, \dots, n$. Eftersom samtliga egenvärden är olika är egenvektorerna linjärt oberoende och utgör därför en bas för \mathbb{R}^n . För varje vektor $v \in \mathbb{R}^n$ existerar det entydigt bestämda $a_1, a_2, \dots, a_n \in \mathbb{R}$ sådana att

$$v = a_1 v_1 + a_2 v_2 + \dots + a_n v_n.$$

Eftersom $v \mapsto Av$ är en linjär avbildning är

$$\begin{aligned} Av &= A(a_1 v_1 + a_2 v_2 + \dots + a_n v_n) = a_1 Av_1 + a_2 Av_2 + \dots + a_n Av_n \\ &= a_1 \lambda_1 v_1 + a_2 \lambda_2 v_2 + \dots + a_n \lambda_n v_n \end{aligned}$$

och

$$A^m v = a_1 \lambda_1^m v_1 + a_2 \lambda_2^m v_2 + \dots + a_n \lambda_n^m v_n,$$

eftersom $A^m v_k = \lambda_k^m v_k$ för varje positivt heltalet m . Vi har att

$$\left| \frac{\lambda_k}{\lambda_1} \right| < 1$$

då $k = 2, 3, \dots, n$. Det ger att

$$A^m v = \lambda_1^m \left(a_1 v_1 + a_2 \frac{\lambda_2^m}{\lambda_1^m} v_2 + \dots + a_n \frac{\lambda_n^m}{\lambda_1^m} v_n \right) \approx a_1 \lambda_1^m v_1$$

för tillräckligt stora m . Genom att välja vektorn v så att $a_1 \neq 0$ ser vi att $\mathbf{A}^m v$ närmar sig en vektor parallell med egenvektorn v_1 . Om λ är ett egenvärde till \mathbf{A} med tillhörande egenvektor v , så är

$$\lambda v = \mathbf{A}v \Leftrightarrow \lambda v^T v = v^T \mathbf{A}v \Leftrightarrow \lambda = \frac{v^T \mathbf{A}v}{v^T v}.$$

Den sista likheten kallas för *Reyleighs kvot*. Notera att $v^T v = \|v\|_2^2$.

Algoritm 5.3 (Potensmetoden). Låt \mathbf{A} vara en kvadratisk matris av ordning n som har n olika egenvärden. Följande iteration bestämmer en approximation av den dominerande egenparet (λ_1, v_1) till \mathbf{A} .

1. Tag $x_0 \in \mathbb{R}^n$ sådan att $\|x_0\|_2 = 1$.
2. För $k = 1, 2, 3, \dots$ sätt

$$y_k = \mathbf{A}x_{k-1}, \quad x_k = \frac{1}{\|y_k\|_2} y_k \quad \text{och} \quad l_k = x_k^T \mathbf{A}x_k.$$

Avbryt då $\|x_{k-1} - \text{sign}(l_k)x_k\|_2 < \varepsilon$.

3. Returnera $(\lambda_1, v_1) = (l_k, x_k)$.

Anmärkning. Om egenvärdet λ_1 är negativt är $\|x_{k-1} - x_k\|_2 \approx 2$, vilket är skälet till signumfunktionen i stoppkriteriet. För komplexa egenvärden måste en komplex vektor väljas om matrisen \mathbf{A} är reell, och stopkriteriet måste anpassas för komplexa tal, tex $\text{sign}(z) = z/|z|$ då $z \neq 0$.

Exempel 5.28. Låt

$$\mathbf{A} = \begin{pmatrix} 0 & -1 & 7 \\ 4 & 0 & 8 \\ 3 & 5 & 1 \end{pmatrix} \quad \text{och} \quad x_0 = \frac{1}{\sqrt{3}}(1, 1, 1).$$

Då är

$$y_1 = \mathbf{A}x_0 = (3.4641, 6.9282, 5.19615)$$

och

$$x_1 = \frac{1}{\|y_1\|_2} y_1 = (0.371391, 0.742781, 0.557086)$$

samt

$$l_1 = x_0^T \mathbf{A}x_0 = 8.58621.$$

Det ger att

$$\|x_0 - \text{sign}(l_1)x_1\|_2 = 0.264948.$$

Med $\varepsilon = 10^{-12}$ krävs det 55 iterationer innan stoppkriteriet är uppfyllt. Vi finner approximationerna

$$\lambda_1 \approx l_{55} = 8.90945 \quad \text{och} \quad v_1 \approx x_{55} = (0.386722, 0.706176, 0.593094)$$

för det dominerade egenparet till \mathbf{A} . ◊

Exempel 5.29. Egenvärdena till matrisen \mathbf{A} i exempel 5.28 är nollställen till polynomet

$$p(\lambda) = \det(\mathbf{A} - \lambda I) = 120 + 57\lambda + \lambda^2 - \lambda^3.$$

Förvisso kan vi bestämma det karakteristiska polynomet för en matris och sedan tex med Newtons metod lösa ekvationen $p(\lambda) = 0$. Men för stora matriser är de tidskrävande att beräkna determinanten och förenkla uttrycket. Om

$$\mathbf{A} = \begin{pmatrix} -19 & 11 & 10 & 1 & -4 & -10 & -12 & 20 & 12 & 0 \\ 8 & -8 & 20 & 3 & -11 & -7 & -17 & 19 & 7 & -13 \\ 12 & -10 & -7 & 14 & 18 & 13 & 11 & -19 & -2 & 11 \\ 17 & 9 & -9 & 3 & 11 & 11 & 4 & -18 & -8 & 7 \\ -8 & 10 & -4 & -16 & 7 & -6 & 13 & 17 & -7 & -8 \\ -15 & 13 & -14 & -17 & -2 & -1 & -1 & -13 & -20 & 13 \\ -7 & -15 & -16 & -10 & -11 & -8 & -2 & 13 & -14 & 1 \\ 11 & -16 & 13 & 14 & -10 & -3 & 12 & 13 & -15 & 18 \\ -3 & -1 & -19 & 6 & 12 & 7 & 19 & -7 & 5 & -3 \\ -7 & -18 & -2 & 1 & -11 & 16 & -3 & 5 & -3 & 17 \end{pmatrix},$$

så är

$$\begin{aligned} \det(\mathbf{A} - \lambda \mathbf{I}) = & 15\,470\,169\,376\,525 - 306\,583\,517\,315\lambda - 26\,351\,278\,073\lambda^2 \\ & - 938\,744\,435\lambda^3 + 170\,758\,733\lambda^4 + 40\,214\,547\lambda^5 \\ & - 162\,975\lambda^6 - 24\,961\lambda^7 + 617\lambda^8 - 8\lambda^9 + \lambda^{10}. \end{aligned}$$

Egenvärdena till matrisen ges av

$$\begin{array}{ll} \lambda_1 = -24.2543 & \lambda_2 = -15.0567 \\ \lambda_3 = -9.09046 - 38.4524i & \lambda_4 = -9.09046 + 38.4524i \\ \lambda_5 = -4.9182 - 12.4888i & \lambda_6 = -4.9182 + 12.4888i \\ \lambda_7 = 10.2719 - 6.37766i & \lambda_8 = 10.2719 + 6.37766i \\ \lambda_9 = 27.3923 - 16.7311i & \lambda_{10} = 27.3923 + 16.7311i. \end{array}$$

Endast två av egenvärdena är reella. \diamond

Algoritm 5.4 (Potensmetoden med skalning). Låt \mathbf{A} vara en kvadratisk matris av ordning n som har n olika egenvärden. Följande metod finner en approximation av det dominerande egenparet $(\lambda_1, \mathbf{v}_1)$ till \mathbf{A} .

1. Låt $\mathbf{x}_0 \in \mathbb{R}^n$ sådan att $\|\mathbf{x}_0\|_2 = 1$.
2. För $k = 1, 2, 3, \dots$ sätt

$$\mathbf{y}_k = \mathbf{A}\mathbf{x}_{k-1} \quad \text{och} \quad \mathbf{x}_k = \frac{1}{c_k} \mathbf{y}_k,$$

där c_k är det till beloppet största elementet i \mathbf{y}_k , dvs om $\mathbf{y}_k = (y_1, y_2, \dots, y_n)$, bestäm heltalet i så att $|y_i| \geq |y_j|$ för $j = 1, 2, \dots, n$ och sätt sedan $c_i = y_i$.

3. Avbryt då $\|\mathbf{x}_{k-1} - \mathbf{x}_k\|_2 < \varepsilon$.
4. Returnera $(\lambda_1, \mathbf{v}_1) = (c_k, \mathbf{x}_k)$.

Låt \mathbf{A} vara en symmetrisk matris av ordning n . Enligt spektralsatsen existerar det därmed en ortonormal bas av egenvektorer $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ till \mathbf{A} och respektive egenvärde $\lambda_1, \lambda_2, \dots, \lambda_n$. Speciellt gäller att $\mathbf{v}_i^T \mathbf{v}_i = 1$ och $\mathbf{v}_i^T \mathbf{v}_j = 0$ då $i \neq j$. Sätt

$$\mathbf{A}_i = \mathbf{A} - \lambda_i \mathbf{v}_i \mathbf{v}_i^T.$$

Då är

$$\mathbf{A}_i \mathbf{v}_i = (\mathbf{A} - \lambda_i \mathbf{v}_i \mathbf{v}_i^T) \mathbf{v}_i = \mathbf{A} \mathbf{v}_i - \lambda_i \mathbf{v}_i \mathbf{v}_i^T \mathbf{v}_i = \lambda_i \mathbf{v}_i - \lambda_i \mathbf{v}_i = \mathbf{0} = 0 \mathbf{v}_i$$

och

$$\mathbf{A}_i \mathbf{v}_j = (\mathbf{A} - \lambda_i \mathbf{v}_i \mathbf{v}_i^T) \mathbf{v}_j = \mathbf{A} \mathbf{v}_j - \lambda_i \mathbf{v}_i \mathbf{v}_i^T \mathbf{v}_j = \lambda_j \mathbf{v}_j - \mathbf{0} = \lambda_j \mathbf{v}_j$$

då $i \neq j$. Alltså har \mathbf{A} och \mathbf{A}_i exakt samma egenvektorer, men egenvärdet λ_i till \mathbf{A} har ersatts med 0 för \mathbf{A}_i . Alltså är den näst dominerande egenparet $(\lambda_2, \mathbf{v}_2)$ till \mathbf{A} det dominerande egenparet till \mathbf{A}_1 . På liknande sätt kan vi komma fram till att det dominerande egenparet till

$$\mathbf{A} - \lambda_1 \mathbf{v}_1 \mathbf{v}_1^T - \cdots - \lambda_{k-1} \mathbf{v}_{k-1} \mathbf{v}_{k-1}^T \quad (5.9)$$

är $(\lambda_k, \mathbf{v}_k)$.

Algoritm 5.5 (Deflation). Låt \mathbf{A} vara en symmetrisk matris av ordning n . Algoritmen nedan bestämmer samtliga egenvärden och egenvektorer till \mathbf{A} .

1. Bestäm egenparet $(\lambda_1, \mathbf{v}_1)$ med potensmetoden.
2. För $k = 2, 3, \dots, n$ bestäm egenparet $(\lambda_k, \mathbf{v}_k)$ med hjälp av potensmetoden på matrisen (5.9).

Exempel 5.30. Låt

$$\mathbf{A} = \begin{pmatrix} -1 & -1 & 7 \\ -1 & 0 & -2 \\ 7 & -2 & 4 \end{pmatrix} \quad \text{och} \quad \mathbf{x}_0 = \frac{1}{\sqrt{3}}(1, 1, 1).$$

Med $\varepsilon = 10^{-12}$ ger potensmetoden efter 58 iterationer att

$$\lambda_1 = 9.45065 \quad \text{och} \quad \mathbf{v}_1 = (0.556821, -0.227953, 0.798741).$$

Det näst dominerande egenparet bestämmer vi genom att bilda matrisen

$$\begin{aligned} \mathbf{A} - \lambda_1 \mathbf{v}_1 \mathbf{v}_1^T &= \begin{pmatrix} -1 & -1 & 7 \\ -1 & 0 & -2 \\ 7 & -2 & 4 \end{pmatrix} \\ &\quad - 9.450065 \begin{pmatrix} 0.556821 \\ -0.227953 \\ 0.798741 \end{pmatrix} \begin{pmatrix} 0.556821 & -0.227953 & 0.798741 \end{pmatrix} \\ &= \begin{pmatrix} -3.93017 & 0.199562 & 2.79677 \\ 0.199562 & -0.49108 & -0.279269 \\ 2.79677 & -0.279269 & -2.02939 \end{pmatrix} \end{aligned}$$

och bestämma det dominerande egenparet till denna med potensmetoden. Det är dock möjligt att utnyttja att $\mathbf{v}_1^T \mathbf{x}_k$ är en skalär, dvs

$$\mathbf{y}_k = \mathbf{A} \mathbf{x}_k - \lambda_1 (\mathbf{v}_1^T \mathbf{x}_k) \mathbf{v}_1,$$

dvs en matrismultiplikation, en skalärprodukt, en skalärmultiplikation och slutligen en vektorsubtraktion. Efter 13 iterationer får vi att

$$\lambda_2 = -5.95295 \quad \text{och} \quad \mathbf{v}_2 = (-0.810859, 0.0593959, 0.58222).$$

Notera att $\mathbf{v}_1^T \mathbf{v}_2 \approx 0$, dvs de två funna egenvektorerna är i princip ortogonala. För att bestämma det trede och minst dominerande egenparet bildar vi matrisen

$$\mathbf{A} - \lambda_1 \mathbf{v}_1 \mathbf{v}_1^T - \lambda_2 \mathbf{v}_2 \mathbf{v}_2^T = \begin{pmatrix} -0.0161542 & -0.087142 & -0.013608 \\ -0.087142 & -0.470079 & -0.0734072 \\ -0.013608 & -0.0734072 & -0.0114632 \end{pmatrix}.$$

Redan efter två iterationer uppnår vi önskad precision och avbryter. Då är

$$\lambda_3 = -0.497696 \quad \text{och} \quad v_3 = (0.180161, 0.971859, 0.151765).$$

Vi har bestämt samtliga egenpar till A . \diamond

Med potensmetoden kan vi bara finna det dominerande egenparet till en matris. Antag att A är inverterbar och att (λ, v) är ett egenpar till A , där $\lambda \neq 0$. Då är

$$Av = \lambda v \Leftrightarrow A^{-1}Av = \lambda A^{-1}v \Leftrightarrow A^{-1}v = \frac{1}{\lambda}v.$$

Alltså är $1/\lambda$ ett egenvärde till A^{-1} med egenvektorn v . Det minst dominerande egenvärdet skilt från 0 till A är således det dominerande egenvärdet till A^{-1} . Denna observation kombinerad med potensmetoden ger oss ett sätt att bestämma (λ_n, v_n) . Metoden kallas *invers iteration*. Men bestämmer vi de övriga egenparen? Låt $\sigma \in \mathbb{R}$ och sätt $B = A - \sigma I$. Då är

$$Bv_k = (A - \sigma I)v_k = Av_k - \sigma v_k = \lambda_k v_k - \sigma v_k = (\lambda_k - \sigma)v_k,$$

dvs $\lambda_1 - \sigma, \lambda_2 - \sigma, \dots, \lambda_n - \sigma$ är egenvärden till $B = A - \sigma I$ med v_1, v_2, \dots, v_n som respektive egenvektorer. Genom att välja talet σ så att den är närmast egenvärdet λ_k fås att $1/(\lambda_k - \sigma)$ är den dominerande egenvärdet till $(A - \sigma I)^{-1}$.

Algoritm 5.6 (Invers iteration). Låt A vara en inverterbar matris. Följande metod bestämmer det minst dominerade egenparet (λ_n, v_n) till A .

1. LU-faktorisera A , dvs $PA = LU$.
2. Välj $x_0 \in \mathbb{R}^n$ så att $\|x_0\|_2 = 1$.
3. För $k = 1, 2, 3, \dots$ lös i tur och ordning ekvationssystemen

$$Ly_k = Px_{k-1} \quad \text{och} \quad Uz_k = y_k.$$

Sätt

$$x_k = \frac{1}{\|z_k\|_2} z_k \quad \text{och} \quad l_k = x_k^T A x_k.$$

4. Avbryt då $\|x_{k-1} - \text{sign}(l_k)x_k\|_2 < \varepsilon$.
5. Returnera $(\lambda_n, v_n) = (l_k, x_k)$.

Exempel 5.31. Låt

$$A = \begin{pmatrix} 0 & -1 & 7 \\ 4 & 0 & 8 \\ 3 & 5 & 1 \end{pmatrix}.$$

I exempel 5.28 fann vi det dominerande egenparet till A . LU-faktorisering av A ger att

$$P = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \quad L = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 4/3 & 20/3 & 1 \end{pmatrix} \quad \text{och} \quad U = \begin{pmatrix} 3 & 5 & 1 \\ 0 & -1 & 7 \\ 0 & 0 & -40 \end{pmatrix}.$$

Med tolerans en $\varepsilon = 10^{-12}$ krävs det 34 iterationer innan vi får att

$$\lambda_3 \approx l_{34} = -2.48129 \quad \text{och} \quad v_3 \approx x_{34} = (0.88964, -0.285792, -0.356179)$$

är det minst dominerande egenparet till A . \diamond

Algoritm 5.7 (Invers iteration med skift). Låt A vara en inverterbar matris och $\sigma \in \mathbb{R}$. Följande algoritm finner en approximation av det egenpar vars egenvärde är närmast σ .

1. LU-faktorisera $A - \sigma I$, dvs $P(A - \sigma I) = LU$.
2. Välj $x_0 \in \mathbb{R}^n$ så att $\|x_0\|_2 = 1$.
3. För $k = 1, 2, 3, \dots$ lös i tur och ordning ekvationssystemen

$$L y_k = P x_{k-1} \quad \text{och} \quad U z_k = y_k.$$

Sätt

$$x_k = \frac{1}{\|z_k\|_2} z_k \quad \text{och} \quad l_k = x_k^T A x_k.$$

4. Avbryt då $\|x_{k-1} - \text{sign}(l_k)x_k\|_2 < \varepsilon$.
5. Returnera $(\lambda_n, v_n) = (l_k, x_k)$.

Anmärkning. Notera att

$$l_k = x_k^T (A - \sigma I) x_k + \sigma = x_k^T A x_k - \sigma x_k^T x_k + \sigma = x_k^T A x_k$$

eftersom x_k är normerad, dvs $x_k^T x_k = 1$.

Exempel 5.32. Låt A vara samma matris som i exempel 5.28 och 5.31. Vi har så här långt funnit att

$$|\lambda_1| > |\lambda_2| \geq |\lambda_3| \Leftrightarrow 8.90945 > |\lambda_2| \geq 2.48129.$$

Låt oss anta att egenvärdet λ_2 är nära $\sigma = -5$. Välj samma startvektor x_0 och tolerans ε som tidigare. Efter 17 iteration med invers iteration med skift får vi att

$$\lambda_2 \approx l_{17} = -5.42816 \quad \text{och} \quad v_2 \approx x_{17} = (0.780445, 0.261725, -0.567808).$$

Vi har därmed bestämt samtliga egenpar till matrisen. \diamond

Algoritm 5.8 (Rayleighkvotiteration). Låt A vara en inverterbar matris och $\sigma_0 \in \mathbb{R}$. Följande algoritm finner en approximation av det egenpar vars egenvärde är närmast σ_0 .

1. Välj $x_0 \in \mathbb{R}^n$ så att $\|x_0\|_2 = 1$.
2. För $k = 1, 2, 3, \dots$ LU-faktorisera $A - \sigma_{k-1} I$ och lös i tur och ordning

$$L y_k = P x_{k-1} \quad \text{och} \quad U z_k = y_k.$$

Sätt

$$x_k = \frac{1}{\|z_k\|_2} z_k \quad \text{och} \quad \sigma_k = x_k^T A x_k.$$

3. Avbryt då $\|x_{k-1} - \text{sign}(\sigma_k)x_k\|_2 < \varepsilon$.
4. Returnera $(\lambda_n, v_n) = (\sigma_k, x_k)$.

För att välja lämpliga värden på σ och σ_0 i invers iteration med skift respektive Rayleighkvotiteration behöver vi kunna lokalisera egenvärdena. Låt $A = (a_{i,j})$ vara en kvadratisk matris av ordning n . Bilda mängderna

$$D_i = \{z \in \mathbb{C} : |z - a_{i,i}| \leq r_i\}$$

där

$$r_i = \sum_{\substack{j=1 \\ j \neq i}}^n |a_{i,j}|,$$

dvs summan av beloppet av respektive element på rad i i \mathbf{A} utom elementet i diagonalen. Mängden D_i kallas för den i :te Gerschgorinskivan till \mathbf{A} .

Sats 5.15 (Gerschgorins cirkelsats). *Alla egenvärden till en matris tillhör unionen av alla Gerschgorinskivor till matrisen.*

Bevis. Låt λ vara ett egenvärde till $\mathbf{A} = (a_{i,j})$ med $\mathbf{v} = (v_1, v_2, \dots, v_n)$ som en tillhörande egenvektor. Eftersom $\mathbf{v} \neq \mathbf{0}$ är minst en koordinat i \mathbf{v} skilt från 0. Tag i så att

$$|v_i| = \max_{1 \leq k \leq n} |v_k|.$$

Alltså är $v_i \neq 0$ och $|v_k/v_i| \leq 1$ för alla $k = 1, 2, \dots, n$. Vidare ges det i :te elementet i likheten $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$ av

$$a_{i,1}v_1 + a_{i,2}v_2 + \cdots + a_{i,n}v_n = \lambda v_i$$

eller ekvivalent

$$a_{i,1}v_1 + \cdots + a_{i,i-1}v_{i-1} + a_{i,i+1}v_{i+1} + \cdots + a_{i,n}v_n = \lambda v_i - a_{i,i}v_i.$$

Delar vi båda led med v_i får vi att

$$\begin{aligned} |\lambda - a_{i,i}| &= \left| a_{i,1} \frac{v_1}{v_i} + \cdots + a_{i,i-1} \frac{v_{i-1}}{v_i} + a_{i,i+1} \frac{v_{i+1}}{v_i} + \cdots + a_{i,n} \frac{v_n}{v_i} \right| \\ &\leq |a_{i,1}| \cdot \left| \frac{v_1}{v_i} \right| + \cdots + |a_{i,i-1}| \cdot \left| \frac{v_{i-1}}{v_i} \right| + |a_{i,i+1}| \cdot \left| \frac{v_{i+1}}{v_i} \right| + \cdots + |a_{i,n}| \cdot \left| \frac{v_n}{v_i} \right| \\ &\leq |a_{i,1}| + \cdots + |a_{i,i-1}| + |a_{i,i+1}| + \cdots + |a_{i,n}| \\ &= \sum_{\substack{j=1 \\ j \neq i}}^n |a_{i,j}| = r_i \end{aligned}$$

enligt triangelolikheten. Det visar att λ tillhör D_i och därmed unionen av samtliga Gerschgorinskivor. \square

5.9 Övningsuppgifter

1. Lös ekvationssystemet i exempel 5.2 utan pivotering. Använd fyra signifikanta siffror vid avrundning i varje operation.
2. Lös ekvationssystemet i exempel 5.2 med två siffrors avrundning samt
 - (a) med pivotering
 - (b) utan pivotering.

L 3. Lös det linjära ekvationssystemet

$$\begin{cases} 0.4x - 0.5y + 0.8z = 1 \\ -2.8x + 2.0y + 1.6z = 2 \\ 0.2x + 3.2y - 1.2z = 3 \end{cases}$$

med hjälp av Gausselimination och pivotering.

(20140110)

4. Låt $n = 5$. Bestäm $\mathbf{P}_{1,2}$ och $\mathbf{P}_{2,5}$.
5. Faktorisera permutationsmatrisen

$$\mathbf{P} = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

som en produkt av elementära permutationsmatriser på minst två olika sätt.

- L 6. Bestäm de reella talen a, b, c och d så att kurvan $y = a + bx + cx^2 + dx^3$ går genom punkterna $(0, 0), (1, 1), (2, 2)$ och $(3, 2)$.
- L 7. Lös ekvationssystemet

$$\begin{cases} 2x_1 - 3x_2 + 100x_3 = 1 \\ x_1 + 10x_2 - 0.001x_3 = 0 \\ 3x_1 - 100x_2 + 0.01x_3 = 0 \end{cases}$$

med Gausselimination och

- (a) partiell pivotering (b) skalad partiell pivotering.

Använd fyra siffrors avrundning i varje operation under beräknignarna, se exempel 1.16.

- L 8. Lös $\mathbf{Ly} = \mathbf{b}$ och $\mathbf{Ux} = \mathbf{y}$ samt verifiera att $\mathbf{Ax} = \mathbf{b}$ för
- (a) $\mathbf{b} = (-4, 10, 5)$ (b) $\mathbf{b} = (20, 49, 32)$
då $\mathbf{A} = \mathbf{LU}$ ges av

$$\begin{pmatrix} 2 & 4 & -6 \\ 1 & 5 & 3 \\ 1 & 3 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1/2 & 1 & 0 \\ 1/2 & 1/3 & 1 \end{pmatrix} \begin{pmatrix} 2 & 4 & -6 \\ 0 & 3 & 6 \\ 0 & 0 & 3 \end{pmatrix},$$

- L 9. LU-faktorisera följande matriser enligt $\mathbf{A} = \mathbf{LU}$.

$$(a) \mathbf{A} = \begin{pmatrix} 4 & 2 & 1 \\ 2 & 5 & -2 \\ 1 & -2 & 7 \end{pmatrix} \quad (b) \mathbf{A} = \begin{pmatrix} 1 & -2 & 7 \\ 4 & 2 & 1 \\ 2 & 5 & -2 \end{pmatrix}$$

- L 10. Låt

$$\mathbf{A} = \begin{pmatrix} 2 & 1 & -1 \\ 2 & 5 & 3 \\ 1 & -1 & 3 \end{pmatrix}.$$

LU-faktorisera \mathbf{A} enligt $\mathbf{A} = \mathbf{LU}$. (20120603)

11. LU-faktorisera matrisen

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 1 & 1 \\ 5 & 4 & 2 \end{pmatrix},$$

på formen $\mathbf{A} = \mathbf{LU}$. (20150108)

- L 12. LU-faktorisera matrisen

$$A = \begin{pmatrix} 2 & 1 & 0 & 0 & 0 \\ 1 & 2 & 1 & 0 & 0 \\ 0 & 1 & 2 & 1 & 0 \\ 0 & 0 & 1 & 2 & 1 \\ 0 & 0 & 0 & 1 & 2 \end{pmatrix}.$$

Ingen pivotering krävs. Tips: skriv inte talen på decimalform. (20130607)

- L 13. Låt

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

LU-faktorisera A med partiell pivotering. (20140603)

14. LU-faktorisera matrisen

$$A = \begin{pmatrix} -1 & 0 & 1 \\ 4 & -1 & 1 \\ 2 & -1 & 1 \end{pmatrix}$$

med pivotering, dvs bestäm permutationsmatris P , undertriangulär enhetsmatris L och övertriangulär matris U sådana att $PA = LU$. (20150822)

15. Låt $A \in \text{Mat}_{n,n}(\mathbb{R})$. Visa att $x \mapsto \|Ax\|$ är kontinuerlig.

16. Låt A vara en kvadratisk matris med komplexa matriselement. Då definieras det *hermiteska konjugatet* A^* av A som den matris som erhålls vid transponering av A samt komplexkonjugering av samtliga matriselement. Om $A^* = A$, så säges A vara *hermitesk*. Visa att

- (a) $(A^*)^* = A$
- (b) $(A^*)^{-1} = (A^{-1})^*$, då A är inverterbar
- (c) $(A + B)^* = A^* + B^*$
- (d) $(aA)^* = \bar{a}A^*$, där $a \in \mathbb{C}$
- (e) $(AB)^* = B^*A^*$.

- L 17. Låt A vara en kvadratisk matris samt låt u och v vara kolonnmatrider, samtliga med komplexa element. Visa att

- (a) den Euklidiska inre produkten på \mathbb{C}^n ges av $(u, v) = v^*u$.
- (b) $(Au, v) = (u, A^*v)$.
- (c) om A är hermitesk så är $(Au, v) = (u, Av)$.
- (d) egenvärdena till en hermitesk matris är reella.
- (e) A^*A är hermitesk.
- (f) egenvärdena till A^*A är icke-negativa.

- L 18. Låt A vara en $n \times m$ -matris och bilda mängden $V = \{Ax : x \in \mathbb{R}^m\}$. Visa att V är ett underrum till \mathbb{R}^n .

- L 19. Låt u och v vara ortogonala vektorer i \mathbb{R}^n . Visa att

$$\|u + v\|_2^2 = \|u\|_2^2 + \|v\|_2^2,$$

dvs Pythagoras sats.

- L 20. Låt A vara en matris. Visa att $A^T A$ är symmetrisk.
- L 21. Anpassa linjen $y = f(x) = ax + b$ i minsta kvadratmetodens mening till följande data. Beräkna också $E_2(f)$.

	x_k	y_k		x_k	y_k		x_k	y_k
(a)	-2	1		-6	7		-4	-3
	-1	2	(b)	-2	5		-1	-1
	0	3		0	3		0	0
	1	3		2	2		2	1
	2	4		6	0		3	2

- L 22. Antag att $f(x) = a + bx$ bäst approximerar, i minsta kvadratmetodens mening, data på formen (x_i, y_i) , där $i = 1, 2, \dots, n$. Bestäm uttryck för a och b med avseende på x_i och y_i .
- L 23. Anpassa potensfunktionen $f(x) = ax$ till följande data i enlighet med minsta kvadratmetoden samt beräkna den kvadratiska medelvärdesfelet $E_2(f)$.

	x_k	y_k		x_k	y_k		x_k	y_k
(a)	-4	-3		3	1.6		1	1.6
	-1	-1	(b)	4	2.4		2	2.8
	0	0		5	2.9		3	4.7
	2	1		6	3.4		4	6.4
	3	2		8	4.6		5	8.0

24. Anpassa $f(x) = ax + b$ enligt minsta kvadratmetoden till punkterna

$$(1.1, 1.5), (2.4, 3.2), (3.3, 5.3), (3.5, 6.1), (1.7, 2.9), \\ (2.8, 4.6), (-0.5, -0.6), (0.4, 0.3) \text{ och } (0.4, 1.2).$$

Beräkna även det kvadratiska medelvärdesfelet $E_2(f)$.

25. Anpassa $f(x) = ax + b$ enligt minsta kvadratmetoden till punkterna

$$(1.1, 0.5), (2.4, 4.2), (3.3, 1.3), (3.5, 2.1), (1.7, 4.9), (2.8, 4.6), (-0.5, 2.6), \\ (0.4, 2.3), (0.4, -0.2), (2, 1.1), (1.7, 2.3) \text{ och } (1.5, 3).$$

Beräkna även det kvadratiska medelvärdesfelet $E_2(f)$.

26. Bestäm a , b och c så att $f(x) = a + bx$ enligt minsta kvadratmetoden anpassas till punkterna

$$(1, 2), (2, 1), (3, 3) \text{ och } (4, c)$$

samt att linjen $y = f(x)$ är horisontell. (20140822)

- L 27. Bestäm det andragradspolynomet $p(x) = ax^2 + bx + c$ som bäst anpassar till punkterna

$$(-2, 1), (-1, 1), (0, 0), (1, 2) \text{ och } (2, 1)$$

i minsta kvadratmetodens mening. (20130607)

- L 28. Finn det andragradspolynomet $f(x) = ax^2 + bx + c$ som bäst anpassas till punkterna

$$(x_1, y_1) = (-1, 1), (x_2, y_2) = (0, -1), (x_3, y_3) = (1, 0), \\ (x_4, y_4) = (2, 0) \text{ och } (x_5, y_5) = (3, 2).$$

i minsta kvadratmetodens mening. (20140603)

L 29. Låt

$$D = n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2$$

Visa att

$$a = \frac{1}{D} \left(n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i \right)$$

och

$$b = \frac{1}{D} \left(\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i - \sum_{i=1}^n x_i \sum_{i=1}^n x_i y_i \right)$$

är lösningen till ekvationssystemet för a och b i den linje $y = ax + b$ som bäst anpassar $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ i minsta kvadratmetodens mening.

L 30. Anpassa $f(x) = ax^2 + bx + c$ till följande data i enlighet med minsta kvadratmetoden.

	x_k	y_k		x_k	y_k
	-3	15		-3	-1
(a)	-1	5	(b)	-1	25
	1	1		1	25
	3	5		3	1

31. Anpassa $f(x) = ax^2 + bx + c$ enligt minsta kvadratmetoden till

$$(-2.1, 3.1), (-1.4, 1.2), (-0.7, 0.2), (0.2, -0.2), (1.3, 0.6) \text{ och } (2.1, 1.9).$$

Beräkna även det kvadratiska medelvärdesfelet $E_2(f)$.

32. Låt $(x_k y_k)$, där $k = 1, 2, 3, 4, 5, 6$ vara punkterna

$$(-2.1, 3.1), (-1.4, 1.2), (-0.7, 0.2), (0.2, 0.4), (1.3, 0.5) \text{ respektive } (2.1, 0.5).$$

Anpassa i enlighet med minsta kvadratmetoden ett polynom $f_n(x)$ av grad n till dessa punkter och beräkna det kvadratiska-medelvärdesfelet $E_3(f_n)$.

$$(a) \quad n = 1 \quad (b) \quad n = 2 \quad (c) \quad n = 3 \quad (d) \quad n = 4 \quad (e) \quad n = 5$$

33. Givet följande data.

$$(0.1, 0.3), (0.3, 1.2), (0.5, 1.8), (0.8, 2.2), (1.2, 2.3), (1.6, 2.1), \\ (2.1, 2.1), (3.5, 1.1), (4.0, 0.7) \text{ och } (5.2, 0.4).$$

Approximera denna data enligt minsta kvadratmetoden med $f(x) = bxe^{-ax}$, där du linjäriserar data.

L 34. Finn det variabelbyte som linjäriserar sambandet

$$y = \frac{x}{a + bx'}$$

dvs skriv om det på formen $Y = AX + B$. Beskriv också hur konstanterna a och b beror på A och B .

Kapitel 6

Ordinära differentialekvationer

6.1 Numerisk derivering

Antag att $f: \mathbb{R} \rightarrow \mathbb{R}$ är deriverbar i $x \in \mathbb{R}$. Att bestämma $f'(x)$ kan vara svårt av olika skäl, tex kanske uttrycket för $f(x)$ är komplicerat eller rent av okänt. Vi ska här studera metoder att beräkna $f'(x)$, dvs hur man finner en numerisk approximation av $f'(a)$. Algoritmer för att symboliskt derivera en funktion ingår inte i kursen.

20160511

6.1.1 Central differensapproximation

Låt $f \in C^3[a, b]$ och $x \in \mathbb{R}$. Antag att vi vill beräkna derivatan av f i x , dvs $f'(x)$. Taylorutvecklingen av f kring $t = x$ ges av

$$f(t) = f(x) + f'(x)(t - x) + \frac{f''(x)}{2}(t - x)^2 + \frac{f^{(3)}(\xi)}{6}(t - x)^3,$$

för något ξ mellan t och x . Låt $h > 0$. Då är

$$f(x - h) = f(x) - f'(x)h + \frac{f''(x)}{2}h^2 - \frac{f^{(3)}(\xi_1)}{6}h^3$$

och

$$f(x + h) = f(x) + f'(x)h + \frac{f''(x)}{2}h^2 + \frac{f^{(3)}(\xi_2)}{6}h^3,$$

där $x - h \leq \xi_1 \leq x \leq \xi_2 \leq x + h$. Det följer då att

$$\begin{aligned} f(x + h) - f(x - h) &= 2f'(x)h + \frac{f^{(3)}(\xi_1) + f^{(3)}(\xi_2)}{6}h^3 \\ &\Leftrightarrow \\ f'(x) &= \frac{f(x + h) - f(x - h)}{2h} - \frac{f^{(3)}(\xi)}{6}h^2 \end{aligned}$$

för något $\xi \in [\xi_1, \xi_2]$ enligt satsen om mellanliggande värden eftersom $f^{(3)}$ är kontinuerlig. En approximation av $f'(x)$ ges av

$$D_0(h) = \frac{f(x + h) - f(x - h)}{2h}$$

h	$D_0(h)$	h	$D_0(h)$
0.30	-1.60407	0.30	12.0108
0.25	-1.66778	0.25	17.0536
0.20	-1.71923	0.20	26.781
0.15	-1.75878	0.15	49.9978
0.10	-1.78677	0.10	139.333
0.05	-1.80344	0.05	-1428.06
0.03	-1.80699	0.03	-416.849
0.02	-1.80810	0.02	-341.166
0.01	-1.80876	0.01	-307.623
10^{-10}	-1.80898	10^{-10}	-297.857
10^{-15}	-1.72065	10^{-15}	-288.428
10^{-16}	-1.66533	10^{-16}	-337.508
10^{-17}	0	10^{-17}	0

Tabell 6.1

Figur 6.1

och felet är $|E_h| \leq Ch^2$, för något positivt reellt tal C . Vi säger att approximationen är av ordning h^2 , vilket vi betecknar $O(h^2)$. En förbättring av approximationen av $f'(x)$ ges av

$$D_1(h) = \frac{-f(x+2h) + 8f(x+h) - 8f(x-h) + f(x-2h)}{12h},$$

som är av ordning $O(h^4)$.

Exempel 6.1. Låt $f(t) = \cos e^t$. Då är $f'(0.75) \approx -1.80898$. Om $h = 0.3$, så är

$$D_0(0.3) = \frac{f(0.75 + 0.3) - f(0.75 - 0.3)}{2 \cdot 0.3} \approx -1.60407.$$

I vänstra tabellen i tabell 6.1 ser vi motsvarande $D_0(h)$ för olika steg h . Notera att $D_0(h)$ är lutningen på den linje som går genom punkterna $(x-h, f(x-h))$ och $(x+h, f(x+h))$, se figur 6.1. \diamond

Exempel 6.2. Låt

$$f(t) = \frac{t^3 + 1}{t^2 - 0.65}.$$

Då är $f'(0.75) \approx -297.86$. Notera att denna funktion inte är kontinuerlig. I högra tabellen i tabell 6.1 ser vi $D_0(h)$ för olika steg h . \diamond

6.1.2 Richardsons extrapolation

Låt k vara ett heltal och sätt $f_k = f(x + kh)$. Då är

$$D_0(h) = \frac{f_1 - f_{-1}}{2h} \quad \text{och} \quad D_0(2h) = \frac{f_2 - f_{-2}}{4h}.$$

För någon konstant C är

$$f'(x) \approx D_0(h) + Ch^2 \quad \text{och} \quad f'(x) \approx D_0(2h) + 4Ch^2.$$

Det ger att

$$\begin{aligned} 3f'(x) &\approx 4D_0(h) - D_0(2h) = 2\frac{f_1 - f_{-1}}{h} - \frac{f_2 - f_{-2}}{4h} \\ &= \frac{-f_2 + 8f_1 - 8f_{-1} - f_{-2}}{4h}. \end{aligned}$$

eller ekvivalent

$$f'(x) \approx \frac{4D_0(h) - D_0(2h)}{3} = D_1(h).$$

Det ger oss en approximation av $f'(x)$ vilken är proportionell mot h^4 , istället för h^2 , dvs vi har gjort en förbättring av $f'(x)$ där vi interpolerar med de redan kända värdena $D_0(h)$ och $D_0(2h)$. Antag att $D_{k-1}(h)$ och $D_{k-1}(2h)$ är approximationer av $f'(x)$ av ordning $O(h^{2k})$, sådana att

$$f'(x) = D_{k-1}(h) + c_1 h^{2k} + c_2 h^{2k+2} + \dots$$

och

$$f'(x) = D_{k-1}(2h) + 4^k c_1 h^{2k} + 4^{k+1} c_2 h^{2k+2} + \dots$$

Då är

$$D_k(h) = \frac{4^k D_{k-1}(h) - D_{k-1}(2h)}{4^k - 1}$$

en approximation av $f'(x)$ av ordning $O(h^{2k+1})$. Den rekursiva formeln kan skrivas

$$D_k(h) = D_{k-1}(h) + \frac{D_{k-1}(h) - D_{k-1}(2h)}{4^k - 1}, \quad k \geq 1.$$



Bevis

Vi erhåller en triangulär tabell:

$$\begin{array}{cccc} D_0(2^i h) \\ D_0(2^{i-1} h) & D_1(2^{i-1} h) \\ D_0(2^{i-2} h) & D_1(2^{i-2} h) & D_2(2^{i-2} h) \\ D_0(2^{i-3} h) & D_1(2^{i-3} h) & D_2(2^{i-3} h) & D_3(2^{i-3} h) \\ \dots & & & \end{array}$$

Dessa värden beräknas tills

$$|D_k(2^i h) - D_k(2^{i+1} h)| < \varepsilon,$$

och då används $D_k(2^i h)$ som en approximation av $f'(x)$.

Exempel 6.3. Låt $f(t) = \cos e^t$. Då är

$$D_0(h) = \frac{f_1 - f_{-1}}{2h} = f'(x) + a_1 h^2 + a_2 h^4 + \dots,$$

där tex

$$a_1 = \frac{e^x}{6} ((e^{2x} - 1) \sin e^x - 3e^x \cos e^x).$$

Låt $h = 0.01$. Då är

$$2^3 h = 0.08, \quad 2^2 h = 0.04 \quad \text{och} \quad 2h = 0.02$$

och

$$4^3 - 1 = 63, \quad 4^2 - 1 = 15 \quad \text{och} \quad 4^1 - 1 = 3.$$

Antag att vi vill beräkna $f'(0.75)$. Vi får följande tabell.

	$D_0 + \Delta_1/3$	$D_1 + \Delta_2/15$	$D_2 + \Delta_3/63$
$D_0(2^3 h)$			
$D_0(2^2 h)$	$D_1(2^2 h)$		
$D_0(2h)$	$D_1(2h)$	$D_2(2h)$	
$D_0(h)$	$D_1(h)$	$D_2(h)$	$D_3(h)$

Beteckningen Δ_k motsvarar differensen $D_{k-1}(2^i h) - D_{k-1}(2^{i+1} h)$. Med värden ges tabellen av följande.

	$D_0 + \Delta_1/3$	$D_1 + \Delta_2/15$	$D_2 + \Delta_3/63$
-1.794782129			
-1.805438163	-1.808990175		
-1.808096524	-1.808982645	-1.808982143	
-1.808760759	-1.80898217	-1.808982139	-1.808982139

När vi avbryter beräkningarna beror det på vald precision ε . Om tex $\varepsilon = 0.5 \cdot 10^{-8}$, så avbryter då vi når $D_2(h)$, eftersom

$$|D_2(h) - D_2(2h)| \approx 0.4018 \cdot 10^{-8} < \varepsilon.$$

I så fall kan vi konstatera att $f'(0.75) \approx -1.80898214$. Jämför med exempel 6.1 där det krävdes steget $h = 10^{-10}$ för att få samma approximation av $f'(0.75)$. \diamond

6.2 Begynnelsevärdesproblem

Låt $S = \{(x, y) : a \leq x \leq b \text{ och } -\infty < y < \infty\}$ och antag att funktionen $f: S \rightarrow \mathbb{R}$ är kontinuerlig, där a och b är ändliga. Om det existerar en konstant L sådan att

$$|f(x, y) - f(x, z)| \leq L|y - z|$$

för alla $(x, y) \in S$ och $(x, z) \in S$, så säges f uppfylla ett *Lipschitzvillkor* och L kallas för en *Lipschitzkonstant*.

Exempel 6.4. Låt $f(x, y) = (\sin x - x) \cdot y$, $a = 0$ och $b = 4$. Då gäller att

$$\begin{aligned} |f(x, y) - f(x, z)| &= |(\sin x - x) \cdot y - (\sin x - x) \cdot z| \\ &= |\sin x - x| \cdot |y - z| \leq |\sin 4 - 4| \cdot |y - z|. \end{aligned}$$

Alltså uppfyller f ett Lipschitzvillkor för alla x sådana att $0 \leq x \leq 4$ och med Lipschitzkonstanten $L = |\sin 4 - 4| \leq 5$. \diamond

Anmärkning. För att visa att uttrycket $|\sin x - x|$ är som stört då $x = 4$. Sätter vi först derivatan av $\sin x - x$ lika med noll och löser denna ekvation. Det ger att lokala extrempunkter hittar vi i $x = 2k\pi$, där k är ett heltal. Endast $k = 0$ ger ett x som ligger i intervallet $[0, 4]$. Vidare måste vi studera $\sin x - x$ i ändpunkterna, varav den vänstra är i detta fall också en extrempunkt. Vi finner att $|\sin x - x|$ är som störst då $x = 4$ och som minst då $x = 0$, dvs $0 \leq |\sin x - x| \leq |\sin 4 - 4| \approx 4.7568$.

Låt $\alpha \in \mathbb{R}$. Ett *begynnelsevärdesproblem* har formen

$$y' = f(x, y) \quad \text{och} \quad y(a) = \alpha, \quad (a \leq x \leq b) \quad (6.1)$$

där y betraktas som en funktion med avseende på variabeln x . Notera att $y' = f(x, y)$ också kan skrivas på formen $y'(x) = f(x, y(x))$.

Sats 6.1. Om den reellvärda funktionen f är kontinuerlig på remsan S och uppfyller ett Lipschitzvillkor på S , så har begynnelsevärdesproblemet (6.1) en entydig lösning.

Exempel 6.5. Från satsen och tidigare exempel följer att begynnelsevärdesproblemet

$$y' = (\sin x - x) \cdot y \quad \text{och} \quad y(0) = 2 \quad (0 \leq x \leq 4)$$

har en entydig lösning, nämligen

$$y(x) = 2 \exp\left(1 - \frac{x^2}{2} - \cos x\right).$$

Det lämnas som övning att kontrollera att det stämmer. \diamond

Låt $y(x)$ vara lösningen till $y' = f(x, y)$ och antag att $y(c) = d$, där $a \leq c \leq b$. Då är lutning på tangenten till kurvan $y = y(x)$ i punkten (c, d) lika med $y'(c)$, dvs $f(c, d)$.

Exempel 6.6. För $f(x, y) = (\sin x - x) \cdot y$ har vi *riktningsfältet* i figur 6.2. \diamond

Låt N vara ett positivt heltal och sätt $h = (b - a)/N$. Då delar punkterna

$$x_n = a + nh, \quad n = 0, 1, \dots, N$$

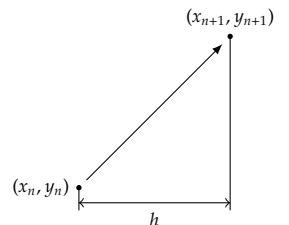
upp intervallet $[a, b]$ i N delintervall av lika längd h . Vi vill approximera lösningen y till begynnelsevärdesproblemet (6.1) i punkterna x_n . Det gör vi iterativt genom att bestämma punkterna (x_n, y_n) sådana att $y_n \approx y(x_n)$. Antag attlösningen $y = y(x)$ passera genom (x_n, y_n) och att vektorn är parallell med tangenten i (x_n, y_n) . Då ges lutningen av

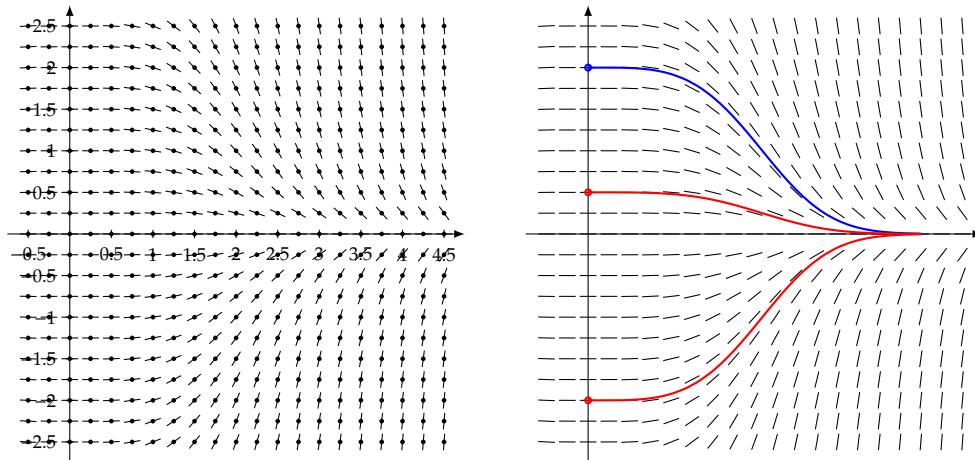
$$y'(x_n) = f(x_n, y_n) = \frac{y_{n+1} - y_n}{x_{n+1} - x_n} = \frac{y_{n+1} - y_n}{h},$$

se figur i marginalen. Vi har härlett *Eulers metod*:

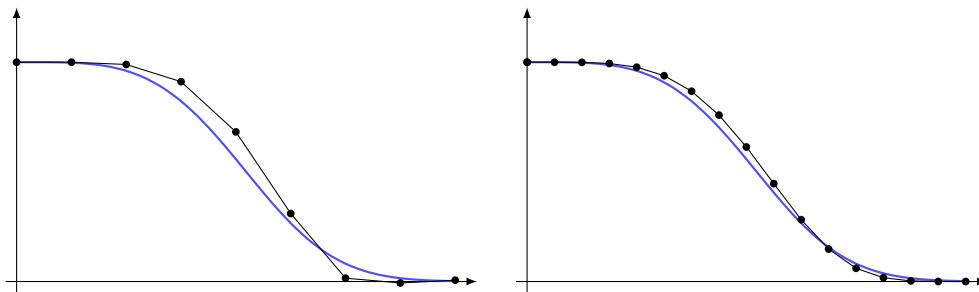
$$(x_0, y_0) = (a, \alpha), \quad x_{n+1} = x_n + h, \quad \text{och} \quad y_{n+1} = y_n + h f(x_n, y_n),$$

för alla $n = 0, 1, \dots, N - 1$. Vi kan interpolera punkterna (x_n, y_n) med tex en splinefunktion för att representera lösningen y med en kontinuerlig funktion.





Figur 6.2

Figur 6.3. Eulers metod då $N = 8$ respektive $N = 15$. Blå kurva motsvarar den exakta lösningen, se exempel 6.5.

n	x_n	y_n
0	0.0	2.0000
1	0.5	2.0000
2	1.0	1.9794
3	1.5	1.8225
4	2.0	1.3646
5	2.5	0.6204
6	3.0	0.0305
7	3.5	-0.0131
8	4.0	0.0121

Exempel 6.7. Studera begynnelsevärdesproblemet

$$y' = f(x, y) = (\sin x - x) \cdot y \quad \text{och} \quad y(0) = 2,$$

i intervallet $[0, 4]$. Då är $a = 0$, $b = 4$ och $\alpha = 2$. Om $N = 8$, så är $h = (b - a)/N = 0.5$. Vidare är $(x_0, y_0) = (0, 2)$ och $y_{n+1} = y_n + h(\sin x_n - x_n) \cdot y_n$, se tabell i marginalen. Vi får en bättre approximation om vi väljer ettn kortare steglängd, se figur 6.3. Det globala trunkeringsfelet: $y(x_n) - y_n = O(h)$. \diamond

6.3 Runge-Kuttas metod

Heuns metod För att bestämma y_{n+1} , då $n = 0, 1, \dots, N - 1$, beräknar vi först

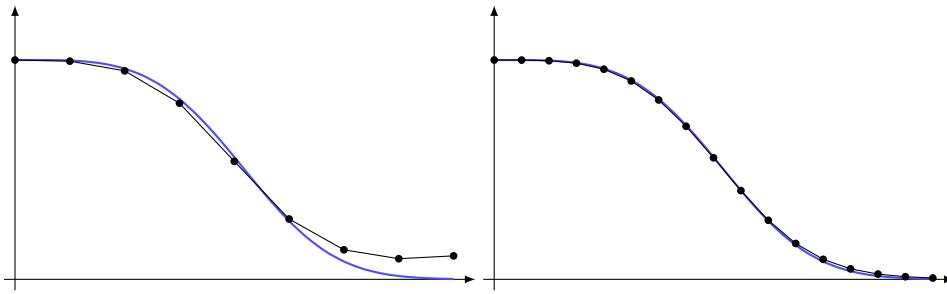
$$k_1 = f(x_n, y_n) \quad \text{och} \quad k_2 = f(x_n + h, y_n + hk_1).$$

Då är

$$y_{n+1} = y_n + \frac{h}{2}(k_1 + k_2).$$

För att "bevisa" detta konstaterar vi först att för $x_n \leq x \leq x_{n+1}$ är

$$f(x, y) = y' \Leftrightarrow \int_{x_n}^x f(t, y(t)) dt = \int_{x_n}^x y'(t) dt = y(x) - y(x_n).$$



Figur 6.4. Heuns metod då $N = 8$ respektive $N = 15$.

Applicerar vi trapetsmetoden på vänsterledet får vi att

$$\begin{aligned} \int_{x_n}^{x_{n+1}} f(t, y(t)) dt &= \frac{h}{2} \{f(x_n, y(x_n)) + f(x_{n+1}, y(x_{n+1}))\} \\ &= \frac{h}{2} \{f(x_n, y_n) + f(x_n + h, y_{n+1})\} \end{aligned}$$

Om vi ersätter y_{n+1} med formeln från Eulers metod, så får vi att

$$\begin{aligned} \int_{x_n}^{x_{n+1}} f(t, y(t)) dt &= \frac{h}{2} \{f(x_n, y_n) + f(x_n + h, y_n + f(x_n, y_n))\} \\ &= \frac{h}{2}(k_1 + k_2). \end{aligned}$$

Alltså är

$$\int_{x_n}^{x_{n+1}} f(t, y(t)) dt = y(x_{n+1}) - y(x_n)$$

ekvivalent med

$$\frac{h}{2}(k_1 + k_2) = y_{n+1} - y_n,$$

vilket "visar" Heuns metod. Det globala trunkeringsfelet: $y(x_n) - y_n = O(h^2)$.

Exempel 6.8. Löser vi samma begynnelsevärdesproblem som i exempel 6.7 med Heuns metod får resultatet i figur 6.4. ◇

Runge-Kuttas metod av ordning $O(h^4)$ För att bestämma y_{n+1} beräknar man i tur och ordning

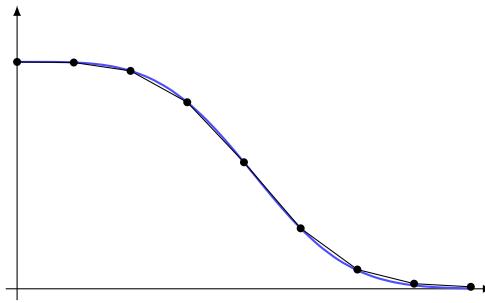
$$\begin{aligned} k_1 &= f(x_n, y_n) \\ k_2 &= f\left(x_n + \frac{h}{2}, y_n + \frac{h}{2}k_1\right) \\ k_3 &= f\left(x_n + \frac{h}{2}, y_n + \frac{h}{2}k_2\right) \\ k_4 &= f(x_n + h, y_n + hk_3). \end{aligned}$$

Då är

$$y_{n+1} = y_n + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4).$$

Som tidigare är $x_n = a + hn$, för $n = 0, 1, \dots, N$.

Exempel 6.9. Resultatet av Runge-Kuttas metod för samma begynnelsevärdesproblem som i exempel 6.7 ser vi i figur 6.5. ◇

Figur 6.5. Runge-Kuttas metod då $N = 8$.

6.4 System av differentialekvationer

En differentialekvation av högre ordning kan skrivas om till ett system av ett system av differentialekvationer av första ordningen.

Exempel 6.10. Studera differentialekvationen

$$y^{(3)} = xyy' + y'y'' - \sin(yy')$$

Låt $y_1 = y$, $y_2 = y'$ och $y_3 = y''$. Då kan differentialekvationen skrivas som systemet

$$\begin{cases} y'_1 = y_2 \\ y'_2 = y_3 \\ y'_3 = xy_1y_2 + y_2y_3 - \sin(y_1y_2), \end{cases}$$

eller som vektorer på formen

$$\mathbf{y}' = f(x, \mathbf{y}),$$

där $\mathbf{y} = (y_1, y_2, y_3)$. Notera att $\mathbf{y}(x) = (y_1(x), y_2(x), y_3(x))$. Om vi också har begynnelsevillkoren

$$y(a) = \alpha_1, \quad y'(a) = \alpha_2, \quad \text{och} \quad y''(a) = \alpha_3,$$

så är $\mathbf{y}(a) = (y_1(a), y_2(a), y_3(a)) = (y(a), y'(a), y''(a)) = (\alpha_1, \alpha_2, \alpha_3)$. \diamond

Exempel 6.11. Studera följande differentialekvation av andra ordningen,

$$y'' = f(x, y, y') = y' \sin x - xy.$$

Sätt $\mathbf{y} = (y_1, y_2)$, där $y_1 = y$ och $y_2 = y'$. Då är

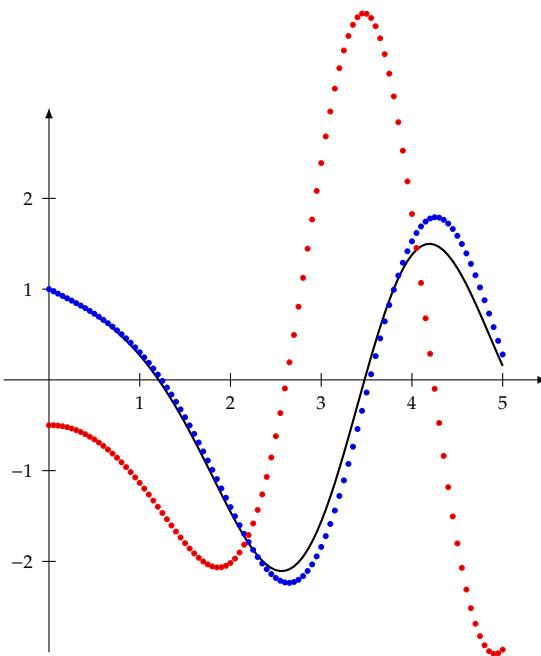
$$\begin{cases} y'_1 = y_2 \\ y'_2 = y_2 \sin x - xy_1 \end{cases} \Leftrightarrow \mathbf{y}' = f(x, \mathbf{y}),$$

där $f: \mathbb{R} \times \mathbb{R}^2 \rightarrow \mathbb{R}^2$ ges av $f(x, \mathbf{y}) = (y_2, y_2 \sin x - xy_1)$. Eulers metod för ett system av första ordningens differentialekvationer ges av

$$\mathbf{y}_{n+1} = \mathbf{y}_n + hf(x_n, \mathbf{y}_n).$$

Om vi har begynnelsevillkoren $\mathbf{y}(0) = \mathbf{y}_0$ och $\mathbf{y}'(0) = \mathbf{y}'_0$, så är $a = 0$ och

$$\mathbf{y}_0 = (y_{1,0}, y_{2,0}) = (y_1(0), y_2(0)) = (1, -0.5).$$



Figur 6.6. Svart kurva representerar den exakta lösningen. Blå punkter motsvarar $y = y_{1,n} \approx y(x_n)$ och röda punkter motsvarar $y = y_{2,n} \approx y'(x_n)$.

Vidare låt $b = 5$ och $N = 100$. Då är $h = 0.05$. Vektorn $\mathbf{y}_1 = (y_{1,1}, y_{2,1})$ beräknas enligt

$$\begin{aligned}\mathbf{y}_1 &= \mathbf{y}_0 + h f(x_1, \mathbf{y}_0) = (1, -0.5) + 0.05 f(0, (1, -0.5)) \\ &= (1, -0.5) + 0.05(-0.5, -0.5 \sin 0 - 0 \cdot 1) = (0.975, -0.5).\end{aligned}$$

Vektorerna $\mathbf{y}_2, \mathbf{y}_3, \dots, \mathbf{y}_{100}$ beräknas på liknande sätt. Första koordinaten $y_{1,n}$ i \mathbf{y}_n är approximationen av den sökta lösningen y i x_n . Den andra koordinaten $y_{2,n}$ är en approximation av y' i x_n , se figur 6.6. ◇

Exempel 6.12 (Lotka-Volterras rov- och bytesdjursmodell). Vi ska studera en enkel version av en populationsmodell som beskriver hur två olika djurarter, en rovdjur och en bytesdjur, interagerar. Låt $x(t)$ och $y(t)$ beteckna antalet bytesdjur respektive antalet rovdjur vid tiden t . Om tillgången till föda och utrymme för bytesdjuren vore obegränsad och om vi bortser från rovdjurens, så skulle x' vara proportionellt mot x , dvs $x' = \alpha x$, där α är en positiv konstant. Denna formel kallas *Malthus lag*. Vi får då en ohämmad tillväxt av x . Men på grund av rovdjurens fås en tendens åt andra hållet. Vi antar att antalet möten mellan rov- och bytesdjur är proportionellt mot både x och y , dvs mot xy . Alltså har vi att $x' = \alpha x - \beta xy$, där β är en positiv konstant som bla beskriver hur ofta ett mötet slutar med att rovdjuret fångar bytesdjuret.

Om inga bytesdjur fanns, så skulle beständet av rovdjur minska enligt $y' = -\gamma y$, där γ är en positiv konstant. Med liknande resonemang som ovan får vi att förändringen av rovdjur kan beskrivas av $y' = -\gamma y + \delta xy$, där δ är en positiv konstant. Man kan tycka att β och δ borde vara lika, men vi vet inte vilka parametrar som styr deras värden och därför bör vi använda två olika konstanter. När man studerar ett specifikt

fall, så kanske det visar sig att $\beta = \delta$. Vi har härlett differentialekvationssystemet

$$\begin{cases} x' = \alpha x - \beta xy \\ y' = -\gamma y + \delta xy. \end{cases} \quad (6.2)$$

Detta är ett icke-linjärt system för vilket det tyvärr inte går att finna en exakt analytisk lösning, se anmärkning efter exemplet för mer information.

Låt $v = (x, y)$. Eftersom x och y beror på t , är v en vektorvärd funktion med avseende på t , dvs $v: \mathbb{R} \rightarrow \mathbb{R}^2$, där $v(t) = (x(t), y(t))$. Vidare, låt $f: \mathbb{R} \times \mathbb{R}^2 \rightarrow \mathbb{R}^2$ enligt

$$f(t, v) = f(t, (x, y)) = (\alpha x - \beta xy, -\gamma y + \delta xy).$$

Notera att f beror indirekt på t . Ekvationsystemet (6.2) kan nu skrivas

$$v' = f(t, v).$$

I frotsättningen låter vi $\alpha = \beta = \gamma = \delta = 1$, dvs $f(t, v) = (x - xy, -y + xy)$. Studera begynnelsevärdesproblemets

$$v' = f(t, v) \quad \text{och} \quad v(0) = v_0, \quad t \geq 0,$$

där $v_0 = (x_0, y_0)$ samt $x_0 = x(0)$ och $y_0 = y(0)$. Eulers metod ges av

$$\begin{aligned} v_{n+1} &= v_n + hf(t_n, v_n) = (x_n, y_n) + h(x_n - x_n y_n, -y_n + x_n y_n) \\ &= (x_n + hx_n - hx_n y_n, y_n - hy_n + hx_n y_n), \end{aligned}$$

där $t_{n+1} = t_n + h$. Rekursivt bestämmer vi punkterna $v_n = (x_n, y_n)$ vars koordinater är approximationer av lösningarna $x(t)$ och $y(t)$ till (6.2), dvs $x_n \approx x(t_n)$ och $y_n \approx y(t_n)$. Antag att vi vill numeriskt bestämma en approximation av lösningen till

$$v' = f(t, v) \quad \text{och} \quad v(0) = v_0 = (0.5, 0.5), \quad 0 \leq t \leq 12.$$

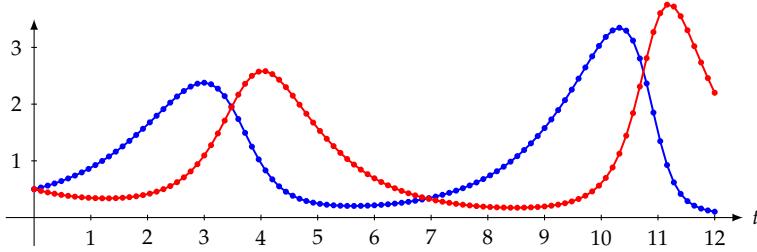
Med $N = 100$, så är $h = (12 - 0)/100 = 0.12$. Då är bla

$$\begin{aligned} t_1 &= t_0 + h = 0 + 0.12 = 0.12 \\ v_1 &= (x_0 + hx_0 - hx_0 y_0, y_0 - hy_0 + x_0 y_0) \\ &= (0.5 + 0.12 \cdot 0.5 - 0.12 \cdot 0.5 \cdot 0.5, 0.5 - 0.12 \cdot 0.5 + 0.12 \cdot 0.5 \cdot 0.5) \\ &= (0.53, 0.47), \end{aligned}$$

dvs $x(0.12) \approx 0.53$ och $y(0.12) \approx 0.47$. Om vi tolkar x och y som densiteten, dvs antal individer per areaenhet, så ser vi att bytesdjuren ökar något medan rovdjuren minskar efter 0.12 tidsenheter. I tur och ordning bestämmer man v_2, v_3, \dots, v_{100} , där $x(12) \approx x_{100}$ och $y(12) \approx y_{100}$, se figur 6.7. Tillsammans bildar lösningarna $x(t)$ och $y(t)$ en kurva $v(t) = (x(t), y(t))$ i xy -planet, se vänstra bilden i figur 6.8. Denna typa av kurva kallas för ett *fasporträtt*. Tyvärr ger Eulers metod en mycket dålig approximation av x och y , ty man kan visa att fasporträttet till Lotka-Volterra-ekvationssystemet i första kvadranten av xy -planet är en sluten kurva – medan vi enligt figur 6.8 erhåller en spiralliknande kurva. Ökar vi N till tex 500, så blir det något bättre, se högra bilden i figur 6.8.

För att få en bättre approximation av lösningarna går vi över till Heuns respektive Runge-Kuttas metod. Heuns metod i detta exempel ges av

$$t_{n+1} = t_n + h$$

Figur 6.7. Blå kurva motsvarar $x(t)$ och röd kurva motsvarar $y(t)$.

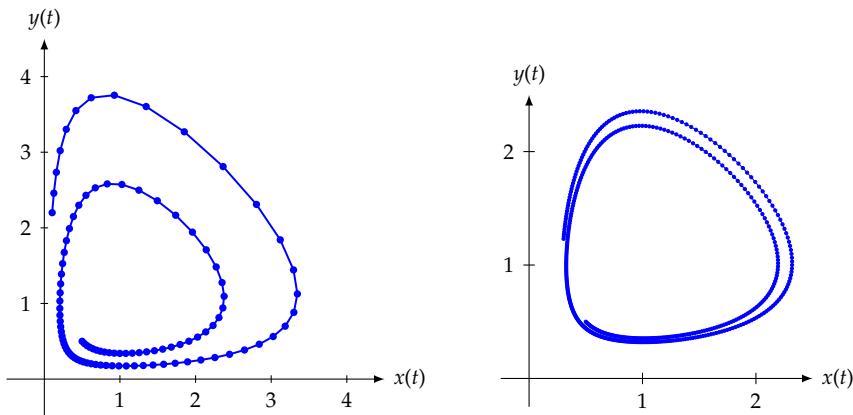
$$\begin{aligned} k_1 &= f(t_n, v_n) = (x_n - x_n y_n, -y_n + x_n y_n) \\ k_2 &= f(t_n + h, v_n + hk_1) = f(t_{n+1}, (x_n + h x_n - h x_n y_n, y_n - h y_n + h x_n y_n)) \\ v_{n+1} &= v_n + \frac{h}{2}(k_1 + k_2). \end{aligned}$$

Det lämnas som övning att bestämma uttrycken för komponenterna i vektorerna k_2 respektive y_{n+1} . Det lämnas som också som övning att ta fram motsvarande formler enligt Runge-Kuttas metod. Enligt figur 6.9 får vi nu ett bättre resultat eftersom approximationerna av lösningarna ser ut att ge en sluten kurva (vi går runt varv efter varv i samma kurva). Punkter i figur 6.9 som ligger nära varandra hör till olika varv. ♦

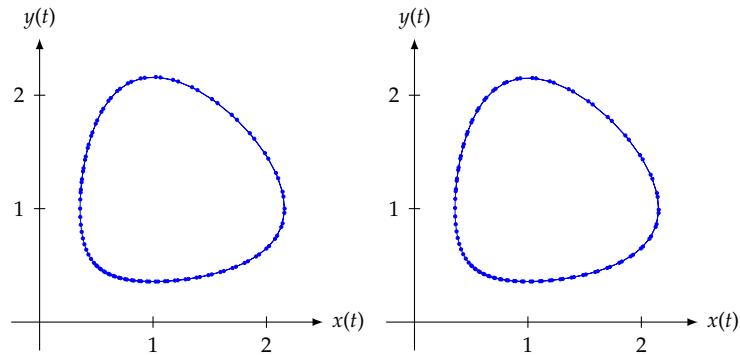
Anmärkning. Antag att $x > 0$ och $y > 0$, dvs att vi studerar en situationer där det existerar minst ett bytesdjur och minst ett rovdjur i respektive population. Ekvationssystem (6.2) är ekivalent med

$$\begin{cases} x' = \beta x \left(\frac{\alpha}{\beta} - y \right) \\ y' = \delta y \left(x - \frac{\gamma}{\delta} \right). \end{cases} \quad (6.3)$$

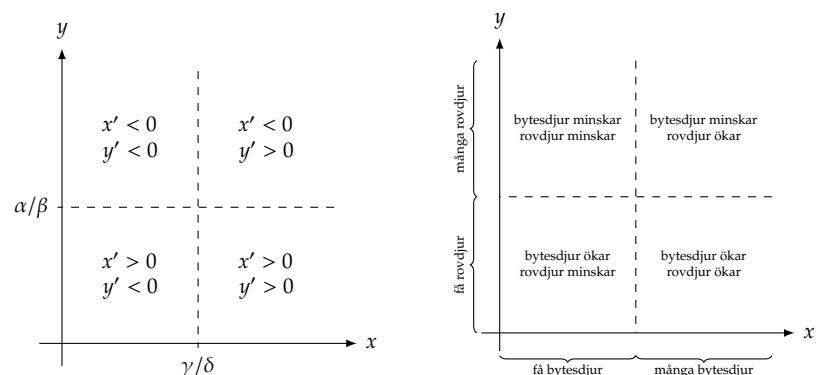
Vi ser att $x(t) = \gamma/\delta$ och $y(t) = \alpha/\beta$ är en lösning, som kallas *jämviktspunkt*. Att denna lösning kallas för jämviktspunkt beror på att x och y är konstanta funktioner, dvs deras derivator är lika med 0, vilket betyder att ingen förändring av antal rovdjur och bytesdjur äger rum. De två linjerna $x = \gamma/\delta$ och $y = \alpha/\beta$ delar in första kvadranten



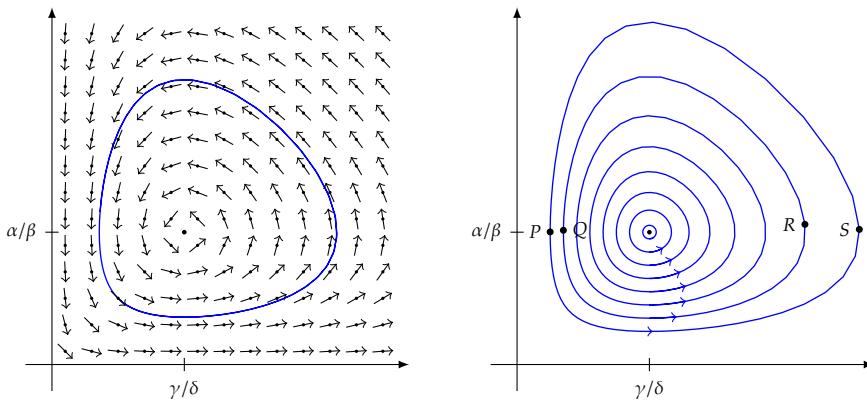
Figur 6.8



Figur 6.9. Fasporträtt för Heuns respektive Runge-Kuttas metod.



Figur 6.10. Teckenstudie och tolkning.



Figur 6.11

av xy -planet i fyra områden där x' och y' har konstant tecken, se figur 6.10. Jämför även med vänstra bilden i figur 6.11. Om $x = 0$ eller $y = 0$, så följer det från (6.3) att $x' = 0$ respektive $y' = 0$, dvs enligt Lotka-Volterras matematiska modell kan en utrotad art aldrig komma tillbaka.¹ Multiplicerar vi första och andra ekvationen i (6.3) med $(\gamma - \delta x)/x$ respektive $(\alpha - \beta y)/y$ erhåller vi

$$\begin{cases} \frac{1}{x}(\gamma - \delta x)x' = (\gamma - \delta x)(\alpha - \beta y) \\ \frac{1}{y}(\alpha - \beta y)y' = -(\gamma - \delta x)(\alpha - \beta y). \end{cases}$$

Alltså är

$$\frac{1}{x}(\gamma - \delta x)x' + \frac{1}{y}(\alpha - \beta y)y' = 0$$

eller ekvivalent

$$\frac{d}{dt}(\gamma \log x - \delta x + \alpha \log y - \beta y) = 0.$$

Sätt $h(x, y) = \delta x - \gamma \log x + \beta y - \alpha \log y$. Vi har visat att derivatan av h med avseende på t är lika med 0, dvs lösningarna till Lotka-Volterras ekvationssystem utgör en nivåkurva till h . En något mer ingående analys av h i första kvadranten visar att denna har ett lokalt minimum i jämnviktpunkten, att h är konvex och att samtliga nivåkurvor är slutna, se högra bilden i figur 6.11. Vi kan ge en ekologisk förklaring till fasporträttet. Låt $x(t)$ beskriva antalet skadeinsekter och $y(t)$ antalet småfåglar som lever på skadeinsekterna. Antag att vi befinner oss i punkten Q i högra bilden i figur 6.11 och att vi efter tex ett mänskligt ingripande hamnar i punkten P . En tid senare kan detta få som effekt att man befinner sig i R i stället för S , dvs antalet skadeinsekter har ökat – vilket troligtvis inte var avsikten med insatsen mot skadeinsektern.

6.5 Randvärdesproblem

Exempel 6.13 (Differensmetoden). Studera randvärdesproblemet

20160517

$$y'' + xy = \cos x, \quad y(0) = 1 \quad \text{och} \quad y(2) = 0.5.$$

¹Bortsett från en nyck i evolutionen.

Låt $a = 0$, $b = 2$ och $N = 8$. Då är $h = (b - a)/N = 0.25$. Vi vill bestämma en approximation $y_n \approx y(x_n)$, där $x_n = a + hn$ då $n = 0, 1, \dots, 8$. Vidare noterar vi att vi vet att $y_0 = 1$ och $y_8 = 0.5$. För $n = 1, 2, \dots, 7$ är

$$y''(x_n) \approx \frac{y(x_{n-1}) - 2y(x_n) + y(x_{n+1})}{h^2} \approx \frac{y_{n-1} - 2y_n + y_{n+1}}{h^2}.$$

Vi har härlett differensekvationerna

$$\begin{aligned} \frac{y_{n-1} - 2y_n + y_{n+1}}{h^2} + x_n y_n &= \cos x_n, \quad n = 1, 2, \dots, 7 \\ \Leftrightarrow \\ y_{n-1} - (2 - h^2 x_n) y_n + y_{n+1} &= h^2 \cos x_n, \quad n = 1, 2, \dots, 7. \end{aligned}$$

Om tex $n = 1$, så är

$$\begin{aligned} y_0 - (2 - h^2 x_1) y_1 + y_2 &= h^2 \cos x_1 \\ \Leftrightarrow \\ -(2 - 0.25^2 \cdot 0.25) y_1 + y_2 &= 0.25^2 \cos 0.25 - 1 \\ \Leftrightarrow \\ -1.98 y_1 + y_2 &= -0.94, \end{aligned}$$

och om $n = 2$, så

$$\begin{aligned} y_1 - (2 - h^2 x_2) y_2 + y_3 &= h^2 \cos x_2 \\ \Leftrightarrow \\ y_1 - 1.97 y_2 + y_3 &= 0.05. \end{aligned}$$

Tillsammans med de övriga fem ekvationerna erhåller vi det linjära ekvationssystemet

$$\left(\begin{array}{ccccccc} -1.98 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & -1.97 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1.95 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1.94 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1.92 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1.91 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1.89 \end{array} \right) \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \end{pmatrix} = \begin{pmatrix} -0.94 \\ 0.05 \\ 0.05 \\ 0.03 \\ 0.02 \\ 0.004 \\ 0.51 \end{pmatrix}.$$

Lösningen ges av $(0.941, 0.928, 0.941, 0.955, 0.944, 0.878, 0.735)$, se figur 6.12. \diamond

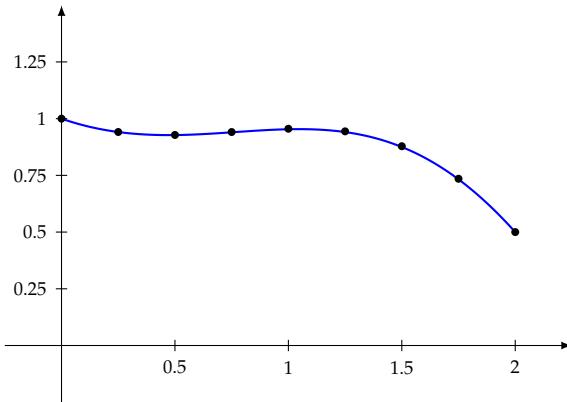
Exempel 6.14 (Inskjutningsmetoden). Antag att vi vill lösa randvärdesproblemet

$$y'' = -y \sin x - xy', \quad y(0) = 0 \quad \text{och} \quad y(3) = 1.5.$$

Låt $v = y'$ och antag att vi vet $y'(0) = v(0) = \gamma$. Då kan det ursprungliga problemet skrivas som

$$\begin{cases} y' = v \\ v' = -y \sin x - xv \end{cases} \quad \text{och} \quad \begin{cases} y(0) = 0 \\ v(0) = \gamma, \end{cases} \quad (6.4)$$

vilket är ett begynnelsevärdesproblem. Om vi kände till värdet på γ , så skulle vi kunna använda tex Runge-Kuttas metod för att finna en numerisk approximation till y . Låt $y(\cdot, \gamma)$ beteckna lösningen till ekvationssystemet ovan. Vi vill bestämma

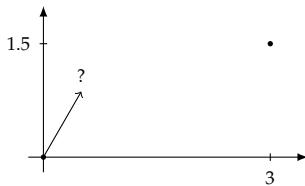


Figur 6.12. Punkterna motsvarar (x_n, y_n) medan blå kurva den exakta lösningen.

det reella tal γ sådan att $y(3, \gamma) = 1.5$. Om $g(\gamma) = y(3, \gamma) - 1.5$, så är γ lösningen till ekvationen $g(\gamma) = 0$. Låt oss använda sekantmetoden, dvs

$$\gamma_{k+1} = \gamma_k - g(\gamma_k) \frac{\gamma_k - \gamma_{k-1}}{g(\gamma_k) - g(\gamma_{k-1})},$$

där k är ett positivt heltal. Newton-Raphson metoden kräver att vi kan derivera g , vilket är svårt eftersom vi inte har ett explicit uttryck för g . Vi måste välja startvärden γ_0 och γ_1 innan vi kan använda sekantmetoden. Eftersom $\gamma = v(0) = y'(0)$ så motsvarar γ



Figur 6.13

riktningen på lösningskurvans tangent i punkten $(0, y(0)) = (0, 0)$, se figur 6.18. En naturlig gissning är att välja den riktning som siktar mot $(3, y(3)) = (3, 1.5)$, dvs

$$\gamma_0 = \frac{1.5 - 0}{3 - 0} = 0.5.$$

Runge-Kuttas metod med $N = 20$ och $h = 0.15$ ger resultatet i vänstra bilden i figur 6.14. Vi missar målet grovt och eftersom $y_{20} = 0.263428 < 1.5$ så är det rimligt att utgå från att riktningen måste ökas, dvs att vi sätter γ_1 till ett större värde än γ_0 , tex

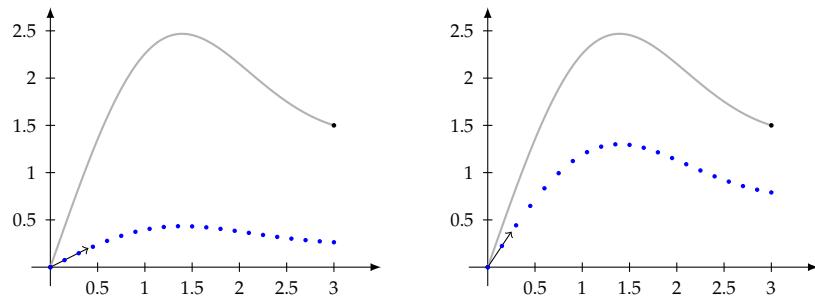
$$\gamma_1 = \gamma_0 + 1 = 1.5.$$

Med Runge-Kuttas metod får vi att $y_{20} = 0.790284 < 1.5$, se högra bilden i figur 6.14. Återigen missar vi målet $y_{20} = 1.5$. Vi har att

$$g(\gamma_0) = 0.263428 - 1.5 = -1.23657 \quad \text{och} \quad g(\gamma_1) = 0.790284 - 1.5 = -0.709716,$$

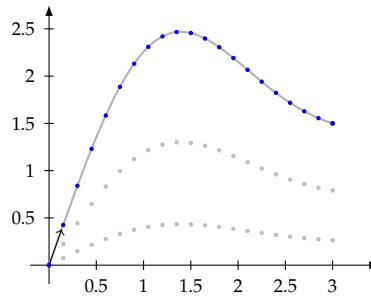
vilket ger att

$$\gamma_2 = \gamma_1 - g(\gamma_1) \frac{\gamma_1 - \gamma_0}{g(\gamma_1) - g(\gamma_0)} = 2.84708.$$



Figur 6.14. Grå kurva motsvarar den sökta lösningen.

Med hjälp av Runge-Kuttas metod och $\gamma = \gamma_2 = 2.84708$ för begynnelsevärdesproblem (6.4) får vi en bra approximation, se figur 6.15. \diamond



Figur 6.15

6.6 Övningsuppgifter

L

1. Låt $f(x) = \sin(x)$.

- Approximera derivatan $f'(0.8)$ med centraldifferens av ordning $O(h^2)$ och stegen $h = 0.1, h = 0.01$ och $h = 0.001$.
- Jämför med $f'(0.8) = \cos(0.8)$.
- Uppskatta trunkeringsfelet, där du använder

$$|f^{(3)}(c)| \leq \cos(0.7) \approx 0.764842187$$

i samtliga fall.

2. Låt $f(x) = xe^x$. Approximera $f'(0.2)$ med $D_0(0.1)$ och $D_1(0.1)$, samt beräkna det relativa felet i båda fallen. (20120821)

L

3. Låt $D(t)$ beteckna sträckan som ett objekt har färdats efter t tidsenheter. Man har samlat in följande data.

t	$D(t)$
8.0	17.453
9.0	21.460
10.0	25.752
11.0	30.301
12.0	35.084

Bestäm hastigheten $V(10)$ med numerisk differentiering. Jämför sedan resultatet med $D(t) = -70 + 7t + 70e^{-t/10}$.

4. Låt

$$f(x) = \frac{x^2}{x+1}.$$

Approximera $f'(0.25)$ med $D_0(0.05)$, dvs central differensapproximation, samt beräkna det absoluta felet. (20140822)

5. Låt f vara den reellvärda funktion som ges av

$$f(x) = \frac{1}{x^2 - 1},$$

där $x \in \mathbb{R} \setminus \{-1, 1\}$.

- (a) Approximera $f'(1.1)$ med $D_0(0.025)$ och $D_1(0.025)$.
 (b) Bestäm det relativa felet för $D_0(0.025)$ och $D_1(0.025)$. (20150108)

- L 6. Låt $f(x) = x \exp(\sqrt{x}) = xe^{\sqrt{x}}$ och $h = 0.01$.

- (a) Vilka centraldifferenser $D_0(2^k h)$ krävs för att man med Richardsons extrapolation ska kunna bestämma $D_4(h)$?
 (b) Approximera $f'(1.2)$ med $D_4(h)$.
 (c) Bestäm absolutfelet. (20130607)

- L 7. Låt $f(t, y) = 3y + 3t$ och $R = \{(t, y) : 0 \leq t \leq 3 \text{ och } 0 \leq y \leq 5\}$.

- (a) Visa att $y(t) = Ce^{3t} - t - 1/3$ är en lösning till differentialekvationen $y' = f(t, y)$.
 (b) Visa att f uppfyller ett Lipschitzvillkor på rektangeln R och bestäm motsvarande Lipschitzkonstant.

8. Låt $f(x, y) = x^2 y$, där $0 \leq x \leq 2$. Visa att f uppfyller Lipschitzvillkoret och bestäm Lipschitzkonstanten L . (20120821)

- L 9. Låt

$$f(x, y) = \frac{xy}{x+1},$$

där $0 \leq x \leq 2$.

- (a) Visa att f uppfyller ett Lipschitzvillkor.
 (b) Approximera $y(2)$ med Eulers metod då $N = 4$, där y är lösningen till begynnelsevärdesproblemet $y' = f(x, y)$ och $y(0) = 0.5$. (20150108)

10. Låt C vara en konstant. Visa att $y(x) = x - 1 + Ce^{-x}$ är en lösning till differentialekvationen $y' = x - y$. Bestäm därefter C då differentialekvationen ingår ett begynnelsevärdesproblem med var och ett av följande begynnelsevillkor.

- (a) $y(0) = 1$ (b) $y(0) = 0$ (c) $y(0) = -1$ (d) $y(0) = -2$

- L 11. Approximera $y(0.4)$ med Eulers metod, där y är lösningen till begynnelsevärdesproblemet

$$y' = -ty, \quad y(0) = 1.$$

Använd stegen $h = 0.2$ respektive $h = 0.1$. Jämför sedan resultaten med den exakta lösningen $y(t) = e^{-t^2/2}$.

- 12.** Låt $y(t)$ beteckna en lösning till begynnelsevärdesproblemet

$$y'' = \frac{y - 5y'}{1 + t}, \quad y(0) = 0.8, \quad y'(0) = 1.5, \quad 0 \leq t \leq 0.8.$$

Skriv om differentialekvationen till ett system av differentialekvationer av första ordningen och bestäm sedan en numerisk approximation av $y(t)$ med hjälp av Eulers metod, med steget $h = 0.2$. (20130823)

- 13.** Låt $f(x, y) = x^2y - x$.

(a) Visa att f uppfyller ett Lipschitzvillkor för alla $x \in [0, 2]$ samt ange det minsta möjliga värdet på Lipschitzkonstanten L .

- (b) Låt $y(x)$ beteckna lösningen till begynnelsevärdesproblemet

$$y' = f(x, y) \quad \text{och} \quad y(0) = 1.5, \quad \text{där} \quad 0 \leq x \leq 2.$$

Approximera $y(2)$ med hjälp av Eulers metod och steget $h = 0.5$. (20140822)

- 14.** Låt $y(x)$ vara lösningen till begynnelsevärdesproblemet

$$e^x y' + xy^2 = 0 \quad \text{där} \quad y(0.0) = 2.3.$$

Approximera $y(1.0)$ med Eulers metod och $N = 5$. (20150822)

- L 15.** Approximera lösningen till begynnelsevärdesproblemet

$$y' = \frac{\cos y}{x^2 + 1} \quad \text{och} \quad y(0) = 1 \quad (0 \leq x \leq 3)$$

med Heuns metod då $N = 3$. (20140603)

- L 16.** Approximera $y(0.4)$ med Heuns metod, där y är lösningen till begynnelsevärdesproblemet

$$y' = e^{-2x} - 2y, \quad y(0) = 0.1.$$

Använd steget $h = 0.2$ respektive $h = 0.1$. Jämför sedan resultaten med den exakta lösningen $y(x) = (0.1 + x)e^{-2x}$.

- L 17.** Approximera $y(0.6)$ med Heuns metod och steget $h = 0.2$, där y är lösningen till begynnelsevärdesproblemet

$$y' = xy - 2x, \quad y(0) = 1.5.$$

Jämför resultatet med den exakta lösningen $y(x) = 2 - 0.5e^{x^2/2}$. (20120603)

- 18.** Approximera $y(0.2)$ med Heuns metod, där y är lösningen till begynnelsevärdesproblemet

$$y' = \sin(x - 0.4y), \quad y(0) = 0.4.$$

Använd steget $h = 0.05$. (20130109)

- L 19.** Använd Heuns metod och steget $h = 0.25$ för att finna en approximation av $y(1)$, där y är lösningen till begynnelsevärdesproblemet

$$(1 + x^2)y' + y - x = 0, \quad y(0) = 0.8 \quad (0 \leq x \leq 1). \quad (20130607)$$

- L 20. Approximera $y(1.0)$ med Heuns metod och steget $h = 0.25$, där y är lösningen till begynnelsevärdesproblemets

$$y' = x - y^2, \quad y(0) = 0.6. \quad (20140110)$$

- L 21. Modifiera Runge-Kuttas metod av ordning $O(h^4)$ så att den kan användas för att lösa ett system av differentialekvationer. Lös med denna differentialekvationssystemet i exempel 6.11.

Facit

1 Felanalys och datoraritmetik

1. De positiva flyttalen:

$1.000 \cdot 2^{-2} = 0.25$	$1.001 \cdot 2^{-2} = 0.28125$	$1.010 \cdot 2^{-2} = 0.3125$
$1.011 \cdot 2^{-2} = 0.34375$	$1.100 \cdot 2^{-2} = 0.375$	$1.101 \cdot 2^{-2} = 0.40625$
$1.110 \cdot 2^{-2} = 0.4375$	$1.111 \cdot 2^{-2} = 0.46875$	$1.000 \cdot 2^{-1} = 0.5$
$1.001 \cdot 2^{-1} = 0.5625$	$1.010 \cdot 2^{-1} = 0.625$	$1.011 \cdot 2^{-1} = 0.6875$
$1.100 \cdot 2^{-1} = 0.75$	$1.101 \cdot 2^{-1} = 0.8125$	$1.110 \cdot 2^{-1} = 0.875$
$1.111 \cdot 2^{-1} = 0.9375$	$1.000 \cdot 2^0 = 1$	$1.001 \cdot 2^0 = 1.125$
$1.010 \cdot 2^0 = 1.25$	$1.011 \cdot 2^0 = 1.375$	$1.100 \cdot 2^0 = 1.5$
$1.101 \cdot 2^0 = 1.625$	$1.110 \cdot 2^0 = 1.75$	$1.111 \cdot 2^0 = 1.875$
$1.000 \cdot 2^1 = 2$	$1.001 \cdot 2^1 = 2.25$	$1.010 \cdot 2^1 = 2.5$
$1.011 \cdot 2^1 = 2.75$	$1.100 \cdot 2^1 = 3$	$1.101 \cdot 2^1 = 3.25$
$1.110 \cdot 2^1 = 3.5$	$1.111 \cdot 2^1 = 3.75$	$1.000 \cdot 2^2 = 4$
$1.001 \cdot 2^2 = 4.5$	$1.010 \cdot 2^2 = 5$	$1.011 \cdot 2^2 = 5.5$
$1.100 \cdot 2^2 = 6$	$1.101 \cdot 2^2 = 6.5$	$1.110 \cdot 2^2 = 7$
$1.111 \cdot 2^2 = 7.5$	$1.000 \cdot 2^3 = 8$	$1.001 \cdot 2^3 = 9$
$1.010 \cdot 2^3 = 10$	$1.011 \cdot 2^3 = 11$	$1.100 \cdot 2^3 = 12$
$1.101 \cdot 2^3 = 13$	$1.110 \cdot 2^3 = 14$	$1.111 \cdot 2^3 = 15$

2. (a) $11001_{\text{två}}$

(b) $110.01_{\text{två}}$

(c) $10.11_{\text{två}}$

3. (a) $\frac{45}{8} = 5.625$

(b) $\frac{27}{32} = 0.84375$

(c) $\frac{1}{32} = 0.03125$

5. (a) $11011_{\text{två}}$

(b) $1.11_{\text{två}}$

(c) $0.111111_{\text{två}}$

7. (a) $1.1000 \cdot 2^1$

(b) $1.0011 \cdot 2^{-1}$

(c) $1.1010 \cdot 2^{-3}$

8. $\hat{x} = 1.110 \cdot 2^0$ och $\delta x = \hat{x} - x \approx 0.0179492$. Notera att

$$\hat{x} = \left(1 + \frac{1}{2} + \frac{1}{2^2} + \frac{0}{2^3}\right) \cdot 2^0 = \frac{7}{4} \approx 1.75,$$

vilket kan jämföras med $\sqrt{3} \approx 1.73205$.

- 9.** (a) $\text{fl}(a) = 1.1011 \cdot 2^3$ och $\text{fl}(b) = 1.1101 \cdot 2^{-3}$
(b) $\delta a = 0.1$ och $\delta b = 0.0005625$
(c) $\text{fl}(a) + \text{fl}(b) = 1.1011 \cdot 2^3$ och $\delta(a+b) = 0.126$

10. $5.2 \leq x \leq 5.3$

2 Icke-linjära ekvationer

Med startvärdet $x_0 = 0.5$ fås

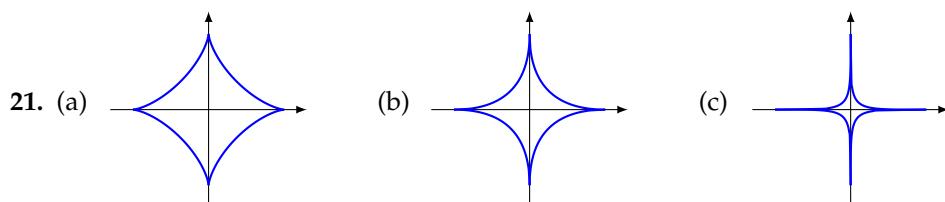
$$x_1 = 0.656735, \quad x_2 = 0.652919, \quad x_3 = 0.652919, \\ x_4 = 0.652919 \quad \text{och} \quad x_5 = 0.652919$$

Observera att valet av startvärde är inte givet i uppgiften och därför finns det inget entydigt bestämt svar.

- $$18. x_{n+1} = \frac{x_n^2 + e^{-x_n}}{1 + x_n}; 0.68394, 0.577454, 0.56723, 0.567143, 0.567143$$

19. $\sqrt{7} \approx 2.64575$

- 20.** $(x_1, x_2, x_3, x_4) = (0.827551, 0.824139, 0.824132, 0.824132, 0.824132)$



24. (a) $\frac{\partial f}{\partial x} = y^2 + ye^{xy}$ och $\frac{\partial f}{\partial y} = 2xy + xe^{xy}$

- (b) $\frac{\partial f}{\partial x} = y^2 - y^2 e^{xy}$ och $\frac{\partial f}{\partial y} = 2xy - (1 + xy)e^{xy}$
- (c) $\frac{\partial f}{\partial x} = x^{\sin(xy)-1} \{xy \cos(xy) \ln(x) - \sin(xy)\}$ och $\frac{\partial f}{\partial y} = x^{\sin(xy)-1} \cos(xy) \ln(x)$
- (d) $\frac{\partial f}{\partial x} = -z \left\{ 1 + \ln \frac{x}{y} \right\} \sin \left(xz \ln \frac{x}{y} \right)$, $\frac{\partial f}{\partial y} = \frac{xz}{y} \sin \left(xz \ln \frac{x}{y} \right)$
 och $\frac{\partial f}{\partial z} = -x \ln \frac{x}{y} \sin \left(xz \ln \frac{x}{y} \right)$

28. $2 < |a| < 3 < |b| < 4$

29. $p_1 = (2, -1)$, $p_2 = (1.5, -0.7)$, $p_3 = (1.65, -0.79)$; Ja metoden konvergerar eftersom matrisen är strikt diagonaldominant.

31. Iterationsformel:

$$\mathbf{v}_n = (x_n, y_n) = g(\mathbf{v}_{n-1}) = g(x_{n-1}, y_{n-1}) = \left(\frac{1 - y_{n-1}}{2}, \frac{2 + x_n}{3} \right).$$

Efter iterationer får vi

$$\mathbf{v}_1 = (0.25, 0.75), \quad \mathbf{v}_2 \approx (0.125, 0.7083) \quad \text{och} \quad \mathbf{v}_3 \approx (0.1458, 0.7153).$$

Exakt lösning är $\mathbf{v} = (1/7, 5/7)$. Det absoluta felet är $\|\mathbf{v} - \mathbf{v}_3\|_2 \approx 0.0031$.

3 Interpolation

8. $1.21987x(1-x)$

11. Vi får följande tabell.

k	x_k	$f[\cdot]$	$f[\cdot, \cdot]$	$f[\cdot, \cdot, \cdot]$
0	2.00	1.09861		
			0.320171	
1	2.25	1.17865		-0.0474779
			0.296432	
2	2.50	1.25276		

Alltså är $p_2(x) = 1.09861 + 0.320171(x - 2) - 0.0474779(x - 2)(x - 2.25)$, vilket kan förenklas till $p_2(x) = 0.24462 + 0.521952x - 0.0474779x^2$.

12. $f[x_0] = x_0^3$, $f[x_0, x_1] = x_0^2 + x_0 x_1 + x_1^2$ och $f[x_0, x_1, x_2] = x_0 + x_1 + x_2$

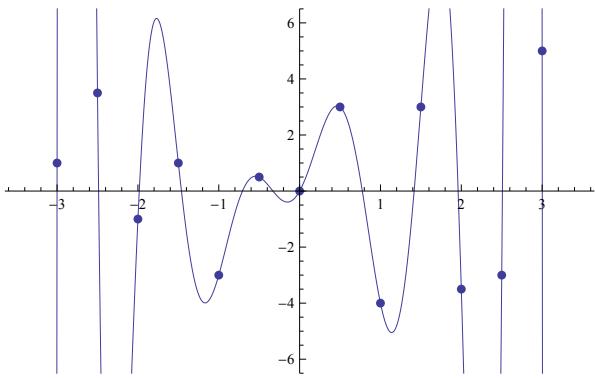
13. (a) $L_{12,0}(x) = \frac{4}{467775} \prod_{k=-5}^6 \left(x - \frac{k}{2} \right)$ och $L_{12,5}(x) = -\frac{32}{4725} \prod_{k=-6, k \neq -1}^6 \left(x - \frac{k}{2} \right)$

(b) Se tabell 6.2.

(c) $p_{12}(x) = -\frac{2696}{42525}x^{12} - \frac{16}{825}x^{11} + \frac{58676}{42525}x^{10} + \frac{19}{45}x^9 - \frac{300077}{28350}x^8 - \frac{10121}{3150}x^7 + \frac{2935921}{85050}x^6 + \frac{7387}{720}x^5 - \frac{30621791}{680400}x^4 - \frac{5257}{400}x^3 + \frac{1228799}{75600}x^2 + \frac{47983}{9240}x$, se figur 6.16

-3	1	5	-14	18	$-\frac{29}{3}$	$-\frac{203}{15}$	$-\frac{412}{15}$	$-\frac{754}{35}$	$-\frac{1432}{315}$	$-\frac{2038}{405}$	$-\frac{2696}{4225}$
-2	-1	$\frac{7}{2}$	-9	13	$-\frac{50}{3}$	$-\frac{203}{15}$	$-\frac{68}{5}$	$-\frac{76}{9}$	$-\frac{1432}{315}$	$-\frac{241}{105}$	$-\frac{2696}{13525}$
-1	$-\frac{3}{2}$	1	4	-12	$-\frac{50}{3}$	$-\frac{203}{15}$	$-\frac{59}{3}$	$-\frac{176}{15}$	$-\frac{112}{15}$	$-\frac{22}{63}$	$-\frac{632}{367}$
0	$-\frac{1}{2}$	$-\frac{1}{2}$	-8	-15	$-\frac{46}{3}$	$-\frac{38}{3}$	$-\frac{32}{3}$	$-\frac{394}{15}$	$-\frac{857}{15}$	$-\frac{4042}{315}$	$-\frac{6536}{31185}$
1	-1	-3	7	15	-18	$-\frac{59}{3}$	$-\frac{112}{15}$	$-\frac{203}{15}$	$-\frac{656}{105}$	$-\frac{388}{105}$	$-\frac{3046}{405}$
2	$-\frac{1}{2}$	$-\frac{1}{2}$	-1	-8	10	-14	25	$-\frac{118}{9}$	$-\frac{128}{45}$	$-\frac{228}{405}$	$-\frac{4725}{405}$
3	0	0	0	6	-20	28	32	$-\frac{103}{2}$	$-\frac{272}{15}$	$-\frac{288}{315}$	$-\frac{272}{315}$
4	$-\frac{1}{2}$	$-\frac{1}{2}$	3	-14	28	-14	-27	$-\frac{110}{3}$	$-\frac{32}{3}$	$-\frac{134}{9}$	$-\frac{1384}{315}$
5	1	-4	14	3	-20	32	-13	$-\frac{103}{2}$	14	1	15
6	$-\frac{1}{2}$	$-\frac{1}{2}$	3	-13	32	32	-27	$-\frac{103}{2}$	$-\frac{272}{15}$	$-\frac{288}{315}$	$-\frac{272}{315}$
7	2	$-\frac{7}{2}$	1	1	14	14	1	$-\frac{40}{3}$	$-\frac{40}{3}$	15	16
8	$-\frac{5}{2}$	3	5	5	15	16	3	$-\frac{40}{3}$	$-\frac{40}{3}$	15	16

Tabell 6.2. "Newtons gran"



Figur 6.16

15. $a = 2, b = 4$ och $c = 12$

16. $k_0 - k_n = 0$ och

$$4h_n k_0 + 2h_n k_1 + 2h_1 k_{n-1} + 4h_1 k_n = 6 \left(h_n \frac{y_0 - y_1}{h_1} + h_1 \frac{y_n - y_{n-1}}{h_n} \right)$$

17. $S_2(x) = -2 + 8x - 7x^2 + 2x^3$

18. (a) $(3t, 3t)$

(b) $(6t - 6t^2 + 3t^3, 15t - 24t^2 + 9t^3)$

(c) $(4.5t - 2.5t^3, 9t - 15t^2 + 7t^3)$

(d) $(3t, 6t - 9t^2 + 6t^3)$

(e) $(1.5 - 4.5t + 13.5t^2 - 9t^3, 6t - 4.5t^2 - 1.5t^3)$

(f) $(3t - 3t^2 + 3t^3), 6t - 6t^2)$

(g) $(6t - 9t^2 + 6t^3, 6t - 6t^2)$

(h) $(1 + 6t - 15t^2 + 10t^3, 9t - 9t^2)$

20. $b_3(t) = (-1 + 6t - 3t^2 - 3t^3, 2 - 6t + 15t^2 - 11t^3)$

22. $(1, 2)$

24. (a) $(-8t^2 + 8t - 1, t^2 + 4t - 2)$

(b) $t = \frac{1}{4}(2 \pm \sqrt{2})$, och $y = \frac{1}{8}(3 \pm 10\sqrt{2})$

25. $P_0 = (-1, 1), P_1 = (1, 0), P_2 = (2, 3)$ och $P_3 = (0, 1)$

4 Numerisk integration

1. $\frac{7}{2}$

2. 0.674865

4. $S(f, 0.15) \approx 8.38874$ och $R_I \approx 0.00004$

5. 0.65933

6. 1.44465

5 Numerisk linjär algebra

1. $x = 0.1892, y = -0.1513, z = 0.05766$

2. (a) $(0.19, -0.15, 0.058)$ (b) $(0.18, -0.15, 0.060)$

3. $(x, y, z) = (1.20321, 1.44385, 1.5508)$

$$4. \mathbf{P}_{12} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \text{ och } \mathbf{P}_{25} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix}.$$

5. Två alternativ är $P = P_{1,2}P_{2,5}P_{3,5}P_{5,6}$ och $P = P_{5,6}P_{3,5}P_{2,3}P_{1,3}$.

10. $L = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0.5 & -0.375 & 1 \end{pmatrix}$ och $U = \begin{pmatrix} 2 & 1 & -1 \\ 0 & 4 & 4 \\ 0 & 0 & 5 \end{pmatrix}$

11. $L = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 5 & 6 & 1 \end{pmatrix}$ och $U = \begin{pmatrix} 1 & 2 & 3 \\ 0 & -1 & -2 \\ 0 & 0 & -1 \end{pmatrix}$

13. $L = \begin{pmatrix} 1 & 0 & 0 \\ 1/2 & 1 & 0 \\ 1/2 & 1/3 & 1 \end{pmatrix}$, $U = \begin{pmatrix} 2 & 1 & 1 \\ 0 & 9/2 & 3/2 \\ 0 & 0 & -2 \end{pmatrix}$ och $P = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$

14. $L = \begin{pmatrix} 1 & 0 & 0 \\ 1/2 & 1 & 0 \\ -1/4 & 1/2 & 1 \end{pmatrix}$, $U = \begin{pmatrix} 4 & -1 & 1 \\ 0 & -1/2 & 1/2 \\ 0 & 0 & 1 \end{pmatrix}$ och $P = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$

24. $f(x) = 1.60333x + 0.0321883$ och $E_2(f) = 0.369771$

25. $f(x) = 0.313351x + 1.86158$ och $E_2(f) = 1.48654$ (dålig anpassning)

26. $a = \frac{11}{6}$, $b = 0$ och $c = \frac{4}{3}$

27. $p(x) = \frac{x^2}{14} + \frac{x}{10} + \frac{6}{7}$

28. $\frac{1}{2}x^2 - \frac{7}{10}x - \frac{2}{5}$

31. $f(x) = 0.615504x^2 - 0.245539x - 0.224813$ och $E_2(f) = 0.0873087$

32. (a) $f_1(x) = -0.461x + 0.937$
 $E_2(f_1) = 0.731$

(b) $f_2(x) = 0.374x^2 - 0.473x + 0.127$
 $E_2(f_2) = 0.351$

(c) $f_3(x) = -0.176x^3 + 0.360x^2 + 0.150x + 0.193$
 $E_2(f_3) = 0.0970$

(d) $f_4(x) = 0.025x^4 - 0.175x^3 + 0.241x^2 + 0.146x + 0.261$
 $E_2(f_4) = 0.0839$

(e) $f_5(x) = 0.060x^5 + 0.028x^4 - 0.526x^3 + 0.219x^2 + 0.533x + 0.289$
 $E_2(f_5) = 0.996 \cdot 10^{-15}$

33. $f(x) = 4.684xe^{-0.788x}$, se figur 6.17

6 Ordinära differentialekvationer

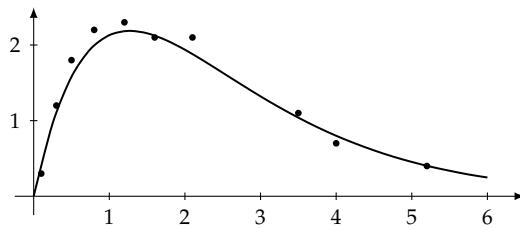
2. $D_0(0.1) \approx 1.4722$ med felet $R \approx 0.00444806$

$D_1(0.1) \approx 1.46566$ med felet $R \approx 0.00001447$

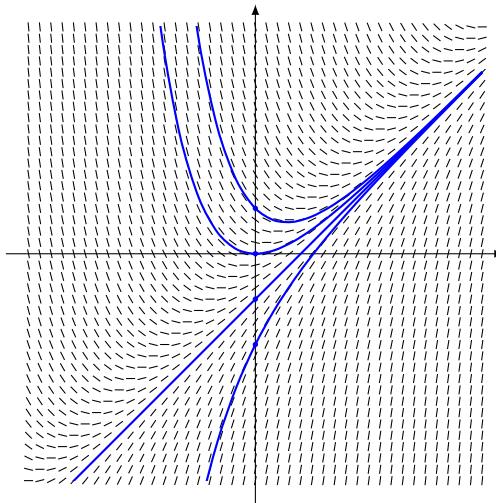
$f'(x) = (x+1)e^x$, $f'(0.2) \approx 1.46568$

4. $D_0(0.05) = 0.358974$ och $E = 0.00102564$, där $f'(0.25) = 0.36$

5. (a) $D_0(0.025) = -53.2199$ och $D_1(0.025) = -48.7755$



Figur 6.17



Figur 6.18

- (b) 0.0668179 respektive 0.0222727
6. (a) $D_0(h), D_0(2h), D_0(2^2h), D_0(2^3h)$ och $D_0(2^4h)$
 (b) $D_4(0.01) = 4.6284852173903$
 (c) $E_x = -7.99361 \cdot 10^{-15}$
8. $L = 4$

9. (b) $y(2) \approx 0.947917$; Exakt svar: $y(x) = \frac{e^x}{2(x+1)}$

10. (a) $C = 2$ (b) $C = 1$ (c) $C = 0$ (d) $C = -1$

För riktningsfält och de fyra lösningarna, se figur 6.18.

12. Med $x_1 = y$ och $x_2 = y'$, så är

$$\begin{cases} x'_1 = x_2 \\ x'_2 = \frac{x_1 - 5x_2}{1+t} \end{cases}$$

Eulers metod ger punkterna $y(0) = 0.8, y(0.2) = 1.1, y(0.4) = 1.132, y(0.6) = 1.174$ och $y(0.8) = 1.21834$.

13. (a) $L = 4$
 (b) $y(2) \approx 2.76953$

- 14.** $y(1.0) \approx 1.4547$, där $y(0.2) \approx 2.3$, $y(0.4) \approx 2.12676$, $y(0.6) \approx 1.8842$, $y(0.8) \approx 1.65039$
- 15.** $y(0) = 1$, $y(1) \approx 1.27777$, $y(2) \approx 1.36479$ och $y(3) \approx 1.39346$
- 17.** $y(0.6) \approx 1.40184$ och $E = 0.000452$
- 18.** 0.388783
- 19.** 0.64186
- 20.** 0.742269

Lösningsförslag

1 Felanalys och datoraritmetik

Uppgift 1. De aktuella talen är de som kan skrivas på formen

$$x = \pm 1.d_1d_2d_3 \cdot 2^e = \pm(1 + d_1 \cdot 2^{-1} + d_2 \cdot 2^{-2} + d_3 \cdot 2^{-3}) \cdot 2^e,$$

där $d_1 = 1, d_2, d_3 \in \{0, 1\}$, och $-2 \leq e \leq 3$, tex har vi att

$$1.101 \cdot 2^1 \quad \text{och} \quad -1.011 \cdot 2^{-2}$$

motsvarar decimaltalen 1.625 respektive -0.34375.

Uppgift 2. (a) Vi får följande uppställning.

$$\begin{aligned} 25 &= 12 \cdot 2 + 1 \\ 12 &= 6 \cdot 2 + 0 \\ 6 &= 3 \cdot 2 + 0 \\ 3 &= 1 \cdot 2 + 1 \\ 1 &= 0 \cdot 2 + 1 \end{aligned}$$

Alltså är $25_{\text{tio}} = 11001_{\text{två}}$. (b) Från föregående deluppgifte följer att

$$\begin{aligned} \frac{25}{4} &= 2^{-2}(1 \cdot 2^4 + 1 \cdot 2^3 + 0 \cdot 2^2 + 0 \cdot 2^1 + 1) \\ &= 1 \cdot 2^2 + 1 \cdot 2 + 0 + 0 \cdot 2^{-1} + 1 \cdot 2^{-2}, \end{aligned}$$

dvs $25/4 = 110.01_{\text{två}}$. (c) Låt x vara ett icke-negativt reellt tal samt låt $\lfloor x \rfloor$ och $\{x\}$ beteckna heltalsdelen respektive bråktalsdelen av x . Då är

$$h = \lfloor 2.75 \rfloor = 2 = 10_{\text{två}} \quad \text{och} \quad f_0 = \{2.75\} = 0.75$$

Sätt

$$d_1 = \lfloor 2f_0 \rfloor = \lfloor 2 \cdot 0.75 \rfloor = \lfloor 1.5 \rfloor = 1 \quad \text{och} \quad f_1 = \{2 \cdot 0.75\} = \{1.5\} = 0.5.$$

Eftersom $d_1 \neq 0$, så fortsätter vi och får att

$$d_2 = \lfloor 2f_1 \rfloor = \lfloor 2 \cdot 0.5 \rfloor = \lfloor 1.0 \rfloor = 1 \quad \text{och} \quad f_2 = \{2 \cdot 0.5\} = \{1.0\} = 0.$$

Alltså är $f = 11_{\text{två}}$ och $2.75 = h + f = 10.11_{\text{två}}$. Alternativ lösning. Notera först att

$$2.75 = 2 + 0.75 = 2 + \frac{3}{4} = \frac{11}{4}.$$

Från uppställningen

$$\begin{aligned} 11 &= 5 \cdot 2 + 1 \\ 5 &= 2 \cdot 2 + 1 \\ 2 &= 1 \cdot 2 + 0 \\ 1 &= 0 \cdot 2 + 1 \end{aligned}$$

följer att $11 = 1011_{\text{två}}$. När vi multiplicerar 11 med $1/4 = 2^{-2}$ flyttar vi punkten för bråktalsdelen två steg åt vänster, dvs $2.75 = 10.11_{\text{två}}$.

Uppgift 4. (a) Vi har att

$$\begin{aligned} 1.0110101_{\text{två}} &= 1 + 0 \cdot 2^{-1} + 1 \cdot 2^{-2} + 1 \cdot 2^{-3} + 0 \cdot 2^{-4} + 1 \cdot 2^{-5} + 0 \cdot 2^{-6} + 1 \cdot 2^{-7} \\ &= \frac{181}{128} = 1.4140625. \end{aligned}$$

Om vi multiplicerar $1.0110101_{\text{två}}$ med $2^7 = 128$, så får vi att $10110101_{\text{två}} = 181_{\text{tio}}$. (b) På samma sätt som i (a) får vi att

$$\begin{aligned} 11.0010010001_{\text{två}} &= 1 \cdot 2 + 1 + 1 \cdot 2^{-3} + 1 \cdot 2^{-6} + 1 \cdot 2^{-10} \\ &= \frac{3217}{1024} = 3.1416015625. \end{aligned}$$

Uppgift 6. Först noterar vi att samtliga nämnare är tvåpotenser, dvs lika med 2^k för något heltal k . Då kan vi bestämma den binära representationen av respektive täljare och sedan multiplicera med 2^{-k} . Vi får följande uppställningar.

(a) $7 = 3 \cdot 2 + 1$	(b) $15 = 7 \cdot 2 + 1$	(c) $23 = 11 \cdot 2 + 1$	(d) $75 = 37 \cdot 2 + 1$
$3 = 1 \cdot 2 + 1$	$7 = 3 \cdot 2 + 1$	$11 = 5 \cdot 2 + 1$	$37 = 18 \cdot 2 + 1$
$1 = 0 \cdot 2 + 1$	$3 = 1 \cdot 2 + 1$	$5 = 2 \cdot 2 + 1$	$18 = 9 \cdot 2 + 0$
	$1 = 0 \cdot 2 + 1$	$2 = 1 \cdot 2 + 0$	$9 = 4 \cdot 2 + 1$
			$4 = 2 \cdot 2 + 0$
			$2 = 1 \cdot 2 + 0$
			$1 = 0 \cdot 2 + 1$

Alltså är $7_{\text{tio}} = 111_{\text{två}}$, $15_{\text{tio}} = 1111_{\text{två}}$, $23_{\text{tio}} = 1011_{\text{två}}$ och $75_{\text{tio}} = 1001011_{\text{två}}$. Det ger oss att

$$\frac{7}{16} = 2^{-4}(1 \cdot 2^2 + 1 \cdot 2 + 1) = 1 \cdot 2^{-2} + 1 \cdot 2^{-3} + 1 \cdot 2^{-4} = 0.0111_{\text{två}}.$$

Med andra ord, flytta punkten för bråktalsdelen i $7 = (111.000\dots)_{\text{två}}$ fyra steg åt vänster. På samma sätt finner vi att

$$\frac{15}{16} = 0.1111_{\text{två}}, \quad \frac{23}{32} = 0.10111_{\text{två}} \quad \text{och} \quad \frac{75}{128} = 0.1001011_{\text{två}},$$

eftersom $32 = 2^5$ och $128 = 2^7$.

Uppgift 9. (a) För att bestämma den binära representationen för heltalsdelen av a använder vi divisionsalgoritmen enligt följande uppställning.

$$\begin{aligned} 13 &= 6 \cdot 2 + 1 \\ 6 &= 3 \cdot 2 + 0 \\ 3 &= 1 \cdot 2 + 1 \end{aligned}$$

$$1 = 0 \cdot 2 + 1$$

Alltså är $13 = 1101_{\text{två}}$. Det behövs en siffra till. Sätt $f = \{13.4\} = 0.4$. Algoritmen för att i tur och ordning bestämma siffrorna i den binära bråktalsdelen av a ger då att

$$\begin{aligned}s_1 &= \lfloor 2f \rfloor = \lfloor 0.8 \rfloor = 0 \\f &= \{2f\} = \{0.8\} = 0.8 \\s_2 &= \lfloor 2f \rfloor = \lfloor 1.6 \rfloor = 1 \\f &= \{2f\} = \{1.6\} = 0.6 \\s_3 &= \lfloor 2f \rfloor = \lfloor 1.2 \rfloor = 1.\end{aligned}$$

Alltså är $13.4 = (1101.011\dots)_{\text{två}}$. Med $t = 5$ siffrors noggrannhet är

$$\begin{aligned}13.4 &\approx 1101.1_{\text{två}} = 1 \cdot 2^3 + 1 \cdot 2^2 + 0 \cdot 2 + 1 + 1 \cdot 2^{-1} \\&= (1 \cdot 2^{-1} + 1 \cdot 2^{-2} + 0 \cdot 2^{-3} + 1 \cdot 2^{-4} + 1 \cdot 2^{-5}) \cdot 2^4 \\&= 1.1011 \cdot 2^3\end{aligned}$$

Låt nu $f = 0.226$. På samma sätt finner vi att

$$\begin{aligned}s_1 &= \lfloor 2f \rfloor = \lfloor 0.452 \rfloor = 0 \\f &= \{2f\} = \{0.452\} = 0.452 \\s_2 &= \lfloor 2f \rfloor = \lfloor 0.904 \rfloor = 0 \\f &= \{2f\} = \{0.904\} = 0.904 \\s_3 &= \lfloor 2f \rfloor = \lfloor 1.808 \rfloor = 1 \\f &= \{2f\} = \{1.808\} = 0.808 \\s_4 &= \lfloor 2f \rfloor = \lfloor 1.616 \rfloor = 1 \\f &= \{2f\} = \{1.616\} = 0.616 \\s_5 &= \lfloor 2f \rfloor = \lfloor 1.232 \rfloor = 1 \\f &= \{2f\} = \{1.232\} = 0.232 \\s_6 &= \lfloor 2f \rfloor = \lfloor 0.464 \rfloor = 0 \\f &= \{2f\} = \{0.464\} = 0.464 \\s_7 &= \lfloor 2f \rfloor = \lfloor 0.928 \rfloor = 0 \\f &= \{2f\} = \{0.928\} = 0.928 \\s_8 &= \lfloor 2f \rfloor = \lfloor 1.856 \rfloor = 1 \\f &= \{2f\} = \{1.856\} = 0.856 \\s_9 &= \lfloor 2f \rfloor = \lfloor 1.712 \rfloor = 1.\end{aligned}$$

Alltså är $0.226 = (0.001110011\dots)_{\text{två}}$ och därmed $0.226 \approx 0.0011101_{\text{två}}$. Som flyttal representas b av $\text{fl}(b) = 1.1101 \cdot 2^{-3}$. (b) Vi har att

$$\text{fl}(a) = \frac{27}{2} = 13.5 \quad \text{och} \quad \text{fl}(b) = \frac{29}{128} = 0.2265625.$$

Därmed är

$$\delta a = 13.5 - 13.4 = 0.1$$

och

$$\delta b = 0.2265625 - 0.226 = 0.0005625.$$

(c) Den exakta summan är $13.4 + 0.226 = 13.626$ och motsvarande summa i flyttalssystemet är

$$\begin{aligned}\text{fl}(a) + \text{fl}(b) &= (1.1011 \cdot 2^3) + (1.1101 \cdot 2^{-3}) \\ &= (1101100 \cdot 2^{-3}) + (1.1101 \cdot 2^{-3}) \\ &= 1101100 \cdot 2^{-4} = 1.1011 \cdot 2^3,\end{aligned}$$

dvs $\text{fl}(a) + \text{fl}(b) = \text{fl}(a)$. Det absoluta felet är således

$$\delta(a + b) = a + b - \text{fl}(a) = 13.626 - 13.5 = 0.126.$$

Precisionen är alltför dålig för att \bar{b} ska någon påverkn vid addition med $\text{fl}(a)$.

Uppgift 10. Vi har att $\hat{x} = 21/4 = 5.25$, dvs entalssiffran 5 och tiondelssiffran 2 är de signifika siffrorna i \hat{x} . Med andra ord söker vi de x för vilka $|\delta x| \leq 0.5 \cdot 10^{-1}$, dvs

$$-0.5 \cdot 10^{-1} \leq \hat{x} - x \leq 0.5 \cdot 10^{-1} \Leftrightarrow 5.2 \leq x \leq 5.3.$$

Uppgift 11. Från

$$x = d_1 \cdot b^{-1} + d_2 \cdot b^{-2} + \cdots + d_n \cdot b^{-n} = \frac{1}{b^n}(d_1 \cdot b^{n-1} + d_2 \cdot b^{n-2} + \cdots + d_n)$$

följer att

$$x = \frac{\alpha}{\beta} = \frac{d_1 \cdot b^{n-1} + d_2 \cdot b^{n-2} + \cdots + d_n}{b^n},$$

där täljaren α och nämnaren β är heltal.

Uppgift 12. Tag tex $\sqrt{2} = 1.414213562373095048801688724\dots$, som är irrationellt. Enligt föregående övningsuppgift vet vi att reella tal med ändlig bråktalsdel är rationella. Bilda en talföljd av rationella tal på formen $x_n = (1.d_1d_2\dots d_n)$, där d_i är den i :te decimalsiffran i $\sqrt{2}$. De fem första talen i följen ges av

$$\begin{aligned}x_0 &= 1 \\ x_1 &= 1.4 \\ x_2 &= 1.41 \\ x_3 &= 1.414 \\ x_4 &= 1.4142 \\ x_5 &= 1.41421.\end{aligned}$$

Alltså har vi att $x_n \rightarrow \sqrt{2}$ då $n \rightarrow \infty$. Denna process kan vi göra för alla irrationella tal. Eftersom tex derivata och integral är definierade som gränsvärden, så kan inget flyttalssystem ge oss tillräckligt hög noggrannhet för att "simulera" metoder i analys.

Uppgift 13. Det absoluta felet är

$$\begin{aligned}\delta x &= \hat{x} - x = 2.7182 - 2.71828182 = -0.00008182 \\ \delta y &= \hat{y} - y = 98.000 - 98.350 = -0.350 \\ \delta z &= \hat{z} - z = 0.00006 - 0.000068 = -0.000008.\end{aligned}$$

Det relativata felet är

$$\begin{aligned}\frac{\hat{x} - x}{x} &= \frac{2.7182 - 2.71828182}{2.71828182} = -0.0000300999 \\ \frac{\hat{y} - y}{y} &= \frac{98.000 - 98.350}{98.350} = -0.00355872 \\ \frac{\hat{z} - z}{z} &= \frac{0.00006 - 0.000068}{0.000068} = -0.117647.\end{aligned}$$

Från

$$|\delta x| \leq 0.000082 \leq 0.5 \cdot 10^{-3} \quad \text{och} \quad |\delta y| \leq 0.35 \leq 0.5 \cdot 10^{(-1)}$$

samt

$$|\delta z| \leq 0.000008 \leq 0.5 \cdot 10^{-4}$$

följer det att antal signifikanta siffror är 5, 1 och 0 i respektive deluppgift.

2 Icke-linjära ekvationer

Uppgift 1. Ekvationen

$$g(x) = x \Leftrightarrow x^2 + x - 4 = x \Leftrightarrow x^2 = 4$$

har lösningarna -2 och 2 . Eftersom $g'(x) = 2x + 1$ så är

$$|g'(-2)| = |2 \cdot (-2) + 1| = 3 > 1 \quad \text{och} \quad |g'(2)| = |2 \cdot 2 + 1| = 5 > 1.$$

Från olikheterna kan vi dra slutsatsen fixpunktmetoden inte kommer att fungera, utan genererar istället en talföljd som divergerar från lösningarna (se sats 2.5).

Uppgift 2. Vi visar (a) och (b) samtidigt med hjälp av induktion över n . Från

$$p_1 = g(p_0) = 1 - 0.0001 \cdot 1^2 = 0.9999$$

följer att $p_0 > p_1$ samt $p_1 > 0$. Vi vet redan att $p_0 > 0$. Antag att

$$p_n < p_{n-1} < \dots < p_1 < p_0 = 1 \quad \text{och} \quad p_k > 0$$

för alla $k = 0, 1, \dots, n$. Då följer det att

$$p_n - p_{n+1} = p_n - g(p_n) = p_n - (p_n - 0.0001p_n^2) = 0.0001p_n^2 > 0,$$

dvs $p_n - p_{n+1} > 0$ eller ekvivalent $p_n > p_{n+1}$. Notera att strikt olikhet följer från antagandet att $p_n > 0$. Det räcker egentligen med att veta att $p_n \neq 0$. För att slutföra induktionsbeviset måste vi visa att $p_{n+1} > 0$. Enligt antagandet är $p_n \leq 1$. Det ger oss att $0.0001p_n \leq 0.0001$ och

$$p_{n+1} = g(p_n) = p_n(1 - 0.0001p_n) \geq p_n(1 - 0.0001) = 0.9999p_n > 0.$$

Därmed följer (a) och (b) enligt induktionsprincipen. (c) Följden $(p_n)_{n=0}^\infty$ har ett gränsvärde a och enligt sats 2.1 är a en fixpunkt till g . För att bestämma eventuella fixpunkter till g löser vi ekvationen $g(x) = x$, dvs

$$x - 0.0001x^2 = x \Leftrightarrow x^2 = 0 \Leftrightarrow x = 0.$$

Alltså måste $a = 0$, dvs $p_n \rightarrow 0$ då $n \rightarrow \infty$.

Alternativ lösning. (a) Antag motsatsen, dvs att det finns ett positivt heltalet n sådant att $p_{n+1} \geq p_n$. Det ger att

$$p_{n+1} \geq p_n \Leftrightarrow p_n - 0.0001p_n^2 \geq p_n \Leftrightarrow -0.0001p_n^2 \geq 0,$$

som endast är uppfyllt då $p_n = 0$. Det ger i sin tur att

$$0 = p_n = g(p_{n-1}) \Leftrightarrow 0 = p_{n-1}(1 - 0.0001p_{n-1})$$

dvs $p_{n-1} = 0$ eller $p_{n-1} = 10\,000$. Men $g(x) \leq 2500$, vilket lämnas som övning att visa. Alltså måste $p_{n-1} = 0$. Det visar att om ett tal i följen är 0, så föregående tal också det. Men det skulle betyda att även $p_0 = 0$, vilket motsäger att $p_0 = 1$. Vårt antagandet inledningsvis att $p_{n+1} \geq p_n$ leder således till en motsägelse och därmed är $p_{n+1} < p_n$. (b) Antag att det finns positiva heltalet k sådana att $p_k \leq 0$. Låt n vara det minsta sådant heltalet, dvs $p_n \leq 0$ och $p_i > 0$ för alla $i = 0, 1, \dots, n-1$. Det ger att

$$\begin{aligned} p_n \leq 0 &\Leftrightarrow g(p_{n-1}) \leq 0 \Leftrightarrow p_{n-1}(1 - 0.0001p_{n-1}) \leq 0 \\ &\Leftrightarrow 1 - 0.0001p_{n-1} \leq 0 \Leftrightarrow p_{n-1} \geq 10\,000, \end{aligned}$$

vilket återigen motsäger $g(x) \leq 2500$. Alltså finns det inget heltalet n för vilket $p_n \leq 0$.

Uppgift 3. (a) Vi får att $g(p) = g(3) = 0.5 \cdot 3 + 1.5 = 3 = p$, vilket skulle visas. (b) Vi har att

$$\begin{aligned} 2|p - p_n| &= |2p - 2p_n| = |2p - 2g(p_{n-1})| = |p - 2(0.5p_{n-1} + 1.5)| \\ &= |2p - p_{n-1} - 3| = |p - p_{n-1} + p - 3| = |p - p_{n-1}|, \end{aligned}$$

eftersom $p - 3 = 3 - 3 = 0$. Alltså har vi funnit att

$$2|p - p_n| = |p - p_{n-1}| \Leftrightarrow |p - p_n| = \frac{|p - p_{n-1}|}{2},$$

vilket skulle visas. (c) Vi visar påståendet med induktion över n och med hjälp av resultatet i föregående deluppgift. Från (b) vet vi att

$$|p - p_1| = \frac{|p - p_0|}{2}.$$

Antag att $|p - p_{n-1}| = |p - p_0|/2^{n-1}$. Då får vi att

$$|p - p_n| = \frac{|p - p_{n-1}|}{2} = \frac{|p - p_0|/2^{n-1}}{2} = \frac{|p - p_0|}{2^n}.$$

Induktionsaxiomet ger nu att påståendet är sant för alla naturliga tal n .

Uppgift 5. (a) Funktionen för fixpunktmetoden ges av

$$g(x) = f(x) + x = \frac{2}{3}x^2 - x - \frac{2}{3}.$$

(b) Vi får att

$$x_1 = g(x_0) = g(1) = \frac{2}{3} \cdot 1^2 - 1 - \frac{2}{3} = -1$$

och

$$x_2 = g(x_1) = g(-1) = \frac{2}{3} \cdot (-1)^2 - (-1) - \frac{2}{3} = 1.$$

Därmed följer att $x_3 = g(x_2) = g(1) = -1$ och $x_4 = g(x_3) = g(-1) = 1$. (c) Fixpunktterna till g är lösningarna till ekvationen $g(x) = x$, dvs

$$\begin{aligned} \frac{2}{3}x^2 - x - \frac{2}{3} &= x \\ \Leftrightarrow \\ x^2 - 3x - 1 &= 0 \\ \Leftrightarrow \\ x = \frac{1}{2}(3 - \sqrt{13}) \quad \text{eller} \quad x &= \frac{1}{2}(3 + \sqrt{13}). \end{aligned}$$

(d) Lösningarna till $f(x) = 0$ är fixpunktterna till g och vice versa. Vi har att

$$g'(x) = \frac{4}{3}x - 1$$

Låt α_1 och α_2 beteckna fixpunktterna vi fann i föregående deluppgift. Då är

$$|g'(\alpha_1)| = \left| \frac{2\sqrt{13} - 3}{3} \right| \approx 1.4037 > 1$$

och

$$|g'(\alpha_2)| = \left| \frac{2\sqrt{13} + 3}{3} \right| \approx 3.4037 > 1,$$

dvs oavsett val av initialvärde så kommer den genererade talföljden inte att konvergerar mot någon av lösningarna till $f(x) = 0$.

Uppgift 7. Från definitionen av g följer att

$$\begin{aligned} x_1 &= g(x_n) = f(x_n) + x_n = f(x_n) + g(x_{n-1}) \\ &= f(x_n) + f(x_{n-1}) + x_{n-1} = \dots \\ &= f(x_n) + f(x_{n-1}) + \dots + f(x_2) + f(x_1) + x_1. \end{aligned}$$

Uppgift 8. (a) Vi misslyckas eftersom $f(3) > 0$ och $f(7) > 0$, dvs de har samma tecken och därför avbryter vi processen redan i första steget. (b) Eftersom $f(x) < 0$ för alla $x \in [1, 2)$ och $f(x) > 0$ för alla $x \in (2, 7]$, så kommer metoden att genererar intervall på formen $[a_n, b_n]$, där a_n och b_n konvergerar mot 2 från vänster respektive höger. Det kommer att ge ett intryck av att f har ett nollställe i $x = 2$. Men funktionen är inte definierad i $x = 2$.

Uppgift 9. Noggrannheten i resultatet mäter vi genom att beräkna halva längden på det interval som returneras (vi approximerar lösningen r med mittpunkten c_n och största felet inträffar då om lösningen råkar ligga i en av ändpunktterna). I varje steg halveras intervallets längd. Efter n steg har vi därför noggrannheten

$$|r - c_n| \leq \frac{b - a}{2^{n+1}} = \frac{7 - 2}{2^{n+1}} = \frac{5}{2^{n+1}},$$

se sats 2.4 på sidan 54 i kursboken. Alltså söker vi det minsta heltalet n sådant att

$$\begin{aligned} \frac{5}{2^{n+1}} < \frac{5}{10^9} &\Leftrightarrow \frac{1}{2^{n+1}} < \frac{1}{10^9} \Leftrightarrow \frac{1}{2^n \cdot 2} < \frac{1}{2^9 \cdot 5^9} \Leftrightarrow 2^{8-n} < 5^{-9} \\ &\Leftrightarrow (8-n) \ln 2 < -9 \ln 5 \Leftrightarrow n > 8 + 9 \frac{\ln 5}{\ln 2} \approx 28.8974. \end{aligned}$$

Det ger att $n = 29$ är det heltalet vi söker. Alltså ska vi upprepa intervallhalveringsmetoden 29 gånger för att uppnå den önskade noggrannheten.

Uppgift 11. (b) Eftersom första mittpunkten är $c = 1.5$ samt där

$$f(-4)f(1.5) > 0 \quad \text{och} \quad f(1.5)f(7) < 0$$

kommer man med algoritmen fortsätta med intervallet $[1.5, 7]$ och däri finns endast nollstället 4.

(c) Eftersom intervallet innehåller ett jämnt antal nollställen till f så har funktionsvärdena $f(-4)$ och $f(3)$ samma tecken och det ge intycket av att det inte finns något nollställe till f i intervallet. Det finns faktiskt två – en i varje delintervall efter halvering.

Uppgift 14. (a) Först deriverar vi f , dvs $f'(x) = 4(x-2)^3$. Sätt

$$g(x) = x - \frac{f(x)}{f'(x)} = x - \frac{(x-2)^4}{4(x-2)^3} = x - \frac{1}{4}(x-2) = \frac{1}{4}(3x+2).$$

Alltså är

$$x_{k+1} = g(x_k) = \frac{1}{4}(3x_k + 2).$$

(b) Med $x_0 = 2.1$ får vi att

$$\begin{aligned} x_1 &= g(x_0) = g(2.1) = 2.075 \\ x_2 &= g(x_1) = g(2.075) = 2.05625 \\ x_3 &= g(x_2) = g(2.05625) = 2.04219 \\ x_4 &= g(x_3) = g(2.04219) = 2.03164. \end{aligned}$$

(c) Från definitionen av f följer att $f(2) = 0$, dvs $x = 2$ är ett nollställe till f . Men 2 är en multipelt nollställe och då är konvergensen linjär. Det följer av att

$$|x_{k+1} - x_k| = \left| \frac{1}{4}(3x_k + 2) - \frac{1}{4}(3x_{k-1} + 2) \right| = \frac{3}{4} |x_k - x_{k-1}|$$

eller ekvivalent

$$\frac{|x_{k+1} - x_k|}{|x_k - x_{k-1}|} = \frac{|\delta x_{k+1}|}{|\delta x_k|} = \frac{3}{4},$$

dvs skillnaden mellan två på varandra efterföljande tal i följen minskar med en faktor av $3/4$, jämfört med paret innan.

Uppgift 15. (a) Vi deriverar f och får att $f'(x) = -\sin(x)$. Sätt

$$g(x) = x - \frac{f(x)}{f'(x)} = x + \frac{\cos(x)}{\sin(x)} = x + \frac{1}{\tan(x)} = x + \cot(x).$$

Då är

$$x_{k+1} = g(x_k) = x + \frac{1}{\tan(x_k)} = x_k + \cot(x_k).$$

(b) Nej, vilket vi ser genom att generera talföljden (x_0, x_1, x_2, \dots) som visar sig konvergerar mot $-3\pi/2 \approx -4.71239$. (c) Ja, ty denna gång konvergerar motsvarande talföld mot $3\pi/2 \approx 4.71239$.

Uppgift 16. Sätt $f(x) = e^{-x} - x$. Då är

$$e^{-x} = x \Leftrightarrow f(x) = 0.$$

Newton's formel ges av $x_{n+1} = g(x_n)$ där

$$g(x) = x - \frac{f(x)}{f'(x)} = x - \frac{e^{-x} - x}{-e^{-x} - 1} = \frac{e^{-x}(x+1)}{e^{-x} + 1} = \frac{x+1}{e^x + 1}.$$

Alltså är

$$x_{n+1} = \frac{x_n + 1}{e^{x_n} + 1}.$$

Med $x_0 = 0.25$ får vi att

$$\begin{aligned} x_1 &= \frac{0.25 + 1}{e^{0.25} + 1} \approx 0.547279 \\ x_2 &= \frac{0.547279 + 1}{e^{0.547279} + 1} \approx 0.567071 \\ x_3 &= \frac{0.567071 + 1}{e^{0.567071} + 1} \approx 0.567143 \\ x_4 &= \frac{0.567143 + 1}{e^{0.567143} + 1} \approx 0.567143. \end{aligned}$$

Uppgift 18. Sätt $f(x) = xe^x - 1$. Då är

$$ex^x = 1 \Leftrightarrow f(x) = 0.$$

Från $f'(x) = (1+x)e^x$ följer att

$$g(x) = x - \frac{f(x)}{f'(x)} = x - \frac{xe^x - 1}{(1+x)e^x} = \frac{x^2 + e^{-x}}{1+x}.$$

Med $x_0 = 1$ får vi att

$$\begin{aligned} x_1 &= g(x_0) \approx 0.68394 \\ x_2 &= g(x_1) \approx 0.577454 \\ x_3 &= g(x_2) \approx 0.56723 \\ x_4 &= g(x_3) \approx 0.567143 \\ x_5 &= g(x_4) \approx 0.567143. \end{aligned}$$

Uppgift 23. (a) Eftersom absolutbeloppet av ett reellt tal är icke-negativt, så är summan

$$|x_1| + |x_2| + \cdots + |x_n|$$

alltid icke-negativ, dvs $\|\mathbf{x}\|_1 \geq 0$.

(b) Först noterar vi att om $x_k \neq 0$ för något $k = 1, 2, \dots, n$, så är $|x_k| > 0$ och därmed följer det att $\|\mathbf{x}\|_1 \neq 0$. Antag att $\|\mathbf{x}\|_1 = 0$. Då är samtliga koordinater x_k lika med 0, dvs $\mathbf{x} = \mathbf{0}$. Antag att $\mathbf{x} \neq \mathbf{0}$. Då är $\|\mathbf{x}\|_1 = \|\mathbf{0}\|_1 = |0| + |0| + \cdots + |0| = 0$.

(c) Eftersom absolutbelopp för reella tal uppfyller $|ab| = |a| \cdot |b|$ följer det att

$$\begin{aligned} \|\mathbf{cx}\|_1 &= \|(cx_1, cx_2, \dots, cx_n)\|_1 = |cx_1| + |cx_2| + \cdots + |cx_n| \\ &= |c| \cdot |x_1| + |c| \cdot |x_2| + \cdots + |c| \cdot |x_n| \\ &= |c| (|x_1| + |x_2| + \cdots + |x_n|) = |c| \cdot \|\mathbf{x}\|_1. \end{aligned}$$

Uppgift 27. Först noterar vi att

$$J(x, y) = \begin{pmatrix} \frac{\partial f_1}{\partial x} & \frac{\partial f_1}{\partial y} \\ \frac{\partial f_2}{\partial x} & \frac{\partial f_2}{\partial y} \end{pmatrix}$$

och därmed är

$$J(x, y)^{-1} = \frac{1}{\det(J(x, y))} \begin{pmatrix} \frac{\partial f_2}{\partial y} & -\frac{\partial f_1}{\partial y} \\ -\frac{\partial f_2}{\partial x} & \frac{\partial f_1}{\partial x} \end{pmatrix}$$

Sätt $\mathbf{f}(\mathbf{x}) = (f_1(x, y), f_2(x, y))$. Newtons metod för ekvationssystem, dvs

$$\mathbf{x}_k = \mathbf{x}_{k-1} - J(\mathbf{x}_{k-1})^{-1} \mathbf{f}(\mathbf{x}_{k-1})$$

motsvarar ett fixpunktmetod på formen

$$\mathbf{x}_k = \mathbf{g}(\mathbf{x}_{k-1}).$$

Alltså är

$$\begin{aligned} \mathbf{g}(\mathbf{x}) &= \mathbf{x} - J(\mathbf{x})^{-1} \mathbf{f}(\mathbf{x}) \\ &= \begin{pmatrix} x \\ y \end{pmatrix} - \frac{1}{\det(J(x, y))} \begin{pmatrix} \frac{\partial f_2}{\partial y}(x, y) & -\frac{\partial f_1}{\partial y}(x, y) \\ -\frac{\partial f_2}{\partial x}(x, y) & \frac{\partial f_1}{\partial x}(x, y) \end{pmatrix} \begin{pmatrix} f_1(x, y) \\ f_2(x, y) \end{pmatrix} \\ &= \begin{pmatrix} x \\ y \end{pmatrix} - \frac{1}{\det(J(x, y))} \begin{pmatrix} f_1(x, y) \frac{\partial f_2}{\partial y}(x, y) - f_2(x, y) \frac{\partial f_1}{\partial y}(x, y) \\ f_2(x, y) \frac{\partial f_1}{\partial x}(x, y) - f_1(x, y) \frac{\partial f_2}{\partial x}(x, y) \end{pmatrix}. \end{aligned}$$

Sätter vi $\mathbf{g}(\mathbf{x}) = (g_1(x, y), g_2(x, y))$ så följer formlerna i uppgiften.

Uppgift 30. (a) Vi skriver om ekvationssystemet till formen

$$\begin{cases} x_k = 3y_{k-1} - 1 \\ y_k = 3x_{k-1} - 1. \end{cases}$$

Alltså är

$$\mathbf{p}_k = g(\mathbf{p}_{k-1}) = (3y_{k-1} - 1, 3x_{k-1} - 1),$$

där $\mathbf{p}_k = (x_k, y_k)$. Det ger att

$$\begin{aligned} \mathbf{p}_1 &= (3 \cdot 0 - 1, 3 \cdot 0 - 1) = (-1, -1), \\ \mathbf{p}_2 &= (3 \cdot (-1) - 1, 3 \cdot (-1) - 1) = (-4, -4), \\ \mathbf{p}_3 &= (3 \cdot (-4) - 1, 3 \cdot (-4) - 1) = (-13, -13). \end{aligned}$$

Vi ser att koordinaterna i \mathbf{p}_k ökar med en faktor 3 för varje iteration och därför divergerar metoden för detta ekvationssystem. (b) Vi modifierar föregående ekvationssystem till

$$\begin{cases} x_k = 3y_{k-1} - 1 \\ y_k = -3x_k - 1. \end{cases}$$

Då är

$$\begin{array}{ll} x_1 = 3 \cdot 0 - 1 = -1 & y_1 = 3 \cdot (-1) - 1 = -4 \\ x_2 = 3 \cdot (-4) - 1 = -13 & y_2 = 3 \cdot (-13) - 1 = -40 \\ x_3 = 3 \cdot (-40) - 1 = -121 & y_3 = 3 \cdot (-121) - 1 = -364. \end{array}$$

Alltså är

$$\mathbf{p}_1 = (-1, -4), \quad \mathbf{p}_2 = (-13, -40) \quad \text{och} \quad \mathbf{p}_3 = (-121, -364).$$

Även här ser vi att metoden divergerar.

Uppgift 32. Vi får att

k	\mathbf{x}_k	$\ \mathbf{x}_{k-1} - \mathbf{x}_k\ _2$
0	(1.000000, 1.000000)	
1	(0.909297, 0.639175)	0.372051
2	(0.999751, 0.792537)	0.17805
3	(0.975571, 0.752877)	0.0464492
4	(0.987599, 0.768924)	0.0200537
5	(0.982802, 0.762674)	0.00787821
6	(0.984782, 0.765198)	0.00320835
7	(0.983989, 0.764187)	0.00128504
8	(0.984309, 0.764594)	0.000517793
9	(0.984181, 0.764431)	0.000208091
10	(0.984233, 0.764496)	0.0000837113

Jämför med tabell 2.3.

Uppgift 33. (a) Ekvationen $x = g(x)$ är ekvivalent med

$$\begin{cases} x = x - y^2 \\ y = -x + 6y \end{cases} \Leftrightarrow \begin{cases} y^2 = 0 \\ x - 5y = 0 \end{cases} \Leftrightarrow \begin{cases} x = 0 \\ y = 0. \end{cases}$$

(b) Vi får ekvationssystemet

$$\begin{cases} x = (x^2 - y^2 - x - 3)/3 \\ y = (-x + y - 1)/3 \end{cases} \Leftrightarrow \begin{cases} x^2 - y^2 - 4x - 3 = 0 \\ x + 2y + 1 = 0. \end{cases}$$

Från andra ekvationen följer att

$$x = -2y - 1. \quad (6.5)$$

Insättning i första ekvationen ger att

$$(-2y - 1)^2 - y^2 - 4(-2y - 1) - 3 = 0 \Leftrightarrow 3y^2 + 12y + 2 = 0.$$

Denna andragradsekvation har lösningarna

$$y = -2 \pm \sqrt{2^2 - \frac{2}{3}} = -2 \pm \sqrt{\frac{10}{3}} = \frac{1}{3}(-6 \pm \sqrt{30}).$$

Motsvarande x fås genom insättning i (6.5), dvs

$$x = -\frac{2}{3}(-6 \pm \sqrt{30}) - 1 = \frac{1}{3}(9 \mp 2\sqrt{30}).$$

Vi har funnit två fixpunkter, nämligen

$$\frac{1}{3}(9 - 2\sqrt{30}, -6 + \sqrt{30}) \text{ och } \frac{1}{3}(9 + 2\sqrt{30}, -6 - \sqrt{30}).$$

(c) Vi får att

$$x = g(x) \Leftrightarrow \begin{cases} x = \sin(y) \\ y = -6x + y \end{cases} \Leftrightarrow \begin{cases} x = \sin(y) \\ 6x = 0. \end{cases}$$

Alltså är $x = 0$ och därmed $\sin(y) = 0$. Det betyder att $y = \pi n$, där n är ett heltal. Vi har funnit oändligt många fixpunkter, nämligen $(0, \pi n)$, där $n \in \mathbb{Z}$. (d) Vi har att ekvationen $x = g(x)$ är ekvivalent med

$$\begin{cases} x = 9 - 3y - 2z \\ y = 2 - x + z \\ z = -9 + 3x + 4y - z \end{cases} \Leftrightarrow \begin{cases} x + 3y + 2z = 9 \\ x + y - z = 2 \\ 3x + 4y - 2z = 9. \end{cases}$$

Gausselimination steg för steg:

$$\begin{array}{l} \left\{ \begin{array}{l} x + 3y + 2z = 9 \\ x + y - z = 2 \\ 3x + 4y - 2z = 9 \end{array} \right. \xrightarrow{-1} \left\{ \begin{array}{l} x + 3y + 2z = 9 \\ y - z = -7 \\ 3x + 4y - 2z = 9 \end{array} \right. \xrightarrow{-3} \left\{ \begin{array}{l} x + 3y + 2z = 9 \\ -2y - 3z = -7 \\ 3x + 4y - 2z = 9 \end{array} \right. \xrightarrow{-2} \left\{ \begin{array}{l} x + 3y + 2z = 9 \\ -2y - 3z = -7 \\ 10y + 16z = 36 \end{array} \right. \\ \left. \begin{array}{l} \\ \\ \end{array} \right\} \xrightarrow{5} \left\{ \begin{array}{l} x + 3y + 2z = 9 \\ -2y - 3z = -7 \\ z = 1. \end{array} \right. \end{array}$$

Bakåtsubstitution ger lösningen $(1, 2, 1)$ till det linjära ekvationssystemet.

Uppgift 34. (a) Fixpunktmetoden ges av $x_k = g(x_{k-1})$, dvs

$$\begin{cases} x_k = g_1(x_{k-1}, y_{k-1}) = \frac{y_{k-1} - x_{k-1}^3 + 3x_{k-1}^2 + 3x_{k-1}}{7} \\ y_k = g_2(x_{k-1}, y_{k-1}) = \frac{y_{k-1}^2 + 2y_{k-1} - x_{k-1} - 2}{2}. \end{cases}$$

Tre iterationer ger följande resultat.

$$\begin{aligned} & \begin{cases} x_1 = g_1(x_0, y_0) = g_1(-0.3, -1.3) = -0.271857 \\ y_1 = g_2(x_0, y_0) = g_2(-0.3, -1.3) = -1.305 \end{cases} \\ & \begin{cases} x_2 = g_1(x_1, y_1) = g_1(-0.271857, -1.305) = -0.268394 \\ y_2 = g_2(x_1, y_1) = g_2(-0.271857, -1.305) = -1.31756 \end{cases} \\ & \begin{cases} x_3 = g_1(x_2, y_2) = g_1(-0.268394, -1.31756) = -0.269614 \\ y_3 = g_2(x_2, y_2) = g_2(-0.268394, -1.31756) = -1.31538. \end{cases} \end{aligned}$$

(b) För Seidels metod får vi systemet

$$\begin{cases} x_k = g_1(x_{k-1}, y_{k-1}) = \frac{y_{k-1} - x_{k-1}^3 + 3x_{k-1}^2 + 3x_{k-1}}{7} \\ y_k = g_2(x_k, y_{k-1}) = \frac{y_{k-1}^2 + 2y_{k-1} - x_k - 2}{2} \end{cases}$$

Tre iterationer ger följande resultat.

$$\begin{aligned} & \begin{cases} x_1 = g_1(x_0, y_0) = g_1(-0.3, -1.3) = -0.271857 \\ y_1 = g_2(x_1, y_0) = g_2(-0.271857, -1.3) = -1.31907 \end{cases} \\ & \begin{cases} x_2 = g_1(x_1, y_1) = g_1(-0.271857, -1.31907) = -0.270405 \\ y_2 = g_2(x_2, y_1) = g_2(-0.270405, -0.271857) = -1.31389 \end{cases} \\ & \begin{cases} x_3 = g_1(x_2, y_2) = g_1(-0.270405, -1.31389) = -0.269426 \\ y_3 = g_2(x_3, y_2) = g_2(-0.269426, -1.31389) = -1.31602. \end{cases} \end{aligned}$$

Uppgift 35. Man kan skriva om ekvationssystemet på flera sätt, som tex

$$\begin{cases} x_n = \frac{1}{3y_{n-1}} \\ y_n = \sin(x_n y_{n-1}) \end{cases}$$

Det ger resultatet

$$\begin{aligned} x_1 &= (0.342898, 0.97211) & x_2 &= (0.327195, 1.01876) \\ x_3 &= (0.327195, 1.01876) & x_4 &= (0.327195, 1.01876). \end{aligned}$$

3 Interpolation

Uppgift 1. Funktionen $f(x) = (1+x)^{-1}$ har derivatorna

$$\begin{aligned} f'(x) &= -(1+x)^{-2} & f'(0) &= -1 \\ f''(x) &= 2(1+x)^{-3} = 2!(1+x)^{-3} & f''(0) &= 2! \end{aligned}$$

$$\begin{aligned}
 f^{(3)}(x) &= -6(1+x)^{-4} = -3!(1+x)^{-4} & f^{(3)}(0) &= -3! \\
 f^{(4)}(x) &= 24(1+x)^{-5} = 4!(1+x)^{-5} & f^{(4)}(0) &= 4! \\
 f^{(5)}(x) &= -120(1+x)^{-6} = -5!(1+x)^{-6} & f^{(5)}(0) &= -5! \\
 f^{(6)}(x) &= 720(1+x)^{-7} = 6!(1+x)^{-7},
 \end{aligned}$$

Vi noterar också att $f(0) = 1$. Alltså är

$$\begin{aligned}
 p_5(x) &= 1 - 1(x-0) + \frac{2!}{2!}(x-0)^2 + \frac{-3!}{3!}(x-0)^3 + \frac{4!}{4!}(x-0)^4 + \frac{-5!}{5!}(x-0)^5 \\
 &= 1 - x + x^2 - x^3 + x^4 - x^5.
 \end{aligned}$$

Lagranges restterm är

$$E_5(x) = \frac{f^{(6)}(\xi)}{6!}(x-0)^6 = (1+\xi)^{-7}x^6 = \frac{x^6}{(1+\xi)^7},$$

där ξ ligger mellan 0 och x .

Uppgift 2. Låt k vara ett positivt heltal. Då är $n = 2k - 1$ och $m = n + 1 = 2k$ ett udda heltal respektive ett jämnt heltal. Vi ska med induktion över k visa att

$$f^{(n)}(x) = (-1)^k(x \sin x - n \cos x) \quad \text{och} \quad f^{(m)}(x) = (-1)^k(x \cos x + m \sin x). \quad (6.6)$$

Låt $k = 1$. Då är $n = 1$ och $m = 2$. Eftersom $f(x) = x \cos x$, så är

$$f^{(1)}(x) = f'(x) = -x \sin x + \cos x = (-1)^k(x \sin x + n \cos x)$$

och

$$f^{(2)}(x) = f''(x) = -x \cos -2 \sin x = (-1)^k(x \cos x + m \sin x).$$

Det visar (6.6) då $k = 1$. Antag att (6.6) är sann för $n = 2k - 1$ och $m = 2k$, där $k \geq 1$. Då är $2(k+1) - 1 = 2k + 1 = n + 2$ och $2(k+1) = 2k + 2 = m + 2$. Vi sak alltså visa att likheterna i (6.6) är uppfyllda för $(n+2)$:te respektive $(m+2)$:te derivatan. Vi får att

$$\begin{aligned}
 f^{(n+2)}(x) &= \frac{d}{dx} f^{(n+1)}(x) = \frac{d}{dx} f^{(m)}(x) = \frac{d}{dx} (-1)^k(x \cos x + m \sin x) \\
 &= (-1)^k(\cos x - x \sin x + m \cos x) = (-1)^{k+1}(x \sin x - (m+1) \cos x) \\
 &= (-1)^{k+1}(x \sin x - (n+2) \cos x),
 \end{aligned}$$

eftersom $m+1 = n+1+1 = n+2$. Det ger i sin tur att

$$\begin{aligned}
 f^{(m+2)}(x) &= \frac{d}{dx} f^{(m+1)}(x) = \frac{d}{dx} f^{(n+2)}(x) = \frac{d}{dx} (-1)^{k+1}(x \sin x - (n+2) \cos x) \\
 &= (-1)^{k+1}(\sin x + x \cos x + (n+2) \sin x) \\
 &= (-1)^{k+1}(x \cos x + (n+3) \sin x) \\
 &= (-1)^{k+1}(x \cos x + (m+2) \sin x),
 \end{aligned}$$

eftersom $n+3 = n+2+1 = m+1+1 = m+2$. Därmed följer (6.6) enligt induktionsprincipen. Notera att

$$f^{(n)}(0) = (-1)^{k+1}n \quad \text{och} \quad f^{(m)}(0) = 0$$

för alla udda heltal $n = 2k - 1$ och alla jämnna heltal m .

Uppgift 3. För att kunna använda Horners metod skriver vi om polynomet enligt

$$p(x) = ((-0.02x + 0.1)x - 0.2)x + 1.66.$$

(a) Sätt $c_0 = 1.66$, $c_1 = -0.2$, $c_2 = 0.1$ och $c_3 = -0.02$. Då är

$$p(x) = ((c_3x + c_2)x + c_1)x + c_0.$$

Med följande algoritm beräknar man $b = p(a)$.

```

 $b_n \leftarrow c_n$ 
for  $k = n - 1, \dots, 1, 0$  do
     $b_k \leftarrow b_{k+1}a + c_k$ 
end for
 $b \leftarrow b_0$ 
```

För det givna polynomet är $n = 3$. Då $a = 4$ får vi att

$$\begin{aligned} b_3 &= c_3 = -0.02 \\ b_2 &= b_3a + c_2 = -0.02 \cdot 4 + 0.1 = 0.02 \\ b_1 &= b_2a + c_1 = 0.02 \cdot 4 - 0.2 = -0.12 \\ b_0 &= b_1a + c_0 = -0.12 \cdot 4 + 1.66 = 1.18. \end{aligned}$$

Alltså är $p(4) = b_0 = 1.18$. (b) Derivatan av polynomet är

$$p'(x) = 3c_3x^2 + 2c_2x + c_1 = (3c_3x + 2c_2)x + c_1$$

Motsvarande algoritm för att beräkna $b = p'(a)$ ges av följande.

```

 $b_{n-1} \leftarrow nc_n$ 
for  $k = n - 1, \dots, 2, 1$  do
     $b_{k-1} \leftarrow b_k a + kc_k$ 
end for
 $b \leftarrow b_0$ 
```

För $a = 4$ får vi att

$$\begin{aligned} b_2 &= 3c_3 = 3 \cdot (-0.02) = -0.06 \\ b_1 &= b_2a + 2c_2 = -0.06 \cdot 4 + 2 \cdot 0.1 = -0.04 \\ b_0 &= b_1a + 1c_1 = -0.04 \cdot 4 + 1 \cdot (-0.2) = -0.36. \end{aligned}$$

Alltså är $p'(4) = b_0 = -0.36$. (c) Vi har att

$$\begin{aligned} p(x) &= \int p(x) dx = \frac{c_3}{4}x^4 + \frac{c_2}{3}x^3 + \frac{c_1}{2}x^2 + c_0x + c \\ &= \left(\left(\left(\frac{c_3}{4}x + \frac{c_2}{3} \right)x + \frac{c_1}{2} \right)x + c_0 \right)x + c, \end{aligned}$$

där c är en godtycklig konstant. Med följande algoritm beräknar vi $b = p(a)$.

```

 $b_{n+1} \leftarrow c_n/(n+1)$ 
for  $k = n, \dots, 2, 1$  do
     $b_k \leftarrow b_{k+1}a + c_{k-1}/k$ 
end for
 $b_0 \leftarrow b_1a + c$ 
 $b \leftarrow b_0$ 
```

Vi ska beräkna

$$\int_1^4 p(x) dx = [p(x)]_1^4 = p(4) - p(1).$$

Då $a = 1$ får vi att

$$\begin{aligned} b_4 &= c_3/4 = -0.02/4 = -0.005 \\ b_3 &= b_4a + c_2/3 = -0.005 \cdot 1 + 0.1/3 \approx 0.0283333 \\ b_2 &= b_3a + c_1/2 \approx 0.0283333 \cdot 1 - 0.2/2 \approx -0.0716667 \\ b_1 &= b_2a + c_0/1 \approx -0.0716667 \cdot 1 + 1.66 \approx 1.58833 \\ b_0 &= b_1a + c \approx 1.58833 \cdot 1 + c \approx 1.58833 + c. \end{aligned}$$

Alltså är $p(1) \approx 1.58833 + c$. Då $a = 4$ får vi

$$\begin{aligned} b_4 &= c_3/4 = -0.02/4 = -0.005 \\ b_3 &= b_4a + c_2/3 = -0.005 \cdot 4 + 0.1/3 \approx 0.0133333 \\ b_2 &= b_3a + c_1/2 \approx 0.0133333 \cdot 4 - 0.2/2 \approx -0.0466667 \\ b_1 &= b_2a + c_0/1 \approx -0.0466667 \cdot 4 + 1.66 \approx 1.47333 \\ b_0 &= b_1a + c \approx 1.47333 \cdot 4 + c \approx 5.89333 + c. \end{aligned}$$

Alltså är $p(4) \approx 5.89333 + c$. Det ger att

$$\int_1^4 p(x) dx = p(4) - p(1) \approx 5.89333 + c - 1.58833 - c \approx 4.305.$$

(d) Samma algoritm som i (a) fast med $a = 5.5$, dvs

$$\begin{aligned} b_3 &= c_3 = -0.02 \\ b_2 &= b_3a + c_2 = -0.02 \cdot 5.5 + 0.1 = -0.01 \\ b_1 &= b_2a + c_1 = -0.01 \cdot 5.5 - 0.2 = -0.255 \\ b_0 &= b_1a + c_0 = -0.255 \cdot 5.5 + 1.66 = 0.2575. \end{aligned}$$

Alltså är $p(5.5) = 0.2575$. (e) Polynomets koefficienter c_0, c_1, c_2 och c_3 är lösningarna till det linjära ekvationssystemet

$$\begin{cases} p(1) = 1.54 \\ p(2) = 1.5 \\ p(3) = 1.42 \\ p(5) = 0.66 \end{cases} \Leftrightarrow \begin{cases} c_0 + c_1 + c_2 + c_3 = 1.54 \\ c_0 + 2c_1 + 4c_2 + 8c_3 = 1.5 \\ c_0 + 3c_1 + 9c_2 + 27c_3 = 1.42 \\ c_0 + 5c_1 + 25c_2 + 125c_3 = 0.66. \end{cases}$$

Kontrollera gärna att så är faller, dvs lös ekvationssystemet eller sätt in de givna värdena i ekvationssystemet.

Uppgift 4. (a) Vi har att $y_0 = f(x_0) = -1$ och $y_1 = f(x_1) = 0$. Det ger att

$$\begin{aligned} p_1(x) &= y_0 L_{1,0}(x) + y_1 L_{1,1}(x) = y_0 \frac{x - x_1}{x_0 - x_1} + y_1 \frac{x - x_0}{x_1 - x_0} \\ &= -1 \frac{x - 0}{-1 - 0} + 0 \frac{x - (-1)}{0 - (-1)} = x. \end{aligned}$$

(b) Vi har att $y_0 = f(x_0) = -1$, $y_1 = f(x_1) = 0$ och $y_2 = f(x_2) = 1$. Det ger att

$$\begin{aligned} p_2(x) &= \sum_{k=0}^2 y_k L_{2,k}(x) \\ &= y_1 \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} + y_1 \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} + y_1 \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} \\ &= -1 \frac{(x-0)(x-1)}{(-1-0)(-1-1)} + 0 \frac{(x-(-1))(x-1)}{(0-(-1))(0-1)} + 1 \frac{(x-(-1))(x-0)}{(1-(-1))(1-0)} \\ &= -\frac{1}{2}x(x-1) + \frac{1}{2}x(x+1) = x. \end{aligned}$$

(c) Lagranges koefficientpolynom ges i detta fall av

$$\begin{aligned} L_{3,0}(x) &= \frac{(x-x_1)(x-x_2)(x-x_3)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)} = \frac{(x-0)(x-1)(x-2)}{(-1-0)(-1-1)(-1-2)} \\ &= -\frac{1}{6}x(x-1)(x-2) = -\frac{1}{6}(x^3 - 3x^2 + 2x), \\ L_{3,1}(x) &= \frac{(x-x_0)(x-x_2)(x-x_3)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)} = \frac{(x-(-1))(x-1)(x-2)}{(0-(-1))(0-1)(0-2)} \\ &= \frac{1}{2}(x+1)(x-1)(x-2) = \frac{1}{2}(x^3 - 2x^2 - x + 2), \\ L_{3,2}(x) &= \frac{(x-x_0)(x-x_1)(x-x_3)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)} = \frac{(x-(-1))(x-0)(x-2)}{(1-(-1))(1-0)(1-2)} \\ &= -\frac{1}{2}x(x+1)(x-2) = -\frac{1}{2}(x^3 - x^2 - 2x), \\ L_{3,3}(x) &= \frac{(x-x_0)(x-x_1)(x-x_2)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)} = \frac{(x-(-1))(x-0)(x-1)}{(2-(-1))(2-0)(2-1)} \\ &= \frac{1}{6}x(x+1)(x-1) = \frac{1}{6}(x^3 - x). \end{aligned}$$

Vidare är

$$y_0 = f(x_0) = -1, \quad y_1 = f(x_1) = 0, \quad y_2 = f(x_2) = 1 \quad \text{och} \quad y_3 = f(x_3) = 8.$$

Det ger att

$$\begin{aligned} p_3(x) &= \sum_{k=0}^3 y_k L_{3,k}(x) \\ &= (-1) \left(-\frac{1}{6}(x^3 - 3x^2 + 2x) \right) + 0 \cdot \frac{1}{2}(x^3 - 2x^2 - x + 2) \\ &\quad + 1 \cdot \left(-\frac{1}{2}(x^3 - x^2 - 2x) \right) + 8 \cdot \frac{1}{6}(x^3 - x) = x^3. \end{aligned}$$

Resultatet är det förväntade. Vi utgår från fyra punkter som vi vet ligger på tredjegradskurvan $y = x^3$ och just för fyra punkter existerar det ett entydigt bestämt polynom av grad 3. Alltså måste $p_3(x) = x^3$. (d) Den här gången har vi att $y_0 = f(x_0) = 1$ och $y_1 = f(x_1) = 8$. Alltså är

$$p_1(x) = y_0 L_{1,0}(x) + y_1 L_{1,1}(x) = y_0 \frac{x-x_1}{x_0-x_1} + y_1 \frac{x-x_0}{x_1-x_0}$$

$$= 1 \frac{x-2}{1-2} + 8 \frac{x-1}{2-1} = 7x - 6.$$

(e) Vi har att $y_0 = f(x_0) = 0$, $y_1 = f(x_1) = 1$ och $y_2 = f(x_2) = 8$. Det ger att

$$\begin{aligned} p_2(x) &= \sum_{k=0}^2 y_k L_{2,k}(x) \\ &= y_1 \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} + y_1 \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} \\ &\quad + y_1 \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} \\ &= 0 \frac{(x-1)(x-2)}{(0-1)(0-2)} + 1 \frac{(x-0)(x-2)}{(1-0)(1-2)} + 8 \frac{(x-0)(x-1)}{(2-0)(2-1)} \\ &= -x(x-2) + 4x(x-1) = 3x^2 - 2x. \end{aligned}$$

Uppgift 5. Först noterar vi att

$$f(1.2) \approx 2.86667 \quad \text{och} \quad f(1.5) \approx 2.83333.$$

(a) Vi har att $y_0 = f(x_0) = 3$, $y_1 = f(x_1) = 3$ och $y_2 = f(x_2) = 3.3$. Det ger att

$$\begin{aligned} p_2(x) &= \sum_{k=0}^2 y_k L_{2,k}(x) \\ &= y_0 \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} + y_1 \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} + y_2 \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} \\ &= 3 \frac{(x-2)(x-2.5)}{(1-2)(1-2.5)} + 3 \frac{(x-1)(x-2.5)}{(2-2)(1-2.5)} + 3.3 \frac{(x-1)(x-2)}{(2.5-1)(2.5-2)} \\ &= 0.4x^2 - 1.2x + 3.8. \end{aligned}$$

Alltså är $f(1.2) \approx p_2(1.2) = 2.936$ och $f(1.5) \approx p_2(1.5) = 2.9$. (b) Lagranges koefficient-polynom ges av

$$\begin{aligned} L_{3,0}(x) &= \frac{(x-x_1)(x-x_2)(x-x_3)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)} = \frac{(x-1)(x-2)(x-2.5)}{(0.5-1)(0.5-2)(0.5-2.5)} \\ &\approx -0.666667x^3 + 3.66667x^2 - 6.33333x + 3.33333, \\ L_{3,1}(x) &= \frac{(x-x_0)(x-x_2)(x-x_3)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)} = \frac{(x-0.5)(x-2)(x-2.5)}{(1-0.5)(1-2)(1-2.5)} \\ &\approx 1.33333x^3 - 6.66667x^2 + 9.66667x - 3.33333, \\ L_{3,2}(x) &= \frac{(x-x_0)(x-x_1)(x-x_3)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)} = \frac{(x-0.5)(x-1)(x-2.5)}{(2-0.5)(2-1)(2-2.5)} \\ &\approx -1.33333x^3 + 5.33333x^2 - 5.66667x + 1.66667, \\ L_{3,3}(x) &= \frac{(x-x_0)(x-x_1)(x-x_2)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)} = \frac{(x-0.5)(x-1)(x-2)}{(2.5-0.5)(2.5-1)(2.5-2)} \\ &\approx 0.666667x^3 - 2.33333x^2 + 2.33333x - 0.666667. \end{aligned}$$

Vidare är

$$y_0 = f(x_0) = 4.5, \quad y_1 = f(x_1) = 3, \quad y_2 = f(x_2) = 3 \quad \text{och} \quad y_3 = f(x_3) = 3.3.$$

Det ger att

$$\begin{aligned}
 p_3(x) &= \sum_{k=0}^3 y_k L_{3,k}(x) \\
 &\approx 4.5(-0.66667x^3 + 3.66667x^2 - 6.33333x + 3.33333) \\
 &\quad + 3(1.33333x^3 - 6.66667x^2 + 9.66667x - 3.33333) \\
 &\quad + 3(-1.33333x^3 + 5.33333x^2 - 5.66667x + 1.66667) \\
 &\quad + 3.3(0.66667x^3 - 2.33333x^2 + 2.33333x - 0.66667) \\
 &\approx -0.8x^3 + 4.8x^2 - 8.8x + 7.8.
 \end{aligned}$$

Vi får att $f(1.2) \approx p_3(1.2) \approx 2.7696$ och $f(1.5) \approx p_3(1.5) \approx 2.7$.

Uppgift 6. Polynomen ges av

$$\begin{aligned}
 p_1(x) &= a_0 + a_1(x - x_0) = 4 - 1(x - 1) = 5 - x, \\
 p_2(x) &= a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1) \\
 &= 4 - (x - 1) + 0.4(x - 1)(x - 3) = 6.2 - 2.6x + 0.4x^2, \\
 p_3(x) &= a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1) + a_3(x - x_0)(x - x_1)(x - x_2) \\
 &= 4 - (x - 1) + 0.4(x - 1)(x - 3) + 0.01(x - 1)(x - 3)(x - 4) \\
 &= 6.08 - 2.41x + 0.32x^2 + 0.01x^3
 \end{aligned}$$

och

$$\begin{aligned}
 p_4(x) &= a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1) + a_3(x - x_0)(x - x_1)(x - x_2) \\
 &\quad + a_4(x - x_0)(x - x_1)(x - x_2)(x - x_3) \\
 &= 4 - (x - 1) + 0.4(x - 1)(x - 3) + 0.01(x - 1)(x - 3)(x - 4) \\
 &\quad - 0.002(x - 1)(x - 3)(x - 4)(x - 4.5) \\
 &= 5.972 - 2.215x + 0.21x^2 + 0.035x^3 - 0.002x^4.
 \end{aligned}$$

Notera att vi kan förenkla ovanstående en del genom att utnyttja att

$$p_n(x) = p_{n-1}(x) + a_n \prod_{k=0}^{n-1} (x - x_k),$$

för $n = 1, 2, 3, \dots$ och $p_0(x) = a_0$. Avslutningsvis får vi att

$$p_1(2.5) = 2.5, \quad p_2(2.5) = 2.2, \quad p_3(2.5) = 2.21125 \quad \text{och} \quad p_4(2.5) = 2.21575.$$

Uppgift 7. Polynomen ges av

$$\begin{aligned}
 p_1(x) &= a_0 + a_1(x - x_0) = 7 + 3(x + 1) = 10 + 3x, \\
 p_2(x) &= p_1(x) + a_2(x - x_0)(x - x_1) = 10 + 3x + 0.1(x + 1)(x - 0) \\
 &= 10 + 3.1x + 0.1x^2, \\
 p_3(x) &= p_2 + a_3(x - x_0)(x - x_1)(x - x_2) \\
 &= 10 + 3.1x + 0.1x^2 + 0.05(x + 1)(x - 0)(x - 1) \\
 &= 10 + 3.05x + 0.1x^2 + 0.05x^3
 \end{aligned}$$

och

$$\begin{aligned} p_4(x) &= p_3(x) + a_4(x - x_0)(x - x_1)(x - x_2)(x - x_3) \\ &= 10 + 3.05x + 0.1x^2 + 0.05x^3 - 0.04(x + 1)(x - 0)(x - 1)(x - 4) \\ &= 10 + 2.89x + 0.14x^2 + 0.21x^3 - 0.04x^4. \end{aligned}$$

Det ger att

$$p_1(3) = 19, \quad p_2(3) = 20.2, \quad p_3(3) = 21.4 \quad \text{och} \quad p_4(3) = 22.36.$$

Uppgift 9. Vi har att

$$\begin{aligned} L_{3,0}(x) &= \frac{(x - 0.5)(x - 1.0)(x - 1.5)}{(0.0 - 0.5)(0.0 - 1.0)(0.0 - 1.5)} = 1 - \frac{11}{3}x + 4x^2 - \frac{4}{3}x^3 \\ L_{3,1}(x) &= \frac{(x - 0.0)(x - 1.0)(x - 1.5)}{(0.5 - 0.0)(0.5 - 1.0)(0.5 - 1.5)} = 6x - 10x^2 + 4x^3 \\ L_{3,2}(x) &= \frac{(x - 0.0)(x - 0.5)(x - 1.5)}{(1.0 - 0.0)(1.0 - 0.5)(1.0 - 1.5)} = -3x + 8x^2 - 4x^3 \\ L_{3,3}(x) &= \frac{(x - 0.0)(x - 0.5)(x - 1.0)}{(1.5 - 0.0)(1.5 - 0.5)(1.5 - 1.0)} = \frac{2}{3}x - 2x^2 + \frac{4}{3}x^3. \end{aligned}$$

och därmed

$$p_3(x) = 1 + 0.0193991x - 0.626484x^2 + 0.273451x^3.$$

Uppgift 10. (a) Vi får följande tabell.

k	x_k	$f[x_k]$	$f[\cdot, \cdot]$	$f[\cdot, \cdot, \cdot]$	$f[\cdot, \cdot, \cdot, \cdot]$	$f[\cdot, \cdot, \cdot, \cdot, \cdot]$
0	4.0	2.00000		0.23607		
1	5.0	2.23607		-0.011325		
2	6.0	2.44949	0.21342		0.000915	
3	7.0	2.64575	0.19626	-0.008580		-0.00008
4	8.0	2.82843	0.18268	-0.006790	0.000597	

(b) Alltså är

$$\begin{aligned} a_0 &= f[x_0] = 2.00000, \\ a_1 &= f[x_0, x_1] \approx 0.23607, \\ a_2 &= f[x_0, x_1, x_2] \approx -0.011325, \\ a_3 &= f[x_0, x_1, x_2, x_3] \approx 0.000915, \\ a_4 &= f[x_0, x_1, x_2, x_3, x_4] \approx -0.00008. \end{aligned}$$

Det ger att

$$\begin{aligned} p_1(x) &= a_0 + a_1(x - x_0) \approx 2.0000 + 0.23607(x - 4) \\ &\approx 1.05572 + 0.23607x \end{aligned}$$

$$p_2(x) = p_1(x) + a_2(x - x_0)(x - x_1) \approx p_1(x) - 0.011325(x - 4)(x - 5)$$

$$\begin{aligned}
&\approx 0.82922 + 0.337995x - 0.011325x^2 \\
p_3(x) &= p_2(x) + a_3(x - x_0)(x - x_1)(x - x_2) \\
&\approx p_2(x) + 0.000915(x - 4)(x - 5)(x - 6) \\
&\approx 0.71942 + 0.405705x - 0.02505x^2 + 0.000915x^3 \\
p_4(x) &= p_3(x) + a_4(x - x_0)(x - x_1)(x - x_2)(x - x_3) \\
&\approx p_3(x) - 0.00008(x - 4)(x - 5)(x - 6)(x - 7) \\
&\approx 0.65257 + 0.45648x - 0.039295x^2 + 0.00267x^3 - 0.00008x^4.
\end{aligned}$$

(c) Vi får att

$$\begin{aligned}
p_1(\alpha) &= p_1(4.5) \approx 2.11804 & p_2(\alpha) &= p_2(4.5) \approx 2.12087 \\
p_3(\alpha) &= p_3(4.5) \approx 2.12121 & p_4(\alpha) &= p_4(4.5) \approx 2.12128
\end{aligned}$$

och

$$\begin{aligned}
p_1(\beta) &= p_1(7.5) \approx 2.82624 & p_2(\beta) &= p_2(7.5) \approx 2.72715 \\
p_3(\beta) &= p_3(7.5) \approx 2.73916 & p_4(\beta) &= p_4(7.5) \approx 2.73864.
\end{aligned}$$

(d) Vi jämför resultaten i föregående deluppgift med

$$f(\alpha) = f(4.5) = \sqrt{4.5} \approx 2.12132$$

och

$$f(\beta) = f(7.5) = \sqrt{7.5} \approx 2.73861.$$

Ju större värde på n desto bättre approximation med p_n .

Uppgift 14. För S ska vara en kubisk spline S måste

- (1) $\deg S_i \leq 3$ för $i = 1, 2$
- (2) $S_1(2) = S_2(2)$
- (3) $S'_1(2) = S'_2(2)$
- (4) $S''_1(2) = S''_2(2)$

Vi kan bortse från villkoret $S(x_i) = y_i$, för alla $i = 0, 1, \dots, n$, då det i uppgiften inte förekommer några y_i . Även första villkoret är uppfyllt – alla S_1 och S_2 är polynom av grad 3. Det återstår att kontrollera villkoren (2)–(4). (a) Från

$$S_1(2) = 3 = S_2(2), \quad S'_1(2) = \frac{3}{4} = S'_2(2) \quad \text{och} \quad S''_1(2) = -9 = S''_2(2)$$

följer att S är en kubisk spline. (b) Från $S_1(2) = 3 \neq 2 = S_2(2)$ följer att S inte är en kubisk spline. Det räcker inte att villkoren (3) och (4), dvs

$$S'_1(2) = 0 = S'_2(2) \quad \text{och} \quad S''_1(2) = -12 = S''_2(2)$$

är uppfyllda. (c) Från

$$S_1(2) = 3 = S_2(2), \quad S'_1(2) = \frac{1}{2} = S'_2(2) \quad \text{och} \quad S''_1(2) = -14 = S''_2(2)$$

följer att S är en kubisk spline. (d) Både från

$$S'_1(2) = 1 \neq -1 = S'_2(2) \quad \text{och} \quad S''_1(2) = -14 \neq -8 = S''_2(2)$$

följer att S inte är en kubisk spline. Notera att $S_1(2) = 3 = S_2(2)$.

Uppgift 16. Låt

$$S_1(x) = a_1 + b_1 u_1 + c_1 u_1^2 + d_1 u_1^3$$

och

$$S_n(x) = a_n + b_n u_n + c_n u_n^2 + d_n u_n^3,$$

där

$$u_1 = \frac{x - x_0}{h_1} \quad \text{och} \quad u_n = \frac{x - x_{n-1}}{h_n}$$

samt $h_1 = x_1 - x_0$ och $h_n = x_n - x_{n-1}$. Vidare är $k_i = S'(x_i)$, dvs $S'_1(x_0) = k_0$ och $S'_n(x_n) = k_n$. Det ger att

$$S'_1(x_0) = S'_n(x_n) \Leftrightarrow k_0 = k_n.$$

Då återstår andra ekvationen i randvillkoret. Eftersom $u'_i = 1/h_i$, så är

$$S'_1(x) = b_1 u'_1 + 2c_1 u_1 u'_1 + 3d_1 u_1^2 u'_1 = \frac{1}{h_1} b_n + \frac{2}{h_1} c_n u_n + \frac{3}{h_1} d_n u_n^2$$

och

$$S''_1(x) = \frac{2}{h_1} c_1 u'_1 + \frac{6}{h_1} d_n u_1 u'_1 = \frac{2}{h_1^2} c_1 + \frac{6}{h_1^2} d_n u_1.$$

På samma sätt finner vi att

$$S''_n(x) = \frac{2}{h_n^2} c_n + \frac{6}{h_n^2} d_n u_n.$$

Eftersom $u_1(x_0) = 0$ och $u_n(x_n) = 1$ så har vi at

$$S''_1(x_0) = \frac{2c_1}{h_1^2} \quad \text{och} \quad S''_n(x_n) = \frac{2c_n}{h_n^2} + \frac{6d_n}{h_n^2}.$$

Från

$$\begin{cases} a_i = y_{i-1} \\ b_i = h_i k_{i-1} \\ c_i = 3(y_i - y_{i-1}) - h_i(2k_{i-1} + k_i) \\ d_i = 2(y_{i-1} - y_i) + h_i(k_{i-1} + k_i) \end{cases}$$

följer att

$$\begin{aligned} S''_1(x_0) &= \frac{2}{h_1^2} (3(y_1 - y_0) - h_1(2k_0 + k_1)) \\ &= \frac{6}{h_1^2} (y_0 - y_1) - \frac{2}{h_1} (2k_0 + k_1). \end{aligned}$$

och

$$\begin{aligned} S''_n(x_n) &= \frac{1}{h_n^2} (2 \{ 3(y_n - y_{n-1}) - h_n(2k_{n-1} + k_n) \} \\ &\quad + 6 \{ 2(y_{n-1} - y_n) + h_n(k_{n-1} + k_n) \}) \\ &= \frac{1}{h_n^2} (6(y_{n-1} - y_n) + 2h_n(k_{n-1} + 2k_n)) \\ &= \frac{6}{h_n^2} (y_{n-1} - y_n) + \frac{2}{h_n} (k_{n-1} + 2k_n). \end{aligned}$$

Därmed är $S_1''(x_0) = S_n''(x_n)$ ekvivalent med

$$\begin{aligned} \frac{6}{h_1^2}(y_0 - y_1) - \frac{2}{h_1}(2k_0 + k_1) &= \frac{6}{h_n^2}(y_{n-1} - y_n) + \frac{2}{h_n}(k_{n-1} + 2k_n) \\ \Leftrightarrow \\ \frac{6h_n}{h_1}(y_0 - y_1) - 2h_n(2k_0 + k_1) &= \frac{6h_1}{h_n}(y_{n-1} - y_n) + 2h_1(k_{n-1} + 2k_n) \\ \Leftrightarrow \\ 4h_n k_0 + 2h_n k_1 + 2h_1 k_{n-1} + 4h_1 k_n &= r_n, \end{aligned}$$

där

$$r_n = 6 \left(h_n \frac{y_0 - y_1}{h_1} + h_1 \frac{y_n - y_{n-1}}{h_n} \right).$$

Uppgift 19. Bernsteinpolynomen definieras enligt

$$B_{i,n}(t) = \binom{n}{i} t^i (1-t)^{n-i}.$$

Vi får att

$$(a) \quad B_{2,4}(t) = \binom{4}{2} t^2 (1-t)^{4-2} = 6t^2(1-2t+t^2) = 6t^2 - 12t^3 + 6t^4$$

$$(b) \quad B_{3,5}(t) = \binom{5}{3} t^3 (1-t)^{5-3} = 10t^3(1-2t+t^2) = 10t^3 - 20t^4 + 10t^5$$

$$(c) \quad B_{5,7}(t) = \binom{7}{5} t^5 (1-t)^{7-5} = 21t^5(1-2t+t^2) = 21t^5 - 42t^6 + 21t^7$$

där binomialkoefficienterna ges av

$$\binom{4}{2} = \frac{4!}{2! \cdot (4-2)!} = 6, \quad \binom{5}{3} = \frac{5!}{3! \cdot (5-3)!} = 10 \quad \text{och} \quad \binom{7}{5} = \frac{7!}{5! \cdot (7-5)!} = 21.$$

Uppgift 20. Bernsteinpolynomen av grad 3 är

$$\begin{aligned} B_{0,3}(t) &= 1 - 3t + 3t^2 - t^3 & B_{1,3}(t) &= 3t - 6t^2 + 3t^3 \\ B_{2,3}(t) &= 3t^2 - 3t^3 & B_{3,3}(t) &= t^3. \end{aligned}$$

Den sökta Bézierkurvan ges då av

$$\begin{aligned} b_3(t) &= \sum_{i=0}^3 B_{i,3}(t) P_i \\ &= (-1, 2)B_{0,3}(t) + (1, 0)B_{1,3}(t) + (2, 3)B_{2,3}(t) + (-1, 0)B_{3,3}(t) \\ &= (-1, 2)(1 - 3t + 3t^2 - t^3) + (1, 0)(3t - 6t^2 + 3t^3) \\ &\quad + (2, 3)(3t^2 - 3t^3) + (-1, 0)t^3 \\ &= (-1 + 6t - 3t^2 - 3t^3, 2 - 6t + 15t^2 - 11t^3). \end{aligned}$$

Uppgift 21. (a) Bernsteinpolynomen av grad 3:

$$\begin{aligned} B_{0,3}(t) &= 1 - 3t + 3t^2 - t^3 & B_{1,3}(t) &= 3t - 6t^2 + 3t^3 \\ B_{2,3}(t) &= 3t^2 - 3t^3 & B_{3,3}(t) &= t^3. \end{aligned}$$

Den sökta Bézierkurvan ges då av

$$\begin{aligned} b_3(t) &= (1, 3)B_{0,3}(t) + (3, -1)B_{1,3}(t) + (2, 4)B_{2,3}(t) + (3, 0)B_{3,3}(t) \\ &= (1 + 6t - 9t^2 + 5t^3, 3 - 12t + 27t^2 - 18t^3). \end{aligned}$$

(b) Bernsteinpolynomen av grad 4:

$$\begin{aligned} B_{0,4}(t) &= 1 - 4t + 6t^2 - 4t^3 + t^4 & B_{1,4}(t) &= 4t - 12t^2 + 12t^3 - 4t^4 \\ B_{2,4}(t) &= 6t^2 - 12t^3 + 6t^4 & B_{3,4}(t) &= 4t^3 - 4t^4 \\ B_{4,4}(t) &= t^4. \end{aligned}$$

Bézierkurvan är

$$b_4(t) = (-2 + 4t + 18t^2 - 28t^3 + 10t^4), 3 + 12t^2 - 20t^3 + 8t^4).$$

(c) Bernsteinpolynomen av grad 5:

$$\begin{aligned} B_{0,5}(t) &= 1 - 5t + 10t^2 - 10t^3 + 5t^4 - t^5 \\ B_{1,5}(t) &= 5t - 20t^2 + 30t^3 - 20t^4 + 5t^5 \\ B_{2,5}(t) &= 10t^2 - 30t^3 + 30t^4 - 10t^5 \\ B_{3,5}(t) &= 10t^3 - 20t^4 + 10t^5 \\ B_{4,5}(t) &= 5t^4 - 5t^5 \\ B_{5,5}(t) &= t^5. \end{aligned}$$

Bézierkurvan är

$$b_5(t) = (1 + 5t, 1 + 5t + 10t^2 - 30t^3 + 15t^4).$$

Uppgift 25. Utnyttja först att $P_0 = b_3(0)$ och $P_3 = b_3(1)$. Ansätt därefter $P_1 = (a, b)$ och $P_2 = (c, d)$. Sätt in i den generella formeln för $b_3(t)$ och identifiera koefficienter. Det ger ett ekvationssystem med avseende på a, b, c och d .

Uppgift 26. Vi deriverar Bernsteinpolynomen två gånger. Först får vi att

$$\begin{aligned} B'_{i,n}(t) &= \frac{d}{dt} \binom{n}{i} t^i (1-t)^{n-i} \\ &= \frac{n!}{i! \cdot (n-i)!} (it^{i-1}(1-t)^{n-i} - (n-i)t^i(1-t)^{n-i-1}) \\ &= n \frac{(n-1)!}{(i-1)! \cdot (n-i)!} t^{i-1}(1-t)^{n-i} - n \frac{(n-1)!}{i! \cdot (n-i-1)!} t^i(1-t)^{n-i-1} \\ &= n \binom{n-1}{i-1} t^{i-1}(1-t)^{n-1-(i-1)} - n \binom{n-1}{i} t^i(1-t)^{n-1-i} \\ &= n B_{i-1,n-1}(t) - n B_{i,n-1}(t), \end{aligned}$$

är ett rekursivt samband. Utnyttjar vi denna får vi att

$$\begin{aligned} B''_{i,n}(t) &= n \frac{d}{dt} \{B_{i-1,n-1}(t) - B_{i,n-1}(t)\} \\ &= n(n-1) \{B_{i-2,n-2}(t) - B_{i-1,n-2}(t) - B_{i-1,n-2}(t) + B_{i,n-2}(t)\} \\ &= n(n-1) \{B_{i-2,n-2}(t) - 2B_{i-1,n-2}(t) + B_{i,n-2}(t)\}. \end{aligned}$$

Från definitionen av Bernsteinpolynom följer det att

$$B_{i,n}(0) = \begin{cases} 0 & \text{om } i \neq 0 \\ 1 & \text{om } i = 0 \end{cases} \quad \text{och} \quad B_{i,n}(1) = \begin{cases} 0 & \text{om } i \neq n \\ 1 & \text{om } i = n. \end{cases}$$

Vidare definierar man $B_{i,n}(t) = 0$ för alla t om $i < 0$ eller $i > n$. Om $n = 2$, så är

$$B_{0,2}(t) = 1 - 2t + t^2, \quad B_{1,2}(t) = 2t - 2t^2 \quad \text{och} \quad B_{2,2}(t) = t^2.$$

Det ger att

$$B''_{0,2}(t) = 2, \quad B''_{1,2}(t) = -4 \quad \text{och} \quad B''_{2,2}(t) = 2.$$

Alltså är

$$\begin{aligned} b''_2(t) &= \sum_{i=0}^2 B_{i,2}(t)P_i = B_{0,2}(t)P_0 + B_{1,2}(t)P_1 + B_{2,2}(t)P_2 \\ &= 2P_0 - 4P_1 + 2P_2 = 2(2 - 1)(P_2 - 2P_1 + P_0), \end{aligned}$$

vilket visar påståendena i båda deluppgifterna i fallet $n = 2$. I fortsättningen antar vi att $n > 2$. (a) Vi har att

$$B_{i-2,n-2}(0) = \begin{cases} 0 & \text{om } i \neq 2 \\ 1 & \text{om } i = 2 \end{cases} \quad \text{och} \quad B_{i-1,n-2}(0) = \begin{cases} 0 & \text{om } i \neq 1 \\ 1 & \text{om } i = 1. \end{cases}$$

Alltså är

$$\begin{aligned} b''_n(0) &= n(n-1) \sum_{i=0}^n \{B_{i-2,n-2}(0) - 2B_{i-1,n-2}(0) + B_{i,n-2}(0)\} P_i \\ &= n(n-1)(P_2 - 2P_1 + P_0), \end{aligned}$$

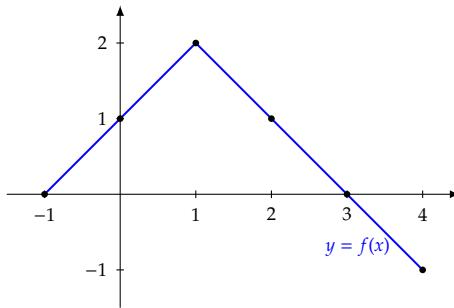
vilket skulle visas. (b) Vi har att

$$B_{i-1,n-2}(1) = \begin{cases} 0 & \text{om } i \neq n-1 \\ 1 & \text{om } i = n-1 \end{cases} \quad \text{och} \quad B_{i,n-2}(0) = \begin{cases} 0 & \text{om } i \neq n-2 \\ 1 & \text{om } i = n-2. \end{cases}$$

Alltså är

$$\begin{aligned} b''_n(1) &= n(n-1) \sum_{i=0}^n \{B_{i-2,n-2}(1) - 2B_{i-1,n-2}(1) + B_{i,n-2}(1)\} P_i \\ &= n(n-1)(P_n - 2P_{n-1} + P_{n-2}), \end{aligned}$$

vilket skulle visas.



Figur L.1

4 Numerisk integration

Uppgift 1. Sätt $f(x) = 2 - |1 - x|$. Förutsättningarna ger att $h = (4 - (-1))/5 = 1$. Då är $x_k = -1 + k$ för $k = 0, 1, \dots, 5$, dvs

$$x_0 = -1, x_1 = 0, x_2 = 1, x_3 = 2, x_4 = 3 \text{ och } x_5 = 4.$$

Sätt $f_k = f(x_k)$. Det ger att

$$f_0 = 0, f_1 = 1, f_2 = 2, f_3 = 1, f_4 = 0 \text{ och } f_5 = -1.$$

Trapetsmetoden ger att

$$\begin{aligned} \int_{-1}^4 f(x) dx &\approx \frac{h}{2}(f_0 + 2f_1 + 2f_2 + 2f_3 + 2f_4 + f_5) \\ &= \frac{1}{2}(0 + 2 + 4 + 2 + 0 - 1) = \frac{7}{2}. \end{aligned}$$

Notera att detta faktiskt är det exakta värdet av integralen eftersom

$$\int_{-1}^4 f(x) dx = \int_{-1}^1 (2 - (1 - x)) dx + \int_1^4 (2 + (2 - x)) dx = \frac{7}{2}.$$

Figur L.1 illustrerar hur trapeterna för delintervallen (x_k, x_{k+1}) tillsammans övertäcker samma area som integralen motsvarar.

Uppgift 3. Låt $f_k = f(0 + hk) = f(hk)$. Hur stort steget h ska vara för respektive metod beror på hur många punkter x_0, x_1, \dots, x_n som behövs.

Metod	n	h
Trapetsregeln	1	1
Simpsons formel	2	1/2
Simpsons 3/8-formel	3	1/3
Booles formel	4	1/4

Här ges steget av $h = (b - a)/n = 1/n$. (a) Trapetsregeln:

$$I \approx \frac{h}{2}(f_0 + f_1) = \frac{1}{2}(\sin(0) + \sin(\pi)) = 0.$$

Simpsons formel:

$$I \approx \frac{h}{3}(f_0 + 3f_1 + f_2) = \frac{1}{6} \left(\sin(0) + 2 \sin\left(\frac{\pi}{2}\right) + \sin(\pi) \right) \approx 0.666667$$

Simpsons 3/8-formel:

$$I \approx \frac{3h}{8}(f_0 + 3f_1 + 3f_2 + f_3) \approx 0.649519.$$

Booles formel:

$$I \approx \frac{2h}{45}(7f_0 + 32f_1 + 12f_2 + 32f_3 + 7f_4) \approx 0.636165.$$

(b) och (c) På samma sätt som i (a) får man följande resultat.

Metod	(b)	(c)
Trapetsregeln	1.37977	0.420735
Simpsons formel	0.95832	0.573336
Simpsons 3/8-formel	0.98693	0.583143
Booles formel	1.00876	0.593376

Detaljerna lämnas som övning.

Uppgift 4. Sätt

$$h = \frac{b-a}{2n} = \frac{1.4-0.2}{2 \cdot 4} = 0.15 \quad \text{och} \quad x_0 = 0.2 + kh = 0.2 + 0.15k,$$

där $k = 0, 1, \dots, 2 \cdot 4$. Låt $f_k = f(x_k)$. Då är

$$\begin{aligned} \int_{0.2}^{1.4} (x^2 - e^{2x}) dx &\approx S(f, 0.15) = \frac{h}{3} \sum_{k=1}^4 (f(x_{2k-2}) + 4f(x_{2k-1}) + f(x_{2k})) \\ &= \frac{0.15}{3} (f_0 + 4f_1 + 2f_2 + 4f_3 + \dots + 2f_6 + 4f_7 + f_8) \\ &\approx 0.05(1.53182 + 4 \cdot 2.13625 + 2 \cdot 2.96828 + \dots \\ &\quad + 2 \cdot 10.235 + 4 \cdot 13.745 + 18.4046) \\ &\approx 8.38874. \end{aligned}$$

Förs att bestämma det relativa felet beräknar vi först integralen, dvs

$$I = \int_{0.2}^{1.4} (x^2 + e^{2x}) dx = \left[\frac{x^3}{3} + \frac{e^{2x}}{2} \right]_{0.2}^{1.4} \approx 8.38841.$$

Det realtiva felet är således

$$R_I = \frac{|I - S(f, 0.15)|}{|I|} \approx \frac{|8.38841 - 8.38874|}{8.38841} \approx 0.00004.$$

Uppgift 6. Låt $x_k = a + kh$ och $f_k = f(x_k) = \cos(\sin x_k)$, där

$$h = \frac{2-0}{2 \cdot 5} = 0.2.$$

Då är

$$\int_0^2 f(x) dx \approx \frac{h}{3} \sum_{k=1}^5 (f_{2k-2} + 4f_{2k-1} + f_{2k}) = 1.44465.$$

Uppgift 7. (a) Trapetsregeln: Sätt

$$h = \frac{b-a}{n} = \frac{1-(-1)}{10} = 0.2 \quad \text{och} \quad x_0 = -1 + kh = -1 + 0.2k,$$

där $k = 0, 1, \dots, 10$. Då är

$$\begin{aligned} \int_{-1}^1 \frac{dx}{1+x^2} &\approx T(f, 0.2) = \frac{h}{2} \sum_{k=1}^{10} (f(x_{k-1}) + f(x_k)) \\ &= \frac{0.2}{2} (f(x_0) + 2f(x_1) + 2f(x_2) + \dots + 2f(x_9) + f(x_{10})) \\ &\approx 0.1(0.5 + 2 \cdot 0.609756 + \dots + 2 \cdot 0.609756 + 0.5) \\ &\approx 1.5674630569. \end{aligned}$$

Simpsons formel: Sätt

$$h = \frac{b-a}{2n} = \frac{1-(-1)}{2 \cdot 5} = 0.2 \quad \text{och} \quad x_0 = -1 + kh = -1 + 0.2k,$$

där $k = 0, 1, \dots, 2 \cdot 5$. Låt $f_k = f(x_k)$. Då är

$$\begin{aligned} \int_{-1}^1 \frac{dx}{1+x^2} &\approx S(f, 0.2) = \frac{h}{3} \sum_{k=1}^5 (f(x_{2k-2}) + 4f(x_{2k-1}) + f(x_{2k})) \\ &= \frac{0.2}{3} (f_0 + 4f_1 + 2f_2 + 4f_3 + \dots + 2f_8 + 4f_9 + f_{10}) \\ &\approx 0.667(0.5 + 4 \cdot 0.609756 + 2 \cdot 0.735294 + \dots \\ &\quad + 2 \cdot 0.862069 + 4 \cdot 0.609756 + 0.5) \\ &\approx 1.570795388. \end{aligned}$$

Låt $F(x) = \arctan(x)$. Då är $F'(x) = 1/(1+x^2)$, dvs F är en primitiv funktion till integranden. Ovanstående kan jämföras med

$$I = \int_{-1}^1 \frac{dx}{1+x^2} = [F(x)]_{-1}^1 = F(1) - F(-1) = \frac{\pi}{2} \approx 1.570796327,$$

De övriga deluppgifterna:

	(b)	(c)	(d)	(e)	(f)
$T(f, h)$	2.857409	3.041920	1.521628	0.783304	0.366951
$S(f, h)$	2.865601	3.007622	1.524599	0.805005	0.382793
I	2.870796	3.000000	1.523793	0.804896	0.382714

där respektive primitiv funktion är

- | | |
|-----|----------------------------------------------------------------------|
| (b) | $F(x) = 2x - \sqrt{x} \cos(2\sqrt{x}) + \frac{1}{2} \sin(2\sqrt{x})$ |
| (c) | $F(x) = 2\sqrt{x}$ |
| (d) | $F(x) = -e^{-x}(x^2 + 2x + 1)$ |
| (e) | $F(x) = 2\cos(x) + 2x\sin(x)$ |
| (f) | $F(x) = -\frac{e^{-x}}{5}(2\cos(2x) + \sin(2x))$ |

Uppgift 8. Låt j vara ett icke-negativt heltal och sätt $h = (b - a)/2^j$. Vi ska fylla i tabellen

j	h
0	$R_0(0)$
1	$R_0(1) \quad R_1(1)$
2	$R_0(2) \quad R_1(2) \quad R_2(2)$

där $R_0(j) = T(f, h)$ and

$$R_{k+1}(j) = R_k(j) + \frac{R_k(j) - R_k(j-1)}{4^{k+1} - 1},$$

(a) Om $j = 0$, så är $h = (3 - 0)/2^0 = 3$. Det ger att

$$R_0(0) = T(f, 3) = \frac{h}{2}(f_0 + f_1) = \frac{3}{2}(f(0) + f(3)) \approx -0.041912.$$

Om $j = 1$, så är $h = (3 - 0)/2^1 = 1.5$ och

$$R_0(1) = T(f, 1.5) = \frac{h}{3}(f_0 + 2f_1 + f_2) \approx 0.0441761.$$

Om $j = 2$, så är $h = (3 - 0)/2^2 = 0.75$ och

$$R_0(2) = T(f, 0.75) = \frac{h}{2}(f_0 + 2f_1 + 2f_2 + 2f_3 + f_4) \approx 0.379954.$$

Då är

$$\begin{aligned} R_1(1) &= R_0(1) + \frac{R_0(1) - R_0(0)}{4^1 - 1} \\ &\approx 0.0441761 + \frac{0.0441761 - (-0.041912)}{3} \approx 0.0728723, \\ R_1(2) &= R_0(2) + \frac{R_0(2) - R_0(1)}{3} \approx 0.49188 \end{aligned}$$

och

$$R_2(2) = R_1(2) + \frac{R_1(2) - R_1(1)}{4^2 - 1} \approx 0.519814.$$

Det ger oss följande tabell.

j	h			
0	3			
1	1.5	0.044176	0.07287	
2	0.75	0.379954	0.49188	0.519814

Övriga deluppgifter lämnas som övning.

j	h		
0	3		
1	1.5	-0.0019951	-0.0218644
2	0.75	-0.0284876	0.0161175

j	h			
0	0.96			
1	0.48	2.10564	1.84752	-0.0287782
2	0.24	1.78168	1.67369	0.0325959

	<i>j</i>	<i>h</i>			
(d)	0	2	10.24390		
	1	1	6.03104	4.6268	
	2	0.5	4.65686	4.1988	4.17027
	<i>j</i>	<i>h</i>			
(e)	0	1.84085	0.441274		
	1	0.92042	0.956419	1.12813	
	2	0.46021	1.216290	1.30291	1.31456
	<i>j</i>	<i>h</i>			
(f)	0	2	2		
	1	1	2.73205	2.97607	
	2	0.5	2.99571	3.0836	3.09076

Uppgift 11. Induktion över k . Om $k = 1$, så vet vi att $I(f) = R_1(h) + O(h^4)$ eller ekvivalent $R_1(h) = I(f) + O(h^{2 \cdot 1 + 1})$. Antag att

$$R_{k-1}(h) = I(f) + O(h^{2(k-1)+2}) = I(f) + a_{k-1}h^{2(k-1)+2} + a_k h^{2k+2} + \dots$$

Då är

$$R_{k-1}(2h) = I(f) + a_{k-1}(2h)^{2(k-1)+2} + a_k(2h)^{2k+2} + \dots$$

och

$$2^{2k}R_{k-1}(h) - R_{k-1}(2h) = (2^{2k} - 1)I(f) + \underbrace{(2^{2k} - 2^{2k+2})a_k h^{2k+2} + \dots}_{=O(h^{2k+2})}$$

Således är

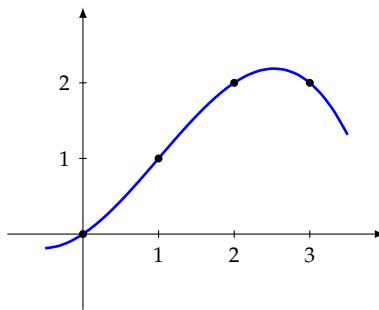
$$\frac{2^{2k}R_{k-1}(h) - R_{k-1}(2h)}{2^{2k} - 1} = I(f) + O(h^{2k+2}),$$

dvs $R_k(h) = I(f) + O(h^{2k+2})$.

5 Numerisk linjär algebra

Uppgift 3. Gaußeliminering med pivotering ger

$$\begin{aligned} & \left\{ \begin{array}{l} 0.4x - 0.5y + 0.8z = 1 \\ -2.8x + 2.0y + 1.6z = 2 \\ 0.2x + 3.2y - 1.2z = 3 \end{array} \right. \quad \left. \begin{array}{c} \square \\ \square \\ \square \end{array} \right. \\ \Leftrightarrow & \left\{ \begin{array}{l} -2.8x + 2.0y + 1.6z = 2 \\ 0.4x - 0.5y + 0.8z = 1 \\ 0.2x + 3.2y - 1.2z = 3 \end{array} \right. \quad \left. \begin{array}{c} \square \\ \square \\ \square \end{array} \right. \quad \left. \begin{array}{c} a_1 \\ a_2 \\ a_3 \end{array} \right. \\ \Leftrightarrow & \left\{ \begin{array}{l} -2.8x + 2.0y + 1.6z = 2 \\ -0.214286y + 1.02857z = 1.28571 \\ 3.34286y - 1.08571z = 3.14286 \end{array} \right. \quad \left. \begin{array}{c} \square \\ \square \\ \square \end{array} \right. \\ \Leftrightarrow & \left\{ \begin{array}{l} -2.8x + 2.0y + 1.6z = 2 \\ 3.34286y - 1.08571z = 3.14286 \\ -0.214286y + 1.02857z = 1.28571 \end{array} \right. \quad \left. \begin{array}{c} \square \\ \square \\ \square \end{array} \right. \quad \left. \begin{array}{c} a_3 \\ a_2 \\ a_1 \end{array} \right. \\ \Leftrightarrow & \left\{ \begin{array}{l} -2.8x + 2.0y + 1.6z = 2 \\ 3.34286y - 1.08571z = 3.14286 \\ 0.958974z = 1.48718 \end{array} \right. \end{aligned}$$



Figur L.2

där

$$\begin{aligned}a_1 &= -0.4/2.8 = 0.142857 \\a_2 &= -0.2/2.8 = -0.0714286 \\a_3 &= -0.214286/3.34286 = -0.0641026.\end{aligned}$$

Bakåtsubstitution ger lösningen $(x, y, z) = (1.20321, 1.44385, 1.5508)$.

Uppgift 6. De fyra punkterna ger oss fyra linjära ekvationer, nämligen

$$\begin{cases} a &= 0 \\ a + b + c + d &= 1 \\ a + 2b + 4c + 8d &= 2 \\ a + 3b + 9c + 27d &= 2. \end{cases}$$

Vi ser att $a = 0$ och kan därför skriva om ekvationssystemet till ett med tre ekvationer och tre obekanta. Vi löser det erhållna ekvationssystemet med gausselimination:

$$\begin{array}{l} \left\{ \begin{array}{l} b + c + d = 1 \\ 2b + 4c + 8d = 2 \\ 3b + 9c + 27d = 2 \end{array} \right. \xrightarrow[-2]{\quad} \left\{ \begin{array}{l} b + c + d = 1 \\ 2c + 6d = 0 \\ 6c + 24d = -1 \end{array} \right. \xrightarrow[-3]{\quad} \left\{ \begin{array}{l} b + c + d = 1 \\ 2c + 6d = 0 \\ 6d = -1 \end{array} \right. \\ \Leftrightarrow \\ \left\{ \begin{array}{l} b + c + d = 1 \\ 2c + 6d = 0 \\ 6d = -1 \end{array} \right. \end{array}$$

Bakåtsubstitution ger att $d = -1/6$, sedan att $2c = -6d = 1$, dvs $c = 1/2$. Slutligen får vi att $b = 1 - c - d = 1 - 1/2 + 1/6 = 2/3$. Den sökta kurvan är således

$$y = \frac{2}{3}x + \frac{1}{2}x^2 - \frac{1}{6}x^3,$$

se figur L.2.

Uppgift 7. Den exakta lösningen ges av

$$\begin{cases} x_1 = 0 \\ x_2 = \frac{1}{999\,997} = 0.00000100000300000749\dots \\ x_3 = \frac{10\,000}{999\,997} = 0.0100000300000899992\dots \end{cases}$$

Jämför ovanstående med de approximationer av lösningen vi bestämmer nedan. För mer information om funktionen $\text{fl} = \text{fl}_{\text{round}}$ se exempel 1.16.

(a) Ekvationssystemet kan representeras med totalmatrisen

$$(\mathbf{A} \mid \mathbf{b}) = \left(\begin{array}{ccc|c} 2 & -3 & 100 & 1 \\ 1 & 10 & -0.001 & 0 \\ 3 & -100 & 0.01 & 0 \end{array} \right).$$

Eftersom

$$\max(|a_{1,1}|, |a_{2,1}|, |a_{3,1}|) = \max(|2|, |1|, |3|) = 3,$$

så ska vi byta plats på rad ett och tre, dvs

$$\left(\begin{array}{ccc|c} 3 & -100 & 0.01 & 0 \\ 1 & 10 & -0.001 & 0 \\ 2 & -3 & 100 & 1 \end{array} \right).$$

Det ger oss multiplikatorerna

$$m_{2,1} = \text{fl} \frac{1}{3} = 0.3333 \quad \text{och} \quad m_{3,1} = \text{fl} \frac{2}{3} = 0.6667.$$

Därefter multiplicerar vi första raden med $-m_{i,1}$ och adderar resultatet till rad i , dvs elementet på rad i och kolonn j ges av

$$\text{fl}(a_{i,j} - \text{fl}(m_{i,1}a_{1,j})) \quad \text{respektive} \quad \text{fl}(b_i - \text{fl}(m_{i,1}b_1))$$

där $i = 2, 3$ och $j = 1, 2, 3$. Det ger oss ekvationssystemet

$$\left(\begin{array}{ccc|c} 3 & -100 & 0.01 & 0 \\ 0.0001 & 43.33 & -0.004333 & 0 \\ 0.0000 & 63.67 & 99.99 & 1 \end{array} \right).$$

Notera att avrundningarna ger ett uppenbart fel i första elementet på andra raden. Efter Gausseliminationen förväntas vi oss en 0:a där! En lösning är att helt sonika strunta i beräkningarna i första kolonnen och sätta första elementet på andra och tredje raden till 0, dvs

$$\left(\begin{array}{ccc|c} 3 & -100 & 0.01 & 0 \\ 0 & 43.33 & -0.004333 & 0 \\ 0 & 63.67 & 99.99 & 1 \end{array} \right).$$

Härnäst studerr vi andra kolonnen. Från

$$\max(|a_{2,2}|, |a_{3,2}|) = \max(|43.33|, |63.67|) = 63.67$$

följer att vi ska byta plats på rad två och tre. Alltså får vi

$$\left(\begin{array}{ccc|c} 3 & -100 & 0.01 & 0 \\ 0 & 63.67 & 99.99 & 1 \\ 0 & 43.33 & -0.004333 & 0 \end{array} \right).$$

Multiplikatorn är

$$m_{3,2} = \text{fl} \frac{43.33}{63.67} = 0.6805.$$

Multiplicerar vi andra raden med $-m_{3,2}$ och adderar resultatet till tredje raden får vi

$$\left(\begin{array}{ccc|c} 3 & -100 & 0.01 & 0 \\ 0 & 63.67 & 99.99 & 1 \\ 0 & 0 & -68.04 & -0.6805 \end{array} \right).$$

Bakåtsubstitution ger nu i tur och ordning att

$$x_3 = \text{fl} \frac{-0.6805}{-68.04} = 0.01,$$

$$x_2 = \text{fl} \frac{\text{fl}(1 - \text{fl}(99.99 \cdot 0.01))}{63.67} = 0.000001571$$

och

$$x_1 = \text{fl} \frac{\text{fl}(0 - \text{fl}(\text{fl}(-100 \cdot 0.000001571) + \text{fl}(0.01 \cdot 0.01)))}{3} = 0.00001903,$$

jämför med den korrekta lösningen.

(b) Återigen studerar vi ekvationssystemet

$$(A | b) = \left(\begin{array}{ccc|c} 2 & -3 & 100 & 1 \\ 1 & 10 & -0.001 & 0 \\ 3 & -100 & 0.01 & 0 \end{array} \right).$$

Först bestämmer vi det till beloppet största elementet på varje rad i matrisen A, dvs

$$s_1 = \max(|a_{1,1}|, |a_{1,2}|, |a_{1,3}|) = \max(|2|, |-3|, |100|) = 100,$$

$$s_2 = \max(|a_{2,1}|, |a_{2,2}|, |a_{2,3}|) = \max(|1|, |10|, |-0.001|) = 10$$

och

$$s_3 = \max(|a_{3,1}|, |a_{3,2}|, |a_{3,3}|) = \max(|3|, |-100|, |0.01|) = 100.$$

För att finna pivotelementet studerar vi

$$\max\left(\frac{|a_{1,1}|}{s_1}, \frac{|a_{2,1}|}{s_2}, \frac{|a_{3,1}|}{s_3}\right) = \max\left(\frac{|2|}{100}, \frac{|1|}{10}, \frac{|3|}{100}\right) = 0.1,$$

vilket härrör från andra raden. Alltså ska vi byta plats på första och andra raden. Multiplikatorerna ges av $m_{2,1} = 2$ och $m_{3,1} = 3$ och efter Gausselimination erhåller vi ekvationssystemet

$$\left(\begin{array}{ccc|c} 1 & 10 & -0.001 & 0 \\ 0 & -23 & 100 & 1 \\ 0 & -130 & 0.013 & 0 \end{array} \right).$$

Notera speciellt att

$$a_{2,3} = \text{fl}(100 - \text{fl}(2 \cdot (-0.001))) = \text{fl}(100.002) = 100.$$

För nästa pivoting studerar vi andra och tredje raden, dvs

$$s_2 = \max(|-23|, |100|) = 100 \quad \text{och} \quad s_3 = \max(|-130|, |0.013|) = 130.$$

Det ger att

$$\max\left(\frac{|a_{2,2}|}{s_2}, \frac{|a_{2,3}|}{s_3}\right) = \max\left(\frac{|-23|}{100}, \frac{|-130|}{130}\right) = 1,$$

dvs pivotelementet hittar vi tredje raden. Efter att bytt plats på andra och tredje raden får vi att $m_{3,2} = 0.1769$ och Gausseliminationen ger därmed ekvationssystemet

$$\begin{array}{ccc|c} 1 & 10 & -0.001 & 0 \\ 0 & -130 & 0.013 & 0 \\ 0 & 0 & 100 & 1 \end{array}$$

Bakåstsubstitution ger att

$$(x_1, x_2, x_3) = (0, 0.000001, 0.01).$$

De utelämnade detaljerna av beräkningarna ovan lämnas som övning.

Uppgift 8. (a) Sätt $x = (x_1, x_2, x_3)$ och $y = (y_1, y_2, y_3)$. Då är

$$Ly = b \Leftrightarrow \begin{cases} y_1 = -4 \\ y_1/2 + y_2 = 10 \\ y_1/2 + y_2/3 + y_3 = 5 \end{cases} \Leftrightarrow \begin{cases} y_1 = -4 \\ y_2 = 12 \\ y_3 = 3. \end{cases}$$

och

$$Ux = y \Leftrightarrow \begin{cases} 2x_1 + 4x_2 - 6x_3 = -4 \\ 3x_2 + 6x_3 = 12 \\ 3x_3 = 3 \end{cases} \Leftrightarrow \begin{cases} x_1 = -3 \\ x_2 = 2 \\ x_3 = 1. \end{cases}$$

Slutligen kontrollerar vi lösningen, dvs

$$Ax = \begin{pmatrix} 2 & 4 & -6 \\ 1 & 5 & 3 \\ 1 & 3 & 2 \end{pmatrix} \begin{pmatrix} -3 \\ 2 \\ 1 \end{pmatrix} = \begin{pmatrix} -4 \\ 10 \\ 5 \end{pmatrix} = b.$$

(b) Tillvägagångssättet är detsamma som i föregående deluppgift. Du ska komma fram till att $y = (20, 39, 9)$ och $x = (5, 7, 3)$.

Uppgift 9. (a) Steg för steg bestämmer vi U respektive L:

Steg 0	$\begin{pmatrix} 4 & 2 & 1 \\ 2 & 5 & -2 \\ 1 & -2 & 7 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 & 0 \\ * & 1 & 0 \\ * & * & 1 \end{pmatrix}$
Steg 1	$\begin{pmatrix} 4 & 2 & 1 \\ 2 & 5 & -2 \\ 1 & -2 & 7 \end{pmatrix} \xrightarrow{-0.5}$	$\begin{pmatrix} 1 & 0 & 0 \\ 0.5 & 1 & 0 \\ * & * & 1 \end{pmatrix}$
Steg 2	$\begin{pmatrix} 4 & 2 & 1 \\ 0 & 4 & -2.5 \\ 1 & -2 & 7 \end{pmatrix} \xrightarrow{-0.25}$	$\begin{pmatrix} 1 & 0 & 0 \\ 0.5 & 1 & 0 \\ 0.25 & * & 1 \end{pmatrix}$
Steg 3	$\begin{pmatrix} 4 & 2 & 1 \\ 0 & 4 & -2.5 \\ 1 & -2.5 & 6.75 \end{pmatrix} \xrightarrow{0.625}$	$\begin{pmatrix} 1 & 0 & 0 \\ 0.5 & 1 & 0 \\ 0.25 & -0.625 & 1 \end{pmatrix}$

Steg 4 $\begin{pmatrix} 4 & 2 & 1 \\ 0 & 4 & -2.5 \\ 0 & 0 & 5.1875 \end{pmatrix}$

Alltså är

$$\begin{pmatrix} 4 & 2 & 1 \\ 2 & 5 & -2 \\ 1 & -2 & 7 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0.5 & 1 & 0 \\ 0.25 & -0.625 & 1 \end{pmatrix} \begin{pmatrix} 4 & 2 & 1 \\ 0 & 4 & -2.5 \\ 0 & 0 & 5.1875 \end{pmatrix}.$$

(b) Steg för steg:

Steg 0	$\begin{pmatrix} 1 & -2 & 7 \\ 4 & 2 & 1 \\ 2 & 5 & -2 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 & 0 \\ * & 1 & 0 \\ * & * & 1 \end{pmatrix}$
Steg 1	$\begin{pmatrix} 1 & -2 & 7 \\ 4 & 2 & 1 \\ 2 & 5 & -2 \end{pmatrix} \xrightarrow{-4}$	$\begin{pmatrix} 1 & 0 & 0 \\ 4 & 1 & 0 \\ * & * & 1 \end{pmatrix}$
Steg 2	$\begin{pmatrix} 1 & -2 & 7 \\ 0 & 10 & -27 \\ 2 & 5 & -2 \end{pmatrix} \xrightarrow{-2}$	$\begin{pmatrix} 1 & 0 & 0 \\ 4 & 1 & 0 \\ 2 & * & 1 \end{pmatrix}$
Steg 3	$\begin{pmatrix} 1 & -2 & 7 \\ 0 & 10 & -27 \\ 0 & 9 & -16 \end{pmatrix} \xrightarrow{-0.9}$	$\begin{pmatrix} 1 & 0 & 0 \\ 4 & 1 & 0 \\ 2 & 0.9 & 1 \end{pmatrix}$
Steg 4	$\begin{pmatrix} 1 & -2 & 7 \\ 0 & 10 & -27 \\ 0 & 0 & 8.3 \end{pmatrix}$	

Alltså är

$$\begin{pmatrix} 1 & -2 & 7 \\ 4 & 2 & 1 \\ 2 & 5 & -2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 4 & 1 & 0 \\ 2 & 0.9 & 1 \end{pmatrix} \begin{pmatrix} 1 & -2 & 7 \\ 0 & 10 & -27 \\ 0 & 0 & 8.3 \end{pmatrix}.$$

Uppgift 10. Steg för steg:

Steg 0	$\begin{pmatrix} 2 & 1 & -1 \\ 2 & 5 & 3 \\ 1 & -1 & 3 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 & 0 \\ * & 1 & 0 \\ * & * & 1 \end{pmatrix}$
Steg 1	$\begin{pmatrix} 2 & 1 & -1 \\ 2 & 5 & 3 \\ 1 & -1 & 3 \end{pmatrix} \xrightarrow{-1}$	$\begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ * & * & 1 \end{pmatrix}$
Steg 2	$\begin{pmatrix} 2 & 1 & -1 \\ 0 & 4 & 4 \\ 1 & -1 & 3 \end{pmatrix} \xrightarrow{-0.5}$	$\begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0.5 & * & 1 \end{pmatrix}$
Steg 3	$\begin{pmatrix} 2 & 1 & -1 \\ 0 & 4 & 4 \\ 0 & -1.5 & 3.5 \end{pmatrix} \xrightarrow{0.375}$	$\begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0.5 & -0.375 & 1 \end{pmatrix}$
Steg 4	$\begin{pmatrix} 2 & 1 & -1 \\ 0 & 4 & 4 \\ 0 & 0 & 5 \end{pmatrix}$	

Alltså är

$$\begin{pmatrix} 2 & 1 & -1 \\ 2 & 5 & 3 \\ 1 & -1 & 3 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0.5 & -0.375 & 1 \end{pmatrix} \begin{pmatrix} 2 & 1 & -1 \\ 0 & 4 & 4 \\ 0 & 0 & 5 \end{pmatrix}.$$

Uppgift 12. Eftersom A är tridiagonal behöver vi endast "eliminera" ettorna under huvuddiagonalen för att erhålla matrisen U . Det ger oss följande uppställning.

	$\begin{pmatrix} 2 & 1 & 0 & 0 & 0 \\ 1 & 2 & 1 & 0 & 0 \\ 0 & 1 & 2 & 1 & 0 \\ 0 & 0 & 1 & 2 & 1 \\ 0 & 0 & 0 & 1 & 2 \end{pmatrix}$	$L = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ * & 1 & 0 & 0 & 0 \\ 0 & * & 1 & 0 & 0 \\ 0 & 0 & * & 1 & 0 \\ 0 & 0 & 0 & * & 1 \end{pmatrix}$
Steg 0:		
Steg 1:	$\begin{pmatrix} 2 & 1 & 0 & 0 & 0 \\ 1 & 2 & 1 & 0 & 0 \\ 0 & 1 & 2 & 1 & 0 \\ 0 & 0 & 1 & 2 & 1 \\ 0 & 0 & 0 & 1 & 2 \end{pmatrix} \xrightarrow{-1/2}$	$L = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 1 & 0 & 0 & 0 \\ 0 & * & 1 & 0 & 0 \\ 0 & 0 & * & 1 & 0 \\ 0 & 0 & 0 & * & 1 \end{pmatrix}$
Steg 2:	$\begin{pmatrix} 2 & 1 & 0 & 0 & 0 \\ 0 & \frac{3}{2} & 1 & 0 & 0 \\ 0 & 1 & 2 & 1 & 0 \\ 0 & 0 & 1 & 2 & 1 \\ 0 & 0 & 0 & 1 & 2 \end{pmatrix} \xrightarrow{-2/3}$	$L = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 1 & 0 & 0 & 0 \\ 0 & \frac{2}{3} & 1 & 0 & 0 \\ 0 & 0 & * & 1 & 0 \\ 0 & 0 & 0 & * & 1 \end{pmatrix}$
Steg 3:	$\begin{pmatrix} 2 & 1 & 0 & 0 & 0 \\ 0 & \frac{3}{2} & 1 & 0 & 0 \\ 0 & 0 & \frac{4}{3} & 1 & 0 \\ 0 & 0 & 1 & 2 & 1 \\ 0 & 0 & 0 & 1 & 2 \end{pmatrix} \xrightarrow{-3/4}$	$L = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 1 & 0 & 0 & 0 \\ 0 & \frac{2}{3} & 1 & 0 & 0 \\ 0 & 0 & \frac{3}{4} & 1 & 0 \\ 0 & 0 & 0 & * & 1 \end{pmatrix}$
Steg 4:	$\begin{pmatrix} 2 & 1 & 0 & 0 & 0 \\ 0 & \frac{3}{2} & 1 & 0 & 0 \\ 0 & 0 & \frac{4}{3} & 1 & 0 \\ 0 & 0 & 0 & \frac{5}{4} & 1 \\ 0 & 0 & 0 & 1 & 2 \end{pmatrix} \xrightarrow{-4/5}$	$L = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 1 & 0 & 0 & 0 \\ 0 & \frac{2}{3} & 1 & 0 & 0 \\ 0 & 0 & \frac{3}{4} & 1 & 0 \\ 0 & 0 & 0 & \frac{4}{5} & 1 \end{pmatrix}$
Steg 5:	$\begin{pmatrix} 2 & 1 & 0 & 0 & 0 \\ 0 & \frac{3}{2} & 1 & 0 & 0 \\ 0 & 0 & \frac{4}{3} & 1 & 0 \\ 0 & 0 & 0 & \frac{5}{4} & 1 \\ 0 & 0 & 0 & 0 & \frac{6}{5} \end{pmatrix}$	

Alltså är

$$\begin{pmatrix} 2 & 1 & 0 & 0 & 0 \\ 1 & 2 & 1 & 0 & 0 \\ 0 & 1 & 2 & 1 & 0 \\ 0 & 0 & 1 & 2 & 1 \\ 0 & 0 & 0 & 1 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 1 & 0 & 0 & 0 \\ 0 & \frac{2}{3} & 1 & 0 & 0 \\ 0 & 0 & \frac{3}{4} & 1 & 0 \\ 0 & 0 & 0 & \frac{4}{5} & 1 \end{pmatrix} \begin{pmatrix} 2 & 1 & 0 & 0 & 0 \\ 0 & \frac{3}{2} & 1 & 0 & 0 \\ 0 & 0 & \frac{4}{3} & 1 & 0 \\ 0 & 0 & 0 & \frac{5}{4} & 1 \\ 0 & 0 & 0 & 0 & \frac{6}{5} \end{pmatrix}.$$

Uppgift 13. Kolonn 1: Vi måste byta plats på första och andra raden, vilket vi gör med permutationsmatrisen

$$\mathbf{P}_1 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

dvs

$$\mathbf{P}_1 \mathbf{A} = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & -1 \\ 1 & 5 & 2 \end{pmatrix}.$$

Multiplikatorerna ges av

$$m_{2,1} = m_{3,1} = \frac{1}{2}.$$

Gausstransformationen för första kolonnen är således

$$\mathbf{L}_1 = \begin{pmatrix} 1 & 0 & 0 \\ 1/2 & 1 & 0 \\ 1/2 & 0 & 1 \end{pmatrix}$$

Notera att

$$\mathbf{P}_1^{-1} = \mathbf{P}_1 \quad \text{och} \quad \mathbf{L}_1^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ -1/2 & 1 & 0 \\ -1/2 & 0 & 1 \end{pmatrix}$$

Det ger att

$$\mathbf{A}_1 = \mathbf{L}_1^{-1} \mathbf{P}_1 \mathbf{A} = \begin{pmatrix} 2 & 1 & 1 \\ 0 & 3/2 & -3/2 \\ 0 & 9/2 & 3/2 \end{pmatrix}.$$

Kolonn 2: Vi byter plats på andra och tredje raden i \mathbf{A}_1 med permutationsmatrisen

$$\mathbf{P}_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix},$$

dvs

$$\mathbf{P}_2 \mathbf{A}_1 = \begin{pmatrix} 2 & 1 & 1 \\ 0 & 9/2 & 3/2 \\ 0 & 3/2 & -3/2 \end{pmatrix}.$$

Multiplikatorn och Gausstransformationen är därmed

$$m_{3,2} = \frac{3/2}{9/2} = \frac{1}{3} \quad \text{respektive} \quad \mathbf{L}_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1/3 & 1 \end{pmatrix},$$

där

$$\mathbf{L}_2^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1/3 & 1 \end{pmatrix}.$$

Matrisen \mathbf{P}_2 är sin egen invers. Då följer att

$$\mathbf{U} = \mathbf{L}_2^{-1} \mathbf{P}_2 \mathbf{A}_1 = \begin{pmatrix} 2 & 1 & 1 \\ 0 & 9/2 & 3/2 \\ 0 & 0 & -2 \end{pmatrix}.$$

Sammanfattningsvis har vi att

$$L_2^{-1}P_2L_1^{-1}P_1A = U \Leftrightarrow P_1A = L_1P_2L_2U.$$

Men $L_1P_2L_2$ är inte en undertriangulär enhetsmatris på grund av P_2 . Genom att multiplicera båda led med P_2 , dvs

$$P_2P_1A = P_2L_1P_2L_2U,$$

finner vi att

$$P = P_2P_1 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \quad \text{och} \quad L = P_2L_1P_2L_2 = \begin{pmatrix} 1 & 0 & 0 \\ 1/2 & 1 & 0 \\ 1/2 & 1/3 & 1 \end{pmatrix}$$

är de matriser vi söker. Notrera att utan pivotering erhåller man

$$\begin{pmatrix} 1 & 2 & -1 \\ 2 & 1 & 1 \\ 1 & 5 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & -1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & -1 \\ 0 & -3 & 3 \\ 0 & 0 & 6 \end{pmatrix},$$

dvs permutationsmatrisen är lika med enhetsmatrisen.

Uppgift 17. (a) Låt

$$\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix} \quad \text{och} \quad \mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix}.$$

Då är

$$\mathbf{v}^* \mathbf{u} = (\bar{v}_1 \quad \bar{v}_2 \quad \cdots \quad \bar{v}_n) \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix} = u_1\bar{v}_1 + u_2\bar{v}_2 + \cdots + u_n\bar{v}_n = (\mathbf{u}, \mathbf{v}).$$

(b) Från övningsuppgift 16 följer det att

$$(\mathbf{u}, \mathbf{A}^* \mathbf{v}) = (\mathbf{A}^* \mathbf{v})^* \mathbf{u} = \mathbf{v}^* (\mathbf{A}^*)^* \mathbf{u} = \mathbf{v}^* (\mathbf{A} \mathbf{u}) = (\mathbf{A} \mathbf{u}, \mathbf{v}).$$

(c) Följer direkt från föregående deluppgift och definitionen av hermitesk matris.

(d) Låt $\lambda \in \mathbb{C}$ vara ett egenvärde till \mathbf{A} och låt $\mathbf{v} \neq \mathbf{0}$ vara en egenvektor som hör till λ , dvs $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$. Då gäller att

$$(\mathbf{A}\mathbf{v}, \mathbf{v}) = (\lambda\mathbf{v}, \mathbf{v}) = \lambda(\mathbf{v}, \mathbf{v}) = \lambda\|\mathbf{v}\|^2.$$

och

$$(\mathbf{v}, \mathbf{A}\mathbf{v}) = (\mathbf{v}, \lambda\mathbf{v}) = \bar{\lambda}(\mathbf{v}, \mathbf{v}) = \bar{\lambda}\|\mathbf{v}\|^2.$$

Eftersom \mathbf{A} är hermitesk så är

$$\lambda\|\mathbf{v}\|^2 = (\mathbf{A}\mathbf{v}, \mathbf{v}) = (\mathbf{v}, \mathbf{A}\mathbf{v}) = \bar{\lambda}\|\mathbf{v}\|^2 \Leftrightarrow (\lambda - \bar{\lambda})\|\mathbf{v}\|^2 = 0.$$

Men från $\mathbf{v} \neq \mathbf{0}$ följer att då måste $\lambda = \bar{\lambda}$, dvs λ är reell. (e) Vi har att

$$(\mathbf{A}^* \mathbf{A})^* = \mathbf{A}^* (\mathbf{A}^*)^* = \mathbf{A}^* \mathbf{A},$$

vilket skulle visas. (f) Låt λ vara ett egenvärde till $\mathbf{A}^* \mathbf{A}$ med $\mathbf{v} \neq \mathbf{0}$ som en egenvektor.

Eftersom $\mathbf{A}^* \mathbf{A}$, så vet vi att λ är reell. Vidare är

$$\lambda\|\mathbf{v}\|^2 = \lambda(\mathbf{v}, \mathbf{v}) = (\lambda\mathbf{v}, \mathbf{v}) = (\mathbf{A}^* \mathbf{A}\mathbf{v}, \mathbf{v}) = (\mathbf{A}\mathbf{v}, \mathbf{A}\mathbf{v}) = \|\mathbf{A}\mathbf{v}\|^2,$$

dvs $\lambda = \|\mathbf{A}\mathbf{v}\|^2/\|\mathbf{v}\|^2 \geq 0$, vilket skulle visas.

Uppgift 18. Tag $a \in \mathbb{R}$ och $\mathbf{u}, \mathbf{v} \in V$ godtyckligt. Vi vill visa att $a\mathbf{v} \in V$ och $\mathbf{u} + \mathbf{v} \in V$. Eftersom $\mathbf{u}, \mathbf{v} \in V$, så finns det $\mathbf{x}, \mathbf{y} \in \mathbb{R}^m$ sådana att $\mathbf{Ax} = \mathbf{u}$ och $\mathbf{Ay} = \mathbf{v}$. Det ger att

$$a\mathbf{v} = a\mathbf{Ay} = \mathbf{A}(a\mathbf{y}) \in V \quad \text{och} \quad \mathbf{u} + \mathbf{v} = \mathbf{Ax} + \mathbf{Ay} = \mathbf{A}(\mathbf{x} + \mathbf{y}) \in V,$$

vilket skulle visas.

Uppgift 19. Att \mathbf{u} och \mathbf{v} är ortogonala betyder att $\mathbf{u} \cdot \mathbf{v} = 0$. Eftersom $\|\mathbf{u}\|^2 = \mathbf{u} \cdot \mathbf{u}$ för alla vektorer \mathbf{u} , så gälelr att

$$\|\mathbf{u} + \mathbf{v}\|_2^2 = (\mathbf{u} + \mathbf{v}) \cdot (\mathbf{u} + \mathbf{v}) = \mathbf{u} \cdot \mathbf{u} + 2\mathbf{u} \cdot \mathbf{v} + \mathbf{v} \cdot \mathbf{v} = \|\mathbf{u}\|_2^2 + \|\mathbf{v}\|_2^2,$$

vilket skulle visas.

Uppgift 20. Låt $\mathbf{B} = \mathbf{A}^T \mathbf{A}$. Vi vill visa att $\mathbf{B}^T = \mathbf{B}$. Räknelagarna för transponat ger att

$$\mathbf{B}^T = (\mathbf{A}^T \mathbf{A})^T = \mathbf{A}^T (\mathbf{A}^T)^T = \mathbf{A}^T \mathbf{A} = \mathbf{B},$$

vilket skulle visas.

Uppgift 21. (a) Vi bestämmer a och b genom att lösa ekvationssystemet

$$\begin{cases} a \sum_{i=1}^n x_i^2 + b \sum_{i=1}^n x_i = \sum_{i=1}^n x_i y_i \\ a \sum_{i=1}^n x_i + b n = \sum_{i=1}^n y_i. \end{cases}$$

Från given data följer att $n = 5$ samt att

$$\begin{aligned} \sum_{i=1}^5 x_i &= (-2) + (-1) + 0 + 1 + 2 = 0, \\ \sum_{i=1}^5 x_i^2 &= (-2)^2 + (-1)^2 + 0^2 + 1^2 + 2^2 = 10, \\ \sum_{i=1}^5 y_i &= 1 + 2 + 3 + 3 + 4 = 13, \\ \sum_{i=1}^5 x_i y_i &= (-2) \cdot 1 + (-1) \cdot 2 + 0 \cdot 3 + 1 \cdot 3 + 2 \cdot 4 = 7. \end{aligned}$$

Det ger oss ekvationssystemet

$$\begin{cases} 10a + 0b = 7 \\ 0a + 5b = 13 \end{cases} \Leftrightarrow \begin{cases} 10a = 7 \\ 5b = 13 \end{cases} \Leftrightarrow \begin{cases} a = 0.7 \\ b = 2.6. \end{cases}$$

Den sökta linjen är $y = f(x) = 0.7x + 2.6$. För bestämma det kvadratiska medelvärdesfelet beräknar vi först $f(x_k)$ för alla $k = 1, 2, 3, 4, 5$. Vi får att

$$f(x_1) = 1.2, f(x_2) = 1.9, f(x_3) = 2.6, f(x_4) = 3.3 \quad \text{och} \quad f(x_5) = 4.0.$$

Därmed är

$$\begin{aligned} E_2(f) &= \left(\frac{1}{5} \sum_{k=1}^5 (f(x_k) - y_k)^2 \right)^{1/2} \\ &= \frac{1}{\sqrt{5}} \left((1.2 - 1)^2 + (1.9 - 2)^2 + (2.6 - 3)^2 \right. \\ &\quad \left. + (3.3 - 3)^2 + (4.0 - 4)^2 \right)^{1/2} \approx 0.244949. \end{aligned}$$

(b) En alternativ metod till den i föregående deluppgift är att först härleda det ekvationssystem som genereras av $f(x_k) = y_k$ för $k = 1, 2, 3, 4, 5$. Vi får att

$$\begin{cases} -6a + b = 7 \\ -2a + b = 5 \\ b = 3 \\ 2a + b = 2 \\ 3a + b = 0 \end{cases} \Leftrightarrow \begin{pmatrix} -6 & 1 \\ -2 & 1 \\ 0 & 1 \\ 2 & 1 \\ 3 & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 7 \\ 5 \\ 3 \\ 2 \\ 0 \end{pmatrix} \Leftrightarrow \mathbf{Ax} = \mathbf{y}.$$

Normalekvationerna ges av $\mathbf{A}^T \mathbf{A} \mathbf{x} = \mathbf{A}^T \mathbf{y}$, dvs

$$\begin{pmatrix} -6 & -2 & 0 & 2 & 6 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} -6 & 1 \\ -2 & 1 \\ 0 & 1 \\ 2 & 1 \\ 3 & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} -6 & -2 & 0 & 2 & 6 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 7 \\ 5 \\ 3 \\ 2 \\ 0 \end{pmatrix} \Leftrightarrow \begin{pmatrix} 80 & 0 \\ 0 & 5 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} -48 \\ 17 \end{pmatrix},$$

som har lösningen $(a, b) = (-0.6, 3.4)$. Alltså är $y = f(x) = -0.6x + 3.4$. På samma sätt som i (a) får vi att $E_2(f) \approx 0.282843$. (c) Detaljerna lämnas som övning. Du ska erhålla linjen $y = f(x) = 0.7x - 0.2$ och det kvadratiska medelvärdesfelet $E_2(f) \approx 0.141421$.

Uppgift 22. Lokheterna $f(x_i) = y_i$, där $i = 1, 2, \dots, n$, motsvarar ett linjärt ekationsystem $\mathbf{Ax} = \mathbf{y}$, dvs

$$\begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \dots & \dots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}.$$

Vi har att

$$\mathbf{A}^T \mathbf{A} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \dots & \dots \\ 1 & x_n \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n 1 & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix}$$

och

$$\mathbf{A}^T \mathbf{y} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{pmatrix}.$$

Notera att

$$\sum_{i=1}^n 1 = n.$$

Lösningen till ekvationssystemet $\mathbf{A}^T \mathbf{A} \mathbf{x} = \mathbf{A}^T \mathbf{y}$ finner vi tex med Gausselimination och bakåtsubstitution. Det ger att

$$b = \frac{\sum_{i=1}^n x_i y_i - \frac{1}{n} \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2}$$

och

$$a = \frac{1}{n} \left(\sum_{i=1}^n y_i - b \sum_{i=1}^n x_i \right).$$

DEssa formler för a och b kanske du känner igen från teorin för linjär regression.

Uppgift 23. Sätt

$$E(a) = \sum_{k=1}^5 (f(x_k) - y_k)^2 = \sum_{k=1}^5 (ax_k - y_k)^2.$$

Då är

$$\frac{dE}{da} = \frac{d}{da} \sum_{k=1}^5 (ax_k - y_k)^2 = \sum_{k=1}^5 \frac{d}{da} (ax_k - y_k)^2 = \sum_{k=1}^5 2x_k(ax_k - y_k).$$

Vi vill minimera E och löser därför ekvationen

$$\frac{dE}{da} = 0 \Leftrightarrow 2 \sum_{k=1}^5 (ax_k^2 - x_k y_k) = 0 \Leftrightarrow a \sum_{k=1}^5 x_k^2 = \sum_{k=1}^5 x_k y_k.$$

Alltså är

$$a = \sum_{k=1}^5 x_k y_k \Big/ \sum_{k=1}^5 x_k^2.$$

(a) Vi har att

$$\sum_{k=1}^5 x_k^2 = (-4)^2 + (-1)^2 + 0^2 + 2^2 + 3^2 = 30$$

och

$$\sum_{k=1}^5 x_k y_k = (-4) \cdot (-3) + (-1) \cdot (-1) + 0 \cdot 0 + 2 \cdot 1 + 3 \cdot 2 = 21.$$

Det ger att $a = 22/30 = 0.7$ och $f(x) = 0.7x$. Då är

$$f(x_1) = -2.8, f(x_2) = -0.7, f(x_3) = 0, f(x_4) = 1.4 \quad \text{och} \quad f(x_5) = 2.1.$$

Det kvadratiska medelvärdesfelet är

$$\begin{aligned} E_2(f) &= \left(\frac{1}{5} \sum_{k=1}^5 (f(x_k) - y_k)^2 \right)^{1/2} \\ &= \frac{1}{\sqrt{5}} \left((-2.8 - (-3))^2 - (-0.7 - (-1))^2 - (0 - 0)^2 \right. \\ &\quad \left. - (1.4 - 1)^2 - (2.1 - 2)^2 \right)^{1/2} \approx 0.244949. \end{aligned}$$

(b) På samma sätt som ovan får man att $f(x) = 0.574x$ och $E_2(f) \approx 0.0756307$. Detaljerna lämnas som övning. (c) Du ska erhålla $f(x) = 1.58x$ och $E_2(f) \approx 0.172047$.

Uppgift 27. De givna punkterna ger oss ekvationssystemet

$$\begin{cases} p(-2) = 1 \\ p(-1) = 1 \\ p(0) = 0 \\ p(1) = 2 \\ p(2) = 1 \end{cases} \Leftrightarrow \begin{cases} 4a - 2b + c = 1 \\ a - b + c = 1 \\ c = 0 \\ a + b + c = 2 \\ 4a + 2b + c = 1. \end{cases}$$

Sätt

$$\mathbf{A} = \begin{pmatrix} 4 & -2 & 1 \\ 1 & -1 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \\ 4 & 2 & 1 \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} a \\ b \\ c \end{pmatrix} \quad \text{och} \quad \mathbf{y} = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 2 \\ 1 \end{pmatrix}.$$

Då ges ovanstående ekvationssystem av $\mathbf{Ax} = \mathbf{y}$. Normalekvationerna är

$$\mathbf{A}^t \mathbf{Ax} = \mathbf{A}^t \mathbf{y} \quad \Leftrightarrow \quad \begin{pmatrix} 34 & 0 & 10 \\ 0 & 10 & 0 \\ 10 & 0 & 5 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} 11 \\ 1 \\ 5 \end{pmatrix}.$$

som har lösningen $(a, b, c) = (1/14, 1/10, 6/7)$, vilket vi bestämmer tex med Gausselimination och bakåtsubstitution. *Alternativ lösning* Sätt

$$E(a, b, c) = \sum_{i=1}^5 (p(x_i) - y_i)^2 = \sum_{i=1}^5 (ax_i^2 + bx_i + c - y_i)^2,$$

där (x_i, y_i) är den i :te punkten given i uppgiften. Då är

$$\begin{aligned} \frac{\partial E}{\partial a} &= 2 \sum_{i=1}^5 x_i^2 (ax_i^2 + bx_i + c - y_i) \\ \frac{\partial E}{\partial b} &= 2 \sum_{i=1}^5 x_i (ax_i^2 + bx_i + c - y_i) \end{aligned}$$

$$\frac{\partial E}{\partial c} = 2 \sum_{i=1}^5 (ax_i^2 + bx_i + c - y_i).$$

Vi sätter de tre partiella derivatorna lika med 0 och erhåller efter omskrivning ekvationssystemet

$$\begin{cases} a \sum x_i^4 + b \sum x_i^3 + c \sum x_i^2 = \sum x_i^2 y_i \\ a \sum x_i^3 + b \sum x_i^2 + c \sum x_i = \sum x_i y_i \\ a \sum x_i^2 + b \sum x_i + c \sum 1 = \sum y_i, \end{cases} \quad (6.7)$$

där varje summa tas över $i = 1, 2, 3, 4, 5$. Vi har att

$$\begin{aligned} \sum x_i^4 &= (-2)^4 + (-1)^4 + 0^4 + 1^4 + 2^4 = 34 \\ \sum x_i^3 &= (-2)^3 + (-1)^3 + 0^3 + 1^3 + 2^3 = 0 \\ \sum x_i^2 &= (-2)^2 + (-1)^2 + 0^2 + 1^2 + 2^2 = 10 \\ \sum x_i &= (-2) + (-1) + 0 + 1 + 2 = 0 \\ \sum 1 &= 1 + 1 + 1 + 1 + 1 = 5 \\ \sum x_i^2 y_i &= (-2)^2 \cdot 1 + (-1)^2 \cdot 1 + 0^2 \cdot 0 + 1^2 \cdot 2 + 2^2 \cdot 1 = 11 \\ \sum x_i y_i &= (-2) \cdot 1 + (-1) \cdot 1 + 0 \cdot 0 + 1 \cdot 2 + 2 \cdot 1 = 1 \\ \sum y_i &= 1 + 1 + 0 + 2 + 1 = 5. \end{aligned}$$

Insättning i (6.7) ger oss ekvationssystemet

$$\begin{cases} 34a + 10c = 11 \\ 10b = 1 \\ 10a + 5c = 5. \end{cases}$$

Jämför med den första lösningen.

Uppgift 28. Ansätt $f(x) = ax^2 + bx + c$. Likheterna $f(x_k) = y_k$ för $k = 1, 2, \dots, 5$ motsvarar ekvationssystemet

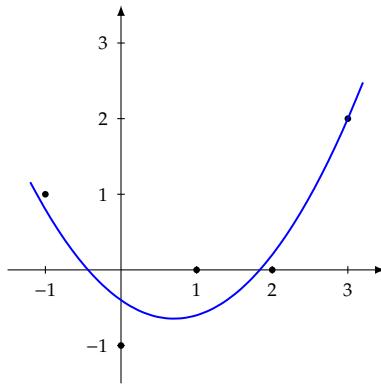
$$\begin{cases} a - b + c = 1 \\ c = -1 \\ a + b + c = 0 \\ 4a + 2b + c = 0 \\ 9a + 3b + c = 2, \end{cases}$$

vilket kan skrivas på matrisform enligt $\mathbf{Ax} = \mathbf{b}$, där

$$\mathbf{A} = \begin{pmatrix} 1 & -1 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \\ 4 & 2 & 1 \\ 9 & 3 & 1 \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} a \\ b \\ c \end{pmatrix} \quad \text{och} \quad \mathbf{b} = \begin{pmatrix} 1 \\ -1 \\ 0 \\ 0 \\ 2 \end{pmatrix}$$

Då är

$$\mathbf{A}^T \mathbf{A} = \begin{pmatrix} 99 & 35 & 15 \\ 35 & 15 & 5 \\ 15 & 5 & 5 \end{pmatrix} \quad \text{och} \quad \mathbf{A}^T \mathbf{b} = \begin{pmatrix} 19 \\ 5 \\ 2 \end{pmatrix}$$



Figur L.3

Normalekvationerna ges således av

$$\mathbf{A}^T \mathbf{A} \mathbf{x} = \mathbf{A}^T \mathbf{b} \Leftrightarrow \begin{cases} 99 + 35 + 15 = 19 \\ 35 + 15 + 5 = 5 \\ 15 + 5 + 5 = 2. \end{cases}$$

Med tex Gausselimination följt av bakåtsubsitution får vi att detta ekvationsstsrem har lösningen

$$a = \frac{1}{2}, \quad b = -\frac{7}{10} \quad \text{och} \quad c = -\frac{2}{5}.$$

Det andragradspolynom som bäst anpassar till de givna punkterna är

$$f(x) = \frac{1}{2}x^2 - \frac{7}{10}x - \frac{2}{5},$$

se figur L.3.

Uppgift 29. Sätt

$$A_1 = \sum_{i=1}^n x_i^2, \quad A_2 = \sum_{i=1}^n x_i, \quad B_1 = \sum_{i=1}^n x_i y_i \quad \text{och} \quad B_2 = \sum_{i=1}^n y_i.$$

Då är $D = nA_1 - A_2^2$ och

$$\begin{cases} a \sum_{i=1}^n x_i^2 + b \sum_{i=1}^n x_i = \sum_{i=1}^n x_i y_i \\ a \sum_{i=1}^n x_i + bn = \sum_{i=1}^n y_i \end{cases} \Leftrightarrow \begin{cases} A_1 a + A_2 b = B_1 \\ A_2 a + nb = B_2. \end{cases}$$

Gausselimination ger

$$\begin{cases} A_1 a + A_2 b = B_1 \\ A_2 a + nb = B_2 \end{cases} \times A_1 \Leftrightarrow \begin{cases} A_1 a + A_2 b = B_1 \\ A_1 A_2 a + n A_1 b = A_1 B_2 \end{cases} \stackrel{-A_2}{\Leftrightarrow}$$

$$\begin{cases} A_1a + A_2b = B_1 \\ (nA_1 - A_2^2)b = A_1B_2 - A_2B_1 \end{cases} \Leftrightarrow \begin{cases} A_1a + A_2b = B_1 \\ Db = A_1B_2 - A_2B_1. \end{cases}$$

Bakåtsubstitution ger att

$$b = \frac{1}{D}(A_1B_2 - A_2B_1) = \frac{1}{D} \left(\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i - \sum_{i=1}^n x_i \sum_{i=1}^n x_i y_i \right)$$

och

$$\begin{aligned} a &= \frac{1}{A_1}(B_1 - A_2b) = \frac{1}{A_1} \left(B_1 - \frac{A_2}{D}(A_1B_2 - A_2B_1) \right) \\ &= \frac{1}{A_1D} (B_1D - A_2(A_1B_2 - A_2B_1)) \\ &= \frac{1}{A_1D} (B_1(nA_1 - A_2^2) - A_1A_2B_2 + A_2^2B_1) = \frac{1}{D}(nB_1 - A_2B_2) \\ &= \frac{1}{D} \left(n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i \right), \end{aligned}$$

vilket skulle visas.

Uppgift 30. (a) Sätt

$$E(a, b, c) = \sum_{k=1}^4 (f(x_k) - y_k)^2 = \sum_{k=1}^4 (ax_k^2 + bx_k + c - y_k)^2.$$

För att minimera E bestämmer vi samtliga partiella derivator och sätter dessa lika med 0. Från

$$\begin{aligned} \frac{\partial E}{\partial a} &= 2 \sum_{k=1}^4 x_k^2 (ax_k^2 + bx_k + c - y_k), \\ \frac{\partial E}{\partial b} &= 2 \sum_{k=1}^4 x_k (ax_k^2 + bx_k + c - y_k), \\ \frac{\partial E}{\partial c} &= 2 \sum_{k=1}^4 (ax_k^2 + bx_k + c - y_k). \end{aligned}$$

följer ekvationssystemet

$$\begin{cases} a \sum_{k=1}^4 x_k^4 + b \sum_{k=1}^4 x_k^3 + c \sum_{k=1}^4 x_k^2 = \sum_{k=1}^4 x_k^2 y_k \\ a \sum_{k=1}^4 x_k^3 + b \sum_{k=1}^4 x_k^2 + c \sum_{k=1}^4 x_k = \sum_{k=1}^4 x_k y_k \\ a \sum_{k=1}^4 x_k^2 + b \sum_{k=1}^4 x_k + cn = \sum_{k=1}^4 y_k. \end{cases}$$

För den data vi har given får vi ekvationssystemet

$$\begin{cases} 164a + 20c = 186 \\ 20b = -34 \\ 20a + 4c = 26 \end{cases}$$

Gausselimination följt av bakåtsubstitution ger lösningen

$$a = \frac{7}{8}, \quad b = -\frac{17}{10} \quad \text{och} \quad c = \frac{17}{8}.$$

Alltså är

$$f(x) = \frac{7}{8}x^2 - \frac{17}{10}x + \frac{17}{8}.$$

(b) Notera att vi har samma värden på x_k för alla k som deluppgift (a), dvs vänsterledet i ekvationssystemet är detsamma. Vi ska studera en alternativ metod. Varje $f(x_k) = y_k$ motsvarar en ekvation med a , b och c som obekanta, dvs

$$\begin{cases} 9a - 3b + c = -1 \\ a - b + c = 25 \\ a + b + c = 25 \\ 9a + 3b + c = 1 \end{cases} \Leftrightarrow \begin{pmatrix} 9 & -3 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & 1 \\ 9 & 3 & 1 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} -1 \\ 25 \\ 25 \\ 1 \end{pmatrix},$$

eller på matrisform $\mathbf{A}\mathbf{x} = \mathbf{y}$. Normalekvationerna $\mathbf{A}^T \mathbf{A}\mathbf{x} = \mathbf{A}^T \mathbf{y}$ ges av

$$\begin{cases} 164a + 20c = 50 \\ 20b = 6 \\ 20a + 4c = 50, \end{cases}$$

som har lösningen

$$a = -\frac{25}{8}, \quad b = \frac{3}{10} \quad \text{och} \quad c = \frac{225}{8}.$$

Alltså är

$$f(x) = -\frac{25}{8}x^2 + \frac{3}{10}x + \frac{225}{8}.$$

Uppgift 34. Vi har att

$$y = \frac{x}{a+bx} \Leftrightarrow \frac{1}{y} = \frac{a+bx}{x} = \frac{a}{x} + b.$$

Sätt $X = 1/x$ och $Y = 1/y$. Då är

$$y = \frac{x}{a+bx} \Leftrightarrow Y = AX + B,$$

där $A = a$ och $B = b$.

6 Ordinära differentialekvationer

Uppgift 1. (a) För steget $h = 0.1$ får vi

$$D_0(0.1) = \frac{f(0.8 + 0.1) - f(0.8 - 0.1)}{2 \cdot 0.1} \approx 0.695546112.$$

På samma sätt får vi att $D_0(0.01) \approx 0.6966951$ och $D_0(0.001) \approx 0.696707$. (b) Absolutfelet för steget $h = 0.1$ är

$$|D_0(0.1) - \cos(0.8)| \approx |0.695546112 - 0.6967067093| \approx 0.001160597.$$

För de två andra stegen är felet 0.0000116 respektive 0.0000116. (c) Trunkeringsfelet ges av

$$E_T = -\frac{h^2 f^{(3)}(c)}{6}.$$

För $h = 0.1$ får vi att

$$|E_T| \leq \frac{0.1^2 \cdot 0.764842187}{6} \approx 0.00127474.$$

Övre gräns för trunkeringsfelet för de två andra stegen finner man på samma sätt, som ger värdena 0.0000127 respektive 0.000000127.

Uppgift 3. Notera att $V(10) = D'(10)$. Centraldifferens med steget $h = 1$ ger att

$$D_0(1) = \frac{D(10 + 1) - D(10 - 1)}{2 \cdot 1} = 4.4205$$

och

$$D_1(1) = \frac{-D(10 + 2) + 8D(10 + 1) - 8D(10 - 1) + D(10 - 2)}{12 \cdot 1} = 4.42475.$$

Deriverar vi det uttryck för D som ges i uppgiften får vi att

$$D'(t) = 7 - 7e^{-t/10}$$

och $D'(10) \approx 4.42484$.

Uppgift 6. (a) Studera nedanstående uppställning.

$D_0(2^4 h)$					
$D_0(2^3 h)$	$D_1(2^3 h)$				
$D_0(2^2 h)$	$D_1(2^2 h)$	$D_2(2^2 h)$			
$D_0(2 h)$	$D_1(2 h)$	$D_2(2 h)$	$D_3(2 h)$		
$D_0(h)$	$D_1(h)$	$D_2(h)$	$D_3(h)$	$D_4(h)$	

Vi söker $D_4(h)$. Med andra ord är det elementen i första kolumnen som efterfrågas i uppgiften. (b) Notera att

$$\begin{aligned} D_0(h) &= D_0(0.01), & D_0(2h) &= D_0(0.02), & D_0(2^2 h) &= D_0(0.04), \\ D_0(2^3 h) &= D_0(0.08) \quad \text{och} \quad D_0(2^4 h) &= D_0(0.16). \end{aligned}$$

Vi har bla att

$$D_0(0.01) = \frac{f(1.2 + 0.01) - f(1.2 - 0.01)}{2 \cdot 0.01}.$$

På samma sätt bestämmer vi samtliga element i första kolumnen i tabellen från (a). De övriga kolumnerna fås nu i tur och ordning med

$$D_1(2^k h) = D_0(2^k h) + \frac{D_0(2^k h) - D_0(2^{k+1} h))}{3}$$

$$\begin{aligned} D_2(2^k h) &= D_1(2^k h) + \frac{D_1(2^k h) - D_1(2^{k+1} h)}{15} \\ D_3(2^k h) &= D_2(2^k h) + \frac{D_2(2^k h) - D_2(2^{k+1} h)}{63} \\ D_4(2^k h) &= D_3(2^k h) + \frac{D_3(2^k h) - D_3(2^{k+1} h)}{255}. \end{aligned}$$

Vi har tex att

$$D_2(2h) = D_2(0.02) = D_1(0.02) + \frac{D_1(0.02) - D_1(0.04)}{15}.$$

Till slut får vi följande tabell.

4.63028					
4.62893	4.62848605				
4.62859	4.62848526	4.62848521700			
4.62851	4.62848522	4.62848521738	4.62848521739038		
4.62849	4.62848521	4.62848521739	4.62848521739030	4.6284852173903	

Värdena är trunkerade efter första decimal som ger olika element i respektive kolumn.
(c) Derivatan är

$$f'(x) = e^{\sqrt{x}} + xe^{\sqrt{x}} \frac{1}{2\sqrt{x}} = \left(1 + \frac{\sqrt{x}}{2}\right)e^{\sqrt{x}}.$$

Alltså är $f'(1.2) \approx 4.62849$. Det absokuta felet är

$$E_x = |f'(1.2) - D_4(0.01)| = -7.99361 \cdot 10^{-15}.$$

Uppgift 7. (a) Först deriverar vi y , dvs bestämmer vänsterledet i differentialekvationen. Vi får att

$$y'(t) = 3Ce^{3t} - 1.$$

Högerledet är

$$f(t, y) = 3y + 3t = 3\left(Ce^{3t} - t - \frac{1}{3}\right) + 3t = 3Ce^{3t} - 1.$$

Det visar att funktionen uppfyller $y' = f(t, y)$. (b) Vi har att

$$|f(t, y) - f(t, z)| = |(3y + 3t) - (3z + 3t)| = 3|y - z|,$$

vilket visar att f uppfyller ett Lipschitzvillkor med Lipschitzkonstanten $L = 3$.

Uppgift 9. (a) Det följer från

$$|f(x, y) - f(x, z)| = \left| \frac{xy}{x+1} - \frac{xz}{x+1} \right| = \left| \frac{x}{x+1} \right| \cdot |y - z|$$

att

$$|f(x, y) - f(x, z)| \leq \frac{2}{3}|y - z|.$$

Uppgift 11. Sätt $f(t, y) = -ty$. Från begynnelsevärdesvillkoret $y(0) = 1$ följer att $t_0 = 0$ och $y_0 = 1$. Eulers metod ges av

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

där $t_k = t_0 + hk = hk$. För $h = 0.2$ får vi att

$$\begin{aligned} t_1 &= 0.2 & y_1 &= y_0 + hf(t_0, y_0) = 1 + 0.2(-0 \cdot 1) = 1 \\ t_2 &= 0.4 & y_2 &= y_1 + hf(t_1, y_1) = 1 + 0.2(-0.2 \cdot 1) = 0.96. \end{aligned}$$

Alltså är $y(0.4) \approx 0.96$. För $h = 0.1$ får vi att

$$\begin{aligned} t_1 &= 0.1 & y_1 &= y_0 + hf(t_0, y_0) = 1 + 0.1(-0 \cdot 1) = 1 \\ t_2 &= 0.2 & y_2 &= y_1 + hf(t_1, y_1) = 1 + 0.1(-0.1 \cdot 1) = 0.99 \\ t_3 &= 0.3 & y_3 &= y_2 + hf(t_2, y_2) = 0.99 + 0.1(-0.2 \cdot 0.99) = 0.9702 \\ t_4 &= 0.4 & y_4 &= y_3 + hf(t_3, y_3) = 0.9702 + 0.1(-0.3 \cdot 0.9702) = 0.941094. \end{aligned}$$

Alltså är $y(0.4) \approx 0.941094$. Den exakta lösningen ger $y(0.4) \approx 0.923116$.

Uppgift 15. Steglängen ges av $h = (3-0)/3 = 1$ och därmed är $x_n = n$, för $n = 0, 1, 2, 3$.

Sätt

$$f(x, y) = \frac{\cos y}{x^2 + 1} \quad \text{och} \quad y_0 = 1.$$

Då ger Heuns metod följande resultat.

x_n	k_1	k_2	y_n
0			1
1	0.540302	0.0152446	1.27777
2	0.144424	0.0296106	1.36479
3	0.040910	0.0164346	1.39346

Uppgift 16. Definiera funktionen $f(x, y) = e^{-2x} - 2y$. Från $y(0) = 0.1$ följer att $x_0 = 0$ och $y_0 = 0.1$. Heuns metod ges av

$$k_1 = f(x_k, y_k), \quad k_2 = f(x_k + h, y_k + hk_1) \quad \text{och} \quad y_{k+1} = y_k + \frac{h}{2}(k_1 + k_2),$$

där $x_k = x_0 + hk = hk$. För $h = 0.2$ får vi att

$$\begin{aligned} x_1 &= 0.2 & k_1 &= f(x_0, y_0) = 0.8 \\ && k_2 &= f(x_0 + h, y_0 + hk_1) \approx 0.15032 \\ && y_1 &= y_0 + \frac{h}{2}(k_1 + k_2) \approx 0.195032 \\ x_2 &= 0.4 & k_1 &= f(x_1, y_1) \approx 0.280256 \\ && k_2 &= f(x_1 + h, y_1 + hk_1) \approx -0.0528375 \\ && y_2 &= y_1 + \frac{h}{2}(k_1 + k_2) \approx 0.217774. \end{aligned}$$

Alltså är $y(0.4) \approx 0.217774$. På samma sätt som ovan får vi för $h = 0.1$ att

$$x_1 = 0.1 \quad y_1 \approx 0.162937$$

$$\begin{array}{ll} x_2 = 0.2 & y_2 \approx 0.199873 \\ x_3 = 0.3 & y_3 \approx 0.218149 \\ x_4 = 0.4 & y_4 \approx 0.223301. \end{array}$$

Alltså är $y(0.4) \approx 0.223301$. Den exakta lösningen ger $y(0.4) \approx 0.224664$.

Uppgift 17. Låt $f(x, y) = xy - 2x$. Från begynnelsevillkoret $y(0) = 1.5$ följer att $x_0 = 0$ och $y_0 = 1.5$. Heuns metod ges av

$$k_1 = f(x_k, y_k), \quad k_2 = f(x_k + h, y_k + hk_1) \quad \text{och} \quad y_{k+1} = y_k + \frac{h}{2}(k_1 + k_2),$$

där $x_k = t_0 + hk = hk$. För $h = 0.2$ får vi att

$$\begin{aligned} x_1 &= 0.2 & k_1 &= f(x_0, y_0) = 0 \\ && k_2 &= f(x_0 + h, y_0 + hk_1) = -0.1 \\ && y_1 &= y_0 + \frac{h}{2}(k_1 + k_2) = 1.49 \\ x_2 &= 0.4 & k_1 &= f(x_1, y_1) \approx -0.102 \\ && k_2 &= f(x_1 + h, y_1 + hk_1) \approx -0.21216 \\ && y_2 &= y_1 + \frac{h}{2}(k_1 + k_2) \approx 1.45858 \\ x_3 &= 0.6 & k_1 &= f(x_2, y_2) \approx -0.216566 \\ && k_2 &= f(x_2 + h, y_2 + hk_1) \approx -0.350838 \\ && y_3 &= y_2 + \frac{h}{2}(k_1 + k_2) \approx 1.40184 \end{aligned}$$

Alltså är $y(0.6) \approx 1.40184$. Den exakta lösningen ger

$$y(0.6) = 2 - 0.5e^{0.6^2/2} \approx 1.40139.$$

Då är det absoluta felet $E = |1.40139 - 1.40184| = 0.000452$.

Uppgift 19. Sätt

$$f(x, y) = \frac{x - y}{1 + x^2}.$$

Då är

$$(1 + x^2)y' + y - x = 0 \quad \Leftrightarrow \quad y' = \frac{x - y}{1 + x^2} = f(x, y).$$

Låt $x_0 = 0$ och $y_0 = 0.8$. Då ges Heuns metod av

$$\begin{aligned} k_1 &= f(x_n, y_n) \\ k_2 &= f(x_{n+1}, y_n + 0.25k_1) \\ y_{n+1} &= y_n + \frac{0.25}{2}(k_1 + k_2), \end{aligned}$$

där $n = 0, 1, 2, 3$ och $x_{n+1} = x_n + 0.25$. Heuns metod:

n	k_1	k_2	x_{n+1}	y_{n+1}
0	-0.8	-0.329412	0.25	0.658824
1	-0.384775	-0.0501038	0.5	0.604464
2	-0.0835709	0.106515	0.75	0.607332
3	0.0913078	0.184921	1	0.64186

Alltså är $y(1) \approx 0.64186$.

Uppgift 20. Vi får i tur och ordning

$$\begin{array}{ll} x_0 = 0 & y_0 = 0.6 \\ x_1 = 0.25 & y_1 = 0.553738 \\ x_2 = 0.5 & y_2 = 0.572766 \\ x_3 = 0.75 & y_3 = 0.640615 \\ x_4 = 1 & y_4 = 0.742269, \end{array}$$

dvs $y(1.0) \approx 0.742269$.

Uppgift 21. Låt $\mathbf{y}_n = (y_{1,n}, y_{2,n}, \dots, y_{m,n})$. Runge-Kuttas metod för ett system av differentialekvationer av första ordningen ges av

$$\begin{aligned} x_n &= a + hn \\ k_1 &= f(x_n, \mathbf{y}_n) \\ k_2 &= f\left(x_n + \frac{h}{2}, \mathbf{y}_n + \frac{h}{2}k_1\right) \\ k_3 &= f\left(x_n + \frac{h}{2}, \mathbf{y}_n + \frac{h}{2}k_2\right) \\ k_4 &= f\left(x_n + h, \mathbf{y}_n + hk_3\right) \\ \mathbf{y}_{n+1} &= \mathbf{y}_n + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4), \end{aligned}$$

där $n = 0, 1, \dots, N$ och där $h = (b - a)/N$.

Litteraturförteckning

- [1] Tom M. Apostol, *Mathematical Analysis*, second edition, Addison-Wesley, Reading, Massachusetts, 1974.
- [2] William W. Hager, *Applied Numerical Linear Algebra*, Prentice Hall, Englewood Cliffs, 1988.
- [3] Lennart Hellström, Staffan Morander och Anders Tengstrand, *En variabelanalys*, Studentlitteratur, Lund, 1991.
- [4] John H. Mathews and Kurtis D. Fink, *Numerical Methods Using Matlab*, fourth edition, Pearson Education, Prentice Hall, London, 2004.
- [5] Jan Thompson, *Matematiken i historien*, Studentlitteratur, Lund, 1996.

Erkännande

Ett varmt tack till
David Danielsson
för hjälpen med korrekturläsningen.



That's all Folks!

Numerisk analys

Hörelässningssattekuniga!

Right Now