Läsanvisningar till Analysens grunder

1 Bevis och definitioner

Det viktigaste målet med kursen är att lära sig läsa och förstå matematik. Detta är ingen lätt sak och kursen betraktas som rätt "tung". En anledning till att det är svårt att lära sig läsa matematik är att matematisk notation utvecklats under en mycket lång tid med resultatet att det finns ett stort antal konventioner som är avsedda att hjälpa läsaren men som förstås inte gör det om denne inte känner till dem!

Det finns dock en viktigare anledning till att det är svårt. Problemet är att matematik inte handlar om något speciellt som man kan hänvisa till (som t ex de olika naturvetenskaperna). Av psykologiska/pedagogiska skäl är det å andra sidan viktigt ha sina egna bilder (som inte behöver vara visuella) av olika matematiska begrepp. Just för att det inte finns någon "verklighet" att hänvisa till blir dessa bilder mycket individuella och det är nästan omöjligt att använda sig av dem för att diskutera matematik med andra eller att skriva om matematik.

Lösningen på detta problem, framförallt när det gäller skriven matematik, är att man inskränker sig till att använda ett mycket precist och ganska formellt språk vars tolkning alla lär sig att vara överens om. Detta betyder att man för att kunna läsa matematik måste

- lära sig förstå det formella, precisa, gemensamma språket,
- skaffa sig sitt eget språk eller bilder samt
- lära sig översätta från det gemensamma formella till sitt eget och från sitt eget till det gemensamma.

Observera att det är absolut nödvändigt att skaffa sig sitt *eget* språk. Det betyder att man inte ska se lärarens försök att förklara en matematisk text som den "rätta" utan på sin höjd som ett tänkbart förslag till en sådan förklaring och man måste alltid komma på sin egen. I början kan detta vara rätt svårt eftersom man ännu inte hunnit skaffa sig sitt eget språk och måste utveckla det. Det enda sättet att utveckla ett sådant språk och samtidigt lära sig hur man gör översättning fram och tillbaka från detta och det gemensamma språket är att arbeta med matematiska definitioner, påståenden och bevis. I början får man räkna med att arbeta mycket med dessa; så mycket att man kommer att ha lärt sig dem utantill. Det är viktigt att vara klar över att detta inte är målet utan målet är just att utveckla sitt eget språk som man kan uttrycka allting i. När man gjort detta kommer man inte längre behöva kunna t ex ett bevis utantill.

Det är som sagt viktigt att vara klar över att det språk och de bilder man slutligen själv kommer att utveckla är ens eget och att andras förslag till bilder inte behöver passa en själv. Dock finns det några principer som verkar hjälpa de flesta att lära sig analysera definitioner, påståenden och bevis och vi ska här ta upp några sådana.

Det allra första man bör göra när man läser en matematisk definition eller påstående är att *förstå vad det säger*. Detta innebär t ex att man kontrollerar att man förstår alla termer som förekommer, att man är klar över vad som antas och vad som påstås vara sant under förutsättningarna. Det innebär också att man klargör för sig själv de exakta formuleringarna som förekommer; det är till exempel skillnad på "för alla x så finns det ett y" och "det finns ett y så att för alla x".

Om det sedan är en fråga om en definition så kan man fortsätta med att

- försöka hitta några exempel som uppfyller definitionens villkor,
- försöka hitta några exempel som *inte* uppfyller de definitionens villkor samt
- försöka förstå avsikten med definitionen.

Vad beträffar exempel och icke-exempel på en definition så är det viktigt att hitta sådana inte bara för att förstå en definition utan även för framtiden. När definitionen sedan används kanske som term i en ny definition, ett påstående eller ett bevis så gäller det ju att förstå detta nya. Det är då bra att kunna ha sina exempel på den urspungliga definitionen i åtanke när man försöker använda sig av definitionen. Det är därför viktigt att efter att man i första skedet lyckats hitta ett antal exempel välja ut några bra exempel. Vad som menas med bra exempel varierar förstås men det viktiga är å ena sidan att man inte har ett alltför stort antal exempel å andra sidan att de "täcker" alla aspekter av definitionen. Om man bara har ett exempel så riskerar man att fixera sig för mycket på detta enda exempel och riskerar att man i sin bild av det begrepp som definitionen definierar låter speciella egenskaper hos exemplet smyga in, egenskaper som inte finns med i definitionen. För många exempel riskerar å andra sidan att bli oanvändbara.

Icke-exempel är till för att markera gränserna för definitionen och det är därför viktigt att välja dem så nära något som verkligen uppfyller definitionen.

Slutligen, att förstå avsikten med definitionen är en ganska vag uppgift. Å ena sidan finns det ganska säkert (får man hoppas i varje fall) en sådan för annars skulle man inte bry sig om att införa definitionen. Å andra sidan behöver avsikten dock inte bli klar med detsamma. Det kan röra sig om att man först senare i texten har användning för det nya begreppet.

Låt oss titta på ett exempel på en definition. Vi börjar med att ge en version av den som har till avsikt att göra det så svårt som möjligt att förstå vad avsikten är.

Definition 1.1 *En* kropp *är en femtuppel* (K, p, m, a, b) *där*

- K *är en mängd,*
- p *är en funktion* $p\colon K \times K \to K$,
- m *är en funktion* $m\colon K \times K \to K$,
- a *och* b *är element i* K.

Dessa data ska uppfylla följande villkor för alla $x, y, z \in K$*:*

1. $p(x, y) = p(y, x)$.
2. $p(x, p(y, z)) = p(p(x, y), z)$.
3. $m(x, y) = m(y, x)$.
4. $m(x, m(y, z)) = m(m(x, y), z)$.
5. $m(p(x + y), z) = p(m(x, z), m(y, z))$.
6. $p(a, x) = x$.
7. $m(b, x) = x$.
8. *Det finns* x' *sådant att* $p(x, x') = a$.
9. *Om* $x \neq a$ *så finns det* x'' *sådant att* $m(x, x'') = a$.

Denna definition är svår att tränga igenom. Det är inte ens lätt att hitta exempel (icke-exempel är lättare!) som uppfyller den och avsikten verkar inte klar. Vi gör ett nytt försök där vi formellt ger precis samma definition men formulerad på ett sätt som gör den lättare att förstå.

Definition 1.2 *En* kropp *är en femtuppel* $(K, +, \cdot, 0, 1)$ *där*

- K *är en mängd,*
- $+$ *är en funktion* $+\colon K \times K \to K$,

- $\cdot$ *är en funktion* $\cdot\colon K \times K \to K$,
- 0 *och* 1 *är element i* K.

Dessa data ska uppfylla följande villkor för alla $x, y, z \in K$:

1. *(Kommutativitet för addition)* $x + y = y + x$.
2. *(Associativitet för addition)* $x + (y + z) = (x + y) + z$.
3. *(Kommutativitet för multiplikation)* $x \cdot y = y \cdot x$.
4. *(Associativitet för multiplikation)* $x \cdot (y \cdot z) = (x \cdot y) \cdot z$.
5. *(Distributivitet)* $(x + y) \cdot z = x \cdot z + y \cdot z$.
6. *(Enhetselement för addition)* $0 + x = x$.
7. *(Enhetselement för multiplikation)* $1 \cdot x = x$.
8. *(Additiv invers) Det finns* x' *sådant att* $x + x' = 0$.
9. *(Multiplikativ invers) Om* $x \neq 0$ *så finns det* x'' *sådant att* $x \cdot x'' = 1$.

Nu ser det betydligt mer förståeligt ut. Observera till att börja med att vi skrivit funktionsvärdena i så kallad *infix-form*, dvs det som egentligen borde vara $+(x, y)$ (precis som vi använde $p(x, y)$ i vår första definition) skriver vi $x+y$. Detta är ett exempel på en matematisk konvention; om man använder vissa speciella namn på funktioner som $+$ och $\cdot$ så kan man (och gör det nästan alltid) använda infix-notation.

Valet av just namnet '+' för den första funktionen, '$\cdot$' för den andra och 0 och 1 för de två elementen är heller ingen tillfällighet. Det är ju så att med dessa val blir det helt plötsligt mycket lätt att hitta exempel på en femtuppel som uppfyller dessa villkor. Vi har dessutom gjort detta ännu tydligare genom att skriva små kommentarer framför varje villkor.

Övning 1: i) Finn några exempel på kroppar.
ii) Finn femtuppler som uppfyller alla villkor utom

- det sista,
- de två sista,
- det näst sista,
- det tredje och det sista samt
- det tredje.

(Det sista fallet är svårt.)

Det finns fler exempel än dem man kanske först tänker på och då ett av dem är lite ovanligt är det bra att ha med bland sin lilla samling av exempel.

Övning 2: Låt K vara mängden $\{0, 1\}$ och definiera addition genom

$+$	0	1
0	0	1
1	1	0

och multiplikation genom

$\cdot$	0	1
0	0	0
1	0	1

.

Visa att femtuppeln $(K, +, \cdot, 0, 1)$ är en kropp.

Om man sedan funderar över avsikten med denna definition så är den nog ganska klar. Man kan sammanfatta den med att säga att en kropp är en mängd i vilken man har tillgång till de fyra räknesätten och alla de räknelagar för dem som vi är vana vid. Å andra sidan visar exemplet från den sista övningen att man får vara lite försiktig. Man kan till exempel inte utgå från att $1 + 1$ är skilt från 0.

Anmärkning: Det finns ytterligare en aspekt på syftet med en definition. Man vill också att definition ska vara användbar, att den ska ha intressanta konsekvenser. I fallet kroppar är det definitivt så. Till exempel så kan det mesta av linjär algebra (som man först möter bara över de reella, eller kanske komplexa, talen) är meningsfull och sann över en godtycklig kropp. Detta har många intressanta tillämpningar.

Vi ska nu titta på påståenden och deras bevis. För påståenden gäller precis som för definitioner att man först måste förstå vad de säger och det man behöver göra för detta är ungefär samma sak; förstå alla termer, reda ut vad som är antaganden och vad som är slutsatser osv. Nästa steg är att förstå vad de innebär men detta steg skiljer sig en hel del från motsvarande steg för definitioner. I början är det så inte så stor skillnad. Det kan vara bra att göra följande.

- Försök hitta några exempel på situationer där antagandena är uppfyllda och försök förstå vad påståendet då ger för konsekvenser i de fallen.

- Försök hitta några exempel på situationer där antagandena inte är uppfyllda (och för att detta steg ska vara nyttigt ska villkoren nästan vara uppfyllda). Kan man till och med komma på ett exempel där villkoren nästan är uppfyllda men slutsatsen är falsk så är det ännu bättre.

Vad man sedan kan försöka göra är att se om man kan se vad påståendet ska vara bra för. Detta kan vara rätt svårt och är kanske inte ens klart just när man läser det.

2 Ett analyserat bevis

Vi ska visa hur man använder sig av supremumaxiomet för att visa att det finns ett positivt reellt tal r sådant att $r^2 = 2$. Vi vill samtidigt visa hur ett bevis kan läggas upp. Precis som i "verkligheten" börjar vi därmed med ett lemma vars betydelse blir klart först lite senare.

Lemma 2.1 *i) Varje nedåt begränsad mängd har en största undre begränsning.*
ii) Om r och s är positiva reella tal och $r^2 < s^2$ så gäller att $r < s$.
iii) För alla reella tal $r > 0$ gäller att $(r^2 + 4/r^2)/4 \geq 1$.

Proof. För det första påståendet antar vi att S är en mängd av reella tal som är nedåt begränsad av M dvs om $s \in S$ så har vi att $M \leq s$. Betrakta nu mängden $S' := \{-s \mid s \in S\}$. Eftersom $-s \leq -M$ för alla $s \in S$ så gäller att S' är uppåt begränsad av $-M$ och enligt supremumaxiomet finns en minsta övre begränsning N till S'. Då är $-N$ en största undre begränsning till S.

För den andra delen kan vi anta att $0 \leq s \leq r$. Från detta följer att $s^2 \leq r^2$ vilket är en motsägelse.

För det sista påståendet har vi att

$$0 \leq (r - 2/r)^2 = r^2 - 4 + 4/r^2$$

vilket ger $1 \leq (r^2 + 4/r^2)/4$ genom att flytta över fyran och sedan dividera med 4. □

Ett problem med bevis är att de är anpassade till en tänkt läsare så att ett bevis riktat till en församling kan hoppa över fler detaljer än ett riktat till en annan. Detta är oftast bra ty ju fler onödiga detaljer som kan hoppas över ju lättare är beviset att läsa. Det krävs dock att en läsare som tränar sig i att läsa bevis hela tiden är medveten om detta.

Övning 3: Redogör för vilka detaljer som hoppats över i beviset och skriv ett mer noggrannt bevis som fyller i dessa.

Vi är nu färdiga för att visa vårt resultat.

Proposition 2.2 *Det finns ett positivt reellt tal r sådant att $r^2 = 2$.*

Proof. Betrakta mängden $S := \{s \in \mathbf{R} \mid s > 0, s^2 \geq 2\}$. Denna mängd är nedåt begränsad av t ex 0 så den har enligt (2.1:i) en största undre begränsning som vi kallar r som är positivt då 0 är en undre begränsning till S. Vi ska nu visa följande två påståenden.

- Det är inte sant att $r^2 < 2$.
- Det är inte sant att $r^2 > 2$.

Tillsammans visar de att $r^2 = 2$.

Om vi börjar med det första påståendet så antar vi att $r^2 < 2$. Välj nu $\epsilon := (2 - r^2)/5$ som enligt antagandet är ett positivt reellt tal. Enligt (2.1:ii) är $r < 2$ ty $r^2 < 2 < 2^2$ och $\epsilon \leq 1$ så att $2r + \epsilon \leq 5$ och därmed är $\epsilon(2r + \epsilon) \leq (2 - r^2)/5 \cdot 5 \leq 2 - r^2$. Vi har därför att $(r + \epsilon)^2 = r^2 + \epsilon(2r + \epsilon) \leq r^2 + 2 - r^2 = 2$ men detta ger att för $s \in S$ har vi $(r + \epsilon)^{\leq}2 \leq s^2$ och enligt (2.1:ii) får vi att $r + \epsilon \leq s$ så att $r + \epsilon$ är en undre begränsning men då $\epsilon > 0$ är det en större undre begränsning än r vilket är en motsägelse.

Om vi istället antar att $r^2 > 2$ så har vi att $2/r < r$ så att om vi sätter $t := (r + 2/r)/2$ så gäller att $t < r$. Vidare har vi att $t^2 = (r^2+4+4/r^2)/4 \geq 1+(r^2+4/r^2)/4 \geq 1+1 = 2$, där vi för den sista olikheten använt (2.1:iii), så att $t \in S$ men då $t < r$ är r ingen undre begränsning. $\square$

När man analyserar ett bevis så är misstänksamhet (gränsande till paranoia) en god utgångspunkt. Det är nödvändigt att ifrågasätta *alla* påståenden som görs i beviset och förstå varför de är sanna. Det är också här som ett bevis som är passande för en oerfaren läsare kan komma att störa en mer erfaren sådan som omedelbart kan fylla i små detaljer.

Övning 4: i) Redogör för vilka detaljer som hoppats över i beviset och skriv ett mer noggrannt bevis som fyller i dessa.

ii) Markera var de olika resultaten från lemmat används i beviset. Skriv om beviset så att det som bevisas i lemmat istället bevisas inne i beviset och jämför läsbarheten.

iii) Försök förklara hur man kommer fram till valet $\epsilon := (2 - r^2)/5$.

iv) Försök lista ut varför man väljer $t := (r + 2/r)/2$.

v) Visa att för varje $a > 0$ så finns $r > 0$ sådant att $r^2 = a$.

3 Kardinalitet

En kommentar i avsnitt 2.3 gör några påståenden som inte är så klara som författarens "clearly" antyder. Faktum är att det är ganska instruktivt att gå genom det argument som döljer sig bakom detta ord. Avsnittet definierar en relation $A \sim B$ mellan mängder nämligen att $A \sim B$ om det finns en bijektion $f : A \to B$ och påstår att följande gäller för denna relation:

- $A \sim A$ för alla mängder A.
- Om $A \sim B$ så $B \sim A$.
- Om $A \sim B$ och $B \sim C$ så $A \sim C$.

Det är ofta en bra idé att tänka på matematiska påståenden och deras bevis i termer av ett spel mellan två personer där ett bevis är en vinnande strategi för den förste spelaren. För det första påståendet ($A \sim A$) ser en spelomgång ut så här (i varje steg kan den spelare som ska göra ett drag erkänna sig besegrad men vi nämner inte detta):

1. Spelare 1 gör påståendet $A \sim A$ för alla A.
2. Spelare 2 väljer nu en specifik mängd A.
3. Spelare 1 måste nu visa att för spelare 2:s val av A så gäller att $A \sim A$. Detta betyder enligt definitionen av relationen $A \sim A$ att denne väljer en funktion $f\colon A \to A$ samt hävdar att denna är en bijektion.
4. Spelare 2 kan nu satsa på att vinna genom att demonstrera att f inte är på eller inte är 1-1. I det förra fallet väljer spelaren att $a \in A$.
5. Spelare 1 måste nu välja ett $b \in A$ sådant att $f(b) = a$. Om denne kan göra detta är omgången vunnen.
6. Den andra möjligheten är att spelare 2 satsar på att ifrågasätta att f är 1-1. Då ska spelaren välja två element $a, a' \in A$ sådana att $f(a) = f(a')$.
7. Spelare 1 vinner nu om $a = a'$ och spelare 2 om $a \neq a'$.

Om i en sådan omgång spelare 2 skulle välja $A = \{1, 2, 3\}$ så kan spelare 2 välja många olika (6 stycken för att vara exakt) bijektioner från A till A och vinner då. Detta görs dock genom "improvisation" dvs det fungerar för just denna mängd men utnyttjar egenskaper hos just det valet av A. För att bevisa $\forall A\colon A \sim A$ måste vi komma på en strategi som garanterar att spelare 1 vinner för alla möjliga val av A (och det finns *många* konstiga mängder). Att spelare 1 har specificerat en funktion f betyder att om spelare 2 väljer ett element $a \in A$ så måste spelare 1 hitta ett element $b \in A$ och kunna beräkna $f(b)$ och se att det är lika med a. Det enda spelare 1 vet om mängden A är att den innehåller elementet a så det verkar nästan omöjligt att tänka sig att spelare 1 kan hitta något annat värde $b \in A$ än just a (det enda element som med säkerhet finns i A). (Spelare 1 har naturligtvis möjligheten att välja f på ett mycket smart sätt men då denne inte i förväg kan veta något om A verkar det osannolikt att detta alltid ska lyckas.) Detta argument är inte något som är en del av en vinnande strategi utan istället en del av ett argument som förhoppningsvis ska leda till en vinnande strategi. Vi har alltså kommit på en kandidat till en vinnande strategi inte för att vi är säkra på att den är bra utan för att vi inte kan komma på en bättre (och vi inte heller kan se hur man skulle komma på någon annan). Strategin är att vilken mängd A som spelare 2 kommer med så svarar spelare 1 med att ge tillbaka *identitetsfunktion*, id, som definieras av att $\mathrm{id}(a) = a$. Detta är dock ingen strategi utan bara första steget i en sådan; vi måste också tala om hur hur spelare 1 ska svara i fortsättningen.

- Om spelare 2 väljer att ifrågasätta att funktionen är på och skickar tillbaka ett $a \in A$ så vinner spelare 1 eftersom denne kan skicka tillbaka a; vi har ju att $\mathrm{id}(a) = a$.
- Om spelare 2 väljer att ifrågasätta att funktionen är 1-1 och skickar tillbaka $a, a' \in A$ med $\mathrm{id}(a) = \mathrm{id}(a')$ så vinner spelare 1 eftersom $a = \mathrm{id}(a) = \mathrm{id}(a') = a'$.

Vi ser alltså att vi har hittat en vinnande strategi för spelare 1. När man skriver ett matematiskt bevis så formulerar man dock inte detta i termer av spel och strategier men det är inte svårt att se hur nära en sådan formulering det ligger. Detta kan vi se om vi skriver ner ett bevis för $A \sim A$ i traditionella termer.

Givet en mängd A så betraktar vi identitetsfunktionen $\mathrm{id}\colon A \to A$. *Denna är en bijektion ty givet $a \in A$ så har vi* $\mathrm{id}(a) = a$ *så att* id *är på och givna $a, a' \in A$ med* $\mathrm{id}(a) = \mathrm{id}(a')$ *så har vi att* $a = \mathrm{id}(a) = \mathrm{id}(a') = a'$ *så att* id *är* 1-1.

Vi ser t.o.m. att orden givet/givna svarar precis mot att motståndaren, spelare 2, ger oss, spelare 1, vissa data som del i ett speldrag.

Som en vidare illustration ser vi på det tredje villkoret $A \sim B \wedge B \sim C \Rightarrow A \sim C$. I detta spel består spelare 2:s första drag av att ge ifrån sig tre mängder A, B och C samt två bijektioner $f\colon A \to B$ och $g\colon B \to C$. Observera att plötsligt befinner sig spelare 2 i en utsatt situation;

spelare 1 kan nu hävda att f eller g är bijektioner och spelare 2 kan inte ifrågasätta detta. Hur som helst spelare 1 är nu tvingad att på något sätt koka ihop en funktion $h\colon A \to C$ och ska också kunna försvara att denna är en bijektion. Spelare 1 kan rimligtvis bara använda sig av f och g och det är svårt att se hur man från dessa skulle kunna konstruera en funktion $A \to B$ annat än genom att sätta ihop dem; $h = g \circ f$. Om detta nu är spelare 1:s drag så kan spelare 2 välja att ifrågasätta att den är 1-1 eller att den är på. I det första fallet kommer spelare 2 alltså att ge tillbaka $a, a' \in A$ sådana att $h(a) = g(f(a))$ är lika med $h(a') = g(f(a'))$. Nu har ju spelare 2 garanterat att g är 1-1 och då $g(f(a)) = g(f(a'))$ så tvingas spelare 2 hålla med när spelare 1 hävdar att $f(a) = f(a')$. Då spelare 2 även har garanterat att f är 1-1 så tvingas spelare 2 acceptera att $a = a'$ och spelare 1 vinner.

Om istället spelare 2 ifrågasätter att h är på och sänder till spelare 1 ett $c \in C$ så kan nu spelare 1 insistera på att spelare 2 ska leverera ett $b \in B$ sådant att $g(b) = c$ eftersom det ju är spelare 2 som är garant för att g är på. Spelare 1 kan sedan kräva att spelare 2 väljer ett $a \in A$ med $f(a) = b$ eftersom spelare 2 garanterat att f är på. Spelare 1 vinner nu genom att avleverera a eftersom $h(a) = g(f(a)) = g(b) = c$. Vi har alltså hittat en vinnande strategi.

Övning 5: Skriv om denna vinnande strategi till ett bevis i vanlig mening.

Övning 6: Finn en vinnande strategi för $A \sim B \Rightarrow B \sim A$ och skriv denna som ett vanligt bevis.

Övning 7: Att en funktion $f\colon A \to B$ är en bijektion är samma sak som att den har invers $g\colon B \to A$ dvs sammansättningarna $g \circ f$ och $f \circ g$ är identitetsfunktionen på A respektive B. När en spelare hävdar att en funktion är en bijektion kan den andre spelaren alltså istället kräva att den förste spelaren levererar en invers. Finn vinnande strategier för de tre spelen där spelarna gör detta istället och skriv också om dessa som bevis.

4 Matematisk notation

Matematisk notation är komplicerad eftersom den dels ämnar vara extremt precis men samtidigt försöker vara så läsbar som möjligt. Detta innebär till exempel att man ofta använder vissa konventioner som man dock inte kan bara kan förutsätta. Ett exempel är att man ofta använder m eller n för att beteckna heltal och r eller s för att beteckna reella tal. Detta betyder *inte* att man kan skriva "Betrakta r" och sedan utgå från att man har gjort klart att r är ett reellt tal. Man måste alltså säga t ex "Betrakta $r \in \mathbf{R}$". Det är å andra sidan en bra idé att använda just r (eller s...) för ett reellt tal eftersom det gör det lättare för läsaren att hänga med.

Anmärkning: Ett slående exempel på hur djupt rotade dessa konventioner är är i integralen

$$\int_1^2 e^x \, de.$$

Det är helt legitimt att använda vilket namn som helst för integrationsvariabeln och när väl e har valts så har den inte längre någon annan speciell betydelse (som att vara bas för den naturliga logaritmen). Det är helt klart att det är en utomordentligt dålig idé att använda e på detta sätt och det bästa sättet att handskas med denna integral är förmodligen att omedelbart ersätta integrationsvariabeln med en annan så att man (till exempel) får

$$\int_1^2 y^x \, dy.$$

4.1 Följder

Ett något allvarligare exempel är notation för följder. Den allmänna principen är att om a är en följd av element i en mängd S så är a_n det n:te elementet i följden. Notera att man också talar om "Följden (a_n)". Detta är egentligen samma sak, man kan t.o.m. skriva "Följden $a = (a_n)$" och då är det exakt samma sak; man ger följden ett namn a och dessutom ger man ett namn a_n åt det n:te elementet i följden, nämligen a_n. Fördelen med med (a_n)-notationen är flexibilitet. Man kan till att börja med tala om följden (n^2) som är en följd a med $a_n = n^2$. Man kan sedan reindexera följden och tala om följden (a_{2n}) som med den givna definitionen har $(2n)^2 = 4n^2$ som n:te element. Notationen (a_n) har dock ibland en del nackdelar. Det första som kan förvirra är att (a_p) är precis samma följd som (a_n); i det första fallet har man en följd sådan att det p:te elementet är lika med a_p och i det andra en följd vars n:te element är lika med a_n. Man kan alltså inte använda olika indexvariabler för att definiera olika följder. Saker och ting kan bli ännu mer komplicerade om S har en speciell form. Ett förvirrande exempel är när S självt är en mängd av följder i en annan mängd T. Detta betyder att om a är en följd i S så är varje individuellt element a_n också en följd (av element i T). Det är troligt att man behöver någon notation för elementen i denna följd. Lyckligtvis behöver vi inte hitta på någon speciell sådan notation. Vi har ju redan sagt att om x betecknar en följd så betecknar x_n det n:te elementet i följden och x_m det m:te elementet. Så det m:te elementet i följden a_n bör betecknas a_{nm}. Det är dock lite svårt att se att m är ett index för a_n så man stoppar in parenteser för att göra detta tydligt; $(a_n)_m$. Man skulle naturligtvis också kunna använda $a_{n,m}$ för att beteckna samma sak men då måste man explicit tala om detta. Fördelen med $(a_n)_m$ är att man inte behöver säga något speciellt eftersom dess betydelse följer från en enda regel. Denna regel kan sedan kombineras med våra andra regler så att till exempel $(n^2)_4$ är det fjärde elementet i följden (n^2) dvs $(n^2)_4 = 4^2 = 16$.

4.2 Existensnotation

Framförallt i analys är påståenden av typen "För alla ... existerar ... så att ... ". Det är extremt viktigt att lära sig förstå sådana påståenden och framförallt förstå hur små variationer i formuleringarna kan ha stor betydelse. Ett exempel på sådana variationer är skillnaden mellan punktvis och likformig konvergens:

- *Om $f, f_n \colon X \to Y$ är funktioner där Y är ett metriskt rum så är f det* punktvisa gränsvärdet *av (f_n) om*

$$\forall \epsilon > 0 \forall x \in X \exists N \colon d_Y(f_n(x), f(x)) < \epsilon \text{ om } n > N.$$

- *Om $f, f_n \colon X \to Y$ är funktioner där Y är ett metriskt rum så är f det* likformiga gränsvärdet *av (f_n) om*

$$\forall \epsilon > 0 \exists N \forall x \in X \colon d_Y(f_n(x), f(x)) < \epsilon \text{ om } n > N.$$

Dessa två definitioner ser ju verkligen mycket lika ut även om en noggrann titt visar att det finns en skillnad; för punktvis konvergens har vi "för alla x existerar ett N" medan för likformig konvergens har vi "det existerar ett N så att för alla x". Det är en skillnad men det är kanske fortfarande inte klart vad skillnaden betyder. Ett sätt att försöka göra det klarare är att tänka i termer av ett spel mellan två personer som ovan.

Punktvis konvergens:

1. Jag påstår att (f_n) konvergerar punktvis mot f.

2. Min motståndare ger mig ett $\epsilon > 0$ *och* ett $x \in X$.

3. Jag vinner om jag kan producera ett N sådant att för alla $n > N$ så har vi $d(f_n(x), f(x)) < \epsilon$.

Likformig konvergens:

1. Jag påstår att (f_n) konvergerar likformigt mot f.
2. Min motståndare ger mig ett $\epsilon > 0$.
3. Jag producerar nu ett N.
4. Min motståndare vinner nu om det finns ett $x \in X$ och ett $n > N$ så att $d(f_n(x), f(x)) \geq \epsilon$.

Detta gör det lite klarare men det finns ett ännu bättre sätt att klargöra skillnaden: I ett "för alla ... så existerar ... " så väljer man först något och sedan ska något annat existera, man tillåter alltså att det som existerar kan bero på det första valet. Man kan då i notationen göra detta klart. Detta ger för punktvis konvergens

$$\forall \epsilon > 0 \forall x \in X \exists N_{\epsilon,x} \colon d_Y(f_n(x), f(x)) < \epsilon \text{ om } n > N_{\epsilon,x},$$

där man gjort klart att N får bero på både ϵ och x medan man för likformig konvergens har

$$\forall \epsilon > 0 \exists N_\epsilon \forall x \in X \colon d_Y(f_n(x), f(x)) < \epsilon \text{ om } n > N_\epsilon,$$

där det nu är klart att N bara tillåts bero på ϵ. Om uttrycken är komplicerade kan man mycket väl tillåta sig att skippa indexen så att man t ex kan skriva

$$\forall \epsilon > 0 \forall x \in X \exists N_{\epsilon,x} \colon d_Y(f_n(x), f(x)) < \epsilon \text{ om } n > N,$$

där alltså $N_{\epsilon,x}$ och N är samma tal. Man kan dock *inte* skriva

$$\forall \epsilon > 0 \forall x \in X \exists N \colon d_Y(f_n(x), f(x)) < \epsilon \text{ om } n > N_{\epsilon,x},$$

det måste vara helt klart att N kan bero på ϵ och x och man måste därför ha med dem när man *introducerar* N.

5 Fler analyserade bevis

Vi ska nu titta på några av bokens bevis och försöka analysera dem.

5.1 Weierstrass approximationssats

Beviset för Weierstrass approximationssats (Sats 7.26) är instruktivt eftersom det ganska väl kan delas upp i mindre steg.

- En inledande reduktion till intervallet $[0, 1]$ görs. (Boken säger inte hur men det är fråga om att betrakta $t \mapsto f((1-t)a + tb)$ på intervallet $[0, 1]$.)
- En andra inledande reduktion görs till fallet när $f(0) = f(1) = 0$. Detta tillåter oss att definiera f för alla reella tal genom att sätta den till 0 utanför $[0, 1]$ och f är då fortfarande likformigt kontinuerlig.
- Beviset består sedan i att man skriver ner en explicit formel för en approximerande följd:

$$P_n(x) := \int_{-1}^{1} f(x+t) Q_n(t)\, dt.$$

- Man måste nu visa två saker; att P_n är ett polynom och att $P_n \to f$ likformigt. Intressant nog är dessa två fakta sanna av helt olika orsaker.
- För att visa att P_n är ett polynom konstaterar vi först att eftersom f är 0 utanför $[0,1]$ så får vi samma integral om vi ändrar lite på integrationsintervallet.

$$\int_{-1}^{1} f(x+t)Q_n(t)\,dt = \int_{-x}^{1-x} f(x+t)Q_n(t)\,dt$$

och vi kan sedan göra ett variabelbyte $t \to t - x$ så att vi får

$$P_n(x) = \int_0^1 f(t)Q_n(t-x)\,dt.$$

Eftersom Q_n är ett polynom så kan $Q_n(t-x)$ skrivas som

$$Q_n(t-x) = \sum_i Q_n^i(x)t^i$$

där $Q_n^i(x)$ är polynom vilket ger

$$P_n(x) = \sum_i Q_n(x) \int_0^1 f(t)t^i\,dt$$

som i sin tur visar att P_n är ett polynom.

- För att visa att $P_n \to F$ likformigt använder man tre egenskaper hos Q_n:
  1. $Q_n(x) \geq 0$ för alla x.
  2. $\int_{-1}^{1} Q_n(t)\,dt = 1$.
  3. För varje $\delta > 0$ så konverger Q_n likformigt mot 0 på mängden $[-1,1]\setminus]-\delta,\delta[$.
- Låt oss visa att för en följd av (integrerbara) funktioner Q_n med dessa egenskaper och för en kontinuerlig funktion f på $[0,1]$ med $f(0) = f(1) = 0$ så gäller att $P_n \to f$ likformigt, där

$$P_n(x) = \int_{-1}^{1} f(t+x)Q_n(t)\,dt,$$

där vi sätter f till att vara 0 på $[-1,1]$ utanför $[0,1]$. Detta gör f till en kontinuerlig funktion på $[-1,1]$. Vi kan därför sätta $M := \sup_{x\in[-1,1]} |f(x)|$ och välja $\delta > 0$ så att $|f(s) - f(t)| < \epsilon$ om $|s - t| \leq 2\delta$ (för $s,t \in [-1,1]$). Slutligen väljer vi N så att om $n \geq N$ så är $Q_n(x) < \epsilon$ om $|x| \geq \delta$. Då har vi för $n \geq N$

$$|P_n(x) - f(x)| \overset{1)}{=} \left|\int_0^1 (f(t+x) - f(x))Q_n(t)\,dt\right| \overset{2)}{\leq}$$
$$\left|\int_{|t|\geq\delta} (f(t+x) - f(x))Q_n(t)\,dt\right| + \left|\int_{-\delta}^{\delta} (f(t+x) - f(x))Q_n(t)\,dt\right| \overset{3)}{\leq}$$
$$\int_{|t|\geq\delta} |f(t+x) - f(x)|Q_n(t)\,dt + \int_{-\delta}^{\delta} |f(t+x) - f(x)|Q_n(t)\,dt \overset{4)}{<}$$
$$4M\epsilon + \epsilon\int_{-\delta}^{\delta} Q_n(t)\,dt \leq 2M\epsilon + \epsilon = (2M+1)\epsilon$$

och högerledet går mot 0 när $\epsilon \to 0$. Här följer likhet 1) av att

$$f(x) = f(x)\cdot 1 = f(x)\int_{-1}^{1} Q_n(t)\,dt = \int_{-1}^{1} f(x)Q_n(t)\,dt$$

och likhet 2) följer av att vi delar upp integrationsintervallet i $\{t \in [-1,1] \mid |t| \geq \delta\}$ (som består av två intervall) och $[-\delta, \delta]$ och använder triangelolikheten. Sedan kommer olikhet 3) av att man kan uppskatta en integral med integralen av absolutbeloppet av integranden samt att $Q_n \geq 0$. Olikhet 4) fås då vi har $|f(t+x) - f(x)| \leq 2M$, $Q_n(t) < \delta$ för $|t| \geq \delta$ och längden av $\{t \in [-1,1] \mid |t| \geq \delta\}$ är < 2 samt att $|f(t+x) - f(x)| < \epsilon$ när $|t| \leq \delta$. Den sista olikheten följer av att av att

$$\int_{-\delta}^{\delta} Q_n(t) \leq \int_{-1}^{1} Q_n(t) = 1$$

eftersom $Q_n(t) \geq 0$.

- Det enda lite trixiga kvar är att konstruera polynom Q_n med de tre egenskaperna. Polynomen $((1-x^2)^n)$ uppfyller 1) och 3) men inte 2) och vi måste *normalisera* dem genom att multiplicera med en konstant c_n så att 2) blir uppfylld. Problemet är att om c_n är för stor så kan vi förstöra 3). Närmare bestämt begränsas $c_n(1-x^2)^n$ av $c_n(1-\delta^2)^n$ i $[0,1] \setminus]-\delta, \delta[$ så att kravet är att $c_n a^n \to n$ när $n \to \infty$ och för alla $0 \leq a < 1$. Bokens uppskattning ger $c_n \leq \sqrt{n}$ vilket medför detta villkor (med råge).

Det kan vara intressant att fundera över vad $\int_{-1}^{1} f(x+t)Q(t)\,dt$ är för slags konstruktion. För att ge den en form som liknar den vi får efter variabelbytet gör man ofta ett variabelbyte som gör den lika med $\int_{-1}^{1} f(x-t)Q(-t)\,dt$ och sedan använder man $Q(-t)$ i stället för $Q(t)$. Dessutom utvidgar man integralen till $]-\infty, \infty[$. (Detta förändrar ju inte integralen eftersom f är 0 utanför $[0,1]$.) Detta ger ett exempel på den sk *faltningen*

$$(f * g)(x) := \int_{-\infty}^{\infty} f(x-t)g(t)\,dt.$$

(Man måste göra vissa antaganden för att denna integral ska konvergera. Det räcker mer än väl att f är 0 utanför ett ändligt intervall.) Om man gör variabelbytet $s = x - t$ så får vi

$$(f * g)(x) = \int_{-\infty}^{\infty} f(x-t)g(t)\,dt = \int_{-\infty}^{\infty} f(s)g(x-s)\,ds = (g * f)(x).$$

(Återigen måste man göra vissa antaganden för att variabelbytet ska gå att göra men att f eller g försvinner utanför ett ändligt intervall räcker.) Detta är ju precis den relation vi använde för att visa att P_n var ett polynom. Om vi nu vill förstå approximationsdelen så är det nyttigt att betrakta en diskret version av faltning. Låt därför (a_n) vara en följd av reella tal där n löper över *alla* heltal och låt (b_n) vara en annan sådan följd. Vi antar att endast ändligt många b_n är skilda från 0 och definierar faltningsföljden $((a * b)_n)$ genom

$$(a * b)_n := \sum_i a_{n-i} b_i.$$

Detta ser ut som en oändlig summa men eftersom bara ändligt många b_i är skilda från 0 så är den i verkligheten ändlig (det räcker även med att bara ändligt många a_i är skilda från 0).

Övning 8: Visa att om a och b är två följder där den ena bara har ändligt många värden skilda från 0 så har vi $a * b = b * a$.

Den diskreta faltningen har många användningar. I vissa sådana användningar så tänker man sig att (a_n) specificerar någon form av signal (t ex en ljudsignal) och faltningen är tänkt att "jämna ut" signalen med hjälp av (b_n). Detta betyder

- endast värden som är nära en given position ska påverka det nya värdet i denna position så att b_i ska vara skilt från 0 bara om $|i|$ är litet,

- vi vill inte vända på signalen så att alla $b_n \geq 0$ och
- vi vill bara jämna ut signalen inte minska eller öka styrkan vilket betyder att vi vill ha $\sum_i b_i = 1$.

Man behöver inte göra en faltning i bara en dimension utan man kan till exempel betrakta följder $(a_{m,n})$ och $(b_{m,n})$ i två diskreta variabler med faltning given av

$$(a * b)_{m,n} := \sum_{i,j} a_{m-i,n-j} b_{i,j}.$$

I detta fall kan $a_{i,j}$ ge t ex gråskalan i en bild och $b_{i,j}$ kan användas för att försöka förbättra upplösningen. En viss upplösning svarar mot en viss pixelstorlek dvs man delar upp bilden i kvadrater av en viss storlek där svärtan är konstant i denna (se Fig. 1). Man kan sedan dela

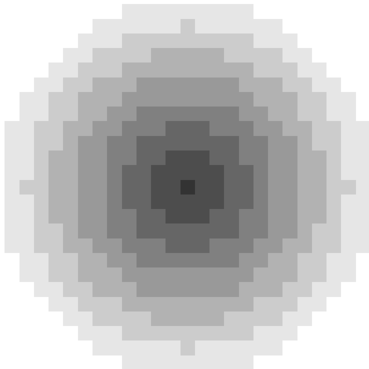

Fig. 1: Bild med dålig upplösning.

upp varje sådan kvadrat i (t ex) fyra delkvadrater. Man modifierar sedan svärtan i varje mindre kvadrat genom att vikta den med svärtan i näraliggande kvadrater. Viktfaktorerna kan t ex ges av följande av följande matris.

0,05	0,1	0,05
0,1	0,4	0,1
0,05	0,1	0,05

där den centrala faktorn 0,4 är vikten som ges till svärtan i den givna kvadraten, 0,05 är vikten som ges till svärtan i kvadraterna upp till höger, upp till vänster, ned till höger, ned till vänster osv. Den transformerade bilden kommer då att ha jämnare övergångar vilket normalt svarar mot en bild med bättre upplösning (se Fig. 2). I termer av faltning kan övergången från den första bilden till den andra ges av en faltning $a * b$, där svärtan i kvadrat (m, n) ges av $a_{m,n}$ och b ges av matrisen ovan dvs $m_{0,0} = 0{,}4$, $m_{1,1} = 0{,}05$, $m_{1,-1} = 0{,}05$, $m_{-1,1} = 0{,}05$, $m_{-1,-1} = 0{,}05$ osv.

Det är klart i diskreta fallet (och genom analogi i det kontinuerliga) att ju mer man koncentrar värdena i b till $(0, 0)$ (eller 0 i det endimensionella fallet) ju närmare kommer den transformerade följden att vara den ursprungliga. I det diskreta fallet kan man t.o.m. välja b med $b_{0,0} = 1$ och resten av $b_{m,n} = 0$ och då kommer $a * b$ att vara lika med a. I det kontinuerliga fallet finns ingen funktion g sådan att $f * g = f$ för alla f men resultatet ovan ger precisa villkor för att $f * Q_n \to f$ och det är klart att dessa villkor säger att Q_n blir mer och mer koncentrerad till 0. Om vi ritar upp olika Q_n för vårt specifika exempel $Q_n(x) = c_n(1 - x^2)^n$ så ser (Fig. 3) vi detta i konkreta termer.

Övning 9: Ett *trigonometriskt polynom* är en funktion på formen $\sum_{n=0}^{M} a_n \cos(nt) + b_n \sin(nt)$.

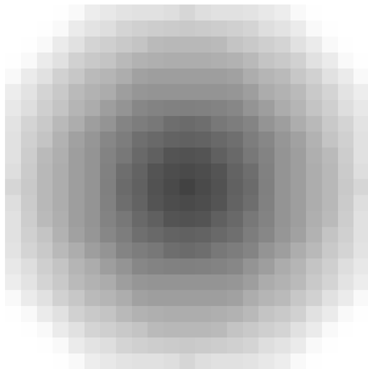

Fig. 2: Faltning av ursprunglig bild.

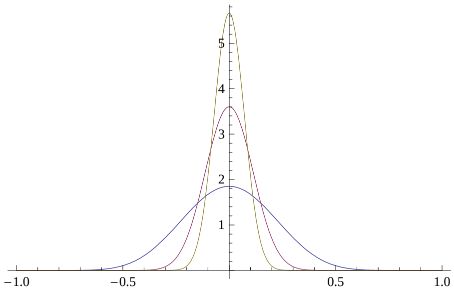

Fig. 3: $Q_n(x)$ för $n = 10, 40, 100$.

i) Visa att produkter och summor av trigonometriska polynom är trigonometriska polynom.

ii) Visa att varje kontinuerlig funktion f på $[-\pi, \pi]$ med $f(-\pi) = f(\pi)$ kan likformigt approximeras med trigonometriska polynom.

Ledning: Betrakta faltning med $Q_n(t) = d_n \cos^{2n}(t)$ för lämpliga d_n.

5.2 Kedjeregeln

Beviset av kedjeregeln för differentialen av funktioner (Sats 9.15) är rätt "upputsat" så att det är lite svårt att se hur man kan komma fram till det. Så här kan det se ut om man analyserar det steg för steg.

- Att en funktion $h \colon U \to \mathbf{R}^n$, U öppen i $\mathbf{R}^m$, är differentierbar i en punkt $z_0 \in U$ betyder att det finns en funktion $\epsilon \colon V \to \mathbf{R}^n$ definierad i en omgivning V till $0 \in \mathbf{R}^m$ så att $h(z_0 + h) = h(z_0) + Ch + |h|\epsilon(h)$ för en linjär avbildning $C \colon \mathbf{R}^m \to \mathbf{R}^n$ och så att $\epsilon(h) \to 0$ när $h \to 0$.

- Om vi anänder bokens beteckningar så finns alltså ϵ och η så $f(x_0+h) = f(x_0)+Ah+|h|\epsilon(h)$ och $g(y_0 + k) = f(y_0) + Bk + |k|\eta(k)$ med $\epsilon(h) \to 0$ när $h \to 0$ och $\eta(k) \to 0$ när $k \to 0$.

- Om vi sätter $F = g \circ f$ så har vi $F(x_0 + h) = g(f(x_0 + h))$. Vi kan utveckla $f(x_0 + h)$ som ovan så att vi får
$$g(f(x_0 + h)) = g(y_0 + Ah + |h|\epsilon(h)).$$

- När h är litet så är också $Ah + |h|\epsilon(h)$ litet och därmed verkar det rimligt att låta $k = Ah + |h|\epsilon(h)$ och tillämpa utvecklingen av g på detta:

$$g(y_0 + Ah + |h|\epsilon(h)) = g(y_0 + k) = g(y_0) + Bk + |k|\eta(k).$$

- Tittar vi närmare på Bk så blir det $B(Ah + |h|\epsilon(h))$ men B är linjär så detta blir lika med

$$BAh + |h|B(\epsilon(h))$$

och om vi skriver ut allt får

$$F(x_0 + h) = F(x_0) + BAh + |h|B(\epsilon(h)) + |k|\eta(k).$$

Detta i sin tur ser ut som att F är differentierbar med differential BA förutsatt att $|h|B(\epsilon(h)) + |k|\eta(k)$ går mot 0 fortare än h.

- Vi jämför därför med $|h|$

$$|h|B(\epsilon(h)) + |k|\eta(k) = |h|\left(B(\epsilon(h)) + \frac{|k|}{|h|}\eta(k)\right)$$

så att vad som är kvar är att visa att

$$B(\epsilon(h)) + \frac{|k|}{|h|}\eta(k) \to 0$$

när $h \to 0$.

- Vi har att $\epsilon(h) \to 0$ och B är kontinuerlig så $B(\epsilon(h)) \to 0$ vilket tar hand om första termen.

- Vi har

$$|k| = |Ah + |h|\epsilon(h)| \leq |Ah| + |h||\epsilon(h)| \leq \|A\||h| + |h||\epsilon(h)|$$

vilket ger att

$$\frac{|k|}{|h|}\eta(k) \leq (\|A\| + |\epsilon(h)|)\,\eta(k).$$

Den första termen, $\|A\| + |\epsilon(h)|$, är mindre än $\|A\| + 1$ om h är litet och $\eta(k) \to 0$ när $h \to 0$ eftersom $k \to 0$.

6 Specifika kommentarer

6.1 Ornade mängder

I Definition 1.5 av ordnade mängder finns (åtminstone i vissa tryckningar) ett feltryck: Den korrekt formuleringen av (ii) är "If $x, y, z \in S$, if $x < y$ and $y < z$, then $x < z$." (Det andra z:at har felaktigt blivit x.)

6.2 Rationell täthet

Den andra delen av Thm 1.20 är lättare att förstå om den delas upp i två delar (som bägge är intressanta, den andra är 1.20 (b)):

- Om $x, y \in \mathbf{R}$ med $y - x > 1$, så finns ett heltal m med $x < m < y$.

- Om $x < y$, så finns $p \in \mathbf{Q}$ med $x < p < y$.

För den första skulle vi vilja att m vore det minsta heltalet med $x < m$. Existensen av ett sådant m påstås i bokens bevis utan kommentar. Det är ett specialfall (av en version) av induktionsprincipen. Det kan dock vara värt att notera att denna följer från supremumaxiomet. (Beviset för detta är inte helt enkelt då det till att börja med kräver att vi talar om precis vad vi menar med heltalen som en delmängd till de reella talen och ges i en övning nedan.) Hur som helst om m är det minsta heltalet $> x$ so måste vi ha att $m - 1 \leq x$ dvs $m \leq x + 1 < y$ vilket visar första delen. För den andra delen så gör vi som i boken och använder (a) till att få ett heltal $n > 0$ sådant att $n(y - x) > 1$ det vill säga $ny - nx > 1$. Enligt den först delen finns då ett heltal m sådant att $nx < m < ny$ vilket ger $x < m/n < y$.

Övning 10: i) Vi säger att en delmängd S till $\mathbf{R}$ är *stabil* om $0 \in S$ och $m \in S \Rightarrow m + 1 \in S$. Visa att snittet av alla stabila delmängder är stabil som är en stabil delmängd innehållen i alla stabila delmängder. Beteckna detta snitt $\mathbf{N}$.

ii) Visa att om $0 \neq m \in \mathbf{N}$ så har vi $m - 1 \in \mathbf{N}$. (Ledning: Visa att mängden $\{0\} \cup \{m \in \mathbf{R} \mid m - 1 \in \mathbf{N}\}$ är stabil.)

iii) Visa att om $n, m \in \mathbf{N}$ så har vi $n < m \Rightarrow n \leq m - 1$. (Ledning: Använd ii) för att visa för ett fixt n att mängden av m som uppfyller implikationen är stabil.)

iv) Visa att om S är en icke-tom delmängd till $\mathbf{N}$ så innehåller den ett minsta element. (Ledning: Om x är infimum av S så finns ett $m \in S$ sådant att $m < x + 1$. Använd iii) för att visa att m är ett minimalt element.)

v) Visa att en icke-tom delmängd till $\mathbf{N}$ som har en övre gräns innehåller ett största element.

vi) Definiera heltalen $\mathbf{Z}$ som unionen av $\mathbf{N}$ och $-\mathbf{N} := \{-m \mid m \in \mathbf{N}\}$. Visa att varje delmängd till $\mathbf{Z}$ som har en undre gräns innehåller ett minimalt element.

vii) Visa att $\mathbf{Z}$ är sluten under addition och multiplikation. (Ledning: Reducera till lämpliga påståenden för $\mathbf{N}$ och använd samma tekning som tidigare.)

6.3 Ekvivalensrelationer och -klasser

I definition 2.3 introduceras begreppet ekvivalensrelation. Senare (såväl som tidigare) används det associerade begreppet ekvivalensklass utan vidare kommentar. Givet en ekvivalensrelation $\sim$ på en mängd S, så är *ekvivalensklassen* till vilken ett element $s \in S$ tillhöre delmängden $\overline{s} := \{t \in S \mid s \sim t\}$. Villkoren för att vara en ekvivalensrelation garanterar till att börja med att s verkligen tillhör $\overline{s}$, sedan att S är den disjunkta unionen av de olika ekvivalensklasserna samt att t och t' tillhör samma ekvivalensklass precis när $t \sim t'$.

Övning 11: Verifiera dessa påstående.

Poängen med en ekvivalensrelation är att man ofta vill se två ekvivalenta element som väsentligen lika. Detta kan man formalisera genom att betrakta mängden $\overline{S}$ av ekvivalensklasser så att $s \mapsto \overline{s}$ som en funktion $S \to \overline{S}$. Ekvivalens mellan element kan sedan ersättas av faktisk likhet mellan ekvivalensklasser.

6.4 Uppräknelig union av uppräkneliga mängder

Beviset av sats 2.12 är lite svåröverskådligt. Ett bättre sätt att organisera det är att börja med ett lemma.

Lemma 6.1 *Om $f\colon S \to T$ är en funktion och S är uppräknelig så är bilden $f(S)$ uppräknelig eller ändlig.*

Proof. Låt $g\colon \mathbf{Z}_+ \to S$ vara en bijektion (där $\mathbf{Z}_+$ är de positiva heltalen) och låt h vara sammansättningen $f \circ g$. Per definition är h en surjektiv avbildning till $f(S)$. Sätt

$$T := \{n \in \mathbf{Z}_+ \mid m < n \Rightarrow h(m) \neq h(n)\}.$$

Då gäller att restriktionen av h till T fortfarande är surjektiv och den är nu också injektiv dvs h ger en bijektion mellan T och $f(S)$. Enligt sats 2.8 i boken är T antingen ändlig eller uppräknelig och därmed är också $f(S)$ det. $\square$

Beviset av sats 2.12 är nu att vi (om vi använder beteckningarna från beviset) först hittar en surjektiv avbildning från $\mathbf{N}$ till S genom uppräkningen $x_{11}; x_{21}, x_{12}; x_{31}, x_{22} \dots$ och lemmat ger då att S är ändlig eller uppräknelig men då $E_1 \subseteq S$ är den uppräknelig.

Observera också att uppräkningen som görs i beviset inte är ett matematiskt bevis utan läsaren förväntas förvandla den grafiska beskrivningen till ett matematiska bevis. Det som ska bevisas är att det finns en bijektion mellan mängden av positiva heltal $\mathbf{Z}_+$ och mängden av par av positiva heltal $\mathbf{Z}_+^2$. Vi kan använda den grafiska beskrivningen som utgångspunkt men måste förvandla den till konkreta formler. Nyckelberäkningen är att man måste räkna ut vilken position man kommit till när man avslutat en diagonal dvs hur många element det finns i mängden $\{(i,j) \in \mathbf{Z}_+^2 \mid i+j \leq n\}$. Man kan räkna varje diagonal för sig och kommer fram till att detta antal är lika med $1+2+\cdots+n-1$ vilket vi vet är lika med $n(n-1)/2$. Med detta som startpunkt kan vi ge ett formellt bevis.

Proposition 6.2 *Avbildningen $f\colon \mathbf{Z}_+^2 \to \mathbf{Z}_+$ given av $f(i,j) = (i+j-1)(i+j-2)/2+j$ är en bijektion.*

Proof. Vi börjar med att bevisa ett påstående som vi kommer att se är ekvivalent med det vi vill bevisa.

För varje $m \in \mathbf{Z}_+$ finns unika $n, j \in \mathbf{Z}_+$ med $0 < j < n$ och $m = (n-1)(n-2)/2 + j$.

För att visa detta låter vi n vara det största heltalet för vilket $(n-1)(n-2)/2 < m$. Ett sådant n existerar eftersom $(n-1)(n-2)/2 < m$ medför att $n \leq m$ så att vi kan använda supremumaxiomet. Vi har då att $n+1$ inte uppfyller villkoret så att $n(n-1)/2 = (n+1-1)(n+1-2)/2 \geq m$. Detta ger att $m - n(n-1)/2 \leq n(n-1)/2 - (n-1)(n-2)/2 = n-1$ så att om vi sätter $j := m - n(n-1)/2$ så har vi $0 < j < n$ och $m = (n-1)(n-2)/2 + j$. Om vi istället antar att vi har skrivit m som $m = (n-1)(n-2)/2 + j$ med $0 < j < n$ så kan vi gå baklänges och ser att n är det största heltalet n så att $(n-1)(n-2)/2 < m$ vilket betyder att n är entydigt bestämt av m och då är j det också.

Vi definierar nu en funktion $g\colon \mathbf{Z}_+ \to \mathbf{Z}_+^2$ genom att för givet $m \in \mathbf{Z}_+$ skriva m som $(n-1)(n-2)/2 + j$ med $0 < j < n$ och sedan sätta $g(m) = (n-j, j)$. Av vad vi redan bevisat ser vi att g är väldefinierad och man ser sedan lätt att f och g är inverser till varandra. $\square$

Observera att detta gör det fullständiga beviset för satsen lite mer komplicerat än man skulle vilja. Det finns dock andra sätt att konstruera en bijektion mellan $\mathbf{Z}_+ \times \mathbf{Z}_+$ och $\mathbf{Z}_+$ eller, vilket är samma sak, mellan $\mathbf{N} \times \mathbf{N}$ och $\mathbf{N}$. Följande övning ger ett.

Övning 12: i) Visa att varje naturligt n tal kan entydigt skrivas på formen $n = \sum_{k=0}^{\infty} a_k 10^k$, där $0 \leq a_k < 10$ och alla utom ändligt många a_k är lika med 0.

ii) Visa att avbildningen

$$(\sum_{k=0}^{\infty} a_k 10^k, \sum_{k=0}^{\infty} b_k 10^k) \mapsto \sum_{k=0}^{\infty} c_k 10^k,$$

där $c_k = a_{k/2}$ om k är jämnt och $c_k = b_{(k-1)/2}$ om k är udda, ger en bijektion $\mathbf{N} \times \mathbf{N} \to \mathbf{N}$.

6.5 Uppräknelighet av de rationella talen

Cor. 2.13 är väldigt kortfattat så att det kan finnas anledning att titta lite närmare på det. Vi börjar med att visa uppräkneligheten av de positiva rationella talen $\mathbf{Q}_+$. Vi har en avbildning $\mathbf{Z}_+^2 \to \mathbf{Q}_+$ som tar ett par (m,n) till talet m/n. Denna avbildning är på och då vi har en bijektion $\mathbf{Z}_+ \to \mathbf{Z}_+^2$ så får vi en sammansättning $\mathbf{Z}_+ \to \mathbf{Z}_+^2 \to \mathbf{Q}_+$ som också den är på (eftersom sammansättningen av två funktioner som bägge är på också är på). Enligt Lemma 6.1 så är $\mathbf{Q}_+$

uppräknelig eller ändlig men $\mathbf{Q}_+$ innehåller $\mathbf{Z}_+$ som inte är ändlig och då kan inte $\mathbf{Q}_+$ heller vara ändlig.

Övning 13: Vi har en explicit avbildning $\mathbf{Z}_+ \to \mathbf{Q}_+$ och Lemma 6.1 ger en explicit uppräkning given en avbildning som är på. Bestäm bilden av 20 under denna explicita uppräkning.

6.6 Konvergens av Cauchyföljder

Sats 3.11(b) kan bevisas med ett lite annorlunda argument än bokens som kanske är tydligare: Man använder sats 3.6(a) för att dra slutsatsen att det finns ett $p \in X$ och en delföljd (p_{n_k}) sådan att $p_{n_k} \to p$. Det räcker sedan att visa att hela följden konvergerar mot p. Låt därför $\epsilon > 0$ vara godtyckligt och välj N så att vi både har $d(p_{n_k}, p) < \epsilon/2$ (detta gör vi genom att först välja ett K så att det gäller för $k \geq K$ och sedan välja ett N så att $n_k \geq N \Rightarrow k \geq K$) om $n_k \geq N$ och $d(p_m, p_n) < \epsilon/2$ om $m, n \geq N$ (vilket vi kan göra då delföljden konvergerar och hela följden är en Cauchyföljd). Välj ett $n_k \geq N$ och låt $m \geq N$. Vi får då genom triangelolikheten; $d(p_m, p) \leq d(p_m, p_{n_k}) + d(p_{n_k}, p) < \epsilon/2 + \epsilon/2 = \epsilon$.

6.7 Sammanhängande mängder

Beviset av sats 4.22 blir mer svåröverskådligt än nödvändigt på grund av den definition av sammanhängande delmängd som boken använder. Man kan göra beviset enklare (mot att man behöver göra lite mer förberedande arbete). Vad man börjar med att göra är att definiera begreppet sammanhängande metriskt rum (boken definierar ju vad som menas med en sammanhängande delmängd till ett metriskt rum).

Definition 6.3 *Ett metriskt rum X är* sammanhängande *om det* inte *kan skrivas som en disjunkt union av två öppna icke-tomma mängder.*

Anmärkning: i) Detta är ju en negativ definition (i det att den snarare definierar vad en icke sammanhängande mängd är). Den positiva versionen (och den som används i praktiken) är att om man har skrivit ett sammanhängande rum som en disjunkt union av två öppna mängder så är en av dessa mängder tom.

ii) Observera att om det metriska rummet X är den disjunkta unionen av de öppna delmängderna U och V så är U komplementet av V så att U (och även V) också är sluten. Omvänt om U är både öppen och sluten så är X unionen av U och dess komplement V som också är öppen. Ett rum är därför sammanhängande precis när det inte innehåller några äkta icke-tomma *slöppna* (= *sl*uten + *öppen*, *clopen* på engelska) delmängder.

iii) Man behöver inte bokens definition av sammanhängande delmängd utan det går utmärkt att bara använda sig av den vi givit (fast vi kommer nedan att visa att de två definitionerna är ekvivalenta).

Det resultat som är analogt med sats 4.22 för denna definition har ett mycket klarare bevis.

Sats 6.4 *Om $f\colon X \to Y$ är en kontinuerlig funktion mellan metriska rum så är bilden $f(X)$ sammanhängande (betraktat som metriskt rum genom metriken som kommer från Y) om X är det.*

Proof. Funktionen f ger upphov till en funktion (som vi också kallar för f) $f\colon X \to f(X)$. Man kan nu lätt visa att f som funktion från X till Y är kontinuerlig precis när den är kontinuerlig som funktion från X till $f(X)$ (där alltså $f(X)$ är ett metriskt rum genom metriken från Y). Vi kan därför anta att $Y = f(X)$. Antag nu att Y är den disjunkta unionen av de icke-tomma öppna delmängderna U och V. Då är X den disjunkta unionen av $f^{-1}(U)$ och $f^{-1}(V)$. Dessa mängder är icke-tomma då $f(X) = Y$ och de är öppna då f är kontinuerlig och U och V är öppna. Eftersom X antags vara sammanhängande är detta en motsägelse. $\square$

Övning 14: Visa att om X och Y är metriska rum och $Z \subseteq Y$ så är en funktion $f\colon X \to Z$ kontinuerlig (där Z betraktas som metriskt rum genom metriken från Y) precis när funktionen $X \to Y$ given av f är kontinuerlig.

För att få ett enklare bevis för sats 4.22 så kvarstår att visa hur vår definition av sammanhängande rum är kopplad till bokens definition av sammanhängande delmängd.

Proposition 6.5 *Låt X vara ett metriskt rum och $E \subseteq X$ en delmängd. Då är är E sammanhängande som metriskt rum (med metrik given av den metrik som finns på X) precis när den är en sammanhängande delmängd till X (dvs enligt definition 2.45).*

Proof. Antag att E är unionen av två separerade mängder A och B och låt $U' := X \setminus \overline{A}$ och $V' := X \setminus \overline{B}$ samt $U := U' \cap E$ och $V := V' \cap E$. Det är klart att U' och V' är öppna mängder i X och därmed är U och V öppna i E (som alltid betraktat som metriskt rum genom X:s metrik). Vi har att U innehåller B eftersom $B \cap \overline{A} = \emptyset$ och U är uppenbarligen disjunkt från A så att $U = B$ eftersom E är den disjunkta unionen av A och B. På samma sätt är $V = A$ så att E är den disjunkta unionen av de två öppna icke-tomma delmängderna U och V.

Antag nu att E är den disjunkta unionen av de öppna icke-tomma delmängderna U och V. Detta betyder att det för varje $x \in U$ finns ett $r_x > 0$ så att klotet $B_E(x, r_x) := \{y \in E \mid d(x, y) < r_x\}$ ligger i U och därmed har vi att $B_E(x, r_x) \cap V = \emptyset$. Detta betyder att $B_X(x, r_x) \cap V = \emptyset$ eftersom $B_X(x, r_x) \cap E = B_E(x, r_x)$. Om vi sätter $U' := \cup_{x \in U} B_X(x, r_x)$ så har vi att U är öppen i X och därmed att $\overline{V} \subseteq X \setminus X'$. Vidare har vi uppenbarligen att $U \subseteq U'$ så att $\overline{V} \cap U = \emptyset$ och på samma sätt får vi att $V \cap \overline{U} = \emptyset$ så att U och V är separerade. □

För att vidare demonstrera att vår definition av sammanhängande rum kan användas utan att man involverar bokens definition så bevisar vi karakteriseringen av sammanhängande delmängder till $\mathbf{R}$.

Proposition 6.6 *En delmängd E till den utvidgade reella linjen är sammanhängande precis när $z \in E$ om $x, y \in E$ och $x < z < y$.*

Proof. Antag till att börja med att det finns x, y, z med $x < z < y$ och $x, y \in E$ men $z \notin E$. Om vi sätter $U := \{t \in E \mid t < z\}$ och $V := \{t \in E \mid z < t\}$. Då gäller att U och V är öppna i E, E är den disjunkta unionen av U och V eftersom $z \notin E$ och U och V är icke-tomma eftersom $x \in U$ och $y \in V$. Alltså är E inte sammanhängande.

För omvändningen kan vi anta, med målet att få en motsägelse, att E är den disjunkta unionen av de öppna icke-tomma mängderna U och V samt att om $x, y \in E$ så gäller att alla z med $x < z < y$ tillhör E. Vi kan vidare välja $x \in U$ och $y \in V$. Efter att eventuellt ha bytt plats på U och V kan vi anta att $x < y$ och vårt antagande ger att $[x, y] \subseteq E$. Låt z vara supremum av $U \cap [x, y]$ (som existerar som utvidgat reellt tal eftersom x tillhör denna mängd) så att om $t \in U \cap [x, y]$ så gäller att $t \leq z$. Antag att $z \in U$. Eftersom U är öppen finns ett $\delta > 0$ så att $\{t \in E \mid |t - z| < \delta\}$ ligger i U. Eftersom $z \neq y$ (då $y \in V$) och $[x, y] \subseteq E$ så ger detta att det finns $t \in U$ med $t > z$ vilket motsäger att $t \leq z$. Därför måste vi ha att $z \in V$ men igen får vi att det finns $\delta > 0$ så att $\{t \in E \mid |t - z| < \delta\}$ ligger i V. Detta ger att $t \leq z - \delta$ om $t \in U$ vilket motsäger att z är supremum av $U \cap [x, y]$ (eftersom $x \in U$). □

6.8 Uppräknelighet av diskontinuitetspunkter

Det finns ett alternativt bevis för Sats 4.30 (att mängden av diskontinuitetspunkter av en monoton funktion är högst uppräknelig) som nog är lite lättare att förstå. Vi antar att $f\colon [a, b] \to \mathbf{R}$ är monoton och låter E vara mängden av dess diskontinuitetspunkter. Genom att eventuellt betrakta $-f$ istället kan vi anta att f är växande. Sätt nu, för ett positivt heltal n, $E_n := \{x \in E \mid f(x+) - f(x-) > 1/n\}$. Om vi kan visa att E_n är ändlig och att $E = \cup_n E_n$ så följer från Korollarium 2.12 att E är som mest uppräknelig. Vi har att $x \in E$ precis när $f(x+) - f(x-) > 0$. För ett $x \in E$ kan vi då välja n så att $f(x+) - f(x-) > 1/n$ vilket betyder att $x \in E_n$ så att

E är unionen av E_n:en. Å andra sidan, låt $\{x_1 < \cdots < x_k\}$ vara en ändlig delmängd till E_n. Då gäller att

$$\begin{aligned} f(b) - f(a) = f(b) - f(x_k+) + f(x_k+) - f(x_k-) + f(x_k-) - f(x_{k-1}+) + \\ f(x_{k-1}) + \cdots - f(x_1-) + f(x_1-) - f(a) = \\ (f(b) - f(x_k+)) + (f(x_k+) - f(x_k-)) + \\ (f(x_k-) - f(x_{k-1}+)) + (f(x_{k-1} + \cdots - f(x_1-)) + (f(x_1-) - f(a)) \,. \end{aligned}$$

Alla termerna i den sista summan är ≥ 0 då f är växande och $f(x_i+) - f(x_i-) > 1/n$ för $i = 1, \ldots, k$. Detta betyder att hela summan är $> k/n$ vilket ger att $k < n(f(b) - f(a))$. Detta betyder att varje ändlig delmängd till E_n innehåller mindre än $< n(f(b) - f(a))$ element vilket betyder att E_n självt innehåller högst så många element och därmed är ändlig.

6.9 Def 4.33

I definition 4.33 ska "such that $V \cap E$ is not empty" ersättas med "such that $V \cap E$ contains points different from x".

Övning 15: Ge ett exempel på att definition 4.33 som den står i boken inte är ekvivalent med definition 4.1 (i förekommande fall).

6.10 Sats 6.10

I början av beviset konstateras att man kan välja $[u_j, v_j]$ runt punkterna i E så att $\sum_j \alpha(v_j) - \alpha(u_j) < \epsilon$. Anledningen till detta är att om p_j är punkten i $E \cap [u_j, v_j]$ så kan vi, genom att krympa intervallet $[u_j, v_j]$, göra $\alpha(v_j) - \alpha(p_j)$ och $\alpha(p_j) - \alpha(u_j)$ godtyckligt litet (då α är kontinuerlig i p_j) och därmed kan $\alpha(v_j) - \alpha(u_j) = \alpha(v_j) - \alpha(p_j) + \alpha(p_j) - \alpha(u_j)$ göras godtyckligt litet. Nu är antalet element i E fixt så att $\sum_j \alpha(v_j) - \alpha(u_j)$ kan göras godtyckligt liten.

Observera att beviset har stora likheter med beviset av Weierstrass approximationssats. I bägge fallen delar man upp det relevanta intervallet i två bitar och behandlar dessa två bitar på helt olika sätt. I det ena fallet använder man bara att den godtyckliga funktionen f är begränsad och en annan faktor ($\alpha(v_j) - \alpha(u_j)$ resp. Q_n) är litet, i det andra fallet använder man att f är kontinuerlig och därför har liten variation på små intervall.

6.11 Normen av en linjär avbildning

Det förmodligen bästa sättet att tänka på normen av en linjär avbildning A är att den är det minsta tal λ sådant att $|Ax| \leq \lambda|x|$ för alla vektorer x. Om man alltså ska bevisa en övre uppskattning av $\|A\|$ så gäller det alltså att uppskatta $|Ax|$ som ett fixt tal gånger $|x|$. Till exempel har vi att $|(A+B)x| = |Ax + Bx| \leq |Ax| + |Bx| \leq \|A\||x| + \|B\||x| = (\|A\| + \|B\|)|x|$ vilket ger $\|A + B\| \leq \|A\| + \|B\|$ (som är ett av resultaten i sats 9.7).

Det finns dock ett annat sätt att se på normen: Om $x \neq 0$ så är $|Ax| \leq \lambda|x|$ ekvivalent med att $A(x/|x|) \leq \lambda$ och längden av $x/|x|$ är lika med 1 så att

$$\|A\| = \sup_{|x|=1} |Ax|.$$

Nu kan vi direkt se att A som en funktion $\mathbf{R}^m \to \mathbf{R}^n$ är kontinuerlig; om $x = \sum_i \lambda_i e_i$ (där e_i är elementen i standardbasen) så har vi $Ax = \sum_i \lambda_i A e_i$ så att kontinuiteten följer direkt av kontinuitet för addition och multiplikation. Från detta följer att funktionen $x \mapsto |Ax|$ är kontinuerlig (eftersom $y \mapsto |y|$ är kontinuerlig). Vidare så är enhetssfären $\{x \in \mathbf{R}^m \mid |x| = 1\}$ sluten och begränsad och därmed kompakt så att den kontinuerliga funktionen $x \mapsto |Ax|$ är begränsad och därmed är $\sup_{|x|=1} |Ax| < \infty$.

Vi kan vara ännu mer precisa och först konstatera att

$$\|A\|^2 = \sup_{|x|=1} |Ax|^2$$

och $x \mapsto |Ax|^2$ är en kvadratisk form. Mer precist har vi att

$$|Ax|^2 = (Ax)^t(Ax) = x^t A^t Ax,$$

där vi identifierat den linjära avbildningen med matrisen (i standardbasen för $\mathbf{R}^n$). Detta betyder att den kvadratiska formen ges av den symmetriska matrisen A^tA. Vi vet nu från teorin för kvadratiska former att $\sup_{|x|=1} |Ax|^2$ är lika med det största egenvärdet till den symmetriska matrisen A^tA så att $\|A\|$ är lika med kvadratroten av detta största egenvärde.

Exempel: Antag att $A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$. Då gäller att $A(x, y) = y$ och normen $\|A\|$ är alltså det minsta tal λ så att $|y| \leq \lambda\sqrt{x^2 + y^2}$. Om vi sätter $(x, y) = (0, 1)$ så får vi att $1 \leq \lambda$. Å andra sidan har vi att $|y| \leq \sqrt{x^2 + y^2}$ så att $\lambda = 1$ fungerar. Detta ger att $\|A\| = 1$.

Vi kan istället beräkna $A^tA = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$ som uppenbarligen har egenvärden 0 och 1 och vi får att $\|A\| = \sqrt{1} = 1$.

7 Inversa funktionssatsen

Bokens bevis för inversa funktionssatsen är ganska svåröverskådligt. Jag kommer att ge ett bevis som innehåller samma grundidéer men som förhoppningsvis är lite lättare att förstå. Det tar sin utgångspunkt i Newton-Raphsons metod.

Innan vi går in på detaljer låt oss starta med att notera några gemensamma drag mellan de två metoderna.

I bägge fallen vill vi lösa en ekvation $f(x) = y$, där x är det värde vi söker. Å andra sidan är y ett variabelt värde nära ett b där vi som startdata har ett a sådant att $f(a) = b$ och som ett ytterligare villkor letar vi efter en lösning x som är nära a. I bägge fallen försöker vi hitta ett *iterationsschema* $x_{n+1} = G_y(x_n)$, där vi använder notationen G_y för att betona att G_y kommer att bero på y. Idén är att följden (x_n) ska konvergera till en lösning x. Som alltid med iterationsschema finns det två saker att göra; visa att följden konvergerar och givet att den konvergerar mot en lösning till det ursprungliga problemet. Givet att följden konvergerar mot något x så kan låta n gå mot ∞ i relationen $x_{n+1} = G_y(x_n)$. Vi får då, om vi antar att G_y är kontinuerlig, $x = G_y(x)$ och ett första villkor på G_y är att denna relation ska medföra att $f(x) = y$ och om så är fallet har vi tagit hand om den andra delen. För konvergensen är den allmänna idén att att G_y ska ha egenskapen att x_{n+1} och x_n ska komma närmare och närmare varandra så fort att (x_n) kommer att bli en Cauchyföljd. (Observera att för detta kommer det inte riktigt att räcka med att $x_{n+1} - x_n \to 0$ när $n \to \infty$ eftersom vi behöver det starkare villkoret att $x_i - x_j \to 0$ när $i, j \to \infty$.) En viktig del i att få detta att fungera för ett givet iterationsschema är att man måste ha ett bra startvärde x_0, närmare bestämt ska det vara nära en faktisk lösning. Detta är i allmänhet ett svårt problem. I vårt fall förväntas y vara nära b och vi hoppas att x kommer att vara nära a. Detta antyder att vi borde välja a som startvärde och ger oss åtminstone kandidat till ett startvärde. Att hela metoden verkligen fungerar beror till väsentlig del på att vi har friheten att välja hur nära b som y ska vara. Vi väljer detta avstånd tillräckligt litet för att få ett antal uppskattningar som tillsammans får alla delar att fungera.

(Note that for this it will not quite be enough that $x_{n+1} - x_n \to 0$ when $n \to \infty$ as we need $x_i - x_j \to 0$ when $i, j \to \infty$.) A crucial part to make this work for an iteration scheme is that one must have a good initial value x_0, in fact it should be close to an actual solution. This is in general a difficult problem. In our case y is supposed to be close to b and we are hoping that x will be close a. This suggests that we should pick a as the initial value and at least it gives us a candidate for initial value.

7.1 Newton-Raphson i en variabel

Newton-Raphsons metod försöker lösa en ekvation $g(x) = 0$, där vi börjar med att anta att g är en funktion från en öppen delmängd i $\mathbf{R}$ till $\mathbf{R}$, genom att starta med ett närmevärde x_0 och sedan steg för steg försöka hitta bättre och bättre approximationer. I ett försök att hitta en bättre approximation till ett nollställe till g byter man ut g mot dess tangent i punkten och hoppas att ett nollställe till tangenten är en bättre approximation. Sedan upprepar man denna procedur (se Fig. 4). Om x_n är resultatet av det n:te steget så har dess tangent ekvationen

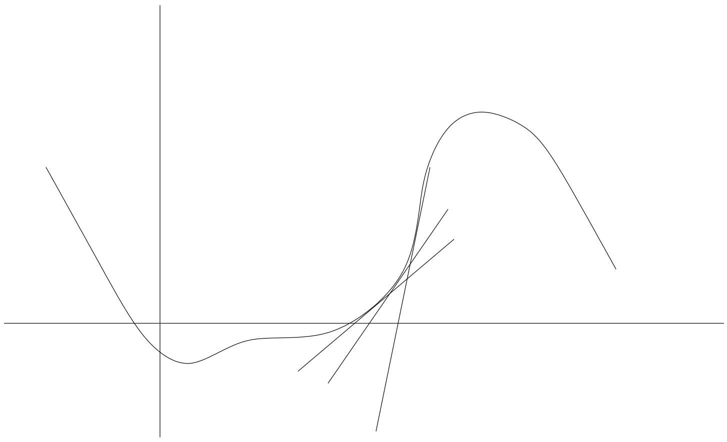

Fig. 4: Newton-Raphsons metod

$y = g(x_n) + g'(x_n)(x - x_n)$ och om vi löser ut ett nollställe får vi

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

Om vi antar att denna följd konvergerar mot ett x så kan vi ta gränsvärdet i denna definition och får att $x = x - g(x)/g'(x)$ vilket ger $g(x) = 0$ så att vi har hittat ett nollställe. Här har vi ju gjort ett antal antaganden och nästa steg är att hitta villkor under vilka dessa är uppfyllda.

Exempel: Låt $g(x) = x^2 - 2$. Då får vi $x_{n+1} = x_n - (x_n^2 - 2)/2x_n$. Detta kan skrivas om som $x_{n+1} = (x_n + 2/x_n)/2$ vilket kan uttryckas som att vi låter x_{n+1} vara medelvärdet av x_n och det tal vars produkt med x_n blir 2. Om vi startar med $x_0 = 1$ får vi $x_1 = 1{,}5$, $x_2 \approx 1{,}4167$, $x_3 \approx 1{,}414216$, $x_4 \approx 1{,}414213562375$, $x_5 \approx 1{,}4142135623709504880169$ vilket som synes konvergerar mycket snabbt mot $\sqrt{2} \approx 1{,}4142135623730950488016887242$.

Till att börja måste vi förstås anta att derivatan till $g(x)$ existerar men vi ser att för att få det avslutande gränsvärdesargumentet att fungera måste vi också anta att derivatan är kontinuerlig. Detta är också vad som behövs för att få en del uppskattningar vi behöver att fungera. Närmare bestämt har vi följande lemma som säger att det relativa felet i approximationen av en funktion med dess tangent är likformigt begärnsat.

Lemma 7.1 *Låt $g\colon]a,b[\to \mathbf{R}$ vara en funktion med kontinuerlig derivata. Då gäller att det för varje slutet intervall I i $]a,b[$ och varje $\epsilon > 0$ finns ett $\delta > 0$ så att*

$$|g(x) - g(y) - g'(y)(x - y)| < \epsilon |x - y|$$

för alla $x, y \in I$ med $|x - y| < \delta$.

Proof. Vi kan på grund av likformig kontinuitet välja ett $\delta > 0$ så att $|g'(x) - g'(y)| < \epsilon$ om $x, y \in I$ och $|x - y| < \delta$. VI kan använda medelvärdessatsen för att finna ett ξ som ligger i $]x, y[$

så att $g(x) - g(y) = g'(\xi)(x - y)$. Detta ger

$$|g(x) - g(y) - g'(y)(x - y)| = |g'(\xi) - g'(y)||x - y| < \epsilon|x - y|,$$

den sista olikheten då $|\xi - y| < \delta$. □

Om vi nu betraktar vår rekursion så ser vi att vi valt x_n precis så att $g(x_{n-1}) + g'(x_{n-1})(x_n - x_{n-1}) = 0$ dvs så att

$$g(x_n) = g(x_n) - g(x_{n-1}) - g'(x_{n-1})(x_n - x_{n-1}).$$

Enligt lemmat kan vi göra detta litet i förhållande till $|x_n - x_{n-1}|$ och då kommer enligt rekursionsformeln $|x_{n+1} - x_n|$ att bli litet i förhållande till $|x_n - x_{n-1}|$ vilket betyder att x_n och x_{n+1} kommer närmare varandra och borde betyda att $\{x_i\}$ bildar en Cauchyföljd. Det är nu ganska lätt att skriva ner villkor som gör detta sant. Vi vill att två villkor ska vara uppfyllda. För det första vill vi att $|x_{n+1} - x_n|$ ska uppskattas uppifrån av en fix faktor som är mindre än 1 gånger $|x_n - x_{n-1}|$. Detta kommer att innebära att skillnaderna inte bara mellan x_n och x_{n+1} går mot 0 utan även skillnaderna mellan x_i och x_j gör det vilket betyder att vi får en Cauchyföljd. För det andra behöver vi en uppskattning av $|x_1 - x_0|$ som garanterar att x_n:en håller sig tillräckligt nära x_0 så att den uppskattning som behövs i första delen gäller. Vi antar att $g'(x_0) \neq 0$ och kan på grund av kontinuiteten hitta ett slutet intervall I sådant att x_0 ligger i det inre av I och sådant att $|g'(x)| \geq d > 0$ för $x \in I$. Vi använder sedan lemmat till att finna ett $\delta > 0$ så att om $|x-y| < \delta$ så gäller att $|g(x)-g(y)-g'(y)(x-y)| < 1/2d|x-y|$ och vidare så att om $|x-x_0| < \delta$ så ligger x i I. Vi definierar nu x_n genom den induktiva formeln

$$x_{n+1} = x_n - \frac{g(x_n)}{g'(x_n)}$$

och vill visa att $|x_{n+1} - x_n| < 1/2|x_n - x_{n-1}|$ vilket genom induktion ger att $|x_{n+1} - x_n| < 2^{-n}|x_1 - x_0|$ och upprepad användning av triangelolikheten ger

$$|x_{n+1} - x_0| < 1 + \frac{1}{2} + \cdots + \frac{1}{2^n} = 2\left(1 - \frac{1}{2^{n+1}}\right)|x_1 - x_0| < 2|x_1 - x_0|.$$

Om vi till att börja med antar att $|x_1 - x_0| < \delta/2$ så ger induktionen att $|x_n - x_0| < \delta$ och dessutom att $|x_n - x_{n-1}| < \delta$ så att $x_n \in I$ och

$$|g(x_n)| = |g(x_n) - g(x_{n-1}) - g'(x_{n-1})(x_n - x_{n-1})| < \frac{1}{2d}|x_n - x_{n-1}|$$

vilket ger att

$$|x_{n+1} - x_n| = \left|\frac{g(x_n)}{g'(x_n)}\right| < \frac{d}{2d}|x_n - x_{n-1}| = \frac{1}{2}|x_n - x_{n-1}|$$

vilket ger nästa steg i induktionen. Hela induktionen fungerar nu förutsatt att vi vet att $|x_1 - x_0| < \delta/2$. Nu har vi att $|x_1 - x_0| = |g(x_0)/g'(x_0)|$ så detta gäller om vi kan anta att x_0 är en så god approximation att $|g(x_0)| < d\delta/2$, vilket är vad vi gör.

Om nu $i > j$ så kan vi återigen använda att vi vet att $|x_{n+1} - x_n| < 2^{-n}|x_1 - x_0|$ till att få

$$|x_i - x_j| < \left(\frac{1}{2^j} + \frac{1}{2^{j+1}} + \cdots + \frac{1}{2^{i-1}}\right)|x_1 - x_0| < 2\frac{1}{2^j}|x_1 - x_0|$$

som visar att $\{x_i\}$ är en Cauchyföljd. Detta leder till följande resultat:

Proposition 7.2 *Låt $g\colon I \to \mathbf{R}$ vara en funktion från ett öppet intervall I med kontinuerlig derivata i intervallet. Låt $x_0 \in I$ och antag givet $\delta > 0$ och $d > 0$ sådana att*

1. $]x_0 - \delta, x_0 + \delta[\subseteq I$,

2. $|g'(y)| \geq d$ *för alla* $y \in I$,

3. $|g(x) - g(y) - g'(y)(x-y)| < d/2|x-y|$ *om* $x, y \in I$ *och* $|x-y| < \delta$,

4. *och* $|g(x_0)| \leq d\delta/2$.

Då finns ett $x \in I$ *med* $|x - x_0| \leq \delta$ *och* $g(x) = 0$.

Proof. Beviset har redan givits, låt oss bara påpeka några saker. För det första att villkoret att $|g(x_0)| \leq d\delta/2$ var till för att göra $|x_1 - x_0|$ vilket i sin tur var till för att vi skulle vara säkra på att $|x_n - x_0| < \delta$ så att vi sedan kunde använda de andra olikheterna för att uppskatta $|x_{n+1} - x_n|$. Uppskattningen $|x_n - x_0| < \delta$ ger oss sedan att $|x - x_0| \leq \delta$ genom att ta gränsvärde. Denna extra information är användbar i olika sammanhang. □

Man kan nu förstås undra över hur man hittar startvärdet x_0. Det svar som kommer att intressera oss för inversa funktionssatsen är att $g(x)$ är på formen $f(x) - y'$; vi försöker lösa ekvationen $f(x) = y'$ där y' är nära $y_0 := g(x_0)$. Genom att bara tillåta värden på y' sådana att $|y' - y_0| \leq d\delta/2$ så kan vi, eftersom $g(x_0) = y_0 - y'$, få en lösning x' med $f(x') = y'$. (Observera att $g(x) - g(y)$ och $g'(y)$ inte beror på y' så d och δ endast beror på f och ej på y'.)

Anmärkning: För våra syften (dvs inversa funktionssatsen) räcker det att visa att det finns en omgivning till y_0 i vilken det finns en lösning men observera att man sedan kan använda x' som nytt startvärde och hitta en omgivning till y' i vilken man kan lösa ekvationen. I många fall kan man genom att upprepa detta hitta lösningar där y inte nödvändigtvis ligger nära y_0. (Detta går inte alltid, man kan t ex ha otur och träffa på ett nollställe till $f'(x)$ på vägen.)

Vi ska avsluta denna del med några kommentarer som inte är direkt relevanta för inversa funktionssatsen. Att $g(x)$ är på formen $f(x) - y'$ med y' nära $f(x_0)$ är inte det enda sättet att få ett närmevärde till ett nollställe till g. Det finns nämligen en annan metod för att hitta en följd som konvergerar till ett nollställe till g, intervallinkapsling. Antag därför att vi har hittat x_0 och t_0 så att $g(x_0) < 0$ och $g(t_0) > 0$. Betrakta $g((x_0 + t_0)/2)$. Om detta värde är 0 är vi färdiga, om det är > 0 sätter vi $x_1 := x_0$ och $t_1 := (x_0 + t_0)/2$ och om det är < 0 sätter vi $x_1 := (x_0 + t_0)/2$ och $t_1 := t_0$. Då har vi att $g(x_1) < 0$ och $g(t_1) > 0$ och $|x_1 - t_1| = 1/2|x_0 - t_0|$. Upprepar vi denna procedur så får vi (x_n, t_n) med $g(x_n) < 0$ och $g(t_n) > 0$ samt $|x_n - y_n| = 2^{-n}|x_0 - t_0|$ så att x_n och t_n konvergerar mot ett gemensamt värde x som med nödvändighet är ett nollställe till g.

Man kan nu fråga sig vad det är för poäng med Newton-Raphsons metod när denna metod finns speciellt som vi ser att i båda fallen har vi att $|x_n - x|$ är i storleksordningen 2^{-n} så att vi i bägge fallen måste göra i storleksordningen n iterationer för att få n siffrors noggrannhet. Svaret är att i praktiken är ofta den uppskattning på felet vi gjort i Newton-Raphsons metod grovt överskattad. Om vi antar inte bara att g är deriverbar med kontinuerlig derivata utan två gånger deriverbar med begränsad andraderivata så får vi en mycket bättre uppskattning som väsentligen säger att antalet korrekta siffror fördubblas i varje iteration.

Övning 16: i) Antag att $g\colon I \to \mathbf{R}$ är en funktion på det öppna intervallet I med första- och andraderivata i varje punkt samt att vi har att $|f''(x)| \leq M$ för alla $x \in I$. Visa att

$$|g(x) - g(y) - g'(y)(x-y)| \leq \frac{M}{2}|x-y|^2$$

och

$$|g'(x) - g'(y)| \leq M|x-y|$$

för alla $x, y \in I$.

ii) Visa att om x_0 används som startpunkt för Newton-Raphsoniteration och om $x_n \in I$ för alla n så gäller att det finns en konstant C sådan att $|x_{n+1} - x_n| \leq (C|x_1 - x_0|)^{2^n}$. Ange villkor för att $C|x_1 - x_0| < 1$.

Anmärkning: En konvergens som denna där $|x_n - x| = O(\epsilon^{2^n})$ för något $\epsilon < 1$ kallas för *kvadratisk konvergens* och betyder att i stort antalet korrekta siffror fördubblas.

7.2 Newton-Raphsons metod i flera variabler

För att kunna generalisera Newton-Raphsons metod till flera variabler blir vi tvungna att glömma tolkningen som använde sig av tangenter. Detta är egentligen ganska naturligt eftersom vi såg i beviset av att Newton-Raphsons metod fungerade så använde vi direkt att definitionen av derivatan gav att $g(x) - g(y) - g'(y)(x-y)$ skulle vara litet (fast vi behövde viss likformighet i uppskattningen som går utanför definitionen). Så vi kan formulera det som att vi istället för att lösa ekvationen $g(x) = 0$ försöker lösa (i första steget) den ekvation vi får genom att approximera $g(x)$ med de två första termerna i dess Taylorutveckling dvs $g(x_0)+g'(x_0)(x-x_0) = 0$. Här är det inget som hindrar att g är en funktion från en öppen delmängd i $\mathbf{R}^n$ till $\mathbf{R}^n$. Den enda skillnaden är att vi istället för en linjär ekvation får ett linjärt ekvationssystem. Detta kan vi dock fortfarande lösa om vi antar att $g'(x_0)$ är en inverterbar linjär avbildning; vi får $x = x_0 - g'(x_0)^{-1}(g(x_0))$. Vi ska nu gå genom samma steg som vi gjorde i en variabel och se att det fungerar.

Lemma 7.3 *Låt $g\colon U \to \mathbf{R}^n$ vara en funktion från en öppen delmängd till $\mathbf{R}^n$ till $\mathbf{R}^n$ med kontinuerlig derivata. Då gäller att det för varje kompakt delmängd K i U och varje $\epsilon > 0$ finns ett $\delta > 0$ så att*

$$|g(x) - g(y) - g'(y)(x-y)| < \epsilon|x-y|$$

för alla $x, y \in K$ med $|x-y| < \delta$.

Proof. Vi kan på grund av likformig kontinuitet välja ett $\delta > 0$ så att $\|g'(x) - g'(y)\| < \epsilon$ om $x, y \in K$ och $|x-y| < \delta$. Fixera $y \in K$ och betrakta funktionen $h(x) := g(x) - g'(y)(x)$. Då gäller att $h'(x) = g'(x) - g'(y)$ så att $\|h'(z)\| < \epsilon$ om $|z-y| < \delta$. Enligt sats 9.19 har vi, för $|x-y| < \delta$

$$|g(x) - g(y) - g'(y)(x-y)| = |h(x) - h(y)| \leq \epsilon|x-y|.$$

□

Övning 17: Använd beviset i lemmat för att ge ett nytt bevis för Sats 9.21.

Vi kan använda detta lemma för att ge ett kriterium för att Newton-Raphsons metod i flera variabler ska fungera som är nästan identiskt med envariabelfallet.

Proposition 7.4 *Låt $g\colon U \to \mathbf{R}^n$ vara en funktion från en öppen delmängd $U \subseteq \mathbf{R}^n$ med kontinuerlig derivata. Låt $x_0 \in U$ och antag givet $\delta > 0$ och $d > 0$ sådana att*

1. *Klotet $D(x_0, \delta)$ ligger i U,*
2. *$g'(y)$ är inverterbar och $\|g'(y)^{-1}\| \leq 1/d$ för alla $y \in U$,*
3. *$|g(x) - g(y) - g'(y)(x-y)| \leq d/2|x-y|$ om $x, y \in U$ och $|x-y| < \delta$,*
4. *och $|g(x_0)| < d\delta/2$.*

Då finns ett $x \in U$ med $|x - x_0| < \delta$ och $g(x) = 0$.

Proof. Vi gör som i 1-variabelfallet och sätter

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

där det förstås ingår i vårt bevis att visa att $x_n \in U$ för alla U (om så är fallet ger alla förutsättningarna att $g'(x_n)$ är inverterbar för alla n). Vi konstaterar först att $|x_1 - x_0| \leq \|g'(x_0)^{-1}\||g(x_0)| < 1/d \cdot d\delta/2 = \delta/2$. Vi visar genom induktion att $|x_{n+1} - x_n| < 1/2|x_n - x_{n-1}|$. Om detta är sant för n så får vi genom induktion att $|x_{n+1} - x_n| < 2^{-n}|x_1 - x_0|$ och därmed att

$$|x_n - x_0| = |x_n - x_{n-1} + x_{n-1} - \ldots x_1 - x_0| \leq |x_n - x_{n-1}| + \ldots |x_1 - x_0| < (2 - 2^{-n+1})|x_1 - x_0| < \delta$$

så att x_n ligger i U. Vi har nu valt x_n så att $g(x_{n-1}) + g'(x_{n-1})(x_n - x_{n-1})$ vilket ger

$$|g(x_n)| = |g(x_n) - g(x_{n-1}) - g'(x_{n-1})(x_n - x_{n-1})| < d/2|x_n - x_{n-1}|$$

och detta ger

$$|x_{n+1} - x_n| = |g'(x_n)(g(x_n))| \leq \|g'(x_n)\||g(x_n)| < \frac{1}{d}\frac{d}{2}|x_n - x_{n-1}| = \frac{1}{2}|x_n - x_{n-1}|.$$

Samma argument som gav $|x_n - x_0| < (2 - 2^{-n+1})|x_1 - x_0|$ ger nu för $i > j$

$$|x_i - x_j| < 2\left(1 - \frac{1}{2^{i-j}}\right)\frac{1}{2^j}|x_1 - x_0|,$$

vilket visar att $\{x_i\}$ är en Cauchyföljd. Låt x vara gränsvärdet till denna följd. Då vi har att $|x_n - x_0| < (2 - 2^{-n+1})|x_1 - x_0|$ får vi $|x - x_0| \leq 2|x_1 - x_0| < \delta$ genom att låta n gå mot oändligheten och speciellt har vi att $x \in U$. Om vi låter $n \to \infty$ i ekvationen $x_{n+1} = x_n - g'(x_n)^{-1}(g(x_n))$ och utnyttjar att g och g' är kontinuerliga får vi $x = x - g'(x)(g(x))$ vilket ger $g'(x)(g(x)) = 0$, vilket då $g'(x)$ är inverterbar i sin tur ger att $g(x) = 0$. Vi har därmed visat propositionen. □

Anmärkning: Notera att det är väsentligen samma idé i detta bevis som i sats 9.23 i boken.

7.3 Inversa funktionssatsen

Inversa funktionssatsen är nu en ganska direkt konsekvens av Newton-Raphsons metod.

Sats 7.5 *Antag att $f \colon W \to \mathbf{R}^n$ är en funktion från en öppen delmängd W av $\mathbf{R}^n$ till $\mathbf{R}^n$ med kontinuerlig derivata. Antag att $f'(a)$ är inverterbar för något $a \in W$ och sätt $b := f(a)$. Då gäller att*

1. *det finns öppna delmängder U och V i $\mathbf{R}^n$ sådana att $a \in U$, $b \in V$ och f är bijektion från U till V och*
2. *om $g \colon V \to U$ är inversen till f (som existerar enligt första delen) så gäller att g har kontinuerlig derivata i hela V.*

Proof. Eftersom $f'(x)$ är kontinuerlig och de inverterbara matriserna är en öppen mängd i $L(\mathbf{R}^n)$ är de x för vilka $f'(x)$ är inverterbar en öppen mängd. Därför kan vi, genom att eventuellt ersätta U med en mindre mängd anta att $g'(x)$ är inverterbar. Vi har också att $x \mapsto \|f'(x)^{-1}\|$ är en kontinuerlig funktion så vi kan, genom att igen möjligen krympa U anta att $\|f'(x)^{-1}\| \leq 1/d$ för något fixt d. Vi kan vidare ersätta U med $D(a, \alpha)$ för något $\alpha > 0$ sådant att det slutna klotet $\overline{D}(a, \alpha)$ ligger i U. Detta betyder att vi kan tillämpa lemma 7.3 på den kompakta mängden $\overline{D}(a, \alpha)$ och hitta ett $\delta > 0$ sådant att

$$|f(x) - f(y) - f'(y)(x - y)| < \frac{d}{2}|x - y|$$

om $|x - y| < \delta$. Vi vill nu använda Newton-Raphsons metod på funktionen $g(x) := f(x) - y$ för olika y. Observera att $g'(x) = f'(x)$ och $g(x) - g(x') = f(x) - f(x')$ så att det enda vi behöver visa för att kunna använda proposition 7.4 på $g(x)$ med startvärde x_0 är att $g(x_0) = f(x_0) - y = y_0 - y$ uppfyller $|y_0 - y| < d\delta'/2$ där $\delta' \leq \delta$ valts så att $D(x_0, \delta') \subseteq U$ och $y_0 := f(x_0)$. Detta betyder att om vi väljer y tillräckligt nära y_0 så kommer detta alltid att vara fallet. Beviset fortgår nu som i sats 9.24. Vi kan dessutom konstatera att f är injektiv i U, ty om $f(x) = f(y)$ så har vi att

$$|f'(y)(x-y)| = |f(x) - f(y) - f'(y)(x-y)| < \frac{d}{2}|x-y| = \frac{d}{2}|f'(y)^{-1}f'(y)(x-y)| \leq \frac{1}{2}|f'(y)(x-y)|,$$

vilket ger $f'(y)(x - y) = 0$ vilket i sin tur ger $x - y = 0$ då $f'(y)$ är inverterbar. Slutligen för att visa att inversen h till f har kontinuerlig differential så räcker det att visa att $h'(y) = f'(h(y))^{-1}$, eftersom h' då är sammansättningen av kontinuerliga funktioner. Nu har vi för $y, y' \in V$ att

$$h(y') - h(y) - f'(h(y))^{-1}(y' - y) = f'(h(y))^{-1}\left(f'(h(y))(h(y') - h(y) - (f(h(y')) - f(h(y))))\right)$$

vilket ger

$$|h(y') - h(y) - f'(h(y))^{-1}(y' - y)| \leq \|f'(h(y))^{-1}\| |f(x') - f(x) - f(x)(x' - x)|,$$

där $x = h(y)$ och $x' = h(y')$. Det följer från proposition 7.4 och att $g(x') - g(x) = y' - y$ att det finns en konstant C sådan att $|x' - x| \leq C|y' - y|$ vilket ger

$$\frac{|h(y') - h(y) - f'(h(y))^{-1}(y' - y)|}{|y' - y|} \leq \frac{\|f'(h(y))^{-1}\|}{C} \frac{|f(x') - f(x) - f(x)(x' - x)|}{|x' - x|}$$

och då $|x' - x| \leq C|y' - y|$ ger att $x' \to x$ när $y' \to y$ så får vi att högerledet går mot 0 när $y' \to y$ (eftersom f är differentierbar) vilket visar att $h'(y) = f'(h(y))^{-1}$. $\square$

Om vi jämför detta bevis med bokens bevis så ser vi att skillnaden är att vi i bokens bevis använder oss av iterationen

$$x_{n+1} = x_n + f'(x_0)^{-1}(y - f(x_n))$$

medan iterationen i Newtons metod är

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

Om man sedan tittar på beviset att dessa två metoder konvergerar så ser man att bevisen är rätt analoga och speciellt att uppskattningarna av konvergenshastigheten är ungefär densamma. Närmare bestämt får vi *linjär konvergens* dvs $x_n - x = \mathcal{O}(\epsilon^n)$ (vilket ungefär betyder att varje iteration ger en ny korrekt siffra). Det får en att undra vad det är för poäng med Newtons metod eftersom den verkar mer komplicerad, man måste ju i varje steg beräkna inversen till $f'(x_n)$ medan i bokens metod räcker det att beräkna $f'(x_0)^{-1}$ en gång för alla. Anledningen till att Newtons metod verkligen är intressant är den samma som i 1-variabelfallet: Om man lägger på extra villkor på f så kan man få kvadratisk konvergens, vilket visas i följande övningar.

Övning 18: Antag att för $f\colon [a, b] \to \mathbf{R}^n$ så är $f^{(n-1)}$ är kontinuerlig på hela $[a, b]$ och att $f^{(n)}(t)$ existerar för alla $t \in]a, b[$. Visa att det finns $x \in]a, b[$ sådant att

$$\left| f(b) - \sum_{k=0}^{n-1} \frac{(b-a)^k}{k!} f^k(a) \right| \leq \frac{b-a}{n!} |f^{(n)}(x)|.$$

Ledning: Imitera beviset av Sats 5.19 kombinerat med Taylors sats.

References