Innhold

Kapittel 5

Ikke-lineære ligningssystemer

5.1 Litt topologi i $\mathbb{R}^m$

Når vi arbeider med funksjoner av én variabel, vil definisjonsområdet som regel være et intervall eller en enkel sammensetning av intervaller. For funksjoner av flere variable finnes det mange flere muligheter for hvordan definisjonsområdet kan være, og vi må derfor være litt mer formelle i vår omgang med mengder. I denne seksjonen skal vi innføre *åpne* og *lukkede* mengder. Disse mengdene spiller på mange måter den samme rollen i flervariabelteori som åpne og lukkede intervaller spiller i teorien for funksjoner av en variabel. En liten bemerkning om notasjon før vi starter: I dette kapitlet vil vi stort sett referere til det underliggende rommet som $\mathbb{R}^m$ og ikke $\mathbb{R}^n$ slik vi gjør i de andre kapitlene. Det er rett og slett fordi vi kommer til å arbeide mye med følger, og ønsker å ha bokstaven n ledig for å referere til det n-te leddet x_n i en følge.

Vi begynner med å repetere litt fra kapittel 2. Husk at *(den åpne) kulen om* $\mathbf{a} \in \mathbb{R}^m$ med radius r er mengden

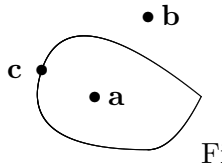
$$B(\mathbf{a}, r) = \{\mathbf{x} \in \mathbb{R}^m \ : \ |\mathbf{x} - \mathbf{a}| < r\}$$

bestående av de punktene i $\mathbb{R}^m$ som har avstand mindre enn r til punktet $\mathbf{a}$. I $\mathbb{R}^3$ er dette virkelig en (åpen) kule i tradisjonell forstand, mens det i $\mathbb{R}^2$ er en (åpen) sirkelskive og i $\mathbb{R}$ et åpent intervall. Vi velger å bruke "kule" som et fellesord i alle dimensjoner selv om det til å begynne med kan virke litt uvant når vi arbeider i planet eller på tallinjen. De fleste illustrasjonene våre vil være i planet, og der vil kuler fremstå som sirkler. Noen ganger skal vi også ha bruk for de *lukkede kulene*

$$\overline{B}(\mathbf{a}, r) = \{\mathbf{x} \in \mathbb{R}^m \ : \ |\mathbf{x} - \mathbf{a}| \leq r\},$$

men det er de åpne som vil spille hovedrollen til å begynne med.

Figur 1 viser en mengde i planet (området innenfor kurven) og tre punkter $\mathbf{a}$, $\mathbf{b}$ og $\mathbf{c}$.



Figur 1

De tre punktene ligger på forskjellig måte i forhold til mengden — **a** ligger klart på innsiden av mengden, **b** ligger klart på utsiden, mens **c** ligger på grensen mellom mengden og omgivelsene. Vi kaller **a** et *indre punkt*, **b** et *ytre punkt* og **c** et *randpunkt* for mengden. I høyere dimensjoner kan vi ikke støtte oss på figurer, og vi trenger derfor en mer formell definisjon av indre punkter, ytre punkter og randpunkter:

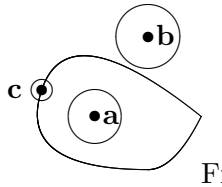
Definisjon 5.1.1 *La A være en delmengde av $\mathbb{R}^m$.*

(i) Et punkt $\mathbf{a} \in \mathbb{R}^m$ kalles et indre punkt *for A dersom det finnes en kule $B(\mathbf{a}, r)$ om $\mathbf{a}$ som bare inneholder punkter som er med i A.*

(ii) Et punkt $\mathbf{b} \in \mathbb{R}^m$ kalles et ytre punkt *for A dersom det finnes en kule $B(\mathbf{b}, r)$ om $\mathbf{b}$ som ikke inneholder noen punkter som er med i A.*

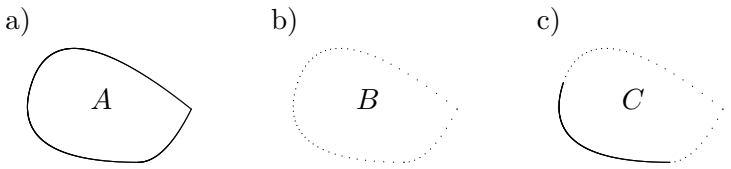
(iii) Et punkt $\mathbf{c} \in \mathbb{R}^m$ kalles et randpunkt *for A dersom enhver kule $B(\mathbf{c}, r)$ om $\mathbf{c}$ inneholder både punkter som er med i A og punkter som ikke er med i A.*

Legg merke til at de tre delene av definisjonen uttømmer alle muligheter, så et punkt i $\mathbb{R}^m$ må enten være et indre punkt, et ytre punkt eller et randpunkt for A. I figur 2 har vi illustrert definisjonen ved å legge "kuler" (dvs. sirkler) rundt punktene **a**, **b** og **c**.



Figur 2

Et indre punkt hører alltid med til mengden A, mens et ytre punkt aldri hører med til mengden. For randpunkter er det ingen generell regel; de vil noen ganger høre med til mengden og andre ganger ikke. Figur 3 viser dette for tre mengder i planet. I punkt a) er randen tegnet med en hel strek — det markerer at alle punktene på randen hører med til mengden A. En slik mengde kalles *lukket*. I punkt b) er randen stiplet — det markerer at ingen av punktene på randen hører med til mengden B. En slik mengde kalles *åpen*. I punkt c) er noe av randen heltrukken og resten stiplet — det markerer at noen av punktene på randen hører med til mengden C, mens andre ikke gjør det. En slik mengde er hverken åpen eller lukket (på figuren har vi kalt den "halvåpen", men det er ingen offisiell betegnelse).



Figur 3: Lukket, åpen og "halvåpen" mengde

La oss skrive opp definisjonen av åpne og lukkede mengder litt mer formelt:

Definisjon 5.1.2 *En mengde $A \subset \mathbb{R}^m$ er* lukket *dersom den inneholder alle sine randpunkter, og* åpen *dersom den ikke inneholder noen randpunkter.*

Åpne og lukkede mengder spiller omtrent samme rolle i teorien for funksjoner av flere variable som åpne og lukkede intervaller gjør i teorien for funksjoner av én variabel. Det er imidlertid en viktig forskjell — åpne og lukkede mengder kan ha mange forskjellige former, og det gjør at flervariabelteorien har en del geometriske komplikasjoner som envariabelteorien ikke har.

Følger i $\mathbb{R}^m$

I resten av denne seksjonen skal vi se litt på konvergens av følger i $\mathbb{R}^m$. Dette er et begrep som kommer til å stå sentralt i de neste seksjonene.

En *følge* i $\mathbb{R}^m$ er bare en uendelig sekvens

$$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_n, \ldots$$

av elementer $\mathbf{x}_n \in \mathbb{R}^m$. Akkurat som for tallfølger bruker vi $\{\mathbf{x}_n\}$ som en kortfattet skrivemåte for disse følgene. Vi tillater at følgen starter med andre indekser enn 1, f.eks.

$$\mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5, \ldots, \mathbf{x}_n, \ldots$$

eller

$$\mathbf{x}_{-3}, \mathbf{x}_{-2}, \mathbf{x}_{-1}, \ldots, \mathbf{x}_n, \ldots$$

Dersom det er viktig å vite hvor følgen starter, kan vi markere det i den kortfattede skrivemåten ved å skrive $\{\mathbf{x}_n\}_{n=3}^{\infty}$ og $\{\mathbf{x}_n\}_{n=-3}^{\infty}$ for de to følgene ovenfor.

Intuitivt sier vi at en følge $\{\mathbf{x}_n\}$ i $\mathbb{R}^m$ nærmer seg punktet $\mathbf{a}$ som grenseverdi dersom vi kan få avstanden mellom $\mathbf{x}_n$ og $\mathbf{a}$ så liten vi måtte ønske ved å velge n tilstrekkelig stor. Vi kan formulere dette på akkurat samme måte som i det en-dimensjonale tilfellet (sammenlign med definisjon 4.3.1 i *Kalkulus*):

Definisjon 5.1.3 *Følgen* $\{\mathbf{x}_n\}$ *i* $\mathbb{R}^m$ konvergerer mot *punktet* $\mathbf{a} \in \mathbb{R}^m$ *dersom det til enhver* $\epsilon > 0$ *finnes en* $N \in \mathbb{N}$ *slik at* $|\mathbf{x}_n - \mathbf{a}| < \epsilon$ *for alle* $n \geq N$. *Vi skriver*

$$\lim_{n\to\infty} \mathbf{x}_n = \mathbf{a}$$

Vi har akkurat de samme regnereglene som for tallfølger:

Setning 5.1.4 *Anta at* $\{\mathbf{x}_n\}$ *og* $\{\mathbf{y}_n\}$ *er to følger i* $\mathbb{R}^m$ *som konvergerer mot henholdsvis* $\mathbf{x}$ *og* $\mathbf{y}$. *Da har vi:*

(i) Følgen $\{c\,\mathbf{x}_n\}$ *konvergerer for ethvert tall* c, *og* $\lim_{n\to\infty}(c\,\mathbf{x}_n) = c\,\mathbf{x}$

(ii) Følgen $\{\mathbf{x}_n + \mathbf{y}_n\}$ *konvergerer, og* $\lim_{n\to\infty}(\mathbf{x}_n + \mathbf{y}_n) = \mathbf{x} + \mathbf{y}$

(iii) Følgen $\{\mathbf{x}_n - \mathbf{y}_n\}$ *konvergerer, og* $\lim_{n\to\infty}(\mathbf{x}_n - \mathbf{y}_n) = \mathbf{x} - \mathbf{y}$

(iv) Følgen $\{\mathbf{x}_n \cdot \mathbf{y}_n\}$ *konvergerer, og* $\lim_{n\to\infty}(\mathbf{x}_n \cdot \mathbf{y}_n) = \mathbf{x} \cdot \mathbf{y}$ *(legg merke til at dette er en* tallfølge *og ikke en følge av vektorer).*

Bevis: Bevisene er nesten identiske med de tilsvarende bevisene for tallfølger (se 4.3.3 i *Kalkulus*), den eneste forskjellen er at vi nå må bruke Schwarz' ulikhet og trekantulikheten for vektorer istedenfor de tilsvarende ulikhetene for tall. For å illustrere bruken av disse ulikhetene, tar vi med bevisene for (ii) og (iv).

(ii) Vi må vise at gitt en $\epsilon > 0$, kan vi alltid finne en $N \in \mathbb{N}$ slik at $|(\mathbf{x}+\mathbf{y}) - (\mathbf{x}_n + \mathbf{y}_n)| < \epsilon$ for alle $n \geq N$. Det første vi gjør er å omgruppere leddene slik at vi kan behandle $\{\mathbf{x}_n\}$ og $\{\mathbf{y}_n\}$ hver for seg:

$$|(\mathbf{x} + \mathbf{y}) - (\mathbf{x}_n + \mathbf{y}_n)| = |(\mathbf{x} - \mathbf{x}_n) + (\mathbf{y} - \mathbf{y}_n)| \leq |\mathbf{x} - \mathbf{x}_n| + |\mathbf{y} - \mathbf{y}_n|$$

der vi i det siste skrittet har brukt trekantulikheten (setning 1.2.4). Siden $\lim_{n\to\infty} \mathbf{x}_n = \mathbf{x}$, må det finnes en $N_1 \in \mathbb{N}$ slik at $|\mathbf{x} - \mathbf{x}_n| < \frac{\epsilon}{2}$ for alle $n \geq N_1$, og siden $\lim_{n\to\infty} \mathbf{y}_n = \mathbf{y}$, må det finnes en $N_2 \in \mathbb{N}$ slik at $|\mathbf{y} - \mathbf{y}_n| < \frac{\epsilon}{2}$ for alle $N \geq N_2$. Velger vi N lik det største av tallene N_1, N_2, ser vi at når $n \geq N$, er

$$|(\mathbf{x} + \mathbf{y}) - (\mathbf{x}_n + \mathbf{y}_n)| \leq |\mathbf{x} - \mathbf{x}_n| + |\mathbf{y} - \mathbf{y}_n| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$$

Dermed er (ii) bevist.

(iv) Vi må vise at gitt en $\epsilon > 0$, kan vi alltid finne en $N \in \mathbb{N}$ slik at $|\mathbf{x} \cdot \mathbf{y} - \mathbf{x}_n \cdot \mathbf{y}_n| < \epsilon$ for alle $n \geq N$. Vi bruker først trikset med å legge til og trekke fra leddet $\mathbf{x} \cdot \mathbf{y}_n$, og benytter deretter trekantulikheten og Schwarz' ulikhet (setning 1.2.3):

$$|\mathbf{x} \cdot \mathbf{y} - \mathbf{x}_n \cdot \mathbf{y}_n| = |\mathbf{x} \cdot \mathbf{y} - \mathbf{x} \cdot \mathbf{y}_n + \mathbf{x} \cdot \mathbf{y}_n - \mathbf{x}_n \cdot \mathbf{y}_n| \leq$$

$$\leq |\mathbf{x}\cdot\mathbf{y}-\mathbf{x}\cdot\mathbf{y}_n|+|\mathbf{x}\cdot\mathbf{y}_n-\mathbf{x}_n\cdot\mathbf{y}_n|\leq|\mathbf{x}||\mathbf{y}-\mathbf{y}_n|+|\mathbf{x}-\mathbf{x}_n||\mathbf{y}_n|$$

Vi skal nå vise at vi kan få hvert av de to leddene $|\mathbf{x}||\mathbf{y}-\mathbf{y}_n|$ og $|\mathbf{x}-\mathbf{x}_n||\mathbf{y}_n|$ mindre enn $\frac{\epsilon}{2}$ ved å velge n stor nok. Det første er opplagt mindre enn $\frac{\epsilon}{2}$ hvis $|\mathbf{x}|=0$, så vi kan konsentrere oss om tilfellet $\mathbf{x}\neq 0$. Siden $\lim_{n\to\infty}\mathbf{y}_n=\mathbf{y}$, må det finnes en $N_1\in\mathbb{N}$ slik at $|\mathbf{y}-\mathbf{y}_n|<\frac{\epsilon}{2|\mathbf{x}|}$ for alle $n\geq N_1$. Dermed er $|\mathbf{x}||\mathbf{y}-\mathbf{y}_n|<|\mathbf{x}|\cdot\frac{\epsilon}{2|\mathbf{x}|}=\frac{\epsilon}{2}$ når $N\geq N_1$.

Det andre leddet $|\mathbf{x}-\mathbf{x}_n||\mathbf{y}_n|$ er litt verre. Vi observerer først at siden $\lim_{n\to\infty}\mathbf{y}_n=\mathbf{y}$, finnes det et tall N_2 slik at $|\mathbf{y}_n-\mathbf{y}|\leq 1$ når $n\geq N_2$. Dermed er

$$|\mathbf{y}_n|=|\mathbf{y}+(\mathbf{y}_n-\mathbf{y})|\leq|\mathbf{y}|+|\mathbf{y}_n-\mathbf{y}|\leq|\mathbf{y}|+1$$

Siden $\lim_{n\to\infty}\mathbf{x}_n=\mathbf{x}$, må det finnes en $N_3\in\mathbb{N}$ slik at $|\mathbf{x}-\mathbf{x}_n|<\frac{\epsilon}{2(|\mathbf{y}|+1)}$ for alle $n\geq N_3$. Hvis n er større enn eller lik både N_2 og N_3, er dermed

$$|\mathbf{x}-\mathbf{x}_n||\mathbf{y}_n|<\frac{\epsilon}{2(|\mathbf{y}|+1)}\cdot(|\mathbf{y}|+1)=\frac{\epsilon}{2}$$

Velger vi nå N til å være det største av tallene N_1,N_2,N_3, ser vi at for $n\geq N$ er

$$|\mathbf{x}\cdot\mathbf{y}-\mathbf{x}_n\cdot\mathbf{y}_n|\leq|\mathbf{x}||\mathbf{y}-\mathbf{y}_n|+|\mathbf{x}-\mathbf{x}_n||\mathbf{y}_n|<\frac{\epsilon}{2}+\frac{\epsilon}{2}=\epsilon$$

Dermed er (iv) bevist. □

Vi tok med bevisene ovenfor for å demonstrere hvordan trekantulikheten og Schwarz' ulikhet ofte brukes. Det viser seg at vi kan bevise disse resultatene vel så enkelt ved å føre dem tilbake til tilsvarende resultater for tallfølger. Det neste resultatet vil gi oss det redskapet vi trenger. Litt notasjon før vi begynner: Dersom $\{\mathbf{x}_n\}$ er en følge i $\mathbb{R}^m$, skriver vi koordinatene til $\mathbf{x}_n$ slik:

$$\mathbf{x}_n=(x_1^{(n)},x_2^{(n)},\ldots,x_m^{(n)})$$

Vi skriver indeksen n oppe for ikke å blande den sammen med koordinatene til $\mathbf{x}_n$, og vi putter den inn i en parentes for å gjøre det klart at den ikke er en eksponent.

Setning 5.1.5 *Anta at $\{\mathbf{x}\}$ er en følge i $\mathbb{R}^m$ med komponenter*

$$\mathbf{x}_n=(x_1^{(n)},x_2^{(n)},\ldots,x_m^{(n)})$$

og at $\mathbf{x}\in\mathbb{R}^m$ har komponenter $\mathbf{x}=(x_1,x_2,\ldots,x_m)$. Da er

$$\lim_{n\to\infty}\mathbf{x}_n=\mathbf{x}$$

hvis og bare hvis

$$\lim_{n\to\infty}x_i^{(n)}=x_i \text{ for alle } i=1,2,\ldots m$$

Med andre ord: $\{\mathbf{x}_n\}$ konvergerer mot $\mathbf{x}$ hvis og bare hvis hver komponent i $\mathbf{x}_n$ konvergerer mot tilsvarende komponent i $\mathbf{x}$.

Dersom du synes det er vanskelig å forstå hva setning sier, er det lurt å ta en kikk på eksemplet nedenfor før du går videre.

Bevis: Anta først at $\lim_{n\to\infty} \mathbf{x}_n = \mathbf{x}$. Vi skal vise at $\lim_{n\to\infty} x_i^{(n)} = x_i$. Det betyr at gitt en $\epsilon > 0$, må vi vise at det alltid finnes en $N \in \mathbb{N}$ slik at $|x_i^{(n)} - x_i| < \epsilon$ for alle $n \geq N$. Siden $\lim_{n\to\infty} \mathbf{x}_n = \mathbf{x}$, finnes det en $N \in \mathbb{N}$ slik at $|\mathbf{x}_n - \mathbf{x}| < \epsilon$ når $n \geq N$. Siden

$$|x_i^{(n)} - x_i| = \sqrt{(x_i^{(n)} - x_i)^2} \leq$$

$$\leq \sqrt{(x_1^{(n)} - x_1)^2 + \cdots + (x_i^{(n)} - x_i)^2 + \cdots + (x_m^{(n)} - x_m)^2} = |\mathbf{x}_n - \mathbf{x}|$$

medfører dette at $|x_i^{(n)} - x_i| < \epsilon$ for alle $n \geq N$, og det er akkurat det vi skulle vise.

Anta så at $\lim_{n\to\infty} x_i^{(n)} = x_i$ for $i = 1, 2, \ldots, m$. Vi skal vise at at $\lim_{n\to\infty} \mathbf{x}_n = \mathbf{x}$. Det betyr at gitt en $\epsilon > 0$, må vi produsere en $N \in \mathbb{N}$ slik at $|\mathbf{x}_n - \mathbf{x}| < \epsilon$ når $n \geq N$. Siden $\lim_{n\to\infty} x_i^{(n)} = x_i$, finnes det for hver i en $N_i \in \mathbb{N}$ slik at $|x_i^{(n)} - x_i| < \frac{\epsilon}{\sqrt{m}}$ når $n \geq N_i$. La N være den største av $N_1, N_2, \ldots, N_m$. For $n \geq N$ er da

$$|\mathbf{x}_n - \mathbf{x}| = \sqrt{(x_1^{(n)} - x_1)^2 + \cdots + (x_i^{(n)} - x_i)^2 + \cdots + (x_m^{(n)} - x_m)^2} \leq$$

$$\leq \sqrt{\left(\frac{\epsilon}{\sqrt{m}}\right)^2 + \cdots + \left(\frac{\epsilon}{\sqrt{m}}\right)^2 + \cdots + \left(\frac{\epsilon}{\sqrt{m}}\right)^2} = \sqrt{m\frac{\epsilon^2}{m}} = \sqrt{\epsilon^2} = \epsilon$$

□

Eksempel 1: Finn grenseverdien

$$\lim_{n\to\infty} \begin{pmatrix} \frac{n^2}{n^2+1} \\ n\sin\left(\frac{1}{n}\right) \\ \left(1+\frac{2}{n}\right)^n \end{pmatrix}$$

Ifølge setningen ovenfor behøver vi bare å regne ut grensen til hver komponent. Den første er enkel:

$$\lim_{n\to\infty} \frac{n^2}{n^2+1} \overset{\text{L'H}}{=} \lim_{n\to\infty} \frac{2n}{2n} = 1$$

Den andre går også greit:

$$\lim_{n\to\infty} n\sin\left(\frac{1}{n}\right) = \lim_{n\to\infty} \frac{\sin\left(\frac{1}{n}\right)}{\frac{1}{n}} \overset{\text{L'H}}{=} \lim_{n\to\infty} \frac{\cos\left(\frac{1}{n}\right)\left(-\frac{1}{n^2}\right)}{\left(-\frac{1}{n^2}\right)} = \lim_{n\to\infty} \cos\left(\frac{1}{n}\right) = 1$$

Den tredje skriver vi først om:

$$\left(1+\frac{2}{n}\right)^n = e^{n\log\left(1+\frac{2}{n}\right)}$$

og bruker deretter L'Hôpitals regel på eksponenten:

$$\lim_{n\to\infty} n\log\left(1+\frac{2}{n}\right) = \lim_{n\to\infty}\frac{\log\left(1+\frac{2}{n}\right)}{\frac{1}{n}} \overset{\text{L'H}}{=} \lim_{n\to\infty}\frac{\frac{1}{\left(1+\frac{2}{n}\right)}\cdot\left(-\frac{2}{n^2}\right)}{-\frac{1}{n^2}} = 2$$

Altså er

$$\lim_{n\to\infty}\left(1+\frac{2}{n}\right)^n = e^2$$

Kombinerer vi alt dette, får vi

$$\lim_{n\to\infty}\begin{pmatrix} \frac{n^2}{n^2+1} \\ n\sin\left(\frac{1}{n}\right) \\ \left(1+\frac{2}{n}\right)^n \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ e^2 \end{pmatrix}$$

♣

Det neste resultatet gir et nytt innblikk i hvorfor lukkede mengder kalles "lukkede":

Setning 5.1.6 *Anta at $A \subset \mathbb{R}^m$ er lukket, og at $\{\mathbf{x}_n\}$ er en følge fra A som konvergerer mot et punkt $\mathbf{x}$. Da er $\mathbf{x} \in A$.*

Bevis: Anta for motsigelse at $\mathbf{x} \notin A$. Siden A er lukket, må $\mathbf{x}$ da være et ytre punkt (alle randpunktene hører jo med til A når A er lukket). Det betyr at det finnes en kule $B(\mathbf{x}, r)$ om $\mathbf{x}$ som ikke inneholder noe punkt fra A. Spesielt kan ingen av leddene i følgen $\{\mathbf{x}_n\}$ ligge i $B(\mathbf{x}, r)$. Det betyr at $|\mathbf{x}_n - \mathbf{x}| \geq r$ for alle n, og følgelig kan ikke $\{\mathbf{x}_n\}$ konvergere mot $\mathbf{x}$. Dette gir oss den selvmotsigelsen vi er på jakt etter. □

Det siste resultatet i denne seksjonen binder sammen konvergens og kontinuitet (se *Kalkulus*, setning 5.1.10 for den endimensjonale versjonen).

Setning 5.1.7 *Anta at $\mathbf{F} : A \to \mathbb{R}^m$ er en funksjon av flere variable, og at $\mathbf{a}$ er et punkt i definisjonsområdet A til $\mathbf{F}$. Da er $\mathbf{F}$ kontinuerlig i $\mathbf{a}$ hvis og bare hvis $\mathbf{F}(\mathbf{x}_n) \to \mathbf{F}(\mathbf{a})$ for* alle *følger $\{\mathbf{x}_n\}$ fra A slik at $\mathbf{x}_n \to \mathbf{a}$.*

Bevis: Anta først at $\mathbf{F}$ er kontinuerlig i $\mathbf{a}$. Gitt en følge $\{\mathbf{x}_n\}$ av punkter i A slik at $\mathbf{x}_n \to \mathbf{a}$, må vi vise at det for enhver $\epsilon > 0$, finnes en $N \in \mathbb{N}$ slik

at $|\mathbf{F}(\mathbf{x}_n) - \mathbf{F}(\mathbf{a})| < \epsilon$ for alle $n \geq N$. Siden $\mathbf{F}$ er kontinuerlig i $\mathbf{a}$, finnes det en $\delta > 0$ slik at $|\mathbf{F}(\mathbf{y}) - \mathbf{F}(\mathbf{a})| < \epsilon$ for alle $\mathbf{y} \in A$ slik at $|\mathbf{y} - \mathbf{a}| < \delta$. Siden $\mathbf{x}_n \to \mathbf{a}$, finnes det en $N \in \mathbb{N}$ slik at $|\mathbf{x}_n - \mathbf{a}| < \delta$ når $n \geq N$. Men dette betyr at når $n \geq N$, så er $|\mathbf{x}_n - \mathbf{a}| < \delta$ og følgelig $|\mathbf{F}(\mathbf{x}_n) - \mathbf{F}(\mathbf{a})| < \epsilon$.

Anta så at $\mathbf{F}$ *ikke* er kontinuerlig i $\mathbf{a}$. Vi må vise at det finnes i hvert fall én følge $\{\mathbf{x}_n\}$ fra A slik at $\mathbf{x}_n \to \mathbf{a}$, men $\mathbf{F}(\mathbf{x}_n) \not\to \mathbf{F}(\mathbf{a})$. Siden $\mathbf{F}$ ikke er kontinuerlig i $\mathbf{a}$, må det finnes en $\epsilon > 0$ slik at uansett hvor liten vi velger $\delta > 0$, så eksisterer det en $\mathbf{x} \in A$ slik at $|\mathbf{x} - \mathbf{a}| < \delta$, men $|\mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{a})| \geq \epsilon$. Velger vi $\delta = \frac{1}{n}$, finner vi på denne måten et punkt $\mathbf{x}_n \in A$ slik at $|\mathbf{x}_n - \mathbf{a}| < \frac{1}{n}$, men $|\mathbf{F}(\mathbf{x}_n) - \mathbf{F}(\mathbf{a})| \geq \epsilon$. Følgen $\{\mathbf{x}_n\}$ konvergerer mot $\mathbf{a}$, men $\{\mathbf{F}(\mathbf{x}_n)\}$ kan ikke konvergere mot $\mathbf{F}(\mathbf{a})$ siden $|\mathbf{F}(\mathbf{x}_n) - \mathbf{F}(\mathbf{a})| \geq \epsilon$ for alle n. □

Vi avslutter denne seksjonen med et eksempel som antyder hvordan grenseverdier til følger dukker opp i mer virkelighetsnære problemstillinger.

Eksempel 2: I eksempel 4 i seksjon 1.5 så vi på fordelingen av handlevogner i tre stativer. Dersom fordelingen om morgenen er gitt ved en vektor

$$\mathbf{x}_0 = \begin{pmatrix} x_0 \\ y_0 \\ z_0 \end{pmatrix},$$

vil fordelingen om kvelden være gitt ved vektoren $\mathbf{x}_1 = A\mathbf{x}_0$ der A er matrisen

$$A = \begin{pmatrix} 0.7 & 0.3 & 0.4 \\ 0.2 & 0.5 & 0.2 \\ 0.1 & 0.2 & 0.4 \end{pmatrix}$$

Dersom vi ikke rører handlevognene i løpet av natten, vil vi starte neste morgen med fordelingen $\mathbf{x}_1$, og fordelingen neste kveld blir $\mathbf{x}_2 = Ax_1$. Lar vi denne stå urørt til morgenen etter, vil vi neste kveld ha fordelingen $\mathbf{x}_3 = A\mathbf{x}_2$ osv. På denne måten får vi en følge av fordelinger $\{\mathbf{x}_n\}$. Dersom du faktisk regner ut disse vektorene, vil du raskt få en mistanke om at de konvergerer mot en likevektstilstand $\mathbf{b}$, dvs. en fordeling som er den samme om kvelden som om morgenen. Legg også merke til at $\mathbf{x}_1 = A\mathbf{x}_0$, $\mathbf{x}_2 = A\mathbf{x}_1 = A(A\mathbf{x}_0) = A^2\mathbf{x}_0$ osv., slik at $\mathbf{x}_n = A^n\mathbf{x}_0$. Følger som oppstår ved *iterasjon* (gjentagelse) på denne måten er viktige i mange anvendelser. Vi skal møte dem igjen i seksjon 5.4.

♣

5.2 Kompletthet av $\mathbb{R}^m$

I denne seksjonen skal vi se på konvergens av følger fra en teoretisk synsvinkel, men resultatene vi kommer frem til, har stor praktisk nytte — blant

annet når vi skal bruke datamaskiner til å finne numeriske løsninger på matematiske problemer. Du vil få se flere eksempler på dette i de neste seksjonene.

Vi starter med en følge av punkter i $\mathbb{R}^m$

$$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6, \mathbf{x}_7, \mathbf{x}_8, \mathbf{x}_9, \mathbf{x}_{10}, \mathbf{x}_{11}, \ldots, \mathbf{x}_n, \ldots$$

og tenker oss at vi plukker ut uendelig mange av punktene i følgen (men sannsynligvis ikke alle). Vi kan for eksempel begynne med å plukke ut de elementene som vi har satt en strek under her:

$$\mathbf{x}_1, \mathbf{x}_2, \underline{\mathbf{x}_3}, \mathbf{x}_4, \underline{\mathbf{x}_5}, \underline{\mathbf{x}_6}, \mathbf{x}_7, \mathbf{x}_8, \mathbf{x}_9, \underline{\mathbf{x}_{10}}, \mathbf{x}_{11}, \ldots, \mathbf{x}_n, \ldots$$

På denne måten får vi en ny følge som begynner

$$\mathbf{x}_3, \mathbf{x}_5, \mathbf{x}_6, \mathbf{x}_{10}, \ldots$$

Denne nye følgen kalles en *delfølge* av den opprinnelige følgen.

La oss se litt mer formelt på dette. Dersom de leddene som vi plukker ut i den opprinnelige følgen, har nummer $n_1, n_2, n_3, \ldots, n_k, \ldots$, vil den nye følgen (delfølgen) ha elementene $x_{n_1}, x_{n_2}, x_{n_3}, \ldots, x_{n_k}, \ldots$. Kaller vi delfølgen $\{\mathbf{y}_k\}$, har vi altså $\{\mathbf{y}_k\} = \{\mathbf{x}_{n_k}\}$. Vi kan nå gi den presise definisjonen av en delfølge:

Definisjon 5.2.1 *Anta at $\{\mathbf{x}_n\}$ er en følge av punkter i $\mathbb{R}^m$ og at*

$$n_1 < n_2 < n_3 < n_4 < \ldots < n_k < \ldots$$

er en strengt voksende følge av naturlige tall. Da kalles følgen $\{\mathbf{y}_k\}$ der $\mathbf{y}_k = \mathbf{x}_{n_k}$ en delfølge *av $\{\mathbf{x}_n\}$.*

Vi skal være interessert i samspillet mellom konvergens av følger og konvergens av delfølger. Det første resultatet er enkelt.

Setning 5.2.2 *Anta at en følge $\{\mathbf{x}_n\}$ i $\mathbb{R}^m$ konvergerer mot et punkt $\mathbf{x}$. Da konvergerer også alle delfølger av $\{\mathbf{x}_n\}$ mot $\mathbf{x}$.*

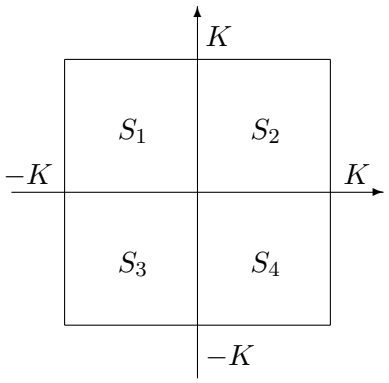
Bevis: Beviset overlates til leserne. □

Vi sier at en følge $\{\mathbf{x}_n\}$ i $\mathbb{R}^m$ er *begrenset* dersom det finnes et tall K slik at $|\mathbf{x}_n| \leq K$ for alle n. Det neste resultatet er nøkkelen til resten av denne seksjonen.

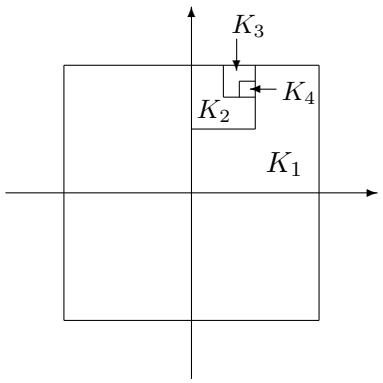
Teorem 5.2.3 (Bolzano-Weierstrass' teorem) *Alle begrensede følger i $\mathbb{R}^m$ har en konvergent delfølge.*

Bevis: Beviset er lettest å forstå for følger i $\mathbb{R}^2$, men når man først har forstått det der, ser man lett at det også fungerer i andre dimensjoner. Vi skal derfor bevise resultatet for en følge $\{\mathbf{x}_n\}$ i $\mathbb{R}^2$ og overlate resten til leserne.

Siden følgen er begrenset, vet vi at det finnes et tall K slik at $|\mathbf{x}_n| < K$ for alle n. Det betyr spesielt at alle elementene i følgen ligger innenfor det store kvadratet i figur 1.



Vi deler kvadratet opp i fire mindre kvadrater S_1, S_2, S_3, S_4 som vist på figuren (disse kvadratene er lukkede og utgjør "hvert sitt hjørne" av det opprinnelige kvadratet). Siden følgen har uendelig mange ledd $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots$, må minst ett av kvadratene S_1, S_2, S_3, S_4 også inneholde uendelig mange ledd. Kall dette kvadratet K_1 (hvis flere av kvadratene S_1, S_2, S_3, S_4 inneholder uendelig mange ledd, velger vi bare ett av dem). Vi deler nå dette kvadratet K_1 i fire nye kvadrater på samme måte som før, og observerer at minst én av delene må inneholde uendelig mange ledd fra følgen. Dette kvadratet kaller vi K_2. Fortsetter vi på denne måten, får vi en følge av kvadrater $K_1, K_2, K_3, \ldots$ som ligger inni hverandre og som alle inneholder uendelig mange ledd fra følgen. Figur 2 viser hvordan en slik følge kan se ut:



Ideen er nå å plukke en delfølge $\{\mathbf{x}_{n_k}\}$ av$\{\mathbf{x}_n\}$ slik at $\mathbf{x}_{n_k} \in K_k$ for alle k. Dette er mulig siden hvert kvadrat inneholder uendelig mange av leddene i den opprinnelige følgen, og det er intuitivt rimelig at en slik delfølge må

konvergere siden leddene "er fanget i" mindre og mindre kvadrater. La oss se på detaljene.

For å konstruere delfølgen $\{\mathbf{x}_{n_k}\}$ lar vi først n_1 være det minste tallet slik at $\mathbf{x}_{n_1} \in K_1$. Deretter velger vi n_2 til å være det første tallet etter n_1 slik at $\mathbf{x}_{n_2}$ er med i K_2 — et slikt element må finnes siden K_2 inneholder uendelig mange ledd fra følgen. På tilsvarende måte velger vi n_3 til å være det første tallet etter n_2 slik at $\mathbf{x}_{n_3} \in K_3$, osv. Vi har nå funnet en delfølge $\{\mathbf{x}_{n_k}\}$ av $\{\mathbf{x}_n\}$ slik at $\mathbf{x}_{n_k} \in K_k$, og det gjenstår å vise at den konvergerer.

La (a_k, b_k) være koordinatene til det nedre, venstre hjørne til kvadratet K_k. Det følger fra konstruksjonen at $\{a_k\}$ og $\{b_k\}$ er voksende (dvs. ikke avtagende), begrensede følger. Ifølge teorem 4.3.9 i *Kalkulus* må de da konvergere mot hver sin grenseverdi a og b, og følgelig vil punktene $\mathbf{z}_k = (a_k, b_k)$ konvergere mot $\mathbf{z} = (a, b)$ (husk setning 5.1.5). Siden både $\mathbf{z}_k$ og $\mathbf{x}_{n_k}$ ligger i kvadratet K_k, og størrelsen av dette kvadratet går mot null, må da også følgen $\{\mathbf{x}_{n_k}\}$ konvergere mot $\mathbf{z}$. Dermed er teoremet bevist. □

Som du snart vil få anledning til å se, er teoremet ovenfor et usedvanlig nyttig redskap når man arbeider med følger. Et enda nyttigere redskap er *Cauchy-følger*.

Definisjon 5.2.4 *En følge $\{\mathbf{x}_n\}$ i $\mathbb{R}^m$ er en* Cauchy-følge *dersom det for enhver $\epsilon > 0$, finnes en $N \in \mathbb{N}$ slik at $|\mathbf{x}_n - \mathbf{x}_k| < \epsilon$ for alle $n, k \geq N$*

En følge er altså en Cauchy-følge dersom vi kan få avstanden mellom to ledd til å bli vilkårlig liten ved å gå tilstrekkelig langt ut i følgen. Det er ikke vanskelig å se at alle konvergente følger er Cachy-følger.

Lemma 5.2.5 *Enhver konvergent følge i $\mathbb{R}^m$ er en Cauchy-følge*

Bevis: Anta at $\mathbf{x}_n$ konvergerer mot $\mathbf{x}$. Gitt et tall $\epsilon > 0$, vet vi da at det finnes et tall $N \in \mathbb{N}$ slik at $|\mathbf{x}_n - \mathbf{x}| < \frac{\epsilon}{2}$ når $n \geq N$. Hvis både n og k er større enn eller lik N, har vi dermed ved trekantulikheten

$$|\mathbf{x}_n - \mathbf{x}_k| = |(\mathbf{x}_n - \mathbf{x}) + (\mathbf{x} - \mathbf{x}_k)| \leq |\mathbf{x}_n - \mathbf{x}| + |\mathbf{x} - \mathbf{x}_k| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$$

Dermed har vi vist at $\mathbf{x}_n$ er en Caychy-følge. □

Det neste resultatet er mer overraskende — og atskillig mer nyttig.

Teorem 5.2.6 *Alle Cauchy-følger i $\mathbb{R}^m$ konvergerer.*

Bevis: Gangen i beviset er som følger: Først viser vi at enhver Cauchy-følge er begrenset, deretter bruker vi teorem 5.2.3 til å plukke ut en konvergent delfølge, og til slutt viser vi at Cauchy-følgen konvergerer mot det samme punktet som delfølgen.

Anta at $\{\mathbf{x}_n\}$ er en Cauchy-følge. For å vise at $\{\mathbf{x}_n\}$ er begrenset, velger vi en $\epsilon > 0$ (f.eks. $\epsilon = 1$). Da finnes det en N slik at $|\mathbf{x}_n - \mathbf{x}_k| < \epsilon$ når $n, k \geq N$. Spesielt må $|\mathbf{x}_n - \mathbf{x}_N| < \epsilon$ for alle $n \geq N$, og ved trekantulikheten betyr det at

$$|\mathbf{x}_n| = |\mathbf{x}_N + (\mathbf{x}_n - \mathbf{x}_N)| \leq |\mathbf{x}_N| + |\mathbf{x}_n - \mathbf{x}_N| < |\mathbf{x}_N| + \epsilon$$

når $n \geq N$. Dette betyr at følgen er begrenset av det største av tallene

$$|\mathbf{x}_1|, |\mathbf{x}_2|, |\mathbf{x}_3|, \ldots, |\mathbf{x}_{N-1}|, |\mathbf{x}_N| + \epsilon$$

Siden følgen $\{\mathbf{x}_n\}$ er begrenset, har den ifølge teorem 5.2.3 en konvergent delfølge $\{\mathbf{x}_{n_k}\}$ med en grenseverdi $\mathbf{x}$. Vi skal vise at også den opprinnelige følgen $\{\mathbf{x}_n\}$ konvergerer mot $\mathbf{x}$. Gitt en $\epsilon > 0$, må vi da vise at det finnes en N slik at $|\mathbf{x}_n - \mathbf{x}| < \epsilon$ når $n \geq N$. Det er ikke så vanskelig: Siden $\{\mathbf{x}_n\}$ er en Cauchy-følge, finnes det en N slik at $|\mathbf{x}_n - \mathbf{x}_k| < \frac{\epsilon}{2}$ når $n, k \geq N$. Siden delfølgen $\{\mathbf{x}_{n_k}\}$ konvergerer mot $\mathbf{x}$, finnes det et element $\mathbf{x}_{n_K}$ i denne følgen slik at $n_K \geq N$ og $|\mathbf{x}_{n_K} - \mathbf{x}| < \frac{\epsilon}{2}$. Hvis $n \geq N$, er dermed

$$|\mathbf{x}_n - \mathbf{x}| = |(\mathbf{x}_n - \mathbf{x}_{n_K}) + (\mathbf{x}_{n_K} - \mathbf{x})| \leq |\mathbf{x}_n - \mathbf{x}_{n_K}| + |\mathbf{x}_{n_K} - \mathbf{x}| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$$

Dermed er teoremet bevist. □

Kombinerer vi teoremet ovenfor med det foregående lemmaet, får vi denne eksakte sammenhengen:

Korollar 5.2.7 *En følge i $\mathbb{R}^m$ konvergerer hvis og bare hvis den er en Cauchy-følge.*

Bemerkning: Noen synes kanskje at resultatene ovenfor er så opplagte at det ikke er noen grunn til å bevise dem — det er da en selvfølge at Cauchy-følger og konvergente følger er det samme! Det er imidlertid lett å overbevise seg om at dette ikke nødvendigvis er tilfellet. Anta at vi foretrakk å gjøre vår matematikk i mengden $\mathbb{Q}$ av rasjonale tall istedenfor i mengden $\mathbb{R}$ av reelle tall. Følgen

$$x_0 = 1, x_2 = 1.4, x_3 = 1.41, x_4 = 1.4142, x_5 = 1.41421, \ldots$$

(bestående av lengre og lengre desimaltallstilnærminger til $\sqrt{2}$) er en Cauchy-følge i $\mathbb{Q}$, men den konvergerer ikke i $\mathbb{Q}$ siden grensen er det irrasjonale tallet $\sqrt{2}$.

Det er mulig å definere konvergens og Caychy-følger i andre mengder (matematikere kaller dem gjerne *metriske rom*) enn $\mathbb{R}^m$ — det viser seg at alt vi trenger, er et mål for avstanden mellom to punkter (en *metrikk*). Et slikt metrisk rom kalles *komplett* dersom alle Cauchy-følger konvergerer. Teoremet ovenfor forteller oss altså at $\mathbb{R}^m$ er komplett, mens eksemplet i

begynnelsen av denne bemerkningen viser at $\mathbb{Q}$ *ikke* er komplett. Du har støtt på ordet *komplett* før i forbindelse med reelle tall — i seksjon 2.3 i *Kalkulus* studerte vi kompletthetsprinsippet for $\mathbb{R}$ (det sier at alle begrensede, ikke-tomme delmengder av $\mathbb{R}$ har en minste øvre skranke). Det viser seg å være en nær sammenheng mellom disse to formene for kompletthet, og vi kunne ha basert vår diskusjon av $\mathbb{R}$ på konvergens av Cauchy-følger istedenfor eksistens av minste øvre skranker.

Selv om det kankje ikke er så lett å se ved første øyekast, er teoremet ovenfor et meget viktig redskap når man skal studere konvergens av følger — det er ofte mye lettere å vise at en følge $\{\mathbf{x}_n\}$ konvergerer ved å vise at den er en Cauchy-følge enn ved å bruke definisjonen av konvergens. Grunnen er at for å bruke definisjonen, må vi først finne grensepunktet $\mathbf{x}$ og så vise at vi kan få $|\mathbf{x}_n - \mathbf{x}|$ mindre enn ϵ ved å velge N stor nok. For å vise at $\{\mathbf{x}_n\}$ er en Cauchy-følge, trenger vi ikke kjenne grensepunktet — alt vi skal sjekke, er at vi kan få $|\mathbf{x}_n - \mathbf{x}_k|$ mindre enn ϵ ved å velge n og k store nok. Vi arbeider altså med de *gitte* verdiene $\mathbf{x}_n$ og $\mathbf{x}_k$ og trenger ikke å vite noe om grensepunktet på forhånd. Vi skal se slående eksempler på denne teknikken både i avsnittet nedenfor og i seksjonene 5.4 og 5.5.

*Operatornorm og inverterbarhet

I dette avsnittet skal vi se på en litt overraskende anvendelse av teorien ovenfor; vi skal bruke den til å utlede et nyttig kriterium for inverterbarhet av matriser! Vi skal få bruk for dette kriteriet når vi studerer Newtons metode senere i kapitlet.

Inspirasjonskilden til kriteriet er summeformelen for en geometrisk rekke

$$\frac{1}{1-x} = 1 + x + x^2 \cdots + x^n + \cdots \qquad \text{for } |x| < 1$$

Vi skal vise en tilsvarende formel for $m \times m$-matriser A, nemlig at hvis $|A| < 1$ (vi skal forklare hva dette betyr om et øyeblikk), så er $I_m - A$ inverterbar og

$$(I_m - A)^{-1} = I_m + A + A^2 + \ldots + A^n + \ldots$$

En slik geometrisk rekke av matriser kalles ofte en *Neumann-rekke*. Formelen ovenfor er nyttig for å vise inverterbarhet av matriser som ikke avviker for mye fra identitetsmatrisen.

La oss begynne med å forklare hva vi mener med $|A|$. I slutten av seksjon 1.6 innførte vi *normen* $\|A\|$ til en $m \times m$-matrise

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mm} \end{pmatrix}$$

ved

$$\|A\| = \sqrt{\sum_{1\leq i,j\leq m} a_{ij}^2}$$

og beviste (setning 1.6.3) at

$$|A\mathbf{x}| \leq \|A\||x| \quad \text{for alle } \mathbf{x} \in \mathbb{R}^m$$

Deler vi med $|\mathbf{x}|$ i den siste ligningen, får vi

$$\frac{|A\mathbf{x}|}{|\mathbf{x}|} \leq \|A\| \quad \text{for alle } \mathbf{x} \neq \mathbf{0} \tag{5.2.1}$$

Vi skal nå innføre et annet mål på størrelsen til en matrise som ofte er mer effektivt enn normen.

Definisjon 5.2.8 *Anta at A er en $m \times m$-matrise.* Operatornormen $|A|$ *til A er definert ved*

$$|A| = \sup\left\{ \frac{|A\mathbf{x}|}{|\mathbf{x}|} \ : \ \mathbf{x} \in \mathbb{R}^m, \mathbf{x} \neq \mathbf{0} \right\}$$

Legg merke til at ifølge formel (5.2.1) ovenfor er mengden i definisjonen begrenset av $\|A\|$, så den minste øvre skranken $|A|$ finnes. Av samme grunn er

$$|A| \leq \|A\|$$

Vi understreker igjen at normen $\|A\|$ og operatornormen $|A|$ er to forskjellige måter å måle størrelsen til en matrise på. Normen $\|A\|$ er som regel lettest å beregne, mens operatornormen $|A|$ ofte er lettere å bruke i teoretiske argumenter. På grunn av ulikheten $|A| \leq \|A\|$ er det ofte mulig å erstatte operatornormen med normen i beregningsproblemer og dermed få enklere beregninger.

Det følger fra definisjonen av operatornorm at

$$|A\mathbf{x}| \leq |A||\mathbf{x}| \quad \text{for alle } \mathbf{x} \in \mathbb{R}^m$$

Ved hjelp av denne ulikheten kan vi vise våre første resultater.

Lemma 5.2.9 *For enhver kvadratisk matrise A er*

$$|a_{ij}| \leq |A|$$

for alle elementer a_{ij} i matrisen.

Bevis: Ganger vi A med enhetsvektoren $\mathbf{e}_j$, får vi

$$A\mathbf{e}_j = \begin{pmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{mj} \end{pmatrix}$$

Tar vi normen på begge sider og bruker formelen ovenfor, får vi (husk at $|\mathbf{e}_j| = 1$):

$$|A| = |A||\mathbf{e}_j| \geq |A\mathbf{e}_j| = \left| \begin{pmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{mj} \end{pmatrix} \right| = \sqrt{a_{1j}^2 + a_{2j}^2 + \cdots + a_{mj}^2} \geq |a_{ij}| \quad \square$$

Lemma 5.2.10 *Hvis A, B er $m \times m$-matriser, så er*

(i) $|AB| \leq |A||B|$

(ii) $|A + B| \leq |A| + |B|$

Bevis: (i) For enhver $\mathbf{x} \neq \mathbf{0}$ i $\mathbb{R}^m$ har vi

$$|(AB)\mathbf{x}| = |A(B\mathbf{x})| \leq |A||B\mathbf{x}| \leq |A||B||\mathbf{x}|$$

der vi har brukt ulikheten $|C\mathbf{y}| \leq |C||\mathbf{y}|$ to ganger, først med $C = A$ og $\mathbf{y} = B\mathbf{x}$, og så med $C = B$ og $\mathbf{y} = \mathbf{x}$. Deler vi på $|\mathbf{x}|$ i ulikheten ovenfor, får vi

$$\frac{|(AB)\mathbf{x}|}{|\mathbf{x}|} \leq |A||B|$$

for alle $\mathbf{x} \neq \mathbf{0}$. Fra definisjonen av operatornorm får vi dermed at $|AB| \leq |A||B|$

(ii) For enhver $\mathbf{x} \neq \mathbf{0}$ i $\mathbb{R}^m$ har vi

$$|(A + B)\mathbf{x}| = |A\mathbf{x} + B\mathbf{x}| \leq |A\mathbf{x}| + |B\mathbf{x}| \leq |A||\mathbf{x}| + |B||\mathbf{x}| = (|A| + |B|)|\mathbf{x}|$$

Deler vi på $|\mathbf{x}|$, får vi

$$\frac{|(A + B)\mathbf{x}|}{|\mathbf{x}|} \leq |A| + |B|$$

for alle $\mathbf{x} \neq \mathbf{0}$. Fra definisjonen av operatornorm får vi dermed at $|A + B| \leq |A| + |B|$. $\square$

Resultatet ovenfor gjelder selvfølgelig også for flere matriser enn to; vi har

$$|A_1 A_2 \dots A_n| \leq |A_1||A_2| \dots |A_n|$$

og

$$|A_1 + A_2 + \cdots + A_n| \leq |A_1| + |A_2| + \cdots + |A_n|$$

Spesielt er

$$|A^n| \leq |A|^n$$

som vi snart skal få bruk for.

Vi er nå nesten ferdig med forberedelsene, alt som gjenstår er å definere grenseverdien til en følge $\{A_n\}$ av matriser. Det er lett; vi sier rett og slett at $\{A_n\}$ konvergerer mot B dersom hvert element i A_n konvergerer mot tilsvarende element i B.

I det neste beviset får vi bruk for det vi vet om Cauchy-følger.

Lemma 5.2.11 *Dersom A er en $m \times m$-matrise med $|A| < 1$, så konvergerer den geometriske rekken*

$$I_m + A + A^2 + \ldots + A^n + \ldots$$

mot en matrise B i den forstand at $B = \lim_{n\to\infty}(I_m + A + A^2 + \ldots + A^n)$.

Bevis: La $a_{ij}^{(k)}$ være det ij-te elementet til A^k (vi skriver k-en i $a_{ij}^{(k)}$ i parentes for å understreke at det ikke er snakk om en potens). Vi må vise at følgen av delsummer $\{s_n\}$, der $s_n = \sum_{k=0}^{n} a_{ij}^{(k)}$, konvergerer. Ifølge teorien vår er det nok å vise at denne følgen er en Cauchy-følge. Dersom $N > n$, har vi ifølge de to lemmaene ovenfor

$$|s_N - s_n| = |\sum_{k=n+1}^{N} a_{ij}^{(k)}| \leq \sum_{k=n+1}^{N} |a_{ij}^{(k)}| \leq \sum_{k=n+1}^{N} |A^k| \leq \sum_{k=n+1}^{N} |A|^k =$$

$$= \frac{|A|^{n+1}(1 - |A|^{N-n})}{1 - |A|} \leq \frac{|A|^{n+1}}{1 - |A|}$$

der vi har summert en geometrisk rekke (husk at $|A|$ er et tall slik at det er en vanlig geometrisk rekke av tall vi summerer her). Siden $|A| < 1$, kan vi få uttrykket på høyre side så lite vi vil ved å velge n tilstrekkelig stor. Følgelig er $\{s_n\}$ en Cauchy-følge og må konvergere. □

Bemerkning: Argumentet ovenfor er ganske typisk for hvordan man bruker Cauchy-følger i praksis. I dette tifellet har vi ikke full kontroll over hvordan elementene $s_n = \sum_{k=0}^{n} a_{ij}^{(k)}$ i følgen ser ut (du kan prøve å regne dem ut, men uttrykkene blir utrolig kompliserte), men vi har en ulikhet som forteller oss noe om størrelsen. Ved hjelp av denne ulikheten kan vi vise at følgen er en Cauchy-følge, og dermed er konvergensen etablert til tross for at vi fortsatt ikke vet noe særlig hverken om leddene i følgen eller grenseverdien!

Vi kan nå bevise hovedresultatet vårt:

Teorem 5.2.12 *Anta at A er en $m \times m$-matrise og at $|A| < 1$. Da er $I_m - A$ inverterbar og*

$$(I_m - A)^{-1} = I_m + A + A^2 + \ldots + A^n + \ldots$$

Bevis: La $B_n = (I_m + A + A^2 + \ldots + A^n)$ og $B = \lim_{n\to\infty} B_n = I_m + A + A^2 + \ldots + A^n + \ldots$. Multipliserer vi ut og forkorter, ser vi at

$$(I_m - A)B_n = (I_m - A)((I_m + A + A^2 + \ldots + A^n) = I_n - A^{n+1}$$

Lar vi n gå mot uendelig i denne ligningen, får vi

$$(I_m - A)B = \lim_{n\to\infty}(I_m - A)B_n = \lim_{n\to\infty}(I_n - A^{n+1}) = I_m$$

der vi har brukt at siden $|A| < 1$, går $|A|^{n+1}$ mot null (tenk gjennom hva du egentlig bruker i denne overgangen!). □

Vi tar med en reformulering av dette resultatet som ofte er mer naturlig å bruke:

Korollar 5.2.13 *Anta at C er en $m \times m$-matrise slik at $|I_m - C| < 1$. Da er C inverterbar og*

$$C^{-1} = I_m + (I_m - C) + (I_m - C)^2 + \ldots + (I_m - C)^n + \ldots$$

Bevis: Sett $A = I_m - C$ og bruk teoremet ovenfor. □

Til slutt skal vi utvide resultatet vårt slik at det ikke bare gjelder for matriser som ligger nær identitetsmatrisen, men for matriser som ligger nær en hvilket som helst inverterbar matrise.

Teorem 5.2.14 (Banachs lemma) *Anta at B er en inverterbar $m \times m$-matrise og at A er en annen $m \times m$-matrise. Dersom*

$$|B - A| < |B^{-1}|^{-1}$$

så er også A inverterbar og

$$|A^{-1}| \leq \frac{|B^{-1}|}{1 - |B^{-1}||B - A|}$$

Bevis: Siden B er inverterbar, kan vi skrive

$$A = B - (B - A) = B\left(I_m - B^{-1}(B - A)\right)$$

Ifølge teorem 5.2.12 er matrisen $I_m - B^{-1}(B - A)$ inverterbar dersom

$$|B^{-1}(B - A)| < 1$$

Siden $|B^{-1}(B-A)| \leq |B^{-1}||B-A|$, følger det at $I_m - B^{-1}(B-A)$ er inverterbar dersom $|B^{-1}||B-A| < 1$, dvs. dersom

$$|B-A| < |B^{-1}|^{-1}$$

(legg merke til at dette er betingelsen i teoremet). I så fall er også

$$A = B\left(I_m - B^{-1}(B-A)\right)$$

inverterbar med

$$A^{-1} = \left(I_m - B^{-1}(B-A)\right)^{-1} B^{-1}$$

Det gjenstår å estimere normen til A^{-1}. Siden

$$|A^{-1}| \leq |\left(I_m - B^{-1}(B-A)\right)^{-1}||B^{-1}|$$

og (ifølge teorem 5.2.12)

$$|\left(I_m - B^{-1}(B-A)\right)^{-1}| = |\sum_{k=0}^{\infty}(B^{-1}(B-A))^k| \leq$$

$$\leq \sum_{k=0}^{\infty}(|B^{-1}||(B-A)|)^k \leq \frac{1}{1-|B^{-1}||B-A[}$$

får vi

$$|A^{-1}| \leq \frac{|B^{-1}|}{1-|B^{-1}||B-A|}$$

□

5.3 Iterasjon av funksjoner

Anta at vi har en funksjon $\mathbf{F} : \mathbb{R}^m \to \mathbb{R}^m$. Dersom vi velger et startpunkt $\mathbf{x}_0 \in \mathbb{R}^m$, kan vi bruke $\mathbf{F}$ til å skaffe oss en følge $\{\mathbf{x}_n\}$ på denne måten:

$$\mathbf{x}_1 = \mathbf{F}(\mathbf{x}_0),\ \mathbf{x}_2 = \mathbf{F}(\mathbf{x}_1),\ \mathbf{x}_3 = \mathbf{F}(\mathbf{x}_2),\ \ldots,\ \mathbf{x}_{n+1} = \mathbf{F}(\mathbf{x}_n),\ \ldots$$

Hvert ledd i følgen fremkommer altså ved at vi bruker $\mathbf{F}$ på det foregående leddet. Vi sier at følgen $\{\mathbf{x}_n\}$ oppstår ved *iterasjon* av $\mathbf{F}$.

Når man tenker på iterasjoner, er det ofte lurt å forestille seg at punktene i følgen representerer tilstander ved forskjellige tidspunkt; $\mathbf{x}_0$ er tilstanden ved tiden 0, $\mathbf{x}_1$ er tilstanden ved tiden 1, osv. Funksjonen $\mathbf{F}$ blir da en mekanisme som oppdaterer tilstanden fra et tidspunkt til det neste (i kapittel 1 tenkte vi på funksjoner gitt av matriser på denne måten). La oss se på et lite eksempel som kanskje gjør tankegangen lettere å forstå.

Eksempel 1: To dyreslag, et byttedyr og et rovdyr, lever i det samme området. Dersom det ett år er x_n byttedyr og y_n rovdyr i området, tenker man seg at antall dyr året etter er gitt ved

$$\begin{aligned} x_{n+1} &= ax_n - bx_n y_n \\ y_{n+1} &= cy_{n+1} + dx_n y_n \end{aligned}$$

der a, b, c, d er positive tall. Legg merke til logikken; "kryssleddene" $x_n y_n$ representerer møter mellom byttedyr og rovdyr, og slike møter reduserer veksten av byttedyr, men bidrar til vekst i rovdyrbestanden. Hvis vi innfører funksjonen $\mathbf{F} : \mathbb{R}^2 \to \mathbb{R}^2$ ved

$$\mathbf{F}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} ax - bxy \\ cy + dxy \end{pmatrix}$$

og lar $\mathbf{x}_n = \begin{pmatrix} x_n \\ y_n \end{pmatrix}$, ser vi at systemet ovenfor kan skrives som $\mathbf{x}_{n+1} = \mathbf{F}(\mathbf{x}_n)$.

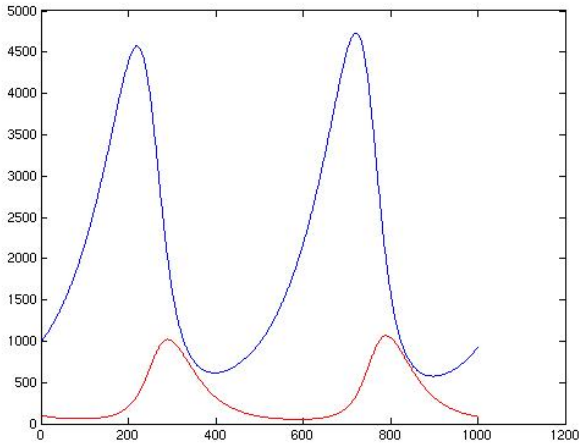
Det er ikke lett å gjette hvordan et system av denne typen vil utvikle seg i tiden, så la oss bruke MATLAB som hjelpemiddel til å se på et spesielt tilfelle. Vi velger $a = 1.01, b = 3 \cdot 10^{-5}, c = 0.98, d = 10^{-5}$. Følgende m-fil regner ut utviklingen når vi starter med m byttedyr og k rovdyr og gjennomfører N iterasjoner. Vær oppmerksom på at det er en liten forskyvning i nummereringen av leddene i følgen; i teoretisk arbeid får vi ofte penest uttrykk om vi begynner iterasjonen med punkt nummer 0 (altså $\mathbf{x}_0$ som ovenfor), men i programmet nedenfor har vi tatt hensyn til MATLABs forkjærlighet for å la startpunktet være nummer 1 (og ikke nummer 0).

```
function C=byttedyr(m,k,N)
x=[m];      %med disse linjene forteller vi
y=[k];      %MATLAB at iterasjonen starter i punktet (m,k)
for n=1:N % starter løkken som utfører iterasjonene
    x(n+1)=1.01*x(n)-3*10^(-5)*x(n)*y(n);
    y(n+1)=0.98*y(n)+10^(-5)*x(n)*y(n);
end         %avslutter for-løkken
C=[x;y]; %lager en matrise med x som første rad, y som andre osv.
           %Rutinen returnerer denne matrisen
```

Dersom vi ønsker å se grafisk på utviklingen når vi starter med 1000 byttedyr og hundre rovdyr, kan vi gi kommandoene

```
>> C=byttedyr(1000,100,1000);
>> x=C(1,:);
>> y=C(2,:);
>> plot(x)
>> hold on
>> plot(y,'r')
```

Vi får dette resultatet:



Figur 1: Utviklingen av byttedyr (øverst) og rovdyr (nederst).

Vi ser at bestandene følger et bølgemønster med klare topper og bunner. Logikken er ikke så vanskelig å forstå; til å begynne med er det relativt få rovdyr, og byttedyrbestanden vokser. Dette fører til gode betingelser for rovdyrbestanden som også begynner å vokse kraftig. Til slutt gjør rovdyrene så kraftig innhogg at byttedyrbestanden begynner å avta. Etter hvert fører dette til dårligere forhold for rovdyrene, og rovdyrbestanden begynner også å avta. Dette gir etter hvert bedre forhold for byttedyrene som begynner å ta seg opp igjen osv. ♣

Bemerkning: La oss smette inn en liten bemerkning om effektivitet i MATLAB-beregninger. Programmet ovenfor er kort, og det fungerer utmerket for de små datamengdene vi har i dette eksemplet. Det er imidlertid lite effektivt fordi MATLAB hele tiden må endre dimensjonen på vektorene `x` og `y`. Det viser seg at MATLAB arbeider mye mer effektivt dersom vi gir disse vektorene den "riktige" dimensjonen helt fra starten av. Det kan vi gjøre ved å endre programmet til

```
function C=byttedyr(m,k,N)
x=zeros(1,N);
y=zeros(1,N);
x(1)=m;
y(1)=k;
for n=1:N
    x(n+1)=1.01*x(n)-3*10^(-5)*x(n)*y(n);
    y(n+1)=0.98*y(n)+10^(-5)*x(n)*y(n);
end
C=[x;y];
```

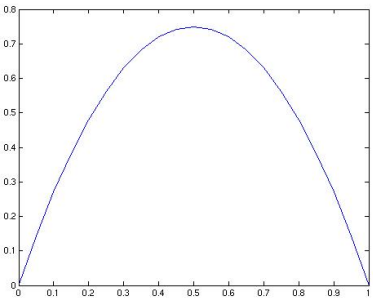
Kommandoen `zeros(1,N)` gir oss et N-tuppel (radvektor) med bare 0'er, og sørger derfor for at vektorene våre har den riktige dimensjonen helt fra starten av.

Bølgemønsteret ovenfor er bare ett av mange man kan støte på når man itererer en funksjon. For å bli litt kjent med de forskjellige mulighetene skal vi bruke MATLAB til å gjøre noen eksperimenter. Selv om vi i dette heftet hovedsakelig er interessert i funksjoner av flere variable, skal vi gjennomføre disse eksperimentene for en funksjon $f : \mathbb{R} \to \mathbb{R}$ av én variabel for å få oversiktlige figurer.

Vi skal stort sett arbeide med funksjoner $f : [0, 1] \to [0, 1]$ gitt ved

$$f(x) = bx(1 - x)$$

der b er en konstant mellom 0 og 4 (når b ligger utenfor dette intervallet, vil ikke f avbilde intervallet $[0, 1]$ inn i $[0, 1]$). Funksjoner av denne typen ser ut som grafen på figur 2 (større b-verdier gir høyere topp).

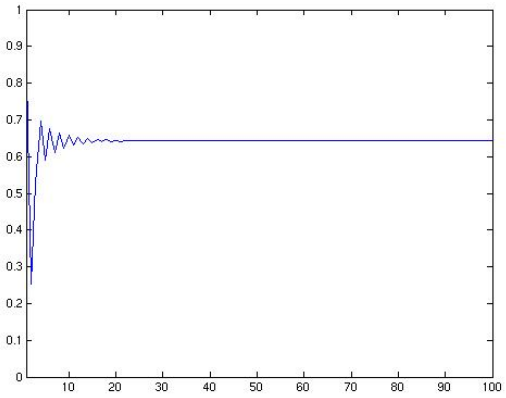


Figur 2: En graf av typen $f(x) = bx(1 - x)$

For å gjennomføre iterasjonen, lager vi et lite program `iterasjon.m` (uten å bry oss for mye om beregningseffektivitet):

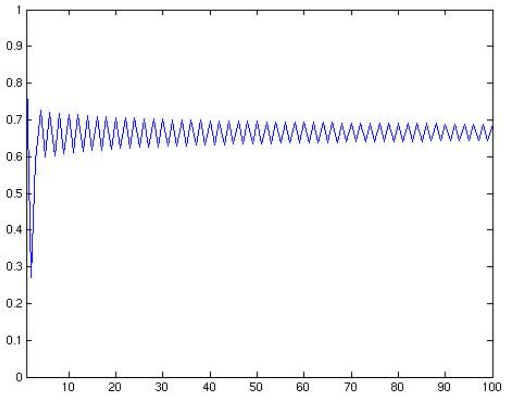
```
function x=iterasjon(a,b,N)
x=[a];
for n=1:N
    x(n+1)=b*x(n)*(1-x(n));
end
```

Input-parametrene a, b og N angir henholdsvis startverdien, parameteren b i ligningen ovenfor og antall iterasjonen. Gir vi MATLAB kommandoen `>>iterasjon(0.9,2.8,100)` etterfulgt av `>>plot(x)`, får vi figuren nedenfor:



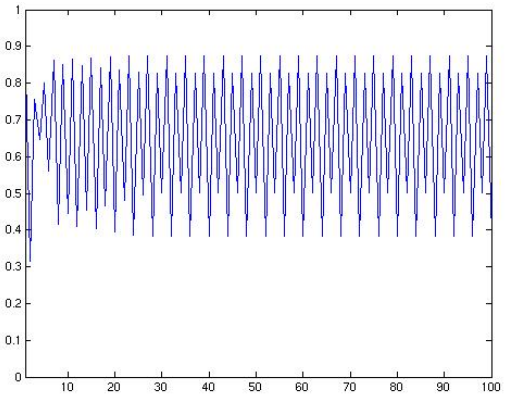
Figur 3: $b = 2.8$.

Vi ser at etter noen innledende fluktuasjoner, slår følgen seg til ro og nærmer seg en grenseverdi $x \approx 0.64$. Dette punktet x er et likevektspunkt (eller *fikspunkt* som matematikere liker å si) i den forstand at $f(x) = x$. La oss nå endre b-verdien og prøve med $b = 3.0$. Vi får denne figuren:



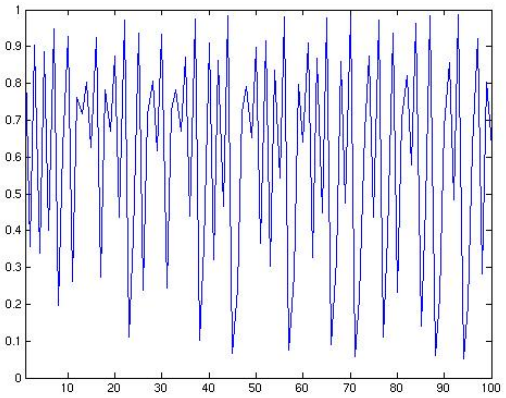
Figur 4: $b = 3.0$.

I dette tilfellet nærmer vi oss også en slags grenseverdi, men dette er ikke et fast punkt, men en svingebevegelse mellom to faste punkter — vi nærmer oss en "stabil bane med periode 2". Går vi et skritt videre og velger $b = 3.5$, får vi denne figuren:



Figur 5: $b = 3.5$.

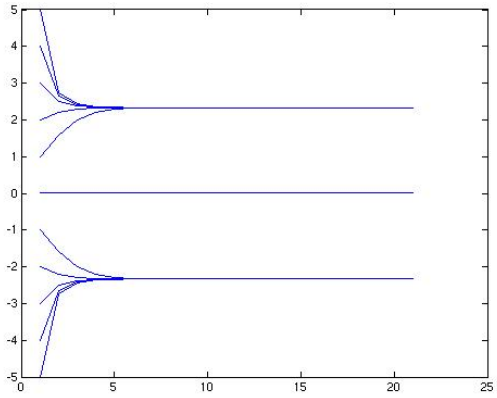
Denne gangen nærmer vi oss en "stabil bane med periode 4", altså en svingning mellom fire faste punkter. Til slutt setter vi $b = 3.95$ og får denne figuren:



Figur 6: $b = 3.95$.

Her er det tilsynelatende ingen orden i det hele tatt, bare usystematiske svingninger opp og ned.

Eksemplene ovenfor viser noen av de fenomenene vi kan støte på når vi itererer en funksjon. Vær oppmerksom på at oppførselen også kan avhenge av startverdien x_0 (altså parameteren a i programmet). I figuren nedenfor har vi iterert funksjonen $g(x) = 2 \arctan x$ med forskjellige startverdier.



Figur 6: Iterasjon av $g(x) = 2 \arctan x$ for forskjellige startverdier.

Vi ser at dersom vi har en positiv startverdi, konvergerer følgen mot et positivt fikspunkt $y \approx 2.3311$, men dersom vi starter i et negativt punkt, konvergerer den mot et negativt fikspunkt $z \approx -2.3311$. Det finnes også et tredje fikspunkt $u = 0$, men det er *frastøtende* i den forstand at følgen bare konvergerer mot det dersom x_0 (og dermed alle x_n) er lik 0. På tilsvarende måte kan lengden til en periodisk bane avhenge av startpunktet; ett startpunkt kan f.eks. lede til baner med periode 3, mens et annet startpunkt leder til baner med periode 14 (for samme b-verdi).

Det viser seg at itererte følger kan ha enda mer komplisert oppførsel enn det vi har sett eksempler på her, spesielt i høyeredimensjonale systemer der vi itererer en funksjon $\mathbf{F} : \mathbb{R}^m \to \mathbb{R}^m$ for $m > 1$. Dette får vi komme tilbake til en annen gang; hensikten med denne seksjonen er bare å gi deg en viss følelse for de fenomenene som kan oppstå ved iterasjon.

5.4 Konvergens mot et fikspunkt

I forrige seksjon så vi på noen av de oppførslene vi kan få når vi itererer en funksjon $\mathbf{F}$. En av de enkleste og viktigste er at følgen konvergerer mot en likevektstilstand $\mathbf{x}$, det vil si en tilstand slik at $\mathbf{F}(\mathbf{x}) = \mathbf{x}$ (se figur 3 og 6 i forrige seksjon). Matematisk kaller vi en slik tilstand for et *fikspunkt* for $\mathbf{F}$:

Definisjon 5.4.1 *Anta at A er en delmengde av $\mathbb{R}^m$ og at $\mathbf{F}$ er en funksjon fra A til $\mathbb{R}^m$. Vi sier at $\mathbf{x} \in A$ er et* fikspunkt *for $\mathbf{F}$ dersom $\mathbf{F}(\mathbf{x}) = \mathbf{x}$.*

Det er en del naturlige spørsmål knyttet til iterasjon og fikspunkter: Når har en funksjon $\mathbf{F}$ et fikspunkt? Dersom $\mathbf{F}$ har et fikspunkt $\mathbf{x}$, når vil en følge dannet ved iterasjon av $\mathbf{F}$ konvergere mot $\mathbf{x}$? Dersom en funksjon har flere fikspunkter, hvilket av disse vil en iterert følge konvergere mot? Hvordan vil dette avhenge av hvilket punkt $\mathbf{x}_0$ vi startet iterasjonen i, altså av

begynnelsestilstanden til systemet? Dette er vanskelige spørsmål som vi ikke kan gi utfyllende svar på her, men vi skal i hvert fall bevise et resultat som er nyttig i mange sammenhenger. Dette resultatet gjelder for *kontraksjoner* som er en spesielt enkel type funksjoner å ha med å gjøre. I neste seksjon skal vi se hvordan vi kan bruke fikspunktiterasjon til å løse ikke-lineære ligningssystemer.

Definisjon 5.4.2 *Anta at A er en ikke-tom delmengde av $\mathbb{R}^m$. En funksjon $\mathbf{F} : A \to A$ kalles en* kontraksjon *av mengden A dersom det finnes et positivt tall $C < 1$ slik at*

$$|\mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{y})| \leq C|\mathbf{x} - \mathbf{y}|$$

for alle $\mathbf{x}, \mathbf{y} \in A$. Vi kaller C en kontraksjonsfaktor *for F.*

Siden C er mindre enn 1, ser vi at kontraksjoner reduserer avstanden mellom punkter — avstanden mellom $\mathbf{F}(\mathbf{x})$ og $\mathbf{F}(\mathbf{y})$ er mindre enn avstanden mellom $\mathbf{x}$ og $\mathbf{y}$. Legg også merke til at vi krever at alle verdiene til $\mathbf{F}$ skal ligge i A; funksjonen $\mathbf{F}$ skal altså avbilde mengden A inn i A selv. Når man skal vise at en funksjon $\mathbf{F}$ er en kontraksjon av en mengde A, er dette ofte det vanskeligste punktet å sjekke.

Vi trenger litt notasjon før vi går videre. Dersom følgen $\{\mathbf{x}_n\}$ oppstår når vi itererer $\mathbf{F}$ med $\mathbf{x}_0$ som startpunkt, ser vi at

$$\mathbf{x}_1 = \mathbf{F}(\mathbf{x}_0),\ \mathbf{x}_2 = \mathbf{F}(\mathbf{x}_1) = \mathbf{F}(\mathbf{F}(\mathbf{x}_0)),\ \mathbf{x}_3 = \mathbf{F}(\mathbf{x}_2) = \mathbf{F}(\mathbf{F}(\mathbf{F}(\mathbf{x}_0))), \ldots$$

Generelt er $\mathbf{x}_n = \mathbf{F}(\mathbf{F}(\ldots(\mathbf{F}(\mathbf{x}_0))\ldots))$ der vi har n $\mathbf{F}$'er etter hverandre. For å få en mer oversiktelig notasjon, skriver vi $\mathbf{x}_n = \mathbf{F}^{\circ n}(\mathbf{x}_0)$; med andre ord er $\mathbf{F}^{\circ n}$ den funksjonen vi får når vi setter $\mathbf{F}$ sammen med seg selv n ganger.

Lemma 5.4.3 *Anta at $\mathbf{F} : A \to A$ er en kontraksjon med kontraksjonsfaktor C. For alle $\mathbf{x}, \mathbf{y} \in A$ og alle $n \in \mathbb{N}$ er da*

$$|\mathbf{F}^{\circ n}(\mathbf{x}) - \mathbf{F}^{\circ n}(\mathbf{y})| \leq C^n|\mathbf{x} - \mathbf{y}|$$

$\mathbf{F}^{\circ n}$ er altså en kontraksjon med kontraksjonsfaktor C^n.

Bevis: For $n = 1$ er dette bare definisjonen av kontraksjon. For $n = 2$ har vi:

$$|\mathbf{F}^{\circ 2}(\mathbf{x}) - \mathbf{F}^{\circ 2}(\mathbf{y})| = |\mathbf{F}(\mathbf{F}(\mathbf{x})) - \mathbf{F}(\mathbf{F}(\mathbf{y}))| \leq C|\mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{y})| \leq C^2|\mathbf{x} - \mathbf{y}|$$

Ved å fortsette på sammen måte, får vi resultatet for alle $n \in \mathbb{N}$ (du kan føre et formelt induksjonsbevis hvis du vil). □

Legg spesielt merke til at dersom $\mathbf{x}_k$ og $\mathbf{x}_{k+1}$ er to ledd som følger etter hverandre i følgen, så er

$$|\mathbf{x}_k - \mathbf{x}_{k+1}| = |\mathbf{F}^{\circ k}(\mathbf{x}_0) - \mathbf{F}^{\circ k}(\mathbf{x}_1)| \leq C^k|\mathbf{x}_0 - \mathbf{x}_1| \tag{5.4.1}$$

Vi er nå klare til å vise hovedresultatet i denne seksjonen.

Teorem 5.4.4 (Banachs fikspunktsteorem) *Anta at A er en ikke-tom, lukket delmengde av $\mathbb{R}^m$ og at $\mathbf{F} : A \to A$ er en kontraksjon av A med kontraksjonsfaktor C. Da har $\mathbf{F}$ nøyaktig ett fikspunkt $\mathbf{x}$ i A. Uansett hvilket punkt $\mathbf{x}_0$ i A vi starter iterasjonen i, vil følgen $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n, \ldots$ der $\mathbf{x}_n = \mathbf{F}^{\circ n}(\mathbf{x}_0)$ konvergere mot $\mathbf{x}$, og for alle $n \in \mathbb{N}$ er*

$$|\mathbf{x}_n - \mathbf{x}| \leq \frac{C^n}{1-C}|\mathbf{x}_0 - \mathbf{x}_1|$$

Bevis: La oss først vise at $\mathbf{F}$ kan ha høyst ett fikspunkt. Dersom både $\mathbf{x}$ og $\mathbf{y}$ er fikspunkter, har vi nemlig

$$|\mathbf{x} - \mathbf{y}| = |\mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{y})| \leq C|\mathbf{x} - \mathbf{y}|$$

Det betyr at $|\mathbf{x} - \mathbf{y}| \leq C|\mathbf{x} - \mathbf{y}|$, og siden $C < 1$ er det bare mulig dersom $|\mathbf{x} - \mathbf{y}| = 0$, dvs. dersom $\mathbf{x} = \mathbf{y}$.

Neste skritt er å vise at følgen $\{\mathbf{x}_n\}$ er en Cauchy-følge uansett hvilket punkt $\mathbf{x}_0 \in A$ vi begynner iterasjonen i. Dersom vi har to ledd i følgen $\mathbf{x}_n$ og $\mathbf{x}_k$ med $n < k$, har vi

$$\begin{aligned}
|\mathbf{x}_n - \mathbf{x}_k| &= |(\mathbf{x}_n - \mathbf{x}_{n+1}) + (\mathbf{x}_{n+1} - \mathbf{x}_{n+2}) + \ldots + (\mathbf{x}_{k-2} - \mathbf{x}_{k-1}) + (\mathbf{x}_{k-1} - \mathbf{x}_k)| \leq \\
&\leq |\mathbf{x}_n - \mathbf{x}_{n+1}| + |\mathbf{x}_{n+1} - \mathbf{x}_{n+2}| + \ldots + |\mathbf{x}_{k-2} - \mathbf{x}_{k-1}| + |\mathbf{x}_{k-1} - \mathbf{x}_k| \leq \\
&\leq C^n|\mathbf{x}_0 - \mathbf{x}_1| + C^{n+1}|\mathbf{x}_0 - \mathbf{x}_1| + \ldots + C^{k-2}|\mathbf{x}_0 - \mathbf{x}_1| + C^{k-1}|\mathbf{x}_0 - \mathbf{x}_1| \leq \\
&\leq \left(C^n + C^{n+1} + \ldots + C^{k-2} + C^{k-1} + \ldots\right)|\mathbf{x}_0 - \mathbf{x}_1| = \frac{C^n}{1-C}|\mathbf{x}_0 - \mathbf{x}_1|
\end{aligned}$$

der vi først har brukt trekantulikheten, så ulikheten i formel (5.4.1) ovenfor og til slutt summeformelen for en geometrisk rekke. Siden $C < 1$, kan vi få uttrykket $\frac{C^n}{1-C}|\mathbf{x}_0 - \mathbf{x}_1|$ så lite vi måtte ønske oss ved å velge n tilstrekkelig stor. Dette betyr at $\{\mathbf{x}_n\}$ er en Cauchy-følge, og ifølge teorem 5.2.6 konvergerer den mot en grense $\mathbf{x}$ (siden A er lukket, ligger $\mathbf{x}$ i A — husk setning 5.1.6). Det er lett å vise at grensepunktet $\mathbf{x}$ er et fikspunkt for $\mathbf{F}$; siden enhver kontraksjon er kontinuerlig, har vi nemlig ifølge setning 5.1.7:

$$\mathbf{x} = \lim_{n\to\infty} \mathbf{x}_{n+1} = \lim_{n\to\infty} \mathbf{F}(\mathbf{x}_n) = \mathbf{F}(\lim_{n\to\infty} \mathbf{x}_n) = \mathbf{F}(\mathbf{x})$$

Siden vi allerede vet at $\mathbf{F}$ ikke kan ha mer enn ett fikspunkt, betyr dette at følgen $\{\mathbf{x}_n\}$ konvegerer mot det samme punktet $\mathbf{x}$ uansett hvilket startpunkt $\mathbf{x}_0$ man bruker.

Det gjenstår å vise den siste formelen i teoremet. Ifølge utledningen ovenfor, er

$$|\mathbf{x}_n - \mathbf{x}_k| \leq \frac{C^n}{1-C}|\mathbf{x}_0 - \mathbf{x}_1|$$

når $k > n$. Lar vi $k \to \infty$, får vi

$$|\mathbf{x}_n - \mathbf{x}| \leq \frac{C^n}{1-C}|\mathbf{x}_0 - \mathbf{x}_1|$$

og beviset er fullført. □

Legg merke til at beviset ovenfor illustrerer filosofien vår fra seksjon 5.2 — det er mye lettere å vise at en følge konvergerer ved å sjekke at den er en Cauchy-følge enn ved å konstruere grenseelementet!

Bemerkning: Situasjonen i teoremet ovenfor er den best tenkelige — funksjonen $\mathbf{F}$ har nøyaktig ett fikspunkt $\mathbf{x}$, og itererer vi $\mathbf{F}$, vil følgen konvergere mot $\mathbf{x}$ uansett hvilket punkt $\mathbf{x}_0$ vi starter i. Det er ett poeng til som er viktig, og det er at vi har kontroll på *hvor fort* følgen $\{\mathbf{x}_n\}$ konvergerer mot $\mathbf{x}$. I praksis er det nemlig sjelden vi kan finne et fikspunkt helt nøyaktig, alt vi kan gjøre er å iterere funksjonen så mange ganger at $\mathbf{x}_n$ kommer så nær fikspunktet som vi trenger. Ulikheten

$$|\mathbf{x}_n - \mathbf{x}| \leq \frac{C^n}{1-C}|\mathbf{x}_0 - \mathbf{x}_1|$$

gir oss kontroll over dette — ønsker vi at $\mathbf{x}_n$ skal gi oss fikspunktet med en nøyaktighet bedre enn ϵ, må vi velge n så stor at

$$\frac{C^n}{1-C}|\mathbf{x}_0 - \mathbf{x}_1| < \epsilon$$

Det betyr at med en gang vi har regnet ut $\mathbf{x}_1$, har vi den informasjonen vi trenger for å beregne konvergeringsraten.

Vi skal nå se på et eksempel som viser at vi ikke kan fjerne betingelsen om at $\mathbf{F}$ er en kontraksjon fra teoremet ovenfor.

Eksempel 1: La $A = \{\mathbf{x} \in \mathbb{R}^2 \,:\, 1 \leq |\mathbf{x}| \leq 2\}$ være området mellom to sirkler i planet, og la $\mathbf{F}$ være avbildningen som dreier hele området A en vinkel $\theta < 2\pi$ om origo. Da er $|\mathbf{x} - \mathbf{y}| = |\mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{y})|$ (siden alle punkter dreies samme vinkel), men det finnes ingen fikspunkter siden alle punkter er rotert i forhold til utgangspunktet.

Legg merke til at dersom vi erstatter A med området $A' = \{\mathbf{x} \in \mathbb{R}^2 \,:\, |\mathbf{x}| \leq 2\}$, der "hullet" i midten er fjernet, så har $\mathbf{F}$ et fikspunkt, nemlig $\mathbf{0}$. Dette fikspunktet kan vi imidlertid ikke nå frem til ved iterasjon; enhver iterasjon av $\mathbf{F}$ sender punkter i sirkelbaner rundt origo. ♣

For å bruke Banachs fikspunktteorem trenger vi en metode for å vise at $\mathbf{F}$ er en kontraksjon. Vi skal nå arbeide oss frem mot et kriterium som ofte er nyttig, men først trenger vi en hjelpesetning som er av interesse også i andre sammenhenger:

Setning 5.4.5 (Middelverdisetning for funksjoner av flere variable) *Anta at $f : A \to \mathbb{R}$ er en funksjon av m variable, og at f er deriverbar i et*

område som inneholder linjestykket mellom punktene $\mathbf{a}, \mathbf{b} \in \mathbb{R}^m$*. Da finnes det et punkt* $\mathbf{c}$ *på linjestykket fra* $\mathbf{a}$ *til* $\mathbf{b}$ *slik at*

$$f(\mathbf{b}) - f(\mathbf{a}) = \nabla f(\mathbf{c}) \cdot (\mathbf{b} - \mathbf{a})$$

Bevis: Definer en funksjon $g : [0, 1] \to \mathbb{R}$ av én variabel ved

$$g(t) = f(\mathbf{r}(t))$$

der $\mathbf{r}(t) = \mathbf{a} + t(\mathbf{b} - \mathbf{a})$, $t \in [0, 1]$, er en parametrisering av linjestykket fra $\mathbf{a}$ til $\mathbf{b}$. Ved kjerneregelen er

$$g'(t) = \nabla f(\mathbf{r}(t)) \cdot \mathbf{r}'(t) = \nabla f(\mathbf{r}(t)) \cdot (\mathbf{b} - \mathbf{a})$$

Etter den vanlige middelverdisetningen (*Kalkulus*, setning 6.2.3) finnes det et tall c mellom 0 og 1 slik at

$$g(1) - g(0) = g'(c) = \nabla f(\mathbf{r}(c)) \cdot (\mathbf{b} - \mathbf{a})$$

Setter vi $\mathbf{c} = \mathbf{r}(c)$, har vi dermed

$$f(\mathbf{b}) - f(\mathbf{a}) = \nabla f(\mathbf{c}) \cdot (\mathbf{b} - \mathbf{a})$$

og setningen er bevist. □

Den neste setningen bringer oss enda et skritt nærmere resultatet vårt:

Setning 5.4.6 *Anta at* $\mathbf{F} : A \to \mathbb{R}^m$ *er en funksjon av* m *variable, og at* $\mathbf{F}$ *er deriverbar i et område som inneholder linjestykket mellom punktene* $\mathbf{a}, \mathbf{b} \in \mathbb{R}^m$*. Da finnes det punkter* $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_m$ *på linjestykket fra* $\mathbf{a}$ *til* $\mathbf{b}$*, slik at*

$$|\mathbf{F}(\mathbf{b}) - \mathbf{F}(\mathbf{a})| \leq |\mathbf{b} - \mathbf{a}| \sqrt{\nabla F_1(\mathbf{c}_1)^2 + \cdots + \nabla F_m(\mathbf{c}_m)^2}$$

der $F_1, F_2, \ldots, F_m$ *er komponentene til* $\mathbf{F}$*.*

Bevis: Bruker vi forrige setning på den i-te komponenten F_i, får vi et punkt $\mathbf{c}_i$ på linjestykket fra $\mathbf{a}$ til $\mathbf{b}$ slik at

$$F_i(\mathbf{b}) - F_i(\mathbf{a}) = \nabla F_i(\mathbf{c}_i) \cdot (\mathbf{b} - \mathbf{a})$$

Ved Schwarz' ulikhet er da

$$|F_i(\mathbf{b}) - F_i(\mathbf{a})| \leq |\nabla F_i(\mathbf{c}_i)||\mathbf{b} - \mathbf{a}|$$

Dermed er

$$|\mathbf{F}(\mathbf{b}) - \mathbf{F}(\mathbf{a})| = \sqrt{(F_1(\mathbf{b}) - F_1(\mathbf{a}))^2 + \cdots + (F_m(\mathbf{b}) - F_m(\mathbf{a}))^2} \leq$$

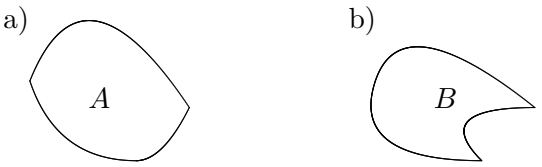
$$\leq \sqrt{|\nabla F_1(\mathbf{c}_1)|^2|\mathbf{b}-\mathbf{a}]^2 + \cdots + |\nabla F_m(\mathbf{c}_m)|^2|\mathbf{b}-\mathbf{a}]^2}$$
$$= |\mathbf{b}-\mathbf{a}|\sqrt{|\nabla F_1(\mathbf{c}_1)|^2 + \cdots + |\nabla F_m(\mathbf{c}_m)|^2}$$

og setningen er bevist. □

Denne setningen forteller oss at dersom det finnes et tall $C < 1$ slik at

$$\sqrt{|\nabla F_1(\mathbf{c}_1)|^2 + \cdots + |\nabla F_m(\mathbf{c}_m)|^2} \leq C$$

for alle punkter $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_m$ vi kan komme borti, så er $\mathbf{F}$ en kontraksjon. For å formulere dette som en setning på en grei måte, trenger vi begrepet *konveks mengde*. En delmengde A av $\mathbb{R}^m$ kalles *konveks* dersom det er slik at hver gang $\mathbf{a}$ og $\mathbf{b}$ er med i A, så er hele linjestykket mellom $\mathbf{a}$ og $\mathbf{b}$ også med i A. Intuitivt er en mengde konveks dersom randen "buler utover". Figuren nedenfor viser en konveks mengde A og en ikke-konveks mengde B. Legg merke til at det er lett å finne to punkter i B slik at ikke hele linjestykket som forbinder dem, ligger i B.



Figur 1: Konveks mengde A og ikke-konveks mengde B

Setning 5.4.7 *Anta at A er en ikke-tom, lukket, konveks delmengde av $\mathbb{R}^m$ og at $\mathbf{F} : A \to A$ er en avbildning som er deriverbar i A. Anta at det finnes et tall $C < 1$ slik at*

$$\sqrt{|\nabla F_1(\mathbf{c}_1)|^2 + \cdots + |\nabla F_m(\mathbf{c}_m)|^2} \leq C$$

for alle punkter $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_m \in A$. Da er $\mathbf{F}$ en kontraksjon og har et entydig fikspunkt. Vi kan iterere oss frem til fikspunktet ved å starte i et hvilket som helst punkt $\mathbf{x}_0$ i A.

Bevis: Dette er bare å kombinere setning 5.4.6 med teorem 5.4.4. Legg merke til at siden mengden A er konveks, vil ethvert punkt $\mathbf{c}_i$ på linjestykket mellom to punkter $\mathbf{a}, \mathbf{b} \in A$ selv ligge i A. □

Bemerkning: C-verdien vi får fra setningen ovenfor er ofte langt unna den beste (dvs. den minste) kontraksjonsfaktoren til $\mathbf{F}$, og det kan godt hende at $\mathbf{F}$ er en kontraksjon selv om

$$\sqrt{|\nabla F_1(\mathbf{c}_1)|^2 + \cdots + |\nabla F_m(\mathbf{c}_m)|^2} > 1$$

La oss se på et (ganske dårlig!) eksempel:

Eksempel 2: Definer en avbildning $\mathbf{F} : \mathbb{R}^3 \to \mathbb{R}^3$ ved

$$F(x,y,z) = \begin{pmatrix} \frac{x}{8} - \frac{y}{2} + \frac{z}{4} + 1 \\ \frac{x}{4} + \frac{z}{4} + 2 \\ \frac{x}{2} + \frac{y}{2} - 1 \end{pmatrix}$$

Her er Jacobi-matrisen

$$\mathbf{F}(x,y,z) = \begin{pmatrix} \frac{1}{8} & -\frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & 0 & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}$$

Gradienten til F_1 er den første linjen i matrisen, altså

$$\nabla F_1(x,y,z) = \frac{1}{8}\,\mathbf{i} - \frac{1}{2}\,\mathbf{j} + \frac{1}{4}\,\mathbf{k}$$

Tilsvarende er gradientene til F_2 og F_3 gitt av de neste linjene i matrisen

$$\nabla F_2(x,y,z) = \frac{1}{4}\,\mathbf{i} + \frac{1}{4}\,\mathbf{k}$$

og

$$\nabla F_3(x,y,z) = \frac{1}{2}\,\mathbf{i} + \frac{1}{2}\,\mathbf{j}$$

Uansett hvilke punkter vi evaluerer gradientene i, har vi dermed

$$|\nabla F_1|^2 + |\nabla F_2|^2 + |\nabla F_3|^2 =$$

$$= \left(\frac{1}{8}\right)^2 + \left(-\frac{1}{2}\right)^2 + \left(\frac{1}{4}\right)^2 + \left(\frac{1}{4}\right)^2 + 0^2 + \left(\frac{1}{4}\right)^2 + \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 + 0^2 =$$

$$= \frac{61}{64}$$

Dermed er

$$\sqrt{|\nabla F_1|^2 + |\nabla F_2|^2 + |\nabla F_3|^2} = \sqrt{\frac{61}{64}} = \frac{\sqrt{61}}{8} < 1$$

og $\mathbf{F}$ er altså en kontraksjon og har et entydig fikspunkt.

For å finne (en tilnærmet verdi for) fikspunktet, starter vi en iterasjon. Følgende MATLAB-program `fikspunkt.m` starter med punktet $\mathbf{x}_0 = (a,b,c)$ og gjennomfører N iterasjoner. Vær oppmerksom på den vanlige forskyvningen i nummereringen av punktene; i MATLAB-programmet lar vi startpunktet være (x_1, y_1, z_1). Husk også at vi kan gjøre programmet mer effektivt ved å gi `x`, `y` og `z` riktig lengde fra starten av (se bemerkningen etter eksempel 1 i seksjon 5.3).

```
function C=fikspunkt(a,b,c,N)
x=[a];     %med disse linjene forteller vi
y=[b];     %MATLAB at iterasjonen starter i
z=[c];     %punktet (a,b,c)
for n=1:N % starter for-løkken som utfører iterasjonene
    x(n+1)=x(n)/8-y(n)/2+z(n)/4+1;
    y(n+1)=x(n)/4+z(n)/4+2;
    z(n+1)=x(n)/2+y(n)-1;
end        %avslutter for-løkken
C=[x;y;z]; %lager en matrise med x som første rad, y som andre osv.
           %Rutinen returnerer denne matrisen
```

For å regne ut de første 10 verdiene med startpunkt $\mathbf{x}_0 = (0, 0, 0)$, gir vi nå kommandoen

```
>>fikspunkt(0,0,0,9)
```

Output er:

$x_0 =$ 0.0000	$x_1 =$ 1.0000	$x_2 =$ -0.1250	$x_3 =$ 0.3594	$x_4 =$ 0.1074	$x_5 =$ 0.2322	$x_6 =$ 0.1696	$x_7 =$ 0.2009	$x_8 =$ 0.1853	$x_9 =$ 0.1931
$y_0 =$ 0.000	$y_1 =$ 2.0000	$y_2 =$ 2.0000	$y_3 =$ 2.3438	$y_4 =$ 2.3242	$y_5 =$ 2.4077	$y_6 =$ 2.4025	$y_7 =$ 2.4234	$y_8 =$ 2.4221	$y_9 =$ 2.4273
$z_0 =$ 0.0000	$z_1 =$ -1.0000	$z_2 =$ 1.5000	$z_3 =$ 0.9375	$z_4 =$ 1.5234	$z_5 =$ 1.3779	$z_6 =$ 1.5238	$z_7 =$ 1.4874	$z_8 =$ 1.5238	$z_9 =$ 1.5147

Vi ser at følgen $\mathbf{x}_n$ ser ut til å stabilisere seg rundt $(0.2, 2.4, 1.5)$, men at det fortsatt er ganske store fluktuasjoner. Vi gir derfor kommandoen

```
>>fikspunkt(0,0,0,19)
```

for også å få de neste ti verdiene. De er:

$x_{10} =$ 0.1892	$x_{11} =$ 0.1911	$x_{12} =$ 0.1902	$x_{13} =$ 0.1906	$x_{14} =$ 0.1904	$x_{15} =$ 0.1905	$x_{16} =$ 0.1905	$x_{17} =$ 0.1905	$x_{18} =$ 0.1905	$x_{19} =$ 0.1905
$y_{10} =$ 2.4269	$y_{11} =$ 2.4282	$y_{12} =$ 2.4282	$y_{13} =$ 2.4285	$y_{14} =$ 2.4285	$y_{15} =$ 2.4286	$y_{16} =$ 2.4285	$y_{17} =$ 2.4286	$y_{18} =$ 2.4286	$y_{19} =$ 2.4286
$z_{10} =$ 1.5238	$z_{11} =$ 1.5215	$z_{12} =$ 1.5238	$z_{13} =$ 1.5232	$z_{14} =$ 1.5238	$z_{15} =$ 1.5237	$z_{16} =$ 1.5238	$z_{17} =$ 1.5238	$z_{18} =$ 1.5238	$z_{19} =$ 1.5238

Nå ser vi en tydelig konvergens mot et fikspunkt med (tilnærmet) verdi $(0.1905, 2.4286, 1.5239)$.

Programmet `fikspunkt.m` er ganske primitivt; vi må bestemme på forhånd hvor mange iterasjoner vi ønsker. I praksis bruker man ofte mer avanserte programmer (gjerne med en `while`-løkke) som avbryter iterasjonen når man har fått en viss nøyaktighet.

Vi innledet med å si at dette er et ganske dårlig eksempel. Grunnen er at vi kunne ha funnet fikspunktet direkte ved å løse ligningssystemet $\mathbf{F}(x, y, z) = (x, y, z)$ ved regning. Dette er et lineært ligningssystem som er lett å løse med metodene i forrige kapittel. Metoden ovenfor har imidlertid også sine fordeler; vi har vist at fikspunktet er tiltrekkende (dvs. at iterasjoner av $\mathbf{F}$ alltid konvergerer). Det viser seg også at når man får store lineære

ligningssystemer med tusenvis av ligninger og ukjente, så er det ofte raskere å løse dem ved en eller annen form for iterasjon enn ved radoperasjoner. ♣

Et problem med teorien i denne seksjonen er at det ofte kan være vanskelig å vise at den funksjonen **F** som vi arbeider med, virkelig er en kontraksjon. Ofte er ikke **F** en kontraksjon i hele $\mathbb{R}^m$, men bare i mindre områder i nærheten av fikspunktet, og det kan ofte være vanskelig å finne et slikt område A (for at det skal fungere, må **F** avbilde A inn i A). I slike tilfeller kan det ofte være lurt å prøve seg med en fikspunktiterasjon uten å være sikker på at **F** er en kontraksjon. I praktisk arbeid er ofte et fikspunkt funnet på denne måten mer enn godt nok, og i teoretisk arbeid kan informasjon om hvor fikspunktet (sannsynligvis) ligger, gjøre det lettere å lokalisere området der **F** er en kontraksjon.

Det kan være greit å vite at det finnes mer generelle fikspunktteoremer enn Banachs. *Brouwers fikspunktteorem* sier for eksempel at dersom A er en ikke-tom, lukket, begrenset, konveks delmengde av $\mathbb{R}^m$, så vil enhver kontinuerlig funksjon $\mathbf{F} : A \to A$ ha ett eller flere fikspunkt. Det er imidlertid ingen garanti for at disse fikspunktene kan finnes ved iterasjon.

5.5 Newtons metode i flere variable

I forrige kapittel så vi hvordan vi kan løse lineære ligningssystemer med m ligninger og m ukjente:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1m}x_m &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2m}x_m &= b_2 \\ \vdots \qquad\qquad\qquad\qquad\quad &\vdots \quad \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mm}x_m &= b_m \end{aligned}$$

Lineære ligningssystemer er viktige i mange sammenhenger, men de fleste ligningssystemene som dukker opp i praksis, er ikke av denne typen, men inneholder mer kompliserte funksjonsuttrykk. Vi kan f.eks. ha tre mer kompliserte ligninger slik som her:

$$\begin{aligned} 3x^2y + 2e^{z+z} &= 0 \\ 2z^2\cos(xy^2 + z) + e^x &= 0 \\ x^3(y^2 + z) &= 0 \end{aligned}$$

Slike ligningssystemer er som regel umulig å løse eksakt, og man må derfor bruke datamaskin for å finne tilnærmede løsninger. Vi skal se hvordan Newtons metode, som du kjenner for funksjoner av én variabel (se *Kalkulus*, seksjon 7.3), kan utvides til ligningssystemer.

Vi observerer først at det å løse et ligningssystem med m ligninger og m ukjente, er det samme som å finne nullpunktene til en funksjon $\mathbf{F} : \mathbb{R}^m \to \mathbb{R}^m$. Lar vi f.eks. $\mathbf{F} : \mathbb{R}^3 \to \mathbb{R}^3$ være funksjonen

$$\mathbf{F}(x, y, z) = \begin{pmatrix} 3x^2y + 2e^{z+z} \\ 2z^2\cos(xy^2 + z) + e^x \\ x^3(y^2 + z) \end{pmatrix}$$

ser vi at løsningene til ligningene ovenfor er de samme som nullpunktene til funksjonen $\mathbf{F}$. Vi er altså interessert i å finne nullpunkter til funksjoner $\mathbf{F} : \mathbb{R}^m \to \mathbb{R}^m$.

La oss anta at vi vet eller mistenker at funksjonen har et nullpunkt i nærheten av et punkt $\mathbf{x}_0 \in \mathbb{R}^m$. Vi skal forsøke å finne en bedre tilnærming til nullpunktet enn $\mathbf{x}_0$. Husk at i nærheten av $\mathbf{x}_0$ er $\mathbf{F}$ godt tilnærmet av sin linearisering i $\mathbf{x}_0$, altså av den affine funksjonen

$$T_{\mathbf{x}_0}\mathbf{F}(\mathbf{x}) = \mathbf{F}(\mathbf{x}_0) + \mathbf{F}'(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)$$

der $\mathbf{F}'(\mathbf{x}_0)$ er Jacobi-matrisen til $\mathbf{F}$ i punktet $\mathbf{x}_0$ (se seksjon 2.9). Istedenfor å løse den kompliserte ligningen $\mathbf{F}(\mathbf{x}) = \mathbf{0}$, løser vi den enklere ligningen $T_{\mathbf{x}_0}\mathbf{F}(\mathbf{x}) = \mathbf{0}$, dvs. ligningen

$$\mathbf{F}(\mathbf{x}_0) + \mathbf{F}'(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) = \mathbf{0}$$

Dette er et lineært ligningssystem, og hvis Jacobi-matrisen $\mathbf{F}'(\mathbf{x}_0)$ er inverterbar, har det løsningen

$$\mathbf{x} = \mathbf{x}_0 - \mathbf{F}'(\mathbf{x}_0)^{-1}\mathbf{F}(\mathbf{x}_0)$$

Som allerede påpekt er dette en løsning av det forenklede ligningssystemet $T_{\mathbf{x}_0}\mathbf{F}(\mathbf{x}) = \mathbf{0}$ og ikke av det opprinnelige ligningssystemet $\mathbf{F}(\mathbf{x}) = \mathbf{0}$ som vi egentlig vil løse, men med litt flaks er $\mathbf{x}$ en bedre tilnærming til løsningen enn den gjetningen $\mathbf{x}_0$ som vi startet med. Vi lar derfor $\mathbf{x}_1 = \mathbf{x}$ være vår andre tilnærming til løsning.

Vi gjentar nå prosedyren med $\mathbf{x}_1$ som input istedenfor $\mathbf{x}_0$ og får en ny tilnærming

$$\mathbf{x}_2 = \mathbf{x}_1 - \mathbf{F}'(\mathbf{x}_1)^{-1}\mathbf{F}(\mathbf{x}_1)$$

Deretter bruker vi $\mathbf{x}_2$ som input og får en ny tilnærming $\mathbf{x}_3$:

$$\mathbf{x}_3 = \mathbf{x}_2 - \mathbf{F}'(\mathbf{x}_2)^{-1}\mathbf{F}(\mathbf{x}_2)$$

Slik fortsetter vi å følge mønsteret

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \mathbf{F}'(\mathbf{x}_n)^{-1}\mathbf{F}(\mathbf{x}_n)$$

og håper å nærme oss en løsning av det opprinnelige ligningssystemet $\mathbf{F}(\mathbf{x}) = 0$.

Definisjon 5.5.1 *Anta at* $\mathbf{F} : A \to \mathbb{R}^m$ *er en deriverbar funksjon av* m *variable.* Newtons metode *anvendt på* $\mathbf{F}$ *med startpunkt* $\mathbf{x}_0$ *gir oss følgen* $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n, \ldots$ *der*

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \mathbf{F}'(\mathbf{x}_n)^{-1}\mathbf{F}(\mathbf{x}_n)$$

Legg merke til at Newtons metode ikke er noe annet enn fikspunktsiterasjon av funksjonen

$$\mathbf{G}(x) = \mathbf{x} - \mathbf{F}'(\mathbf{x})^{-1}\mathbf{F}(\mathbf{x})$$

Legg også merke til at prosedyren forutsetter at Jacobi-matrisen $\mathbf{F}'(\mathbf{x}_n)$ er inverterbar for alle n.

La oss se på et eksempel.

Eksempel 1: Vi skal bruke Newtons metode til å finne en løsning av ligningssystemet

$$\begin{aligned} x^2y + 1 &= 0 \\ e^x + y &= 0 \end{aligned}$$

Sagt på en annen måte skal vi finne et nullpunkt for funksjonen

$$\mathbf{F}(x, y) = \begin{pmatrix} x^2y + 1 \\ e^x + y \end{pmatrix}$$

Denne funksjonen har Jacobi-determinant

$$\mathbf{F}'(x, y) = \begin{pmatrix} 2xy & x^2 \\ e^x & 1 \end{pmatrix}$$

Lar vi $\mathbf{x}_0 = \begin{pmatrix} x_0 \\ y_0 \end{pmatrix}$, $\mathbf{x}_1 = \begin{pmatrix} x_1 \\ y_1 \end{pmatrix}$, $\mathbf{x}_2 = \begin{pmatrix} x_2 \\ y_2 \end{pmatrix}, \ldots$ være en følge som fremkommer når vi bruker Newtons metode på $\mathbf{F}$, ser vi at iterasjonsformelen

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \mathbf{F}'(\mathbf{x}_n)^{-1}\mathbf{F}(\mathbf{x}_n)$$

kan skrives

$$\begin{pmatrix} x_{n+1} \\ y_{n+1} \end{pmatrix} = \begin{pmatrix} x_n \\ y_n \end{pmatrix} - \begin{pmatrix} 2x_ny_n & x_n^2 \\ e_n^x & 1 \end{pmatrix}^{-1} \begin{pmatrix} x_n^2y_n + 1 \\ e^{x_n} + y_n \end{pmatrix}$$

For å regne ut punktene i følgen skriver vi et lite MATLAB-program. Legg merke til at variablen `u` alltid inneholder det *siste* punktet $\mathbf{x}_n = \begin{pmatrix} x_n \\ y_n \end{pmatrix}$ vi har regnet ut, mens variablene `x` og `y` lagrer hele listen av x- og y-koordinater. Legg også merke til at vi bruker kommandoen `A\v` istedenfor `A^-1v` for å spare regnearbeid (dette er mer effektivt enn å tvinge MATLAB til å regne ut A^{-1}).

```
function [x,y]=newtonfler(a,b,N)
x=zeros(1,N); % sørger for at vektoren x har "riktig" lengde
y=zeros(1,N); % sørger for at vektoren y har "riktig" lengde
x(1)=a;     %laster inn x-koordinaten til startpunktet
y(1)=b;     %laster inn y-koordinaten til startpunktet
u=[a;b];   %setter u lik startpunktet
for n=1:N
A=[2*x(n)*y(n) x(n)^2;exp(x(n)) 1]; %setter A lik Jacobi-matrisen
v=[x(n)^2*y(n)+1;exp(x(n))+y(n)];   %setter v lik funksjonsverdien
u=u-A\v;     %oppdaterer u
x(n+1)=u(1); %finner x(n+1) som førstekoordinaten til u
y(n+1)=u(2); %finner y(n+1) som annenkoordinaten til u
    end
```

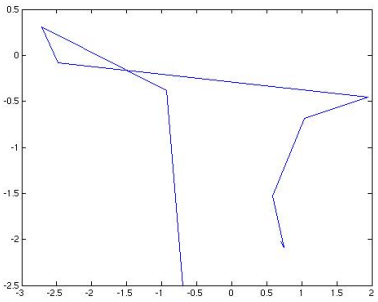
For å kjøre programmet med startpunkt $(-0.7, -2.5)$, gir vi kommandoen

```
>> [x,y]=newtonfler(-.7,-2.5,20);
```

Vi kan få ut x- og y-verdiene ved å skrive henholdsvis `>> x` og `>> y`. Tabellen nedenfor viser de første verdiene:

$x_0 =$ -0.7000	$x_1 =$ -0.9323	$x_2 =$ -2.7167	$x_3 =$ -2.4802	$x_4 =$ 1.9425	$x_5 =$ 1.0409	$x_6 =$ 0.5810	$x_7 =$ 0.7475	$x_8 =$ 0.7056	$x_9 =$ 0.7035	$x_{10} =$ 0.7035
$y_0 =$ -2.5000	$y_1 =$ -0.3812	$y_2 =$ 0.3088	$y_3 =$ -0.0817	$y_4 =$ -0.4540	$y_5 =$ -0.6865	$y_6 =$ -1.5295	$y_7 =$ -2.0855	$y_8 =$ -2.0233	$y_9 =$ -2.0208	$y_{10} =$ -2.0207

Følgen ser altså ut til å konvergere mot et nullpunkt med koordinater tilnærmet lik $(0.7035, -2.0207)$ (de neste tallene i utskriften bekrefter dette inntrykket). For å få bedre oversikt hvordan følgen oppfører seg, kan vi plotte den med kommandoen `>> plot(x,y)`.



Figur 1: Konvergens av Newtons metode

Resultatet ser du i figur 1. Legg merke til at de siste skrittene er så små at du ikke kan se dem på figuren. ♣

Programmet i eksemplet ovenfor er ganske primitivt siden det er skreddersydd for akkurat den funksjonen vi skal studere. Skal man bruke Newtons metode i flere sammenhenger (eller som del av et større program), bør man lage en versjon der man kan laste inn de funksjonene man skal arbeide med.

Konvergens av Newtons metode

I eksemplet ovenfor konvergerte Newtons metode mot et nullpunkt, men i andre eksempler er dette slett ikke tilfellet — vi kan f.eks. få en følge som går mot det uendelig fjerne. Ja, selv i eksemplet ovenfor er vi ikke hundre prosent sikre på at vi virkelig er i nærheten av et nullpunkt; det går an å lage eksempler der det ser ut som vi har konvergens mot et nullpunkt som i virkeligheten ikke finnes. Et tilleggsproblem får vi dersom funksjonen har mer enn ett nullpunkt, og vi ikke er sikre på hvilket nullpunkt Newtons metode konvergerer mot — det kan godt være et helt annet enn det vi er interessert i.

Ser du etter hva som står i *Kalkulus* (seksjon 7.3) eller i de fleste andre lærebøker om Newtons metode, vil du finne en setning av denne typen:

Setning 5.5.2 (Newtons metode i én variabel) *Anta $f : \mathbb{R} \to \mathbb{R}$ har et nullpunkt i a. Dersom $f'(a) \neq 0$, og $f''(x)$ eksisterer og er kontinuerlig i en omegn rundt a, så finnes det en $\delta > 0$ slik at hvis $x_0 \in (a - \delta, a + \delta)$, så konvegerer følgen $\{x_n\}$ i Newtons metode mot a.*

Setninger av denne typen (det finnes tilsvarende for funksjoner av flere variable) forteller deg at dersom du starter nær nok et nullpunkt, så vil Newtons metode (så sant betingelsene er oppfylt) konvergere mot dette nullpunktet. Slike setninger er teoretisk beroligende, men de er til liten nytte når man bruker Newtons metode i praksis — da kjenner man jo ikke nullpunktene, og kan umulig vite om man er "nær nok" eller ikke. Vi skal isteden se på et teorem som har praktisk nytte — det gir nemlig en mulighet til å sjekke konvergens allerede før vi begynner beregningene. Betingelsene i teoremet ser ganske kompliserte ut, men vi skal forklare dem etter at vi har skrevet opp teoremet. Én ting bør vi imidlertid klargjøre på forhånd. I seksjon 5.2 definerte vi *operatornormen* $|A|$ til en kvadratisk matrise A ved

$$|A[= \sup \left\{ \frac{|A\mathbf{x}|}{|\mathbf{x}|} \ : \ \mathbf{x} \in \mathbb{R}^m \right\}$$

og understreket at dette er et (av flere mulige) mål på hvor stor en matrise er. Når vi i teoremet nedenfor har ulikheten

$$|\mathbf{F}'(\mathbf{u}) - \mathbf{F}'(\mathbf{v})| \leq M|\mathbf{u} - \mathbf{v}| \quad \text{for alle } \mathbf{u}, \mathbf{v} \in U_0$$

betyr dette at størrelsen til differensen mellom Jacobi-matrisen i punktet $\mathbf{u}$ og Jacobi-matrisen i punktet $\mathbf{v}$ er mindre enn M ganger størrelsen til

differensen mellom $\mathbf{u}$ og $\mathbf{v}$ — med andre ord: Jacobi-matrisen endrer seg ikke veldig mye raskere enn avstanden mellom punktene. Som vi snart skal se. vil to ganger deriverbare funksjoner oppfylle en slik betingelse, men kanskje med en M som er ganske stor.

La oss også ta med litt notasjon. Vi har tidligere arbeidet med de *åpne* kulene

$$B(\mathbf{x}, r) = \{\mathbf{y} \in \mathbb{R}^m \,:\, |\mathbf{y} - \mathbf{x}| < r\}$$

Nå får vi også bruk for de *lukkede* kulene

$$\overline{B}(\mathbf{x}, r) = \{\mathbf{y} \in \mathbb{R}^m \,:\, |\mathbf{y} - \mathbf{x}| \leq r\}$$

Teorem 5.5.3 (Kantorovitsj' teorem) *La $\mathbf{F} : U \to \mathbb{R}^m$ være en deriverbar funksjon definert på en åpen, konveks delmengde U av $\mathbb{R}^m$, og anta at det finnes en konstant M slik at*

$$|\mathbf{F}'(\mathbf{x}) - \mathbf{F}'(\mathbf{y})| \leq M|\mathbf{x} - \mathbf{y}| \quad \text{for alle } \mathbf{x}, \mathbf{y} \in U$$

La $\mathbf{x}_0$ være et punkt i U, og anta at Jacobi-matrisen $\mathbf{F}'(\mathbf{x}_0)$ i $\mathbf{x}_0$ er inverterbar med

$$|\mathbf{F}'(\mathbf{x}_0)^{-1}| \leq K$$

Anta videre at

$$|\mathbf{F}'(\mathbf{x}_0)^{-1}\mathbf{F}(\mathbf{x}_0)| \leq \epsilon$$

der $h = KM\epsilon \leq \frac{1}{2}$. Anta til slutt at den lukkede kulen $\overline{B}(\mathbf{x}_0, \frac{1}{KM})$ er inneholdt i definisjonsmengden U. Da er $\mathbf{F}'(\mathbf{x})$ inverterbar for alle $\mathbf{x}$ i den åpne kulen $B(\mathbf{x}_0, \frac{1}{KM})$. Starter vi Newtons metode i $\mathbf{x}_0$, vil alle punktene $\mathbf{x}_n$ ligge i $B(\mathbf{x}_0, \frac{1}{KM})$, og de vil konvergere mot et punkt $\mathbf{x} \in \overline{B}(\mathbf{x}_0, \frac{1}{KM})$ der $\mathbf{F}(\mathbf{x}) = \mathbf{0}$.

La oss forsøke å forklare logikken i teoremet. Det er de tre ulikhetene

$$|\mathbf{F}'(\mathbf{x}) - \mathbf{F}'(\mathbf{y})| \leq M|\mathbf{x} - \mathbf{y}|$$

$$|\mathbf{F}'(\mathbf{x}_0)^{-1}| \leq K$$

$$|\mathbf{F}'(\mathbf{x}_0)^{-1}\mathbf{F}(\mathbf{x}_0)| \leq \epsilon$$

som får teoremet igang, og hovedbetingelsen er at produktet $KM\epsilon$ av de tre skrankene M, K og ϵ skal være mindre enn $\frac{1}{2}$. Som allerede nevnt er den første ulikheten ($|\mathbf{F}'(\mathbf{x}) - \mathbf{F}'(\mathbf{y})| \leq M|\mathbf{x} - \mathbf{y}|$) som regel oppfylt (i hvert fall på begrensede områder), men det er ingen grunn til å tro at M er liten. Heller ikke K behøver å være liten; selv om vi er i nærheten av et nullpunkt, kan $|\mathbf{F}'(\mathbf{x}_0)^{-1}|$ godt være stor. Den tredje betingelsen har vi imidlertid bedre kontroll over: Når $\mathbf{x}_0$ nærmer seg et nullpunkt, vil $\mathbf{F}(\mathbf{x}_0)$ gå mot null, og vi kan normalt få $|\mathbf{F}'(\mathbf{x}_0)^{-1}\mathbf{F}(\mathbf{x}_0)|$ så liten vi måtte ønske, f.eks. så liten at $h = KM\epsilon \leq \frac{1}{2}$. Teoremet forteller oss altså at under milde betingelser vil Newtons metode konvergere mot et nullpunkt dersom vi starter iterasjonen

i et fornuftig punkt $\mathbf{x}_0$. Styrken i teoremet er at vi kan undersøke om betingelsene er oppfylt uten å vite noe om eventuelle nullpunkter — betingelsene kan sjekkes med en gang vi har valgt startpunkt $\mathbf{x}_0$ (en annen sak er at det er mye lettere å finne et startpunkt der betingelsene er oppfylt dersom vi har en viss anelse om hvor nullpunktene ligger!) Vær forøvrig oppmerksom på at følgen $\{\mathbf{x}_n\}$ godt kan konvergere uten at betingelsene er oppfylt. I mange tilfeller der Newtons metode konvergerer, vil betingelsene først være oppfylt etter at vi har gjennomført noen iterasjoner. Når man bruker metoden i praksis, lønner det seg derfor å la den foreta noen iterasjoner før man forkaster håpet om å finne et nullpunkt.

Før vi går videre, tar vi med to tilleggsresultater — ett om entydighet av nullpunktet og ett om konvergenshastighet. Begge er egentlig del av Kantorovitsj' resultat, men vi velger å presentere dem separat for ikke å få et altfor overlesset teorem. Vi kommer tilbake med bevisene etter at vi har bevist Kantorovitsj' teorem.

Setning 5.5.4 *Anta at betingelsene i Kantorovitsj' teorem er oppfylt. Da har* $\mathbf{F}$ *nøyaktig ett nullpunkt i den lukkede kulen* $\overline{B}(\mathbf{x}_0, \frac{1}{KM})$*, nemlig det vi finner ved å starte Newtons metode i* $\mathbf{x}_0$*.*

Setning 5.5.5 *Anta at betingelsene i Kantorovitsj' teorem er oppfylt, og la* $\mathbf{x}$ *være grensepunktet for følgen* $\{\mathbf{x}_n\}$ *gitt av Newtons metode med startpunkt* $\mathbf{x}_0$*. Da er*

$$|\mathbf{x} - \mathbf{x}_n| \leq \frac{1}{KM}\left(\frac{(1-\sqrt{1-2h})^{2^n}}{2^n}\right)$$

(husk at $h = KM\epsilon \leq \frac{1}{2}$*).*

Det siste resultatet ser litt mystisk ut før man har lest beviset for Kantorovitsj's teorem, men observer at når $h < \frac{1}{2}$, så er $1 - \sqrt{1-2h} < 1$. Uttrykket $(1-\sqrt{1-2h})^{2^n}$ går derfor svært raskt mot 0 (man kaller ofte fenomenet *superkonvergens*).

Før vi går løs på bevisene for Kantorovitsj' teorem og de to tilleggssetningene ovenfor, er det et lite problem vi bør ta oss av — hvordan finner vi i praksis et tall M slik at $|\mathbf{F}'(\mathbf{u}) - \mathbf{F}'(\mathbf{v})| \leq M|\mathbf{u} - \mathbf{v}|$ for alle $\mathbf{u}, \mathbf{v}$ i en mengde U? Setningen nedenfor gir oss en nyttig metode, men vær klar over at den ofte gir en M-verdi som er mye større enn nødvendig. Legg merke til at setningen (og beviset) er en litt mer komplisert variant av setning 5.4.6.

Setning 5.5.6 *Anta at* U *er en åpen, konveks delmengde av* $\mathbb{R}^m$ *og at* $\mathbf{F} : U \to \mathbb{R}^m$ *har kontinuerlige annenderiverte. Anta at det for alle tripler* (i, j, k)*,* $1 \leq i, j, k \leq m$*, finnes tall* $m_{i,j,k}$ *slik at*

$$\left|\frac{\partial^2 F_i}{\partial x_k \partial x_j}(\mathbf{x})\right| \leq m_{i,j,k} \qquad \text{for alle } \mathbf{x} \in U$$

Sett

$$M = (\sum_{1\leq i,j,k\leq m} m_{i,j,k}^2)^{\frac{1}{2}}$$

Da er

$$|\mathbf{F}'(\mathbf{u}) - \mathbf{F}'(\mathbf{v})| \leq M|\mathbf{u}-\mathbf{v}| \qquad \textit{for alle } \mathbf{u},\mathbf{v} \in U$$

Bevis: Det i,j-te elementet i matrisen $\mathbf{F}'(\mathbf{u}) - \mathbf{F}'(\mathbf{v})$ er $\frac{\partial F_i}{\partial x_j}(\mathbf{u}) - \frac{\partial F_i}{\partial x_j}(\mathbf{v})$. Ifølge middelverdisetningen for funksjoner av flere variable (setning 5.4.5), finnes det et punkt $\mathbf{c}_{i,j}$ på linjestykket mellom $\mathbf{u}$ og $\mathbf{v}$ slik at

$$\frac{\partial F_i}{\partial x_j}(\mathbf{u}) - \frac{\partial F_i}{\partial x_j}(\mathbf{v}) = \nabla\left(\frac{\partial F_i}{\partial x_j}\right)(\mathbf{c}_{i,j}) \cdot (\mathbf{u}-\mathbf{v})$$

Ifølge Schwartz' ulikhet er dermed

$$\left|\frac{\partial F_i}{\partial x_j}(\mathbf{u}) - \frac{\partial F_i}{\partial x_j}(\mathbf{v})\right| \leq \left|\nabla\left(\frac{\partial F_i}{\partial x_j}\right)(\mathbf{c}_{i,j})\right| |\mathbf{u}-\mathbf{v}|$$

Bruker vi at

$$\left|\nabla\left(\frac{\partial F_i}{\partial x_j}\right)(\mathbf{c}_{i,j})\right| = \sqrt{\left(\frac{\partial^2 F_i}{\partial x_1 \partial x_j}(\mathbf{c}_{i,j})\right)^2 + \cdots + \left(\frac{\partial^2 F_i}{\partial x_m \partial x_j}(\mathbf{c}_{i,j})\right)^2} \leq \sqrt{\sum_{k=1}^m m_{i,j,k}^2}\ ,$$

får vi

$$\left|\frac{\partial F_i}{\partial x_j}(\mathbf{u}) - \frac{\partial F_i}{\partial x_j}(\mathbf{v})\right| \leq \sqrt{\sum_{k=1}^m m_{i,j,k}^2}\, |\mathbf{u}-\mathbf{v}|$$

Husk at *normen* $\|A\|$ til en $m \times m$-matrise A med komponenter a_{ij} er gitt ved

$$\|A\| = \sqrt{\sum_{1\leq i,j\leq m} a_{ij}^2}$$

Dermed er

$$\|\mathbf{F}'(\mathbf{u}) - \mathbf{F}'(\mathbf{v})\| = \sqrt{\sum_{i,j=1}^m \left(\frac{\partial F_i}{\partial x_j}(\mathbf{u}) - \frac{\partial F_i}{\partial x_j}(\mathbf{v})\right)^2} \leq$$

$$\leq \sqrt{\sum_{1\leq i,j\leq m} \left(\sqrt{\sum_{k=1}^m m_{i,j,k}^2}\, |\mathbf{u}-\mathbf{v}|\right)^2} = \sqrt{\sum_{1\leq i,j,k\leq m} m_{i,j,k}^2}|\mathbf{u}-\mathbf{v}| = M|\mathbf{u}-\mathbf{v}|$$

Siden *operatornormen* $|\mathbf{F}'(\mathbf{u}) - \mathbf{F}'(\mathbf{v})|$ er mindre enn *normen* $\|\mathbf{F}'(\mathbf{u}) - \mathbf{F}'(\mathbf{v})\|$ (se seksjon 5.2), følger det at

$$|\mathbf{F}'(\mathbf{u}) - \mathbf{F}'(\mathbf{v})| \leq M|\mathbf{u}-\mathbf{v}|$$

□

*Bevis for Kantorovitsj' teorem

Vi skal nå bevise Kantorovitsj' teorem. Før vi begynner, gjør vi oppmerksom på at argumentet er både langt og krevende sammenlignet med de fleste andre bevisene i dette heftet. Det gir imidlertid en flott illustrasjon av hvordan teknikkene våre kan brukes til å bevise avanserte matematiske resultater. Før vi kommer til selve beviset for Kantorovitsj' teorem, skal vi bevise tre lemmaer. I tillegg skal vi gjøre utstrakt bruk av Banachs lemma fra seksjon 5.2, så har du ikke lest det før, bør du gjøre det nå.

Det første lemmaet vårt vil hjelpe oss å utnytte betingelsen $|\mathbf{F}'(\mathbf{u}) - \mathbf{F}'(\mathbf{v})| \leq M|\mathbf{u}-\mathbf{v}|$ i Kantorovitsj' teorem.

Lemma 5.5.7 *Anta at U er en åpen, konveks delmengde av $\mathbb{R}^m$, og at $\mathbf{F} : U \to \mathbb{R}^m$ er en deriverbar funksjon slik at*

$$|\mathbf{F}'(\mathbf{u}) - \mathbf{F}'(\mathbf{v})| \leq M|\mathbf{u}-\mathbf{v}| \quad \textit{for alle } \mathbf{u}, \mathbf{v} \in U$$

Da er

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

for alle $\mathbf{x}, \mathbf{y} \in U$.

Bevis: Siden U er konveks, ligger alle punkter på linjestykket mellom $\mathbf{x}$ og $\mathbf{y}$ i U (dette er definisjonen av konveksitet). Linjestykket er parametrisert ved

$$\mathbf{r}(t) = \mathbf{x} + t(\mathbf{y}-\mathbf{x}) \qquad \text{der } t \in [0,1],$$

og deriverer vi $\mathbf{G}(t) = \mathbf{F}(\mathbf{r}(t))$, gir kjerneregelen

$$\mathbf{G}'(t) = \mathbf{F}'(\mathbf{r}(t))(\mathbf{y}-\mathbf{x})$$

som vi kan skrive om til

$$\mathbf{G}'(t) = \mathbf{F}'(\mathbf{x})(\mathbf{y}-\mathbf{x}) + \big(\mathbf{F}'(\mathbf{r}(t)) - \mathbf{F}'(\mathbf{x})\big)(\mathbf{y}-\mathbf{x})$$

Integrerer vi på begge sider, får vi

$$\mathbf{F}(\mathbf{y}) - \mathbf{F}(\mathbf{x}) = \mathbf{G}(1) - \mathbf{G}(0) = \int_0^1 \mathbf{G}'(t)\,dt =$$

$$= \int_0^1 \big(\mathbf{F}'(\mathbf{x})(\mathbf{y}-\mathbf{x}) + \big(\mathbf{F}'(\mathbf{r}(t)) - \mathbf{F}'(\mathbf{x})\big)(\mathbf{y}-\mathbf{x})\big)\,dt =$$

$$= \mathbf{F}'(\mathbf{x})(\mathbf{y}-\mathbf{x}) + \int_0^1 \big(\mathbf{F}'(\mathbf{r}(t)) - \mathbf{F}'(\mathbf{x})\big)(\mathbf{y}-\mathbf{x})\,dt$$

der vi i det siste skrittet har brukt at $\mathbf{F}'(\mathbf{x})(\mathbf{y}-\mathbf{x})$ er konstant. Dette betyr at

$$\big|\mathbf{F}(\mathbf{y}) - \mathbf{F}(\mathbf{x}) - \mathbf{F}'(\mathbf{x})(\mathbf{y}-\mathbf{x})\big| = \left|\int_0^1 \big(\mathbf{F}'(\mathbf{r}(t)) - \mathbf{F}'(\mathbf{x})\big)(\mathbf{y}-\mathbf{x})\,dt\right| \leq$$

$$\leq \int_0^1 \left|\left(\mathbf{F}'(\mathbf{r}(t)) - \mathbf{F}'(\mathbf{x})\right|\,|\mathbf{y}-\mathbf{x}|\right)\,dt \leq M\int_0^1 |\mathbf{r}(t)-\mathbf{x}||\mathbf{y}-\mathbf{x}|\,dt$$

Bruker vi at $\mathbf{r}(t)-\mathbf{x} = (\mathbf{x}+t(\mathbf{y}-\mathbf{x}))-\mathbf{x} = t(\mathbf{y}-\mathbf{x})$, får vi

$$\left|\mathbf{F}(\mathbf{y})-\mathbf{F}(\mathbf{x})-\mathbf{F}'(\mathbf{x})(\mathbf{y}-\mathbf{x})\right| \leq M\int_0^1 t|\mathbf{y}-\mathbf{x}|^2\,dt = \frac{M}{2}|\mathbf{y}-\mathbf{x}|^2 \qquad \square$$

Vi skal bevise Kantorovitsj' teorem ved å sammenligne følgen $\{\mathbf{x}_n\}$ med en atskillig enklere tallfølge. Her er det grunnleggende sammenligningsprinsippet:

Lemma 5.5.8 *La $\{\mathbf{x}_n\}$ være en følge av punkter i $\mathbb{R}^m$. Anta at det finnes en voksende følge $\{t_n\}$ av punkter på tallinjen som konvergerer mot et tall t, og som er slik at*

$$|\mathbf{x}_{n+1}-\mathbf{x}_n| \leq t_{n+1}-t_n$$

for alle n. Da konvergerer $\{\mathbf{x}_n\}$ mot et punkt $\mathbf{x}\in\mathbb{R}^m$ og

$$|\mathbf{x}-\mathbf{x}_n| \leq t-t_n$$

for alle n.

Bevis: Vi begynner med å vise at $\{\mathbf{x}_n\}$ er en Cauchy-følge. Hvis $k>n$, har vi

$$|\mathbf{x}_k-\mathbf{x}_n| = |(\mathbf{x}_k-\mathbf{x}_{k-1})+(\mathbf{x}_{k-1}-\mathbf{x}_{k-2})+\cdots+(\mathbf{x}_{n+1}-\mathbf{x}_n)| \leq$$

$$\leq |\mathbf{x}_k-\mathbf{x}_{k-1}|+|\mathbf{x}_{k-1}-\mathbf{x}_{k-2}|+\cdots+|\mathbf{x}_{n+1}-\mathbf{x}_n)| \leq$$

$$\leq (t_k-t_{k-1})+(t_{k-1}-t_{k-2})+\cdots+(t_{n+1}-t_n) = t_k-t_n$$

Siden følgen $\{t_n\}$ konvergerer, er den en Caychy-følge, og følgelig kan vi få t_k-t_n så liten vi vil ved å velge n og k store nok. Men dermed kan vi også få $|\mathbf{x}_k-\mathbf{x}_k|$ så liten vi vil, og følgelig er $\{\mathbf{x}_n\}$ en Cauchy-følge og må konvergere mot et punkt $\mathbf{x}$.

Fra ulikhetene ovenfor vet vi at $|\mathbf{x}_k-\mathbf{x}_n| \leq t_k-t_n$. Holder vi n fast og lar $k\to\infty$, får vi $|\mathbf{x}-\mathbf{x}_n| \leq t-t_n$. $\square$

I lemmaet ovenfor sier vi at tallfølgen $\{t_n\}$ *majoriserer* den opprinnelige følgen $\{\mathbf{x}_n\}$. I beviset for Kantorovitsj' teorem skal vi majorisere den opprinnelige følgen $\{\mathbf{x}_n\}$ ved hjelp av en tallfølge $\{t_n\}$ som fremkommer ved å bruke Newtons metode på en enkel annengradsfunksjon $P(t)$ på tallinjen. Det siste lemmaet (som ser atskillig verre ut enn det er) vil gi oss de nødvendige verktøyene for å gjennomføre denne majoriseringen.

Lemma 5.5.9 *Anta at a, b og c er positive reelle tall, og la $P : \mathbb{R} \to \mathbb{R}$ være annengradspolynomet*

$$P(t) = at^2 - bt + c$$

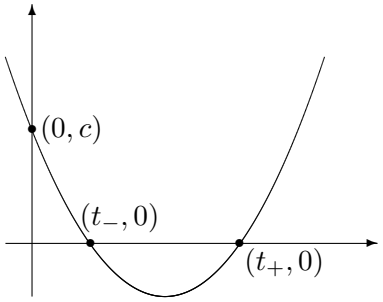
Anta at P har reelle røtter

$$t_\pm = \frac{b \pm \sqrt{b^2 - 4ac}}{2a}$$

og la t_- være den minste roten $t_- = \frac{b-\sqrt{b^2-4ac}}{2a}$. Anta at $\{t_n\}$ er den følgen vi får når vi bruker Newtons metode på $P(t)$ med startverdi $t_0 = 0$. Da er $\{t_n\}$ en voksende følge som konvergerer mot t_-. Punktene i følgen tilfredstiller ligningen

$$t_{n+1} - t_n = \frac{a(t_n - t_{n-1})^2}{b - 2at_n} \tag{5.5.1}$$

Bevis: Siden a, b og c er positive og polynomet har reelle røtter, er det lett å sjekke at grafen må se ut som på figur 2. Det er også lett å se grafisk at følgen vi får ved å starte Newtons metode i $t_0 = 0$, må være voksende og konvergere mot t_- (se oppgavene til denne seksjonen).



Figur 2: En parabel $P(t) = at^2 - bt + c$

Det gjenstår å vise ligning (5.5.1). Her trenger vi et litt fiffig regnestykke. Siden $P'(t) = 2at - b$, er Newtons metode gitt ved

$$t_n = t_{n-1} - \frac{P(t_{n-1})}{P'(t_{n-1})} = t_{n-1} - \frac{at_{n-1}^2 - bt_{n-1} + c}{2at_{n-1} - b} = \frac{at_{n-1}^2 - c}{2at_{n-1} - b}$$

Ganger vi denne likheten $t_n = \frac{at_{n-1}^2 - c}{2at_{n-1} - b}$ med $2at_{n-1} - b$, får vi

$$2at_nt_{n-1} - bt_n = at_{n-1}^2 - c$$

som igjen gir

$$-bt_n + c = -2at_nt_{n-1} + at_{n-1}^2 \tag{5.5.2}$$

Etter Newtons metode er også

$$t_{n+1} - t_n = -\frac{P(t_n)}{P'(t_n)} = -\frac{at_n^2 - bt_n + c}{2at_n - b}$$

Bruker vi (5.5.2) og annen kvadratsetning, kan dette skrives som

$$t_{n+1} - t_n = -\frac{at_n^2 - 2at_nt_{n-1} + at_{n-1}^2}{2at_n - b} = \frac{a(t_n - t_{n-1})^2}{b - 2at_n}$$

□

Vi er nå klar til å bevise Kantorovitsj' teorem. For å gjøre det enklere å lese beviset, skrive vi opp teoremet på nytt.

Teorem 5.5.10 (Kantorovitsj' teorem) *La $\mathbf{F} : U \to \mathbb{R}^m$ være en deriverbar funksjon definert på en åpen, konveks delmengde U av $\mathbb{R}^m$, og anta at det finnes en konstant M slik at*

$$|\mathbf{F}'(\mathbf{x}) - \mathbf{F}'(\mathbf{y})| \leq M|\mathbf{x} - \mathbf{y}| \quad \textit{for alle } \mathbf{x}, \mathbf{y} \in U$$

La $\mathbf{x}_0$ være et punkt i U, og anta at Jacobi-matrisen $\mathbf{F}'(\mathbf{x}_0)$ i $\mathbf{x}_0$ er inverterbar med

$$|\mathbf{F}'(\mathbf{x}_0)^{-1}| \leq K$$

Anta videre at

$$|\mathbf{F}'(\mathbf{x}_0)^{-1}\mathbf{F}(\mathbf{x}_0)| \leq \epsilon$$

der $h = KM\epsilon \leq \frac{1}{2}$. Anta til slutt at den lukkede kulen $\overline{B}(\mathbf{x}_0, \frac{1}{KM})$ er inneholdt i definisjonsmengden U. Da er $\mathbf{F}'(\mathbf{x})$ inverterbar for alle $\mathbf{x}$ i den åpne kulen $B(\mathbf{x}_0, \frac{1}{KM})$. Starter vi Newtons metode i $\mathbf{x}_0$, vil alle punktene $\mathbf{x}_n$ ligge i $B(\mathbf{x}_0, \frac{1}{KM})$, og de vil konvergere mot et punkt $\mathbf{x} \in \overline{B}(\mathbf{x}_0, \frac{1}{KM})$ der $\mathbf{F}(\mathbf{x}) = \mathbf{0}$.

Bevis: Ifølge Banachs lemma (5.2.14) er $\mathbf{F}'(\mathbf{x})$ inverterbar dersom $|\mathbf{F}'(\mathbf{x}) - \mathbf{F}'(\mathbf{x}_0)| < |\mathbf{F}'(\mathbf{x}_0)^{-1}|^{-1}$. Siden $|\mathbf{F}'(\mathbf{x}) - \mathbf{F}'(\mathbf{x}_0)| \leq M|\mathbf{x} - \mathbf{x}_0|$ og $K^{-1} \leq |\mathbf{F}'(\mathbf{x}_0)^{-1}|^{-1}$, er denne betingelsen oppfylt dersom $M|\mathbf{x} - \mathbf{x}_0| < K^{-1}$, dvs. når $|\mathbf{x} - \mathbf{x}_0| < \frac{1}{KM}$. Altså er $\mathbf{F}'(\mathbf{x})$ inverterbar når $\mathbf{x} \in B(\mathbf{x}_0, \frac{1}{KM})$. Fra Banachs lemma vet vi også at

$$|\mathbf{F}'(\mathbf{x}_n)^{-1}| \leq \frac{|\mathbf{F}'(\mathbf{x}_0)^{-1}|}{1 - |\mathbf{F}'(\mathbf{x}_0)^{-1}||\mathbf{F}'(\mathbf{x}_n) - \mathbf{F}(\mathbf{x}_0)|} \leq \frac{K}{1 - KM|\mathbf{x}_n - \mathbf{x}_0|}$$

Vi har dermed

$$|\mathbf{x}_{n+1} - \mathbf{x}_n| = |-\mathbf{F}'(\mathbf{x}_n)^{-1}\mathbf{F}(\mathbf{x}_n)| \leq |\mathbf{F}'(\mathbf{x}_n)^{-1}||\mathbf{F}(\mathbf{x}_n)| \leq \frac{K|\mathbf{F}(\mathbf{x}_n)|}{1 - KM|\mathbf{x}_n - \mathbf{x}_0|}$$

Bruker vi at $\mathbf{F}(\mathbf{x}_{n-1}) + \mathbf{F}'(\mathbf{x}_{n-1})(\mathbf{x}_n - \mathbf{x}_n) = \mathbf{0}$ (dette er selve utgangspunktet for Newtons metode), ser vi at

$$|\mathbf{F}(\mathbf{x}_n)| = |\mathbf{F}(\mathbf{x}_n) + \mathbf{F}(\mathbf{x}_{n-1}) - \mathbf{F}'(\mathbf{x}_{n-1})(\mathbf{x}_n - \mathbf{x}_n)| \leq \frac{M}{2}|\mathbf{x}_n - \mathbf{x}_{n-1}|^2$$

ifølge lemma 5.5.7. Setter vi dette inn i ulikheten ovenfor, får vi

$$|\mathbf{x}_{n+1} - \mathbf{x}_n| \leq \frac{KM|\mathbf{x}_n - \mathbf{x}_{n-1}|^2}{2(1 - KM|\mathbf{x}_n - \mathbf{x}_0|)}$$

Denne ulikheten minner om likheten (5.5.1) i lemma 5.5.6, og det er denne observasjonen som er utgangspunktet for vårt majoriseringstriks.

La $P(t) = \frac{KM}{2}t^2 - t + \epsilon$. Løser vi annengradsligningen $P(t) = 0$, får vi

$$t = \frac{1 \pm \sqrt{1 - 4\frac{KM}{2}\epsilon}}{2\frac{KM}{2}} = \frac{1 \pm \sqrt{1-2h}}{KM}$$

der h er størrelsen $h = KM\epsilon < \frac{1}{2}$ i teoremet. Dette betyr at betingelsene i lemma 5.5.6 oppfylt, og vi vet dermed at hvis vi bruker Newtons metode på $P(t)$ med startpunkt $t_0 = 0$, så vil følgen $\{t_n\}$ vokse mot grenseverdien

$$t_- = \frac{1 - \sqrt{1-2h}}{KM}$$

Legg merke til at $t_- \leq \frac{1}{KM}$.

Fra lemma 5.5.6 vet vi også at

$$t_{n+1} - t_n = \frac{KM(t_n - t_{n-1})^2}{2(1 - KMt_n)}$$

Ved induksjon er det nå lett å vise at $|\mathbf{x}_{n+1} - \mathbf{x}_n| \leq t_{n+1} - t_n$. For $n = 0$ følger dette av at $t_1 - t_0 = -\frac{P(0)}{P'(0)} = -\frac{\epsilon}{-1} = \epsilon$, mens $|\mathbf{x}_1 - \mathbf{x}_0| = |-\mathbf{F}'(\mathbf{x}_0)^{-1}\mathbf{F}(\mathbf{x}_0)| \leq \epsilon$ ifølge en av betingelsene i teoremet. For å gjennomføre induksjonstrinnet antar vi at $|\mathbf{x}_{k+1} - \mathbf{x}_k| \leq t_{k+1} - t_k$ for alle $k < n$. Da er $|\mathbf{x}_n - \mathbf{x}_0| \leq t_n - t_0 = t_n$, og vi har

$$|\mathbf{x}_{n+1} - \mathbf{x}_n| \leq \frac{KM|\mathbf{x}_n - \mathbf{x}_{n-1}|^2}{2(1 - KM|\mathbf{x}_n - \mathbf{x}_0|)} \leq \frac{KM(t_n - t_{n-1})^2}{2(1 - KMt_n)} = t_{n+1} - t_n$$

Dette viser at $\{t_n\}$ majoriserer $\{\mathbf{x}_n\}$. Legg merke til at siden $|\mathbf{x}_n - \mathbf{x}_0| \leq t_n < t_- < \frac{1}{KM}$, er følgen $\{\mathbf{x}_n\}$ hele tiden innenfor den kulen $B(\mathbf{x}_0, \frac{1}{MK})$ der vi vet at $\mathbf{F}'(x)$ er inverterbar, og vi har derfor ingen problemer med å utføre Newton-iterasjonen uendelig mange ganger. Siden $\{t_n\}$ majoriserer $\{\mathbf{x}_n\}$, konvergerer $\mathbf{x}_n$ mot et punkt $\mathbf{x}$ ifølge lemma 5.5.5. Siden $\mathbf{x}_n \in B(\mathbf{x}_0, \frac{1}{MK})$, må grensepunktet $\mathbf{x}$ ligge i den lukkede kulen $\overline{B}(\mathbf{x}_0, \frac{1}{MK})$ (som regel vil det ligge i den åpne kulen $B(\mathbf{x}_0, \frac{1}{MK})$).

Det gjenstår å vise at $\mathbf{F}(\mathbf{x}) = \mathbf{0}$. Vi har allerede vist at $|\mathbf{F}(\mathbf{x}_n)| \leq \frac{M}{2}|\mathbf{x}_n - \mathbf{x}_{n-1}|^2$. Siden $\{\mathbf{x}_n\}$ konvergerer, går uttrykket på høyre side mot null, og følgelig er $\lim_{n\to\infty} |\mathbf{F}(\mathbf{x}_n)| = 0$. Siden $\mathbf{F}$ er kontinurlig, følger det at $\mathbf{F}(\mathbf{x}) = 0$. □

Bevis for setning 5.5.4: Fra Kantorovitsj' teorem vet vi at $\mathbf{F}$ har minst ett nullpunkt $\mathbf{x}$ som er grensen for følgen $\{\mathbf{x}_n\}$ generert av Newtons metode med startpunkt $\mathbf{x}_0$. Vi skal vise at dersom $\mathbf{y}$ er et vilkårlig nullpunkt for $\mathbf{F}$ i $\overline{B}(\mathbf{x}_0, \frac{1}{KM})$, så vil $|\mathbf{y} - \mathbf{x}_n| \to 0$. Dette medfører åpenbart at $\mathbf{x} = \mathbf{y}$, og følgelig finnes det bare ett nullpunkt.

Det viser seg at i dette beviset lønner det seg å ha ϵ så stor som mulig. Har vi hittil greid oss med en ϵ slik at $h = KM\epsilon < \frac{1}{2}$, bytter vi den nå ut med en større slik at $h = KM\epsilon = \frac{1}{2}$. Siden den eneste betingelsen ϵ ellers skal oppfylle er $|\mathbf{F}'(\mathbf{x}_0)^{-1}\mathbf{F}(\mathbf{x}_0)| \leq \epsilon$, kan vi alltid gjøre dette byttet uten å ødelegge forutsetningen i Kantorovitsj' teorem.

Fra lemma 5.5.6 vet vi at hvis

$$\mathbf{r}_n = \mathbf{F}(\mathbf{y}) - \mathbf{F}(\mathbf{x}_n) - \mathbf{F}'(\mathbf{x}_n)(\mathbf{y} - \mathbf{x}_n)$$

så er

$$|\mathbf{r}_n| \leq \frac{M}{2}|\mathbf{y} - \mathbf{x}_n|^2$$

Siden $\mathbf{F}(\mathbf{y}) = \mathbf{0}$, får vi også fra ligningen ovenfor at

$$\mathbf{y} - \mathbf{x}_n + \mathbf{F}'(\mathbf{x}_n)^{-1}\mathbf{F}(\mathbf{x}_n) = -\mathbf{F}'(\mathbf{x}_n)^{-1}\mathbf{r}_n$$

Siden $\mathbf{x}_{n+1} = \mathbf{x}_n - \mathbf{F}'(\mathbf{x}_n)^{-1}\mathbf{F}(\mathbf{x}_n)$, gir dette

$$\mathbf{y} - \mathbf{x}_{n+1} = -\mathbf{F}'(\mathbf{x}_n)^{-1}\mathbf{r}_n$$

Fra beviset for Kantorovitsj' teorem vet vi at

$$|\mathbf{F}'(\mathbf{x}_n)^{-1}| \leq \frac{K}{1 - KM|\mathbf{x}_n - \mathbf{x}_0|}$$

og dermed er

$$|\mathbf{y} - \mathbf{x}_{n+1}| \leq \frac{KM|\mathbf{y} - \mathbf{x}_n|^2}{1 - KM|\mathbf{x}_n - \mathbf{x}_0|} \leq \frac{KM|\mathbf{y} - \mathbf{x}_n|^2}{1 - KM|t_n|}$$

der $\{t_n\}$ er den majoriserende tallfølgen fra beviset for Kantorovitsj' teorem. Siden

$$t_{n+1} - t_n = \frac{KM(t_n - t_{n-1})^2}{2(1 - KMt_n)}$$

er tanken nå å vise ved induksjon at $|\mathbf{y} - \mathbf{x}_n| \leq t_n - t_{n-1}$ for alle n. Siden $(t_n - t_{n-1}) \to 0$, vil da $|\mathbf{y} - \mathbf{x}_n| \to 0$. Induksjonen er enkel hvis vi bare kan få startbetingelsene til å stemme. Ved å bruke Newtons metode baklengs, ser vi at $t_{-1} = -\sqrt{\frac{2\epsilon}{KM}} = -\frac{1}{KM}$ (husk at vi har valgt ϵ slik at $KM\epsilon = \frac{1}{2}$). Dermed er $t_0 - t_{-1} = \frac{1}{KM}$. Siden $\mathbf{y} \in \overline{B}(\mathbf{x}_0, \frac{1}{KM})$, er $|\mathbf{y} - \mathbf{x}_0| \leq \frac{1}{KM}$. Med dette har vi $|\mathbf{y} - \mathbf{x}_0| \leq t_0 - t_{-1}$, og siden resten av induksjonen går greit, er setningen bevist □

Før vi beviser setning 5.5.5, skal vi se på et spesialtilfelle vi får bruk for underveis.

Lemma 5.5.11 *La $P(t)$ være polynomet $\frac{KM}{2}t^2 - t + \epsilon$ i det tilfellet hvor $h = KM\epsilon = \frac{1}{2}$, dvs. at*

$$P(t) = \frac{KM}{2}t^2 - t + \frac{1}{2KM}$$

La $\{t_n\}$ være følgen vi får når vi bruker Newtons metode på $P(t)$ med $t_0 = 0$, og la t_- være nullpunktet denne følgen konvergerer mot. Da er $t_- - t_n = \frac{2^{-n}}{KM}$.

Bevis: Siden

$$P(t) = \frac{KM}{2}t^2 - t + \frac{1}{2KM} = \frac{KM}{2}\left(t - \frac{1}{KM}\right)^2$$

er $t_- = \frac{1}{KM}$. Vi har også $P'(t) = KM\left(t - \frac{1}{KM}\right)$, så Newtons metode gir

$$t_{n+1} = t_n - \frac{\frac{KM}{2}\left(t_n - \frac{1}{KM}\right)^2}{KM\left(t_n - \frac{1}{KM}\right)} = \frac{1}{2}t_n + \frac{1}{2KM}$$

En liten omforming leder til

$$\frac{1}{KM} - t_{n+1} = \frac{1}{2}\left(\frac{1}{KM} - t_n\right)$$

som ved induksjon gir

$$\frac{1}{KM} - t_n = 2^{-n}(\frac{1}{KM} - t_0) = \frac{2^{-n}}{KM}$$

der vi har brukt at $t_0 = 0$. Siden $t_- = \frac{1}{KM}$, følger setningen □

Bevis for setning 5.5.5: La $\{t_n\}$ være tallfølgen som majoriserer $\{\mathbf{x}_n\}$ i beviset for Kantorovitsj' teorem. Ifølge lemma 5.5.7 er $|\mathbf{x} - \mathbf{x}_n| \leq t_- - t_n$, så det holder å vise at

$$t_- - t_n \leq \frac{1}{KM}\left(\frac{(1 - \sqrt{1 - 2h})^{2^n}}{2^n}\right)$$

Vi begynner med å utlede identiteten

$$(t_- - t_{n+1}) = \frac{KM(t_- - t_n)^2}{2(1 - KMt_n)}$$

Bruker vi formelen for Newtons metode på $P(t) = \frac{KM}{2} - t + \epsilon$, får vi

$$t_- - t_{n+1} = t_- - t_n + \frac{\frac{KM}{2}t_n^2 - t_n + \epsilon}{KMt_n - 1} = \frac{KMt_-t_n - t_- + \frac{KM}{2}t_n^2 - t_n + \epsilon}{KMt_n - 1}$$

Siden t_- er en løsning av ligningen $\frac{KM}{2}t^2 - t + \epsilon = 0$, er

$$t_- = \frac{KM}{2}t_-^2 + \epsilon$$

Setter vi dette inn for (den ene forekomsten av) t_- i formelen ovenfor, får vi

$$t_- - t_{n+1} = \frac{KMt_-t_n - t_- + \frac{KM}{2}t_n^2 - t_n + \epsilon}{KMt_n - 1} =$$

$$= \frac{KMt_-t_n - \frac{KM}{2}t_-^2 - \epsilon + \frac{KM}{2}t_n^2 - t_n + \epsilon}{KMt_n - 1} = \frac{KM(t_- - t_n)^2}{2(1 - KMt_n)}$$

som er identiteten vi skulle vise.

Neste skritt er å kontrollere faktoren $1 - KMt_n$ i nevneren. Når $\epsilon = \frac{1}{2KM}$ (som i lemmaet ovenfor), er $t_n = \frac{1}{KM}\left(1 - \frac{1}{2^n}\right)$, og følgelig $1 - KMt_n = 2^{-n}$ i dette tilfellet. Vi skal snart vise at dette er den minste verdien $1 - KMt_n$ kan ha (vi har altså $1 - KMt_n \geq 2^n$), og dermed får vi

$$t_- - t_{n+1} = \frac{KM(t_- - t_n)^2}{2(1 - KMt_n)} \leq KM2^{n-1}(t - t_n)^2$$

Det er lett å vise ved induksjon at dersom $\{r_n\}$ er en tallfølge slik at

$$r_{n+1} \leq c2^{n-1}r_n^2$$

for alle $n \geq 0$, så er $r_n \leq \frac{(cr_0)^{2^n}}{c2^n}$. Bruker vi dette på tallfølgen $r_n = t_- - t_n$ (og husker at $r_0 = t_- - 0 = \frac{1-\sqrt{1-2h}}{KM}$), får vi

$$t_-t_n \leq \frac{\left(KM\frac{1-\sqrt{1-2h}}{KM}\right)^{2^n}}{KM2^n} = \frac{1}{KM}\frac{\left(1 - \sqrt{1-2h}\right)^{2^n}}{2^n}$$

som er formelen vi skulle frem til.

Det gjenstår å vise at vi virkelig har ulikheten $1 - KMt_n \geq 2^{-n}$ for alle valg av ϵ (og ikke bare for $\epsilon = \frac{1}{2KM}$). Vi tenker oss derfor at vi bruker Newtons metode med $t_0 = 0$ for forskjellige valg av $\epsilon \in [0, \frac{1}{2KM}]$, og at vi lar $t_n(\epsilon)$ være det n-te punktet i iterasjonen. Ser vi grafisk på hvordan Newtons metode fungerer, virker det naturlig at $t_n(\epsilon)$ er voksende som funksjon av ϵ. Kan vi bevise dette, er vi ferdig fordi vi da har $1 - KMt_n(\epsilon) \geq 1 - KMt_n(\frac{1}{2KM}) = 2^{-n}$.

Siden $t_0(\epsilon) = 0, t_1(\epsilon) = \epsilon$, ser vi at både $t_0(\epsilon), t_1(\epsilon)$ og $t_1(\epsilon) - t_0(\epsilon)$ er voksende i ϵ. Ifølge formel (5.5.1) i lemma 5.5.8 er

$$t_{n+1} - t_n = \frac{KM(t_n - t_{n-1})}{2(1 - KMt_n)}$$

Vi ser at hvis både $t_n(\epsilon)$ og $t_n(\epsilon) - t_{n-1}(\epsilon)$ er voksende i ϵ, så forteller denne likheten oss at $t_{n+1}(\epsilon) - t_n(\epsilon)$ er voksende i ϵ, noe som igjen medfører at $t_{n+1}(\epsilon)$ er voksende. Ved induksjon må da $t_n(\epsilon)$ være voksende i ϵ for alle n, og setningen er bevist. □

5.6 Omvendte og implisitte funksjoner

Det første temaet vi skal ta opp i denne seksjonen, er omvendte (inverse) funksjoner. Fra teorien om funksjoner av én variabel husker du sikkert at hvis en funksjon f tar oss fra en x-verdi til en y-verdi, så bringer den omvendte funksjonen f^{-1} oss tilbake fra y-verdien til x-verdien — er $y = f(x)$, så er altså $x = f^{-1}(y)$. Du husker sikkert også at for å få teorien til å fungere ordentlig, må vi anta at funksjonen f er *injektiv*, dvs. at det til hver y finnes høyst én x slik at $y = f(x)$. I noen tilfeller må vi sørge for at funksjonen blir injektiv ved å redusere definisjonsområdet; dette var trikset vi brukte for å få definert de omvendte trigonometriske funksjonene arcsin, arccos og arctan.

Teorien for omvendte funksjoner av én variabel er vanskelig nok, men den har én stor fordel; siden kontinuerlige, injektive funksjoner på et intervall er strengt monotone, er det som regel lett å avgjøre om en funksjon er injektiv eller ikke. Det er også lett å se hvor mye vi må innskrenke definisjonsområdet for å gjøre en ikke-injektiv funksjon injektiv. I høyere dimensjoner er det flere geometriske muligheter, og det er slett ikke lett å få oversikt over når en funksjon er injektiv. Heldigvis finnes det et teorem som sier at dersom Jacobi-matrisen til $\mathbf{F} : \mathbb{R}^m \to \mathbb{R}^m$ er inverterbar i punktet $\bar{\mathbf{x}}$, så er $\mathbf{F}$ injektiv når vi innskrenker den til en (tilstrekkelig liten) omegn rundt $\bar{\mathbf{x}}$, og den har en invers funksjon $\mathbf{G}$ som er definert i en omegn rundt punktet $\bar{\mathbf{y}} = \mathbf{F}(\bar{\mathbf{x}})$. Vi kan også regne ut den deriverte til den omvendte funksjonen $\mathbf{G}$ dersom vi kjenner den deriverte til $\mathbf{F}$; akkurat som i det endimensjonale tilfellet har vi

$$\mathbf{G}'(\bar{\mathbf{y}}) = \mathbf{F}'(\bar{\mathbf{x}})^{-1}$$

Alt dette kalles *omvendt funksjonsteorem*, og dette teoremet er vårt første mål i denne seksjonen.

La oss begynne med noen definisjoner. Når vi arbeider med inverse funksjoner, er det viktigere enn ellers å holde styr på definisjonsmengder og verdimengder. Vi lar $\mathrm{D}_{\mathbf{F}}$ betegne *definisjonsmengden* til $\mathbf{F}$ (dvs. de $\mathbf{x}$ som $\mathbf{F}(\mathbf{x})$ er definert for), og vi lar

$$\mathrm{V}_{\mathbf{F}} = \{\mathbf{F}(\mathbf{x}) \,:\, \mathbf{x} \in \mathrm{D}_{\mathbf{F}}\}$$

være *verdimengden* til $\mathbf{F}$.

Definisjon 5.6.1 *Funksjonen* $\mathbf{F} : \mathrm{D}_{\mathbf{F}} \to \mathrm{V}_{\mathbf{F}}$ *kalles* injektiv *dersom det til hver* $\mathbf{y} \in \mathrm{V}_{\mathbf{F}}$ *bare finnes én* $\mathbf{x} \in \mathrm{D}_{\mathbf{F}}$ *slik at* $\mathbf{y} = \mathbf{F}(\mathbf{x})$. *I så fall er den* omvendte funksjonen $\mathbf{G} : \mathrm{V}_{\mathbf{F}} \to \mathrm{D}_{\mathbf{F}}$ *definert ved*

$$\mathbf{G}(\mathbf{y}) = \mathbf{x} \quad \textit{dersom} \quad \mathbf{F}(\mathbf{x}) = \mathbf{y}$$

Den omvendte funksjonen $\mathbf{G}$ *betegnes ofte med* $\mathbf{F}^{-1}$.

Anta at vi har en funksjon $\mathbf{F} : U \to \mathbb{R}^m$ og at U_0 er en delmengde av U. Med *restriksjonen* av $\mathbf{F}$ til U_0 mener vi den funksjonen vi får når vi innskrenker definisjonsområdet til $\mathbf{F}$ til å være U_0 (men ellers ikke gjør noen endringer). Vi sier at U_0 er en *omegn* om punktet $\mathbf{x}$ dersom $\mathbf{x}$ er et indre punkt i U_0, dvs. dersom U_0 inneholder en åpen kule med sentrum i $\mathbf{x}$. Vi kan nå gi en presis formulering av resultatet vi er på jakt etter:

Teorem 5.6.2 (Omvendt funksjonsteorem) *Anta at U er en åpen mengde i $\mathbb{R}^m$ og at $\mathbf{F} : U \to \mathbb{R}^m$ har kontinuerlige partiellderiverte. Anta at $\bar{\mathbf{x}} \in U$ og at Jacobi-matrisen $\mathbf{F}'(\bar{\mathbf{x}})$ er inverterbar. Da finnes det en omegn $U_0 \subset U$ om $\bar{\mathbf{x}}$ slik at $\mathbf{F}$ restriktert til U_0 er injektiv. Verdimengden V til denne restriksjonen er en omegn om $\bar{\mathbf{y}} = \mathbf{F}(\bar{\mathbf{x}})$, og den omvendte funksjonen $\mathbf{G} : V \to U_0$ er deriverbar i $\bar{\mathbf{y}}$ med Jacobi-matrise*

$$\mathbf{G}'(\bar{\mathbf{y}}) = \mathbf{F}'(\bar{\mathbf{x}})^{-1}$$

Beviset for omvendt funksjonsteorem er ganske vanskelig, og vi skal utsette det til slutten av seksjonen. Det vanskeligste punktet er å vise at det i det hele tatt finnes en omvendt funksjon. Legg merke til at dette er det samme som å vise at ligningen $\mathbf{y} = \mathbf{F}(\mathbf{x})$ (med $\mathbf{x}$ som ukjent) har en løsning når $\mathbf{y}$ er i nærheten av $\bar{\mathbf{y}}$. Vi skal bruke Banachs fikspunktteorem til å vise at dette alltid er tilfellet.

Et annet tidkrevende punkt i beviset er å vise at den omvendte funksjonen $\mathbf{G}$ er deriverbar i punktet $\bar{\mathbf{y}}$. Når dette er vist, er det imidlertid ikke vanskelig å finne ut hva den deriverte er. Siden $\mathbf{F}$ og $\mathbf{G}$ er omvendte funksjoner, har vi nemlig

$$\mathbf{G}(\mathbf{F}(\mathbf{x})) = \mathbf{x}$$

for alle $\mathbf{x}$ i nærheten av $\bar{\mathbf{x}}$. Deriverer vi venstresiden av dette uttrykket, får vi ved kjerneregelen

$$\mathbf{G}'(\mathbf{F}(\mathbf{x}))\mathbf{F}'(\mathbf{x})$$

mens den deriverte av høyresiden er identitetsmatrisen I_m (hvorfor det?). Dermed er

$$\mathbf{G}'(\mathbf{F}(\mathbf{x}))\mathbf{F}'(\mathbf{x}) = I_m$$

og følgelig er

$$\mathbf{G}'(\mathbf{F}(\mathbf{x})) = \mathbf{F}'(\mathbf{x})^{-1}$$

Setter vi inn $\mathbf{x} = \bar{\mathbf{x}}$ og bruker at $\mathbf{F}(\bar{\mathbf{x}}) = \bar{\mathbf{y}}$, får vi formelen i teoremet.

La oss se på et eksempel.

Eksempel 1: Vi skal vise at funksjonen

$$\mathbf{F}(x_1, x_2) = \begin{pmatrix} e^{x_1} + x_2 \\ x_2 \cos x_1 \end{pmatrix}$$

er injektiv når den restrikteres til en passende omegn om $(0,0)$, og at denne restriksjonen har en omvendt funksjon $\mathbf{G}$ definert i en omegn om $(1,0)$. Vi skal også finne de partiellderiverte til $\mathbf{G}$ i punktet $(1,0)$.

Vi observerer først at

$$\mathbf{F}(0,0) = \begin{pmatrix} e^0 + 0 \\ 0\cos 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

Videre er

$$\mathbf{F}'(x_1,x_2) = \begin{pmatrix} e^{x_1} & 1 \\ -x_2\sin x_1 & \cos x_1 \end{pmatrix}$$

som gir

$$\mathbf{F}'(0,0) = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

Siden $\det(\mathbf{F}'(0,0)) = 1$, er $\mathbf{F}'(0,0)$ inverterbar, og følgelig er $\mathbf{F}$ injektiv i en omegn om $(0,0)$ og har en omvendt funksjon $\mathbf{G}$ definert i en omegn om $\mathbf{F}(0,0) = (1,0)$. Siden

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}$$

er

$$\mathbf{G}'(1,0) = \mathbf{F}(0,0)^{-1} = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}$$

Det betyr at $\frac{\partial G_1}{\partial y_1}(1,0) = 1$, $\frac{\partial G_1}{\partial y_2}(1,0) = -1$, $\frac{\partial G_2}{\partial y_1}(1,0) = 0$ og $\frac{\partial G_2}{\partial y_2}(1,0) = 1$. ♣

Omvendt funksjonsteorem brukes mye både i teori og anvendelser. I anvendelser har man ofte behov for å bytte om på hva som er uavhengige og avhengige størrelser — istedenfor å tenke på etterspørselen som en funksjon av prisene, har man plutselig lyst til å tenke på prisene som funksjon av etterspørselen. Omvendt funksjonsteorem (og særlig formelen med de deriverte) gjør det enkelt å bytte synsvinkel på denne måten. I teoretisk arbeid er det ofte betingelsene som er viktige — vi trenger garantier for at de funksjonene vi skal arbeide med, virkelig finnes og har de egenskapen vi ønsker oss.

Vi skal nå se på en viktig konsekvens av omvendt funksjonsteorem, nemlig det som kalles *implisitt funksjonsteorem*. Det kan være lurt å ta utgangspunkt i et konkret eksempel. Vi vet at kuleflaten med radius 1 er beskrevet av ligningen

$$x^2 + y^2 + z^2 = 1$$

Denne beskrivelsen av kuleflaten er nyttig for en del formål, men for andre hadde det vært mer effektivt å tenke på kuleflaten som en funksjonsgraf. Dette er ikke mulig hvis vi vil ha med både øvre og nedre halvkule, men

dersom vi er fornøyd med (for eksempel) øvre halvkule, kan vi løse ligningen ovenfor for z og få

$$z = \sqrt{1 - x^2 - y^2}$$

Dette viser at i hvert fall lokalt (vi innskrenker oss til øvre halvkule) vil ligningen $x^2 + y^2 + z^2 = 1$ definere z som en funksjon $z = g(x, y) = \sqrt{1 - x^2 - y^2}$ av x og y. Dette eksemplet kan vi generalisere. Anta at vi har en funksjon $f(x_1, x_2, \ldots, x_m, y)$ av $m+1$-variable, og at vi er interessert i mengden av punkter $(x_1, x_2, \ldots, x_m, y)$ som tilfredsstiller ligningen

$$f(x_1, x_2, \ldots, x_m, y) = 0$$

(det er lurt å tenke på dette som en generalisert flate). Akkurat som ovenfor kan vi tenke oss at vi løser denne ligningen for y og får et uttrykk

$$y = g(x_1, x_2, \ldots, x_m)$$

Da har vi beskrevet den generaliserte flaten vår som grafen til en funksjon g med m variable. Vi kaller en slik funksjon g en *implisitt gitt* (eller bare *implisitt*) funksjon.

Det er flere grunner til at denne planen kanskje ikke lar seg gjennomføre. Dersom f er funksjonen

$$f(x_1, x_2, y) = x_1^2 + x_2^2 + y^2 + 1$$

så har for eksempel ikke ligningen $f(x_1, x_2, y) = 0$ løsninger i det hele tatt. Et annet problem er at ligningen kan ha flere løsninger. I halvkuleeksemplet ovenfor har vi to løsninger

$$z = \pm\sqrt{1 - x^2 - y^2}$$

Disse er greie å skille mellom dersom vi vet at vi vil være på enten øvre eller nedre halvkule, men det finnes andre eksempler der det er vanskeligere å skjøte ting sammen. Et tredje problem er at det kan være vanskelig å løse ligningen $f(x_1, x_2, \ldots, x_m, y) = 0$ for y selv i de tilfellene der det finnes en løsning. Det er i den siste situasjonen vi virkelig får bruk for teoremet nedenfor. Det forteller oss når ligningen har en løsning, og gir oss viktig informasjon om løsningsfunksjonen i de tilfellene vi ikke greier å regne den ut.

Teorem 5.6.3 (Implisitt funksjonsteorem) *Anta at U er en åpen delmengde av $\mathbb{R}^{m+1}$ og la $f : U \to \mathbb{R}$ være en funksjon med kontinuerlige partiellderiverte. Anta at $(\bar{\mathbf{x}}, \bar{y}) = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n, \bar{y})$ er et punkt i U der $f(\bar{\mathbf{x}}, \bar{y}) = 0$. Anta videre at $\frac{\partial f}{\partial y}(\bar{\mathbf{x}}, \bar{y}) \neq 0$. Da finnes det en omegn U_0 om $\bar{\mathbf{x}}$, og en deriverbar funksjon $g : U_0 \to \mathbb{R}$ slik at $\mathbf{g}(\bar{\mathbf{x}}) = \bar{y}$ og*

$$f(\mathbf{x}, g(\mathbf{x})) = 0$$

for alle $\mathbf{x} \in U_0$*. Den deriverte til* g *er gitt ved*

$$\frac{\partial g}{\partial x_i}(\bar{\mathbf{x}}) = -\frac{\frac{\partial f}{\partial x_i}(\bar{\mathbf{x}}, \bar{y})}{\frac{\partial f}{\partial y}(\bar{\mathbf{x}}, \bar{y})}$$

Bevis: La oss begynne med det som kanskje ser vanskeligst ut, men som faktisk er lettest, nemlig formelen for $\frac{\partial g}{\partial x_i}(\bar{\mathbf{x}})$. Anta derfor at vi har greid å vise at det finnes en deriverbar funksjon g slik at

$$f(\mathbf{x}, g(\mathbf{x})) = 0$$

i en omegn om $\bar{\mathbf{x}}$. Partiellderiverer vi begge sider mhp. x_i, får vi ifølge kjerneregelen

$$\frac{\partial f}{\partial x_i}(\mathbf{x}, g(\mathbf{x})) + \frac{\partial f}{\partial y}(\mathbf{x}, g(\mathbf{x}))\frac{\partial g}{\partial x_i}(\mathbf{x}) = 0$$

Setter vi inn $\mathbf{x} = \bar{\mathbf{x}}$, $\bar{y} = g(\bar{\mathbf{x}})$ og bruker at $\frac{\partial f}{\partial y}(\bar{\mathbf{x}}, \bar{y}) \neq 0$, får vi

$$\frac{\partial g}{\partial x_i}(\bar{\mathbf{x}}) = -\frac{\frac{\partial f}{\partial x_i}(\bar{\mathbf{x}}, \bar{y})}{\frac{\partial f}{\partial y}(\bar{\mathbf{x}}, \bar{y})}$$

som er den formelen vi skulle vise.

For å vise at funksjonen g virkelig finnes og er deriverbar, bruker vi et triks — vi anvender omvendt funksjonsteorem på funksjonen $\mathbf{F} : \mathbb{R}^{m+1} \to \mathbb{R}^{m+1}$ gitt ved

$$\mathbf{F}(x_1, x_2, \ldots, x_m, y) = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \\ f(x_1, x_2, \ldots, x_m, y) \end{pmatrix}$$

Denne funksjonen har Jacobi-matrise

$$\mathbf{F}'(x_1, x_2, \ldots, x_m, y) = \begin{pmatrix} 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \ldots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & 0 \\ \frac{\partial f}{\partial x_1}(\mathbf{x}, y) & \frac{\partial f}{\partial x_2}(\mathbf{x}, y) & \ldots & \frac{\partial f}{\partial x_m}(\mathbf{x}, y) & \frac{\partial f}{\partial y}(\mathbf{x}, y) \end{pmatrix}$$

Regner vi ut determinanten til Jacobi-matrisen, får vi

$$\det(\mathbf{F}'(x_1, x_2, \ldots, x_m, y)) = \frac{\partial f}{\partial y}(\mathbf{x}, y)$$

Setter vi inn $(\bar{\mathbf{x}}, \bar{y})$, ser vi at

$$\det(\mathbf{F}'(\bar{\mathbf{x}}, \bar{y})) = \frac{\partial f}{\partial y}(\bar{\mathbf{x}}, \bar{y}) \neq 0$$

ifølge antagelsen i teoremet. Jacobi-matrisen $\mathbf{F}'(\bar{\mathbf{x}}, \bar{y})$ er dermed inverterbar, og vi kan bruke omvendt funksjonsteorem på $\mathbf{F}$: Restrikterer vi $\mathbf{F}$ til en tilstrekkelig liten omegn om $(\bar{\mathbf{x}}, \bar{y})$, har den en omvendt funksjon $\mathbf{G}$ definert på en åpen mengde $V \subset \mathbb{R}^{m+1}$ som inneholder $\mathbf{F}(\bar{\mathbf{x}}, \bar{y}) = (\bar{\mathbf{x}}, f(\bar{\mathbf{x}}, \bar{y})) = (\bar{\mathbf{x}}, 0)$. På grunn av den spesielle formen til $\mathbf{F}$, er det lett å se at $\mathbf{G}$ må ha formen

$$\mathbf{G}(x_1, x_2 \ldots, x_m, z) = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \\ h(x_1, x_2, \ldots x_m, z) \end{pmatrix}$$

der h er en deriverbar funksjon. Siden G er den omvendte funksjonen til $\mathbf{F}$, har vi videre

$$\mathbf{F}(x_1, x_2, \ldots, x_m, h(x_1, x_2, \ldots, x_m, z)) = \mathbf{F}(\mathbf{G}(x_1, x_2, \ldots, x_m, z)) = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \\ z \end{pmatrix}$$

På den annen side vet vi fra definisjonen av $\mathbf{F}$ at

$$\mathbf{F}(x_1, x_2, \ldots, x_m, h(x_1, x_2, \ldots, x_m, z)) =$$
$$= \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \\ f(x_1, x_2, \ldots, x_m, h(x_1, x_2, \ldots, x_m, z)) \end{pmatrix}$$

Sammenligner vi disse to formlene, ser vi at

$$z = f(x_1, x_2, \ldots, x_m, h(x_1, x_2, \ldots, x_m, z))$$

Setter vi inn $z = 0$, har vi dermed

$$f(x_1, x_2, \ldots, x_m, h(x_1, x_2, \ldots, x_m, 0)) = 0$$

Vi lar nå $U_0 = \{(x_1, x_2, \ldots, x_m) \mid (x_1, x_2, \ldots, x_m, 0) \in V\}$, og definerer $g : U_0 \to \mathbb{R}$ ved $g(x_1, x_2, \ldots, x_m) = h(x_1, x_2, \ldots, x_m, 0)$. Det følger fra utledningene ovenfor at $g(\bar{\mathbf{x}}) = \bar{y}$, at $f(\mathbf{x}, g(\mathbf{x})) = 0$ for alle $\mathbf{x} \in U_0$, og at g er deriverbar i $\mathbf{y}$ med

$$\frac{\partial g}{\partial x_i}(\bar{\mathbf{x}}) = -\frac{\frac{\partial f}{\partial x_i}(\bar{\mathbf{x}}, \bar{y})}{\frac{\partial f}{\partial y}(\bar{\mathbf{x}}, \bar{y})}$$

Det gjenstår én liten detalj. Teoremet påstår at g er deriverbar i hele omegnen U_0, men vi har bare vist deriverbarhet i punktet $\bar{\mathbf{x}}$. Dette er lett å fikse med et lite triks. Vi bruker argumentene ovenfor på nytt, men erstatter punktet $(\bar{\mathbf{x}}, \bar{y})$ med et fritt valgt punkt $(\mathbf{x}, g(\mathbf{x}))$ der $\mathbf{x} \in U_0$ (det kan hende vi må innsnevre U_0 noe for å være sikre på at $\frac{\partial f}{\partial y}(\mathbf{x}, g(\mathbf{x})) \neq 0$). □

Eksempel 2: La

$$f(x, y) = e^{x+y} + y - 1$$

Vi skal vise at det finnes en funksjon $g(x)$ definert i en omegn om 0 slik at $g(0) = 0$ og $f(x, g(x)) = 0$. Vi skal også regne ut $g'(0)$.

Vi observerer først at $f(0, 0) = e^{0+0} + 0 - 1 = 0$. Videre er $\frac{\partial f}{\partial y}(x, y) = e^{x+y} + 1$, så $\frac{\partial f}{\partial y}(0, 0) = e^{0+0} + 1 = 2 \neq 0$. Dette betyr at det finnes en (implisitt definert) funksjon g slik at $g(0) = 0$ og $f(x, g(x)) = 0$ i en omegn om 0. Den deriverte til g er gitt ved

$$g'(0) = -\frac{\frac{\partial f}{\partial x}(0, 0)}{\frac{\partial f}{\partial y}(0, 0)} = -\frac{e^{0+0}}{e^{0+0} + 1} = -\frac{1}{2}$$

♣

Ofte kan det være lønnsomt å bruke implisitt derivasjon også i situasjoner der vi faktisk *kan* finne et uttrykk for den implisitt gitte funksjonen:

Eksempel 3: Vi skal finne stigningstallene i x- og y-retning i et punkt (x, y, z) på kuleflaten

$$x^2 + y^2 + z^2 = R^2$$

Istedenfor å referere eksplisitt til teoremet ovenfor, skal vi bruke implisitt derivasjon slik det som regel gjøres i praksis. Vi tenker på z som en funksjon av x og y og deriverer ligningen ved hjelp av kjerneregelen. Deriverer vi mhp. x, får vi

$$2x + 2z\frac{\partial z}{\partial x} = 0$$

og deriverer vi mhp. y, får vi

$$2y + 2z\frac{\partial z}{\partial y} = 0$$

Dette gir $\frac{\partial z}{\partial x} = -\frac{x}{z}$ og $\frac{\partial z}{\partial y} = -\frac{y}{z}$ (forutsatt at $z \neq 0$). Du kan selv bruke implisitt funksjonsteorem til å rettferdiggjøre disse regningene. ♣

Vi tar også med et eksempel som viser hvordan implisitt derivasjon dukker opp i andre fag.

Eksempel 4: Når man studerer gasser, er det tre naturlige variable; trykket p, temperaturen T og volumet V. I gassmodeller er de knyttet sammen gjennom en ligning $f(p, V, T) = 0$. Funksjonen f varierer fra en modell til en annen — i teorien for såkalte *ideelle gasser* er f.eks. $f(p, v, T) = pV - kT$ der k er en konstant. Siden variablene er knyttet sammen gjennom en ligning, tenker man ofte på hver av disse størrelsene p, V og T som en funksjon av de to andre. Tenker vi f.eks. på trykket som en funksjon av volumet og temperaturen, har vi en funksjon $p(V, T)$ som oppfyller ligningen

$$f(p(V, T), V, T) = 0$$

altså en implisitt gitt funksjon. Vi kan finne uttrykk for de partiellderiverte til p ved å derivere implisitt. Deriverer vi først mhp. V, får vi

$$\frac{\partial f}{\partial p}\frac{\partial p}{\partial V} + \frac{\partial f}{\partial V} = 0,$$

mens derivasjon mhp. T gir

$$\frac{\partial f}{\partial p}\frac{\partial p}{\partial T} + \frac{\partial f}{\partial T} = 0,$$

Løser vi for de partiellderiverte til p, ser vi at

$$\frac{\partial p}{\partial V} = -\frac{\frac{\partial f}{\partial V}}{\frac{\partial f}{\partial p}} \qquad \text{og} \qquad \frac{\partial p}{\partial T} = -\frac{\frac{\partial f}{\partial T}}{\frac{\partial f}{\partial p}}$$

Regninger av denne typen (gjerne litt mer innfløkt enn disse) spiller en sentral rolle i mange anvendelser. Som oftest argumenterer man uformelt slik som her; man deriverer ved bruk av kjerneregelen uten å bry seg for mye om deriverbarhet og andre matematiske finurligheter. Ønsker man bedre matematiske begrunnelser, er implisitt funksjonsteorem det riktige redskapet. ♣

Det finnes også en vektorvaluert versjon av implisitt funksjonsteorem. Her tenker vi oss at vi skal finne k variable $y_1, y_2, \ldots, y_k$ uttrykt ved m andre variable $x_1, x_2, \ldots, x_m$. Siden vi har k ukjente, trenger vi også k ligninger, så vi starter med en funksjon $\mathbf{F} : \mathbb{R}^{m+k} \to \mathbb{R}^k$ og ønsker å løse ligningssystemet

$$\mathbf{F}(x_1, \ldots, x_m, y_1, \ldots, y_k) = \mathbf{0}$$

for $y_1, \ldots, y_k$ (legg merke til at ligningssystemet har k ligninger og k ukjente). Sagt med andre ord ønsker vi å finne funksjoner $g_1(x_1, \ldots, x_m)$, $\ldots$, $g_k(x_1, \ldots, x_k)$ slik at

$$\mathbf{F}(x_1, \ldots, x_m, g_1(x_1, \ldots, x_m), \ldots, g_k(x_1, \ldots, x_m)) = \mathbf{0}$$

Disse uttrykkene blir enklere om vi bruker vektornotasjon. Skriver vi $\mathbf{x} = (x_1, \ldots, x_m)$, $\mathbf{y} = (y_1, \ldots, y_k)$ og $\mathbf{G}(\mathbf{x}) = (g_1(\mathbf{x}), \ldots, g_k(\mathbf{x}))$, kan formelen ovenfor skrives

$$\mathbf{F}(\mathbf{x}, \mathbf{G}(\mathbf{x})) = \mathbf{0}$$

Før vi presenterer den vektorvaluerte versjonen av implisitt funksjonsteorem, trenger vi litt notasjon. Anta at vi har en funksjon $\mathbf{F} : \mathbb{R}^{m+k} \to \mathbb{R}^k$ som ovenfor, og at vi har delt inn variablene i to grupper $\mathbf{x} = (x_1, \ldots, x_m)$ og $\mathbf{y} = (y_1, \ldots, y_k)$. Vi lar $\frac{\partial \mathbf{F}}{\partial \mathbf{x}}$ være den Jacobi-matrisen vi får når vi bare tenker på $\mathbf{F}$ som en funksjon av x-variablene, dvs.

$$\frac{\partial \mathbf{F}}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial F_1}{\partial x_1} & \frac{\partial F_1}{\partial x_2} & \cdots & \frac{\partial F_1}{\partial x_m} \\ \frac{\partial F_2}{\partial x_1} & \frac{\partial F_2}{\partial x_2} & \cdots & \frac{\partial F_2}{\partial x_m} \\ \vdots & \vdots & \ldots & \vdots \\ \frac{\partial F_k}{\partial x_1} & \frac{\partial F_k}{\partial x_2} & \cdots & \frac{\partial F_k}{\partial x_m} \end{pmatrix}$$

Tilsvarende lar vi $\frac{\partial \mathbf{F}}{\partial \mathbf{y}}$ være den Jacobi-matrisen vi får, når vi bare tenker på $\mathbf{F}$ som en funksjon av y-variablene, dvs.

$$\frac{\partial \mathbf{F}}{\partial \mathbf{y}} = \begin{pmatrix} \frac{\partial F_1}{\partial y_1} & \frac{\partial F_1}{\partial y_2} & \cdots & \frac{\partial F_1}{\partial y_k} \\ \frac{\partial F_2}{\partial y_1} & \frac{\partial F_2}{\partial y_2} & \cdots & \frac{\partial F_2}{\partial y_k} \\ \vdots & \vdots & \ldots & \vdots \\ \frac{\partial F_k}{\partial y_1} & \frac{\partial F_k}{\partial y_2} & \cdots & \frac{\partial F_k}{\partial y_k} \end{pmatrix}$$

Teorem 5.6.4 (Vektorvaluert versjon av implisitt funksjonsteorem) *Anta at U er en åpen delmengde av $\mathbb{R}^{m+k}$ og la $\mathbf{F} : U \to \mathbb{R}^k$ være en funksjon med kontinuerlige partiellderiverte. Anta at $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ er et punkt i U der $\mathbf{F}(\bar{\mathbf{x}}, \bar{\mathbf{y}}) = \mathbf{0}$. Anta videre at $k \times k$-matrisen $\frac{\partial \mathbf{F}}{\partial \mathbf{y}}(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ er inverterbar. Da finnes det en omegn U_0 om $\bar{\mathbf{x}}$, og en deriverbar funksjon $\mathbf{G} : U_0 \to \mathbb{R}^k$ slik at $\mathbf{g}(\bar{\mathbf{x}}) = \bar{\mathbf{y}}$ og*

$$\mathbf{F}(\mathbf{x}, \mathbf{G}(\mathbf{x})) = \mathbf{0}$$

for alle $\mathbf{x} \in U_0$. Jacobi-matrisen til $\mathbf{G}$ er gitt ved

$$\mathbf{G}'(\bar{\mathbf{x}}) = -\left(\frac{\partial \mathbf{F}}{\partial \mathbf{y}}(\bar{\mathbf{x}}, \bar{\mathbf{y}})\right)^{-1} \left(\frac{\partial \mathbf{F}}{\partial \mathbf{x}}(\bar{\mathbf{x}}, \bar{\mathbf{y}})\right)$$

Vi overlater beviset til leserne. Ideen er akkurat den samme som i det skalarvaluerte tilfellet, men det er litt flere partiellderiverte å holde styr på.

*Bevis for omvendt funksjonsteorem

Vi skal dele beviset opp i lemmaer, og det er ikke så lett å se hvordan alt passer sammen før du har vært gjennom hele resonnementet. Det første lemmaet er enkelt, men inneholder det grunnleggende eksistensresultatet (det som sikrer at en omvendt funksjon finnes). Utgangspunktet for lemmaet er en selvfølgelighet, nemlig at identitetsavbildningen $I(\mathbf{x}) = \mathbf{x}$ avbilder kulen $\overline{B}(\mathbf{x}, r)$ injektivt på $\overline{B}(\mathbf{x}, r)$. Lemmaet viser at en avbildning som ikke avviker for mye fra I, har en lignende egenskap. Husk at

$$\overline{B}(\mathbf{a}, r) = \{\mathbf{x} \in \mathbb{R}^m \,:\, |\mathbf{x} - \mathbf{a}| \leq r\}$$

betegner en *lukket* kule i $\mathbb{R}^m$.

Lemma 5.6.5 (Perturbasjonslemma) *La $\overline{B}(\mathbf{0}, r)$ være en lukket kule i $\mathbb{R}^m$, og anta at funksjonen $\mathbf{H} : \overline{B}(\mathbf{0}, r) \to \mathbb{R}^m$ er slik at $\mathbf{H}(\mathbf{0}) = \mathbf{0}$ og*

$$|\mathbf{H}(\mathbf{u}) - \mathbf{H}(\mathbf{v})| \leq \frac{1}{2}|\mathbf{u} - \mathbf{v}| \qquad \text{for alle } \mathbf{u}, \mathbf{v} \in \overline{B}(\mathbf{0}, r)$$

Da er funksjonen $\mathbf{L} : \overline{B}(\mathbf{0}, r) \to \mathbb{R}^m$ definert ved $\mathbf{L}(\mathbf{x}) = \mathbf{x} + \mathbf{H}(\mathbf{x})$ injektiv, og kulen $\overline{B}(\mathbf{0}, \frac{r}{2})$ er inneholdt i verdimengden til $\mathbf{L}$.

Bevis: La oss først vise at $\mathbf{L}$ er injektiv. Vi antar at $\mathbf{L}(\mathbf{x}) = \mathbf{L}(\mathbf{y})$, og må vise at $\mathbf{x} = \mathbf{y}$. Per definisjon av $\mathbf{L}$ er

$$\mathbf{x} + \mathbf{H}(\mathbf{x}) = \mathbf{y} + \mathbf{H}(\mathbf{y})$$

dvs.

$$\mathbf{x} - \mathbf{y} = \mathbf{H}(\mathbf{y}) - \mathbf{H}(\mathbf{x})$$

som gir

$$|\mathbf{x} - \mathbf{y}| = |\mathbf{H}(\mathbf{x}) - \mathbf{H}(\mathbf{y})|$$

Ifølge antagelsene er $|\mathbf{H}(\mathbf{x}) - \mathbf{H}(\mathbf{y})| \leq \frac{1}{2}|\mathbf{x} - \mathbf{y}|$, så likheten ovenfor er bare mulig hvis $|\mathbf{x} - \mathbf{y}| = 0$, dvs. hvis $\mathbf{x} = \mathbf{y}$.

Det gjenstår å vise at $\overline{B}(\mathbf{0}, \frac{r}{2})$ er inneholdt i verdimengden til $\mathbf{L}$. Vi må da vise at for alle $\mathbf{y} \in \overline{B}(\mathbf{0}, \frac{r}{2})$, har ligningen $\mathbf{L}(\mathbf{x}) = \mathbf{y}$ en løsning i $\overline{B}(\mathbf{0}, r)$. Denne ligningen kan skrives

$$\mathbf{x} = \mathbf{y} - \mathbf{H}(\mathbf{x}),$$

så det er nok å vise at funksjonen $\mathbf{K}(\mathbf{x}) = \mathbf{y} - \mathbf{H}(\mathbf{x})$ har et fikspunkt i $\overline{B}(\mathbf{0}, r)$. Dette vil følge av Banachs fikspunktteorem dersom vi kan vise at $\mathbf{K}$ er en kontraksjon av $\overline{B}(\mathbf{0}, r)$. La oss først vise at $\mathbf{K}$ avbilder $\overline{B}(\mathbf{0}, r)$ inn i $\overline{B}(\mathbf{0}, r)$. Det følger av at

$$|\mathbf{K}(\mathbf{x})| = |\mathbf{y} - \mathbf{H}(\mathbf{x})| \leq |\mathbf{y}| + |\mathbf{H}(\mathbf{x})| \leq \frac{r}{2} + \frac{r}{2} = r$$

der vi har brukt at ifølge betingelsene på $\mathbf{H}$ er

$$|\mathbf{H}(\mathbf{x})| = |\mathbf{H}(\mathbf{x}) - \mathbf{H}(\mathbf{0})| \leq \frac{1}{2}|\mathbf{x} - \mathbf{0}| \leq \frac{r}{2}$$

Til slutt sjekker vi kontraksjonsbetingelsen:

$$|\mathbf{K}(\mathbf{u}) - \mathbf{K}(\mathbf{v})| = |\mathbf{H}(\mathbf{u}) - \mathbf{H}(\mathbf{v})| \leq \frac{1}{2}|\mathbf{u} - \mathbf{v}|$$

Dermed har vi vist at $\mathbf{K}$ er en kontraksjon, og følgelig har den et entydig fikspunkt i $\overline{B}(\mathbf{0}, r)$. □

I det neste lemmaet skal vi vise at omvendt funksjonsteorem gjelder for funksjoner $\mathbf{L}$ slik at $\mathbf{L}(\mathbf{0}) = \mathbf{0}$ og $\mathbf{L}'(\mathbf{0}) = I_m$. Dette kan høres veldig spesielt ut, men det viser seg at det generelle tilfellet følger ved et enkelt variabelskifte.

Lemma 5.6.6 *Anta U er et område i $\mathbb{R}^m$ som inneholder $\mathbf{0}$ og at $\mathbf{L} : U \to \mathbb{R}^m$ er en funksjon med kontinuerlige partiellderiverte slik at $\mathbf{L}(\mathbf{0}) = \mathbf{0}$ og $\mathbf{L}'(\mathbf{0}) = I_m$. Da finnes det en $r > 0$ slik at $\mathbf{L}$ er injektiv når den restrikteres til $\overline{B}(\mathbf{0}, r)$ og har en omvendt funksjon $\mathbf{M}$ definert på et område som inneholder $\overline{B}(\mathbf{0}, \frac{r}{2})$. Den omvendte funksjonen $\mathbf{M}$ er deriverbar i $\mathbf{0}$ og har Jacobi-matrise $\mathbf{M}'(\mathbf{0}) = I_m$.*

Bevis: La $\mathbf{H}(\mathbf{x}) = \mathbf{L}(\mathbf{x}) - \mathbf{x}$. Vi skal først vise at $\mathbf{H}$ tilfredsstiller betingelsene i lemmaet over. Observer først at siden $\mathbf{L}(\mathbf{0}) = \mathbf{0}$ og $\mathbf{L}'(\mathbf{0}) = I_m$, så er $\mathbf{H}(\mathbf{0}) = \mathbf{0}$ og alle de partiellderiverte $\frac{\partial H_i}{\partial x_j}(\mathbf{0})$ er lik 0. Spesielt er $\nabla H_i(\mathbf{0}) = \mathbf{0}$ for alle i. Bruker vi middelverdisetningen for funksjoner av flere variable (setning 5.4.5) på H_i, får vi dermed

$$H_i(\mathbf{x}) - H_i(\mathbf{y}) = \nabla H_i(\mathbf{c}_i) \cdot (\mathbf{x} - \mathbf{y}) = (\nabla H_i(\mathbf{c}_i) - \nabla H_i(\mathbf{0})) \cdot (\mathbf{x} - \mathbf{y})$$

Siden de partiellderiverte til $\mathbf{H}$ er kontinuerlige, kan vi få $(\nabla H_i(\mathbf{c}_i) - \nabla H_i(\mathbf{0}))$ så liten vi måtte ønske ved å velge $\mathbf{x}$ og $\mathbf{y}$ tilstrekkelig nær $\mathbf{0}$. Spesielt finnes det en $r > 0$ slik at hvis $|\mathbf{x}|, |\mathbf{y}| \leq r$, så er $|\nabla H_i(\mathbf{c}_i) - \nabla H_i(\mathbf{0})| \leq \frac{1}{2\sqrt{m}}$ for alle i. Bruker vi Schwarz' ulikhet, ser vi at

$$|H_i(\mathbf{x}) - H_i(\mathbf{y})| \leq |\nabla H_i(\mathbf{c}_i) - \nabla H_i(\mathbf{0})|\,|\mathbf{x} - \mathbf{y}| \leq \frac{1}{2\sqrt{m}}|\mathbf{x} - \mathbf{y}|$$

og

$$|\mathbf{H}(\mathbf{x}) - \mathbf{H}(\mathbf{y})| = \sqrt{(H_1(\mathbf{x}) - H_1(\mathbf{y}))^2 + \ldots + (H_m(\mathbf{x}) - H_m(\mathbf{y}))^2} \leq$$

$$\leq \sqrt{m\left(\frac{1}{2\sqrt{m}}|\mathbf{x} - \mathbf{y}|\right)^2} = \frac{1}{2}|\mathbf{x} - \mathbf{y}|$$

Dermed har vi vist at $\mathbf{H}$ tilfredsstiller betingelsene i forrige lemma, og siden

$$\mathbf{L}(\mathbf{x}) = \mathbf{x} + \mathbf{H}(\mathbf{x}),$$

vet vi fra lemmaet at $\mathbf{L}$ restriktert til $\overline{B}(\mathbf{0}, r)$ er injektiv og at verdimengden inneholder $\overline{B}(\mathbf{0}, \frac{r}{2})$. Dette betyr at $\mathbf{L}$ (restriktert til $\overline{B}(\mathbf{0}, r)$) har en omvendt funksjon $\mathbf{M}$ som er definert på et område som omfatter $\overline{B}(\mathbf{0}, \frac{r}{2})$.

Det gjenstår å vise at $\mathbf{M}$ er deriverbar i $\mathbf{0}$ med Jacobi-matrise I_m, men før vi går løs på deriverbarheten, trenger vi et lite estimat. Ifølge trekantulikheten har vi

$$|\mathbf{x}| = |\mathbf{L}(\mathbf{x}) - \mathbf{H}(\mathbf{x})| \leq |\mathbf{L}(\mathbf{x})| + |\mathbf{H}(\mathbf{x})| \leq |\mathbf{L}(\mathbf{x})| + \frac{1}{2}|\mathbf{x}|$$

som gir

$$\frac{1}{2}|\mathbf{x}| \leq |\mathbf{L}(\mathbf{x})|$$

når vi flytter over.

Vi er nå klar til å vise at den omvendte funksjonen $\mathbf{M}$ til $\mathbf{L}$ er deriverbar i $\mathbf{0}$ med Jacobi-matrise I_m. Ifølge Setning 2.6.4 er det nok å vise at

$$\lim_{|\mathbf{y}|\to 0} \frac{\mathbf{M}(\mathbf{y}) - \mathbf{M}(\mathbf{0}) - I_m\mathbf{y}}{|\mathbf{y}|} = \mathbf{0}$$

Siden $\mathbf{M}(\mathbf{0}) = \mathbf{0}$ og $I_m\mathbf{y} = \mathbf{y}$, er dette det samme som

$$\lim_{|\mathbf{y}|\to 0} \frac{\mathbf{M}(\mathbf{y}) - \mathbf{y}}{|\mathbf{y}|} = \mathbf{0}$$

Siden vi er interessert i grensen når $|\mathbf{y}| \to 0$, kan vi nøye oss med å betrakte $\mathbf{y} \in \overline{B}(\mathbf{0}, \frac{r}{2})$. For hver slik $\mathbf{y}$ vet vi at det finnes en entydig bestemt $\mathbf{x}$ i $\overline{B}(\mathbf{0}, r)$ slik at $\mathbf{y} = \mathbf{L}(\mathbf{x})$ og $\mathbf{x} = \mathbf{M}(\mathbf{y})$. Setter vi dette inn i uttrykket ovenfor, får vi

$$\lim_{|\mathbf{y}|\to 0} \frac{\mathbf{M}(\mathbf{y}) - \mathbf{y}}{|\mathbf{y}|} = \lim_{|\mathbf{y}|\to 0} \frac{\mathbf{x} - \mathbf{L}(\mathbf{x})}{|\mathbf{y}|} = -\lim_{|\mathbf{y}|\to 0} \frac{\mathbf{L}(\mathbf{x}) - I_m\mathbf{x}}{|\mathbf{x}|} \cdot \frac{|\mathbf{x}|}{|\mathbf{y}|}$$

Siden $\frac{1}{2}|\mathbf{x}| \leq |\mathbf{L}(\mathbf{x})| = |\mathbf{y}|$, vil $|\mathbf{x}|$ gå mot null når $|\mathbf{y}|$ går mot null. Siden $\mathbf{L}$ er deriverbar i $\mathbf{0}$ med Jacobi-matrise I_m, er derfor

$$\lim_{|\mathbf{y}|\to 0} \frac{\mathbf{L}(\mathbf{x}) - I_m\mathbf{x}}{|\mathbf{x}|} = \lim_{|\mathbf{x}|\to 0} \frac{\mathbf{L}(\mathbf{x}) - \mathbf{L}(\mathbf{0}) - I_m\mathbf{x}}{|\mathbf{x}|} = \mathbf{0}$$

Siden faktoren $\frac{|\mathbf{x}|}{|\mathbf{y}|} = \frac{|\mathbf{x}|}{|\mathbf{L}(\mathbf{x})|} \leq \frac{|\mathbf{x}|}{\frac{1}{2}|\mathbf{x}|} = 2$ er begrenset, følger det at

$$\lim_{|\mathbf{y}|\to 0} \frac{\mathbf{M}(\mathbf{y}) - \mathbf{y}}{|\mathbf{y}|} = -\lim_{|\mathbf{y}|\to 0} \frac{\mathbf{L}(\mathbf{x}) - I_m\mathbf{x}}{|\mathbf{x}|} \cdot \frac{|\mathbf{x}|}{|\mathbf{y}|} = \mathbf{0}$$

Dermed er lemmaet bevist. □

Vi kan nå bevise omvendt funksjonsteorem. For at du skal slippe å bla for mye, skriver vi det opp på nytt.

Teorem 5.6.7 (Omvendt funksjonsteorem) *Anta at U er en åpen mengde i $\mathbb{R}^m$ og at $\mathbf{F} : U \to \mathbb{R}^m$ har kontinuerlige partiellderiverte. Anta at $\bar{\mathbf{x}} \in U$ og at Jacobi-matrisen $\mathbf{F}'(\bar{\mathbf{x}})$ er inverterbar. Da finnes det en omegn $U_0 \subset U$ om $\bar{\mathbf{x}}$ slik at $\mathbf{F}$ restriktert til U_0 er injektiv. Verdimengden V til denne restriksjonen er en omegn om $\bar{\mathbf{y}} = \mathbf{F}(\bar{\mathbf{x}})$, og den omvendte funksjonen $\mathbf{G} : V \to U_0$ er deriverbar i $\bar{\mathbf{y}}$ med Jacobi-matrise*

$$\mathbf{G}'(\bar{\mathbf{y}}) = \mathbf{F}'(\bar{\mathbf{x}})^{-1}$$

Bevis: Planen er å omdanne $\mathbf{F}$ til en funksjon $\mathbf{L}$ som tilfredsstiller betingelsene i foregående lemma. Denne funksjonen $\mathbf{L}$ har da en omvendt funksjon $\mathbf{M}$ som vi kan omdanne til en omvendt funksjon $\mathbf{G}$ for $\mathbf{F}$. Når vi har funnet $\mathbf{G}$, er det lett å sjekke at den har de egenskapene som teoremet angir.

Vi begynner med å definere funksjonen $\mathbf{L}$ ved

$$\mathbf{L}(\mathbf{x}) = A\left(\mathbf{F}(\mathbf{x} + \bar{\mathbf{x}}) - \bar{\mathbf{y}}\right)$$

der $A = \mathbf{F}'(\bar{\mathbf{x}})^{-1}$. Siden $\mathbf{F}$ er definert i en omegn rundt $\bar{\mathbf{x}}$, er $\mathbf{L}$ definert i en omegn rundt $\mathbf{0}$. Vi ser også at

$$\mathbf{L}(\mathbf{0}) = A\left(\mathbf{F}(\bar{\mathbf{x}}) - \bar{\mathbf{y}}\right) = \mathbf{0}$$

siden $\mathbf{F}(\bar{\mathbf{x}}) = \bar{\mathbf{y}}$. Videre gir kjerneregelen

$$\mathbf{L}'(\mathbf{x}) = A\mathbf{F}'(\mathbf{x} + \bar{\mathbf{x}}),$$

så

$$\mathbf{L}'(\mathbf{0}) = A\mathbf{F}'(\bar{\mathbf{x}}) = I_m$$

siden $A = \mathbf{F}'(\bar{\mathbf{x}})^{-1}$.

Dette betyr at $\mathbf{L}$ oppfyller betingelsene i lemmaet ovenfor, og at $\mathbf{L}$ restriktert til en kule $\overline{B}(\mathbf{0}, r)$ har en invers funksjon $\mathbf{M}$ definert på en mengde som inneholder $\overline{B}(\mathbf{0}, \frac{r}{2})$. For å finne en invers funksjon til $\mathbf{F}$ observerer vi at dersom vi snur litt på ligningen $\mathbf{L}(\mathbf{x}) = A\left(\mathbf{F}(\mathbf{x} + \bar{\mathbf{x}}) - \bar{\mathbf{y}}\right)$, får vi

$$\mathbf{F}(\mathbf{x}) = A^{-1}\mathbf{L}(\mathbf{x} - \bar{\mathbf{x}}) + \bar{\mathbf{y}}$$

for alle $\mathbf{x} \in \overline{B}(\bar{\mathbf{x}}, r)$. Siden $\mathbf{L}$ er injektiv og A^{-1} er inverterbar, følger det at $\mathbf{F}$ er injektiv på $\overline{B}(\bar{\mathbf{x}}, r)$. For å finne den omvendte funksjonen, løser vi ligningen

$$\mathbf{y} = A^{-1}\mathbf{L}(\mathbf{x} - \bar{\mathbf{x}}) + \bar{\mathbf{y}}$$

med hensyn på $\mathbf{y}$ og får

$$\mathbf{x} = \bar{\mathbf{x}} + \mathbf{M}(A(\mathbf{y} - \bar{\mathbf{y}}))$$

Dette betyr at $\mathbf{F}$ restriktert til $\mathbf{x} \in \overline{B}(\bar{\mathbf{x}}, r)$ har en invers funksjon $\mathbf{G}$ definert ved

$$\mathbf{G}(\mathbf{y}) = \bar{\mathbf{x}} + \mathbf{M}(A(\mathbf{y} - \bar{\mathbf{y}}))$$

Siden definisjonsmengden til $\mathbf{M}$ omfatter hele $\overline{B}(\mathbf{0}, \frac{r}{2})$, vil definisjonsmengden til $\mathbf{G}$ omfatte alle $\mathbf{y}$ slik at $|A(\mathbf{y}-\bar{\mathbf{y}})| \leq \frac{r}{2}$. Siden $|A(\mathbf{y}-\bar{\mathbf{y}})| \leq |A||\mathbf{y}-\bar{\mathbf{y}}|$, inkluderer dette $\overline{B}(\bar{\mathbf{y}}, \frac{r}{2|A|})$, og følgelig er $\mathbf{G}$ definert i en omegn om $\bar{\mathbf{y}}$.

Resten er lett. Siden $\mathbf{M}$ er deriverbar og $\mathbf{G}(\mathbf{y}) = \bar{\mathbf{x}} + \mathbf{M}(A(\mathbf{y} - \bar{\mathbf{y}}))$, forteller kjerneregelen oss at $\mathbf{G}$ er deriverbar med Jacobi-matrise

$$\mathbf{G}'(\mathbf{y}) = \mathbf{M}'(A(\mathbf{y} - \bar{\mathbf{y}}))A$$

Setter vi inn $\mathbf{y} = \bar{\mathbf{y}}$ og bruker at $\mathbf{M}'(\mathbf{0}) = I_m$, får vi

$$\mathbf{G}'(\bar{\mathbf{y}}) = I_m A = \mathbf{F}'(\bar{\mathbf{x}})^{-1}$$

siden A per definisjon er lik $\mathbf{F}'(\bar{\mathbf{x}})^{-1}$. □

5.7 Ekstremalverdisetningen

I resten av kapitlet skal vi konsentrere oss om maksimums- og minimumsproblemer for funksjoner av flere variable. Fra teorien for funksjoner av én variabel husker vi *ekstremalverdisetningen* (se *Kalkulus*, seksjon 5.3) som sier at en kontinuerlig funksjon definert på et lukket, begrenset intervall er begrenset og har maksimums- og minimumspunkter. I denne seksjonen skal vi bevise et tilsvarende resultat for kontinuerlige funksjoner av flere variable definert på lukkede, begrensede mengder. Vi begynner med noen definisjoner.

Definisjon 5.7.1 *Anta $f : A \to \mathbb{R}$ er en funksjon av m variable. Vi sier at f er* begrenset *dersom det finnes tall K, M slik at*

$$K \leq f(\mathbf{x}) \leq M \qquad \textit{for alle } \mathbf{x} \in A$$

Vi sier at $\mathbf{c} \in A$ er et (globalt) maksimumspunkt *for f dersom*

$$f(\mathbf{c}) \geq f(\mathbf{x}) \qquad \textit{for alle } \mathbf{x} \in A$$

og vi sier at $\mathbf{d} \in A$ er et (globalt) minimumspunkt *for f dersom*

$$f(\mathbf{d}) \leq f(\mathbf{x}) \qquad \textit{for alle } \mathbf{x} \in A$$

Dersom en funksjon har maksimums- og minimumspunkter, er den åpenbart begrenset, men det finnes mange eksempler på funksjoner som er begrenset, men ikke har maksimums- og/eller minimumspunkter. Det er heller ikke uvanlig at en funksjon er ubegrenset til tross for at den er definert på en begrenset mengde. For kontinuerlige funksjoner definert på lukkede, begrensede mengder er det imidlertid orden i sysakene:

Setning 5.7.2 (Ekstremalverdisetningen) *Anta at A er en lukket, begrenset delmengde av $\mathbb{R}^m$ og at $f : A \to \mathbb{R}$ er kontinuerlig. Da har f minimumspunkter og maksimumspunkter og er følgelig begrenset.*

Bevis: Vi skal vise at f har et maksimumspunkt. Beviset for minimumspunkt er helt likt og overlates til leserne. La

$$M = \sup\{f(\mathbf{x}) \ : \ \mathbf{x} \in A\}$$

der vi er enig om å sette $M = \infty$ dersom f ikke er oppad begrenset. Velg en følge $\{\mathbf{x}_n\}$ i A slik at $f(\mathbf{x}_n) \to M$ når $n \to \infty$ (dette er mulig uansett om M er endelig eller uendelig). Siden A er lukket og begrenset, har $\{\mathbf{x}_n\}$ en konvergent delfølge $\{\mathbf{x}_{n_k}\}$ ifølge teorem 5.2.3 (Bolzano-Weierstrass' teorem). Siden A er lukket, ligger grensepunktet $\mathbf{c}$ til denne delfølgen i A (setning 5.1.6), og ifølge setning 5.1.7 er

$$\lim_{k\to\infty} f(\mathbf{x}_{n_k}) = f(\mathbf{c})$$

På den annen side er $\lim_{k\to\infty} f(\mathbf{x}_{n_k}) = M$ (siden $f(\mathbf{x}_n) \to M$ når $n \to \infty$). Dermed må

$$f(\mathbf{c}) = M$$

Dette viser at M er endelig (f kan ikke ha verdien ∞ i et punkt $\mathbf{c}$) og at $\mathbf{c}$ er et maksimumspunkt for f. □

Bemerkning: Beviset ovenfor er det samme som du finner for det éndimensjonale tilfellet i *Kalkulus*, men siden vi har gjort en del forarbeid (spesielt teorem 5.2.3), er det atskillig kortere.

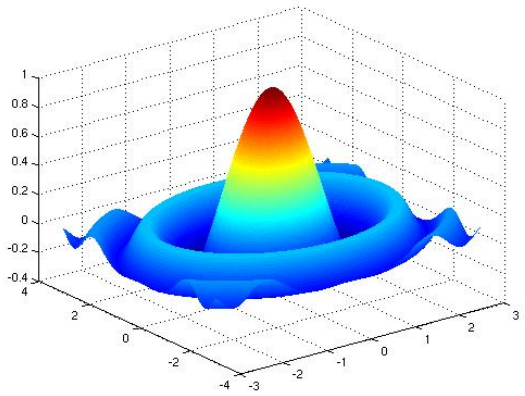
5.8 Maksimums- og minimumspunkter

Hvordan finner vi maksimums- og minimumsverdier for funksjoner av flere variable? For funksjoner av én variabel vet vi at vi først må finne de punktene der den førstederiverte er null, og deretter undersøke hva slags punkter dette er ved enten å se på fortegnskiftet til den førstederiverte eller på fortegnet til den annenderiverte. I dette kapitlet skal vi se at det er en tilsvarende teori for funksjoner av flere variable. Hovedideene er de samme som i det envariable tilfellet, men siden geometrien er rikere, finnes det flere muligheter å holde styr på i den flervariable teorien.

I forrige seksjon definerte vi (globale) maksimums- og minimumspunkter for funksjoner av flere variable. Vi kan imidlertid ikke regne med å finne de globale ekstremalpunktene direkte, men må gå veien om lokale maksima og minima. Før vi skriver opp definisjonen, minner vi om at *snittet* $C \cap D$ av to mengder C og D består av de punktene som er med i *både* C og D.

Definisjon 5.8.1 *La $f\colon A \to \mathbb{R}$ være en funksjon av m variable. Vi sier at f har et* lokalt maksimum *i punktet $\mathbf{a} \in A$ dersom det finnes en kule $B(\mathbf{a}, r)$ med sentrum i $\mathbf{a}$ slik at $f(\mathbf{a}) \geq f(\mathbf{y})$ for alle $\mathbf{y} \in B(\mathbf{a}, r) \cap A$. Tilsvarende kalles $\mathbf{a}$ et* lokalt *minimum dersom det finnes en kule $B(\mathbf{a}, r)$ slik at $f(\mathbf{a}) \leq f(\mathbf{y})$ for alle $y \in B(\mathbf{a}, r) \cap A$.*

Lokale maksimumspunkter ser litt forskjellige ut ettersom $\mathbf{a}$ er et indre punkt eller et randpunkt. Figur 1 viser noen av mulighetene. Et lokalt maksimum i det indre kan f.eks. være en "fjelltopp" som den høyeste toppen på figuren, eller det kan være et punkt på en "åskam" som de andre lokale maksimumspunktene i det indre. I begge disse tilfelle ville alle de partiellderiverte i punktet være 0. Dette behøver imidlertid ikke være tilfellet for lokale maksimumspunkter på randen. Grafen i figur 1 har lokale maksimumspunkter i hjørnene av definisjonsområdet (de fire "flippene" i kanten av figuren), men de partiellderiverte i disse punktene er ikke 0 — punktene ligger i en "skråning" der funksjonen hadde fortsatt å stige hvis den var blitt forlenget utover definisjonsområdet sitt.



Figur 1: Lokale ekstremalpunkter i det indre og på randen

I denne seksjonen skal vi stort sett konsentrere oss om jakten på lokale ekstremalpunkter i det indre av definisjonsområdet. I neste seksjon skal vi se på en teknikk som (blant annet) kan brukes til å finne mulige ekstremalpunkter på randen.

Setning 5.8.2 *Anta at en funksjon $f\colon A \to \mathbb{R}$ har et lokalt maksimum eller minimum i et indre punkt $\mathbf{a}$. Dersom f er deriverbar i $\mathbf{a}$, må $\nabla f(\mathbf{a}) = 0$, dvs. at $\frac{\partial f}{\partial x_i}(\mathbf{a}) = 0$ for alle i.*

Bevis: Vi fører resultatet tilbake til det tilsvarende resultatet for funksjoner av én variabel (*Kalkulus*, setning 6.2.1). Anta at en av de partiellderiverte ikke er null, f.eks. at $\frac{\partial f}{\partial x_i}(\mathbf{a}) \neq 0$. La g være funksjonen av én variabel definert

ved

$$g(x_i) = f(a_1, a_2, \ldots, x_i, \ldots, a_m)$$

(vi holder altså alle variablene konstant unntatt x_i). Da er

$$g'(a_i) = \frac{\partial f}{\partial x_i}(\mathbf{a}) \neq 0$$

Følgelig har g hverken et lokalt maksimum eller et lokalt minimum i a_i, og dermed kan heller ikke f ha et lokalt maksimum eller et lokalt minimum i $\mathbf{a}$. □

Ved hjelp av setningen ovenfor kan vi innskrenke jakten på mulige maksimums- og minimumspunkter betraktelig.

Eksempel 1: La oss forsøke å lokalisere eventuelle maksimums- og minimumspunkter for funksjonen

$$f(x, y) = 3xy - 3x + 9y$$

Vi deriverer:

$$\frac{\partial f}{\partial x} = 3y - 3 \qquad \text{og} \qquad \frac{\partial f}{\partial y} = 3x + 9$$

Ifølge setningen ovenfor bør vi se etter punkter hvor begge de partiellderiverte er null. Dette gir ligningssystemet

$$3y - 3 = 0 \qquad \text{og} \qquad 3x + 9 = 0$$

som har løsningen $x = -3$, $y = 1$. Dette betyr at det eneste mulige maksimums- eller minimumspunktet til f er $(-3, 1)$.

Neste spørsmål er om $(-3, 1)$ virkelig er et lokalt maksimums- eller minimumspunkt. For å avgjøre dette skal vi bruke et triks som av og til er nyttig. Vi innfører nye variable x' og y' slik at punktet $(-3, 1)$ blir det nye origo, det vil si at vi setter

$$\begin{aligned} x' &= x - (-3) = x + 3 \\ y' &= y - 1 \end{aligned}$$

Legg merke til at $(x', y') \to (0, 0)$ når $(x, y) \to (-3, 1)$. Siden $x = x' - 3$, $y = y' + 1$, ser vi at

$$\begin{aligned} f(x, y) &= 3xy - 3x + 9y = 3(x' - 3)(y' + 1) - 3(x' - 3) + 9(y' + 1) \\ &= 3x'y' + 3x' - 9y' - 9 - 3x' + 9 + 9y' + 9 \\ &= 9 + 3x'y' \end{aligned}$$

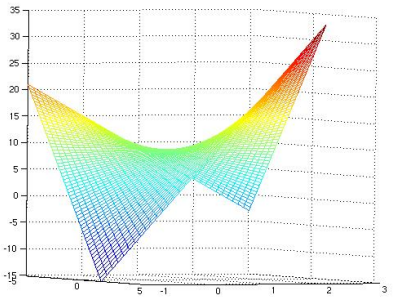
Vi ser at hvis x' og y' har samme fortegn, så vil $f(x, y)$ være større enn $f(-3, 1) = 9$. Har derimot x' og y' motsatt fortegn, vil $f(x, y)$ være mindre

enn 9. Altså kan $(-3,1)$ hverken være et lokalt maksimum eller et lokalt minimum.

La oss til slutt bruke MATLAB til å få et bedre inntrykk av funksjonen. Skriver vi

```
>> x=-5:0.1:1;
>> y=-1:0.1:3;
>> [x,y]=meshgrid(x,y);
>> z=3.*x.*y-3*x+9*y;
>> mesh(x,y,z)
```

svarer MATLAB med figur 2 (etter at vi har rotert litt på aksene for å få et oversiktlig bilde).



Figur 2: Et sadelpunkt

Legg merke til at grafen minner litt om en sal (på en hest), og at "vårt punkt " $(-3,1)$ ligger på det stedet hvor man naturlig sitter på salen. Som vi snart skal komme tilbake til, kalles slike punkter "sadelpunkter". ♣

Eksemplet ovenfor peker på det som skal være hovedproblemstillingen i resten av dette kapitlet: Hvis $\nabla f(\mathbf{a}) = 0$, hvordan avgjør vi da på en effektiv måte om $\mathbf{a}$ er et lokalt maksimum, minimum eller ingen av delene? Teknikken med å skifte variable er nyttig i en del enkle tilfeller, men vi trenger tyngre skyts for å kunne behandle mer kompliserte uttrykk.

La oss begynne med å innføre litt terminologi. Et punkt $\mathbf{a}$ der $\nabla f(\mathbf{a}) = 0$ vil vi kalle et *stasjonært* punkt for funksjonen f. Et stasjonært punkt som hverken er et lokalt maksimum eller et lokalt minimum, vil vi kalle et *sadelpunkt* (se figur 2 ovenfor). Som vi allerede har vært inne på, er det ikke vanskelig å forstå hvor det siste navnet kommer fra – det punktet du sitter på når du rir på en hest, er et typisk eksempel på et sadelpunkt; det er et minimum når du beveger deg i hestens lengderetning og et maksimum nå du beveger deg på tvers av hesten.

Vi tar med et eksempel til på hvordan man finner stasjonære punkter.

Eksempel 2: Finn de stasjonære punktene til

$$f(x,y) = x^2 - y^2 + 4xy - 7x + 3y$$

Vi deriverer:

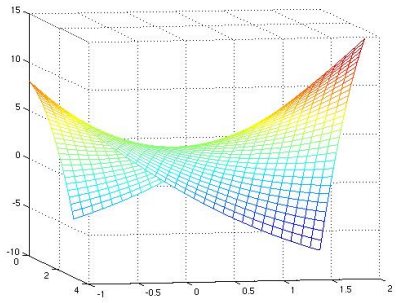
$$\frac{\partial f}{\partial x} = 2x + 4y - 7$$
$$\frac{\partial f}{\partial y} = -2y + 4x + 3$$

Dette gir ligningen

$$2x + 4y = 7$$
$$4x - 2y = -3$$

Løser vi dette ligningssystemet, får vi $x = \frac{1}{10}$, $y = \frac{17}{10}$. Det betyr at punktet $(\frac{1}{10}, \frac{17}{10})$ er et stasjonært punkt for f.

Kjører vi MATLAB på samme måte som i forrige eksempel, får vi grafen på figur 3.



Figur 3: Grafen til $f(x,y) = x^2 - y^2 + 4xy - 7x + 3y$

Den viser at vi også i dette tilfellet har et sadelpunkt. ♣

Bemerkning: Legg merke til at når vi skal finne de stasjonære punktene til en funksjon f av m variable $x_1, x_2, \ldots, x_m$, må vi løse et ligningssystem med m ukjente og m ligninger:

$$\begin{aligned}
\frac{\partial f}{\partial x_1}(x_1, x_2, \ldots, x_m) &= 0 \\
\frac{\partial f}{\partial x_2}(x_1, x_2, \ldots, x_m) &= 0 \\
\vdots \qquad \vdots \qquad &= \vdots \\
\frac{\partial f}{\partial x_m}(x_1, x_2, \ldots, x_m) &= 0
\end{aligned}$$

Dette er et system av den typen som vi brukte Newtons metode til å løse i seksjon 5.5. Når vi får et maks/min-problem i flere variable som går ut over det aller enkleste, må vi regne med å bruke numeriske metoder for å finne løsningen. Newton's metode er én mulighet, men som vi skal se senere (seksjon 5.10), finnes det også mer direkte metoder.

Taylors formel

Når vi arbeider med funksjoner av to variable slik som i eksemplene ovenfor, kan vi ofte bruke MATLAB eller et lignende verktøy til å undersøke om de stasjonære punktene våre er minimumspunkter, maksimumspunkter eller sadelpunkter. Hvis funksjonsgrafen er svært flat i området rundt det stasjonære punktet, kan det imidlertid være vanskelig å avgjøre visuelt hva slags punkt vi har med å gjøre. Arbeider vi med funksjoner av flere enn to variable, er det atskillig verre å bruke visuelle hjelpemidler. Vi trenger derfor en teori som kan hjelpe oss i klassifiseringen av stasjonære punkter.

For funksjoner av én variabel har vi et slikt hjelpemiddel, nemlig *annenderiverttesten*. Den sier at hvis f er en funksjon av én variabel med $f'(a) = 0$, så er a er et lokalt minimum dersom $f''(a) > 0$ og at a er et lokalt maksimum dersom $f''(a) < 0$. Når $f''(a) = 0$, gir testen ingen konklusjon. Vårt mål er å lage en tilsvarende test for funksjoner av flere variable. Dette arbeidet er ganske komplisert fordi en funksjon av flere variable har så mange forskjellige annenderiverte, og de må kombineres på riktig måte for å få en test som virker. Heldigvis skal vi få hjelp av det vi vet om lineær algebra og basiser av egenverdier.

Dersom $f(x_1, \ldots, x_m)$ er en funksjon av n variable, kan vi skrive opp alle de annenordens partiellderiverte som en $m \times m$ matrise:

$$Hf(\mathbf{a}) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2}(\mathbf{a}) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(\mathbf{a}) & \cdots\cdots & \frac{\partial^2 f}{\partial x_1 \partial x_m}(\mathbf{a}) \\ \frac{\partial^2 f}{\partial x_2 \partial x_1}(\mathbf{a}) & \frac{\partial^2 f}{\partial x_2^2}(\mathbf{a}) & \cdots\cdots & \frac{\partial^2 f}{\partial x_2 \partial x_m}(\mathbf{a}) \\ \vdots & & & \vdots \\ \frac{\partial^2 f}{\partial x_m \partial x_1}(\mathbf{a}) & \frac{\partial^2 f}{\partial x_m \partial x_2}(\mathbf{a}) & \cdots\cdots & \frac{\partial^2 f}{\partial x_m^2}(\mathbf{a}) \end{pmatrix}$$

Vi kaller dette *Hesse-matrisen* til f i punktet $\mathbf{a}$ (ikke bland Hesse-matrisen, som er en matrise av annenderiverte til et skalarfelt, sammen med Jacobimatrisen, som er en matrise av førstederiverte til en vektorvaluert funksjon!). Foreløpig er Hesse-matrisen bare en grei måte å skrive opp de annenordens partiellderiverte på, men vi skal snart se at den også har matematisk betydning.

Vi husker fra seksjon 2.5 at dersom de blandede partiellderiverte $\frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{a})$ og $\frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{a})$ er kontinuerlige, så er de like. Det betyr at Hesse-matrisen $Hf(\mathbf{a})$ er symmetrisk. Fra lineær algebra (spektralteoremet 4.10.6) vet vi at

$Hf(\mathbf{a})$ da må ha m reelle egenverdier $\lambda_1(\mathbf{a}), \lambda_2(\mathbf{a}), \ldots, \lambda_m(\mathbf{a})$ (flere av dem kan være like). Det viser seg at det er fortegnet til disse egenverdiene som avgjør om et stasjonært punkt er et lokalt maksimum, et lokalt minimum eller et sadelpunkt. Dersom alle egenverdiene er positive, så er $\mathbf{a}$ et lokalt minimum, dersom alle er negative, så er $\mathbf{a}$ et lokalt maksimum, og dersom det finnes egenverdier med motsatte fortegn, så er $\mathbf{a}$ et sadelpunkt.

For å forstå hva Hesse-matrisen har å gjøre med lokale maksimums- og minimumspunkter, må vi først studere Taylors formel for funksjoner av flere variable.

Setning 5.8.3 (Taylors formel) *La f være en funksjon av m variable. Anta at de annenordens partiellderiverte til f er kontinuerlige i en kule $B(\mathbf{a}, r)$ om $\mathbf{a}$. For enhver $\mathbf{y} \in \mathbb{R}^m$ med $\|\mathbf{y}\| < r$ finnes det et tall $c \in (0,1)$ slik at*

$$f(\mathbf{a}+\mathbf{y}) = f(\mathbf{a}) + \nabla f(\mathbf{a}) \cdot \mathbf{y} + \frac{1}{2}(Hf(\mathbf{a}+c\mathbf{y})\mathbf{y}) \cdot \mathbf{y}$$

I dette uttrykket er $Hf(\mathbf{a}+c\mathbf{y})\mathbf{y}$ matriseproduktet av $Hf(\mathbf{a}+c\mathbf{y})$ og (søylevektoren) $\mathbf{y}$, mens $(Hf(\mathbf{a}+c\mathbf{y})\mathbf{y}) \cdot \mathbf{y}$ er skalarproduktet mellom to vektorer.

Bevis: Definer en funksjon g av én variabel ved

$$g(t) = f(\mathbf{a}+t\mathbf{y}).$$

Bruker vi kjerneregelen for funksjoner av flere variable, ser vi at

$$g'(t) = \sum_{i=1}^{m} \frac{\partial f}{\partial x_i}(\mathbf{a}+t\mathbf{y})y_i = \nabla f(\mathbf{a}+t\mathbf{y}) \cdot \mathbf{y}$$

Bruker vi kjerneregelen på nytt, får vi

$$g''(t) = \sum_{i=1}^{m}\sum_{j=1}^{m} \frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{a}+t\mathbf{y})y_i y_j$$

Dette uttrykket kan også skrives på matriseform

$$g''(t) = (Hf(\mathbf{a}+t\mathbf{y})\mathbf{y}) \cdot \mathbf{y}$$

(skriv opp uttrykket til høyre på koordinatform og skjekk at dette stemmer).

Taylors formel for funksjoner av én variabel (se *Kalkulus*, seksjon 11.2) forteller oss at det finnes et tall mellom 0 og 1 slik at

$$g(1) = g(0) + g'(0) + \frac{1}{2}g''(c)$$

Setter vi inn uttrykkene ovenfor, får vi

$$f(\mathbf{a}+\mathbf{y}) = f(\mathbf{a}) + \nabla f(\mathbf{a}) \cdot \mathbf{y} + \frac{1}{2}(Hf(\mathbf{a}+c\mathbf{y})\mathbf{y}) \cdot \mathbf{y}$$

□

Taylors formel forteller oss at i et stasjonært punkt $\mathbf{a}$ er det fortegnet til $(Hf(\mathbf{a}+c\mathbf{y})\mathbf{y})\cdot\mathbf{y}$ som avgjør om $f(\mathbf{a}+\mathbf{y})$ er større enn eller mindre enn $f(\mathbf{a})$ (husk at i et stasjonært punkt er $\nabla f(\mathbf{a})=0$). Det hadde vært fint om vi kunne ha byttet ut $(Hf(\mathbf{a}+c\mathbf{y})\mathbf{y})\cdot\mathbf{y}$ med $(Hf(\mathbf{a})\mathbf{y})\cdot\mathbf{y}$ i dette utsagnet, siden vi har mye bedre kontroll over punktet $\mathbf{a}$ enn over $\mathbf{a}+c\mathbf{y}$. Den neste versjonen av Taylors formel sier at vi kan gjøre dette byttet uten å måtte betale altfor mye.

Setning 5.8.4 (Taylors formel, versjon 2) *La f være en funksjon av m variable. Anta at de annenordens partiellderiverte til f er kontinuerlige i en omegn om $\mathbf{a}$. Da finnes det en funksjon ε av m variable slik at*

$$\lim_{\mathbf{y}\to 0}\varepsilon(\mathbf{y})=0$$

og

$$f(\mathbf{a}+\mathbf{y})=f(\mathbf{a})+\nabla f(\mathbf{a})\cdot\mathbf{y}+\frac{1}{2}(Hf(\mathbf{a})\mathbf{y})\cdot\mathbf{y}+\varepsilon(\mathbf{y})\|\mathbf{y}\|^2$$

for alle tilstrekkelige små $\mathbf{y}\in\mathbb{R}^m$.

Bemerkning: Setningen ovenfor forteller oss at hvis vi tilnærmer $f(\mathbf{a}+\mathbf{y})$ med $f(\mathbf{a})+\nabla f(\mathbf{a})\cdot\mathbf{y}+\frac{1}{2}(Hf(\mathbf{a})\mathbf{y})\cdot\mathbf{y}$, så gjør vi (for små $\mathbf{y}$) en feil som er liten sammenlignet med $\|\mathbf{y}\|^2$. Siden $\frac{1}{2}(Hf(\mathbf{a})\mathbf{y})\cdot\mathbf{y}$ typisk er av samme størrelse som $\|\mathbf{y}\|^2$, betyr dette at det nest siste leddet vil dominere over feilleddet $\varepsilon(\mathbf{y})\|\mathbf{y}\|^2$.

Bevis for versjon 2 av Taylors formel: Tar vi utgangspunkt i den første versjonen av Taylors formel og legger til og trekker fra $\frac{1}{2}(Hf(\mathbf{a})\mathbf{y})\cdot\mathbf{y}$ på høyresiden, får vi

$$\begin{aligned}f(\mathbf{a}+\mathbf{y})=f(\mathbf{a})+\nabla f(\mathbf{a})\cdot\mathbf{y}+\frac{1}{2}(Hf(\mathbf{a})\mathbf{y})\cdot\mathbf{y}\\+\frac{1}{2}[(Hf(\mathbf{a}+c\mathbf{y})-Hf(\mathbf{a}))\cdot\mathbf{y}]\mathbf{y}\end{aligned}$$

Vi forenkler notasjonen ved å sette

$$A(\mathbf{y})=\frac{1}{2}(Hf(\mathbf{a}+c\mathbf{y})-Hf(\mathbf{a}))$$

Legg merke til at $A(\mathbf{y})$ er en $m\times m$-matrise der koeffisientene $a_{ij}(\mathbf{y})$ går mot null når $\mathbf{y}$ går mot $\mathbf{0}$.

Uttrykket for $f(\mathbf{a}+\mathbf{y})$ kan nå skrives

$$f(\mathbf{a}+\mathbf{y})=f(\mathbf{a})+\nabla f(\mathbf{a})\cdot\mathbf{y}+\frac{1}{2}(Hf(\mathbf{a})\mathbf{y})\cdot\mathbf{y}+(A(\mathbf{y})\mathbf{y})\cdot\mathbf{y}$$

Sammenligner vi dette med formelen i setningen, ser vi at vi må sette

$$\varepsilon(y) = \frac{(A(\mathbf{y})\mathbf{y}) \cdot \mathbf{y}}{\|\mathbf{y}\|^2} = \Big(A(\mathbf{y})\frac{\mathbf{y}}{\|\mathbf{y}\|}\Big) \cdot \frac{\mathbf{y}}{\|\mathbf{y}\|}$$

Det gjenstår å vise at $\lim_{\mathbf{y}\to\mathbf{0}} \varepsilon(\mathbf{y}) = 0$, men det er lett:

$$\lim_{\mathbf{y}\to\mathbf{0}} |\varepsilon(\mathbf{y})| = \lim_{\mathbf{y}\to\mathbf{0}} \Big| \sum_{i,j=1}^{m} a_{ij}(\mathbf{y}) \frac{y_i}{\|\mathbf{y}\|} \cdot \frac{y_j}{\|\mathbf{y}\|} \Big| \leq$$

$$\leq \lim_{\mathbf{y}\to\mathbf{0}} \sum_{i,j=1}^{m} |a_{ij}(\mathbf{y})| = 0$$

hvor vi har brukt at $\frac{|y_i|}{\|\mathbf{y}\|}, \frac{|y_j|}{\|\mathbf{y}\|} \leq 1$ og at $\lim_{\mathbf{y}\to\mathbf{0}} a_{ij}(\mathbf{y}) = 0$. □

For å utnytte Taylors formel trenger vi et enkelt resultat fra lineær algebra.

Lemma 5.8.5 *La A være en symmetrisk $m \times m$-matrise.*

a) *Anta at alle egenverdiene til A er positive: $0 < \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_m$. Da er*

$$(A\mathbf{y}) \cdot \mathbf{y} \geq \lambda_1 \|\mathbf{y}\|^2$$

for alle $\mathbf{y} \in \mathbb{R}^m$

b) *Anta at alle egenverdiene til A er negative: $0 > \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m$. Da er*

$$(A\mathbf{y}) \cdot \mathbf{y} \leq \lambda_1 \|\mathbf{y}\|^2$$

for alle $\mathbf{y} \in \mathbb{R}^m$

Bevis: a) La $\mathbf{v}_1, \mathbf{v}_2 \ldots, \mathbf{v}_m$ være en ortonormal basis av egenvektorer for A slik at $\mathbf{v}_i$ har egenverdi λ_i. Observer først at hvis $\mathbf{y} = c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \ldots + c_m\mathbf{v}_m = \sum_{i=1}^m c_i\mathbf{v}_i$, så er

$$|\mathbf{y}|^2 = \mathbf{y} \cdot \mathbf{y} = (\sum_{i=1}^{m} c_i\mathbf{v}_i) \cdot (\sum_{j=1}^{m} c_j\mathbf{v}_j) =$$

$$= \sum_{i,j=1}^{m} c_i c_j (\mathbf{v}_i \cdot \mathbf{v}_j) = c_1^2 + c_2^2 + \ldots + c_m^2$$

der vi har brukt at $\mathbf{v}_i \cdot \mathbf{v}_j$ er 1 dersom $i = j$ og 0 ellers. Videre er

$$A\mathbf{y} = A(c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \ldots + c_m\mathbf{v}_m) = c_1\lambda_1\mathbf{v}_1 + c_2\lambda_2\mathbf{v}_2 + \ldots + c_m\lambda_m\mathbf{v}_m$$

Ved en tilsvarende regning som ovenfor ser vi at

$$\begin{aligned}(A\mathbf{y}) \cdot \mathbf{y} &= (c_1\lambda_1\mathbf{v}_1 + c_2\lambda_2\mathbf{v}_2 + \ldots + c_m\lambda_m\mathbf{v}_m) \cdot (c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \ldots + c_m\mathbf{v}_m) \\ &= c_1^2\lambda_1 + c_2^2\lambda_2 + \ldots + c_m^2\lambda_m \geq \lambda_1(c_1^2 + c_2^2 + \ldots + c_m^2) \geq \lambda_1\|y\|^2\end{aligned}$$

b) Bruk punkt a) på matrisen $(-A)$. □

Annenderiverttesten

Vi er nå klare til å vise hovedresultatet i dette kapitlet.

Teorem 5.8.6 (Annenderiverttesten) *La* $\mathbf{a}$ *være et stasjonært punkt for en funksjon* f *av* m *variable. Anta at de annenordens partiellderiverte til* f *er kontinuerlige i en omegn om* $\mathbf{a}$*. Da gjelder:*

a) *Hvis alle egenverdiene til* $Hf(\mathbf{a})$ *er (strengt) positive, så er* $\mathbf{a}$ *et lokalt minimumspunkt.*

b) *Hvis alle egenverdiene til* $Hf(\mathbf{a})$ *er (strengt) negative, så er* $\mathbf{a}$ *et lokalt maksimumspunkt.*

c) *Hvis* $Hf(\mathbf{a})$ *har både (strengt) positive og (strengt) negative egenverdier, så er* $\mathbf{a}$ *et sadelpunkt.*

Dersom noen av egenverdiene til $Hf(\mathbf{a})$ *er null og de andre har samme fortegn, så gir testen ingen konklusjon.*

Bevis: a) La λ_1 være den minste egenverdien til $Hf(\mathbf{a})$. Ifølge versjon 2 av Taylors formel er da (husk at $\nabla f(\mathbf{a}) = \mathbf{0}$ siden $\mathbf{a}$ er et stasjonært punkt):

$$\begin{aligned} f(\mathbf{a}+\mathbf{y}) - f(\mathbf{a}) &= \nabla f(\mathbf{a}) \cdot \mathbf{y} + \tfrac{1}{2}(Hf(\mathbf{a})\mathbf{y}) \cdot \mathbf{y} + \varepsilon(\mathbf{y})\|\mathbf{y}\|^2 \\ &= \tfrac{1}{2}(Hf(\mathbf{a})\mathbf{y}) \cdot \mathbf{y} + \varepsilon(\mathbf{y})\|\mathbf{y}\|^2 \\ &\geq \tfrac{1}{2}\lambda_1\|\mathbf{y}\|^2 + \varepsilon(\mathbf{y})\|\mathbf{y}\|^2 \end{aligned}$$

der vi har brukt lemma 5.8.5 i den siste overgangen. Siden λ_1 er positiv og $\varepsilon(\mathbf{y}) \to 0$, vil også $\frac{1}{2}\lambda_1 + \varepsilon(\mathbf{y})$ være positiv når $\mathbf{y}$ er tilstrekkelig liten. Dermed er

$$f(\mathbf{a}+\mathbf{y}) - f(\mathbf{a}) \geq (\tfrac{1}{2}\lambda_1 + \varepsilon(\mathbf{y}))\|\mathbf{y}\|^2 \geq 0$$

som viser at $f(\mathbf{a})$ er et lokalt minimum.

b) La λ_1 være den største egenverdien til $Hf(\mathbf{a})$. Da er

$$\begin{aligned} f(\mathbf{a}+\mathbf{y}) - f(\mathbf{a}) &= \nabla f(\mathbf{a}) \cdot \mathbf{y} + \tfrac{1}{2}(Hf(\mathbf{a})\mathbf{y}) \cdot \mathbf{y} + \varepsilon(\mathbf{y})\|\mathbf{y}\|^2 \\ &= \tfrac{1}{2}(Hf(\mathbf{a})\mathbf{y}) \cdot \mathbf{y} + \varepsilon(\mathbf{y}(\|\mathbf{y}\|^2 \\ &\leq \tfrac{1}{2}\lambda_1\|\mathbf{y}\|^2 + \varepsilon(\mathbf{y})\|\mathbf{y}\|^2 \end{aligned}$$

Siden λ_1 er negativ og $\varepsilon(\mathbf{y}) \to 0$, vil også $\frac{1}{2}\lambda_1 + \varepsilon(\mathbf{y})$ være negativ når $\mathbf{y}$ er tilstrekkelig liten. Dermed er

$$f(\mathbf{a}+\mathbf{y}) - f(\mathbf{a}) \leq (\tfrac{1}{2}\lambda_1 + \varepsilon(\mathbf{y}))\|\mathbf{y}\|^2 \leq 0$$

som viser at $f(\mathbf{a})$ er et lokalt maksimum.

c) For å vise at $\mathbf{a}$ ikke er et lokalt maksimum, lar vi $\mathbf{y}$ være en egenvektor for $Hf(\mathbf{a})$ med en positiv egenverdi λ. Da er

$$\begin{aligned} f(\mathbf{a}+\mathbf{y}) - f(\mathbf{a}) &= \nabla f(\mathbf{a})\mathbf{y} + \tfrac{1}{2}(Hf(\mathbf{a})\mathbf{y})\cdot\mathbf{y} + \varepsilon(\mathbf{y})\|\mathbf{y}\|^2 \\ &= \tfrac{1}{2}\lambda\|\mathbf{y}\|^2 + \varepsilon(y)\|\mathbf{y}\|^2 \end{aligned}$$

Har vi valgt $\mathbf{y}$ tilstrekkelig liten, vil $\frac{1}{2}\lambda + \varepsilon(\mathbf{y}) > 0$, og følgelig er $f(\mathbf{a}+\mathbf{y}) > f(\mathbf{a})$. Dette viser at $\mathbf{a}$ ikke er et lokalt maksimum.

For å vise at $\mathbf{a}$ heller ikke er et lokalt minimum, lar vi isteden $\mathbf{y}$ være en egenvektor med en *negativ* egenverdi og gjennomfører det samme resonnementet. □

For funksjoner av to variable har annenderiverttesten også en annen form som er enklere å bruke i praksis.

Korollar 5.8.7 (Annenderiverttesten i to variable) *La* $\mathbf{a}$ *være et stasjonært punkt for en funksjon* f *av to variable. Anta at de annenordens partiellderiverte er kontinuerlige i en omegn om* $\mathbf{a}$*. La*

$$A = \frac{\partial^2 f}{\partial x^2}(\mathbf{a}), \qquad B = \frac{\partial^2 f}{\partial x \partial y}(\mathbf{a}), \qquad C = \frac{\partial^2 f}{\partial y^2}(\mathbf{a})$$

og la D *være determinanten til Hesse-matrisen:* $D = \begin{vmatrix} A & B \\ B & C \end{vmatrix} = AC - B^2$*. Da gjelder*

(i) *Hvis* $D < 0$*, så er* $\mathbf{a}$ *et sadelpunkt.*

(ii) *Hvis* $D > 0$ *og* $A > 0$*, så er* $\mathbf{a}$ *et lokalt minimum.*

(iii) *Hvis* $D > 0$ *og* $A < 0$*, så er* $\mathbf{a}$ *et lokalt maksimum.*

Hvis $D = 0$*, gir testen ingen konklusjon.*

Bevis: La λ_1 og λ_2 være de to egenverdiene til Hesse-matrisen. Vi vet fra lineær algebra (korollar 4.9.10) at determinanten er produktet av egenverdiene, så $D = \lambda_1\lambda_2$.

(i) Hvis $D < 0$, må λ_1 og λ_2 ha motsatt fortegn. Ifølge den generelle annenderiverttesten er $\mathbf{a}$ et sadelpunkt.

(ii) Hvis $D > 0$, har de to egenverdiene samme fortegn, og ifølge den generelle annenderiverttesten må $\mathbf{a}$ enten være et lokalt maksimum eller minimum. Siden $A = \frac{\partial^2 f}{\partial x^2}(\mathbf{a}) > 0$, har f et lokalt minimum i $\mathbf{a}$ når vi ser på den som en funksjon av x alene. Men da må det være et lokalt minimum f har i $\mathbf{a}$.

(iii) Helt analogt til (ii) □

La oss se hvordan annenderiverttesten virker på noen eksempler. Vi tar først for oss funksjonen fra eksempel 1 på nytt.

Eksempel 3: Vi ser altså på funksjonen $f(x,y) = 3xy - 3x + 9y$ som vi allerede har vist har partiellderiverte

$$\frac{\partial f}{\partial x} = 3y - 3 \qquad \text{og} \qquad \frac{\partial f}{\partial y} = 3x + 9$$

Vi vet også at begge de partiellderiverte er null i punktet $(-3,1)$. For å bruke annenderiverttesten regner vi ut

$$A = \frac{\partial^2 f}{\partial x^2} = 0, \qquad B = \frac{\partial^2 f}{\partial x \partial y} = 3, \qquad C = \frac{\partial^2 f}{\partial y^2} = 0$$

som gir

$$D = \begin{vmatrix} A & B \\ B & C \end{vmatrix} = \begin{vmatrix} 0 & 3 \\ 3 & 0 \end{vmatrix} = -9.$$

Ifølge annenderiverttesten er $(-3,1)$ et sadelpunkt. ♣

La oss se på et litt mer komplisert eksempel:

Eksempel 4: Vi skal finne de stasjonære punktene til

$$f(x,y) = xye^{x-y^2}$$

og avgjøre om de er lokale maksimums-, minimums- eller sadelpunkter.

Derivasjon gir

$$\frac{\partial f}{\partial x} = 1 \cdot ye^{x-y^2} + xye^{x-y^2} = y(1+x)e^{x-y^2}$$
$$\frac{\partial f}{\partial y} = x \cdot 1 \cdot e^{x-y^2} + xye^{x-y^2}(-2y) = x(1-2y^2)e^{x-y^2}.$$

Siden e^{x-y^2} ikke kan være null, er det nok å løse ligningene

$$y(1+x) = 0$$
$$x(1-2y^2) = 0$$

for å finne de stasjonære punktene. Den første ligningen har to løsninger $x = -1$ og $y = 0$. Setter vi $x = -1$ inn i den andre ligningen, får vi $y = \pm\frac{1}{\sqrt{2}} = \pm\frac{\sqrt{2}}{2}$. Setter vi $y = 0$ inn i den andre ligningen, får vi $x = 0$. Vi har altså tre stasjonære punkter $(-1, \frac{\sqrt{2}}{2}), (-1, -\frac{\sqrt{2}}{2})$ og $(0,0)$.

Neste skritt er å regne ut de annenderiverte:

$$\frac{\partial^2 f}{\partial x^2} = y \cdot 1 \cdot e^{x-y^2} + y(1+x)e^{x-y^2} = y(2+x)e^{x-y^2}$$
$$\frac{\partial^2 f}{\partial x \partial y} = 1 \cdot (1+x) \cdot e^{x-y^2} + y(1+x)e^{x-y^2}(-2y) = (1+x)(1-2y^2)e^{x-y^2}$$
$$\frac{\partial^2 f}{\partial y^2} = x(-4y)e^{x-y^2} + x(1-2y^2)e^{x-y^2}(-2y) = -2xy(3-2y^2)e^{x-y^2}$$

Vi må undersøke de stasjonære punktene hver for seg.

Det stasjonære punktet $(0,0)$**:** Her er

$$A = \frac{\partial^2 f}{\partial x^2}(0,0) = 0(2+0)e^{0-0^2} = 0$$
$$B = \frac{\partial^2 f}{\partial x \partial y}(0,0) = (1+0)(1-2\cdot 0^2)e^{0-0^2} = 1$$
$$C = \frac{\partial^2 f}{\partial y^2}(0,0) = -2\cdot 0\cdot 0(3-2\cdot 0^2)e^{0-0^2} = 0.$$

Dette gir $D = \begin{vmatrix} 0 & 1 \\ 1 & 0 \end{vmatrix} = 0^2 - 1^2 = -1$. Altså er $(0,0)$ et sadelpunkt.

Det stasjonære punktet $(-1, \frac{\sqrt{2}}{2})$**:** Her er

$$A = \frac{\partial^2 f}{\partial x^2}\Big(-1,\frac{\sqrt{2}}{2}\Big) = \frac{\sqrt{2}}{2}(2+(-1))e^{-1-(\frac{\sqrt{2}}{2})^2} = \frac{\sqrt{2}}{2}e^{-\frac{3}{2}}$$
$$B = \frac{\partial^2 f}{\partial y \partial x}\Big(-1,\frac{\sqrt{2}}{2}\Big) = (1+(-1))\Big(1-2\Big(\frac{\sqrt{2}}{2}\Big)^2\Big)e^{-1-(\frac{\sqrt{2}}{2})^2} = 0$$
$$C = \frac{\partial^2 f}{\partial y^2}\Big(-1,\frac{\sqrt{2}}{2}\Big) = -2(-1)\frac{\sqrt{2}}{2}\Big(3-2\Big(\frac{\sqrt{2}}{2}\Big)^2\Big)e^{-1-(\frac{\sqrt{2}}{2})^2} = 2\sqrt{2}\,e^{-\frac{3}{2}}$$

Dette gir

$$D = \begin{vmatrix} \frac{\sqrt{2}}{2}\,e^{-3/2} & 0 \\ 0 & 2\sqrt{2}\,e^{-3/2} \end{vmatrix} = 2e^{-3}$$

Siden $D > 0$, $A > 0$, forteller annenderiverttesten oss at $(-1, \frac{\sqrt{2}}{2})$ er et lokalt minimum.

Det stasjonære punktet $(-1, -\frac{\sqrt{2}}{2})$**:** Her er

$$A = \frac{\partial^2 f}{\partial x^2}\Big(-1,-\frac{\sqrt{2}}{2}\Big) = -\frac{\sqrt{2}}{2}(2+(-1))e^{-1-(-\frac{\sqrt{2}}{2})^2} = -\frac{\sqrt{2}}{2}e^{-\frac{3}{2}}$$
$$B = \frac{\partial^2 f}{\partial y \partial x}\Big(-1,-\frac{\sqrt{2}}{2}\Big) = (1+(-1))\Big(1-2\Big(-\frac{\sqrt{2}}{2}\Big)^2\Big)e^{-1-(-\frac{\sqrt{2}}{2})^2} = 0$$
$$C = \frac{\partial^2 f}{\partial y^2}\Big(-1,-\frac{\sqrt{2}}{2}\Big) = -2(-1)\Big(-\frac{\sqrt{2}}{2}\Big)\Big(3-2\Big(-\frac{\sqrt{2}}{2}\Big)^2\Big)e^{-1-(-\frac{\sqrt{2}}{2})^2} = -2\sqrt{2}\,e^{-\frac{3}{2}}$$

Dette gir

$$D = \begin{vmatrix} -\frac{\sqrt{2}}{2}\,e^{-3/2} & 0 \\ 0 & -2\sqrt{2}\,e^{-3/2} \end{vmatrix} = 2e^{-3}$$

Siden $D > 0$, $A < 0$, må $(-1, -\frac{\sqrt{2}}{2})$ være et lokalt maksimum. ♣

Uoppstilte problemer

Til slutt i denne seksjonen skal vi se på noen eksempler på uoppstilte minimums- og maksimumsproblemer.

Eksempel 5: Vi har en 1 meter lang ståltråd som skal deles i maksimalt tre biter. Hver bit skal så bøyes sammen til et kvadrat. Hva er det største og minste samlede areal disse rektanglene kan ha?

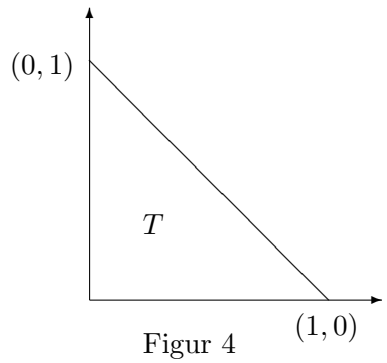
Hvis vi sier at de to første bitene har lengde x og y, må den tredje ha lengde 1-x-y. Det totale arealet er dermed

$$A(x,y) = x^2 + y^2 + (1-x-y)^2$$

Det er fristende å sette igang å derivere med en gang, men la oss først se hvilke verdier x og y kan ha. Vi må åpenbart ha $x \geq 0$, $y \geq 0$ og $x + y \leq 1$ (legg merke til at vi godt kan ha $x = 0$, $y = 0$ eller $x + y = 1$ — det svarer bare til at vi deler opp ståltråden i færre enn tre biter). Dette betyr at vi ønsker å maksimere og minimere funksjonen A på området

$$T = \{(x,y) \ : \ x \geq 0, y \geq 0, x + y \leq 1\}$$

Dette er trekanten med hjørner $(0,0)$, $(1,0)$ og $(0,1)$ som vist på figur 4.



Figur 4

Siden T er en lukket mengde, vet vi fra ekstremalverdisetningen at A har en maksimums- og en minimumsverdi.

Partiellderiverer vi A, får vi

$$\frac{\partial A}{\partial x}(x,y) = 2x + 2(1-x-y)(-1) = 4x + 2y - 2$$

$$\frac{\partial A}{\partial y}(x,y) = 2y + 2(1-x-y)(-1) = 2x + 4y - 2$$

Ligningssystemet $4x + 2y - 2 = 0$, $2x + 4y - 2 = 0$ er lett å løse og gir $x = y = \frac{1}{3}$. Bruker vi annenderiverttesten (gjør det!), ser vi at $(\frac{1}{3}, \frac{1}{3})$ er et lokalt minimum.

Dette betyr at det eneste potensielle ekstremalpunktet vi har i det indre av T, er et lokalt minimum i $(\frac{1}{3}, \frac{1}{3})$. For å finne andre kandidater må vi se på randen til T. Den faller naturlig i tre deler, og vi ser på hver del for seg.

Linjestykket fra $(0,0)$ til $(1,0)$: På dette linjestykket er $y = 0$, og vi får

$$A(x,0) = x^2 + (1-x)^2 \qquad \text{for } 0 \leq x \leq 1$$

Drøfter vi dette uttrykket som en vanlig funksjon av én variabel, finner vi et minimum for $x = \frac{1}{2}$ og maksima for $x = 0$ og $x = 1$. Vi har dermed et mulig minimumspunkt i $(\frac{1}{2}, 0)$ og mulige maksimumspunkter i $(0,0)$ og $(1,0)$.

Linjestykket fra $(0,0)$ til $(0,1)$: På dette linjestykket er $x = 0$, og vi får

$$A(0,y) = y^2 + (1-y)^2 \qquad \text{for } 0 \leq y \leq 1$$

Dette er samme uttrykk som ovenfor bare med x byttet ut med y. Vi får derfor et mulig minimumspunkt i $(0, \frac{1}{2})$ og mulige maksimumspunkter i $(0,0)$ og $(0,1)$.

Linjestykket fra $(1,0)$ til $(0,1)$: På dette linjestykket er $y = 1 - x$, og vi får

$$A(x,1-x) = x^2 + (1-x)^2 \qquad \text{for } 0 \leq x \leq 1$$

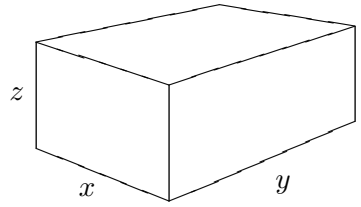
Dette er samme uttrykk som i det første punktet ovenfor, og vi finner et minimum for $x = \frac{1}{2}$ og maksima for $x = 0$ og $x = 1$. Vi har dermed et mulig minimumspunkt i $(\frac{1}{2}, \frac{1}{2})$ og mulige maksimumspunkter i $(1,0)$ og $(0,1)$.

La oss ta en liten oppsummering: Vi har potensielle minimumspunkter i $(\frac{1}{3}, \frac{1}{3})$, $(\frac{1}{2}, 0)$, $(0, \frac{1}{2})$ og $(\frac{1}{2}, \frac{1}{2})$. Regner vi ut funksjonsverdiene, får vi $A(\frac{1}{3}, \frac{1}{3}) = \frac{1}{3}$ og $A(\frac{1}{2}, 0) = A(0, \frac{1}{2}) = A(\frac{1}{2}, \frac{1}{2}) = \frac{1}{2}$. Dette betyr at minimumspunktet er $(\frac{1}{3}, \frac{1}{3})$ og minimumsverdien $\frac{1}{3}$. Vi får altså mimimalt areal når vi deler opp ståltråden i tre like store deler.

På tilsvarende måte har vi de potensielle maksimumspunktene $(0,0)$, $(1,0)$ og $(0,1)$. Regner vi ut funksjonsverdiene, får vi $A(0,0) = A(1,0) = A(1,1) = 1$. Maksimumsverdiene er altså 1 oppnådd i hjørnene $A(0,0) = A(1,0) = A(1,1) = 1$. Geometrisk representerer disse hjørnene den samme løsningen — vi deler ikke opp ståltråden i det hele tatt, men bøyer den sammen til ett stort kvadrat. ♣

Eksemplet ovenfor viser at vi ikke kan neglisjere punktene på randen av definisjonsområdet — det kan hende at det er der det interessante foregår! Det neste eksemplet illustrerer (blant annet) de problemene vi kan støte på når definisjonsområdet *ikke* er begrenset.

Eksempel 6: Vi skal lage en boks med volum V. Hvordan skal vi ordne oss for at overflatearealet A skal bli minst mulig?



Figur 5

Kaller vi sidekantene x, y og z som vist på figuren, ser vi at arealet blir

$$A = 2xy + 2xz + 2yz$$

Siden volumet $V = xyz$ er gitt, kan vi eliminere en av variablene

$$z = \frac{V}{xy}$$

Dette gir

$$A(x, y) = 2xy + \frac{2V}{y} + \frac{2V}{x}$$

I dette uttrykket kan x og y være vilkårlige, positive tall.

Vi deriverer A:

$$\frac{\partial A}{\partial x} = 2y - \frac{2V}{x^2}$$

$$\frac{\partial A}{\partial y} = 2x - \frac{2V}{y^2}$$

For å finne de stasjonære punktene, må vi løse ligningen

$$y = \frac{V}{x^2} \qquad \text{og} \qquad x = \frac{V}{y^2}\,.$$

Setter vi det første uttrykket inn i det andre, ser vi at

$$x = \frac{V}{(\frac{V}{x^2})^2} = \frac{x^4}{V}$$

som gir $x = \sqrt[3]{V}$. Dette gir $y = \frac{V}{x^2} = \frac{V}{(\sqrt[3]{V})^2} = \sqrt[3]{V}$. Vi har altså ett stasjonært punkt $(\sqrt[3]{V}, \sqrt[3]{V})$.

La oss regne ut de annenderiverte:

$$\frac{\partial^2 A}{\partial x^2} = \frac{4V}{x^3}\,; \qquad \frac{\partial^2 A}{\partial y \partial x} = 2\,; \qquad \frac{\partial^2 A}{\partial y^2} = \frac{4V}{y^3}$$

Dette gir

$$\frac{\partial^2 A}{\partial x^2}(\sqrt[3]{V}, \sqrt[3]{V}) = 4$$
$$\frac{\partial^2 A}{\partial y \partial x}(\sqrt[3]{V}, \sqrt[3]{V}) = 2$$
$$\frac{\partial^2 A}{\partial y^2}(\sqrt[3]{V}, \sqrt[3]{V}) = 4$$

og $D = \begin{vmatrix} 4 & 2 \\ 2 & 4 \end{vmatrix} = 16 - 4 = 12$. Følgelig er $(\sqrt[3]{V}, \sqrt[3]{V})$ et lokalt minimum. Legg merke til at siden

$$z = \frac{V}{xy} = \frac{V}{\sqrt[3]{V} \cdot \sqrt[3]{V}} = \sqrt[3]{V}$$

svarer dette lokale minimumet til at vi lar kassen være en kube (alle sider like lange) med overflateareal

$$A = 6 \cdot V^{2/3}$$

La oss oppsummere våre resultater så langt: Vi har vist at funksjonen A bare har ett stasjonært punkt, og det er et lokalt minimum for $x = \sqrt[3]{V}$ og $y = \sqrt[3]{V}$. Dersom det finnes et globalt minimum, må dette være i punktet $(\sqrt[3]{V}, \sqrt[3]{V})$. Mange vil nok slå seg til ro med at dette betyr at $(\sqrt[3]{V}, \sqrt[3]{V})$ er et globalt minimum, men det finnes faktisk en annen mulighet – det kan tenkes at A nærmer seg en lavere "minimalverdi" enn $6V^{2/3}$ uten noen gang å nå frem til den. Vi skal nå vise at dette ikke kan skje og at $A_0 = 6V^{2/3}$ faktisk er den minste verdien arealet kan ha.

Vi tar utgangspunkt i at uttrykket for arealet

$$A(x, y) = 2xy + \frac{2V}{y} + \frac{2V}{x}$$

består av tre positive ledd $2xy$, $\frac{2V}{y}$ og $\frac{2V}{x}$. Skal arealet være mindre enn A_0, må i hvert fall hvert av disse leddene være mindre enn A_0. Begynner vi bakfra, ser vi at skal $\frac{2V}{x}$ være mindre enn A_0, må

$$x > \frac{2V}{A_0} = \tfrac{1}{3}V^{1/3},$$

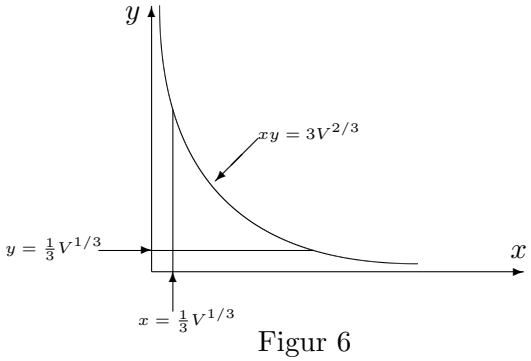
dvs. punktet (x, y) må ligge til høyre for den vertikale linjen $x = \frac{1}{3}V^{1/3}$. Tilsvarende ser vi at skal $\frac{2V}{y}$ være mindre enn A_0, må

$$y > \tfrac{1}{3}V^{1/3}$$

dvs. punktet (x, y) må ligge over for den horisontale linjen $y = \frac{1}{3}V^{1/3}$. Til slutt ser vi at skal $2xy$ være mindre enn A_0, må

$$xy \leq \frac{A_0}{2} = 3V^{2/3}$$

dvs. punktet (x, y) må ligge under hyperbelen $xy = 3V^{2/3}$.



Figur 6

Kombinerer vi disse kravene, ser vi at (x, y) må ligge i det avgrensede området på figur 6 dersom det skal være noe håp om at $A(x, y) < A_0 = 6V^{2/3}$. Vi legger også merke til at på randen av det avgrensede området vil $A(x, y) > A_0$ – her er nemlig ett av de tre leddene $2xy$, $\frac{2V}{y}$ og $\frac{2V}{x}$ lik A_0 mens de to andre er positive.

Inkluderer vi randen, er det avgrensede området lukket og begrenset. Ifølge ekstremalverdisetningen i forrige seksjon har A et (globalt) minimumspunkt i dette området. Dette minimumspunktet kan ikke ligge på randen siden verdien på randen hele tiden er større enn verdien A_0 i det indre punktet $(\sqrt[3]{V}, \sqrt[3]{V})$. Altså må minimumspunkt ligge i det indre, og ifølge setning 5.8.2 må det være et stasjonært punkt. Siden det eneste stasjonære punktet er $(\sqrt[3]{V}, \sqrt[3]{V})$, må dette være et globalt minimum for A på det avgrensede området. Siden funksjonsverdien utenfor dette området alltid er større enn $A(\sqrt[3]{V}, \sqrt[3]{V})$, må $(\sqrt[3]{V}, \sqrt[3]{V})$ være det globale minimumet for A på hele definisjonsområdet. ♣

Det siste resonnementet i eksemplet ovenfor er ganske komplisert, og i praktiske oppgaver hender det ofte at man utelater argumenter av denne typen. Isteden argumenterer man for at det ut i fra oppgavens praktiske tolkning, må finnes et maksimum eller minimum. I eksemplet ovenfor virker det imidlertid ikke så lett å gi et slikt argument.

5.9 Lagranges multiplikatormetode

I forrige seksjon så vi hvordan vi kan finne de lokale maksimums- og minimumspunktene til en funksjon $f(x_1, x_2, \ldots, x_m)$ av flere variable når x_1, x_2, …, x_m får lov til å ha *alle* verdiene i definisjonsområdet til f. Vi skal

nå se hva som skjer når vi har tilleggsbetingelser (såkalte *bibetingelser*) på variablene. La oss begynne med et enkelt (men langt!) eksempel.

Eksempel 1: Vi skal finne maksimums- og minimumsverdien til funksjonen $f(x, y) = xy$ på sirkelen $x^2 + y^2 = 1$ (det er dette som er bibetingelsen). Det er mange måter å løse dette problemet på. Den mest naturlige er kanskje å løse ligningen $x^2 + y^2 = 1$ for y og sette inn (substituere) resultatet i f. Da får vi to ekstremalverdiproblemer med én variabel,

$$h(x) = x\sqrt{1 - x^2}$$

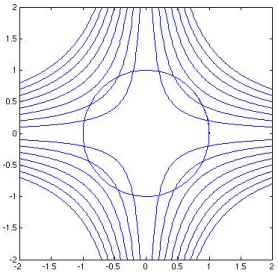
for øvre halvsirkel, og

$$k(x) = -x\sqrt{1 - x^2}$$

for nedre halvsirkel. Det er lett å finne maksimumspunktene til disse funksjonene ved vanlige metoder.

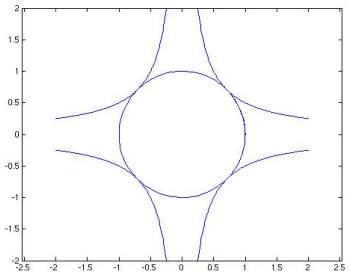
Selv om "substitusjonsmetoden" fungerer bra i dette eksemplet, har den en fundamental svakhet. Dersom bibetingelsen er mer komplisert enn $x^2 + y^2 = 1$, klarer vi ikke å løse ligningen for én av variablene, og hele forsøket vårt bryter sammen. Vi ønsker derfor å finne en metode som ikke er basert på at vi løser ligninger og substituerer.

I figur 1 har vi tegnet opp punktene som tilfredsstiller bibetingelseskurven (sirkelen) sammen med noen av nivåkurvene til funksjonen $f(x, y) = xy$. Nivåkurvene tilsvarer funksjonsverdiene fra -1.6 til 1.6 med trinn på 0.2. Absoluttverdien til funksjonen vokser med x og y, så det er de ytterste nivåkurvene som svarer til høye positive og negative funksjonsverdier (legg merke til at f har positive verdier i første og tredje kvadrant, og negative verdier i annen og fjerde kvadrant).



Figur 1: Nivåkurver og bibetingelseskurve

Vi ser at det er noen nivåkurver som ikke skjærer bibetingelseskurven i det hele tatt — de tilsvarer verdier som funksjonen ikke kan ha så lenge vi innskrenker oss til punkter på sirkelen. Nivåkurver som skjærer sirkelen, tilsvarer verdier som funksjonen har på sirkelen. De største og minste verdiene får vi når nivåkurvene bare berører sirkelen og går ut igjen. Figur 2 viser denne situasjonen.



Figur 2: Optimale nivåkurver

Nivåkurvene i figur 2 tilsvarer verdiene $\frac{1}{2}$ (i første og tredje kvadrant) og $-\frac{1}{2}$ (i annen og fjerde kvadrant), så maksimumsverdien til funksjonen på sirkelen er $\frac{1}{2}$ og minimumsverdien er $-\frac{1}{2}$.

Legg merke til at nivåkurvene i figur 2 tangerer bibetingelseskurven. Det er lett å innse at dette er et generelt fenomen som ikke bare gjelder i dette eksemplet — dersom nivåkurven *krysser* bibetingelseskurven, vil vi normalt ha større verdier på den ene siden av skjæringspunktet og mindre på den andre. Dette betyr at når vi leter etter våre maksimums- og minimumspunkter, så må vi lete etter punkter der nivåkurven tangerer bibetingelseskurven, eller — sagt med andre ord — der normalen til nivåkurven er parallell med normalen til bibetingelseskurven. Disse normalene er lette å finne siden vi vet at gradienter alltid står normalt på nivåkurver (husk setning 3.7.2), og bibetingelseskurven er en nivåkurve for funksjonen $g(x, y) = x^2+y^2$. Vi leter altså etter punkter der de to gradientene

$$\nabla f(x,y) = \begin{pmatrix} y \\ x \end{pmatrix} \quad \text{og} \quad \nabla g(x,y) = \begin{pmatrix} 2x \\ 2y \end{pmatrix}$$

er parallelle, dvs. punkter der det finnes et tall λ slik at

$$\begin{pmatrix} y \\ x \end{pmatrix} = \lambda \begin{pmatrix} 2x \\ 2y \end{pmatrix}$$

Skriver vi ut denne ligningen komponentvis, får vi

$$y = 2\lambda x$$

$$x = 2\lambda y$$

Dette gir oss to ligninger med tre ukjente, x, y og λ (den nye ukjente λ som har sneket seg inn i regnestykket, kalles en *Lagrangemultiplikator* og har gitt navn til hele metoden). I tillegg har vi en tredje ligning siden punktet vårt må tilfredsstille bibetingelsen:

$$x^2 + y^2 = 1$$

Dette ligningssystemet kan løses på mange måter. La oss først observere at ingen av de ukjente x, y kan være 0, for hvis den ene er det, må den andre også være det, og da får vi ikke oppfylt ligningen $x^2+y^2 = 1$. Dette medfører at heller ikke λ kan være 0. Dermed kan vi dele den første av ligningene våre på den andre, og få

$$\frac{y}{x} = \frac{x}{y}$$

som gir $y^2 = x^2$. Setter vi dette inn i den tredje ligningen, får vi $2x^2 = 1$ som gir $x = \pm\frac{\sqrt{2}}{2}$. Siden $y^2 = x^2$, får vi også $y = \pm\frac{\sqrt{2}}{2}$. Dermed har vi fire punkter vi må se videre på: $(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2})$, $(-\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2})$,$(-\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2})$ og $(\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2})$. Setter vi inn i funksjonen $f(x, y) = xy$, får vi

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

og

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

Dette stemmer svært godt med våre grafiske undersøkelser ovenfor, og det er derfor rimelig å tro at vi har en maksimumsverdi $\frac{1}{2}$ som oppnås i punktene $(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2})$ og $(-\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2})$ og en minimumsverdi $-\frac{1}{2}$ som oppnås i punktene $(-\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2})$ og $(\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2})$. ♣

La oss oppsummere eksemplet ovenfor i litt mer generelle vendinger. Vi har en funksjon $f(x, y)$ som vi ønsker å maksimere eller minimere under bibetingelsen $g(x, y) = b$, der b er en konstant. Da må vi lete etter punkter på bibetingelseskurven der $\nabla f(x, y) = \lambda\nabla g(x, y)$. Skriver vi ut denne ligningen komponentvis, får vi to ligninger med to ukjente

$$\frac{\partial f}{\partial x}(x, y) = \lambda\frac{\partial g}{\partial x}(x, y)$$

$$\frac{\partial f}{\partial y}(x, y) = \lambda\frac{\partial g}{\partial y}(x, y)$$

I tillegg har vi bibetingelsen

$$g(x, y) = b$$

slik at vi får tre ligninger med tre ukjente. Løser vi dette ligningssystemet, vil vi (under svært generelle betingelser) ha funnet alle potensielle maksimums- og minimumspunkter for problemet vårt.

Vi kan generalisere enda litt lenger. Anta at vi har en funksjon

$$f(x_1, x_2, \ldots, x_m)$$

av m variable og en bibetingelse

$$g(x_1, x_2, \ldots, x_m) = b$$

Setter vi opp den samme ligningen $\nabla f(x_1, x_2, \ldots, x_m) = \lambda \nabla g(x_1, x_2, \ldots, x_m)$ som før og skriver den ut komponentvis, får vi m ligninger med $m+1$ ukjente x_1, x_2, ...,x_m, λ:

$$\begin{aligned}
\frac{\partial f}{\partial x_1}(x_1, x_2, \ldots, x_m) &= \lambda \frac{\partial g}{\partial x_1}(x_1, x_2, \ldots, x_m) \\
\frac{\partial f}{\partial x_2}(x_1, x_2, \ldots, x_m) &= \lambda \frac{\partial g}{\partial x_2}(x_1, x_2, \ldots, x_m) \\
\vdots \qquad & \vdots \qquad \vdots \\
\frac{\partial f}{\partial x_m}(x_1, x_2, \ldots, x_m) &= \lambda \frac{\partial g}{\partial x_m}(x_1, x_2, \ldots, x_m)
\end{aligned}$$

Legger vi til bibetingelsen

$$g(x_1, x_2, \ldots, x_m) = b$$

har vi $m + 1$ ligninger med $m + 1$ ukjente. Igjen viser det seg at dette ligningssystemet gir oss alle mulige maksimumsl- og minimumspunkter.

Bemerkning: Ser du i litteraturen, vil du finne at bibetingelsene formuleres litt forskjellig — i noen bøker finner du alltid formen $g(x_1, x_2, \ldots, x_m) = 0$, mens andre tillater $g(x_1, x_2, \ldots, x_m) = b$ for en vilkårlig $b \in \mathbb{R}$. Egentlig er det ikke noen forskjell på disse formene — har vi et krav av typen $g(x_1, x_2, \ldots, x_m) = b$, innfører vi bare en ny funksjon

$$\tilde{g}(x_1, x_2, \ldots, x_m) = g(x_1, x_2, \ldots, x_m) - b,$$

og dermed har vi en bibetingelse av typen

$$\tilde{g}(x_1, x_2, \ldots, x_m) = 0$$

Siden $\nabla \tilde{g} = \nabla g$, blir betingelsene for maksimums-/minimumspunkt uforandret. Vi skal derfor formulere resultatene våre for tilfellet $g(x_1, x_2, \ldots, x_m) = b$, men i bevisene nøye oss med å se på tilfellet $g(x_1, x_2, \ldots, x_m) = 0$ som ofte er notasjonsmessig enklere.

Teorem 5.9.1 (Lagranges multiplikatormetode med én bibetingelse) *Anta at U er en åpen delmengde av $\mathbb{R}^m$, og at $f, g : U \to \mathbb{R}$ er to funksjoner med kontinuerlige partiellderiverte. La b være et reelt tall, og anta at $\bar{\mathbf{x}} = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_m)$ er et lokalt maksimums- eller minimumspunkt for f på mengden*

$$A = \{\mathbf{x} \in U \mid g(\mathbf{x}) = b\}$$

Da er enten $\nabla g(\bar{\mathbf{x}}) = 0$, eller det finnes en konstant $\lambda \in \mathbb{R}$ slik at

$$\nabla f(\bar{\mathbf{x}}) = \lambda \nabla g(\bar{\mathbf{x}})$$

Legg merke til at det i teoremet er kommet inn en ekstra mulighet som vi ikke har hatt med tidligere, nemlig at $\nabla g(\bar{\mathbf{x}}) = 0$. Dette er bare naturlig — dersom $\nabla g(\bar{\mathbf{x}}) = 0$, bryter vårt geometriske resonnement sammen, og hva som helst kan hende. Vi understreker også at teoremet bare hjelper oss å finne *potensielle* maksimums- og minimumspunkter — et punkt som tilfredsstiller betingelsene, behøver ikke å være noen av delene, men kan være et generalisert sadelpunkt.

Før vi beviser teoremet, tar vi med et eksempel på bruken.

Eksempel 2: Vi skal finne minimumsverdien til funksjonen

$$f(x, y, z) = (x - 3)^2 + y^2 + z^2$$

under bibetingelsen $x^2 + 4y^2 - z = 0$. Legg merke til at problemet har en geometrisk tolkning — vi ønsker å finne det punktet (x, y, z) på flaten $z = x^2 + 4y^2$ som har kortest avstand til punktet $(3, 0, 0)$.

Lar vi $g(x, y, z) = x^2 + 4y^2 - z = 0$, ser vi at

$$\nabla g(x, y, z) = \begin{pmatrix} 2x \\ 8y \\ -1 \end{pmatrix}$$

Siden ∇g aldri er null, slipper vi å bry oss om tilfellet $\nabla g(\bar{\mathbf{x}}) = 0$. Vi ser videre at

$$\nabla f(x, y, z) = \begin{pmatrix} 2x - 6 \\ 2y \\ 2z \end{pmatrix}$$

Skriver vi ligningen $\nabla f(x, y, z) = \lambda \nabla g(x, y, z)$ på komponentform, får vi (etter å ha forkortet litt)

$$\begin{aligned} x - 3 &= \lambda x \\ y &= 4\lambda y \\ 2z &= -\lambda \end{aligned}$$

I tillegg har vi bibetingelsen

$$x^2 + 4y^2 - z = 0$$

En av utfordringene ved Lagranges multiplikatormetode er å løse ligningene vi kommer frem til. De kan være av forskjellig type, og det er ikke lett å gi generelle råd om hvordan det er lurt å gå frem. I dette tilfellet ser det ut til å være lurest å starte med ligning nummer to, $y = 4\lambda y$. Her er det to muligheter. Dersom $y \neq 0$, må $\lambda = \frac{1}{4}$. Dersom $y = 0$, kan derimot λ være hva som helst. Vi ser på disse tilfellene hver for seg:

Tilfellet $\lambda = \frac{1}{4}$: Den øverste ligningen blir nå til $x-3 = \frac{1}{4}x$, som gir $x = 4$, og den tredje ligningen gir $z = -\frac{1}{8}$. Setter vi dette inn i den nederste ligningen, får vi

$$16 + 4y^2 + \frac{1}{8} = 0$$

som åpenbart ikke har noen løsning. Tilfellet $\lambda = \frac{1}{4}$ fører derfor ikke frem.

Tilfellet $y = 0$: Vi sitter nå igjen med tre ligninger for x, z og λ, nemlig

$$\begin{aligned} x - 3 &= \lambda x \\ 2z &= -\lambda \\ x^2 - z &= 0 \end{aligned}$$

Eliminerer vi z fra de to siste, ser vi at $\lambda = -2x^2$, og setter vi dette inn i den øverste ligningen, sitter vi igjen med $x - 3 = -2x^3$, dvs.

$$2x^3 + x - 3 = 0$$

Vi ser at $x = 1$ er en løsning av denne ligningen. For å undersøke om det finnes flere løsninger, polynomdividerer vi $2x^3 + x - 3$ med $x - 1$, og får $2x^2 + 2x + 3$ som ikke har reelle røtter. Dermed har vi bare én løsning for x, nemlig $x = 1$. Siden $x^2 - z = 0$, følger det at $z = 1$, (det følger også at $\lambda = -2$, men λ er vi egentlig ikke interessert i).

Vi har dermed sett at den eneste løsningen av ligningssystemet er $x = 1, y = 0, z = 1$. Siden den geometriske tolkningen forteller oss at funksjonen må ha et minimumspunkt, er det dette vi har funnet. ♣

Vi er nå klar til å bevise Lagranges multiplikatorteorem for én bibetingelse. Ideen i beviset er den samme som vi hadde i begynnelsen av eksempel 1: vi løser ligningen $g(x_1, x_2, \ldots, x_m) = b$ for én av de variable og substituerer inn i f. På denne måten får vi et "fritt ekstremalverdiproblem" (dvs. et ekstremalverdiproblem uten bibetingelser). I praksis er dette en ubrukelig metode fordi vi ikke greier å løse ligningen $g(x_1, x_2, \ldots, x_m) = 0$, men teoretisk fungerer den fordi implisitt funksjonsteorem forteller oss at det finnes en løsning med de egenskapene vi trenger.

Bevis for teorem 5.9.1: Som påpekt ovenfor er det nok å vise teoremet for $b = 0$. Vi skal vise at dersom $\nabla g(\bar{\mathbf{x}}) \neq 0$, så finnes det en konstant λ slik at $\nabla f(\bar{\mathbf{x}}) = \lambda \nabla g(\bar{\mathbf{x}})$. Siden vi antar at $\nabla g(\bar{\mathbf{x}}) \neq 0$, finnes det minst én variabel x_i slik at $\frac{\partial g}{\partial x_i}(\bar{\mathbf{x}}) \neq 0$. Ved eventuelt å bytte om på variablene, kan vi anta at dette er den siste variabelen x_m. Siden denne variabelen kommer til å spille en litt spesiell rolle i beviset, bytter vi navn på den og kaller den y istedenfor x_m (dette er bare for å gjøre beviset lettere å lese). Vi har dermed funksjoner $f(x_1, \ldots, x_{m-1}, y)$ og $g(x_1 \ldots, x_{m-1}, y)$, der $\frac{\partial g}{\partial y}(\bar{x}_1, \ldots, \bar{x}_{m-1}, \bar{y}) \neq 0$.

Siden $\frac{\partial g}{\partial y}(\bar{x}_1, \ldots, \bar{x}_{m-1}, \bar{y}) \neq 0$, finnes det ifølge implisitt funksjonsteorem (teorem 5.6.3) en deriverbar funksjon ϕ definert i en omegn om $(\bar{x}_1, \ldots, \bar{x}_{m-1})$ slik at $\phi(\bar{x}_1, \ldots, \bar{x}_{m-1}) = \bar{y}$ og

$$g(x_1, \ldots, x_{m-1}, \phi(x_1, \ldots, x_{m-1})) = 0$$

i denne omegnen. Dette betyr at funksjonen

$$h(x_1, \ldots, x_{m-1}) = f((x_1, \ldots, x_{m-1}, \phi(x_1, \ldots, x_{m-1}))$$

har et vanlig ekstremalpunkt (uten bibetingelser) i punktet $(\bar{x}_1, \ldots, \bar{x}_{m-1})$. Dermed er alle de partiellderivete $\frac{\partial h}{\partial x_i}$ lik null i dette punktet. Bruker vi kjerneregelen, er dermed

$$\begin{aligned} 0 = \frac{\partial h}{\partial x_i}(\bar{x}_1, \ldots, \bar{x}_{m-1}) &= \frac{\partial f}{\partial x_i}(\bar{x}_1, \ldots, \bar{x}_{m-1}, \bar{y}) + \\ &+ \frac{\partial f}{\partial y}(\bar{x}_1, \ldots, \bar{x}_{m-1}, \bar{y}) \frac{\partial \phi}{\partial x_i}(\bar{x}_1, \ldots, \bar{x}_{m-1}) \end{aligned}$$

Implisitt funksjonsteorem gir oss at

$$\frac{\partial \phi}{\partial x_i}(\bar{x}_1, \ldots, \bar{x}_{m-1}) = -\frac{\frac{\partial g}{\partial x_i}(\bar{x}_1, \ldots, \bar{x}_{m-1}, \bar{y})}{\frac{\partial g}{\partial y}(\bar{x}_1, \ldots, \bar{x}_{m-1}, \bar{y})}$$

Setter vi dette inn i den foregående ligningen (og bruker at $\bar{\mathbf{x}} = (\bar{x}_1, \ldots, \bar{x}_{m-1}, \bar{y})$), får vi

$$\frac{\partial f}{\partial x_i}(\bar{\mathbf{x}}) = \frac{\frac{\partial f}{\partial y}(\bar{\mathbf{x}})}{\frac{\partial g}{\partial y}(\bar{\mathbf{x}})} \frac{\partial g}{\partial x_i}(\bar{\mathbf{x}})$$

Setter vi $\lambda = \frac{\frac{\partial f}{\partial y}(\bar{\mathbf{x}})}{\frac{\partial g}{\partial y}(\bar{\mathbf{x}})}$, har vi dermed

$$\frac{\partial f}{\partial x_i}(\bar{\mathbf{x}}) = \lambda \frac{\partial g}{\partial x_i}(\bar{\mathbf{x}})$$

for $i = 1, \ldots, m-1$. Det gjenstår å vise at formelen også holder for den siste variabelen, dvs. at $\frac{\partial f}{\partial y}(\bar{\mathbf{x}}) = \lambda \frac{\partial g}{\partial y}(\bar{\mathbf{x}})$. Dette er bare en triviell utregning:

$$\lambda \frac{\partial g}{\partial y}(\bar{\mathbf{x}}) = \frac{\frac{\partial f}{\partial y}(\bar{\mathbf{x}})}{\frac{\partial g}{\partial y}(\bar{\mathbf{x}})} \frac{\partial g}{\partial y}(\bar{\mathbf{x}}) = \frac{\partial f}{\partial y}(\bar{\mathbf{x}})$$

Dermed har vi vist at $\nabla f(\bar{\mathbf{x}}) = \lambda \nabla g(\bar{\mathbf{x}})$, og beviset er fullført. □

Lagranges multiplikatormetode har mange anvendelser. Vi skal komme tilbake til noen av disse senere i seksjonen, men for øyeblikket skal vi nøye oss med å følge opp en problemstilling fra forrige seksjon. I den seksjonen

viste vi hvordan man kan finne lokale maksimums- og minimumspunkter i *det indre* av et område ved å lete etter punkter der alle de partiellderiverte er null. Dersom området vi er interessert i er lukket, er det også mulig at noen av maksimums- eller minimumspunktene ligger på randen (husk eksempel 5 i seksjon 5.8). Disse kan vi finne ved hjelp av Lagranges multiplikatormetode. Vi illustrerer fremgangsmåten med et enkelt eksempel.

Eksempel 3: Vi skal finne maksimums- og minimumsverdien til funksjonen

$$f(x,y) = x^2 - y^3$$

på området

$$A = \{(x,y) \in \mathbb{R}^2 \mid x^2 + y^2 \leq 1\}$$

Siden området er lukket og funksjonen er kontinuerlig, vet vi at den har (globale) maksimums- og minimumspunkter. Hvis et slikt punkt ligger i det indre av området, vet vi at de partiellderiverte må være null i punktet. Siden

$$\frac{\partial f}{\partial x} = 2x \qquad \text{og} \qquad \frac{\partial f}{\partial y} = -3y^2,$$

ser vi at det eneste stasjonære punktet er $(0,0)$ og at $f(0,0) = 0$. Dette er vår første kandidat til tittelen som maksimums- og minimumspunkt. De andre kandidatene må ligge på randen

$$\{(x,y) \in \mathbb{R}^2 \mid x^2 + y^2 = 1\},$$

så vi bruker Lagranges multiplikatormetode med $f(x,y) = x^2 - y^3$ og $g(x,y) = x^2 + y^2$. Vi har

$$\nabla f(x,y) = \begin{pmatrix} 2x \\ -3y^2 \end{pmatrix} \qquad \text{og} \qquad \nabla g(x,y) = \begin{pmatrix} 2x \\ 2y \end{pmatrix}$$

og skriver vi ligningen $\nabla f(x,y) = \lambda \nabla g(x,y)$ på komponentform, får vi (etter litt forkorting) ligningssystemet

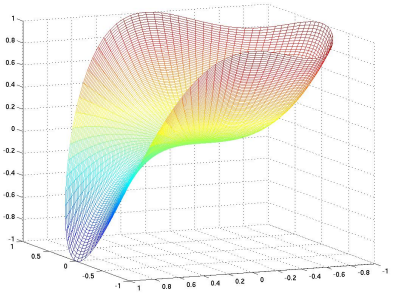
$$\begin{aligned} x &= \lambda x \\ -3y^2 &= 2\lambda y \\ x^2 + y^2 &= 1 \end{aligned}$$

Den første ligningen kan oppfylles på to måter, enten er $x = 0$ eller så er $\lambda = 1$. Vi ser på tilfellene hver for seg. Hvis $x = 0$, følger det fra den siste ligningen at $y = \pm 1$. Dette betyr at $(0, \pm 1)$ er mulige ekstremalpunkter. Setter vi isteden $\lambda = 1$, får den andre ligningen i systemet formen $-3y^2 = 2y$. Denne ligningen har to løsninger, $y = 0$ og $y = -\frac{2}{3}$. Setter vi disse løsningene inn i den tredje ligningen, ser vi at $y = 0$ gir $x = \pm 1$ og at $y = -\frac{2}{3}$ gir $x = \pm\frac{\sqrt{5}}{3}$.

Ialt har vi dermed sju kandidater: $(0,0)$, $(0,\pm 1)$, $(\pm 1,0)$ og $(\pm\frac{\sqrt{5}}{3},-\frac{2}{3})$. For å finne maksimum og minimum, regner vi ut alle funksjonsverdiene:

$$f(0,0)=0,\ f(0,\pm 1)=\mp 1,\ f(\pm 1,0)=1,\ f(\pm\frac{\sqrt{5}}{3},-\frac{2}{3})=\frac{22}{27}$$

Dette viser at maksimumsverdien 1 finner vi i punktene $(0,-1),(\pm 1,0)$, mens minimumsverdien -1 finner vi i punktet $(0,1)$.



Figur 3: Grafisk fremstilling av flaten $f(x,y)=x^2-y^3$

Figur 3 viser grafen. Du ser tydelig de tre maksimumspunktene og det ene minimumspunktet. Punktene $(\pm\frac{\sqrt{5}}{3},-\frac{2}{3})$ er lokale minimumspunkter når du går fra topp til topp langs randen. Det indre punktet $(0,0)$ er et sadelpunkt (vi kunne ha vist dette ved annenderiverttesten og dermed ha eliminert punktet før sluttrunden, men ville neppe ha tjent noe tid på dette). ♣

Lagranges multiplikatormetode med flere bibetingelser

Vi skal nå se på Lagranges multiplikatormetode når vi ønsker å maskimere eller minimere en funksjon

$$f(x_1,x_2,\ldots,x_m)$$

under *flere* bibetingelser

$$\begin{aligned} g_1(x_1,x_2,\ldots,x_m) &= 0\\ g_2(x_1,x_2,\ldots,x_m) &= 0\\ \vdots \qquad\qquad &\ \ \vdots\ \ \vdots\\ g_k(x_1,x_2,\ldots,x_m) &= 0 \end{aligned}$$

Normalt må vi ha $k<m$ for å få et fornuftig ekstremalproblem, og vi skal derfor anta at dette alltid er tilfellet.

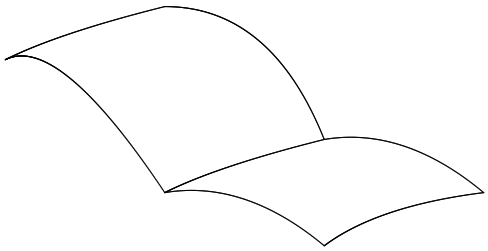
For å få en følelse for problemet ser vi først på tilfellet der vi ønsker å maksimere en funksjon

$$f(x,y,z)$$

av tre variable under to bibetingelser

$$\begin{aligned} g_1(x,y,z) &= 0 \\ g_2(x,y,z) &= 0 \end{aligned}$$

De to ligningene $g_1(x,y,z)=0$ og $g_2(x,y,z)=0$ vil normalt definere to flater i rommet som skjærer hverandre langs en kurve (se figur 3). Problemet er altså å finne den største verdien til f langs denne kurven.



Figur 3: To flater skjærer hverandre i en kurve

Husk at gradienten til f peker i den retningen hvor f vokser raskest. Dersom ∇f ikke står normalt på kurven, er det rimelig å tro at funksjonen langs kurven stiger i den retningen hvor ∇f peker. Skal vi derfor ha maksimum i et punkt, må ∇f i dette punktet stå normalt på kurven, dvs. den må ligge i normalplanet til kurven. Dette normalplanet et utspent av normalvektorene til flatene (prøv å forstå dette geometrisk!), og ∇f må derfor være en lineærkombinasjon av normalvektorene ∇g_1 og ∇g_2 til de to flatene. Vi venter derfor å finne maksimalverdien i et punkt $(\bar{x},\bar{y},\bar{z})$ der det finnes konstanter λ_1 og λ_2 slik at

$$\nabla f(\bar{x},\bar{y},\bar{z}) = \lambda_1 \nabla g_1(\bar{x},\bar{y},\bar{z}) + \lambda_2 \nabla g_2(\bar{x},\bar{y},\bar{z})$$

Følgende teorem forteller oss at denne geometriske intuisjonen er riktig.

Teorem 5.9.2 (Lagranges multiplikatormetode med flere bibetingelser) *Anta at U er en åpen delmengde av $\mathbb{R}^m$, og at $f, g_1, g_2, \ldots, g_k : U \to \mathbb{R}$ er funksjoner med kontinuerlige partiellderiverte. Dersom b_1, b_2, …, b_k er reelle tall og $\bar{\mathbf{x}} = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_m)$ er et lokalt maksimums- eller minimumspunkt for f på mengden*

$$A = \{\mathbf{x} \in U \mid g_1(\mathbf{x}) = b_1, g_2(\mathbf{x}) = b_2, \ldots \text{ og } g_k(\mathbf{x}) = b_k\}$$

så er enten $\nabla g_1(\bar{\mathbf{x}}), \nabla g_2(\bar{\mathbf{x}}), \ldots, \nabla g_k(\bar{\mathbf{x}})$ lineært avhengige, eller det finnes konstanter $\lambda_1, \lambda_2, \ldots, \lambda_k$ slik at

$$\nabla f(\bar{\mathbf{x}}) = \lambda_1 \nabla g_1(\bar{\mathbf{x}}) + \lambda_2 \nabla g_2(\bar{\mathbf{x}}) + \cdots + \lambda_k \nabla g_k(\bar{\mathbf{x}})$$

Før vi ser på beviset, skal vi ta en nærmere kikk på hva teoremet sier. Legg merke til at vi nå har et ligningssystem med $m+k$ ukjente x_1, x_2,...,x_m, λ_1, λ_2 ,..., λ_k, men at vi også har $m+k$ ligninger: Skriver vi ut ligningen

$$\nabla f(\bar{\mathbf{x}}) = \lambda_1 \nabla g_1(\bar{\mathbf{x}}) + \lambda_2 \nabla g_2(\bar{\mathbf{x}}) + \cdots + \lambda_k \nabla g_k(\bar{\mathbf{x}})$$

komponentvis, får vi m ligninger, og bibetingelsene

$$\begin{aligned} g_1(x_1, x_2, \ldots, x_m) &= 0 \\ g_2(x_1, x_2, \ldots, x_m) &= 0 \\ \vdots \qquad\qquad &\vdots \\ g_k(x_1, x_2, \ldots, x_m) &= 0 \end{aligned}$$

gir oss de k siste. La oss se på et enkelt eksempel.

Eksempel 4: Vi skal minimalisere funksjonen $f(x, y, z) = x^2 + y^2 + z^2$ under bibetingelsene

$$x + 2y - z = 2$$

$$-x + y + 2z = 1$$

(Dette er ekvivalent med å finne det punktet på skjæringslinjen mellom planene $x + 2y - z = 2$ og $-x + y + 2z = 1$ som ligger nærmest origo, så det er klart at problemet har en løsning). Vi regner ut gradientene til f og funksjonene $g_1(x, y, z) = x + 2y - z$, $g_2(x, y, z) = -x + y + 2z$:

$$\nabla f(x, y, z) = \begin{pmatrix} 2x \\ 2y \\ 2z \end{pmatrix}, \quad \nabla g_1(x, y, z) = \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix}, \quad \nabla g_2(x, y, z) = \begin{pmatrix} -1 \\ 1 \\ 2 \end{pmatrix}$$

Ifølge teoremet ovenfor leter vi etter punkter der

$$\begin{pmatrix} 2x \\ 2y \\ 2z \end{pmatrix} = \lambda_1 \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix} + \lambda_2 \begin{pmatrix} -1 \\ 1 \\ 2 \end{pmatrix}$$

Skriver vi ut ligningen komponentvis, får vi

$$2x = \lambda_1 - \lambda_2$$

$$2y = 2\lambda_1 + \lambda_2$$

$$2z = -\lambda_1 + 2\lambda_2$$

og i tillegg har vi bibetingelsene

$$x + 2y - z = 2$$

$$-x + y + 2z = 1,$$

altså fem ligninger med fem ukjente. Ligningssystemet er lineært og kan løses ved våre standardmetoder, men vi velger en snarvei. Fra de tre første ligningene, får vi uttrykkene $x = \frac{\lambda_1}{2} - \frac{\lambda_2}{2}$, $y = \lambda_1 + \frac{\lambda_2}{2}$, $z = -\frac{\lambda_1}{2} + \lambda_2$, som vi setter inn i de to siste ligningene. Resultatet er

$$3\lambda_1 - \frac{\lambda_2}{2} = 2$$

$$-\frac{\lambda_1}{2} + 3\lambda_2 = 1$$

Løser vi dette ligningssystemet, får vi $\lambda_1 = \frac{26}{35}$ og $\lambda_2 = \frac{16}{35}$. Setter vi inn i uttrykkene for x, y og z, får vi $x = \frac{1}{7}$, $y = \frac{34}{35}$, $z = \frac{3}{35}$. Siden det geometriske minimaliseringsproblemet vårt åpenbart har en løsning, og $x = \frac{1}{7}$, $y = \frac{34}{35}$, $z = \frac{3}{35}$ er den eneste kandidaten, er problemet løst. ♣

Vi skal nå se på beviset for teorem 5.9.2 (Lagranges multiplikatorteorem for flere bibetingelser). Ideen er akkurat den samme som for én bibetingelse, men utførelsen blir litt mer komplisert fordi matriser erstatter tall en del steder i argumentet. Vi får blant annet bruk for *rangteoremet* for matriser (teorem 4.7.9) som sier at vi alltid kan finne like mange lineært uavhengige søyler som lineært uavhengige rader i en matrise.

**Bevis for teorem 5.9.2:* Det er nok å vise at hvis gradientene $\nabla g_1(\bar{\mathbf{x}})$, $\nabla g_2(\bar{\mathbf{x}})$, ..., $\nabla g_k(\bar{\mathbf{x}})$ er lineært uavhengige, så finnes det konstanter λ_1, λ_2,..., λ_k slik at

$$\nabla f(\bar{\mathbf{x}}) = \lambda_1 \nabla g_1(\bar{\mathbf{x}}) + \lambda_2 \nabla g_2(\bar{\mathbf{x}}) + \cdots + \lambda_k \nabla g_k(\bar{\mathbf{x}})$$

Som i tilfellet med én bibetingelse er det også her nok å se på tilfellet der alle b_i-ene er 0. Lar vi $\mathbf{G} : \mathbb{R}^m \to \mathbb{R}^k$ være funksjonen

$$\mathbf{G}(\mathbf{x}) = \begin{pmatrix} g_1(\mathbf{x}) \\ g_2(\mathbf{x}) \\ \vdots \\ g_k(\mathbf{x}) \end{pmatrix},$$

kan vi dermed sammenfatte alle bibetingelsene i formelen $\mathbf{G}(\mathbf{x}) = \mathbf{0}$. Deriverer vi, ser vi at radene i Jacobi-matrisen $\mathbf{G}'(\mathbf{x})$ rett og slett er gradientene $\nabla g_1(\mathbf{x}), \nabla g_2(\mathbf{x}), \ldots, \nabla g_k(\mathbf{x})$. Siden disse gradientene er lineært uavhengige i punktet $\bar{\mathbf{x}}$, vet vi fra *rangteoremet* 4.7.9 for matriser at Jacobi-matrisen $\mathbf{G}(\bar{\mathbf{x}})$ har k lineært uavhengige søyler. Ved eventuelt å bytte om på rekkefølgen til variablene kan vi anta at dette er de k siste søylene. Siden de $m - k$ første variablene og de k siste kommer til å spille ulike roller i resten av beviset, bytter vi navn på dem for å gjøre forskjellen tydeligere. Vi setter

$$\mathbf{z} = (x_1, x_2, \ldots, x_n)$$

der $n = m - k$, og

$$\mathbf{y} = (x_{n+1}, x_{n+2}, \ldots, x_m)$$

Den "partielle" Jacobi-matrisen $\frac{\partial \mathbf{G}}{\partial \mathbf{y}}(\bar{\mathbf{x}})$ består av de siste k søylene i den "fulle" Jacobi-matrisen $\mathbf{G}'(\bar{\mathbf{x}})$, og siden disse søylene er lineært uavhengige, er $\frac{\partial \mathbf{G}}{\partial \mathbf{y}}(\bar{\mathbf{x}})$ inverterbar. Ifølge teorem 5.6.4 (den vektorvaluerte versjonen av implisitt funksjonsteorem) finnes det da en deriverbar funksjon $\mathbf{\Phi}$ definert i en omegn om $\bar{\mathbf{z}} = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n)$ slik at

$$\mathbf{G}(\mathbf{z}, \mathbf{\Phi}(\mathbf{z})) = \mathbf{0}$$

og $\mathbf{\Phi}(\bar{\mathbf{z}}) = \bar{\mathbf{y}} = (\bar{x}_{n+1}, \bar{x}_{n+2}, \ldots, \bar{x}_m)$. Vi vet også at

$$\mathbf{\Phi}'(\bar{\mathbf{z}}) = -\left(\frac{\partial \mathbf{G}}{\partial \mathbf{y}}(\bar{\mathbf{x}})\right)^{-1} \frac{\partial \mathbf{G}}{\partial \mathbf{z}}(\bar{\mathbf{x}}) \tag{5.9.1}$$

Fra konstruksjonen ovenfor og betingelsene i teoremet følger det at funksjonen

$$h(\mathbf{z}) = f(\mathbf{z}, \mathbf{\Phi}(\mathbf{z}))$$

har et lokalt ekstremalpunkt i $\bar{\mathbf{z}}$, og følgelig er $\nabla h(\bar{\mathbf{z}}) = \mathbf{0}$. Vi skal nå bruke kjerneregelen til å regne ut ∇h. Det er da viktig å huske på at Jacobi-matrisen til et skalarfelt er det samme som gradienten *forutsatt at gradienten oppfattes som en radvektor*. Alle gradienter i dette beviset må derfor oppfattes som radvektorer. Vi får også bruk for at "kjernefunksjonen" $\mathbf{\Psi}(\mathbf{z}) = \begin{pmatrix} \mathbf{z} \\ \mathbf{\Phi}(\mathbf{z}) \end{pmatrix}$ har Jacobi-matrise

$$\mathbf{\Psi}'(\mathbf{z}) = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & \ldots & 1 \\ \frac{\partial \phi_1}{\partial z_1} & \frac{\partial \phi_1}{\partial z_2} & \cdots & \frac{\partial \phi_1}{\partial z_n} \\ \frac{\partial \phi_2}{\partial z_1} & \frac{\partial \phi_2}{\partial z_2} & \cdots & \frac{\partial \phi_2}{\partial z_n} \\ \vdots & \vdots & \ldots & \vdots \\ \frac{\partial \phi_k}{\partial z_1} & \frac{\partial \phi_k}{\partial z_2} & \cdots & \frac{\partial \phi_k}{\partial z_n} \end{pmatrix}$$

Legg merke til at den nedre delen av denne matrisen er Jacobi-matrisen $\mathbf{\Phi}'(\mathbf{z})$. Ifølge kjerneregelen har vi

$$\nabla h(\mathbf{z}) = \nabla f(\mathbf{z}, \mathbf{\Phi}(\mathbf{z}))\mathbf{\Psi}'(\mathbf{z})$$

Dersom vi lar

$$\nabla_{\mathbf{z}} f(\mathbf{z}, \mathbf{y}) = \left(\frac{\partial f}{\partial z_1}(\mathbf{z}, \mathbf{y}), \ldots, \frac{\partial f}{\partial z_n}(\mathbf{z}, \mathbf{y})\right)$$

være den gradienten vi får dersom vi bare tenker på f som en funksjon av z-variablene, og

$$\nabla_{\mathbf{y}} f(\mathbf{z},\mathbf{y}) = \left(\frac{\partial f}{\partial y_1}(\mathbf{z},\mathbf{y}),\ldots,\frac{\partial f}{\partial z_k}(\mathbf{z},\mathbf{y})\right)$$

være den gradienten vi får dersom vi bare tenker på f som en funksjon av y-variablene, så kan ligningen ovenfor skrives:

$$\nabla h(\mathbf{z}) = \nabla_{\mathbf{z}} f(\mathbf{z},\mathbf{\Phi}(\mathbf{z})) + \nabla_{\mathbf{y}} f(\mathbf{z},\mathbf{\Phi}(\mathbf{z}))\mathbf{\Phi}'(\mathbf{z})$$

Siden $\nabla h(\bar{\mathbf{z}}) = \mathbf{0}$, har vi dermed

$$\mathbf{0} = \nabla_{\mathbf{z}} f(\bar{\mathbf{z}},\mathbf{\Phi}(\bar{\mathbf{z}})) + \nabla_{\mathbf{y}} f(\bar{\mathbf{z}},\mathbf{\Phi}(\bar{\mathbf{z}}))\mathbf{\Phi}'(\bar{\mathbf{z}})$$

Fra ligning (5.9.1) ovenfor vet vi at

$$\mathbf{\Phi}'(\bar{\mathbf{z}}) = -\left(\frac{\partial \mathbf{G}}{\partial \mathbf{y}}(\bar{\mathbf{x}})\right)^{-1}\frac{\partial \mathbf{G}}{\partial \mathbf{z}}(\bar{\mathbf{x}})$$

og setter vi dette inn i ligningen ovenfor, får vi (husk at $\bar{\mathbf{x}} = (\bar{\mathbf{z}},\mathbf{\Phi}(\bar{\mathbf{z}}))$):

$$\mathbf{0} = \nabla_{\mathbf{z}} f(\bar{\mathbf{x}}) - \nabla_{\mathbf{y}} f(\bar{\mathbf{x}})\left(\frac{\partial \mathbf{G}}{\partial \mathbf{y}}(\bar{\mathbf{x}})\right)^{-1}\frac{\partial \mathbf{G}}{\partial \mathbf{z}}(\bar{\mathbf{x}})$$

Dette ser komplisert ut, men er egentlig ikke så ille. Observer at $\nabla_{\mathbf{y}} f(\bar{\mathbf{x}})$ er en $1\times k$-matrise og at $\frac{\partial \mathbf{G}}{\partial \mathbf{y}}(\bar{\mathbf{x}})^{-1}$ er en $k\times k$-matrise. Ganger vi sammen disse, får vi en $1\times k$-matrise $\Lambda = (\lambda_1,\lambda_2,\ldots,\lambda_k)$. Dermed kan ligningen ovenfor skrives

$$\nabla_{\mathbf{z}} f(\bar{\mathbf{x}}) = \Lambda\frac{\partial \mathbf{G}}{\partial \mathbf{z}}(\bar{\mathbf{z}})$$

Ganger vi ut høyresiden og bruker at radene i $\frac{\partial \mathbf{G}}{\partial \mathbf{z}}(\bar{\mathbf{z}})$ er gradientene $\nabla_{\mathbf{z}} g_i(\bar{\mathbf{z}})$, får vi

$$\nabla_{\mathbf{z}} f(\bar{\mathbf{x}}) = \lambda_1\nabla_{\mathbf{z}} g_1(\bar{\mathbf{x}}) + \lambda_2\nabla_{\mathbf{z}} g_2(\bar{\mathbf{x}}) + \cdots + \lambda_k\nabla_{\mathbf{z}} g_k(\bar{\mathbf{x}})$$

Dette er nesten det vi skulle vise, det eneste problemet er at vi har de *begrensede* gradientene $\nabla_{\mathbf{z}} f$, $\nabla_{\mathbf{z}} g_1$ $\ldots$,$\nabla_{\mathbf{z}} g_k$ og ikke de *fulle* gradientene ∇f, $\nabla g_1 \ldots \nabla g_k$. Det gjenstår derfor å vise at

$$\nabla_{\mathbf{y}} f(\bar{\mathbf{x}}) = \lambda_1\nabla_{\mathbf{y}} g_1(\bar{\mathbf{x}}) + \lambda_2\nabla_{\mathbf{y}} g_2(\bar{\mathbf{x}}) + \cdots + \lambda_k\nabla_{\mathbf{y}} g_k(\bar{\mathbf{x}})$$

Dette er ekvivalent med å vise at

$$\nabla_{\mathbf{y}} f(\bar{\mathbf{x}}) = \Lambda\frac{\partial \mathbf{G}}{\partial \mathbf{y}}(\bar{\mathbf{x}}),$$

og siden $\Lambda = \nabla_{\mathbf{y}} f(\bar{\mathbf{x}})\left(\frac{\partial \mathbf{G}}{\partial \mathbf{y}}(\bar{\mathbf{x}})\right)^{-1}$, er dette en enkel utregning:

$$\Lambda\frac{\partial \mathbf{G}}{\partial \mathbf{y}}(\bar{\mathbf{x}}) = \nabla_{\mathbf{y}} f(\bar{\mathbf{x}})\left(\frac{\partial \mathbf{G}}{\partial \mathbf{y}}(\bar{\mathbf{x}})\right)^{-1}\frac{\partial \mathbf{G}}{\partial \mathbf{y}}(\bar{\mathbf{x}}) = \nabla_{\mathbf{y}} f(\bar{\mathbf{x}})$$

□

Vi skal ikke komme nærmere inn på det her, men nevner i forbifarten at det også finnes annenderiverttester for ekstremalverdiproblemer med bibetingelser.

Økonomisk tolkning av Lagrangemultiplikatorer

Lagranges multiplikatormetode brukes mye i økonomiske fag. Det er ikke så vanskelig å forstå hvorfor — i økonomi er man opptatt av maksimums- og minimumsproblemer (man ønsker f.eks. å maksimere inntektene og minimere utgiftene), men samtidig har man naturlige bibetingelser — man kan f.eks. ha en begrenset sum å kjøpe råvarer for, eller man har et begrenset antall arbeidstimer å fordele på ulike oppgaver.

I de eksemplene vi har sett på hittil, har Lagrangemultiplikatorene spilt en underordnet rolle; de har vært hjelpestørrelser vi har trengt for å løse problemet vårt, men de har ikke hatt noen selvstendig betydning. I en del økonomiproblemer spiller imidlertid Lagrangemultiplikatorene en viktig rolle.

La oss tenke oss av vi ønsker å maksimere en inntektsfunksjon $f(\mathbf{x})$ under bibetingelsene $g_1(\mathbf{x}) = b_1$, $g(\mathbf{x}) = b_2$, $\ldots$,$g_k(\mathbf{x}) = b_k$. Dersom vi endrer verdiene b_1, b_2, $\ldots$, b_k, må vi selvfølgelig regne med at både maksimalpunktet $\bar{\mathbf{x}}$ og maksimalverdien $\bar{y} = f(\bar{\mathbf{x}})$ endrer seg. Vi kan derfor tenke på disse som funksjoner av $\mathbf{b} = (b_1, b_2, \ldots, b_k)$, altså $\bar{\mathbf{x}}(\mathbf{b})$, $y(\mathbf{b}) = f(\bar{\mathbf{x}}(\mathbf{b}))$. Dersom λ_1, λ_2, $\ldots$,λ_k er Lagrangemultiplikatorene som gir maksimumspunktet $\bar{\mathbf{x}}(\mathbf{b})$, må vi regne med at også disse avhenger av $\mathbf{b}$, altså $\lambda_1(\mathbf{b})$, $\lambda_2(\mathbf{b})$, $\ldots$, $\lambda_k(\mathbf{b})$. Det er lurt å tenke på b_1, b_2 , $\ldots$,b_k som *innsatsfaktorer* i produksjonen — b_1 er kanskje det totalet beløpet vi er villige til å kjøpe råvarer for, b_2 er det totale antall arbeidstimer vi er villige til å bruke i produksjonen, b_3 beløpet vi bruker på å videreutvikle produktene osv.

Et naturlig spørsmål er hvordan en endring i innsatsfaktorene vil påvirke inntektene — hvor mye vil vi f.eks. tjene på å øke arbeidsinnsatsen med 10%? Disse endringene måles av de partiellderiverte

$$\frac{\partial y}{\partial b_i}(\mathbf{b}) = \frac{\partial f(\bar{\mathbf{x}}(\mathbf{b}))}{\partial b_i}$$

Som vi snart skal se, er

$$\frac{\partial y}{\partial b_i}(\mathbf{b}) = \lambda_i(\mathbf{b})$$

Dette betyr at dersom vi gir innsatsfaktoren b_i en liten økning Δb_i, så øker inntektene med $\lambda_i(\mathbf{b})\Delta b_i$. Dersom kostnadene ved å øke b_i én enhet er mindre enn $\lambda_i(\mathbf{b})$, så lønner det seg altså å øke innsatsfaktoren b_i, men dersom kostnadene er større enn $\lambda_i(\mathbf{b})$, lønner det seg å redusere b_i. Av denne grunn kalles $\lambda_i(\mathbf{b})$ *likevektsprisen* til innsatsfaktor b_i (den kalles også *skyggeprisen*

for å understreke at den ikke nødvendigvis har noe med den virkelige prisen å gjøre).

La oss nå vise at

$$\frac{\partial y}{\partial b_i}(\mathbf{b}) = \lambda_i(\mathbf{b})$$

Vi skal ikke gjennomføre et fullstendig matematisk resonnement, men vise at denne formelen følger dersom vi antar at de involverte funksjonene er deriverbare (det går an å vise at dette er tilfellet under svært rimelige betingelser). La oss begynne med å se på bibetingelsene. Siden de alltid er oppfylt, har vi

$$g_j(\bar{\mathbf{x}}(\mathbf{b})) = b_j$$

Deriverer vi dette uttrykket mhp. b_i, får vi (husk kjerneregelen på venstresiden!):

$$\sum_{n=1}^{m} \frac{\partial g_j}{\partial x_n}(\bar{\mathbf{x}}(\mathbf{b}))\frac{\partial \bar{\mathbf{x}}_n}{\partial b_i}(\mathbf{b}) = \begin{cases} 1 & \text{hvis i=j} \\ \\ 0 & \text{ellers} \end{cases} \tag{5.9.2}$$

Deriverer vi inntektsfunksjonen mhp. b_i, får vi tilsvarende

$$\frac{\partial y}{\partial b_i}(\mathbf{b}) = \frac{\partial}{\partial b_i} f(\bar{\mathbf{x}}(\mathbf{b})) = \sum_{n=1}^{m} \frac{\partial f}{\partial x_n}(\bar{\mathbf{x}}(\mathbf{b}))\frac{\partial \bar{\mathbf{x}}_n}{\partial b_i}(\mathbf{b})$$

Ifølge Lagrangebetingelsene er

$$\frac{\partial f}{\partial x_n}(\bar{\mathbf{x}}(\mathbf{b})) = \sum_{j=1}^{k} \lambda_j(\mathbf{b})\frac{\partial g_j}{\partial x_n}(\bar{\mathbf{x}}(\mathbf{b}))$$

og setter vi dette inn i uttrykket ovenfor, får vi

$$\frac{\partial y}{\partial b_i}(\mathbf{b}) = \sum_{n=1}^{m}\sum_{j=1}^{k} \lambda_j(\mathbf{b})\frac{\partial g_j}{\partial x_n}(\bar{\mathbf{x}}(\mathbf{b}))\frac{\partial \bar{\mathbf{x}}_n}{\partial b_i}(\mathbf{b}) =$$

$$= \sum_{j=1}^{k} \lambda_j(\mathbf{b}) \sum_{n=1}^{m} \frac{\partial g_j}{\partial x_n}(\bar{\mathbf{x}}(\mathbf{b}))\frac{\partial \bar{\mathbf{x}}_n}{\partial b_i}(\mathbf{b}) = \lambda_i(\mathbf{b})$$

der vi i siste overgang har brukt formel (5.9.2).

5.10 Gradientmetoden

I de foregående seksjonene har vi studert optimeringsproblemer for funksjoner av flere variable. Vi har sett at både problemer *med* bibetingelser og problemer *uten* bibetingelser leder til ligningssystemer som skal løses. Disse ligningssystemene blir fort så kompliserte at de ikke kan løses for hånd, og man må derfor bruke datamaskiner til å finne tilnærmede løsninger. Newtons

metode er et utmerket redskap for numerisk løsning av ligningssystemer, og det er selvfølgelig utviklet metoder der man kobler optimeringsproblemer direkte til Newtons metode. Men det finnes også andre metoder der man går direkte løse på optimeringsproblemene uten å gå veien om stasjonære punkter. I denne seksjonen skal vi se kort på en slik metode — "gradient-metoden" eller "den bratteste nedstigningsmetoden" ("method of steepest descent").

Grunnideen i denne metoden er enkel. Vi vet at gradienten til en funksjon peker i den retningen hvor funksjonen stiger raskest, og ønsker vi å finne et minimumspunktet for funksjonen (metoden presenteres gjerne som en metode for å finne minimumspunkter), er det naturlig å gå i *motsatt* retning av gradienten. Etter å ha gått i denne retningen et stykke, stopper vi opp, regner ut gradienten på nytt, og fortsetter i motsatt retning av den nye gradienten osv. Denne metoden bringer oss stadig lenger ned, og med litt flaks burde den lede oss til et (lokalt) minimumspunkt.

Et viktig spørsmål er hvor langt vi skal gå i én retning før vi stopper opp og regner ut en ny gradient. Går vi for kort, blir metoden ineffektiv fordi vi må regne ut nye gradienter oftere enn nødvendig, og går vi for langt, risikerer vi å passere minimumspunktet og gå langt ut på den andre siden. Et naturlig valg er å fortsette i den retningen vi har begynt så lenge det går nedover, og først beregne en ny gradient når vi kommer til et sted der det begynner å gå oppover.

La oss se hvordan dette ser ut matematisk. Anta at ønsker å finne et (lokalt) minimumspunkt for funksjonen $f(\mathbf{x})$, og at vi har et startpunkt $\mathbf{x}_0$ som vi tror ikke ligger altfor langt unna det minimalpunktet vi er på jakt etter. Gradienten $\nabla f(\mathbf{x}_0)$ gir oss den retningen hvor funksjonen vokser brattest, og vi ønsker å gå i motsatt retning, altså langs linjen

$$\mathbf{r}(t) = \mathbf{x}_0 - \nabla f(\mathbf{x}_0)t$$

Setter vi dette uttrykket inn i f, får vi en funksjon g av én variabel

$$g(t) = f(\mathbf{r}(t))$$

Vi ønsker å finne minimumspunktet til g, så vi regner ut den deriverte ved hjelp av kjerneregelen

$$g'(t) = \nabla f(\mathbf{r}(t)) \cdot \mathbf{r}'(t) = \nabla f(\mathbf{x}_0 - \nabla f(\mathbf{x}_0)t) \cdot \nabla f(\mathbf{x}_0)$$

Vi er på jakt etter en $t > 0$ slik at $g(t) = 0$, dvs.

$$\nabla f(\mathbf{x}_0 - \nabla f(\mathbf{x}_0)t) \cdot \nabla f(\mathbf{x}_0) = 0$$

(finnes det flere slike t'er, velger vi den første). Selv om dette bare er én ligning med én ukjent, er det slett ikke sikkert vi kan løse den for hånd, og

da må vi bruke f.eks. Newtons metode for å finne en tilnærmet verdi. Når vi har funnet en løsning t_0, setter vi

$$\mathbf{x}_1 = \mathbf{x}_0 - \nabla f(\mathbf{x}_0)t_0$$

og gjentar hele prosedyren med $\mathbf{x}_1$ som utgangspunkt. På denne måten får vi en følge $\{\mathbf{x}_n\}$ som (forhåpentligvis) konvergerer mot et lokalt minimum.

Det går an å analysere gradientmetoden teoretisk (omtrent som vi tidligere har analysert Newtons metode) og komme frem til sikre kriterier for konvergens og gode estimater for konvergenshastighet. Vi skal ikke gjøre dette her, men nøye oss med et enkelt eksempel som viser metoden i praksis.

Eksempel 1: Vi skal bruke gradientmetoden på funksjonen

$$f(x, y) = x^2 + 4y^2$$

Det er lett å se at f har ett eneste minimumspunkt, nemlig $(0, 0)$, så poenget med eksemplet er ikke å finne minimumspunktet, men å studere hvordan gradientmetoden virker.

Vi trenger åpenbart gradienten til f, så la oss regne den ut med en gang:

$$\nabla f(x, y) = (2x, 8y)$$

Anta at vi starter iterasjonen i et punkt $\mathbf{x}_0 = (x_0, y_0)$. Ifølge teorien ovenfor, leter vi etter en $t > 0$ som løser ligningen

$$\nabla f(\mathbf{x}_0 - \nabla f(\mathbf{x}_0)t) \cdot \nabla f(\mathbf{x}_0)$$

Bruker vi at $\nabla f(x, y) = (2x, 8y)$ og $\mathbf{x}_0 - \nabla f(\mathbf{x_0})t = (x_0 - 2x_0t, y_0 - 8y_0t)$, får vi ligningen

$$(2(x_0 - 2x_0t), 8(y_0 - 8y_0t)) \cdot (2x_0, 8y_0) = 0$$

Ganger vi ut og forkorter, gir dette ligningen

$$x_0^2 - 2x_0t + 16y_0^2 - 128y_0^2t = 0$$

som har løsningen

$$t_0 = \frac{x_0^2 + 16y_0^2}{2x_0^2 + 128y_0^2}$$

Dermed får vi

$$\mathbf{x}_1 = \mathbf{x}_0 - \nabla f(\mathbf{x}_0)t_0 = (x_0, y_0) - (2x_0, 8y_0)\frac{x_0^2 + 16y_0^2}{2x_0^2 + 128y_0^2}$$

dvs.

$$x_1 = x_0\left(1 - \frac{x_0^2 + 16y_0^2}{x_0^2 + 64y_0^2}\right) = \frac{48x_0y_0^2}{x_0^2 + 64y_0^2}$$

$$y_1 = y_0\left(1 - \frac{4x_0^2 + 64y_0^2}{x_0^2 + 64y_0^2}\right) = -\frac{3x_0^2 y_0}{x_0^2 + 64y_0^2}$$

Tilsvarende formler gjelder selvfølgelig senere i iterasjonen; vi har

$$x_{n+1} = \frac{48x_n y_n^2}{x_n^2 + 64y_n^2}$$

$$y_{n+1} = -\frac{3x_n^2 y_n}{x_n^2 + 64y_n^2}$$

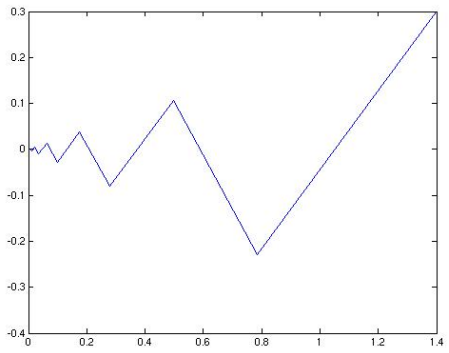
La oss skrive et lite MATLAB-program for å se hvordan iterasjonen forløper:

```
function [x,y]=gradient(a,b,N)
x=zeros(1,N);
y=zeros(1,N);
x(1)=a;
y(1)=b;
for n=1:N
x(n+1)=48*x(n)*y(n)^2/(x(n)^2+64*y(n)^2));
y(n+1)=-3*x(n)^2*y(n)/(x(n)^2+64*y(n)^2));
end
```

Gir vi nå kommandoene

```
>> [x,y]=gradient(1.4,.3,10);
>> plot(x,y)
```

får vi figuren nedenfor som viser hvordan gradientmetoden gir en følge som nærmer seg nullpunktet $(0, 0)$ når vi starter i punktet $(1.4, 0.3)$.



Figur 1: Gradientmetoden ♣

Legg merke til at i figuren ovenfor ser gradientmetoden ut til å "overskyte" ved å gå litt for langt i hvert skritt. Dette er ganske vanlig og gjelder ikke bare eksemplet vi nå har studert. Gradientmetoden har også andre svakheter, og det finnes derfor en rekke videreutviklinger av metoden, men disse skal vi ikke komme inn på her.