Kalkulus og lineær algebra

Utfyllende stoff

Arne Hole

Versjon oppdatert: 10.06.2025

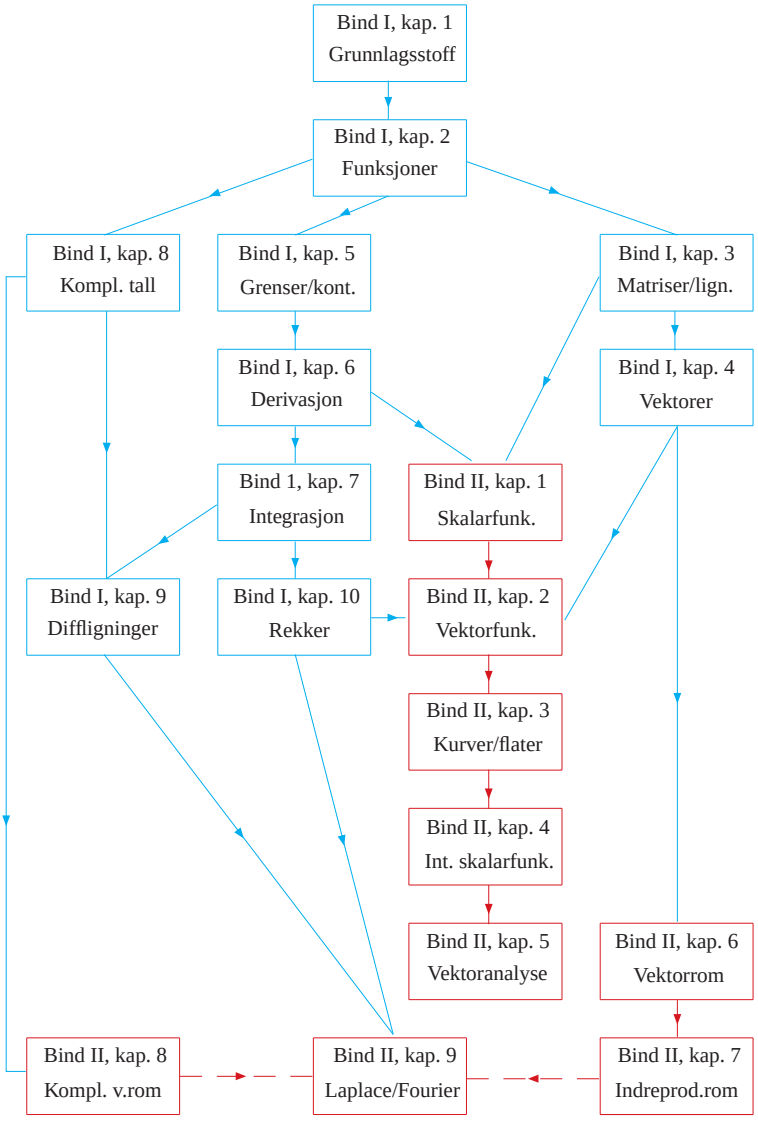
Universitetsforlaget

1. utgave 2023

Henvendelser om publikasjonen kan rettes til
Universitetsforlaget AS
Postboks 508 Sentrum
0105 Oslo
www.universitetsforlaget.no

INNHOLD

Nedenfor er et diagram som viser avhengigheten mellom kapitlene i hovedteksten til *Kalkulus og lineær algebra*, 2. utgave 2025 (bind I og II). I dette heftet referer vi til disse kapitlene ved kapittel nummer og bindnummer.



KAPITTEL 1

LINEÆR ALGEBRA

1.1 Jacobi-metoden

Jacobi-metoden er en iterativ metode for å løse lineære ligningssystemer med n ligninger og n ukjente. Vi starter med et eksempel på hvordan metoden fungerer. Betrakt ligningssystemet

$$\begin{cases} 10x_1 - x_2 + 2x_3 = 6 \\ 2x_1 - x_2 + 10x_3 = -10 \\ -x_1 + 11x_2 - x_3 = 22 \end{cases} \tag{1}$$

For å bruke Jacobi-metoden på dette skriver vi systemet om slik at det blir **strengt diagonaldominant**. Dette betyr at vi for alle $i = 1, \ldots, n$ skal ha

$$|a_{ii}| > \sum_{j \neq i} |a_{ij}|$$

der a_{ij} er elementene i systemets koeffisientmatrise. Absoluttverdien av elementene langs diagonalen i koeffisientmatrisen skal altså være strengt større enn summen av absoluttverdeiene til de øvrige elementene på samme linje i matrisen. For systemet ovenfor kan vi få dette til ved å bytte ligning II og III:

$$\begin{cases} 10x_1 - x_2 + 2x_3 = 6 \\ -x_1 + 11x_2 - x_3 = 22 \\ 2x_1 - x_2 + 10x_3 = -10 \end{cases}$$

Selvsagt kan ikke alle ligningssystemer skrives slik at de blir strengt diagonaldominante, men det gikk altså i dette tilfellet. Etter å ha fått systemet på strengt diagonaldominant form, kan vi for enkelthets skyld dividere hver ligning på diagonalkoeffisienten a_{ii}. I eksemplet vårt gir dette

$$\begin{cases} x_1 - \frac{1}{10}x_2 + \frac{2}{10}x_3 = \frac{3}{5} \\ -\frac{1}{11}x_1 + x_2 - \frac{1}{11}x_3 = 2 \\ \frac{1}{5}x_1 - \frac{1}{10}x_2 + x_3 = -1 \end{cases}$$

Så løser vi med hensyn på leddene langs diagonalen:

$$\begin{cases} x_1 &= \frac{3}{5} + \frac{1}{10}x_2 - \frac{2}{10}x_3 \\ x_2 &= 2 + \frac{1}{11}x_1 + \frac{1}{11}x_3 \\ x_3 &= -1 - \frac{1}{5}x_1 + \frac{1}{10}x_2 \end{cases}$$

Vi er nå klare til å starte den iterative løsningsprosessen. Vi starter med en tilfeldig gjetting, vi velger $x_1^{(0)} = x_2^{(0)} = x_3^{(0)} = 0$, og oppdaterer ved hjelp av

$$\begin{cases} x_1^{(k+1)} &= \frac{3}{5} + \frac{1}{10}x_2^{(k)} - \frac{2}{10}x_3^{(k)} \\ x_2^{(k+1)} &= 2 + \frac{1}{11}x_1^{(k)} + \frac{1}{11}x_3^{(k)} \\ x_3^{(k+1)} &= -1 - \frac{1}{5}x_1^{(k)} + \frac{1}{10}x_2^{(k)} \end{cases}$$

I første steg får vi

$$\begin{cases} x_1^{(1)} &= \frac{3}{5} + \frac{1}{10}\cdot 0 - \frac{2}{10}\cdot 0 &= \frac{3}{5} \\ x_2^{(1)} &= 2 + \frac{1}{11}\cdot 0 + \frac{1}{11}\cdot 0 &= 2 \\ x_3^{(1)} &= -1 - \frac{1}{5}\cdot 0 + \frac{1}{10}\cdot 0 &= -1 \end{cases}$$

Så fortsetter vi på samme måte. Vi bestemmer oss for et visst antall desimaler, og iterterer til vi har stabilitet opp til nøyaktigheten dette tilsvarer. Hvis vi velger fire desimaler, får vi

$$\begin{cases} x_1^{(2)} &= \frac{3}{5} + \frac{1}{10}\cdot 2 - \frac{2}{10}\cdot(-1) &\approx 1 \\ x_2^{(2)} &= 2 + \frac{1}{11}\cdot\frac{3}{5} + \frac{1}{11}\cdot(-1) &\approx 1.9636 \\ x_3^{(2)} &= -1 - \frac{1}{5}\cdot\frac{3}{5} + \frac{1}{10}\cdot 2 &\approx -0.92 \end{cases}$$

$$\begin{cases} x_1^{(3)} &= \frac{3}{5} + \frac{1}{10}\cdot 1.9636 - \frac{2}{10}\cdot(-0.92) &\approx 0.9804 \\ x_2^{(3)} &= 2 + \frac{1}{11}\cdot 1 + \frac{1}{11}\cdot(-0.92) &\approx 2.0073 \\ x_3^{(3)} &= -1 - \frac{1}{5}\cdot 1 + \frac{1}{10}\cdot 1.9636 &\approx -1.0036 \end{cases}$$

$$\begin{cases} x_1^{(4)} &= \frac{3}{5} + \frac{1}{10}\cdot 2.0073 - \frac{2}{10}\cdot(-1.0036) &\approx 1.0015 \\ x_2^{(4)} &= 2 + \frac{1}{11}\cdot 0.9804 + \frac{1}{11}\cdot(-1.0036) &\approx 1.9979 \\ x_3^{(4)} &= -1 - \frac{1}{5}\cdot 0.9804 + \frac{1}{10}\cdot 2.0073 &\approx -0.9954 \end{cases}$$

Regner du ut flere steg, vil du se at løsningen stadig nærmer seg

$$(x_1, x_2, x_3) = (1, 2, -1),$$

som er den eksakte løsningen av systemet (1). Faktum er at siden systemet vårt kunne skrives om slik at det ble trengt diagonaldominant, vil dette skje uansett hvilken gjetting $(x_1^{(0)}, x_2^{(0)}, x_3^{(0)})$ vi starter med. Vi skal nå se hvorfor.

Konvergens av Jacobi-algoritmen

La oss studere hva som skjer når vi bruker Jacobi-metoden på et ligningssystem

$$A\mathbf{x} = \mathbf{b} \tag{2}$$

der A er en strengt diagonaldominant $(n \times n)$-matrise. Her kan vi uten tap av generalitet anta at alle diagonalelementene a_{ii} i matrisen A er 1, jamfør divisjonstrikset i eksemplet. Likning (2) kan skrives

$$\mathbf{x} = \mathbf{b} + (I - A)\mathbf{x} \tag{3}$$

der I er identitetsmatrisen av størrelse $(n \times n)$. Sjekk dette ved å multiplisere $\mathbf{x}$ inn i parentesen. Likning (3) gir grunnlaget for iterasjonen i Jacobi-metoden. Vi starter med en gjetting $\mathbf{x}^{(0)}$ og regner så ut $\mathbf{x}^{(k+1)}$ for $k \geq 0$ ved å bruke

$$\mathbf{x}^{(k+1)} = \mathbf{b} + (I - A)\mathbf{x}^{(k)} \tag{4}$$

At en vektor $\mathbf{x} \in \mathbb{R}^n$ løser (2) er ekvivalent med at $\mathbf{x}$ er et **fikspunkt** den iterative prosessen: Hvis $\mathbf{x}_n$ løser (2), får vi $\mathbf{x}^{k+1} = \mathbf{x}^k$. Sagt på en annen måte: Hvis vi definerer

$$\mathbf{F} : \mathbb{R}^n \to \mathbb{R}^n$$

ved

$$\mathbf{F}(\mathbf{x}) = \mathbf{b} - (I - A)\mathbf{x} \tag{5}$$

så løser $\mathbf{x}$ ligning (2) hvis og bare hvis

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

Konvergensresultatet vi er ute etter, følger nå fra en variant av *Banachs fikspunktteorem* fra bind II (teorem 2.8.1). I dette teoremet brukte vi den vanlige normen

$$||\mathbf{x}|| \stackrel{\text{def}}{=} (x_1^2 + \cdots + x_n^2)^{1/2}$$

for vektorer i $\mathbb{R}^n$, som vanligvis bare skrives $|\mathbf{x}|$. Her og nå er det imidlertid bedre å bruke den såkalte **sup-normen** $||\mathbf{x}||_\infty$ for vektorer. Med denne normen er «størrelsen» av en vektor $\mathbf{x} = (x_1, \ldots, x_n)$ definert som det største tallet man får når man tar absoluttverdien $|x_i|$ av komponentene i $\mathbf{x}$. Altså

$$||\mathbf{x}||_\infty \stackrel{\text{def}}{=} \max_i\{|x_i, \ldots, |x_n|\}$$

Vi kan så definere begrepet **kontraksjon** slik vi gjorde i definisjon 2.8.1 i bind II, bortsett fra at vi nå bruker sup-normen. Per definisjon blir da en funksjon $\mathbf{F} : \mathbb{R}^n \to \mathbb{R}^n$ en kontraksjon av $\mathbb{R}^n$ hvis og bare hvis det fins et reellt tall $c \in (0, 1)$ slik at

$$||\mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{y})||_\infty \leq c \cdot ||\mathbf{x} - \mathbf{y}||_\infty \tag{6}$$

for alle $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. Tallet c kalles en **kontraksjonsfaktor** for $\mathbf{F}$.

TEOREM 1

Fikspunktteorem for kontraksjoner under sup-normen

Anta at $\mathbf{F} : \mathbb{R}^n \to \mathbb{R}^n$ er en kontraksjon med kontraksjonsfaktor c under sup-normen på $\mathbb{R}^n$. Da har $\mathbf{F}$ et unikt fikspunkt $\mathbf{x}^* \in \mathbb{R}^n$, og uansett hvilket startpunkt $\mathbf{x}_0 \in \mathbb{R}^n$ vi velger for iterasjonen

$$\mathbf{x}_{k+1} = \mathbf{F}(\mathbf{x}_k)$$

vil følgen $\{\mathbf{x}_k\}_{k=0}^{\infty}$ konvergere mot $\mathbf{x}^*$. For alle $k \in \mathbb{N}$ har vi feilestimatet

$$||\mathbf{x}^* - \mathbf{x}_k||_\infty \leq \frac{c^k}{1-c} \cdot ||\mathbf{x}_1 - \mathbf{x}_0||_\infty$$

BEVIS Den eneste ikke-trivielle egenskapen ved den vanlige normen $||\mathbf{x}||$ som ble brukt i beviset for Banachs fikspunktteorem i bind II, var trekantulikheten: $||\mathbf{x}+\mathbf{y}|| \leq ||\mathbf{x}|| + ||\mathbf{y}||$. Men denne holder opplagt for sup-normen også: Siden

$$|(\mathbf{x}+\mathbf{y})_i| \leq |x_i| + |y_i|,$$

følger at

$$||\mathbf{x}+\mathbf{y}||_\infty \leq ||x||_\infty + ||y||_\infty$$

Beviset fra bind II går så gjennom akkurat som før. ■

For å vise konvergens av Jacobi-algoritmen holder det nå å bevise at funksjonen $\mathbf{F}$ gitt i (5) oppfyller (6).

TEOREM 2

Jacobi-iterasjonen er en kontraksjon under sup-normen

La A være en strengt diagonaldominant $(n \times n)$-matrise med diagonalelementer lik 1, og la $\mathbf{b} \in \mathbb{R}^n$. Da er funksjonen $\mathbf{F} : \mathbb{R}^n \to \mathbb{R}^n$ gitt ved

$$\mathbf{F}(\mathbf{x}) = \mathbf{b} + (I - A)\mathbf{x}$$

en kontraksjon under sup-normen på $\mathbb{R}^n$.

BEVIS La $B = I - A$. Da er

$$\mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{y}) = (\mathbf{b} + B\mathbf{x}) - (\mathbf{b} + B\mathbf{y}) = B\mathbf{x} - B\mathbf{y} = B(\mathbf{x} - \mathbf{y})$$

For å vise (6) holder det altså å vise

$$||B(\mathbf{x} - \mathbf{y})||_\infty \leq c \cdot ||\mathbf{x} - \mathbf{y}||_\infty,$$

og dette følger direkte hvis vi kan vise at

$$||B(\mathbf{v})||_\infty \leq c \cdot ||\mathbf{v}||_\infty$$

for alle $\mathbf{v} \in \mathbb{R}^n$. Siden A er strengt diagonaldominant og har enere langs diagonalen, oppfyller elementene b_{ij} i matrisen $B = I - A$ ulikheten

$$|b_{i1}| + \cdots + |b_{in}| < 1$$

for hver $i = 1, \ldots, n$. La c være det største av tallene $|b_{i1}| + \cdots + |b_{in}|$, der $i = 1, \ldots, n$. Da er $c < 1$, og vi får

$$\begin{aligned} |(B(\mathbf{x}))_i| &= |b_{i1}x_1 + \cdots + b_{in}x_n| \\ &\leq |b_{i1}| \cdot |x_1| + \cdots + |b_{in}| \cdot |x_n| \\ &\leq (|b_{i1}| + \cdots + |b_{in}|) \cdot ||\mathbf{x}||_\infty \\ &\leq c \cdot ||\mathbf{x}||_\infty \end{aligned}$$

Siden dette holder for $i = 1, \ldots, n$, følger at

$$||B(\mathbf{x})||_\infty \leq c \cdot ||\mathbf{x}||_\infty \quad \blacksquare$$

Kombinasjon av teoremene 1.1.1 og 1.1.2 gir oss nå følgende:

TEOREM 3

Tilstrekkelig betingelse for konvergens av Jacobi-algoritmen

La A være en strengt diagonaldominant $(n \times n)$-matrise. Da konvergerer Jacobi-metoden for ligningssystemet $A\mathbf{x} = \mathbf{b}$ mot en entydig løsning av systemet for alle gitte høyresider $\mathbf{b} \in \mathbb{R}^n$ og alle startverdier $\mathbf{x}^{(0)}$.

Vi har også teoremet under, som vi dropper beviset for. Dette teoremet gir en nødvendig og tilstrekkelig betingelse for konvergens av Jacobi-metoden. Teoremet bruker begrepet **spektralradius** $\rho(A)$ for en kvadratisk matrise A, som per definisjon er den største absoluttverdien (modulusen) som forekommer blant alle egenverdier for A. Merk at vi her må regne komplekst, slik at vi tillater komplekse egenverdier. Merk også at i teoremet inngår matrisen $B = I - A$, som gir iterasjonen i Jacobi-algoritmen (3).

TEOREM 4

Kriterium for konvergens av Jacobi-algoritmen

La A være en $(n \times n)$-matrise. Da konvergerer Jacobi-metoden for ligningssystemet $A\mathbf{x} = \mathbf{b}$ hvis og bare hvis $\rho(I - A) < 1$.

1.2 Gauss–Seidel-metoden

Vi skal nå se på annen iterativ metode for å løse lineære ligningsysstemer med n ligninger og n ukjente, nemlig **Gauss–Seidel-metoden**. Forskjellen fra Jacobi-metoden er at vi nå hele veien bruker de nyeste verdiene av komponentene i den ukjente vektoren $\mathbf{x}$. Vi venter altså ikke til alle komponentene har gått gjennom det aktuelle iterasjonssteget før de tas i bruk. La oss se på det samme eksemplet som vi brukte for Jacobi-metoden:

$$\begin{cases} 10x_1 - x_2 + 2x_3 = 6 \\ 2x_1 - x_2 + 10x_3 = -10 \\ -x_1 + 11x_2 - x_3 = 22 \end{cases} \tag{1}$$

Prepareringen av systemet er akkurat som i Jacobi-metoden. Vi bytter først om ligning 2 og 3, slik at systemet blir strengt diagonaldominant:

$$\begin{cases} 10x_1 - x_2 + 2x_3 = 6 \\ -x_1 + 11x_2 - x_3 = 22 \\ 2x_1 - x_2 + 10x_3 = -10 \end{cases}$$

Så dividerer vi hver ligning med diagonalkoeffisienten og flytter over de øvrige leddene:

$$\begin{cases} x_1 - \frac{1}{10}x_2 + \frac{2}{10}x_3 = \frac{3}{5} \\ -\frac{1}{11}x_1 + x_2 - \frac{1}{11}x_3 = 2 \\ \frac{1}{5}x_1 - \frac{1}{10}x_2 + x_3 = -1 \end{cases}$$

$$\begin{cases} x_1 = \frac{3}{5} + \frac{1}{10}x_2 - \frac{2}{10}x_3 \\ x_2 = 2 + \frac{1}{11}x_1 + \frac{1}{11}x_3 \\ x_3 = -1 - \frac{1}{5}x_1 + \frac{1}{10}x_2 \end{cases}$$

Da er vi klare for å starte den iterative prosessen. Som før starter vi med $x_1^{(0)} = x_2^{(0)} = x_3^{(0)} = 0$, men nå oppdaterer vi slik:

$$\begin{cases} x_1^{(k+1)} = \frac{3}{5} + \frac{1}{10}x_2^{(k)} - \frac{2}{10}x_3^{(k)} \\ x_2^{(k+1)} = 2 + \frac{1}{11}x_1^{(k+1)} + \frac{1}{11}x_3^{(k)} \\ x_3^{(k+1)} = -1 - \frac{1}{5}x_1^{(k+1)} + \frac{1}{10}x_2^{(k+1)} \end{cases}$$

Merk forskjellen fra Jacobi-metoden: Vi bruker den nye verdien $x_1^{(k+1)}$ til å finne $x_2^{(k+1)}$, og så bruker vi de nye verdiene $x_1^{(k+1)}$ og $x_2^{(k+1)}$ til å finne $x_3^{(k+1)}$. I første iterasjon får vi

$$\begin{cases} x_1^{(1)} = \frac{3}{5} + \frac{1}{10}\cdot 0 - \frac{2}{10}\cdot 0 = \frac{3}{5} \\ x_2^{(1)} = 2 + \frac{1}{11}\cdot\frac{3}{5} + \frac{1}{11}\cdot 0 = 2.0545 \\ x_3^{(1)} = -1 - \frac{1}{5}\cdot\frac{3}{5} + \frac{1}{10}\cdot 2.0545 = -0.9146 \end{cases}$$

Igjen velger vi avrunding til fire desimaler. Neste iterasjon blir

$$\begin{cases} x_1^{(2)} &= \frac{3}{5} + \frac{1}{10}\cdot 2.0545 - \frac{2}{10}\cdot(-0.9146) &\approx 0.9884 \\ x_2^{(2)} &= 2 + \frac{1}{11}\cdot 0.9884 + \frac{1}{11}\cdot(-0.9146) &\approx 2.0067 \\ x_3^{(2)} &= -1 - \frac{1}{5}\cdot 0.9884 + \frac{1}{10}\cdot 2.0067 &\approx -0.9970 \end{cases}$$

Her ser vi at iterasjonene allerede er på god vei til å konvergere mot den eksakte løsningen

$$(x_1, x_2, x_3) = (1, 2, -1)$$

Faktum er at Gauss–Seidel-metoden i de fleste tilfeller konvergerer raskere enn Jacobi-metoden, noe som intuitivt sett er rimelig fordi vi tar snarveier ved å bruke de nyeste oppdaterte verdiene underveis.

Konvergens av Gauss–Seidel-algoritmen

Vi skal nå se nærmere på hva som skjer når vi bruker Gauss–Seidel-metoden på et ligningssystem

$$A\mathbf{x} = \mathbf{b} \tag{1}$$

For enkelhets skyld ser vi først på tilfellet der A er en (3×3)-matrise, tilsvarende eksemplet vårt. Iterasjonen i eksemplet kan skrives slik:

$$\begin{bmatrix} x_1^{(k+1)} \\ x_2^{(k+1)} \\ x_3^{(k+1)} \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} + \begin{bmatrix} 0 - a_{12}x_2^{(k)} - a_{13}x_3^{(k)} \\ -a_{21}x_1^{(k+1)} + 0 - a_{23}x_3^{(k)} \\ -a_{31}x_1^{(k+1)} - a_{32}x_2^{(k+1)} + 0 \end{bmatrix}$$

$$= \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ -a_{21} & 0 & 0 \\ -a_{31} & -a_{32} & 0 \end{bmatrix} \begin{bmatrix} x_1^{(k+1)} \\ x_2^{(k+1)} \\ x_3^{(k+1)} \end{bmatrix} + \begin{bmatrix} 0 & -a_{12} & -a_{13} \\ 0 & 0 & -a_{23} \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1^{(k)} \\ x_2^{(k)} \\ x_3^{(k)} \end{bmatrix}$$

Dette kan vi også skrive slik:

$$\begin{bmatrix} 1 & 0 & 0 \\ a_{21} & 1 & 0 \\ a_{31} & a_{32} & 1 \end{bmatrix} \begin{bmatrix} x_1^{(k+1)} \\ x_2^{(k+1)} \\ x_3^{(k+1)} \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} - \begin{bmatrix} 0 & -a_{12} & -a_{13} \\ 0 & 0 & -a_{23} \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1^{(k)} \\ x_2^{(k)} \\ x_3^{(k)} \end{bmatrix}$$

Altså

$$L\mathbf{x}^{(k+1)} = \mathbf{b} - U\mathbf{x}^{(k)} \tag{2}$$

der

$$L = \begin{bmatrix} 1 & 0 & 0 \\ a_{21} & 1 & 0 \\ a_{31} & a_{32} & 1 \end{bmatrix} \quad \text{og} \quad U = \begin{bmatrix} 0 & -a_{12} & -a_{13} \\ 0 & 0 & -a_{23} \\ 0 & 0 & 0 \end{bmatrix}$$

Matrisen L er **nedre triangulær**, mens U kalles **strengt øvretriangulær** fordi den er øvretrianulær og har kun nuller langs diagonalen. Merk at

$$A = L + U$$

Matrisen L har determinant 1 og er inverterbar. Multiplikasjon med L^{-1} på begge sider i (2) gir

$$\mathbf{x}^{(k+1)} = L^{-1}(\mathbf{b} - U\mathbf{x}^{(k)}) = L^{-1}\mathbf{b} - (L^{-1}U)\mathbf{x}^{(k)}$$

Her har vi fått skrevet iterasjonen i Gauss-Seidel-metoden på formen

$$\mathbf{x}^{(k+1)} = \mathbf{b}' - B\mathbf{x}^{(k)}$$

tilsvarende ligning (4) i seksjonen om Jacobi-metoden, men nå med $\mathbf{b}' = L^{-1}\mathbf{b}$ og

$$B = -L^{-1}U$$

som iterasjonsmatrise i stedet for $B = I - A$ som vi hadde i Jacobi-metoden. I full analogi med teorem 1.1.4 for Jacobimetoden har vi nå dette teoremet, som vi dropper beviset for:

TEOREM 1

Kriterium for konvergens av Gauss–Seidel-algoritmen

La A være en $(n \times n)$-matrise. Da konvergerer Gauss–Seidel-metoden for ligningssystemet $A\mathbf{x} = \mathbf{b}$ hvis og bare hvis $\rho(-L^{-1}U) < 1$, der vi har splittet $A = L + U$ med L nedretriangulær og U strengt øvretriangulær.

Videre har vi følgende analogi til teorem 1.1.3. Også her dropper vi beviset.

TEOREM 2

Tilstrekkelig betingelse for konvergens av Gauss–Seidel

La A være en strengt diagonaldominant $(n \times n)$-matrise. Da konvergerer Gauss-Seidel-metoden for ligningssystemet $A\mathbf{x} = \mathbf{b}$ mot en entydig løsning av systemet for alle gitte høyresider $\mathbf{b} \in \mathbb{R}^n$ og alle startverdier $\mathbf{x}^{(0)}$.

1.3 Banachs lemma

Vi starter med litt repetisjon fra seksjon 2.7 i bind II. **Normen** $||A||$ av en $(n \times n)$-matrise A er definert ved

$$||A|| = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}}$$

Normen $||A||$ er altså det samme som lengden av vektoren du får hvis du leser matrisen A linjevis og slår alle linjene sammen til en vektor med n^2 komponenter. Hvis $\mathbf{x}$ er en vilkårlig vektor i $\mathbb{R}^n$, har vi da (oppgave 1.3.1)

$$|A\mathbf{x}| \leq ||A|| \cdot |x|$$

Altså: Ingen vektor $\mathbf{x}$ kan få lengden sin strukket med mer enn faktoren $||A||$ når den multipliseres med matrisen A. Hvis $\mathbf{x} \neq \mathbf{0}$, gir divisjon med $|\mathbf{x}|$ i ligningen ovenfor

$$\frac{|A\mathbf{x}|}{|\mathbf{x}|} \leq ||A||$$

Denne øvre begrensningen gjør at følgende definisjon blir meningsfull:

DEFINISJON 1

Operatornormen til en matrise

Operatornormen $|A|$ til en $(n \times n)$-matrise er definert ved

$$|A| = \sup\left\{ \frac{|A\mathbf{x}|}{|\mathbf{x}|} \;\middle|\; \mathbf{x} \in \mathbb{R}^n \text{ slik at } \mathbf{x} \neq 0 \right\}$$

Merk at vi skriver operatornormen $|A|$ med vertikale streker, akkurat som vi også gjør med lengden av vektorer. I definisjonen ovenfor er altså $|A\mathbf{x}|$ lengden av vektoren $A\mathbf{x}$, mens $|A|$ er operatornormen til matrisen A. Hva de vertikale strekene betyr, kommer altså an på om de står rundt en matrise eller en vektor.

La $A_0, A_1, A_2, \ldots$ være en følge av $(n \times n)$-matriser. Vi sier at **matriserekken** $A_0 + A_1 + A_2 + \cdots$ konvergerer mot $(n \times n)$-matrisen B hvis vi har

$$\lim_{N\to\infty} \Big((A_0)_{ij} + (A_1)_{ij} + (A_2)_{ij} + \cdots + (A_N)_{ij} \Big) = B_{ij}$$

for alle $1 \leq i, j \leq n$. Vi skriver i så fall

$$A_0 + A_1 + A_2 + \cdots = B$$

La I være identitetsmatrisen av størrelse $(n \times n)$. I oppgave 1.3.3 blir du bedt om å vise at hvis A er en matrise med operatornorm $|A| < 1$, så er matrisen $I - A$ inverterbar, og vi har

$$(I - A)^{-1} = I + A + A^2 + A^3 + \cdots \tag{1}$$

Merk analogien med teorem 10.1.1 i bind I om summen av en vanlig geometrisk rekke.

TEOREM I

Banachs lemma

La A og B være $(n \times n)$-matriser, der B er inverterbar. Hvis vi har $|B - A| < |B^{-1}|^{-1}$, så er A også inverterbar. Videre har vi

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

BEVIS Ved oppgave 1.3.2 i første ulikhet og antakelsen

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

i andre ulikhet, får vi

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

Dermed vet vi at matrisen

$$I - B^{-1}(B - A)$$

er inverterbar. Innsetting i (1) gir

$$\begin{aligned} |(I - B^{-1}(B - A))^{-1}| &= \left| I + B^{-1}(B - A) + \left[B^{-1}(B - A) \right]^2 + \cdots \right| \\ &\leq 1 + |B^{-1}||B - A| + \left[|B^{-1}||B - A| \right]^2 + \cdots \\ &= \frac{1}{1 - |B^{-1}||B - A|} \end{aligned}$$

Her brukte vi oppgave 1.3.2 ved andre ulikhet og formelen for sum av en vanlig geometrisk rekke (teorem 10.1.1, bind I) ved siste likhet. Ved antakelsen om at B^{-1} fins, kan vi skrive

$$A = B - (B - A) = B(I - B^{-1}(B - A))$$

Altså

$$A^{-1} = (I - B^{-1}(B - A))^{-1} B^{-1},$$

der vi brukte formelen

$$(MN)^{-1} = N^{-1}M^{-1}$$

fra oppgave 3.6.5. Bruk av oppgave 1.3.2 enda en gang gir da

$$|A^{-1}| \leq |(I - B^{-1}(B - A))^{-1}| \cdot |B^{-1}|$$

Kombineres dette med ulikheten for $|(I-B^{-1}(B-A))^{-1}|$ som vi fant ovenfor, faller teoremet rett ut. ■

1.3 OPPGAVER

1. La A være en $(n \times n)$-matrise. Vis at for alle $\mathbf{x} \in \mathbb{R}^n$ gjelder

$$|A\mathbf{x}| \leq ||A|| \cdot |\mathbf{x}|$$

(Hint: Bruk at hver komponent i vektoren $A\mathbf{x}$ er skalarproduktet av $\mathbf{x}$ med en linjevektor i matrisen A.) Vis så at $|A| \leq ||A||$.

2. La A og B være $(n \times n)$-matriser.

a) Vis at hvis $\mathbf{x} \in \mathbb{R}^n$, så er $|(AB)\mathbf{x}| \leq |A| \cdot |B| \cdot |\mathbf{x}|$.

b) Vis at hvis $\mathbf{x} \in \mathbb{R}^n$, så er $|(A + B)\mathbf{x}| \leq (|A| + |B|) \cdot |\mathbf{x}|$.

c) Vis at $|AB| \leq |A| \cdot |B|$ og $|A + B| \leq |A| + |B|$.

3. La A være en kvadratisk matrise slik at $|A| < 1$, og la I være identitetsmatrisen. For $n \geq 1$, la $S_n = I + A + A^2 + \cdots + A^n$.

a) Vis at $(I - A)S_n = I - A^{n+1}$.

b) Vis at hvis $m > n$, har vi $|(S_m)_{ij} - (S_n)_{ij}| \leq \dfrac{|A|^{n+1}}{1 - |A|}$.

c) Vis at for hver kombinasjon ij er følgen $\{(S_m)_{ij}\}_{n=1}^{\infty}$ en Cauchy-følge. Bruk dette til å vise at det fins en matrise S slik at $\lim_{n\to\infty} S_n = S$.

d) Ved å la $n \to \infty$ i punkt a), vis at $(I - A)S = I$.

e) Vis at matrisen $I - A$ er inverterbar, og at

$$(I - A)^{-1} = I + A + A^2 + A^3 + \cdots$$

4. I denne oppgaven skal vi utlede en m-dimensjonal versjon av resultatet i oppgave 3.3.2. Hvis du ikke har gjort den, kan det være lurt å gjøre den før du starter på denne. La $U \subseteq \mathbb{R}^m$ være åpen og konveks, og la $\mathbf{F} : U \to \mathbb{R}^m$ være en deriverbar funksjon slik at

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

for alle $\mathbf{x}, \mathbf{y} \in U$, der M er et konstant tall. De verikale strekene på venstre side er *operatornormen* til en matrise, se definisjon 2.7.1.

a) La $\mathbf{r}(t) = \mathbf{x} + t(\mathbf{y} - \mathbf{x})$ og $\mathbf{G}(t) = \mathbf{F}(\mathbf{r}(t))$. Vis at

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

og bruk dette til å vise at

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

b) Vis at når vi definerer integralet komponentvis, er

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

c) Vis at $\int_0^1 \mathbf{G}'(t)dt$ kan skrives

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

d) Vis at $|\mathbf{F}(\mathbf{y}) - \mathbf{F}(\mathbf{x}) - \mathbf{F}'(\mathbf{x})(\mathbf{y} - \mathbf{x})| = I$, der

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

e) Vis at

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

der $|\mathbf{F}'(\mathbf{r}(t)) - \mathbf{F}'(\mathbf{x})|$ er operatornormen til $\mathbf{F}'(\mathbf{r}(t)) - \mathbf{F}'(\mathbf{x})$.

f) Vis at

$$I \leq M \int_0^1 |\mathbf{r}(t) - \mathbf{x}| \cdot |\mathbf{y} - \mathbf{x}|\ dt$$

g) Forklar hvorfor $\mathbf{r}(t) - \mathbf{x} = t(\mathbf{y} - \mathbf{x})$, og bruk dette til å vise at

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

h) Vis at vi for alle $\mathbf{x}, \mathbf{y} \in U$ har

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

1.4 Normale transformasjoner

Alle eksterne referanser i denne seksjonen er til bind II.

Fra teorem 7.4.2 vet vi at symmetriske, reelle matriser har en ortonormal egenbasis. Dette betyr at alle reelle, symmetriske matriser er ortogonalt diagonaliserbare.

I det komplekse tilfellet er hermitiske transformasjoner/matriser en naturlig analogi til reelle, symmetriske transformasjoner/matriser. Videre er unitære transformasjoner/matriser en naturlig analogi til ortogonale, reelle transformasjoner/matriser. Fra teorem 8.3.9 vet vi at alle hermitiske transformasjoner/matriser er unitært diagonaliserbare, noe som klaffer med disse analogiene.

Likevel er det en forskjell. Hvis A er en reell matrise som er ortogonalt diagonaliserbar, har vi

$$A = PDP^{\mathrm{T}},$$

der P er en ortogonal matrise. Dette gir

$$A^{\mathrm{T}} = (PDP^{\mathrm{T}})^{\mathrm{T}} = (P^{\mathrm{T}})^{\mathrm{T}} D^{\mathrm{T}} P^{\mathrm{T}} = PDP^{\mathrm{T}} = A,$$

så A er nødvendigvis symmetrisk. Med andre ord har vi følgende teorem:

TEOREM 1

Ortogonal diagonaliserbarhet for reelle matriser

En reell kvadratisk matrise er ortogonalt diagonaliserbar hvis og bare hvis den er symmetrisk.

En analogi til denne ekvivalensen mangler vi for hermitiske matriser. Selv om alle hermitiske matriser er unitært diagonaliserbare, er de ikke de *eneste* komplekse matrisene med denne egenskapen. Det viser seg at det fins matriser som er unitært diagonaliserbare og ikke hermitiske.

For å få en naturlig avrunding av den komplekse teorien vår skal vi derfor innføre nye begreper. Vi skal definere begrepet **normal** lineærtransformasjon og det tilhørende begrepet **normal matrise**. Klassen av normale transformasjoner/matriser inneholder hermitiske transformasjoner/matriser som spesialtilfeller. Hovedresultatet i denne seksjonen er teorem 1.4.6, som sier at en transformasjon har en ortonormal egenbasis hvis og bare hvis den er normal. Dette teoremet viser dermed også teorem 8.3.9 fra forrige seksjon, som vi utsatte beviset for.

Men før vi kommer til teorem 1.4.6, må vi utvikle litt mer teori. Først skal vi beskrive den transformasjonen som tilsvarer den adjungerte av en gitt matrise. Vi begrenser oss i hele denne seksjonen til endeligdimensjonale vektorrrom.

TEOREM 2

Den adjungerte av en transformasjon

La $T : V \to V$ være en lineærtransformasjon av et endeligdimensjonalt komplekst indreproduktrom inn i seg selv. Da fins det en entydig bestemt lineærtransformasjon $T^* : V \to V$, kalt den **adjungerte** til T, som oppfyller

$$\langle \mathbf{u}, T^*(\mathbf{v}) \rangle = \langle T(\mathbf{u}), \mathbf{v} \rangle \qquad \text{for alle } \mathbf{u} \text{ og } \mathbf{v} \text{ i } V.$$

Hvis B er en ortonormal basis for V, så er matrisen $[T^*]_B$ til T^* i basisen B den adjungerte av matrisen til T, altså

$$[T^*]_B = ([T]_B)^*$$

BEVIS Siden V er endeligdimensjonalt, vet vi at V har en ortonormal basis B. La $T^* : V \to V$ være transformasjonen definert ved at

$$[T^*]_B = ([T]_B)^*$$

Hvis $\mathbf{u}$ og $\mathbf{v}$ er elementer i V, så er

$$\langle \mathbf{u}, \mathbf{v} \rangle = [\mathbf{u}]_B^* [\mathbf{v}]_B$$

ved teorem 7.1.3. Dette gir

$$\begin{aligned} \langle T(\mathbf{u}), \mathbf{v} \rangle &= [T(\mathbf{u})]_B^* [\mathbf{v}]_B = ([T]_B [\mathbf{u}]_B)^* [\mathbf{v}]_B \\ &= [\mathbf{u}]_B^* ([T]_B)^* [\mathbf{v}]_B \\ &= [\mathbf{u}]_B^* [T^*(\mathbf{v})]_B = \langle \mathbf{u}, T^*(\mathbf{v}) \rangle \end{aligned}$$

Dermed har vi vist at det eksisterer en lineærtransformasjon med egenskapen beskrevet i teoremet. For å vise unikhet, anta at $S : V \to V$ er en annen lineærtransformasjon som i likhet med T^* oppfyller

$$\langle \mathbf{u}, S(\mathbf{v}) \rangle = \langle T(\mathbf{u}), \mathbf{v} \rangle \qquad \text{for alle } \mathbf{u} \text{ og } \mathbf{v} \text{ i } V.$$

Da er

$$\langle \mathbf{u}, S(\mathbf{v}) \rangle = \langle T(\mathbf{u}), \mathbf{v} \rangle = \langle \mathbf{u}, T^*(\mathbf{v}) \rangle$$

for alle $\mathbf{u}$ og $\mathbf{v}$ i V. Ved teorem 7.1.2 betyr dette spesielt at vektorene $S(\mathbf{v})$ og $T^*(\mathbf{v})$ vil ha samme komponenter i en ortonomal basis for V. Da er de like. Siden $\mathbf{v} \in V$ var vilkårlig, har vi vist at $S = T^*$. ■

Transformasjonen T^*, som forrige teorem gir oss eksistensen av, kalles altså for den **adjungerte** transformasjonen til T. Merk at siden

$$(M^*)^* = M$$

for matriser, følger det fra dette teoremet at

$$(T^*)^* = T$$

for alle transformasjoner T. Det går også an å vise dette direkte fra betingelsen

$$\langle \mathbf{u}, T^*(\mathbf{v}) \rangle = \langle T(\mathbf{u}), \mathbf{v} \rangle$$

uten å trekke inn matriser.

Neste teorem viser at klassene av hermitiske og unitære lineærtransformasjoner begge kan beskrives ved enkle kriterier knyttet til den adjungerte av transformasjonen.

TEOREM 3

Hermitiske, unitære og adjungerte transformasjoner

La $T : V \to V$ være en lineærtransformasjon på et endeligdimensjonalt komplekst indreproduktrom V. Da gjelder

(1) T er hermitisk hvis og bare hvis $T = T^*$.

(2) T er unitær hvis og bare hvis $T^{-1} = T^*$.

BEVIS (1) Dette følger fra teorem 1.4.2 kombinert med definisjonen 8.3.1 av en hermitisk transformasjon.

(2) Dette følger fra teorem 8.3.6 kombinert med teorem 8.3.7. ■

Vi skal nå definere en klasse transformasjoner som inkluderer både hermitiske og unitære transformasjoner som spesialtilfeller.

DEFINISJON 1

Normale transformasjoner og matriser

En lineærtransformasjon $T : V \to V$ på et endeligdimensjonalt komplekst indreproduktrom kalles **normal** hvis

$$T^* \circ T = T \circ T^*,$$

altså hvis $T^*(T(\mathbf{v})) = T(T^*(\mathbf{v}))$ for alle $\mathbf{v}$ i V. En $(n \times n)$-matrise M kalles **normal** hvis

$$M^* M = M M^*$$

Nå har vi følgende sammenhenger:

TEOREM 4

Betingelser som sikrer normalitet

La $T : V \to V$ være en lineærtransformasjon på et endeligdimensjonalt komplekst indreproduktrom V, og la B være en ortonormal basis for V. Da gjelder:

(1) Hvis T er hermitisk, så er T normal.

(2) Hvis T er unitær, så er T normal.

(3) T er normal hvis og bare hvis $[T]_B$ er normal.

BEVIS (1) Hvis T er hermitisk, så gir teorem 1.4.3 at

$$T^* \circ T = T \circ T = T \circ T^*$$

(2) Hvis T er unitær, så gir teorem 1.4.3 at

$$\begin{aligned} T^* \circ T &= T^{-1} \circ T \\ &= I = T \circ T^{-1} = T \circ T^* \end{aligned}$$

(3) Ved teorem 1.4.2 i fjerde overgang fås

$$\begin{aligned} T \text{ er normal} &\iff T^* \circ T = T \circ T^* \\ &\iff [T^* \circ T]_B = [T \circ T^*]_B \\ &\iff [T^*]_B[T]_B = [T]_B[T^*]_B \\ &\iff ([T]_B)^*[T]_B = [T]_B([T]_B)^* \\ &\iff [T]_B \text{ er normal.} \quad \blacksquare \end{aligned}$$

TEOREM 5

Egenvektorer for normale transformasjoner

La $T : V \to V$ være en normal transformasjon på et n-dimensjonalt komplekst indreproduktrom V, og anta at $\mathbf{v}$ er en egenvektor for T med tilhørende egenverdi λ. Da gjelder:

(1) Vektoren $\overline{\mathbf{v}}$ er en egenvektor for T^* med egenverdi $\overline{\lambda}$.

(2) Mengden

$$U = \{\mathbf{u} \in V \mid \langle \mathbf{u}, \mathbf{v} \rangle = 0\}$$

er et underrom av V med dimensjon $n - 1$. Hvis $\mathbf{u} \in U$, er $T(\mathbf{u}) \in U$ og $T^*(\mathbf{u}) \in U$ også.

BEVIS (1) Vi har

$$\begin{aligned}\|T(\mathbf{v})-\lambda\mathbf{v}\|^2 &= \langle T(\mathbf{v})-\lambda\mathbf{v}, T(\mathbf{v})-\lambda\mathbf{v}\rangle \\ &= \langle T(\mathbf{v}), T(\mathbf{v})\rangle + \overline{\lambda}\lambda\langle\mathbf{v},\mathbf{v}\rangle - \lambda\langle T(\mathbf{v}),\mathbf{v}\rangle - \overline{\lambda}\langle\mathbf{v}, T(\mathbf{v})\rangle,\end{aligned}$$

og på samme måte

$$\begin{aligned}\|T^*(\mathbf{v})-\overline{\lambda}\mathbf{v}\|^2 &= \langle T^*(\mathbf{v})-\lambda\mathbf{v}, T^*(\mathbf{v})-\lambda\mathbf{v}\rangle \\ &= \langle T^*(\mathbf{v}), T^*(\mathbf{v})\rangle + \overline{\lambda}\lambda\langle\mathbf{v},\mathbf{v}\rangle - \overline{\lambda}\langle T^*(\mathbf{v}),\mathbf{v}\rangle - \lambda\langle\mathbf{v}, T^*(\mathbf{v})\rangle\end{aligned}$$

Teorem 1.4.2 gir for alle $\mathbf{w}\in V$ relasjonene

$$\begin{aligned}\langle T(\mathbf{v}),\mathbf{w}\rangle &= \langle\mathbf{v}, T^*(\mathbf{w})\rangle \\ \langle\mathbf{v}, T(\mathbf{w})\rangle &= \langle T^*(\mathbf{v}),\mathbf{w}\rangle\end{aligned}$$

Bruk av disse samt normalitet av T gir nå, med $\mathbf{w}=T(\mathbf{v})$ i første overgang:

$$\begin{aligned}\langle T(\mathbf{v}), T(\mathbf{v})\rangle &= \langle\mathbf{v}, T^*(T(\mathbf{v}))\rangle \\ &= \langle\mathbf{v}, T(T^*(\mathbf{v}))\rangle \\ &= \langle T^*(\mathbf{v}), T^*(\mathbf{v})\rangle,\end{aligned}$$

der vi i siste overgang brukte $\langle\mathbf{v}, T(\mathbf{w})\rangle = \langle T^*(\mathbf{v}),\mathbf{w}\rangle$ med $\mathbf{w}=T^*(\mathbf{v})$. Setter du disse tre resultatene inn i de uttrykkene vi fant rett ovenfor, får du at

$$\|T(\mathbf{v})-\lambda\mathbf{v}\|^2 = \|T^*(\mathbf{v})-\overline{\lambda}\mathbf{v}\|^2$$

Det følger direkte av dette at $T(\mathbf{v})-\lambda\mathbf{v}=\mathbf{0}$ hvis og bare hvis $T^*(\mathbf{v})-\overline{\lambda}\mathbf{v}=\mathbf{0}$, så (1) er bevist.

(2) La S være underrommet utspent av kun $\mathbf{v}$. Dette rommet er 1-dimensjonalt, og U er det ortogonale komplementet $S^\perp$ til S. Det følger nå fra teorem 7.3.1 i kompleks tolkning at U er et underrom av V, og at dimensjonen til U er $n-1$. (Se oppgave 7.3.10). Hvis $\mathbf{u}\in U$, får vi

$$\begin{aligned}\langle\mathbf{v}, T(\mathbf{u})\rangle &= \overline{\langle T(\mathbf{u}),\mathbf{v}\rangle} \\ &= \overline{\langle\mathbf{u}, T^*(\mathbf{v})\rangle} \\ &= \overline{\langle\mathbf{u}, \overline{\lambda}\mathbf{v}\rangle} && \text{(ved punkt 1)} \\ &= \langle\overline{\lambda}\mathbf{v},\mathbf{u}\rangle = \lambda\langle\mathbf{v},\mathbf{u}\rangle = 0,\end{aligned}$$

siste overgang siden $\langle\mathbf{v},\mathbf{u}\rangle=0$. Så $T(\mathbf{u})\in U$. Siden $\langle T^*(\mathbf{u}),\mathbf{v}\rangle = \overline{\langle\mathbf{v}, T(\mathbf{u})\rangle} = 0$, følger også at $T^*(\mathbf{u})\in U$. ■

TEOREM 6

Normale matriser og ortonormal egenbasis

La $T : V \to V$ være en lineærtransformasjon på et n-dimensjonalt komplekst indreproduktrom V. Da har T en ortonormal egenbasis hvis og bare hvis T er normal.

BEVIS Anta at T har en ortonormal egenbasis B. Da er matrisen $[T]_B$ en diagonal matrise, og alle diagonale $(n \times n)$-matriser er normale.

For å vise den motsatte implikasjonen skal vi bruke induksjon på dimensjonen n av vektorrommet V. For $n = 1$ er påstanden triviell, fordi da vil enhver vektor $\mathbf{v} \neq \mathbf{0}$ være en egenvektor for T.

Anta nå at implikasjonen er vist for for $n = k$. La $T : V \to V$ være en normal transformasjon på et vektorrom av dimensjon $n = k + 1$. Ved teorem 8.1.1 har T en egenverdi λ og en tilhørende egenvektor $\mathbf{v}$. Siden vi kan dele $\mathbf{v}$ på dens lengde, kan vi anta at $\|\mathbf{v}\| = 1$. La U være det ortogonale komplementet til underrommet utspent av $\mathbf{v}$. Ved teorem 1.4.5 er da dimensjonen til U lik k, og hvis $\mathbf{u} \in U$, er $T(\mathbf{u}) \in U$. Det siste medfører at vi kan definere en lineærtransformasjon $S : U \to U$ ved å sette $S(\mathbf{u}) = T(\mathbf{u})$ for alle $\mathbf{u} \in U$. Vi lar indreproduktet på U være det samme som vi har når U betraktes som et underrom av V. Ved teorem 1.4.2 har vi da

$$\langle S(\mathbf{u}), \mathbf{v}\rangle = \langle T(\mathbf{u}), \mathbf{v}\rangle = \langle \mathbf{u}, T^*(\mathbf{v})\rangle$$

for alle $\mathbf{u}$ og $\mathbf{v}$ i U, og det følger ved unikhetsegenskapen i nevnte teorem at $S^*(\mathbf{u}) = T^*(\mathbf{u})$ for alle $\mathbf{u} \in U$. Siden $T = T^*$, er dermed $S = S^*$, så S er normal. Ved induksjonshypotesen fins derfor en ortonormal basis $\{\mathbf{b}_1, \ldots, \mathbf{b}_k\}$ for U bestående av egenvektorer for S. Men siden $\mathbf{v}$ har lengde 1 og står vinkelrett på alle vektorene $\mathbf{b}_i$, blir da samlingen

$$\{\mathbf{v}, \mathbf{b}_1, \ldots, \mathbf{b}_k\}$$

en ortonormal basis for V bestående av egenvektorer for T. ■

1.4 OPPGAVER

1. Vis at hvis M er en normal matrise, så er også matrisen M^2 normal.

2. En kvadratisk, kompleks matrise A kalles *skjev-hermitisk* hvis $A^* = -A$.

a) Vis at matrisen

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

er skjev-hermitisk. Finn en ortonormal egenbasis for A. Angi også en unitær matrise U som diagonaliserer A.

b) Vis at hvis A er en skjev-hermitisk matrise, så er A normal.

c) Vis at alle skjev-hermitiske matriser M er unitært diagonaliserbare.

d) Vis at alle egenverdier for en skjev-hermitisk matrise er rent imaginære.

3. Vis at en normal matrise er hermitisk hvis og bare hvis alle egenverdiene dens er reelle.

KAPITTEL 2

DIVERSE

2.1 Kvantemekanikk

Partikler og bølger i klassisk fysikk

I klassisk mekanikk (fra Newtons tid og fremover) har man formelen

$$E = \frac{1}{2}mv^2 \tag{1}$$

for den **kinetiske energien** (bevegelsesenergien) til objekt med masse m (kg) og fart v (m/s) i et gitt referansesystem/koordinatsystem. Denne energien gir mening å diskutere i forbindelse med bevaring av *total mekanisk energi*, for eksempel når en kule slippes fra høyde h og faller fritt nedover med tyngdens akselerasjon g. Se eksempel 7.1.2 i bind I. Når kulen har nådd høyde 0, har dens **potensielle energi** mgh gått over til kinetisk energi, så vi har $mgh = (1/2)mv^2$. Dette gir fart $v = \sqrt{2gh}$, som stemmer med det man får dersom man setter $(1/2)gt^2 = h$, løser med hensyn på t og setter inn i formelen $v(t) = s'(t) = gt$ fra eksempel 7.1.2.

På den annen side regner man i klassisk mekanikk med at et objekt med masse m og fart v har en **bevegelsesmengde** (impuls) p gitt ved

$$p = mv \tag{2}$$

I perfekte (elastiske) kollisjoner mellom objekter ser man at dersom v regnes med fortegn, er summen av bevegelsesmengdene for de involverte objektene bevart. Kombinasjon av (1) og (2) gir at vi i klassisk mekanikk har sammenhengen

$$E = \frac{p^2}{2m} \tag{3}$$

mellom kinetisk energi E og bevegelsesmengde p.

Bølger er i klassisk fysikk noe helt annet enn **materie**, altså materielle objekter som kuler og partikler. En bølge med bølgelengde λ og frekvens f kan representere noe som forplanter seg. Den kan modelleres ved en trigonometrisk funksjon, for eksempel

$$\cos\left[2\pi\left(\frac{x}{\lambda} - f \cdot t\right)\right] \tag{4}$$

Her er x posisjon (meter), t er tid (sekunder) og f er frekvensen, altså antall bølgelengder som passerer et gitt punkt x i løpet av ett sekund. Eksempler på bølger behandlet i klassisk fysikk er lysbølger, lydbølger og vannbølger.

Empirisk grunnlag for kvantemekanikken

Rundt år 1900 hadde ulike eksperimenter tydelig vist at dersom man zoomer ned til en lengde mindre enn omlag 10^{-10} meter, så er ikke lenger de materielle objektene vi har rundt oss "kontinuerlige". I stedet så man at de består av byggeklosser man etterhvert kalte **atomer**, i sin tur bygd opp av enda mindre **partikler** kalt elektroner, protoner, nøytroner og så videre.

Imidlertid var det mange uløste problemer. Et av disse var knyttet til den **fotoelektriske effekten**, nemlig at lys kan slå elektroner løs fra metalloverflater. Eksperimenter viste at farten til de frigjorte elektronene kun avhang av lysets frekvens, og at lys med frekvens under en viss grense ikke slo løs *noen* elektroner. Som forklaring på dette fremsatte Albert Einstein i 1905 hypotesen om at lyset består av små "lyskvanter" kalt **fotoner**, der hvert foton har energi

$$E = hf \tag{5}$$

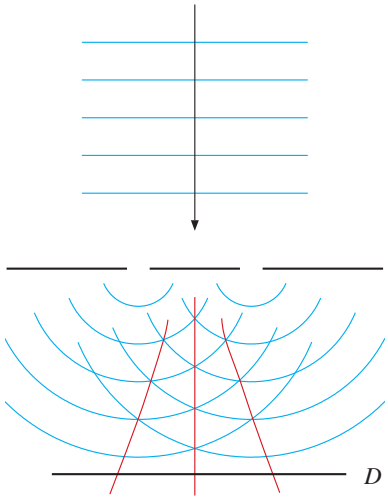
Her er f er frekvensen til lyset, og h er en konstant kalt **Plancks konstant** etter fysikeren Max Planck. Forklaringen på at lavfrekvent lys ikke klarer å slå løs elektroner, var da at slikt lys representerer en strøm av "slag" som alle er for svake til å slå løs et elektron. Målinger viste at $h \approx 6.63 \cdot 10^{-34}$ Js, der J (Joule) er enheten for energi og s er sekund. Senere viste eksperimenter at hvert foton i lys med bølgelengde λ også har *bevegelsesmengde*

$$p = h/\lambda \tag{6}$$

i kollisjoner med elektroner. Dette viste at lys har en **partikkel/bølgedualitet**: Noen ganger oppfører det seg som *bølger*, f.eks. ved at det gir interferensmønstre som bølger på vann. I andre sammenhenger oppfører det seg som en strøm av *partikler* med energi $E = hf$ og bevegelsesmengde $p = h/\lambda$.

Kort etter oppdaget man at dette også gjaldt omvendt vei: Partikler med masse, som for eksempel elektroner, viste seg å ha bølgeegenskaper! Eksperimenter viste at en strøm av elektroner med bevegelsesmengde p produserte **interferensmønstre** tilsvarende en bølge med bølgelengde

$$\lambda = h/p$$



Figur 2.1.1 Prinsippskisse for dobbelspalteksperiment med elektroner. Elektroner med bevegelsesmengde p sendes inn mot spalten ovenfra på figuren, og interfererer med seg selv bak dobbelspalten. Langs de røde kurvene er det konstruktiv interferens. Der disse treffer detektorskjermen D, blir det mange svarte prikker fra elektrontreff. Nær midten mellom punktene der de røde kurvene treffer skjermen er det destruktiv interferens og nesten ingen treff.

og frekvens $f = E/h$, der vi har $E = p^2/2m$ med m som elektronmassen. Videre viste eksperimenter med ekstremt lav intensitet at elektronene ikke interfererer med *hverandre*, i istedet interfererer hvert elektron *med seg selv*! Dette viser at når elektroner med kjent bevegelsesmengde p går gjennom noe som produserer et interferensmønster, for eksempel en dobbelspalte, så "er" ikke elektronet noe bestemt sted, det er en utstrakt *bølge*. Først når elektronene senere treffer detektoren bak spalten, blir deres posisjon bestemt. Da blir hvert elektron en partikkel og produserer en prikk på skjermen. Etterhvert som det kommer flere elektroner, vil prikkene bygge opp et interferensmønster der det er striper med mange treff og få treff vekselvis. Se figur 2.1.1.

Schrödingerligningen

Innsetting av $\lambda = h/p$ and $f = E/h$ i (4) gir bølgefunksjonen

$$\cos\frac{px - Et}{\hbar}$$

der $\hbar = h/2\pi$. Hvis vi regner komplekst, tilsvarer dette

$$\Psi(x,t) = e^{i(px-Et)/\hbar} = e^{ipx/\hbar}e^{-iEt/\hbar} \tag{7}$$

Denne bølgefunksjonen tenker vi oss er en "materiebølge" som representerer en partikkel (f.eks. et elektron) med bevegelsesmengde p og masse m, der vi fortsatt har sammenhengen $E = p^2/2m$ fra (3). Derivasjon av (7) gir

$$\frac{\partial\Psi}{\partial t} = -\frac{iE}{\hbar}\Psi \quad \text{og} \quad \frac{\partial^2\Psi}{\partial x^2} = \left(\frac{ip}{\hbar}\right)^2\Psi$$

Ved å løse disse to ligningene med hensyn på Ψ og sette uttrykkene lik hverandre, får vi

$$-\frac{\hbar}{iE}\frac{\partial\Psi}{\partial t} = \frac{\hbar^2}{i^2p^2}\frac{\partial^2\Psi}{\partial x^2}$$

Med $E = p^2/2m$ innsatt blir dette

$$-i\hbar\frac{2m}{p^2}\frac{\partial\Psi}{\partial t} = -\frac{\hbar^2}{p^2}\frac{\partial^2\Psi}{\partial x^2}$$

Multiplikasjon med p^2 og divisjon med $2m$ gir

$$i\hbar\frac{\partial\Psi}{\partial t} = -\frac{\hbar^2}{2m}\frac{\partial^2\Psi}{\partial x^2}$$

Denne ligningen kalles **schrödingerligningen** for en fri partikkel, oppkalt etter fysikeren Erwin Schrödinger (1887-1961).

Men hva er det som "bølger"? Dette er et stort filosofisk problem, og egentlig er dette problemet forsatt ikke løst den dag i dag. Imidlertid foreslo Max Born i 1927 en tolkning som regneteknisk sett fungerer, og som gir mening til de eksperimentelle resultatene. Han foreslo at

$$|\Psi(x,t)|^2$$

tolkes som **sannsynlighetstettheten** for å finne partikkelen i punktet x ved tid t hvix vi måler posisjonen til partikkelen, for eksempel ved en detektor bak en dobbelspalte, som nevnt ovenfor. Hvis dette skal fungere, må vi ha

$$\int_{-\infty}^{\infty} |\Psi(x,t)|^2 dx = 1 \qquad \text{ved alle tidspunkter } t \tag{8}$$

fordi sannsynligheten integrert over alle mulige posisjoner x må være 1. Bølgen (7) gir på sin side

$$|\Psi(x,t)|^2 = |e^{i(px-Et)/\hbar}|^2 = 1^2 = 1$$

for alle (x,t), så den tilfredsstiller ikke dette. Det hjelper ikke å multiplisere (7) med en normaliseringskonstant, integralet divergerer uansett. Den rene, monokromatiske bølgen (7) representerer altså et "ufysisk" grensetilfelle der posisjonen til partikkelen er helt uskarp, altså helt ubestemt. I praksis vet man alltid *noe* om hvor en partikkel befinner seg. For å beskrive en slik delvis lokalisert partikkel ved et valgt tidspunkt, kan vi bruke en kombinasjon av monokromatiske bølger $e^{ipx/\hbar}$ for ulike verdier av p, altså et integral

$$\psi(x) = \frac{1}{\sqrt{2\pi\hbar}} \int_{-\infty}^{\infty} \phi(p) e^{ipx/\hbar}\, dp \tag{9}$$

Her kommer teorien for fouriertransform inn i bildet. Normaliseringsfaktoren $1/\sqrt{2\pi\hbar}$ gir pen regning, som du øyeblikkelig skal få se. Hvis vi substituerer

$$\omega = \frac{2\pi}{\lambda} = \frac{2\pi p}{h} = \frac{p}{\hbar}$$

i integralet ovenfor, får vi $dp = \hbar\, d\omega$ og

$$\psi(x) = \frac{1}{\sqrt{2\pi}\sqrt{\hbar}} \cdot \hbar \int_{-\infty}^{\infty} \phi(\hbar\omega) e^{i\omega x}\, d\omega$$

Dette ser vi fra fourierteorien at blir korrekt hvis

$$\phi(\hbar\omega) = \frac{1}{\sqrt{\hbar}} \mathcal{F}\{\psi(x)\}(\omega)$$

Med definisjonen av $\mathcal{F}\{\psi(x)\}$ innsatt gir dette

$$\phi(\hbar\omega) = \frac{1}{\sqrt{\hbar}} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \psi(x) e^{-i\omega x}\, d\omega$$

Innsetting av $\omega = p/\hbar$ gir så

$$\phi(p) = \frac{1}{\sqrt{2\pi\hbar}} \int_{-\infty}^{\infty} \psi(x) e^{-ipx/\hbar}\, dp \tag{10}$$

Likningene (9) og (10) viser at $\psi(x)$ og $\phi(p)$ er et fouriertransformpar, skalert med den felles faktoren $1/\sqrt{\hbar}$.

Bølgepakker

Schrödingerligningen er en partiell differensialligning. Hvis vi kjenner starttilstanden $\Psi(x, 0)$, kan vi finne $\Psi(x, t)$ for $t > 0$ ved å løse ligningen. Man kan vise at dersom $\Psi(x, 0)$ oppfyller normaliseringskravet (8) vil $\Psi(x, t)$ også gjøre det, for alle $t > 0$.

For å se hvordan dette fungerer, la oss ta et konkret eksempel. Den monokromatiske bølgen (7) med en valgt bevegelsesmengde $p = p_0$ gir

$$\Psi(x, 0) = e^{ip_0 x/\hbar}$$

Dette er ikke en fysisk realiserbar initialtilstand, fordi (8) ikke er oppfylt. En initialtilstand $\psi(x)$ som oppfyller dette kan vi derimot få dersom vi multipliserer med en dempende gaussfunksjon for passende $\sigma > 0$, slik:

$$\psi(x) = \frac{1}{(2\pi\sigma^2)^{1/4}} e^{-x^2/4\sigma^2} \cdot e^{ip_0 x/\hbar} \tag{11}$$

Substitusjonen $u = x\sigma$ og oppgave 4.8.22 gir nemlig

$$\int_{-\infty}^{\infty} |\psi(x)|^2 dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(x/\sigma)^2} dx = \frac{1}{\sigma\sqrt{2\pi}} \cdot \sigma \cdot \sqrt{2\pi} = 1$$

Vi antar nå at $\psi(x)$ gitt ved (11) er initialtilstanden vår, og vi ønsker å finne en løsning $\Psi(x, t)$ av schrödingerligningen slik at

$$\Psi(x, 0) = \psi(x)$$

Dette blir analogt til sitasjonen i seksjon 9.10. Fra (7) vet vi at

$$\Psi(x, t) = e^{i(px - Et)/\hbar} = e^{i[px - (p^2/2m)t]/\hbar}$$

løser schrödingerligningen for hver konstant verdi av p. Fra dette får vi, analogt med seksjon 9.10, at den "uendelige lineærkombinasjonen" av slike løsninger gitt ved

$$\Psi(x, t) = \frac{1}{\sqrt{2\pi\hbar}} \int_{-\infty}^{\infty} \phi(p) e^{i[px - (p^2/2m)t]/\hbar}\, dp \tag{12}$$

også løser schrödingerligningen. I kvantemekanikk kalles dette en **bølgepakke**. Per konstruksjon oppfyller bølgepakken $\Psi(x, 0) = \psi(x)$, se (9).

Innsetting av (11) i (10) gir (se oppgave 2.1.1)

$$\phi(p) = \left(\frac{2\sigma^2}{\pi\hbar^2}\right)^{1/4} e^{(p-p_0)^2\sigma^2/\hbar^2} \tag{13}$$

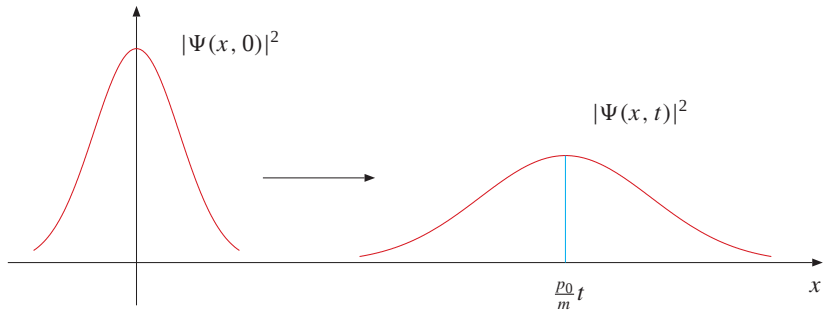
Innsetting av (13) i (12), beregning av $\Psi(x,t)$ og deretter utregning av $|\Psi(x,t)|^2$ gir (se oppgave 2.1.1)

$$|\Psi(x,t)|^2 = \frac{1}{\sqrt{2\pi\beta(t)}}\, e^{-[x-(p_0/m)t]^2/2\beta(t)} \tag{14}$$

der

$$\beta(t) = \sigma^2 + \frac{\hbar^2 t^2}{4m^2\sigma^2}$$

kalles **variansen** til $|\Psi(x,t)|^2$. Merk analogien med eksempel 9.9.1 i bind II. Vi ser at sannsynlighetsfordelingen $|\Psi(x,t)|^2$ har sin topp for $x = (p_0/m)t$, slik at den mest sannsynlige posisjonen for partikkelen beveger seg med farten p_0/m. Dette stemmer med formelen (2) for sammenhengen mellom bevegelsesmengde og fart. Vi ser at også $\beta(0) = \sigma^2$, og at $\beta(t)$ øker med t. Fordelingen $|\Psi(x,t)|^2$ blir dermed lavere og flatere med tiden. Se figur 2.1.2.



Figur 2.1.2 Tidsutvikling for bølgepakken fra $t = 0$ til et tidspunkt $t > 0$.

Heisenbergs uskarphetstrelasjon

Fra (11) får vi

$$|\psi(x)|^2 = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-x^2/2\sigma^2}$$

I statistikk og sannynslighetsteori kalles denne funksjonen **normalfordelingen** med standardavvik $\sigma > 0$, varians σ^2 og sentrum 0. Fra (13) får vi

$$|\phi(p)|^2 = \phi(p) = \left(\frac{2\sigma^2}{\pi\hbar^2}\right)^{1/2} e^{2(p-p_0)^2\sigma^2/\hbar^2}$$

Hvis vi setter $\sigma_p = \hbar/2\sigma$ kan dette skrives

$$|\phi(p)|^2 = \frac{1}{\sqrt{2\pi\sigma_p^2}}\, e^{-(p-p_0)^2/2\sigma_p^2}$$

Dette betyr at $|\phi(p)|^2$ også er en normalfordeling. Denne har sentrum p_0 og standardavvik

$$\sigma_p = \frac{\hbar}{2\sigma}$$

Hvis vi setter $\Delta x = \sigma$ og $\Delta p = \sigma_P$, har vi

$$\Delta x \cdot \Delta p = \hbar/2 \tag{15}$$

Så hvis standardavviket Δx til posisjonsfordelingen $|\psi(x)|^2$ blir større, blir standardavviket til fordelingen $|\phi(p)|^2$ av bevegelsesmengde mindre, og vice versa. For partikkelen representert ved bølgepakken på figur 2.1.2 er Δx ganske liten ved $t = 0$, og deretter vokser den. Omvendt er Δp stor til å begynne med, og reduseres når t øker. Merk her at valget av nullpunkt for tiden t er helt uavhengig av regningen vi har gjort. Hvis vi definerer $\Phi(p,t)$ ved

$$\Phi(p,t) = \frac{1}{\sqrt{2\pi\hbar}} \int_{-\infty}^{\infty} \Psi(x,t) e^{-ipx/\hbar}\, dp$$

for et gitt tidspunkt $t \geq 0$, så erstattes (9) av

$$\Psi(x,t) = \frac{1}{\sqrt{2\pi\hbar}} \int_{-\infty}^{\infty} \Phi(p,t) e^{ipx/\hbar}\, dp$$

Bølgefunksjonen $\Phi(p,t)$ for bevegelsesmengde er altså fouriertransformen av posisjonsbølgefunksjonen $\Phi(x,t)$ for partikkelen ved alle tidspunkter $t \geq 0$.

I vår utledning av (15) brukte vi (11) som initialtilstand. **Heisenbergs uskarphetsrelasjon**, oppkalt etter fysikeren Werner Heisenberg (1901-1976), sier at verdien $\hbar/2$ av produktet $\Delta x \Delta p$ som vi fikk frem her, er den laveste man kan få uansett initialtilstand. Uskarphetsrelasjonen sier altså

$$\boxed{\Delta x \cdot \Delta p \geq \hbar/2}$$

Poenget er, som beskrevet over, at posisjon x og bevegelsesmengde p tilsvarer et fouriertransformpar. Se også oppgave 9.9.1. Skal du ha stor nøyaktighet i posisjonen x til en partikkel, altså Δx liten, trenger du et bredt spekter av bevegelsesmengder p. Altså Δp stor. Skal du ha stor nøyaktighet i bevegelsesmengden p, altså Δp liten, blir nødvendigvis posisjonen til partikkelen mer uskarpt bestemt, altså Δx stor. I grensen $\Delta P \to 0$, når vi nærmer oss en ren planbølge (7), har vi $\Delta x \to \infty$. Partikkelens posisjon blir da helt ubestemt.

2.1 OPPGAVER

1. I denne oppgaven er målet å fylle inn mellomregningene som leder frem til uttrykket (14) for $|\Psi(x,t)|^2$.

a) Vis at innsetting av (11) i (10) gir

$$\phi(p) = \frac{(2\pi\sigma^2)^{-1/4}}{\sqrt{2\pi\hbar}} \int_{-\infty}^{\infty} e^{-x^2/(4\sigma^2) - i(p-p_0)x/\hbar}\, dx$$

b) Bruk integralformelen

$$\int_{-\infty}^{\infty} e^{-Ax^2+Bx}\, dx = \sqrt{\frac{\pi}{A}} e^{B^2/4A}$$

til å vise at

$$\phi(p) = \left(\frac{2\sigma^2}{\pi\hbar^2}\right)^{1/4} e^{-\sigma^2(p-p_0)^2/\hbar^2}$$

c) Bruk b) sammen med (12) til å vise at

$$\Psi(x,t) = C \int_{-\infty}^{\infty} e^{-\sigma^2(p-p_0)^2/\hbar^2 + ipx/\hbar - i(p^2/2m)t/\hbar}\, dp$$

der

$$C = \frac{(2\sigma^2/\pi\hbar^2)^{1/4}}{\sqrt{2\pi\hbar}}$$

d) Bruk substitusjonen

$$u = \frac{p - p_0}{\hbar}$$

i integralet fra c) etterfulgt av integralformelen fra b) til å utlede (14).

KAPITTEL 3

TEORI UTELATT I HOVEDTEKSTEN

3.1 Cauchy–kriteriet for konvergens av følger

Vi skal her gi et bevis for teorem 5.1.4 i bind I.

Anta først at følgen $\{x_n\}$ konvergerer, la oss si mot x. Gitt et tall $\varepsilon > 0$. Ved definisjonen av konvergens vet vi da at det fins et naturlig tall N slik at

$$|x_n - x| < \varepsilon/2$$

når $n \geq N$. Hvis vi antar at både n og p er større enn eller lik K, får vi ved hjelp av trekantulikheten

$$\begin{aligned} |x_n - x_p| &= |(x_n - x) + (x - x_p)| \\ &\leq |x_n - x| + |x_p - x| = \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon \end{aligned}$$

Altså er følgen $\{x_n\}$ Cauchy.

Omvendt, anta at $\{x_n\}$ er Cauchy, altså at den oppfyller Cauchy–kriteriet for konvergens. Ved å la $\varepsilon = 1$ i Cauchy-kriteriet, får vi da at det fins N slik at

$$|x_n - x_p| \leq 1$$

for alle $n, p \geq N$. Dette betyr at *alle* leddene x_n i følgen oppfyller

$$|x_n| \leq M + 1$$

der M er den største absoluttverdien som forekommer blant tallene $x_1, \ldots, x_N$. Altså er følgen vår begrenset, og ved Bolzano–Weierstrass har den dermed en delfølge $\{y_k\}$ som konvergerer mot et tall y. La $\varepsilon > 0$ være gitt. Siden $\{x_n\}$ er Cauchy, fins det N slik at

$$|x_n - x_p| < \varepsilon$$

for alle $n, p \geq N$. Siden delfølgen $\{y_k\}$ konvergerer mot y, kan vi velge et ledd y_p slik at

$$|y_p - y| < \varepsilon/2$$

og $p \geq N$. Hvis $n \geq N$, gir trekantulikheten nå

$$\begin{aligned} |x_n - y| &\leq |(x_n - y_p) + (y_p - y)| \\ &\leq |x_n - y_p| + |y_p - y| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon \end{aligned}$$

Siden $\varepsilon > 0$ var vilkårlig, har vi vist at følgen $\{x_n\}$ konvergerer mot y. ■

3.2 Resten av grenselovene

Her er bevisene for resten av grenselovene i teorem 5.3.2, bind I.

Lov nr. 3 i tilfellet $a \in \mathbb{R}$

Gitt $\varepsilon > 0$. Ved trekantulikheten har vi

$$\begin{aligned} |f(x)g(x) - rs| &= \big|[f(x) - r][g(x) - s] + r[g(x) - s] + s[f(x) - r]\big| \\ &\leq |f(x) - r| \cdot |g(x) - s| + r|g(x) - s| + s|f(x) - r| \end{aligned}$$

Vi vil ha dette mindre enn ε, dvs. vi må få hvert av de tre uttrykkene i siste sum mindre enn en passende ε-brøk. La t være det største av tallene r, s og 1. Ved antakelsene kan vi finne $\delta > 0$ slik at når

$$|x - a| < \delta,$$

er $|f(x) - r|$ og $|g(x) - s|$ begge mindre enn

$$\frac{\varepsilon}{3t},$$

og dessuten mindre enn 1. Innsetting gir nå at

$$|f(x)g(x) - rs| \leq \frac{\varepsilon}{3t} \cdot 1 + r \cdot \frac{\varepsilon}{3t} + s \cdot \frac{\varepsilon}{3t} < \varepsilon.$$

Siden $\varepsilon > 0$ var vilkårlig, følger resultatet.

Lov nr. 4 i tilfellet $a \in \mathbb{R}$

Vi antar først $f(x) = 1$ for alle x. Da er $r = 1$. Gitt $\varepsilon > 0$. Vi får

$$\left|\frac{1}{g(x)} - \frac{1}{s}\right| = \left|\frac{s - g(x)}{s \cdot g(x)}\right| = \frac{|s - g(x)|}{|s| \cdot |g(x)|}$$

Vi må ha dette mindre enn ε. Siden $g(x) \to s$, kan vi finne $\delta > 0$ slik at når

$$|x - a| < \delta,$$

så er

$$|g(x)| > \frac{1}{2}s$$

og

$$|g(x) - s| < \frac{1}{2}\varepsilon s^2$$

Innsetting gir at uttrykket ovenfor er mindre enn

$$(\frac{1}{2}\varepsilon s^2)/(s \cdot \frac{1}{2}s) = \varepsilon$$

Hvis $f(x) \neq 1$, kan vi nå skrive

$$\frac{f(x)}{g(x)} = f(x) \cdot \frac{1}{g(x)}$$

og bruke lov 3 til å få at

$$\frac{f(x)}{g(x)} \to r \cdot \frac{1}{t} = \frac{r}{t}$$

Lov nr. 5 i tilfellet $a \in \mathbb{R}$

Gitt $\varepsilon > 0$. Siden f er kontinuerlig i $x = r$, må $f(x) \to f(r)$ når $x \to r$. Ergo fins $t > 0$ slik at når $|x - r| < t$, er

$$|f(x) - f(r)| < \varepsilon$$

Og siden $g(x) \to r$, fins $\delta > 0$ slik at når $|x - a| < \delta$, er

$$|g(x) - r| < t$$

Men i så fall er jo $|f(g(x)) - f(r)| < \varepsilon$.

Lov nr. 1–5 i tilfellet $a = \pm\infty$

Dette er helt tilsvarende bevisene i tilfellet $a \in \mathbb{R}$. Endringene som må gjøres, er de samme hele veien: De fleste δ-er må erstattes med A-er, de fleste utsagn av typen

$$|x - a| < \delta$$

må erstattes med $x > A$ (i tilfellet $x \to +\infty$) og $x < A$ (i tilfellet $x \to -\infty$), og alle antakelser om $x \neq a$ kan droppes. Som eksempel tar vi lov 5 i tilfellet $x \to -\infty$.

Gitt $\varepsilon > 0$. Siden f er kontinuerlig i $x = r$, må $f(x) \to f(r)$ når $x \to r$. Ergo fins $t > 0$ slik at når $|x - r| < t$, er

$$|f(x) - f(r)| < \varepsilon$$

Og siden $g(x) \to r$, fins et tall A slik at når $x < A$, er

$$|g(x) - r| < t$$

Men i så fall er jo $|f(g(x)) - f(r)| < \varepsilon$. ■

3.3 Teoremet til Kantorovitsj

I denne seksjonen skal vi bevise teorem 6.12.1 i bind I. Beviset står i stil med teoremet; det er kronglete og svært langt. Det er flott hvis du vil prøve å komme deg gjennom det, men vær advart om at du må jobbe en del for å forstå noen av overgangene.

Vi skal starte med et konkret eksempel som vil vise seg å være fint å sammenligne med. Anta at funksjonen vi vil bruke Newtons metode på, er

$$f(t) = \tfrac{1}{2}KMt^2 - t + \varepsilon,$$

der K, M og ε er positive tall. Forklaring på de sære koeffisientene kommer senere. Nå vet vi jo hvordan vi kan finne nullpunkter eksakt for denne funksjonen, det blir å løse en andregradsligning. Ved abc-formelen får vi at

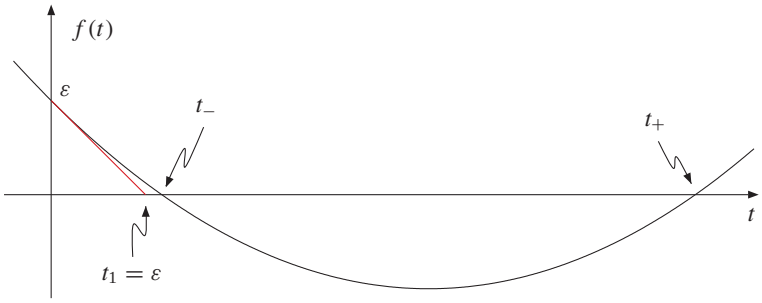
$$t_- = \frac{1 - \sqrt{1 - 2KM\varepsilon}}{KM} \tag{1}$$

er det minste nullpunktet for $f(t)$, gitt at $2KM\varepsilon < 1$. Det andre nullpunktet t_+ fremkommer ved å velge pluss foran rottegnet, så dette er større. Vi ser at begge nullpunktene er positive, og at $f(0) = \varepsilon$. Anta nå at vi bruker Newtons metode på $f(t)$, med $t_0 = 0$. Vi har $f'(t) = KMt - 1$. Dermed blir

$$t_1 = t_0 - \frac{f(0)}{f'(0)} = 0 - \frac{\varepsilon}{0 - 1} = \varepsilon$$

Alt i alt får vi nå figur 3.3.1. Grafen må se slik ut fordi koeffisienten foran t^2 er positiv, og begge nullpunktene er positive. Det første estimatet $t_1 = \varepsilon$ for nullpunktet er tegnet inn. Ved å tenke grafisk ser vi at følgen t_n fra Newtons

metode i dette tilfellet vil være voksende og oppad begrenset av nullpunktet t_-. Altså konvergerer følgen $\{t_n\}$ mot t_-.



Figur 3.3.1 Første iterasjon av Newtons metode for funksjonen $f(t)$

La oss gjøre litt griseregning her. Formelen fra Newtons metode gir

$$\begin{aligned} t_n &= t_{n-1} - \frac{f(t_{n-1})}{f'(t_{n-1})} \\ &= t_{n-1} - \frac{\frac{1}{2}KMt_{n-1}^2 - t_{n-1} + \varepsilon}{KMt_{n-1} - 1} = \frac{\frac{1}{2}KMt_{n-1}^2 - \varepsilon}{KMt_{n-1} - 1} \end{aligned}$$

Ganger vi med $KMt_{n-1} - 1$ på begge sider, får vi

$$t_n(KMt_{n-1} - 1) = \tfrac{1}{2}KMt_{n-1}^2 - \varepsilon$$

Altså

$$KMt_nt_{n-1} - t_n = \tfrac{1}{2}KMt_{n-1}^2 - \varepsilon$$

Dette kan vi skrive

$$\tfrac{1}{2}KMt_{n-1}^2 - KMt_nt_{n-1} = -t_n + \varepsilon$$

Newtons metode brukt på steget fra n til $n+1$ gir

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

Overflytting av t_n og innsetting av forrige ligning gir

$$\begin{aligned} t_{n+1} - t_n &= \frac{\frac{1}{2}KMt_n^2 - KMt_nt_{n-1} + \frac{1}{2}KMt_{n-1}^2}{1 - KMt_n} \\ &= \tfrac{1}{2}KM \cdot \frac{t_n^2 - 2t_nt_{n-1} + t_{n-1}^2}{1 - KMt_n} = \tfrac{1}{2}KM \cdot \frac{(t_n - t_{n-1})^2}{1 - KMt_n} \end{aligned} \tag{2}$$

En ting vi kan se fra (2), er t_n er størst når ε har sin maksimale verdi

$$\varepsilon = \frac{1}{2KM}$$

Grunnen er at $t_1 = \varepsilon$, slik at $t_1 - t_0$ blir større jo større ε er. Dette gjør også at nevneren

$$1 - KMt_1$$

i likning (2) er minst for maksimal epsilon, noe som gjør brøken større. Dermed vil $t_2 - t_1$, og dermed også t_2, bli størst for maksimal epsilon. Resonnementet kan så gjentas induktivt.

Vi skal nå se nærmere på spesialtilfellet $\varepsilon = 1/(2KM)$, som altså tilsvarer 0 under rottegnet i formelen for nullpunktene til f. I dette tilfellet har $f(t)$ kun ett nullpunkt

$$t_- = \frac{1}{KM}$$

Vi får

$$f(t) = \tfrac{1}{2}KM\Big(t^2 - \frac{2}{KM}t + \frac{1}{(KM)^2}\Big) = \tfrac{1}{2}KM\Big(t - \frac{1}{KM}\Big)^2$$

Videre er

$$f'(t) = KM(t - \frac{1}{KM})$$

så formelen i Newtons metode gir

$$\begin{aligned} t_{n+1} &= t_n - \frac{f(t_n)}{f'(t_n)} \\ &= t_n - \frac{1}{2}\Big(t_n - \frac{1}{KM}\Big) = \frac{1}{2}t_n + \frac{1}{2KM} \end{aligned}$$

Innsetting av $t_- = 1/(KM)$ gir at dette kan skrives

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

Dette betyr at differansen $t_- - t_n$ halveres for hvert steg. Siden

$$t_- - t_0 = t_-$$

følger ved induksjon at

$$t_- - t_n = \frac{t_-}{2^n}$$

Setter vi inn $t_- = \frac{1}{KM}$ her, får vi til slutt at

$$1 - KMt_n = \frac{1}{2^n} \tag{3}$$

Bevis for teoremet til Kantorovitsj

Etter denne forpostfektningen er vi klare for å begynne på selve beviset for Kantorovitsjs teorem. Merk først at forutsetningene gir $|F'(x_0)| \geq 1/K$ og

$$|F'(x) - F'(x_0)| \leq M \cdot |x - x_0| < M \cdot \frac{1}{KM} = \frac{1}{K}$$

for alle $x \in (x_0 - \frac{1}{KM}, x_0 + \frac{1}{KM})$. Det følger at $F'(x) \neq 0$ på dette intervallet. Fra trekantulikheten får vi for alle $x \in U$, se oppgave 1.4.12

$$|F'(x_0)| - |F'(x)| \leq |F'(x_0) - F'(x)|$$

Dette kan skrives om til (sjekk det ved baklengs regning)

$$|F'(x)|^{-1} \leq \frac{|F'(x_0)|^{-1}}{1 - |F'(x_0)|^{-1} \cdot |F'(x) - F'(x_0)|}$$

Ved å bruke antakelsene om K og M, får vi nå

$$\frac{1}{|F'(x)|} \leq \frac{K}{1 - KM|x - x_0|} \tag{4}$$

Formelen i Newtons metode kombinert med dette gir

$$|x_{n+1} - x_n| = \left|\frac{F(x_n)}{F'(x_n)}\right| \leq \frac{K|F(x_n)|}{1 - KM|x_n - x_0|} \tag{5}$$

Newtons metode gir også

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

Så

$$|F(x_n)| = |F(x_n) - F(x_{n-1}) - F'(x_{n-1})(x_n - x_{n-1})| \tag{6}$$

Fra oppgave 3.3.2 får vi at

$$|F'(x) - F'(y)| \leq M \cdot |x - y|$$

for alle $x, y \in U$ medfører

$$|F(y) - F(x) - F'(y)(y - x)| \leq \frac{M}{2}|y - x|^2 \tag{7}$$

for alle $x, y \in U$. Setter du $y = x_n$ og $x = x_{n-1}$ og kombinerer (7) med (6), fås

$$|F(x_n)| \leq \frac{M}{2}|x_n - x_{n-1}|^2 \tag{8}$$

Setter du så dette inn på høyre side i (5), får du

$$|x_{n+1} - x_n| \leq \tfrac{1}{2}KM \cdot \frac{|x_n - x_{n-1}|^2}{1 - KM|x_n - x_0|}$$

Dette resultatet er identisk med ligningen (2) for vår eksempelfunksjon $f(t)$, bortsett fra at vi nå har en ulikhet. Husk at for $f(t)$ er $t_0 = 0$ og $t_n \geq 0$, så $t_n = |t_n - t_0|$. Vi hadde også $t_1 - t_0 = \varepsilon - 0 = \varepsilon$. Hvis vi velger $\varepsilon = 1/(2KM)$, får vi ved forutsetningene i teoremet at

$$|x_1 - x_0| \leq t_1$$

La $n > 1$ være gitt. Anta at

$$|x_{k+1} - x_k| \leq t_{k+1} - t_k$$

for alle $k < n$. Da fås $|x_n - x_0| \leq t_n - t_0 = t_n$, og

$$|x_{n+1} - x_n| \leq \tfrac{1}{2}KM\frac{|x_n - x_{n-1}|^2}{1 - KM|x_n - x_0|} \leq \tfrac{1}{2}KM\frac{(t_n - t_{n-1})^2}{1 - KMt_n} = t_{n+1} - t_n$$

Ved induksjon holder da denne ulikheten for alle n. Det følger fra vår regning med eksempelfunksjonen $f(t)$ at x_n vil ligge innenfor avstand $t_- \leq 1/(KM)$ for alle n, så x_n holder seg i intervallet $(x_0 - \frac{1}{KM}, x_0 + \frac{1}{KM})$ for alle n. Hvis $k > n$, gir trekantulikheten

$$\begin{aligned} |x_k - x_n| &\leq |x_k - x_{k-1}| + |x_{k-1} - x_{k-2}| + \cdots + |x_{n+1} - x_n| \\ &\leq (t_k - t_{k-1}) + (t_{k-1} - t_{k-2}) + \cdots + (t_{n+1} - t_n) = t_k - t_n \end{aligned} \tag{9}$$

Vi vet at følgen $\{t_n\}$ konvergerer, og dermed kan vi få $t_k - t_n$ mindre enn en gitt ε ved å velge en tilstrekkelig stor N og ta $k > n > N$. Altså er $\{x_n\}$ en Cauchy-følge, så den konvergerer mot et punkt x^* som da må ligge i intervallet

$$[x_0 - \frac{1}{KM}, x_0 + \frac{1}{KM}]$$

At $f(x) = 0$ følger ved å la $n \to \infty$ på begge sider av (7). Venstre side nærmer seg $f(x^*)$ fordi f er kontinuerlig. Høyre side går mot 0, fordi vi nå vet at $|x_n - x_{n-1}|$ går mot 0 når $n \to \infty$.

La nå y^* være et vilkårlig nullpunkt for F i intervallet $[x_0 - \frac{1}{KM}, x_0 + \frac{1}{KM}]$. Vi skal vise at $x_n \to y$ når $n \to \infty$. Siden følgen $\{x_n\}$ kun kan konvergere mot ett punkt, følger av dette at $y^* = x^*$. Da har vi vist at x^* er det *eneste* nullpunktet for F i intervallet. For å få frem dette setter vi $y = y^*$ og $x = x_n$. Ved å bruke at $F(y^*) = 0$, får vi da

$$|F(x_n) + F'(x_n)(y^* - x_n)| \leq \frac{M}{2}|y^* - x_n|^2 \tag{10}$$

Videre har vi (sjekk det ved baklengs regning)

$$y^* - [x_n - F'(x_n)^{-1}F(x_n)] = F'(x_n)^{-1}[F(x_n) + F'(x_n)(y^* - x_n)]$$

Ved formelen i Newtons metode er uttrykket i hakeparentesen på venstre side x_{n+1}. Setter vi dette inn, tar absoluttverdi på begge sider og bruker (4) og (10), får vi

$$|y^* - x_{n+1}| \leq \frac{\frac{1}{2}KM \cdot |y^* - x_n|^2}{1 - KM \cdot |x_n - x_0|}$$

Samtidig vet vi fra (2) at vår eksempelfølge $\{t_n\}$ oppfyller

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

og $t_n \geq |x_n - x_0|$ for $n \geq 0$.
Med andre ord: Hvis vi kan vise for en konkret verdi av n at $|y^* - x_n| \leq t_n - t_{n-1}$, får vi

$$|y^* - x_{n+1}| \leq \frac{\frac{1}{2}KM \cdot (t_n - t_{n-1})^2}{1 - KM \cdot t_n} = t_{n+1} - t_n$$

Dette gir oss induksjonssteget i et induksjonsbevis for at $|y^* - x_n| \leq t_n - t_{n-1}$ gjelder for alle $n \geq 1$. Siden $t_n - t_{n-1}$ går mot 0 når $n \to \infty$, vil det følge fra dette at $x_n \to y^*$ når $n \to \infty$, som var det vi ønsket å vise. Men vi har et problem: Vi har $t_1 - t_0 = \varepsilon - 0 = \varepsilon$, der $\varepsilon \leq 1/(2KM)$, mens vi bare vet at $|y^* - x_1| \leq 1/KM$. Dette betyr at selv om vi velger den maksimale verdien $\varepsilon = 1/(2KM)$ i eksempelfunksjonen, er det en faktor 2 i forskjell! Vi får altså ikke startet induksjonen vår på $n = 1$. Trikset er å tenke oss at vi startet Newtons metode for eksempelfunksjonen med $\varepsilon = 1/(2KM)$ for $n = -1$ istedenfor $n = 0$. Med denne ε-verdien vet vi at nullpunktet t_- er $1/KM$, og at avstanden til dette halveres for hvert steg i Newtons metode. Det betyr at avstanden fra t_{-1} til t_0 må være lik avstanden fra t_0 til t_-, som er $1/KM$. Altså $t_0 - t_{-1} = 1/KM$, noe som gjør at $|y^* - x_0| \leq t_0 - t_{-1}$. Dermed kan vi starte induksjonen på $n = 0$ i stedet, og vi er i mål.

Det gjenstår å vise estimatet nederst i Kantorovitsjs teorem. Fra (9) vet vi at $|x_k - x_n| \leq t_k - t_n$ for alle k og n. Holder vi k fast og lar $n \to \infty$ her, får vi at

$$|x - x_n| \leq t_- - t_n$$

Derfor holder det å vise at følgen $\{t_n\}$ for eksempelfunksjonen $f(t)$ oppfyller

$$t_- - t_n \leq \frac{1}{KM} \cdot \frac{(1 - \sqrt{1 - 2h})^{2^n}}{2^n} \tag{11}$$

der $h = KM\varepsilon$. Her tenker vi oss at $\varepsilon > 0$ er gitt. Siden vi har $\varepsilon \leq \frac{1}{2MK}$, har vi $h \leq \frac{1}{2}$. I oppgave 3.3.1 blir du bedt om å vise at

$$t_- - t_{n+1} = \frac{\frac{1}{2}KM \cdot (t_- - t_n)^2}{1 - KM \cdot t_n} \tag{12}$$

for alle $n \geq 0$. I spesialtilfellet $\varepsilon = \frac{1}{2MK}$ har vi $1 - KMt_n = \frac{1}{2^n}$ fra (3), og vi fant også at x_n er størst når $\varepsilon = \frac{1}{2MK}$. Altså har vi

$$1 - KMt_n \geq \frac{1}{2^n}$$

uansett verdi av ε. Med dette innsatt kan (12) skrives om til

$$KM(t_- - t_{n+1}) \leq 2^{n-1} \cdot (KM(t_- - t_n))^2 \tag{13}$$

for $n \geq 0$. Man kan sjekke ved induksjon at hvis en tallfølge $\{a_n\}_{n=0}^{\infty}$ oppfyller

$$a_{n+1} \leq 2^{n-1} a_n$$

for alle $n \geq 0$, så er $a_n \leq 2^{-n} a_0^n$. Med $a_n = KM(t_- - t_n)$ får vi

$$a_0 = KM(t_- - 0) = 1 - \sqrt{1 - 2h}$$

Altså gir (13)

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

Divisjon med KM på begge sider gir ligningen (11), og beviset er komplett. ■

3.3 OPPGAVER

1. La $f(t) = \frac{1}{2}KMt^2 - t + \varepsilon$, der $0 < \varepsilon \leq \frac{1}{2MK}$. La t_n være følgen vi får ved å bruke Newtons metode på f med startpunkt $t_0 = 0$, og la t_- være det minste nullpunktet for f.

a) Bruk formelen fra Newtons metode til å vise at

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

b) Forklar hvorfor $t_- = \frac{1}{2}KM(t_-)^2 + \varepsilon$

c) Vis, ved å sette inn uttrykket for t_- fra b) for den andre forekomsten av t_- på høyre side av ligningen fra a), at

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

2. La U være et åpent intervall, og la $F : U \to \mathbb{R}$ være en deriverbar funksjon slik at

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

for alle $x, y \in U$, der M er en konstant.

a) La $r(t) = x + t(y - x)$ og $G(t) = F(r(t))$. Vis at

$$G'(t) = F'(r(t))(y - x),$$

og bruk dette til å vise at

$$G'(t) = (F'(r(t)) - F'(x))(y - x) + F'(x)(y - x)$$

b) Vis at $F(y) - F(x) = \int_0^1 G'(t)dt$.

c) Vis at $\int_0^1 G'(t)dt$ kan skrives

$$F'(x)(y - x) + \int_0^1 (F'(r(t)) - F'(x))(y - x)\, dt$$

d) Vis at $|F(y) - F(x) - F'(x)(y - x)| = I$, der

$$I = \left| \int_0^1 (F'(r(t)) - F'(x))(y - x)\, dt \right|$$

e) Vis at

$$I \leq M \int_0^1 |r(t) - x| \cdot |y - x|\, dt$$

f) Forklar hvorfor $r(t) - x = t(y - x)$, og bruk dette til å vise at

$$I \leq \frac{M}{2}|y - x|^2$$

g) Vis at vi for alle $x, y \in U$ har

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

3.4 Algebraens fundamentalteorem

I denne seksjonen skal vi bevise algebraens fundamentalteorem fra side 270 i bind I. Fremstillingen bygger kun på kapittel 8 i bind I samt ekstremalverdisetningen 1.4.1 i bind II.

Det viser seg at fundamentalteoremet ganske greit lar seg utlede hvis vi vet at ethvert komplekst polynom av grad 1 eller høyere har minst én rot. Dette er budskapet i teoremet under. Utledningen av fundamentalteoremet på basis av dette tas opp i oppgave 3.4.3 ved slutten av seksjonen.

TEOREM 1

Nullpunkt for komplekse polynomer

La $P(z) = c_0 + c_1 z + c_2 z^2 + \cdots + c_n z^n$ være et komplekst polynom av grad $n \geq 1$. Da fins en $z_0 \in \mathbb{C}$ slik at $P(z_0) = 0$.

BEVIS Definer den kontinuerlige, reelle funksjonen $f : \mathbb{R}^2 \to \mathbb{R}$ ved

$$f(x, y) = |P(x + iy)|$$

La m være infimum av verdimengden til f. Tallet m fins fordi verdimengden er nedad begrenset av 0. Da fins det (se oppgave 3.4.2) et reelt tall $R > 0$ slik at $f(x, y) > m + 1$ utenfor den lukkede sirkelskiven $x^2 + y^2 \leq R^2$. Ved ekstremalverdisetningen 1.4.1 i bind II brukt på sirkelskiven følger at f har et globalt minimumspunkt (a, b) på sirkelskiven, og dette vil da også være et globalt minimum på hele planet. Vi har $f(a, b) = m$. Det eneste vi trenger å vise, er at $m = 0$.

Anta at $m > 0$. Vi skal utlede en selvmotsigelse fra dette. La $z_0 = a + ib$ og $w = z - z_0$. Da er $z = z_0 + w$, så

$$P(z) = c_0 + c_1(z_0 + w) + c_2(z_0 + w)^2 + \cdots + c_n(z_0 + w)^n$$

Ved å gange ut parenteser og samle ledd som har felles potens w^k, kan man skrive dette om på formen

$$P(z) = d_0 + d_1 w + d_2 w^2 + \cdots + d_n w^n,$$

der koeffisientene d_i er komplekse tall uavhengige av z. Merk at $f(z_0) = d_0$. Ergo er $|d_0| = m \neq 0$. La d_k være første koeffisient etter d_0 som er ulik 0. Da kan vi skrive

$$P(z) = d_0 + d_k w^k + q \cdot w^k,$$

der $q = d_{k+1} w + \cdots + d_n w^{n-k}$. Vi skriver nå på polarform: $w = re^{i\theta}$, $d_k = se^{i\phi}$ og $d_0 = me^{iv}$. Merk at siden $m = |d_0|$, stemmer det siste. Innsatt fås

$$\begin{aligned} P(z) &= me^{iv} + se^{i\phi} r^k e^{ik\theta} + qr^k e^{ik\theta} \\ &= me^{iv} + sr^k e^{i(\phi + k\theta)} + qr^k e^{ik\theta} \end{aligned}$$

Vi velger så z slik at θ oppfyller $\phi + k\theta = v + \pi$. Da er

$$e^{i(\phi+k\theta)} = e^{i(v+\pi)} = e^{iv}e^{i\pi} = -e^{iv}$$

For r så liten at $sr^k < m$ gjelder da

$$\begin{aligned}|P(z)| &= |(m - sr^k)e^{iv} + qr^k e^{ik\theta}| \\ &\le |(m - sr^k)e^{iv}| + |qr^k e^{ik\theta}| \\ &= |m - sr^k| \cdot |e^{iv}| + |q| \cdot r^k \cdot |e^{ik\theta}| \\ &= (m - sr^k) \cdot 1 + |q| \cdot r^k \cdot 1 = m - (s - |q|)r^k,\end{aligned}$$

der vi brukte trekantulikheten for komplekse tall (se oppgave 3.4.1) i andre overgang. Men slik q er definert, vil automatisk $|q|$ gå mot 0 når r går mot 0. For $r > 0$ tilstrekkelig liten, er derfor $s > |q|$. Da er

$$|P(z)| \le m - (s - |q|)r^k < m$$

Dette er selvmotsigelsen vi var på jakt etter, og teoremet er bevist. ■

3.4 OPPGAVER

1. *Trekantulikheten for komplekse tall*. Vis at for alle komplekse tall z og w gjelder

$$|z + w| \le |z| + |w|$$

(Hint: Vi har tidligere vist et slikt resultat for reelle vektorer i planet. Er det noen forskjell?)

2. I denne oppgaven skal vi utlede et resultat om funksjonen $f(x, y) = |P(x + iy)|$ i beviset ovenfor.

a) Vis at polynomet $P(z)$ i teoremet oppfyller
$|P(z)| = |z^n| \cdot |c_0 z^{-n} + \cdots + c_{n-1}z^{-1} + c_n|$

b) La k være et reelt tall. Vis at det fins $R > 0$ slik at hvis $|z| > R$, så er $|P(z)|$ større enn k.
(Hint: Når $|z| \to \infty$, går $|z^n|$ mot ∞. Den andre faktoren i uttrykket for $|P(z)|$ går mot $|c_n|$. Hvorfor beviser dette resultatet om f som vi var ute etter?)

3. Vi skal nå se hvordan man kan bevise fundamentalteoremet på grunnlag av teorem 3.4.1. Først skal vi vise følgende utsagn ved induksjon: «Ethvert komplekst polynom
$P(z) = c_0 + c_1 z + \cdots + c_n z^n$ av grad $n \ge 1$ kan skrives

$$P(z) = c_n(z - r_1)(z - r_2)\cdots(z - r_n)$$

der $r_1, \ldots, r_n$ er komplekse tall.»

a) Begrunn at utsagnet holder for $n = 1$.

Anta nå at utsagnet holder for $n = k$, og la $P(z)$ være et polynom av grad $k + 1$. Fra teorem 3.4.1 vet vi at $P(z)$ har en rot r.

b) Begrunn med inspirasjon av teorem 1.9.3 i bind I om polynomdivisjon at det fins et komplekst polynom $Q(z)$ av grad k slik at

$$P(z) = Q(z) \cdot (z - r)$$

c) Vis at utsagnet vårt holder for alle $n \ge 1$.

Vi har nå *nesten* bevist fundamentalteoremet side 270 i bind I. Det eneste som mangler, er unikheten av faktoriseringen. Denne tar vi oss av i neste punkt.

d) Vis at hvis et komplekst polynom $P(z)$ kan skrives på to måter

$$\begin{aligned}P(z) &= c_n(z - r_1)(z - r_2)\cdots(z - r_n) \\ &= c_n(z - s_1)(z - s_2)\cdots(z - s_n),\end{aligned}$$

så må $r_1, \ldots, r_n$ være de samme n tallene som $s_1, \ldots, s_n$, bortsett eventuelt fra rekkefølgen. (Hint: Se når uttrykkene er null.)

3.5 Leddvis derivasjon og integrasjon av potensrekker

I denne seksjonen skal vi vise gyldigheten av triks 3 fra seksjon 10.5 i bind I.

Vi starter med en potensrekke som konvergerer mot en sum $S(x)$ på åpent intervall rundt potensrekkens sentrum, altså

$$\sum_{n=0}^{\infty} c_n(x-a)^n = S(x) \qquad \text{for } x \in U$$

der $U = (a-R, a+R)$ er et åpent intervall. La S_N være summen av de N første leddene i rekken ovenfor.

Påstand 1

La $J \subseteq U$ være et lukket intervall. Da konvergerer $\sum_{n=0}^{\infty} c_n(x-a)^n$ absolutt for alle $x \in J$. For hver $\varepsilon > 0$ fins N slik at for alle $x \in J$ er

$$|S(x) - S_N(x)| < \varepsilon$$

Bevis for påstand 1

Velg $b \in U$ og $r < 1$ slik at $|x-a| \le r \cdot |b-a|$ for alle $x \in J$. Siden rekken konvergerer i b, fins K slik at $|c_n(b-a)^n| < K$ for alle n. Da er

$$|c_n(x-a)^n| \le |c_n(b-a)^n| \cdot r^n \le K \cdot r^n \qquad \text{for alle } x \in J.$$

Rekken $\sum Kr^n$ er geometrisk og konvergent, så sammenligningstesten gir at potensrekken vår er absolutt konvergent i punktet x. Videre er

$$|S(x) - S_N(x)| = \left| \sum_{n=N}^{\infty} c_n(x-a)^n \right| \le \left| \sum_{n=N}^{\infty} K \cdot r^n \right|$$

Her er uttrykket lengst til høyre «resten» av en konvergent geometrisk rekke, og det er derfor mindre enn en gitt $\varepsilon > 0$ for N stor nok. ■

Påstand 2

Funksjonen $S(x)$ er kontinuerlig på U.

Bevis for påstand 2

Gitt $p \in U$, velg et lukket intervall $J \subseteq U$ slik at $p \in J^*$. Gitt $\varepsilon > 0$. Ved kontinuitet av polynomet $S_N(x)$ fins en omegn V om p inneholdt i J slik at $|S_N(x) - S_N(p)| < \varepsilon/3$ hvis $x \in V$. Ved påstand 1 fins N slik at $|S(x) - S_N(x)| < \varepsilon/3$ for alle $x \in V$. Hvis $x \in V$, fås da (trekantulikheten)

$$\begin{aligned} |S(x) - S(p)| &\le |S(x) - S_N(x)| + |S_N(x) - S_N(p)| + |S_N(p) - S(p)| \\ &< \varepsilon/3 + \varepsilon/3 + \varepsilon/3 = \varepsilon \end{aligned}$$

■

Påstand 3

Rekken $\sum_{n=1}^{\infty} nc_n(x-a)^{n-1}$ konvergerer for alle $x \in U$.

Bevis for påstand 3

Gitt $x \in U$, velg $b \in U$ slik at $|x-a| = r|b-a|$, der $r < 1$. Da er

$$|nc_n(x-a)^{n-1}| \le B_n \cdot |c_n(b-a)^n|, \qquad \text{der} \qquad B_n = \frac{nr^n}{r|b-a|}$$

Her går $nr^n = ne^{n(\ln r)}$ mot 0 når $n \to \infty$ fordi $r < 1$ (l'Hôpital), så det fins K slik at $B_n < K$ for alle n. Siden rekken $\sum c_n(b-a)^n$ konvergerer absolutt ved påstand 1, gir sammenligningstesten at $\sum nc_n(x-a)^{n-1}$ konvergerer. ■

Påstand 4

$$\int_b^c S(x)\,dx = \lim_{N\to\infty}\left[\int_b^c S_N(x)\,dx\right] \text{ for alle } b, c \in U \text{ slik at } b < c.$$

Bevis for påstand 4

Gitt $\varepsilon > 0$. Ved påstand 1, velg N slik at $|S(x) - S_N(x)| < \varepsilon/(c-b)$ for alle $x \in [b, c]$. Da fås

$$\left|\int_b^c S(x)\,dx - \int_b^c S_N(x)\,dx\right| = \left|\int_b^c [S(x) - S_N(x)]\,dx\right| \le \frac{\varepsilon(c-b)}{c-b} = \varepsilon,$$

ved påstand 2. Påstand 4 er vist. ■

Fra påstand 4 følger at leddvis bestemt integrasjon innenfor U er ok. For siden vi kan integrere polynomet $S_N(x)$ ledd for ledd, har vi

$$\lim_{N\to\infty}\left[\int_b^c S_N(x)\,dx\right] = \lim_{N\to\infty}\sum_{i=0}^{N-1}\left[\int_b^c c_n(x-a)^n\,dx\right],$$

og uttrykket til høyre er summen av den rekken vi får ved å integrere rekken for $S(x)$ leddvis. At leddvis ubestemt integrasjon er lovlig følger også, for ved fundamentalteoremet er det bestemte integralet $\int_a^x S(t)\,dt$ en antiderivert av $S(x)$, og dette bestemte integralet vet vi nå at vi kan beregne leddvis. Så var det leddvis derivasjon. Ved påstand 3 kan vi døpe $H(t) = \sum_{n=1}^{\infty} nc_n(t-a)^{n-1}$ for $t \in U$. Leddvis integrasjon gir da

$$\int_a^x H(t)\,dt = \sum_{n=1}^{\infty}\Big[c_n(t-a)^n\Big]_a^x = \sum_{n=1}^{\infty} c_n(x-a)^n = S(x) - c_0,$$

for alle $x \in U$. Fundamentalteoremet gir $S'(x) = H(x)$ for alle $x \in U$.

3.6 Punktvis og uniform konvergens

Epsilon-betingelsen fra påstand 1 i forrige seksjon har fått et eget navn. En potensrekke

$$\sum_{n=0}^{\infty} c_n(x-a)^n$$

med sum $S(x)$ for alle x i et intervall $I \subseteq \mathbb{R}$, sies å konvergere **uniformt** mot sumfunksjonen $S(x)$ på intervallet I hvis det for hver gitt $\varepsilon > 0$ fins et naturlig tall N slik at

$$|S_N(x) - S(x)| < \varepsilon$$

for alle $x \in I$, der S_N er summen av de N første leddene i potensrekken.

Betingelsen ovenfor er et *sterkere* krav enn at potensrekken skal konvergere mot $S(x)$ på I. Grunnen er at det naturlige tallet N tilhørende hver ε her skal fungere uniformt, altså være «felles» for *alle* $x \in I$. Til sammenlikning krever den «vanlige» definisjonen av konvergens på intervallet I at det for hver valgt $x \in I$ og hver $\varepsilon > 0$ fins N slik at

$$|S_N(x) - S(x)| < \varepsilon$$

Her tillates det naturlige tallet N som skal finnes til hver ε, å variere med x. Dette konvergenbegrepet, som i mer presis terminologi kalles **punktvis konvergens**, betyr altså ganske enkelt at vi for alle $x \in I$ har

$$\lim_{N\to\infty} S_N(x) = S(x)$$

Jamfør definisjonen av konvergens for en rekke i seksjon 10.1, som i sin tur refererer til definisjonen av konvergens for en følge i seksjon 5.1.

Geometrisk tolkning

At en potensrekke konvergerer uniformt mot en sumfunksjon

$$S(x)$$

på et intervall I, betyr grafisk at det for enhver «ε-pølse» om grafen til $S(x)$ finnes en N slik at grafen til delsumfunksjonen $S_N(x)$ ligger innenfor pølsen på hele intervallet I. Påstand 1 i forrige seksjon viser altså at dersom I er et lukket intervall $[a, b]$ innholdt i konvergensområdet til en potensrekke, så vil dette være tilfelle: Rekken vil da konvergere uniformt mot sin sumfunksjon $S(x)$ på intervallet I.

Kikker du på figur 10.3.4 og 10.3.5, som illustrerer konvergens av taylorrekken til sin x, så kan du se at dette er rimelig. Hvis vi velger et intervall $[-a, a]$ og har gitt en $\varepsilon > 0$, så kan vi ved å velge N stor nok oppnå at delsummen

$$S_N(x) = T_N(x)$$

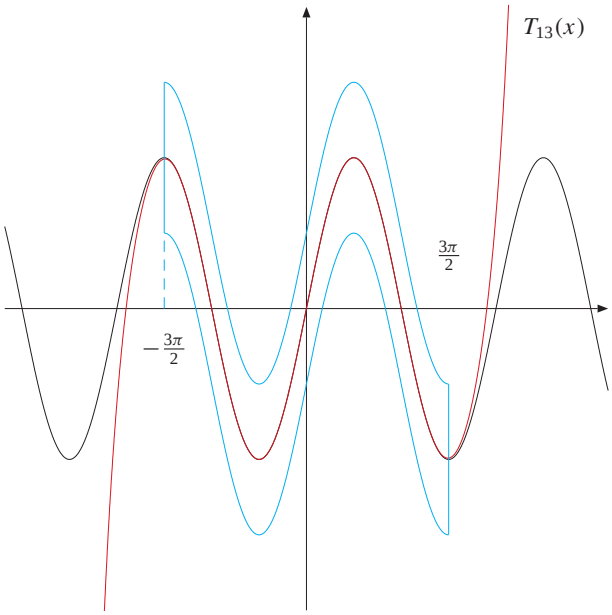
av denne taylorrekken ligger innenfor pluss/minus ε fra

$$S(x) = \sin x$$

på intervallet. Dette er illustrert for intervallet $[-\frac{3\pi}{2}, \frac{3\pi}{2}]$ med $\varepsilon = \frac{1}{2}$ på figur 3.6.1. Her ser vi at delsummen

$$S_{13}(x) = T_{13}(x)$$

ligger innenfor ε-pølsen med $\varepsilon = \frac{1}{2}$ på $[-\frac{3\pi}{2}, \frac{3\pi}{2}]$. Det gjør derimot ikke $T_5(x)$, som vi kan se fra figur 10.3.4.



Figur 3.6.1 Taylorpolynomet $T_{13}(x)$ ligger inne i ε-pølsen med $\varepsilon = \frac{1}{2}$ om $f(x) = \sin x$ på intervallet $[-\frac{3\pi}{2}, \frac{3\pi}{2}]$

Merk også at taylorrekken ikke konvergerer uniformt mot $S(x) = \sin x$ på hele det uendelige intervallet $(-\infty, \infty)$. Ingen N er stor nok til å tvinge $S_N(x)$ innenfor en gitt avstand ε fra grafen til

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

på hele tallinjen. Derimot konvergerer taylorrekken **punktvis** mot $\sin x$ på hele tallinjen, noe som altså kort og godt betyr at vi har

$$\lim_{N\to\infty} T_N(x) = \sin x$$

for hver fastholdt $x \in \mathbb{R}$.

Generelle definisjoner

Foreløpig har vi kun diskutert uniform og punktvis konvergens for potensrekker. Siden konvergens av en rekke per definisjon betyr at følgen av delsummer konvergerer, er det vi har gjort egentlig å definere hva det vil si at følgen

$$S_N(x)$$

av delsumfunksjoner for en potensrekke konvergerer uniformet og punktvis mot en funksjon $S(x)$ på et gitt intervall I. Mer generelt har vi definisjonen nedenfor.

DEFINISJON 1

Uniform og punktvis konvergens av funksjonsfølger

La $U \subseteq \mathbb{R}$, og la $f : U \to \mathbb{R}$ være en gitt funksjon. At en følge $\{f_N\}$ av funksjoner definert på U konvergerer **uniformt** mot f på U, betyr at det for alle $\varepsilon > 0$ fins et naturlig tall N slik at vi for alle $x \in U$ har

$$|f_N(x) - f(x)| < \varepsilon$$

At følgen konvergerer **punktvis** mot f på U, betyr at vi for hver fastholdt $x \in U$ har

$$\lim_{N\to\infty} f_N(x) = f(x)$$

I tilfellet potensrekker består altså følgen $\{f_N\}$ av delsummene S_N til potensrekken. Merk at hvis en funksjonsfølge konvergerer uniformt mot en funksjon f på U, så konvergerer den også punktvis mot f på U. Videre har vi følgende generalisering av påstand 2 i forrige seksjon:

TEOREM 1

Uniform konvergens bevarer kontinuitet

Anta at $\{f_N\}$ er en følge av kontinuerlige funksjoner som konvergerer uniformt mot funksjonen f på et intervall U. Da er er grensefunksjonen f også kontinuerlig på U.

BEVIS Dette er en relativt ren oversettelse av beviset for påstand 2 i forrige seksjon. Detaljene droppes. ■

3.7 Kompakthetsteoremet

I denne seksjonen skal vi gi et bevis for kompakthetsteoremet, altså teorem 1.2.1 i bind II.

Påstand 1

Teoremet holder for lukkede intervaller $[a, b] \subseteq \mathbb{R}$.

Bevis for påstand 1

Gitt en overdekning O av $[a, b]$ bestående av åpne mengder i $\mathbb{R}$. Anta at

$$q = \inf \left\{ x \in [a, \infty) \;\middle|\; \text{Intet endelig utvalg fra } O \text{ dekker } [a, x] \right\}$$

fins. Siden en mengde fra O må inneholde a, og dermed også $[a, a + \delta]$ for en passende δ, må $q > a$. Anta

$$q \leq b$$

Velg en $U \in O$ slik at $q \in U$. La $\varepsilon > 0$ være så liten at

$$[q - \varepsilon, q + \varepsilon] \subseteq U$$

samtidig som $q - \varepsilon > a$. Pr. definisjon av q fins da et endelig utvalg $U_1, \ldots, U_k$ fra O som dekker $[a, q - \varepsilon]$. Men da dekker jo samlingen

$$U_1, \ldots, U_k, U$$

intervallet $[a, q + \varepsilon]$, i strid med definisjonen av q. Altså $q > b$, så hvis q fins, holder påstanden. Men hvis q *ikke* fins må det skyldes at mengden den er infimum av, er tom. Spesielt er ikke b med i denne mengden! Påstand 1 er vist.

Påstand 2

Teoremet holder for lukkede rektangler $R = [a_1, b_1] \times \cdots \times [a_n, b_n] \subseteq \mathbb{R}^n$.

Bevis for påstand 2

Teoremet holder klart for R når $a_j = b_j$ for alle j, for da består R av ett punkt. Anta at det holder hvis

$$a_j = b_j \text{ for } j > k.$$

Anta så at vi har gitt en R med $a_j = b_j$ kun for

$$j > k + 1$$

Hvis vi kan vise at teoremet holder for R, følger påstand 2 ved induksjon. La O være en åpen overdekning av R bestående av åpne *rektangler*. Da er O en åpen overdekning av

$$R_x = [a_1, b_1] \times \cdots \times [a_k, b_k] \times \{x\} \times \{a_{k+2}\} \times \cdots \times \{a_n\}$$

også, for hver $x \in [a_{k+1}, b_{k+1}]$. Ved induksjonshypotesen fins et endelig utvalg

$$O_x \subseteq O$$

som også dekker R_x. Men da må det (se oppgave 2) fins en åpen omegn U_x om x i $\mathbb{R}$ slik at mengdene i O_x tilsammen også dekker

$$S_x = [a_1, b_1] \times \cdots \times [a_k, b_k] \times U_x \times \{a_{k+2}\} \times \cdots \times \{a_n\}$$

Samlingen

$$O' = \{U_x \mid x \in [a_{k+1}, b_{k+1}]\}$$

er en åpen overdekning av intervallet

$$[a_{k+1}, b_{k+1}],$$

og dermed fins

$$x_1, \ldots, x_N \in [a_{k+1}, b_{k+1}]$$

slik at samlingen $U_{x_1}, \ldots, U_{x_N}$ også dekker $[a_{k+1}, b_{k+1}]$. Men da dekker den kombinerte samlingen

$$O_{x_1}, \ldots, O_{x_N}$$

hele R, og denne samlingen er endelig. Vi har dermed vist at enhver åpen overdekning av R bestående av *rektangler* inneholder et endelig utvalg som dekker. At *enhver* åpen overdekning av R inneholder et endelig utvalg som dekker, følger direkte fra dette (se oppgave 1). Påstand 2 er vist.

Gitt påstand 2 følger teoremet lett. Hvis $K \subseteq \mathbb{R}^n$ er kompakt, fins et lukket rektangel R slik at

$$K \subseteq R$$

La O være en åpen overdekning av K. Da er O kombinert med

$$\mathbb{R}^n \setminus K$$

en åpen overdekning av R. Ved påstand 2 fins et endelig utvalg $U_1, \ldots, U_N$ fra O som, muligens sammen med $\mathbb{R}^n \setminus K$, dekker R. Da dekkes K av $U_1, \ldots, U_N$. ■

3.7 OPPGAVER

1. Anta at mengden $A \subseteq \mathbb{R}^n$ har den egenskap at enhver åpen overdekning av A bestående av åpne rektangler innholder et endelig utvalg som dekker A. Vis at *enhver* åpen overdekning av A da inneholder et endelig utvalg som dekker A.

2. Begrunn at det er riktig som det påstås i linje 5 av beviset for påstand 2 på denne siden, at det fins U_x slik at samlingen O_x dekker S_x. (Hint: Tegn en figur i tilfellet $n = 2$, $k = 1$.)

3.8 Koordinatskifteteoremet

I denne seksjonen skal vi bevise koordinatskifteteoremet for dobbeltintegraler (teorem 4.3.1 i bind II). Vi deler beviset opp i to biter. Først skal vi bevise en «lokal» utgave som sier at konklusjonen i teoremet holder for funksjoner som er ulik null bare i en liten omegn rundt et gitt punkt.

TEOREM 1

Lokal variant av koordinatskifteteoremet

La D og A være åpne, begrensede områder i $\mathbb{R}^2$, og la T være en C^1, injektiv koordinattransformasjon i $\mathbb{R}^2$ slik at determinanten $\det T'(u,v)$ er ulik 0 for alle $(u,v) \in A$, og $T(A) = D$. La $(\overline{u},\overline{v})$ være et gitt punkt i A. Da fins et åpent rektangel $E \subseteq D$ med sentrum $T(\overline{u},\overline{v})$ slik at hvis $f : D \to \mathbb{R}$ er kontinuerlig og begrenset og oppfyller $f(x,y) = 0$ utenfor E, så gjelder

$$\iint_D f(x,y)\,dxdy = \iint_A f(T(u,v)) \cdot |\det T'(u,v)|\,dudv$$

BEVIS Vi har

$$\det T'(\overline{u},\overline{v}) = \begin{vmatrix} \frac{\partial T_1}{\partial u}(\overline{u},\overline{v}) & \frac{\partial T_1}{\partial v}(\overline{u},\overline{v}) \\ \frac{\partial T_2}{\partial u}(\overline{u},\overline{v}) & \frac{\partial T_2}{\partial v}(\overline{u},\overline{v}) \end{vmatrix} \neq 0$$

Dermed må

$$\frac{\partial T_2}{\partial u}(\overline{u},\overline{v}) \neq 0 \quad \text{eller} \quad \frac{\partial T_2}{\partial v}(\overline{u},\overline{v}) \neq 0$$

Å bytte om søyler i determinanten svarer bare til å bytte rekkefølgen på koordinatene $\mathbb{R}^2$, og slik ombytting endrer ingen av integralene i koordinatskifteteoremet. Dermed kan vi anta at

$$\frac{\partial T_2}{\partial v}(\overline{u},\overline{v}) \neq 0$$

Definer $H : A \to \mathbb{R}^2$ ved

$$H(u,v) = (u, T_2(u,v))$$

Da er

$$\det H'(\overline{u},\overline{v}) = \begin{vmatrix} 1 & 0 \\ \frac{\partial T_2}{\partial u}(\overline{u},\overline{v}) & \frac{\partial T_2}{\partial v}(\overline{u},\overline{v}) \end{vmatrix} = \frac{\partial T_2}{\partial v}(\overline{u},\overline{v}) \neq 0$$

Ved inversfunksjonsteoremet fins dermed en åpen omegn V om $H(\overline{u},\overline{v})$ slik at H har en C^1 invers på V, og

$$\det(H^{-1})' \neq 0 \text{ på } V, \quad \det(H') \neq 0 \text{ på } H^{-1}(V).$$

Definer $K : V \to \mathbb{R}^2$ ved

$$K(x, y) = (T_1(H^{-1}(x, y)), y)$$

For alle $(u, v) \in H^{-1}(V)$ har vi nå

$$T(u, v) = K(H(u, v))$$

Her har vi fått skrevet T som en sammensetning av to funksjoner H og K som hver bare endrer én av koordinatene. Strategien videre er å bruke vanlig integrasjon ved substitusjon for funksjoner av én variabel på hver av disse. Før vi starter med det, merk at

$$K(x, y) = T(H^{-1}(x, y))$$

for alle $(x, y) \in V$. Det følger ved kjerneregelen at K er C^1 på V, og ved produktregelen for determinanter fås dessuten $\det(K') \neq 0$ på V.

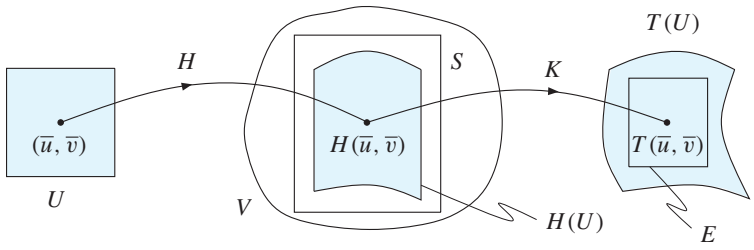
La nå $S \subseteq V$ være et åpent rektangel med sentrum $H(\overline{u}, \overline{v})$. Siden H er kontinuerlig, fins et åpent rektangel

$$U \subseteq H^{-1}(V)$$

med sentrum $(\overline{u}, \overline{v})$ slik at $H(U) \subseteq S$. Velg et åpent rektangel

$$E \subseteq T(U)$$

med sentrum $T(\overline{u}, \overline{v})$. Figur 3.8.1 viser situasjonen.



Figur 3.8.1 Transformasjonen T som sammensetningen av H og K.

Skriv

$$S = [a, b] \times [\alpha, \beta] \quad \text{og} \quad U = [c, d] \times [\gamma, \delta].$$

For gitt $x \in [c, d]$ og $y \in [\alpha, \beta]$, la

$$K_1^y(x) = K_1(x, y) \quad \text{og} \quad H_2^x(y) = H_2(x, y).$$

La $f : D \to \mathbb{R}$ være kontinuerlig, begrenset og med $f(x, y) = 0$ utenfor E. Vi får da:

$$\begin{aligned}
\iint_D f(x,y)\,dxdy &\overset{1}{=} \int\int_{K(S)} f(x,y)\,dxdy \\
&\overset{2}{=} \int_\alpha^\beta \Big[\int_{K_1^y([a,b])} f(x,y)\,dx \Big]\,dy \\
&\overset{3}{=} \int_\alpha^\beta \Big[\int_a^b f(K_1^y(u),y)\cdot |(K_1^y)'(u)|\,du \Big]\,dy \\
&\overset{4}{=} \int_S f(K(u,y))\cdot |\det K'(u,y)|\,dudy \\
&\overset{5}{=} \int_{H(U)} f(K(u,y))\cdot |\det K'(u,y)|\,dudy \\
&\overset{6}{=} \int_c^d \Big[\int_{H_2^u([\gamma,\delta])} f(K(u,y))\cdot |\det K'(u,y)|\,dy \Big] du \\
&\overset{7}{=} \int_c^d \Big[\int_\gamma^\delta f(K(u,H_2^u(v)))\cdot |\det K'(u,H_2^u(v))|\cdot |(H_2^u)'(v)|\,dv \Big]\,du \\
&\overset{8}{=} \iint_U f(K(H(u,v)))\cdot |\det K'(H(u,v))|\cdot |\det H'(u,v)|\,dudv \\
&\overset{9}{=} \iint_U f(T(u,v))\cdot |\det T'(u,v)|\,dudv \\
&\overset{10}{=} \iint_A f(T(u,v))\cdot |\det T'(u,v)|\,dudv
\end{aligned}$$

Overgang (1) gjelder fordi $f(x, y) = 0$ utenfor $K(S)$. Overgang (2) er Fubinis teorem brukt på beskrivelsen

$$y \in [\alpha, \beta],\ \ x \in K_1^y([a, b])$$

av området $K(S)$. I overgang (3) bruker vi integrasjon ved substitusjon for funksjoner av én variabel (teorem 7.4.2, bind I). Absoluttverdien rundt den deriverte av kjernen skyldes at intervallet $K_1^y([a, b])$ ikke er angitt med grenser. Vi vet altså ikke om $(K_1^y)'(u)$ er negativ eller positiv for $u \in [\alpha, \beta]$. Vi vet kun at den ulik 0 overalt. Overgang (4) er fordi $K(u, y) = (K_1^y(u), y)$. Overgang (5) er fordi $f(K(u, y))$ er 0 utenfor $H(U)$. Overgang (6) er Fubini brukt på beskrivelsen $u \in [c, d]$, $y \in H_2^u[\gamma, \delta]$ av området $H(U)$. I overgang (7) brukes igjen integrasjon ved substitusjon i form av teorem 7.4.2. Overgang (8) er fordi hvis $H(u, v) = (u, H_2^u(v))$. I overgang (9) bruker vi at $T(\mathbf{u}) = K(H(\mathbf{u}))$, samt kjerneregelen og produktregelen for determinanter. Overgang (10) er fordi $f(T(u, v))$ er 0 utenfor U. ■

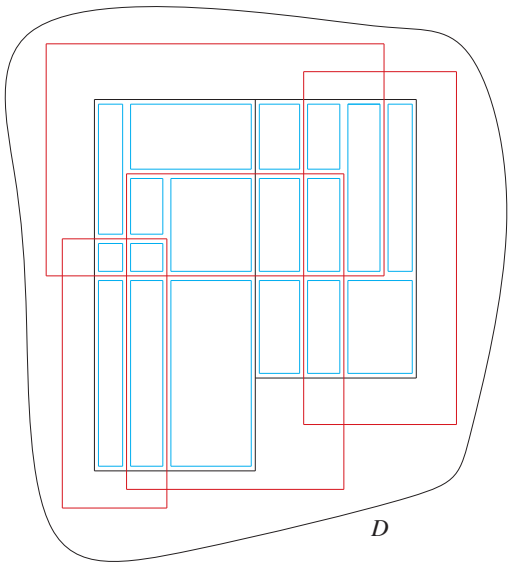
Da er vi klare for å bevise den fulle versjonen av koordinatskifteteoremet. Ved oppgave 3.8.1 kan vi anta at f er ikke-negativ. Gitt $\varepsilon > 0$, velg en nedresum $\mathcal{S}$ for f på et lukket rektangel som inneholder D slik at

$$\mathcal{S} \geq \iint_D f(x, y)\, dxdy - \varepsilon$$

Siden f er ikke-negativ og den utvidede funksjonen f_D er 0 i alle punkter utenfor D, kan vi droppe alle ledd i $\mathcal{S}$ som ikke tilsvarer rektangler inneholdt i D og likevel beholde ulikheten ovenfor. Vi har da

$$\mathcal{S} = \sum_{i=1}^{\alpha} c_i |R_i|,$$

der alle rektanglene R_i er innenfor D, og $c_i \leq f(x, y)$ for alle $(x, y) \in R_i$. For hvert punkt $(u, v) \in A$, velg et rektangel $E_{(u,v)}$ som i teorem 3.8.1. Da er $\{E_{(u,v)} \mid (u, v) \in A\}$ en åpen overdekning av den kompakte mengden $R_1 \cup \cdots \cup R_\alpha$, så ved kompakthetsteoremet fins et endelig utvalg $E_1, \ldots, E_\beta$ av $E_{(u,v)}$-ene som også dekker. Vi kan nå finne disjunkte, åpne rektangler $B_1, \ldots, B_m$ slik at hvert rektangel B_j er helt inneholdt i nøyaktig én av mengdene R_i og én av mengdene E_j, samtidig som avstanden mellom mengdene B_i er nedad begrenset av et positivt tall $\mu > 0$, og vi har $|B_1| + \cdots + |B_m| > |R_1| + \cdots + |R_\alpha| - \varepsilon$. Se figur 3.8.2.



Figur 3.8.2 Området D, rektanglene R_i (svarte), rektanglene E_i (røde) samt rektanglene B_i (lyseblå). Her er $\alpha = 2$.

Ideen videre er nå enkel: Vi erstatter funksjonen $f(x, y)$ med et «lappeteppe» av funksjoner

$$f(x, y) \cdot g_i(x, y)$$

for $i = 1, \ldots, m$, der $g_i = 1$ på rektanglet B_i og så går lineært ned til 0 som funksjon av avstanden til B_i, slik at g_i er kontinuerlig og $g_i = 0$ når avstanden til B_i er større enn eller lik $\mu/2$. Hvis $B_i \subseteq R_j$, så la $d_i = c_j$. Hvis M er slik at $|f(x, y)| < M$ på D, er da $\sum_{i=1}^{m} d_i |B_i| \geq \mathcal{S} - \varepsilon M$. Vi får nå

$$\begin{aligned}
&\iint_A f(T(u, v))|\det T'(u, v)|\, du dv \\
&\quad \geq \iint_A \sum_{i=1}^{m} f(T(u, v)) g_i(T(u, v))|\det T'(u, v)|\, du dv \\
&\quad = \sum_{i=1}^{m} \iint_A f(T(u, v)) g_i(T(u, v))|\det T'(u, v)|\, du dv \\
&\quad = \sum_{i=1}^{m} \iint_D f(x, y) g_i(x, y)\, dx dy \geq \sum_{i=1}^{m} \iint_{B_i} f(x, y)\, dx dy \\
&\quad \geq \sum_{i=1}^{m} d_i |B_i| \geq \mathcal{S} - \varepsilon M \geq \iint_D f(x, y)\, dx dy - \varepsilon - \varepsilon M
\end{aligned}$$

I første overgang brukte vi at f er ikke-negativ, i den andre teorem 4.1.4 (1) i bind II, og i tredje overgang den lokale varianten 3.8.1 av teoremet med $f(x, y) \cdot g_i(x, y)$ i rollen som $f(x, y)$. Siden $\varepsilon > 0$ var vilkårlig, har vi vist teoremet med ulikhet én vei.

For å vise teoremet med ulikhet motsatt vei, anvender vi resultatet vi nettopp viste, på den inverse koordinattransformasjonen. Vi kjører baklengs, altså. La $\hat{D} = A$, $\hat{A} = D$, $\hat{T} = T^{-1}$ og definer

$$\hat{f}(x, y) = f(T(x, y)) \cdot |\det T'(x, y)| \qquad \text{for alle } (x, y) \in \hat{D}.$$

Ved inversfunksjonsteoremet er $|\det \hat{T}'(x, y)| = |\det T'(\hat{T}(x, y))|^{-1}$ og $\hat{T}$ er C^1. Setter du inn de «hattede» versjonene i teoremet med ulikhet den veien vi har vist, faller ulikheten motsatt vei ut. ■

3.8 OPPGAVER

1. La $f(x, y)$ være en kontinuerlig skalarfunksjon definert på en åpen, begrenset mengde $D \subseteq \mathbb{R}^2$. For hver $(x, y) \in D$, la $f_+(x, y)$ være det største av tallene $f(x, y)$ og 0, og la $f_-(x, y)$ være det største av tallene $-f(x, y)$ og 0.

a) Vis at f_+ og f_- begge er ikke-negative og kontinuerlige på D, og at $f(x, y) = f_+(x, y) - f_-(x, y)$ for alle $(x, y) \in D$.

b) Bevis deretter at

$$\int_D f(x, y)\, dx dy = \int_D f_+(x, y)\, dx dy - \int_D f_-(x, y)\, dx dy$$

c) Anta at koordinatskifteteoremet holder med den ekstra forutsetningen at f er ikke-negativ på D. Vis at teoremet da holder generelt.

STIKKORDLISTE