TMA4212 Numerisk løsning av partielle differensialligninger med endelig differensemetoder

Brynjulf Owren

8. januar 2007

Forord

Dette notatet ble påbegynt vinteren 2004 til bruk i undervisningen i den ene halvdelen av kurset TMA4210 Numerisk løsning av partielle differensialligninger med differensemetoden. Vinteren 2006 ble det oppdatert med flere nye delkapitler og tilpasset kurset TMA4212. Jeg skylder en takk til studenter som fulgte kurset i de aktuelle semestrene for inspirasjon til skriving og for å ha gjort meg oppmerksom på mange trykkfeil i tidlige versjoner av notatet.

Innhold

1 Innledning **1**

2 Bakgrunnsstoff **3**
2.1 Litt repetisjon av matriseteori . . . 3
2.1.1 Jordanformen . . . 3
2.1.2 Symmetriske matriser . . . 4
2.1.3 Positiv definitte matriser . . . 4
2.1.4 Gershgorins teorem . . . 4
2.1.5 Vektor- og matrisenormer . . . 5
2.1.6 Forenlige og tilordnede matrisenormer . . . 6
2.1.7 Matrisenormer og spektralradius . . . 7
2.2 Differenseformler . . . 8
2.2.1 Taylorutvikling . . . 8
2.2.2 Stor $\mathcal{O}$-notasjon . . . 9
2.2.3 Differenseapproksimasjoner til deriverte . . . 9
2.2.4 Differensoperatorer og andre operatorer . . . 11
2.2.5 Differensialoperatoren . . . 12

3 Diskretisering av varmeledningsligningen **15**
3.1 Om utledning av varmeledningsligningen . . . 15
3.2 Numerisk løsning av start/randverdiproblemet . . . 16
3.2.1 Numerisk approksimasjon på gitter . . . 16
3.2.2 Euler, Baklengs Euler og Crank–Nicolson . . . 17
3.2.3 Løsning av ligningene i Baklengs Euler og Crank–Nicolson . . . 21
3.2.4 Løsning av ligninger med Matlab . . . 22
3.2.5 θ-metoden . . . 23
3.3 Semidiskretisering . . . 24
3.3.1 Semidiskretisering av varmeledningsligningen . . . 24
3.3.2 Semidiskretiseringsprinsippet generelt . . . 25
3.3.3 Formalisering . . . 26
3.3.4 $u_t = Lu$ med forskjellige valg av L . . . 27
3.4 Randkrav med derivert . . . 29
3.4.1 Ulike typer randkrav . . . 29
3.4.2 Diskretisering av randkrav . . . 30
3.5 Ikke-lineære parabolske differensialligninger . . . 32

4 Stabilitet, konsistens og konvergens **35**
4.1 Egenskaper ved det kontinuerlige problemet . . . 35
4.2 Konvergens av numerisk metode . . . 36
4.3 Avhengighetsområde for en numerisk metode . . . 37
4.4 Konvergensbevis for Eulers metode på (S/R) med $r \leq \frac{1}{2}$ . . . 38
4.5 Stabilitet på ubegrenset tidsintervall (F-stabilitet) . . . 39
4.6 Stabilitet på $[0,T]$ når $h \to 0$, $k \to 0$ . . . 41
4.7 Stabilitet og avrundingsfeil . . . 45
4.8 Konsistens og Lax' ekvivalensteorem . . . 46
4.9 von Neumanns stabilitetskriterium . . . 47

5 Elliptiske differensialligninger **51**
5.1 Elliptisk ligning i planet . . . 51
5.2 Differensmetoder via Taylor . . . 52
5.2.1 Diskretisering av en selvadjungert ligning . . . 54
5.3 Randkrav av Neumanns og Robins type . . . 55
5.4 Gitterlignende nett og varierende skrittlengder . . . 57
5.5 Generelt rektangulært nett . . . 58
5.6 Diskretisering via Taylor på fullstendig gerenelt nett . . . 59
5.7 Differensformler utledet via integrasjon . . . 60
5.8 Nett basert på trekanter . . . 63
5.9 Differensligningene . . . 63
5.10 Konvergens av metoder for elliptiske ligninger . . . 65
5.10.1 Konvergensbevis for 5-punktsformelen på et Dirichletproblem . . . 65
5.10.2 Noen generelle kommentarer om konvergens . . . 66
5.11 Kommentarer om løsningsmetoder . . . 67

6 En introduksjon til endelige elementmetoder **69**
6.1 Introduksjon . . . 69
6.2 Tre ekvivalente problemer . . . 70
6.2.1 Noen definisjoner . . . 70
6.2.2 Ekvivalente problemer . . . 70
6.3 Tilnærmet løsning av variasjonsproblemet . . . 73
6.3.1 Generell framgangsmåte . . . 73
6.3.2 En viktig egenskap ved løsning av $(\mathbf{V}_h)$ . . . 74
6.3.3 Minimaliseringsproblemet og navnekonvensjoner . . . 74
6.3.4 Endimensjonalt problem . . . 74
6.3.5 Todimensjonalt problem – Endelig elementmetode med triangulære elementer og lineære elementfunksjoner . . . 76
6.3.6 Konstruksjon av formfunksjonene . . . 77
6.3.7 Basisfunksjoner for S_h, pyramidefunksjoner . . . 78
6.3.8 Elementmetodeløsning av (D) . . . 79
6.3.9 Beregning av stivhetsmatrisen ved innaddering . . . 79
6.3.10 Et omfattende eksempel uten innaddering . . . 81
6.3.11 Feilen i U . . . 83
6.4 Problem med inhomogene randkrav . . . 83
6.4.1 Tilnærmet løsning av inhomogent problem . . . 84
6.5 Andre elementer og elementfunksjoner . . . 84
6.6 Generelle 2. ordens differensialligninger . . . 87

7 Hyperbolske ligninger **89**
7.1 Eksempler på ligninger . . . 89
7.2 Karakteristikker . . . 90
7.3 Eksplisitte differenseformler for $u_t + au_x = 0$ . . . 92
7.4 Stabilitet . . . 93
7.5 Implisitte metoder for $u_t + au_x = 0$ . . . 96
7.6 Hyperbolske systemer av første ordens ligninger . . . 97
7.7 Dissipasjon og dispersjon . . . 104

Kapittel 1

Innledning

Numerisk approksimasjon av partielle differensialligninger utgjør en betydelig del av all virksomhet innen simulering av prosesser i naturen eller i samfunnet forøvrig. Som eksempler på fagfelt der slike ligninger brukes kan vi ta med kjemiske prosesser, fluidmekanikk, strukturmekanikk, kvantefysiske prosesser, elektromagnetisme, finans, osv. Når vi snakker om partielle differensialligninger mener vi en ligning hvis løsning er en funksjon (eller en vektor av funksjoner) av minst 2 variable, disse kalles uavhengige variable. Ligningen beskriver en relasjon der løsningen og dens partiellderiverte av forskjellig orden inngår. Men spesifikasjonen av en matematisk modell i anvendelser involverer langt mer en kun denne relasjonen eller ligningen. Først og fremst er modellen gjerne knyttet til en *geometri*. Dette betyr at vi spesifiserer et domene i rommet av de uavhengige variable der diffligningen skal gjelde. Dette domenet kan være endelig eller uendelig i utstrekning. I 2 dimensjoner kan det for eksempel være et delområde av planet, men det kan også være helt andre geometrier som overflaten på en sylinder eller en kule. Vanligvis er det slik at differensialligningen selv har uendelig mange løsninger. Spesifikasjonen av en ligning på et domene må derfor suppleres med et sett av randbetingelser (randkrav). Disse kravene kan se litt forskjellige ut, men generelt kan en si at de spesifiserer hva løsningen (og/eller dens deriverte) skal være på randen av domenet. Dersom en av de uavhengige variablene er fysisk tid, så kaller man betingelsen ved starttidspunktet for initialbetingelsen, dette er kun en spesiell form for randkrav.

De fleste partielle differensialligninger har det til felles at man ikke kan skrive opp den eksakte løsningen på noen enkel måte. Man trenger derfor approksimative metoder som kan implementeres på en datamaskin. Det fins flere hovedklasser av metoder, en av dem er differensemetoder som behandles i dette kurset. Andre teknikker er spektralmetoder og endelig elementmetoder. Disse lærer man om i TMA4220.

Partielle differensialligninger kan være lineære eller ikke-lineære. Siden dette er en elementær introduksjon, skal vi stort sett holde oss til det lineære tilfellet. Slike lineære partielle differensialligninger kan deles inn i tre kategorier, parabolske, elliptiske og hyperbolske differensialligninger. I dette kurset ser vi kun på de første to typene, men doktorgradskurset MA8103 Ikke-lineære partielle differensialligninger handler mest om hyperbolske ligninger. Som en prototype på parabolske ligninger vil vi holde oss mye til varmeledningsligningen, og kapittel 3 dreier seg mest om løsning av denne. Her arbeider vi stort sett med 2 uavhengige variable, tid og en romdimensjon. På grunn av dette er det få utfordringer relatert til problemets geometri som diskutert her.

Kapittel 4 tar for seg noen generelle egenskaper for differenseapproksimasjoner, som er av betydning når man løser partielle differensialligninger i alle de tre nevnte kategorier ovenfor. Det dreier seg i siste instans om konvergens, men vi diskuterer også begrepene

stabilitet og konsistens som brukes i diskusjonen av konvergens av differenseskjemaet.

I kapittel 5 diskuteres elliptiske ligninger, og prototyper er her Laplace's ligning og Poisson's ligning. Vi arbeider fremdeles med todimensjonale domener, men her er begge de tilhørende variable fysiske romvariable. Derfor er det interessant å se på mer komplekse geometrier, der domenet ikke er et rektangel slik det typisk var i kapittel 3. Konsekvensen blir at mye av diskusjonen dreier seg om hvordan man skal approksimere ulike randbetingelser for domener som ikke er rektangulære.

Men før vi begynner å se på partielle differensialligninger skal vi friske opp noe grunnleggende matriseteori samt litt Taylorutvikling, og definere differenseoperatorer i kapitlet om bakgrunnsstoff.

Kapittel 2

Bakgrunnsstoff

2.1 Litt repetisjon av matriseteori

La A være en $n \times n$-matrise av reelle (eller komplekse) tall, vi skriver $A \in \mathbf{R}^{n\times n}$ (eller $A \in \mathbf{C}^{n\times n}$). Vi sier at A *er diagonaliserbar* hvis det fins en matrise $X \in \mathbf{C}^{n\times n}$ slik at

$$\Lambda = X^{-1}AX = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix}$$

Hver $\lambda_i \in \mathbf{C}$ kalles en *egenverdi* til A. Matrisen X består av n kolonner $X = [x_1, \ldots, x_n]$ der hver $x_i \in \mathbf{C}^n$ kalles *egenvektoren* (*tilhørende egenverdien* λ_i). En diagonalmatrise som Λ ovenfor, skrives av og til som

$$\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n).$$

2.1.1 Jordanformen

Til enhver $A \in \mathbf{R}^{n\times n}$ (eller $A \in \mathbf{C}^{n\times n}$) fins en matrise $M \in \mathbf{C}^{n\times n}$ slik at

$$M^{-1}AM = J = \begin{bmatrix} J_1 & & & \\ & J_2 & & \\ & & \ddots & \\ & & & J_k \end{bmatrix} \quad \text{(blokkdiagonal)} \tag{2.1}$$

Her er J_i en $m_i \times m_i$-matrise, og $\sum_{i=1}^{k} m_i = n$. *Jordanblokkene* J_i er av formen

$$J_i = \begin{bmatrix} \lambda_i & 1 & & \\ & \lambda_i & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_i \end{bmatrix}, \quad \text{hvis } m_i \geq 2$$

og $J_i = [\lambda_i]$ hvis $m_i = 1$. Så hvis alle $m_i = 1$, er $k = n$ og matrisen diagonaliserbar. Hvis A har n forskjellige egenverdier, er den alltid diagonaliserbar. Men det motsatte utsagnet gjelder ikke, dvs en matrise kan være diagonaliserbar selv om den har sammenfallende egenverdier.

2.1.2 Symmetriske matriser

Når vi snakker om symmetriske matriser, mener vi som regel *reelle* symmetriske matriser. Den *transponerte* A^T av en $m \times n$-matrise A, er $n \times m$-matrisen med (ij)-element a_{ji} (en matrise hvis kolonner er radene i A. En $n \times n$ matrise er symmetrisk hvis $A^T = A$.

En symmetrisk $n \times n$ matrise har reelle egenverdier $\lambda_1, \ldots, \lambda_n$ og et sett av reelle ortonormale egenvektorer $x_1, \ldots, x_n$. Lar vi standard indreprodukt på $\mathbf{C}^n$ betegnes $\langle \cdot, \cdot \rangle$, gjelder altså $\langle x_i, x_j \rangle = \delta_{ij}$ (Kronecker-delta).

En konsekvens av dette blir at egenvektormatrisen $X = [x_1, \ldots, x_n]$ blir reell og ortogonal og dens inverse er derfor den transponerte

$$X^{-1} = X^T.$$

Diagonaliseringen er gitt som

$$\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n), \quad X = [x_1, \ldots, x_n], \quad X^T X = I, \quad X^T A X = \Lambda \Leftrightarrow A = X \Lambda X^T$$

2.1.3 Positiv definitte matriser

Hvis A er symmetrisk og $\langle x, Ax \rangle = x^T Ax > 0$ for alle $0 \neq x \in \mathbf{R}^n$ kalles A *positiv definitt*.

A (symmetrisk) er positiv semidefinitt hvis $\langle x, Ax \rangle \geq 0$ for alle $x \in \mathbf{R}^n$ og $\langle x, Ax \rangle = 0$ for minst en $x \neq 0$.

A positiv definitt $\Leftrightarrow$ A har bare positive egenverdier

A positiv semidefinitt $\Leftrightarrow$ A har bare ikke-negative egenverdier, og minst en 0-egenverdi.

2.1.4 Gershgorins teorem

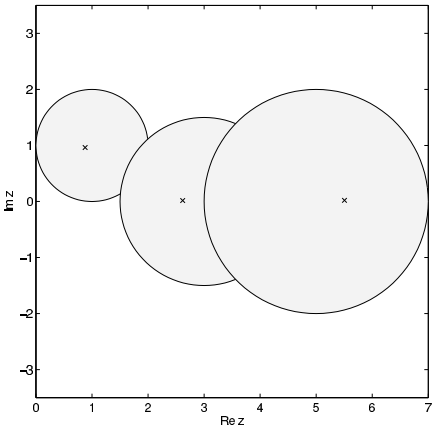
Gershgorins teorem. Gitt $A = (a_{ik}) \in \mathbf{C}^{n \times n}$. Definer n sirkelflater S_j i det komplekse plan ved

$$S_j = \left\{ z \in \mathbf{C} : |z - a_{jj}| \leq \sum_{k \neq j} |a_{jk}| \right\}.$$

Unionen $S = \bigcup_{j=1}^{n} S_j$ inneholder alle egenverdiene til A. Det vil si at for enhver egenverdi λ til A fins en j slik at $\lambda \in S_j$.

Eksempel.

$$A = \begin{bmatrix} 1+i & 1 & 0 \\ 0.5 & 3 & 1 \\ 1 & 1 & 5 \end{bmatrix}$$



,

□

Bevis av Gershgorins teorem: La λ være en egenverdi med tilhørende egenvektor $x = [\xi_1, \ldots, \xi_n]^T \neq 0$. Velg deretter ℓ blant indeksene $1, \ldots, n$ slik at $|\xi_\ell| \geq |\xi_k|$, $k = 1, \ldots, n$, klart at $|\xi_\ell| > 0$. Ligningen $Ax = \lambda x$ har komponent ℓ:

$$\sum_{k=1}^{n} a_{\ell k} \xi_k = \lambda \xi_\ell \;\Rightarrow\; (\lambda - a_{\ell\ell}) \xi_\ell = \sum_{k \neq \ell} a_{\ell k} \xi_k$$

Divider med $|\xi_\ell|$ på hver side og ta absoluttverdier

$$|\lambda - a_{\ell\ell}| \quad = \quad \left| \sum_{k \neq \ell} a_{\ell k} \frac{\xi_k}{\xi_\ell} \right| \quad \leq \quad \sum_{k \neq \ell} |a_{\ell k}| \frac{|\xi_k|}{|\xi_\ell|} \quad \leq \quad \sum_{k \neq \ell} |a_{\ell k}|$$

Altså er $\lambda \in S_\ell$.

Eksempel. Diagonaldominante matriser med positiv diagonal er positiv definitt. Hvorfor?

2.1.5 Vektor- og matrisenormer

Gitt et vektorrom X (reelt eller komplekst). En norm $\| \cdot \| : X \to \mathbf{R}$ oppfyller følgende aksiomer

1. $\|x\| \geq 0$ for alle x, $\|x\| = 0 \Leftrightarrow x = 0$.
2. $\|\alpha x\| = |\alpha| \|x\|$ $\quad (\alpha \in \mathbf{R}\ (\mathbf{C}))$
3. $\|x + y\| \leq \|x\| + \|y\|$

Eksempler. $x = (\xi_k)$, $X = \mathbf{R}^n$.

$$\|x\|_1 = \sum_{k=1}^{n} |\xi_k|, \qquad \|x\|_2 = \left(\sum_{k=1}^{n} |\xi_k|^2 \right)^{1/2}, \qquad \|x\|_\infty = \max_{1 \leq k \leq n} |\xi_k|$$

Matriserommene $\mathbf{R}^{n \times n}$ og $\mathbf{C}^{n \times n}$ er også vektorrom over $\mathbf{R}$ ($\mathbf{C}$). Vi kaller $\| \cdot \|$ en matrisenorm hvis for alle $A, B \in \mathbf{R}^{n \times n}$ ($\mathbf{C}^{n \times n}$)

1. $\|A\| > 0$ for alle A, $\|A\| = 0 \Leftrightarrow A = 0$.
2. $\|\alpha A\| = |\alpha| \|A\| \quad (\alpha \in \mathbf{R}\ (\mathbf{C}))$
3. $\|A + B\| \leq \|A\| + \|B\|$
4. $\|A \cdot B\| \leq \|A\| \cdot \|B\|$

Merknad. Det siste punktet krever at et matrise-matrise produkt er definert (dette er ingen operasjon som generelt er definert i vektorrom). I abstrakte termer er aksiomene 1–4 et eksempel på en *Banach-algebra*.

Eksempel. Frobeniusnormen til en matrise er definert som

$$\|A\|_F = \left(\sum_{j=1}^{n} \sum_{k=1}^{n} |a_{jk}|^2 \right)^{1/2}$$

2.1.6 Forenlige og tilordnede matrisenormer

En gitt matrisenorm er *forenlig* med en gitt vektornorm på $\mathbf{R}^n$ hvis

$$\|Ax\| \leq \|A\| \cdot \|x\| \quad \text{for alle} \quad A \in \mathbf{R}^{n \times n},\ x \in \mathbf{R}^n.$$

En gitt matrisenorm er *tilordnet* en gitt vektornorm på $\mathbf{R}^n$ hvis

$$\|A\| = \max_{x \neq 0} \frac{\|Ax\|}{\|x\|}$$

Eksempler. Vi gjengir her noen av de mest vanlige tilordnede matrisenormene, vi ser på hvilke matrisenormer som er tilordnet de tre vektornormene $\|\cdot\|_1$, $\|\cdot\|_2$ og $\|\cdot\|_\infty$.

1. Vi lar $\|\cdot\|_1$ betegne matrisenormen som tilordnes vektornormen $\|\cdot\|_1$. Det er mulig å vise at for $A \in \mathbf{R}^{n \times n}$ $(\mathbf{C}^{n \times n})$ er

   $$\|A\|_1 = \max_{1 \leq k \leq n} \sum_{i=1}^{n} |a_{ik}|$$

   Med ord kan man si at $\|A\|_1$ er "maksimal kolonnesum i A".

2. For å skrive ned matrisenormen som tilordnes vektornormen $\|\cdot\|_2$ må vi først definere *spektralradien* til en matrise $M \in \mathbf{R}^{n \times n}$ $(\mathbf{C}^{n \times n})$. Hvis $\lambda_1, \ldots, \lambda_n$ er egenverdiene til M, betegnes spektralradien til M for $\rho(M)$, og den er definert som

   $$\rho(M) = \max_{1 \leq k \leq n} |\lambda_k| \tag{2.2}$$

   Om vi plotter egenverdiene til M i det komplekse plan, er spektralradien til M radien til den minste mulige sirkelskive, sentrert i origo, som inneholder alle egenverdiene til M.

   Vi definerer nå 2-normen til en matrise A ved

   $$\|A\|_2 = \sqrt{\rho(A^T A)}$$

   Merk at $A^T A$ er positiv (semi)definitt, så alle egenverdiene er reelle og positive. Tar vi kvadratroten av den største av dem, finner vi $\|A\|_2$. Merk også at spektralradien til A kan være svært forskjellig fra (kvadratroten av) spektralradien til $A^T A$. Er A imidlertid symmetrisk vil $\|A\|_2 = \rho(A)$.

3. La $\|\cdot\|_\infty$ betegne matrisenormen som tilordnes vektornormen $\|\cdot\|_\infty$. En har da at

   $$\|A\|_\infty = \max_{1 \leq i \leq n} \sum_{k=1}^{n} |a_{ik}|$$

   Det vil si at $\|A\|_\infty$ er "maksimal linjesum i A". Merk forøvrig at $\|A\|_1 = \|A^T\|_\infty$.

2.1.7 Matrisenormer og spektralradius.

For vilkårlig matrisenorm $\|\cdot\|$ gjelder at

$$\|A\| \geq \rho(A) \tag{2.3}$$

Bevis: La x være en egenvektor til A tilhørende en egenverdi λ slik at

$$Ax = \lambda x$$

La nå $y \in \mathbf{C}^n$ være vilkårlig. Da er

$$A(xy^T) = (Ax)y^T = \lambda(xy^T)$$

slik at

$$\|A(xy^T)\| \leq \|A\| \, \|xy^T\|$$

Så

$$|\lambda| \, \|xy^T\| = \|\lambda(xy^T)\| = \|A(xy^T)\| \leq \|A\| \, \|xy^T\|$$

Dermed er $|\lambda| \leq \|A\|$, og siden dette gjelder for hver av egenverdiene til A, må $\rho(A) \leq \|A\|$.

Spørsmål til leseren: Hva er galt med argumentet

$$|\lambda| \, \|x\| = \|\lambda x\| = \|A x\| \leq \|A\| \, \|x\| \quad \text{osv}$$

Konvergent matrise. En matrise A kalles *konvergent* (mot null) hvis

$$A^k \to 0 \quad \text{når} \quad k \to \infty$$

Tilstrekkelig betingelse. Hvis $\|A\| < 1$ for en eller annen matrisenorm er A konvergent.

Bevis.

$$\|A^k\| = \|A \cdot A^{k-1}\| \leq \|A\| \cdot \|A^{k-1}\| \leq \cdots \leq \|A\|^k \ \to 0 \qquad \text{hvis } \|A\| < 1$$

Nødvendig og tilstrekkelig betingelse. A er konvergent hvis og bare hvis spektralradien $\rho(A)$, definert ved (2.2), oppfyller $\rho(A) < 1$.

Bevis (ikke pensum): Vi bruker Jordanformen, og lar $A = MJM^{-1}$ der $M \in \mathbf{C}^{n \times n}$ og J er som i (2.1). Da er $A^2 = MJM^{-1}MJM^{-1} = MJ^2M^{-1}$, og ved induksjon er $A^k =$

MJ^kM^{-1}. At $A^k \to 0$ er ekvivalent med at $J^k \to 0$. At $J^k \to 0$ er ekvivalent med at hver enkelt Jordanblokk $J_i^k \to 0$. Anta at en slik Jordanblokk har diagonalelement λ og at $m_i \times m_i$-matrisen F har sitt $(i, i+1)$-element lik 1, og de øvrige elementer er null. Da er $J_i = \lambda I + F$ der I er identitetsmatrisen. Matrisen F er nilpotent, dvs $F^m = 0,\ m \geq n$. Vi antar at $k \geq n-1$ og regner ut

$$J_i^k = (\lambda I + F)^k = \sum_{m=0}^{k} \binom{k}{m} \lambda^{k-m} F^m = \sum_{m=0}^{n-1} \binom{k}{m} \lambda^{k-m} F^m = \sum_{m=0}^{n-1} \varphi_k^{(m)}(\lambda) F^m$$

der $\varphi_k(\lambda) = \lambda^k/k!$. Når $k \to \infty$ vil $\varphi_k^{(m)}(\lambda) \to 0$ for $0 \leq m \leq n-1$ hvis og bare hvis $|\lambda| < 1$. Dette må gjelde for alle Jordanblokkene (dvs egenverdiene til A) og resultatet følger. □

2.2 Differenseformler

2.2.1 Taylorutvikling

1 fri variabel. La $u \in C^{n+1}(I)$ hvor $I \subset \mathbf{R}$ er et intervall på tallinja. Dette betyr at den $n+1$'te deriverte av u eksisterer og er kontinuerlig på intervallet I. Da gjelder

Taylors formel med restledd. Med $x \in I$, $x + h \in I$ er

$$u(x+h) = \sum_{m=0}^{n} \frac{h^m}{m!} u^{(m)}(x) + r_n$$

der

$$r_n = \frac{h^{n+1}}{(n+1)!} u^{(n+1)}(x + \theta h), \quad 0 < \theta < 1.$$

2 frie variabler. Anta nå at $u \in C^{n+1}(\Omega)$ hvor $\Omega \subset \mathbf{R}^2$. Det er gunstig for oss å bruke operatornotasjon for partiellderiverte. Vi skriver $\mathbf{h} = [h, k]$, og lar

$$\mathbf{h} \cdot \nabla := h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y} \quad \text{dvs} \quad \mathbf{h} \cdot \nabla u = h\frac{\partial u}{\partial x} + k\frac{\partial u}{\partial y}$$

Det operatoren gjør er å derivere en funksjon i retning $\mathbf{h} = [h, k]$, man finner altså den *retningsderiverte*.

Vi kan også definere potenser av operatoren ved f.eks.

$$(\mathbf{h} \cdot \nabla)^2 = \left(h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y}\right)^2 = h^2\frac{\partial^2}{\partial x^2} + 2hk\frac{\partial^2}{\partial x \partial y} + k^2\frac{\partial^2}{\partial y^2}$$

Utvidelse til m'te potens er opplagt. Da kan vi skrive opp

Taylors formel med restledd for funksjoner av to variable.

$$u(x+h, y+k) = \sum_{m=0}^{n} \frac{1}{m!} \left(h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y}\right)^m u(x,y) + r_n \tag{2.4}$$

der

$$r_n = \frac{1}{(n+1)!} \left(h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y}\right)^{n+1} u(x+\theta h, y+\theta k), \quad 0 < \theta < 1.$$

Vi har her forutsatt at det rette linjestykket mellom (x, y) og $(x+h, y+k)$ tilhører Ω.

Om utledning. En ser på en funksjon av 1 variabel $\mu(t) = u(x+th, y+tk)$ for fikserte x, y, h, k. Bruk Taylorutvikling med restledd i en variabel for $\mu(t)$ omkring $t = 0$ og sett til slutt inn $t = 1$.

2.2.2 Stor $\mathcal{O}$-notasjon

La ϕ være en funksjon av h og p et positivt heltall. Da er

$$\phi(h) = \mathcal{O}(h^p) \quad \text{når} \quad h \to 0$$

hvis det fins konstanter $C, H > 0$ slik at

$$|\phi(h)| \leq C|h|^p \quad \text{når} \quad 0 < |h| < H$$

Hvis dette holder, sier man gjerne at $\phi(h)$ er av *orden* p i variabelen h.

Den typiske bruken vi har for slik stor $\mathcal{O}$-notasjon er i forbindelse med lokal avbruddsfeil i numeriske metoder. For eksempel i Taylorutvikling med en variabel vil

$$|r_n| = \left|\frac{h^{n+1}}{(n+1)!} u^{(n+1)}(x+\theta h)\right| \leq \frac{M}{(n+1)!}|h|^{n+1}, \quad M = \max_{y \in I} |u^{(n+1)}(y)|$$

der vi vet at dette maksimumet eksisterer under de ovenfor nevnte forutsetninger hvis I er et lukket, begrenset intervall. Så i dette tilfellet følger at $r_n = \mathcal{O}(h^{n+1})$.

Merk at med definisjonen ovenfor er det slik at for positive heltall p vil

$$\phi(h) = \mathcal{O}(h^{p+1}) \quad \Rightarrow \quad \phi(h) = \mathcal{O}(h^p).$$

Derfor er det av og til underforstått at når vi sier at $\phi(h) = \mathcal{O}(h^p)$ mener vi at p er det største mulige heltall slik at dette gjelder. Ofte er det hensiktsmessig å skrive $\mathcal{O}(h^p)$ i uttrykk med summer, f.eks. kunne vi for Taylorrekka til u ovenfor erstattet r_n med $\mathcal{O}(h^{n+1})$ slik at

$$u(x+h) = \sum_{k=0}^{n} \frac{h^k}{k!} u^{(m)}(x) + \mathcal{O}(h^{n+1})$$

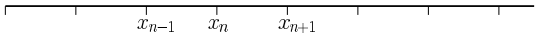
Generelt har vi at dersom $\phi(h) = \mathcal{O}(h^{p_\phi})$ og $\psi(h) = \mathcal{O}(h^{p_\psi})$, så gjelder at

$$\psi(h) + \phi(h) = \mathcal{O}(h^q), \quad \text{der} \quad q = \min(p_\phi, p_\psi)$$

Men noen ganger kan man oppnå høyere potens. Et opplagt eksempel er hvis $\phi(h) = h^2$, $\psi(h) = h^3 - h^2$, hver for seg er de $\mathcal{O}(h^2)$ mens summen er $\mathcal{O}(h^3)$. Om man multipliserer en funksjon $\phi(h)$ med en konstant ($\neq 0$), blir ordenen uendret.

2.2.3 Differenseapproksimasjoner til deriverte

Vi introduserer et *gitter* på $\mathbf{R}$ dvs en monoton tallfølge $\{x_n\}$ der hver $x_n \in \mathbf{R}$.



Anta at $u(x)$ er en gitt funksjon, $u \in C^q(I)$ for en q som spesifiseres senere. La

$$u_n := u(x_n), \qquad u_n^{(m)} := u^{(m)}(x_n)$$

Anta at gitterpunktene x_n er ekvidistante, dvs $x_{n+1} = x_n + h$ for alle n, der $h \in \mathbf{R}$ kalles *skrittlengden*. Vi ønsker å tilnærme $u_n^{(m)}$ med et uttrykk av formen

$$\sum_{\ell=p}^{q} a_\ell u_{n+\ell}$$

$p \le q$ er heltall, og typisk er $p \le 0$ og $q \ge 0$.

Avbruddsfeil. Vi definerer her

$$\tau_n(h) = \sum_{\ell=p}^{q} a_\ell u_{n+\ell} - u_n^{(m)}$$

Strategien er å velge p og q, og deretter beregne de $q-p+1$ parametrene $a_p, \ldots, a_q$ slik at τ_n blir "liten".

Ved Taylorutvikling får vi

$$u_{n+\ell} = u(x_n + \ell h) = \sum_{k=0}^{\nu} \frac{(\ell h)^k}{k!} u_n^{(k)} + r_\nu$$

der $r_\nu = \mathcal{O}(h^{\nu+1})$ slik at

$$\tau_n = \sum_{\ell=p}^{q} a_\ell \sum_{k=0}^{\nu} \frac{1}{k!} (\ell h)^k u_n^{(k)} - u_n^{(m)} + \mathcal{O}(h^{\nu+1})$$

Vi ønsker at $\tau_n = \mathcal{O}(h^r)$ med r så stor som mulig. Skal vi approksimere $u_n^{(m)}$ må vi sette krav på p og q. Sett $j := q-p$. Vi må kreve $j > m$. Vi velger så $a_p, \ldots, a_q$ slik at

$$\frac{1}{k!} \sum_{\ell=p}^{q} (\ell h)^k a_\ell = \begin{cases} 0 & 0 \le k \le m-1 \\ 1 & k = m \\ 0 & m+1 \le k \le j \end{cases} \tag{2.5}$$

Merk at vi har $q-p+1 = j+1$ fri parametre $a_p, \ldots, a_q$ til rådighet, og betingelsene i (2.5) gjelder for $0 \le k \le j$, det vil si totalt $j+1$ krav. Ligningssystemet har entydig løsning for alle $h \neq 0$. Velger vi a_ℓ fra (2.5), og antar $\nu \ge j$, får vi for avbruddsfeilen

$$\tau_n = \sum_{k=j+1}^{\nu} \frac{h^k}{k!} u_n^{(k)} \sum_{\ell=p}^{q} a_\ell \ell^k + \mathcal{O}(h^{\nu+1})$$

Metoden kalles *ubestemte koeffisienters metode*.

Eksempel. $m = 1$ (u_n'). Velg $p = -1$, $q = 1$, $j = 2$. Skal bestemme a_{-1}, a_0, a_1. Vi setter opp $j+1 = 3$ ligninger dvs $k = 0, 1, 2$ i (2.5).

$$\left.\begin{array}{rrcrcrcl} k=0 & a_{-1} & + & a_0 & + & a_1 & = & 0 \\ k=1 & -h\,a_{-1} & + & 0\cdot a_0 & + & h\,a_1 & = & 1 \\ k=2 & h^2\,a_{-1} & + & 0\cdot a_0 & + & h^2\,a_1 & = & 0 \end{array}\right] \Rightarrow \begin{array}{rcl} a_{-1} & = & -\frac{1}{2h} \\ a_0 & = & 0 \\ a_1 & = & \frac{1}{2h} \end{array}$$

Ser vi på de første leddene i avbruddsfeilen får vi

$$\tau_n = \sum_{k=3} \frac{h^k}{k!} u_n^{(k)} \left(-\frac{1}{2h}(-1)^k + \frac{1}{2h} 1^k \right) = \sum_{s=1} \frac{u_n^{(2s+1)}}{(2s+1)!} h^{2s}$$

I den siste likheten, har vi brukt at leddene med like k forsvinner, slik at vi kan sette $k = 2s+1$ og la s løpe fra 1 og oppover. Vi har med hensikt utelatt øvre grense i summen fordi antall ledd vi tar med før restleddet avhenger av omstendighetene. Siden det første leddet i uttrykket for τ_n er av type h^2, sier vi at formelen er av *orden* 2.

Noen flere formler. Andre populære differenseapproksimasjoner er

$$\begin{array}{lll} m=1 & \dfrac{u_{n+1}-u_n}{h} & = u_n' + \frac{1}{2!} h\, u_n'' + \cdots \\ m=1 & \dfrac{u_n-u_{n-1}}{h} & = u_n' - \frac{1}{2!} h\, u_n'' + \cdots \\ m=2 & \dfrac{u_{n+1}-2u_n+u_{n-1}}{h^2} & = u_n'' + \frac{1}{12} h^2 u_n^{(4)} + \cdots \end{array}$$

Hva slags orden har de tre formlene?

2.2.4 Differensoperatorer og andre operatorer

Foroverdifferens:	$\Delta u(x) = u(x+h) - u(x)$
Bakoverdifferens:	$\nabla u(x) = u(x) - u(x-h)$
Sentraldifferens:	$\delta u(x) = u(x+\frac{h}{2}) - u(x-\frac{h}{2})$
Middelverdi:	$\mu u(x) = \frac{1}{2}\left(u(x+\frac{h}{2}) - u(x-\frac{h}{2})\right)$
Forskyvning:	$E u(x) = u(x+h)$
Enhetsoperator	$1\, u(x) = u(x)$

Linearitet. Alle operatorene

$$\Delta,\ \nabla,\ \delta,\ \mu,\ E,\ 1,$$

er lineære. Dette betyr at for $\alpha \in \mathbf{R}$, og med funksjoner $u(x)$ og $v(x)$ vil

$$F(\alpha u(x) + v(x)) = \alpha F u(x) + F v(x).$$

der F kan være hvilken som helst av operatorene ovenfor. La oss sjekke for eksempel for $F = \Delta$.

$$\begin{array}{rcl} \Delta(\alpha u(x) + v(x)) & = & (\alpha u(x+h) + v(x+h)) - (\alpha u(x) + v(x)) \\ & = & \alpha(u(x+h) - u(x)) + (v(x+h) - v(x)) = \alpha \Delta u(x) + \Delta v(x) \end{array}$$

Potenser av operatorer. La igjen F være en vilkårlig valgt operator blant dem vi definerte ovenfor. Vi kan definere potenser av F som følger

$$F^0 = 1, \qquad F^k u(x) = F(F^{k-1}u(x))$$

Eksempel.

$$\begin{aligned}
\delta u(x) &= u(x+\tfrac{h}{2}) - u(x-\tfrac{h}{2}) \\
\delta^2 u(x) &= \delta(\delta u(x)) = \delta u(x+\tfrac{h}{2}) - \delta u(x-\tfrac{h}{2}) = u(x+h) - u(x) - (u(x) - u(x-h)) \\
&= u(x+h) - 2u(x) + u(x-h)
\end{aligned}$$

Et annet interessant eksempel er forskyvningsoperatoren. Vi ser at $E^k u(x) = u(x+kh)$. Her kan vi faktisk utvide potenser fra å være ikke-negative heltall til alle mulige reelle tall, simpelthen ved å definere $E^s u(x) = u(x+sh)$ for alle $s \in \mathbf{R}$. For eksempel er $E^{-1}u(x) = u(x-h)$ og dette er den inverse til E siden $Eu(x-h) = E^{-1}u(x+h) = u(x)$.

Sammenhenger mellom differenseoperatorer.

$$\begin{aligned}
\Delta u(x) &= u(x+h) - u(x) = Eu(x) - 1\,u(x) = (E-1)\,u(x) \\
\nabla u(x) &= u(x) - u(x-h) = 1\,u(x) - E^{-1}u(x) = (1-E^{-1})\,u(x) \\
\delta u(x) &= u(x+\frac{h}{2}) - u(x-\frac{h}{2}) = (E^{1/2} - E^{-1/2})\,u(x) \\
\mu u(x) &= \frac{1}{2}\left(u(x+\frac{h}{2}) + u(x-\frac{h}{2})\right) = \frac{1}{2}(E^{1/2} + E^{-1/2})\,u(x)
\end{aligned}$$

I mer kompakt notasjon skriver vi

$$\begin{aligned}
\Delta &= (E-1) \\
\nabla &= (1-E^{-1}) \\
\delta &= (E^{1/2} - E^{-1/2}) \\
\mu &= \frac{1}{2}(E^{1/2} + E^{-1/2})
\end{aligned}$$

Nå er for eksempel

$$\Delta^k = (E-1)^k = \sum_{\ell=0}^{k} \binom{k}{\ell} (-1)^{k-\ell} E^\ell$$

slik at

$$\Delta^k u(x) = \sum_{\ell=0}^{k} \binom{k}{l} (-1)^{k-\ell} E^\ell u(x) = \sum_{\ell=0}^{k} \binom{k}{l} (-1)^{k-\ell} u(x+\ell h)$$

2.2.5 Differensialoperatoren.

Definer

$$D = \frac{\mathrm{d}}{\mathrm{d}x} \quad \text{slik at} \quad Du(x) = u'(x)$$

La tilsvarende ovenfor

$$D^m u(x) = u^{(m)}(x)$$

Hvis $u(x)$ er analytisk[1] i et intervall som inneholder $x, x+h$ har vi

$$u(x+h) = \sum_{m=0}^{\infty} \frac{h^m}{m!} D^m u(x) = \left(\sum_{m=0}^{\infty} \frac{1}{m!}(hD)^m\right) u(x) = \mathrm{e}^{hD}\, u(x)$$

Vi tenker på dette kun som *notasjon*, en elegant skrivemåte. Vi har altså

$$Eu(x) = \mathrm{e}^{hD} u(x)$$

så vi setter $E = \mathrm{e}^{hD}$.

Sammenheng mellom D og andre operatorer.

$$\begin{aligned}
\Delta &= E - 1 = \mathrm{e}^{hD} - 1 = \sum_{m=1}^{\infty} \frac{1}{m!}(hD)^m \\
\Delta &= hD + \frac{1}{2!}(hD)^2 + \cdots
\end{aligned}$$

Vi skal videre se at under forutsetningen om at u er analytisk kan vi manipulere med analytiske funksjoner slik vi er vant med, betydningen er alltid at sluttresultatet utvikles i en Taylorrekke og tolkes som en sum av potenser av operatorer som anvendes på en glatt funksjon. Analytisitetskravet kan alltid erstattes i etterkant med en Taylor restleddsformel som kun forutsetter at funksjonen vi anvender resultatet på er et endelig antall ganger deriverbar.

Ser vi på potenser av Δ, får vi

$$\Delta^k = \left(\sum_{m=1}^{\infty} \frac{(hD)^m}{m!}\right)^k = h^k D^k + \frac{k}{2!} h^{k+1} D^{k+1} + \cdots$$

eller

$$\Delta^k u(x) = h^k D^k u(x) + \frac{k}{2!} h^{k+1} D^{k+1} u(x) + \cdots$$

som viser at Δ^k / h^k er en første ordens approksimasjon (avbruddsfeil $\mathcal{O}(h)$) til operatoren D^k.

Merk at for $s \in \mathbf{R}$ er

$$E^s u(x) = u(x+sh) = \sum_{k=0}^{\infty} \frac{(sh)^k}{k!} D^k u(x) = \mathrm{e}^{shD} u(x)$$

som stemmer godt overens med vante regneregler. For sentraldifferensen kan vi dermed skrive

$$\delta = E^{1/2} - E^{-1/2} = \mathrm{e}^{\frac{1}{2}hD} - \mathrm{e}^{-\frac{1}{2}hD} = 2\sinh\frac{hD}{2}$$

Videre kan vi beregne

$$\delta^k = \left(2\sinh\frac{hD}{2}\right)^k = \left(hD + \frac{2}{3!}\left(\frac{hD}{2}\right)^3 + \cdots\right)^k = (hD)^k + \frac{k}{24}(hD)^{k+2} + \cdots$$

[1] Som definisjon på at en funksjon er analytisk i et intervall kan en her rett og slett ta at dens Taylorrekke konvergerer i en omegn om hvert punkt i intervallet

det vil si

$$\delta^k u(x) = h^k D^k u(x) + \frac{k}{24} h^{k+2} D^{k+2} u(x) + \cdots$$

noe som viser at δ^k/h^k er en andreordens approksimasjon til D^k.

Spesielt finner vi som tidligere at

$$\delta^2 u(x) = u(x+h) - 2u(x) + u(x-h) = h^2 u''(x) + \frac{1}{12} h^4 u^{(4)}(x) + \cdots \tag{2.6}$$

Det er fristende å manipulere videre med de analytiske funksjonene. Vi har sett at

$$\frac{\delta}{2} = \sinh \frac{hD}{2}$$

Vi skriver dermed formelt

$$D = \frac{2}{h} \sinh^{-1} \frac{\delta}{2}$$

En kan utvikle $\sinh^{-1} z$ i en Taylorrekke

$$\sinh^{-1} z = z - \frac{1}{6} z^3 + \frac{3}{40} z^5 - \frac{5}{112} z^7 + \cdots$$

så hvis vi setter $z = \delta/2$ og multipliserer med $2/h$ får vi

$$D = \frac{1}{h} \left(\delta - \frac{1}{24} \delta^3 + \frac{3}{640} \delta^5 - \frac{5}{7168} \delta^7 + \cdots \right)$$

Siden vi vet at $\delta^k = \mathcal{O}(h^k)$ så ser vi at vi kan finne approksimasjoner til differensialoperatoren D av vilkårlig høy orden bare ved å ta med nok ledd i rekka. Den litt dårlig underbygde manipuleringen med symboler som vi har gjort her, viser seg å være korrekt. For mer detaljert diskusjon av manipulering av differenseoperatorer, se læreboka til Iserles.

Kapittel 3

Diskretisering av varmeledningsligningen

3.1 Om utledning av varmeledningsligningen

Vi skal bruke et standardeksempel for å illustrere numerisk løsning av parabolske differensialligninger gjennom hele kurset. Det dreier seg om den lineære varmeledningsligningen i en romdimensjon, som for eksempel kan brukes til å modellere varmetransport i en rett homogen stav over tid.

Staven på figuren har lengde $L = 1$, og vi lar koordinaten x beskrive et punkt på staven. Ved tid $t \geq 0$ har staven en temperatur $u(x,t)$ i punktet x. Vi kan utlede differensialligningen ved hjelp av Fouriers lov. Fluksen av varme ϕ gjennom et tverrsnitt av staven ved punktet x er proporsjonal med temperaturgradienten, slik at

$$\phi = -\lambda u_x, \quad \lambda > 0,$$

samtidig har vi konserveringsloven

$$\rho c u_t + \phi_x = 0, \quad \rho \text{ er stavens tetthet.}$$

Disse to ligningene impliserer til sammen at

$$u_t = a\, u_{xx}, \quad a = \frac{\lambda}{\rho c}.$$

Ved å innføre skalaer for tid, rom og temperatur

$$w = \frac{u}{u_0}, \quad y = \frac{x}{L}, \quad \tau = \frac{at}{L^2},$$

der w, y og τ er dimensjonsløse variable, u_0 er karakteristisk temperatur, og L er stavens lengde, får man

$$w_\tau = w_{yy}, \quad 0 < y < 1.$$

Vi har demonstrert at etter skalering kan man alltid anta at intervallet for romvariabelen er $[0,1]$ og at koeffisienten a kan sette til 1. Fra nå av ser vi vanligvis på problemet

$$u_t = u_{xx}, \quad 0 < x < 1,\ t > 0.$$

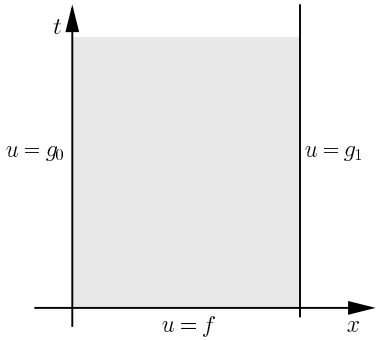
I tillegg til den partielle differensialligningen må man alltid supplere med ulike typer start og/eller randbetingelser. Hva slags betingelser som er nødvendig og tilstrekkelig for at hele problemet skal ha en entydig løsning varierer fra differensialligning til differensialligning. Vi skal se på noen varianter for varmeledningsligningen.

Rent startverdiproblem. Her antar at vi at staven er uendelig lang.

$$\begin{aligned} u_t &= u_{xx}, && x \in \mathbf{R}, \quad t > 0, \\ u(x,0) &= f(x), && x \in \mathbf{R}. \end{aligned}$$

Start/Randverdiproblem (S/R). Dette dekker tilfellet med varmetransport i en homogen stav med lengde 1. Vi må angi startverdi, samt randverdi i de to endene av staven.

$$\begin{aligned} u_t &= u_{xx}, && 0 < x < 1,\ t > 0, \\ u(x,0) &= f(x), && 0 \le x \le 1, \\ u(0,t) &= g_0(t), && t > 0, \\ u(1,t) &= g_1(t), && t > 0. \end{aligned} \tag{3.1}$$



3.2 Numerisk løsning av start/randverdiproblemet

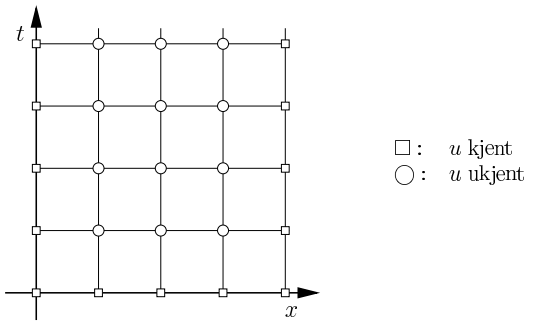
3.2.1 Numerisk approksimasjon på gitter

Vi innfører først en *skrittlengde* i x-retning som vi kaller h, og en i t-retning som vi kaller k. Vi antar i første omgang at $h = 1/(M+1)$ for en heltallig M.

Vi definerer så *gitterpunkter* eller *noder* (x_m, t_n) ved

$$x_m = mh,\ 0 \le m \le M+1, \qquad t_n = nk, \quad n = 0,1,2,\ldots.$$

Merk at dette betyr at $x_0 = 0$ og $x_{M+1} = 1$ blir randpunkter. Den eksakte løsningen i punktet (x_m, t_n) betegnes $u_m^n := u(x_m, t_n)$. I alt hva som kommer skal vi betegne med U_m^n den approksimasjonen som vår numeriske metode gir til løsningen i (x_m, t_n).



3.2.2 Euler, Baklengs Euler og Crank–Nicolson

Vi presenterer her tre forskjellige differenseskjemaer for varmeledningsligningen.

Eulers metode. Vi forenkler nå notasjonen litt for de deriverte og skriver rett og slett

$$\partial_x u = u_x = \frac{\partial u}{\partial x}, \quad \partial_x^k u = \frac{\partial^k u}{\partial x^k}, \quad \partial_t u = u_t = \frac{\partial u}{\partial t}.$$

Vi rekkeutvikler $u_m^{n+1} = u(x_m, t_n + k)$ for konstant $x = x_m$, omkring $t = t_n$, og får

$$u_m^{n+1} = u_m^n + k\partial_t u_m^n + \varphi_m^n, \qquad \varphi_m^n = \frac{1}{2}k^2\partial_t^2 u_m^n + \cdots.$$

Men vi kan nå bruke varmeledningsligningen som spesielt sier at $\partial_t u_m^n = \partial_x^2 u_m^n$, deretter approksimerer vi denne andrederiverte med sentraldifferens som i (2.6)

$$u_m^{n+1} = u_m^n + \frac{k}{h^2}\delta_x^2 u_m^n - \psi_m^n + \varphi_m^n$$

der indeksen på δ betyr at vi anvender den i x-retning dvs

$$\delta_x^2 u_m^n = u_{m+1}^n - 2u_m^n + u_{m-1}^n$$

Fra uttrykket i (2.6) finner vi at

$$\psi_m^n = \frac{1}{12}\,k\,h^2\,\partial_x^4 u_m^n + \cdots$$

Oppsummert har vi

$$u_m^{n+1} = u_m^n + \frac{k}{h^2}\,\delta_x^2 u_m^n + \tau_m^n = u_m^n + \frac{k}{h^2}(u_{m+1}^n - 2u_m^n + u_{m-1}^n) + \tau_m^n$$

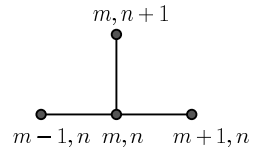
der

$$\tau_m^n = \varphi_m^n - \psi_m^n = \left(\frac{1}{2}k^2\,\partial_t^2 - \frac{1}{12}k\,h^2\,\partial_x^4\right)\,u_m^n + \cdots$$

Eulers formel framkommer ved å erstatte alle små (eksakte) u-verdier med (approksimative) verdier U i formelen ovenfor, og se bort fra *avbruddsfeilen* τ_m^n.

Eulers metode

$$U_m^{n+1} = U_m^n + r\,\delta_x^2 U_m^n \qquad (3.2)$$
$$r = \frac{k}{h^2}$$



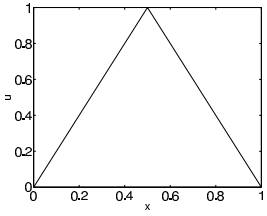
Figuren til høyre ovenfor kalles et beregningsmolekyl, det er et slags lokalt kart over gitteret som forteller hvilke gitterpunkter som er involvert i formelen. Ideen er nå at man starter på $n = 0$ som tilsvarer $t_0 = 0$ der $u(x, t_0) = u(x, 0) = f(x)$ som er kjent. Man kan dermed tilordne verdiene $U_m^0 = f(x_m)$, $m = 0, \ldots, M+1$. Så setter man $n = 1$ og forsyner seg først av de oppgitte randverdiene og får $U_0^1 = g_0(k)$ og $U_0^{M+1} = g_1(k)$. For de øvrige verdiene brukes formelen (3.2) ovenfor, en ser at det øvre gitterpunktet i molekylet kan beregnes ut fra kjente verdier.

Algoritme (Eulers metode for varmeledningsligningen)

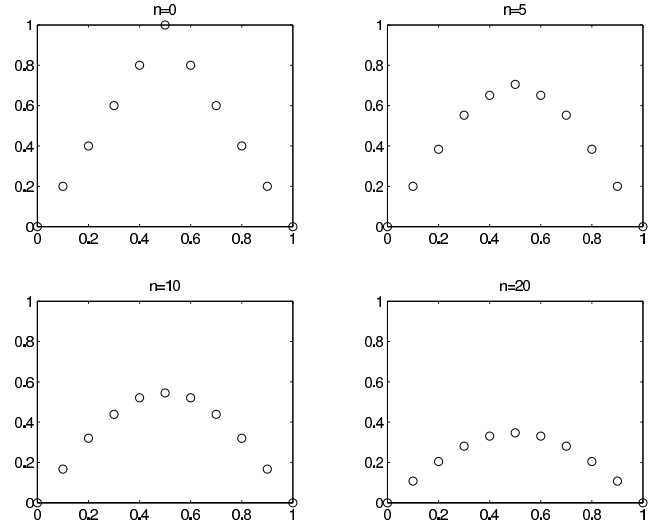
$U_m^0 := f(x_m), \quad m = 0, \ldots, M+1$
for $n = 0, 1, 2, \ldots$
  $U_0^{n+1} := g_0(t_{n+1})$
  $U_{M+1}^{n+1} := g_1(t_{n+1})$
  $U_m^{n+1} := U_m^n + r\,\delta_x^2 U_m^n$, $m = 1, \ldots, M$
end

Eksempel.

$$u_t = u_{xx} \qquad 0 < x < 1,\ t > 0,$$
$$f(x) = \begin{cases} 2x, & 0 \le x \le \frac{1}{2}, \\ 2(1-x), & \frac{1}{2} < x \le 1, \end{cases}$$
$$g_0(t) = g_1(t) = 0, \qquad t > 0.$$



I plottet ovenfor ser du initialfunksjonen. Vi kjører en simulering i Matlab med dette tilfellet, der vi lar $h = 0.1$ ($M = 9$), og $k = 0.0045$. Hvorfor k er så liten i forhold til h, er et spørsmål vi kommer tilbake til. Figure 3.1 viser den numeriske løsningen i gitterpunktene som sirkler, ved tidsskritt 0, 5, 10 og 20.



Figur 3.1: Matlabsimulering av Eulers metode på varmeledningsligningen

Baklengs Euler. Vi rekkeutvikler nå istedet u_m^n omkring $x = x_m$, $t = t_{n+1}$, og får

$$\begin{aligned} u_m^n &= u(x_m, t_{n+1} - k) \\ &= u_m^{n+1} - k\,\partial_t u_m^{n+1} + \tfrac{1}{2} k^2 \partial_t^2 u_m^{n+1} + \cdots \\ &= u_m^{n+1} - k\,\partial_x^2 u_m^{n+1} + \tfrac{1}{2} k^2 \partial_t^2 u_m^{n+1} + \cdots \\ &= u_m^{n+1} - k\,(\tfrac{1}{h^2}\delta_x^2 u_m^{n+1} - \tfrac{1}{12} h^2 \partial_x^4 u_m^{n+1} + \cdots) + \tfrac{1}{2} k^2 \partial_t^2 u_m^{n+1} + \cdots \\ &= u_m^{n+1} - r\delta_x^2 u_m^{n+1} + \tau_m^n, \end{aligned}$$

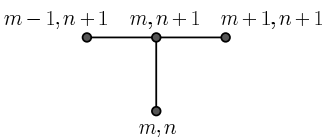
der

$$\tau_m^n = \left(\frac{1}{12} k h^2 \partial_x^4 + \frac{1}{2} k^2 \partial_t^2 \right) u_m^{n+1} + \cdots .$$

Ved å erstatte u'er med U'er og se bort fra avbruddsfeilen τ_m^n får vi

Baklengs Eulers metode

$$U_m^{n+1} - r\,\delta_x^2 U_m^{n+1} = U_m^n, \qquad (3.3)$$
$$r = \frac{k}{h^2}.$$



Baklengs Euler er en implisitt metode. Dette betyr at i hvert tidsskritt må man løse et

lineært ligningssystem for å få ut U_m^{n+1}, $m = 1, \ldots, M$. Vi kommer tilbake til å diskutere løsning av ligningssystemet, først skal vi presentere ytterligere en implisitt metode.

Crank–Nicolsons metode. Denne metoden er tuftet på trapesregelen, en kan representere avbruddsfeil i trapesregelen ved [1]

$$\int_0^k f(t)\,\mathrm{d}t = \frac{1}{2}k\,(f(0)+f(k)) - \frac{1}{12}k^3 f''\left(\frac{k}{2}\right) + \cdots .$$

For å utlede vår metode benytter vi den opplagte formelen

$$u(x_m, t_{n+1}) - u(x_m, t_n) = \int_{t_n}^{t_{n+1}} u_t(x_m, t)\,\mathrm{d}t,$$

og approksimerer integralet med trapesregelen der vi benytter notasjonen $u_m^{n+1/2} = u(x_m, t_n + \frac{1}{2}k)$.

$$\begin{aligned} u_m^{n+1} &= u_m^n + \frac{1}{2}k\,(\partial_t u_m^n + \partial_t u_m^{n+1}) - \frac{1}{12}k^3\,\partial_t^3 u_m^{n+1/2} + \cdots \\ &= u_m^n + \frac{1}{2}k\,(\partial_x^2 u_m^n + \partial_x^2 u_m^{n+1}) - \frac{1}{12}k^3\,\partial_t^3 u_m^{n+1/2} + \cdots \\ &= u_m^n + \frac{1}{2}k\left(\frac{1}{h^2}\delta_x^2 u_m^n + \frac{1}{h^2}\delta_x^2 u_m^{n+1}\right) - \frac{1}{2}k\left(\frac{1}{12}h^2\partial_x^4 u_m^n + \frac{1}{12}h^2\partial_x^4 u_m^{n+1} + \cdots\right) \\ &\quad -\frac{1}{12}k^3\,\partial_t^3 u_m^{n+1/2} + \cdots . \end{aligned}$$

Vi forenkler og oppsummerer

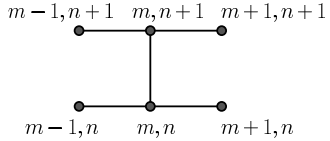
$$\begin{aligned} u_m^{n+1} &= u_m^n + \frac{r}{2}(\delta_x^2 u_m^n + \delta_x^2 u_m^{n+1}) + \tau_m^n, \\ \tau_m^n &= -\frac{1}{12}k^3\partial_t^3 u_m^{n+1/2} - \frac{1}{12}k\,h^2\,\partial_x^4 u_m^{n+1/2} + \cdots , \end{aligned}$$

hvor vi har benyttet at

$$\frac{1}{2}(\partial_x^4 u_m^n + \partial_x^4 u_m^{n+1}) = \partial_x^4 u_m^{n+1/2} + \mathcal{O}(k^2).$$

Crank–Nicolsons metode

$$(1 - \tfrac{r}{2}\,\delta_x^2)\,U_m^{n+1} = (1 + \tfrac{r}{2}\,\delta_x^2)\,U_m^n, \qquad r = \tfrac{k}{h^2}. \tag{3.4}$$



[1] Akkurat denne formen av avbruddsfeil i trapesregelen er nok ikke gjennomgått i TMA4215, men vi skal ta den for gitt likevel.

Vi oppsummerer og skriver alle tre formlene på kompakt form

$$\begin{aligned} \textbf{(E)} \quad & U_m^{n+1} = (1 + r\,\delta_x^2)U_m^n \quad \text{eller} \quad \tfrac{1}{k}\Delta_t U_m^n = \tfrac{1}{h^2}\delta_x^2 U_m^n, \\ \textbf{(BE)} \quad & (1 - r\,\delta_x^2)U_m^{n+1} = U_m^n \quad \text{eller} \quad \tfrac{1}{k}\nabla_t U_m^{n+1} = \tfrac{1}{h^2}\delta_x^2 U_m^{n+1}, \\ \textbf{(CN)} \quad & (1 - \tfrac{r}{2}\,\delta_x^2)U_m^{n+1} = (1 + \tfrac{r}{2}\,\delta_x^2)U_m^n \quad \text{eller} \quad \tfrac{1}{k}\delta_t U_m^{n+1/2} = \tfrac{1}{h^2}\,\delta_x^2\mu_t U_m^{n+1/2}. \end{aligned}$$

Merk at **(E)** er eksplisitt mens både **(BE)** og **(CN)** er implisitte.

Bemerkning om avbruddsfeil. I metodene ovenfor har vi definert den lokale avbruddsfeilen τ_m^n i punktet (x_m, t_n). Gitt en formel generelt, finner vi dens lokale avbruddsfeil ved å sette den eksakte løsningen inn i formelen og flytte alle ledd til en side. Avbruddsfeil uttrykkes ved potenser av skrittlengdene og deriverte av eksakt løsning gjennom Taylorutviklingen.

3.2.3 Løsning av ligningene i Baklengs Euler og Crank–Nicolson

Utgangspunktet er at vi kjenner U_m^n, $0 \le m \le M+1$, samt at U_0^{n+1} og U_{M+1}^{n+1} er gitt av randverdiene. Vi må bestemme U_m^{n+1}, $1 \le m \le M$. Vi ser først på Crank–Nicolson. Høyresiden av ligningen er kjent, og vi setter

$$d_m^{n+1} = \left(1 + \frac{r}{2}\delta_x^2\right)U_m^n = \frac{r}{2}U_{m-1}^n + (1-r)U_m^n + \frac{r}{2}U_{m+1}^n, \quad 1 \le m \le M.$$

For venstresiden får vi komponentvis

$$\left(1 - \frac{r}{2}\,\delta_x^2\right)U_m^{n+1} = -\frac{r}{2}U_{m-1}^{n+1} + (1+r)U_m^{n+1} - \frac{r}{2}U_{m+1}^{n+1}, \quad 1 \le m \le M,$$

der vi setter inn $U_0^{n+1} = g_0^{n+1} = g_0(t_{n+1})$ og $U_{M+1}^{n+1} = g_1^{n+1} = g_1(t_{n+1})$. Vi kan formulere ligningene på matriseform

$$\begin{bmatrix} 1+r & -\frac{r}{2} & & & \\ -\frac{r}{2} & 1+r & -\frac{r}{2} & & \\ & \ddots & \ddots & \ddots & \\ & & -\frac{r}{2} & 1+r & -\frac{r}{2} \\ & & & -\frac{r}{2} & 1+r \end{bmatrix} \begin{bmatrix} U_1^{n+1} \\ U_2^{n+1} \\ \vdots \\ U_{M-1}^{n+1} \\ U_M^{n+1} \end{bmatrix} = \begin{bmatrix} d_1^{n+1} + \frac{r}{2}g_0^{n+1} \\ d_2^{n+1} \\ \vdots \\ d_{M-1}^{n+1} \\ d_M^{n+1} + \frac{r}{2}g_1^{n+1} \end{bmatrix}.$$

En helt tilsvarende utledning gir følgende ligningsystem for Baklengs Euler

$$\begin{bmatrix} 1+2r & -r & & & \\ -r & 1+2r & -r & & \\ & \ddots & \ddots & \ddots & \\ & & -r & 1+2r & -r \\ & & & -r & 1+2r \end{bmatrix} \begin{bmatrix} U_1^{n+1} \\ U_2^{n+1} \\ \vdots \\ U_{M-1}^{n+1} \\ U_M^{n+1} \end{bmatrix} = \begin{bmatrix} U_1^n + r\,g_0^{n+1} \\ U_2^n \\ \vdots \\ U_{M-1}^n \\ U_M^n + r\,g_1^{n+1} \end{bmatrix}.$$

De to matrisene i Crank–Nicolson og Baklengs Euler er eksempler på *tridiagonale* matriser. Disse spesielle tridiagonale matrisene er dessuten *Toeplitzmatriser*, dvs at elementene langs hver av de tre diagonalene er like. I Toeplitzmatriser trenger vi bare å spesifisere første rad og første kolonne, så vil resten være gitt. Hvis man i tillegg vet at den er symmetrisk, så holder det faktisk med å spesifisere første rad (kolonne).

Generelt når man løser partielle differensialligninger med differansemetoder vil man få matriser som er glisne. Ligninger med en romdimensjon og høyst andreordens deriverte resulterer typisk i tridiagonale matriser. Høyere ordens deriverte øker typisk båndbredden i matrisen. Flere romdimensjoner gir opphav til blokkstrukturerte matriser, for eksempel varmeledningsligningen i to romdimensjoner vil typisk gi en blokk-tridiagonal matrise. Toeplitzstruktur mister man i varmeledningsligningen hvis man for eksempel har et inhomogent materiale i staven (se kapittel 3.1), da blir diffligningen av typen

$$u_t = a(x)u_{xx}$$

der $a(x)$ er en gitt funksjon.

Det fins spesielle algoritmer som kan brukes til å løse ligningssystemer med tridiagonale matriser. Av direktemetodene fins det en variasjon av Gausseliminasjon som kalles for Thomasalgoritmen. Vi skal ikke diskutere den nærmere her.

Istedet skal vi gi noen enkle eksempler på hvordan en kan bruke Matlab for å sette opp og løse ligningene ovenfor.

3.2.4 Løsning av ligninger med Matlab

Når man arbeider med glisne matriser, det vil si matriser der storparten av elementene er null, blir det viktig å lagre matrisen på en økonomisk måte i datamaskinminnet. For en $M \times M$-matrise som ovenfor er det dumt å lagre alle elementene, en kan for eksempel istedet lagre en liste over alle indekser tilsvarende ikke-null-elementer med tilhørende verdi. Om man for eksempel setter $M = 1000$ ovenfor blir det $10^6 =$ en million elementer totalt, mens det er kun ca 3000 elementer eller ca 3 promille som er ulik null. En annen sak er at dersom vi multipliserer en stor glissen matrise med en vektor vil vi utføre veldig mange multiplikasjoner og addisjoner med null som vi kunne vært foruten.

Matlab har innebygd støtte for dette. Start Matlab, og forsøk med `> help sparse` eller `> help spdiags`. *Sparse* betyr *glissen* og funksjonen konverterer en full matrise til en sparse matrise (der bare elementer ulik null lagres). Konvertering fra sparse til full gjøres med funksjonen `full`. Ta deg tid til å gå gjennom hjelpeteksten og forsøke noen eksempler. Funksjonen `spdiags` brukes til å generere matriser i sparse format fra dens diagonaler. La oss generere matrisen til Crank–Nicolson metoden i sparse format. Vi antar at $M = 10$ slik at $h = 1/11$. Velg k slik at $r = k/h^2 = 1$. Forsøk følgende sekvens av kommandoer

```
> M=10;
> r=1;
> e=ones(M,1);
> A=spdiags([-r/2*e, (1+r)*e, -r/2*e], -1:1,M,M);
```

Prøv å fjerne ';' i den siste kommandoen for å se hvordan Matlab viser fram en matrise i sparse format. Matlab indekserer diagonalene i en matrise ved å gi hoveddiagonalen indeks 0, subdiagonalen får indeks -1, superdiagonalen indeks 1 osv. Det andre inputargumentet til `spdiags` er en vektor `d` av heltall slik at kolonne j fra matrisen i det første inputargumentet blir diagonal `d(j)` i resultatet. De to siste inputargumentene i kallet ovenfor spesifiserer at resultatet skal være en $M \times M$-matrise.

La oss se hvordan ett tidsskritt med Crank–Nicolson kan utføres i Matlab. Anta derfor at variabelen `U0` har $M+1$ elementer og inneholder den numeriske løsningen ved tid $t = t_n$. Hvis for eksempel $n = 0$ er `U0` generert fra de oppgitte startverdier. Det kan være fornuftig å la `U0` ha dimensjon $M + 2$ og lagre randverdiene i hhv `U0(1)` og `U0(M+2)`

La oss nå sette $n = 0$, og bruke hattfunksjonen som startverdi, dvs $f(x) = 1 - |2x - 1|$. Vi setter opp `U0` ved

```
> h=1/(M+1);      % definer romskrittlengde h
> X=(0:h:1)';     % Gitterpunktene i x-retning er en kolonnevektor
> U0 = 1-abs(2*X-1); % Definer hattfunksjonen som startverdi
```

Anta at disse kommandoene samt de ovenfor er utført, slik at `r,A,U0,X` alle er definert. Vi antar videre at randverdiene er $g_0(t) = g_1(t) = 0$. Da kan vi ta et skritt med Crank–Nicolson som følger

```
> d=r/2*U0(1:end-2)+(1-r)*U0(2:end-1)+r/2*U0(3:end);
> U1=[0;A\d;0];  % Numerisk losning ved tidsskritt n+1
> plot(X,U1,'o') % Plott resultatet
```

Merk at oppsettet av A gjøres en gang, den brukes i alle etterfølgende tidsskritt. Mer optimalt hadde det vært å LU-faktorisere A først, slik at man i etterfølgende tidsskritt kun bruker innsettingsalgoritmen (jfr TMA4200 Numerikk og programmering).

Du kan selv lage et mer fullstendig program som utfører mange skritt med Crank–Nicolson. Forsøk å plotte, eventuelt animere resultatet over tid, se for eksempel `> help movie`. Du kan jo også forsøke med større verdi av M slik at du får bedre oppløsning og mindre numerisk feil.

3.2.5 θ-metoden

Man kan uttrykke alle de tre metodene ovenfor i et felles format ved å skrive

$$(1 - \theta\, r\, \delta_x^2) U_m^{n+1} = (1 + (1-\theta)\, r\, \delta_x^2) U_m^n$$

En har da

(E)	$\theta = 0$
(BE)	$\theta = 1$
(CN)	$\theta = \frac{1}{2}$

Lokal avbruddsfeil i θ-metoden

$$\tau_m^n = (1 - \theta\, r\, \delta_x^2)\, u_m^{n+1} - (1 + (1-\theta)\, r\, \delta_x^2)\, u_m^n = (1 - \theta\, r\, \delta_x^2)(u_m^{n+1} - u_m^n) - r\, \delta_x^2\, u_m^n.$$

Vi utvikler nå alle uttrykk omkring (x_m, t_n) og får

$$\begin{aligned}
\tau_m^n &= \left(1 - \theta\, k\, (\partial_x^2 + \frac{1}{12} h^2\, \partial_x^4 + \cdots)\right)\left(k\, \partial_t + \frac{1}{2} k^2\, \partial_t^2 + \frac{1}{6} k^3\, \partial_t^3\right) u_m^n - k\, \left(\partial_x^2 + \frac{1}{12} h^2\, \partial_x^4 + \cdots\right) u_m^n \\
&= \left(k\, \partial_x^2 + \frac{1}{2} k^2\, \partial_t^2 + \frac{1}{6} k^3\, \partial_t^3 - \theta\, k^2\, \partial_t^2 - \frac{1}{2}\theta\, k^3\, \partial^3 - k\, \partial_x^2 - \frac{1}{12} k\, h^2\, \partial_x^4 + \cdots\right) u_m^n + \cdots \\
&= (\frac{1}{2} - \theta)\, k^2\, \partial_t^2 u_m^n - \frac{1}{12} k\, h^2 \partial_x^4 u_m^n + (\frac{1}{6} - \frac{1}{2}\theta)\, k^3\, \partial_t^3 u_m^n + \cdots .
\end{aligned}$$

Vi konkluderer med at

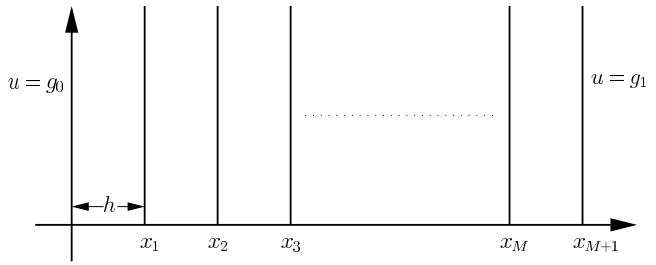
$$\begin{aligned} \tau_m^n &= \mathcal{O}(k^2 + k\,h^2) \quad \text{når } \theta \neq \frac{1}{2}, \\ \tau_m^n &= \mathcal{O}(k^3 + k\,h^2) \quad \text{når } \theta = \frac{1}{2}. \end{aligned}$$

Vi forventer altså at **(CN)** er mer nøyaktig enn **(E)** og **(BE)**.

3.3 Semidiskretisering

3.3.1 Semidiskretisering av varmeledningsligningen

Vi ser igjen på (S/R) problemet med varmeledningsligningen (3.1). La oss nå kun trekke opp vertikale gitterlinjer som på figuren.



Linjene er parallelle med t-aksen og går gjennom $x = x_m,\ m = 0, \ldots, M+1$. Vi betrakter differensialligningen langs en slik linje, der gjelder

$$\begin{aligned} \partial_t u(x_m, t) &= \partial_x^2 u(x_m, t) \\ &= \frac{1}{h^2}\delta_x^2 u(x_m, t) + \varphi(x_m, t), \\ \varphi(x_m, t) &= -\frac{1}{12} h^2 \partial_x^4 u(x_m, t) + \cdots . \end{aligned}$$

Vi innfører nå funksjoner av en variabel $v_m(t),\ m = 0, \ldots, M+1$, som approksimasjoner til $u(x_m, t)$. Vi krever at

$$\begin{aligned} v_0(t) &= g_0(t), \\ v_{M+1}(t) &= g_1(t), \\ \dot{v}_m(t) &= \frac{1}{h^2}\,\delta_x^2 v_m(t), \quad v_m(0) = f(x_m), \quad m = 1, \ldots, M, \end{aligned}$$

der $\dot{v}_m(t) = \dfrac{\mathrm{d}v_m(t)}{\mathrm{d}t}$. For mer kompakt notasjon, la $\boldsymbol{v}(t) \coloneqq [v_1(t), \ldots, v_M(t)]^T$. Vi får

$$\dot{\boldsymbol{v}} = \begin{bmatrix} \dot{v}_1 \\ \dot{v}_2 \\ \vdots \\ \dot{v}_{M-1} \\ \dot{v}_M \end{bmatrix} = \underbrace{\frac{1}{h^2}\begin{bmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{bmatrix}}_{A} \cdot \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_{M-1} \\ v_M \end{bmatrix} + \underbrace{\frac{1}{h^2}\begin{bmatrix} g_0(t) \\ 0 \\ \vdots \\ 0 \\ g_1(t) \end{bmatrix}}_{\boldsymbol{b}(t)}.$$

Så vi har et (lineært) system av ordinære differensialligninger (ODL) av typen

$$\dot{\boldsymbol{v}} = A\boldsymbol{v} + \boldsymbol{b}(t), \qquad \boldsymbol{v}(0) = \boldsymbol{v}_0 = [f(x_1), \ldots, f(x_M)]^T. \tag{3.5}$$

Dette systemet er et spesialtilfelle av det generelle formatet for ordinære differensialligninger som man bruker standard programvare for å løse. Det generelle formatet er

$$\dot{\boldsymbol{v}} = \boldsymbol{F}(t, \boldsymbol{v}), \quad \boldsymbol{v}(0) = \boldsymbol{v}_0, \tag{3.6}$$

der $\boldsymbol{v}$ og $\boldsymbol{F}(t, \boldsymbol{v})$ er vektorer i $\mathbf{R}^M$. 3 av de enkleste metodene for å løse (3.6) med tidsskritt k er

(E)	Euler :	$\boldsymbol{V}^{n+1} = \boldsymbol{V}^n + k\,\boldsymbol{F}(t_n, \boldsymbol{V}^n)$
(BE)	Baklengs Euler :	$\boldsymbol{V}^{n+1} = \boldsymbol{V}^n + k\,\boldsymbol{F}(t_{n+1}, \boldsymbol{V}^{n+1})$
(T)	Trapes :	$\boldsymbol{V}^{n+1} = \boldsymbol{V}^n + \frac{k}{2}\left(\boldsymbol{F}(t_n, \boldsymbol{V}^n) + \boldsymbol{F}(t_{n+1}, \boldsymbol{V}^{n+1})\right)$

Setter man nå inn den spesifikke funksjonen $\boldsymbol{F}(t, \boldsymbol{v}) = A\boldsymbol{v} + \boldsymbol{b}(t)$ fra (3.5), kommer man tilbake til de tre metodene presentert tidligere. Spesielt blir **(T)** til **(CN)**.

3.3.2 Semidiskretiseringsprinsippet generelt

Dette prinsippet fungerer også for andre diffligninger enn varmeledningsligningen. For en slik ligning erstatter vi alle romderiverte med differenseapproksimasjoner. Vi beholder tiden t som kontinuerlig variabel. Resultatet er

$$\text{PDL} \quad \longrightarrow \quad \text{System av ODL}$$

En fordel med å gjøre det slik er at vi kan benytte ferdig programvare for ODL, som har avanserte rutiner for feil- og tidsskrittkontroll. Spesielt kan metoden være interessant dersom PDL'en har ikke-lineære ledd i seg, fordi da blir det tilhørende ODL-systemet også ikke-lineært, og standard ODL-programvare håndterer gjerne slike systemer på en god måte.

Et problem man ofte ser at det resulterende ODL-systemet blir *stivt*. For det lineære semidiskretiserte systemet (3.5) betyr dette gjerne at egenverdiene $\lambda_1, \ldots, \lambda_M$ til A har negativ realdel og at kvotienten

$$\alpha = \frac{\max_i |\mathrm{Re}\,\lambda_i|}{\min_i |\mathrm{Re}\,\lambda_i|},$$

er veldig stor. Hvis A er den tridiagonale matrisen ovenfor vil

$$\lambda_s = -\frac{4}{h^2}\sin^2\frac{s\pi}{2(M+1)}, \quad m = 1,\ldots,M,$$

det vil si at alle egenverdiene er reelle. For små verdier av x er $\sin x \approx x$, slik at for store M er den minste egenverdien (i absoluttverdi)

$$|\lambda_1| \approx \frac{4}{h^2}\frac{\pi^2 h^2}{2^2} = \pi^2.$$

Den største egenverdien (i absoluttverdi) blir

$$|\lambda_M| = -\frac{4}{h^2}\sin^2\frac{M\pi}{2(M+1)} \approx \frac{4}{h^2}\sin^2\frac{\pi}{2} = \frac{4}{h^2}.$$

Så vi finner at $\alpha \approx \frac{4}{\pi h^2} \gg 1$ når h er liten.

Seinere skal vi se at dette gjør at metodene **(BE)** og **(CN)** fungerer bedre enn **(E)**.

3.3.3 Formalisering

Abstrakt kan en skrive en partiell differensialligning (evolusjonsligning) på formen

$$\partial_t u = Lu,$$

der L er en differensialoperator med romderiverte. For eksempel i varmeledningsligningen er $L = \partial_x^2$. Generelt, i en romdimensjon, er

$$L = L(x, t, \partial_x, \partial_x^2, \ldots).$$

Semidiskretisering leder til

$$L \quad \longrightarrow \quad L_h,$$

det vil si, L_h er en diskretisert operator som nå virker på elementene i en vektor av funksjoner av en variabel istedetfor en på en funksjon av to variable. Vi skriver for hver komponent

$$\partial_t u(x_m, t) = L_h\, u(x_m, t) + \varphi(x_m, t),$$

der $\varphi(x_m, t)$ er avbruddsfeilen i romdiskretiseringen. Vi lar nå $v_m(t) \approx u(x_m, t)$ og definerer

$$\dot{v}_m(t) = L_h v_m(t), \quad \text{(ta med randkrav)}.$$

Vi kan videre se på den avbruddsfeilen som skyldes tidsdiskretiseringen, det vil si, *etter* valg av ODL-metode. Vi bruker trapes som eksempel.

La $y(t)$ være eksakt løsning til

$$\dot{y} = F(t, y), \qquad y \in \mathbf{R}^M.$$

En kan vise at med tidsssskrittlengde k blir

$$y_m^{n+1} \coloneqq y_m(t_{n+1}) = y_m^n + \frac{k}{2}\left(F_m(t_n, y^n) + F_m(t_{n+1}, y^{n+1})\right) + \psi_m^n,$$

der

$$\psi_m^n = -\frac{1}{12}\,k^3\, y_m^{(3)}(t_n) + \cdots.$$

Men $u(x_m, t)$ oppfyller $\partial_t u(x_m, t) = L_h u(x_m, t) + \varphi(x_m, t)$. La oss for et øyeblikk la $y_m(t) = u(x_m, t)$ i generell ODL-formulering, slik at

$$F_m(t, y) = L_h u(x_m, t) + \varphi(x_m, t),$$

det vil si, et ODL-system hvis eksakte løsning er eksakt løsning av PDL-problemet langs vertikale linjer ($x = x_m, t$). Vi får da

$$u_m^{n+1} = u_m^n + \frac{k}{2}\left(L_h u_m^n + \varphi_m^n + L_h u_m^{n+1} + \varphi_m^{n+1}\right) + \psi_m^{n+1} = u_m^n + \frac{k}{2}\left(L_h u_m^n + L_h u_m^{n+1}\right) + \tau_m^n,$$

der

$$\tau_m^n = \frac{k}{2}(\varphi_m^n + \varphi_m^{n+1}) + \psi_m^n.$$

Merk at dette gjelder generelt for semidiskretiseringsprinsippet. Spesielt er resultatet i tråd med det vi har sett for de tre metodene **(E)**, **(BE)** og **(CN)** anvendt på varmeledningsligningen.

3.3.4 $u_t = Lu$ med forskjellige valg av L

Tilfelle A.

$$u_t = \underbrace{a(x)\,u_{xx} + b(x)\,u_x + c(x)\,u}_{Lu}.$$

Vi kan også skrive

$$L = a\partial_x^2 + b\partial_x + c.$$

Krav: a, b og c er kontinuerlige i $[0,1]$,
$a(x) > 0$ i $[0,1]$.

Romdiskretisering

		Diskretisering	Avbruddsfeil
u_{xx}	$\longrightarrow$	$\frac{1}{h^2}\delta_x^2 u$	$\mathcal{O}(h^2)$
u_x	$\longrightarrow$	$\frac{1}{h}\Delta_x u$	$\mathcal{O}(h)$
		$\frac{1}{h}\nabla_x u$	$\mathcal{O}(h)$
		$\frac{1}{2h}(u(x+h) - u(x-h)) = \frac{1}{h}\,\mu\,\delta_x u$	$\mathcal{O}(h^2)$

Vi setter altså

$$L_h u = a\,\frac{1}{h^2}\delta_x^2 u + b\left\{\begin{array}{c} \frac{1}{h}\Delta_x u \\ \frac{1}{h}\nabla_x u \\ \frac{1}{h}\,\delta_x\,\mu\,u \end{array}\right\} + c\,u.$$

En har $Lu = L_h u + \varphi$ der

$$\varphi = -\frac{1}{12}\,a\,h^2\partial_x^4 u - b\left\{\begin{array}{c} -\frac{1}{2}\,h\,\partial_x^2 u \\ \frac{1}{2}\,h\,\partial_x^2 u \\ -\frac{1}{6}\,h^2\,\partial_x^3 u \end{array}\right\}$$

Valg av Δ_x versus ∇_x kalles oppstrøms/nedstrøms differensiering. En av disse velges når $b \gg a$ i såkalte konveksjonsdominerte problemer. Fortegnet til b avgjør om en bruker Δ_x eller ∇_x. $b > 0 \to \Delta$, $b < 0 \to \nabla$.

Tilfelle B. Se nå på ligningen

$$u_t = \underbrace{(a(x)\,u_x)_x}_{Lu}, \qquad L = \partial_x(a\partial_x).$$

L er *selvadjungert*. Spesielt betyr dette at hvis man bruker indreprodukt på deriverbare funksjoner som er 0 i endepunktene 0 og 1 definert ved

$$\langle u, v\rangle = \int_0^1 u(x)v(x)\,\mathrm{d}x,$$

så vil $\langle Lu, v\rangle = \langle u, Lv\rangle$ for alle u, v. Ser man på en analog situasjon med indreprodukt på $\mathbf{R}^n$

$$\langle x, y\rangle = y^T x$$

og "bytter ut" L med en matrise, blir dette kravet

$$y^T Ax = \langle Ax, y\rangle = \langle x, Ay\rangle = y^T A^T x$$

for alle $x, y \in \mathbf{R}^n$, noe som impliserer at $A = A^T$, det vil si at A er symmetrisk.

En mulig tanke er å ekspandere L ovenfor ved produktregelen for derivasjon,

$$u_t = au_{xx} + a'\,u_x$$

Da har vi en ligning som i tilfelle A med $b = a'$ og $c = 0$. En annen (og vanligvis bedre) mulighet er å diskretisere direkte den opprinnelige formen, vi lar

$$\begin{aligned}\partial_x(a\,\partial_x)u(x_m,t) \quad \longrightarrow \quad & \frac{1}{h}\,\delta_x(a\,\frac{1}{h}\,\delta_x U)_m = \frac{1}{h^2}\,\delta_x(a_m(U_{m+1/2} - U_{m-1/2})) \\ & = \frac{1}{h^2}\left(a_{m+1/2}(U_{m+1} - U_m) - a_{m-1/2}(U_m - U_{m-1})\right).\end{aligned}$$

Avbruddsfeilen er $\mathcal{O}(h^2)$.

En metode av Tikhonov og Samarski. Skriv på bevaringsform

$$1)\ u_t + w_x = 0, \qquad 2)\ w = -au_x.$$

Vi diskretiserer 1) ved

$$\partial_t u_m = -\partial_x w_m \approx -\frac{1}{h}\delta_x w_m = -\frac{1}{h}(w_{m+1/2} - w_{m-1/2}).$$

For den andre ligningen får vi $u_x = -w/a$ og

$$\int_{x_m}^{x_{m+1}} u_x\,\mathrm{d}x = -\int_{x_m}^{x_{m+1}} \frac{w}{a}\,\mathrm{d}x \approx -w_{m+1/2}\int_{x_m}^{x_{m+1}} \frac{\mathrm{d}x}{a}.$$

Sett nå

$$A_m = \frac{1}{h\int_{x_m}^{x_{m+1}} \frac{\mathrm{d}x}{a}}.$$

Dermed blir

$$\frac{1}{h}w_{m+1/2} \approx -A_m(u_{m+1} - u_m), \quad \frac{1}{h}w_{m-1/2} \approx -A_{m-1}(u_m - u_{m-1}),$$

slik at i diskretiseringen av 1) får vi

$$\partial_t u_m \approx A_m(u_{m+1} - u_m) - A_{m-1}(u_m - u_{m-1}),$$

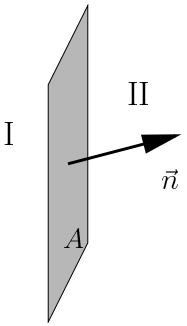
og det semidiskretiserte systemet blir

$$\dot{v}_m = A_m(v_{m+1} - v_m) - A_{m-1}(v_m - v_{m-1}).$$

3.4 Randkrav med derivert

3.4.1 Ulike typer randkrav

Vi ser på typer av randkrav forbundet med varmeledning i 3 romdimensjoner



Figuren illusterer varmefluks ϕ gjennom flaten A med normalvektor $\vec{n}$ fra side I til side II. Denne fluksen er proporsjonal med den retningsderiverte av temperaturen i retning av en normalvektor som peker *ut* av området. Vi skriver

$$\phi = -\lambda\frac{\partial u}{\partial n} = -\lambda\,\vec{n}\cdot\nabla u.$$

En kan tenke seg at flaten A er en del av overflaten (randen) til et området i $\mathbf{R}^3$ der vi løser ligningen. Vi benevner området i rommet for Ω og kaller randen $\partial\Omega$.

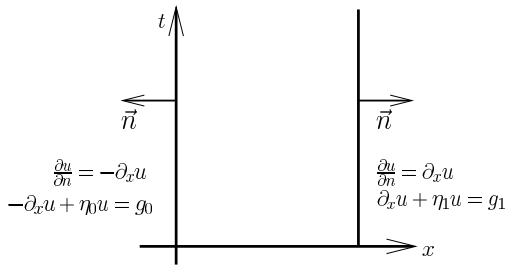
Fysiske situasjoner med deriverte randkrav.

1. Varmefluks gitt (spesifisert) på $\partial\Omega$
$$-\lambda\frac{\partial u}{\partial n} = \phi \quad \text{gitt.}$$
2. Konveksjon
$$-\lambda\frac{\partial u}{\partial n} = \alpha(u - u_0).$$
der en ser bort fra grensesjikt.
3. Stråling (Plancks strålingslov fra statistisk mekanikk)
$$-\lambda\frac{\partial u}{\partial n} = \sigma(u^4 - u_0^4).$$

Vi skal i det videre benytte modellen

$$\frac{\partial u}{\partial n} + \eta\,u = g,$$

der $\eta \in \mathbf{R}$ og funksjonen g er definert på randen $\partial\Omega$. Vi krever $\eta > 0$, og minner om at normalvektoren $\vec{n}$ har retning *ut* av Ω. Vi betrakter nå tilfellet med en romdimensjon.



I en romdimensjon er $\frac{\partial u}{\partial n} = \pm u_x$, fortegnet avhenger av om normalvektoren peker mot venstre eller høyre. Vi kaller det tilsvarende start/randverdiproblemet for (S/RD).

$$
\begin{aligned}
&u_t = u_{xx},\\
&u(x,0) = f(x),\\
&-u_x(0,t) + \eta_0\, u(0,t) = g_0(t),\\
&u_x(1,t) + \eta_1\, u(1,t) = g_1(t),\\
&\eta_0, \eta_1 > 0.
\end{aligned}
$$

NB! $u(0,t)$ og $u(1,t)$ er ukjente.

Semidiskretisering: Vi setter nå $h = 1/M$ og $x_m = mh,\ 0 \le m \le M$. Vi får tilsammen $M+1$ ukjente $v_0, \ldots, v_M$ der $v_m(t) \approx u(x_m, t)$.

3.4.2 Diskretisering av randkrav

Vi skal se hvordan deriverte randkrav kan diskretiseres. En nyttig teknikk er å introdusere "fiktive gitterlinjer"; en til venstre for venstre rand, dvs linjen $x = -h,\ t > 0$, og en til høyre for den høyre randen, nemlig linjen $x = 1+h,\ t > 0$.

Venstre rand. Vi ønsker å benytte en formel med avbruddsfeil $\mathcal{O}(h^2)$ og forsøker med sentraldifferens, som vil involvere løsningen på den fiktive gitterlinja utenfor området vi løser ligningen på.

$$
\begin{gathered}
-\partial_x u(0,t) + \eta_0\, u(0,t) = g_0(t),\\
\downarrow\\
-\frac{u_1 - u_{-1}}{2h} + \eta_0\, u_0 = g_0 + \theta_0,
\end{gathered}
$$

hvor

$$
\theta_0 = -\frac{1}{6}\, h^2\, \partial_x^3 u_0 + \cdots = \text{avbruddsfeil}.
$$

Her er altså $u_{-1} = u(x_{-1}, t) = u(-h, t)$ utenfor området hvor $u(x,t)$ søkes. Dette kan virke litt tvilsomt, men seinere skal vi se at størrelsen u_{-1} elimineres bort

Høyre rand. Tilsvarende får vi

$$
\begin{gathered}
\partial_x u(0,t) + \eta_1\, u(1,t) = g_1(t),\\
\downarrow\\
\frac{u_{M+1} - u_{M-1}}{2h} + \eta_1\, u_M = g_1 + \theta_1,
\end{gathered}
$$

hvor

$$
\theta_1 = \frac{1}{6}\, h^2\, \partial_x^3 u_M + \cdots .
$$

Semidiskretiseringen blir altså

$$
\begin{cases}
\dot{v}_m = \dfrac{1}{h^2}\, \delta_x^2 v_m, \quad 0 \le m \le M,\\
-\dfrac{v_1 - v_{-1}}{2h} + \eta_0\, v_0 = g_0,\\
-\dfrac{v_{M+1} - v_{M-1}}{2h} + \eta_1\, v_M = g_1,
\end{cases}
\tag{3.7}
$$

det vil si $M+3$ ligninger for de $M+3$ ukjente $v_{-1}, v_0, \ldots, v_M, v_{M+1}$. Vi eliminerer umiddelbart bort v_{-1} og v_{M+1}. Fra de to siste ligningene i (3.7) løser vi ut

$$
\begin{aligned}
v_{-1} &= v_1 - 2\, h\, \eta_0\, v_0 + 2hg_0,\\
v_{M+1} &= v_{M-1} - 2h\eta_1 v_M + 2hg_1.
\end{aligned}
$$

Dette settes inn i den første ligningen i (3.7) for $m = 0,\ m = M$

$$
\begin{aligned}
\dot{v}_0 &= \frac{1}{h^2}\, \delta_x^2 v_0 = \frac{1}{h^2}\, (v_{-1} - 2\, v_0 + v_1)\\
&= \frac{1}{h^2}\, \big(-2\, (h\, \eta_0 + 1)\, v_0 + 2\, v_1 \big) + \frac{2}{h}\, g_0,
\end{aligned}
$$

$$
\begin{aligned}
\dot{v}_M &= \frac{1}{h^2}\, \delta_x^2 v_M = \frac{1}{h^2}\, (v_{M-1} - 2\, v_M + v_{M+1})\\
&= \frac{1}{h^2}\, \big(2\, v_{M-1} - 2\, (h\, \eta_1 + 1)\, v_M \big) + \frac{2}{h}\, g_1.
\end{aligned}
$$

Vi kan skrive opp dette systemet med matrise-vektor notasjon. Nå er $\boldsymbol{v}(t) = [v_0(t), \ldots, v_M(t)]^T$, og vi setter

$$
\dot{\boldsymbol{v}} = \frac{1}{h^2}\, Q\, \boldsymbol{v} + \frac{2}{h}\, \boldsymbol{d},
\tag{3.8}
$$

der

$$
Q = \begin{bmatrix}
-2(h\eta_0 + 1) & 2 & & & \\
1 & -2 & 1 & & \\
 & \ddots & \ddots & \ddots & \\
 & & 1 & -2 & 1\\
 & & & 2 & -2(h\eta_1 + 1)
\end{bmatrix}, \qquad
\boldsymbol{d} = \begin{bmatrix} g_0 \\ 0 \\ \vdots \\ 0 \\ g_1 \end{bmatrix}.
\tag{3.9}
$$

Vi merker oss først at denne matrisen ikke er symmetrisk, men den er similær med en symmetrisk matrise. Med diagonalmatrisen $D = \mathrm{diag}(\sqrt{2}, 1, \ldots, 1, \sqrt{2})$ finner vi at $\tilde{Q} = D^{-1} Q D$ er symmetrisk. Derfor har den reelle egenverdier. $\tilde{Q}$ er dessuten *negativ definitt*[2].

Som eksempel på et fulldiskretisert system kan vi bruke trapesmetoden i tid, og får da Crank–Nicolson. La $U^n = (U_0^n, \ldots, U_M^n)^T$.

$$
U^{n+1} = U^n + \frac{k}{2} \left(\frac{1}{h^2}\, Q U^n + \frac{2}{h}\, d^n + \frac{1}{h^2}\, Q U^{n+1} + \frac{2}{h}\, d^{n+1} \right),
$$

eller

$$
(I - \frac{r}{2}\, Q)\, U^{n+1} = (I + \frac{r}{2}\, Q)\, U^n + \frac{k}{h}(d^n + d^{n+1}).
$$

[2] en matrise A er *negativ definitt* hvis og bare hvis $(-A)$ er positiv definitt

Alternativ diskretisering av randkrav. En kan unngå fiktiv gitterlinje, og bruke lavere ordens approksimasjon, dermed mindre nøyaktig diskretisering

$$-\partial_x u_0 + \eta_0\, u_0 = g_0,$$
$$\downarrow$$
$$-\frac{u_1 - u_0}{h} + \eta_0\, u_0 = g_0 + \tilde{\theta}_0,$$

der $\tilde{\theta}_0 = \frac{1}{2}h\partial_x^2 u_0 + \cdots$. Tilsvarende brukes bakoverdifferens for høyre rand.

$$\partial_x u_M + \eta_1\, u_M = g_1 \quad \rightarrow \quad \frac{u_M - u_{M-1}}{h} + \eta_1\, u_M = g_1 + \tilde{\theta}_1.$$

3.5 Ikke-lineære parabolske differensialligninger

Generelt kunne man se på ligninger av formen

$$u_t = f(x, t, u, u_x, u_{xx}), \quad \frac{\partial f}{\partial u_{xx}} > 0, \quad + \text{startkrav \& randkrav.}$$

Vi semidiskretiserer ligningen ved å introdusere $x_m = mh$ og $v_m(t) \approx u(x_m, t)$, $h = 1/(M+1)$.

$$\dot{v}_m = f\left(x_m, t, v_m, \frac{1}{2h}(v_{m+1} - v_{m-1}), \frac{1}{h^2}\delta_x^2 v_m\right).$$

Hvis vi også inkluderer randkravene, får vi et system av ordinære differensialligninger (ODL). Sett $v = [v_1, v_2, \ldots, v_M]^T$ og $F = [F_1, F_2, \ldots, F_M]^T$ der

$$F_m = f\left(x_m, t, v_m, \frac{1}{2h}(v_{m+1} - v_{m-1}), \frac{1}{h^2}\delta_x^2 v_m\right).$$

Vi har altså funnet et ikke-lineært system av ODL

$$\dot{v} = F(t, v),$$

som kan løses med passende ODL-løsere (f. eks. i Matlab).

Burgers' ligning.

$$u_t = \varepsilon\, u_{xx} - u u_x$$

En kan semidiskretisere ved å sette

$$F_m = \frac{\varepsilon}{h^2}(v_{m+1} - 2v_m + v_{m-1}) - v_m \frac{1}{2h}(v_{m+1} - v_{m-1}).$$

Her kan man for eksempel bruke en Runge–Kutta metode på $\dot{v} = F(v)$, se for eksempel `> help ode45` i Matlab. Merk ellers at Burgers' ligning kan skrives på formen

$$\partial_t u = \varepsilon\, \partial_x^2 u - \frac{1}{2}\partial_x u^2,$$

som kan diskretiseres direkte med sentraldifferens, og en får

$$F_m = \frac{\varepsilon}{h^2}(v_{m+1} - 2v_m + v_{m-1}) - \frac{1}{4h}(v_{m+1}^2 - v_{m-1}^2).$$

Spesiell ligningstype. Noen ganger framkommer ikke-lineære partielle differensialligninger på formen

$$b(u)\, u_t = (a(u)\, u_x)_x, \quad b(u) > 0, \quad a(u) > 0.$$

Man kan her bruke teknikken ovenfor på problemet

$$u_t = \frac{a(u)}{b(u)}\, u_{xx} + \frac{a'(u)}{b(u)}\, u_x^2.$$

Men en bedre måte er å la

$$((a(u)\, u_x)_x)_m \longrightarrow \frac{1}{h^2}(\delta_x(a\, \delta_x v))_m = \frac{1}{h^2}(a_{m+1/2}(v_{m+1} - v_m) - a_{m+1/2}(v_m - v_{m-1}))$$

der

$$a_{m\pm 1/2} = a(v_{m\pm 1/2}) = a(v(x_m \pm h/2)),$$

dette er en størrelse som ikke er med i gitteret. Men vi kan benytte at

$$u_{m\pm 1/2} = \frac{1}{2}(u_m + u_{m\pm 1}) + \mathcal{O}(h^2),$$

så en slik approksimasjon har avbruddsfeil av samme orden som vi allerede har fra diskretiseringen av de deriverte. Vi definerer videre

$$\alpha_{m\pm 1/2} = a\left(\frac{v_m + v_{m\pm 1}}{2}\right).$$

Semidiskretiseringen blir nå

$$b(v_m)\dot{v}_m = \frac{1}{h^2}\left(\alpha_{m+1/2}(v_{m+1} - v_m) - \alpha_{m-1/2}(v_m - v_{m-1})\right), \quad \text{avbruddsfeil: } \mathcal{O}(h^2).$$

Crank–Nicolson på ikke-lineært parabolsk problem.

$$U_m^{n+1} = U_m^n + \frac{k}{2}\left(F_m(U^n) + F_m(U^{n+1})\right),$$

der

$$F_m(U^n) = \frac{1}{b(U^n)}(\alpha_{m+1/2}\Delta_x U_m^n - \alpha_{m-1/2}\nabla_x U_m^n).$$

Dette betyr at vi må løse en ikke-lineær ligning for å få ut U_m^{n+1}. Om vi insisterer på å ha samme nøyaktighet som Crank–Nicolson, men vil unngå å løse ikke-lineært system i hvert tidsskritt, kan vi forsøke med en 3-nivå formel. Vi bruker sentraldifferens i tid, og får

$$\frac{U_m^{n+1} - U_m^{n-1}}{2k} = F_m(U^n),$$

eller

$$U_m^{n+1} = U_m^{n-1} + 2k\, F_m(U^n).$$

NB! Denne formelen er alltid ustabil for den aktuelle ligningen.

Modifikasjon. I den ustabile formelen inngår

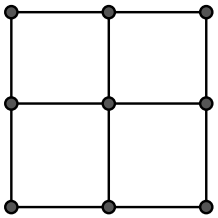
$$F_m(U^n) = \frac{1}{b(U_m^n)} \frac{1}{h^2} \left(\alpha_{m+1/2}\Delta_x U_m^n - \alpha_{m-1/2}\nabla_x U_m^n\right).$$

Erstatt

$$\begin{aligned}
\Delta_x U_m^n &\longrightarrow \frac{1}{3}(\Delta_x U_m^{n-1} + \Delta_x U_m^n + \Delta_x U_m^{n+1}),\\
\nabla_x U_m^n &\longrightarrow \frac{1}{3}(\nabla_x U_m^{n-1} + \nabla_x U_m^n + \nabla_x U_m^{n+1}),
\end{aligned}$$

Resultatet blir formelen

$$\begin{aligned}
U_m^{n+1} = U_m^{n-1} + \frac{2}{3} r \frac{1}{b(U_m^n)} \big(&\alpha_{m+1/2}(\Delta_x U_m^{n-1} + \Delta_x U_m^n + \Delta_x U_m^{n+1})\\
&-\alpha_{m-1/2}(\nabla_x U_m^{n-1} + \nabla_x U_m^n + \nabla_x U_m^{n+1})\big),
\end{aligned}$$



som opprinnelig ble foreslått av Lees. Figuren til venstre viser formelens beregningsmolekyl. Formelen er *lineært implisitt*, det vil si at man løser et lineært ligningssystem i hvert skritt. Som med lineære flerskrittsmetoder for ODL, trenger man også her en startmetode for å komme igang, man må beregne U^1 med en annen metode, som helst bør ha samme ordens avbruddsfeil som hovedmetoden. Slike startmetoder trenger man generelt i p-nivåformler når $p > 2$.

Kapittel 4

Stabilitet, konsistens og konvergens

Vi skal nå analysere den numeriske løsningen av en partiell differensialligning. Noe av denne teorien er gyldig generelt for PDL, men eksempler blir naturlig nok varmeledningsligningen og de metoder vi har introdusert for denne.

4.1 Egenskaper ved det kontinuerlige problemet

Når man skal approksimere løsningen av en partiell differensialligning er det vesentlig at differensialligningen selv har en veldefinert løsning. Et *velformet PDL-problem* oppfyller følgende tre kriterier

1. Det eksisterer en løsning
2. Løsningen er entydig
3. Løsningen avhenger kontinuerlig av problemets data

Eksempel. Vi tar igjen (S/R) problemet for varmeledningsligningen som eksempel

$$\begin{aligned}
u_t &= u_{xx}, \quad 0 < x < 1,\ t > 0,\\
u(x,0) &= f(x), \quad 0 \le x \le 1,\\
u(0,t) &= g_0(t), \quad t > 0,\\
u(1,t) &= g_1(t), \quad t > 0.
\end{aligned}$$

Data er her funksjonene f, g_0 og g_1.

Anta at f, g_0 og g_1 er kontinuerlige og at $f(0) = g_0(0), f(1) = g_1(0)$. Da har (S/R) problemet for varmeledningsligningen en entydig løsning $u(x,t)$ som er kontinuerlig for $0 \le x \le 1,\ t \ge 0$ og som oppfyller et maksimumsprinsipp

$$\max_{0\le x\le 1} |u(x,t)| \le \max\left\{\max_{0\le x\le 1}|f(x)|, \quad \max_{s\le t}|g_0(s)|, \quad \max_{s\le t}|g_1(s)|\right\} \tag{4.1}$$

Mer generelle og avanserte resultater av denne typen finnes i litteraturen. Vi skal ikke bevise dette resultatet heller, men påpeker kun hvordan maksimumsprinsippet (4.1) enkelt impliserer egenskapene (2) og (3) ovenfor. La u_1 og u_2 være løsninger med data

$$f^{(i)},\ g_0^{(i)},\ g_1^{(i)}, \quad i = 1,2.$$

Sett

$$w = u^{(1)} - u^{(2)}, \quad \phi = f^{(1)} - f^{(2)}, \quad \gamma_0 = g_0^{(1)} - g_0^{(2)}, \quad \gamma_1 = g_1^{(1)} - g_1^{(2)}.$$

Vi har $w_t = u_t^{(1)} - u_t^{(2)} = u_{xx}^{(1)} - u_{xx}^{(2)} = w_{xx}$ så w er en løsning av varmeledningsligningen med data

$$w(x,0) = \phi(x), \quad w(0,t) = \gamma_0(t), \quad w(1,t) = \gamma_1(t),$$

og fra (4.1) finner vi at

$$\max_{0\le x\le 1} |w(x,t)| \le \max\left\{ \max_{0\le x\le 1} |\phi(x)|, \quad \max_{s\le t} |\gamma_0(s)|, \quad \max_{s\le t} |\gamma_1(s)| \right\}. \tag{4.2}$$

Så entydigheten (2) følger nå, fordi to løsninger med like data vil ha ϕ, γ_0, γ_1 identisk null, og dermed blir $w(x,t)$ også identisk null. Men også egenskap (3) følger av (4.2). Egenskap (3) ovenfor refereres ofte til som *stabilitet av differensialligningen*. Kan vi overføre dette begrepet til numerisk løsning?

4.2 Konvergens av numerisk metode

La u være løsning av et PDL-problem (f.eks. (S/R) som ovenfor) på rektanglet

$$\Omega_T = [0,1] \times [0,T] = \{(x,t) : 0 \le x \le 1,\ 0 \le t \le T\}.$$

Innfør et gitter

$$G = \{(x_m, t_n),\ 0 \le m \le M,\ 0 \le n \le N\},$$

der

$$x_m = mh,\ h = \frac{1}{M}, \quad t_n = nk,\ k = \frac{T}{N}.$$

La U være definert på gitteret slik at $U_m^n \approx u(x_m, t_n)$. Diskretiseringsfeilen defineres som

$$e_m^n = u_m^n - U_m^n, \qquad u_m^n = u(x_m, t_n).$$

Vi sier at $U \to u$ i Ω_T når $h \to 0$, $k \to 0$ hvis

$$\max_{0\le n\le T/k} \quad \max_{0\le m\le 1/h} |e_m^n| \to 0, \quad h \to 0,\ k \to 0.$$

Mer generelt. Sett

$$U^n = [U_0^n, \ldots, U_M^n]^T,\ u^n = [u_0^n, \ldots, u_M^n]^T,\ e^n = [e_0^n, \ldots, e_M^n]^T, \text{ vektorer i } \mathbf{R}^{M+1}.$$

Vi velger nå en vektornorm $\|\cdot\|$ definert på $\mathbf{R}^{M+1}$ for alle $M \ge 0$, og sier at $U \to u$ i Ω_T hvis

$$\max_{0\le n\le T/k} \|e^n\| \to 0 \quad \text{når } k \to 0,\ h \to 0.$$

Eksempler på normer er

$$\|e^n\|_\infty = \max_m |e_m^n|,$$

og en skalert variant av den vanlige $\|\cdot\|_2$-normen

$$\|e^n\|_{2,h} = \left(h \sum_{m=0}^{M} |e_m^n|^2 \right)^{1/2} = \frac{1}{\sqrt{M}} \|e^n\|_2.$$

Merknader.

1. Dette med normer er litt tricky her. Vi vet at alle normer på et endeligdimensjonalt rom er *ekvivalente* , noe som igjen innebærer at konvergens i en norm er ekvivalent med konvergens i en annen norm. Men her må vi ta hensyn til at dimensjonen til dette rommet går mot uendelig når $h \to 0$. En har for eksempel sammenhengen

   $$\|x\|_{2,h} \le \|x\|_\infty \le \sqrt{M} \|x\|_{2,h}.$$

   For eksempel vektoren med 1 i første element og 0 i resten vil konvergere mot 0 i $\|\cdot\|_{2,h}$ men ikke i $\|\cdot\|_\infty$ når $h \to 0$ (og dermed $M \to \infty$).

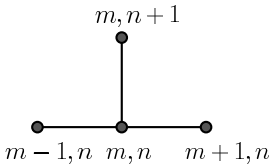
2. Den skalerte normen ovenfor, $\|\cdot\|_{2,h}$ kan en tenke seg approksimerer L_2-normen til en underliggende kontinuerlig funksjon

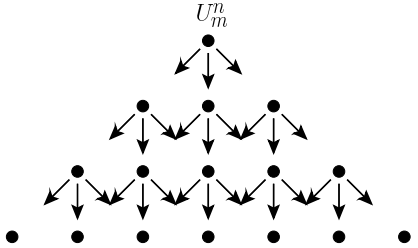
   $$\|\boldsymbol{f}\|_{2,h} = \left(h \sum_{m=0}^{M} |e_m^n|^2 \right)^{1/2} \approx \left(\int_0^1 |f(x)|^2 \,\mathrm{d}x \right)^{1/2} = \|f\|_{L_2},$$

   der $\boldsymbol{f}_m = f(mh)$.

4.3 Avhengighetsområde for en numerisk metode

Avhengighetsområdet til Eulers metode. Ser vi på beregningsmolekylet for Eulers metode presentert tidligere, ser vi at approksimasjonen U_m^n i punktet (x_m, t_n) avhenger av U_{m-1}^{n-1}, U_m^{n-1} og U_{m+1}^{n-1}. Hver av disse avhenger av 3 gitterpunkter på foregående tidsnivå osv. Fortsetter en slik nedover, finner en at U_m^n avhenger indirekte av punktene $U_{m-n}^0, \ldots, U_{m+n}^0$ hvis man løser et rent startverdiproblem (eller hvis $0 \le m-n, m+n \le M$). Avhengighetsområdet er i dette tilfellet trekanten med hjørner i $U_{m-n}^0, U_{m+n}^0, U_m^n$. Generelt inkluderer avhengighetsområdet til U_m^n alle U_μ^ν verdier som har vært involvert i beregningen av U_m^n.





I (S/R) problemet vil man etterhvert få et avhengighetsområde som på figuren nedenfor.

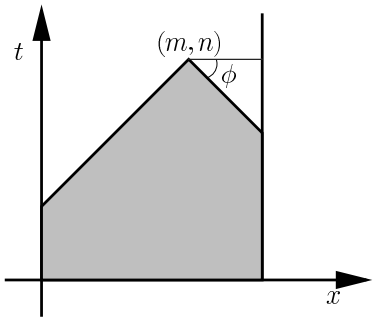
Vinkelen

$$\phi = \arctan\frac{k}{h},$$

karakteriserer dette området for Eulers metode. Det viser seg at den eksakte løsningen $u(x_m, t_n)$ avhenger av hele rektangelet med hjørner i $(0,0),(0,t_n),(1,t_n),(1,0)$. Hvis vi lar $r = \frac{k}{h^2}$ være konstant når $h \to 0$ vil

$$\phi = \arctan\frac{k}{h} = \arctan rh \to 0,$$

slik at i grensen vil vi gjenvinne hele avhengighetsområdet for den eksakte løsningen.



4.4 Konvergensbevis for Eulers metode på (S/R) med $r \le \frac{1}{2}$

Vi ser altså på problemet (3.1) og minner om Eulers metode

$$\begin{aligned} U_m^{n+1} &= U_m^n + r\,\delta_x^2 U_m^n, \quad 1 \le m \le M, n \ge 0, \\ U_0^n &= g_0^n, \quad U_{M+1}^n = g_1^n, \quad n > 0, \\ U_0^m &= f_m, \qquad 0 \le m \le M+1. \end{aligned} \tag{4.3}$$

For den eksakte løsningen gjelder

$$u_m^{n+1} = u_m^n + r\,\delta_x^2 u_m^n + \tau_m^n. \tag{4.4}$$

Vi definerer nå $e_m^n = u_m^n - U_m^n$, og trekker (4.3) fra (4.4). Vi får

$$e_m^{n+1} = e_m^n + r\,\delta_x^2 e_m^n + \tau_m^n,\ n > 0, 1 \le m \le M. \tag{4.5}$$

Dessuten er $e_m^0 = 0$ og $e_0^n = e_{M+1}^n = 0$. Vi vet at for Eulers metode gjelder

$$\tau_m^n = \frac{1}{2}k^2\partial_t^2 u_m^n - \frac{1}{12}kh^2\partial_x^4 u_m^n + \cdots,$$

så det virker rimelig, under visse antagelser om den eksakte løsningens glatthet, å anta at det fins en konstant A slik at

$$|\tau_m^n| \le A(k^2 + kh^2), \quad \text{for alle } m, n.$$

Vi skriver ut (4.5) og får

$$e_m^{n+1} = r\,e_{m-1}^n + (1-2r)e_m^n + r\,e_{m+1}^n + \tau_m^n.$$

Når vi etterpå tar absoluttverdier, skal vi gjøre nytte av antagelsen om at $0 \le r \le \frac{1}{2}$ for dette innebærer at både r og $1-2r$ er ikke-negative. Vi finner derfor

$$\begin{aligned} |e_m^{n+1}| &\le r\,|e_{m-1}^n| + (1-2r)\,|e_m^n| + r\,|e_{m+1}^n| + A\,(k^2 + kh^2) \\ &\le \max_\ell |e_\ell^n|(r + (1-2r) + r) + A\,(k^2 + kh^2) \\ &= \max_\ell |e_\ell^n| + A\,(k^2 + kh^2). \end{aligned}$$

Setter vi nå $E^n = \max_\ell |e_\ell^n|$ finner vi

$$E^{n+1} \le E^n + A\,(k^2 + kh^2),$$

og siden $E^0 = 0$ er

$$E^n \le n\,k\,A\,(k + h^2) = t_n\,A\,(k + h^2) \le T\,A\,(k + h^2),$$

det vil si

$$\max_\ell |e_\ell^n| \le T\,A\,(k + h^2) \quad \text{for alle } n \le T/k.$$

Vi konkluderer med at Eulers metode er konvergent, dvs $U \to u$ når $h \to 0$ og $k \to 0$ for konstant $r \le \frac{1}{2}$.

4.5 Stabilitet på ubegrenset tidsintervall (F-stabilitet)

La oss formulere θ-metoden (som inkluderer **(E)**, **(BE)** **(CN)**) på matriseform. Vi har altså

$$(1 - \theta\,r\,\delta_x^2)U_m^{n+1} = (1 + (1-\theta)\,r\,\delta_x^2)U_m^n.$$

La S være matrisen

$$S = \begin{bmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{bmatrix}. \tag{4.6}$$

Definerer vi vektoren $U^n = [U_1^n, \ldots, U_m^n]^T$ for $n = 0, 1, \ldots$ kan vi skrive θ-metoden på vektorform ved

$$(I - \theta\,r\,S)U^{n+1} = (I + (1-\theta)\,r\,S)U^n + d^n, l$$

hvor

$$d^n = [\theta\,r\,g_0^{n+1} + (1-\theta)\,r\,g_0^n,\ 0,\ \ldots,\ 0,\ \theta\,r\,g_1^{n+1} + (1-\theta)\,r\,g_1^n]^T.$$

Typisk kan differensemetoder skrives som

$$A\,U^{n+1} = B\,U^n + c^n. \tag{4.7}$$

A og B vil avhenge av h og k ved at elementene er funksjoner av h og k, men også ved at matrisenes dimensjon typisk er $M \times M$ der $M = 1/h$ (evt $M = 1/h - 1$ e.l.)

A og B *kan* avhenge av n, men vi antar i denne fremstillingen at de ikke gjør det.

En metode kalles for en *enskrittsmetode* (ODL-terminologi) eller alternativt en *tonivåmetode* (PDL-terminologi) hvis den kan skrives på formen

$$U^{n+1} = C\,U^n + q^n. \tag{4.8}$$

Hvis A i (4.7) er invertibel, kan vi sette $C = A^{-1}B$ og $q^n = A^{-1}d^n$.

Definisjon av F-stabilitet. (stabilitet på $(0,\infty)$)

For en vilkårlig vektor w^0, beregn følgen $w^{n+1} = C\,w^n,\ n = 0, 1, \ldots$. Velg en vektornorm $\|\cdot\|$. Vi sier at (4.8) er F-stabil hvis det fins en konstant L uavhengig av n slik at

$$\|w^n\| \le L\|w^0\| \qquad \text{for enhver } w^0.$$

Merk at stabilitetsbegrepet ikke har noe med løsningen av en differensialligning å gjøre, det er kun en egenskap ved differenseformlen.

Kriterium for F-stabilitet.

$$\rho(C) < 1 \qquad \Rightarrow \qquad F\text{-stabilitet} \qquad \Rightarrow \qquad \rho(C) \le 1.$$

Bevis. Anta at $\rho(C) > 1$, da fins en egenverdi λ med tilhørende egenvektor x slik at[1] $|\lambda| > 1$. Setter vi $w^0 = x$ får vi

$$w^n = C^n w^0 = \lambda^n w^0 \qquad \Rightarrow \qquad \|w^n\| = |\lambda|^n \|w^0\|,$$

så det fins ingen L slik at $|\lambda|^n \le L$ for alle n. Vi konkluderer med at F-stabilitet $\Rightarrow\ \rho(C) \le 1$.

Fra resultat i kapittel 2.1.7 om konvergente matriser konkluderer vi med at $\rho(C) < 1\ \Rightarrow\ C^n \to 0$ når $n \to \infty$. Dette innebærer at det fins N, L_0 slik at $\|C^n\| \le L_0,\ n > N$ for den tilordnede matrisenormen. Vi kan sette $L = \max\{L_0, \|C^0\|, \ldots, \|C^N\|\}$ og får

$$\|w^n\| = \|C^n w^0\| \le \|C^n\|\,\|w^0\| \le L\,\|w^0\|.$$

Merknad. I denne diskusjonen holdes både h og k fast, og stabilitet undersøkes når $n \to \infty$, det vil si også $t_n \to \infty$. Vi kommer tilbake til situasjonen der h og k går mot null, mens vi har en fast øvre grense T slik at $t_n = nk \le T$ dvs vi lar $0 \le n \le T/k$.

Eksempel. Euler anvendt på **(S/R)**. Her er $A = I$ og $B = I + rS$ slik at $C = B = I + rS$ og S er definert i (4.6). Egenverdiene til C er gitt som

$$\lambda_j = 1 + r\,\sigma_j,$$

der σ_j er egenverdiene til S. Disse er reelle og finnes fra øving 1,

$$\sigma_j = 2\left(\cos\frac{j\pi}{M+1} - 1\right) = -4\sin^2\frac{j\pi}{2(M+1)},\ j = 1, \ldots, M.$$

[1] Det oppstår en teknikalitet når en har komplekse egenverdier og egenvektorer, hvis resultatet skal holde kun for reelle w^0. Men hvis største egenverdi er kompleks og C reell, vil også $\bar{\lambda}$ være en egenverdi med egenvektor $\bar{x}$. Beviset kan enkelt modifiseres ved å bruke begge disse.

Vi antar $r = k/h^2 > 0$ og ser da at $\lambda_j < 1$ for alle j. Men vi må også forlange at $\lambda_j \ge -1$.

$$\begin{aligned} -1 &\le 1 - 4\,r\sin^2\frac{j\pi}{2(M+1)}, \qquad j = 1, 2, \ldots, M \\ &\Updownarrow \\ r &\le \frac{1}{2\sin^2\frac{j\pi}{2(M+1)}}, \qquad j = 1, 2, \ldots, M. \end{aligned}$$

Minimum av denne skranken for r inntreffer når $j = M$, en har $\sin^2\frac{M}{M+1}\frac{\pi}{2} = \cos^2\frac{h\pi}{2} < 1$ slik at en får faktisk $\rho(C) < 1$ hvis $r \le \frac{1}{2}$.

Eulers metode er F-stabil for $r = \dfrac{k}{h^2} \le \dfrac{1}{2}$.

4.6 Stabilitet på $[0,T]$ når $h \to 0$, $k \to 0$

Nå endrer vi utgangspunktet, og ser på stabilitet av metoder som approksimerer PDL-en på et rektangel $[0,1] \times [0,T]$. Fremdeles ser vi på prosessen når $n \to \infty$, men nå skjer dette ved at $k \to 0$ samtidig, slik at vi alltid har en øvre grense $T = nk$. Vi antar dessuten at $h \to 0$ samtidig, slik at matrisedimensjonene øker i prosessen.

Vi bruker kun navnet *stabilitet* om dette tilfellet. Vi ser igjen på en beregningsprosess av typen (4.8).

Stabilitetsdefinisjon. Vi sier at (4.8) er stabil hvis og bare hvis det fins en konstant L uavhengig av h og k slik at følgen $w^{n+1} = Cw^n$ oppfyller

$$\|w^n\| \le L\,\|w^0\| \quad \text{for alle } n \le \frac{T}{k} \quad \text{og startvektorer } w^0, \tag{4.9}$$

der $\|\cdot\|$ er en vektornorm.

Ekvivalent definisjon. Skjemaet definert ved (4.8) er stabilt hvis og bare hvis det fins en konstant L uavhengig av h og k slik at

$$\|C^n\| \le L \quad \text{for alle } n \le \frac{T}{k}, \tag{4.10}$$

der matrisenormen er tilordnet vektornormen ovenfor.

Eksempel. Hvis en bruker normen

$$\|w\|_\infty = \max_m |w_m| \qquad \text{i (4.9)},$$

bruker man

$$\|C\|_\infty = \max_\ell \sum_m |C_{\ell m}| \qquad \text{i (4.10)}.$$

Bevis av ekvivalens av (4.9) og (4.10). Anta først at (4.9) holder. La h, k og n være vilkårlig. Fordi matrisenormen i (4.10) er tilordnet, kan vi finne w^0 slik at $\|C^n w^0\| = \|C^n\|\,\|w^0\|$. Dermed blir

$$\|C^n\|\,\|w^0\| = \|C^n w^0\| = \|w^n\| \le L\,\|w^0\| \quad \Rightarrow \quad \|C^n\| \le L$$

Anta deretter at (4.10) holder, og la w^0, h, k og n være vilkårlige.

$$\|w^n\| = \|C^n w^0\| \le \|C^n\|\,\|w^0\| \le L\,\|w^0\|. \quad \square$$

Om ikke annet spesifiseres i det videre, skal vi anta at matrisenormen det refereres til er den som er tilordnet vektornormen brukt i (4.9).

Tilstrekkelig betingelse for stabilitet. Hvis det fins en $\mu \ge 0$ uavhengig av h og k slik at

$$\|C\| \le 1 + \mu\, k,$$

så er (4.8) stabil.

Bevis.

$$\|C^n\| \le \|C\|^n \le (1+\mu\,k)^n \le (1+\mu\,k)^{T/k} = \Big((1+\mu\,k)^{1/(\mu k)}\Big)^{\mu T} \le \mathrm{e}^{\mu T}.$$

Vi minner om at følgen $x_n = (1+1/n)^n$ er monotont voksende og konvergerer mot $e = 2.717\cdots$ når $n \to \infty$. slik at vi kan bruke $L = \mathrm{e}^{\mu T}$ i (4.9).

Nødvendig betingelse for stabilitet. Hvis (4.8) er stabil, fins $\nu \ge 0$ uavhengig av h og k slik at

$$\rho(C) \le 1 + \nu\, k, \tag{4.11}$$

der $\rho(C)$ er spektralradien til C.

Bevis. Siden vi antar at (4.8) er stabil, fins en konstant L slik at $\|C^n\| \le L$ for $n \le T/k$. Videre har vi fra (2.3) at

$$\rho(C)^n = \rho(C^n) \le \|C^n\|.$$

Dermed får vi $\rho(C) \le L^{1/n}$, $n \le T/k$, og spesielt for $n = T/k$ gjelder

$$\rho(C) \le L^{k/T} = \mathrm{e}^{k \ln L/T}.$$

Vi rekkeutvikler nå denne skranken og får

$$\rho(C) \le 1 + \frac{k}{T} \ln L\, \mathrm{e}^{k\theta \ln L/T}, \quad \text{der } 0 < \theta < 1.$$

Vi bruker videre at e^x er en monotont voksende funksjon og at $\theta < 1$ og $k \le T$, dermed blir

$$\rho(C) \le 1 + \frac{k}{T} \ln L \mathrm{e}^{\ln L} = 1 + \frac{k}{T} L \ln L,$$

slik at (4.11) holder med $\nu = \dfrac{L \ln L}{T}$. $\square$

Merknad. Betingelsen $\rho(C) \le 1 + \nu\, k$ er generelt ikke tilstrekkelig for stabilitet. Et (kunstig) moteksempel får vi ved å la $C = C(h) = I + F \in \mathbf{R}^{M\times M}$ og $M = 1/h$. Her har C 1 i element (i,i) og $(i, i-1)$ og 0 ellers (som en Jordanblokk). Det er for eksempel ganske enkelt å vise at $\|C^n\|_\infty = 2^n$ når $0 \le n \le M-1$. Dette i seg selv er nok til å fastslå at (4.8) med dette valget av C ikke kan være stabil. Men siden C er triangulær fins egenverdiene på diagonalen, så $\rho(C) = 1$.

Merk også at stabilitet er normavhengig, det er mulig at et skjema kan være stabilt i en norm, men ikke i en annen.

En fallgrube. Følgende argument er fristende, men galt. Anta at den tilordnede matrisenormen er slik at for diagonalmatriser er $\rho(D) = \|D\|$ (som gjelder for de vanligste normene). Anta også at C er diagonaliserbar, $C = P\Lambda P^{-1}$.

$$\begin{aligned} \|C^n\| &= \|P\Lambda^n P^{-1}\| \le \|P\|\,\|\Lambda^n\|\,\|P^{-1}\| = \|P\|\,\|P^{-1}\|\,\rho(C)^n \\ &\le \|P\|\,\|P^{-1}\|\,(1+\nu k)^n \le \|P\|\,\|P^{-1}\|\,\mathrm{e}^{\nu T}, \end{aligned}$$

så vi har tilsynelatende funnet en skranke for $\|C^n\|$. Problemet ligger i at P kan avhenge av h, k på en slik måte at $\|P\|\,(\|P^{-1}\|) \to \infty$ når $h, k \to 0$.

Viktig spesialtilfelle. Hvis C er symmetrisk er betingelsen $\rho(C) \le 1 + \nu\, k$ både nødvendig og tilstrekkelig for stabilitet når vi benytter $\|\cdot\|_{2,h}$. For symmetriske matriser er nemlig $\|C\|_{2,h} = \rho(C)$.

Stabilitet av θ-metoden på (S/R). Vi minner om skjemaet

$$(I - \theta\, r\, S)U^{n+1} = (1 + (1-\theta)\, r\, S)U^n + d^n,$$

slik at

$$C = (I - \theta\, r\, S)^{-1}\,(I + (1-\theta)\, r\, S).$$

Her er $r = k/h^2$, og S er den symmetriske matrisen definert ved (4.6). En har dermed diagonaliseringen $S = P\Lambda P^T$ der $P^T P = I$. Vi får videre

$$\begin{aligned} I - \theta\, r\, S &= P(I - \theta\, r\, \Lambda)P^T, \\ (I + (1-\theta)\, r\, S) &= P(I + (1-\theta)\, r\, \Lambda)P^T. \end{aligned}$$

Setter vi dette inn i uttrykket for C får vi

$$C = P \underbrace{(I - \theta\, r\, \Lambda)^{-1}(I + (1-\theta)\, r\, \Lambda)}_{\Delta} P^T = P\Delta P^T.$$

Nå er Δ en diagonalmatrise med reelle diagonalelementer

$$\Delta_m = \frac{1 + (1-\theta)\, r\lambda_m}{1 - \theta\, r\lambda_m}$$

Fra diagonaliseringen er det åpenbart at C også er symmetrisk, så det holder å kreve at $\rho(C) \le 1 + \nu\, k$ for en eller annen $\nu \ge 0$. Fra før vet vi at

$$\lambda_m = -4\sin^2\phi_m, \quad \phi_m = \frac{m\pi}{2(M+1)}, \quad m = 1, \ldots, M,$$

og dermed

$$\Delta_m = \frac{1-4(1-\theta)\,r\sin^2\phi_m}{1+4\,\theta\,r\sin^2\phi_m}.$$

Vi antar at $0 \le \theta \le 1$, slik at uttrykket i telleren er ≤ 1, mens nevneren er ≥ 1. Dermed blir $\Delta_m \le 1$ for alle m. Vi forsøker å kreve $\Delta_m \ge -1$, setter inn uttrykket for Δ_m ovenfor i ulikheten, og multipliserer hver side med nevneren (som garantert er positiv)

$$\begin{aligned} 1-4(1-\theta)\,r\sin^2\phi_m &\ge -1-4\,\theta\,r\sin^2\phi_m \\ &\Downarrow \\ 2\,(1-2\theta)\,r\,\sin^2\phi_m &\le 1. \end{aligned}$$

Hvis $\frac{1}{2} \le \theta \le 1$ er venstresiden ≤ 0, så ulikheten er oppfylt uansett verdi av $r \ge 0$. Men hvis $0 \le \theta < \frac{1}{2}$ må vi kreve

$$r \le \frac{1}{2\,r\,(1-2\theta)\sin^2\phi_m}, \quad m = 1,\ldots,M.$$

Høyresiden er minst når $m = M$, det vil si $\phi_m = \frac{M\pi}{2(M+1)} = \frac{\pi}{2} - \frac{h\pi}{2}$. En har

$$\sin^2\left(\frac{\pi}{2} - \frac{h\pi}{2}\right) = \cos^2\frac{h\pi}{2},$$

så kravet må bli

$$r \le \frac{1}{2(1-2\theta)\cos^2(h\pi/2)}.$$

For små verdier av h er $\cos^2(h\pi/2) \approx 1$ og en får et tilstrekkelig krav ved å erstatte den med 1. Oppsummert har vi

Stabilitetskriterium for θ-metoden anvendt på (S/R).

$$\begin{aligned} 0 \le \theta < \tfrac{1}{2} \quad &\Rightarrow \quad \text{Stabil hvis} \quad 0 \le r \le \frac{1}{2(1-2\theta)}, \\ \tfrac{1}{2} \le \theta \le 1 \quad &\Rightarrow \quad \text{Stabil for alle} \quad r \ge 0. \end{aligned}$$

Stabilitet av θ-metode på (S/RD). Om vi anvender θ-metoden på det semidiskretiserte systemet (3.8) får vi et system som kan skrives på formen (4.8) med

$$C = (I - \theta\,r\,Q)^{-1}(I + (1-\theta)\,r\,Q),$$

der Q er gitt av (3.9). Siden Q er usymmetrisk vil også C være usymmetrisk, og en kan ikke uten videre benytte kravet $\rho(C) \le 1 + \nu k$. Men det viser seg at det går bra i dette tilfellet. Vi har sett at ved å innføre matrisen $D = \mathrm{diag}(\sqrt{2}, 1, \ldots, 1, \sqrt{2})$ finner man at $\tilde{Q} = D^{-1}\,Q\,D$ er symmetrisk. Dette innebærer at også $\tilde{C} = D^{-1}CD$ blir symmetrisk. Vi finner at

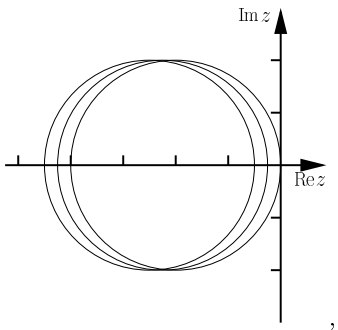
$$\|C^n\| \le \|D\|\,\|D^{-1}\|\,\|\tilde{C}^n\|.$$

For de vanligste normene gjelder at $\|D\| = \sqrt{2}$ og $\|D^{-1}\| = 1$. Siden $\tilde{C}$ er symmetrisk har vi derfor stabilitet hvis $\rho(\tilde{C}) \le 1$. Men C og $\tilde{C}$ er similære, så de har samme egenverdier. Stabilitet oppnås derfor for **(S/RD)** hvis $\rho(C) \le 1$.

Matrisen C har egenverdier

$$\Delta_m = \frac{1+(1-\theta)\,r\,\lambda_m}{1-\theta\,r\,\lambda_m}, \tag{4.12}$$

der λ_m er egenverdiene til Q. Vi kan ikke finne noen formel for egenverdiene til Q, men vi vet at egenverdiene er reelle siden $\tilde{Q}$ er symmetrisk. Vi kan bruke Gershgorins teorem for å finne et tilstrekkelig kriterium for stabilitet.



,

Gershgorinsirklene for alle rader unntatt første og siste er identiske, det er sirkelen lengst til høyre på figuren. Den første og siste sirkelen er tegnet som de to lengst til venstre. Disse er sentrert i punktene $-2(1+\eta_i\,h)$, $i = 0, 1$ og deres venstre skjæringspunkt med den reelle aksen er $-4 - 2\eta_i\,h$, $i = 0, 1$. Egenverdiene til Q ligger altså på den reelle aksen og $-4 - 2\eta\,h \le \lambda \le 0$ der $\eta = \max\{\eta_0, \eta_1\}$. Fra (4.12) følger nå umiddelbart at $\Delta_m \le 1$ for $0 \le \theta \le 1$. Kravet $\Delta_m \ge -1$ gir videre

$$r\,\lambda_m\,(2\theta - 1) \le 2,$$

som er oppfylt for $\theta \ge \frac{1}{2}$, uansett $r > 0$. La $\theta < \frac{1}{2}$.

Om en skulle hatt $\lambda_m = 0$ så holder ulikheten ubetinget. For $\lambda_m < 0$ dividerer vi med den positive størrelsen $-\lambda_m(1-2\theta)$ på hver side. Den kritiske verdien oppstår for egenverdien lengst til venstre, så vi setter inn grensen $\lambda_m \ge -4 - 2\eta\,h$. Til slutt får vi

Stabilitetskriterium for θ-metoden anvendt på (S/RD).

$$\begin{aligned} 0 \le \theta < \tfrac{1}{2} \quad &\Rightarrow \quad \text{Stabil hvis} \quad 0 \le r \le \frac{1}{2(1-2\theta)(1+\frac{\eta h}{2})}, \\ \tfrac{1}{2} \le \theta \le 1 \quad &\Rightarrow \quad \text{Stabil for alle} \quad r \ge 0. \end{aligned}$$

der $\eta = \max\{\eta_0, \eta_1\}$.

4.7 Stabilitet og avrundingsfeil

I numeriske beregninger på datamaskin er det alltid avrundingsfeil fordi maskinen kun regner med et endelig antall siffer. Når vi forsøker å beregne U^{n+1} fra (4.8), er det derfor en annen størrelse vi egentlig finner, definert ved

$$\tilde{U}^{n+1} = C\,\tilde{U}^n + q^n + s^n.$$

Vektoren s^n inneholder avrundingsfeilen som har oppstått i skritt nr n. Om vi definerer feilen som skyldes avrundingsfeil ved $R^n = \tilde{U}^n - U^n$ får vi

$$R^{n+1} = C\,R^n + s^n.$$

Denne løses, og vi får

$$R^n = C^n\,R^0 + C^{n-1}\,s^1 + \cdots + C\,s^{n-2} + s^{n-1}.$$

Anta $R^0 = 0$. Da blir

$$\|R^n\| \le \|C^{n-1}\|\,\|s^0\| + \cdots \|C\|\,\|s^{n-2}\| + \|s^{n-1}\|.$$

Videre skal vi anta at vi har en skranke σ slik at $\|s^\ell\| \le \sigma$ for alle ℓ. Hvis (4.8) er stabil får vi dermed

$$\|R^n\| \le \sigma + \sum_{j=1}^{n-2} L\,\sigma = (1 + (n-2)L)\sigma,$$

så stabilitet sikrer at avrundingsfeilen vokser høyst lineært med n.

4.8 Konsistens og Lax' ekvivalensteorem

Vi minner om (4.7)

$$AU^{n+1} = BU^n + c^n, \tag{4.13}$$

som angir generell form av differenseskjemaer anvendt på en lineær PDL. Om vi setter inn eksakt løsning av PDL'en i formelen, får vi fram lokal avbruddsfeil, vi lar $\tau^n = [\tau_1^n, \ldots, \tau_m^n]^T$, og har pr definisjon

$$Au^{n+1} = Bu^n + c^n + \tau^n.$$

Konsistens. Differenseskjemaet (4.13) sies å være *konsistent* (med differensiallikningen) hvis

$$\frac{1}{k}\tau_m^n \to 0, \quad \text{for alle } m,n \text{ når } h \to 0,\ k \to 0.$$

Merknad. Litteraturen er inkonsistent i definisjonen av lokal avbruddsfeil, og dette påvirker også definisjonen av konsistens. Man lar alternativt avbruddsfeilen være definert som $\widehat{\tau}_m^n = \frac{1}{k}\tau_m^n$ slik at kravet til konsistens blir $\widehat{\tau}_m^n \to 0$ (dette er slik vi gjorde det for ODL i TMA4215). Det er altså ikke nok å kun kreve at $\tau_m^n \leftrightarrow 0$.

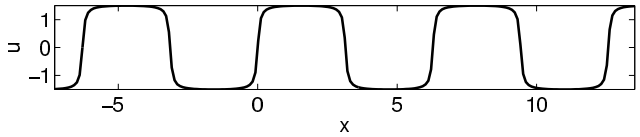
Lax' ekvivalensteorem. En konsistent differensmetode er konvergent hvis og bare hvis den er stabil.

Beviset av Lax' ekvivalensteorem ligger utenfor rammene i dette kurset.

4.9 von Neumanns stabilitetskriterium

Vi ser nå igjen på varmeledningsligningen, men vi bruker en annen type randkrav enn før, nemlig *periodiske*

$$\begin{aligned} u_t &= u_{xx}, \quad -\infty < x < \infty,\ t > 0, \\ u(x,0) &= f(x), \quad -\infty < x < \infty, \\ f(x+2\pi) &= f(x), \quad x \in \mathbf{R}, \\ u(x+2\pi,t) &= u(x,t), \quad x \in \mathbf{R}. \end{aligned}$$



,

En kan utvikle funksjonen $f(x)$ i en Fourierrekke

$$f(x) = \sum_{\beta=-\infty}^{\infty} A_\beta\,\mathrm{e}^{\mathrm{i}\beta x}, \quad A_\beta = \frac{1}{2\pi}\int_0^{2\pi} f(x)\,\mathrm{e}^{-\mathrm{i}\beta x}\,\mathrm{d}x.$$

Ved separasjon av variable, finner en løsninger av formen

$$u_\beta(x,t) = \mathrm{e}^{-\beta^2 t}\,\mathrm{e}^{\mathrm{i}\beta x}, \quad \beta \in \mathbf{Z}.$$

og bruk av initialfunksjonen $f(x)$ gir

$$u(x,t) = \sum_{\beta=-\infty}^{\infty} A_\beta\,\mathrm{e}^{-\beta^2 t}\,\mathrm{e}^{\mathrm{i}\beta x}.$$

La oss nå se på en analog analyse for numerisk løsning, vi bruker først Eulers metode.

$$\begin{aligned} U_m^{n+1} &= (1 + r\,\delta_x^2)U_m^n, \quad h = \frac{2\pi}{M}, \\ U_{m+M}^n &= U_m^n \quad \text{for alle } m \in \mathbf{Z}, \\ U_m^0 &= f(x_m) \quad \text{for alle } m \in \mathbf{Z}. \end{aligned}$$

En kunne nå forsøke å skrive

$$U_m^n = \sum_{\beta=-\infty}^{\infty} A_\beta\,\xi^n\,\mathrm{e}^{\mathrm{i}\beta x_m}, \tag{4.14}$$

denne formelen stemmer for U_m^0. Vi sjekker nå om det er mulig å velge ξ slik at den også er riktig for $n > 0$. Det holder å sjekke ett generelt ledd i rekka (om gangen), så vi setter inn

$$U_m^n = \xi^n\,\mathrm{e}^{\mathrm{i}\beta x_m}.$$

Innsatt i Eulers metode gir dette

$$\xi^{n+1}\,\mathrm{e}^{\mathrm{i}\beta x_m} = \xi^n\,\mathrm{e}^{\mathrm{i}\beta x_m} + r\,(\xi^n\,\mathrm{e}^{\mathrm{i}\beta x_{m-1}} - 2\,\xi^n\,\mathrm{e}^{\mathrm{i}\beta x_m} + \xi^n\,\mathrm{e}^{\mathrm{i}\beta x_{m+1}}).$$

Vi kan forutsette $\xi \neq 0$ og bruke at $x_m = mh$, dermed kan vi dividere hver side med $\xi^n \, \mathrm{e}^{\mathrm{i}\beta x_m}$ og vi får

$$\xi = 1 + r\,(\mathrm{e}^{-\mathrm{i}\beta h} - 2 + \mathrm{e}^{\mathrm{i}\beta h}) = 1 + 2r\,(\cos\beta h - 1) = 1 - 4r\,\sin^2\frac{\beta h}{2}.$$

Vi kan sette $\xi = \xi_\beta$ fra dette uttrykket inn i summen (4.14) og vi finner da at

$$U_m^n = \sum_{\beta=-\infty}^{\infty} A_\beta\, \xi_\beta^n\, \mathrm{e}^{\mathrm{i}\beta x_m}, \quad \xi_\beta = 1 + 2r\,(\cos\beta h - 1),$$

er eksakt løsning av differenslingningen. ξ_β tilsvarer faktoren $\mathrm{e}^{-\beta^2 h}$ i uttrykket for den eksakte løsningen. Vi bør derfor kreve at $|\xi_\beta| \leq 1$ for alle β for at den numeriske løsningen skal være stabil. I dette spesielle tilfellet er ξ_β reell, og vi ser umiddelbart at $\xi_\beta \leq 1$ for alle β. Når vi krever $\xi_\beta \geq -1$ får vi som krav

$$r \leq \frac{1}{2\sin^2\frac{\beta h}{2}}, \quad \text{for alle } \beta,$$

så vi må igjen kreve $r \leq \frac{1}{2}$.

Generelt. Ser på 2-nivås differensformel skrevet på formen

$$\sum_{p=-r}^{r} a_p U_{m+p}^{n+1} = \sum_{p=-s}^{s} b_p U_{m+p}^{n}.$$

Søker en løsning av formen

$$U_m^n \;=\; \xi^n\, \mathrm{e}^{\mathrm{i}\beta x_m},$$

som innsatt i formelen gir

$$\xi^{n+1}\, \mathrm{e}^{\mathrm{i}\beta x_m} \sum_{p=-r}^{r} a_p\, \mathrm{e}^{\mathrm{i}\beta p h} = \xi^n\, \mathrm{e}^{\mathrm{i}\beta x_m} \sum_{p=-s}^{s} b_p\, \mathrm{e}^{\mathrm{i}\beta p h},$$

og dermed

$$\xi \;=\; \frac{\displaystyle\sum_{p=-r}^{r} a_p\, \mathrm{e}^{\mathrm{i}\beta p h}}{\displaystyle\sum_{p=-s}^{s} b_p\, \mathrm{e}^{\mathrm{i}\beta p h}}.$$

Von Neumanns stabilitetskriterium. Det fins en konstant $\mu \geq 0$ slik at

$$|\xi| \leq 1 + \mu\, k.$$

Eksempel. La oss nå bruke differensialligningen

$$u_t = u_{xx} - \lambda u_x.$$

Vi bruker Eulers metode og sentraldifferens også på u_x

$$U_m^{n+1} = U_m^n + \frac{k}{h^2}\,\delta_x^2 U_m^n - \lambda\,\frac{k}{2h}\,(U_{m+1}^n - U_{m-1}^n).$$

Sett $U_m^n = \xi^n\, \mathrm{e}^{\mathrm{i}\beta x_m}$ og merk at $r = \dfrac{k}{h^2}$ impliserer at $\dfrac{k}{2h} = \dfrac{1}{2} r h$.

$$\begin{aligned} \xi &= 1 + r\,(\mathrm{e}^{-\mathrm{i}\beta h} - 2 + \mathrm{e}^{\mathrm{i}\beta h}) - \frac{\lambda r h}{2}(\mathrm{e}^{\mathrm{i}\beta h} - \mathrm{e}^{-\mathrm{i}\beta h}) \\ &= 1 - 4\,r\,\sin^2\frac{\beta h}{2} - i\,\lambda\, r\, h\,\sin\beta h \end{aligned}$$

Vi beregner $|\xi|^2 = (\mathrm{Re}\,\xi)^2 + (\mathrm{Im}\,\xi)^2$,

$$|\xi|^2 = \left(1 - 4\,r\,\sin^2\frac{\beta h}{2}\right)^2 + \lambda^2\, r\, k \sin^2\beta h, \qquad \text{siden } r^2\, h^2 = r\, k.$$

Hvis man skal ha $|\xi| \leq 1 + \mu\, k$ må man nødvendigvis ha

$$\left|1 - 4\,r\,\sin^2\frac{\beta h}{2}\right| \leq 1 + \tilde{\mu}\, k,$$

og fra Euler på $u_t = u_{xx}$ vet vi allerede at dette impliserer $r \leq \frac{1}{2}$. Men hvis $r \leq \frac{1}{2}$ finner vi

$$|\xi|^2 \leq 1 + \frac{1}{2}\lambda^2\, k,$$

og en finner da ved middelverdisetningen at von Neumanns stabilitetskriterium holder med $\mu = \frac{1}{4}\lambda^2$ når $r \leq \frac{1}{2}$.

Sammenheng mellom von Neumann og tidligere definisjon av stabilitet. Anta

1. Differensialligning med konstante koeffisienter og én avhengig variabel (bare én u-komponent)
2. Rent startverdiproblem
3. 2-nivå differenseformel

Da er von Neumann nødvendig og tilstrekkelig for stabilitet.

Likevel brukes von Neumanns stabilitetskriterium som en indikasjon på stabilitet eller instabilitet i langt mer generelle tilfeller.

Eksempel.

$$u_t = a(x,u)\,u_{xx}, \qquad (+\text{randkrav \& startkrav}).$$

Euler

$$U_m^{n+1} = U_m^n + r\,a(x,U_m^n)\,\delta_x^2 U_m^n.$$

Med a konstant har von Neumann kriteriet formen

$$|\xi| = \left|1 - 4\,a\,r\,\sin^2\frac{\beta h}{2}\right| = 1 + \mathcal{O}(k).$$

La oss for eksempel anta at en kan finne en skranke for a slik at $1 \le a(x,u) \le 2$ for alle x, u av interesse. Da kunne man forlange

$$r \le \frac{1}{2\,\max a} = \frac{1}{4},$$

uten at vi med dette har gjort noen matematisk holdbar analyse.

Kapittel 5

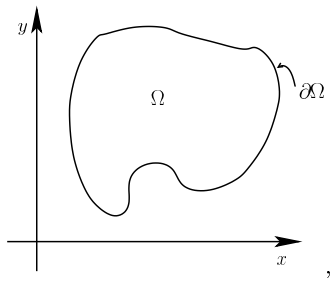
Elliptiske differensialligninger

5.1 Elliptisk ligning i planet

Vi skal nå se på partielle differensialligninger av typen

$$a\,u_{xx} + 2b\,u_{xy} + c\,u_{yy} = d(x,y,u,u_x,u_y), \qquad (x,y) \in \Omega$$

der a, b, c og d kan være funksjoner av x og y.



,

Ellipticitet. Hvis funksjonene a, b og c for alle $(x,y) \in \Omega$ oppfyller

$$ac - b^2 > 0$$

er differensialligningen elliptisk i Ω.

Eksempel. Hvis $a = c = 1, b = d = 0$ får vi Laplaceligningen

$$u_{xx} + u_{yy} = 0.$$

Randkrav. Det fins tre vanlige typer av randkrav

1. Dirichlet randkrav: $u = f$ på $\partial\Omega$.
2. Neumann randkrav: $\frac{\partial u}{\partial n} = \vec{n} \cdot \nabla u = g$ på $\partial\Omega$.

3. Robin randkrav: $\alpha u + \beta \frac{\partial u}{\partial n} = \gamma$ på $\partial\Omega$.

Ofte kan man ha en blanding av 1–3, slik at randen $\partial\Omega$ kan deles opp i flere biter, og med forskjellig type randkrav på de forskjellige bitene.

Maksimumsprinsippet. La Ω være en åpen sammenhengende delmengde av $\mathbf{R}^2$. Definer differensialoperatoren L ved

$$Lu = a\,u_{xx} + 2b\,u_{xy} + c\,u_{yy} + d\,u_x + e\,u_y$$

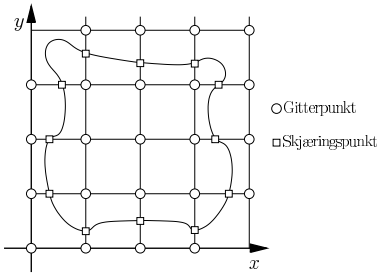
der a, b, c, d og e er funksjoner av x og y, og L er elliptisk ($ac > b^2$) i Ω.

Hvis $Lu = 0$ i Ω, kan ikke u anta noe strengt relativt maksimum eller minimum i Ω medmindre $u =$ konstant i Ω.

Alternativ formulering. Hvis u er kontinuerlig i $\bar{\Omega} = \Omega \cup \partial\Omega$ og $Lu = 0$ i Ω vil u anta maksimum og minimum på $\partial\Omega$.

5.2 Differensmetoder via Taylor

Vi starter med et regulært gitter.



Gitterlinjer: $x = x_\ell,\ y = y_m$

Gitterpunkter: $P = (x_\ell, y_m)$

Vi søker approksimasjon til eksakt løsning på et *nett* som består av gitterpunkter og skjæringspunkter mellom gitterlinjer og $\partial\Omega$.

Vi definerer nå noen delmengder av nettet

$G = \{(x_\ell, y_m)\}$: Hele gitteret

$\mathcal{N}^o$: $G \cap \Omega$, mengden av indre gitterpunkt

$D = \{(x,y) \in \partial\Omega : x = x_\ell \text{ eller } y = y_m\}$

$\mathcal{N} = \mathcal{N}^o \cup D$

Skjæringspunkter kan skape problemer, fordi de har en tendens til å ødelegge nøyaktigheten i den numeriske løsningen og kan også forstyrre matrisestruktur som er nødvendig for å få hurtige ligningsløsere.

Gitterlignende nett. Et nett er gitterlignende hvis alle punkt i mengden D er gitterpunkt. I det videre ser vi på gitterlignende nett med konstante skrittlengder, det vil si at $x_\ell = x_0 + \ell h,\ \ell = 0, 1, \ldots$ og $y_m = y_0 + mh,\ m = 0, 1, \ldots$.

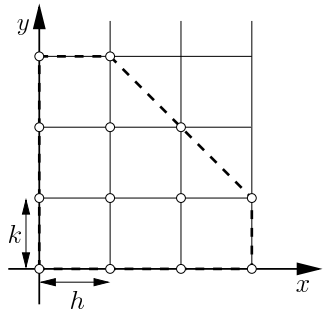
Poisson's ligning.

$\Delta u = u_{xx} + u_{yy} = f$ i Ω, $\qquad u = g(x,y)$ på $\partial\Omega$.

$$\partial_x^2 u_p = \frac{1}{h^2}\,\delta_x^2 u_p + \mathcal{O}(h^2)$$

$$\partial_y^2 u_p = \frac{1}{k^2}\,\delta_y^2 u_p + \mathcal{O}(k^2)$$

der u_p er eksakt løsning av diffligning i $p = (x,y)$.



Om vi diskretiserer differensialligningen i punktet $p = (x,y)$ får vi altså

$$\Delta u_p = f_p \longrightarrow \frac{1}{h^2}\delta_x^2 u_p + \frac{1}{k^2}\delta_y^2 u_p = f_p + \tau_p$$

der f_p er funksjonen f evaluert i punktet p, og avbruddsfeilen τ_p er

$$\tau_p = \frac{1}{12}\,h^2\,\partial_x^4 u_p + \frac{1}{12}\,k^2\,\partial_y^4 u_p + \cdots \tag{5.1}$$

Videre lar vi nå U_p være approksimasjon til u_p og vi får ligningssystemet

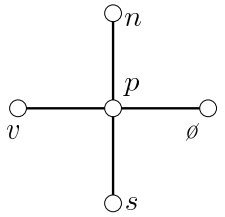
$$\begin{aligned} &\tfrac{1}{h^2}\delta_x^2 U_p + \tfrac{1}{k^2}\delta_y^2 U_p = f_p, && p \in \mathcal{N}^o, \\ &U_p = g_p && p \in D. \end{aligned}$$

Klassisk 5-punkts formel for Poissons ligning. I tilfellet $k = h$ fås

$$\delta_x^2 U_p + \delta_y^2 U_p = h^2\,f_p$$

Vi kan skrive ut dette ved å bruke indekser som refererer til himmelretningene

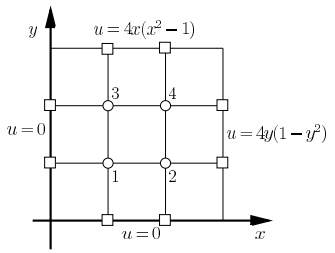
$$U_v + U_s + U_\emptyset + U_n - 4U_p = h^2\,f_p$$



Eksempel. Vi regner gjennom et konkret tilfelle.

$$\Delta u = 0, \quad (x,y) \in \Omega = (0,1) \times (0,1), \qquad u(x,y) = g(x,y), \quad (x,y) \in \partial\Omega$$

der $g(x,y)$ er som beskrevet i figuren.



$$
\begin{aligned}
g(0,y) &= 0, && 0 \le y \le 1\\
g(x,0) &= 0, && 0 \le x \le 1\\
g(1,y) &= 4y(1-y^2), && 0 \le y \le 1\\
g(x,1) &= 4x(x^2-1), && 0 \le x \le 1
\end{aligned}
$$

$$
\begin{array}{lrrrrcr}
p=1: & -4U_1 & +U_2 & +U_3 & & = & 0\\
p=2: & U_1 & -4U_2 & & +U_4 & = & -\frac{32}{27}\\
p=3: & U_1 & & -4U_3 & +U_4 & = & \frac{32}{27}\\
p=4: & & U_2 & +U_3 & -4U_4 & = & 0.
\end{array}
$$

Løser man disse ligningene får man

$$U_1 = U_4 = 0, \quad U_2 = \frac{8}{27}, \quad U_3 = -\frac{8}{27}.$$

Du kan sjekke at eksakt løsning av problemet er

$$u(x,y) = 4xy\,(x^2 - y^2).$$

Ser vi på (5.1), ser vi at avbruddsfeilen er identisk lik null fordi $\partial_x^4 u \equiv 0$ og $\partial_y^4 u \equiv 0$ og dermed alle høyere ordens deriverte. Så i dette tilfellet gir formelen eksakt riktig løsning. Dette er veldig spesielt, med litt andre randbetingelser vil $\tau_p \neq 0$.

5.2.1 Diskretisering av en selvadjungert ligning

Vi studerer problemet

$$Lu = f, \qquad \text{der} \quad Lu = \partial_x(a\partial_x u) + \partial_y(c\partial_y u), \quad a = a(x,y),\ c = c(x,y).$$

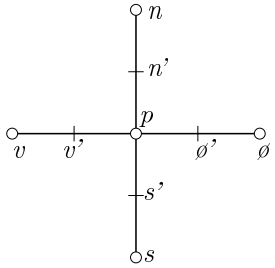
$$\partial_x(a\partial_x u) = \frac{1}{h^2}\,(a_{ø'}(u_ø - u_p) - a_{v'}(u_p - u_v)) + \mathcal{O}(h^2)$$

$$\partial_y(c\partial_y u) = \frac{1}{k^2}\,(c_{n'}(u_n - u_p) - c_{s'}(u_p - u_s)) + \mathcal{O}(k^2)$$

Så $Lu = f$ kan diskretiseres til

$$\alpha_ø U_ø + \alpha_n U_n + \alpha_v U_v + \alpha_s U_s - \alpha_p U_p = f_p,$$

$$\text{der} \quad \alpha_p = \alpha_ø + \alpha_n + \alpha_v + \alpha_s$$

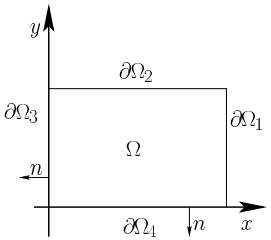


og

$$\alpha_ø = \frac{1}{h^2}a_{ø'},\ \alpha_n = \frac{1}{k^2}c_{n'},\ \alpha_v = \frac{1}{h^2}a_{v'},\ \alpha_s = \frac{1}{k^2}c_{s'}.$$

5.3 Randkrav av Neumanns og Robins type

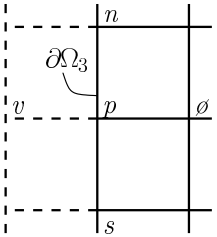
Eksempel.



$$
\begin{aligned}
&\Delta u = 0 && \text{i } \Omega\\
&u = g(x,y) && \text{på } \partial\Omega_1 \cup \partial\Omega_2\\
&\partial_x u - q^{(0)}\,u = f^{(0)} && \text{på } \partial\Omega_3\\
&\partial_y u - q^{(1)}\,u = f^{(1)} && \text{på } \partial\Omega_4
\end{aligned}
$$

Vi krever at $q^{(0)} \ge 0$ og $q^{(1)} \ge 0$. Nå trenger vi ligninger for U_p for alle $p \in \mathcal{N}^o$, samt alle p på $\partial\Omega_3$ og $\partial\Omega_4$. La oss for enkelthets skyld bruke samme skrittlengde i begge retninger, vi setter altså $k = h$. La oss bruke den venstre randen $\partial\Omega_3$ som eksempel.

$$\partial_x u - qu = f$$



Alternativ 1.

$$\partial_x u_p = \frac{u_ø - u_p}{h} + \mathcal{O}(h) \quad \longrightarrow \quad \frac{U_ø - U_p}{h} - q_p^{(0)}\,U_p = f_p^{(0)}$$

der $q_p^{(0)}$ og $f_p^{(0)}$ er de gitte funksjonene $q^{(0)}$ og $f^{(0)}$ evaluert i punktet p.

Alternativ 2. Bruk fiktivt punkt v som ligger utenfor området, se figuren.

$$\partial_x u_p = \frac{u_ø - u_v}{2h} + \mathcal{O}(h^2) \quad \longrightarrow \quad \frac{U_ø - U_v}{2h} - q_p^{(0)}\,U_p = f_p^{(0)} \tag{5.2}$$

Den ekstra ukjente U_v krever en ekstra ligning, og vi bruker skjemaet for selve differensialligningen i punktet p

$$U_ø + U_n + U_v + U_s - 4U_p = 0, \tag{5.3}$$

og vi eliminerer U_v ved hjelp av det diskretiserte randkravet. Fra (5.2) finner vi

$$U_v = U_ø - 2h\,(q_p^{(0)}\,U_p + f_p^{(0)}),$$

som innsatt i (5.3) gir

$$2\,U_ø + U_n + U_s - (4 + 2\,h\,q_p^{(0)})U_p = 2\,h\,f_p^{(0)}.$$

Denne diskretiseringen er mer nøyaktig enn alternativ 1.

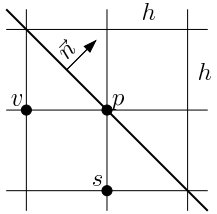
Man gjør helt tilsvarende på randen $\partial\Omega_4$. Vi ser der kun på alternativ 2 med fiktivt randpunkt s, og p på $\partial\Omega_4$:

$$\partial_y u_p = \frac{u_n - u_s}{2h} + \mathcal{O}(h^2) \quad \longrightarrow \quad \frac{U_n - U_s}{2h} - q_p^{(1)} U_p = f_p^{(1)}.$$

Vi kan her også bruke (5.3) og eliminere bort den fiktive verdien U_s, resultatet blir

$$2\,U_n + U_v + U_\o - (4 + 2\,h\,q_p^{(1)})U_p = 2\,h\,f_p^{(1)}.$$

Rand langs gitterdiagonal



$$\partial_n u + qu = \vec{n}\cdot\nabla u + qu = f$$
$$\vec{n} = [n_x, n_y]^T$$
$$n_x = n_y = \tfrac{1}{\sqrt{2}} \quad \text{fordi } k = h$$
$$\Downarrow$$
$$\partial_n u = n_x\partial_x u + n_y\partial_y u = \tfrac{1}{\sqrt{2}}(\partial_x u + \partial_y u)$$

$$\partial_n u_p = \frac{1}{\sqrt{2}}\left(\frac{u_p - u_v}{h} + \frac{u_p - u_s}{h}\right) + \mathcal{O}(h)$$

slik at en førsteordens tilnærmelse blir

$$(2 + \sqrt{2}\,h\,q_p)U_p - U_v - U_s = \sqrt{2}\,h\,f_p$$

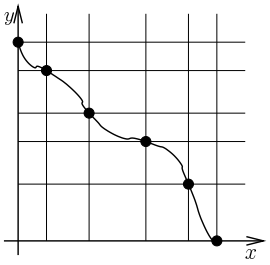
Man nøyer seg ofte med en slik grov tilnærmelse for denne type rand. I prinsippet kan man også her bruke alternativ 2 med fiktivt gitterpunkt, men det blir fort veldig kompliserte formler.

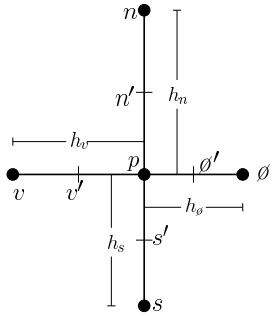
Et problem forbundet med rent Neumann problem for Laplace's ligning. Problemet

$$\begin{aligned} \Delta u &= 0, && \text{i } \Omega \\ \partial_n u &= g, && \text{på } \partial\Omega \end{aligned}$$

har bare løsning hvis $\int_{\partial\Omega} g = 0$. Dette ser man ved å bruke divergensteoremet på $\int_\Omega \Delta u$. Hvis u er en løsning, ser vi umiddelbart at også $u + c$ er en løsning for en vilkårlig konstant c. Dermed har vi intet *velformet* PDL-problem. Dette kommer igjen i det diskretiserte problemet der vi typisk vil få et lineært ligningssystem av typen $AU = b$ hvor A er en kvadratisk matrise som er singulær. Hvis b ikke tilhører kolonnerommet til A har vi ingen løsning. I motsatt fall har vi en løsning, men den er ikke entydig, siden vi kan addere en vilkårlig vektor fra (det ikke-trivielle) nullrommet til A.

5.4 Gitterlignende nett og varierende skrittlengder





Vi ser på approksimasjon av

$$Lu = \partial_x(a\partial_x u) + \partial_y(c\partial_y u)$$

Vi lar

$$\partial_x(a\partial_x u) \quad \longrightarrow \quad L_h^{(x)}U_p = \frac{2}{h_v + h_\o}\left(a_{\o'}\frac{U_\o - U_p}{h_\o} - a_{v'}\frac{U_p - U_v}{h_v}\right)$$

og

$$\partial_y(c\partial_y u) \quad \longrightarrow \quad L_h^{(y)}U_p = \frac{2}{h_s + h_n}\left(c_{n'}\frac{U_n - U_p}{h_n} - c_{s'}\frac{U_p - U_s}{h_s}\right)$$

Vi refererer til figuren ovenfor til høyre for definisjon av punktene $p, v, v', s, s', \o, \o', n, n'$ og de tilhørende skrittlengder. Vi approksimerer L med

$$L_h = L_h^{(x)} + L_h^{(y)}$$

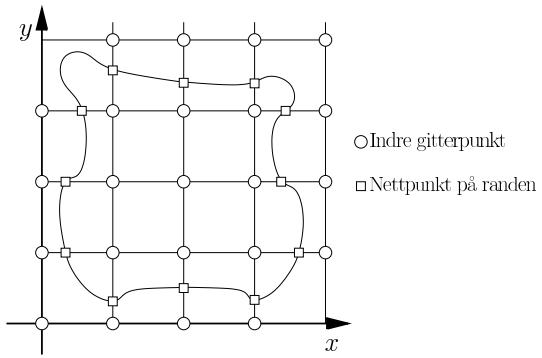
og finner ved

$$\begin{aligned} L_h^{(x)}u_p &= \frac{2}{h_v + h_\o}\Big((1 + \frac{h_\o}{2}\partial_x + \frac{h_\o^2}{8}\partial_x^2 + \frac{h_\o^3}{48}\partial_x^3 + \cdots)(a_p(\partial_x + \frac{1}{24}h_\o^2\partial_x^3 + \cdots)u_p) \\ &\qquad -(1 - \frac{h_v}{2}\partial_x + \frac{h_v^2}{8}\partial_x^2 - \frac{h_v^3}{48}\partial_x^3 + \cdots)(a_p(\partial_x + \frac{1}{24}h_v^2\partial_x^3 + \cdots)u_p)\Big) \\ &= \partial_x(a\partial_x)u_p + \frac{1}{3}(h_\o - h_v)\partial_x^2(a\partial_x)u_p + \frac{1}{24}\frac{h_\o^3 + h_v^3}{h_\o + h_v}(\partial_x a\partial_x^3 + \partial_x^3 a\partial_x)u_p + \ldots \end{aligned}$$

Helt tilsvarende utregning kan man gjøre for $L_h^{(y)}u_p$ og til slutt finner man at

$$Lu - L_h u = \begin{cases} \mathcal{O}((h_\o - h_v) + (h_n - h_s) + h_\o^2 + h_n^2) & \text{generelt} \\ \mathcal{O}(h_\o^2 + h_n^2) & h_\o = h_v,\ h_n = h_s. \end{cases}$$

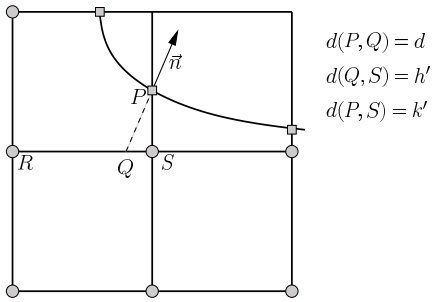
5.5 Generelt rektangulært nett



Vi ser på ligningen $\Delta u = f$ som eksempel.

Dirichletproblem. Vi kan benytte den generelle 5-punktsformelen med $h_\o$, h_v, h_n, h_s for alle indre gitterpunkter.

Robinproblem. Vanskeligheten er her $\partial_n u = \vec{n} \cdot \nabla u$. Vi lar gitteret være rektangulært med skrittlenger h og k i henholdsvis x- og y-retning.



Vi har

$$\partial_n u_P = \frac{u_P - u_Q}{d} + \mathcal{O}(d), \qquad d = \sqrt{h'^2 + k'^2}$$

Problemet er at Q ikke er et gitterpunkt, men vi kan approksimere løsningen i punktet Q ved lineær interpolasjon. Vi finner

$$u_Q = u_R \frac{h'}{h} + u_S \frac{h - h'}{h} + \mathcal{O}(h^2)$$

Dermed blir

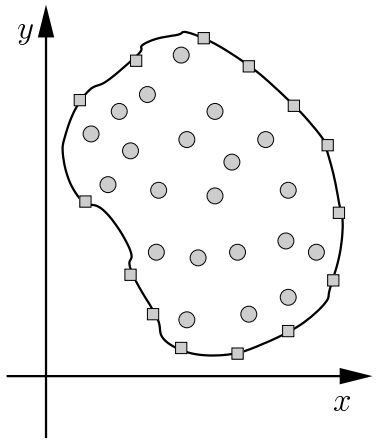
$$\partial_n u_P = \frac{1}{d}\left(u_P - (u_R \frac{h'}{h} + u_S \frac{h - h'}{h})\right) + \mathcal{O}(h^2/d) + \mathcal{O}(d)$$

Men vi bemerker at denne typen problem er generelt vanskelig å sette opp.

5.6 Diskretisering via Taylor på fullstendig gerenelt nett

Vi skal diskretisere operatoren

$$Lu = au_{xx} + 2bu_{xy} + cu_{yy} + du_x + eu_y + fu$$



La P være en vilkårlig indre node. Vi velger ut s noder blant de øvrige punktene i nettet, typisk de s nærmeste naboene til P. La h være en karakteristisk gitterstørrelse. Vi beskriver beliggenheten til punktet Q_i relativt til P ved koordinater

$$\overline{PQ}_i = (\xi_i h, \eta_i h).$$

Vi approksimerer operatoren L med L_h der vi forlanger at

$$L_h U_P = \sum_{i=0}^{s} \alpha_i U_{Q_i} - \alpha_0 U_P.$$

for et valg av konstanter $\alpha_0, \ldots, \alpha_s$.

Som vanlig setter vi inn eksakt løsning u i denne formelen og rekkeutvikler ved å bruke Taylor i 2 dimensjoner som i (2.4). Vi får da

$$u_{Q_i} = u_P + \xi_i h \partial_x u_P + \eta_i h \partial_y u_P + \frac{1}{2}\xi_i^2 h^2 \partial_x^2 u_P + \xi_i \eta_i h^2 \partial_x \partial_y u_P + \frac{1}{2}\eta_i^2 h^2 \partial_y^2 u_P + \cdots,$$

som innsatt i uttrykket for $L_h u_P$ gir

$$\begin{aligned} L_h u_p &= \left(\sum_{i=1}^{s} \alpha_i - \alpha_0\right) u_P + \left(h \sum_{i=1}^{s} \xi_i \alpha_i\right) \partial_x u_P + \left(h \sum_{i=1}^{s} \eta_i \alpha_i\right) \partial_y u_P + \left(\frac{1}{2} h^2 \sum_{i=1}^{s} \xi_i^2 \alpha_i\right) \partial_x^2 u_P \\ &\quad + \left(h^2 \sum_{i=1}^{s} \xi_i \eta_i \alpha_i\right) \partial_x \partial_y u_P + \left(\frac{1}{2} h^2 \sum_{i=1}^{s} \eta_i^2 \alpha_i\right) \partial_y^2 u_P \end{aligned}$$

Dette bør være "mest mulig likt Lu_p", og vi bør derfor kreve at

$$\begin{aligned}
\sum_{i=1}^{s}\alpha_i &= \alpha_0 + f \\
\sum_{i=1}^{s}\xi_i\alpha_i &= \frac{d}{h} \\
\sum_{i=1}^{s}\eta_i\alpha_i &= \frac{e}{h} \\
\sum_{i=1}^{s}\xi_i^2\alpha_i &= \frac{2a}{h^2} \\
\sum_{i=1}^{s}\xi_i\eta_i\alpha_i &= \frac{2b}{h^2} \\
\sum_{i=1}^{s}\eta_i^2\alpha_i &= \frac{2c}{h^2}
\end{aligned}$$

I tillegg bør følgende ligninger være oppfylt for flest mulig verdier av ℓ

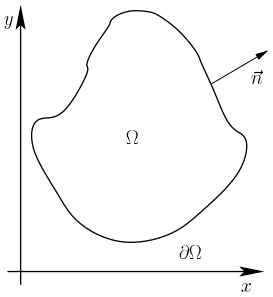
$$\sum_{i=1}^{s}\xi_i^{\ell}\,\eta_i^{\ell-m}\,\alpha_i = 0,\ m = 0,\ldots,\ell,\quad \ell = 3,4,\ldots.$$

Dette ser vi ved å kikke på resten av Taylorrekka til $L_h u_P$, leddene fra h^3 og oppover

$$\sum_{\ell=3}^{\cdots}\frac{h^\ell}{\ell!}\sum_{m=0}^{\ell}\binom{\ell}{m}\sum_{i=1}^{s}(\alpha_i\,\xi_i^m\,\eta_i^{\ell-m})\,\partial_x^m\partial_y^{\ell-m}u_P.$$

5.7 Differensformler utledet via integrasjon

Denne teknikken kalles ofte boksintegrasjon, og er nær relatert til det som i litteraturen kalles for endelige volum-metoder.



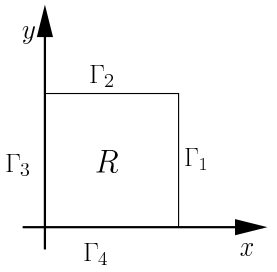
Gauss' divergensteorem. Gitt et område Ω som på figuren med rand $\partial\Omega$ som i hvert punkt har en utadrettet normalvektor $\vec{n}$. La $\vec{p} = \vec{p}(x,y)$ være et vektorfelt i Ω. Da gjelder

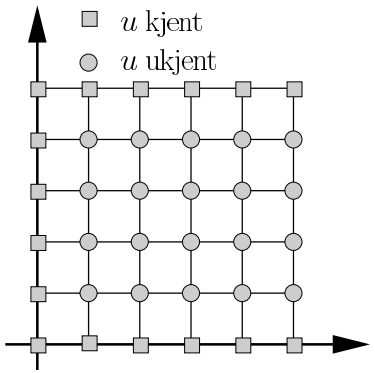
$$\int_\Omega \operatorname{div}\vec{p}\,\mathrm{d}A = \oint_{\partial\Omega}\vec{p}\cdot\vec{n}\,\mathrm{d}s$$

Med $\operatorname{div}\vec{p} = \nabla\cdot\vec{p}$ menes divergensen til vektorfeltet $\vec{p}$. Spesielt hvis $\vec{p} = \nabla u$ (gradientvektorfelt) får man at $\operatorname{div}\vec{p} = \Delta u = \partial_x^2 u + \partial_y^2 u$. Dermed får vi

$$\int_\Omega \Delta u\,\mathrm{d}A = \oint_{\partial\Omega}\nabla u\cdot\vec{n}\,\mathrm{d}s = \oint_{\partial\Omega}\partial_n u\,\mathrm{d}s$$

Vi illusterer boksintegrasjon gjennom et spesifikt eksempel. La $\Omega = R$ være et rektangel med render $\Gamma_1,\ldots,\Gamma_4$ som på figuren.





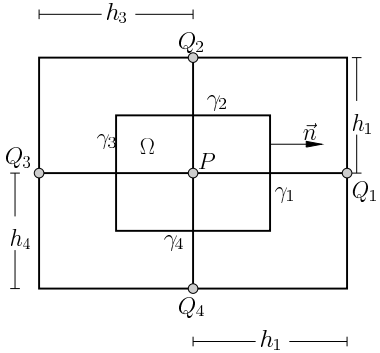
Vi ser på problemet

$$\begin{aligned}
&\Delta u = f, && \text{i } R.\\
&\partial_n u + au = d, && \text{på } \Gamma_1,\\
&u = g, && \text{på } \Gamma_2\cup\Gamma_3\cup\Gamma_4.
\end{aligned}$$

Her har vi innført et rektangulært gitter på R, merk at det kan gjerne være varierende skrittlengder i begge retninger.

La nå P være et indre punkt i R. Vi ser først på rektanglet avgrenset av nabogitterlinjene til de P ligger på (se figuren). Deretter innfører vi et nytt mindre rektangel som på figuren, vi kaller dette Ω. Sidekantene i Ω, som er sentrert mellom gitterlinjene, kalles $\gamma_i,\ i = 1,2,3,4$. Lengden av γ_i kalles ℓ_i. Vi lar skrittlengdene ut fra P være h_1 (mot høyre), h_2 (oppover), h_3 (mot venstre), h_4 (nedover). Dermed blir i følge figuren

$$\ell_1 = \ell_3 = \frac{h_2+h_4}{2},\ \ell_2 = \ell_4 = \frac{h_1+h_3}{2}.$$



Arealet Ω blir $A = \frac{1}{4}(h_1+h_3)(h_2+h_4)$. Nå bruker vi Gauss' divergensteorem på det lille rektanglet Ω og finner

$$\Delta u = f \quad\Rightarrow\quad \int_\Omega \Delta u\,\mathrm{d}A = \underbrace{\oint_{\partial\Omega}\partial_n u\,\mathrm{d}s}_{I} = \underbrace{\int_\Omega f\,\mathrm{d}A}_{II}.$$

Vi forsøker å approksimere I og II.

$$\text{I:}\quad \oint_{\partial\Omega}\partial_n u\,\mathrm{d}s = \sum_{i=1}^{4}\int_{\gamma_i}\partial_n u\,\mathrm{d}s \approx \sum_{i=1}^{4}\ell_i\,\frac{u_{Q_i}-u_P}{h_i}$$

$$\text{II:}\quad \int_\Omega f\,\mathrm{d}A \approx f_P\,A = \frac{1}{4}f_P(h_1+h_3)(h_2+h_4).$$

Dermed blir vår differensformel

$$\sum_{i=1}^{4} \frac{\ell_i}{Ah_i}(U_{Q_i} - U_P) = f_P$$

Vi kan alternativt skrive om formelen til det "gamle formatet"

$$\sum_{i=1}^{4} \alpha_i U_{Q_i} - \alpha_0 U_P = f_P$$

der

$$\alpha_1 = \frac{2}{h_1(h_1+h_3)},\ \alpha_2 = \frac{2}{h_2(h_2+h_4)},\ \alpha_3 = \frac{2}{h_3(h_1+h_3)},\ \alpha_4 = \frac{2}{h_4(h_2+h_4)},\ \alpha_0 = \frac{2}{h_1h_3} + \frac{2}{h_2h_4}.$$

Denne formelen brukes for alle indre punkt i R. Men vi må ha ligninger også for de ukjente på randen Γ_1.
Nå sammenfaller γ_1 med den ytre randen Γ_1 der vi har

$$\partial_n u + au = d$$

Gauss' divergensteorem på Ω gir igjen

$$\sum_{i=1}^{4} \int_{\gamma_i} \partial_n u \,\mathrm{d}s = \int_\Omega f \,\mathrm{d}A$$

Ser vi spesielt på randen γ_1 der $\partial_n u$ er gitt som $d - au$ har vi

$$\int_{\gamma_1} \partial_n u \,\mathrm{d}s = \int_{\gamma_1} (d - au)\,\mathrm{d}s \approx \ell_1(d_P - a_P u_P)$$

der a_P og d_P er funksjonene a og d evaluert i punktet P. På de øvrige rendene setter vi

$$\int_{\gamma_1} \partial_n u \,\mathrm{d}s \approx \ell_i \frac{u_{Q_i} - u_P}{h_i},\ i = 2,3,4.$$

Vi ender dermed opp med følgende differensformel for et punkt P som ligger på randen Γ_1.

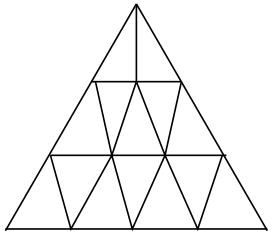
$$\ell_1(d_P - a_P U_P) + \sum_{i=2}^{4} \frac{\ell_i}{h_i}(U_{Q_i} - U_P) = f_P A$$

der

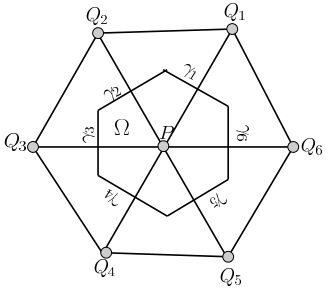
$$A = \frac{1}{4}h_3(h_2+h_4),\quad \ell_2 = \ell_4 = \frac{1}{2}h_3,\quad \ell_3 = \frac{h_2+h_4}{2}$$

5.8 Nett basert på trekanter

Det fine med boksintegrasjon er at det ikke forutsetter at området vårt er delt inn i rektangler, for eksempel trekanter kan brukes like enkelt. Det er ofte lettere å dele opp et område som ikke er rektangulært i trekanter enn i rektangler. Trekantene behøver ikke å være like eller ensformede, men vi krever her at alle vinkler er mindre enn 90 grader. I det videre forutsetter vi også at vi ikke har såkalte "hengende noder", vi krever at ingen noder får plasseres på en trekants sidekant (men kun i hjørnene).



I den neste figuren har vi tegnet et utsnitt av trekantnettet, der s trekanter ($s = 6$ i figuren) har det indre gitterpunktet P som felles hjørne. Sidekanten γ_i i det indre polygonet skjærer ortogonalt og midt på linjestykket $\overline{PQ_i}$. Vi lar γ_i har lengde ℓ_i og linjestykkene $\overline{PQ_i}$ har lengde h_i. Vi antar videre at Ω har areal A. Vi bruker Gauss' divergensteorem på området Ω, og får



$$\sum_{i=1}^{s} \int_{\gamma_i} \partial_n u \,\mathrm{d}s = \int_\Omega f \,\mathrm{d}A$$

Nå approksimerer vi som før

$$\int_{\gamma_i} \partial_n u \,\mathrm{d}s \approx \ell_i \frac{U_{Q_i} - U_P}{h_i}$$

slik at den endelige formelen blir

$$\sum_{i=1}^{s} \frac{\ell_i}{h_i A}(U_{Q_i} - U_P) = f_P$$

Vi har ikke oppgitt spesifikke formler for utregning av arealet A eller noen relasjoner i mellom ℓ_i og h_i (dette er heller ikke mulig i det helt generelle tilfellet vi har drøftet). Det fins et viktig alternativ til boksintegrasjon, nemlig de såkalte endelig elementmetodene. Disse bygger på et helt annerledes matematisk fundament enn det vi har presentert her. Kurset *TMA4220 Numerisk løsning av partielle differensialligninger med elementmetoden* tar for seg slike metoder i stor utstrekning.

5.9 Differensligningene

La oss først skrive elliptisk ligning med randbetingelser på abstrakt form som

$$\begin{aligned} Lu &= f \quad \text{i } \Omega \\ Bu &= g \quad \text{på } \partial\Omega \end{aligned}$$

Vi innfører diskretisering av dette problemet ved

$$\alpha_0 U_P - \sum_{i=1}^{s} \alpha_i U_{Q_i} = \beta_P. \tag{5.4}$$

Vi lar P løpe gjennom alle de punktene hvor u skal approksimeres. Her er Q_i et nabopunkt til P for $i = 1, \ldots, s$. Koeffisientene $\alpha_1, \ldots, \alpha_s$ kan avhenge av P, det samme kan s.

Ønskelige egenskaper for (5.4)**.**

i.
$$\left.\begin{array}{l}\alpha_0 > 0\\ \alpha_i \geq 0\end{array}\right\} \quad \text{for enhver } P.$$

ii.
$$\alpha_0 \geq \sum_{i=1}^{s} \alpha_i \qquad \text{for enhver } P,$$

det vi i lineær ligningsløsning kaller *diagonaldominans*. Vi krever dessuten ekte ulikhet for *minst* en P.

iii. Koeffisientmatrisen er symmetrisk

Om vi sier at koeffisienten som brukes i node P foran node Q er $\alpha_{P,Q}$, betyr symmetriegenskapen at $\alpha_{P,Q} = \alpha_{Q,P}$.

Vi minner om maksimumsprinsippet beskrevet i kapittel 5.1. Det viser seg at et tilsvarende prinsipp gjelder for differensligninger av typen beskrevet ovenfor.

Diskret maksimumsprinsipp. Anta at en differensligning oppfyller (i) og (ii) ovenfor. Anta: Størrelsen V_P er definert for alle $P \in \Omega \cup \partial\Omega$ og at

$$\alpha_0 V_P - \sum_i \alpha_i V_{Q_i} \leq 0$$

for enhver $P \in \Omega$ (indre gitterpunkt).
Da gjelder at

$$V_P \leq \max_{S \in \partial\Omega} V_S \quad \text{for enhver } P \in \Omega.$$

Bevis. Anta det motsatte, nemlig at det fins en $P^* \in \Omega$ slik at

$$V_{P^*} = \max_{P \in \Omega} V_P > \max_{S \in \partial\Omega} V_S \tag{5.5}$$

Anvender vi derfor antagelsen i teoremet for dette punktet får vi

$$\alpha_0 V_{P^*} \leq \sum_i \alpha_i V_{Q_i}$$

$$\Updownarrow \text{ (i)}$$

$$V_p^* \leq \sum_i \frac{\alpha_i}{\alpha_0} V_{Q_i} \overset{\text{(ii)}}{\leq} \frac{1}{\sum_j \alpha_j} \sum_i \alpha_i V_{Q_i} = \sum_i \gamma_i V_{Q_i}$$

hvor

$$\gamma_i = \frac{\alpha_i}{\sum_j \alpha_j} \qquad \text{slik at } \sum_i \gamma_i = 1.$$

Vi får dermed at

$$\sum_i \gamma_i V_{Q_i} \leq \sum_i \gamma_i \max_i V_{Q_i} = \max_i V_{Q_i}$$

og dermed $V_{P^*} \leq \max_i V_{Q_i}$. I følge (5.5) må derfor $V_{Q_i} = V_{P^*}$ for alle i. Vi kan dermed anvende det samme argumentet på hvert nabopunkt V_{Q_i}, osv helt inntil enhver nabo blir et randpunkt S. Dermed har vi at $V_S = V_{P^*}$ for alle $S \in \partial\Omega$ som er en selvmotsigelse (i forhold til (5.5)) og vi konkluderer med at teoremet er sant.

5.10 Konvergens av metoder for elliptiske ligninger

5.10.1 Konvergensbevis for 5-punktsformelen på et Dirichletproblem

La oss se på problemet

$$\begin{aligned} -\Delta u &= f \quad \text{i } R \\ u &= g \quad \text{på } \partial R \end{aligned}$$

der R er kvadratet $(0,1) \times (0,1)$.

Sett $h = \frac{1}{M}$ og anvend 5-punktsformelen $L_h U_p = f_p$ hvor

$$\begin{aligned} L_h U_p &= \frac{1}{h^2}(4U_p - U_\o - U_v - U_n - U_s), \quad p \in R \\ U_p &= g_p, \quad p \in \partial R. \end{aligned}$$

Avbruddsfeilen i node p er gitt som $\tau_p = L_h u_p + f_p$. Via Taylor kan vi vise at

$$|\tau_p| = \frac{1}{6} h^2 K = \bar{\tau}$$

der

$$K = \max_{p \in R} \{|\partial_x^4 u_p|, |\partial_y^4 u_p|\},$$

denne definisjonen krever at disse fjerde-deriverte er begrenset overalt i R. Diskretiseringsfeil (global feil) er definert som

$$e_p = u_p - U_p$$

og vi finner at

$$\begin{aligned} L_h e_p = L_h u_p - L_h U_p = f_p + \tau_p - f_p = \tau_p, &\quad p \in R \\ L_h e_p = 0, &\quad p \in \partial R \end{aligned}$$

Så derfor er

$$|L_h e_p| \leq \bar{\tau} \qquad \text{for alle } p \in R.$$

Vi innfører nå funksjonen $\varphi(x,y) = \frac{1}{2}x^2$ og vi anvender operatoren L_h på denne funksjonen

$$L_h \varphi_p = \frac{1}{h^2}\left(4\,\frac{1}{2}x_p^2 - \frac{1}{2}x_\o^2 - \frac{1}{2}x_v^2 - \frac{1}{2}x_s^2 - \frac{1}{2}x_n^2\right)$$

Her er

$$\left.\begin{array}{l} x_o = x_p + h \\ x_n = x_s = x_p \\ x_v = x_p - h \end{array}\right\} \Rightarrow L_h\varphi_p = \frac{1}{2h^2}\Big(4x_p^2 - (x_p+h)^2 - (x_p - h^2) - x_p^2 - x_p^2\Big) = -1.$$

Vi setter nå $V_p = e_p + \bar{\tau}\varphi_p$ og får

$$L_h V_p = L_h e_p + \bar{\tau} L_h \varphi_p = L_h e_p - \bar{\tau} \leq 0$$

Så L_h oppfyller betingelsen i maksimumsprinsippet med $V_p = e_p + \bar{\tau}\varphi_p$. Derfor er

$$e_p + \bar{\tau}\varphi_p \leq \max_{S\in\partial R}(e_S + \bar{\tau}\varphi_S) \leq \frac{1}{2}\bar{\tau} \quad \text{for alle } p \in R,$$

siden $e_S = 0$ og siden R er kvadratet $(0,1)\times(0,1)$ så er $\varphi(x,y) = \frac{1}{2}x^2 \leq \frac{1}{2}$ for $(x,y) \in R$.

Om vi gjentar argumentet med $V_p = -e_p + \bar{\tau}\varphi_p$ finner vi tilsvarende at

$$-e_p + \bar{\tau}\varphi_p \leq \frac{1}{2}\bar{\tau} \quad \text{for alle } p \in R.$$

Siden $\varphi_p \bar{\tau} \geq 0$ kan vi konkludere at

$$|e_p| \leq \frac{1}{2}\bar{\tau} \leq \frac{1}{12}Kh^2 \quad \text{for alle } p \in R.$$

5.10.2 Noen generelle kommentarer om konvergens

En ser generelt på differensskjemaer av typen

$$\alpha_{pp}u_p - \sum_q \alpha_{pq}u_q = \beta_p + \tau_p$$

Diskretiseringsfeilen er $e_p = u_p - U_p$, og vi finner ved innsetting i formelen at

$$\alpha_{pp}e_p - \sum_q \alpha_{pq}e_q = \tau_p, \quad p \in \Omega$$

Om vi setter opp e_p, τ_p, $p \in R$ i vektorer $\boldsymbol{e}$ og $\boldsymbol{\tau}$, kan vi skrive om systemet på formen

$$A\boldsymbol{e} = \boldsymbol{\tau}$$

Om A er inverterbar fås

$$\boldsymbol{e} = A^{-1}\boldsymbol{\tau}$$

som impliserer

$$\|\boldsymbol{e}\|_\infty \leq \|A^{-1}\|_\infty \cdot \|\boldsymbol{\tau}\|_\infty$$

Stabilitet: Skjemaet sies å være *stabilt* dersom det fins en konstant C slik at

$$\|A^{-1}\| \leq C, \qquad \text{for alle skrittlengder } h.$$

Avbruddsfeilen $\boldsymbol{\tau}$ vil man typisk kunne vise (via Taylor) at oppfyller

$$\|\boldsymbol{\tau}\|_\infty = \mathcal{O}(h^\sigma), \quad \sigma \text{ heltall.}$$

Noe som ved stabilitet vil implisere at

$$\|\boldsymbol{e}\|_\infty = \mathcal{O}(h^\sigma)$$

Det kan tenkes at for eksempel $\tau_p\mathcal{O}(h^2)$ for noen p, mens for andre (typisk nær eller på randen) har en $\tau_p\mathcal{O}(h)$. Da ser vi generelt at den globale feilen $\|\boldsymbol{e}\|_\infty$ ikke kan forventes å være mer enn $\mathcal{O}(h)$. Det hender likevel at en under slike omstendigheter får $\|\boldsymbol{e}\|_\infty = \mathcal{O}(h^2)$.

5.11 Noen kommentarer om løsningsmetoder for de lineære ligningssystemene

Det å løse lineære ligninger av typen (5.4) er et stort fagfelt i seg selv. Det er ofte denne prosessen som er bestemmende for hvor store og kompliserte tilfeller som kan løses på datamaskin. I dette kurset har vi ikke nok tid til å diskutere slike metoder i detalj.

Når man skal velge metode for å løse et stort lineært ligningssystem har man to hovedklasser av metoder å velge mellom, nemlig *direktemetoder* og *iterative* metoder. Førstnevnte metode inkluderer Gausseliminasjon, eller mer spesifikt, Choleskyfaktorisering dersom ligningssystemet er symmetrisk og positiv definitt. De iterative metodene inkluderer Jacobi, Gauss–Seidel, og SOR (suksessiv overrelaksasjon). Men den typen lineære iterative ligningsløsere som har hatt mest suksess i de seinere år, er de såkalte *Krylovmetodene*. For symmetriske matriser inkluderer disse *konjugerte gradienters metode*. Det er ikke lett å gi noe enkelt svar på akkurat når det er raskest med direktemetoder, eller iterative metoder. Fordel iterative metoder har man typisk når

1. Ligningssystemene er meget store og glisne. Et system er glissent hvis det er kun en liten andel av elementene i matrisa som er ulik null.

2. Matrisene har ingen utpreget båndstruktur, det vil si at det fins indekser forholdsvis langt unna diagonalen (ij-element med stor verdi av $|i-j|$) hvis elementer er forskjellig fra null. Båndbredden til en matrise kan for eksempel defineres som

$$b(A) = \max\{|i-j| : a_{ij} \neq 0\}.$$

3. Det fins en god prekondisjonering av systemet. Teoretisk betyr dette at man kan finne matriser T, S, slik at systemet

$$\begin{array}{cccc} T^{-1}AS & S^{-1}x & = & T^{-1}b \\ \widehat{A} & \hat{x} & = & \hat{b} \end{array}$$

er "enklere" å løse enn det opprinnelige systemet $Ax = b$.

I praksis kan det være slik at med to romdimensjoner for differensialligningen, så er direktemetoder og iterative løsere omtrent like effektive, mens i tre romdimensjoner "vinner" de iterative løserne. Typisk er det slik at iterative løsere bruker multiplikasjon av matrisen A (resp $\widehat{A}$) med vilkårlige vektorer som byggesteiner i metoden. Jamfør punkt 1 ovenfor så kan man gjøre multiplikasjon Ax nokså billig hvis matrisen A er lagret på en fornuftig måte. Punkt 2 er viktig fordi dersom båndbredden til A er liten så blir Gausseliminasjon relativt sett mye billigere. Faktorene L og U i LU-faktorisering vil ha samme båndbredde

som A (dersom en ikke pivoterer). Ved stor båndbredde for en dermed også at antall ikke-null-elementer kan bli mye større i L og U enn i A. Dette fenomenet kalles "fill-in".

Det som virkelig gjør den store forskjellen for iterative metoder er punkt 3, nemlig at man kan finne en god prekondisjonering. Merk at transformasjonen $\widehat{A} = T^{-1}AS$ er kun teoretisk. For enkelhets skyld, la oss sette $S = I$. Som nevnt ovenfor bygger de iterative metodene på opersjoner av typen $y = Ax$ for vilkårlige vektorer x. For det prekondisjonerte systemet blir dette til $y = \widehat{A}x = T^{-1}Ax$. Prekondisjoneringen foregår dermed som en "indre løkke" i den iterative metoden. Hver gang en slik metode skal beregne $\widehat{A}x$ foregår dette ved at man først finner $\tilde{y} = Ax$ og deretter $y = T^{-1}\tilde{y}$. Den siste operasjonen foregår ikke som en eksplisitt matrise-vektor multiplikasjon, men er resultatet av en prosess man utfører på vektoren $\tilde{y}$. Et eksempel på en slik prosess, kan være at man utfører en deler av en Gausseliminasjon på systemet $Ay = \tilde{y}$. En slik delvis eller inkomplett Gausseliminasjon kan lages på en slik måte at beregningsarbeidet kan spres utover mange prosessorer på en datamaskin, og kan derfor utføres svært effektivt. En annen prekondisjonering går ut på å splitte opp domenet der diffligningen løses i mange små delområder. Dermed splittes differensligningene opp i mange mindre systemer av ligninger, hvis man ser bort fra koblingen i mellom områdene. Ligningssystemet for hvert område løses da, for eksempel med Gausseliminasjon på hver sin prosessor.

Det finnes et eget kurs *TMA4205 Numerisk lineær algebra* som behandler slike metoder i detalj.

Kapittel 6

En introduksjon til endelige elementmetoder

6.1 Introduksjon

I dette kapitlet skal vi gi en kortfattet introduksjon til endelige elementmetoder. Vi vil bruke Poisson's ligning som prototype, og for det meste bruke Dirichlet randkrav. . Mange bøker begynner med å se på én romdimensjon, men vi skal her ta steget direkte opp til 2 romdimensjoner. Vi ser altså på ligningen

$$-\Delta u = f \qquad \text{i} \quad \Omega, \qquad u = 0 \quad \text{på} \quad \partial\Omega \tag{6.1}$$

der Ω er et åpent begrenset område i $\mathbf{R}^2$ med en stykkevis glatt rand. En kan også tenke seg dette problemet i en romdimensjon

$$-u''(x) = f(x), \quad 0 < x < a, \qquad u(0) = u(a) = 0. \tag{6.2}$$

En viktig prosedyre i utledningen av elementmetoder er å multiplisere differensialligningen med en (nokså) vilkårlig funksjon[1] og integrere over Ω og anvende en form for delvis integrasjon. For eksempel, hvis v er deriverbar, kan vi med $u(x)$ fra (6.2) utlede med bruk av delvis integrasjon

$$\int_0^a u'' \cdot v \, \mathrm{d}x = u'(a)v(a) - u'(0)v(0) - \int_0^a u'v' \, \mathrm{d}x$$

Tilsvarende kan vi gjøre med Δu i (6.1) ved å benytte oss av divergensteoremet . Vi minner om at dersom $\vec{X}$ er et vektorfelt med kontinuerlige partiellderiverte av hver komponent, Ω er et område som ovenfor med en utadrettet normalvektor $\vec{n}$ så gjelder at

$$\oint_{\partial\Omega} \vec{X} \cdot \vec{n} \, ds = \int_\Omega \mathrm{div}\vec{X} \, \mathrm{d}A$$

Vi minner om at hvis $\vec{X}(x,y) = \Big(X_1(x,y), X_2(x,y)\Big)$ så er

$$\mathrm{div}\vec{X} = \frac{\partial X_1}{\partial x} + \frac{\partial X_2}{\partial y}$$

[1] Vi skal straks se at denne funksjonen velges fra et velspesifisert funksjonsrom

La oss benytte divergensteoremet på vektorfeltet $\vec{X} = v\nabla u = (vu_x, vu_y)$. Da blir

$$\operatorname{div}\vec{X} = \frac{\partial}{\partial x}(vu_x) + \frac{\partial}{\partial y}(vu_y) = \nabla u \cdot \nabla v + v\Delta u$$

Bruker vi dette i divergensteoremet så får vi

Greens identitet.

$$\int_\Omega v\Delta u \,\mathrm{d}A = \int_{\partial\Omega} v\nabla u \cdot \vec{n} \,\mathrm{d}s - \int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}A$$

6.2 Tre ekvivalente problemer

6.2.1 Noen definisjoner

Vi introduserer et reelt indreprodukt for funksjoner definert på Ω

$$\langle u, v\rangle = \int_\Omega u(x,y)v(x,y)\,\mathrm{d}x\,\mathrm{d}y$$

i det følgende skriver vi gjerne dA for $dx\,dy$. Indreproduktet er symmetrisk. Man forutsetter naturligvis at integralet eksisterer, og seinere skal vi stort sett jobbe med funksjoner som er kontinuerlige. Det er ganske enkelt å sjekke at denne definisjonen oppfyller de vanlige kravene for et indreprodukt. Merk spesielt at det er bilineært dvs for funksjoner u, v og skalarer α, β gjelder

$$\langle \alpha\, u + \beta\, v, w\rangle = \alpha\langle u, w\rangle + \beta\langle v, w\rangle$$

En annen bilineær form er definert av

$$a(u,v) = \int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}A$$

Merk at her er ∇u og ∇v gradientvektorfelt og prikken er indreprodukt, dvs integranden er $u_x v_x + u_y v_y$.

Vi må nå introdusere et funksjonsrom, for å unngå vanskelige definisjoner skal vi løse litt opp på rigorøsiteten og definere det reelle vektorrommet S bestående av funksjoner $u(x,y)$ slik at

1. u er kontinuerlig på $\bar{\Omega}$
2. u_x, u_y stykkevis kontinuerlige i Ω
3. $u = 0$ på randen $\partial\Omega$.

Naturligvis er det slik at hvis $u, v \in S$ og $\alpha, \beta \in \mathbf{R}$ så vil $\alpha u + \beta v \in S$.

6.2.2 Ekvivalente problemer

Følgende tre problemer er ekvivalente

(D) $-\Delta u = f$ i Ω, $u = 0$ på $\partial\Omega$

(V) Finn $u \in S$ slik at $a(u,v) = \langle f, v\rangle$ for alle $v \in S$

(M) Definer $P(v) = \frac{1}{2}a(v,v) - \langle f, v\rangle$. Finn $u \in S$ slik at

$$P(u) = \min_{v\in S} P(v)$$

Merknad 1. Funksjonen v i problemet **(V)** kalles gjerne en *testfunksjon* .

Merknad 2. Merk at en kun kan vise at **(V)** og **(M)** impliserer **(D)** under antagelse om glatthet av løsningen for i utgangspunktet krever jo Δu at u er to ganger deriverbar, mens funksjonene i S kun er en gang stykkevis kontinuerlig deriverbar. Men det skal vise seg å være veldig nyttig å kunne anta mindre grad av deriverbarhet for funksjoner i S når vi seinere kommer til endelige elementmetoder.

Vi beviser ekvivalensene.

Trinn 1. La oss begynne med å vise at "u løser **(D)**" impliserer "u løser **(V)**". Vi har for $v \in S$

$$\langle f, v\rangle = \int_\Omega fv\,\mathrm{d}A = -\int_\Omega \Delta u \cdot v\,\mathrm{d}A = -\int_{\partial\Omega} v\,(\nabla u \cdot \vec{n})\,\mathrm{d}s + \int_\Omega \nabla u \cdot \nabla v\,\mathrm{d}A$$

Men siden $v \in S$ så er $v(x,y) = 0$ for alle $(x,y) \in \partial\Omega$ så randintegralet forsvinner. Vi står dermed igjen med at $a(u,v) = \langle f, v\rangle$ for alle $v \in S$.

Trinn 2. Nå viser vi at "u løser **(V)**" $\Leftrightarrow$ "u løser **(M)**". La $u, w \in S$ og $\lambda \in \mathbf{R}$. Vi beregner

$$\begin{aligned} P(u+\lambda w) &= \frac{1}{2}a(u+\lambda w, u+\lambda w) - \langle f, u+\lambda w\rangle \\ &= \frac{1}{2}a(u,u) - \langle f, u\rangle + \lambda\big(a(u,w) - \langle f, w\rangle\big) + \frac{1}{2}\lambda^2 a(w,w) \end{aligned}$$

Vi har derfor

$$P(u+\lambda w) = P(u) + \lambda\big(a(u,w) - \langle f, w\rangle\big) + \frac{1}{2}\lambda^2 a(w,w)$$

Nå antar vi at $a(u,v) = \langle f, v\rangle$ for alle $v \in S$ dvs at u løser **(V)**. La $v \in S$ være villårlig og sett $w = v - u \in S$. Da er $v = u + \lambda w$ med $\lambda = 1$. Vi får

$$P(v) = P(u) + \frac{1}{2}\lambda^2 a(w,w)$$

Men

$$a(w,w) = \int_\Omega |\nabla w|^2\,\mathrm{d}A \geq 0 \qquad \text{for enhver} \quad w$$

så vi har vist at u faktisk løser **(M)**. Anta istedet at u *ikke* oppfyller **(V)**. Da fins $w_1 \in S$, $w_1 \neq 0$ slik at $\mu_1 = a(u,w_1) - \langle f, w_1\rangle \neq 0$. Merk at $a(w_1,w_1) > 0$ når $0 \neq w_1 \in S$. Sett

$$\lambda = \lambda_1 = -\frac{\mu_1}{a(w_1,w_1)}$$

Vi finner at

$$P(u+\lambda w_1) = P(u) - \frac{\mu_1}{a(w_1,w_1)}\mu_1 + \frac{1}{2}\frac{\mu_1^2}{a(w_1,w_1)^2}a(w_1,w_1) = P(u) - \frac{1}{2}\frac{\mu_1^2}{a(w_1,w_1)} < P(u)$$

så vi har motbevist at u er et minimum for $P(v)$. Oppsummert: vi har dermed vist at "u løser **(M)**" impliserer "u løser **(V)**"

Trinn 3. Vi vil vise at "u løser **(V)**" impliserer "u løser **(D)**" Dette trinnet forutsetter at u som brukes i **(V)** ikke bare tilhører S, men også at Δu er veldefinert i henhold til Greens identitet. Vi starter altså med

$$a(u,v) = \int_\Omega \nabla u \cdot \nabla v\,\mathrm{d}A = \langle f, v\rangle \qquad \text{for alle} \quad v \in S.$$

Fra Greens identitet samt at $v = 0$ på $\partial\Omega$ fås

$$\int_\Omega (-\Delta u)\, v \, \mathrm{d}A = \int_\Omega f \cdot v \, \mathrm{d}A \qquad \text{for alle} \quad v \in S,$$

eller

$$\int_\Omega (\Delta u + f) v \, \mathrm{d}A = 0 \qquad \text{for alle} \quad v \in S.$$

Om vi antar at $\phi := \Delta u + f$ er kontinuerlig i Ω så må vi ha $\phi = 0$ for alle $(x, y) \in \Omega$. For om det fantes et punkt (x_0, y_0) der $\phi \neq 0$, så sørger kontinuiteten for at $\phi \neq 0$ i en omegn U omkring (x_0, y_0). Det er mulig å finne en $v \in S$ slik at $v(x, y) = 0$ når $(x, y) \notin U$ $v(x, y) \geq 0$ for alle $(x, y) \in \Omega$ og $v(x_0, y_0) = 1$. Dermed blir $\int_\Omega \phi \cdot v \mathrm{d}A \neq 0$, en selvmotsigelse. Vi konkluderer med at "u løser **(V)**" impliserer "u løser **(D)**" når u er tilstrekkelig glatt til at **(D)** er veldefinert.

Trinn 4. Vi viser at løsningen av **(V)** er entydig. Anta at det istedet fins 2 løsninger $u_1 \in S$ og $u_2 \in S$ av **(V)**. Da gjelder altså

$$\begin{aligned} a(u_1, v) &= \langle f, v \rangle \quad \text{for all} \quad v \in S \\ a(u_2, v) &= \langle f, v \rangle \quad \text{for all} \quad v \in S \end{aligned}$$

Vi trekker nå de to ligningene fra hverandre og bruker at a er bilineær

$$a(u_1 - u_2, v) = 0 \quad \text{for all} \quad v \in S$$

Spesielt kan vi sette $v = u_1 - u_2 \in S$ og da får vi med $w = u_1 - u_2$ at

$$a(w, w) = 0 \qquad \Rightarrow \qquad w = u_1 - u_2 \equiv 0$$

så vi konkluderer med at løsningen er entydig. □

Vi kan nå koste på oss å ta en kikk på den éndimensjonale varianten av dette teoremet. Vi har nå enklere definisjoner av indreproduktet og av $a(u, v)$. La $\Omega = (0, 1)$, da blir

$$\langle u, v \rangle = \int_0^1 u(x)\, v(x) \, \mathrm{d}x, \tag{6.3}$$

og

$$a(u, v) = \int_0^1 u'(x)\, v'(x) \, \mathrm{d}x \tag{6.4}$$

Vi formulerer de tre ekvivalente problemer som

(D) $-u'' = f$ i $(0, 1)$, $u(0) = u(1) = 0$

(V) Finn $u \in S$ slik at $a(u, v) = \langle f, v \rangle$ for alle $v \in S$

(M) La $P(v) = \frac{1}{2} \int_0^1 (v')^2 \mathrm{d}x - \int_0^1 v\, f \, \mathrm{d}x$. Da er $P(u) = \min_{v \in S} P(v)$

6.3 Tilnærmet løsning av variasjonsproblemet

6.3.1 Generell framgangsmåte

Velg lineært uavhengige funksjoner φ_j, $j = 1 : n$ som alle tilhører S. Sett

$$S_h = \mathrm{span}\{\varphi_1, \ldots, \varphi_n\}$$

Vi har da $S_h \subset S$. Erstatt nå **(V)** med det enklere problemet

($\mathbf{V}_h$) Finn $U \in S_h$ slik at

$$a(U, V) = \langle f, V \rangle \qquad \text{for alle} \quad V \in S_h$$

Betingelsen er ekvivalent med

($\mathbf{V}'_h$) Finn $U \in S_h$ slik at

$$a(U, \varphi_k) = \langle f, \phi_k \rangle \qquad \text{for } k = 1, \ldots, n$$

Denne ekvivalensen er ikke så vanskelig å forklare. For siden $V \in S_h$ kan den skrives som $V = \sum_{k=1}^n v_k \varphi_k$ der v_k er reelle konstanter. Dermed får vi

$$a(U, V) = \langle f, V \rangle \qquad \Leftrightarrow \qquad \sum_{k=1}^n v_k a(U, \varphi_k) = \sum_{k=1}^n v_k \langle f, \phi_k \rangle$$

dette forklarer at **($\mathbf{V}'_h$)** er ekvivalent med **($\mathbf{V}_h$)**.

Men også $U \in S_h$ og kan skrives som $U = \sum_{k=1}^n u_k \varphi_k$ og fra **($\mathbf{V}'_h$)** får vi dermed

$$\sum_{j=1}^n u_j\, a(\varphi_j, \varphi_k) = \langle f, \varphi_k \rangle, \qquad k = 1, \ldots, n.$$

Dette er et system av n ligninger for $\vec{u} = (u_1, \ldots, u_n)^T$ og kan skrives

$$A\vec{u} = \vec{b}$$

der A er $n \times n$-matrisen med ij-element $a(\varphi_i, \varphi_j)$ og $\vec{b} \in \mathbf{R}^n$ er vektoren med komponenter $b_k = \langle f, \varphi_k \rangle$.

Matrisa A har følgende egenskaper

1. A er symmetrisk
2. A er positiv definitt (gitt at $\varphi_1, \ldots, \varphi_n$ er lineært uavhengige)

Symmetrien følger fra at den bilineære formen $a(u, v)$ er symmetrisk. At A er positiv definitt vises som følger: La $0 \neq \vec{c} \in \mathbf{R}^n$ være vilkårig. Vi beregner

$$\vec{c}^T A \vec{c} = \sum_{i,j} c_i a(\varphi_i, \varphi_j) c_j = a(\sum_i c_i \varphi_i, \sum_j c_j \varphi_j) = a(w, w)$$

der $w = \sum_i c_i \varphi_i$. Siden $\{\varphi_i\}$ er lineært uavhengige og $\vec{c} \neq 0$ så er $w \neq 0$ og dermed er $\vec{c}^T A \vec{c} > 0$ så A er positiv definitt.

6.3.2 En viktig egenskap ved løsning av $(\mathbf{V}_h)$

La u og U være løsning av henholdsvis $(\mathbf{V})$ og $(\mathbf{V}_h)$. Da gjelder

$$a(u-U,u-U) \;\le\; a(u-V,u-V) \quad \text{for alle} \quad V \in S_h \tag{6.5}$$

Dersom vi bruker $a(u-V,u-V)$ som et mål på feilen i V som tilnærmelse til u, så er altså U den beste tilnærmelsen til u vi kan finne i rommet S_h.

Bevis. Vi har $a(u,w) = \langle f, w\rangle$ for alle $w \in S \supset S_h$, og $a(U,w) = \langle f, w\rangle$ for alle $w \in S_h$. Trekker vi disse ligningene fra hverandre får vi

$$a(u-U,w) = 0 \qquad \text{for alle } w \in S_h \tag{6.6}$$

La nå $V \in S_h$ være vilkårlig og sett $w = V - U \in S_h$. Da er

$$a(u-V,u-V) = a(u-U-w,u-U-w) = a(u-U,u-U) - 2a(u-U,w) + a(w,w)$$

På grunn av (6.6) og fordi $a(w,w) \ge 0$ får vi at $a(u-U,u-U) \le a(u-V,u-V)$ for alle $V \in S_h$. □

6.3.3 Minimaliseringsproblemet og navnekonvensjoner

Vi kan naturligvis lage en approksimativ versjon av $(\mathbf{M})$ også ved å la

$(\mathbf{M}_h)$ Finn $U \in S_h$ slik at

$$P(U) = \min_{V \in S_h} P(V)$$

Når vi bruker

$$(\mathbf{D}) \quad \longrightarrow \quad (\mathbf{V}) \quad \longrightarrow \quad (\mathbf{V}_h)$$

kaller vi det *Galerkins metode*

Når vi istedet gjør

$$(\mathbf{D}) \quad \longrightarrow \quad (\mathbf{M}) \quad \longrightarrow \quad (\mathbf{M}_h)$$

kalles det *Rayleigh-Ritz metode*.

Rayleigh-Ritz tankegangen kan bare brukes når problemet $(\mathbf{D})$ er selvadjungert (symmetrisk + en teknisk tilleggsbetingelse som oftest er oppfylt for problemer fra anvendelser). I så fall gir Galerkin og Rayleigh-Ritz det samme ligningssystemet $Ax = b$, så metoden blir gjerne omtalt som Rayleigh-Ritz-Galerkin (RRG). Galerkin kan også brukes på problemer som *ikke* er selvadjungerte.

6.3.4 Endimensjonalt problem

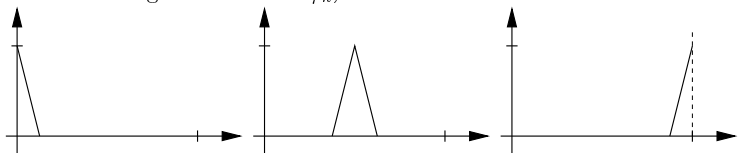
La oss gå tilbake til det endimensjonale problemet med uttrykkene (6.3) og (6.4) for indreprodukt og bilineær form.

Klassisk RRG – $\varphi_j(x) = x^j(1-x),\ j = 1,\dots,n$. I dette tilfelle blir

$$\langle \varphi_i, \varphi_{k-i}\rangle = \frac{2}{(k+1)(k+2)(k+3)}, \quad 1 \le i < k.$$

Du kan jo selv sjekke at disse funksjonene er i S.

Endelig elementmetode Vi skal her velge en basis av pyramidefunksjoner eller hattfunksjoner . Del intervallet $[0,1]$ inn i $N = n+1$ intervaller av lengde $h = 1/N$ og sett $x_k = k\,h$ for heltallige k. Vi definerer $\varphi_k,\ k = 0:N$ ved

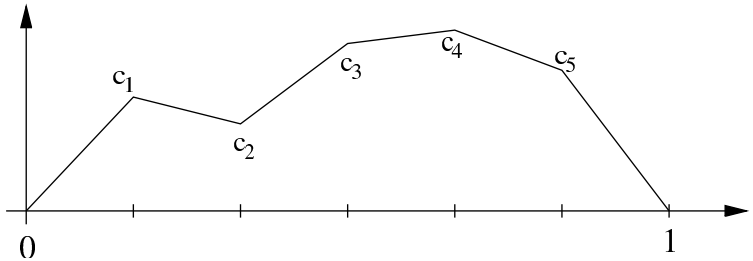


$$\varphi_k(x) = \begin{cases} 1 - \frac{1}{h}|x - x_k|, & x \in [0,1] : x_{k-1} < x < x_{k+1} \\ 0, & x \in [0,1] : x < x_{k-1} \text{ eller } x > x_{k+1} \end{cases}$$

Merk at φ_0 kun er den høyre halvdelen av hatten, mens φ_N kun er den venstre halvdelen, som på figuren. Disse to endefunksjonene er imidlertid ikke med i S siden de ikke er 0 på randen av området. Derimot er $\varphi_k \in S,\ s = 1,\dots,n$, siden de er null på randen og har stykkevis kontinuerlig derivert. Lineærkombinasjoner av φ_k tilhører naturligvis også S

$$U(x) = \sum_{k=1}^{n} c_k \varphi_k(x) \;\in\; S$$

Spesielt merker vi oss at $\varphi_k(c_j) = \delta_{kj}$, der vi bruker kronecker-delta som er 0 for $k \ne j$ og 1 for $k = j$. Dette fører til at $U(x_k) = c_k$.



Vi kan nå beregne koeffisientmatrisen A hvis ij-element er $a(\varphi_i,\varphi_j)$. Dette blir spesielt enkelt fordi φ_k og dens deriverte kun er ulik 0 på de to delintervallene $[x_{k-1},x_k]$ og $[x_k,x_{k+1}]$. Dermed blir produktet $\varphi_i' \cdot \varphi_j'$ identisk null hvis $|i-j| > 1$. Samlet får vi

$$a(\varphi_i,\varphi_j) = \int_0^1 \varphi_i' \cdot \varphi_j' \,\mathrm{d}x = \begin{cases} 0 & \text{hvis } |i-j| > 1 \\ \int_0^h \frac{1}{h}(-\frac{1}{h})\,\mathrm{d}x = -\frac{1}{h} & \text{hvis } |i-j| = 1 \\ \int_0^{2h} \frac{1}{h^2}\,\mathrm{d}x = \frac{2}{h} & \text{hvis } i = j \end{cases}$$

Høyresiden $\vec{b}$ av ligningssystemet har elementer

$$b_j = \int_0^1 f \cdot \varphi_j \,\mathrm{d}x = \int_{x_{j-1}}^{x_{j+1}} f \cdot \varphi_j \,\mathrm{d}x$$

Siden $U_k := U(x_k) = c_k$ kan vi nå lage en differensligning for U ved å skrive

$$\frac{1}{h}(-U_{j-1} + 2U_j - U_{j+1}) = b_j, \quad U_0 = U_N = 0.$$

Dette er ikke helt det samme som differensligningen vi ville fått ved å erstatte $-u''$ med sentraldifferenser ($-u''(x_m) \longrightarrow \frac{1}{h^2}\delta^2 U_m$) fordi høyresiden er et integral som vi ikke uten videre kan beregne eksakt. Men la oss si at vi approksimerer integralet ved å sette

$$b_j \approx f(x_j)\int_{x_{j-1}}^{x_{j+1}} \varphi_j(x)\,\mathrm{d}x = hf(x_j)$$

da er vi tilbake til det nevnte differenseskjemaet.

6.3.5 Todimensjonalt problem – Endelig elementmetode med trianguläre elementer og lineære elementfunksjoner

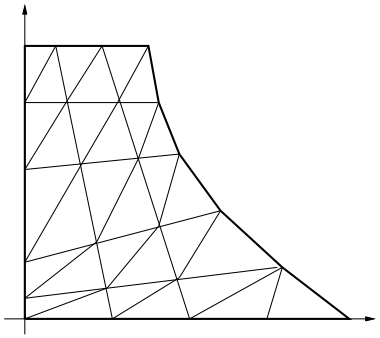
Vi antar at Ω kan settes sammen av trekanter. Om nødvendig må vi da deformere $\partial\Omega$ litt. En trekant kalles et element, og dens hjørner kalles *noder* eller *knutepunkter*.

Notasjon:

T_h: mengden av alle elementer K. Dermed har vi $\Omega = \bigcup_{K\in T_h} K$

diam(K): Lengden av lengste sidekant i K.

h: $\max_{K\in T_h} \mathrm{diam}(K)$



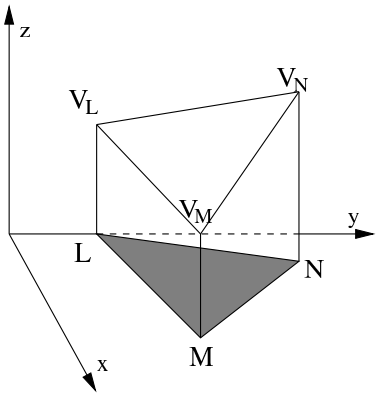
Vi kaller som før funksjonsrommet for tilnærmet løsning for S_h, som vi nå lar bestå av stykkevis lineære funksjoner av x og y. Når vi snakker om lineære funksjoner her mener vi funksjoner av typen $f(x,y) = a + bx + cy$, som strengt tatt ikke er lineær fordi $f(x_1+\lambda x_2, y_1+\lambda y_2) \neq f(x_1,y_1) + \lambda f(x_2,y_2)$. En riktigere betegnelse kunne vært *affine funksjoner* eller *bivariate polynomer av grad 1*. Men vi skal likevel bruke begrepet lineære elementfunksjoner i fortsettelsen. La v være en funksjon som er definert på Ω og la K være et element i Ω. Restriksjonen av v til K, betegnet $v|_K$, er den funksjonen som fremkommer når vi innskrenker definisjonsmengden for v til K. Vi definerer nå

$$S_h = \{v : v \in C(\Omega),\ v = 0 \text{ på } \partial\Omega, \quad v|_K \text{ lineær for enhver } K \in T_h\}$$

Notasjonen $C(\Omega)$ betyr som før "kontinuerlig i Ω". Når vi sier $v|_K$ lineær, betyr dette at $v(x,y)$ har formen

$$v(x,y) = a_K + b_K\,x + c_K\,y \quad \text{for } (x,y) \in K.$$

der a_K, b_K, c_K er konstanter.



Det trengs 3 parametre for å bestemme $v|_K$, og vi kan benytte de tre funksjonsverdiene i hjørnene L, M, N på figuren. Hvis L, M, N har koordinater hhv (x_L, y_L), (x_M, y_M),(x_N, y_N), så kan vi skrive opp de tre ligningene

$$a_K + b_K x_P + c_K y_P = v_P, \quad P = L, M, N$$

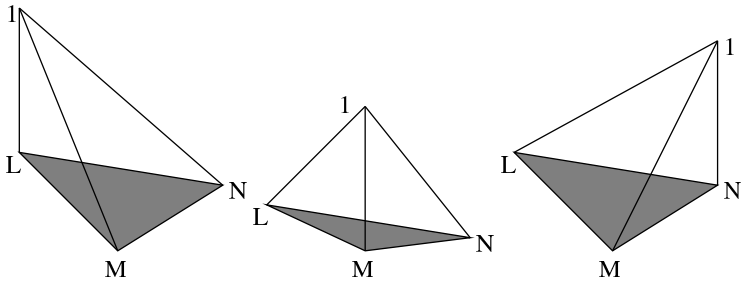
Dette er altså tre ligninger for a_K, b_K, c_K.

Formfunksjoner. Vi kan lage tre lineære funksjoner ψ_L, ψ_M, ψ_N til et gitt element K der

$$\psi_L(x_L, y_L) = 1, \quad \psi_L(x_M, y_M) = 0, \quad \psi_L(x_N, y_N) = 0.$$

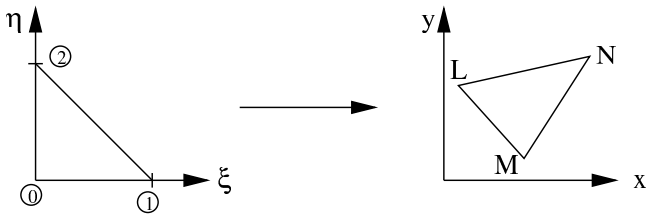
Tilsvarende krav gjelder for ψ_M og ψ_N, en får tre "pyramider" med toppunktet vertikalt over hhv L, M og N som på figuren. Vi kan dermed skrive

$$v|_K = v_L\psi_L + v_M\psi_M + v_N\psi_N$$



6.3.6 Konstruksjon av formfunksjonene

Fordi disse trekantelementene kan befinne seg på vilkårlige posisjoner i xy-planet, kan det være en fornuftig strategi å danne formfunksjonen på et referanseelement, og så transformere definisjonsområdet for hver enkelt trekant. Vi lar referanseelementet ha hjørner $(0,0),\ (1,0),\ (0,1)$ og illusterer med en figur



$$\begin{array}{rcll} \chi_0(\xi,\eta) & = & 1-(\xi+\eta) & \longrightarrow \quad \psi_L(x,y) \\ \chi_1(\xi,\eta) & = & \xi & \longrightarrow \quad \psi_M(x,y) \\ \chi_2(\xi,\eta) & = & \eta & \longrightarrow \quad \psi_N(x,y) \end{array}$$

La nå hjørnene L, M, N ha koordinater (x_L, y_L), (x_M, y_M) og (x_N, y_N). Vi transformerer punkter i $\xi\eta$-planet til punkter i xy-planet gjennom

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x_L \\ y_L \end{pmatrix} + T_K \cdot \begin{pmatrix} \xi \\ \eta \end{pmatrix} \quad T_K = \begin{pmatrix} x_M - x_L & x_N - x_L \\ y_M - y_L & y_N - y_L \end{pmatrix}$$

Matrisen T_K er inverterbar hvis og bare hvis ikke alle punktene L, M, N ligger på samme rette linje. Vi skriver

$$\begin{pmatrix} \xi \\ \eta \end{pmatrix} = T_K^{-1} \begin{pmatrix} x - x_L \\ y - y_L \end{pmatrix} := \begin{pmatrix} r_1(x,y) \\ r_2(x,y) \end{pmatrix}$$

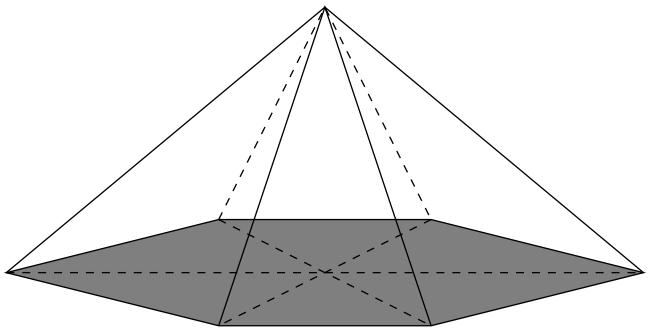
Formfunksjonene kan nå skrives

$$\begin{aligned} \psi_L(x,y) &= \chi_0(r_1(x,y), r_2(x,y)) \\ \psi_M(x,y) &= \chi_1(r_1(x,y), r_2(x,y)) \\ \psi_N(x,y) &= \chi_2(r_1(x,y), r_2(x,y)) \end{aligned}$$

6.3.7 Basisfunksjoner for S_h, pyramidefunksjoner

For alle noder P i trekantnettet definerer vi en basisfunksjon $\varphi_P(x,y)$. φ_P skal være kontinuerlig for alle $(x,y) \in \Omega$. For alle elementer $K \in T_h$ krever vi at $\varphi_P|_K$ er lineær og dessuten

$$\varphi_P(x_Q, y_Q) = \begin{cases} 1 & \text{hvis } Q = P \\ 0 & \text{hvis } Q \neq P \end{cases}$$



P kan gjerne tilhøre $\partial\Omega$, i såfall blir litt av pyramiden borte. Definer

$$\mathcal{Z} = \{P : P \text{ node}, \ P \in \Omega\}$$

det vil si $\mathcal{Z}$ er megden av indre noder. Hvis $P \in \mathcal{Z}$ så er $\varphi_P \in S$.

Mengden av funksjoner $\{\varphi_P : P \in \mathcal{Z}\}$ danner en basis for det vi tar som rommet S_h, en har altså

1. Enhver $v \in S_h$ kan skrives $v = \sum_{p \in \mathcal{Z}} v_P \varphi_P$
2. $\{\varphi_P\}_{P \in \mathcal{Z}}$ er lineært uavhengige

6.3.8 Elementmetodeløsning av (D)

Vi nummererer nodene i $\mathcal{Z}$ fra 1 til n. La φ_P være pyramidefunksjonen knyttet til node P. Sett

$$U(x,y) = \sum_{q=1}^{n} U_q \, \varphi_q(x,y)$$

for konstanter $U_1, \ldots, U_n$. Bestem nå $U_1, \ldots, U_n$ ved at

$$\begin{array}{c} a(U,V) = \langle f, v \rangle \quad \text{for alle } V \in S_h \\ \Updownarrow \\ a(U, \varphi_p) = \langle f, \varphi_p \rangle \quad \text{for } p = 1, \ldots, n \\ \Updownarrow \\ \sum_{q=1}^{n} a(\varphi_p, \varphi_q) U_q = \langle f, \varphi_p \rangle \quad \text{for } p = 1, \ldots, n \end{array}$$

Skriver vi $\vec{U} = (U_1, \ldots, U_n)^T$, $\vec{b} = (b_1, \ldots, b_n)^T$ der $b_p = \langle f, \varphi_p \rangle$, og lar A være den symmetriske matrisen med pq-element $a(\varphi_p, \varphi_q)$ så har vi ligningssystemet

$$A\vec{U} = \vec{b}$$

Tilsvarende et tidligere bevis for det endimensjonale tilfellet kan vi også her konkludere med at A er positiv definitt. Dessuten er A glissen (sparse) fordi $\alpha_{pq} = 0$ hvis $p \neq q$ og nodene tilhørende (p,q) ikke er nabonoder. Matrisen A kalles i litteraturen for *stivhetsmatrisen* (det har ingen ting med stive ODE'er å gjøre).

6.3.9 Beregning av stivhetsmatrisen ved innaddering

Vi lar som før mengden av alle elementer kalles for T_h slik at $\Omega = \bigcup_{K \in T_h} K$. Dermed er

$$a(u,v) = \int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}A = \sum_{K \in T_h} \int_K \nabla u \cdot \nabla v \, \mathrm{d}A$$

For enhver $K \in T_h$ definerer vi

$$a^K(u,v) = \int_K \nabla u \cdot \nabla v \, \mathrm{d}A \quad \Rightarrow \quad a(u,v) = \sum_{K \in T_h} a^K(u,v)$$

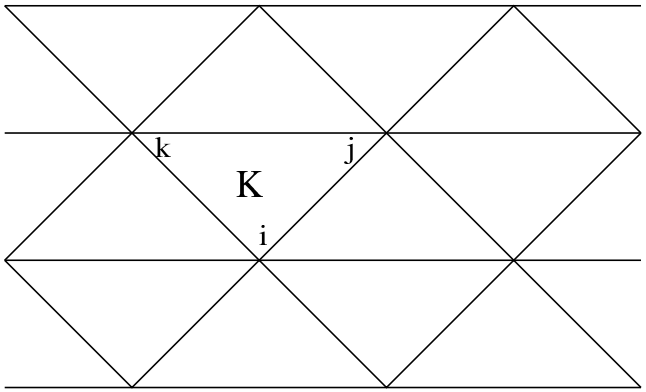
Spesielt blir naturligvis

$$a(\varphi_p, \varphi_q) = \sum_{K \in T_h} a^K(\varphi_p, \varphi_q)$$

La K være et element med hjørner i, j, k og med formfunksjoner ψ_i^K, ψ_j^K, ψ_k^K. Da har vi

$$\varphi|_K = \begin{cases} \psi_p^K & \text{hvis } p = i, j, \text{ eller } k \\ 0 & \text{for alle andre } p \end{cases}$$

Tanken er her at vi holder elementet (trekanten) K fast, og spør oss om hvilke pyramidefunksjoner som er ulik 0 på K.



Vi definerer nå en 3×3 *elementstivhetsmatrise* for K

$$A^K \in \mathbf{R}^{3\times 3} \quad \text{der} \quad A_{pq}^K = a^K(\varphi_p, \varphi_q),\ p, q \in \{i, j, k\}$$

selv om matrisen er 3×3 så bruker vi indekser i, j, k for de tre rader/kolonner. Da gjelder for hele stivhetsmatrisen at

$$A_{pq} = \sum_K A_{pq}^K \quad \text{sum over } K \text{ som har } p, q \text{ som hjørner.}$$

Helt tilsvarende definerer vi det lokale indreproduktet

$$\langle u, v \rangle^K = \int_K u\, v \, \mathrm{d}A$$

Da har vi

$$\langle f, \varphi_p \rangle = \sum_{K \in T_h} \langle f, \varphi_p \rangle^K$$

For hver K lager vi en *elementhøyreside*

$$\vec{b}^K = (b_i^K, b_j^K, b_k^K)^T, \qquad b_p^K = \langle f, \varphi_p \rangle^K = \langle f, \psi_p^K \rangle, \quad p = i, j, k$$

A og $\vec{b}$ kan nå beregnes slik

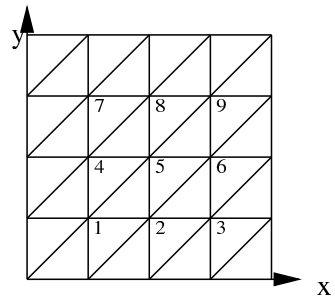
- Nullstill A, b
- La K løpe gjennom T_h og for hver K med hjørner i, j, k beregn A_{pq}^K, $p, q = i, j, k$ og $\vec{b}^K$, $p = i, j, k$.
- Sett $A_{pq} = A_{pq} + A_{pq}^K$ og $b_p = b_p + b_p^K$.

6.3.10 Et omfattende eksempel uten innaddering

$$\Omega = (0, \mu) \times (0, \nu)$$

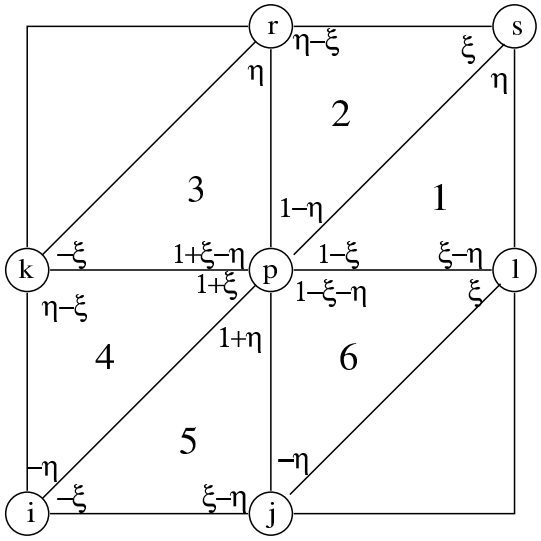
$$h = \frac{\mu}{M} = \frac{\nu}{N}$$

Antall noder: $(M-1)(N-1)$.



Vi regner koeffisienter og høyresider i $A\vec{U} = \vec{b}$, direkte fra definisjonen av A_{pq} og b_p uten bruk av innaddering. Vi vet at $A_{pq} = 0$ hvis p og q *ikke* tilhører samme element, så det holder å se på noder omkring p som vist på figuren, disse nodene er merket i, j, k, ℓ, r, s. Vi setter

$$\xi = \frac{x - x_p}{h}, \qquad \eta = \frac{y - y_p}{h}, \qquad \varepsilon = \frac{1}{h}$$



K	φ_p $\nabla\varphi_p$	φ_i $\nabla\varphi_i$	φ_k $\nabla\varphi_k$	φ_j $\nabla\varphi_j$	φ_ℓ $\nabla\varphi_\ell$	φ_r $\nabla\varphi_r$	φ_s $\nabla\varphi_s$
1	$1-\xi$ $(-\varepsilon,0)$	0	0	0	$\xi-\eta$ $(\varepsilon,-\varepsilon)$	0	η $(0,\varepsilon)$
2	$1-\eta$ $(0,-\varepsilon)$	0	0	0	0	$\eta-\xi$ $(-\varepsilon,\varepsilon)$	ξ $(\varepsilon,0)$
3	$1+\xi-\eta$ $(\varepsilon,-\varepsilon)$	0	$-\xi$ $(-\varepsilon,0)$	0	0	η $(0,\varepsilon)$	0
4	$1+\xi$ $(\varepsilon,0)$	$-\eta$ $(0,-\varepsilon)$	$\eta-\xi$ $(-\varepsilon,\varepsilon)$	0	0	0	0
5	$1+\eta$ $(0,\varepsilon)$	$-\xi$ $(-\varepsilon,0)$	0	$\xi-\eta$ $(\varepsilon,-\varepsilon)$	0	0	0
6	$1-\xi+\eta$ $(-\varepsilon,\varepsilon)$	0	0	$-\eta$ $(0,-\varepsilon)$	ξ $(\varepsilon,0)$	0	0

Vi merker oss at arealet av hver trekant er $\frac{1}{2}h^2$. Vi beregner

$$a(\varphi_p,\varphi_p)=\int_1\varepsilon^2+\int_2\varepsilon^2+\int_3 2\varepsilon^2+\int_4\varepsilon^2+\int_5\varepsilon^2+\int_6 2\varepsilon^2=\frac{1}{2}h^2\,8\varepsilon^2=4$$

Videre er

$$a(\varphi_p,\varphi_i)=\int_4(\varepsilon,0)\cdot(0,-\varepsilon)+\int_5(0,\varepsilon)\cdot(-\varepsilon,0)=0$$

$$a(\varphi_p,\varphi_j)=\int_5(0,\varepsilon)\cdot(\varepsilon,-\varepsilon)+\int_6(-\varepsilon,\varepsilon)\cdot(0,-\varepsilon)=\frac{1}{2}h^2(-\varepsilon^2-\varepsilon^2)=-1$$

$$a(\varphi_p,\varphi_k)=\int_3(\varepsilon,-\varepsilon)\cdot(-\varepsilon,0)+\int_4(\varepsilon,0)\cdot(-\varepsilon,\varepsilon)=\frac{1}{2}h^2(-\varepsilon^2-\varepsilon^2)=-1$$

$$a(\varphi_p,\varphi_\ell)=\int_1(-\varepsilon,0)\cdot(\varepsilon,-\varepsilon)+\int_6(-\varepsilon,\varepsilon)\cdot(\varepsilon,0)=\frac{1}{2}h^2(-\varepsilon^2-\varepsilon^2)=-1$$

$$a(\varphi_p,\varphi_r)=\int_2(0,-\varepsilon)\cdot(-\varepsilon,\varepsilon)+\int_3(\varepsilon,-\varepsilon)\cdot(0,\varepsilon))=\frac{1}{2}h^2(-\varepsilon^2-\varepsilon^2)=-1$$

$$a(\varphi_p,\varphi_s)=\int_1(-\varepsilon,0)\cdot(0,\varepsilon)+\int_2(0,-\varepsilon)\cdot(\varepsilon,0)=0$$

For høyresiden har vi

$$b_p=\langle f,\varphi_p\rangle=\sum_{K=1}^{6}\langle f,\varphi_p\rangle^K$$

og lengre kommer vi ikke uten å kjenne f. Men

$$b_p\approx f(x_p,y_p)\sum_{K=1}^{6}\int_K\varphi_p\mathrm{d}A=h^2f_p$$

Sjekk selv integralet av ϕ_p ovenfor ved å summere volumet av 6 tetraedere. Ligningene har altså formen

$$4U_p-U_j-U_k-U_\ell-U_r=\int f\cdot\varphi_p\mathrm{d}A$$

6.3.11 Feilen i U

Utledning av skranker for $u-U$ bygger på ulikheten (6.5), nemlig at

$$a(u-U,u-U)\le a(u-V,u-V)\quad\text{for alle } V\in S_h$$

dette betyr altså at

$$\int_\Omega|\nabla(u-U)|^2\,\mathrm{d}A\le\int_\Omega|\nabla(u-V)|^2\,\mathrm{d}A\quad\text{for alle } V\in S_h$$

La nå $V=\tilde{u}$ der $\tilde{u}$ er funksjonen fra S_h som interpolerer u i nodene i nettet ($\mathcal{Z}$). Vi kan da skrive

$$u=\sum_{p\in\mathcal{Z}}u_p\varphi_p,\qquad u_p=u(x_p,y_p)$$

dvs u_p er den eksakte løsningen av variasjonsproblemet. Et uvurdelig hjelpemiddel i analyse av elementmetoden er de såkalte Sobolevnormene

$$\|u\|_m=\left(\int_\Omega\sum_{i+j\le m}|\partial_x^i\partial_y^j u|^2\mathrm{d}A\right)^{1/2}$$

Spesielt har man da at $\|\cdot\|_0$ blir den vanlige L^2-normen

$$\|u\|_0=\left(\int_\Omega|u|^2\,\mathrm{d}A\right)^{1/2}$$

Mens $\|\cdot\|_m$, $m=1,2$, også kalt H^m-normen blir

$$\|u\|_1=\left(\int_\Omega(|u|^2+|\partial_x u|^2+|\partial_y u|^2)\,\mathrm{d}A\right)^{1/2}$$

$$\|u\|_2=\left(\int_\Omega(|u|^2+|\partial_x u|^2+|\partial_y u|^2+|\partial_x^2 u|^2+|\partial_y^2 u|^2+2|\partial_x\partial_y u|^2)\,\mathrm{d}A\right)^{1/2}$$

Fra approksimasjonsteorien kan vi hente følgende resultat

$$\left(a(u-\tilde{u},u-\tilde{u})\right)^{1/2}\le K\,|u|_2\,h$$

hvor K er en konstant, og h er som tidligere definert, største avstand i mellom 2 nabonoder. Denne ulikheten kan videre benyttes til å vise at

$$\|u-U\|_0\le C\,|u|_2\,h^2$$

6.4 Problem med inhomogene randkrav

La oss nå se på Poisson's ligning med inhomogene randkrav

$$\begin{aligned}-\Delta u&=f, &&\text{i}\quad\Omega\\ u&=g, &&\text{på}\quad\partial\Omega\end{aligned}$$

Dette problemet kan transformeres til et problem med homogene randkrav (dvs $u = 0$ på $\partial\Omega$) ved at man finner en funksjon w som oppfyller $w = g$ på $\partial\Omega$ og setter $u = v + w$. Dermed må v oppfylle

$$\begin{aligned} -\Delta v &= f + \Delta w \qquad \text{i} \quad \Omega \\ v &= 0, \qquad\qquad \text{på} \quad \partial\Omega \end{aligned}$$

Ekvivalent kan man løse variasjonsproblemet:

$$\text{Finn } v \in S \text{ slik at } a(v,\zeta) = \langle f,\zeta\rangle - a(w,\zeta) \quad \text{for alle } \zeta \in S$$

Fører vi nå u tilbake kan vi skrive: Finn $u \in H^1$ slik at

$$\begin{aligned} u &= g \qquad \text{på } \partial\Omega \\ a(u,\zeta) &= \langle f,\zeta\rangle \qquad \text{for alle } \zeta \in S \end{aligned}$$

Når vi her snakker om H^1 så kan du for enkelthets skyld (og litt upresist) tenke på dette som funksjoner som er kontinuerlige i Ω og med stykkevis kontinuerlige partiellderiverte, men som i motsetning til funksjoner i S ikke behøver å være 0 på $\partial\Omega$. En har naturligvis $S \subset H^1$.

6.4.1 Tilnærmet løsning av inhomogent problem

La oss nå si at
$\mathcal{Z}$: mengden av indre noder i nettet
$\mathcal{B}$: mengden av randnoder i nettet
Sett så

$$U = \sum_{p\in\mathcal{Z}} U_p\varphi_p + \sum_{p\in\mathcal{B}} g_p\varphi_p$$

der den første summen approksimerer v og den andre approksimerer w. Merk at funksjonene φ_p, $p \in \mathcal{B}$ ikke er med i S de er avkuttede pyramidefunksjoner sentrert på randen.

Vi bestemmer nå U_p ved å forlange at

$$a(U,V) = \langle f,V\rangle \qquad \text{for alle } V \in S_h$$

som er ekvivalent med at

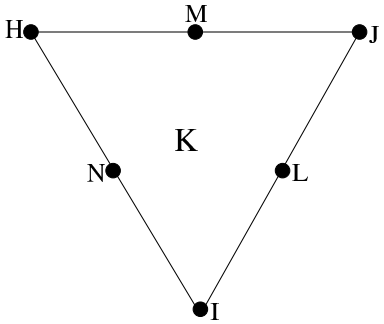
$$\sum_{q\in\mathcal{Z}} a(\varphi_p,\varphi_q)U_q = \langle f,\varphi_p\rangle - \sum_{k\in\mathcal{B}} g_k a(\varphi_p,\varphi_k) \qquad \text{for alle } p \in \mathcal{Z}.$$

6.5 Andre elementer og elementfunksjoner

Triangulære elementer og stykkevis lineære elementfunksjoner var de første elementtyper og funksjoner i bruk. I dag brukes mer kompliserte funksjoner og former. Vi skal gå igjennom noen eksempler.

Triangulære elementer, kvadratiske elementfunksjoner

Vi inkluderer nå noder midt på hver sidekant. S_h skal bestå av kontinuerlige funksjoner v hvor $v|_K$ er et andregradspolynom i x og y for hver K. Vi kan finne 6 formfunksjoner



$$\psi_p^K = a_p^K + b_p^K x + c_p^K y + d_p^K x^2 + e_p^K xy + f_p^K y^2$$

slik at

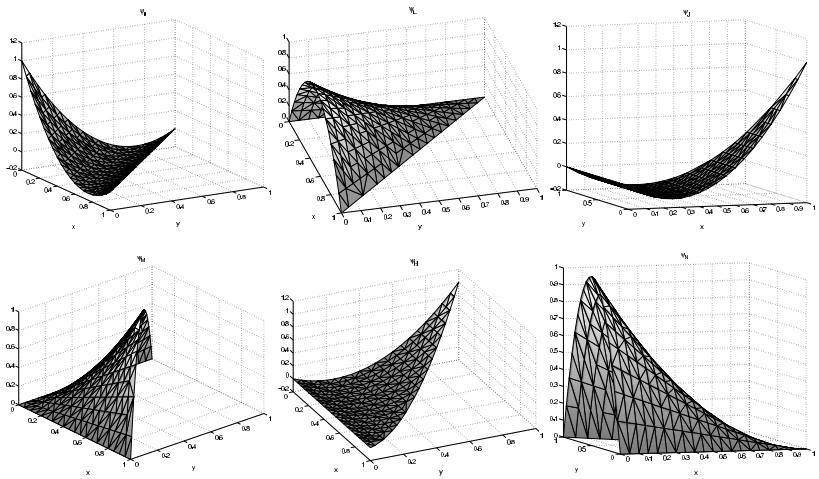
$$\psi_p^K(x_q,y_q) = \begin{cases} 1, & q = p \\ 0, & q \neq p \end{cases}$$

så for $v \in S_h$ gjelder

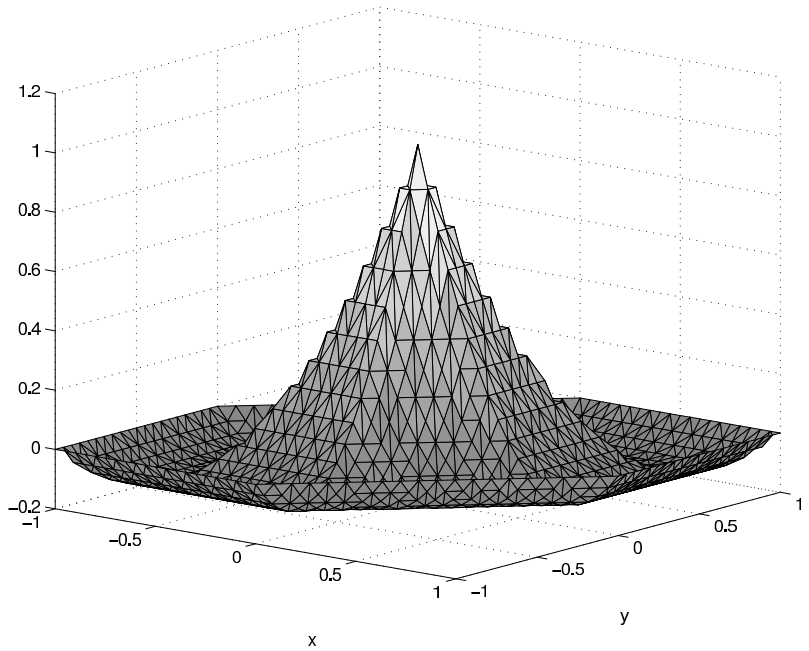
$$v|_K = v_I\psi_I^K + v_J\psi_J^K + v_H\psi_H^K + v_L\psi_L^K + v_M\psi_M^K + v_N\psi_N^K$$

der $v_p = v(x_p,y_p)$ for $p = I,J,H,L,M,N$. Av formfunksjonene lager vi basisfunksjoner som nå blir pyramider med krumme sideflater. Vi skriver som før

$$U = \sum_{p\in\mathcal{Z}} U_p\varphi_p$$



Vi plotter også basisfunksjonen φ_p for trekantnettet brukt som eksempel tidligere.

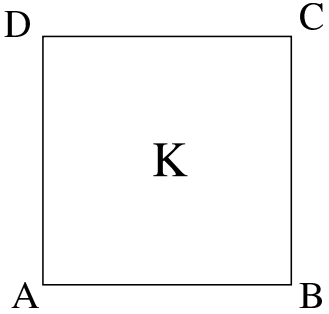


Kvadratiske elementer, bilineære elementfunksjoner

Vi antar nå for enkelthetsskyld at vi har konstante skrittlengder h i begge retninger. En bilineær funksjon har formen

$$f(x,y) = a + bx + cy + dxy$$

vi kaller den bilineær fordi dersom vi holder den ene variablen (av x og y) konstant, så blir det en rett linje i den andre. Vi lar S_h bestå av kontinuerlige funksjoner som er stykkevis bilineære.



Vi innfører nå formfunksjoner relativt til $[0,1]\times[0,1]$ ved

$$\psi_A^K(\xi,\eta) = (1-\xi)(1-\eta)$$
$$\psi_B^K(\xi,\eta) = \xi(1-\eta)$$
$$\psi_C^K(\xi,\eta) = \xi\eta$$
$$\psi_D^K(\xi,\eta) = (1-\xi)\eta$$

og finner tilsvarende funksjoner relativt til vilkårlige kvadrater ved å la

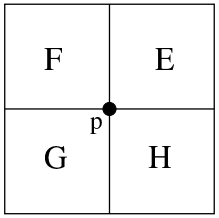
$$\xi = \frac{x-x_A}{h}, \qquad \eta = \frac{y-y_A}{h}$$

Da kan $v|_E$ der $v \in S_h$ skrives

$$v|_E = v_A\,\psi_A^K + v_B\,\psi_B^K + v_C\,\psi_C^K + v_D\,\psi_D^K +$$

hvor $v_p = v(x_p, y_p)$ for $p = A, B, C, D$. Så $v|_E$ er entydig gitt ved sine verdier i hjørnene av K. Fra formfunksjonene finner vi basisfunksjoner

$$\varphi_p = \begin{cases} \psi_p^E &= (1-\xi)(1-\eta) & \text{i } E \\ \psi_p^F &= (1+\xi)(1-\eta) & \text{i } E \\ \psi_p^G &= (1+\xi)(1+\eta) & \text{i } E \\ \psi_p^H &= (1-\xi)(1+\eta) & \text{i } E \end{cases}$$



Vi har her

$$\xi = \frac{x-x_p}{h}, \qquad \eta = \frac{y-y_p}{h}$$

Funksjonen φ_p er karakterisert ved at den er kontinuerlig i Ω, stykkevis bilineær og oppfyller

$$\varphi_p(x_q, y_q) = \begin{cases} 1 & q = p \\ 0 & q \neq p \end{cases}$$

6.6 Generelle 2. ordens differensialligninger

La oss nå se på det mer generelle problemet

$$-Lu = f \qquad \text{i } \Omega \tag{6.7}$$
$$u = g \qquad \text{på } \partial\Omega \tag{6.8}$$

hvor

$$Lu = a\,u_{xx} + 2b\,u_{xy} + c\,u_{yy} + \tilde{d}\,u_x + \tilde{e}\,u_y + k\,u$$

Koeffisientene $a, b, c, \tilde{d}, \tilde{e}, k$ kan avhenge av x og y. Hvis a, b og c er deriverbare kan Lu omskrives til

$$Lu = (a\,u_x + b\,u_y)_x + (b\,u_x + c\,u_y)_y + d\,u_x + e\,u_y + k\,u$$

der

$$d = \tilde{d} - a_x - b_y, \qquad e = \tilde{e} - b_x - c_y$$

Vi modifiserer nå Greens identitet ved å anvende divergensteoremet på vektorfeltet

$$\vec{X}(x,y) = v\,A\nabla u, \qquad A = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$$

slik at vi får

$$\int_\Omega v\,\nabla\cdot(A\nabla u)\,\mathrm{d}A = \int_{\partial\Omega} v\,A\nabla u\cdot\vec{n}\,\mathrm{d}S - \int_\Omega \nabla v\cdot A\nabla u\,\mathrm{d}A$$

der $\vec{n} = (n_x, n_y)$ er en utadrettet normalvektor av lengde 1. Bruker vi dette på Lu og skriver ut uttrykket får vi

$$\int_\Omega v\,Lu\,\mathrm{d}A = \int_{\partial\Omega} v\,((a\,u_x + b\,u_y)\,n_x + (b\,u_x + c\,u_y)\,n_y)\,\mathrm{d}S +$$
$$\int_\Omega (-v_x\,(a\,u_x + b\,u_y) - v_y\,(b\,u_x + c\,u_y) + v\,(d\,u_x + e\,u_y + k\,u))\,\mathrm{d}A$$

Siden randintegralet vil forsvinne dersom vi velger testfunksjoner som forsvinner på $\partial\Omega$ virker det naturlig å definere den bilineære formen

$$\alpha(u,v) = \int_\Omega (v_x\,(a\,u_x + b\,u_y) + v_y\,(b\,u_x + c\,u_y) - v\,(d\,u_x + e\,u_y + k\,u))\,\mathrm{d}A$$

Vi lar S bety det samme som før, og erstatter nå differensialligningsproblemet (6.7) og (6.8) med

Finn $u \in H^1$ (u kontinuerlig og u_x, u_y stykkevis kontinuerlige) slik at

$$\alpha(u,v) = \langle f, v\rangle \qquad \text{for alle } v \in S$$
$$u = g, \qquad \text{på } \partial\Omega$$

Dette problemet løses tilnærmet som før

$$U = \sum_{p\in\mathcal{Z}} U_p\varphi_p + \sum_{p\in\mathcal{B}} g_p\varphi_p$$
$$\alpha(U,\varphi_q) = \langle f, \varphi_q\rangle \quad \text{for alle } q \in \mathcal{Z}$$

Koeffisientmatrisen A med elementer $A_{pq} = \alpha(\varphi_p, \varphi_q)$ er generelt ikke symmetrisk og behøver ikke å være regulær (inverterbar). Vi er sikret regularitet hvis vi forlanger at det fins en konstant $\gamma > 0$ slik at

$$\alpha(u,u) \geq \gamma\|u\|_1^2 \quad \text{for enhver } u \in S.$$

Her er $\|u\|_1^2 = \int_\Omega (u^2 + |\nabla u|^2)\mathrm{d}A$. Egenskapen kalles koersivitet . Differensialoperatorer L slik at α er koersiv kalles S-elliptisk (i rommet S). Hvis L er S-elliptisk, vil L være elliptisk i hvert punkt, det vil si $ac - b^2 > 0$ overalt i Ω.

Kapittel 7

Hyperbolske ligninger

7.1 Eksempler på ligninger

1. Det mest berømte eksemplet på en hyperbolsk differensialligning er kanskje den lineære andre ordens bølgeligningen
$$u_{tt} = c^2 u_{xx}$$
eller i flere romdimensjoner
$$u_{tt} = c^2\Delta u$$

2. Hvis vi ser på en generell andre ordens skalar lineær PDE i to dimensjoner av typen
$$a\,u_{xx} + 2b\,u_{xy} + cu_{yy} + du_x + eu_y + f\,u = 0$$
der a, b, c, d, e, f er funksjoner av x, y, så er denne hyperbolsk dersom
$$b^2 - ac > 0.$$
Merk at her spiller y rollen som t i ligningene ovenfor.

3. En mye studert ligning som fins i mange anvendelser er den såkalte konserveringsloven
$$u_t + c(x,t,u)\,u_x = 0$$
Dette er altså en skalar, første ordens kvasilineær PDE som vi også sier er hyperbolsk. En slik type ligning kan for eksempel brukes til å beskrive trafikkflyt langs en vei.

4. Systemer av første ordens lineære ligninger med konstante koeffisienter kan beskrives ved
$$u_t + A\,u_x = 0 \tag{7.1}$$
der $u \in \mathbf{R}^n$ og A er en reell $n \times n$-matrise. Dette systemet er hyperbolsk hvis A er diagonaliserbar med reelle egenverdier.

5. La oss se på et ikke-lineært hyperbolsk system av ligninger som beskriver en viktig anvendelse, nemlig gruntvannsligningene (shallow water equations på engelsk). Vi skriver ned disse i en romdimensjon, og lar $v(x,t)$ være vannhastigheten i punktet x ved tid t, mens $z(x,t)$ måler (den vertikale) bølgehøyden i forhold til en likevektsposisjon.

Massebevarelse $\quad z_t + (vz)_x = 0$

Impulsbevarelse $\quad v_t + (\frac{1}{2}v^2 + z)_x = 0$

6. Vi skriver nå ned generelle diffligninger på bevarelsesform

$$u_t + (f(u))_x = 0. \tag{7.2}$$

Her er $u \in \mathbf{R}^n$ mens $f : \mathbf{R}^n \to \mathbf{R}^n$ er en ikke-lineær avbildning. Et spesialtilfelle der $f(u) = Au$ for en $n \times n$ matrise A er gitt ved (7.1). For at (7.2) skal være hyperbolsk krever vi at jacobimatrisen $Df = f'(u)$ er diagonaliserbar med reelle egenverdier.

7.2 Karakteristikker

Vi skal introdusere et modellproblem som vi skal bruke ganske mye i det som følger

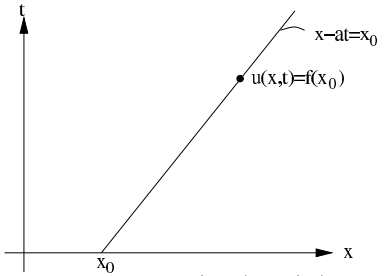
$$u_t + au_x = 0, \quad -\infty < x < \infty, \quad t \geq 0, \qquad a > 0 \tag{7.3}$$

der a altså er en konstant. Initialverdi må oppgis for hele $\mathbf{R}$

$$u(x,0) = f(x), \qquad -\infty < x < \infty.$$

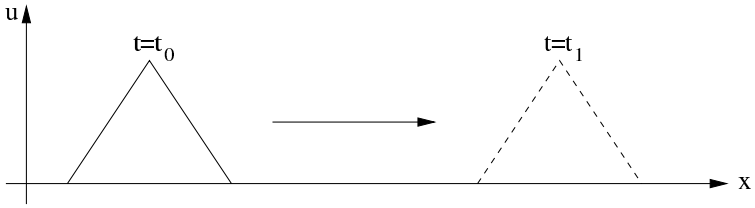
For denne enkle ligningen er det faktisk mulig å skrive ned eksakt løsning, som er

$$u(x,t) = f(x - at).$$



Sjekk selv ved innsetting at denne oppfyller (7.3). I (x,t)-planet er det altså en rett linje der $u(x,t)$ er konstant, nemlig linjen $x = x_0 + at$ der $u(x,t) = u(x_0 + at, t) = f(x_0 + at - at) = f(x_0)$. Denne linjen kalles for en *karakteristikk*.

Konstante verdier av u forplanter seg altså fra initialverdien $u(x_0, 0) = f(x_0)$ og framover i tid. Hastigheten blir det resiproke av stigningstallet til kurven. Er kurven vertikal så er altså stigningstallet uendelig, og hastigheten blir 0. Dersom karakteristikken skrår oppover mot høyre så forplanter verdien $u(x_0, 0)$ seg mot høyre. I et xu-plott kan vi tegne løsningen ved to ulike tidspunkt



La oss nå se på den mer generelle ligningen

$$u_t + a(x,t)\, u_x = b(x,t) \tag{7.4}$$

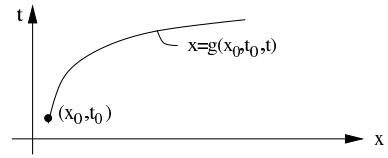
Karakteristisk diffligning for (7.4)

$$\frac{dx}{dt} = a(x,t) \tag{7.5}$$

Anta nå at x_0 og t_0 er gitt og la

$$x = g(x_0, t_0, t)$$

være en løsning av (7.5) som oppfyller $x(t_0) = x_0$. La $u(x,t)$ være en løsning av (7.4) og se på



Karakteristikk γ gjennom (x_0, t_0)

$$v(t) := u(x,t)|_{x=g} = u(g(x_0,t_0,t),t)$$

Her angir v løsningen $u(x,t)$ langs kurven γ. Vi kan derivere v med hensyn på tiden og får ved bruk av kjerneregelen for derivasjon

$$v'(t) = \left(u_t + u_x \frac{\mathrm{d}x}{\mathrm{d}t}\right)\Big|_{x=g} = (u_t + a(x,t)u_x)|_{x=g} = b(x,t)|_{x=g}$$

Vi ser altså at langs karakteristikken er $v'(t)$ en kjent funksjon av t, så vi kan finne v ved integrasjon

$$v(t) = v(t_0) + \int_{t_0}^{t} b(x,s)|_{x=g(x_0,t_0,s)}\, \mathrm{d}s.$$

Hvis vi kjenner $u(x_0,t_0)$ kan vi finne $u(x,t)$ for $(x,t) \in \gamma$ fra formelen

$$u(x,t) = u(x_0,t_0) + \int_{t_0}^{t} b(g(x_0,t_0,s),s)\, \mathrm{d}s \tag{7.6}$$

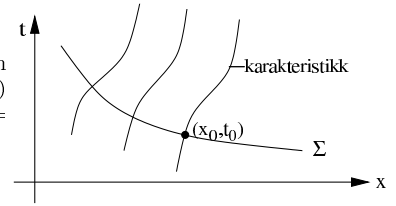
Spesialtilfelle: $b(x,t) \equiv 0$ impliserer $u(x,t) = u(x_0,t_0)$ for alle $(x,t) \in \gamma$.

Startverdiproblem for (7.4)**.** Anta at $u(x,t)$ oppfyller PDE'en (7.4) samt at

$$u(x,t) = f(x,t) \qquad \text{langs en kurve } \Sigma: \; x = \sigma(t)$$

NB! Σ skal *ikke* være en karakteristikk.

Løsningsmetode: Beregn karakteristikken $x = g(x_0,t_0,t)$ gjennom punktet (x_0,t_0) på Σ. Bruk deretter (7.6) med $u(x_0,t_0) = f(x_0,t_0)$.



Et enkelt eksempel. La oss se på det enkleste problemet

$$u_t + au_x = 0 \qquad a \in \mathbf{R}.$$

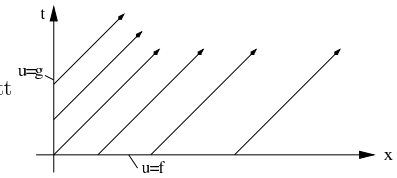
Karakteristikkene blir da

$$\frac{dx}{dt} = a \qquad \Leftrightarrow \qquad x = g(x_0,t_0,t) = x_0 + a(t - t_0)$$

La oss for eksempel si at kurven Σ er x-aksen der $u(x,0) = f(x)$ dvs vi har $t_0 = 0$. Karakteristikken gjennom (x,t) krysser åpenbart x-aksen i $x_0 = x - at$ der løsningen er $u(x_0,0) = f(x_0) = f(x - at)$ og vi har reprodusert eksakt løsning som vi har presentert tidligere.

Start/randverdiproblem for $u_t + au_x = 0$. Vi formulerer problemet for $0 \le x \le \infty$ og $t \ge 0$. Anta nå for enkelhets skyld at $a > 0$.

Startverdi $u(x,0) = f(x), \quad x \ge 0$
Randverdi $u(0,t) = h(t), \quad t \ge 0$



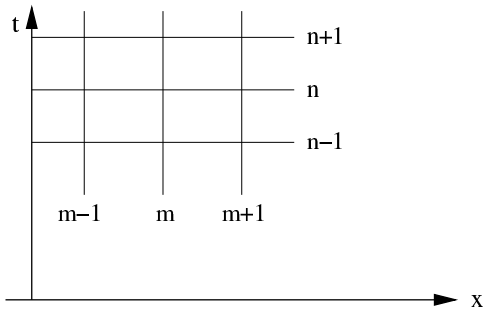
Eksakt løsning til dette problemet er gitt ved

$$u(x,t) = \begin{cases} f(x-at), & x - at \ge 0 \\ h(t - x/a), & x - at < 0. \end{cases}$$

Merknad. Vi kunne godt ha brukt to render, for eksempel $x = 0$ og $x = 1$ ovenfor, men det blir da galt å spesifisere løsningen langs linja $x = 1$.

7.3 Eksplisitte differenseformler for $u_t + au_x = 0$

$x_m = mh, \quad t_n = nk$
$u_m^n = u(x_m, t_n)$



Ulike diskretiseringer

$$\partial_t u_m^n = \frac{1}{k}(u_m^{n+1} - u_m^n) + \mathcal{O}(k) \tag{7.7}$$

$$\partial_x u_m^n = \frac{1}{h}(u_{m+1}^n - u_m^n) + \mathcal{O}(h) \tag{7.8}$$

$$\partial_x u_m^n = \frac{1}{h}(u_m^n - u_{m-1}^n) + \mathcal{O}(h) \tag{7.9}$$

$$\partial_x u_m^n = \frac{1}{2h}(u_{m+1}^n - u_{m-1}^n) + \mathcal{O}(h^2) \tag{7.10}$$

Om vi velger (7.7) og (7.9) får vi

$$\frac{1}{k}(u_m^{n+1} - u_m^n) + \frac{a}{h}(u_m^n - u_{m-1}^n) + \mathcal{O}(k) + \mathcal{O}(h) = 0$$

og dermed for den numeriske metoden

$$U_m^{n+1} = (1 - ap)U_m^n + apU_{m-1}^n, \qquad p = \frac{k}{h} \tag{7.11}$$

Avbruddsfeilen blir $\tau_m^n = \mathcal{O}(k^2) + \mathcal{O}(kh)$. Om vi skulle finne på å velge (7.10) for u_x får vi istedet

$$U_m^{n+1} = U_m^n - \frac{ap}{2}(U_{m+1}^n - U_{m-1}^n) \tag{7.12}$$

Her blir $\tau_m^n = \mathcal{O}(k^2) + \mathcal{O}(kh^2)$, tilsynelatende en framgang i forhold til (7.11) men som vi skal se senere: (7.12) *er alltid ustabil for dette problemet*!

Lax-Wendroff's formel. Fra diffligningen kan vi utlede

$$u_{tt} = (-au_x)_t = -a(u_t)_x = -a(-au_x)_x = a^2 u_{xx}$$

Vi lager nå en formel ved å rekkeutvikle u_m^{n+1} til andre orden, bruke diffligningen, og deretter diskretisere romderiverte med sentraldifferenser. Vi får da

$$\begin{aligned} u_m^{n+1} &= u_m^n + k\,\partial_t u_m^n + \frac{1}{2}k^2\,\partial_t^2 u_m^n + \mathcal{O}(k^3) \\ &= u_m^n - ak\,\partial_x u_m^n + \frac{1}{2}(ak)^2\,\partial_x^2 u_m^n + \mathcal{O}(k^3) \\ &= u_m^n - ak\,\frac{1}{2h}(u_{m+1}^n - u_{m-1}^n) + +\frac{1}{2}(ak)^2\,\frac{1}{h^2}\,\delta_x^2 u_m^n + \mathcal{O}(kh^2) + \mathcal{O}(k^3) \end{aligned}$$

Herfra kommer

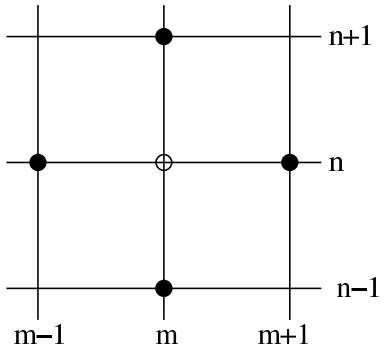
Lax-Wendroff's formel for $u_t + au_x = 0$

$$U_m^{n+1} = U_m^n - \frac{ap}{2}(U_{m+1}^n - U_{m-1}^n) + \frac{1}{2}(ap)^2\,\delta_x^2 U_m^n \tag{7.13}$$

Avbruddsfeil

$$\tau_m^n = \mathcal{O}(k^3) + \mathcal{O}(kh^2)$$

Hoppe-bukk formel (Leap frog).



$$\partial_t u_m^n = \frac{1}{2k}(u_m^{n+1} - u_m^{n-1}) + \mathcal{O}(k^2)$$

$$\partial_x u_m^n = \frac{1}{2h}(u_{m+1}^n - u_{m-1}^n) + \mathcal{O}(h^2)$$

Vi får dermed formelen **Hoppe-bukk formel (leap-frog) for $u_t + au_x = 0$**

$$U_m^{n+1} = U_m^{n-1} - ap(U_{m+1}^n - U_{m-1}^n)$$

Avbruddsfeil: $\tau_m^n = \mathcal{O}(k^3) + \mathcal{O}(kh^2)$

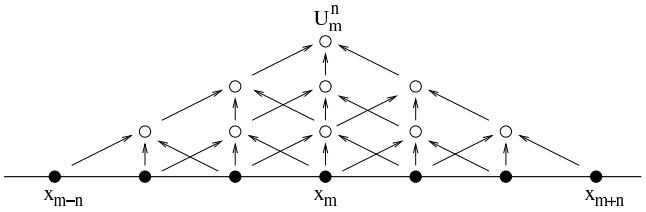
Legg merke til at dette er en to-skritts (tre-nivå) metode i tid. Derfor trengs en spesiell oppstartingsmetode, akkurat som for flerskrittsmetoder for ordinære differensialligninger.

7.4 Stabilitet

Courant-Friedrichs-Levy betingelsen Vi ser fremdeles på ligningen $u_t + au_x = 0$. Anta at vi har en differensformel av typen

$$U_m^{n+1} = \alpha_{-1}U_{m-1}^n + \alpha_0 U_m^n + \alpha_1 U_{m+1}^n d \tag{7.14}$$

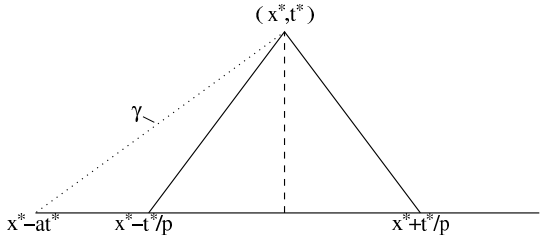
Vi studerer nå avhengighetsområdet for U_m^n. U_m^n avhenger av $U_{m+\ell}^0$ der $-n \le \ell \le n$. Avhengighetsintervall for U_m^n på x-aksen er $I_m^n = [x_{m-n}, x_{m+n}]$.



La oss nå holde $x_m = x^* = mh$ og $t_n = t^* = nk$ fast i det vi lar h og k gå mot null og m, n gå mot uendelig samtidig. Anta også at dette gjøres slik at $p = k/h =$ konstant. Endepunktene for I_m^n blir da

$$x_{m\pm n} = x^* \pm nh = x^* \pm t^* \frac{h}{k} = x^* \pm \frac{t^*}{p}$$

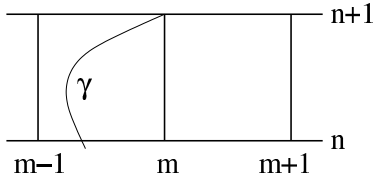
Så $I_m^n = [x^* - \frac{t^*}{p}, x^* + \frac{t^*}{p}]$. Merk at dette intervallet er fiksert når h og k går mot null slik det er beskrevet ovenfor. La γ være en karakteristikk for $u_t + au_x = 0$ som går gjennom punktet (x^*, t^*). Denne har stigningstall a og skjærer x-aksen i $x^* - at^*$ så $u(x^*, t^*) = f(x^* - at^*)$. Dersom $x^* - at^* \notin I_m^n$ betyr det at den beregnende approksimasjonen U_m^n bygger på initialdata som ikke inkluderer $f(x^* - at^*)$ uansett hvor liten h og k er. I situasjonen på figuren kan man derfor ikke ha konvergens mot eksakt løsning for alle initialverdier når $h, k \to 0$.



Nødvendig kriterium for konvergens: CFL-betingelsen. Karakteristikk gjennom (x^*, t^*) må skjære x-aksen i et punkt i avhengighetsintervallet for $x_m = x^*,\ t_n = t^*$. CFL er en forkortelse for *Courant-Friedrichs-Levy*.

Prinsippet gjelder også for krumme karakteristikker, slik som når $a = a(x, t)$.

Karakteristikken gjennom (x_m, t_{n+1}) må skjære linja $t = t_n$ mellom ytterpunktene av de to m-verdiene som brukes i formelen, slik som angitt på figuren. Karakteristikkene må aldri forlate avhengighetsområdet.



La oss se hva kriteriet blir med konstant a og en differensformel av typen presentert ovenfor. Vi har da kravet

$$x^* - t^*/p \le x^* - at^* \le x^* + t^*/p$$

som gir $|a|p < 1$. Dette er altså et nødvendig kriterium for konvergens for alle mulige differenseformler av typen (7.14).

Von Neumann-betingelsen. Vi minner om denne betingelsen, diskutert tidligere i kurset, som bygger på Fourieranalyse. Metoden går ut på å sette

$$U_m^n = \xi^n \mathrm{e}^{\mathrm{i}\beta x_m}, \quad \mathrm{i} = \sqrt{-1} \tag{7.15}$$

inn i differensligningen, og deretter løse denne med hensyn på *forsterkningsfaktoren* ξ. Stabilitetskrav er da

$$|\xi| \le 1 \quad \text{for enhver } \beta \in \mathbf{R}.$$

Stabilitet av Lax-Wendroff's skjema. Vi minner om formelen som gjelder for problemet $u_t + au_x = 0$

$$U_m^n - \frac{1}{2}ap(U_{m+1}^n - U_{m-1}^n) + \frac{1}{2}(ap)^2(U_{m+1}^n - 2U_m^n + U_{m-1}^n)$$

Om vi setter inn (7.15), forkorter med $\xi^n \mathrm{e}^{\mathrm{i}\beta x_m}$ på hver side og bruker $\mathrm{e}^{\mathrm{i}\beta h} = \cos\beta h + \mathrm{i}\sin\beta h$ får vi

$$\xi = 1 - \mathrm{i}\,ap\sin\beta h + (ap)^2(\cos\beta h - 1) = 1 - 2(ap)^2\sin^2\frac{\beta h}{2} - \mathrm{i}\,ap\sin\beta h$$

Vi har her benyttet den trigonometriske identiteten $\cos\beta h = 1 - 2\sin^2\frac{\beta h}{2}$ noe vi gjør også nedenfor. La oss definere $r = ap$ og $q = \sin\frac{\beta h}{2}$.

$$\begin{aligned}|\xi|^2 &= (1 - 2r^2q^2)^2 + r^2(1 - \cos^2\beta h) = (1 - 2r^2q^2)^2 + r^2(1 - (1 - 2q^2)^2) \\ &= 1 - 4r^2q^2 + 4r^4q^4 + r^2 + r^2 - r^2(1 - 4q^2 + 4q^4) \\ &= 1 - 4r^2(1 - r^2)q^4\end{aligned}$$

Vi krever da

$$\begin{gathered}1 - 4r^2(1 - r^2)q^4 \le 1, \qquad 0 \le q \le 1 \\ \Updownarrow \\ 4r^2(1 - r^2) \ge 0 \\ \Updownarrow \\ |r| \le 1 \\ \Updownarrow \\ |a|\,p \le 1\end{gathered}$$

så vi får akkurat det samme kravet som CFL-betingelsen gir.

En enklere formel å sjekke er den "naive" formelen (7.12) som vi advarte mot

$$U_m^{n+1} = U_m^n - \frac{ap}{2}(U_{m+1}^n - U_{m-1}^n)$$

Vi får da

$$\xi = 1 - \mathrm{i}\,ap\sin\beta h \tag{7.16}$$

slik at

$$|\xi|^2 = 1 + (ap)^2\sin^2\beta h > 1$$

for nesten alle β, det vil si den er ustabil for alle skrittlengder. En interessant modifikasjon av denne dårlige metoden kan en få ved å erstatte

$$U_m^n \quad \text{med} \quad \frac{1}{2}\left(U_{m-1}^n + U_{m+1}^n\right)$$

I von Neumann analysen erstattes da uttrykket for ξ i (7.16) med

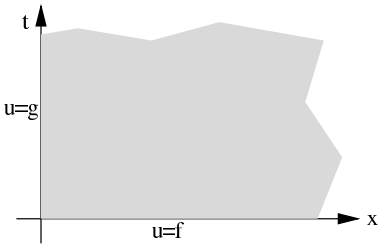
$$\xi = \cos\beta h - \mathrm{i}\,ap\sin\beta h,$$

og en finner at $|\xi||^2 = 1 + (1 - (ap)^2)\sin^2\beta h$ slik at en får stabilitet hvis $|ap| \le 1$. Denne metoden har navnet *Lax-Friedrichs*.

7.5 Implisitte metoder for $u_t + au_x = 0$

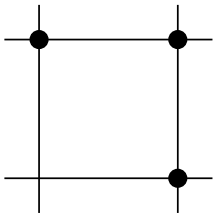
Implisitte metoder kan bare brukes på start/randverdiproblemer. Vi ser altså på problemer av typen

$$\begin{aligned} u_t + au_x &= 0, \quad && x \geq 0,\ t \geq 0 \\ u(x,0) &= f(x), && x \geq 0 \\ u(0,t) &= g(t), && t \geq 0 \end{aligned}$$



Den enkleste implisitte formelen. Vi forsøker med enklest mulig diskretisering av deriverte i (x_m, t_{n+1})

$$\begin{aligned} \partial_t u_m^{n+1} &= \frac{1}{k}(u_m^{n+1} - u_m^n) + \mathcal{O}(k) \\ \partial_x u_m^{n+1} &= \frac{1}{h}(u_m^{n+1} - u_{m-1}^{n+1} + \mathcal{O}(h) \end{aligned}$$



som gir formelen

$$U_m^{n+1} - U_m^n + ap(U_m^{n+1} - U_{m-1}^{n+1}) = 0$$

Vi kan løse eksplisitt med hensyn på U_m^{n+1} og får da

$$U_m^{n+1} = \frac{ap}{1+ap}\, U_{m-1}^{n+1} + \frac{1}{1+ap}\, U_m^n.$$

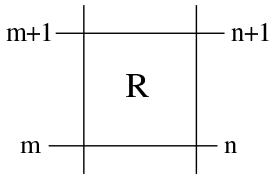
Selv om superskript $n+1$ inngår på høyre side av denne ligningen så er metoden i praksis eksplisitt dersom vi beregner løsningsverdier på tidsnivå $n+1$ fra venstre mot høyre. For beregning av U_1^{n+1} ser vi at denne avhenger kun av U_0^{n+1} på tidsnivå $n+1$ og denne er gitt som $g(t_{n+1})$. Denne beregnede U_1^{n+1} brukes deretter til beregning av U_2^{n+1} osv.

Vi finner at metodens avbruddsfeil er gitt som

$$\tau_m^n = -\frac{1}{2}\left(a^2 k^2 + a\, hk\right) \partial_x^2 u_m^n + \cdots = \mathcal{O}(k^2 + hk)$$

Wendroff's metode. Vi gjør en form for boksintegrasjon

$$\begin{gathered} u_t + au_x = 0 \\ \Downarrow \\ \int_{t_n}^{t_{n+1}} \int_{x_m}^{x_{m+1}} (u_t + au_x)\, \mathrm{d}x\, \mathrm{d}t = 0 \end{gathered}$$



Om vi bytter integrasjonsrekkefølgen for det første av leddene og bruker fundamentalsetningen for kalkulus, finner vi

$$\int_{x_m}^{x_{m+1}} (u^{n+1} - u^n)\, \mathrm{d}x + a \int_{t_n}^{t_{n+1}} (u_{m+1} - u_m)\, \mathrm{d}t = 0 \tag{7.17}$$

der

$$u^n = u(x, t_n), \quad \text{og} \quad u_m = u(x_m, t)$$

Nå minnes vi trapesregelen for integrasjon. Gitt en funksjon $f(x)$ så gjelder

$$\int_r^{r+d} f(s)\, \mathrm{d}s = \frac{d}{2}(f(r) + f(r+d)) - \frac{1}{12} d^3 f''(r + \frac{d}{2}) + \cdots$$

Så om vi anvender trapesregelen på begge integralene i (7.17) får vi

$$\frac{h}{2}\Big((u_m^{n+1} - u_m^n) + (u_{m+1}^{n+1} - u_{m+1}^n)\Big) + \frac{ak}{2}\Big((u_{m+1}^n - u_m^n) + (u_{m+1}^{n+1} - u_m^{n+1})\Big) + \mathcal{O}(k^3 + h^3) = 0$$

Wendroff's metode.

$$(1+ap)\, U_{m+1}^{n+1} + (1-ap)\, U_m^{n+1} - (1-ap)\, U_{m+1}^n - (1+ap)\, U_m^n = 0$$

Avbruddsfeilen kan utvikles omkring midtpunktet $(x_m + h/2, t_n + k/2)$ av rektanglet R, og en får da

$$\tau_m^n = \frac{1}{6}\Big(a^3 k^3 - a\, kh^2\Big) \partial_x^3 u_{m+1/2}^{n+1/2} + \cdots = \mathcal{O}(k^3 + kh^2)$$

For å studere metodens stabilitet bruker vi Von Neumann-betingelsen igjen, med $\gamma = (1-ap)/(1+ap)$ kan vi skrive

$$U_{m+1}^{n+1} + \gamma\, U_m^{n+1} - \gamma\, U_{m+1}^n - U_m^n = 0.$$

Ved å sette $U_m^n = \mathrm{e}^{\mathrm{i}\beta mh}$ får vi

$$\xi(\mathrm{e}^{\mathrm{i}\beta h} + \gamma) = \gamma \mathrm{e}^{\mathrm{i}\beta h} + 1,$$

som vi løser mhp ξ og får

$$\xi = \mathrm{e}^{\mathrm{i}\beta h}\, \frac{\gamma + \mathrm{e}^{-\mathrm{i}\beta h}}{\gamma + \mathrm{e}^{\mathrm{i}\beta h}}.$$

Den første faktoren $\mathrm{e}^{\mathrm{i}\beta h}$ har absoluttverdi 1, og brøken er et uttrykk av form z^*/z så det har også absoluttverdi 1. Dermed er $|\xi| = 1$ for alle β og vi sier at skjemaet til Wendroff er ubetinget stabilt, dvs stabilt for alle h og k.

7.6 Hyperbolske systemer av første ordens ligninger

Definisjon av hyperbolisitet, karakteristikker. Vi ser på systemer av 1. ordens partielle differensialligninger med to uavhengige variable x og t og ℓ avhengige variable

$$\begin{aligned} &u_t + Au_x = 0, \\ &u = (u_1, \ldots, u_\ell)^T,\ A \in \mathbf{R}^{\ell \times \ell}. \end{aligned} \tag{7.18}$$

Til å begynne med lar vi A være konstant. Hvis A er diagonaliserbar og har reelle egenverdier, kalles (7.18) hyperbolsk. I så fall kan A skrives

$$A = T\Lambda T^{-1}, \quad \Lambda = \mathrm{diag}(\lambda_i)_{i=1:\ell}, \quad \lambda_i \quad \text{reelle}$$

Vi gjør en variabeltransformasjon

$$\begin{aligned} u &= Tv \Leftrightarrow v = T^{-1}u \\ u_t &= Tv_t \\ u_x &= Tv_x \end{aligned}$$

hvor T er en konstant matrise. Sett inn i (7.18)

$$Tv_t + (T\Lambda T^{-1})Tv_x = T(v_t + \Lambda v_x) = 0$$

Vi har dermed dekoblet (7.18)

$$\begin{gathered} v_t + \Lambda v_x = 0 \\ \Updownarrow \\ (v_i)_t + \lambda_i (v_i)_x = 0, \quad i = 1, \ldots, \ell \end{gathered} \tag{7.19}$$

For hver ligning har vi en karakteristikkligning

$$\frac{dx}{dt} = \lambda_i, \quad i = 1, \ldots, \ell. \tag{7.20}$$

Dermed ser vi at til (7.18) hører ℓ karakteristikkligninger og ℓ familier av karakteristikker,

$$x = \lambda_i t + \text{konstant}, \quad i = 1, \ldots, \ell. \tag{7.21}$$

Eksempel. Vi omdanner bølgeligningen

$$\phi_{tt} = c^2 \phi_{xx}$$

til et system av 1. ordens ligninger. Sett

$$\begin{aligned} &u_1 = \phi_t, \quad u_2 = \phi_x, \\ &(u_1)_t = \phi_{tt} = c^2 \phi_{xx} = c^2 (u_2)_x \\ &(u_2)_t = \phi_{xt} = \phi_{tx} = (u_1)_x \end{aligned}$$

Eller på formen (7.18)

$$\begin{pmatrix} u_1 \\ u_2 \end{pmatrix}_t + \begin{pmatrix} 0 & -c^2 \\ -1 & 0 \end{pmatrix} \cdot \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}_x = 0 \tag{7.22}$$

Så matrisen A i (7.18) har formen

$$A = -\begin{pmatrix} 0 & c^2 \\ 1 & 0 \end{pmatrix}$$

A har egenverdier og egenvektorer

$$\lambda_1 = c, \quad t_1 = \begin{pmatrix} -c \\ 1 \end{pmatrix}, \quad \lambda_2 = -c, \quad t_2 = \begin{pmatrix} c \\ 1 \end{pmatrix}$$

og er derfor diagonaliserbar:

$$\begin{aligned} &A = T\Lambda T^{-1}, \\ &\Lambda = \begin{pmatrix} c & 0 \\ 0 & -c \end{pmatrix}, \quad T = \begin{pmatrix} -c & c \\ 1 & 1 \end{pmatrix}, \quad T^{-1} = \begin{pmatrix} -1/(2c) & 1/2 \\ 1/(2c) & 1/2 \end{pmatrix} \end{aligned}$$

Karakteristikkligninger er $dx/dt = \pm c$, og (7.22) har to familier av karakteristikker $x = \pm ct +$ konst. Transformasjon til formen (7.19)

$$\begin{aligned} &v = T^{-1}u \Leftrightarrow v_1 = (-u_1/c + u_2)/2, \; v_2 = (u_1/c + u_2)/2, \\ &(v_1)_t - c(v_1)_x = 0, \\ &(v_2)_t + c(v_2)_x = 0. \end{aligned}$$

Generelt lineært system av 1. ordens ligninger.

$$u_t + A(x,t)u_x + B(x,t)u = f(x,t) \tag{7.23}$$

Ligning (7.23) er hyperbolsk hvis A er diagonaliserbar med reelle egenverdier

$$\begin{aligned} A(x,t) &= T(x,t) \cdot \Lambda(x,t) \cdot T^{-1}(x,t) \\ \Lambda(x,t) &= \mathrm{diag}(\lambda_i(x,t))_{i=1:l}. \end{aligned}$$

Variabeltransformasjon

$$\begin{aligned} u &= T\,v \\ u_t &= Tv_t + T_t v, \quad u_x = Tv_x + T_x v \end{aligned}$$

(7.23) $\Rightarrow$

$$Tv_t + T_t v + T\Lambda T^{-1}(Tv_x + T_x v) + BTv = f$$

som gir

$$\begin{aligned} &v_t + \Lambda v_x + \Gamma v = g, \\ &\Gamma = T^{-1}T_t + \Lambda T^{-1}T_x + T^{-1}BT, \\ &g = T^{-1}f. \end{aligned} \tag{7.24}$$

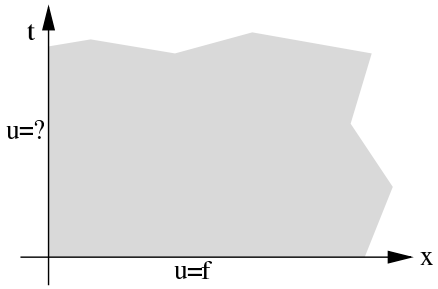
Ligningene (7.24) er ikke et dekoblet system som (7.21)), men leder til karakteristikkligningene for (7.23)

$$\frac{dx}{dt} = \lambda_i(x,t), \quad i = 1, \ldots, \ell.$$

Startverdiproblem og start/randverdiproblem. Startverdiproblemet er gitt som

$$\begin{aligned} &u_t + Au_x = 0, \quad -\infty < x < \infty, \; t > 0 \\ &u(x,0) = f(x), \quad -\infty < x < \infty. \end{aligned}$$

Start/randverdiproblem av type I.



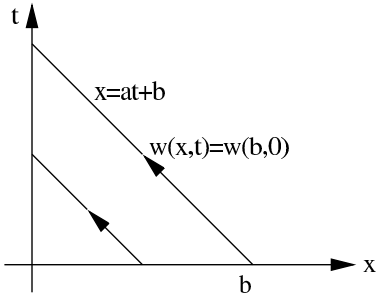
$\mathcal{R} = \{(x,t) : x \geq 0, t \geq 0$
$u_t + Au_x = 0$ i $\mathcal{R}$
$u(x,0) = f(x),\ x \geq 0$

I tillegg trengs det et sett betingelser for u langs t-aksen som vi skal se i det følgende.

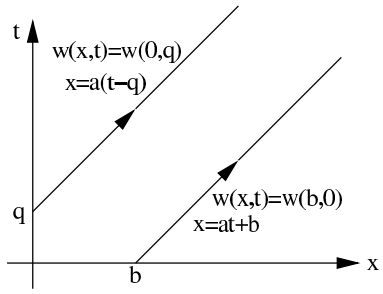
Se først på en skalar ligning

$$w_t + aw_x = 0$$

Karakteristisk ligning: $dx/dt = a$. Karakteristikker: $x = at + \text{konst}$.



$a \leq 0$. $w(x,t)$ er entydig kjent i $\mathcal{R}$ hvis $w(x,0)$ er kjent for $x \geq 0$.



$a > 0$. For å bestemme $w(x,t)$ i $\mathcal{R}$ må vi kjenne $w(x,0),\ x \geq 0$ og $w(0,t),\ t \geq 0$.

Se på et system

$$\left.\begin{array}{l} u_t + Au_x = 0,\ \text{i } \mathcal{R}, \\ u(x,0) = f(x),\ x \geq 0 \end{array}\right\} \quad \begin{array}{c} \Longleftrightarrow \\ A = T\Lambda T^{-1}, \\ u = Tv, \\ h = T^{-1}f. \end{array} \quad \left\{\begin{array}{l} v_t + \Lambda v_x = 0 \text{ i } \mathcal{R}, \\ v(x,0) = h(x), \quad x \geq 0. \end{array}\right.$$

La $\Lambda = \text{diag}(\lambda_i)_{i=1:\ell}$ og anta at ligningene er sortert slik at $\lambda_i > 0,\ i = 1,\ldots,k$ og $\lambda_i \leq 0,\ i = k+1,\ldots,\ell$ (for en eller annen k). Vi får skalare ligninger

$$(v_i)_t + \lambda_i (v_i)_x = 0 \text{ i } \mathcal{R}, \quad v_i(x,0) = h_i(x),\ x \geq 0.$$

Her vil $v_i(x,t)$ være bestemt av $v_i(x,0) = h_i(x)$ for alle $i \geq k+1$ (karakteristikker som peker oppover mot venstre). For å beregne $v_i(x,t)$ i $\mathcal{R}$ med $i \leq k$ trenger vi verdier av v_i langs t-aksen. Vi kan få disse verdiene ved å stille opp k randverdibetingelser.

$$\sum_{j=1}^{l} \gamma_{ij} v_j(0,t) = g_i(t),\ i = 1:k.$$

Ligningene må kunne løses med hensyn på $\{v_i, i = 1,\ldots,k\}$ det vil si $C = [\gamma_{ij}]_{i,j=1:k}$ må være regulær (ikke-singulær). Dette leder til følgende start/randverdiproblem:

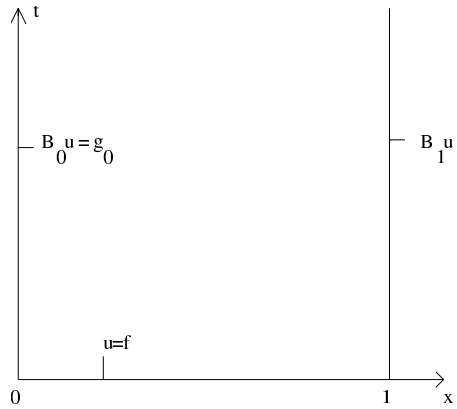
$u_t + Au_x = 0$ i $\mathcal{R}$,	$At_i = \lambda_i t_i,\ i = 1,\ldots,\ell$,
$u(x,0) = f(x),\ x \geq 0$,	$\lambda_i > 0,\ i = 1,\ldots,k$,
$Bu(0,t) = g(t)$.	B er $k \times \ell$,
	$C = B \cdot [t_1, \cdots, t_k]$ er regulær.

Eksempel.

$$\left.\begin{array}{l} (u_1)_t - c^2 (u_2)_x = 0 \\ (u_2)_t - (u_1)_x = 0 \end{array}\right\} \text{ i } \mathcal{R}, \quad \left.\begin{array}{l} u_1(x,0) = f_1(x) \\ u_2(x,0) = f_2(x) \end{array}\right\}, \quad x \geq 0.$$

$$\beta_1 u_1(0,t) + \beta_2 u_2(0,t) = g(t), \quad t \geq 0, \qquad -\beta_1 c + \beta_2 \neq 0.$$

Start/randverdiproblem av type II.
$\mathcal{R} = \{(x,t) : 0 \leq x \leq 1,\ t \geq 0\}$. Anta at A har k positive og m negative egenverdier $(k + m \leq l)$.



$u_t + Au_x = 0$ i $\mathcal{R}$,
$u(x,0) = f(x),\ 0 \leq x \leq 1$,
$B_0 u(0,t) = g_0(t),\ t \geq 0$,
$B_1 u(1,t) = g_1(t),\ t \geq 0$.

Her er B_0 $k \times \ell$ og B_1 er $m \times \ell$.

Lax-Wendroff og Wendroff for systemer. Vi minner om Lax-Wendroff for skalar ligning $u_t + au_x = 0$

$$U_m^{n+1} = U_m^n - ap\mu_x \delta_x U_m^n + \frac{1}{2}(ap)^2 \delta_x^2 U_m^n$$

der $\mu_x u(x,t) = \frac{1}{2}(u(x + h/2, t) + u(x - h/2, t))$ er en midlingsoperator og $\delta_x u(x,t) = u(x + h/2, t) - u(x - h/2, t)$ som alltid. Spesielt merker vi oss at

$$\mu_x \delta_x U_m^n = \mu_x (U_{m+1/2}^n - U_{m-1/2}^n) = \frac{1}{2}(U_{m+1}^n + U_m^n) - \frac{1}{2}(U_m^n + U_{m-1}^n) = \frac{1}{2}(U_{m+1}^n - U_{m-1}^n).$$

Vi formulerer nå først Lax-Wendroff for et system av ligninger $u_t + Au_x = 0$ der matrisen $A \in \mathbf{R}^{\ell \times \ell}$ er konstant og $U_m^n \in \mathbf{R}^\ell$ for alle m og n

$$U_m^{n+1} = U_m^n - pA\mu_x \delta_x U_m^n + \frac{p^2}{2} A^2 \delta_x^2 U_m^n, \tag{7.25}$$

dette er en nokså åpenbar generalisering. Men vi lar nå $A = A(x,t)$ være variabel, og får

Lax-Wendroff for systemer med variabel A, $u_t + A(x,t)u_x = 0$

$$U_m^{n+1} = U_m^n - pA_m^{n+1/2}\mu_x\delta_x U_m^n + \frac{p^2}{2}A_m^{n+1/2}\delta_x\left(A_m^{n+1/2}\delta_x U_m^n\right)$$

der $A_m^{n+1/2} = A(x_m, t_n + k/2)$.

Vi ser nå på stabilitet med konstant A, og generaliserer Von Neumann-betingelsen til systemer. Vi presenterer kun oppskriften, som går ut på å sette

$$U_m^n = \mathrm{e}^{\mathrm{i}\beta x_m} G^n U^0$$

inn i differensskjemaet. Nå er $G \in \mathbf{R}^{\ell\times\ell}$ en matrise, ofte kalt amplifikasjonsmatrisen, $\beta \in \mathbf{R}$ er en vilkårlig frekvens, og $U^0 \in \mathbf{R}^\ell$ en vilkårlig vektor. Vi setter dette uttrykket inn i (7.25), og finner

$$G = I - \mathrm{i}pA\sin\theta + p^2(\cos\theta - 1)A^2, \quad \theta = \beta h. \tag{7.26}$$

Et krav til stabilitet blir nå

$$\rho(G) \le 1 \qquad \text{for alle } \theta \in \mathbf{R}.$$

Siden A er diagonaliserbar, kan vi skrive

$$A = T\Lambda T^{-1}, \quad \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_\ell)$$

Om vi benytter dette i (7.26) får vi

$$G = T\left(I - \mathrm{i}p\sin\theta\Lambda + p^2(\cos\theta - 1)\Lambda^2\right)T^{-1}$$

så matrisen T diagonaliserer ikke bare A, men også G. Dermed fins egenverdiene til G i den midterste diagonale faktoren, de er

$$\mu_j = 1 - \mathrm{i}p\sin\theta\lambda_j + p^2(\cos\theta - 1)\lambda_j^2.$$

Dermed er uttrykket for μ_j akkurat det samme som for ξ i den skalare stabilitetsanalysen av Lax-Wendroff, bare at a er skiftet ut med λ_j på høyre side. Samme analyse gjelder når vi krever $|\mu_j| \le 1$, og vi får kravet $p|\lambda_j| \le 1$ for alle j, det vil si

$$\rho(A)\,p \le 1,$$

som er vårt stabilitetskrav for Lax-Wendroff anvendt på systemer med konstant A.

Et nødvendig krav for stabilitet for varierende A er at $\rho(A(x,t))\,p \le 1$ for alle (x,t) hvor metoden brukes.

Vi ser nå på *generalisering av Wendroff's metode til systemer*. Vi minner om metoden for skalar ligning

$$(1+ap)\,U_{m+1}^{n+1} + (1-ap)\,U_m^{n+1} - (1-ap)\,U_{m+1}^n - (1+ap)\,U_m^n = 0$$

Ved å bruke foroverdifferens i rom kan vi omformulere denne som

$$\left(1 + \frac{1}{2}(1+ap)\Delta_x\right)U_m^{n+1} = \left(1 + \frac{1}{2}(1-ap)\Delta_x\right)U_m^n$$

Dermed kan vi sette opp

Wendroff's metode for systemer med variabel A, $u_t + A(x,t)u_x = 0$

$$\left(I + \frac{1}{2}(I + p\,A_{m+1/2}^{n+1/2})\Delta_x\right)U_m^{n+1} = \left(I + \frac{1}{2}(I - p\,A_{m+1/2}^{n+1/2})\Delta_x\right)U_m^n$$

der $A_{m+1/2}^{n+1/2} = A(x_m + h/2, t_n + k/2)$.

Om vi kun kjenner/ kan beregne A i gitterpunktene, kan vi uten tap av nøyaktighetsorden erstatte

$$A_{m+1/2}^{n+1/2} \quad \text{med} \quad \frac{1}{4}(A_m^n + A_{m+1}^n + A_m^{n+1} + A_{m+1}^{n+1})$$

Stabilitetsanalyse for konstant A følger samme teknikk som før, og sluttresultatet er at Wendroffs metode er stabil for alle p.

Start/randverdiproblemer for systemer av to ligninger. Bølgeligningen $\phi_{tt} = \phi_{xx}$ kan som i (7.22) med $c = 1$ omdannes til systemet

$$\begin{pmatrix} u_1 \\ u_2 \end{pmatrix}_t + \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix} \cdot \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}_x = 0$$

Her har A egenverdier $\lambda_1 = 1$ og $\lambda_2 = -1$, så en karakteristikkfamilie peker mot venstre og en mot høyre som i figuren på side 100. Dermed skal kun en randbetingelse spesifiseres. Rand/initialbetingelser for eksempel være

$$u(x,0) = f(x),\ x \ge 0$$
$$v(x,0) = g(x),\ x \ge 0$$
$$u(0,t) = \phi(t),\ t \ge 0$$

La oss for illustrasjonens del velge Lax-Wendroff som differensmetode. Vi lar

$$W_m^n = \begin{pmatrix} U_m^n \\ V_m^n \end{pmatrix}, \qquad A = \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix}.$$

Metoden blir da altså

$$W_m^{n+1} = \left(I - pA\mu_x\delta_x + \frac{p^2}{2}A^2\,\delta_x^2\right)W_m^n. \tag{7.27}$$

Anta nå at W_m^n er kjent for alle $m \ge 0$ og en $n \ge 0$. Vi kan bruke (7.27) til å finne W_m^{n+1} for $m \ge 1$. Hva med W_0^{n+1}?

$$U_0^{n+1} = \phi(t_{n+1}) \quad \text{OK}, \quad \text{problemet er } V_0^{n+1}.$$

For å bestemme V_0^{n+1} foreslår vi to alternativer

1. Fra den første av diffligningene, $u_t - v_x = 0$, finner vi

   $$v_x(0,t) = u_t(0,t) = \phi'(t).$$

   Approksimer: $v_x(0,t) = \frac{1}{h}(v(h,t) + v(0,t) + \mathcal{O}(h)$. Dermed får vi approksimasjonen

   $$V_0^{n+1} = V_1^{n+1} + h\phi'(t_{n+1}).$$

   Legg merke til at V_1^{n+1} først beregnes fra (7.27). Høyere ordens approksimasjon til den deriverte kan også benyttes.

2. Ta utgangspunkt i den andre diffligningen $v_t - u_x = 0$ og benytt boksintegrasjon som ved utledning av Wendroff's metode,

$$\int_0^h \int_{t_n}^{t_{n+1}} v_t \,\mathrm{d}t\,\mathrm{d}x = \int_{t_n}^{t_{n+1}} \int_0^h u_x \,\mathrm{d}x\,\mathrm{d}t.$$

Fra fundamentalsetningen får vi altså

$$\int_0^h (v(x,t_{n+1}) - v(x,t_n))\,\mathrm{d}x = \int_{t_n}^{t_{n+1}} (u(h,t) - u(0,t))\,\mathrm{d}t.$$

Vi approksimerer begge integralene med trapesregelen

$$\frac{h}{2}(V_0^{n+1} - V_0^n + V_1^{n+1} - V_1^n) = \frac{k}{2}(U_1^n - U_0^n + U_1^{n+1} - U_0^{n+1}).$$

Ligningen løses så med hensyn på V_0^{n+1}

$$V_0^{n+1} = -V_1^{n+1} + V_0^n + V_1^n + p(U_1^n + U_1^{n+1} - U_0^n - U_0^{n+1}).$$

7.7 Dissipasjon og dispersjon

La oss nå igjen gå tilbake til startverdiproblemet

$$u_t + au_x = 0, \quad -\infty < x < \infty, \quad t \geq 0$$
$$u(x,0) = f(x), \quad -\infty < x < \infty.$$

Fouriertransformen til initialfunksjonen $f(x)$ er gitt som

$$\hat{f}(\beta) = \frac{1}{2\pi}\int_{-\infty}^{\infty} f(x)\,\mathrm{e}^{-\mathrm{i}\beta x}\,\mathrm{d}x.$$

Den inverse transformen er

$$f(x) = \int_{-\infty}^{\infty} \hat{f}(\beta)\,\mathrm{e}^{\mathrm{i}\beta x}\,\mathrm{d}\beta.$$

Den eksakte løsningen av startverdiproblemet kan skrives

$$u(x,t) = \int_{-\infty}^{\infty} \hat{f}(\beta)\,\mathrm{e}^{\mathrm{i}\beta(x-at)}\,\mathrm{d}\beta.$$

Om vi lar $f(x) = \mathrm{e}^{\mathrm{i}\beta x}$ får vi løsningen

$$u(x,t) = \mathrm{e}^{\mathrm{i}\beta(x-at)}.$$

Løsningen u er altså en bølge med amplitude 1 og bølgelengde $\lambda = 2\pi/\beta$ som beveger seg med hastighet a langs x-aksen. La oss nå anvende en differensmetode på problemet med samme initialfunksjon. Dette gir en numerisk approksimasjon

$$U_m^n = \xi^n = \mathrm{e}^{\mathrm{i}\beta x_m},$$

slik vi har sett i forbindelse med Von Neumann-stabilitet. Her er ξ et komplekst tall som avhenger av β, slik man finner ved å substituere uttrykket ovenfor inn i differensligningen. I det som følger vil det være lurt å skrive om ξ på polar form

$$\xi = |\xi|\,\mathrm{e}^{-\mathrm{i}\varphi}.$$

La oss definere tallet α ved at $\varphi = \beta\,\alpha\,k$ der k er tidsskrittlengden. Altså blir $\alpha = \varphi/(\beta k)$. Vi finner dermed

$$\xi^n = |\xi|^n\,\mathrm{e}^{-\mathrm{i}n\varphi} = |\xi|^n\,\mathrm{e}^{-\mathrm{i}nk\,\alpha\beta} = |\xi|^n\,\mathrm{e}^{-\mathrm{i}\beta\,\alpha\,t_n}$$

og dermed kan vi sette opp følgende uttrykk for numerisk og eksakt løsning

$$U_m^n = |\xi|^n \mathrm{e}^{\mathrm{i}\,\beta(x_m - \alpha t_n)}$$
$$u_m^n = |1|^n \mathrm{e}^{\mathrm{i}\beta(x_m - at_n)}.$$

Mer generelt kan man bruke en vilkårlig startverdifunksjon $f(x)$ og får i såfall

$$U_m^n = \int_{-\infty}^{\infty} \hat{f}(\beta)\,|\xi|^n\,\mathrm{e}^{\mathrm{i}\beta(x_m - \alpha t_n)}\,\mathrm{d}\beta.$$

Tidligere har vi lært om stabilitetskravet $|\xi| \leq 1$. Ekte ulikhet svarer til

Dissipasjon. Hvis det fins en øvre grense for tidsskrittlengden $k_0 > 0$ og en konstant $\sigma > 0$ slik at

$$|\xi| \leq 1 - \sigma(\beta h)^{2s} \qquad \text{for } |\beta h| \leq \pi$$

og alle $k \leq k_0$, er differensformelen dissipativ av orden $2s$.

Dispersjon. Hvis α avhenger av β er differenseskjemaet *dispersivt*. Da transporteres ulike frekvenser med ulik hastighet. En kan godt si at dissipasjon måler amplitudefeil, mens dispersjon måler fasefeil.

Eksempel. La oss prøve ut definisjonene på Lax-Wendroff sitt skjema. Fra tidligere har vi (med $\theta = \beta h$)

$$\xi = 1 - \mathrm{i}\,ap\,\sin\theta - (ap)^2(1-\cos\theta) = 1 - 2r^2\sin^2\left(\frac{\theta}{2}\right) - \mathrm{i}\,r\,\sin\theta, \quad r = ap.$$

Fra stabilitetsanalysen for Lax-Wendroff fant vi uttrykket

$$|\xi|^2 = 1 - q\,\sin^4\left(\frac{\theta}{2}\right), \qquad q = 4r^2(1-r^2). \tag{7.28}$$

Som tidligere vist, ser vi at $|\xi| \leq 1$ hvis og bare hvis $|r| \leq 1$, men spørsmålet nå blir hvordan $|\xi|$ oppfører seg for $0 < |r| < 1$, da er $0 < q \leq 1$. La oss nå enes om at hvis x er et tall slik at $x \leq 1$ så vil

$$1 - x \leq 1 - x + \frac{1}{4}x^2 = (1 - \frac{x}{2})^2 \qquad \Rightarrow \qquad \sqrt{1-x} \leq 1 - \frac{x}{2}$$

Om vi tar kvadratroten av (7.28) og lar $x = q\,\sin^4\left(\frac{\theta}{2}\right)$ får vi

$$|\xi| \leq 1 - \frac{q}{2}\sin^4\left(\frac{\theta}{2}\right) = 1 - \frac{q}{2}\left(\frac{\sin(\theta/2)}{\theta}\right)^4\theta^4.$$

I henhold til definisjonen av dissipasjonsorden trenger vi kun å se på intervallet $-\pi \leq \theta \leq \pi$. Den minste verdien av $\sin(\theta/2))/\theta$ er $\sin(\pi/2))/\pi = 1/\pi$. Så vi får

$$|\xi| \leq 1 - \frac{q}{2}\left(\frac{1}{\pi}\right)^4\theta^4 = 1 - \frac{q}{2\pi^4}\,\theta^4, \quad |\theta| \leq \pi.$$

Vi konkluderer med at Lax-Wendroff er dissipativ av orden 4 med $\sigma = \frac{2(ap)^2(1-(ap)^2)}{\pi^4}$.
Ser vi på dispersjon, finner vi at

$$\varphi = \alpha\,\beta\,k = \arctan\left(\frac{r\sin\theta}{1-2r^2\sin^2(\theta/2)}\right).$$

Om vi løser med hensyn på α og bruker at $\beta k = \frac{\theta r}{a}$ får vi

$$\alpha = a\,\frac{1}{r\theta}\arctan\left(\frac{r\sin\theta}{1-2r^2\sin^2(\theta/2)}\right),$$

så generelt vil α definitivt avhenge av β og Lax-Wendroff er derfor dispersivt. Det er interessant å se at på stabilitetsgrensen $r = 1$ vil skjemaet ikke være dispersivt, men man får $\alpha = a$ for alle frekvenser β.

Register

F-stabilitet, 40
θ-metoden, 23
2-norm, 6

affine funksjoner, 76
amplitudefeil, 105
analytisk funksjon , 13
avhengighetsområde, 37, 93

bølgeligningen, 89
Baklengs Euler, 17
bakoverdifferens, 11
Banach-algebra, 6
basis, 75
bilineær, 86
bilineær form, 70
bivariate polynomer, 76
blokk-tridiagonal matrise, 22
blokkdiagonal, 3
boksintegrasjon, 60, 96, 104
Burgers' ligning, 32

Choleskyfaktorisering, 67
Courant-Friedrichs-Levy betingelsen, 93
Crank–Nicolson, 17, 20

diagonaldominans, 64
diagonaliserbar, 3
diagonalmatrise, 3
differensformler, 8
differensialoperator, 12
differensoperator, 11
diffligninger på bevarelsesform, 90
dimensjonsløse variable, 15
direktemetoder, 67
Dirichlet randkrav, 51, 69
diskret maksimumsprinsipp, 64
dispersjon, 104
dissipasjon, 104
divergensteorem, 56, 60, 69

egenvektoren, 3
egenverdi, 3
ekvivalente normer, 37
elementhøyreside, 80
elementstivhetsmatrise, 80
elliptiske differensialligninger, 51
endelig elementmetode, 69, 75
endelige volum-metoder, 60
enhetsoperator, 11
enskrittsmetode, 39
entydigheten, 36
Euler, 17
evolusjonsligning, 26

første ordens kvasilineær PDE, 89
fasefeil, 105
fiktive gitterlinjer, 30
forenlig matrisenorm, 6
formfunksjon, 77, 86
foroverdifferens, 11
forskyvningsoperatoren, 12
Fourieranalyse, 95
Fourierrekke, 47
Fouriers lov, 15
Frobeniusnormen, 6

Galerkins metode, 74
Gauss' divergensteorem, 60
Gauss–Seidel, 67
Gershgorins teorem, 4
gitterlignende nett, 57
glisne matriser, 22
gradientvektorfelt, 60
Greens identitet, 70, 87
gruntvannsligningene, 89

hattfunksjoner, 75
hengende noder, 63
homogene randkrav, 84
hoppe-bukk formel, 93
hyperbolske ligninger, 89
hyperbolske systemer av første ordens ligninger, 97

ikke-lineære parabolske differensialligninger, 32
implisitt metode, 19
implisitte metoder, 96
indreprodukt, 70
inhomogene randkrav, 83
innaddering, 79
iterative metoder, 67

Jacobi, 67
jordanblokkene, 3
jordanformen, 3

karakteristikker, 90
knutepunkter, 76
koersivitet, 88
konjugerte gradienters metode, 67
konserveringslov, 15, 89
konveksjonsdominerte problemer, 27
konvergens av numerisk metode, 36
konvergensbevis, 38
konvergent matrise, 7
Kronecker-delta, 4
Krylovmetodene, 67
kvadratiske elementfunksjoner, 84

Laplaceligningen, 51
Lax' ekvivalensteorem, 46
Lax-Friedrichs, 95
Lax-Wendroff, 93, 105
Lax-Wendroff og Wendroff for systemer, 101
lineære elementfunksjoner, 76
lineært system av ordinære differensialligninger, 25
linearitet av operatorer, 11
lokal avbruddsfeil, 9, 46

maksimal kolonnesum, 6
maksimal linjesum, 7
maksimumsprinsipp, 35, 52
massebevarelse, 89
matrisenorm, 5
middelverdi, 11
minimaliseringsproblemet, 74

nødvendig betingelse, 42
nedstrøms differensiering, 27
negativ definitt, 31
Neumann randkrav, 51, 55
noder, 76
norm, 5

oppstrøms differensiering, 27

parabolske differensialligninger, 15
periodiske randkrav, 47
Plancks strålingslov, 29
Poisson's ligning, 53, 69, 83
positiv definitt, 73
Positiv definitte matriser, 4
potenser av operatorer, 8, 12
prekondisjonering, 67
pyramidefunksjoner, 75

randbetingelser, 16
Rayleigh-Ritz metode, 74
referanseelement, 77
retningsderiverte, 8
Robin randkrav, 52, 55

Samarski, 28
selvadjungert, 28
selvadjungert ligning, 54
semidiskretisering, 24
sentraldifferens, 11, 76
separasjon av variable, 47
similær matrise, 31
skalar lineær PDE, 89
skrittlengde, 10
SOR (suksessiv overrelaksasjon), 67
sparse, 22
spdiags, 22
spektralradien, 6
stabilitet, 39
stabilitet av differensialligningen, 36
stabilitetsdefinisjon, 41
stabilt skjema, 66
standard indreprodukt på $\mathbf{C}^n$, 4
startbetingelser, 16
stivhetsmatrisen, 79
stykkevis lineære funksjoner, 76
symmetriske matriser, 4

taylorutvikling, 8
testfunksjon, 71
testfunksjoner, 88
Thomasalgoritmen, 22
Tikhonov, 28
tilordnet matrisenorm, 6
tilstrekkelig betingelse, 42
Toeplitzmatriser, 22
tonivåmetode, 39
trafikkflyt, 89

transponerte, 4
trapesregelen, 20, 97, 104
trekanter, 76
triangulære elementer, 76
tridiagonale matriser, 22

ubestemte koeffisienters metode, 10

variasjonsproblemet, 73
varierende skrittlengder, 57
varmeledningsligningen, 15
vektorrnorm, 5
velformet PDL-problem, 35
Von Neumann-betingelsen, 95
von Neumanns stabilitetskriterium, 47

Wendroff's metode, 96