

TMA4210 Numerisk løsning av partielle differensialligninger
med endelig differensemetoder

Brynjulf Owren

11. januar 2005

Forord

Dette notatet ble skrevet vinteren 2004 til bruk i undervisningen i den ene halvdel av kurset TMA4210 Numerisk løsning av partielle differensialligninger med differensemetoden. Det er ment som et supplement til boka *A. Iserles, A first course in the Numerical Analysis of Differential Equations*. Jeg skylder en takk til studenter som fulgte kurset dette semesteret 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 definite 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	Differensformler	8
2.2.1	Taylorutvikling	8
2.2.2	Stor \mathcal{O} -notasjon	9
2.2.3	Differensapproximasjoner til deriverte	9
2.2.4	Differensoperatører og andre operatører	11
2.2.5	Differensialoperatoren.	12
3	Diskretisering av varmeledningsligningen	15
3.1	Om utledning av varmeledningsligningen	15
3.2	Numerisk løsning av start/randverdi problemet	16
3.2.1	Numerisk approximasjon 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 paraboliske 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 $h \rightarrow 0, k \rightarrow 0$	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 skritt lengder	57
5.5	Generelt rektangulært nett	58
5.6	Diskretisering via Taylor på fullstendig generelt nett	59
5.7	Differensformler utledet via integrasjon	60
5.8	Nett basert på trekanten	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

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 diffiligningen 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 differensmetoder 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, paraboliske, 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 paraboliske 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 differensialligninger, 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 $X = [x_1, \dots, x_n]$ der hver $x_i \in \mathbf{C}^n$ kalles *egenvektoren* (*tilhørende egenverdien* λ_i). En diagonalmatrise som Λ ovenfor, skrives av og til som

$$\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n).$$

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

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

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 $\lambda_1, \dots, \lambda_n$ og et sett av reelle ortonormale egenvektorer x_1, \dots, 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 $X = [x_1, \dots, x_n]$ blir reell og ortogonal og dens inverse er derfor den transponerte

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

Diagonaliseringen er gitt som

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

2.1.3 Positiv definitte matriser

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

A (symmetrisk) er positiv semidefinit 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 semidefinit $\Leftrightarrow A$ har bare ikke-negative egenverdier, og minst en 0-egenverdi.

2.1.4 Gershgorins teorem

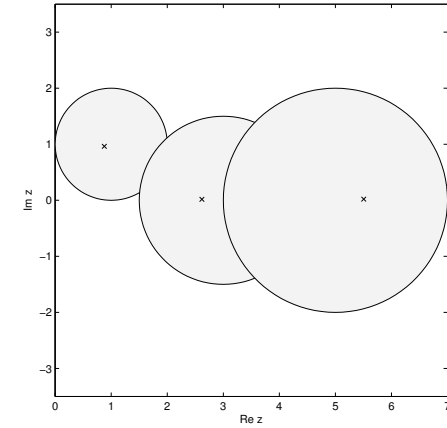
Gershgorins teorem. Gitt $A = (a_{ik}) \in \mathbf{C}^{n \times n}$. Definer n sirkelflater S_j i det komplekse plan ved

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

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 $x = [\xi_1, \dots, \xi_n]^T \neq 0$. Velg deretter ℓ blant indeksene $1, \dots, n$ slik at $|\xi_\ell| \geq |\xi_k|$, $k = 1, \dots, n$, 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

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

Altså er $\lambda \in S_\ell$.

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

2.1.5 Vektor- og matrisenormer

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

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

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

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

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. $\|\alpha A\| = |\alpha| \|A\|$ ($\alpha \in \mathbf{R}$ (\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

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

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 $\lambda_1, \dots, \lambda_n$ er egenverdiene til M , betegnes spektralradien til M for $\rho(M)$, og den er definert som

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

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)definit, 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

$$\|A\| \geq \rho(A) \quad (2.3)$$

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

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

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

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

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

Bevis.

$$\|A^k\| = \|A \cdot A^{k-1}\| \leq \|A\| \cdot \|A^{k-1}\| \leq \dots \leq \|A\|^k \rightarrow 0 \quad \text{hvis } \|A\| < 1$$

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 $A^k =$

MJM^{-1} . At $A^k \rightarrow 0$ er ekvivalent med at $J^k \rightarrow 0$. At $J^k \rightarrow 0$ er ekvivalent med at hver enkelt Jordanblokk $J_i^k \rightarrow 0$. 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 $k \rightarrow \infty$ vil $\varphi_k^{(m)}(\lambda) \rightarrow 0$ 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. \square

2.2 Differensformler

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

Taylorformel 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 $u \in C^{m+1}(\Omega)$ hvor $\Omega \subset \mathbf{R}^2$. Det er gunstig for oss å bruke operatormotasjon 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

Taylorformel med restledd for funksjoner av to variable.

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

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

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

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

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

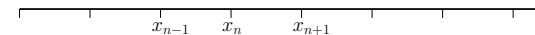
Generelt har vi at dersom $\phi(h) = \mathcal{O}(h^{p\phi})$ og $\psi(h) = \mathcal{O}(h^{p\psi})$, så gjelder at

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

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

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

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

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

$p \leq q$ er heltall, og typisk er $p \leq 0$ og $q \geq 0$.

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, \dots, 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, \dots, a_q slik at

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

Merk at vi har $q - p + 1 = j + 1$ fri parametre a_p, \dots, a_q til rådighet, og betingelsene i (2.5) gjelder for $0 \leq k \leq j$, 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 $\nu \geq j$, 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).

$$\left. \begin{array}{l} k=0 \quad a_{-1} + a_0 + a_1 = 0 \\ k=1 \quad -h a_{-1} + 0 \cdot a_0 + h a_1 = 1 \\ k=2 \quad h^2 a_{-1} + 0 \cdot a_0 + h^2 a_1 = 0 \end{array} \right\} \Rightarrow \begin{array}{l} a_{-1} = -\frac{1}{2h} \\ a_0 = 0 \\ a_1 = \frac{1}{2h} \end{array}$$

Ser vi på de første leddene i avbruddsfeilen får vi

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

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 differenseapproximasjoner er

$$m = 1 \quad \frac{u_{n+1} - u_n}{h} = u_n' + \frac{1}{2!} h u_n'' + \dots$$

$$m = 1 \quad \frac{u_n - u_{n-1}}{h} = u_n' - \frac{1}{2!} h u_n'' + \dots$$

$$m = 2 \quad \frac{u_{n+1} - 2u_n + u_{n-1}}{h^2} = u_n'' + \frac{1}{12} h^2 u_n^{(4)} + \dots$$

Hva slags orden har de tre formelene?

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: $\mu u(x) = \frac{1}{2} (u(x + \frac{h}{2}) - u(x - \frac{h}{2}))$

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 $F = \Delta$.

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

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

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

Eksempel.

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

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 differenseoperatoren.

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

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

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

2.2.5 Differensialoperatoren.

Definer

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

La tilsvarende ovenfor

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

Hvis $u(x)$ er analytisk¹ i et intervall som inneholder $x, x + h$ har vi

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

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

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

så vi setter $E = e^{hD}$.

Sammenheng mellom D og andre operatoren.

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

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 operatoren 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

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

eller

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

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

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

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

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

Videre kan vi beregne

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

¹Som 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

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

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

Spesielt finner vi som tidligere at

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

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

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

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

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

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 varmeledning ligningen

3.1 Om utledning av varmeledning ligningen

Vi skal bruke et standardeksempel for å illustrere numerisk løsning av paraboliske differensialligninger gjennom hele kurset. Det dreier seg om den lineære varmeledning ligningen 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

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

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.

$$u_t = u_{xx}, \quad x \in \mathbf{R}, \quad t > 0,$$

$$u(x, 0) = f(x), \quad x \in \mathbf{R}.$$

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.

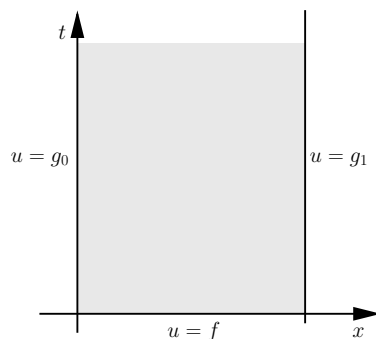
$$u_t = u_{xx}, \quad 0 < x < 1, \quad t > 0,$$

$$u(x, 0) = f(x), \quad 0 \leq x \leq 1,$$

$$u(0, t) = g_0(t), \quad t > 0,$$

$$u(1, t) = g_1(t), \quad t > 0.$$

(3.1)



3.2 Numerisk løsning av start/randverdiproblemet

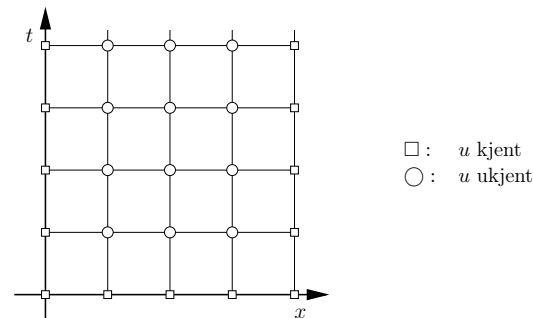
3.2.1 Numerisk approksimasjon på gitter

Vi innfører først en *skritt lengde* 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

$$x_m = mh, \quad 0 \leq m \leq M+1, \quad t_n = nk, \quad n = 0, 1, 2, \dots$$

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

$$u_m^{n+1} = u_m^n + k \partial_t u_m^n + \varphi_m^n, \quad \varphi_m^n = \frac{1}{2} k^2 \partial_t^2 u_m^n + \dots$$

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

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

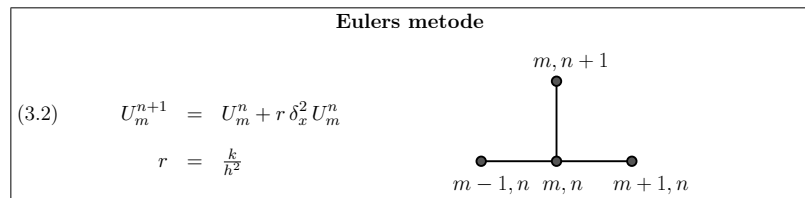
Oppsummert har vi

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

der

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

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 .



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 $U_m^0 = f(x_m)$, $m = 0, \dots, M+1$. Så setter man $n = 1$ og forsyner seg først av de oppgitte randverdiene og får $U_0^1 = g_0(k)$ og $U_{M+1}^1 = g_1(k)$. 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), m = 0, ..., M+1
for n = 0, 1, 2, ...
  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, ..., M
end

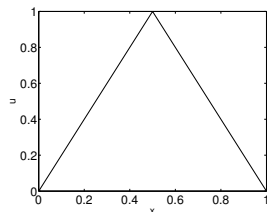
```

Eksempel.

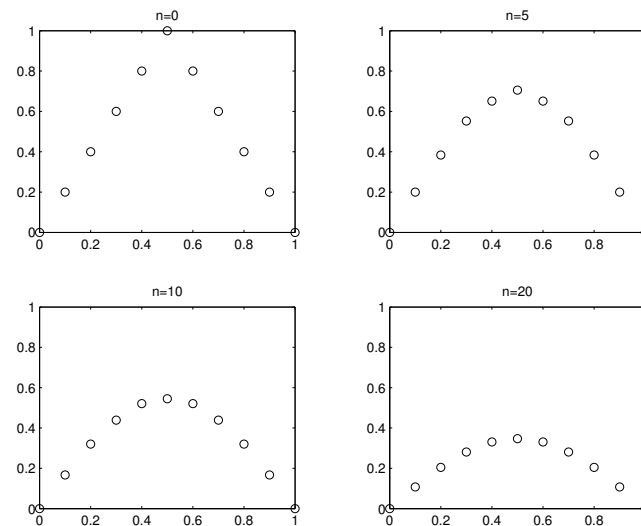
$$u_t = u_{xx} \quad 0 < x < 1, t > 0,$$

$$f(x) = \begin{cases} 2x, & 0 \leq x \leq \frac{1}{2}, \\ 2(1-x), & \frac{1}{2} < x \leq 1, \end{cases}$$

$$g_0(t) = g_1(t) = 0, \quad t > 0.$$



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 tidskritt 0, 5, 10 og 20.



Figur 3.1: Matlabsimulering av Eulers metode på varmeledningsligningen

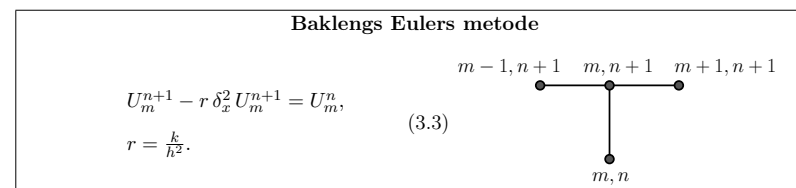
Baklengs Euler. Vi rekketviker nå istedet u_m^n omkring $x = x_m$, $t = t_{n+1}$, og får

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

der

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

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



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

lineært ligningssystem for å få ut U_m^{n+1} , $m = 1, \dots, M$. 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

$$\int_0^k f(t) dt = \frac{1}{2} k (f(0) + f(k)) - \frac{1}{12} k^3 f''\left(\frac{k}{2}\right) + \dots$$

For å utlede vår metode benytter vi den opplagte formelen

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

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

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

Vi forenkler og oppsummerer

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

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

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

$$r = \frac{k}{h^2}.$$

$$(3.4)$$

¹Akkurat 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 formelene på kompakt form

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

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

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

For venstresiden får vi komponentvis

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

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 ligningssystem 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 diffiligningen 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 glisne 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 *glisnen* 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 romskritt lengde 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 innsetningsalgoritmen (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

$$\begin{aligned} \text{(E)} \quad & \theta = 0 \\ \text{(BE)} \quad & \theta = 1 \\ \text{(CN)} \quad & \theta = \frac{1}{2} \end{aligned}$$

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

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

Vi konkluderer med at

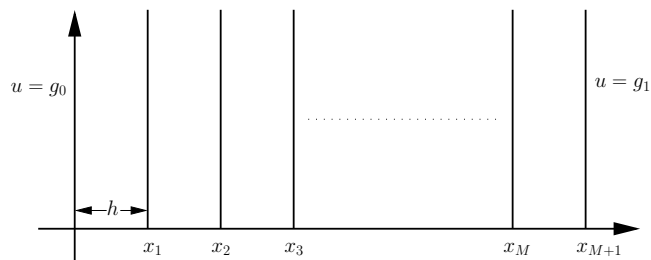
$$\begin{aligned}\tau_m^n &= \mathcal{O}(k^2 + k h^2) \quad \text{når } \theta \neq \frac{1}{2}, \\ \tau_m^n &= \mathcal{O}(k^3 + k h^2) \quad \text{når } \theta = \frac{1}{2}.\end{aligned}$$

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 $x = x_m$, $m = 0, \dots, M + 1$. Vi betrakter differensialligningen langs en slik linje, der gjelder

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

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

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

der $\dot{v}_m(t) = \frac{dv_m(t)}{dt}$. For mer kompakt notasjon, la $\mathbf{v}(t) := [v_1(t), \dots, v_M(t)]^T$. Vi får

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

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

$$\dot{\mathbf{v}} = A\mathbf{v} + \mathbf{b}(t), \quad \mathbf{v}(0) = \mathbf{v}_0 = [f(x_1), \dots, f(x_M)]^T. \quad (3.5)$$

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

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

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

$$\begin{aligned}(\mathbf{E}) \quad \text{Euler} : & \quad \mathbf{V}^{n+1} = \mathbf{V}^n + k \mathbf{F}(t_n, \mathbf{V}^n) \\ (\mathbf{BE}) \quad \text{Baklengs Euler} : & \quad \mathbf{V}^{n+1} = \mathbf{V}^n + k \mathbf{F}(t_{n+1}, \mathbf{V}^{n+1}) \\ (\mathbf{T}) \quad \text{Trapez} : & \quad \mathbf{V}^{n+1} = \mathbf{V}^n + \frac{k}{2} (\mathbf{F}(t_n, \mathbf{V}^n) + \mathbf{F}(t_{n+1}, \mathbf{V}^{n+1}))\end{aligned}$$

Setter man nå inn den spesifikke funksjonen $\mathbf{F}(t, \mathbf{v}) = A\mathbf{v} + \mathbf{b}(t)$ 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 diffiligninger enn varmeledningsligningen. For en slik ligning erstatter vi alle romderiverte med differenseapproximasjoner. 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 $\lambda_1, \dots, \lambda_M$ til A har negativ realdel og at kvotienten

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

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

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

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

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

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

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

En kan vise at med tidsskritt lengde k blir

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

der

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

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

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

der

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

Merkt 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 $[0, 1]$,

$$a(x) > 0 \text{ i } [0, 1].$$

Romdiskretisering

	Diskretisering	Avbruddsfeil
u_{xx}	$\longrightarrow \frac{1}{h^2} \delta_x^2 u$	$\mathcal{O}(h^2)$
u_x	$\longrightarrow \left\{ \begin{array}{l} \frac{1}{h} \Delta_x u \\ \frac{1}{h} \nabla_x u \end{array} \right.$	$\mathcal{O}(h)$
	$\longrightarrow \frac{1}{2h} (u(x+h) - u(x-h)) = \frac{1}{h} \delta_x u$	$\mathcal{O}(h^2)$

Vi setter altså

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

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

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

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 . $b > 0 \rightarrow \Delta$, $b < 0 \rightarrow \nabla$.

Tilfelle B. Se nå på ligningen

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

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

$$\langle u, v \rangle = \int_0^1 u(x)v(x) dx,$$

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

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

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

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

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

En metode av Tikhonov og Samarski. Skriv på bevaringsform

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

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

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

Sett nå

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

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

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

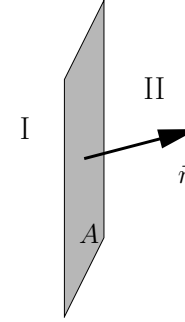
og det semidiskretiserte systemet blir

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

3.4 Randkrav med derivert

3.4.1 Ulike typer randkrav

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



Figuren illustrerer varmekraft 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.

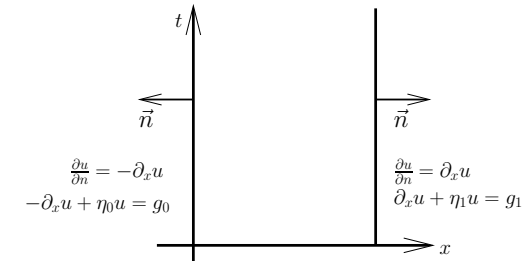
3. Stråling (Plancks strålingslov fra statistisk mekanikk)

$$-\lambda \frac{\partial u}{\partial n} = \sigma(u^4 - u_0^4).$$

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/randverdi problemet 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 $x_m = mh$, $0 \leq m \leq M$. Vi får tilsammen $M + 1$ ukjente v_0, \dots, 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 sentraddifferens, som vil involvere løsningen på den fiktive gitterlinja utenfor området vi løser ligningen på.

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

hvor

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

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

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

hvor

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

Semidiskretiseringen blir altså

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

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

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

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

$$\begin{aligned} \dot{v}_0 &= \frac{1}{h^2} \delta_x^2 v_0 = \frac{1}{h^2} (v_{-1} - 2v_0 + v_1) \\ &= \frac{1}{h^2} (-2(h\eta_0 + 1)v_0 + 2v_1) + \frac{2}{h} g_0, \\ \dot{v}_M &= \frac{1}{h^2} \delta_x^2 v_M = \frac{1}{h^2} (v_{M-1} - 2v_M + v_{M+1}) \\ &= \frac{1}{h^2} (2v_{M-1} - 2(h\eta_1 + 1)v_M) + \frac{2}{h} g_1. \end{aligned}$$

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

$$\dot{\mathbf{v}} = \frac{1}{h^2} \mathbf{Q} \mathbf{v} + \frac{2}{h} \mathbf{d}, \quad (3.8)$$

der

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

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

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

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

eller

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

²en matrise A er *negativ definit* hvis og bare hvis $(-A)$ er positiv definit

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

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

der $\tilde{\theta}_0 = \frac{1}{2}h\partial_x^2 u_0 + \dots$. 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 paraboliske 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 $v = [v_1, v_2, \dots, v_M]^T$ og $F = [F_1, F_2, \dots, F_M]^T$ 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.

$$u_t = \varepsilon u_{xx} - uu_x$$

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 sentralfdifferens, 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

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

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

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

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

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

Crank–Nicolson på ikke-lineært parabolisk problem.

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

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 sentralfdifferens 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

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

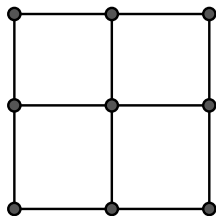
Erstatt

$$\Delta_x U_m^n \longrightarrow \frac{1}{3} (\Delta_x U_m^{n-1} + \Delta_x U_m^n + \Delta_x U_m^{n+1}),$$

$$\nabla_x U_m^n \longrightarrow \frac{1}{3} (\nabla_x U_m^{n-1} + \nabla_x U_m^n + \nabla_x U_m^{n+1}),$$

Resultatet blir formelen

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



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 flerskrittmetoder 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

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

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 $0 \leq x \leq 1$, $t \geq 0$ og som oppfyller et maksimumsprinsipp

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

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

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

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 rektanget

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

Innfør et gitter

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

der

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

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

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

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

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

Mer generelt. Sett

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

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

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

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 $h \rightarrow 0$. En har for eksempel sammenhengen

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

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 $h \rightarrow 0$ (og dermed $M \rightarrow \infty$).

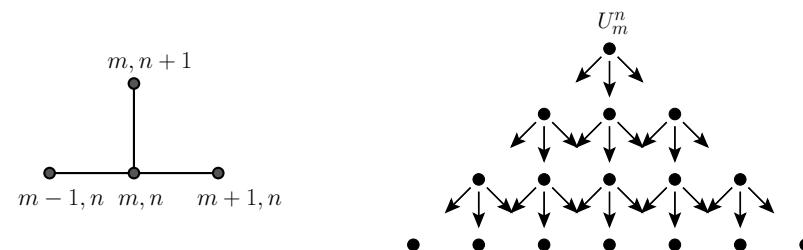
2. Den skalerte normen ovenfor, $\|\cdot\|_{2,h}$ kan en tenke seg approksimerer L_2 -normen til en underliggende kontinuerlig funksjon

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

der $f_m = f(mh)$.

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 $U_{m-n}^0, \dots, U_{m+n}^0$ hvis man løser et rent startverdiproblem (eller hvis $0 \leq m-n, m+n \leq M$). 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.

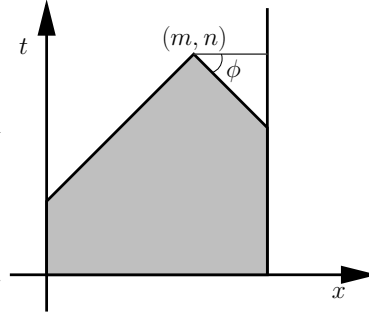
Vinkelen

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

karaktiserer 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 $h \rightarrow 0$ vil

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

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



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

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

$$(4.3) \quad \begin{aligned} U_m^{n+1} &= U_m^n + r \delta_x^2 U_m^n, \quad 1 \leq m \leq M, n \geq 0, \\ U_0^n &= g_0^n, \quad U_{M+1}^n = g_1^n, \quad n > 0, \\ U_0^m &= f_m, \quad 0 \leq m \leq M+1. \end{aligned}$$

For den eksakte løsningen gjelder

$$u_m^{n+1} = u_m^n + r \delta_x^2 u_m^n + \tau_m^n. \quad (4.4)$$

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

$$e_m^{n+1} = e_m^n + r \delta_x^2 e_m^n + \tau_m^n, \quad n > 0, 1 \leq m \leq M. \quad (4.5)$$

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

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

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

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

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 $0 \leq r \leq \frac{1}{2}$ for dette innebærer at både r og $1 - 2r$ er ikke-negative. Vi finner derfor

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

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

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

og siden $E^0 = 0$ er

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

det vil si

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

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

4.5 Stabilitet på ubegrenset tidsintervall (F -stabilitet)

La oss formulere θ -metoden (som inkluderer (E), (BE) (CN)) på matrisform. 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

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

Definerer vi vektoren $U^n = [U_1^n, \dots, U_M^n]^T$ for $n = 0, 1, \dots$ 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

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

Typisk kan differensemetoder skrives som

$$A U^{n+1} = B U^n + c^n. \quad (4.7)$$

A og B vil avhenge av h og k ved at elementene er funksjoner av h og k , men også ved at matrisesens 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

$$U^{n+1} = C U^n + q^n. \quad (4.8)$$

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

$$\|w^n\| \leq L \|w^0\| \quad \text{for enhver } w^0.$$

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

Kriterium for F -stabilitet.

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

Bevis. Anta at $\rho(C) > 1$, da fins en egenverdi λ med tilhørende egenvektor x slik at¹ $|\lambda| > 1$. Setter vi $w^0 = x$ får vi

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

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

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

$$\|w^n\| = \|C^n w^0\| \leq \|C^n\| \|w^0\| \leq L \|w^0\|.$$

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

Eksempel. Euler anvendt på (\mathbf{S}/\mathbf{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,

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

¹Det 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 $\lambda_j \geq -1$.

$$\begin{aligned} -1 &\leq 1 - 4r \sin^2 \frac{j\pi}{2(M+1)}, \quad j = 1, 2, \dots, M \\ &\Downarrow \\ r &\leq \frac{1}{2 \sin^2 \frac{j\pi}{2(M+1)}}, \quad j = 1, 2, \dots, M. \end{aligned}$$

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 $r \leq \frac{1}{2}$.

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

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

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 $n \rightarrow \infty$, men nå skjer dette ved at $k \rightarrow 0$ samtidig, slik at vi alltid har en øvre grense $T = nk$. Vi antar dessuten at $h \rightarrow 0$ 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 $w^{n+1} = C w^n$ oppfyller

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

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

$$\|C^n\| \leq L \quad \text{for alle } n \leq \frac{T}{k}, \quad (4.10)$$

der matrisenormen er tilordnet vektornormen ovenfor.

Eksempel. Hvis en bruker normen

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

bruker man

$$\|C\|_\infty = \max_k \sum_m |C_{km}| \quad \text{i (4.10).}$$

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

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

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

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

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 $\mu \geq 0$ uavhengig av h og k slik at

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

så er (4.8) stabil.

Bevis.

$$\|C^n\| \leq \|C\|^n \leq (1 + \mu k)^n \leq (1 + \mu k)^{T/k} = \left((1 + \mu k)^{1/(\mu k)} \right)^{\mu T} \leq e^{\mu T}.$$

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

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

$$\rho(C) \leq 1 + \nu k, \quad (4.11)$$

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

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

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

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

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

Vi rekkeutvikler nå denne skranken og får

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

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

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

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

Merknad. Betingelsen $\rho(C) \leq 1 + \nu k$ 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 $0 \leq n \leq M-1$. 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, $C = P\Lambda P^{-1}$.

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

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 $\|P\| (\|P^{-1}\|) \rightarrow \infty$ når $h, k \rightarrow 0$.

Viktig spesialtilfelle. Hvis C er symmetrisk er betingelsen $\rho(C) \leq 1 + \nu k$ 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

$$I - \theta r S = P(I - \theta r \Lambda)P^T,$$

$$(I + (1 - \theta) r S) = P(I + (1 - \theta) r \Lambda)P^T.$$

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 $\rho(C) \leq 1 + \nu k$ for en eller annen $\nu \geq 0$. Fra før vet vi at

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

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

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

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

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

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

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

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).

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

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 $\rho(C) \leq 1 + \nu k$. Men det viser seg at det går bra i dette tilfellet. Vi har sett at ved å innføre matrisen $D = \text{diag}(\sqrt{2}, 1, \dots, 1, \sqrt{2})$ finner man at $\tilde{Q} = D^{-1} Q D$ er symmetrisk. Dette innebærer at også $\tilde{C} = D^{-1} C D$ blir symmetrisk. Vi finner at

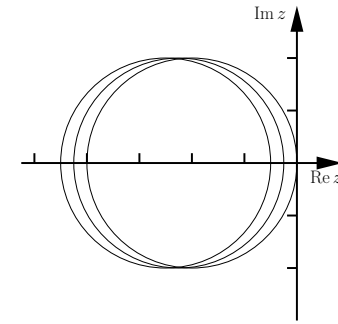
$$\|C^n\| \leq \|D\| \|D^{-1}\| \|\tilde{C}^n\|.$$

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

Matrisen C har egenverdier

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

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.



Gershgorinsirkelene 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 aksens er $-4 - 2\eta_i h$, $i = 0, 1$. Egenverdiene til Q ligger altså på den reelle aksens $-4 - 2\eta h \leq \lambda \leq 0$ der $\eta = \max\{\eta_0, \eta_1\}$. Fra (4.12) følger nå umiddelbart at $\Delta_m \leq 1$ for $0 \leq \theta \leq 1$. Kravet $\Delta_m \geq -1$ gir videre

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

som er oppfylt for $\theta \geq \frac{1}{2}$, 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 $\lambda_m \geq -4 - 2\eta h$. Til slutt får vi

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

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

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

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

Anta $R^0 = 0$. Da blir

$$\|R^n\| \leq \|C^{n-1}\| \|s^0\| + \dots + \|C\| \|s^{n-2}\| + \|s^{n-1}\|.$$

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

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

så stabilitet sikrer at avrundingsfeilen vokser høyst lineært med n .

4.8 Konsistens og Lax' ekvivalensteorem

Vi minner om (4.7)

$$AU^{n+1} = BU^n + c^n, \quad (4.13)$$

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 $\tau^n = [\tau_1^n, \dots, \tau_m^n]^T$, og har pr definisjon

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

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

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

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

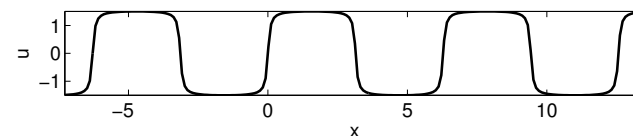
Lax' ekvivalensteorem. En konsistent differensmetode er konvergent hvis og bare hvis den er stabil.

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*

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



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

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

Ved separasjon av variable, finner en løsninger av formen

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

og bruk av initialfunksjonen $f(x)$ gir

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

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

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

En kunne nå forsøke å skrive

$$U_m^n = \sum_{\beta=-\infty}^{\infty} A_{\beta} \xi^{\beta n} e^{i\beta x_m}, \quad (4.14)$$

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

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

Innsatt i Eulers metode gir dette

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

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

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

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

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

er eksakt løsning av differensligningen. ξ_β tilsvare faktoren $e^{-\beta^2 h}$ 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-nivå 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

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

som innsatt i formelen gir

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

og dermed

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

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

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

Eksempel. La oss nå bruke differensligningen

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

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

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

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

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

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

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

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. Differensligning med konstante koeffisienter og én avhengig variabel (bare én u -komponent)
2. Rent startverdiproblem
3. 2-nivå differensformel

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.

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

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

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

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

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

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

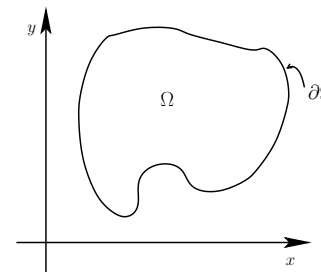
Kapittel 5

Elliptiske differensialligninger

5.1 Elliptisk ligning i planet

Vi skal nå se på partielle differensialligninger av typen

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

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

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

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

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

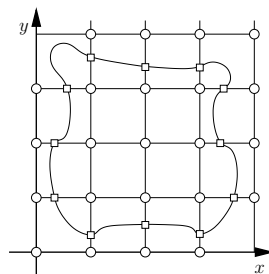
der a, b, c, d og e er funksjoner av x og y , og L er elliptisk ($ac > b^2$) i Ω .

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



○ Gitterpunkt

□ Skjæringspunkt

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

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

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

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

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

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 ligningsløserne.

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

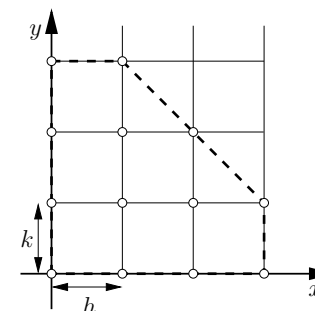
Poisson's ligning.

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

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

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

der u_p er eksakt løsning av diffiligning 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

$$\tau_p = \frac{1}{12} h^2 \partial_x^4 u_p + \frac{1}{12} k^2 \partial_y^4 u_p + \dots \quad (5.1)$$

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

$$\frac{1}{h^2} \delta_x^2 U_p + \frac{1}{k^2} \delta_y^2 U_p = f_p, \quad p \in \mathcal{N}^\circ,$$

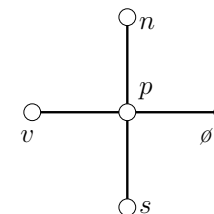
$$U_p = g_p \quad p \in D.$$

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

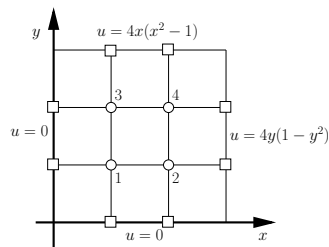
$$U_v + U_s + U_\theta + U_n - 4U_p = h^2 f_p$$



Eksempel. Vi regner gjennom et konkret tilfelle.

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

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



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

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

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 selvdjungert ligning

Vi studerer problemet

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

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

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

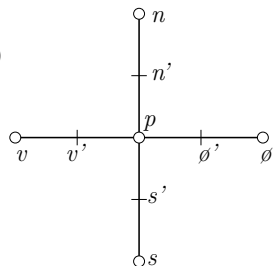
Så $Lu = f$ kan diskretiseres til

$$\alpha_{\theta} U_{\theta} + \alpha_n U_n + \alpha_v U_v + \alpha_s U_s - \alpha_p U_p = f_p,$$

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

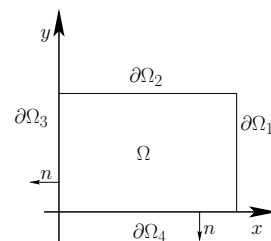
og

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



5.3 Randkrav av Neumanns og Robins type

Eksempel.



$$\Delta u = 0 \quad \text{i } \Omega$$

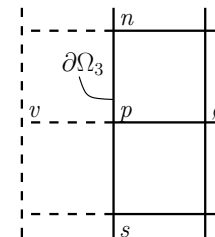
$$u = g(x, y) \quad \text{på } \partial\Omega_1 \cup \partial\Omega_2$$

$$\partial_x u - q^{(0)} u = f^{(0)} \quad \text{på } \partial\Omega_3$$

$$\partial_y u - q^{(1)} u = f^{(1)} \quad \text{på } \partial\Omega_4$$

Vi krever at $q^{(0)} \geq 0$ og $q^{(1)} \geq 0$. 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 skrittengde 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.

$$\partial_x u_p = \frac{u_{\theta} - u_p}{h} + \mathcal{O}(h) \quad \longrightarrow \quad \frac{U_{\theta} - U_p}{h} - q_p^{(0)} U_p = f_p^{(0)}$$

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.

$$\partial_x u_p = \frac{u_{\theta} - u_v}{2h} + \mathcal{O}(h^2) \quad \longrightarrow \quad \frac{U_{\theta} - U_v}{2h} - q_p^{(0)} U_p = f_p^{(0)} \quad (5.2)$$

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

$$U_{\theta} + U_n + U_v + U_s - 4U_p = 0, \quad (5.3)$$

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

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

som innsatt i (5.3) gir

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

Denne diskretiseringen er mer nøyaktig enn alternativ 1.

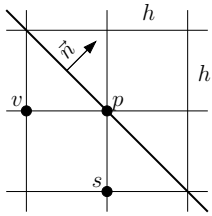
Man gjør helt tilsvarende på randen $\partial\Omega_4$. Vi ser der kun på alternativ 2 med fiktivt randpunkt s , og p på $\partial\Omega_4$:

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

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

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

Rand langs gitterdiagonal



$$\partial_n u + qu = \vec{n} \cdot \nabla u + qu = f$$

$$\vec{n} = [n_x, n_y]^T$$

$$n_x = n_y = \frac{1}{\sqrt{2}} \quad \text{fordi } k = h$$

↓

$$\partial_n u = n_x \partial_x u + n_y \partial_y u = \frac{1}{\sqrt{2}}(\partial_x u + \partial_y u)$$

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

slik at en førsteordens tilnærming blir

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

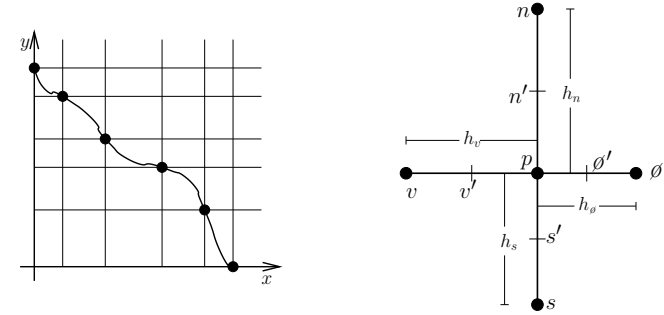
Man nøyer seg ofte med en slik grov tilnærming for denne type rand. I prinsippet kan man også her bruke alternativ 2 med fiktivt gitterpunkt, men det blir fort veldig kompliserte formler.

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

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

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 singular. 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

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

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 $p, v, v', s, s', \theta, \theta', n, n'$ og de tilhørende skrittlengder. Vi approksimerer L med

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

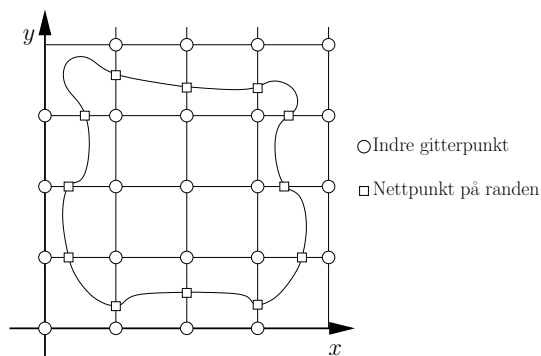
og finner ved

$$\begin{aligned} L_h^{(x)} u_p &= \frac{2}{h_v + h_\theta} \left(\left(1 + \frac{h_\theta}{2} \partial_x + \frac{h_\theta^2}{8} \partial_x^2 + \frac{h_\theta^3}{48} \partial_x^3 + \dots\right) (a_p (\partial_x + \frac{1}{24} h_\theta^2 \partial_x^3 + \dots) u_p) \right. \\ &\quad \left. - \left(1 - \frac{h_v}{2} \partial_x + \frac{h_v^2}{8} \partial_x^2 - \frac{h_v^3}{48} \partial_x^3 + \dots\right) (a_p (\partial_x + \frac{1}{24} h_v^2 \partial_x^3 + \dots) u_p) \right) \\ &= \partial_x(a\partial_x) u_p + \frac{1}{3} (h_\theta - h_v) \partial_x^2 (a\partial_x) u_p + \frac{1}{24} \frac{h_\theta^3 + h_v^3}{h_\theta + h_v} (\partial_x a \partial_x^3 + \partial_x^3 a \partial_x) u_p + \dots \end{aligned}$$

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

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

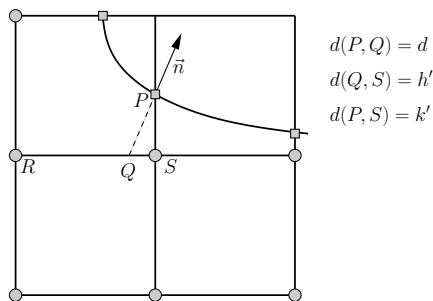
5.5 Generelt rektangulært nett



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

Dirichletproblem. Vi kan benytte den generelle 5-punktsformelen med h_θ, 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

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

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

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

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

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

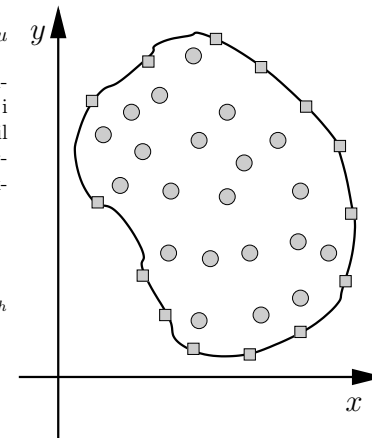
La P være en vilkårlig indre node. Vi velger ut s noder blant de øvrige punktene i nettet, typisk de s nærmeste naboene til P . La h være en karakteristisk gitterstørrelse. Vi beskriver beliggenheten til punktet Q_i relativt til P ved koordinater

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

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 $\alpha_0, \dots, \alpha_s$.



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

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

som innsatt i uttrykket for $L_h u_P$ gir

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

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

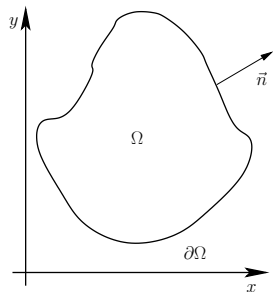
$$\sum_{i=1}^s \xi_i^\ell \eta_i^{\ell-m} \alpha_i = 0, \quad m = 0, \dots, \ell, \quad \ell = 3, 4, \dots$$

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

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

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.



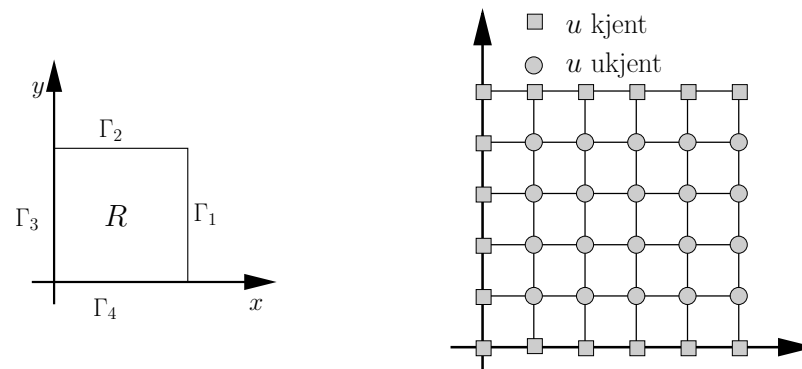
Gauss' divergensteorem. Gitt et område Ω som på figuren med rand $\partial\Omega$ som i hvert punkt har en utadrettet normalvektor \vec{n} . La $\vec{p} = \vec{p}(x, y)$ være et vektorfelt i Ω . Da gjelder

$$\int_{\Omega} \operatorname{div} \vec{p} \, dA = \oint_{\partial\Omega} \vec{p} \cdot \vec{n} \, ds$$

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

$$\int_{\Omega} \Delta u \, dA = \oint_{\partial\Omega} \nabla u \cdot \vec{n} \, ds = \oint_{\partial\Omega} \partial_n u \, ds$$

Vi illustrerer boksintegrasjon gjennom et spesifikt eksempel. La $\Omega = R$ være et rektangel med rander $\Gamma_1, \dots, \Gamma_4$ som på figuren.



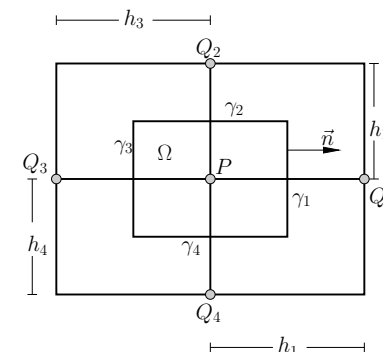
Vi ser på problemet

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

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 rektanget avgrenset av nabo-gitterlinjene 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 γ_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

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



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

$$\Delta u = f \quad \Rightarrow \quad \int_{\Omega} \Delta u \, dA = \underbrace{\oint_{\partial\Omega} \partial_n u \, ds}_I = \underbrace{\int_{\Omega} f \, dA}_{II}$$

Vi forsøker å approksimere I og II.

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

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

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

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

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

$$\sum_{i=1}^4 \int_{\gamma_i} \partial_n u \, ds = \int_{\Omega} f \, dA$$

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

$$\int_{\gamma_1} \partial_n u \, ds = \int_{\gamma_1} (d - au) \, ds \approx \ell_1 (d_P - a_P U_P)$$

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

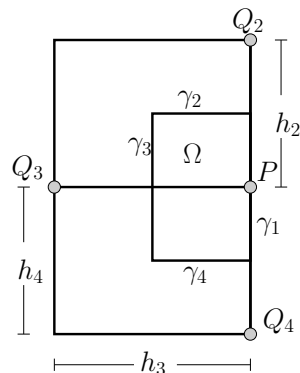
$$\int_{\gamma_i} \partial_n u \, ds \approx \ell_i \frac{u_{Q_i} - u_P}{h_i}, \quad i = 2, 3, 4.$$

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 PQ_i . Vi lar γ_i har lengde ℓ_i og linjestykkene PQ_i har lengde h_i . Vi antar videre at Ω har areal A . Vi bruker Gauss' divergensteorem på området Ω , og får

$$\sum_{i=1}^s \int_{\gamma_i} \partial_n u \, ds = \int_{\Omega} f \, dA$$

Nå approksimerer vi som før

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

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.

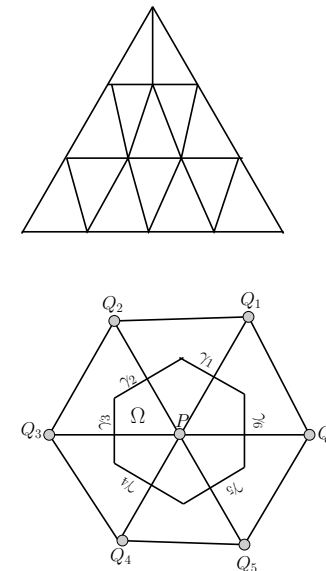
5.9 Differensligningene

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

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



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

Ønskelige egenskaper for (5.4).

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

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

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

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

Avender vi derfor antagelsen i teoremet for dette punktet får vi

$$\begin{aligned} \alpha_0 V_{P^*} &\leq \sum_i \alpha_i V_{Q_i} \\ &\Updownarrow \text{(i)} \\ V_{P^*} &\leq \sum_i \frac{\alpha_i}{\alpha_0} V_{Q_i} \stackrel{\text{(ii)}}{\leq} \frac{1}{\sum_j \alpha_j} \sum_i \alpha_i V_{Q_i} = \sum_i \gamma_i V_{Q_i} \end{aligned}$$

hvor

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

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 selvmotsgivelse (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

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

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

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

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

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

Her er

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

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

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

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 $e_p, \tau_p, p \in R$ i vektorer \mathbf{e} og $\boldsymbol{\tau}$, kan vi skrive om systemet på formen

$$A\mathbf{e} = \boldsymbol{\tau}$$

Om A er inverterbar fås

$$\mathbf{e} = A^{-1} \boldsymbol{\tau}$$

som impliserer

$$\|\mathbf{e}\|_\infty \leq \|A^{-1}\|_\infty \cdot \|\boldsymbol{\tau}\|_\infty$$

Stabilitet: Skjemaet sies å være *stabilt* dersom det fins en konstant C slik at

$$\|A^{-1}\| \leq C, \quad \text{for alle skrittlengder } h.$$

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

$$\|\mathbf{e}\|_\infty = \mathcal{O}(h^\sigma)$$

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 $\|\mathbf{e}\|_\infty$ ikke kan forventes å være mer enn $\mathcal{O}(h)$. Det hender likevel at en under slike omstendigheter får $\|\mathbf{e}\|_\infty = \mathcal{O}(h^2)$.

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 Gausseliminering, eller mer spesifikt, Choleskyfaktoriserings dersom ligningssystemet er symmetrisk og positiv definit. De iterative metodene inkluderer Jacobi, Gauss-Seidel, og SOR (suksessiv overrelaksasjon). Men den typen lineære iterative ligningsløserne som har hatt mest suksess i de senere 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 prekonisjonering av systemet. Teoretisk betyr dette at man kan finne matriser T, S , slik at systemet

$$\begin{array}{l} T^{-1}AS \quad S^{-1}x = T^{-1}b \\ \hat{A} \quad \hat{x} = \hat{b} \end{array}$$

er "enkler" 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 løserne omtrent like effektive, mens i tre romdimensjoner "vinner" de iterative løserne. Typisk er det slik at iterative løserne bruker multiplikasjon av matrisen A (resp \hat{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 Gausseliminering relativt sett mye billigere. Faktorene L og U i LU -faktoriserings 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 $\hat{A} = T^{-1}AS$ er kun teoretisk. For enkelhets skyld, la oss sette $S = I$. Som nevnt ovenfor bygger de iterative metodene på operasjoner av typen $y = Ax$ for vilkårlige vektorer x . For det prekondisjonerte systemet blir dette til $y = \hat{A}x = T^{-1}Ax$. Prekondisjoneringen foregår dermed som en “indre løkke” i den iterative metoden. Hver gang en slik metode skal beregne $\hat{A}x$ 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 Gausseliminering på systemet $Ay = \tilde{y}$. En slik delvis eller inkomplett Gausseliminering 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 differensialligningen løses i mange små delområder. Dermed splittes differensialligningene 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 Gausseliminering på hver sin prosessor.

Det finnes et eget kurs *TMA4205 Numerisk lineær algebra* som behandler slike metoder i detalj.

Register

- F -stabilitet, 40
- θ -metoden, 23
- 2-norm, 6
- analytisk funksjon, 13
- avhengighetsområde, 37
- Baklengs Euler, 17
- bakoverdifferens, 11
- Banach-algebra, 6
- blokk-tridiagonal matrise, 22
- blokkdiagonal, 3
- boksintegrasjon, 60
- Burgers’ ligning, 32
- Choleskyfaktorisering, 67
- Crank–Nicolson, 17, 20
- diagonaldominans, 64
- diagonaliserbar, 3
- diagonalmatrise, 3
- differensformler, 8
- differensialoperatoren, 12
- differensoperatorer, 11
- dimensjonsløse variable, 15
- direktemetoder, 67
- Dirichlet randkrav, 51
- diskret maksimumsprinsipp, 64
- divergensteorem, 56, 60
- egenvektoren, 3
- egenverdi, 3
- ekvivalente normer, 37
- elliptiske differensialligninger, 51
- endelige volum-metoder, 60
- enhetsoperator, 11
- enskrittmetode, 39
- entydigheten, 36
- Euler, 17
- evolusjonsligning, 26
- fiktive gitterlinjer, 30
- forenlig matrisenorm, 6
- foroverdifferens, 11
- forskyvningsoperatoren, 12
- Fourierrekke, 47
- Fouriers lov, 15
- Frobeniusnormen, 6
- Gauss’ divergensteorem, 60
- Gauss–Seidel, 67
- Gershgorins teorem, 4
- gitterlignende nett, 57
- glisne matriser, 22
- gradientvektorfelt, 60
- hengende noder, 63
- ikke-lineære paraboliske differensialligninger, 32
- implisitt metode, 19
- iterative metoder, 67
- Jacobi, 67
- jordanblokkene, 3
- jordanformen, 3
- konjugerte gradienters metode, 67
- konserveringslov, 15
- konveksjonsdominerte problemer, 27
- konvergens av numerisk metode, 36
- konvergensbevis, 38
- konvergent matrise, 7
- Kronecker-delta, 4
- Krylovmetodene, 67
- Laplaceligningen, 51
- Lax’ ekvivalensteorem, 46
- lineært system av ordinære differensialligninger, 25
- linearitet av operatorer, 11
- lokal avbruddsfeil, 9, 46
- maksimal kolonnesum, 6
- maksimal linjesum, 7
- maksimumsprinsipp, 35

- maksimumsprinsippet, 52
- matrisenorm, 5
- middelverdi, 11

- nødvendig betingelse, 42
- nedstrøms differensiering, 27
- negativ definit, 31
- Neumann randkrav, 51, 55
- norm, 5

- oppstrøms differensiering, 27

- parabolske differensialligninger, 15
- periodiske randkrav, 47
- Plancks strålingslov, 29
- Poisson's ligning, 53
- Positiv definte matriser, 4
- potenser av operatører, 8, 12
- prekondisjonering, 67

- randbetingelser, 16
- retningsderiverte, 8
- Robin randkrav, 52, 55

- Samarski, 28
- selvadjungert, 28
- selvadjungert ligning, 54
- semidiskretisering, 24
- sentraldifferens, 11
- separasjon av variable, 47
- similær matrise, 31
- skritt lengde, 10
- SOR (suksessiv overrelaksasjon), 67
- sparse, 22
- spdiags, 22
- spektralradien, 6
- stabilitet, 39
- stabilitet av differensialligningen, 36
- stabilitetsdefinisjon, 41
- standard indreprodukt på \mathbf{C}^n , 4
- startbetingelser, 16
- statilt skjema, 66
- symmetriske matriser, 4

- taylorutvikling, 8
- Thomasalgoritmen, 22
- Tikhonov, 28
- tilordnet matrisenorm, 6
- tilstrekkelig betingelse, 42
- Toeplitzmatriser, 22
- tonivåmetode, 39

- transponerte, 4
- trapesregelen, 20
- tridiagonale matriser, 22

- ubestemte koeffisienters metode, 10

- varierende skritt lengder, 57
- varmeledningsligningen, 15
- vektornorm, 5
- velformet PDL-problem, 35
- von Neumanns stabilitetskriterium, 47