Matematisk modellering og numeriske metoder

Lektion 13 og 14

Morten Grud Rasmussen

28. oktober, 2013

1 Numeriske metoder til løsning af differentialligninger

1.1 Bevarelseslove

[Peiró-Sherwin-noterne]
I det følgende vil vi skrive p for et punkt på linjen, i planen eller i rummet. Da de følgende overvejelser gør sig gældende i alle tilfælde vil vi således ikke specificere, om $p = x$, $p = (x, y)$ eller $p = (x, y, z)$. Mht. sprogbrug vil vi dog bruge den rummelige betegnelse, således at eksempelvis "kube" skal forstås som "rektangel" i to dimensioner og "interval" i en dimension. Ligeledes vil vi skrive $\mathrm{d}V$ (som i "volumen"), når vi integrerer, ligegyldigt om vi er i en, to eller tre dimensioner, og $\mathrm{d}A$ når vi integrerer "overfladen" (som altså i én dimension reducerer til en sum over endepunkterne).

Som I husker fra udledningen af varmligningen, så udnyttede vi, at den samlede termiske energi var *bevaret*, således at en tidsændring i temperaturen i en kube $-\frac{\partial}{\partial t}\int_T u(p,t)\,\mathrm{d}V$ kunne tilskrives, hvad der susede gennem kubens sider $\int_A f(u)\cdot n\,\mathrm{d}A$ (hvor A er overfladen af kuben T, f står for "flux" og n er en ydre normalvektor af længde 1):

$$\frac{\partial}{\partial t}\int_T u(p,t)\,\mathrm{d}V + \int_A f(u)\cdot n\,\mathrm{d}A = 0.$$

Som sagt gjaldt dette under antagelse af, at den samlede termiske energi var *bevaret*, og ovenstående er derfor også et eksempel på en såkaldt *bevarelseslov*. Bemærk, at formuleret som ovenfor behøver u ikke være temperatur, men kan være en hvilken som helst størrelse, som hører under, hvad man kalder kontinuummekanik. Hvad sker der så, hvis vi har en model, hvor vi rent faktisk ønsker at modellere tilførsel eller fjernelse af termisk energi (eksempelvis hvis der et sted i rummet foregik en endo- eller eksotermisk kemisk reaktion)? Lad nu S betegne den tilførte energi (S kunne stå for "source") i et givet punkt i rummet, hvor S eksempelvis også kan afhænge af tid t og temperatur u. Ovenstående bevarelseslov skal da modificeres til følgende:

$$\frac{\partial}{\partial t}\int_T u(p,t)\,\mathrm{d}V + \int_A f(u)\cdot n\,\mathrm{d}A - \int_T S(p,u(p),t) = 0,$$

som også kaldes en *bevarelseslov*, da den igen bygger på bevarelsesprincipper. Bemærk igen, at ovenstående kunne være en bevarelseslov for en hvilket som helst kontinuummekanisk størrelse. Som i tilfældet med varmeligningen, kunne vi nu anvende Gauss' sætning og få

$$\int_T \Big(\frac{\partial u}{\partial t} + \operatorname{div} f(u) - S \Big) \, \mathrm{d}V = 0.$$

Med dette integral kunne vi så argumentere for, at da kuben er vilkårligt valgt, så må integranten også være 0, hvis den ellers er kontinuert:

$$\frac{\partial u}{\partial t} + \operatorname{div} f(u) - S = 0$$

(prøv at overvej, hvad de enkelte led kan tolkes som – og husk, at $f(u)$ altså er fluxet af u, ikke blot u). Dette kaldes den *stærke* udgave af bevarelsesloven. Bemærk, at dette krævede, at integranten var kontinuert, og vi går altså eksempelvis glip af de *generaliserede løsninger*, som vi talte om i lektion 10. Frygt ej! Dette kan afhjælpes. I første omgang konstaterer vi, at integralformen af bevarelsesloven ikke kræver kontinuitet, og den er altså en svagere form (deraf navnet *den stærke udgave*), men vi kan klare os med endnu mindre! I første omgang konstaterer vi, at

$$\int_T \Big(\frac{\partial u}{\partial t} + \operatorname{div} f(u) - S \Big) \, \mathrm{d}V = \int_{\mathbb{R}^d} \Big(\frac{\partial u}{\partial t} + \operatorname{div} f(u) - S \Big) 1_T \, \mathrm{d}V = 0,$$

hvor d er dimensionen og 1_T er den såkaldte *indikatorfunktion* for T, som er givet ved

$$1_T(p) = \begin{cases} 1 & \text{hvis} \quad (p) \in T \\ 0 & \text{ellers} \end{cases}.$$

Den integrale form af bevarelsesloven, som jo skal gælde for alle kuber T, kan altså tolkes som, at vi "tester" udtrykket ($\frac{\partial u}{\partial t} + \operatorname{div} f(u) - S$)'s gennemsnit henover alle tænkelige små bokse. Vi har naturligvis blot valgt kuber af bekvemmelighedsgrunde, men vi kunne lige så vel have valgt eksempelvis kugler. Og her kommer så en ny idé, som ændrer spillereglerne: vi kunne også bruge en familie af *glatte testfunktioner*, hvor "glat" betyder, at alle partielt afledede af alle ordner eksisterer, i stedet for funktioner af typen 1_T! Vi får i givet fald udtrykket

$$\int_{\mathbb{R}^d} \Big(\frac{\partial u}{\partial t} + \operatorname{div} f(u) - S \Big) w \, \mathrm{d}V = 0,$$

hvor w er en funktion af p. Her står w for "weight" (funktionen kaldes også en vægtfunktion). Ved at anvende den rumlige pendant til partiel integration[1], fås

$$\int_{\mathbb{R}^d} \left(\Big(\frac{\partial u}{\partial t} - S \Big) w - f(u) \cdot \nabla w \right) \mathrm{d}V + \int_A f \cdot w \, \mathrm{d}A = 0. \tag{1}$$

Vi bemærker, at vi har "flyttet" en stedafledet fra $f(u)$ til w (som jo er glat), og sidste udtryk kræver derfor mindre af f. Dvs., at hvis (1) gælder for alle w i en klasse af glatte testfunktioner, så er dette et svagere krav end de to andre formuleringer, og (1) kaldes derfor for den *svage* udgave af bevarelsesloven.

[1] Den rumlige pendant til partiel integration er i familie med Gauss' divergenssætning!

1.2 Numeriske overvejelser – punktvis repræsentation

Mange PDE'er kan ikke løses eksakt, og vi er nødsaget til at lave numeriske approksimationer. For at foretage numeriske approksimationer er vi nødt til at repræsentere de uendeligt mange tal, som indgår i en fuldstændig beskrivelse af en kontinuert løsning, vha. endeligt mange tal.

Antag i første omgang at vi har en kontinuert funktion u af én variabel x. Da u er kontinuert, er $u(x)$ tæt på $u(x_0)$, såfremt x er tæt på x_0. Udnytter vi denne tankegang, skulle vi få en nogenlunde repræsentation af u, hvis vi med små mellemrum gennem hele u's definitionsmængde udvælger nogle værdier x_j, $j = 0, \ldots, J$, hvad enten der er jævne mellemrum (altså fast afstand) mellem x_j'erne eller ej, og tilnærmer u ud fra u's værdier i disse faste punkter. Vi vælger u_j for $j = 0, \ldots, J$ så $u_j \approx u(x_j)$, og forsøger at beskrive u vha. x_j'erne og disse u_j'er. Vi kalder x_i'erne for en *maske* eller et *net* (på engelsk bruges betegnelserne *grid* eller *mesh*).

1.3 Finite Difference-metoden

Den simpleste (og ældste) metode til numerisk løsning (hvor "numerisk løsning" selvfølgelig betyder numerisk approksimation af den "rigtige" (analytiske) løsning) er finite difference-metoden. Dens byggeklodser består af *differenstilnærmelser* af afledede. Den grundlæggende idé er, at afledede, ordinære såvel som partielle og af vilkårlig orden, kan approksimeres ved forskellige former for *differenskvotienter*. Vi vil om lidt gennemgå nogle af de mest grundlæggende. For at lette forståelsen, vil vi illustrere begreberne med et konkret eksempel. Vi betragter derfor vores allesammens endimensionelle varmeligning med $c^2 = 1$, $x \in [0, 1]$ (altså $L = 1$), $t \in [0, T]$, Dirichlet-randbetingelser og begyndelsesbetingelse f:

$$\begin{aligned} u_t &= u_{xx} && \text{(varmeligningen)} \\ u(0,t) &= u(1,t) = 0 && \text{(randbetingelsen)} \\ u(x,0) &= f(x) && \text{(begyndelsesbetingelsen)} \end{aligned}$$

Først skal vi vælge en maske i både rum og tid, og disse vil vi vælge såkaldt *uniforme*, dvs. $x_{j+1} - x_j = h$ for alle $j = 0, \ldots, J-1$ og $t_{n+1} - t_n = k$ for alle $n = 0, \ldots, N-1$. Vi vælger i øvrigt $t_0 = 0$, $t_N = T$, $x_0 = 0$ og $x_J = 1$, således at endepunkterne af vores tids- og steddomæne er inkluderet i masken. Løsningen u vil så blive approksimeret med $u_j^n \approx u(x_j, t_n)$. Bemærk, at n'et i u_j^n *ikke* er en potens, men et index, som angiver det aktuelle tidsskridt!

Forskellige førsteordens differenskvotienter

I varmeligningen optræder som bekendt en førsteordens tidsafledet. Vi skal approksimere en tidsafledet ud fra u_j^n'erne (husk, at j relaterer sig til x-variablen, mens n relaterer sig til t-variablen). De tre mest grundlæggende måder at gøre det på er de følgende udtryk:

$$\begin{aligned} u_t(x_j, t_n) &\approx \frac{u_j^{n+1} - u_j^n}{k} && \text{(\textit{forward difference})} \\ u_t(x_j, t_n) &\approx \frac{u_j^n - u_j^{n-1}}{k} && \text{(\textit{backward difference})} \\ u_t(x_j, t_n) &\approx \frac{u_j^{n+1} - u_j^{n-1}}{2k} && \text{(\textit{central difference})} \end{aligned}$$

som hver især har deres fordele og ulemper, både hvad angår implementering og hvad angår præcision. Ideen bag dem alle er imidlertid den samme: brøkerne angiver hældningen mellem de indgående værdier af u_j^m, $m \in \{n-1, n, n+1\}$.

Den centrale andenordens differenskvotient

Vi vil nu finde en approksimation af den andenordens stedafledede, u_{xx}, som indgår i varmeligningen. Vi skal altså finde en differenskvotient mellem to differenskvotienter, eksempelvis

$$u_{xx}(x_j, t_n) \approx \frac{\frac{u_{j+1}^n - u_j^n}{h} - \frac{u_j^n - u_{j-1}^n}{h}}{h} = \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{h^2},$$

som kaldes den *centrale andenordens differenskvotient*.

En eksplicit metode

Vi vil nu gennemgå en finite difference-metode til numerisk løsning af varmeligningen, som kaldes en *eksplicit metode*, fordi man – med udgangspunkt i begyndelsesbetingelserne – for hvert tidsskridt har et eksplicit udtryk for enhvert punkt i det næste tidsskridt. Metoden er kendt som FTCS-metoden, som står for *forward in time, central in space*, hvilket gerne skulle give mening ved nærmere inspektion af nedenstående. Idéen er følgende: idet

$$u_t = u_{xx}, \qquad \text{mens} \qquad u_t(x_j, t_n) \approx \frac{u_j^{n+1} - u_j^n}{k} \qquad \text{og} \qquad u_{xx}(x_j, t_n) \approx \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{h^2},$$

så er

$$\frac{u_j^{n+1} - u_j^n}{k} \approx \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{h^2}, \tag{2}$$

hvor approksimationen er bedst for små h og meget små k. Husker vi nu, at u_j^n blot skulle være *approksimationer* af $u(x_j, t_n)$, så kan vi simpelthen tillade os *definere* vores u_j^n som løsningen til ovenstående ligninger for alle kombinationer af følgende valg af j og n:

$$j \in \{1, \ldots, J-1\} \qquad \text{og} \qquad n \in \{1, \ldots, N\}.$$

De sidste tilfælde bestemmes ud fra begyndelses- og randbetingelserne. For begyndelsesbetingelsernes vedkommende er det oplagt at gøre følgende:

$$u_j^{t_0} = u_j^0 = f(x_j) \qquad \text{for} \qquad j \in \{0, \ldots, J\},$$

mens randbetingelserne giver

$$u_0^n = u_J^n = 0 \qquad \text{for} \qquad n \in \{0, \ldots, N\},$$

hvor vi selvfølgelig for at kunne opfylde begyndelses- og randbetingelser samtidigt er nødt til at antage, at $f(0) = f(1) = 0$, hvormed u_0^0 og u_J^0, som indgår begge steder, altså er 0, uanset hvilken definition, vi vælger.

Som sagt er der tale om en *eksplicit* metode, hvilket betyder, at vi ikke behøver løse nogen ligningssystemer for at nå til næste tidsskridt, vi skal blot isolere u_j^{n+1} i (2):

$$u_j^{n+1} = (1 - 2\frac{k}{h^2})u_j^n + \frac{k}{h^2}(u_{j-1}^n + u_{j+1}^n) = (1-2r)u_j^n + r(u_{j-1}^n + u_{j+1}^n), \qquad \text{hvor} \qquad r = \frac{k}{h^2},$$

som altså angiver u_j^{n+1} som en funktion af u_{j-1}^n, u_j^n og u_{j+1}^n, som alle tilhører et tidligere tidsskridt.

Vi vil nu skrive det hele op i matrixform:

$$\begin{bmatrix} u_0^0 \\ u_1^0 \\ u_2^0 \\ \vdots \\ u_j^0 \\ \vdots \\ u_{J-2}^0 \\ u_{J-1}^0 \\ u_J^0 \end{bmatrix} = \begin{bmatrix} f(x_0) \\ f(x_1) \\ f(x_2) \\ \vdots \\ f(x_j) \\ \vdots \\ f(x_{J-2}) \\ f(x_{J-1}) \\ f(x_J) \end{bmatrix}$$

og

$$\begin{bmatrix} u_0^{n+1} \\ u_1^{n+1} \\ u_2^{n+1} \\ \vdots \\ u_j^{n+1} \\ \vdots \\ u_{J-2}^{n+1} \\ u_{J-1}^{n+1} \\ u_J^{n+1} \end{bmatrix} = \begin{bmatrix} 1 & 0 & & & \cdots & & & & 0 \\ r & 1-2r & r & 0 & & \cdots & & & 0 \\ 0 & r & 1-2r & r & 0 & & \cdots & & 0 \\ & & & & & & & & \\ \vdots & & \ddots & \ddots & \ddots & & & & \vdots \\ & & & & & & & & \\ 0 & & \cdots & & 0 & r & 1-2r & r & 0 \\ 0 & & & \cdots & & 0 & r & 1-2r & r \\ 0 & & & & \cdots & & & 0 & 1 \end{bmatrix} \begin{bmatrix} u_0^n \\ u_1^n \\ u_2^n \\ \vdots \\ u_j^n \\ \vdots \\ u_{J-2}^n \\ u_{J-1}^n \\ u_J^n \end{bmatrix} \qquad \text{for} \qquad n \geq 0,$$

eller kort $u^0 = f$ og $u^n = Au^{n-1} = A^n u^0$ for $n \geq 1$, for passende definition af u^n, A og f. Som det ses, er der ikke nogen ligninger, som skal løses, man skal blot regne matrixprodukter ud, hvilket er årsagen til betegnelsen *eksplicit* metode. Denne eksplicitte metode kan vises at være, hvad man kalder *numerisk stabil* og *konvergent*, når $r = \frac{k}{h^2} \leq \frac{1}{2}$. De numeriske fejl er af en størrelsesorden, som maksimalt er proportional med tidsskridtet k og kvadratet på det rumlige skridt h, hvilket også skrives $E = O(k) + O(h^2)$.

En implicit metode

Vi vil nu gennemgå en anden metode, som er en *implicit metode*, idet man efter at have opstillet modellen, skal løse nogle ligninger for at nå til resultatet. Hvor ovenstående metode kaldtes FTCS-metoden, går denne metode under betegnelsen BTCS-metoden: "backwards in space, central in space." Som dette navn antyder, er årsagen til forskellen på de to metoder, at man tager udgangspunkt i *backward difference*-approksimationen

$$u_t(x_j, t_n) \approx \frac{u_j^n - u_j^{n-1}}{k}$$

fremfor *forward difference*-approksimationen for den tidsafledede. Dette giver med samme argument som før ligningssystemet

$$\frac{u_j^{n+1} - u_j^n}{k} = \frac{u_{j+1}^{n+1} - 2u_j^{n+1} + u_{j-1}^{n+1}}{h^2}.$$

Bemærk, at vi her ikke kan finde u_j^{n+1} alene ud fra information om tidligere tidsskridt; i ligningen indgår også u_{j+1}^{n+1} og u_{j-1}^{n+1} som også er ukendte, og vi ender i stedet med et *system af lineære ligninger*:

$$(1+2r)u_j^{n+1} - ru_{j-1}^{n+1} - ru_{j+1}^{n+1} = u_j^n \qquad \text{hvor} \qquad r = \frac{k}{h^2},$$

eller når vi også tager hensyn til begyndelses- og randbetingelser og skriver det i matrixform:

$$\begin{bmatrix} u_0^0 \\ u_1^0 \\ u_2^0 \\ \vdots \\ u_j^0 \\ \vdots \\ u_{J-2}^0 \\ u_{J-1}^0 \\ u_J^0 \end{bmatrix} = \begin{bmatrix} f(x_0) \\ f(x_1) \\ f(x_2) \\ \vdots \\ f(x_j) \\ \vdots \\ f(x_{J-2}) \\ f(x_{J-1}) \\ f(x_J) \end{bmatrix}$$

og

$$\begin{bmatrix} 1 & 0 & & & \cdots & & & 0 \\ -r & 1+2r & -r & 0 & \cdots & & & 0 \\ 0 & -r & 1+2r & -r & 0 & \cdots & & 0 \\ \vdots & & \ddots & \ddots & \ddots & & & \vdots \\ 0 & & \cdots & & 0 & -r & 1+2r & -r & 0 \\ 0 & & & \cdots & & 0 & -r & 1+2r & -r \\ 0 & & & & \cdots & & & 0 & 1 \end{bmatrix} \begin{bmatrix} u_0^{n+1} \\ u_1^{n+1} \\ u_2^{n+1} \\ \vdots \\ u_j^{n+1} \\ \vdots \\ u_{J-2}^{n+1} \\ u_{J-1}^{n+1} \\ u_J^{n+1} \end{bmatrix} = \begin{bmatrix} u_0^n \\ u_1^n \\ u_2^n \\ \vdots \\ u_j^n \\ \vdots \\ u_{J-2}^n \\ u_{J-1}^n \\ u_J^n \end{bmatrix} \qquad \text{for} \qquad n \geq 0,$$

eller mere kompakt $u^0 = f$ og $u^n = Au^{n+1}$ for passende definition af u^n, f og A. Bemærk her, at vi ikke som før, hvor $u^n = Au^{n-1} = A^n u^0$, kan skrive u^{n+1} direkte på en itereret form, men i stedet skal løse *ligningssystemet* $u^n = Au^{n+1}$. Dette gør metoden mere beregningstung, men belønningen kommer så i form af, at metoden *altid* er numerisk stabil (uanset r), og selvom vi stadig har at fejlen opfører sig som for den eksplicitte metode, $E = O(k) + O(h^2)$, så giver metoden bedre resultater, specielt for store tidsskridt.

1.4 Numeriske overvejelser – elementvis repræsentation

Lad os antage, at vi stadig vil repræsentere en funktion vha. endeligt mange tal, men i stedet for at angive tilnæmelser til funktionen i en given mængde punkter, så deler vi funktionen op i *elementer*[2], hvor vi antager, at hvert element af funktionen på tilfredsstillende vis kan beskrives tilnæmelsesvist ved en funktion af en simplere type, eksempelvis et førstegradspolynomium.

[2]Der er i litteraturen udbred uenighed om, om det er definitionsmængden eller funktionen, som er opdelt i "elementer." Vi forholder os neutrale.

Antag i første omgang, at vi har en kontinuert funktion u af én variabel. Da u er kontinuert, vil vi også gerne have, at tilnærmelsen er kontinuert. Vi deler nu u's definitionsmængde op i en række delintervaller ("elementer"), som kan skrives $[x_{i-1}, x_i]$ for passende valg af x_i, $i = 0, \ldots, N$. På hvert delinterval tilnærmer vi nu u med et førstegradspolynomium (vi kunne også have valgt højeregradspolynomier eller en helt anden klasse af funktioner), men på en sådan måde, at tilnærmelsen stadig er *kontinuert*, dvs. endepunktet for et polynomium på $[x_{i-1}, x_i]$ skal have samme værdi som begyndelsespunktet for et polynomium på $[x_i, x_{i+1}]$.

Til forskel fra tidligere, hvor u blev tilnærmet vha. værdier i en række punkter, tilnærmer vi altså her u vha. en *funktion* fra en bestemt *klasse* V af funktioner (V er mængden af funktioner, som er kontinuerte på u's definitionsmængde og lig et polynomium på delintervallerne $[x_i, x_{i+1}]$), som ikke er differentiable overalt (de er normalt ikke differentiable i x_i), og hvis andenafledede er 0, hvor den eksisterer (den andenafledede af et førstegradspolynomium er 0).

Spørgsmålet er nu, hvordan man på fornuftig vis vælger sin funktion $v \in V$, så v er en god approksimation af u. Af mange grunde, hvoraf nogle af de vigtigste først opstår i højere dimensioner end vores eksempels ene, er finite difference-metoden ikke optimal, hvilket også intuitivt synes oplagt, idet en sådan tilgang slet ikke udnytter, at v rent faktisk er en rigtig funktion med samme definitionsmængde som u, og ikke blot en endelig samling værdier. Hvis man sammensætter de indledende knæbøjninger i afsnittet om bevarelseslove med det faktum, at v ikke er differentiabel overalt, så kan man måske gætte, at en god løsning er at omformulere ens differentialligning i en *svag* udgave.

1.5 Finite element-metoden

Vi vil nu introducere finite element-metoden ved at illustrere den med to eksempler. Det ene eksempel vil være et ODE-eksempel, som egentlig kan løses numerisk på en mere elegant metode, men som kan beskrives meget konkret og som tjener det formål, at det illustrerer idéerne ret klart, mens det andet eksempel er et PDE-eksempel, som vi vil gennemgå mere overfladisk, og som tjener det formål, at vi kan illustrere, hvordan idéerne skal anvendes på mere komplekse problemer. De to problemer er valgt, så de ligger tæt op ad hinanden og kan kort beskrives som Poisson-problemer med Dirichlet-randbetingelser i hhv. én og to dimensioner.

Eksempel 1

Første eksempel er det endimensionelle Poisson-problem med homogene Dirichlet-randbetingelser på $[0, 1]$:

$$\begin{aligned} u''(x) &= f(x) \qquad \text{for} \quad x \in (0,1), \\ u(0) &= u(1) = 0, \end{aligned}$$

hvor $f\colon (0,1) \to \mathbb{R}$ er en given funktion.

Eksempel 2

Andet eksempel er det todimensionelle Poisson-problem med homogene Dirichlet-randbetingelser på den "tilpas pæne" (men ikke nærmere definerede) mængde Ω hvis rand betegnes $\partial\Omega$:

$$\begin{aligned} u_{xx}(x,y) + u_{yy}(x,y) &= f(x,y) \qquad && \text{for} \quad (x,y) \in \Omega, \\ u(x,y) &= 0 && \text{for} \quad (x,y) \in \partial\Omega. \end{aligned}$$

for en given funktion $f\colon \Omega \to \mathbb{R}$.

Den svage formulering af problemet

Som allerede antydet er første skridt at omformulere problemet til den svage form. Som diskuteret i første afsnit så går den svage formulering af en differentialligning ud på at gange begge sider med en testfunktion og integrere udtrykkene. Vi får derfor i Eksempel 1's tilfælde:

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

hvor v er vores testfunktion. Pga. randbetingelserne $u(0) = u(1) = 0$ vælger vi at indskrænke vores valg af testfunktioner til funktioner v, som opfylder, at

$$v(0) = v(1) = 0. \tag{3}$$

Bringes denne betingelse i samspil med partiel integration fås

$$\begin{aligned}\int_0^1 f(x)v(x)\,\mathrm{d}x &= [u'(x)v(x)]_{x=0}^1 - \int_0^1 u'(x)v'(x)\,\mathrm{d}x \\ &= u'(1)\cdot 0 - u'(0)\cdot 0 - \int_0^1 u'(x)v'(x)\,\mathrm{d}x \\ &= -\int_0^1 u'(x)v'(x)\,\mathrm{d}x. \end{aligned} \tag{4}$$

På *helt tilsvarende vis* fås for Eksempel 2:

$$\int_\Omega fv\,\mathrm{d}s = -\int_\Omega \nabla u\cdot\nabla v\,\mathrm{d}s, \tag{5}$$

hvor $\nabla w = \frac{\partial w}{\partial x} + \frac{\partial w}{\partial y}$. For passende valg af testfunktioner er (4) og (5) altså de svage formuleringer af problemet. Næste spørgsmål, vi søger svar på, er derfor hvilken klasse af testfunktioner, vi skal benytte.

Valg af testfunktioner

Det kan vises, at et godt valg af testfunktioner er funktioner af samme type som de (numeriske) løsninger, vi leder efter. Dette vil altså sige, at i Eksempel 1 skal testfunktionerne være kontinuerte på $[0,1]$ og lig et polynomium på delintervaller $[x_i, x_{i+1}]$, hvor $x_0 = 0$, $x_N = 1$ og $x_i < x_{i+1}$ er en inddeling af $[0,1]$. Vi bemærker, at vi allerede har et yderligere krav på testfunktionerne om, at de skal antage værdien 0 i endepunkterne 0 og 1, men i samme ombæring bemærker vi, at dette *ikke* indskrænker mængden af testfunktioner i forhold til mængden af funktioner, hvoriblandt vi søger vores numeriske løsning, idet dette ekstrakrav på mulige løsninger allerede er dikteret af Dirichlet-randbetingelserne! Vi vil betegne denne mængde V_0, altså:

$$V_0 = \{v\colon [0,1]\to\mathbb{R} \mid v \text{ kontinuert}, v(0) = v(1) = 0, \text{ og } v_{|[x_i,x_{i+1}]}(x) = p_i(x), i = 0,\ldots,N-1\},$$

Hvor $v_{|[x_i,x_{i+1}]}$ betegner restriktionen af v til $[x_i, x_{i+1}]$ og p_i betegner et førstegradspolynomium. V_0 er altså både mængden af funktioner, hvoriblandt vi skal finde vores numeriske løsning af problemet, og mængden af testfunktioner.

Hvad vil dette så sige i tilfældet Eksempel 2? I første omgang skal vi have generaliseret, hvad vi mener med et *element* i Eksempel 2. Hvor det i Eksempel 1 kun er muligt at opdele definitionsmængden på én måde, nemlig i delintervaller, så er definitionsmængden i Eksempel 2 todimensionel, og det er klart, at en todimensionel mængde kan deles på mange flere måder. Det viser sig da også, at der her er en stor grad af valgfrihed. En af de mest almindelige måder er dog, at man inddeler definitionsmængden i trekantede områder (bemærk, hvordan dette *ikke* harmonerer med en finite difference-tilgang). Hvad så med kravet om, at de numeriske tilnærmelser skal være førstegradspolynomier på hver trekant? Jo, polynomier findes også i højeredimensionelle udgaver, og et førstegradspolynomium af to variable er en funktion p på formen

$$p(x, y) = a_1 x + a_2 y + b$$

(mens et andengradspolynomium af to variable er på formen $a_{11}x^2 + a_{12}xy + a_{22}y^2 + b_1 x + b_2 y + c$).

Valg af basis for testfunktionerne

Vi begynder igen med tilfældet Eksempel 1. Det er ikke svært at indse, at V_0 er et *endeligdimensionelt vektorrum*, hvilket betyder, at der findes et endeligt antal funktioner ($N - 1$, for at være præcis), så alle elementer i V_0 kan skrives som en linearkombination af disse endeligt mange funktioner. En sådan mængde af funktioner kaldes en *basis*. Da $v(0) = v(1) = 0$ for alle $v \in V_0$ og v skal være kontinuert, skal det samme også gælde for basisfunktionerne. Da v endvidere er lig et førstegradspolynomium på hvert delinterval af typen $[x_i, x_{i+1}]$, så kender vi v på $[x_i, x_{i+1}]$, hvis vi kender (x_i) og $v(x_{i+1})$. Ved snedig udnyttelse af disse fakta kan det vises, at funktionerne givet ved

$$v_k(x) = \begin{cases} \frac{x - x_{k-1}}{x_k - x_{k-1}} & \text{for} \quad x \in [x_{k-1}, x_k] \\ \frac{x_{k+1} - x}{x_{k+1} - x_k} & \text{for} \quad x \in [x_k, x_{k+1}] \\ 0 & \text{ellers,} \end{cases}$$

$k = 1, \ldots, N - 1$, udgør en basis for V_0. En basisfunktion v_k antager således værdien 1 i x_k, 0 i alle andre x_i'er, er kontinuert og er lig et polynomium på hvert delinterval $[x_i, x_{i+1}]$.

Sidstnævnte sproglige beskrivelse af v_k'erne er nøglen til beskrivelsen af en basis i tilfældet Eksempel 2, hvor vi har valgt trekantede elementer. Hver trekant kan angives vha. koordinaterne for de tre hjørner, og mængden af hjørner (alle trekanter inkluderet) kan vi betegne $\{x_k \colon k \in K\}$, for en eller anden passende indeksmængde K. Lad nu K_0 betegne de k, hvorom det gælder, at $x_k \notin \partial\Omega$. Basisfunktionerne er så de v_k, $k \in K_0$, som opfylder, at $v_k(x_k) = 1$, $v_k(x_j) = 0$ for $k \neq j \in K$, og v_k restringeret til ethvert trekantet element er lig et førstegradspolynomium.

Udnyttelse af basisvektorernes begrænsede støtte

Støtten af en funktion f betegner den mindste lukkede mængde S, hvorom det gælder, at f er 0 udenfor S. For basisvektorerne v_k i Eksempel 1 er støtten således $[x_{k-1}, x_{k+1}]$. Hvis N er stor, er disse intervaller typisk små og v_k siges at have en lille støtte. Støtten for en funktion vil indeholde støtten for funktionens afledede, og dette betyder, at følgende to integraler vil være 0 for langt de fleste valg af k og j (igen forudsat at N er stor):

$$\int_0^1 v_j(x) v_k(x) \, \mathrm{d}x \qquad \text{og} \qquad \int_0^1 v_j'(x) v_k'(x) \, \mathrm{d}x$$

(helt konkret vil de være 0 for $|k - j| > 1$).

Tilsvarende vil

$$\int_\Omega v_j v_k \,\mathrm{d}s \qquad \text{og} \qquad \int_\Omega \nabla v_j \cdot \nabla v_k \,\mathrm{d}s$$

være 0, såfremt x_j og x_k ikke er "nabohjørner."

Matrixformulering af problemet

Vi vil nu vise, hvordan man finder en numerisk løsning, og tager igen udgangspunkt i Eksempel 1. Da vi har en basis for V_0, så vil det at finde en numerisk løsning til vores problem svare til at finde koefficienter til basisvektorerne. At finde koefficienter for en basis lyder som et lineær algrbra-problem, og vi vil da også vise, hvordan man kan formulere problemet som et matrixproblem. Vi begynder med at konstatere, at vi leder efter en numerisk løsning v på formen

$$v(x) = \sum_{i=1}^{N-1} u_i v_i(x),$$

hvor u_i er de ukendte koefficienter og v_i er basisvektorerne. Den svage form af problemet (4) tager altså formen

$$\int_0^1 f(x)v_j(x)\,\mathrm{d}x = -\int_0^1 \Big(\sum_{k=1}^{N-1} u_k v_k'(x)\Big) v_j'(x)\,\mathrm{d}x = -\sum_{k=1}^{N-1} u_k \int_0^1 v_k'(x)v_j'(x)\,\mathrm{d}x,$$

som skal gælde for alle $j = 1, \ldots, N-1$. Her kan venstresiden evalueres numerisk, og højresiden er en sum af nogle ukendte koefficienter ganget med nogle integraler, som også kan evalueres numerisk, og som i øvrigt for det meste er 0 (vi kunne have nøjes med at summe over $k \in \{j-1, j, j+1\}$). Skrevet på matrixform giver det

$$-\begin{bmatrix} A_{1,1} & A_{1,2} & \cdots & A_{N-1,1} \\ A_{2,1} & A_{2,2} & \cdots & A_{N-1,2} \\ \vdots & \vdots & \ddots & \vdots \\ A_{1,N-1} & A_{2,N-1} & \cdots & A_{N-1,N-1} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_{N-1} \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_{N-1} \end{bmatrix},$$

eller $-Au = b$ for passende definition af A, u og b, hvor

$$A_{j,k} = \int_0^1 v_j'(x)v_k'(x)\,\mathrm{d}x$$

og

$$b_j = \int_0^1 f(x)v_j(x)\,\mathrm{d}x.$$

Her bemærkes det, at $A_{j,k}$ er 0 undtagen for $|j-k| \leq 1$, eller sagt på en anden måde, A-matricen har 0-indgange undtagen på, lige over og lige under diagonalen jf. diskussionen om den begrænsede støtte, og der findes derfor effektive metoder at løse matrixligningen på, selv for store værdier af N. I tilfældet Eksempel 2 fås noget *helt* parallelt, blot med $A = (A_{j,k})_{j,k\in K_0}$, $u = (u_k)_{k\in K_0}$ og $b = (b_j)_{j\in K_0}$ passende redefineret. Igen er A en såkaldt *tyndt besat* matrix, og det er derfor også i dette tilfælde muligt numerisk at løse matrixligningen for selv et meget stort antal elementer (som svarer til et stort K_0).

Afsluttende kommentar angående finite element-metoden

Det bør kraftigt understreges, at det frie valg af "elementer" i ovenstående metode kan udnyttes endog særdeles effektivt. Hvis et område er specielt interessant at få præcise værdier for, at så kan man blot *lokalt* skrue op for opløsningen. Eller hvis ens model beskriver noget, som består af flere forskellige typer materiale, så overgangen mellem materiale er af speciel vigtighed, så kan man blot skrue op for opløsningen hér – igen lokalt og derfor uden alt for store beregningsmæssige omkostninger. Et sidste eksempel er, at hvis en løsning opfører sig relativt simpelt i store områder af løsningsmængden, mens den i andre områder har en mere involveret opførsel, så skal man naturligvis igen blot skrue op for opløsningen i disse områder. Det er rent faktisk muligt at få den lokale "kvalitet" af den numeriske løsning ud som et biprodukt af beregningerne, og man kan derved automatisere en sådan forfining af elementinddelingen.

1.6 Numeriske overvejelser – bevarelseslove og voluminer

Indtil videre har vi i de numeriske tilgange ikke rigtig beskæftiget os med, om der var tale om en *bevarelseslov*. Der er dog mange tilfælde, hvor netop denne egenskab betyder meget for modellen og kan være en stor hjælp i beregningerne. Antag eksempelvis, at vi har en eller anden form for opløsning i en væske i et eller andet mere eller mindre kompliceret system, og at vi er fuldt ud tilfredse med at kende gennemsnitsværdien af koncentrationen i såkaldte *kontrolvoluminer* men interesserede i, at flowet er så korrekt som muligt, og at eventuelle fejl udjævner hinanden, således, at vi eksempelvis ikke pludselig ender med at have en større mængde opløst materiale, end der blev puttet ind i systemet. I et sådant tilfælde er det oplagt at udnytte bevarelseslovene, og det er denne idé, *finite volume*-metoden bygger på. I finite volume-metoden inddeles det område, hvor det undersøgte problem udfolder sig, i en masse dele kaldet kontrolvoluminer, eksempelvis bestående af tetraedre ("tredimensionelle trekanter"), og den numeriske løsning vil være gennemsnitsværdien i disse kontrolvoluminer.

1.7 Finite volume-metoden

Som nævnt ovenfor spiller bevarelseslovene en stor rolle i finite volume-metoden. Først inddeles rummet i nogle kontrolvoluminer. Idéen er, som allerede antydet ovenfor, at udnytte, at integralet over divergens-led kan konverteres til overfladeintegraler vha. Gauss' divergenssætning. Disse overfladeintegraler på de enkelte kontrolvoluminer bliver så evalueret som flux-værdier på overfladen af de enkelte kontrolvoluminer. Metoden er lokalt konservativ, idet idéen går på at udnytte, at et flux ud af siden på ét kontrolvolumen nødvendigvis må være lig med fluxet ind gennem siden på nabokontrolvoluminet. Vi vil nu kort illustrere idéen vha. Eksempel 1 fra finite element-afsnittet.

Eksempel 1 genfortalt i konservativ form

Det endimensionelle Poisson-problem med homogene Dirichlet-randbetingelser på $[0,1]$ er givet ved:

$$\begin{aligned} u''(x) &= f(x) \qquad \text{for} \quad x \in (0,1) \\ u(0) &= u(1) = 0, \end{aligned}$$

hvor $f\colon (0,1) \to \mathbb{R}$ er en given funktion.

Da vi er i én dimension, så er divergensen blot den x-afledede, og vi kan altså omformulere problemet på følgende måde:

$$\operatorname{div} u'(x) = f(x) \qquad \text{for} \quad x \in (0,1). \tag{6}$$

Inddelingen af området

Da vi både skal kunne referere til midtpunkter og randpunkter af kontrolvoluminerne (som takket være den ene dimension blot er delintervaller), så bliver vores inddeling lidt mere omstændelig end tidligere. Mere præcist skal vi bruge

$$0 = x_0 = x_{\frac{1}{2}} < x_1 < x_{\frac{3}{2}} < \cdots < x_{i-\frac{1}{2}} < x_i < x_{i+\frac{1}{2}} < \cdots < x_N < x_{N+\frac{1}{2}} = x_{N+1} = 1,$$

og vores kontrolvoluminer er så $K_i = (x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}})$, $i = 1, \ldots, N$. Vi skriver $\Delta x_i = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}$ og $\Delta x_{i+\frac{1}{2}} = x_{i+1} - x_i$ for $i = 1, \ldots N$. De ubekendte er nu u_i, $i = 1, \ldots, N$ og skal tolkes som tilnærmelser til u's gennemsnitsværdi over K_i.

Matrixformuleringen af problemet

Vi integrerer (6) over K_i og anvender Gauss' divergenssætning, som i én dimension reducerer til analysens fundamentalsætning:

$$u'(x_{i+\frac{1}{2}}) - u'(x_{i-\frac{1}{2}}) = \int_{K_i} f(x)\,\mathrm{d}x.$$

Vi kan nu approksimere u' med en differenskvotient (*central difference*):

$$\frac{u(x_{i+1}) - u(x_i)}{\Delta x_{i+\frac{1}{2}}} - \frac{u(x_i) - u(x_{i-1})}{\Delta x_{i-\frac{1}{2}}} \approx \int_{K_i} f(x)\,\mathrm{d}x.$$

Da u_i blot skal *approksimere* u's gennemsnitsværdi over K_i, vælger vi at definere u_i'erne ud fra

$$\frac{u_{i+1} - u_i}{\Delta x_{i+\frac{1}{2}}} - \frac{u_i - u_{i-1}}{\Delta x_{i-\frac{1}{2}}} = \int_{K_i} f(x)\,\mathrm{d}x.$$

som er gyldig for alle $i = 1, \ldots, N$, hvis man sætter $u_0 = u_{N+1} = 0$. Dette kan også skrives på matrixform:

$$Au = f,$$

hvor $u = \begin{bmatrix} u_0 & u_1 & \cdots & u_N & u_{N+1} \end{bmatrix}^T$, $f = \begin{bmatrix} 0 & \int_{K_1} f(x)\,\mathrm{d}x & \cdots & \int_{K_i} f(x)\,\mathrm{d}x & \cdots \int_{K_N} f(x)\,\mathrm{d}x & 0 \end{bmatrix}$ og $A = (A_{ij})_{i,j \in \{0,1,\ldots,N,N+1\}}$ er givet ved $(Au)_0 = u_0$, $(Au)_{N+1} = u_{N+1}$ og $(Au)_i = \frac{u_{i+1}-u_i}{\Delta x_{i+\frac{1}{2}}} - \frac{u_i - u_{i-1}}{\Delta x_{i-\frac{1}{2}}}$ for $i = 1, \ldots, N$.

Afsluttende kommentar angående finite volume-metoden

Som for finite element-metoden bør det kraftigt understreges, at det frie valg af kontrolvoluminer i ovenstående metode kan udnyttes effektivt med ca. samme argumenter som før. Desuden betyder den lokalt konservative metode, at upræcisheder tildels ophæver hinanden: den ene celles tab er den næste celles gevinst, således at det overordnede flow forbliver tæt på korrekt.