MC SYLLABUS 27

M. BAKKER
P.W. HEMKER
P.J. VAN DER HOUWEN
S.J. POLAK
M. VAN VELDHUIZEN

COLLOQUIUM DISCRETISERINGSMETHODEN

MATHEMATISCH CENTRUM AMSTERDAM 1976

AMS(MOS) subject classification scheme (1970): 65L05

ISBN 90 6196 124 6

INHOUD

Inhoud v

Voorwoord vii

1. Eindige differentiemethoden
   door P.J. van der Houwen 1

2. De eindige elementenmethode voor het discretiseren van elliptische randwaardeproblemen
   door P.W. Hemker 18

3. Toepassing van de eindige elementenmethode op één- en tweedimensionale randwaardeproblemen
   door M. Bakker 67

4. Gewogen residuenmethoden
   door P.W. Hemker 116

5. Galerkin-methoden voor gewone differentiaalvergelijkingen
   door M. van Veldhuizen 130

6. Eindige elementenmethode voor magnetostatische problemen
   door S.J. Polak 149

VOORWOORD

Het colloquium Discretiseringsmethoden werd van oktober 1974 tot en met januari 1975 gehouden en werd georganiseerd door de afdeling Numerieke Wiskunde van het Mathematisch Centrum.

Het colloquium was bedoeld om een inzicht te geven in bestaande methoden om in het bijzonder elliptische randwaardeproblemen discreet op te lossen. De nadruk kwam hierbij te liggen op de globale of eindige elementenmethoden.

Aan het colloquium namen deel M. Bakker, P.W. Hemker en P.J. van der Houwen van het Mathematisch Centrum, S.J. Polak van Philips en M. van Veldhuizen van de Vrije Universiteit.

De syllabus van het colloquium werd geredigeerd door M. Bakker.

1. EINDIGE DIFFERENTIEMETHODEN

Alhoewel het hoofdthema van dit colloquium gevormd wordt door de *eindige elementenmethode* als discretiseringsmethode voor randwaardeproblemen, willen we eerst kort ingaan op een "concurrerende" methode, namelijk de *eindige differentiemethode*. Verder zullen we de begrippen *consistentie*, *convergentie* en *stabiliteit* voor discrete modellen definiëren. We beginnen echter met de definitie en voorbeelden van randwaardeproblemen.

1.1. Randwaardeproblemen

Definities

Beschouw de Euclidische ruimte $\mathbb{R}^m$ van dimensie m, het gebied G in $\mathbb{R}^m$ en de rand Γ van G. Laten $E(\bar{G})$, E(G) en $E(\Gamma)$ lineaire, genormeerde ruimten van scalar- of vectorfuncties voorstellen, respectievelijk gedefinieerd op $\bar{G}$, G en Γ ($\bar{G}$ stelt de afsluiting van G voor, i.e. $\bar{G} = G \cup \Gamma$). De elementen van $E(\bar{G})$ en E(G) geven we met Latijnse hoofdletters aan, de elementen van $E(\Gamma)$ met Griekse hoofdletters. De ruimten E(G) en $E(\Gamma)$ definieren de ruimte $E(G) * E(\Gamma)$ van elementen (F,Φ), welke lineair is als we de operaties

$$(F,\Phi) + (F',\Phi') = (F+F',\Phi+\Phi')$$

en

$$a.(F,\Phi) = (aF,a\Phi) \quad , \text{ a scalair}$$

definiëren. Verder definiëren we in $E(G) * E(\Gamma)$ de norm

$$\|(F,\Phi)\| = \|F\|_G + \|\Phi\|_\Gamma ,$$

waarin $\|\cdot\|_G$ en $\|\cdot\|_\Gamma$ de normen in E(G) respectievelijk E(Γ) voorstellen.

Definitie 1.1.1

Het probleem om de inverse te vinden van een gegeven afbeelding L van een onbekende functie $U \in E(\bar{G})$ op een gegeven element $(F,\Phi) \in E(G) * E(\Gamma)$ wordt een randwaardeprobleem genoemd.

Randwaardeproblemen zullen we beschrijven met de vergelijking

(1.1.1) $\quad LU = (F,\Phi)$

De componenten F en Φ van het rechterlid in (1.1.1) zullen we respectievelijk de *inwendige functie* en de *randfunctie* noemen.

In het volgende beperken we ons tot zogenaamde *correct gestelde problemen*.

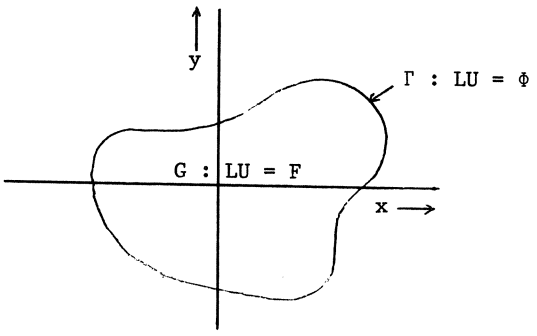
Definitie 1.1.2

Het probleem LU = (F,Φ) wordt correct gesteld genoemd ten opzichte van de normen in $E(\bar{G})$ en $E(G) * E(\Gamma)$ wanneer L een eenduidige inverse L^{-1} heeft die continu is in (F,Φ).

Na deze abstracte definitie van de problemen die onderwerp van dit colloquium zullen zijn, geven we in de volgende paragrafen een drietal voorbeelden van in de practijk voorkomende randwaardeproblemen. Het eerste voorbeeld is standaard, het tweede en derde voorbeeld zijn onderwerp van onderzoek geweest in de afdeling Toegepaste Wiskunde van het Mathematisch Centrum en zijn verre van standaardproblemen.

De vergelijking van Poisson

Beschouw in de $\mathbb{R}^2$ een gebied G waarvan de rand Γ een gesloten Jordankromme is (zie figuur 1.1).



Figuur 1.1 Randwaardeprobleem

In G definieren we de vergelijking (Poisson)

$$(1.1.2) \qquad \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right) U = F \ ,$$

en op Γ de randvoorwaarde

$$(1.1.3) \qquad (\Phi_1 + \Phi_2 \frac{\partial}{\partial x} + \Phi_3 \frac{\partial}{\partial y}) U = \Phi .$$

Hierin zijn Φ_1, Φ_2 en Φ_3 gegeven functies op Γ.

Onder vrij algemene aannamen omtrent de functies F, Φ_j en Φ bezit bovenstaand randwaardeprobleem een eenduidige oplossing, met andere woorden, is een correct gesteld probleem.

De vergelijking van Poisson is van het elliptische type en beschrijft bijvoorbeeld de evenwichtstoestand van membranen. Voor verdere voorbeelden van physische problemen beschreven door elliptische randwaardeproblemen verwijzen we naar FORSYTHE & WASOW [1960].

Het Tricomi-probleem

Zo bekend als de vergelijking van Poisson is, zo weinig bekend is de *vergelijking van Tricomi*:

$$(1.1.4) \qquad \left(y \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right) U = F.$$

Deze vergelijking is van het *gemengde type*, dat wil zeggen in een gebied G zoals in figuur 1.1, is deze vergelijking niet overal van eenzelfde type, maar *elliptisch* voor $y > 0$, *parabolisch* voor $y = 0$ en *hyperbolisch* voor $y < 0$.
In tegenstelling tot elliptische vergelijkingen behoeven bij dit type vergelijkingen niet langs de hele rand Γ van het gebied G randvoorwaarden gedefinieerd te worden om eenduidigheid te verzekeren. In de gasdynamica neemt de vergelijking van Tricomi een centrale plaats in; bijvoorbeeld de transone stroming van een gas door een pijp wordt beschreven door vergelijking (1.1.4) met $F = 0$ en randvoorwaarden

$$(1.1.5) \qquad U = \Phi_e$$

langs het elliptische gedeelte Γ_e van de rand Γ (zie figuur 1.2) en

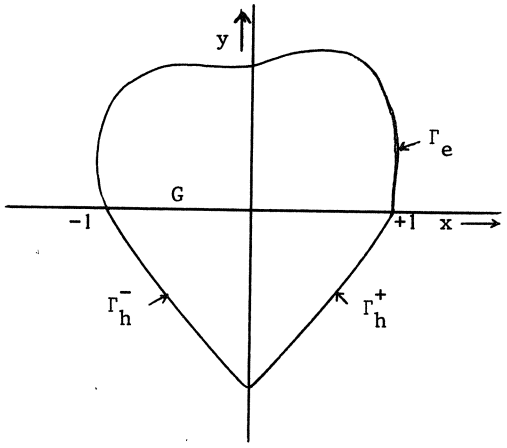
$$(1.1.6) \qquad U = \Phi_h$$

langs Γ_h^-, waarin Γ_h^- gegeven wordt door de karakteristiek

$$x - \frac{2}{3}(-y)^{3/2} = -1 \quad , \quad -1 \le x \le 0$$

van vergelijking (1.1.4). Langs de rest Γ_h^+ van de rand Γ hoeft niets gegeven te worden. Laat Γ_h^+ ook langs een karakteristiek lopen, dan bepalen (1.1.5) en (1.1.6) de oplossing in het door $\Gamma = \Gamma_e + \Gamma_h^+ + \Gamma_h^-$ omrande gebied G. Dit randwaardeprobleem wordt het *Tricomi-probleem* genoemd.

Het Tricomi-probleem kan met behulp van analytische methoden gereduceerd worden tot een (singuliere) integraalvergelijking (zie BITSADZE [1964]) voor $U_y(x,0)$, waarin de oplossing $U(x,y)$ dan weer uitgedrukt kan worden via nogal ingewikkelde integraalrepresentaties. Het is duidelijk dat een directe numerieke oplossingsmethode aantrekkelijker is. In paragraaf 1.3 komen we hier nog op terug



Figuur 1.2 Tricomi-probleem

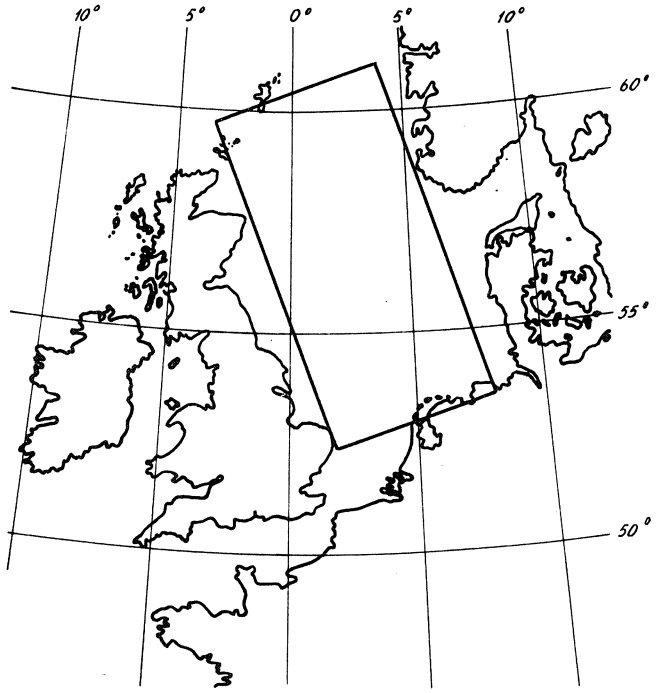
Het Noordzeeprobleem voor tijdsonafhankelijke windvelden

De stationaire waterbeweging in de Noordzee onder een niet van de tijd afhangend windveld $\vec{F} = (F_1, F_2)$, wordt beschreven door de vergelijkingen

$$
\begin{aligned}
-\lambda U + \Omega V - gh\frac{\partial Z}{\partial x} &= F_1 \\
-\Omega U - \lambda V - gh\frac{\partial Z}{\partial y} &= F_2 \qquad (1.1.7) \\
\frac{\partial}{\partial x} U + \frac{\partial}{y} V &= 0
\end{aligned}
$$

met langs de kusten de randvoorwaarden dat de stroom $\vec{W} = (U,V)$ geen component loodrecht op de kust heeft en langs de oceaanbegrenzing de voorwaarde dat de waterhoogte Z constant is, zeg Z = 0. In de vergelijkingen (1.1.7) stellen λ, Ω, g en h respectievelijk de zeebodemwrijving, de Coriolis-coefficient, de gravitatieconstante en de diepte voor.

Van Dantzig en Lauwerier hebben dit randwaardeprobleem diepgaand onderzocht. Door nogal drastische aannamen te maken omtrent de functies



Figuur 1.3 Noordzee-probleem

λ, Ω en h, maar vooral door de kustlijnen sterk te vereenvoudigen (zie figuur 1.3), slaagden zij erin met analytische hulpmiddelen benaderingen te vinden voor U, V en Z. Voor meer realistische Noordzeemodellen is men op numerieke oplossingsmethoden aangewezen.

Tenslotte merken we nog op dat de introductie van een stroomfunctie

$$S(x,y) \; : \; U = -\frac{\partial S}{\partial y} \quad , \quad V = \frac{\partial S}{\partial x} \; ,$$

de vergelijkingen (1.1.7) overvoert in de elliptische vergelijking

$$\left[\lambda\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right) + \left(\frac{\partial\lambda}{\partial x} - h^{-1}(\lambda\frac{\partial}{\partial x} + \Omega\frac{\partial}{\partial y})h\right)\frac{\partial}{\partial x} + \right.$$

$$\left. + \left(\frac{\partial\lambda}{\partial y} - h^{-1}(\lambda\frac{\partial}{\partial y} - \Omega\frac{\partial}{\partial x})h\right)\frac{\partial}{\partial y}\right] S =$$

$$= \text{rot } \vec{F} - h^{-1}(F_1\frac{\partial}{\partial x} - F_2\frac{\partial}{\partial y})h;$$

langs de kusten geldt voor de stroomfunctie S de randvoorwaarde

(1.1.9) $\qquad S = 0,$

terwijl voor een oceaanbegrenzing langs een lijn y = constant de oceaanvoorwaarde wordt (zie VAN DER HOUWEN [1968])

(1.1.10) $\qquad (\lambda\frac{\partial}{\partial y} + \Omega\frac{\partial}{\partial x})\, S = -F_1.$

Voor numerieke behandeling lijkt (1.1.7) echter geschikter dan (1.1.8), alleen al omdat men in het eerste geval direct U, V en Z in handen krijgt, waar men in het tweede geval extra rekenwerk moet doen om deze componenten te verkrijgen.

1.2 Discrete benaderingen, consistentie, convergentie en stabiliteit

We keren terug tot het algemene randwaardeprobleem

$$LU = (F,\Phi).$$

We willen deze afbeelding van *continue* vectorfuncties uit $E(\bar{G})$ op *continue* vectorfunctie-paren uit $E(G) * E(\Gamma)$ vervangen door een afbeelding tussen ruimten van *discrete* functies. Daartoe kiezen we in het gebied G en op de rand Γ respectievelijk de rijen $\{G_h\}_{h=\bar{h}}^{0}$ en $\{\Gamma_h\}_{h=\bar{h}}^{0}$ van eindige puntverzamelingen, met de eigenschap dat $\lim_{h\to 0} G_h$ en $\lim_{h\to 0} \Gamma_h$ dicht liggen in respectievelijk G en Γ.
Analoog aan de definitie van de ruimten $E(\bar{G})$, $E(G)$ en $E(\Gamma)$ definieren we nu de lineaire genormeerde ruimten $E(\bar{G}_h) = E(G_h+\Gamma_h)$, $E(G_h)$ en $E(\Gamma_h)$.
De elementen van deze ruimten van discrete functies (*roosterfuncties*) geven we aan met U_h, F_h en Φ_h.

Analoog aan (1.1.1) beschrijven we een discreet randwaardeprebleem door

middel van de vergelijking

$$L_h U_h = (F_h, \Phi_h), \tag{1.2.1}$$

waarin L_h een operator is met definitiegebied in $E(\bar{G}_h)$ en beeldgebied in $E(G_h) * E(\Gamma_h)$. Omdat L_h op discrete functies werkt, zal deze operator niet een *analytische* operator zijn maar een *algebraische* operator, ook wel differentieoperator genoemd.

Om (1.1.1) en (1.2.1) met elkaar in verband te kunnen brengen, definiëren we een *discretiseringsoperator* $[\ ]_h$ die aan een functie uit bijvoorbeeld $E(\bar{G})$ de waarden van deze functie in $\bar{G}_h$ associeert. Het is duidelijk dat deze waarden een roosterfunctie uit $E(\bar{G}_h)$ definiëren, dus

$$U \in E(\bar{G}) \Rightarrow [U]_h \in E(\bar{G}_h).$$

Op dezelfde wijze liggen $[F]_h$ en $[\Phi]_h$ in $E(G_h)$ respectievelijk $E(\Gamma_h)$ wanneer F en Φ in $E(G)$ respectievelijk $E(\Gamma)$ liggen.

We zijn nu zover om (1.1.1) en (1.2.1) met elkaar in verband te brengen. Laat $\bar{U}$ een oplossing zijn van (1.1) en substitueer deze in de discrete vergelijking

$$L_h U_h = ([F]_h, [\Phi]_h) , \tag{1.2.1a}$$

dan vinden we het residu (ook wel *afbreekfout* genoemd)

$$L_h \tilde{U}_h = ([F]_h, [\Phi]_h) . \tag{1.2.2}$$

Wanneer dit residu naar het nul-element van de ruimte $E(G_h) * E(\Gamma_h)$ convergeert voor $h \to 0$, dan wordt (1.2.1a) een *consistente benadering* van (1.1.1) genoemd. Men spreekt van *consistentie van de orde p* als

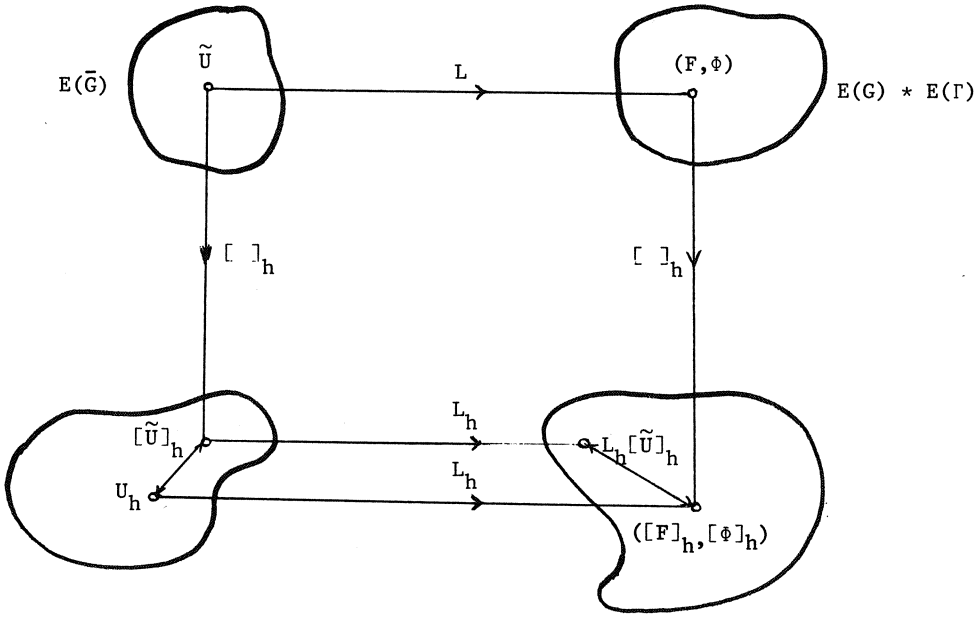
$$\| L_h [\tilde{U}]_h - ([F]_h, [\Phi]_h) \| = O(h^p) \quad \text{voor } h \to 0, \tag{1.2.3}$$

waarin $\| \ \|$ de norm in $E(G_h) * E(\Gamma_h)$ voorstelt.

Voorwaarde (1.2.3) met $p \geq 1$ garandeert nog niet dat de oplossing U_h van (1.2.1a) naar $[\tilde{U}]_h$ convergeert voor $h \to 0$; consistentie betekent alleen maar dat het discrete randwaardeprobleem convergeert naar het continue randwaardeprobleem.

Laten we aannemen dat L_h een eenduidige inverse L_h^{-1} heeft, dan kunnen we schrijven (zie ook figuur 1.4)

$$[\tilde{U}]_h - U_h = L_h^{-1}L_h[\tilde{U}]_h - L_h^{-1}([F]_h,[\Phi]_h). \tag{1.2.4}$$



Figuur 1.4 Afbreekfout en discretiseringsfout

Het verschil $[\tilde{U}]_h - U_h$ noemen we de *discretiseringsfout*; we spreken van *convergentie* wanneer deze fout naar het nul-element van $E(\bar{G}_h)$ convergeert als $h \to 0$. Met andere woorden, wanneer de operator L_h een consistente benadering is van L en wanneer L_h^{-1} uniform continu is in het punt $([F]_h,[\Phi]_h)$ dan convergeert U_h naar $[\tilde{U}]_h$ voor $h \to 0$. De snelheid van convergentie kan eenvoudig geschat worden wanneer de operator L_h^{-1} differentieerbaar is; wanneer we de afgeleide operator van L_h^{-1} in het punt (F_h,Φ_h) met $A_h((F_h,\Phi_h))$ aangeven dan geldt volgens de "middelwaarde-ongelijkheid"

$$\|[\tilde{U}]_h - U_h\| \leq \|A_h((\bar{F}_h, \bar{\Phi}_h))\| \cdot \|L_h[\tilde{U}_h] - ([F]_h,[\Phi]_h)\| , \tag{1.2.5}$$

waarin $(\bar{F}_h,\bar{\Phi}_h)$ een punt is op de "verbindingslijn" tussen $L_h[\tilde{U}]_h$ en $([F]_h,[\Phi]_h)$. Uit deze relatie volgt dat de discretiseringsfout zich gedraagt als

$$(1.2.6) \qquad \|[\tilde{U}]_h - U_h\| \sim \|A_h((\bar{F}_h,\bar{\Phi}_h))\| \cdot O(h^p)$$

als op de orde van consistentie is. De discrete oplossingen convergeren dus naar de oplossing $\tilde{U}$ wanneer de norm van de operator A_h voor $h \to 0$ niet harder dan h^{-p+1} aangroeit.

Tenslotte een enkele opmerking over de *stabiliteit* van (1.2.1) dat wil zeggen de gevoeligheid van de oplossing U_h voor verstoringen van de data (F_h,Φ_h). Analoog aan de afleiding van (1.2.5) kunnen we deduceren dat

$$(1.2.7) \qquad \|U_h' - U_h\| \leq \|A_h((\bar{F}_h,\bar{\Phi}_h))\| \cdot \|(F_h' - F_h,\ \Phi_h' - \Phi_h)\| ,$$

waarin U_h' de bij de data (F_h',Φ_h') behorende oplossing is en $(\bar{F}_h,\bar{\Phi}_h)$ op de "verbindingslijn" tussen (F_h',Φ_h') en (F_h,Φ_h) ligt. De norm van de afgeleide van L_h^{-1} bepaalt dus de gevoeligheid voor verstoringen van de data.

1.3 Eindige differentiemethoden

Het principe van de eindige differentiemethode is eenvoudig: *kies een rooster* $\bar{G}_h$ *in* $\bar{G}$ *en vervang de differentaties in de operator* L *door differentiequotienten gedefinieerd op de roosterpunten van* $\bar{G}_h$. We zullen met een drietal voorbeelden het *kiezen* van het rooster en het *vervangen* van differentaties door differentiequotienten toelichten.

Het Dirichlet-probleem voor de vergelijking van Poisson

Beschouw vergelijking (1.1.2), i.e.

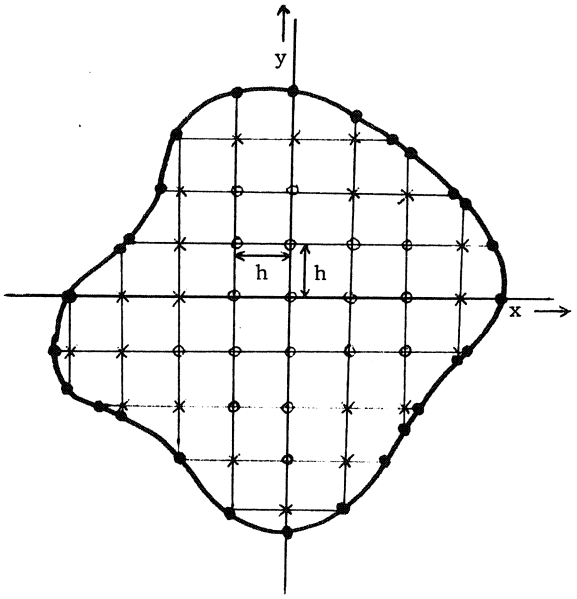
$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right) U = F \ ,$$

met randvoorwaarden van de eerste soort, i.e.

$$U = \Phi \qquad \text{voor } (x,y) \in \Gamma \ .$$

Dit randwaardeprobleem wordt een *Dirichlet-probleem* genoemd.

Beschouw in het gebied G een rooster met vierkante mazen en roosterpuntafstand h (zie figuur 1.5)



Figuur 1.5 Vierkant rooster met maaswijdte h

De roosterpunten in dit rooster kan men verdelen in 1^e *inwendige* roosterpunten, aangegeven met ∘ en gekarakteriseerd door het feit dat in de 8 "windrichtingen N, NO, O,..., W, NW" nog juist een roosterpunt in $\bar{G}$ ligt, 2^e *randpunten* aangegeven met · en op Γ gelegen, en 3^e *niet-inwendige* roosterpunten aangegeven met x.

In de inwendige roosterpunten kan men de Laplace-operator vrij nauwkeurig benaderen door de differentie-operator (z.g. *9-puntsformule*):

$$(1.3.1) \qquad L_h = \frac{1}{6h^2}\left[4(X_+ + X_- + Y_+ + Y_-) + (X_+Y_+ + X_+Y_- + X_-Y_+ + X_-Y_-) - 20\right]$$

waarin $X_\pm$ en $Y_\pm$ translatie-operatoren zijn:

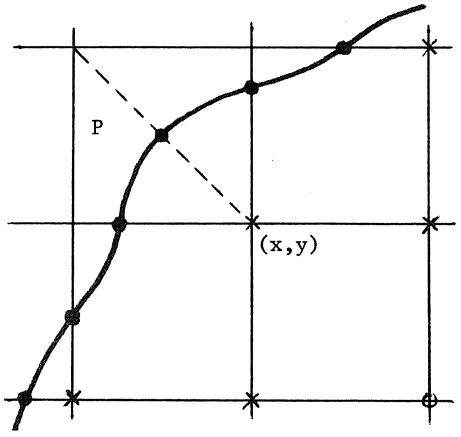
$$X_\pm\, U(x,y) = U(x\pm h, y), \qquad Y_\pm\, U(x,y) = U(x, y\pm h).$$

Men kan eenvoudig bewijzen dat in de inwendige punten geldt:

$$L_h\, \tilde{U}(x,y) - F = O(h^2)\ ,$$

wanneer $\tilde{U}$ aan (1.1.2) voldoet. Dus in deze punten is L_h consistent van de orde 2 met de Laplace-operator. Wanneer $F \equiv 0$ is er zelfs sprake van 6^e orde consistentie (zie e.g. FORSYTHE & WASOW [1960]).

In de niet-inwendige randpunten is de operator (1.3.1) niet gedefinieerd. Bijvoorbeeld in de in figuur 1.6 geschetste situatie kan (1.3.1) niet in

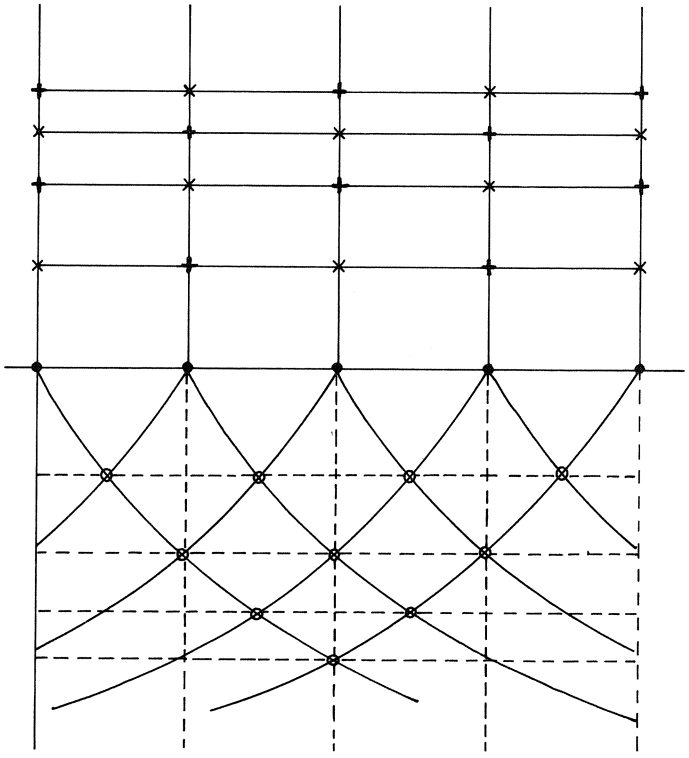


Figuur 1.6 Niet-inwendig randpunt (x,y)

het roosterpunt (x,y) toegepast worden omdat in "noordwestelijke" richting geen roosterpunt binnen $\bar{G}$ meer ligt. Men kan nu het rooster aanvullen met punten van de rand Γ zodanig dat de niet-inwendige punten inwendig worden. In figuur 1.6 zou dit het punt P zijn. Op deze manier kan men in alle niet-randpunten een 9-punts differentie-operator definieren. We zullen hier niet ingaan hoe (1.3.1) gewijzigd moet worden om in deze tot inwendig "gepromoveerde" roosterpunten consistente differentie-operatoren te verkrijgen (zie BRAMBLE & HUBBARD [1962]).

Het Tricomi-probleem

Filippov [1957] was de eerste die een convergente differentiemethode aangaf voor het Tricomi-probleem. Daartoe koos hij het in figuur 1.7 aangegeven rooster. De roosterpunten in het hyperbolische gebied zijn gedefi-



Figuur 1.7 Hyperbolische (o) en elliptische (+,x) roosterpunten in het Tricomi-probleem

nieërd als de snijpunten van de karakteristieken

$$x \pm \frac{2}{3}(-y)^{3/2} = \text{constant},$$

die ontspringen vanuit op onderling gelijke afstand 2h gelegen punten van de parabolische lijn y = 0. De elliptische roosterpunten vindt men dan door spiegeling van de hyperbolische punten ten opzichte van de parabolische lijn en aanvulling tot een rechthoekig rooster.

Laten we hier aannemen dat de "elliptische" rand Γ_e opgebouwd is uit

de zijden en diagonalen van de mazen van het elliptische rooster. We kun nen dan in elk inwendig elliptisch roosterpunt (X,y) de operator $y\partial^2/\partial x^2 + \partial^2/\partial y^2$ benaderen door de differentie-operator

$$L_h = \frac{y}{h^2}(X_+ - 2 + X_-) + \frac{2}{k+1}\left(\frac{1}{k}Y_- + \frac{1}{1}Y_+\right) - \frac{2}{k1}, \tag{1.3.2}$$

waarin $X_\pm$ weer verschuivingen over $\pm h$ in de x-richting, en Y_+ en Y_- verschuivingen in de y-richting over 1 respectievelijk -k voorstellen, waarbij (zie figuur 1.7)

$$k = y - \left[(y)^{3/2} - \frac{3}{2}h\right]^{2/3}, \tag{1.3.3}$$

$$l = \left[(y)^{3/2} + \frac{3}{2}h\right]^{2/3} - y.$$

Op de parabolische lijn geldt eenvoudig ($Y_\pm$ verschuivingen naar het eerst komende roosterpunt)

$$L_h = \left(\frac{3}{2}h\right)^{-4/3}(Y_+ - 2 + Y_-) \tag{1.3.4}$$

in de "oneven" parabolische punten (x/h oneven) en

$$L_h = (h)^{-\frac{4}{3}}(Y_+ - 2 + Y_-) \tag{1.3.4a}$$

in de "even" parabolische punten (x/h even).
Tenslotte kan men in de hyperbolische punten de operator

$$L_h = -\frac{1}{k1}(X_+ + X_-) + \frac{2}{k+1}\left(\frac{1}{k}Y_- + \frac{1}{1}Y_+\right) \tag{1.3.5}$$

gebruiken, wanneer in (1.20) y door -y vervangen wordt.

De consistentie van de operatoren (1.3.2)-(1.3.4a) is eenvoudig te verifiëren, de consistentie van (1.3.3) behoeft echter enige toelichting.
Zij $\tilde{U}$ een oplossing van het Tricomi-probleem dan geldt (ontwikkeling in Taylor-reeksen)

$$L_h[\tilde{U}]_h = \left[\left(-\frac{h^2}{kl}\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)\tilde{U}\right]_h + O(h) \ .$$

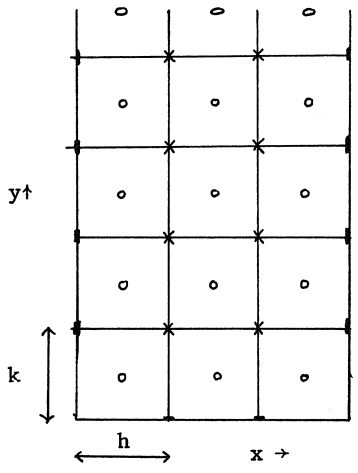
Uit de definitie van k en 1 in het hyperbolische gebied volgt nu dat voor $h \to 0$

$$k \sim 1 \sim -\frac{h}{(-y)^{1/2}} \ ,$$

zodat L_h naar $y\partial^2/\partial x^2 + \partial^2/\partial y^2$ convergeert.

Het Noordzee-probleem

Karakteristiek voor de op het MC ontwikkelde discrete modellen voor het Noordzee-probleem is dat de stroming (U,V) en de waterhoogte Z in verschillende roosterpunten worden berekend (zie figuur 1.8).



Figuur 1.8 Stromingspunten (x,-,1) en verhogingspunten in rechthoekig Noordzeemodel.

Discretiseren we de eerste twee vergelijkingen van (1.1.7) in de stromingspunten en de derde vergelijking in de verhogingspunten, dan vinden we in de interne stromingspunten de vergelijkingen

$$- \lambda U + \Omega V - \sqrt{gh}\left[\frac{(Y_+ + Y_-)(X_+ - X_-)}{4\xi}\right] Z = F_1 \ ,$$

$$- \lambda U - \Omega V - \sqrt{gh}\left[\frac{(X_+ + X_-)(Y_+ - Y_-)}{4\eta}\right] Z = F_2 \ ,$$

en in de inwendige verhogingspunten de vergelijking

$$(Y_+ + Y_-)(X_+ - X_-)\ U + (X_+ + X_-)(Y_+ - Y_-)\ V = 0 \ .$$

Hierin stellen X_+ en $Y_\pm$ weer de gebruikelijke verschuivingsoperatoren voor, in dit geval over de afstanden $\pm\xi$ respectievelijk $\pm\eta$.

Door het rooster verder zo te kiezen dat de kustlijnen door de stromingspunten lopen en de oceaanbegrenzing door de verhogingspunten, kunnen de randvoorwaarden, althans voor rechthoekige Noordzee-modellen, vrij eenvoudig gediscretiseeerd worden. Langs de oceaanbegrenzing geldt eenvoudig

$$Z = 0 \ ,$$

en langs de kusten geldt (zie VAN DER HOUWEN [1968])

$$\begin{aligned} - \lambda U - sgh(sD_x + cD_y)\ Z &= - s(sF_1 + cF_2) \\ - \lambda V - cgh(sD_x + cD_y)\ Z &= - c(sF_1 + cF_2) \end{aligned} \ ,$$

waarin (s,c) de eenheidsvector evenwijdig aan de kust (in positieve zin) voorstelt en D_x en D_y discretiseringen van $\frac{\partial}{\partial x}$ en $\frac{\partial}{\partial y}$ in de kustpunten zijn. Van rechthoekmodellen zijn D_x en D_y eenvoudig te construeren door extrapolatie van Z landinwaarts.

LITERATUUR

BRAMBLE, J.H., HUBBARD, B.E., *On the formulation of finite difference analogs of the Dirichlet problem for Poisson's equation*, Numerische Mathematik 4, p. 313-377 (1962).

FORSYTHE, G., WASAW, W., *Finite Difference Methods for Partial Differential Equations*, John Wiley and Sons, New York, 1960.

VAN DER HOUWEN, P.J., *Finite Difference Methods for solving Partial Differential Equations*, Mathematisch Centrum, Amsterdam, 1968.

BITZADSE, A.V., *Equations of the Mixed Type*, Pergamon, Oxford, 1964.

FILLIPOV, A.F., *On the application of the method of finite differences to the solution of the problem of Tricomi*, Izv. Akad. Nauk. SSSR. Ser. Math. 21, 73-88 (1957)

2. DE EINDIGE ELEMENTENMETHODE VOOR HET DISCRETISEREN VAN ELLIPTISCHE RANDWAARDEPROBLEMEN.

De eindige elementenmethode is een numerieke techniek die het mogelijk maakt de discretisering van randwaardeproblemen systematisch uit te voeren. Aanvankelijk werd de eindige elementenmethode in het begin van de jaren 50 ontwikkeld door civiel ingenieurs om gecompliceerde problemen uit de sterkteleer (structural mechanics) op te lossen. In latere jaren vond de methode een ruimer toepassingsgebied en werd het verband duidelijk tussen de eindige elementenmethode en de klassieke Ritz-methode. Sindsdien (na ± 1965) is het duidelijk geworden dat bepaalde principes uit de methode veel algemener toegepast kunnen worden en numeriek wiskundigen hebben een vrij hechte fundering voor en een groot aantal uitbreidingen van de methode gevonden.

Het is de bedoeling van dit hoofdstuk van het colloquium de wiskundige grondslagen van de methode in zijn eenvoudigste vorm te beschrijven.
Een korte samenvatting van een aantal technieken die in verband met de eindige elementenmethode gebruikt worden, zal in het volgende hoofdstuk gegeven worden.

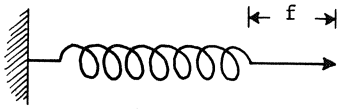
2.1. Inleiding

De eindige elementenmethode kan op twee verschillende manieren benaderd worden, van de fysische kant en van de mathematische kant.

Fysische beschrijving

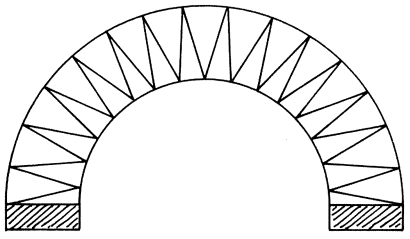
Deze beschrijving heeft een nauw verband met de oorspronkelijke formulering van de eindige elementenmethode in de sterkteleer. Een aantal termen, zoals stijfheidsmatrix en massamatrix, worden hieraan trouwens ontleend.
We beschouwen eerst een eenvoudige lineair elastische structuur. Voor het

ééndimensionale geval (een enkele plaatscoördinaat) beschouwen we een veer.



Hiervoor geldt f = Ku (wet van Hooke), waarin f een kracht is die op de veer wordt uitgeoefend, K een constante en u de verplaatsing van het uiteinde is. De potentiële energie van de uitgetrokken veer bedraagt $E = \frac{1}{2} f u$.

Wanneer we een meer gecompliceerde structuur beschouwen, zoals hieronder aangegeven,

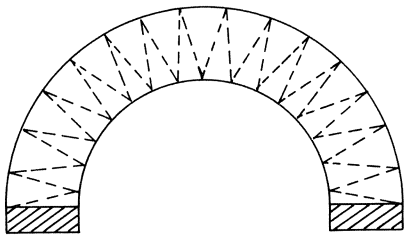


dan bestaat dezelfde relatie f = Ku.
Hierin is f een vector van krachten, uitgeoefend op de verbindingspunten in de structuur; u is een vector van plaatscoördinaten (de afwijkingen van de verbindingspunten uit de oorspronkelijke toestand) en K is een matrix (de stijfheidsmatrix van het systeem). De inverse matrix $M = K^{-1}$ wordt wel flexabiliteitsmatrix genoemd. De potentiële energie van de structuur wordt gegeven door $E = \frac{1}{2} u^t f = \frac{1}{2} u^t Ku$.
De matrix K is symmetrisch (actie = reactie) en positief definiet. Wanneer de constructie in evenwicht is (d.w.z. in een stationaire toestand verkeert), dan wordt u bepaald door de eis

$$E(u) = \frac{1}{2} u^t Ku$$

is minimaal.

Wanneer de eindige elementenmethode gebruikt wordt voor een elastisch continuum, kunnen we als volgt te werk gaan:



(i) Het continuum wordt verdeeld in denkbeeldige *elementen*.

(ii) De waarden van u op de hoekpunten van de elementen (de steunpunten) beschrijven het gedrag van de gehele constructie.

(iii) Binnen de elementen wordt een *interpolatieformule* gekozen om de toestand te beschrijven.

(iv) *Uitwendige krachten* en *randvoorwaarden* worden geconcentreerd op de hoekpunten gedacht.

Bovendien moet nu aan de volgende voorwaarde *(comptability condition)* voldaan worden: de interpolatieformules dienen zodanig te zijn dat op de gemeenschappelijke randen (tussen de elementen) hetzelfde interpolatieresultaat verkregen wordt.

Wiskundige beschrijving.

Een (vector-) functie u(x), gedefinieerd op een gebied $\Omega \subset \mathbb{R}^n$, moet gevonden worden, zodanig dat een convexe functionaal E(u) geminimaliseerd wordt. Deze functie wordt, voor numerieke doeleinden, benaderd door een functie u_h uit een eindigdimensionale functieruimte V_h met basis $\{\phi_i\}$:

$$u_h(x) = \sum_i a_i \phi_i(x).$$

De eindige elementenmethode benadert nu de functie $u \in V$ met die functie $u_h \in V_h$ waarvoor geldt

$$E(u_h) \leq E(v) \text{ , voor alle } v \in V_h.$$

Hiermee is de eindige elementenmethode in grote lijnen equivalent met de klassieke Ritz-methode. De eindige elementenmethode (the *nodal finite element method)* geeft bovendien echter suggesties voor een verstandige keuze van de functies $\{\phi_i\}$. Deze functies ϕ_i worden namelijk als volgt geconstrueerd:

Eerst wordt het gebied Ω in een aantal deelgebieden of elementen (meestal driehoeken of vierhoeken) verdeeld, zodanig dat naast elkaar liggende elementen de hoekpunten (i.h.a. steunpunten op gemeenschappelijke zijden) gemeenschappelijk hebben. Hierna worden de functies ϕ_i gekozen zodat

(i) een voldoende nauwkeurige interpolatie van een functie over het gebied Ω mogelijk wordt;

(ii) de drager van iedere functie ϕ_i in een zo klein mogelijk aantal elementen bevat is;

(iii) in elk steunpunt x_j de functiewaarden en de afgeleiden (tot een zekere orde) van alle ϕ_i gelijk aan nul zijn, behalve voor één ϕ_i waarvan óf de functiewaarde óf een afgeleide gelijk aan 1 is. In een formule kan dit aangegeven worden door

$$D_j\, \phi_i(x_j) = \delta_{ij}.$$

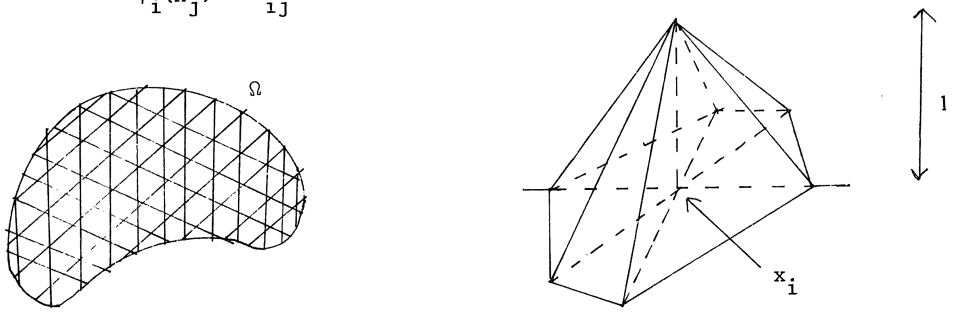
δ_{ij} is hierin de Kronecker-delta. D_j is een differentiaaloperator van nulde orde ($D_j u = u$) of van een hogere orde $\partial/\partial x$, $\partial/\partial y$, $(\partial/\partial x)^2$ e.d. . We merken op dat één steunpunt gecombineerd kan worden met verschillende differentiaaloperatoren; de index j definieert een paar (x_j, D_j). Deze keuze van de functies ϕ_i heeft het voordeel dat de parameters a_i, die in het proces berekend worden, direkt te interpreteren zijn als functiewaarden van $u_h(x)$ of als waarden van een afgeleide.

(iv) afhankelijk van de eigenschappen van de functionaal E(u) moet geëist worden dat alle ϕ_i continu zijn op Ω.

Voorbeeld

Een keuze van $\{\phi_i\}$ die geschikt is voor 2^e orde elliptische randwaardeproblemen en die lineaire interpolatie mogelijk maakt op een 2-dimensionaal gebied Ω, wordt gegeven door de "dak-functies". Hiervoor wordt het gebied overdekt met driehoekige elementen en wordt voor elk steunpunt x_i (hoekpunt) een functie ϕ_i gekozen, lineair op elk element en zodanig dat

$$\phi_i(x_j) = \delta_{ij} \ .$$



Figuur 2.1 Partitie van Ω; grafiek van $\phi_i(x)$.

Hoe een randwaardeprobleem waarin een tweede orde differentiaaloperator voorkomt, benaderd kan worden met functies die niet twee maal differentieerbaar zijn, zal in het volgende hoofdstuk behandeld worden

2.2. Verschillen tussen de eindige differentiemethode en de eindige elementenmethode

Het belangrijkste verschil tussen enerzijds de eindige differentiemethode en anderzijds de eindige elementenmethode en andere Ritz-Galerkin-achtige methoden is de wijze waarop de oplossing gepresenteerd en de continue operator gediscretiseerd wordt.
De oplossing u, van het oorspronkelijke probleem

$$Lu = f$$

kan beschouwd worden als een element van een genormeerde lineaire ruimte V, welke bestaat uit functies die gedefinieerd zijn op een deelgebied $\Omega \subset \mathbb{R}^n$. Bij de eindige differentiemethode worden u en f gediscretiseerd door de discretiseringsoperator $[.]_h$. Deze voegt aan de functie u, gedefinieerd op Ω, een functie $[u]_h$ toe die gedefinieerd is op een discrete puntverzameling, n.l. op het rooster Ω_h. We kunnen zeggen dat het definitiegebied gediscretiseerd wordt en dat de functies u en f door de discretiseringsoperator worden aangepast. De operator L_h wordt hierna (door"intelligent proberen") samengesteld en de doeltreffendheid van L_h wordt nagegaan aan de hand van de begrippen consistentie en stabiliteit.

Bij de Ritz-Galerkin-achtige methoden blijft het definitiegebied Ω in principe ongewijzigd. De genormeerde ruimte V waartoe u behoort, wordt

echter benaderd door een eindigdimensionale functieruimte V_h. Er wordt een restrictie $r_h: V \to V_h$ gedefinieerd en een extensie $p_h: V_h \to V$.

Op analoge wijze wordt de genormeerde ruimte W, waartoe f behoort, benaderd met W_h en worden een restrictie $s_h: W \to W_h$ en een extensie $q_h: W_h \to W$ gedefinieerd.
Elke operator $L: V \to W$ krijgt nu op eenduidige wijze een discreet equivalent $L_h: V_h \to W_h$ door de definitie

$$L_h = s_h L p_h .$$

De doeltreffendheid van het proces wordt nu geanalyseerd aan de hand van de eigenschappen van de restricties en extensies. Bij de Ritz-Galerkin-achtige methoden moet de constructie van de restricties en extensies door "intelligent proberen" tot stand gebracht worden.

Een tweede opvallend verschil tussen de eindige differentiemethode en de eindige elementenmethode is, dat bij de eindige differentiemethode meestal wordt uitgegaan van het randwaardeprobleem in de operatorvorm $Lu = f$ terwijl men bij de eindige elementenmethode uitgaat van een equivalent variatieprobleem. In de paragrafen 2.4 en 2.5 zal hierop verder worden ingegaan.

2.3. Discrete benadering van een genormeerde lineaire ruimte; Convergentie en stabiliteit

Zoals we in de vorige paragraaf reeds zagen, wordt de oplossing $u \in V$ van een randwaardeprobleem, bij de eindige elementenmethode benaderd door een element $u_h \in V_h$. Om de eigenschappen van zo'n benadering te analyseren, willen we niet alleen een element $v \in V$ benaderen met een element v_h uit *een bepaalde* andere genormeerde ruimte, maar we willen nagaan hoe v benaderd kan worden door elementen uit een klasse genormeerde ruimten.

Voorbeeld 2.3.1.

We willen niet alleen weten hoe een bepaalde continue functie op het interval [0,1] benaderd kan worden door een stapfunctie met N vast gekozen sprongpunten, we willen ook weten in welke mate de benadering beter wordt wanneer we het aantal sprongpunten vergroten.

We zullen daarom een genormeerde lineaire ruimte V benaderen met een klasse van genormeerde lineaire ruimten $\{V_h\}_{h \in H}$. Bij de index h denken we

aan een maaswijdte en we laten een limietovergang $h \to 0$ toe. Formeel kan als indexverzameling H bijvoorbeeld gekozen worden: een omgeving van de oorsprong in een eindig dimensionale vectorruimte.

Op twee manieren kunnen we nu een element $u_h \in V_h$ vergelijken met het element $u \in V$. We kunnen een afbeelding $r_h: V \to V_h$, de restrictie van V tot V_h, construeren en r_h u vergelijken met u_h of we kunnen een afbeelding $p_h: V_h \to V$, de extensie van V_h tot V, construeren en p_h u_h vergelijken met u. We voeren beide afbeeldingen, r_h en p_h, in en we definiëren de benadering van een genormeerde ruimte.

Definitie 2.3.1

De benadering van een genormeerde lineaire ruimte V bestaat uit een verzameling tripels $\{V_h, p_h, r_h\}_{h \in H}$ waarin

(i) V_h een genormeerde lineaire ruimte is;

(ii) p_h een begrensde lineaire injectie van V_h is, en

(iii) r_h een begrensde lineaire surjectie van V op V_h is.

Opmerking 2.3.1

Bij het oplossen van randwaardeproblemen zal V dikwijls een Sobolev-ruimte zijn en we zullen voor V_h eindig dimensionale functie ruimten kiezen, waarvan de dimensie afhankelijk is van h. Bovendien kiezen we veelal $V_h \subset V$.

Definitie 2.3.2

Voor een gegeven $h \in H$ heet voor ieder $u \in V$, $u_h \in V_h$

(i) $\|u-p_h u_h\|_V$ het verschil tussen u en u_h;

(ii) $\|u_h-r_h u\|_{V_h}$ het discrete verschil tussen u en u_h;

(iii) $\|u-p_h r_h u\|_V$ de afbreekfout van u.

Definitie 2.3.3 De norm van een operator.

Is p een afbeelding van een genormeerde ruimte V in een genormeerde ruimte W, dan is de norm van p gedefinieerd door

$$\|p\| = \sup_{v \in V,\ \|v\|_V \leq 1} \|pv\|_W .$$

Definitie 2.3.4 Stabiele restricties en extensies.

Restricties r_n en extensies p_h heten stabiel wanneer de normen $\|r_h\|$ resp. $\|p_h\|$ uniform begrensd zijn met betrekking tot h.

Definitie 2.3.5 Stabiele benadering.

We noemen een benadering $\{V_h, p_h, r_h\}_{h \in H}$ van een ruimte V stabiel als alle restricties en extensies stabiel zijn.

Definitie 2.3.6 Convergente benadering.

We noemen een benadering $\{V_h, p_h, r_h\}_{h \in H}$ van een ruimte V convergent als voor elke $u \in V$ geldt

$$\lim_{h \to 0} \|p_h r_h u - u\|_V = 0.$$

Opmerking 2.3.2

Als $\{V_h, p_h, r_h\}_{h \in H}$ een stabiele benadering van V is, dan is deze benadering reeds convergent voor V als hij convergent is voor een dichte deelverzameling van V. Zie hiervoor CÉA [1964].

Definitie 2.3.7

We zeggen dat een rij $\{u_h\}$ *convergeert* naar u, als geldt

$$\lim_{h \to 0} \|u - p_h u_h\|_V = 0.$$

We zeggen dat een rij u_h *discreet convergeert* naar u, als geldt

$$\lim_{h \to 0} \|u_h - r_h u\|_{V_h} = 0.$$

Stelling 2.3.1

Als de benadering $\{V_h, p_h, r_h\}_{h \in H}$ van de ruimte V stabiel en convergent is, dan is elke discreet convergente rij $\{u_h\}$ ook convergent.

Bewijs

$$\|u-p_h u_h\|_V \leq \|u-p_h r_h u\|_V + \|p_h r_h u-p_h u\|_V$$

$$\leq \|u-p_h r_h u\|_V + \|p_h\| \, \|r_h u-u\|_{V_h} .$$

De eerste term nadert tot nul vanwege de convergentie van de benadering; $\|p_h\|$ is uniform begrensd voor $h \to 0$. Hieruit volgt de stelling direkt. □

Voorbeeld 2.3.2

Als V een separabele Hilbert-ruimte is met een basis $\{\phi_i\}_{i \in \mathbb{N}}$, dan kunnen we een benadering $\{V_h, p_h, r_h\}_{h = 1/n, n \in \mathbb{N}}$ construeren als volgt:

(i) V_h is de ruimte opgespannen door $\{\phi_i\}_{i=1}^{n}$;

(ii) p_h is de kanonieke injectie $V_h \to V$;

(iii) r_h is de projectie van V op V_h.

Het is eenvoudig in te zien dat deze benadering stabiel en convergent is.

Definitie 2.3.8 Discrete norm.

Door een extensie $p_h: V_h \to V$ wordt op V_h een norm $\|\cdot\|_h$ geinduceerd door

$$\|u_h\|_h = \|p_h u_h\|_V .$$

Deze norm wordt de *discrete norm* (geinduceerd door p_h) genoemd.
Voor de discrete norm op V_h geldt $\|p_h\| = 1$.

Approximatie van operatoren.

Laten V en W twee genormeerde lineaire ruimten zijn en laat $L: V \to W$ een begrensde lineaire operator zijn. We nemen aan dat voor elke $f \in W$ er één oplossing bestaat voor de vergelijking

$$Lu = f$$

(m.a.w. L is een isomorfisme).
We willen nu een discreet analogon van deze vergelijking construeren. Hiertoe definiëren we een discrete operator en een discreet rechterlid f_h.
We maken hierbij gebruik van de discrete norm $\|\cdot\|_{u_h}$ op V_h.

Definitie 2.3.9

Zij $\{V_h, p_h, r_h\}_{h\in H}$ een benadering van V en zij $\{W_h, q_h, s_h\}_{h\in H}$ een benadering van W, dan wordt de *discrete operator* L_h, behorend bij de continue operator L, gedefinieerd door

$$L_h = s_h L p_h.$$

Het *discrete rechterlid* $f_h \in W_h$, behorend bij het rechterlid f, wordt gedefinieerd door

$$f_h = s_h f.$$

Definitie 2.3.10

De discrete operator L_h heet stabiel als er een constante $S > 0$ bestaat, onafhankelijk van h, zodat

$$\|p_h u_h\|_V \leq S \|L_h u_h\|_{W_h}$$

voor elke $h \in H$ en $u_h \in V_h$.

Opmerking 2.3.3

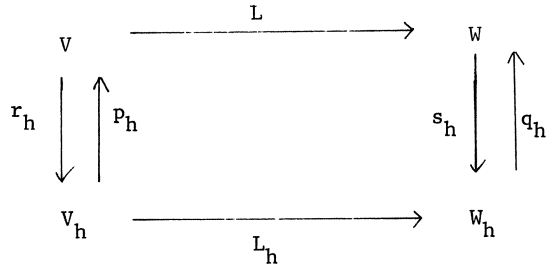
Als $\{W_h, g_h, s_h\}$ een stabiele benadering is van W, dan geldt voor een stabiele L_h

$$S^{-1} \|p_h u_h\|_V \leq \|L_h u_h\|_{W_h} = \|s_h L p_h u_h\|_{W_h}$$
$$\leq \sup_{h\in H} \|s_h\| \cdot \|L\| \cdot \|p_h u_h\|_V.$$

In het bijzonder impliceert dit dat L_h een isomorfisme is van V_h (met de discrete norm geinduceerd door p_h) op W_h.

Gevolg

Als de discrete operator L_h stabiel is en de benadering van W is stabiel, dan heeft de discrete vergelijking $L_h u_h = f_h$ een eenduidige oplossing.



Figuur 2.2

Stelling 2.3.2

Laat $u \in V$ de oplossing zijn van $Lu = f$, $f \in W$ willekeurig, en laat $u_h \in V_h$ de oplossing zijn van de gediscretiseerde vergelijking $L_h u_h = f_h$.

Als (i) de discrete operatoren L_h stabiel zijn;

(ii) $\{W_h, q_h, s_h\}_{h\in H}$ een stabiele benadering is van W;

(iii) $\{V_h, p_h, r_h\}_{h\in H}$ een convergente benadering is van V.

Dan (i) bestaat er een eenduidige oplossing voor de discrete vergelijking;

(ii) convergeert de discrete oplossing naar de continue oplossing, in die zin dat

$$\lim_{h\to 0} p_h u_h = u.$$

Bewijs

(i) volgt direkt uit opmerking 2.3.3.

(ii) leiden we af met behulp van de volgende ongelijkheid

$$\|u_h - r_h u\|_h = \|p_h(u_h - r_h u)\|_V \leq S \|L_h(u_h - r_h u)\|_{W_h} =$$

$$= S \|s_h L u - s_h L p_h r_h u\|_{W_h} \leq S \sup_H \|s_h\| \|L\| \|u - p_h r_h u\|_V.$$

Uit de stabiliteit van de benadering van W en uit de convergentie van de benadering van V volgt dat u_h discreet convergeert naar u. Door toepassen van stelling (2.3.1) volgt de convergentie. □

2.4. Randwaardeproblemen en convexe functionalen

Eén van de belangrijkste principes die aan de klassieke eindige elementenmethode ten grondslag liggen is, dat een lineaire differentiaalvergelij-

king (met homogene randvoorwaarden) van de vorm

(2.4.1) $\quad Lu = f$

in verband gebracht kan worden met de kwadratische functionaal

(2.4.2) $\quad E(v) = (Lv,v) - 2(f,v).$

De oplossing van de vergelijking in operatorvorm (2.4.1), met een positief definiete operator L, is precies het element dat de functionaal (2.4.2) minimaliseert. In veel toepassingen - i.h.b. waar stationaire toestanden van een fysisch systeem beschreven worden - geeft de formulering in termen van de functionaal het primaire fysische principe weer, terwijl de differentiaalvergelijking slechts een gevolg hiervan is. Hierdoor is het niet verbazingwekkend dat, in dat geval, het minimaliseren van de convexe functionaal het eigenlijke probleem is waarvan de oplossing benaderd moet worden. We geven hieronder enige voorbeelden van functionalen en de bijbehorende differentiaalvergelijkingen.

Voorbeeld 2.4.1 De vergelijking van Laplace.

Zij $\Omega \subset \mathbb{R}^2$ met rand $\partial\Omega$, $u \in C^2(\Omega)$ en
zij $u(x) = g(x)$ op $\partial\Omega$. De oplossing van

(2.4.3) $\quad u_{xx} + u_{yy} = 0$

minimaliseert de functionaal

(2.4.4) $\quad E(u) = \int_{\Omega} (u_x^2 + u_y^2)\, d\Omega.$

Voorbeeld 2.4.2 De doorbuiging van een vastgeklemde plaat.

Een kracht loodrecht op de plaat wordt gegeven door $g(x,y)$.
De randwaarden zijn $u = u_n = 0$ op $\partial\Omega$.
De oplossing van

(2.4.5) $\quad u_{xxxx} + 2\, u_{xxyy} + u_{yyyy} = g(x,y)$

minimaliseert de functionaal

$$(2.4.6) \qquad E(u) = \int_{\Omega} u_{xx}^2 + 2\,u_{xy}^2 + u_{yy}^2 - 2\,g(x,y)u\,d\Omega.$$

Voorbeeld 2.4.3 Straling (e^u) en moleculaire diffusie (u^2).

Zij $u(x) = g(x)$ op $\partial\Omega$.
De oplossing van

$$u_{xx} + u_{yy} = e^u$$

resp.

$$u_{xx} + u_{yy} = u^2$$

minimaliseert de functionaal

$$E(u) = \int u_x^2 + u_y^2 + 2\,e^u\,d\Omega$$

resp.

$$E(u) = \int u_x^2 + u_y^2 + \tfrac{2}{3}\,u^3\,d\Omega.$$

Voorbeeld 2.4.4

De oplossing van de lineaire elliptische differentiaalvergelijking

$$(2.4.7) \qquad Lu \equiv (au_x+bu_y)_x + (bu_x+cu_y)_y + fu = g,$$
$$b^2 - ac < 0 \text{ op } \Omega,$$

gedefinieerd op een begrensd gebied $\Omega \subset \mathbb{R}^2$ met gladde rand $\partial\Omega$, met de natuurlijke randvoorwaarde

$$\alpha u + \gamma + \beta\,u_n + \delta\,u_s = 0 \quad \text{op } \partial\Omega,$$

waarin

u_n = de inwendige normaalafgeleide,
u_s = de positieve tangentiële afgeleide,
$\beta = -a\,y_s^2 + 2\,b\,x_s\,y_s - c\,x_s^2$,
$\delta = (a-c)\,x_s\,y_s + b\,(y_s^2 - x_s^2)$,

is equivalent met het minimaliserend element van de kwadratische functionaal

$$E(u) = \int_{\Omega} (au_x^2+2bu_xu_y+cu_y^2-fu^2+2gu)\, dxdy + \int_{\partial\Omega} (\alpha u^2+2\gamma u)\, ds. \tag{2.4.9}$$

Voor het bewijs hiervan, zie FORSYTH & WASOW [1964].

Inhomogene randvoorwaarden

Voor de behandeling van inhomogene randvoorwaarden introduceren we, in het bijzonder, een eendimensionaal modelprobleem. We beschouwen het volgende tweepunts randwaardeprobleem, gedefinieerd op $I = [a,b] \subset \mathbf{R}^1$:

$$Lu \equiv -\frac{d}{dx}\left(p(x)\frac{du}{dx}\right) + q(x)\, u = f(x), \tag{2.4.10}$$

$$p(x) \geq p_0 > 0,\ q(x) \geq 0,\ x \in I,$$

$$-u'(a) + \alpha\, u(a) = \beta \tag{2.4.11}$$

$$u'(b) + \gamma\, u(b) = \delta,\ \alpha, \gamma \geq 0.$$

De bijbehorende kwadratische functionaal wordt gegeven door

$$E(v) = \int (p(v)^2+qv^2)\, dx + \alpha\, p(a)\, v^2(a) + \gamma\, p(b)\, v^2(b) - 2\int fv\, dx \qquad -2\beta\, p(a)\, v(a) - 2\delta\, p(b)v(b).$$

De lezer kan op eenvoudige wijze nagaan dat de oplossing van het tweepunts randwaardeprobleem de functionaal minimaliseert.

De kleinste kwadratenmethode

We moeten opmerken dat de functionaal die bij elk probleem genoemd is, niet de enige is waarvan het minimum met de oplossing van het randwaardeprobleem overeenkomt. Zo kan de oplossing $u \in V$ van ieder randwaardeprobleem (zelfgeadjungeerd of niet)

$$Lu = f$$

beschouwd worden als het minimaliserend element van de functionaal

$$E(u) = \|Lu-f\|^2 .$$

Het is ook zeer wel mogelijk om in dit geval de oplossing u te benaderen met het minimaliserend element $u_h \in V_h$.
Wanneer we de randvoorwaarden even verwaarlozen, is de functionaal

$$(2.4.13) \quad E(u) = \|Lu-f\|^2 = (Lu,Lu) - 2(f,Lu) + (f,f) = \\ = (L^TLu,u) - 2(L^Tf,u) + (f,f).$$

De laatste term is onafhankelijk van u en daarom minimaliseert de kleinste kwadraten methode in feite

$$(2.4.14) \quad (L^TLu,u) - 2(L^Tf,u).$$

Dit is precies de Ritz-functionaal voor het probleem

$$(2.4.15) \quad L^TLu = L^Tf.$$

Het nieuwe probleem is automatisch zelfgeadjungeerd. De orde van de vergelijking is echter verdubbeld. Dit laatste heeft belangrijke consequenties voor de implementatie van de randvoorwaarden. Voor een behandeling van dit probleem zij verwezen naar BRAMBLE & SCHATZ [1970,1971].
Het zal duidelijk zijn dat het kleinste kwadraten probleem ook numeriek slechter geconditioneerd is dan het oorspronkelijke.

2.5. De energieruimte

Het is de bedoeling van deze paragraaf na te gaan op welke wijze het minimaliserend element van een kwadratische functionaal samenhangt met het oplossen van een operatorvergelijking.

Bij de voorbeelden die in de vorige paragraaf gegeven werden, was het eenvoudig in te zien dat de oplossing van de differentiaalvergelijking de bijbehorende functionaal minimaliseert. Wanneer we echter het minimum van een functionaal willen vinden, moeten we ons eerst afvragen *over welke ruimte* we willen minimaliseren. Het is dan de vraag of op deze ruimte het *infimum aangenomen* wordt.

In de eerste plaats gaan we na hoe men het definitiegebied moet kiezen, opdat het infimum bereikt wordt. Als dit gebied uitgebreid is, dan bestaat er een oplossing (althans in gegeneraliseerde zin) en dan moet een effectieve manier worden aangegeven om deze oplossing te berekenen.

Voorbeeld 2.5.1

Om de vergelijking (2.4.3) op te lossen moeten we een functie u vinden waarvoor u_{xx} en u_{yy} bestaan, terwijl we bij het zoeken naar het minimum van (2.4.4) ook functies toelaten waarvoor u_{xx} en u_{yy} niet bestaan. Voor het bestaan van de functionaal behoeven u_x en u_y slechts kwadratisch integreerbaar te zijn en deze voorwaarde is (lang) niet voldoende om het bestaan van u_{xx} en u_{yy} te verzekeren.

We zullen ook zien dat het definitiegebied van de uitgebreide functionaal ook functies kan bevatten die niet aan de randvoorwaarden voldoen. Daarom moet ook duidelijk worden gemaakt wanneer en in welke zin de minimaliserende functie *wél* aan de randvoorwaarde voldoet.

In deze paragraaf is H een Hilbertruimte over het lichaam der reële getallen, met inproduct $(\cdot,\cdot)$ en norm $\|\cdot\|$. D is een dichte deelruimte uit H en L is een lineaire operator $L: D \to H$.

Definitie 2.5.1

Een bilineaire functionaal Φ op D is een afbeelding $\Phi: D \times D \to \mathbb{R}$ zodanig dat voor alle $\alpha,\beta \in \mathbb{R}$ geldt

$$\Phi(\alpha v+\beta w,u) = \alpha\Phi(v,u) + \beta\Phi(w,u), \qquad (2.5.1)$$
$$\Phi(v,\alpha w+\beta u) = \alpha\Phi(v,u) + \beta\Phi(v,w).$$

Definitie 2.5.2

Een bilineaire functionaal heet symmetrisch, als geldt

$$\Phi(u,v) = \Phi(v,u) \quad , \quad \text{voor alle } u,v \qquad (2.5.2)$$

Definitie 2.5.3

Een bilineaire functionaal heet begrensd als er een $\beta > 0$ bestaat zodat

$$|\Phi(u,v)| < \beta \|u\| \|v\| . \qquad (2.5.3)$$

Definitie 2.5.4

Is Φ een bilineaire functionaal en is ℓ een lineaire functionaal, dan heet

(2.5.4) $\quad E(u) = \Phi(u,u) - \ell(u)$

een kwadratische functionaal.

Voorbeelden

De functionalen E(u) die geïntroduceerd werden in de voorbeelden 2.4.1, 2.4.2 en 2.4.4 zijn kwadratische functionalen, evenals de functionaal E(u) in vergelijking (2.4.12).

Definitie 2.5.5.

De bilineaire functionaal Φ: $D \times D \to \mathbb{R}$ heet positief wanneer

(2.5.5) $\quad \Phi(u,u) > 0, u \neq 0$
$\quad\quad\quad \Phi(0,0) = 0.$

Definitie 2.5.6

De bilineaire functionaal heet positief definiet wanneer er een $\gamma > 0$ bestaat zodat voor alle u geldt

(2.5.6) $\quad \Phi(u,u) \geq \gamma^2 \|u\|^2.$

Definitie 2.5.7

De operator L heet symmetrisch (begrensd, positief, positief definiet) als de bilineaire functionaal (Lu,v) symmetrisch (begrensd, positief, positief definiet) is.

Opmerking 2.5.1

Wanneer H een Hilbertruimte is over het lichaam der complexe getallen (i.p.v. over $\mathbb{R}$), dan kan men, met overeenkomstige definities voor positieve en symmetrische operatoren, bewijzen dat iedere positieve operator symmetrisch is. Om deze reden zullen wij in het volgende alle positieve operatoren symmetrisch veronderstellen.

Stelling 2.5.1

Φ is een begrensde lineaire functionaal op D dan en slechts dan als er een éénduidige begrensde lineaire operator L op D bestaat zodat

$$\Phi(u,v) = (Lu,v).$$

Bewijs

Elementair, m.b.v. de stelling van Riesz; zie bijv. STUMMEL [1969] p.12. □

Stelling 2.5.2

Zij $L: D \to H$ een positieve operator, dan bestaat er voor iedere $f \in H$ *hoogstens één* oplossing $u \in D$ zodat

$$Lu = f.$$

Bewijs Elementair. □

Stelling 2.5.3

Zij $L: D \to H$ een positieve operator, dan zijn de volgende twee beweringen equivalent

(i) er bestaat bij gegeven $f \in H$ één $u \in D$ zodat $(Lu,v) = (f,v)$ voor alle $v \in H$;

(ii) $u \in D$ minimaliseert de functionaal $E(v) = (Lv,v) - 2(f,v)$.

Bewijs

(i) $\Rightarrow$ (ii)

$E(u+h) - E(u) = 2(Lu,h) - 2(f,h) + (Lh,h) = (Lh,h) \geq 0.$

(ii) $\Rightarrow$ (i)

$E(u) \leq E(u+\lambda h) = (Lu,u) + 2\lambda(Lu,h) + \lambda^2(Lh,h) - 2(f,u) - 2\lambda(f,h) =$

$= E(u) + \lambda[\lambda(Lh,h)+2(Lu,h)-2(f,h)].$

Dus voor alle $\lambda \in \mathbb{R}$ geldt $\lambda[\lambda(Lh,h)+2(Lu,h)-2(f,h)] \geq 0$ zodat $(Lu,h) = (f,h)$. □

Opmerking 2.5.2

We doen hier geen uitspraak omtrent het *bestaan* van zo'n $u \in D$.

Stelling 2.5.4

Iedere positieve operator L induceert een prehilbertruimte-structuur op het definitiegebied D, door het inproduct

(2.5.7) $(u,v)_L = (Lu,v) = [u,v]$.

Bewijs Elementair. □

Gevolg

D is een genormeerde lineaire ruimte (D_L) met de norm

(2.5.8) $\|u\|_L = (Lu,u)^{\frac{1}{2}} = \|u\|$.

Definitie 2.5.8

De completering van de prehilbertruimte D_L wordt de *energieruimte* H_L genoemd.

Stelling 2.5.5

De energieruimte H_L van een positieve operator L is separabel d.e.s.d. als de oorspronkelijke ruimte H separabel is.

Bewijs MIKHLIN [1970], p.107. □

Opmerking 2.5.3

Aangezien D dicht ligt in H_L kan een aftelbare verzameling basisfuncties $\{\phi_n\}$ uit D gekozen worden.

Stelling 2.5.6

Als de operator L positief definiet is, dan is de prehilbertruimte D_L te completeren met elementen uit H; m.a.w. als L positief is dan geldt de volgende inclusierelatie

$$D_L \subset H_L \subset H.$$

Bewijs YOSIDA [1965], p.317. □

Opmerking 2.5.4

Als L positief definiet is dan volgt uit $\bar{D} = H$ dat ook $\bar{H}_L = H$ ($\bar{\cdot}$ de afsluitingsoperator).

Opmerking 2.5.5

Als L positief definiet is dan geldt voor alle $u \in H_L$

$$\gamma \|u\| \leq \|u\|_L .$$

Evenzo geldt voor alle $u \in D$ en $v \in H_L$

(2.5.7) $\quad (Lu,v) = (u,v)_L .$

Gevolg

Als L positief definiet is en $\ell(u)$ een begrensde lineaire operator op H, dan is $\ell(u)$ ook begrensd op H_L.

Opmerking 2.5.6

Als L positief definiet is en ℓ is begrensd op H_L, dan is ℓ niet noodzakelijk begrensd op H.

Stelling 2.5.7

Als de operator L positief is, maar niet positief definiet dan kan het definitiegebied van (Lu,u) eenduidig worden uitgebreid tot H_L; echter, H_L is niet noodzakelijk een deelverzameling van H. Er kan wel een injectie $H \to H_L$ geconstrueerd worden.

Bewijs. MIKHLIN [1965], p.14. □

Gevolg

Als L positief is, maar niet positief definiet, en ℓ is begrensd op H, dan is ℓ niet noodzakelijk begrensd op H_L.

Het variatieprobleem

Het variatieprobleem luidt als volgt: zij L een positieve operator gedefinieerd op een dichte deelverzameling $D \subset H$ en zij ℓ een lineaire functionaal (begrensd of onbegrensd) op H; bepaal u, zodanig, dat de kwadratische functionaal

$$E(u) = (Lu,u) - 2\ell(u) \tag{2.5.8}$$

minimaal is.

Door de energieruimte H_L in te voeren schrijven we vergelijking (2.5.3) als

$$E(u) = \|u\|_L^2 - 2\ell(u). \tag{2.5.9}$$

We onderscheiden nu twee mogelijkheden:

(i) ℓ is niet begrensd in H_L. In dit geval bestaat er geen minimum voor E(u) want er bestaat een rij $\{u_n\}$ zodat $\|u_n\|_L = 1$ en $\lim_{n\to\infty} \ell(u_n) = \infty$.

(ii) ℓ is begrensd in H_L.
Volgens de stelling van Riesz bestaat er dan één en slechts één $u_0 \in H_L$ zodat $\ell(u) = [u,u_0]$. Nu kunnen we schrijven

$$E(u) = \|u-u_0\|_L^2 - \|u_0\|_L^2, \tag{2.5.10}$$

zodat het minimum wordt bereikt voor $u_0 \in H_L$.

Opmerking 2.5.7

(i) Het definitiegebied D van de functionaal E wordt door de gelijkheid (2.5.10) uitgebreid tot H_L.

(ii) Ligt u_0 niet in D, dan kan men u_0 een *gegeneraliseerde* oplossing van het variatie probleem noemen.

(iii) Als L een positief definiete operator is op D en als ℓ begrensd is op H_L, maar onbegrensd op H, dan is u_0 geen oplossing maar wel een gegeneraliseerde oplossing van het variatieprobleem. Immers, als $u_0 \in D$ dan $\ell(u) = (u_0,u)_L = (Lu_0,u)$ voor alle $u \in H_L$, zodat ℓ begrensd zou zijn op H. Dit is een contradictie, waaruit volgt $u_0 \notin D$.

Opmerking 2.5.8

Zij $f \in H$, dan is $\ell(u) = (f,u)$ een eenvoudig voorbeeld van een begrensde lineaire operator ℓ.

Stelling 2.5.8 (LAX & MILGRAM)

Zij Φ een begrensde, positief definiete, bilineaire vorm gedefinieerd op de Hilbertruimte D, dan bestaat er precies één begrensde lineaire operator S op D zodat

(2.5.11) $(x,y) = \Phi(s,Sy)$ voor alle $x,y \in D$.

Deze operator heeft een inverse S^{-1} en er geldt $\|S\| < \gamma^{-2}$, $\|S^{-1}\| \leq \|\Phi\|$.

Bewijs Zie YOSIDA [1965], p.92-93. □

Gevolg

Zij Φ een begrensde, positief definiete, bilineaire vorm op D, dan bestaat er voor elke begrensde lineaire functionaal $\ell: D \to \mathbb{R}$ e.e.s.e. oplossing van

(2.5.12) $\Phi(x,w) = \ell(x)$, voor alle $x \in D$.

Bewijs

M.b.v. de stellingen van Riesz en Lax-Milgram volgt direkt dat er bij iedere $\ell: D \to \mathbb{R}$ precies één $p \in E$ bestaat met

$$\ell(x) = (x,p) = \Phi(x,Sp) = \Phi(x,w). \quad \square$$

De Ritz-methode

Als L positief definiet is en als H separabel is, dan kan - volgens opmerking (2.5.3) - uit D een rij $\{\phi_n\}$ gekozen worden die volledig is in H_L. De Ritz-methode bestaat nu daaruit dat het minimum van de functionaal E(u) niet gezocht wordt op H_L maar op een n-dimensionale deelruimte opgespannen door $\{\phi_k\}_{k=1}^n$. D.w.z. we schrijven

(2.5.13) $u_n = \sum_{k=1}^{n} a_k \phi_k$

en we zoeken het minimum van de functie in n variabelen

$$E(u_n) = (L \sum_{k=1}^{n} a_k \phi_k, \sum_{k=1}^{n} a_k \phi_k) - 2(f, \sum_{k=1}^{n} a_k \phi_k).$$

We zien direkt dat dit leidt tot het lineaire stelsel

$$\sum_{k=1}^{n} a_k (L\phi_k, \phi_m) = (f, \phi_m) \quad , \qquad m = 1(1)n.$$

Als nu de functies ϕ_k lineair onafhankelijk zijn in H_L dan is dit stelsel eenduidig oplosbaar.

Wanneer in (2.5.13) voor a_k de oplossing van het stelsel genomen wordt, dan heet u_n de *Ritz-benadering*. Wanneer L en f gegeven zijn en $\{\phi_k\}_{k=1}^{n}$ gekozen is, is deze benadering expliciet te berekenen.

Opmerking 2.5.9

Zoals reeds in 2.1 vermeld werd, is de klassieke eindige elementen methode hierin geheel identiek aan de Ritz-methode. Als extra eigenschap geeft de eindige elementen methode bovendien een suggestie voor de keuze van de functies $\{\phi_n\}$.

Stelling 2.5.9

Als L positief definiet is dan convergeren de Ritz-benaderingen u_n van het variatie probleem in H_L en in H naar de oplossing u_0.

Bewijs

Kies $\{\phi_k\}_{k=1}^{n}$ orthonormaal in H_L, dan geldt $a_k = (f,\phi_k)$ en dus $u_n = \sum_{k=1}^{n}(f,\phi_k)\phi_k$. Omdat $\{\phi_n\}$ volledig is volgt

$$u_0 = \sum_{k=1}^{\infty} (u_0,\phi_k)_L \phi_k = \sum_{k=1}^{\infty} (f,\phi_k)\phi_k,$$

waardoor

$$\lim_{n\to\infty} \|u_n - u_0\|_L = 0.$$

Aangezien $\|u_n - u_0\| \leq \gamma^{-1} \|u_n - u_0\|_L$ convergeert u_n naar u ook in H. □

2.6. Sobolev-ruimten

In de vorige paragrafen werd enerzijds gesproken in termen van genormeerde lineaire ruimten en Hilbert-ruimten, anderzijds werden voorbeelden gegeven van differentiaaloperatoren die gedefinieerd waren op ruimten van functies die een aantal malen differentieerbaar moeten zijn.

In deze paragraaf zullen we enkele concrete Hilbert-ruimten definiëren welke bestaan uit functies die - in zekere zin - voldoende differentieerbaar zijn. Dit stelt ons in staat preciese formuleringen van problemen en stellingen te geven. Voor een uitgebreide behandeling van o.a. Sobolev-ruimten zij verwezen naar YOSIDA [1965] en LIONS & MAGENES [1968].

Zij Ω een meetbare open deelverzameling van de eindigdimensionale Euclidische ruimte $\mathbb{R}^n$. De rand van Ω geven we aan met $\partial\Omega$.

Definitie 2.6.1

Met $C^k(\Omega)$, k geheel, $k \geq 0$, geven we aan de verzameling der reëelwaardige functies op Ω met continue (partiële) afgeleiden tot en met de k-de orde. Met $C^k(\bar{\Omega})$ geven we alle functies aaan waarvoor de eigenschap uniform geldt.

Definitie 2.6.2

Met $C_0^k(\Omega)$ geven we aan de verzameling van alle functies uit $C^k(\Omega)$ welke een compacte drager in Ω bezitten.

Opmerking 2.6.1

Met de gebruikelijke definities voor het optellen van twee functies en het vermenigvuldigen van een functie met een scalar zijn $C^k(\Omega)$ en $C_0^k(\Omega)$ lineaire ruimten.

Opmerking 2.6.2

Voor een natuurlijk getal $m \leq k$ en voor een willekeurige compacte drager $K \subset \Omega$ is de grootheid

$$(2.6.1) \qquad \sup_{\substack{|s|=m \\ x\in K}} |D^s f(x)|$$

een seminorm op de lineaire ruimte $C^k(\Omega)$ of $C_0^k(\Omega)$.

Opmerking 2.6.3

We maakten hier gebruik van de z.g. *multiindex-notatie:* $s = (s_1, s_2, \ldots, s_n)$, s_i een niet negatief geheel getal; $|s| = \sum_{i=1}^{n} s_i$;

$$D^s = \left(\frac{\partial}{\partial x_1}\right)^{s_1} \left(\frac{\partial}{\partial x_2}\right)^{s_2} \cdots \left(\frac{\partial}{\partial x_n}\right)^{s_n} .$$

Aldus wordt met D^s een $|s|$-de orde partiële differentiaal operator aangegeven.

Met $\mathcal{D}(\Omega)$ geven we de topologische vectorruimte aan die door formule (2.6.1) op $C_0^\infty(\Omega)$ geïnduceerd wordt.

We zeggen dat een bewering "bijna overal" (b.o.) geldt op een meetbare verzameling S, wanneer ze geldt op die verzameling S met uitzondering van een deelverzameling met maat nul.

In het volgende identificeren we een functie f met de equivalentieklasse van functies welke ontstaat door de equivalentierelatie $f(s) \sim g(s)$, als $f(s) = g(s)$ b.o..

Met $L^p(\Omega)$, $1 \leq p \leq \infty$, geven we aan de verzameling van alle reëelwaardige functies f, b.o. gedefinieerd op Ω, waarvoor

$$\int_\Omega |f(s)|^p \, ds$$

bestaat.

Opmerking 2.6.4

$\|f\|_{0,p} = \left(\int_\Omega |f(s)|^p \, ds\right)^{1/p}$ is een norm op $L^p(\Omega)$ en met deze norm is $L^p(\Omega)$ een Banach-ruimte.

Met $L^\infty(\Omega)$ geven we aan de verzameling van alle reëelwaardige, meetbare functies f, b.o. gedefinieerd op Ω, waarvoor

$$\|f\|_{0,\infty} = \operatorname*{ess\,sup}_{s \in \Omega} |f(s)| = \inf \{z \mid z \in \mathbb{R}, |f(s)| \leq z \text{ b.o.}\}$$

bestaat.

Opmerking 2.6.5

$\|f\|_{0,\infty}$ is een norm op $L^\infty(\Omega)$ en met deze norm is $L^\infty(\Omega)$ een Banach-ruimte.

Opmerking 2.6.6

$(x,y) = \int_\Omega x(s)y(s)ds$ is een inproduct op $L^2(\Omega)$ en met dit inproduct is $L^2(\Omega)$ een Hilbert-ruimte.

Opmerking 2.6.7

We weten dat $\mathcal{D}(\Omega)$ dicht ligt in iedere Banach-ruimte $L^p(\Omega)$, $1 \le p < \infty$.

Voor een willekeurig geheel getal $k \ge 0$ en voor elke p, $1 \le p \le \infty$, wordt een *Sobolev-ruimte* $W^{k,p}(\Omega)$ gedefinieerd als de verzameling van alle functies die, tezamen met hun distributionele afgeleiden tot en met de k-de orde, tot de ruimte $L^p(\Omega)$ behoren; ofwel

$$(2.6.2) \qquad W^{k,p}(\Omega) = \{f \mid D^s f \in L^p(\Omega),\ 0 \le |s| \le k\}.$$

Een seminorm op $W^{k,p}(\Omega)$ wordt gedefinieerd door

$$(2.6.3) \qquad |f|_{k,p} = \left(\sum_{|s|=k} \int_\Omega |D^s f(x)|^p \, dx \right)^{1/p}.$$

Een norm voor $W^{k,p}(\Omega)$ wordt gedefinieerd door

$$(2.6.4) \qquad \|f\|_{k,p} = \left(\sum_{0 \le |s| \le k} \int_\Omega |D^s f(x)|^p \, dx \right)^{1/p}.$$

Met deze norm is $W^{k,p}(\Omega)$ een Banach-ruimte.

In het bijzonder worden voor $p = 2$ de volgende notaties gebruikt

$$W^{k,2}(\Omega) = W^k(\Omega) = H^k(\Omega)$$

en

$$\|f\|_{k,2} = \|f\|_k.$$

Op $W^{k,2}(\Omega)$ kan een inprodukt gedefinieerd worden door

$$(2.6.5) \qquad (x,y)_k = \sum_{0 \le |s| \le k} \int_\Omega D^s x(t)\, D^s y(t)\, dt.$$

Met dit inprodukt is $W^{k,2}(\Omega)$ een Hilbert-ruimte. In het bijzonder noteren we $\|f\| = \|f\|_0$ en $(x,y) = (x,y)_0$. Ook voor de lineaire deelruimte van alle

functies uit $C^k(\Omega)$ of $C_0^k(\Omega)$, $k < \infty$, waarvoor geldt

$$(2.6.6) \qquad \|f\|_k = \left(\sum_{0 \le |s| \le k} \int_\Omega |D^s f(t)|^2 \, dt \right)^{\frac{1}{2}} < \infty$$

is $(x,y)_k$ een inprodukt (en daarmee $\|f\|_k$ een norm).

Uit de definitie van $H^k(\Omega)$ volgt direkt

$$C_0^\infty(\Omega) \subset H^k(\Omega) \subset H^m(\Omega) \subset H^0(\Omega) = L^2(\Omega)$$

en het is duidelijk dat de injectieve afbeeldingen

$$\mathcal{D}(\Omega) \to H^k(\Omega) \to H^m(\Omega) \to L^2(\Omega)$$

continu zijn.

Het kan worden aangetoond dat $C^\infty(\Omega)$ dicht ligt in $H^k(\Omega)$ als Ω een begrensd gebied is. D.w.z. ieder element uit $C^\infty(\Omega)$ waarvoor (2.6.6) geldt, kan worden opgevat als element van $H^k(\Omega)$ en in de norm $\|\cdot\|_k$ kan een element van $H^k(\Omega)$ willekeurig dicht benaderd worden door elementen uit $C^\infty(\Omega)$.

Hieruit volgt dat $H^k(\Omega)$ de afsluiting is in de norm $\|\cdot\|_k$ van alle functies uit $C^\infty(\Omega)$ waarvoor geldt dat de norm $\|\cdot\|_k$ eindig is.

Met $H_0^k(\Omega)$ wordt aangegeven de completering van $C_0^\infty(\Omega)$ in de norm $\|\cdot\|_k$.

In het algemeen geldt

$$H_0^k(\Omega) \subset H^k(\Omega),$$

maar men kan bewijzen dat

$$H_0^k(\mathbb{R}^n) = H^k(\mathbb{R}^n).$$

Ook met een negatief geheel getal als bovenindex geeft men betekenis aan de notatie $H^k(\Omega)$: met $k > 0$ is $H^{-k}(\Omega)$ de duale ruimte van $H_0^k(\Omega)$, d.w.z. $H^{-k}(\Omega)$ is de ruimte van alle begrensde lineaire operatoren op $H_0^k(\Omega)$.

Definitie 2.6.3

We noemen een gebied Ω k-regulier als voor de rand $\partial\Omega$ geldt: $\partial\Omega$ is een k maal continu differentieerbare varieteit van dimensie k-1, en het gebied Ω bevindt zich aan één kant van $\partial\Omega$,

Stelling 2.6.1 (Spoortheorema)

Laat Ω een open deelverzameling van $\mathbb{R}^n$ zijn en k-regulier, dan bestaan er k continue lineaire afbeeldingen $\gamma_0,\gamma_1,\ldots,\gamma_{k-1}$ van $H^k(\Omega)$ op $L^2(\partial\Omega)$ - d.i. de verzameling van alle kwadratisch integreerbare functies op $\partial\Omega$ m.b.t. de oppervlakte maat $d\sigma$ - zodat voor iedere u die k maal continu differentieerbaar is op $\bar{\Omega}$ geldt

$$\gamma_0 u = u \qquad \text{op } \partial\Omega ,$$
$$\gamma_s u = \left(\frac{\partial}{\partial\nu}\right)^s u \text{ op } \partial\Omega \qquad (s = 1,2,\ldots,k-1).$$

Hierin is ν de vector in $\mathbb{R}^n$ welke loodrecht staat op $\partial\Omega$ en naar het buitengebied van Ω wijst.

$H_0^k(\Omega)$ is de kern van de afbeelding

$$\gamma_0 \times \gamma_1 \times \cdots \times \gamma_{k-1},$$

d.w.z. voor $u \in H^k(\Omega)$ geldt

$$u \in H_0^k(\Omega) \iff \forall s \in \{0,1,\ldots,k-1\} \left(\frac{\partial}{\partial\nu}\right)^s u = 0.$$

Bewijs Zie LIONS & MAGENES [1968], p.68, th.11.5. □

Opmerking 2.6.8

$\gamma_j H^k(\Omega)$ is niet de gehele ruimte $L^2(\partial\Omega)$ maar een deelruimte van $L^2(\partial\Omega)$. Van belang is vooral de ruimte $\gamma_0 H^1(\Omega)$ welke wordt aangegeven met $L^{\frac{1}{2}}(\partial\Omega)$.

Stelling 2.6.2 (Sobolev's lemma)

Zij Ω een begrensde open deelverzameling van $\mathbb{R}^n$ en k-regulier, laat m een geheel getal zijn, $m > k + n/2$. Dan geldt voor iedere $u \in H^m(\Omega)$ dat er een $w \in C^k(\bar{\Omega})$ bestaat zodat $u = w$ b.o..

Meestal geven we dit - minder nauwkeurig - aan met

$$H^m(\Omega) \subset C^k(\bar{\Omega}) , \quad \text{als } m > k + n/2.$$

Opmerking 2.6.9

Dergelijke inclusiestellingen kunnen ook gegeven worden voor de algemenere Sobolev-ruimten $W^{k,p}(\Omega)$. Voor niet-negatieve gehele getallen k, j, $k \geq j$ geldt

(i) als $\frac{1}{p} \geq \frac{1}{q} \geq \frac{1}{p} - \frac{k-j}{n} > 0$,

dan $W^{k,p}(\Omega) \subset W^{j,q}(\Omega)$,

d.i. $\|u\|_{j,q} \leq C(j,k,p,q)\ \|u\|_{k,p}$ voor alle $u \in W^{k,p}$.

(ii) Als $\frac{1}{p} < \frac{k-j}{n}$,

dan $W^{k,p}(\Omega) \subset C^{j}(\Omega)$,

d.i. $\|u\|_{C^{j}(\Omega)} \leq C(j,k,p)\ \|u\|_{k,p}$.

2.7. Natuurlijke en essentiële randvoorwaarden

We beschouwen, evenals in paragraaf 2.4, de differentiaalvergelijking

$$-D(pDu) + qu = k, \tag{2.7.1}$$

gedefinieerd op $I = [a,b]$, met de randvoorwaarden

$$\begin{aligned} -Du(a) + \alpha u(a) &= \beta, \\ Du(b) + \gamma u(b) &= \delta. \end{aligned}$$

We nemen aan dat $p \in C^1[a,b]$, $q \in C^0[a,b]$, $k \in L^2[a,b]$ en bovendien dat $p \geq p_0 > 0$, $q \geq 0$; $\alpha,\beta,\gamma,\delta \in \mathbb{R}$, $\alpha \geq 0$, $\gamma \geq 0$ en niet $\alpha = \gamma = 0$. Dit is een randwaardeprobleem met gemengde randvoorwaarden, d.w.z. in de randvoorwaarden komen zowel Du als u voor. Een randvoorwaarde van de vorm $Du(x_b) = \delta$, waarin u niet voorkomt heet *Neumann-randvoorwaarde*.
N.B. Met α (resp. γ) gelijk aan nul zijn in bovenstaande formulering Neuman-randvoorwaarden inbegrepen. Een randvoorwaarde van de vorm $u(x_b) = \nu$, waarin Du niet voorkomt, heet *Dirichlet-randvoorwaarde*. In de bovenstaande formulering is deze soort randvoorwaarden uitgesloten.

We schrijven het randwaardeprobleem (2.7.1)-(2.7.2) in de operatorvorm

$$Lu = f$$

door middel van de operator

(2.7.3) $\quad L\cdot \equiv - D(pD\cdot) + q.$

Deze operator is gedefinieerd voor alle $u \in C^2[a,b]$. De oplossing u van ons probleem moet echter gevonden worden in de deelverzameling van $C^2[a,b]$ die voldoet aan (2.7.2). Deze deelverzameling is echter geen lineaire ruimte wanneer niet $\beta = 0$ en $\delta = 0$. We kiezen nu, alleen voor theoretische doeleinden, een willekeurige functie $w_0 \in C^2[a,b]$ die voldoet aan (2.7.2). We kunnen nu schrijven $u = w + w_0$. Om u te bepalen moet nu een $w \in C^2[a,b]$ gevonden worden die voldoet aan

(2.7.4) $\quad Lw = k - Lw_0 = f,$

(2.7.5) $\quad \begin{cases} -Dw(a) + \alpha w(a) = 0, \\ Dw(b) + \gamma w(b) = 0. \end{cases}$

De deelverzameling van alle $w \in C^2[a,b]$ die voldoen aan (2.7.5), is *wel* een lineaire deelruimte van $C^2[a,b]$; deze geven we aan met $C_B^2[a,b]$. Het probleem is nu gereduceerd tot het vinden van een $w \in C_B^2[a,b]$ zodat $Lw = f$.

Als grondruimte (pivot space)[H] voor ons probleem nemen we $L^2[a,b]$. Voor het definitiegebied D van de operator L nemen we $C_B^2[a,b]$. Dit definitiegebied ligt dicht in H; we weten namelijk dat

(2.7.6) $\quad C_0^\infty[a,b] \subset C_B^2[a,b] \subset L^2[a,b]$

in elkaar dichte deelverzamelingen zijn.

Omdat $p,q \in C^0[a,b]$ bestaan er $p_1, q_1 > 0$ zodat $p_1 \geq p \geq p_0$ en $q_1 \geq q \geq 0$. Hierdoor is L een begrensde lineaire operator $D \to H$.

De operator L is bovendien symmetrisch:

$$(2.7.7) \quad (Lw,v) = \int_a^b [-D(pDw)v + qwv]\, dx =$$

$$= \int_a^b [p\, Dw\, Dv + qwv]\, dx + [\alpha pwv](a) + [\gamma pwv](b).$$

Het rechterlid van (2.7.7) is symmetrisch in u en v zodat

(2.7.8) $\quad (Lw,v) = (w,Lv).$

Wanneer we in de ongelijkheid (2.7.7) w = v nemen, dan blijkt L ook positief definiet:

$$(Lw,w) = \int_a^b p(Dw)^2 + qw^2\,dx + [\alpha pw^2](a) + [\alpha pw^2](b) \geq \text{(Poincaré)} \geq$$

(2.7.9)

$$\geq \frac{p_0}{2(b-a)} * \min(\frac{1}{b-a},\max(\alpha,\gamma)) * \|w\|^2.$$

L is dus positief definiet.

Nu L symmetrisch en positief definiet blijkt te zijn kunnen we de energieruimte H_L introduceren. Volgens stelling (2.5.6) geldt $H_L \subset L_2[a,b]$.

Uit welke functies bestaat nu de ruimte H_L? In de eerste plaats bevat deze ruimte $C_B^2[a,b]$ en vervolgens alle functies die als limietfuncties van $C_B^2[a,b]$-functies beschouwd kunnen worden in de energienorm $\|\cdot\|_L$:

$$\|w\|_L^2 = \int_a^b p(Dw)^2 + qw^2\,dx + [\alpha pw^2](a) + [\gamma pw^2](b). \tag{2.7.10}$$

Lemma 2.7.1

Voor de operator L, gedefinieerd door (2.7.4), (2.7.5) zijn de normen $\|\cdot\|_L$ en $\|\cdot\|_1$ equivalent.

Bewijs

Uit (2.7.9) volgt $\|w\|_L^2 \geq p_0\|Dw\|^2$ en $\|w\|_L^2 \geq \|w\|_L^2 \geq K\|w\|^2$, zodat

$$\|w\|_L^2 > K_1\,\|w\|_1^2. \tag{2.7.11}$$

Anderzijds volgt uit (2.7.10)

$$\|w\|_L^2 \leq \max(p_1,q_1)\,\|w\|_1^2 + |\alpha p(a)w(a)^2| + |\gamma p(b)w(b)^2|$$

en uit opmerking (2.6.9) volgt

$$|\alpha p(a)w(a)^2| + |\gamma p(b)w(b)^2| \leq K_2\,\|w\|_1^2,$$

zodat

$$\|w\|_L^2 \leq K_3\,\|w\|_1^2\,. \tag{2.7.12}$$

Uit (2.7.11) en (2.7.12) volgt de equivalentie. □

Zoals beschreven werd in de vorige paragraaf, kan de betekenis van (Lw,w) nu worden uitgebreid van $w \in C_B^2[a,b]$ tot $w \in H^1[a,b]$. De oplossing van Lw = f is dus het minimaliserend element uit $H^1[a,b]$ van

(2.7.13) $\quad (Lw,w) - 2(f,w).$

De functie w_0 (zie (2.7.4)) behoort tot $C^2[a,b]$ en daarmee tot $H^1[a,b]$, zodat de oplossing u van de vergelijking Lu = k het minimaliserend element is van

$$(L(u-w_0),(u-w_0)) - 2(f,(u-w_0)).$$

Met behulp van (2.7.4) leiden we nu af dat u het minimaliserend element is, in $H^1[a,b]$, van

(2.7.14) $\quad (Lu,u) - 2(k,u) - (Lw_0,w_0) + 2(k,w_0).$

Deze kwadratische functionaal komt overeen met E(u) in vergelijking (2.4.12).

Opmerking 2.7.1

Het is een groot praktisch voordeel dat de klasse van functies zo uitgebreid is. We kunnen nu de benadering van de oplossing u zoeken in een eindigdimensionale deelruimte van $H^1[a,b]$. Het is nu bijvoorbeeld mogelijk u te benaderen met stuksgewijs lineaire functies.

Opmerking 2.7.2

We zien, dat aan de ruimte waarover geminimaliseerd wordt geen beperking wordt opgelegd door de randvoorwaarden (2.7.2). Deze randvoorwaarden zijn z.g. *natuurlijke randvoorwaarden* waaraan blijkbaar automatisch voldaan wordt door het minimum u uit H_L. Anders ligt de zaak, wanneer we aan één (of twee) van de randen i.p.v. een gemengde randvoorwaarde, een Dirichlet-randvoorwaarde opleggen. We merken op dat deze randvoorwaarde niet door (2.7.2) geïmpliceerd wordt. We kunnen zo'n randvoorwaarde probleem met een Dirichlet-voorwaarde, bijv. (2.7.1), (2.7.2)), namelijk voor $\alpha,\beta \to \infty$ met $\beta/\alpha = \nu$. We zien aan de functionaal E(u) (verg. 2.4.12) dat de termen

(2.7.12) $\alpha p(a)u^2(a) - 2\beta p(a)u(a)$

een extra groot gewicht krijgen voor $\alpha,\beta \to \infty$. Opdat E(u) geminimaliseerd wordt zal dus in de eerste plaats voldaan moeten worden aan de eis dat (2.7.12) minimaal is. Dit is equivalent met

(2.7.13) $u(a) = \nu$.

Wanneer hieraan voldaan is, moet u verder zodanig worden gevonden dat

(2.7.14) $$\int_a^b (p(Du)^2+qu^2-2fu)dx + p(b)(\gamma u^2(b) - w\delta u(b))$$

minimaal is.

In dit geval van een Dirichlet-randvoorwaarde zoeken we het minimaliserende element over slechts díe elementen u uit $H^1[a,b]$ waarvoor $u(a) = \nu$ Zo'n randvoorwaarde, die een eis stelt aan de ruimte waarin de oplossing van het minimaliseringsprobleem gezicht moet worden, heet (in tegenstelling tot een natuurlijke) een *essentiële randvoorwaarde*.

Opmerking 2.7.3

Uit het spoor theorema (stelling 2.6.1) volgt dat, voor elliptische problemen van de orde 2m, voorwaarden voor de k-de orde afgeleiden op de rand, optreden als essentiële randvoorwaarden, als $0 \le k \le m-1$ en als natuurlijke randvoorwaarden, als $m \le k \le 2m-1$.

2.8. De eindige elementenmethode voor een tweepunts randwaardeprobleem

In deze sectie laten we, aan de hand van het probleem (2.7.1)-(2.7.2), zien hoe met de eindige elementen methode een randwaardeprobleem in één dimensie opgelost kan worden. We gaan hierbij uit van de kwadratische functionaal (2.4.12). Gemotiveerd door de theorie uit de vorige paragrafen, gebruiken we de Ritz-methode en we minimaliseren over eindigdimensionale deelruimten van $H^1[a,b]$.

De benaderende functie u_h wordt geschreven als

(2.8.1) $$u_h = \sum_{i=0}^{N} a_i\phi_i.$$

We moeten de bij het probleem behorende functionaal, de energie $E(u_h)$, mini-

maliseren. Deze is een functie van de N+1 variabelen $a_0,a_1,\ldots,a_N$. Volgens de vergelijking (2.4.12) is de energie gelijk aan

$$
\begin{aligned}
E(u_h) = \sum_{i,j=0}^{N} a_i a_j [&\alpha p(a)\phi_i(a)\phi_j(a)+\gamma p(b)\phi_i(b)\phi_j(b) + \\
&+ \int p D\phi_i D\phi_j + q\phi_i\phi_j \, dx] \\
- 2 \sum_{j=0}^{N} a_j [&\beta p(a)\phi_j(a)+\delta p(b)\phi_j(b) + \int f\phi_j \, dx].
\end{aligned}
\tag{2.8.2}
$$

Het minimum wordt bepaald door het stelsel lineaire algebraïsche vergelijkingen

$$
\begin{aligned}
&\sum_{i=0}^{N} a_i [\alpha p(a)\phi_i(a)\phi_j(a)+\gamma p(b)\phi_i(b)\phi_j(b) + \int_a^b p D\phi_i D\phi_j + q\phi_i\phi_j \, dx] = \\
&= \beta p(a)\phi_i(a) + \delta p(b)\phi_j(b) + \int_a^b f\phi_j \, dx , \qquad \text{voor } j = 0,1,\ldots,N.
\end{aligned}
\tag{2.8.3}
$$

De keuze van de functies $\{\phi_i\}$ bepaalt nu de oplossing u_h. Volgens de theorie moeten de functies ϕ_i aan drie eisen voldoen:

(i) alle ϕ_i moeten tot $H^1[a,b]$ behoren;

(ii) alle ϕ_i uit $\{\phi_i\}_{i=0}^{\infty}$ moeten lineair onafhankelijk zijn in H_L;

(iii) de rij $\{\phi_i\}_{i=0}^{\infty}$ moet volledig zijn in H_L (opdat $\lim_{h\to 0} u_h = u$).

In de praktijk moet de keuze van de functies ϕ_i ook zodanig zijn dat (1) de elementen van de matrix en van het rechterlid van het stelsel (2.8.3) eenvoudig te berekenen zijn, en (2) dat de oplossing van het stelsel zonder teveel moeite berekend kan worden.

Een groot aantal alternatieven is hier mogelijk voor de keuze $\{\phi_i\}$. Een aantal hiervan zal in het volgende hoofdstuk behandeld worden. We zullen hier een eenvoudige keuze maken: voor de eindigdimensionale deelruimte van $H^1[a,b]$ kiezen we een ruimte van stuksgewijs lineaire functies.

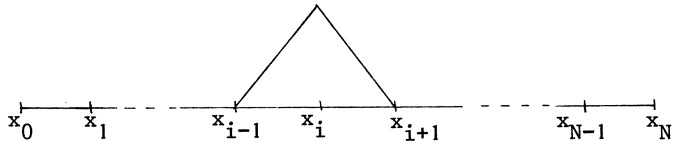
Het interval [a,b] wordt verdeeld in N subintervallen

$$a = x_0 < x_1 < \ldots < x_N = b.$$

Voor iedere i, $0 \le i \le N$, definiëren we ϕ_i, zodat

(i) $\phi_i(x_j) = \delta_{ij}$;

(ii) $\phi_i(x)$ lineair is op ieder subinterval $[x_{j-1},x_j]$, $j = 1(1)N$.



Figuur 2.3 Grafiek van $\phi_i(x)$.

Alle stuksgewijs lineaire functies v_h op [a,b], met knooppunten x_i, kunnen nu geschreven worden als

$$v_h(x) = \sum_{i=0}^{N} a_i \phi_i(x)$$

en er geldt

$$v_h(x_j) = a_j.$$

Deze functies ϕ_i voldoen aan de eisen (i)-(iii) wanneer geldt dat

$$\lim_{N\to\infty} \max_j |x_j - x_{j-1}| = 0.$$

Met deze keuze van $\{\phi_i\}$ kan het lineaire stelsel (2.8.3) geschreven worden als een matrix-vergelijking

$$Ax = b,$$

waarin

$$(2.8.4) \qquad A_{ij} = \alpha p(a)\delta_{i0}\delta_{j0} + \gamma p(b)\delta_{iN}\delta_{jN} + \int_a^b pD\phi_i D\phi_j + q\phi_i\phi_j \, dx \,,$$

$$(2.8.5) \qquad b_j = \beta p(a)\delta_{j0} + \delta p(b)\delta_{jN} + \int_a^b f\phi_j \, dx \,.$$

We zien direkt dat alle matrixelementen o.a. bestaan uit een term

$$a_{ij} = \int_a^b pD\phi_i D\phi_j + q\phi_i\phi_j \, dx.$$

Deze term is gelijk aan nul als $|i-j| > 1$ zodat de matrix (a_{ij}) een tridiagonale matrix is. Slechts op twee plaatsen verschillen A_{ij} en a_{ij}. Aan het 0-de en N-de element van de hoofddiagonaal wordt een randvoorwaardeterm toegevoegd:

$$A_{00} = a_{00} + \alpha p(a), \quad A_{NN} = a_{NN} + \gamma p(b).$$

De term

$$(2.8.6) \qquad a_{ij} = \sum_{k=1}^{N} \int_{x_{k-1}}^{x_k} pD\phi_i D\phi_j + q\phi_i\phi_j \, dx$$

kan worden geschreven als de som van N integralen (over elk element of subinterval één integraal), waarvan er overigens maar één of twee ongelijk aan nul zijn.

Op analoge wijze bestaat het rechterlid uit vector elementen van de vorm

$$(2.8.7) \qquad b_j = \int_{x_{j-1}}^{x_j} f\phi_j \, dx + \int_{x_j}^{x_{j+1}} f\phi_j \, dx.$$

Aan het 0-de en N-de element worden nog de randvoorwaarde-termen $\beta p(a)$ resp. $\delta p(b)$ toegevoegd. Zoals behandeld zal worden in hoofdstuk 3, is de samenstelling van deze matrix op eenvoudige wijze automatisch te verwezenlijken.

Opmerking

De matrix van het lineaire stelsel is symmetrisch en positief definiet.

Opmerking

Ook het *implementeren van een essentiële randvoorwaarde* is eenvoudig automatisch uit te voeren. Wil men bijvoorbeeld de randvoorwaarde $u_h(a) = \nu$ opleggen, dan kan dit gerealiseerd worden door in de matrix en het rechterlid de volgende substituties uit te voeren:

$$b_0 = \nu$$
$$A_{0j} = \delta_{0j}.$$

2.9. Foutschattingen

Nu we in paragraaf 2.8 de constructie van het discrete probleem gegeven hebben, blijft er een zeer belangrijke vraag over: Hoe dicht ligt de eindige elementen benadering u_h bij de exacte oplossing u ? Bovendien willen we weten hoe snel u_h naar u convergeert.

In de volgende stelling zullen we aantonen dat de eindige elementen benadering optimaal is in de norm $\|\cdot\|_L$. Dat wil zeggen dat u_h juist de functie uit V_h is die de afstand $\|u-u_h\|_L$ minimaliseert. We merken op dat $\|\cdot\|_L$ juist een norm is die door de operator L wordt geïnduceerd; $\|u\|_L^2$ is precies het zuiver kwadratisch gedeelte van de energie-functionaal

$$E(v) = \|v\|_L^2 - 2(f,v)_0.$$

Stelling 2.9.1

Als u de functionaal $(Lv,v) - 2(f,v)$ minimaliseert over $V = H_L$ en V_h is een gesloten deelruimte van V, dan gelden de volgende drie beweringen.

(i) Het minimum van E(v) en het minimum van $\|u-v\|_L$, met $v \in V_h$, worden aangenomen door dezelfde functie $u_h \in V_h$, zodat

$$(2.9.1) \qquad \|u-u_h\|_L = \min_{v_h \in V_h} \|u-v_h\|_L.$$

(ii) In H_L is u_h de projectie van u op V_h:

$$(2.9.2) \qquad (u-u_h,v_h)_L = 0 \quad , \qquad \text{voor alle } v_h \in V_h.$$

(iii) De minimaliserende functie u_h voldoet aan

$$(2.9.3) \qquad (u_h,v_h)_L = (f,v_h)_0 \ , \qquad \text{voor alle } v_h \in V_h.$$

Gevolg 2.9.1

In het bijzonder volgt uit (2.9.3) met $V_h = H_L$ dat

$$(2.9.4) \qquad (u,v)_L = (f,v)_0 \quad , \qquad \text{voor alle } v \in H_L.$$

Gevolg 2.9.2

Uit (2.9.2) volgt

$$(2.9.5) \qquad \|u-u_h\|_L^2 = (u,u_h)_L - \|u_h\|_L^2 = \|u\|_L^2 - \|u_h\|_L^2.$$

Bewijs

Het bewijs van (2.9.3) verloopt analoog aan het bewijs van stelling 2.5.3;

voor alle $\lambda \in \mathbb{R}$, $v_h \in V_h$ geldt namelijk

$$0 \leq E(u_h-\lambda v_h) - E(u_h) = \lambda[2(u_h,v_h)_L-2(f,v_h)_0+\lambda(v_h,v_h)_L] ,$$

zodat

$$(u_h,v_h)_L = (f,v_h)_0 \qquad , \qquad \text{voor elke } v_h \in V_h.$$

Dit bewijst (2.9.3). Het bewijs van (2.9.2) volgt direkt door aftrekken van (2.9.3) en (2.9.4) met $v \in V_h$. Nu volgt (2.9.1) vanwege

$$\|u-u_h-v_h\|_L^2 = \|u-u_h\|_L^2 - 2(u-u_h,v_h)_L + \|v_h\|_L^2 = \|u-u_h\|_L^2 + \|v_h\|_L^2,$$

zodat

$$\|u-u_h-v_h\|^2 \geq \|u-u_h\|_L^2. \qquad \square$$

Opmerking 2.9.1

In de praktijk wordt V_h altijd eindigdimensionaal genomen. Hierdoor is V_h gesloten, hetgeen het bestaan van een minimaliserend element garandeert.

Opmerking 2.9.2

Stelling 2.9.1. houdt in dat, met betrekking tot het energie-inproduct $(\cdot,\cdot)_L$ de Ritz-benadering u_h de projectie is van u op V_h. Hierdoor wordt het afschatten van de fout teruggebracht tot het afschatten van de afstand van een functie $u \in V = H^1[a,b]$ tot een functieruimte $V_h \subset V$. We kunnen dit schrijven als

$$\|u-u_h\|_L \leq K_1\|u-u_h\|_1 \leq K \inf_{v_h \in V_h} \|u-v_h\|_1 =$$

$$= K\|[u]\|_{H^1[a,b]/V_h}.$$

Hier is $H^1[a,b]/V_h$ de quotientruimte van $H^1[a,b]$ over V_h.
De convergentiesnelheid van de eindige elementen methode hangt af van de orde waarmee de oplossing kan worden benaderd door de "trial space" V_h, die bestaat uit stuksgewijs polynomen. Zo wordt het oorspronkelijk probleem

gereduceerd tot een approximatie-theoretisch probleem: Hoe goed kan een ruimte V_h een ruimte $H^k[a,b]$ benaderen. Aan dit probleem is paragraaf 2.10 gewijd.

We willen hier stelling 2.9.1 toepassen om een foutschatting te vinden voor de eindige elementen benadering u_h van de oplossing van het tweepunts randwaardeprobleem uit paragraaf 2.4. We nemen aan dat u_h berekend is met behulp van de "dak-functies" zoals beschreven in paragraaf 2.8. Hiertoe definiëren we de foutfunctie e_h door

$$e_h = u_h - u.$$

We weten

$$\|e_h\|_L \leq \|u-v_h\|_L , \qquad \text{voor alle } v_h \in V_h.$$

In de H_L-norm wordt u dus niet slechter benaderd dan de beste interpolerende, stuksgewijs lineaire, functie. Op grond hiervan kunnen we eerst trachten met conventionele middelen een foutschatting te geven.

Wanneer een functie $u \in C^2[a,b]$ benaderd wordt door een interpolerende, stuksgewijs lineaire functie u_I, is het aan de hand van een locale Taylorreeks-ontwikkeling, zeer eenvoudig te zien dat

$$\max_{x\in[a,b]} |u(x)-u_I(x)| \leq \tfrac{1}{2}h^2 \max_{x\in[a,b]} |u''(x)| ,$$

$$\max_{x\in[a,b]} |u'(x)-u_I'(x)| \leq h \max_{x\in[a,b]} |u''(x)|.$$

Wanneer we dit substitueren in (2.7.10), blijkt dat

$$\|u-u_I\|_L^2 = \int_a^b p(u'-u_I')^2 + q(u-u_I)^2 \, dx +$$

$$+ [\alpha p(u-u_I)^2](a) + [\gamma p(u-u_I)^2](b) \leq \tag{2.9.6}$$

$$\leq K h^2 \max |u''(x)|^2.$$

Voor de H_L-norm van de fout vinden we dus

$$\|u-u_I\|_L \leq K\ h\ \max\ |u''(x)|^2.$$

Waar hier een globale schatting gegeven wordt voor de fout (de linkerzijde van de ongelijkheid zegt niets over de puntsgewijze fout) lijkt de puntsgewijze norm aan de rechterzijde misplaatst. Om een foutschatting in termen van een globale norm te krijgen moeten we echter gebruik maken van stelling 2.10.1.

Stelling 2.9.2

Voor de eindige elementenbenadering met stuksgewijs lineaire functies, van het probleem (2.4.10), geldt de foutschatting

$$\|u-u_h\|_L \leq K \cdot \|u''\|_0 \cdot h. \tag{2.9.7}$$

Bewijs

Uit stelling 2.10.1 volgt, met $p = 2$, $k = 1$, $0 \leq m \leq 2$

$$\|u-u_I\|_m \leq K\ |u|_2\ h^{2-m},$$

zodat

$$\|u-u_I\|_1 \leq K\ \|u''\|_0\ h.$$

Wanneer we gebruik maken van het feit dat de norm $\|\cdot\|_1$ equivalent is met $\|\cdot\|_L$ (zie lemma 2.7.1) dan blijkt (2.9.8) direkt. □

Gevolg 2.9.3

We kunnen ook gebruik maken van de regulariteit van de oplossing (zie bijv. FRIEDMAN [1969])

$$\|u\|_2^2 \leq K(\|f\|_0^2+\beta^2+\delta^2), \tag{2.9.8}$$

waarmee we de fout in de berekening van u kunnen schatten in termen van het rechterlid:

$$\|u-u_h\|_L \leq K\cdot h\cdot\{\|f\|_0+\beta+\delta\}. \tag{2.9.9}$$

Opmerking 2.9.3

Dat de eindige elementenbenadering voor elke $f \in H^0$ convergeert in de energienorm is een direkt gevolg van het feit dat een volledige rij functies $\{\phi_i\}$ in H_L gekozen wordt als basisfuncties voor de methode. Over de convergentiesnelheid kan echter niet direkt een uitspraak worden gedaan.

Stelling 2.9.3

Voor *elke* $u \in H^1[a,b]$ (d.w.z. $f \in H^0$) convergeert de eindige elementenmethode in de energienorm (dus ook als u'' niet overal bestaat).

Bewijs.

$C_B^2[a,b]$ ligt dicht in H_L. Voor elke $u \in H^1[a,b]$ bestaat er dus een rij $\{v_N\} \subset C_B^2[a,b]$ zodat $\lim_{N\to\infty} \|v_N-u\|_L = 0$. Uit stelling (2.9.3) blijkt $\lim_{\substack{h\to 0 \\ N\to\infty}} \|v_N^h-v_N\|_L = 0$ en dus geldt, vanwege

$$\|u_h-u\|_L \leq \|v_N^h-u\|_L \leq \|v_N^h-v_N\|_L + \|v_N-u\|_L,$$

dat

$$\lim_{h\to 0} \|u_h-u\|_L = 0. \qquad \square$$

De fout in de H^0-norm

We hebben nu schattingen in H_L, en daarmee in H^1, gevonden. We willen nu nagaan welke schatting in H^0 gevonden kan worden.
Met de ongelijkheid van Poincaré

$$|v(x_0)-v(a)| \leq \sqrt{b-a}\, \|v\|_1, \qquad x_0 \in [a,b],$$

kunnen we eenvoudig zien dat, *uniform* op $[a,b]$,

$$|u-u_h| = O(h).$$

In de volgende stelling tonen we echter aan dat in de H^0-norm de convergentie niet $O(h)$ is, maar $O(h^2)$.

Opmerking 2.9.4

We weten wel dat, voor een interpolerende functie u_I,

$$\|u-u_I\|_0 \leq \frac{h^2}{(b-a)^2} \|u''\|_0.$$

De Ritz-methode is echter niet optimaal in H^0, zodat *niet* geldt

$$\|u-u_h\|_0 \leq \|u-v\|_0 \quad , \qquad \text{voor alle } v \in H^0[a,b].$$

Stelling 2.9.2

Voor de stuksgewijs lineaire eindige elementenbenadering geldt de volgende foutschatting

$$\|u-u_h\|_0 \leq K \|u''\|_0 h^2. \tag{2.9.10}$$

Bewijs (zie ook STRANG & FIX [1973], p.49)

Voor het bewijs introduceren we een funtie z, die de oplossing is van $Lz = u_h-u$. Nu geldt

$$(Lz,v) = (u_h-u,v) \quad , \qquad \text{voor alle } v \in H_L,$$

i.h.b.

$$\|u_h-u\|_0^2 = (Lz,u_h-u) = (z,u_h-u)_L.$$

Uit stelling 2.9.1 volgt

$$(v_h,u_h-u)_L = 0 \quad , \qquad \text{voor alle } v_h \in V_h,$$

zodat

$$\|u_h-u\|_0^2 = (z-v_h,u_h-u)_L \leq \|z-v_h\|_L \|u-u_h\|_L \text{ voor alle } v_h \in V_h.$$

Uit (2.9.7) volgt

$$\|u-u_h\|_L^2 \leq K^2 \|u''\|_0^2 h^2 ,$$

$$\inf_{v_h \in V_h} \|z-v_h\|_L^2 \leq K^2 \|z''\|_0^2 h^2 ,$$

zodat

$$\|u-u_h\|_0^2 \leq K^2 \|u''\|_0 \|z''\|_0 h^2.$$

Uit de regulariteits eigenschap (2.9.8) volgt

$$\|z''\|_0^2 \leq \|z\|_2^2 \leq K_R^2 \|u_h-u\|_0^2 ,$$

zodat

$$\|u-u_h\|_0 \leq K^2 \|u''\|_0 K_R h^2. \qquad \square$$

Opmerking 2.9.5

Wanneer we de regulariteits eigenschap (2.9.8) nogmaals toepassen krijgen we

$$\|u-u_h\|_0 \leq K^2 K_R^{-2} h^2 (\|f\|_0^2+\beta^2+\delta^2)^{\frac{1}{2}}.$$

2.10. Interpolatie voor Sobolev-ruimten

In deze paragraaf vermelden we hoe goed Sobolev-ruimten door ruimten van stuksgewijs polynomen benaderd kunnen worden. We geven hier de theorie op een abstrakte wijze. Enerzijds heeft dit misschien het nadeel dat niet direkt de belangrijkste praktische implicaties gezien worden. Anderzijds heeft dit het voordeel dat alle speciale gevallen, zoals die in de praktijk gebruikt worden, in deze theorie ondergebracht kunnen worden. In hoofdstuk 3 zullen toepassingen van deze theorie behandeld worden, in het bijzonder Lagrange en Hermite interpolatie. De theorie die hier gegeven wordt is voornamelijk gebaseerd op het werk van BRAMBLE & HILBERT [1970] en vele anderen. In grote lijnen volgen wij hier de uiteenzetting zoals die gegeven wordt door CIARLET & RAVIART [1972].

Het belangrijkste resultaat van deze paragraaf is de volgende stelling.

Stelling 2.10.1

Wanneer een approximatie-schema als volgt wordt gedefinieerd: bij een gegeven open begrensde deelverzameling Ω uit $\mathbb{R}^n$, voegen we aan een willekeurige functie uit de Sobolev-ruimte $W^{k+1,p}(\Omega)$ een eenduidige "interpolatie" Πu toe. Wanneer Π alle polynomen van een graad kleiner of gelijk aan k invariant laat, dan geldt voor alle m, $0 \leq m \leq k+1$,

$$(2.10.1) \qquad \|u-\Pi u\|_{m,p} \leq K(k,p)\ |u|_{k+1,p}\ \frac{h^{k+1}}{\rho^m}\ ,$$

waarin

(2.10.2) h = de diameter van Ω,

(2.10.3) ρ = het supremum van de diameters van de ingeschreven bollen in Ω.

Opmerking 2.10.1

Deze stelling beschrijft in welke mate een benadering met een bepaalde soort elementen (driehoeken, vierhoeken e.d.) met een bepaalde soort benaderingen daarop (bijv. stuksgewijs lineaire of stuksgewijs kwadratische), beter wordt wanneer het gebied in kleinere elementen verdeeld wordt. Bovendien wordt beschreven in welke mate vormveranderingen van de elementen (bijv. vervormingen van de driehoeken) invloed hebben.

Definitie 2.10.1

De verzameling van alle polynomen in de $\mathbb{R}^n$ van de graad kleiner of gelijk aan k, wordt aangegeven met P_k.

We zullen nu, aan de hand van twee lemma's, aangeven hoe de stelling bewezen wordt.

Lemma 2.10.1

Zij Ω een begrensde open deelverzameling uit $\mathbb{R}^n$ met een continue rand. Laat p gegeven zijn met $1 \leq p \leq \infty$ en laat $k \geq 0$ een vast geheel getal zijn. Laat f een begrensde lineaire functionaal zijn op $W^{k+1,p}(\Omega)$ zodat

$$(2.10.4) \qquad \langle f,u\rangle = 0 \quad , \quad \text{voor alle } u \in P_k.$$

Dan bestaat er een constante $C = C(n,k,p,\Omega)$ zodanig dat

$$(2.10.5) \qquad |<f,u>| \leq C \, \|f\|_{W^{k+1,p} \to \mathbb{R}} \, |u|_{k+1,p} , \qquad \text{voor alle } u \in W^{k+1,p}(\Omega).$$

Bewijs

Volgens de definitie van de norm van een operator is $\|f\|_{W^{k+1,p} \to \mathbb{R}}$ gedefinieerd door

$$(2.10.6) \qquad \|f\|^*_{k+1,p} = \|f\|_{W^{k+1,p} \to \mathbb{R}} = \sup_{v \neq 0, v \in W^{m,p}(\Omega)} \frac{|<f,v>|}{\|v\|_{m,p}} .$$

Op de equivalentieklassen $[u]$ uit de quotientruimte $W^{k+1,p}(\Omega)/P_k$ wordt een norm gedefinieerd door

$$(2.10.7) \qquad \|[u]\|_{W^{k+1,p}(\Omega)/P_k} = \inf_{v \in P_k} \|u+v\|_{k+1,p} .$$

NEČAS [1967, p.112, Stelling 7.2] bewijst dat de seminorm $|\cdot|_{k+1,p}$ in $W^{k+1,p}(\Omega)$ een norm is op de bovengenoemde quotientruimte en dat deze equivalent is met (2.10.7).
Voor iedere f die aan (2.10.4) voldoet geldt nu dat $<f,u> = <f,u+v>$ voor alle $u \in W^{k+1,p}(\Omega)$ en alle $v \in P_k$,

zodat

$$|<f,u>| \leq \|f\|^*_{k+1,p} \inf_{v \in P_k} \|u-v\|_{k+1,p} \leq C \, \|f\|^*_{k+1,p} \, |u|_{k+1,p}. \qquad \square$$

Voordat we laten zien hoe het bewijs verloopt moeten we het begrip *equivalent gebied* invoeren.

Definitie 2.10.2

Laten Ω en $\hat{\Omega}$ begrensde open deelverzamelingen van $\mathbb{R}^n$ zijn. We zeggen dat Ω en $\hat{\Omega}$ equivalent zijn d.e.s.d. als er een affiene bijectie $x = B\hat{x}+b$ bestaat die Ω in $\hat{\Omega}$ overvoert.

Wanneer Ω en $\hat{\Omega}$ twee equivalente gebieden zijn, kunnen we aan iedere functie u op Ω een functie $\hat{u}$ op $\hat{\Omega}$ toevoegen door

$$(2.10.11) \qquad \hat{u}(\hat{x}) = u(B\hat{x}+b) = u(x) , \qquad \text{voor alle } x \in \Omega.$$

De affiene afbeelding induceert zo een isomorfisme tussen $W^{m,p}(\Omega)$ en $W^{m,p}(\hat{\Omega})$ voor alle m en p.

Op dezelfde wijze kunnen we nu aan elke Π: $W^{k+1,p}(\Omega) \to W^{m,p}(\Omega)$ een $\hat{\Pi}$: $W^{k+1,p}(\hat{\Omega}) \to W^{m,p}(\hat{\Omega})$ toevoegen door

$$\hat{\Pi}\hat{u} = \widehat{\Pi u} \qquad \text{voor alle } u \in W^{k+1,p}(\Omega). \tag{2.10.12}$$

Het is eenvoudig in te zien dat P_k invariant is voor Π d.e.s.d. als P_k invariant is voor $\hat{\Pi}$.

Zijn h, ρ, $\hat{h}$ en $\hat{\rho}$ gedefinieerd door (2.10.2) en (2.10.3) voor Ω respectievelijk $\hat{\Omega}$, dan kan men bewijzen (CIARLET & RAVIART [1972])

$$\|B\| \leq \frac{h}{\hat{\rho}} \quad \text{en} \quad \|B^{-1}\| \leq \frac{\hat{h}}{\rho} \tag{2.10.13}$$

en dat, voor h klein genoeg,

$$1 \leq \frac{\hat{\rho}}{h} \leq \frac{1}{\|B\|} \leq \|B^{-1}\|. \tag{2.10.14}$$

Lemma 2.10.2

Zij Ω een begrensde open deelverzameling uit $\mathbb{R}^n$ met een continue rand. Laat p gegeven zijn met $1 \leq p \leq \infty$, laat $k \geq 0$ een vast geheel getal zijn en laat m een geheel getal zijn, $0 \leq m \leq k+1$. Zij Π een lineaire afbeelding

$$\Pi: W^{k+1,p}(\Omega) \to W^{m,p}(\Omega), \tag{2.10.15}$$

zodat

$$\Pi u = u \quad , \qquad \text{voor alle } u \in P_k. \tag{2.10.16}$$

Dan geldt voor alle $u \in W^{k+1,p}(\Omega)$

$$\|u-\Pi u\|_{m,p} \leq C \, \|I-\Pi\|_{W^{k+1,p} \to W^{m,p}} \, |u|_{k+1,p} \quad , \tag{2.10.17}$$

waarin C dezelfde constante is als in lemma 2.10.1.

Bewijs

Zij g een gegeven begrensde lineaire functionaal op $W^{m,p}(\Omega)$. Voor de begrensde lineaire functionaal op $W^{k+1,p}(\Omega)$ gedefinieerd door

$$<f,u> = <g,u-\Pi u>$$

geldt $<f,u> = 0$ voor alle $u \in P_k$. Bovendien

$$\|f\|^*_{k+1,p} \leq \|I-\Pi\|_{W^{k+1,p}(\Omega) \to W^{m,p}(\Omega)} \|g\|^*_{m,p}$$

en

$$\|u-\Pi u\|_{m,p} = \sup_g \frac{|<g,u-\Pi u>|}{\|g\|^*_{m,p}} .$$

Aangezien $|<g,u-\Pi u>| = |<f,u>| < C \|f\|^*_{k+1,p} |u|_{k+1,p}$ volgt het lemma nu direkt.

Na enig rekenwerk blijkt nu

$$|\hat{u}|_{\ell,p} \leq \|B\|^{\ell} |\det(B)|^{-1/p} |u|_{\ell,p} \qquad (2.10.18)$$

en analoog

$$|u|_{\ell,p} \leq \|B^{-1}\|^{\ell} |\det(B)|^{1/p} |\hat{u}|_{\ell,p}. \qquad (2.10.19)$$

Hieruit volgt

$$\|u\|^p_{m,p} \leq |\det(B)| \, \|B^{-1}\|^{mp} \|\hat{u}\|^p_{mp}. \qquad (2.10.20)$$

Als h aan (2.10.14) voldoet dan volgt uit lemma 2, door gebruik te maken van (2.10.20) en (2.10.19) en door de schattingen (2.10.13) te substitueren, dat

$$\|u-\Pi u\|_{m,p} \leq |\det(B)|^{1/p} \|B^{-1}\|^m \|\hat{u}-\hat{\Pi}\hat{u}\|_{m,p} \leq$$

$$\leq C(u,k,p,\hat{\Omega}) \frac{\hat{h}^m}{\hat{\rho}^{k+1}} \|I-\hat{\Pi}\|_{W^{k+1,p}(\hat{\Omega}) \to W^{im,p}(\hat{\Omega})} *$$

$$* |u|_{k+1,p} \frac{h^{k+1}}{\rho^m}.$$

Nu is $C(u,k,p,\hat{\Omega})\ \hat{h}^m\ \hat{\rho}^{-k-1}\ \|I-\hat{\Pi}\|$ een konstante die uitsluitend afhangt van het type element $(\hat{\Omega})$ en van de gebruikte benadering op dat element $(\hat{\Pi})$. Samenvattend kunnen we zeggen dat voor een vast type interpolatie (vaste $u,\hat{\Omega},\Pi$) geldt

$$\|u-\Pi u\|_{m,p} \leq K(k,p)\ |u|_{k+1,p}\ \frac{h^{k+1}}{\rho^m}.$$

Hiermee is stelling 2.10.1 bewezen. □

Gevolg vgl. STRANG [1972] theorem 1.

Zij Ω een begrensde open deelverzameling uit $\mathbb{R}^n$, zij $u \in C^{k+1}(\Omega)$ dan geldt, wanneer Π alle polynomen uit P_k invariant laat voor alle m, $0 \leq m \leq k+1$, dat

$$(2.10.18)\quad \sup_{\substack{x\in\Omega \\ |\ell|=m}} \|D^\ell u(x)-D^\ell \Pi u(x)\| \leq K(h) \sup_{\substack{x\in\Omega \\ |\ell|=k+1}} \|D^\ell u(x)\|\ \frac{h^{k+1}}{\rho^m}.$$

Voorbeeld 2.10.1

Een voorbeeld van een afbeelding Π wordt gegeven door iedere Lagrange of Hermite interpolatie formule op een element $\hat{\Omega}$. De steunpunten van deze interpolatie formules worden door de affiene afbeelding op overeenkomstige punten van de equivalente elementen Ω afgebeeld.

LITERATUUR

BRAMBLE, J.H. & S.R. HILBERT, *Estimation of linear functionals on Sobolev spaces with applications to Fourier transforms and spline interpolations*, SIAM Journal of Analysis 7, p.113-124 (1970).

BRAMBLE, J.H. & A.H. SCHATZ, *Rayleigh-Ritz-Galerkin methods for Dirichlets problem using subspaces without boundary conditions*, Comm. Pure Appl. Math. 23, p.653-675 (1970).

BRAMBLE, J.H. & A.H. SCHATZ, *On the numerical solution of elliptic boundary value problems by least square approximations of the data*, Uit: B. HUBBARD, editor, *Numerical Solution of Partial Differential Equations* (SYNSPADE), Academic Press, New York, 1971.

CÉA, J., *Approximation variatonnelle des problèmes aux limites*, Ann. Just. Fourier 14, p.345-444 (1964).

CIARLET, P.G. & P.A. RAVIART, *General Lagrange and Hermite interpolation in $\mathbb{R}^n$ with applications to the finite element method*, Arch. Rat. Mech. Anal. 46, p.177-199 (1972).

FORSYTH, G. & W. WASOW, *Finite Difference Methods for Partial Differential Equations*, John Wiley and Sons, New York, 1960.

FRIEDMANN, A., *Partial Differential Equations*, Rinehart, Holt and Winston, New York, 1969.

LIONS, J.L. & E. MAGÈNES, *Problèmes aux limites non homogènes et applications*, Dunod, Paris, 1968.

MIKHLIN, S.G., *The Problem of the Minimum of a Quadratic Functional*, Holden-Day, San Francisco, 1965.

MIKHLIN, S.G., *The numerical Performance of Variational Methods*, Wolters-Noordhoff, Groningen, 1971.

NEČAS, J., *Les méthodes directes en theorie des équations elliptiques*, Academia, Praag, 1967.

STRANG, G. & G. FIX, *An Analysis of the Finite Element Method*, Prentice-Hall, Englewood Cliffs, N.J., 1973.

STUMMEL, F., *Rand und Eigenwertaufgaben in Sobolewschen Räumen*, Springer-Verlag, Berlin, Göttingen, Heidelberg, 1969.

YOSIDA, K., *Functional Analysis*, Springer-Verlag, Berlin, Göttingen, Heidelberg, 1965.

TÉMAM, R., *Numerical Analysis*, Reidel, Dordrecht, 1973.

3. TOEPASSING VAN DE EINDIGE ELEMENTENMETHODE OP ÉÉN- en TWEEDIMENSIONALE RANDWAARDEPROBLEMEN

3.1. Inleiding

In dit hoofdstuk zullen de eenvoudigste eindige elementenruimtes worden behandeld, alsmede toepassingen op randwaardeproblemen. In de §§2-5 behandelen we de stuksgewijze Lagrange-interpolatie met toepassingen op tweede orde randwaardeproblemen. In de §§6-8 behandelen we de stuksgewijze Hermite-interpolatie met toepassingen op vierde orde randwaardeproblemen. In de §§9-12 behandelen we de tweedimensionale generalisatie van Lagrange- en Hermite-interpolatie. Tot slot worden in §13 de isoparametische elementen en in §14 een globale foutschatting behandeld.

3.2. Stuksgewijze Lagrange-interpolatie

We verdelen een segment $[a,b]$ in N (niet noodzakelijk equidistante) deelsegmenten $e_j = [x_{j-1},x_j]$, $j = 1,\ldots,N$, en geven deze verdeling aan met

$$\pi : a = x_0 < x_1 < \ldots < x_N = b. \tag{3.2.1}$$

Verder definiëren we

$$\begin{aligned} h_j &= x_j - x_{j-1}, \\ h &= \max_j h_j \quad , \end{aligned} \tag{3.2.2}$$

waarbij we h de maaswijdte van π noemen.

Definitie 3.2.1 Onder $\mathbb{P}_k(\pi)$ verstaan we de verzameling van alle funkties die

(i) kontinu zijn op $[a,b]$;

(ii) op ieder segment e_j een polynoom van de graad $\leq k$ zijn.

Het is duidelijk dat $\mathbb{P}_k(\pi)$ onder de definities van inprodukt

$$(u,v)_1 = \int_a^b \left[\frac{du}{dx}\frac{dv}{dx} + u(x)v(x)\right] dx,$$

en norm

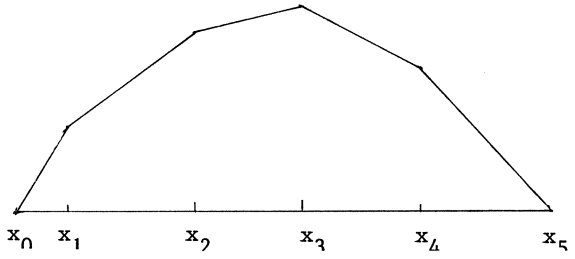
$$\|u\|_1 = \sqrt{(u,u)_1} \quad ,$$

een deelruimte is van de separabele Hilbertruimte $H^1[a,b]$, de ruimte van funkties waarvan de nulde en eerste afgeleide kwadratische integreerbaar zijn over $[a,b]$. Verder bevat $\mathbb{P}_k(\pi)$ alle polynomen van de graad $\leq k$. We geven $\mathbb{P}_k(\pi)$ soms ook aan met $\mathbb{P}_k$.

Om te bepalen hoe groot de dimensie van $\mathbb{P}_k$ is en hoe de basisfunkties eruitzien, bekijken we de gevallen $k = 1,2,3$.

$\underline{k = 1}$

Een element uit $\mathbb{P}_2$ is een stuksgewijs lineaire funktie (zie fig. 3.1).

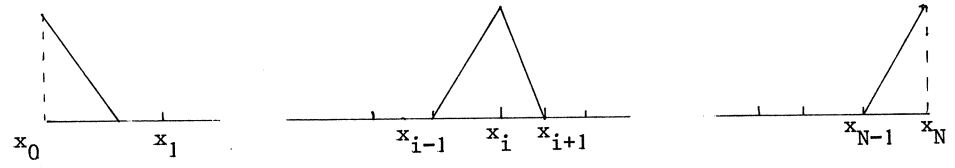


Figuur 3.1 Stuksgewijs lineaire funktie

Het is evident dat ieder element uit $\mathbb{P}_2$ volkomen bepaald wordt door de waarden die het in de punten x_j aanneemt. Er zijn dus N+1 basisfuncties nodig, die we met ϕ_i aangeven. Een voor de hand liggende definitie van ϕ_i is de volgende:

(3.2.3) $\quad \phi_i(x_j) = \delta_{ij}, \qquad 0 \leq i, \quad j \leq N,$

waarbij δ_{ij} het Kronecker-symbool is. De funkties die hieraan voldoen, zijn de zgn. "dakfunkties" (zie fig. 3.2).



Figuur 3.2 Dakfunkties $\phi_i(x)$

De basisfunkties $\phi_i(x)$ worden gegeven door de formules

$$\phi_0(x) = \begin{cases} \dfrac{x-x_1}{x_0-x_1} = -(x-x_1)/h_1, & x_0 \le x \le x_1, \\ 0, & \text{elders;} \end{cases}$$

$$\text{(3.2.4)} \qquad \phi_i(x) = \begin{cases} \dfrac{x-x_{i-1}}{x_i-x_{i-1}} = (x-x_{i-1})/h_i, & x_{i-1} \le x \le x_i, \\ \dfrac{x-x_{i+1}}{x_i-x_{i+1}} = -(x-x_{i+1})/h_{i+1}, & x_i \le x \le x_{i+1}, \\ 0, & \text{elders, } i = 1,\ldots,N-1; \end{cases}$$

$$\phi_N(x) = \begin{cases} \dfrac{x-x_{N-1}}{x_N-x_{N-1}} = (x-x_{N-1})/h_N, & x_{N-1} \le x \le x_N \\ 0, & \text{elders.} \end{cases}$$

Uit bovenstaande formule blijkt dat $\phi_i(x)$ alleen $\not\equiv 0$ is op die deelsegmenten waartoe x_i behoort. Hieruit volgt dat er op het deelsegment e_i slechts twee basisfunkties $\not\equiv 0$ zijn, namelijk $\phi_{i-1}(x)$ en $\phi_i(x)$. Deze twee funkties worden *op* e_i gegeven door de formules

$$\text{(3.2.5)} \qquad \begin{aligned} \phi_{i-1}(x) &= 1 - (x-x_{i-1})/h_i; \\ \phi_i(x) &= (x-x_{i-1})/h_i, \qquad x_{i-1} \le x \le x_i. \end{aligned}$$

Uit (3.2.5) blijkt dat $\phi_i(x)$ en $\phi_{i-1}(x)$ op e_i lineaire funkties zijn van

$$t = (x-x_{i-1})/h_i.$$

We merken nog op dat

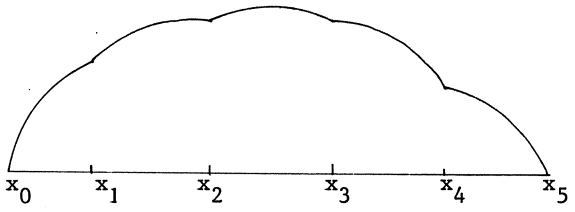
$$\phi_i(x) = \phi_{i-1}(x_i+x_{i-1}-x) ,$$

(3.2.6) $$\phi_{i-1}(x) = \phi_i(x_i+x_{i-1}-x) ,$$

$$1 = \phi_i(x) + \phi_{i-1}(x) , \qquad x \in e_i .$$

k = 2

Een element uit $\mathbb{P}_2$ is een stuksgewijs kwadratische funktie (zie fig. 3.3).



Figuur 3.3 Stuksgewijs kwadratische funktie

Een element uit $\mathbb{P}_2$ is nu niet meer alleen bepaald door de waarden die het in de punten $x_0,\ldots,x_N$ aanneemt, aangezien een parabool drie vrijheidsgraden heeft. We moeten op ieder segment e_j nog één extra punt kiezen en nemen daarvoor het middelpunt

$$x_{j-\frac{1}{2}} = \tfrac{1}{2}(x_{j-1}+x_j).$$

Andermaal kiezen we de basisfunkties $\phi_i(x)$ zo, dat geldt

$$\phi_i(x_j) = \delta_{ij}, \qquad i,j = 0,\tfrac{1}{2},\ldots,N.$$

Het aantal basisfunkties van $\mathbb{P}_2$ is dus 2N+1. Evenals in het geval k = 1 is $\phi_i(x)$ alleen $\neq 0$ op die segmenten waartoe x_i behoort (zie figuur 3.4).

We bekijken nu het segment $e_i = [x_{i-1},x_i]$. Op dat segment zijn alleen $\phi_{i-1}(x)$, $\phi_{i-\frac{1}{2}}$ en $\phi_i(x)$ relevant. Passen we de transformatie

(3.2.7) $$x = h_i t + x_{i-1}, \qquad 0 \le t \le 1$$

toe, dan vinden we de formules

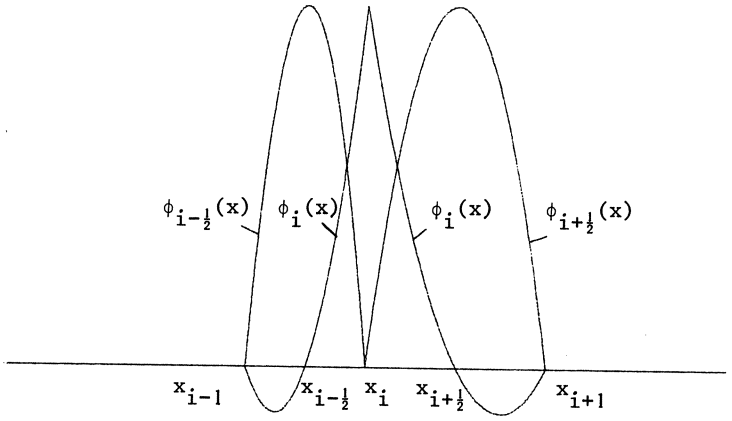
$$\phi_{i-1}(x) = (t-1)(2t-1) = \Phi_0(t),$$

$$\phi_{i-\frac{1}{2}}(x) = 4t(1-t) \quad = \Phi_{\frac{1}{2}}(t),$$

(3.2.8)

$$\phi_i(x) \quad = t(2t-1) \quad = \Phi_1(t),$$

$$t \quad = (x-x_{i-1})/h_i \qquad , \qquad x_{i-1} \le x \le x_i.$$



Figuur 3.4 Grafiek van $\phi_{i-\frac{1}{2}}(x)$, $\phi_i(x)$ en $\phi_{i+\frac{1}{2}}(x)$

Blijkbaar zijn op het segment e_i de relevante basisfunkties polynoom in de variabele $t = (x-x_{i-1})/h_i$. Merk nog op dat (zie (3.2.6))

$$\Phi_i(t) = \Phi_{1-i}(1-t), \qquad i = 0,\tfrac{1}{2},1;$$

(3.2.9)

$$\Phi_0(t) + \Phi_{\frac{1}{2}}(t) + \Phi_1(t) = 1.$$

k = 3

Dit gaat geheel analoog met k = 2. Op het segment e_i kiezen we nu *twee* inwendige punten, namelijk

$$x_{i-\frac{2}{3}} = x_{i-1} + \tfrac{1}{3}h_i,$$

$$x_{i-\frac{1}{3}} = x_{i-1} + \tfrac{2}{3}h_i.$$

Op het segment $[x_{i-1},x_i]$ zijn de relevante basisfunkties nu als volgt gedefiniëerd:

$$\phi_{i-1}(x) = \Phi_0(t) = -\tfrac{1}{2}(3t-1)(3t-2)(t-1),$$

$$\phi_{i-\frac{2}{3}}(x) = \Phi_{\frac{1}{3}}(t) = \tfrac{9}{2}t(3t-2)(t-1) \quad ,$$

$$\phi_{i-\frac{1}{3}}(x) = \Phi_{\frac{2}{3}}(t) = -\tfrac{9}{2}t(3t-1)(t-1) \quad ,$$

$$\phi_i(x) = \Phi_1(t) = \tfrac{1}{2}t(3t-1)(3t-2) \quad ,$$

$$t = (x-x_{i-1})/h_i \quad , \quad x_{i-1} \le x \le x_i.$$

Merk op dat ook hier geldt:

$$\Phi_i(t) = \Phi_{1-i}(1-t) \quad , \quad i = 0,\tfrac{1}{3},\tfrac{2}{3},1,$$

(3.2.11)

$$1 = \Phi_0(t) + \Phi_{\frac{1}{3}}(t) + \Phi_{\frac{2}{3}}(t) + \Phi_1(t).$$

Algemeen kiezen we de basis van $\mathbb{P}_K$ als volgt:

(i) Op het segment e_i kiezen we k-1 inwendige punten, die e_i in k *equidistante* deelsegmenten verdelen. Deze punten geven we aan met

$$x_{i-1+\frac{j}{k}} = x_{i-1} + \frac{j}{k} * h_i \quad , \quad \begin{array}{l} i = 1,\ldots,N; \\ j = 1,\ldots,k-1. \end{array}$$

(ii) Op e_i is $\phi_{i-1+\frac{j}{k}}(x)$ als volgt gedefinieerd:

$$\phi_{i-1+\frac{j}{k}}(x) = \Phi_{\frac{j}{k}}(t) = \prod_{\substack{\ell=0 \\ \ell\neq j}}^{k} \frac{kt-\ell}{\ell-j} \quad , \quad \begin{array}{l} x_{i-1} \le x \le x_i, \\ t = (x-x_{i-1})/h_i, \\ i = 1,\ldots,N; \\ j = 0,\ldots,k. \end{array}$$

Ook hier gelden de relaties

$$\Phi_{\frac{j}{k}}(t) = \Phi_{1-\frac{j}{k}}(1-t) \quad , \quad j = 0,\ldots,k;$$

(3.2.13)

$$1 = \sum_{j=0}^{k} \Phi_{\frac{j}{k}}(t).$$

3.3. Een variatieprobleem

We bekijken het volgende variatieprobleem: Minimaliseer de kwadratische funktionaal

$$(3.3.1a) \qquad I(v) = \int_a^b \left[p(x)\left(\frac{dv}{dx}\right)^2 + q(x)[v(x)]^2 - 2f(x)v(x) \right] dx,$$

onder de nevenvoorwaarden

$$(3.3.1b) \qquad \begin{aligned} &v(a) = 0;\ q(x) \text{ en } f(x) \text{ kontinu op } [a,b]; \\ &p(x) \geq p_0 > 0;\ q(x) \geq 0;\ p(x) \text{ kontinu differentieerbaar op } [a,b]. \end{aligned}$$

De oplossing u van (3.3.1) ligt in de separabele energieruimte H_L van funkties waarvan de nulde en eerste afgeleide kwadratisch integreerbaar zijn over [a,b], en die in a de waarde nul aannemen. In deze Hilbert-ruimte zijn het *energieinprodukt*

$$(3.3.2a) \qquad (u,v)_L = \int_a^b \left[p(x)\,\frac{du}{dx}\,\frac{dv}{dx} + q(x)u(x)v(x) \right] dx$$

en de *energienorm*

$$(3.3.2b) \qquad \|u\|_L = \sqrt{(u,u)_L}$$

gedefinieerd.

Voor de oplossing u van (3.3.1) geldt wegens

$$I(u+\varepsilon v) \geq I(u), \qquad \varepsilon \text{ willekeurig, } v \in H_L,$$

dat

$$\lim_{\varepsilon \to 0} \frac{I(u+\varepsilon v)-I(u)}{\varepsilon} = 0, \qquad v \in H_L.$$

Uitwerking hiervan geeft de zwakke Galerkin-vorm

$$(3.3.3) \qquad \begin{aligned} &\int_a^b \left[p(x)\,\frac{du}{dx}\,\frac{dv}{dx} + q(x)uv \right] dx = \int_a^b f(x)v\,dx; \\ &(u,v)_L = (f,v), \qquad \text{voor iedere } v \in H_L. \end{aligned}$$

We beschouwen bij een verdeling

$$\pi = a = x_0 < x_1 < \ldots < x_N = b$$

met maaswijdte h de Hilbert-ruimte $\mathbb{P}_k(\pi)$ en definiëren

$$V_h = \{v \mid v \in \mathbb{P}_k(\pi),\ v(a) = 0\}.$$

Het is evident dat V_h een eindigdimensionale deelruimte is van H_L met dimensie kN. We zoeken nu een approximatie van de oplossing van (3.3.1) door het volgende probleem op te lossen: Minimaliseer $I(v_h)$, $v_h \in V_h$. De oplossing u_h van dit probleem voldoet eveneens aan de Galerkin-vorm

(3.3.3a) $\quad (u_h,v_h)_L = (f,v_h), \qquad$ voor iedere $v_h \in V_h$.

We passen dit toe voor k=2. De basisfunkties $\phi_{\frac{1}{2}}(x),\phi_1(x),\ldots,\phi_{N-\frac{1}{2}}(x)$, $\phi_N(x)$ van V_h zijn door de formules (3.2.7) - (3.2.8) gegeven. Voor de oplossing u_h geldt dan

$$(u_h,\phi_i)_L = (f,\phi_i), \qquad i = \tfrac{1}{2},1,\ldots,N.$$

Schrijven we

$$u_h(x) = \sum_{j=\frac{1}{2}}^{N} q_j\phi_j(x),$$

dan wordt $(q_{\frac{1}{2}},q_1,\ldots,q_N)^T$ bepaald door het lineaire stelsel

(3.3.4) $\quad \sum_{j=\frac{1}{2}}^{N} (\phi_i,\phi_j)_L q_j = (f,\phi_i), \qquad i = \tfrac{1}{2},1,\ldots,N.$

Het gaat dus in de eerste plaats om de berekening van de matrix $A = (a_{ij})$, gedefinieerd door

$$a_{ij} = (\phi_i,\phi_j)_L = \int_a^b p(x)\,\frac{d\phi_i}{dx}\,\frac{d\phi_i}{dx}\,dx + \int_a^b q(x)\phi_i(x)\phi_i(x)dx =$$

(3.3.5)

$$= s_{ij} + m_{ij}, \qquad i,j = \tfrac{1}{2},1,\ldots,N,$$

en de vektor $F = (F_i)$, gedefinieerd door

$$(3.3.6)\qquad F_i = \int_a^b f(x)\phi_i(x)dx, \qquad i = \tfrac{1}{2},1,\ldots,N.$$

De matrices $S = (s_{ij})$ en $M = (m_{ij})$ heten respectievelijk stijfheidsmatrix en massamatrix. De vektor F wordt de belastingsvektor genoemd. Deze termen zijn afkomstig uit de strukturele analyse.

3.4. Berekening van stijfheidsmatrix, massamatrix en belastingsvektor

Bij de berekening van de matrices S en M en de vektor F maken we gebruik van het feit dat er op ieder segment *hoogstens drie* basisfunkties $\neq 0$ zijn. Dit betekent dat het stelsel $\{\phi_i(x)\}$ "bijna orthogonaal" is, want $\phi_i(x)\phi_j(x)$ en $\phi_i'(x)\phi_j'(x)$ zijn alleen $\neq 0$, als x_i en x_j tot hetzelfde segment behoren. We bekijken nu de stijfheidsmatrix $S = (s_{ij})$, gedefinieerd door

$$s_{ij} = \int_a^b p(x)\phi_i'(x)\phi_j'(x)dx.$$

We kunnen S hier schrijven als

$$S = \sum_{k=1}^{N} S^{(k)},$$

waarbij

$$(3.4.1)\qquad s_{ij}^{(k)} = \int_{e_k} p(x)\phi_i'(x)\phi_j'(x)dx.$$

$S^{(k)}$ noemen we de k de *elementstijfheidsmatrix*. Aangezien op e_k alleen de basisfunkties $\phi_{k-1}, \phi_{k-\frac{1}{2}}$ en ϕ_k relevant zijn, bestaat $S^{(k)}$, op één klein vierkantje na, geheel uit nullen:

$$S^{(k)} = \begin{bmatrix} 0 & 0 & 0 \\ & * * * & \\ 0 & * * * & 0 \\ & * * * & \\ 0 & 0 & 0 \end{bmatrix} \begin{matrix} \\ \leftarrow\ k-1 \\ \\ \leftarrow\ k-\frac{1}{2} \\ \\ \leftarrow\ k \end{matrix} .$$

Om de notatie te vereenvoudigen, geven we in het vervolg $S^{(k)}$ aan met

$$S^{(k)} = \begin{bmatrix} * & * & * \\ * & * & * \\ * & * & * \end{bmatrix} \begin{matrix} \\ \leftarrow\ k-1 \\ \leftarrow\ k-\frac{1}{2} \\ \leftarrow\ k \end{matrix} .$$

Op een analoge manier definiëren we de *elementmassamatrix*

$$(3.4.2) \qquad M^{(k)} = \left(\int_{e_k} q(x)\phi_i(x)\phi_j(x)dx \right)$$

en de *elementbelastingsvektor*

$$(3.4.3) \qquad F^{(k)} = \left(\int_{e_k} f(x)\phi_i(x)dx \right).$$

Bij de berekening van de elementstijfheidsmatrix $S^{(k)}$, de elementmassamatrix $M^{(k)}$ en de elementbelastingsvektor $F^{(k)}$ maken we gebruik van de lineaire transformatie $x = h_k t + x_{k-1}$, $0 \le t \le 1$. Met behulp van formule (3.2.8) verkrijgen we dan de volgende formules (we gebruiken nu *lokale* indices)

$$(3.4.4) \qquad S^{(k)} = \left(\frac{1}{h_k} \int_0^1 p(h_k t+x_{k-1}) \frac{d\Phi_i}{dt} \frac{d\Phi_j}{dt} dt \right);$$

$$(3.4.5) \qquad M^{(k)} = \left(h_k \int_0^1 q(h_k t+x_{k-1})\Phi_i(t)\Phi_j(t)dt \right);$$

$$(3.4.6) \qquad F^{(k)} = \left(h_k \int_0^1 f(h_k t+x_{k-1})\Phi_i(t)dt \right),$$

waarbij $\Phi_i(t)$ $(i=0,\frac{1}{2},1)$ door (3.2.8) gegeven is. Voor k=1 krijgen we 2*2 i.p.v. 3*3 matrices $S^{(1)}$ en $M^{(1)}$ en een tweedimensionale vektor $F^{(1)}$, omdat $\Phi_0(t)$ dan niet meedoet.

Bij de berekening van de elementstijfheidsmatrix maken we gebruik van het feit dat

$$\Phi_0(t) + \Phi_{\frac{1}{2}}(t) + \Phi_1(t) = 1,$$

zodat voor $\Phi_{\frac{1}{2}}(t)$ geldt

$$\frac{d}{dt}\Phi_{\frac{1}{2}}(t) = -\frac{d\Phi_0}{dt} - \frac{d\Phi_1}{dt}.$$

Als gevolg hiervan is de matrix $S^{(k)}$ te schrijven als

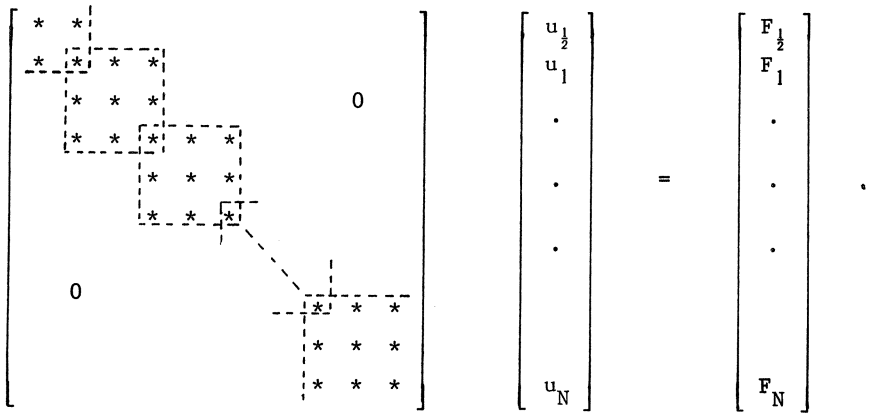
$$(3.4.7)\qquad S^{(k)} = \begin{bmatrix} s_{00}^{(k)} & -(s_{00}^{(k)}+s_{01}^{(k)}) & s_{01}^{(k)} \\ -(s_{00}^{(k)}+s_{01}^{(k)}) & s_{00}^{(k)}+2s_{01}^{(k)}+s_{11}^{(k)} & -(s_{01}^{(k)}+s_{11}^{(k)}) \\ s_{01}^{(k)} & -(s_{01}^{(k)}+s_{11}^{(k)}) & s_{11}^{(k)} \end{bmatrix},$$

waarbij

$$(3.4.8)\qquad s_{ij}^{(k)} = \frac{1}{h_k}\int_0^1 p(h_k t+x_{k-1})\,\frac{d\Phi_i}{dt}\,\frac{d\Phi_j}{dt}\,dt, \qquad i,j = 0,1,$$

zodat we voor de berekening van $S^{(k)}$ slechts drie integraties hoeven uit te voeren.

We berekenen nu de stijfheidsmatrix S, de massamatrix M (en daarmee A) en de belastingsvektor F, door *segment voor segment* de *elementmatrices* $S^{(k)}$ en $M^{(k)}$ en de *elementbelastingsvektor* $F^{(k)}$ te berekenen. De aldus berekende komponenten worden op de juiste plaats in de *globale* matrix A en de *globale* belastingsvektor F ingepast. We krijgen dan een lineair stelsel van de vorm



$$= \begin{bmatrix} F_{\frac{1}{2}} \\ F_1 \\ \cdot \\ \cdot \\ \cdot \\ F_N \end{bmatrix}.$$

We besluiten deze sektie door voor het geval $p(x) \equiv p_0 > 0$ en $q(x) \equiv q_0 \geq 0$ de elementmatrices $S^{(k)}$ en $M^{(k)}$ te geven:

$$(3.4.9)\qquad S^{(k)} = \frac{p_0}{3h_k}\begin{bmatrix} 7 & 8 & 7 \\ -8 & 16 & -8 \\ 1 & -8 & 7 \end{bmatrix}, \qquad k = 1,\ldots,N;$$

$$S^{(1)} = \frac{p_0}{3h_1}\begin{bmatrix} 16 & -8 \\ -8 & 7 \end{bmatrix};$$

$$M^{(k)} = \frac{q_0 h_k}{30}\begin{bmatrix} 4 & 2 & -1 \\ 2 & 16 & 2 \\ -1 & 2 & 4 \end{bmatrix}, \qquad k > 1,$$

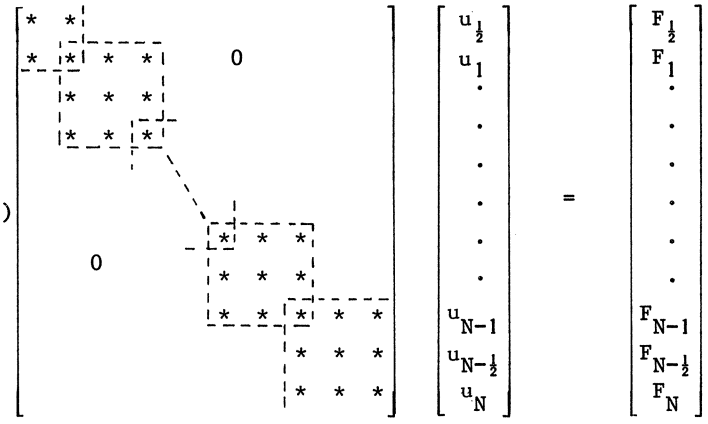
(3.4.10)

$$M^{(1)} = \frac{q_0 h_1}{30}\begin{bmatrix} 16 & 2 \\ 2 & 4 \end{bmatrix} .$$

3.5. Statische kondensatie

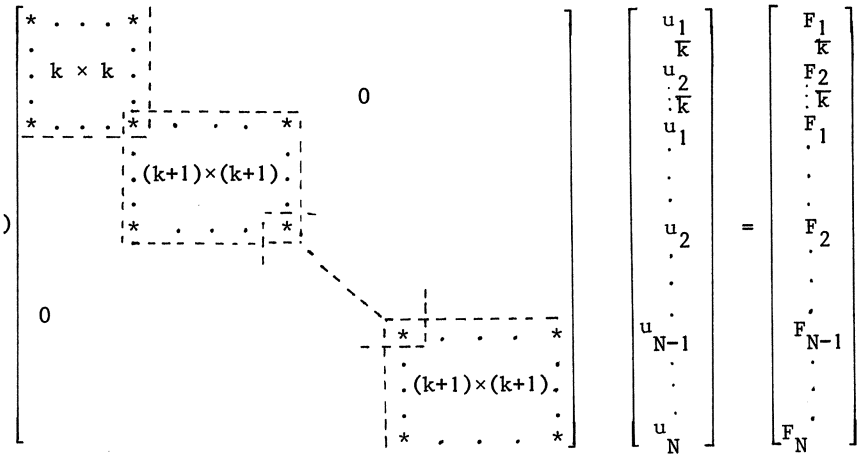
We hebben in 3.4 gezien dat we voor de oplossing van (3.3.1) een lineair stelsel van de vorm

(3.5.1)



moeten oplossen. Algemeen moeten we, als we de oplossing van (3.3.1) in $\mathbb{P}_k(\pi)$ benaderen, een stelsel vergelijkingen oplossen van de vorm

(3.5.2)



We bespreken nu een methode om lineaire stelsels van de vorm (3.5.2) voor k > 1 te reduceren tot *tridiagonale* stelsels, door u_i met gebroken indices *vooraf* te elimineren. Deze methode wordt door fysici *statistische kondensatie* genoemd. We illustreren e.e.a. voor k=2.

Voor komponenten met halftallige index luidt vergelijking (3.5.1)

$$a_{\frac12,\frac12}u_{\frac12} + a_{\frac12,1}u_1 = F_{\frac12},$$

$$a_{i-\frac12,i-1}u_{i-1} + a_{i-\frac12,i-\frac12}u_{i-\frac12} + a_{i-\frac12,i}u_i = F_{i-\frac12}, \quad i = 2,\ldots,N,$$

ofwel

$$(3.5.3)\qquad \begin{aligned} u_{\frac12} &= \frac{F_{\frac12}-a_{\frac12,1}u_1}{a_{\frac12,\frac12}}, \\ u_{i-\frac12} &= \frac{F_{i-\frac12}-a_{i-\frac12,i-1}u_{i-1}-a_{i-\frac12,i}u_i}{a_{i-\frac12,i-\frac12}}, \quad i > 1. \end{aligned}$$

Vergelijking (3.5.1) voor komponenten met geheeltallige indices is

$$(3.5.4)\qquad \begin{aligned} &a_{1,\frac12}u_{\frac12} + a_{1,1}u_1 + a_{1,1\frac12}u_{1\frac12} + a_{1,2}u_2 = F_1; \\ &a_{i,i-1}u_{i-1} + a_{i,i-\frac12}u_{i-\frac12} + a_{i,i}u_i + \\ &\qquad + a_{i,i+\frac12}u_{i+\frac12} + a_{i,i+1}u_{i+1} = F_i, \quad 1 < i < N; \\ &a_{N,N-1}u_{N-1} + a_{N,N-\frac12}u_{N-\frac12} + a_{N,N}u_N = F_N. \end{aligned}$$

Substitueren we (3.5.3) in (3.5.4), dan krijgen we

$$(3.5.5)\qquad \begin{aligned} &\left(a_{1,1} - \frac{a_{1,\frac12}a_{\frac12,1}}{a_{\frac12,\frac12}} - \frac{a_{1,1\frac12}a_{1\frac12,1}}{a_{1\frac12,1\frac12}}\right)u_1 + \\ &+ \left(a_{1,2} - a_{1,1\frac12}a_{1\frac12,2}\right)u_2 = F_1 - \frac{a_{1,\frac12}F_{\frac12}}{a_{\frac12,\frac12}}; \\ &\left(a_{i,i-1} - \frac{a_{i,i-\frac12}a_{i-\frac12,i-1}}{a_{i-\frac12,i-\frac12}}\right)u_{i-1} \\ &+ \left(a_{i,i} - \frac{a_{i,i-\frac12}a_{i-\frac12,i}}{a_{i-\frac12,i-\frac12}} - \frac{a_{i,i+\frac12}a_{i+\frac12,i}}{a_{i+\frac12,i+\frac12}}\right)u_i \end{aligned}$$

$$+\left(a_{i,i+1} - \frac{a_{i,i+\frac{1}{2}}a_{i+\frac{1}{2},i+1}}{a_{i+\frac{1}{2},i+\frac{1}{2}}}\right)u_{i+1} =$$

$$= F_i - \frac{a_{i,i-\frac{1}{2}}F_{i-\frac{1}{2}}}{a_{i-\frac{1}{2},i-\frac{1}{2}}} - \frac{a_{i,i+\frac{1}{2}}F_{i+\frac{1}{2}}}{a_{i+\frac{1}{2},i+\frac{1}{2}}}, \qquad 1 < i < N;$$

$$\left(a_{N,N-1} - \frac{a_{N,N-\frac{1}{2}}a_{N-\frac{1}{2},N-1}}{a_{N-\frac{1}{2},N-\frac{1}{2}}}\right)u_{N-1}$$

$$+\left(a_{N,N} - \frac{a_{N,N-\frac{1}{2}}a_{N-\frac{1}{2},N}}{a_{N-\frac{1}{2},N-\frac{1}{2}}}\right)u_N = F_N - \frac{a_{N,N-\frac{1}{2}}F_{N-\frac{1}{2}}}{a_{N-\frac{1}{2},N-\frac{1}{2}}}.$$

De matrixelementen a_{ij} en de vektorkomponenten F_i met geheeltallige i en j ondergaan dus transformaties van de vorm

$$\begin{aligned}
a_{i,i} &\to a_{i,i} + (\ldots)a_{i-\frac{1}{2},i} + (\ldots)a_{i+\frac{1}{2},i}, \\
a_{i,i-1} &\to a_{i,i-1} + (\ldots)a_{i-\frac{1}{2},i-1}, \\
a_{i,i+1} &\to a_{i,i+1} + (\ldots)a_{i+\frac{1}{2},i+1}, \\
F_i &\to F_i + (\ldots)F_{i-\frac{1}{2}} + (\ldots)F_{i+\frac{1}{2}},
\end{aligned}$$

waarbij de koefficienten (...) uit (3.3.5) afgeleid kunnen worden. Het interessante van deze transformatie is dat ze, evenals de evaluatie van de matrix (a_{ij}) en de belastingsvektor (F_i), segmentsgewijs uitgevoerd kan worden. We illustreren dit met de volgende afbeeldingen, waarbij we lokale indices gebruiken:

$$A^{(k)} = \begin{bmatrix} a_{0,0} & a_{0,\frac{1}{2}} & a_{0,1} \\ a_{\frac{1}{2},0} & a_{\frac{1}{2},\frac{1}{2}} & a_{\frac{1}{2},1} \\ a_{1,0} & a_{1,\frac{1}{2}} & a_{1,1} \end{bmatrix} \begin{matrix} \leftarrow k-1 \\ \leftarrow k-\frac{1}{2} \\ \leftarrow k \end{matrix},$$

$$F^{(k)} = \begin{bmatrix} F_0 \\ F_{\frac{1}{2}} \\ F_1 \end{bmatrix} \begin{matrix} \leftarrow k-1 \\ \leftarrow k-\frac{1}{2} \\ \leftarrow k \end{matrix},$$

$$A^{(1)} = \begin{bmatrix} a_{\frac{1}{2},\frac{1}{2}} & a_{\frac{1}{2},1} \\ a_{1,\frac{1}{2}} & a_{1,1} \end{bmatrix} \begin{matrix} \leftarrow \frac{1}{2} \\ \leftarrow 1 \end{matrix},$$

$$F^{(1)} = \begin{bmatrix} F_{\frac{1}{2}} \\ F_1 \end{bmatrix} \begin{matrix} \leftarrow \frac{1}{2} \\ \leftarrow 1 \end{matrix}.$$

De elementmatrices en -vektoren ondergaan dan de volgende verschuivingen:

$$A^{(k)} \to \begin{bmatrix} a_{0,0} - \frac{a_{0,\frac{1}{2}}a_{\frac{1}{2},0}}{a_{\frac{1}{2},\frac{1}{2}}} & a_{0,1} - \frac{a_{0,\frac{1}{2}}a_{\frac{1}{2},1}}{a_{\frac{1}{2},\frac{1}{2}}} \\ a_{1,0} - \frac{a_{1,\frac{1}{2}}a_{\frac{1}{2},0}}{a_{\frac{1}{2},\frac{1}{2}}} & a_{1,1} - \frac{a_{1,\frac{1}{2}}a_{\frac{1}{2},1}}{a_{\frac{1}{2},\frac{1}{2}}} \end{bmatrix} = C^{(k)}, \quad k > 1,$$

$$F^{(k)} \to \begin{bmatrix} F_0 - \frac{a_{0,\frac{1}{2}}F_{\frac{1}{2}}}{a_{\frac{1}{2},\frac{1}{2}}} \\ F_1 - \frac{a_{1,\frac{1}{2}}F_{\frac{1}{2}}}{a_{\frac{1}{2},\frac{1}{2}}} \end{bmatrix} = G^{(k)}, \quad k > 1,$$

$$A^{(1)} \to \begin{bmatrix} a_{1,1} - \frac{a_{1,\frac{1}{2}}a_{\frac{1}{2},1}}{a_{\frac{1}{2},\frac{1}{2}}} \end{bmatrix} = C^{(1)},$$

$$F^{(1)} \to \begin{bmatrix} F_1 - \frac{a_{1,\frac{1}{2}}F_{\frac{1}{2}}}{a_{\frac{1}{2},\frac{1}{2}}} \end{bmatrix} = G^{(1)}.$$

We kunnen nu (3.5.5) als volgt schrijven:

$$Cv = G,$$

met

$$v = (u_1, u_2, \ldots, u_N)^T,$$

$$C = \sum_{k=1}^{N} C^{(k)},$$

$$G = \sum_{k=1}^{N} G^{(k)}.$$

Voor $k > 2$ gaat de kondensatie analoog, zij het ingewikkelder. We gaan hier niet op in omdat het een al te technische kwestie is.

3.6. Stuksgewijze Hermite-interpolatie

We brengen weer een verdeling π aan van het interval $[a,b]$. Deze verdeling heeft de maaswijde $h = \max(x_j - x_{j-1})$.

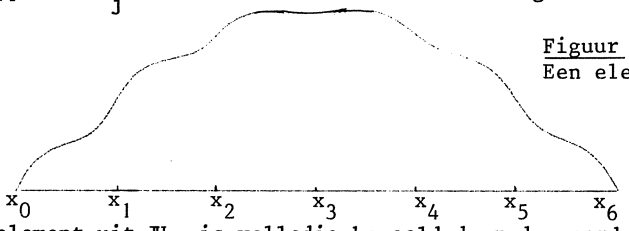
DEFINITIE 3.6.1 Onder $\mathbb{H}_k(\pi)$ verstaan we de verzameling van alle funkties die

(i) een kontinue nulde en eerste afgeleide op [a,b] hebben;

(ii) op ieder segment e_j een polynoom van de graad $\leq k$ zijn.

Uit het feit dat een element van $\mathbb{H}_k(\pi)$ op ieder segment e_j *vier* vrijheidsgraden heeft, volgt dat $k \geq 3$ moet zijn. We bekijken in deze sektie alleen $\mathbb{H}_3(\pi)$, ook wel aangegeven met $\mathbb{H}_3$.

Een element $\mathbb{H}_3$ is een stuksgewijs kubische funktie, die in de knooppunten x_j een kontinue nulde en eerste afgeleide heeft.



Figuur 3.5
Een element uit $\mathbb{H}_3$

Een element uit $\mathbb{H}_3$ is volledig bepaald door de waarden van de nulde en eerste afgeleide in de punten $x_0,\ldots,x_N$. Aangezien in ieder punt aan *twee* kondities moet worden voldaan, worden de punten $x_0,\ldots,x_N$ in het geval van de $\mathbb{H}_3$ *dubbele knooppunten* genoemd.

Het is eenvoudig in te zien dat de $\mathbb{H}_3$ een eindigdimensionale deelruimte van de $H^2[a,b]$ is, i.e. de klasse van funkties waarvan de nulde, eerste en tweede afgeleide kwadratisch integreerbaar zijn over [a,b]. In deze ruimte zijn het inprodukt

$$(3.6.1a) \qquad (u,v)_2 = \int_a^b [u''v''+u'v'+uv]dx$$

en de norm

$$(3.6.1b) \qquad \|u\|_2 = \sqrt{(u,u)_2}$$

gedefinieerd.

De basisfunkties van $\mathbb{H}_3$ zijn verdeeld in twee klassen:

(i) de funkties $\phi_i(x)$ die in alle punten de afgeleide nul hebben, in x_i de funktiewaarde 1 hebben en in alle overige punten de funktiewaarde nul hebben: $\phi_i(x_j) = \delta_{ij}$ en $\phi_i'(x_j) = 0$, $0 \leq i,j \leq N$;

(ii) de funkties $\psi_i(x)$ die in alle punten de funktiewaarde nul hebben, in het punt x_i de afgeleide 1 hebben en in alle overige punten de funktiewaarde 0: $\psi_i(x_j) = 0$ en $\psi_i'(x_j) = \delta_{ij}$, $0 \leq i,j \leq N$.

Iedere $v_h \in \mathbb{H}_3$ kan nu als volgt worden gerepresenteerd:

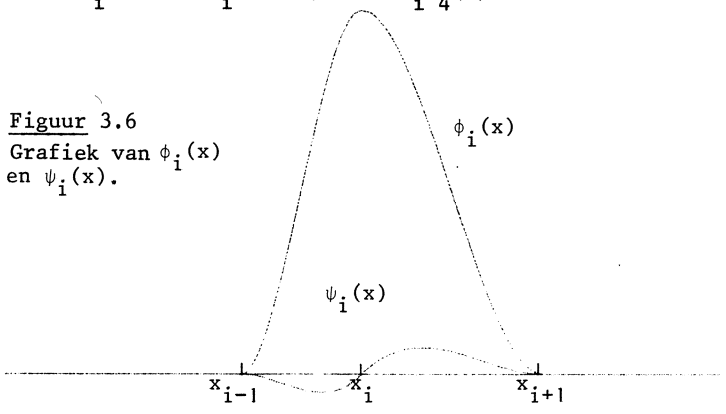
$$
\begin{aligned}
v_h(x) &= \sum_{j=0}^{N} [q_j\phi_j(x)+r_j\psi_j(x)], \\
q_j &= v_h(x_j), \\
r_j &= v_h'(x_j), \quad j = 0,\ldots,N.
\end{aligned}
\tag{3.6.2}
$$

Ook hier zijn de basisfunkties $\phi_i(x)$ en $\psi_i(x)$ alleen $\neq 0$ op die segmenten e_i waartoe x_i behoort, dus op hooguit twee segmenten. Op ieder segment $e_i = [x_{i-1},x_i]$ zijn zo alleen de basisfunkties $\phi_{i-1}(x)$, $\psi_{i-1}(x)$, $\phi_i(x)$ en $\psi_i(x)$ relevant. We geven hieronder de formules van deze basisfunkties:

$$
\begin{aligned}
\phi_{i-1}(x) &= \left(\frac{x-x_i}{h_i}\right)^2\left(2\,\frac{x-x_{i-1}}{h_i} + 1\right), \\
\psi_{i-1}(x) &= \left(\frac{x-x_i}{h_i}\right)^2\left(x-x_{i-1}\right), \\
\phi_i(x) &= \left(\frac{x-x_{i-1}}{h_i}\right)^2\left(2\,\frac{x-x_i}{h_i} - 1\right), \\
\psi_i(x) &= \left(\frac{x-x_{i-1}}{h_i}\right)^2\left(x-x_i\right), \qquad x_{i-1} \le x \le x_i.
\end{aligned}
\tag{3.6.3}
$$

Substitueren we $x = h_i t + x_{i-1}$, $0 \le t \le 1$, dan gaat formule (3.6.3) over in

$$
\begin{aligned}
\phi_{i-1}(x) &= (t-1)^2(2t+1) = \Phi_1(t), \\
\psi_{i-1}(x) &= h_i t(t-1)^2 = h_i\Phi_2(t), \\
\phi_i(x) &= t^2(3-2t) = \Phi_3(t), \\
\psi_i(x) &= h_i t^2(t-1) = h_i\Phi_4(t).
\end{aligned}
\tag{3.6.4}
$$



Figuur 3.6
Grafiek van $\phi_i(x)$ en $\psi_i(x)$.

We zien nu een belangrijk verschil tussen de basisfunkties ϕ_i en ψ_i. De eerste zijn derdegraadspolynomen in t, terwijl de laatste eveneens derdegraadspolynomen in t zijn, maar *vermenigvuldigd met* h_i. Die vermenigvuldiging is nodig om de invloed van de transformatie op de differentiatie ongedaan te maken.

We merken nog op dat

$$\Phi_3(t) = \Phi_1(1-t) = 1 - \Phi_1(t),$$
$$\Phi_4(t) = -\Phi_2(1-t). \tag{3.6.5}$$

Bij een gegeven verdeling π van [a,b] is de basis van $\mathbb{H}_3(\pi)$ dus volledig bepaald door de "superfunkties" Φ_1 en Φ_2 en door de getallen $h_i = x_i - x_{i-1}$, $i = 1,\ldots,N$.

3.7. Tweede orde variatieproblemen

Gegeven het variatieprobleem: minimaliseer

$$I(v) = \int_a^b \left[p(x)[v''(x)]^2 + q(x)[v'(x)]^2 + r(x)[v(x)]^2 - 2f(x)v(x) \right] dx, \tag{3.7.1}$$

onder de nevenvoorwaarden

$$v(a) = v'(a) = 0,$$
$$p \in H^2[a,b],\ q(x) \in H^1[a,b],\ r(x), f(x) \in H^0[a,b], \tag{3.7.1a}$$
$$p(x) \geq p_0 > 0,\quad q(x), r(x) \geq 0.$$

Volgens de theorie van hoofdstuk 2 heeft (3.7.1) een oplossing in de separabele energieruimte

$$H_M = \{v \mid v \in H^2[a,b];\ v(a) = v'(a) = 0\}.$$

In deze ruimte definiëren we het energieinprodukt

$$(u,v)_M = \int_a^b \left[p(x)u''(x)v''(x) + q(x)u'(x)v'(x) + r(x)u(x)v(x) \right] dx \tag{3.7.2a}$$

en de energienorm

(3.7.2b) $\|u\|_M = \sqrt{(u,u)_M}$.

We willen nu voor de oplossing u van (3.7.1) een approximatie u_h in een eindigdimensionale deelruimte V_h van H_M berekenen. We definiëren V_h als volgt:

$$V_h = \{v \mid v \in \mathbb{H}_3(\pi);\ v(a) = v'(a) = 0\},$$

waarbij π een verdeling van [a,b] is met maaswijdte h. Het is gemakkelijk na te gaan dat V_h met definitie (3.7.2) van inprodukt en norm een eindigdimensionale deelruimte van H_M is.

We kunnen nu voor probleem (3.7.1) op twee manieren een approximatie berekenen:

(i) minimaliseer

(3.7.3) $I(v_h), \qquad v_h \in V_h;$

(ii) minimaliseer

(3.7.4) $I(v_h), \qquad v_h \in \mathbb{H}_3(\pi),$

onder de voorwaarden

(3.7.4a) $v_h(a) = v_h'(a) = 0.$

De laatste aanpak is equivalent aan de eerste, aangezien V_h een deelruimte van $\mathbb{H}_3(\pi)$ is. Evenwel kan (ii) ook toegepast worden in het geval van inhomogene randvoorwaarden, zodat (ii) programmeertechnisch te verkiezen is.

Voor het gemak worden alle basisfunkties van $\mathbb{H}_3$ nu aangegeven met de letter ϕ. Met $\phi_{2i}(x)$ geven we de oorspronkelijke funkties $\phi_i(x)$ uit 3.6 aan, met $\phi_{2i+1}(x)$ geven we de funkties $\psi_i(x)$ uit 3.6 aan. Voor $\phi_i(x)$ gelden dan de relaties

$$\begin{aligned} &\phi_{2i}(x_j) = \phi'_{2i+1}(x_j) = \delta_{ij}, \\ &\phi'_{2i}(x_j) = \phi_{2i+1}(x_j) = 0 \quad , \qquad 0 \le i,j \le N. \end{aligned} \tag{3.7.5}$$

Dit betekent dat op het interval $[x_{i-1},x_i]$ de funkties ϕ_{2i-2}, ϕ_{2i-1}, ϕ_{2i} en ϕ_{2i+1} als volgt zijn gedefinieerd (zie ook 3.6.4):

$$
\begin{aligned}
&\phi_{2i-2}(x) = \Phi_1(t) && ,\\
&\phi_{2i-1}(x) = h_i\Phi_2(t) && ,\\
&\phi_{2i}(x) = 1 - \Phi_1(t) = \Phi_3(t) && ,\\
&\phi_{2i+1}(x) = h_i\Phi_4(t) && .
\end{aligned}
\tag{3.7.6}
$$

Schrijven we nu

$$v_h(x) = \sum_{i=0}^{2N+1} a_i\phi_i(x),$$

dan gaat (3.7.4) over in: minimaliseer

$$I(v_h) = G(a_0,a_1,\ldots,a_{2N+1})$$

onder de nevenvoorwaarden

$$a_0 = a_1 = 0.$$

De oplossingsvektor $(q_0,q_1,\ldots,q_{2N+1})^T$ wordt gegeven door

$$\left.\frac{\partial G}{\partial a_i}\right|_{a_i=q_i} = 0, \qquad i = 2,\ldots,2N+1,$$

$$q_0 = q_1 = 0.$$

Uitwerking hiervan geeft het lineaire stelsel

$$
\begin{aligned}
&\sum_{j=0}^{2N+1} (\phi_i,\phi_j)_M\, q_j = (f,\phi_i), && i = 2,\ldots,2N+1,\\
&q_i = 0, && i = 0,1.
\end{aligned}
\tag{3.7.7a}
$$

De matrix van het stelsel (3.7.7a) is asymmetrisch. We maken hem symmetrisch door (3.7.7a) om te zetten in

$$
\begin{aligned}
&\sum_{j=2}^{2N+1} (\phi_i,\phi_j)_M\, u_j = (f,\phi_i) - \sum_{j=0}^{1} (\phi_i,\phi_j)_M\, u_j, \quad i = 2,\ldots,2N+1.\\
&u_0 = u_1 = 0.
\end{aligned}
\tag{3.7.7b}
$$

Formule (3.7.7b) doet nogal overbodig aan. Immers $u_0 = u_1 = 0$, zodat er in feite niets wordt opgeteld bij het rechterlid van (3.7.7a). Hebben we evenwel te doen met *inhomogene* randvoorwaarden, dan blijkt (3.7.7b) een vrij zinvolle formule. We gaan dan als volgt te werk:

(i) eerst worden de matrix $A^* = ((\phi_i,\phi_j)_M)$ en de vektor $F^* = ((f,\phi_i))$ segmentsgewijs geëvalueerd;

(ii) vervolgens worden de inhomogene randvoorwaarden $q_0 = \alpha$, $q_1 = \beta$ geïmplementeerd door de matrix $A = (a_{ij})$ en de vektor $F = (F_i)$ als volgt te definiëren (zie ook 2.8):

$$
\begin{aligned}
a_{ij} &= \begin{cases} \delta_{ij}, & i = 0,1,\ j = 0,\ldots,2N+1, \\ (\phi_i,\phi_j)_M, & i = 2,\ldots,2N+1,\ j=0,\ldots,2N+1, \end{cases} \\
F_i &= \begin{cases} (f,\phi_i) - (\phi_i,\phi_0)q_0 - (\phi_i,\phi_i)q_1, & i > 1, \\ u_i, & i = 0,1. \end{cases}
\end{aligned} \tag{3.7.8}
$$

De oplossing van (3.7.7b) is nu identiek met de oplossing van

$$Aq = F,$$

waarbij A en F door (3.7.8) worden gegeven.

De hierboven beschreven methode is vooral goed hanteerbaar in het geval van meerdimensionale minimaliseringsproblemen, als er sprake is van verschillende randvoorwaarden.

Berekening van A

We bouwen A elementsgewijs op. Voor $A^{(k)}$ geldt de formule

$$a_{ij}^{(k)} = \int_{e_k} \left[p(x)\phi_i''\phi_j'' + q(x)\phi_i'\phi_j' + r(x)\phi_i\phi_j\right]dx,$$

vervolgens passen we de transformatie $x = h_K t + x_{K-1}$ toe en we gebruiken weer lokale indices. Met behulp van formule (3.6.4) krijgen we

$$A^{(k)} = \begin{bmatrix} a_{11} & h_k a_{12} & a_{13} & h_k a_{14} \\ h_k a_{12} & h_k^2 a_{22} & h_k a_{23} & h_k^2 a_{24} \\ a_{13} & h_k a_{13} & a_{33} & h_k a_{34} \\ h_k a_{14} & h_k^2 a_{24} & h_k a_{34} & h_k^2 a_{44} \end{bmatrix},$$

waarbij,

$$a_{ij} = \frac{1}{h_k^3} \int_0^1 p(h_k t+x_{k-1}) \frac{d^2\Phi_i}{dt^2} \frac{d^2\Phi_j}{dt^2} dt$$
$$+ \frac{1}{h_k} \int_0^1 q(h_k t+x_{k-1}) \frac{d\Phi_i}{dt} \frac{d\Phi_j}{dt} dt$$
$$+ h_k \int_0^1 r(h_k t+x_{k-1}) \Phi_i(t)\Phi_j(t)dt$$
$$= \frac{1}{h_k^3} b_{ij} + \frac{1}{h_k} s_{ij} + h_k m_{ij}, \qquad 1 \le i, \ j \le 4. \tag{3.7.9}$$

De matrices (b_{ij}), (s_{ij}) en (m_{ij}) heten respektievelijk *elementbuigingsmatrix*, *elementstijfheidsmatrix* en *elementmassamatrix*. Voor de elementbelastingsvektor $F^{(k)}$ kunnen we de formule

$$F^{(k)} = h_k \begin{bmatrix} F_1 \\ h_k F_2 \\ F_3 \\ h_k F_4 \end{bmatrix}$$

afleiden, waarbij

$$F_i = \int_0^1 f(h_k t+x_{k-1})\Phi_i(t)dt, \quad i = 1,2,3,4. \tag{3.7.10}$$

We merken op dat de elementbuigingsmatrix $B^{(k)}$ dubbel singulier is (0 is dubbele eigenwaarde). Er geldt namelijk

$$\Phi_1(t) + \Phi_3(t) = 1,$$

$$\Phi_2(t) + \Phi_3(t) + \Phi_4(t) = t,$$

waaruit na toepassing van (3.7.9) volgt dat

$$b_{i1} + b_{i3} = 0,$$

$$b_{i2} + b_{i3} + b_{i4} = 0, \qquad i = 1,\ldots,4.$$

Verder is de elementstijfheidsmatrix $S^{(k)}$ enkelvoudig singulier.

Tot slot van deze sektie geven we de formules voor $B^{(k)}$, $S^{(k)}$ en $M^{(k)}$ in het geval $p(x) \equiv p_0$, $q(x) \equiv q_0$, $r(x) \equiv r_0$ (zie ook STRANG & FIX [1973]). We veronderstellen $h_k = h$, $k = 1,\ldots,N$.

$$B^{(k)} = \frac{p_0}{h^3}\begin{bmatrix} 12 & 6h & -12 & 6h \\ 6h & 4h^2 & -6h & 2h^2 \\ -12 & -6h & 12 & -6h \\ 6h & 2h^2 & -6h & 4h^2 \end{bmatrix};$$

(3.7.11)

$$S^{(k)} = \frac{q_0}{30h}\begin{bmatrix} 36 & 3h & -36 & 3h \\ 3h & 4h^2 & -3h & -h^2 \\ -36 & -3h & 36 & -3h \\ 3h & -h^2 & -3h & 4h^2 \end{bmatrix};$$

$$M^{(k)} = \frac{hr_0}{420}\begin{bmatrix} 156 & 22h & 54 & -13h \\ 22h & 4h^2 & 13h & -3h^2 \\ 54 & 13h & 156 & -22h \\ -13h & -3h^2 & -22h & 4h^2 \end{bmatrix}.$$

3.8. Voorbeeld

We willen het volgende probleem oplossen: minimaliseer

$$I(v) = \int_0^1 [v''(x)]^2+2[v'(x)]^2+[v(x)]^2-2v(x)\,dx,$$

(3.8.1)

$$v(0) = v(1) = v'(0) = v'(1) = 0.$$

Het daaruit voortkomende randwaardeprobleem

$$u^{iv} - 2u'' + u = 1, \qquad 0 < x < 1,$$

$$u(0) = u(1) = u'(0) = u'(1) = 0$$

heeft de analytische oplossing

$$u(x) = 1 + e^{x}(c_1+c_2x) + e^{-x}(c_3+c_4x),$$

$$c_1 = -\frac{(2e-1)(e^2-2e-1)}{e^4-6e^2+1},$$

$$c_2 = \frac{(e-1)(e^2-2e-1)}{e^4-6e^2+1}, \qquad (3.8.2)$$

$$c_3 = -1 - c_1,$$

$$c_4 = c_3 - c_1 - c_2 .$$

Zij

$$\pi : 0 = x_0 < x_1 < \ldots < x_N = 1$$

een uniforme partitie van $[0,1]$, i.e. $x_i = hi$, $i = 0,\ldots,N$, $h = 1/N$. We benaderen de oplossing van (3.8.1) door het volgende probleem op te lossen: Minimaliseer

$$I(v_h), \qquad v_h \in \mathbb{H}_3(\pi),$$
$$v_h(0) = v_h'(0) = v_h(1) = v_h'(1). \qquad (3.8.3)$$

De oplossing u_h van (3.8.3) wordt gegeven door

$$u_h(x) = \sum_{j=0}^{2N+1} q_j\phi_j(x),$$

waarbij $(q_0,q_1,\ldots,q_{2N},q_{2N+1})$ bepaald wordt door het lineaire stelsel

$$\sum_{j=0}^{2N+1} a_{ij}q_j = F_i, \qquad i = 0,\ldots,2N+1. \qquad (3.8.4)$$

(a_{ij}) en (F_i) worden gegeven door

$$a_{ij} = \begin{cases} \delta_{ij}, & i = 0,1,2N,2N+1,\ j = 0,\ldots,2N+1; \\ \int_0^1 [\phi_i''\phi_j''+2\phi_i'\phi_j'+\phi_i\phi_j]dx, & i = 2,\ldots,2N-1,\ j = 0,\ldots,2N+1. \end{cases}$$

$$F_i = \begin{cases} 0, & i = 0,1,2N,2N+1, \\ \int_0^1 \phi_i dx, & i = 2,\ldots,2N-1. \end{cases}$$

We bouwen (a_{ij}) en (F_i) elementsgewijs op met behulp van de formules (3.7.8) - (3.7.12). We vinden dan voor de belastingsvektor F en de *bovendriehoek* van de symmetrische matrix A de volgende waarden:

$$a_{2i,2i} = \begin{cases} 1, & i = 0,N, \\ \frac{24}{h^3} + \frac{24}{5h} + \frac{26}{35}h, & i = 1,\ldots,N-1; \end{cases}$$

$$a_{2i,2i+1} = \begin{cases} 0, & i = 0,N-1,N, \\ -\frac{12}{h^3} - \frac{12}{5h} + \frac{9}{70}h, & i = 1,\ldots,N-2. \end{cases}$$

$$a_{2i,2i+2} = \begin{cases} 0, & i = 0,N-1,N, \\ \frac{6}{h^2} + \frac{1}{5} - \frac{13}{420}h^2, & i = 1,\ldots,N-2; \end{cases}$$

$$a_{2i-1,2i-1} = \begin{cases} 1 & i = 1,N, \\ \frac{8}{h} + \frac{8}{15}h + \frac{2}{105}h^3, & i = 2,\ldots,N-1; \end{cases}$$

$$a_{2i-1,2i+1} = \begin{cases} 0 & i = 1,N-1, \\ \frac{2}{h} - \frac{h}{15} - \frac{h^3}{140}, & i = 2,\ldots,N-2; \end{cases}$$

$$a_{i,j} = 0, \text{ alle overige waarden van i en j;}$$

$$F_i = \begin{cases} 0, & i = 0,1,3,5,\ldots,2N-1,2N, \\ h, & i = 2,4,6,\ldots,2N-2. \end{cases}$$

Enige resultaten:

N	$\max_i \lvert u(x_i)-u_h(x_i)\rvert$	$\max_i \lvert u'(x_i)-u_h(x_i)\rvert$
10	$9.47_{10}{}^{-10}$	$3.07_{10}{}^{-9}$
20	$6.15_{10}{}^{-11}$	$1.91_{10}{}^{-10}$
40	$3.40_{10}{}^{-12}$	$1.07_{10}{}^{-11}$

Merk op dat de maximale fout in de funktiewaarde en de afgeleide met ca. 16 afneemt, als N wordt verdubbeld (en h gehalveerd). Blijkbaar is de orde van nauwkeurigheid in de roosterpunten gelijk aan 4. Deze *lokale* fouten zijn nauwkeuriger dan de *globale* fout die voor de funktiewaarde van de orde 4 is en voor de afgeleide van de orde 3. Samengevat:

$$\text{(3.8.5)}\qquad \begin{aligned} &\|u-u_h\|_\ell \le C\|u\|_4\, h^{4-\ell}, && \ell = 0,1;\\ &\left|\frac{d^\ell}{dx^\ell}(u-u_h)(x_i)\right| \le C\|u\|_4\, h^4, && \ell = 0,1;\ i = 1,\ldots,N-1. \end{aligned}$$

Dit wordt bewezen in BAKKER [1976]. Zie voor meer "superkonvergentie" het volgende hoofdstuk.

3.9. Tweedimensionale variatieproblemen

Bij het oplossen van tweedimensionale randwaardeproblemen of daarmee ekwivalente variatieproblemen door middel van de eindige elementenmethode doen zich een aantal problemen voor die in het eendimensionale geval onbekend zijn.

(i) Een gebied Ω kan op verschillende manieren in kleine deelgebieden of elementen worden verdeeld, bijv. in driehoeken, vierhoeken of elementen met kromme randen.

(ii) Wanneer Ω een kromme rand heeft, is een partitie van Ω in driehoeken niet mogelijk. We kunnen in dat geval gebruik maken van de zgn. isoparametrische elementen. Behalve in § 3.13, waar we de isoparametrische

behandelen, veronderstellen we Ω steeds polygonaal, zodat Ω in driehoekjes verdeeld kan worden.

(iii) Een partitie van Ω moet aan zekere voorwaarden voldoen. Zo mogen twee driehoekjes óf één zijde óf één hoekpunt óf niets gemeen hebben. Verder moet in ieder driehoekje de verhouding tussen de grootste zijde en de straal van de grootste ingeschreven cirkel aan zekere regulariteitseisen voldoen. We komen daar in §3.10 op terug.

(iv) Voor de opbouw van een eindige elementenruimte V_h gebruiken we stuksgewijze polynomen als basisfunkties. De basis van V_h kan evenwel ook *onvolledig* zijn: hoewel de basisfunkties polynomen van de graad t zijn, behoort niet elk polynoom van de graad $\leq$ t tot V_h.

Voorbeeld Zij V_h opgespannen door stuksgewijze polynomen van de vorm

$$(a_0+\ldots+a_k x^k)(b_0+\ldots+b_k y^k), \quad a_k b_k \neq 0.$$

Dan zijn de basisfunkties polynomen van de graad 2k, maar de ruimte V_h bevat niet alle polynomen van de graad t ($k+1 \leq t \leq 2k$). Deze zgn. *bikomplete* ruimten worden gebruikt bij partities in vierhoeken. We zullen ons in dit hoofdstuk beperken tot eerste orde zelfgeadjungeerde, positief definiete variatieproblemen, d.w.z. problemen van de vorm:

Minimaliseer

$$I(v) = \iint_{\Omega} [(\bar{\nabla}v \cdot P\bar{\nabla}v) + q(x,y)v^2 - 2f(x,y)v]dxdy,$$

onder de randvoorwaarden

$$v = 0 \quad \text{op } \partial\Omega,$$

waarbij $P = P(x,y)$ een 2×2 symmetrische, positief definiete matrix is en $q \geq 0$. We zullen aandacht besteden aan stuksgewijs Lagrange- en Hermite-interpolaties, aan de assemblage van de globale stijfheidsmatrix, massamatrix en belastingsvektor en aan de isoparametrische elementen. Tot slot leiden we beknopt een schatting of van de globale fout.

3.10. Lagrange-interpolatie in twee dimensies

Laat $\Omega \subset \mathbb{R}^2$ een begrensd polygonaal gebied zijn. We verdelen Ω in M driehoekjes e_ℓ en geven deze verdeling aan met

$$(3.10.1) \qquad \pi = \{e_\ell\}_{\ell=1}^{M}, \quad \bigcup_{\ell=1}^{M} e_\ell = \Omega.$$

We definiëren voor $\ell = 1,\ldots,M$

$$\begin{aligned} & h_\ell = \text{diameter } (e_\ell); \\ (3.10.2) \qquad & \rho_\ell = \text{straal van de grootste in } e_\ell \text{ geschreven cirkel}; \\ & h = \max h_\ell, \text{ de maaswijdte van } \pi. \end{aligned}$$

Deze partitie π moet aan zekere eisen van regulariteit voldoen.

Definitie 3.10.1 Een familie van partities π met maaswijdte h van Ω heet regulier, als er een konstante $C > 0$ onafhankelijk van M en h is met

$$(3.10.3) \qquad \rho_\ell \geq C\, h_\ell, \quad \ell = 1,\ldots,M.$$

Volgens ZLAMÀL [1970] is deze definitie ekwivalent met de eigenschap dat er een konstante ϕ_0 bestaat, $0 < \phi_0 \leq \frac{\pi}{3}$, ϕ_0 onafhankelijk van M en h, met de eigenschap

$$\phi_0 \leq \alpha_\ell, \beta_\ell, \gamma_\ell, \quad \ell = 1,\ldots,M,$$

waarbij α_ℓ, β_ℓ en γ_ℓ de hoeken van e_ℓ zijn.

Definitie 3.10.2 Zij $\pi = \{e_\ell\}_{\ell=1}^{M}$ een reguliere partitie van Ω. Dan wordt $\mathbb{P}_k(\pi)$ gedefinieerd als de ruimte van funkties die

(i) kontinu zijn op $\bar{\Omega}$;

(ii) op ieder driehoekje e_ℓ ($\ell = 1,\ldots,M$) een polynoom in x en y van de graad $\leq k$ zijn.

Het is duidelijk dat $\mathbb{P}_k(\pi)$ een eindigdimensionale deelruimte is van

$$H^1(\Omega) = \{v \mid v,\ v_x, v_y \in L^2(\Omega)\}.$$

We geven de basisfunkties van $\mathbb{P}_k(\pi)$ voor $k = 1,2,3$.

$\underline{k = 1}$

Een element van $\mathbb{P}(\pi)$ is een stuksgewijs lineaire funktie in x en y en is geheel bepaald door de waarden die het in de hoekpunten van de driehoekjes aanneemt. Deze N hoekpunten kiezen we als steunpunten voor een basis van $\mathbb{P}_1(\pi)$. De steunpunten geven we aan met

$$z_i = (x_i, y_i)^T, \qquad i = 1,\ldots,N.$$

De basisfunkties $\phi_i(x,y)$ kiezen we nu lineair op elke e_ℓ en wel zo, dat

$$\phi_i(z_j) = \delta_{ij}, \qquad 1 \le i,\ j \le N. \tag{3.10.4}$$

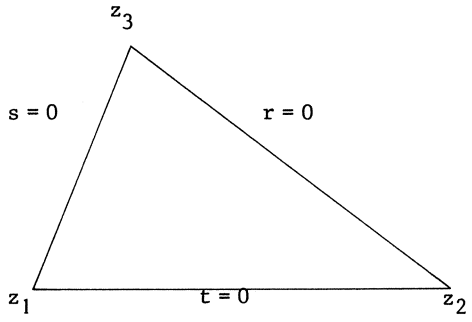
Kennelijk is $\phi_i(x,y)$ alleen $\neq 0$ op die driehoeken e_ℓ ($\ell = 1,\ldots,M$) waartoe z_i behoort. Hieruit volgt dat op iedere driehoek e_ℓ met hoekpunten z_1, z_2 en z_3 (we gebruiken lokale nummering) slechts *drie* basisfunkties $\neq 0$ zijn. Het is natuurlijk mogelijk om ϕ_1, ϕ_2 en ϕ_3 rechtstreeks in x en y uit te drukken. Het is evenwel handiger met *barycentrische* of *zwaartepuntskoördinaten* te werken. Hiertoe nemen we drie parameters r, s en t met r+s+t = 1. Door middel van de transformatie

$$\begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = \begin{pmatrix} x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} r \\ s \\ t \end{pmatrix} \tag{3.10.5}$$

wordt het gebied

$$T: \quad \begin{aligned} &0 \le r,s,t \le 1; \\ &1 = r + s + t, \end{aligned}$$

één-eenduidig in het driehoekje e_ℓ overgevoerd. We merken op dat de punten $w_1 = (1,0,0)^T$, $w_2 = (0,1,0)^T$ en $w_3 = (0,0,1)^T$ overeenkomen met de hoekpunten van e_ℓ en dat de vergelijkingen r = 0, s = 0 of t = 0 de driehoekszijden weergeven (zie figuur 3.7).



Figuur 3.7

Voor de meetkundige betekenis van zwaartepuntskoördinaten zie BELL [1969].

We moeten nu drie funkties vinden, lineair in r, s en t en aan te geven met $\Phi_i(r,s,t)$ die voldoen aan

$$\Phi_i(w_j) = \delta_{ij}, \qquad 1 \leq i, \ j \leq 3.$$

Deze funkties zijn:

$$\begin{aligned} &\Phi_1(r,s,t) = r; \\ &\Phi_2(r,s,t) = s; \\ &\Phi_3(r,s,t) = t. \end{aligned} \tag{3.10.6}$$

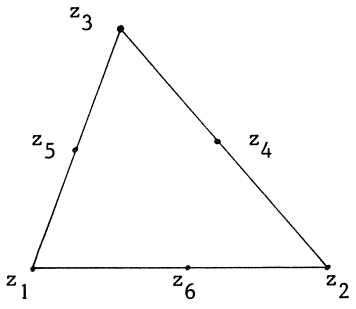
Omdat een lineaire transformatie polynomen in polynomen vervoert, geldt:

$$\begin{aligned} &\phi_1(x,y) = r; \\ &\phi_2(x,y) = s; \\ &\phi_3(x,y) = t. \end{aligned} \tag{3.10.6a}$$

k = 2

Een volledig polynoom van de tweede graad in twee variabelen heeft *zes* vrijheidsgraden, zodat op ieder driehoekige e_ℓ een element uit $\mathbb{P}_2(\pi)$ volledig bepaald is door de waarden die het in zes punten van e_ℓ aanneemt.

We kiezen daarvoor de hoekpunten en de middens van de driehoekszijden.



Figuur 3.8

De lezer ga zelf na dat de kontinuïteit van een element uit $\mathbb{P}_2(\pi)$ op de hoekpunten en de middens der driehoekszijden kontinuïteit op het gehele gebied impliceert.

Een element in $\mathbb{P}_2(\pi)$ is nu volledig bepaald door de waarden die het in de hoekpunten en op de middens der zijden aanneemt. We kiezen dus de N steunpunten z_i van de basisfunkties in de hoekpunten en op de middens der zijden. De basisfunkties $\phi_i(x,y)$ $(i = 1,\ldots,N)$ worden gedefinieerd als funkties die tweedegraadspolynomen op ieder driehoekje e_ℓ zijn en die voldoen aan de betrekkingen

$$\phi_i(z_j) = \delta_{ij}, \qquad i,j = 1,\ldots,N.$$

We bekijken nu een driehoekje e_ℓ met hoekpunten z_1, z_2 en z_3 en middens z_4, z_5 en z_6. Op dit driehoekje zijn alleen de basisfunkties ϕ_1, ϕ_2, ϕ_3, ϕ_4, ϕ_5 en $\phi_6 \not\equiv 0$. Om deze basisfunkties te bepalen maken we gebruik van zwaartepuntskoördinaten r, s en t. Geven we de hoekpunten en middens der zijden in het driehoekje

$$T: \quad \begin{array}{l} 0 \le r,s,t \le 1; \\ 1 = r+s+t, \end{array}$$

aan met $w_1 = (1,0,0)^T$, $w_2 = (0,1,0)^T$, $w_3 = (0,0,1)^T$, $w_4 = (0,\frac{1}{2},\frac{1}{2})^T$, $w_5 = (\frac{1}{2},0,\frac{1}{2})^T$ en $w_6 = (\frac{1}{2},\frac{1}{2},0)^T$, dan moeten we tweedegraadpolynomen Φ_i in r, s en t vinden, die voldoen aan:

$$\Phi_i(w_j) = \delta_{ij}, \qquad 1 \leq i,\ j \leq 6.$$

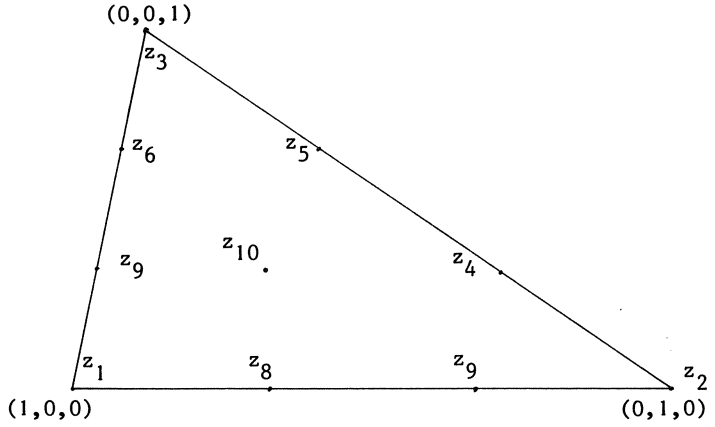
Deze polynomen zijn (zie MITCHELL [1973]):

$$\begin{aligned}
&\Phi_1(r,s,t) = r(2r-1) \ ;\\
&\Phi_2(r,s,t) = s(2s-1) \ ;\\
&\Phi_3(r,s,t) = t(2t-1) \ ;\\
&\Phi_4(r,s,t) = 4st \quad\ ;\\
&\Phi_5(r,s,t) = 4rt \quad\ ;\\
&\Phi_6(r,s,t) = 4st \quad\ ,
\end{aligned} \tag{3.10.7}$$

waarmee lokaal op e_ℓ de basis van $\mathbb{P}_2(\pi)$ is bepaald.

<u>k = 3</u>

We kiezen de steunpunten in de hoekpunten van de driehoekjes en verder twee op elke driehoekszijde, bij voorkeur zodanig, dat iedere driehoekszijde in drie gelijke stukken wordt verdeeld. In ieder driehoekje moet nog een inwendig steunpunt worden gekozen. We kunnen daarvoor het zwaartepunt kiezen.



<u>Figuur</u> 3.9

We kunnen de basisfunkties het beste in de zwaartepuntskoördinaten r, s en t uitdrukken. We moeten dan derdegraadspolynomen $\Phi_i(r,s,t)$ vinden met

$$\Phi_i(z_j) = \delta_{ij}, \qquad 1 \le i,\ j \le 10.$$

Deze polynomen zijn (zie MITCHELL [1973]):

$$\begin{aligned}
&\Phi_1(r,s,t) = \tfrac{1}{2}r(3r-1)(3r-2) \ ;\\
&\Phi_2(r,s,t) = \tfrac{1}{2}s(3r-1)(3s-2) \ ;\\
&\Phi_3(r,s,t) = \tfrac{1}{2}t(3t-1)(3t-2) \ ;\\
&\Phi_4(r,s,t) = \tfrac{9}{2}st(3s-1) \ ;\\
&\Phi_5(r,s,t) = \tfrac{9}{2}st(3t-1) \ ;\\
&\Phi_6(r,s,t) = \tfrac{9}{2}rt(3t-1) \ ;\\
&\Phi_7(r,s,t) = \tfrac{9}{2}rt(3r-1) \ ;\\
&\Phi_8(r,s,t) = \tfrac{9}{2}rs(3r-1) \ ;\\
&\Phi_9(r,s,t) = \tfrac{9}{2}rs(3s-1) \ ;\\
&\Phi_{10}(r,s,t) = 27rst \ ,
\end{aligned} \tag{3.10.9}$$

waarmee $\phi_i(x,y)$ lokaal bepaald is. Voor een verdere beschrijving zie MITCHELL [1973].

3.11. Hermite-interpolatie in twee dimensies

We bespreken in deze sektie een deelruimte van $H^1(\Omega)$ die iets specialer is dan de $\mathbb{P}_k(\pi)$ uit de vorige sektie.

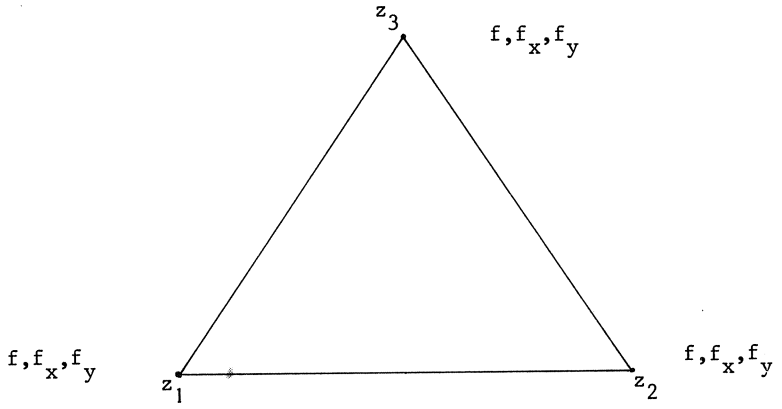
Laat $\pi = \{e_1,\ldots,e_M\}$ een reguliere partitie van Ω in driehoekjes zijn met maaswijdte h.

Definitie 3.11.1 Onder $\mathbb{H}_k(\pi)$ verstaan we de ruimte van funkties f die

(i) tot $\mathbb{P}_k(\pi)$ behoren;

(ii) in alle hoekpunten z_i van de driehoekjes kontinue partiële afgeleiden f_x en f_y hebben.

We gaan nu na hoe groot k moet zijn. Daartoe beschouwen we een driehoekje e_ℓ. We geven de hoekpunten aan met z_1, z_2 en z_3.



Figuur 3.10

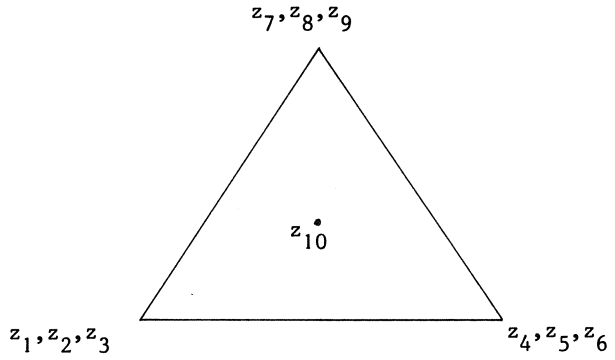
Ieder hoekpunt levert drie kondities op (kontinuïteit van f, f_x en f_y), zodat per driehoek aan *negen* kondities moet worden voldaan. Hieruit volgt dat $k \geq 3$.

We behandelen hier uitsluitend $k = 3$. Aangezien een derdegraadspolynoom in twee variabelen *tien* vrijheidsgraden heeft, moeten we op het driehoekje e_ℓ nog een *inwendig* steunpunt kiezen en nemen daarvoor het zwaartepunt z_c. De steunpunten van de basisfunkties van $\mathbb{H}_3(\pi)$ zijn nu de hoekpunten (drievoudig) en de zwaartepunten (enkelvoudig). Een element uit $\mathbb{H}_3(\pi)$ is nu volledig bepaald door (i) de funktiewaarden en partiële afgeleiden in de hoekpunten en (ii) de funktiewaarden in de zwaartepunten. Om de basisfunkties van $\mathbb{H}_3(\pi)$ te definiëren voeren we het begrip *knooppunt* in.

Definitie 3.11.2. Onder een knooppunt verstaan we het koppel (D^α, z_i), waarbij $D^\alpha = (\frac{\partial}{\partial x})^{\alpha_1} (\frac{\partial}{\partial y})^{\alpha_2}$ een differentiaaloperator van de orde $|\alpha| = \alpha_1 + \alpha_2$ is en z_i een steunpunt. D^α kan de identiteit, $\partial/\partial x$, $\partial/\partial y$ etc. zijn.

In ieder *hoekpunt* z_i van de partitie π zijn dus drie knooppunten gedefinieerd: (I, z_i), $(\partial/\partial x, z_i)$ en $(\partial/\partial y, z_i)$; in ieder *zwaartepunt* z_c is slechts het knooppunt (I, z_c) gedefinieerd.

We gaan nu de steunpunten opnieuw nummeren, waarbij de multipliciteit wordt meegerekend; de steunpunten op een driehoekje worden nu (lokaal) genummerd: $z_1, \ldots, z_{10}$.



Figuur 3.11

We gaan vervolgens de knooppunten nummeren: $(D_1,z_1),D_2,z_2),\ldots,(D_N,z_N)$, waarbij D_i een operator is die de identiteit $\partial/\partial x$ of $\partial/\partial y$ kan zijn. In figuur 3.11 wordt dus als volgt (lokaal) genummerd: $(D_1,z_1),(D_2,z_2),\ldots,$ (D_{10},z_{10}) met $D_1 = D_4 = D_7 = D_{10} = I$, $D_2 = D_5 = D_8 = \partial/\partial x$, $D_3 = D_6 = D_9 = \partial/\partial y$.

Na deze definitie van knooppunten kunnen we de basisfunkties $\phi_i(x,y)$ definiëren door de relaties

$$D_j\phi_i(z_j) = \delta_{ij}, \qquad 1 \le i,\ j \le N.$$

We gaan nu per driehoek een formule voor ϕ_i geven.

Op ieder driehoekje e_ℓ zijn slechts tien basisfunkties $\not\equiv 0$. We geven deze basisfunkties weer m.b.v. zwaartepuntskoördinaten r, s en t. Als (x_1,y_1), (x_2,y_2) en (x_3,y_3) de hoekpunten van e_ℓ zijn, dan definiëren we

$$x_{ij} = x_i - x_j,$$

$$y_{ij} = y_i - y_j, \qquad i,j = 1,2,3.$$

De formules van de basisfunkties zijn nu (zie FELIPPA [1970] en MITCHELL [1973]):

$$\begin{aligned}
&\Phi_1(r,s,t) = r^2(3-2r) - 7rst \qquad ;\\
&\Phi_2(r,s,t) = r^2(x_{21}s - x_{13}t) + (x_{13}-x_{21})rst;\\
&\Phi_3(r,s,t) = r^2(y_{21}s - y_{13}t) + (y_{13}-y_{21})rst;\\
&\Phi_4(r,s,t) = s^2(3-2s) - 7rst \qquad ;
\end{aligned} \tag{3.11.1}$$

$$\Phi_5(r,s,t) = s^2(x_{32}t-x_{21}r) + (x_{21}-x_{32})rst;$$

$$\Phi_6(r,s,t) = s^2(y_{32}t-y_{21}r) + (y_{21}-y_{32})rst;$$

$$\Phi_7(r,s,t) = t^2(3-2t) - 7rst \quad ;$$

$$\Phi_8(r,s,t) = t^2(x_{13}r-x_{32}s) + (x_{32}-x_{13})rst;$$

$$\Phi_9(r,s,t) = t^2(y_{13}r-y_{32}s) + (y_{32}-y_{13})rst;$$

$$\Phi_{10}(r,s,t) = 27rst.$$

3.12. Een variatieprobleem

Gegeven het probleem: Minimaliseer

$$I(v) = \iint_\Omega [p_1 v_x^2 + p_2 v_y^2 + qv^2 - 2fv]dxdy,$$

(3.12.1)

$$v = 0 \quad \text{op} \quad \partial\Omega,$$

waarbij $p_1, p_2 \in C^1(\bar{\Omega})$, $f, q \in C^0(\bar{\Omega})$, $p_1, p_2 \geq p_{min} > 0$, $q \geq 0$. De oplossing u van (3.12.1) ligt in

$$H_L = \{v \mid v \in H^1(\Omega);\ v = 0,\ (x,y)^T \in \partial\Omega\}.$$

In H_L zijn energieinprodukt en energienorm gedefinieerd door

$$(u,v)_L = \iint_\Omega [p_1 u_x v_x + p_2 u_y v_y + quv]dxdy;$$

(3.12.1)

$$\|u\|_L = \sqrt{(u,u)_L}.$$

Zij $\pi = \{e_1, \ldots, e_M\}$ een reguliere partitie van Ω. We definiëren $\mathbb{P}_k(\Omega)$ als in §3.10 met N steunpunten z_i en basisfunkties $\phi_i(x,y)$. We benaderen nu (3.12.1) door het probleem:

Minimaliseer

$$I(v_h), \quad v_h \in \mathbb{P}_k(\pi);$$

(3.12.3)

$$v_h = 0, \quad (x,y) \in \partial\Omega.$$

Aangezien $\partial\Omega$ stuksgewijs recht is, is het voldoende de randvoorwaarde te vervangen door

$$v_h(z_i) = 0, \quad z_i \in \partial\Omega.$$

We stellen nu $N = N_I + N_B$, waarbij N_B het aantal steunpunten is dat tot $\partial\Omega$ behoort en N_I het aantal *inwendige* steunpunten. We nummeren de inwendige steunpunten met $z_1,\ldots,z_{N_I}$ en de randsteunpunten met $z_{N_I+k},\ldots,z_N$. Schrijven we voor een element $v_h \in \mathbb{P}_k(\pi)$

$$v_h(x,y) = \sum_{i=1}^{N} q_i \phi_i(x,y),$$

dan gaat (3.12.3) over in:

Minimaliseer

$$F(q_1,\ldots,q_N) \equiv I(v_h);$$

$$q_i = 0, \quad i = N_I+1,\ldots,N.$$

De oplossing hiervan wordt gegeven door het stelsel

$$\frac{\partial F}{\partial q_i} = 0, \quad i = 1,\ldots,N_I;$$

(3.12.4)

$$q_i = 0, \quad i > N_I.$$

De oplossing u_h van (3.12.3) wordt nu gerepresenteerd door

$$u_h(x,y) = \sum_{i=1}^{N_I} a_i \phi_i(x,y),$$

waarbij $(a_1,\ldots,a_{N_I})$ de oplossing is van het stelsel

$$(3.12.5)\qquad \sum_{j=1}^{N_I} (\phi_i,\phi_j)_L\, a_j = \iint_\Omega f\,\phi_i\,dxdy, \qquad i = 1,\ldots,N_I.$$

We moeten dus de matrix $((\phi_i,\phi_j)_L)$ berekenen met

$$(3.12.6)\qquad \begin{aligned} (\phi_i,\phi_j)_L &= s_{ij} + m_{ij};\\ s_{ij} &= \iint_\Omega [p_1 \frac{\partial\phi_i}{\partial x}\frac{\partial\phi_j}{\partial x} + p_2 \frac{\partial\phi_i}{\partial y}\frac{\partial\phi_j}{\partial y}]dxdy;\\ m_{ij} &= \iint q\,\phi_i\,\phi_j\,dxdy, \quad 1 \le i, j \le N_I. \end{aligned}$$

Ook moeten we de vektor b berekenen met

$$b_i = \iint_\Omega f\,\phi_i\,dxdy, \qquad i = 1,\ldots,N_I.$$

We berekenen $A = ((\phi_i,\phi_j)_L)$ en b als volgt:

(i) We berekenen de $N \times N$ matrix

$$\begin{aligned} A^* &= \sum_{\ell=1}^{M} A^{(\ell)} = \sum_{\ell=1}^{M} (S^{(\ell)} + M^{(\ell)}) \quad ;\\ S^{(\ell)} &= (\iint_{e_\ell} [p_1 \frac{\partial\phi_i}{\partial x}\frac{\partial\phi_j}{\partial x} + p_2 \frac{\partial\phi_i}{\partial y}\frac{\partial\phi_j}{\partial y}]dxdy);\\ M^{(\ell)} &= (\iint_{e_\ell} q\,\phi_i\,\phi_j\,dxdy). \end{aligned}$$

(ii) We berekenen de vektor

$$\begin{aligned} b^* &= \sum_{\ell=1}^{M} b^{(\ell)} \quad ;\\ b^{(\ell)} &= (\iint_{e_\ell} f\,\phi_i\,dxdy). \end{aligned}$$

(iii) Vervolgens konstrueren we A uit A^* en b uit b^* door de komponenten met index groter dan N_I te schrappen:

$$A^* = \begin{bmatrix} \leftarrow N_I \rightarrow & \leftarrow N_B \rightarrow \\ A_I & A_{BI} \\ A_{BI}^T & A_B \end{bmatrix} \longrightarrow A_I \equiv A ;$$

$$b^* = \begin{bmatrix} b_I \\ \\ b_B \end{bmatrix} \begin{matrix} \uparrow \\ N_I \\ \downarrow \\ \uparrow \\ N_B \\ \downarrow \end{matrix} \longrightarrow b_I \equiv b .$$

In het geval van *inhomogene* koördinaten treedt een kleine modifikatie op:

$$A^* \rightarrow A_I \equiv A;$$

$$b^* \rightarrow b_I - A_{BI} b_B \equiv b.$$

$S^{(\ell)}$, $M^{(\ell)}$ en $b^{(\ell)}$ moeten veelal met behulp van numerieke kwadratuur worden geëvalueerd. Een voorbeeld daarvan is de Newton-Cotes-kwadratuur.

Zij

$$w_i^{(\ell)} = \iint_{e_\ell} \phi_i(x,y)dxdy,$$

dan kunnen we voor $S^{(\ell)}$, $M^{(\ell)}$ en $b^{(\ell)}$ de volgende benaderingen gebruiken:

$$s_{ij}^{(\ell)} \simeq \sum_m w_m^{(\ell)} \left[p_1 \frac{\partial \phi_i}{\partial x} \frac{\partial \phi_j}{\partial x} + p_2 \frac{\partial \phi_i}{\partial y} \frac{\partial \phi_j}{\partial y}\right](x_m,y_m);$$

$$(3.12.7) \qquad m_{ij}^{(\ell)} \simeq \sum_m w_m^{(\ell)} [q\ \phi_i\ \phi_j](x_m,y_m) = w_i^{(\ell)}\ q(x_i,y_i)\delta_{ij};$$

$$b_i^{(\ell)} \simeq \sum_m w_m^{(\ell)} [f\ \phi_i](x_m,y_m) = w_i^{(\ell)}\ f(x_i,y_i).$$

De gewichten $w_i^{(\ell)}$ worden gegeven door de formule

$$(3.12.8) \qquad w_i^{(\ell)} = 2\mu(e_\ell) \iiint_T \Phi_i(r,s,t)\ dr\ ds\ dt,$$

waarbij $\Phi_i(r,s,t)$ de barycentrische representatie van $\phi_i(s,y)$ is en $\mu(e_\ell)$ de oppervlakte van e_ℓ. Formule (3.12.8) kan verder uitgewerkt worden met behulp van de formule

$$(3.12.9) \qquad \iiint_T r^m s^n t^p\ dr\ ds\ t = \frac{m!n!p!}{(m+n+p+r)!} \quad .$$

Zie verder STRANG & FIX [1973] voor een analyse van het effekt van de kwadratuurfout en BELL [1969] voor de berekening van (3.12.9).

3.13. Isoparametrische elementen

In de vorige sekties was Ω steeds een polygonaal gebied, zodat Ω precies in driehoeken kon worden opgesplitst. Indien Ω een *kromme* rand heeft, kunnen we gebruik maken van driehoeken met kromme randen, de zgn. *isoparametrische elementen.*

Om dit begrip te kunnen definiëren voeren we eerst het begrip *unisolvente verzameling* in. We veronderstellen in deze paragraaf steeds dat $N_k = \frac{1}{2}(k+1)(k+2)$.

Definitie 3.13.1 Zij $v = \{z_1,\ldots,z_{N_k}\}$ een verzameling van N_k verschillende punten in $\mathbb{R}^2$. Dan heet v een *k-unisolvente verzameling*, als er bij iedere verzameling $\{\alpha_1,\ldots,\alpha_{N_k}\}$ van N_k reële getallen precies één polynoom $P_k(x,y)$ van de graad k in x en y bestaat, met

$$P_k(z_i) = \alpha_i, \quad i = 1,\ldots,N_k.$$

Voorbeelden van dergelijke verzamelingen zijn:

(i) $k = 1$; de hoekpunten van een driehoek;

(ii) $k = 2$; de hoekpunten van een driehoek en de middens der zijden;

(iii) $k = 3$; de hoekpunten van een driehoek, twee verschillende punten op iedere zijde en één inwendig punt.

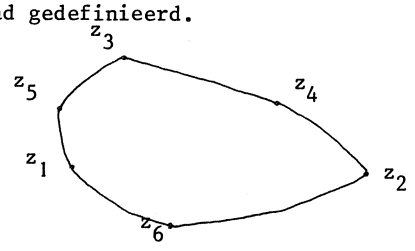
We gaan nu een *isoparametrisch element van de tweede graad* definiëren.

Definitie 3.13.2 Zij $v = \{z_1, z_2, \ldots, z_6\}$ een 2-unisolvente verzameling in $\mathbb{R}^2$. Laten de polynomen $\Phi_1(r,s,t), \ldots, \Phi_6(r,s,t)$ zijn gedefinieerd door (3.10.7). Dan heet de verzameling punten $(x,y)^T$ gegeven door

$$(3.13.1) \qquad \begin{aligned} (x,y)^T &= \sum_{i=1}^{6} z_i \, \Phi_i(r,s,t); \\ 0 &\leq r,s,t \leq 1 \quad ; \\ 1 &= r + s + t \quad , \end{aligned}$$

een isoparametrisch element van de tweede graad (zie figuur 3.12).

Op analoge manier worden isoparametrische elementen van de eerste en derde graad gedefinieerd.



Figuur 3.12

We zien dat iedere k-unisolvente puntenverzameling een isoparametrisch element van de k^e graad ondubbelzinnig definiëert. Het is verder evident dat iedere driehoek met hoekpunten z_1, z_2 en z_3 een isoparametrisch element van de eerste graad is, dat iedere driehoek met hoekpunten z_1, z_2 en z_3 en met zijdemiddens z_4, z_5 en z_6 een isoparametrisch element van de 2^e graad is, etc. We noemen deze elementen, die we reeds in de vorige paragraaf behandelden *rechte isoparametrische elementen.*

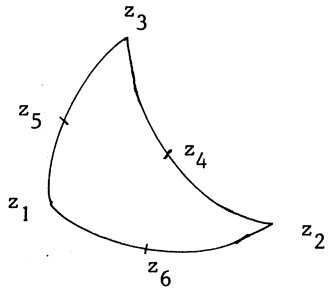
We geven nu een voorbeeld van een *krom* isoparametrisch element van de tweede graad. We nemen drie punten z_1, z_2 en z_3 die tezamen een driehoek opspannen. Vervolgens kiezen we z_4, z_5 en z_6 zo, dat

(i) z_4, z_5 en z_6 *niet alle* op de driehoekszijden liggen;

(ii) $\|z_4-\frac{1}{2}(z_2+z_3)\| << \|z_2-z_3\|$;

$\|z_5-\frac{1}{2}(z_1+z_3)\| << \|z_1-z_3\|$;

$\|z_6-\frac{1}{2}(z_1+z_2)\| << \|z_1-z_2\|$.



Figuur 3.13

Het kan bewezen worden dat in dat geval $\{z_1,\ldots,z_6\}$ een 2-unisolvente verzameling is. Voor meer voorbeelden van kromme isoparametrische elementen, zie CIARLET & RAVIART [1972a].

Benadering van Ω d.m.v. isoparametrische elementen

Zij Ω een konvex gebied met een (stuksgewijs) kromme rand in $\mathbb{R}^2$. Indien we probleem (3.12.1) door middel van de eindige elementenmethode willen oplossen, kunnen we dit doen door Ω door een polygonaal gebied Ω_h te benaderen en vervolgens het probleem (3.12.1) te benaderen door:

Minimaliseer

$$I_h(v) = \iint_{\Omega_h} [p_1(\frac{\partial v}{\partial x})^2 + p_2(\frac{\partial v}{\partial y})^2 + qv^2 - 2fv]dxdy;$$

(3.13.2)

$$v = 0 \quad \text{op } \partial\Omega_h.$$

Vervolgens benaderen we (3.13.2) door:

Minimaliseer

$$I_h(v_h), \quad v_h \in \mathbb{P}_k(\pi);$$

(3.13.3)

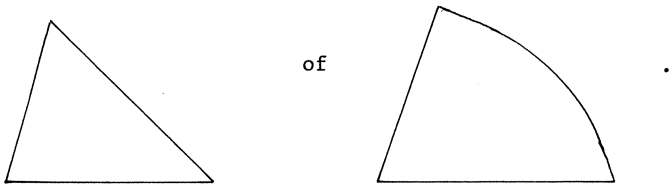
$$v_h = 0 \quad \text{op } \partial\Omega_h.$$

Deze aanpak evenwel werkt alleen goed voor k = 1, omdat voor k > 1 de orde van nauwkeurigheid van de benadering van (3.12.1) te sterk wordt beïnvloed door de onnauwkeurigheid waarmee Ω wordt benaderd door Ω_h. We moeten in dat geval Ω benaderen door een gebied Ω_h dat is opgebouwd uit isoparametrische elementen van de graad k. We illustreren dit voor k = 2.

We voeren de benadering in twee stappen uit:

(i) We splitsen Ω op in M "driehoekjes" $\tilde{e}_\ell$, waarbij elk driehoekje $\tilde{e}_\ell$ twee of drie rechte zijden heeft.

Aangezien $\tilde{e}_\ell$ zelf eveneens konvex is, ziet $\tilde{e}_\ell$ er altijd zo uit:



of .

We nemen aan dat de partitie regulier is, i.e. dat er een c > 0, onafhankelijk van de partitie, bestaat met

$$\tilde{h}_\ell = \text{diameter } (\tilde{e}_\ell) \leq h;$$

$$\tilde{\rho}_\ell = \sup \text{ (straal ingeschr. cirkels in } \tilde{e}_\ell) \geq Ch_\ell.$$

(ii) We konstrueren nu Ω_h door iedere $\tilde{e}_\ell$ te benaderen door een isoparametrisch element e_ℓ van de tweede graad en te definiëren

$$\Omega_h = \bigcup_{\ell=1}^{M} e_\ell.$$

Deze approximatie gaat als volgt: We kiezen in iedere $\tilde{e}_\ell$ de hoekpunten en de middens der zijden als een 2-unisolvente verzameling en passen formule (3.13.3) toe.

Het is evident dat driehoeken met drie rechte zijden invariant blijven, terwijl driehoeken met één bolle rand konvex blijven. Evenwel hoeft Ω_h niet meer konvex te zijn voor $k > 1$.

Het variatieprobleem

We keren terug naar (3.12.1) en proberen dit minimaliseringsprobleem aan te pakken door de ruimte $H^1(\Omega_h)$ op een speciale manier te benaderen. We nemen aan dat Ω_h is opgebouwd uit isoparametrische elementen van de tweede graad en dat de partitie regulier is. We definiëren de ruimte $V_h \subset H^2(\Omega_h)$ door de basisfunkties *op ieder element* e_ℓ te definiëren. Zij gegeven het element e_ℓ. Laten z_1, z_2 en z_3 de hoekpunten zijn en z_3, z_4 en z_5 de middens der zijden (lokale nummering). Dan wordt *op* e_ℓ $\phi_i(x,y)$ als volgt gedefinieerd:

$$
\begin{aligned}
&\phi_i(x,y) = \Phi_i(r,s,t), && i = 1,\ldots,6;\\
&(x,y)^T = \sum_{i=1}^{6} z_i\, \Phi_i(r,s,t) && ;\\
&1 = r + s + t && ;\\
&0 \le r,s,t \le 1 && ,
\end{aligned}
\qquad (3.13.5)
$$

waarbij $\Phi_i(r,s,t)$ is gegeven door (3.10.7). Kennelijk voldoen de basisfunkties eveneens aan de betrekking

$$\phi_i(z_j) = \delta_{ij}.$$

Door bovenstaande *elementsgewijze* definitie van de basisfunkties is de ruimte V_h precies vastgelegd. Aangezien de basisfunkties kontinu zijn op de randen van de "driehoeken" (de gemeenschappelijke randen van de driehoeken zijn altijd recht, waardoor de basisfunkties cih op die randen als tweedegraadspolynomen gedragen), is V_h een deelruimte van $H^1(\Omega_h)$. We benaderen nu (3.13.4) door:

Minimaliseer

$$
\begin{aligned}
&I_h(v_h), \quad v_h \in V_h;\\
&v_h = 0 \quad \text{op} \quad \partial\Omega_h.
\end{aligned}
$$

De aanpak van (3.12.1) gaat nu verder als in §3.12. Voor een analyse van het effekt van de vervanging van Ω door Ω_h, zie CIARLET & RAVIART [1972a].

3.14. Foutschatting

We gaan in deze paragraaf een schatting afleiden van de fout die optreedt als we een $2m^e$ orde zelfgeadjungeerd Dirichlet-randwaardeprobleem oplossen door middel van de eindige elementenmethode.

Zij $\Omega \subset \mathbb{R}^n$ een begrensd konvex gebied met een stuksgewijs rechte rand $\partial\Omega$.

Zij $L: H^{2m}(\Omega) \to H^0(\Omega)$ een lineaire operator gedefinieerd door

$$(3.14.1) \qquad Lu = \sum_{|\alpha|\leq m} (-1)^{|\alpha|} D^\alpha (p_\alpha(x) D^\alpha u),$$

waarbij $\alpha = (\alpha_1,\ldots,\alpha_n)$ een multiindex is met $|\alpha| = \alpha_1+\ldots+\alpha_n$, $|\alpha| \leq m$ en waarbij D^α een differentiaaloperator van de orde $|\alpha|$ is.

We definiëren de ruimte $H_0^m(\Omega)$ als volgt:

$$H_0^m(\Omega) = \{v \mid v \in H^m(\Omega),\ v = \frac{\partial v}{\partial n} = \ldots = \frac{\partial^m v}{\partial n^m} = 0 \text{ op } \partial\Omega\}.$$

L is dan zelfgeadjungeerd op $H_0^m(\Omega) \cap H^{2m}(\Omega)$:

$$(Lu,v) = \int_\Omega Luv\, dx_1\ldots dx_n =$$

$$= \sum_{|\alpha|\leq m} \int_\Omega p_\alpha(x)\, D^\alpha u D^\alpha v\, dx_1\ldots dx_n = (Lv,u),$$

$$u,v \in H_0^m(\Omega) \cap H^{2m}(\Omega).$$

We definiëren nu de bilineaire symmetrische operator $B: H_0^m(\Omega) \times H_0^m(\Omega) \to \mathbb{R}$ als volgt:

$$(3.14.2) \qquad B(u,v) = (Lu,v) = \sum_{|\alpha|\leq m} \int_\Omega p_\alpha(x) D^\alpha u D^\alpha v d\Omega.$$

We veronderstellen dat B sterk koërcief is:

$$(3.14.3) \qquad \begin{aligned} &|B(u,v)| \leq C\|u\|_m \|v\|_m, \quad u,v \in H_0^m(\Omega); \\ &C_1\|u\|_m^2 \leq B(u,u) \leq C_2\|u\|_m^2. \end{aligned}$$

We weten dat in dat geval het randwaardeprobleem

$$\begin{aligned} &Lu = f, \quad f \in H^0(\Omega); \\ &\qquad\qquad u \in H_0^m(\Omega) \end{aligned} \tag{3.14.4}$$

precies één oplossing u heeft. Deze u heeft de eigenschap dat hij de funktionaal

$$I(v) = B(v,v) - 2(f,v) \tag{3.14.5}$$

minimaliseert over $H_0^m(\Omega)$ en aan de zwakke Galerkin-vorm

$$B(u,v) = (f,v), \qquad v \in H_0^m(\Omega) \tag{3.14.6}$$

voldoet. We zoeken nu een approximatie van u door I(v) over een geschikte deelruimte van $H_0^m(\Omega)$ te minimaliseren.

Zij $\pi = \{e_\ell\}_{\ell=1}^M$ een reguliere partitie van Ω in simplexen. Bij een vast natuurlijk getal $k \geq 2m-1$ definiëren we V_h als de ruimte van funkties die

(i) tot $H_0^m(\Omega)$ behoren;

(ii) op ieder simplex e_ℓ een polynoom van de graad kleiner dan $k+1$ zijn.

Volgens CIARLET & RAVIART [1972b] bestaat er dan een konstante C onafhankelijk van $h = \max_\ell (\text{diameter}(e_\ell))$ zodat voor iedere $v \in H^q(\Omega) \cap H_0^m(\Omega)$ de volgende ongelijkheid geldt:

$$\inf_{v_h \in V_h} \|v-v_h\|_s \leq Ch^{d-s}\|v\|_d, \qquad \begin{aligned} &s = 0,\ldots,m; \\ &d = \min(k+1,q),\ q \geq m. \end{aligned} \tag{3.14.7}$$

We benaderen nu de oplossing van (3.14.4) c.q. (3.14.6) door I(v) over V_h te minimaliseren. De unieke funktie $u_h \in V_h$ die I(v) minimaliseert over V_h voldoet aan de volgende eigenschappen (zie STRANG & FIX [1973]):

$$\begin{aligned} &B(u_h,v_h) = (f,v_h), \\ &B(u-u_h,v_h) = 0 \quad , \\ &B(u-u_h,u-u_h) \leq B(u-v_h,u-v_h), \qquad v_h \in V_h. \end{aligned} \tag{3.14.8}$$

Om nu een schatting te maken van de fout $\|u-u_h\|_s$ $(s = 0,\ldots,m)$, introduceren we het volgende *hulpprobleem* (Nitsche's truuk):

Zij $w \in H^{2m-s}(\Omega) \cap H_0^m(\Omega)$ de oplossing van de zwakke Galerkin-vorm

$$B(w,v) = (g,v), \qquad v \in H_0^m(\Omega), \tag{3.14.9}$$

waarbij $g \in H^{-s}(\Omega)$ een begrensde lineaire funktionaal op $H^s(\Omega)$ is $(s = 0,\ldots,m)$. Nemen we $v = e_h = u-u_h$, dan krijgen we na toepassing van (3.14.8) wegens de symmetrie van B

$$B(w-v_h,e_h) = (g,e_h), \qquad v \in V_h. \tag{3.14.10}$$

Wegens (3.14.3) krijgen we dan

$$|(g,e_h)| \leq C\|w-v_h\|_m \|e_h\|_m, \qquad v_h \in V_h. \tag{3.14.11}$$

We moeten dus bovengrenzen afleiden van $\|e_h\|_m$ en $\inf_{v_h}\|w-v_h\|_m$.

Uit (3.14.7) volgt dat

$$\inf_{v_h \in V_h} \|w-v_h\|_m \leq Ch^{d-m}\|w\|_d;$$

$$d = \min(k+1,2m-s).$$

Aangezien $k+1 \geq 2m$, volgt daaruit dat $d = 2m-s$, waaruit volgt dat

$$\inf\|w-v_h\|_m \leq Ch^{m-s}\|w\|_{2m-s}. \tag{3.14.12}$$

Voor $\|e_h\|_m$ geldt wegens (3.14.3) en (3.14.8), aangezien $u \in H^{k+1}(\Omega) \cap H_0^m(\Omega)$,

$$\|e_h\|_m^2 \leq C\ B(e_h,e_h) \leq C\ B(u-v_h,u-v_h) \leq C'\|u-v_h\|_m^2,$$

waaruit volgt wegens (3.14.8) dat

$$\|e_h\|_m \leq Ch^{d-m}\|u\|_d\ ,$$

$$d = \min(k+1,k+1) = k+1. \tag{3.14.13}$$

Uit (3.14.12) en (3.14.13) volgt dat

$$(3.14.14) \quad |(g,e_h)| \leq Ch^{k+1-s}\|u\|_{k+1}\ \|w\|_{2m-s}.$$

Wegens de stabiliteit van L geldt dat

$$(3.14.15) \quad \|w\|_{2m-s} \leq C\|g\|_{-s},$$

waaruit volgt dat

$$|(g,e_h)| \leq Ch^{k+1-s}\ \|u\|_{k+1}\ \|g\|_{-s}.$$

Wegens de formule

$$\|v\|_s = \sup_{g\in H^{-s}(\Omega)} |(g,v)|/\|g\|^{-s}$$

geldt

$$(3.14.16) \quad \|e_h\|_s \leq Ch^{k+1-s}\|u\|_{k+1}.$$

LITERATUUR

BAKKER, M., *The Galerkin's method for the solution of certain nonlinear two-point boundary value problems* (verschijnt in [1976]).

BELL, K., *A refined triangular plate bending finite element*, Int. Journ for Num. Meth. in Engineering 1, 101-122 (1969).

CIARLET, P.G. & P.A. RAVIART, *The combined effect of curved boundaries and numerical integration in isoparametric finite element methods*. Uit: A.K. Aziz, editor, The mathematical foundations of the finite element method with application to partial differential equations, Academic Press, New York, London, 1972.

CIARLET, P.G. & P.A. RAVIART, *General Lagrange and Hermite interpolation in $\mathbb{R}^n$ with applications to finite element methods*, Arch. Rat. Mech. An. 46, 177 ff. (1972).

FELIPPA, C. & R.W. CLOUGH, *The finite element methods in solid mechanics.* Uit: G. Birkhoff & R.S. Varga, editors, Numerical Solution of Field Problems in Continuum Physics, A.M.S., Providence, Rhode Island, 1970.

MITCHELL, A.R., *Introduction to mathematics of finite elements.* Uit: J.R. Whiteman, editor, The mathematics of finite elements and applications, Academic Press, London, New York, 1973.

NITSCHE, J.A., *Lineare Splinefunktionen und die Methode von Ritz für elliptische Randwertaufgaben*, Arch. Rat. Mech. An. 46, 348 ff. (1972).

STRANG, G. & G. FIX, *An Analysis of the Finite Element Method*, Prentice-Hall, Englewood Cliffs, New Jersy, 1973.

ZLAMÁL, M., *A finite element procedure of the second order of accuracy*, Num. Math. 14, 394-402 (1970).

4. GEWOGEN RESIDUEN METHODEN

4.1. Inleiding

In hoofdstuk 2 behandelden we randwaardeproblemen waarvan de oplossing beschouwd kan worden als het minimaliserend element van een convexe functionaal. Dit impliceert dat we ons beperkten tot symmetrische, positief definiete operatoren. In dit hoofdstuk zullen we zien dat de in hoofdstuk 2 behandelde Ritz-methode een bijzonder geval is van een uitgebreidere klasse van methoden: de gewogen residuen methoden. Tot deze klasse zullen nog een aantal bekende discretiseringsmethoden blijken te behoren zoals de collocatie- en de Galerkin-methode.

Voor een gewogen residuen methode wordt uitgegaan van de differentiaalvergelijking

(4.1.1) $\quad Lu = f$

in een gegeneraliseerde vorm. Dit kan op verschillende manieren gebeuren en hiervoor zijn verschillende varianten mogelijk. Deze hebben echter alle gemeen dat de vergelijking vermenigvuldigd wordt met een "functie" v en geïntegreerd wordt over het gebied Ω. Formeel kunnen we dit als volgt beschrijven. Zij D een dichte deelverzameling van een separabele Banachruimte H en zij L een operator $L: D \to H$, dan heet u een "gegeneraliseerde" oplossing van (4.1.1) als elke lineaire operator $\ell: H \to \mathbb{R}$ het residu Lu-f nul maakt. We kunnen dit schrijven als

(4.1.2a) $\quad \ell(Lu-f) = 0 \quad , \quad$ voor alle $\ell \in H^D$,

of als

(4.1.2b) $\quad \langle Lu-f, \ell \rangle = 0 \quad , \quad$ voor alle $\ell \in H^D$.

Hierin is H^D een deelverzameling van de duale ruimte van H.

Voorbeeld 4.1.1

Zijn de elementen van H begrensde functies, dan kunnen we het "inprodukt nemen" met alle delta-functies op Ω:

$$(4.1.3a) \quad \langle Lu, \delta_{x_0} \rangle = \langle f, \delta_{x_0} \rangle \quad , \text{ voor alle } x_0 \in \Omega,$$

of

$$(4.1.3b) \quad \int_\Omega (Lu(x)-f(x))\delta(x-x_0)\, d\Omega = 0, \quad \text{voor alle } x_0 \in \Omega,$$

of

$$(4.1.3c) \quad Lu(x_0) = f(x_0) \quad , \text{ voor alle } x_0 \in \Omega.$$

Een "gegeneraliseerde" oplossing komt in dit geval overeen met de klassieke oplossing.

Voorbeeld 4.1.2

Als H een Hilbert-ruimte is, kunnen we H^D identificeren met H en we kunnen schrijven

$$(4.1.4a) \quad (Lu-f,v)_H = 0 \quad , \text{ voor alle } v \in H,$$

ofwel

$$(4.1.4b) \quad \int_\Omega Luv\, d\Omega = \int_\Omega fv\, d\Omega \quad , \text{ voor alle } v \in H.$$

Dit is de *zwakke vorm* van de differentiaalvergelijking.

De discretisering van een continu probleem

De discretisering van het probleem Lu = f wordt nu als volgt tot stand gebracht:

(i) Een eindigdimensionale ruimte $S_h \subset H$ wordt gekozen, opgespannen door de basiselementen $\{\phi_i\}_{i=1}^N$. De oplossing $u \in D$ wordt benaderd door een element $u_h \in S_h$. Dit wil zeggen dat de benadering u_h wordt beschreven als

$$(4.1.5) \quad u_h = \sum_{i=1}^{N} a_i \phi_i.$$

S_h heet de *trial space*.

(ii) Een eindigdimensionale ruimte $V_h \subset H^D$ wordt gekozen, opgespannen door de *testfuncties* $\{\psi_i\}_{i=1}^N$. V_h heet de *test space*.

(iii) De benadering u_h wordt bepaald door het discreet analogon van (4.1.3). Hierdoor ontstaat een stelsel van N algebraïsche (of transcendente) vergelijkingen met N onbekenden

$$(4.1.6) \qquad \langle L(\sum_{i=1}^{N} a_i\phi_i),\psi_j\rangle = \langle f,\psi_j\rangle \quad , \qquad j = 1,2,\ldots,N.$$

In het resterende gedeelte van dit hoofdstuk zullen we aannemen dat de operator L lineair is. Hierdoor kan het stelsel worden geschreven als

$$(4.1.7) \qquad \sum_{i=1}^{N} a_i \langle L\phi_i,\psi_j\rangle = \langle f,\psi_j\rangle \quad , \qquad j = 1,2,\ldots,N.$$

4.2. De collocatie methode

De collocatie methode ontstaat door discretisering van de gegeneraliseerde vergelijking zoals deze werd gegeven in voorbeeld 4.1.1. De functies $\{\phi_i\}_{i=1}^N$ worden gekozen uit het definitiegebied D van de operator L. Als functies $\{\psi_j\}_{j=1}^N$ worden N verschillende delta-functies δ_{x_0}, $x_0 \in \Omega$, gekozen. Het stelsel vergelijkingen luidt dan

$$(4.2.1) \qquad \sum_{i=1}^{N} a_i L\phi_i(x_j) = f(x_j) \quad , \qquad j = 1,2,\ldots,N.$$

In welke mate de oplossing van het discrete probleem (4.2.1) de oplossing van het continue probleem (4.1.1) benadert, hangt geheel af van de keuze van $\{\phi_i\}$ en $\{x_j\}$. In de eerste plaats is het direkt duidelijk dat de matrix $(L\phi_i(x_j))$ niet singulier mag zijn.

De eindige-elementen-techniek kan, in het geval van een collocatie methode, worden toegepast in die zin dat voor $\{\phi_i\}$ stuksgewijze polynomen gekozen kunnen worden met een zo klein mogelijke drager. Hoewel het niet strikt noodzakelijk is, zal men in het algemeen het definitiegebied Ω in een aantal elementen verdelen en op elk element een gelijke wijze van discretiseren kiezen.

De collocatie methode lijkt niet bijzonder geschikt voor problemen in meer dimensies (n > 1) vanwege de strenge continuiteits-eisen die gesteld worden aan de functies ϕ_i. Voor tweepunts randwaardeproblemen is de methode

interessanter. Enkele belangrijke foutschattingen worden gegeven door RUSSEL & SHAMPINE [1972] en DE BOOR & SWARTZ [1973].

Wij zullen hier niet verder ingaan op de collocatie methoden, maar toch is de volgende stelling zeker het vermelden waard.

Stelling 4.2.1

Zij L een (lineaire) m-de orde differentiaaloperator, zodanig dat Lu = f eenduidig oplosbaar is. Laat $S_h \subset C^{m-1}[a,b]$ bestaan uit stuksgewijs (k-1)-de graads polynomen op een partitie π van het definitiegebied [a,b]

$$(4.2.2) \qquad \pi : \{a = x_0 < x_1 < x_2 < \ldots < x_N = b\}.$$

Zij h gedefinieerd door

$$(4.2.3) \qquad h = \max_{i=1,\ldots,N} |x_i - x_{i-1}|.$$

(i) Dan zijn k-m steunpunten noodzakelijk op elk interval, en

(ii) er bestaat de foutschatting

$$\|D^j(u-u_h)\|_{0,\infty} \leq K\, h^{k-m}\, \|D^k u\|_{0,\infty}, \qquad j = 0,1,\ldots,m.$$

(iii) Als op $[x_{i-1},x_i]$ de nulpunten van het (k-m)-de verschoven Legendre-polynoom gekozen worden als steunpunten voor de collocatie, dan geldt voor de steunpunten $x_i \in \pi$ zelfs

$$(4.2.5) \qquad |D^j(u-u_h)(x_i)| = O(h^{2(k-m)}) \quad , \quad j = 0,1,\ldots,m-1.$$

Bewijs Zie DE BOOR & SCHWARTZ [1973].

4.3. De Galerkin-methode

De Galerkin-methode ontstaat door het discretiseren van de gegeneraliseerde vergelijking zoals die gegeven is in voorbeeld 4.1.2. Hier neemt men $S_h = V_h \subset H$ een eindigdimensionale deelruimte opgespannen door $\{\phi_i\}_{i=1}^N$. De discrete vergelijking luidt

$$(4.3.1) \qquad \sum_{i=1}^{N} a_i\, (L\phi_i,\phi_j) = (f,\phi_j) \quad , \quad j = 1,\ldots,N.$$

We zien, dat de Galerkin-methode samenvalt met de Ritz-methode, wanneer L een symmetrische, positief definiete operator is. De eindige elementen technieken zijn weer toepasbaar en leiden ook hier tot het gebruik van stuksgewijs polynomen ϕ_i met een kleine drager. De inproduct-vorm $(L\phi_i,\phi_j)$ maakt het mogelijk om, door partieel integreren, de differentiaal operator L over ϕ_i en ϕ_j te "verdelen". Hierdoor kan men, bij een operator van de orde 2m, basisfuncties $\phi_i \in H$ kiezen welke niet behoren tot $D = C^{2m}(\Omega)$. Het is voldoende, als de basisfuncties ϕ_i behoren tot $H^m(\Omega)$.

Op deze wijze kunnen, via de Galerkin-methode, de technieken van de klassieke eindige elementen methode, in principe ook worden toegepast op niet-symmetrische en niet-positief-definiete problemen. Convergentie en foutschattingen leveren echter meer moeilijkheden en hiervoor zullen aan de operator L en aan de discretisering toch bepaalde voorwaarden moeten worden opgelegd.

4.4. Een verband tussen de Galerkin- en de collocatie methode

In tegenstelling tot de collocatie methode, moet bij de Galerkin-methode een integraal uitgerekend worden voor de berekening van een element van de discrete operator $(L\phi_i,\phi_j)$. Deze integraal zal echter in de praktijk altijd benaderd worden met behulp van een integratie-regel. Wanneer we de differentiaal-operator L niet "verdelen" over ϕ_i en ϕ_j, wordt een matrix-element benaderd door

$$(4.4.1) \qquad (L\phi_i,\phi_j) \approx \sum_{k=1}^{M} w_k \, L\phi_i(x_k)\phi_j(x_k).$$

Passen we dezelfde integratieregel toe op het rechterlid, dan luidt de discrete vergelijking

$$(4.4.2) \qquad \sum_{i=1}^{N} a_i \sum_{k=1}^{M} w_k \, L\phi_i(x_k)\phi_j(x_k) = \sum_{k=1}^{M} w_k \, f(x_k)\phi_j(x_k).$$

Wanneer de matrix $(w_k\phi_j(x_k))$ vierkant is en niet-singulier, dan is het stelsel (4.4.2) equivalent met

$$(4.4.3) \qquad \sum_{i=1}^{N} a_i \, L\phi_i(x_k) = f(x_k) , \qquad k = 1,\ldots,N.$$

Dit is precies de discrete operator voor een collocatie-methode.

Opmerking 4.4.1.

Om de transformatie van een Galerkin- naar een collocatie-methode mogelijk te maken, kunnen we de operator L (van orde 2m) *niet* "verdelen" over ϕ_i en ϕ_j. Het is daarom noodzakelijk dat $\phi_i \in H^{2m}[a,b]$. Volgens Sobolov's lemma moet derhalve $\phi_i \in C^{2m-1}$.

Ter illustratie passen we de transformatie toe op het tweepunts randwaardeprobleem uit paragraaf 4.2. De integratie (4.4.1) wordt, zoals gebruikelijk bij een eindige elementen methode, elements-gewijs uitgevoerd. Willen we op elk element de nauwkeurigste kwadratuur-formule met steunpunten toepassen, dan leidt dit tot de Gauss-Legendre kwadratuur-formules welke een fout O(h) bezitten. Worden voor de functies ϕ_i (k-1)-de graads stuksgewijs polynomen gekozen, dan moet (opdat de matrix $w_k\phi_i(x_k)$ vierkant en niet-singulier is) gelden K = k-m.

De matrix $(L\phi_i,\phi_j)$ en het rechterlid (f,ϕ_j) kunnen, wanneer Gauss-Legendre-kwadratuur gebruikt wordt, berekend worden met een fout $O(h^{2(k-m)})$. Dit geeft ons een aanwijzing waarom de collocatie-methode nauwkeurig is wanneer juist de nulpunten van een Legendre-polynoom gekozen worden als steunpunten voor de collocatie.

4.5. Een globale foutschatting voor een Galerkin-methode

We beschouwen hier een tweepunts randwaardeprobleem op het interval [a,b]

$$(4.5.1) \qquad Lu \equiv u'' + pu' + qu = f,$$

met Dirichlet-randvoorwaarden. We nemen aan dat p, q en f voldoende glad zijn en begrensd

$$|p| \le P, \quad |q| \le Q \quad \text{op } [a,b].$$

Voor dit probleem geven we in deze paragraaf een foutschatting in een globale norm. Het bewijs laat zich zonder veel moeite uitbreiden voor het geval van een meerdimensionaal elliptisch probleem.

Stelling 4.5.1 (DOUGLAS & DUPONT [1974]).

Laat $u_h \in V_h \subset H^1[a,b]$ de Galerkin-oplossing zijn van Lu = f. Als V_h alle

(stuksgewijs) polynomen van de graad kleiner dan k bevat, dan geldt

$$(4.5.2) \qquad \|u-u_h\|_0 + h\|u-u_h\|_1 \leq K\ |u|_k\ h^k.$$

Bewijs

Uit stelling 2.10.1 volgt dat, voor een voldoende gladde functie w,

$$(4.5.3) \qquad \inf_{v_h \in V_h} \|w-v_h\|_0 + h\|w-v_h\|_1 \leq K\ |w|_k\ h^k.$$

We zullen nu achtereenvolgens aantonen dat

$$(4.5.4) \qquad \|u-u_h\|_0 \leq K_1 \|u-u_h\|_1\ h$$

en dat

$$(4.5.5) \qquad \|u-u_h\|_1 \leq K_2\ |u|_k\ h^{k-1}.$$

Uit deze twee ongelijkheden volgt (4.5.2) direkt, wanneer gebruik gemaakt wordt van (4.5.3).

Is u de oplossing van Lu = f, dan geldt

$$(4.5.6) \qquad B(u,v) = (f,v) \quad , \qquad \text{voor alle } v \in H_0^1[a,b].$$

Hierin is

$$(4.5.7) \qquad B(u,v) = \int_a^b - u'v' + pu'v + quv\ dx$$

een begrensde lineaire operator $H^1[a,b] \times H_0^1[a,b] \to \mathbb{R}$.

De Galerkin-benadering $u_h \in V_h$ wordt zodanig bepaald dat

$$(4.5.8) \qquad B(u_h,v_h) = (f,v_h) \quad , \qquad \text{voor alle } v_h \in V_h \cap H_0^1[a,b].$$

Voor het bewijs van de stelling introduceren we een functie z, die de oplossing is van het probleem

$$L^T z = u_h - u \qquad \text{op } [a,b]$$

met homogene Dirichlet-randvoorwaarden. Hier is L^T de geadjungeerde operator van L, zodat voor alle $w \in H^1[a,b]$ geldt

$$B(w,z) = (w,u_h-u).$$

In het bijzonder geldt

$$\|u_h-u\|_0^2 = B(u_h-u,z).$$

Uit (4.5.6) en (4.5.8) volgt dat $B(u-u_h,v_h) = 0$ voor alle $v_h \in V_h \cap H_0^1[a,b] = V_h$, zodat

$$\|u_h-u\|_0^2 = B(u_h-u,z-v_h) \leq K_0\|u_h-u\|_1 \|z-v_h\|_1 \text{ voor alle } v_h \in V_h.$$

Derhalve geldt

$$\|u_h-u\|_0^2 \leq K_0\|u_h-u\|_1 \inf_{v_h\in V_h} \|z-v_h\|_1 .$$

Omdat $u-u_h \in H^0[a,b]$ is $z \in H^2[a,b]$ en bestaat er een K_R, zodat

$$\|z\|_2 \leq K_R\|u_h-u\|_0.$$

Met behulp van (4.5.3) krijgen we

$$\inf_{v_h\in V_h} \|z-v_h\|_1 \leq K_I\|z\|_2 \; h \leq K_I \; K_R\|u_h-u\|_0 \; h ,$$

zodat

$$\|u_h-u\|_0 \leq K_0 \|u_h-u\|_1 \; K_I \; K_R \; h = K\|u_h-u\|_1 \; h. \tag{4.5.9}$$

Hiermee is voldaan aan de ongelijkheid (4.5.4).

Voor elke $v \in H^1[a,b]$ geldt

$$|v|_1^2 \leq (P+Q) \|v\|_0 \|v\|_1 + |B(u,v)|$$

en in het bijzonder geldt

$$|u-u_h|_1^2 \leq (P+Q)\,\|u_h-u\|_0\,\|u_h-u\|_1 + \inf_{v_h\in V_h} |B(u-u_h,u-v_h)|.$$

Samen met (4.5.9) levert dit

$$\|u-u_h\|_1^2 \leq \|u_h-u\|_1\,\{(P+Q+Kh)\|u_h-u\|_0 + K_0 \inf_{v_h\in V_h} \|u-v_h\|_1\},$$

zodat

$$\|u-u_h\|_1 \leq Kh(P+Q+Kh)\|u-u_h\|_1 + K_0 \inf_{v_h\in V_h} \|u-v_h\|_1.$$

Hieruit volgt dat voor voldoende kleine h, namelijk als

$$Kh(P+Q+Kh) < 1,$$

geldt dat

$$\|u-u_h\|_1 \leq \frac{K_0 \inf_{v_h\in V_h} \|u-v_h\|_1}{1 - Kh(P+Q+Kh)} \leq \frac{K_0\,K_3|u|_k\,h^{k-1}}{1 - Kh(P+Q+Kh)}.$$

Hiermee is voldaan aan (4.5.5), waarmee de stelling bewezen is.

4.6. Een puntsgewijze foutschatting voor een Galerkin-methode

In deze paragraaf geven we een puntsgewijze foutschatting voor het probleem (4.5.1).

Zoals we reeds zagen (paragraaf 2.10), kunnen stuksgewijze polynomen van de graad k-1 een voldoende gladde functie *over het gehele interval* [a,b] benaderen met een globale fout $O(h^k)$. De puntsgewijze foutschatting, die we hier zullen afleiden, geeft echter het verschil tussen de berekende en de exacte oplossing van de differentiaalvergelijking *op een aantal - van te voren bekende - punten*. We zullen zien dat het, met de Galerkin-methode berekende, stuksgewijs polynoom dat de oplossing van het probleem (4.5.1) benadert, op de steunpunten $x_i \in \pi$ veel nauwkeuriger is dan $O(h^k)$, namelijk $O(h^{2k-2})$. We merken op dat een overeenkomstig verschijnsel te zien was bij de collocatie-methode: vergelijk de foutschattingen (4.2.4), (4.2.5) en (4.2.6).

Stelling 4.6.1 (DOUGLAS & DUPONT [1974]).

Laat $u_h \in V_h$ de Galerkin-oplossing zijn van Lu = f (zie paragraaf 4.5). Als

V_h alle stuksgewijs polynomen van de graad kleiner dan k over een partitie $\pi : \{a = x_0 < x_1 < \ldots < x_N = b\}$ bevat en als de stuksgewijze polynomen uit V_h bovendien discontinue afgeleiden in de steunpunten $x_i \in \pi$ toelaten, dan geldt

$$(4.6.1) \qquad |(u-u_h)(x_i)| \leq K \|u\|_k \, h^{2k-2}.$$

Bewijs

Uit (4.5.6) en (4.5.8) volgt

$$B(u_h-u,v_h) = 0 \qquad \text{, voor alle } v_h \in V_h.$$

Laat $G(x,\xi)$ de Greense functie zijn voor de vergelijking (4.5.1), dan geldt

$$w(x) = - (Lw,G(x,\cdot)).$$

De Greense functie heeft de volgende eigenschappen:

(i) $G(x,\xi) = G(\xi,x)$;

(ii) $G(x,\cdot) \in C_o^0[a,b] \cap H^1[a,x] \cap H^1[x,b]$;

(iii) $L^T G(x,\cdot) = 0$ op $[a,x)$ en op $(x,b]$;

(iv) $(\partial/\partial\xi)G(x,\xi)\big|_{\xi=x+0} - (\partial/\partial\xi)G(x,\xi)\big|_{\xi=x-0} = 1.$

We kunnen nu voor een vast punt $x \in (a,b)$ de volgende foutschatting geven:

$$|(u_h-u)(x)| = |(L(u_h-u),G(x,\cdot))| = |B(u_h-u,G(x,\cdot))| =$$

$$= |B(u_h-u,G(x,\cdot)-v_h)| \leq$$

$$\leq K_0 \|u-u_h\|_1 \, \|G(x,\cdot)-v_h\|_1, \text{ voor alle } v_h \in V_h.$$

Aangezien $G(x,\cdot)$ glad is op $[a,x)$ en $(x,b]$ en aangezien de discontinue afgeleide op een steunpunt door de functies uit V_h gerepresenteerd kan worden, geldt voor elke $x_i \in \pi$

$$\inf_{v_h \in V_h} \|G(x_i,\cdot)-v_h\|_1 \leq K_I \|G(x_i,\cdot)\|_k^* \, h^{k-1},$$

$$\|G(x_i,\cdot)\|_k^* = \|G(x_i,\cdot)\|_{H^k[a,x_i]} + \|G(x_i,\cdot)\|_{H^k[x_i,b]}.$$

Gecombineerd met het resultaat van stelling 4.5.1 levert dit

$$|(u_h-u)(x_i)| \leq K_0\, K\, \|u\|_k\, h^{k-1}\, K_I\, \|G(x_i,\cdot)\|^*_k\, h^{k-1}.$$

zodat de stelling bewezen is.

Opmerking

We merken op dat de Lagrange-polynomen voldoen aan de voorwaarden die in deze stelling gesteld worden aan een basisfunctie voor V_h. De Hermite-polynomen voldoen niet, omdat deze op de steunpunten geen discontinuïteiten in de afgeleide kunnen representeren.

4.7. De invloed van numerieke kwadratuur op de nauwkeurigheid van een Galerkin-methode

Is u_h de Galerkin-benadering van de vergelijking

$$B(u,v) = \ell(v) \qquad , \quad \text{voor alle } v \in V \tag{4.7.1}$$

met B een bilineaire en ℓ een lineaire operator, dan wordt u_h beschreven door

$$u_h = \sum_{i=1}^{N} a_i\, \phi_i \qquad ,$$

waarin (a_i) de oplossing is van het lineaire stelsel

$$\sum_{i=1}^{N} B(\phi_i,\phi_j)\, a_i = \ell(\phi_j) \qquad , \quad j = 1,\ldots,N. \tag{4.7.2}$$

De matrix-elementen $B(\phi_i,\phi_j)$ en de elementen van het rechterlid $\ell(\phi_i)$ zijn integralen. Voor de praktijk is het een vraag van fundamenteel belang, met welke nauwkeurigheid deze integralen berekend moeten worden. Het is een redelijke eis, te verlangen dat de integralen zó nauwkeurig worden uitgerekend dat de maximale orde van nauwkeurigheid in de berekening van u_h gegarandeerd is. Anderzijds is het nauwkeurig berekenen van de integralen een overbodige moeite, wanneer daardoor de berekende oplossing u_h geen nauwkeuriger benadering wordt van de exacte oplossing u. Deze twee overwegingen geven een duidelijk criterium voor de nauwkeurigheid waarmee de kwadratuur uitgevoerd moet worden.

We geven het discrete Galerkin-probleem aan met

(4.7.3) $B(u_h,v_h) = \ell(v_h)$, voor alle $v_h \in V_h$.

Het lineaire stelsel dat in feite wordt opgelost - d.w.z. het stelsel waarin de matrix $B(\cdot,\cdot)$ en de vector $\ell(\cdot)$ gestoord zijn door kwadratuurfouten - geven we aan met

(4.7.4) $B^*(u_h^*,v_h) = \ell^*(v_h)$, voor alle $v_h \in V_h$.

Adaptieve kwadratuurprocedures waarmee de integralen met een gewenste absolute of relatieve nauwkeurigheid worden uitgerekend, verrichten in het algemeen meer werk dan voor het doel noodzakelijk is. We kunnen namelijk, na enige analyse, van te voren aangeven *met welke orde van nauwkeurigheid* de integraal berekend moet worden.

Beschouwen we bijvoorbeeld het randwaardeprobleem

$$u'' + p(x)u' + q(x)u = f(x), \quad a \le x \le b,$$

(4.5.7) $$u(a) = u(b) = 0,$$

$$p,\ q \text{ en } f \text{ voldoende glad.}$$

Zij $\{\phi_i\}$ een basis van V_h, de ruimte van continue functies die op ieder deelsegment van de partitie

$$\pi : a = x_0 < x_1 < \ldots < x_N = b$$

een polynoom van de graad kleiner dan k zijn èn die in a en b de waarde 0 aannemen. Dan wordt de oplossing van (4.7.3) bepaald door (4.7.2), waarin

$$\begin{aligned} B(\phi_i,\phi_j) &= \sum_{m=1}^{N} B_m(\phi_i,\phi_j) = \\ &= \sum_{m=1}^{N} \int_{x_{m-1}}^{x_m} [-\phi_i'\phi_j' + p\phi_i'\phi_j + q\phi_i\phi_j]dx; \\ \ell(\phi_j) &= \sum_{m=1}^{N} \ell_m(\phi_j) = \\ &= \sum_{m=1}^{N} \int_{x_{m-1}}^{x_m} f\phi_i \, dx. \end{aligned} \tag{4.7.6}$$

In de praktijk nemen we genoegen met de oplossing u_h^* van (4.7.4), waarin

$$
\begin{aligned}
B^*(\phi_i,\phi_j) &= \sum_{m=1}^{N} B_m^*(\phi_i,\phi_j); \\
\ell^*(\phi_j) &= \sum_{m=1}^{N} \ell_m^*(\phi_j).
\end{aligned}
\tag{4.7.7}
$$

Hierin zijn $B_m^*(\cdot,\cdot)$ en $\ell_m^*(\cdot)$ benaderingen van $B_m(\cdot,\cdot)$, respectievelijk $\ell_m(\cdot)$, die verkregen worden door het toepassen van een kwadratuurregel op het interval $[x_{m-1},x_m]$.

Stelling 4.7.1 Zij $B(u,v)$ een begrensde, sterk coërcieve bilineaire vorm, i.e. er bestaan positieve constantes c_1 en c_2 met

$$|B(u,v)| \le c_1 \|u\|_1 \|v\|_1, \qquad \text{voor alle } u,v \in H_0^1[a,b];$$

$$|B(u,u)| \ge c_2 \|u\|_1^2, \qquad \text{voor alle } u \in H_0^1[a,b].$$

Zij $\pi : a = x_0 < x_1 < \ldots < x_N = b$ een kwasi-uniforme partitie van $[a,b]$, i.e. er bestaat een positieve constante λ, onafhankelijk van de partitie, met

$$h = \max_i (x_i - x_{i-1}) \le \lambda \min_i (x_i - x_{i-1}).$$

Zij u_h de oplossing van (4.7.4), waarin $B^*(\cdot,\cdot)$ en $\ell^*(\cdot)$ door (4.7.7) zijn bepaald.

Indien nu $B_m^*(\cdot,\cdot)$ en $\ell_m^*(\cdot)$ worden verkregen door een kwadratuur van de orde $t \ge 2k-2$ op het interval $[x_{m-1},x_m]$ toe te passen (i.e. een kwadratuur die alle polynomen van de graad kleiner dan $2k-2$ exact integreert), dan gelden voor voldoend kleine h de volgende foutschattingen:

$$|u(x_i) - u_h^*(x_i)| = O(h^{2k-2}), \qquad i = 1,\ldots,N-1;$$

$$\|u - u_h^*\|_j = O(h^{k-j}), \qquad j = 0,1.$$

Bewijs Zie DOUGLAS & DUPONT [1974] en HEMKER [1975]. □

LITERATUUR

DE BOOR, C. & B. SWARTZ, *Collocation at Gaussian points*, SIAM J. Num. Anal., 10, p.582-606 (1973).

DOUGLAS, J. & T. DUPONT, *Galerkin Approximations for the Two Point Boundary Problem using Continuous Piecewise Polynomial Spaces*, Num. Math. 22, p.99-109 (1974).

HEMKER, P.W., *Galerkin's Method and Lobatto points*, NW 24/75, MC, 1975.

RUSSEL, R.D. & L.F. SHAMPINE, *A collocation method for boundary value problems, Parts I and II*, Num. Math. 19, p.1-28 (1972).

5. GALERKIN-METHODEN VOOR GEWONE DIFFERENTIAALVERGELIJKINGEN

5.1. Inleiding

Deze bijdrage geeft een samenvatting van enkele resultaten behaald in het onderzoek van projectiemethoden. Omdat het een overzicht, een "survey" betreft, worden geen bewijzen gegeven.

De laatste paragraaf, §6, wijkt af van een eerder manuscript, waarvan copieën tijdens de voordracht zijn rondgedeeld. De formulering is geheel gewijzigd n.a.v. gerechtvaardigde kritiek van enkele aanwezigen, met name P.W. Hemker.

5.2. De "midpoint" regel

Zij gegeven de gewone dvgl.

$$\frac{dx}{dt} = f(t,x) \tag{5.2.1}$$

met $x \in \mathbb{R}^n$, f voldoende glad. Zij $a < b$, en laat gevraagd worden naar de oplossing van (5.2.1) onder de nevenconditie

$$Ax(a) + Bx(b) = g, \tag{5.2.2}$$

met A, B matrices (i.h.a. singulier) en g een vector uit $\mathbb{R}^n$.

Een probleem van deze vorm is b.v.;

(5.2.3) Een randwaarde probleem voor een gewone tweede orde dvgl., indien de 2e orde dvgl. wordt geschreven als een equivalente 1e orde dvgl. Zo wordt het probleem

$$x'' + ax' + bx = f,$$

(5.2.4)

$$x(a) = g_1 \ , \quad x(b) = g_2$$

geschreven als

$$\frac{d}{dt}\begin{pmatrix} y \\ x \end{pmatrix} = \begin{pmatrix} -a & -b \\ 1 & 0 \end{pmatrix}\begin{pmatrix} y \\ x \end{pmatrix} + \begin{pmatrix} f \\ 0 \end{pmatrix},$$

(5.2.5)

$$\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} y(a) \\ x(a) \end{pmatrix} + \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} y(b) \\ x(b) \end{pmatrix} = \begin{pmatrix} g_2 \\ g_1 \end{pmatrix}.$$

Nu lijkt het niet slim om (5.2.5) numeriek aan te vatten, omdat (5.2.4) ogenschijnlijk een efficiënter algorithme zal opleveren. Onderzocht is dit door VARAH [1974]. Heel globaal zullen wij zijn artikel bespreken, dat aantoont, dat via (5.2.5) een methode is te verkrijgen die kan wedijveren qua efficiëncy met een oplosmethode verkregen via (5.2.4).

Varah heeft de "midpoint rule" in gedachten, n.a.v. een aantal artikelen van KELLER [1969,1974]. Kies in het segment [a,b] een partitie

$$\delta_N: \quad a = t_0 < t_1 < \ldots < t_{n-1} < t_N = b.$$

Zij x_i een benadering voor $\tilde{x}(t_i)$, $\tilde{x}$ de unieke oplossing van (5.2.1), (5.2.2) (uniciteit van $\tilde{x}$ en existentie wordt altijd zonder meer verondersteld). Dan vinden we de x_i als oplossing van

$$x_{i+1} - x_i = h_i f(t_i + h_i/2, \frac{x_i + x_{i+1}}{2}), \qquad i = 0,1,\ldots,N-1.$$

(5.2.6)

$$Ax_0 + Bx_N = g.$$

Door KELLER [1969] is aangetoond dat de oplossing van (5.2.6) bestaat en uniek is indien $\max_i h_i$ voldoende klein is. Bovendien toont hij aan dat de f $x_i - x(t_i)$ te ontwikkelen is in een asymptotische reeks naar machten van $h = \max_i h_i$ (ook indien niet equidistante t_i zijn gekozen).

De bewering van VARAH luidt nu:
de methode (5.2.6) met 1 maal Richardson extrapolatie (d.w.z. berekend worden benaderingen voor een rooster δ_N en een rooster δ_{2N}, waarbij δ_{2N}

uit δ_N ontstaat door als extra steunpunten de middens van (t_i,t_{i+1}) er bij te nemen) is efficiënter (toegepast op (5.2.5)) dan de eindige elementen methode met "Hermite cubics" op het rooster δ_N en 2-punts Gauss-kwadratuur (toegepast op (5.2.4)) voor niet te ingewikkelde a, b en f.

Dit resultaat van Varah kan men zien als voorlopige motivatie om een 2-punts randwaarde probleem in de vorm (5.2.5) aan te vatten. We zullen daarbij de "midpoint rule" (5.2.6) generalizeren en tenslotten een 4de orde methode vinden die wellicht efficiënter is dan de "midpoint rule" plus één extrapolatie.

5.3. Een projectiemethode

Eenvoudigheidshalve beperken we ons tot een lineaire vergelijking

$$\frac{d}{dt}x - Ax = Lx = f. \tag{5.3.1}$$

Hier zullen we aannemen $t \in [0,1]$. We veronderstellen

$$f \in L^2[0,1], \tag{5.3.2}$$

A een begrensde lineaire afb. van $L^2[0,1]$ in zichzelf.

Opmerking

Met $L^2[0,1]$ duiden we aan de equivalentieklassen van afbeeldingen f van $[0,1]$ in $\mathbb{R}^n$ met

$$\|f\|_0 = \left\{ \int_0^1 | f(x) |^2 \, dx \right\}^{1/2} < \infty,$$

waarbij voor $z \in \mathbb{R}^n$ de euclidische norm van z aangeduid wordt door $|z|$.

Voorbeeld 5.3.1 Zie (5.2.5); dan $A = \begin{pmatrix} -a & -b \\ 1 & 0 \end{pmatrix}$.

Voorbeeld 5.3.2

$$\frac{d}{dt}x - \int_0^t q(s)x(s)ds = \int_0^t f(s)ds.$$

Dit is equivalent met

$$x'' - qx = f \quad , \; x'(0) = 0$$

Zij V een gesloten lineaire deelruimte van $H^1[0,1]$.

Veronderstelling De vergelijking (5.3.1) bezit een unieke oplossing $\tilde{x}$ in V en de inverse van L van $L^2[0,1]$ in V is compact.

Voorbeeld 5.3.3

Kies

$$V = \{z \mid z \in H^1[0,1];\ z(t) \in \mathbb{R}^2;\ z_2(0) = z_2(1) = 0\}.$$

(Hier is $z_2(t)$ de tweede coördinaat van $z(t)$).
Een oplossing van

$$\frac{d}{dt} x - \begin{pmatrix} -a & -b \\ 1 & 0 \end{pmatrix} x = \begin{pmatrix} f \\ 0 \end{pmatrix}$$

in V is dan oplossing van

$$x'' + ax' + bx = f$$

met de randcondities $x(0) = x(1) = 0$. Het is bekend dat de inverse van L in dit geval compact is.

We komen nu aan een beschrijving van de projectiemethode. Zij δ_N het rooster

$$\delta_N = \{t_i \mid 0 = t_0 < t_1 < \ldots < t_{N-1} < t_N = 1\}. \tag{5.3.4}$$

We stellen

$$h_i = t_{i+1} - t_i, \quad I_i = [t_i, t_{i+1}], \tag{5.3.5}$$

en nemen aan

$$h = \max_i h_i \lesssim \rho h_i, \quad \forall i,\ \rho \text{ onafhankelijk van de partitie.} \tag{5.3.6}$$

We proberen een benaderde oplossing te vinden in

(5.3.7) $E(\delta_N;k) = \{z \in V \mid z \text{ op } I_i \text{ polynoom van graad} \leq k, \forall i\}.$

Omdat $V \subset H^1[0,1]$ is $E(\delta_N;k)$ de verzameling van continue functies waarvan de beperking tot I_i ($\forall i$) een polynoom van graad $\leq k$ is. We hebben nog nodig

(5.3.8) $F(\delta_N;k-1) = \{z \in L^2[0,1] \mid z \text{ op } I_i \text{ polynoom van graad} \leq k-1, \forall i\}.$

Voorbeeld

Voor V als in (5.3.7) geldt

$$\dim E(\delta_N;k) = \dim F(\delta_N;k-1) = 2kN.$$

Zij P_N de projectie (orthogonaal) van $L^2[0,1]$ op $F(\delta_N;k-1)$. (Eigenlijk zouden we moeten noteren $p(\delta_N;k-1)$ of iets dergelijks i.p.v. P_N). Dus voor $z \in L^2[0,1]$, is $P_N z$ bepaald door

$$z - P_N z \perp F(\delta_N;k-1)$$

of equivalent

$$(z-P_N z, v) = 0, \qquad \forall v \in F(\delta_N;k-1).$$

Hier is $(.,.)$ het inproduct in $L^2[0,1]$.

De projectiemethode wordt gedefiniëerd door de vergelijking

(5.3.9) $P_N L x_N = P_N f, \qquad x_N \in E(\delta_N;k).$

Omdat

$$\frac{d}{dt} E(\delta_N;k) \subset F(\delta_N;k-1),$$

gaat (5.3.9) over in

(5.3.10) $\frac{d}{dt} x_N - P_N A x_N = P_N f, \qquad x_N \in E(\delta_N;k)$

of

$$\left(\frac{d}{dt}x_N - Ax_N, v\right) = (f,v), \qquad \forall v \in F(\delta_N;k-1).$$

5.4. Enkele eigenschappen van de benadering

De in deze paragraaf te noemen stellingen worden bewezen in VAN VELDHUIZEN [1975]; dat we ons daar beperken tot het geval van periodieke randcondities is niet essentieel. A(t) zal steeds een n × n matrix aanduiden maar dit is niet overal nodig.

Zij de operator B_N (eigenlijk $B(\delta_N;k)$) gedefinieerd door

$$(5.4.1) \qquad B_N = [I + (I-P_N)AH]^{-1}, \qquad H = L^{-1}.$$

Stelling 5.4.1

Onder de condities van §2 geldt:

(i) B_N is "1-1 en op" van $L^2[0,1]$ op zichzelf voor alle $N \geq N_0$, en $\|B_N\| \leq$ constante onafhankelijk van N.

(ii) Voor alle $N \geq N_0$ bezit (5.2.6) een unieke oplossing $\tilde{x}_N \in E(\delta_N;k)$.

(iii) Als $\tilde{x}_N$ bestaat, dan

$$(5.4.2) \qquad \tilde{x} - \tilde{x}_N = HB_N(I-P_N)\frac{d}{dt}\tilde{x}$$

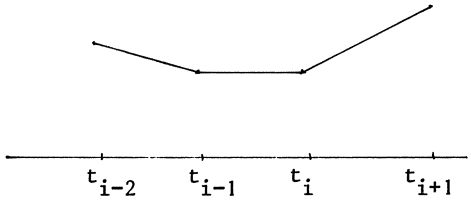
en

$$\|\tilde{x} - \tilde{x}_N\|_{L^2[0,1]} \to 0, \qquad \text{voor } N \to \infty.$$

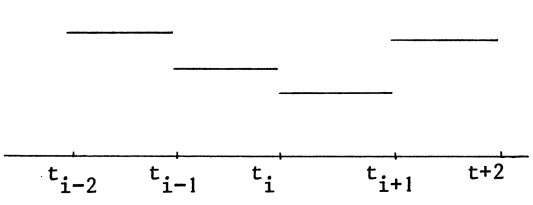
Vervolgens vragen we ons af hoe groot de fout $\tilde{x} - \tilde{x}_N$ is, afhankelijk van h (dus N). Het zou jammer zijn om zo zonder meer het resultaat te geven, omdat we hier met Nitsche's trick te maken krijgen, op een erg doorzichtige manier. We bezien daarom het geval k = 1, het meest eenvoudige geval.

Voorbeeld

Voor k = 1, is $E(\delta_N;1)$ de lineaire ruimte opgespannen door de "roof functions", en $F(\delta_N;0)$ is de ruimte der op I_i ($\forall i$) constante functies. Dus:



Figuur 5.1 Element uit $E(\delta_N;1)$



Figuur 5.2 Element uit $F(\delta_N;0)$

Zij bovendien $\tilde{x} \in H^2[0,1]$. Zij $z_N \in E(\delta_N;1)$ zo gekozen dat

$$z_N(t_i) = \tilde{x}(t_i), \quad i = 0,1,2,\ldots,N.$$

D.w.z. z_N is het stuksgewijze lineaire interpolatiepolynoom van $\tilde{x}$. Kennelijk geldt:

$$\int_{I_i} (\frac{d}{dt} z_N - \frac{d}{dt} \tilde{x})dt = 0, \qquad , \forall i.$$

of equivalent

$$(\frac{d}{dt} z_N - \frac{d}{dt} \tilde{x}, v_i) = 0 \qquad , \forall i,$$

met v_i gelijk aan 1 op I_i en 0 elders. De v_i vormen echter een basis van $F(\delta_N;0)$. M.a.w.

$$\frac{d}{dt} z_N - \frac{d}{dt} \tilde{x} \mid F(\delta_N;0)$$

en dus

$$\frac{d}{dt} z_N = P_N \frac{d}{dt} \tilde{x}.$$

Bijgevolg: (alle normen zijn $L^2[0,1]$ normen)

$$\|\tilde{x} - \tilde{x}_N\|_0 \leq \| H \| \|B_N\| \|\frac{d}{dt}(z_N - \tilde{x})\|_0 = O(h)$$

We hebben echter ook, als tenminste de normen bestaan

$$\| \tilde{x} - \tilde{x}_N \| \leq \| HB_N \frac{d}{dt} \| \, \| z_N - \tilde{x} \|_0 .$$

Nu geldt (via narekenen)

$$\| HB_N \frac{d}{dt} \| \leq c, \text{ onafh. van N, zodat de foutschatting wordt}$$

$$\| \tilde{x} - \tilde{x}_N \|_0 = O(h^2).$$

Het algemene resultaat is:

Stelling 5.4.2

Onder de condities van §3 en $\tilde{x} \in H^{k+1}[0,1]$ geldt

$$\| \tilde{x} - \tilde{x}_N \|_0 + h \| \frac{d}{dt} \tilde{x} - \frac{d}{dt} \tilde{x}_N \|_0 \leq c \, h^{k+1} \| \frac{d^{k+1}}{dt^{k+1}} \tilde{x} \|$$

(in $L^2[0,1]$- norm) met c onafh. van h en $\tilde{x}$. □

Puntsgewijze foutschatting in de t_i worden verkregen m.b.v. de truuk van DOUGLAS & DUPONT [1974]. Men bedenke dat H in de beschouwde gevallen een integraaloperator is

$$(Hz)(t) = \int_0^1 H(t,s)z(s)ds$$

met H(t,s) (dus de kern) discontinu voor s = t, maar met H(t,.) glad op [0,t) en (t,1]. Omdat

$$(I-P_N)B_N(I-P_N) = B_N(I-P_N)$$

en omdat P_N en dus $I - P_N$ een zelfgeadjugeerde operator is, geldt

$$(\tilde{x}-\tilde{x}_N)(t) = ((I-P_N)H(t,.), \, B_N(I-P_N) \frac{d}{dt} \tilde{x}) \frac{d}{dx} \tilde{x})_{L^2[0,1]},$$

waaruit volgt

Stelling 5.4.3

Onder de condities van §2, met $\tilde{x}$ en A voldoende glad, geldt voor $i = 0,1,2,\ldots,N$:

$$|\tilde{x}(t_i) - \tilde{x}_N(t_i)| \leq c\, h^{2k} \left\| \frac{d^{k+1}}{dt^{k+1}} \tilde{x} \right\|_0$$

c onaf. van i,h en $\tilde{x}$.
Voor de praktijk is van belang:

Stelling 5.4.4

Laat iedere integraal over I_i benaderd worden m.b.v. de s-punt Gauss-Legendre ($s \geq k$) kwadratuur formule, aangepast aan I_i. Dan bestaat er voor $N \geq N_0$ een unieke $\hat{x}_N \in E(\delta_N;k)$ als oplossing van (2.15) (met G-L kwadratuur),

$$\|\tilde{x} - \hat{x}_N\| \to 0, \qquad N \to \infty.$$

Bovendien verandert de orde in h van de foutschattingen uit stelling 5.4.7 en 5.4.3 niet, mits $\tilde{x}$, A voldoende glad zijn.

5.5. Een verfijningsproces

In deze § bespreken we ons tot het probleem

$$\frac{d}{dt}\begin{pmatrix} y \\ z \end{pmatrix} = \begin{pmatrix} -a & -b \\ 1 & 0 \end{pmatrix}\begin{pmatrix} y \\ z \end{pmatrix} + \begin{pmatrix} f \\ 0 \end{pmatrix},$$

(5.5.1)

$$z(0) = z(1) = 0.$$

We bekijken een interval I_i. Op I_i kunnen we y schrijven als

(5.5.2) $$y = a_0 L_{0,1} + a_1 L_{1,i} + \ldots + a_k L_{k,i} + \ldots ,$$

met $L_{j,i}$ het j-de Legendre polynoom, aangepast aan I_i. We nemen zelfs aan dat de $\{L_{j,i}\}$ een *orthonormale* verz. polynomen in $L^2(I_i)$ is. Dan wordt $P_N y$ op I_i gegeven door

(5.5.3) $\quad a_0 L_{0,i} + \ldots + a_{k-1} L_{k-1,i}.$

Dus voor $t \in I_i$:

$$(5.5.4) \qquad z_N(t) = z_N(t_i) + \int_{t_i}^{t} (a_0 L_{0,i} + \ldots + a_{k-1} L_{k-1,i}) ds.$$

Definieer nu

$$(5.5.5) \qquad \bar{z}_N(t) = \tilde{z}_N(t_i) + \int_{t_i}^{t} \tilde{y}_N(s) ds, \qquad t \in I_i.$$

Dan

$$(5.5.6) \qquad z_N(t) - z(t) = [\tilde{z}_N(t_i) - \tilde{z}(t_i)] + \int_{t_i}^{t} [\tilde{y}_N(s) - \tilde{y}(s)] ds.$$

M.b.v. deze formule kan men aantonen:

Stelling 5.5.1

Voor $\tilde{y} \in H^{k+2}[0,1]$ en $k \geq 2$:

$$\| \bar{z}_N - \tilde{z} \|_{L^2[0,1]} \leq c\, h^{k+2} \| \tilde{z} \|_{H^{k+3}[0,1]},$$

c een constante onafh. van $\tilde{z}$ en h.

Corollarium 5.5.2

De conclusie van de vorige stelling blijft juist, als men uitgaat van $\hat{z}_N$, $\hat{y}_N$ i.p.v. $\tilde{z}_N$, $\tilde{y}_N$.

Dat in stelling 5.5.1 wordt geeist $k \geq 2$ is niet toevallig. De truck berust namelijk op de extra nauwkeurigheid in de knooppunten en die is er voor $k = 1$ niet, voor $k \geq 2$ wel.

Een andere methode is iets ingewikkelder. De benadering $\bar{\bar{z}}_N$ wordt nu verkregen als oplossing van het 2-punts randwaarde probleem

$$(5.5.7) \qquad \begin{aligned} z'' &= Q_{i,k}(-a\tilde{y}_N - b\tilde{z}_N + f), && t \in I_i \\ z(t_i) &= \tilde{z}_N(t_i), \quad z(t_{i+1}) = \tilde{z}_N(t_{i+1}). \end{aligned}$$

Hier is $Q_{i,k}$ de orthogonale projectie in $L^2(I_i)$ op de polynomen van graad $\leq k$. Het ingewikkelde bewijs geven we niet, maar er komt uit:

Stelling 5.5.3

Voor a, b, f, $\tilde{z}$ voldoende glad en $k \geq 3$ geldt

$$\| z - \bar{\bar{z}}_N \|_{L^2[0,1]} = O(h^{k+3}).$$

Corollarium 5.5.4

De vorige stelling blijft juist, uitgaande van $\hat{z}_N$, $\hat{y}_N$, indien de projecties $Q_{i,k}$ benaderd worden d.m.v. (k+1)-punts Gauss-Legendre-kwadratuur.

We bezien nu nog $\tilde{z}_N - \bar{z}_N$, $\bar{z}_N - \bar{\bar{z}}_N$ op I_i. We hebben (men mag dit zelf aantonen)

Lemma 5.5.5

Op ieder der subintervallen I_i geldt:

$$\bar{\bar{z}}_N = \bar{z}_N + \bar{u}_N$$

met $\bar{u}_N$ oplossing van

$$z'' = (Q_{i,k} - Q_{i,k-1})(-a\tilde{y}_N - b\tilde{z}_N + f), \qquad t \in I_i,$$

$$\bar{u}_N(t_i) = \bar{u}_N(t_{i+1}) = 0$$

(dit geldt ook met G - L-kwadratuur en $\hat{y}_N$, $\hat{z}_N$ als in corollarium 8).

Bijgevolg:

$$\bar{z}_N(t) = \tilde{z}_N(t) + \text{veelvoud} \int_{t_i}^{t} L_{k,i}\, ds \quad ,$$

(5.5.8)

$$\bar{\bar{z}}_N(t) = \bar{z}_N(t) + \text{veelvoud} \int_{I_i} G_i(t,s)\, L_{k,i}\, ds$$

(met in beide gevallen i.h.a. een verschillend veelvoud).

Hier

$$(5.5.9)\qquad G_i(t,s) = \begin{cases} (t-t_{i+1})(s-t_i)/h_i, & s < t;\ s,t \in I_i \\ (t-t_i)(s-t_{i+1})/h_i, & s > t;\ s,t \in I_i \end{cases}.$$

Na inkrimpen en verschuiven van I_i naar $[-1, +1]$ (een lineaire transformatie) vinden we, op een constante na, voor $k = 3$:

$$(5.5.10)\qquad \int_{t_i}^{t} L_{k,i}\, ds \to (\tau^2-1)(\tau^2-1/5)\ ;$$

$$(5.5.11)\qquad \int_{I_i} G_i(t,s)\, L_{k,i}\, ds \to \tau(\tau^2-1)^2$$

en voor $k = 2$:

$$(5.5.12)\qquad \int_{t_i}^{t} L_{k,i}\, ds \to \tau(\tau^2-1)$$

(steeds $\tau \in [-1,+1]$ corresponderend met $t \in I_i$).
We kunnen nu stelling 5.4.3 als volgt uitbreiden (voor k=2, k=3):

Stelling 5.5.6

Voor $k = 2$ geldt:

$$|\tilde{z}_N(t_i) - \tilde{z}(t_i)| = O(h^4)$$

$$|\tilde{z}_N(\tfrac{t_i+t_{i+1}}{2}) - \tilde{z}(\tfrac{t_i+t_{i+1}}{2})| = O(h^4)$$

en men mag hier $\tilde{z}_N$ door $\hat{z}_N$ vervangen.
Voor $k = 3$ geldt:

$$|\tilde{z}_N(t_i) - \tilde{z}(t_i)| = O(h^6)$$

$$|\tilde{z}_N(\tfrac{t_i+t_{i+1}}{2}) - \tilde{z}(\tfrac{t_i+t_{i+1}}{2})| = O(h^6)$$

met weer eventueel $\hat{z}_N$ i.p.v. $\tilde{z}_N$.

5.6. Enkele resultaten

Voor k = 1,2 en een equidistant rooster zijn de methoden uit de vorige §§ geimplementeerd (in Algol). We gaan hier niet op in, omdat een uitvoerige beschrijving van een efficiënte programmering een hoofdstuk apart is. Voor k = 3 zijn enkele ad hoc berekeningen uitgevoerd voor een probleem met constante coefficiënten.

In tabel I worden enkele resultaten gegeven voor het probleem:

$$\frac{d}{dt}\begin{pmatrix} y \\ x \end{pmatrix} = \begin{pmatrix} -5 & -4 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} y \\ x \end{pmatrix} + \begin{pmatrix} -5 & -4t \\ & 0 \end{pmatrix}, \qquad t \in (0,1),$$

(5.6.1)

$$x(0) = x(1) = 0.$$

De oplossing is

$$(5.6.2) \qquad x(t) = \frac{e^{-4t}-e^{-t}}{e^{-4}-e^{-1}} - t, \qquad y(t) = \frac{-4e^{-4t}+e^{-t}}{e^{-4}-e^{-1}} - 1.$$

Uit table I blijkt dat het gedrag van de daar vermelde fouten in de $L^2[0,1]$-norm conform de theorie is. Voor de supnorm volgt uit tabel I eenzelfde foutengedrag; i.h.b. is de orde in h gelijk aan die in de $L^2[0,1]$-norm. Voor de fouten $x - x_N$ is dit bewezen in RUSSELL [1974]. Merk op dat voor het eenvoudige probleem dat we hier bezien geldt $\hat{x}_N = \tilde{x}_N$.

VARAH [1974] vergelijkt methoden op grond van tellingen van het aantal benodigde arithmetische operaties. Op grond van zulke tellingen zou het proces dat voor k = 2 de benadering $\bar{x}_N$ bepaalt (met 2-punts Gauss-kwadratuur) efficiënter zijn dan Keller's "midpoint" regel met één keer Richardson extrapolatie. Dit is reeds een indicatie voor het praktische nut van deze methode. Veel belangrijker lijkt echter het feit dat deze methode op zeer goedkope wijze een foutschatting toestaat (nl. door $\|\bar{x}_N - \hat{x}_N\|$ te berekenen). Er zijn mij geen andere methoden voor randwaarde problemen bekend die op zo'n goedkope wijze een foutschatting toestaan (alleen Richardson extrapolatie en Pereyra's deferred corrections proces zijn mij als realistische methoden bekend). Deze goedkope foutschattingen zou men eventueel kunnen gebruiken om een algorithme te ontwerpen dat automatisch een rooster bepaalt, rekening houdend met het locale foutgedrag. Een en ander zou eerst eens degelijk doordacht en getest moeten worden,

maar zo'n proces lijkt op dit moment zeer wel mogelijk.

Als illustratie van (5.5.8) - (5.5.12) zijn de grafieken I, II, III opgenomen. Merk op dat 9/16 en 10/16 het 10de en 11de steunpunt zijn. Omdat $\bar{x}$ uit $\tilde{x}$ ontstaat door er een veelvoud van

$$\int_{t_i}^{t} L_{k,i}(s)ds, \qquad t \in [t_i, t_{i+1}]$$

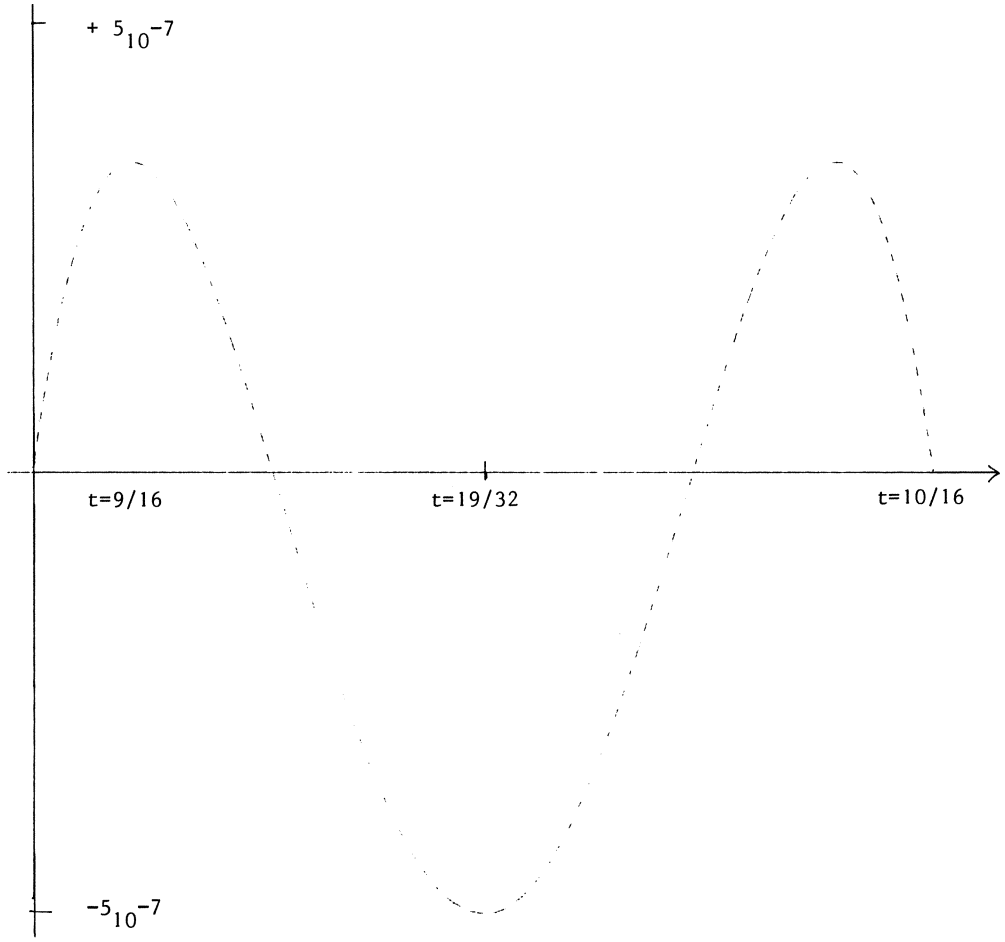
bij op te tellen, blijkt uit grafiek I en II samen dat tussen de steunpunten de fout in dit voorbeeld goed wordt gerepresenteerd door $\tilde{x}_{16} - \bar{x}_{16}$. Evenzo blijkt uit grafiek II en III dat de fout in $\bar{x}_{16}$ in redelijke mate wordt gerepresenteerd door $\bar{\bar{x}}_{16} - \bar{x}_{16}$. Het is te hopen dat deze fouten, tussen de steunpunten in, ook iets zeggen over de fout in de steunpunten. Dit is een onderwerp voor nader onderzoek.

Het is duidelijk dat voorbeeld (5.6.1) een zeer eenvoudig voorbeeld is. Daarom zij vermeld dat voor k = 2 de gegeven getallen representatief genoemd kunnen worden voor problemen met niet-constante coefficiënten. Dit is gebleken uit enkele andere doorgerekende voorbeelden, waarvan de oplossingen zich tamelijk "tam" gedroegen. Voor snel variërende oplossingen is een niet-equidistant rooster een vereiste, zodat op dit moment geen indicatie voor het succes van de methode in dit soort gevallen gegeven kan worden.

	k = 1		k = 2		k = 3			
N	$\tilde{x}_N - \tilde{x}$	$\bar{x}_N - \tilde{x}$	$\tilde{x}_N - \tilde{x}$	$\bar{x}_N - \tilde{x}$	$\tilde{x}_N - \tilde{x}$	$\bar{x}_N - \tilde{x}$	$\bar{\bar{x}}_N - \tilde{x}$	
1	-	-	-	-	$4.3_{10}-2$	$3.1_{10}-2$	$2.7_{10}-2$	$L^2[0,1]$-norm
2	-	-	$3.8_{10}-2$	$1.8_{10}-2$	$4.2_{10}-3$	$1.3_{10}-3$	$5.2_{10}-4$	
4	$6.1_{10}-2$	$4.6_{10}-2$	$5.3_{10}-3$	$1.3_{10}-3$	$3.1_{10}-4$	$4.7_{10}-5$	$9.8_{10}-6$	
8	$1.6_{10}-2$	$1.1_{10}-2$	$7.0_{10}-4$	$8.2_{10}-5$	$2.1_{10}-5$	$1.6_{10}-6$	$1.6_{10}-7$	
16	$4.0_{10}-3$	$2.8_{10}-3$	$8.8_{10}-5$	$5.2_{10}-6$	$1.3_{10}-6$	$4.9_{10}-8$	$2.6_{10}-9$	
32	$9.9_{10}-4$	$6.9_{10}-4$	$1.1_{10}-5$	$3.3_{10}-7$	$8.1_{10}-8$	-	$4.1_{10}-11$	
64	$2.5_{10}-4$	$1.7_{10}-4$	$1.4_{10}-6$	$2.0_{10}-8$	-	-	-	
1	-	-	-	-	7.7 -2	5.7 -2	4.5 -2	supremum-norm
2	-	-	$8.3_{10}-2$	$4.0_{10}-2$	$8.7_{10}-3$	$3.1_{10}-3$	$1.3_{10}-3$	
4	$1.7_{10}-1$	$8.8_{10}-2$	$1.5_{10}-2$	$3.9_{10}-3$	$9.0_{10}-4$	$1.4_{10}-4$	$4.7_{10}-5$	
8	$5.7_{10}-2$	$1.9_{10}-2$	$2.3_{10}-3$	$3.3_{10}-4$	$7.2_{10}-5$	$5.5_{10}-6$	$1.6_{10}-6$	
16	$1.7_{10}-2$	$4.6_{10}-3$	$3.2_{10}-4$	$2.4_{10}-5$	$5.1_{10}-6$	$1.9_{10}-7$	$4.9_{10}-8$	
32	$4.7_{10}-3$	$1.1_{10}-3$	$4.2_{10}-5$	$1.7_{10}-6$	$3.4_{10}-7$	-	$2.2_{10}-10$	
64	$1.2_{10}-3$	$2.8_{10}-4$	$5.4_{10}-6$	$1.1_{10}-7$	-	-	-	

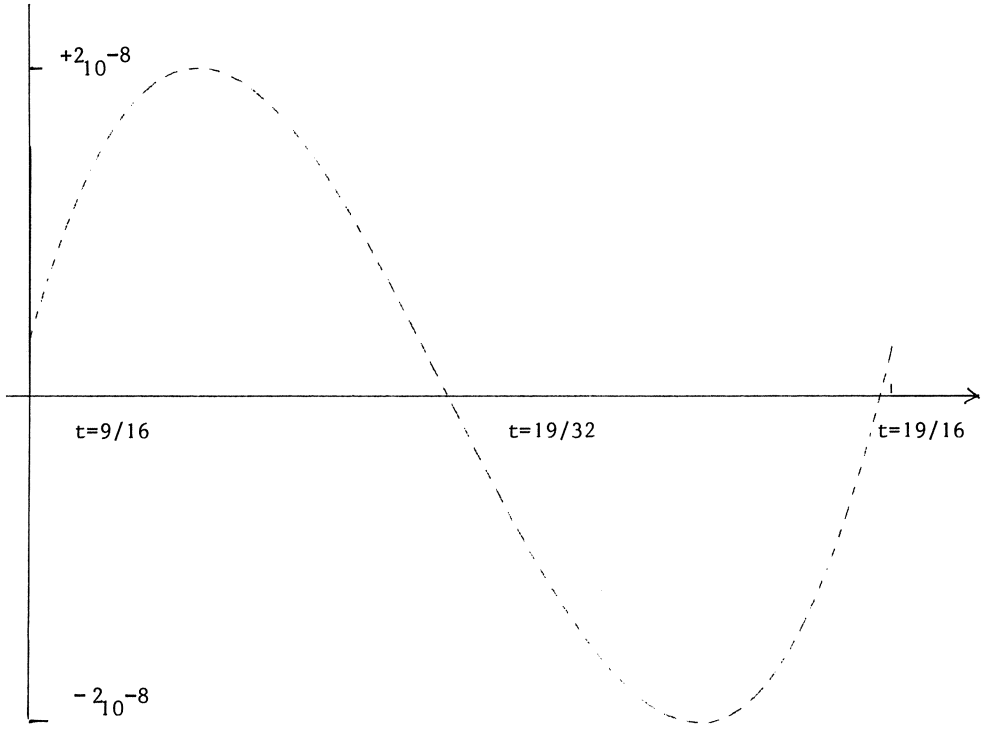
Tabel I.

Overzicht van de locale en globale fout in de oplossing van (5.6.1).



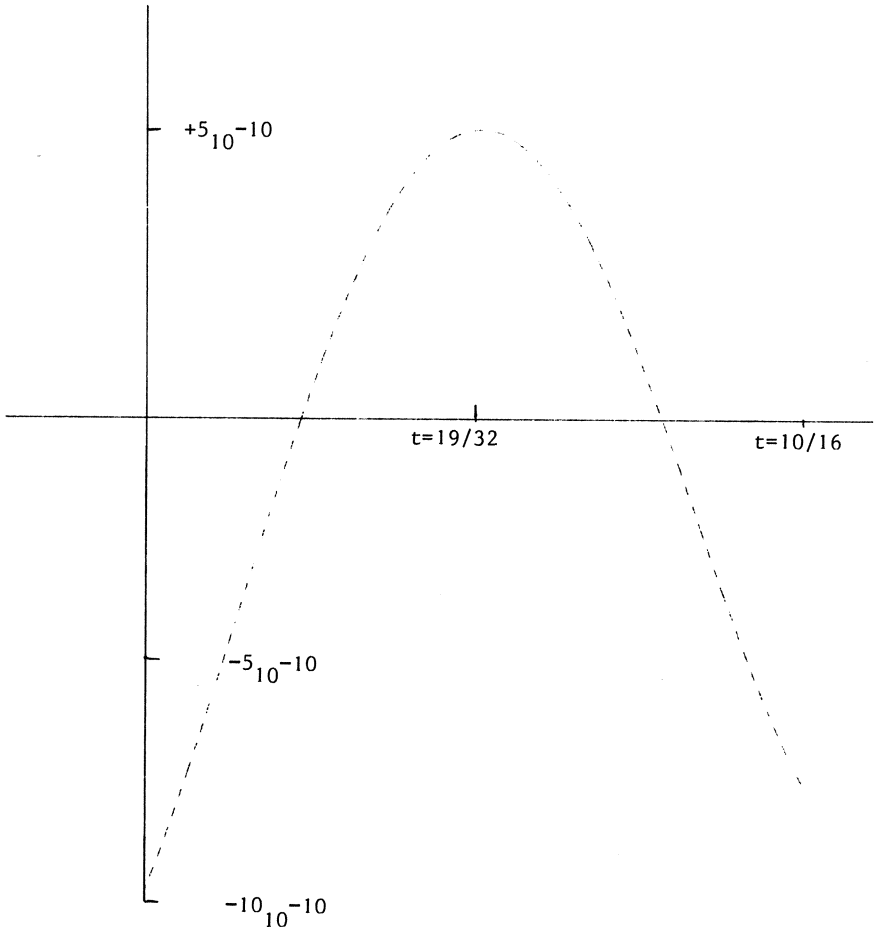
Figuur 5.3 grafiek I.

$\tilde{x} - \tilde{x}_{16}$, $t \in [9/16, 10/16]$, dvgl. (5.5.4), k=3.



Figuur 5.4 grafiek II.

$\bar{x}_{16} - \tilde{x}$, $t \in [9/16, 10/16]$, dvgl (5.5.4), k=3.



Figuur 5.5 grafiek III.

$\tilde{x} - \bar{\bar{x}}_{16}$, $t \in [9/16, 10/16]$, dvgl (5.5.4), k=3.

LITERATUUR

DOUGLAS, J., Jr., DUPONT, T., *Galerkin Approximations for the Two Point Boundary Problem using Continuous Piecewise Polynomial Spaces*, Numer. Math. 22, 99-109 (1974).

KELLER, H.B., *Accurate Difference Methods for linear ordinary differential systems subject to linear constraints*, SIAM J. Num. Anal. 6, 8-30 (1969).

_____, *Accurate Difference Methods for non linear Two-point Boundary Value Problems*, SIAM J. Num. Anal. 11, 305-320 (1974).

RUSSEL, R.D., *Collocation for Systems of Boundary Value Problems*, Numer. Math. 23, 119-133 (1974).

VARAH, J.M., *A Comparison of Some Numerical methods for two-point boundary value problems*, Math. Comp. 28, 743-755 (1974).

VELDHUIZEN, M. VAN, *Projection Methods and Ordinary Differential Equations*, part I, II report Vrije Universiteit (1975).

6. EINDIGE ELEMENTENMETHODE VOOR MAGNETOSTATISCHE PROBLEMEN

6.1. Probleemstelling

De EEM wordt al een tiental jaren niet meer alleen voor mechanische problemen gebruikt. Een nieuw toepassingsgebied vormen ook de magnetotstatische problemen (Munro, Ph.D. th. Cambridge 1972, M.V. Chair, Ph.D. th. McGill University of Montreal, Canada 1970, R. Denin & S. van der Meer, CERN report 67-7 (1967), Iselon Fahsina, CERN, 1975).

We beschouwen alleen een beperkte klasse tweedimensionale problemen en gaan na hoe de EEM daarvoor gebruikt kan worden. We zien daarbij af van enerzijds die aspecten die uit physische overwegingen nog niet in programmatuur te verwezenlijken zijn, anderzijds van die aspecten die niet essentieel zijn in de realisatie.

We zullen ons voornamelijk bezig houden met de praktische realisatie van de EEM zonder ons in het waarom (existentie- en eenduidigheidsbewijzen) te verdiepen. Dit laatste eenvoudig omdat, wegens de nietlineariteit van het probleem en ook het niet compact zijn, deze bewijzen niet zonder meer gegeven kunnen worden.

De operationele vorm van het probleem bestaat uit

(i) een gebied $\Omega \subset \mathbb{R}^2$;

(ii) de Maxwell-vergelijking rot H = J;

(iii) randvoorwaarden.

We veronderstellen Ω op een speciale manier in vierhoeken te verdelen en komen daar in §6.2 op terug.

De afbeelding $H : \mathbb{R}^3 \to \mathbb{R}^3$, de magnetische veldsterkte is 1 × differentieerbaar (i.h.a. niet continu differentieerbaar) in alle componenten.

De beperking tot tweedimensionale problemen betekent hier dat H maar twee componenten in Ω ongelijk nul heeft. De stroomdichtheid J is een vector $\perp \Omega$ en is stuksgewijs continu (meestal stuksgewijs constant). Verder geldt

$$B = \mu(H)H + B_0 \qquad \text{of} \qquad H = \mu(B)^{-1}(B-B_0),$$

waarbij B, de magnetische fluxdichtheid, weer een afbeelding is van $\mathbb{R}^3 \to \mathbb{R}^3$ met de component $\perp \Omega$ identiek nul, $\mu(H)$ is i.h.a. een pos.def. matrix.

We nemen $B_0 \equiv 0$ (dit is een essentiële vereenvoudiging, die permanent-magnetische problemen uitsluit) en $\mu(H)$ een scalaire vermenigvuldiging (ook een essentiële vereenvoudiging, die anisotropie uitsluit).

We definieren nu de vectorpotentiaal A door B = rot A. We nemen verder A als onbekende functie. Dit heeft het volgende voordeel. De functie H heeft twee componenten $\neq 0$ zodat, wanneer we deze als onbekende nemen we in wezen twee functies moeten bepalen. A heeft maar één component ongelijk nul, n.l. $\perp \Omega$. Dus hoeven we maar één onbekende functie te bepalen. De functie A is door de bovenstaande definitie niet eenduidig vastgelegd, immers voor iedere C geldt rot grad C $\equiv$ 0, dus als A de eigenschap B = rot A heeft dan heeft voor iedere C ook A + grad C deze eigenschap. Evenwel is dit geen bezwaar, we zijn namelijk geïnteresseerd in H eb B zodat we alleen de randvoorwaarden zo moeten kiezen dat er maar één A oplossing is voor het gestelde probleem.

We kiezen Ω altijd zo groot dat de rand "oneindig" is, dan mogen we op physische overwegingen $A \equiv 0$ zetten op de rand $\partial\Omega$ van Ω.

Als Maxwell-vergelijking vinden we dan

$$\text{rot}(\mu(A)^{-1}(\text{rot } A - B_0)) = J.$$

De energieintegraal die bij dit probleem hoort is

$$(6.1.1) \qquad E(A) = \iint_{\Omega} \left[\int_0^B (h,db) - (J,A) \right] d\Omega,$$

waarbij $h(b) = \mu(b)^{-1}(b-B_0)$. De vectoren J en A hebben dezelfde richting en we schrijven dus verder JA i.p.v. (J,A). Een heuristische afleiding van deze integraal en een bewijs voor de equivalentie van Maxwell's vergelijking met de minimumeigenschap van deze energieintegraal vindt men in POLAK [1974].

$$(6.1.2)\qquad \int_0^B (h,db) = \int_0^{B_u} h_u db_u + \int_0^{B_v} h_v db_v + \int_0^{B_w} h_w db_w .$$

6.2. Praktische realisatie

(i) We beschouwen de EEM voor drie verschillende coördinatensystemen, waarvan we de coördinaten aanduiden met u en v.

We hebben dan de mogelijkheden

- cartesisch, u = x, v = y;
- poolcoördinaten, u = r, v = θ;
- cylindercoördinaten, u = z, v = r.

We bespreken de realisatie zoveel mogelijk in termen van u en v en refereren alleen waar nodig naar de verschillen die ontstaan uit de verschillen in de coördinatenstelsels.

Even wel bekijken we eerst de Maxwell-vergelijking en de energieintegraal in de verschillende coördinatenstelsels om een beeld te krijgen van de vorm van deze formules (met $B_0 \equiv 0$ en μ een scalaire vermenigvuldiging).

In het cartesische geval vinden we

$$(6.2.1)\qquad \frac{\partial}{\partial x}\left(\mu(A)^{-1}\frac{\partial}{\partial x}A\right) + \frac{\partial}{\partial y}\left(\mu(A)^{-1}\frac{\partial}{\partial y}A\right) = -J .$$

Wanneer $\mu(A)$ voldoende eigenschappen bezit is dus voor dit probleem existentie, eenduidigheid en convergentie te beschouwen m.b.v. Ciarlet, Schultz en Varga [1969]. (Evenwel is μ meestal alleen stuksgewijs in $C^1(\Omega)$.)

In het geval van cylindercoördinaten vinden we, aangezien alleen $A_\theta \neq 0$,

$$(6.2.2)\qquad \frac{\partial}{\partial z}\left(\mu(A)^{-1}\frac{\partial}{\partial z}A\right) + \frac{\partial}{\partial r}\left(\mu(A)^{-1}\frac{1}{r}\frac{\partial}{\partial r}(rA)\right) = -J,$$

terwijl we voor poolcoördinaten vinden

$$(6.2.3)\qquad \frac{1}{r^2}\frac{\partial}{\partial \theta}\left(\mu(A)^{-1}\frac{\partial}{\partial \theta}A\right) + \frac{1}{r}\frac{\partial}{\partial r}\left(r\mu(A)^{-1}\frac{\partial}{\partial r}A\right) = -J.$$

De factoren $\frac{1}{r}$ veroorzaken bij voldoende symmetrie van de oplossing geen echte singulariteiten aangezien we analoog aan

$$A(r=0) = 0 \Rightarrow \lim_{r\to 0}\frac{1}{r}A = \left[\frac{\partial}{\partial r}A\right]_{r=0}$$

altijd in r = 0 als limiet afgeleiden naar r vinden.
We bekijken nu de energieintegraal, maar om een beeld te krijgen van de vorm van deze integraal alleen voor constante μ. We hebben dan

$$\iint_{\Omega} [\tfrac{1}{2}\mu^{-1}B^2 - JA]d\Omega = \iint_{\Omega} \tfrac{1}{2}\mu^{-1}(\text{rot } A)^2 - JA d\Omega.$$

Dus voor cartesische coördinaten

$$\iint_{\Omega} \left\{+\tfrac{1}{2}\mu^{-1}\left[\left(\frac{\partial A}{\partial y}\right)^2 + \left(\frac{\partial A}{\partial x}\right)^2\right] - JA\right\} dxdy,$$

voor cylindercoördinaten

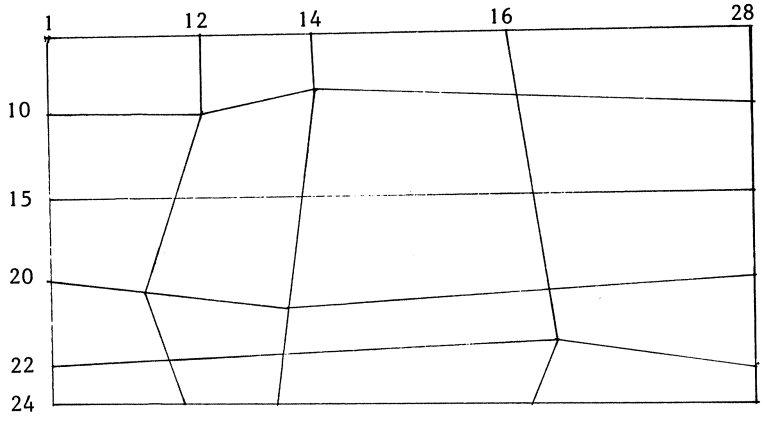
$$\iint_{\Omega} \left\{\tfrac{1}{2}\mu^{-1}\left[\left(\frac{\partial}{\partial z}A\right)^2 + \left(\frac{1}{r}\frac{\partial}{\partial r}(Ar)\right)^2\right] - JA\right\} rdrdz$$

en in poolcoördinaten

$$\iint_{\Omega} \left\{\tfrac{1}{2}\mu^{-1}\left[\left(\frac{1}{r}\right)^2\left(\frac{\partial}{\partial\theta}A\right)^2 + \left(\frac{\partial}{\partial r}A\right)^2\right] - JA\right\} rdrd\theta.$$

(ii) We veronderstellen Ω in vierhoeken te verdelen door twee verzamelingen snijdende lijnen als in figuur 6.1.

In de invoer van het programma geeft dit de mogelijkheid een maas op een eenvoudige manier aan te geven door een grove verdeling te definiëren die een fijne induceert.

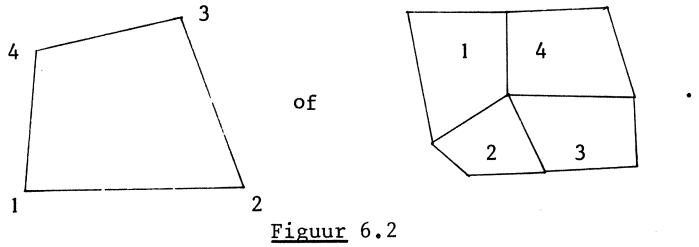


Figuur 6.1

We kunnen nu zowel de hoekpunten als de vierhoeken nummeren van boven naar beneden en van links naar rechts.

Stel er zijn I1 punten van boven naar beneden en J1 van links naar rechts. We hebben dan M1 = (I1-2)(J1-2) onbekende A waarden en M2 = (I1-1)(J1-1) vierhoeken.

We voeren nog twee zgn. locale nummeringen in d.w.z. de punten van een vierhoek nummeren we van 1 tot 4 waar uitsluitend over een vierhoek gesproken wordt en de vierhoeken rond een punt nummeren we waar uitsluitend over een punt gesproken wordt van 1 tot 4, dus



Figuur 6.2

(iii) We gebruiken isoparametrische elementen, d.w.z. iedere vierhoek (met hoekpunten (u_i,v_i), $i = 1,\ldots,4$) wordt door de transformatie

$$u = u_1 + (u_2-u_1)\xi + (u_3-u_1)\eta + (u_4-u_3-u_2-u_1)\xi\eta ,$$
$$v = v_1 + (v_2-v_1)\xi + (v_3-v_1)\eta + (v_4-v_3-v_2-v_1)\xi\eta$$

overgevoerd in $[0,1] \times [0,1]$. We nemen op $[0,1] \times [0,1]$ verder de basisfuncties

$$\psi_1(\xi,\eta) = (1-\xi)(1-\eta) ; \tag{6.2.4}$$
$$\psi_2(\xi,\eta) = \xi(1-\eta) ;$$
$$\psi_3(\xi,\eta) = \xi\eta ;$$
$$\psi_4(\xi,\eta) = (1-\xi)\eta .$$

Dit zijn de vier bilineaire functies met de eigenschap $\psi_i(u_j,v_j) = \delta_{ij}$ op de hoekpunten.

De vectorpotentiaal A wordt dus op een element benaderd door

$$A = \sum_{j=1}^{4} A_j \psi_j(\xi,\eta) \ .$$

Aangezien $A \perp \Omega$ geven we ook ψ_j deze richting. Dan is ook rot $\psi_j = (f_j, g_j, 0)$ gedefinieerd. Zolang de originele vierhoeken in het (u,v)-vlak alle hoeken kleiner dan 180° hebben is de transformatie (6.2.4) niet singulier (b.v. FIX & STRANG [1973], p.158). Ook worden de basisfuncties 1, u en v in de ruimte opgespannen door de $\psi_j(\xi(u,v), \eta(u,v))$ behouden, zodat convergentie in de (ξ,η)-ruimte convergentie in de (u,v)-ruimte impliceert. We hebben nu

$$B = \sum A_j \ (f_j, g_j).$$

De benadering $E(A_1,\ldots,A_{M1})$ met M1 = (I1-2)(J1-2) moet nu minimaal zijn voor de te berekenen vector $(A_1,\ldots,A_{M1})^T$.

Dus we vinden een stelsel niet lineaire vergelijkingen door

$$\frac{\partial}{\partial A_i} E(A_1,\ldots,A_{M1}) = 0, \qquad i = 1,\ldots,M1$$

te stellen. Voor dit stelsel berekenen we dan m.b.v. Newton-Raphson een benaderde oplossing. We moeten dus uiteindelijk

$$\frac{\partial^2}{\partial A_i \partial A_k} E(A_1,\ldots,A_{M1})$$

bepalen.

We bepalen de bijdragen aan de vergelijkingen weer elementsgewijs. Daarvoor realiseren we ons dat

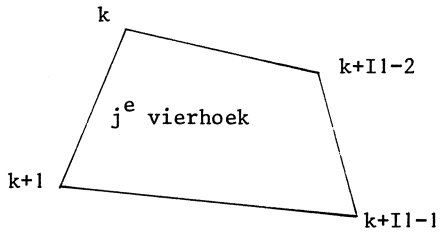
$$E = \sum_{j=1}^{M2} \Delta E_j \ ,$$

waarbij

$$E_j = \int_0^1 \int_0^1 \left[\int_0^B hdb - JA \right] F(\xi,\eta) d\xi d\eta,$$

waarbij $F(\xi,\eta) = G(\xi,\eta)Z(\xi,\eta)$ met $G(\xi,\eta)$ de Jacobiaan van de transformatie en $Z(\xi,\eta)$ ofwel 1, in het cartesische geval, ofwel $r(\xi,\eta)$ in de beide andere gevallen, dus de Jacobiaan van de transformatie $(u,v) \to (r,\theta)$ of $(u,v) \to (r,z)$.

De bijdragen van de vergelijkingen $\frac{\partial}{\partial A_i} E = 0$ in de j^e vierhoek zijn dus aan vier vergelijkingen, b.v. de k, k+1, k+J1-2 en k+I1-1-de vergelijking.



Figuur 6.3

Verder bevat $\frac{\partial}{\partial A_k} \Delta E_j$ dan uitsluitend de onbekenden k, k+1, k+J1-2 en k+I1-1. We merken ook op dat na de berekening van bijdragen uit deze j^e vierhoek de k^e vergelijking af is.

(iv) We bekijken nu $\frac{\partial}{\partial A_k} \Delta E_j$:

$$\frac{\partial}{\partial A_k} \Delta E_j = \int_0^1 \int_0^1 \frac{\partial}{\partial A_k} \left[\int_0^B hdb - JA \right] F(\xi,\eta) d\xi d\eta,$$

$$\frac{\partial}{\partial A_k} JA = J\psi_i, \qquad \text{stel} \qquad P_k = \int_0^1 \int_0^1 J\psi_i F d\xi d\eta.$$

We bekijken verder

$$\frac{\partial}{\partial A_k} \int_0^B hdb = \frac{\partial B}{\partial A_k} \frac{d}{dB} \int_0^B \mu^{-1}(b) bdb = \frac{\partial B}{\partial A_k} \mu(B)^{-1} B$$

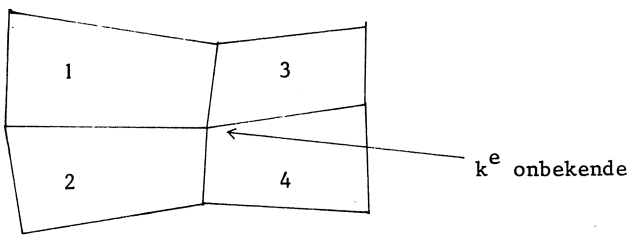
en we bekijken apart

$$\frac{\partial B}{\partial A_k} = \frac{\partial B_u}{\partial A_k} \frac{\partial B}{\partial B_u} + \frac{\partial B_v}{\partial A_k} \frac{\partial B}{\partial B_v} = \frac{1}{B}(f_k B_u + g_k B_v) = \frac{1}{B} \sum_{\ell=1}^{4} (f_k f_\ell + g_k g_\ell) A_\ell$$

waarbij $\ell = 1,\ldots,4$ een telling over de hoekpunten aanduidt en $B = \sqrt{B_u^2 + B_v^2}$. Dus

$$\frac{\partial}{\partial A_k} \Delta E_j = \int_0^1 \int_0^1 \sum_{\ell=1}^{4} (f_k f_\ell + g_k g_\ell) A_\ell \mu(B)^{-1} F d\xi d\eta - P_k = Q_{jk}.$$

Voor $\frac{\partial}{\partial A_k} E$ vinden we dus een som van vier zulke termen Q_{jk}.



Figuur 6.4

We veronderstellen μ langzaam variërend over een maas en gebruiken een Gaussische kwadratuur waarvan de orde bepaald wordt door $(f_k f_\ell + g_k g_\ell)F$, i.h.a. een 9-punts formule

(v) We bekijken nu $\frac{\partial^2}{\partial A_i \partial A_k} \Delta E_j$:

$$\frac{\partial^2 \Delta E_j}{\partial A_i \partial A_k} = \int_0^1 \int_0^1 \frac{\partial}{\partial A_k} \left[\mu(B)^{-1} \sum_{\ell=1}^{4} (f_i f_\ell + g_i g_\ell) A_\ell \right] F d\xi d\eta,$$

$$\frac{\partial}{\partial A_k} \left[\mu(B)^{-1} \sum_{\ell=1}^{4} (f_i f_\ell + g_i g_\ell) A_\ell \right] = \left[\frac{\partial}{\partial A_k} \mu(B)^{-1} \right] \sum_{\ell=1}^{4} A_\ell (f_\ell f_i + g_i g_\ell) +$$

$$+ \left[\mu(B) \right]^{-1} (f_i f_k + g_i g_k);$$

$$\frac{\partial}{\partial A_k} \mu(B)^{-1} = \frac{\partial B}{\partial A_k} * \frac{\partial \mu(B)^{-1}}{\partial B} = \frac{1}{B} \left(\sum_{\ell=1}^{4} (f_k f_\ell + g_k g_\ell) A_\ell \right) \frac{\partial \mu(B)^{-1}}{\partial B} .$$

Dus

$$\frac{\partial^2 \Delta E}{\partial A_i \partial A_k} = \int_0^1 \int_0^1 \left[\frac{1}{B} \frac{\partial \mu(B)^{-1}}{\partial B} \left(\sum_{\ell=1}^{4} (f_k f_\ell + g_k g_\ell) A_\ell \right) * \left(\sum_{n=1}^{4} (f_i f_n + g_i g_n) A_n \right) + \right.$$

$$\left. + \mu(B)^{-1} (f_i f_k + g_i g_k) \right] F d\xi d\eta.$$

(vi) We bekijken nu f en g in de verschillende coordinatenstelsels.
In cartesische coordinaten, u = x, v = y,

$$(6.2.5) \qquad f_\ell = \frac{\partial \psi_\ell}{\partial y} , \qquad g_\ell = - \frac{\partial \psi_\ell}{\partial x} ;$$

in cylindercoordinaten, u = z en v = r,

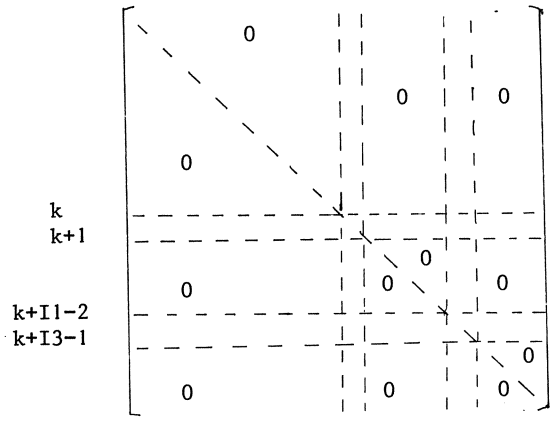
$$(6.2.6) \qquad f_\ell = \frac{\psi_\ell}{v} + \frac{\partial \psi_\ell}{\partial v} , \qquad g_\ell = - \frac{\partial \psi_\ell}{\partial u}$$

en voor u = r, v = θ

(6.2.7) $\quad f_\ell = \frac{1}{u}\frac{\partial \psi_\ell}{\partial v} \qquad$ en $\qquad g_\ell = -\frac{\partial \psi_\ell}{\partial u}$.

In de energieintegraal voor poolcoördinaten vinden we $\frac{1}{r}\frac{\partial}{\partial \theta}$ A. Aangezien $\frac{\partial}{\partial \theta}$ A(r=0) = 0 geldt weer $\lim_{r\to 0} \frac{1}{r}\frac{\partial}{\partial \theta} A = \frac{\partial^2}{\partial r \partial \theta} A(r=0)$. Evenwel in $f_\ell = \frac{1}{u}\frac{\partial \psi_\ell}{\partial r}$ zijn we deze eigenschap kwijt.

Het is zaak in de programma's ervoor te zorgen dat dit geen moeilijkheden geeft. We doen dit door op elementen langs de r = 0-as andere basisfuncties te kiezen met $\frac{\partial \psi_\ell}{\partial v}$ (u=0) = 0 enz.
Aangezien de Gauss-punten binnen de vierhoek liggen is daarmee praktisch het probleem opgelost. Evenwel moet opgemerkt dat dit voor convergentiebewijzen geen rol kan spelen aangezien H_1 niet volledig is t.o.v. deze voorwaarde m.a.w. een rij in H_1 met $\frac{\partial \psi_\ell}{\partial v}$ (u=0) = 0 hoeft geen limiet met deze eigenschap te hebben.
(vii) Wanneer we nu bekijken waar de bijdragen van $\frac{\partial^2 \Delta E_j}{\partial A_k \partial A_i}$ terecht komen in de Jacobiaan dan vinden we

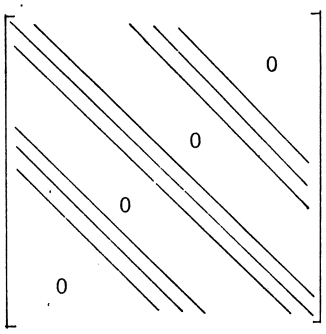


6.3. Programmatechnische overweging

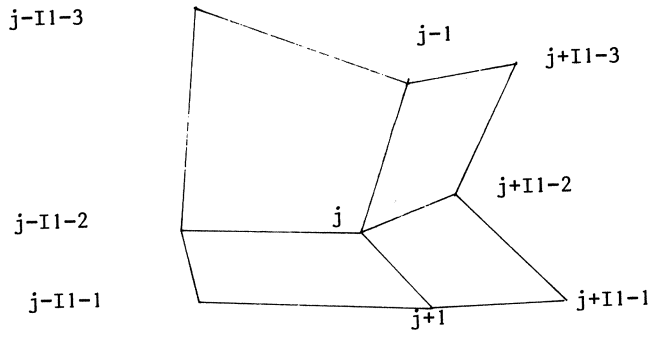
(i) Het stelsel vergelijkingen dat we op moeten lossen voor de Newton-Raphson heeft de volgende vorm

$$\mathbb{Z}a = c$$

met $a = (A_1,\ldots,A_{M1})^T$ en $\mathbf{Z}$ de vorm



waarbij $\mathbf{Z}$ positief definiet is. Bekijken we nu rij j van $\mathbf{Z}$. Deze komt van de $\frac{\partial^2}{\partial A_k \partial A_j} E$ voor k = 1,...,M1.
Evenwel is $\frac{\partial}{\partial A_j}$ alleen van een klein aantal A_j afhankelijk.



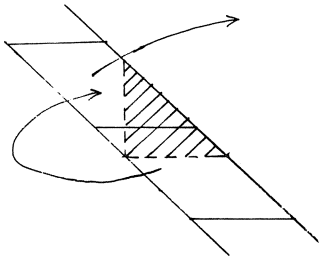
Figuur 6.5

Dus de j^e rij heeft de vorm

j^e rij: 0 ——— 0 *** 0 ——— 0 *** 0 ——— 0 *** 0 ——— 0.
kolom 1 j-I1-2 j j+I1-2 M1

(ii) Voor het opstellen en oplossen van dit stelsel gebruiken we wat we een "block frontal method" genoemd hebben.

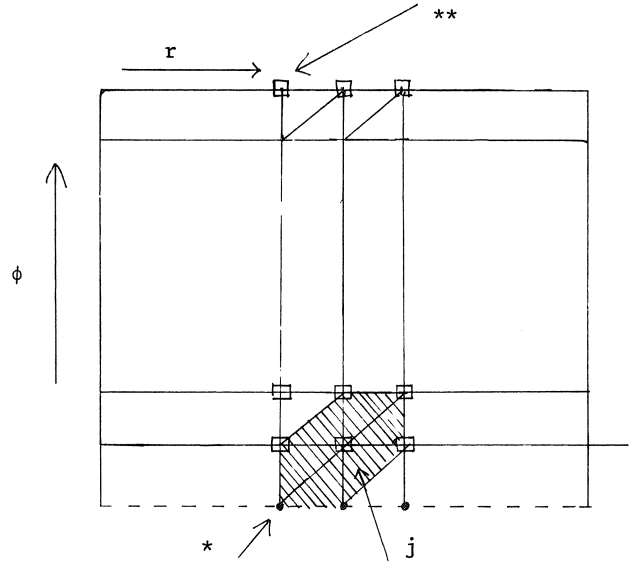
Om deze te begrijpen realiseren we ons dat bij het toepassen van Choleski voor een bandmatrix van dit type voor het berekenen van een nieuwe rij van de $L\,L^T$ decompositie we alleen de al berekende elementen uit een driehoek boven die rij gebruiken. Dus



We creëren dus een blok ter lengte van de bandbreedte, voeren dan m.b.v. het vorige blok de L L^T decompositie uit Ly = c tot het einde van dit blok. Dan schrijven we het voorlaatste blok naar disk en het laatste blok op de plaats an het voorlaatste etc. Dit betekent dat we voortdurend maar twee zulke blokken in het geheugen houden. Om te zien wat dit betekent voor de benodigde geheugenruimte het volgende voorbeeld. Bij een 50 × 50 maas hebben we voor de hele matrix, als we vier octaden per getal gebruiken, $50^3 * 4$ octaden nodig. Dus $500 \cdot 10^3$ octaden. Twee blokken zijn maar $2 * 4 * 50^2 = 20 \cdot 10^3$ octaden.

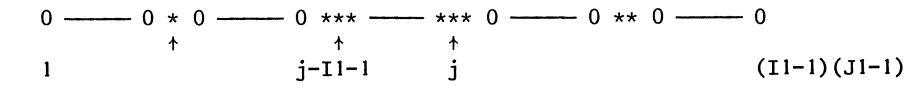
(iii) Een speciaal geval is, wegens de periodiciteitsvoorwaarde, u = r, v = θ.

Wanneer we n.l. gewoon de vergelijkingen opstellen, zoals we dat voor de andere coordinatenstelsels doen, dan vinden we het volgende probleem



Wanneer we de vergelijking voor punt j opstellen dan vinden we de onbekende

van het punt @. Deze onbekende hoort wegens de periodiciteit ook bij het punt @@. Wanneer we in de telling van de onbekenden deze nu bij @@ plaatsen, dan ziet de vergelijking j er zo uit:



De onbekende die bij @@ hoort "springt" in deze vergelijking uit de band. We kunnen daarvoor (minstens) op twee manneren een oplossing vinden.

De ene bestaat uit het verdelen van de vierhoeken in twee driehoeken. Dan kunnen we deze twee driehoeken met lineaire basisfuncties als elementen gebruiken. Dan doet de onbekende bij @@ niet in de j^e vergelijking mee.

Een tweede oplossing, waarbij de vierhoekige elementen behouden blijven is aanzienlijk ingewikkelder. We laten de bespreking daarvan hier achterwege.

6.4. Open problemen

Er zijn aan de zojuist beschreven toepassingen van de EEM twee klassen open problemen.

Enerzijds is de physicus niet in staat de μ's voor de verschillende materialen voldoende te definiëren. Anderzijds zijn convergentie-, existentie- en eenduidigheidsbewijzen voor een groot deel niet beschikbaar (voor zover mij bekend).

Natuurlijk, voor constante μ bij cartesische coördinaten zijn we in een bekende klasse waarvoor bij de gebruikte elementen alles bekend is. Voor niet constante μ is bij voldoende gladheid Ciarlet, Schultz en Varga's artikel [1969] bruikbaar. Meestal zijn evenwel de μ's alleen stuksgewijs "glad". Dan is de basistheorie misschien wel toepasbaar maar de daar gegeven resultaten niet zonder meer.

Voor pool-en cylindercoordinaten zijn bewijzen niet beschikbaar en waarschijnlijk ook moeilijk te geven met de gebruikelijke technieken

De eigenschappen die de vergelijkingen in de operationele vorm betekenis geven zijn niet van toepassing in de variationele aanpak. Bijvoorbeeld is H^2 volledig t.o.v. de eigenschap $\frac{\partial A}{\partial r}(r=0) = 0$, d.w.z. een rij functies met deze eigenschap heeft een limiet met deze eigenschap. Evenwel in H^1 is dit niet het geval zodat we deze eigenschap niet kunnen gebruiken.

LITERATUUR

[1] POLAK, S.J., *Some aspects in developing MAGGY ISCA-MAC-CAD*, UOV-DSA-SCA/74/016/RG, 1974.

[2] CIARLET, P.G., M.H. SCHULTZ & R.S. VARGA, *Numerical methods of high order accuracy for nonlinear boundary value problems*, Numerische Math. 13, 51-77 (1969).

[3] STRANG, G. & G. Fix, *An Analysis of the Finite Element Method*, Prentice Hall, Englewood Cliffs, N.J., 1973.

UITGAVEN IN DE SERIE MC SYLLABUS

Onderstaande uitgaven zijn verkrijgbaar bij het Mathematisch Centrum, 2e Boerhaavestraat 49 te Amsterdam-1005, tel. 020-947272.

MCS 1.1 F. GÖBEL & J. VAN DE LUNE, *Leergang Besliskunde, deel 1: Wiskundige basiskennis*, 1965. ISBN 90 6196 014 2.

MCS 1.2 J. HEMELRIJK & J. KRIENS, *Leergang Besliskunde, deel 2: Kansberekening*, 1965. ISBN 90 6196 015 0.

MCS 1.3 J. HEMELRIJK & J. KRIENS, *Leergang Besliskunde, deel 3: Statistiek*, 1966. ISBN 90 6196 016 9.

MCS 1.4 G. DE LEVE & W. MOLENAAR, *Leergang Besliskunde, deel 4: Markovketens, en wachttijden*, 1966. ISBN 90 6196 017 7.

MCS 1.5 J. KRIENS & G. DE LEVE, *Leergang Besliskunde, deel 5: Inleiding tot de mathematische besliskunde*, 1966. ISBN 90 6196 018 5.

MCS 1.6a B. DORHOUT & J. KRIENS, *Leergang Besliskunde, deel 6a: Wiskundige programmering 1*, 1968. ISBN 90 6196 032 0.

MCS 1.7a G. DE LEVE, *Leergang Besliskunde, deel 7a: Dynamische programmering 1*, 1968. ISBN 90 6196 033 9.

MCS 1.7b G. DE LEVE & H.C. TIJMS, *Leergang Besliskunde, deel 7b: Dynamische programmering 2*, 1970. ISBN 90 6196 055 X.

MCS 1.7c G. DE LEVE & H.C. TIJMS, *Leergang Besliskunde, deel 7c: Dynamische programmering 3*, 1971. ISBN 90 6196 066 5.

MCS 1.8 J. KRIENS, F. GÖBEL & W. MOLENAAR, *Leergang Besliskunde, deel 8: Minimaxmethode, netwerkplanning, simulatie*, 1968. ISBN 90 6196 034 7.

MCS 2.1 G.J.R. FÖRCH, P.J. VAN DER HOUWEN & R.P. VAN DE RIET, *Colloquium stabiliteit van differentieschema's, deel 1*, 1967. ISBN 90 6196 023 1.

MCS 2.2 L. DEKKER, T.J. DEKKER, P.J. VAN DER HOUWEN & M.N. SPIJKER, *Colloquium stabiliteit van differentieschema's, deel 2*, 1968. ISBN 90 6196 035 5.

MCS 3.1 H.A. LAUWERIER, *Randwaardeproblemen, deel 1*, 1967. ISBN 90 6196 024 X.

MCS 3.2 H.A. LAUWERIER, *Randwaardeproblemen, deel 2*, 1968. ISBN 90 6196 036 3.

MCS 3.3 H.A. LAUWERIER, *Randwaardeproblemen, deel 3*, 1968. ISBN 90 6196 043 6.

MCS 4 H.A. LAUWERIER, *Representaties van groepen*, 1968. ISBN 90 6196 037 1.

MCS 5 J.H. VAN LINT, J.J. SEIDEL & P.C. BAAYEN, *Colloquium discrete wiskunde*, 1968. ISBN 90 6196 044 4.

MCS 6 K.K. KOKSMA, *Cursus ALGOL 60*, 1969. ISBN 90 6196 045 2.

MCS 7.1 *Colloquium Moderne rekenmachines, deel 1*, 1969. ISBN 90 6196 046 0.

MCS 7.2 *Colloquium Moderne rekenmachines, deel 2*, 1969. ISBN 90 6196 047 9.

MCS 8 H. BAVINCK & J. GRASMAN, *Relaxatietrillingen*, 1969. ISBN 90 6196 056 8.

MCS 9.1 T.M.T. COOLEN, G.J.R. FÖRCH, E.M. DE JAGER & H.G.J. PIJLS, *Elliptische differentiaalvergelijkingen, deel 1*, 1970. ISBN 90 6196 048 7.

MCS 9.2 W.P. VAN DEN BRINK, T.M.T. COOLEN, B. DIJKHUIS, P.P.N. DE GROEN, P.J. VAN DER HOUWEN, E.M. DE JAGER, N.M. TEMME & R.J. DE VOGELAERE, *Colloquium Elliptische differentiaalvergelijkingen, deel 2*, 1970. ISBN 90 6196 049 5.

MCS 10 J. FABIUS & W.R. VAN ZWET, *Grondbegrippen van de waarschijnlijkheidsrekening*, 1970. ISBN 90 6196 057 6.

MCS 11 H. BART, M.A. KAASHOEK, H.G.J. PIJLS, W.J. DE SCHIPPER & J. DE VRIES, *Colloquium Halfalgebra's en positieve operatoren*, 1971. ISBN 90 6196 067 3.

MCS 12 T.J. DEKKER, *Numerieke algebra*, 1971. ISBN 90 6196 068 1.

MCS 13 F.E.J. KRUSEMAN ARETZ, *Programmeren voor rekenautomaten; De MC ALGOL 60 vertaler voor de EL X8*, 1971. ISBN 90 6196 069 X.

MCS 14 H. BAVINCK, W. GAUTSCHI & G.M. WILLEMS, *Colloquium Approximatietheorie*, 1971. ISBN 90 6196 070 3.

MCS 15.1 T.J. DEKKER, P.W. HEMKER & P.J. VAN DER HOUWEN, *Colloquium Stijve differentiaalvergelijkingen, deel 1*, 1972. ISBN 90 6196 078 9.

MCS 15.2 P.A. BEENTJES, K. DEKKER, H.C. HEMKER, S.P.N. VAN KAMPEN & G.M. WILLEMS, *Colloquium Stijve differentiaalvergelijkingen, deel 2*, 1973. ISBN 90 6196 079 7.

MCS 15.3 P.A. BEENTJES, K. DEKKER, P.W. HEMKER & M. VAN VELDHUIZEN, *Colloquium Stijve differentiaalvergelijkingen, deel 3*, 1975. ISBN 90 6196 118 1.

MCS 16.1 L. GEURTS, *Cursus Programmeren, deel 1: De elementen van het programmeren*, 1973. ISBN 90 6196 080 0.

MCS 16.2 L. GEURTS, *Cursus Programmeren, deel 2: De programmeertaal ALGOL 60*, 1973. ISBN 90 6196 087 8.

MCS 17.1 P.S. STOBBE, *Lineaire algebra, deel 1*, 1974. ISBN 90 6196 090 8.

MCS 17.2 P.S. STOBBE, *Lineaire algebra, deel 2*, 1974. ISBN 90 6196 091 6.

MCS 18 F. VAN DER BLIJ, H. FREUDENTHAL, J.J. DE IONGH, J.J. SEIDEL & A. VAN WIJNGAARDEN, *Een kwart eeuw wiskunde 1946-1971, Syllabus van de Vakantiecursus 1971*, 1974. ISBN 90 6196 092 4.

MCS 19 A. HORDIJK, R. POTHARST & J.TH. RUNNENBURG, *Optimaal stoppen van Markovketens*, 1974. ISBN 90 6196 093 2.

* MCS 20 T.M.T. COOLEN, P.W. HEMKER, P.J. VAN DER HOUWEN & E. SLAGT, *ALGOL 60 procedures voor begin- en randwaardeproblemen*, 1976. ISBN 90 6196 094 0.

MCS 21 J.W. DE BAKKER (red.), *Colloquium Programmacorrectheid*, 1974. ISBN 90 6196 103 3.

* MCS 22 R. HELMERS, F.H. RUYMGAART, M.C.A. VAN ZUYLEN & J. OOSTERHOFF, *Asymptotische methoden in de statistiek*, 1976. ISBN 90 6196 104 1.

* MCS 23.1 J.W. DE ROEVER (red.), *Colloquium Onderwerpen uit de biomathematica, deel 1*, 1976. ISBN 90 6196 105 X.

* MCS 23.2 J.W. DE ROEVER (red.), *Colloquium Onderwerpen uit de biomathematica, deel 2*, 1976. ISBN 90 6196 115 7.

MCS 24.1 P.J. VAN DER HOUWEN, *Numerieke integratie van differentiaalvergelijkingen, deel 1: Eenstapsmethoden*, 1974. ISBN 90 6196 106 8.

* MCS 25 L. GEURTS (red.), *Colloquium Structuur van programmeertalen*, 1976. ISBN 90 6196 116 5.

MCS 26.1 N.M. TEMME (red.), *Nonlinear Analysis, volume 1*, 1976. ISBN 90 6196 117 3.

* MCS 26.2 N.M. TEMME (red.), *Nonlinear Analysis, volume 2*, ISBN 90 6196 121 1.

* MCS 26.3 N.M. TEMME (red.), *Nonlinear Analysis, volume 3*, ISBN 90 6196 122 x.

MCS 27 M. BAKKER, P.W. HEMKER, P.J. VAN DER HOUWEN, S.J. POLAK & M. VAN VELDHUIZEN, *Colloquium Discreteringmethoden*, 1976. ISBN 90 6196 124 6.

De met een * gemerkte uitgaven moeten nog verschijnen.