CWI Syllabi

Centrum voor Wiskunde en Informatica
Centre for Mathematics and Computer Science
P.O. Box 4079, 1009 AB Amsterdam, The Netherlands

The CWI is a research institute of the Stichting Mathematisch Centrum, which was founded on February 11, 1946, as a nonprofit institution aiming at the promotion of mathematics, computer science, and their applications. It is sponsored by the Dutch Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O.).

CWI Syllabus 7

Vacantiecursus 1985
Variatierekening

Centrum voor Wiskunde en Informatica
Centre for Mathematics and Computer Science

ISBN 90 6196 291 9

INHOUD

*

INLEIDING

Variatierekening, in ruime zin opgevat, is de bestudering van *oneindig dimensionale extremaalproblemen*. Voor M een gegeven verzameling van elementen (meestal functies van tijd en/of plaatsvariabelen die aan zekere (neven-)voorwaarden voldoen), en een functie J op M, zijn het bewijzen van het bestaan, en het karakteriseren van dié elementen van M waarvoor J op M een lokaal minimale waarde (of, meer algemeen, een stationaire waarde) heeft belangrijke onderdelen van het bestuderen van zo'n extremaalprobleem.

De bestudering van dit soort problemen begon in de 17e eeuw toen men ontdekte dat veel problemen uit de Mathematische Fysica, i.h.b. de Klassieke Mechanica, beschreven kunnen worden als zo'n extremaalprobleem. Bijvoorbeeld, het *principe van Fermat* luidt dat de voortplanting van een lichtstraal in een optisch medium tussen twee punten plaatsvindt langs dié baan tussen de twee gegeven punten waarvoor geldt dat de benodigde tijd zo klein mogelijk is in vergelijking met de tijd die nodig is langs enig andere baan tussen de twee punten. Ook voor de beweging van massapunten onder invloed van conservatieve krachten werden al snel verschillende *variatie-principes* geformuleerd, en later gegeneraliseerd ter beschrijving van het statisch en dynamisch gedrag van continue media (vloeistoffen, elastica, etc.). Ook andere basiswetten van de theoretische natuurkunde werden later beschreven door, of worden juist afgeleid uit, variatie-principes (speciale en algemene relativiteitstheorie, de electro-magnetische velden theorie en de moderne ijkvelden theorieën zoals die welke leiden tot de Yang-Mills vergelijkingen). Behalve in de Mathematische Fysica komen extremaalproblemen van bovenbeschreven aard ook veelvuldig voor in de economie, speltheorie en meer algemeen in elk probleem waarin van *optimaal handelen* sprake is. Toepassingen in deze gebieden zijn in het bijzonder gedurende de laatste 50 jaar onderzocht.

In deze vakantiecursus kunnen slechts enkele aspecten van de variatierekening aan de orde komen. Van de veelheid van ideeën en resultaten op

dit gebied volgt hier een korte omschrijving van de gekozen onderdelen.
Na een inleiding over het ontstaan van de variatierekening, komen karakteristieke ideeën van zowel locale alsook van globale methoden aan bod.
De locale variatietheorie, onder aanname van differentieerbaarheidsvoorwaarden, omvat o.a. het afleiden van een *stationairiteitsvoorwaarde* waaraan een locaal minimaal element moet voldoen. In het eenvoudigste geval waarin M een lineaire ruimte van functies en J een dichtheidsfunctionaal is, is deze stationairiteitsvoorwaarde te herleiden tot de *Euler-Lagrange vergelijking*, in het algemeen een (stelsel) gewone- of partiële differentiaalvergelijkingen, eventueel vergezeld van natuurlijke randvoorwaarden.
In geval de elementen van M nog aan extra nevenvoorwaarden voldoen, leidt de stationairiteitsvoorwaarde tot veralgemenisering van de bekende *multiplikatorenmethoden van Lagrange* voor functies van een eindig aantal variabelen, of, in geval M een convexe verzameling is, tot een *variatieongelijkheid*. Door bestudering van het globale gedrag van J op M kan de *existentie* van oplossingen van het extremaalprobleem verkregen worden, bijvoorbeeld door aan te tonen dat J op M een eindige, minimale waarde aanneemt. Van zowel conceptueel, alsook van praktisch belang is het associëren van een zogenaamd *duaal probleem* met het oorspronkelijke, primaire, probleem. In het gewenste geval leidt bestudering van het, eenvoudiger, duale probleem tot uitspraken over het primaire probleem.
In het bijzonder voor problemen met nevenvoorwaarden worden deze ideeën behandeld en leiden tot een andere interpretatie van de multiplikatorenmethode. Een concrete toepassing hiervan is het *membraanprobleem* voor de stroming van een stof door een semi-permeabele wand.
Bestudering van het globale gedrag van de functie, en meer in het bijzonder van de topologische eigenschappen van de verzamelingen $\{x \in M \mid J(x) \leq c\}$, $c \in \mathbf{R}$, maakt het mogelijk stationaire punten van J op M (i.e. elementen van M die aan de stationairiteitsvoorwaarde voldoen) te vinden die *niet* corresponderen met locale minima of maxima van de functie. *Mini-max methoden* worden behandeld waarmee de existentie van zo'n "zadelpunt" kan worden bewezen.

***

Veel van de in de Mathematische Fysica voorkomende extremaalproblemen zijn te moeilijk voor directe bestudering of om daarvan expliciete oplossingen te vinden. Twee van de mogelijke reacties op deze constatering worden besproken. Eén daarvan is het zoeken naar een *benadering van het model* dat het oorspronkelijke probleem beschrijft. Benaderen van de oorspronkelijke functie door een eenvoudiger functie die hetzelfde kwalitatieve globale gedrag heeft, levert in veel gevallen een consistente benadering op voor de oorspronkelijke Euler-Lagrange vergelijking. Een andere methode is het *numeriek benaderen* van de oplossing. Dit impliceert dat het oneindig-dimensionale probleem benaderd wordt door een (reeks van) eindig dimensionale problemen. Het gebruik maken van de variationele structuur heeft geleid tot het ontwikkelen van een efficiënte numerieke methode, de methode van Ritz ofwel de *eindige elementen methode*. Na een uiteenzetting hiervan zal deze methode worden toegepast op het bovengenoemde membraanprobleem.

E.W.C. van GROESEN

HOOFDSTUK 1

HET VERHAAL VAN HET ONTSTAAN VAN DE VARIATIEREKENING: DE BIJDRAGEN VAN JAKOB BERNOULLI, JOHANN BERNOULLI EN LEONHARD EULER

T. KOETSIER

1. INLEIDING

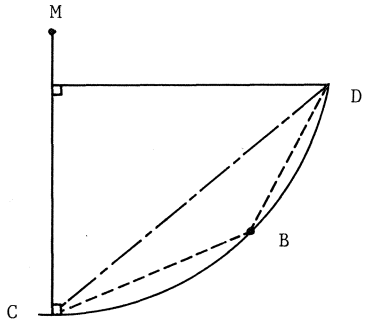
In 1744 verscheen van de hand van LEONHARD EULER (1707 - 1783) een belangrijk werk getiteld *Methodus inveniendi lineas curvas maximi minimive proprietate gaudentis* (Methode om krommen te vinden die maximum of minimum eigenschappen bezitten). In dat boek, dat als het eerste leerboek der variatierekening kan worden beschouwd, behandelt Euler een, rijk met toepassingen op allerlei soorten problemen geïllustreerde, theorie om vraagstukken van het volgende type op te lossen: Gevraagd een kromme y(x) met $a \leq x \leq b$ waarvoor de integraal

$$W = \int_a^b Z \, dx$$

een extreme waarde heeft. Daarbij is Z een uitdrukking waarin y(x) op verschillende manieren kan voorkomen en de verzameling toegelaten krommen y(x) kan op verschillende manieren door één of meer nevenvoorwaarden ingeperkt zijn.
Wij zullen in de volgende paragrafen de belangrijkste onwikkelingen schetsen die tot Euler's *Methodus* hebben geleid. Wij zullen bovendien laten zien hoe Euler in de *Methodus* de bekende nodige voorwaarden voor een extreem, de *vergelijkingen van Euler*, afleidt.

2. GALILEI, FERMAT

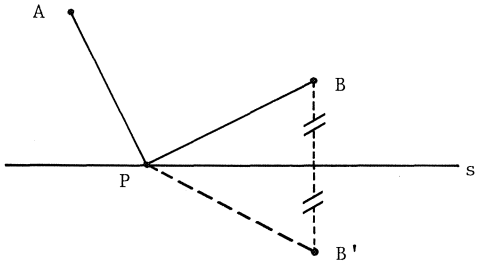
In zijn *Discorsi* van 1638 behandelt GALILEO GALILEI (1564 - 1642) zijn theorie van de valbeweging. Tot die theorie behoort de stelling dat bij wrijvingsloze val uit rust onder invloed van de zwaartekracht de snelheid van het vallend stoffelijk punt gelijk is aan een constante maal de wortel uit de afgelegde *hoogte*. Daarbij maakt het niet uit of de val loodrecht naar beneden plaats vindt of langs een hellend vlak: de snelheid is alleen afhankelijk van de afgelegde hoogte ([GALILEI, 1974], p. 174). In hetzelfde boek vergelijkt Galilei de tijd die een stoffelijk punt nodig heeft om van een punt D naar een punt C te vallen langs het lijnstuk DC (fig. 1) met de tijd die nodig is als de route DBC wordt gevolgd.



Figuur 1

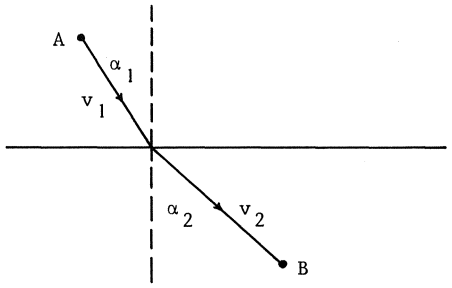
Daarbij is B een willekeurig punt op de boog DC van de cirkel door de punten D en C met middelpunt M loodrecht boven C. Galilei komt tot de conclusie dat de route DBC minder tijd vergt en hij leidt daar (ten onrechte) uit af dat de snelste valweg van D naar C cirkelboog DC is ([GALILEI, 1974], pp. 211 - 213).

Ruim twee decennia later wordt naar dit werk van Galilei verwezen door PIERRE FERMAT (1601 - 1665) in brieven aan DE LA CHAMBRE. In die brieven treffen wij een visie aan die teruggaat tot de antieke natuurfilosofie, inhoudende dat de natuur niets op onredelijke wijze of tevergeefs doet. Vaak heeft die visie de vorm dat de natuur er altijd naar streeft om effecten met zo eenvoudig mogelijke middelen tot stand te brengen. Illustraties hiervan waren in de oudheid de rechtlijnige weg van het licht en ook de stelling van Heron, die zegt dat ook bij weerkaatsing in een spiegel de door het licht afgelegde weg de kortst mogelijke is ([DIJKSTERHUIS & FORBES, 1961], p. 43). De juistheid van de stelling van Heron is onmiddellijk in te zien (fig. 2) omdat elke mogelijke route van het licht APB van A naar B de lengte AP + PB' heeft, waarbij B' het beeld van B is bij spiegeling in s.



Figuur 2

De kortste route van A naar B via de spiegel is klaarblijkelijk die waarbij P op de lijn door A en B' ligt en dat is het geval als de hoek van inval gelijk is aan de hoek van terugkaatsing. Fermat verdedigt in zijn brieven het standpunt dat de natuur niet zozeer altijd de *kortste* weg kiest, maar veeleer die wegen die het *gemakkelijkst en het snelst* zijn. De brieven van Fermat aan de la Chambre zijn belangrijk omdat Fermat op basis hiervan de wet van Snellius bewees. In feite stelde Fermat zichzelf het volgende probleem. Welke route van een punt A in een medium 1 naar een punt B in een medium 2 kost de minste tijd indien de grens tussen de media rechtlijnig is en met de media 1 en 2 resp. de snelheden v_1 en v_2 corresponderen (fig. 3).



Figuur 3

Men gaat gemakkelijk na dat het probleem neerkomt op het minimaliseren van een functie van één variabele. Het gelijk aan nul stellen van de afgeleide leidt tot de conclusie dat de snelste route die is waarvoor geldt dat

$$\frac{\sin \alpha_1}{v_1} = \frac{\sin \alpha_2}{v_2} \,. \tag{2.1}$$

Fermat deed het eigenlijk ook zo. Hij maakte gebruik van zijn "methode van maxima en minima", een voorloper van de differentiaalrekening (zie [GOLDSTINE, 1980], pp. 1 - 6).

3. DE BRACHISTOCHROON, JOHANN BERNOULLI

In de *Acta Eruditorum* van juni 1696 nodigde JOHANN BERNOULLI (1667 - 1748) de wiskundigen van zijn tijd uit om het volgende probleem op te lossen:

Als in een verticaal vlak twee punten A en B gegeven zijn, moet men voor het beweeglijke punt M een baan AMB aanwijzen, waarop het, van A uitgaande, door zijn zwaarte, in de kortste tijd in B aankomt.

Bij het stellen van het probleem van de *brachistochroon* (brachistos = kortste, chronos = tijd) was Johann er zich klaarblijkelijk niet van bewust dat Galilei zich er ook mee had beziggehouden. Het probleem moet ook gezien worden in de context van de ruim 10 jaar eerder door LEIBNIZ geïntroduceerde differentiaalrekening, die talloze nieuwe mogelijkheden opende. In een in januari 1697 te Groningen uitgegeven *Aankondiging* herhaalde Johann de uitdaging en voegde er aan toe dat natuurlijk uitgegaan diende te worden van de hypothese van Galilei dat de snelheden, die een vallend lichaam bereikt zich verhouden als de $2^{\underline{e}}$ machtswortels van de afgelegde hoogten.
Vrijwel alle grote wiskundigen van die tijd hebben het probleem om de brachistochroon te vinden opgelost: Leibniz, Newton, Johann Bernoulli, Jakob Bernoulli e.a.. De oplossingen van de broers Bernoulli verschenen beide in de Acta Eruditorum van mei 1697. De oplossing van Johann berust op de gedachte dat in een uit horizontale lagen van verschillende, goed gekozen dichtheid opgebouwd medium een lichtstraal eenzelfde snelheidsverloop kan hebben als een wrijvingsloos vallend stoffelijk punt. Het bovengenoemde resultaat van Fermat (in feite een speciaal brachistochroon-probleem), dat Johann kende, houdt in dat daarbij

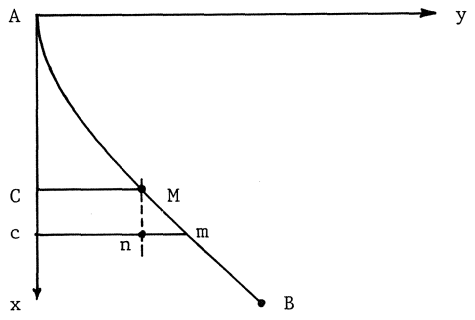
$$\frac{\sin \alpha}{v} = \text{constant.} \tag{3.1}$$

Johann Bernoulli redeneerde nu als volgt. Laat (fig. 4) AB de gezochte kromme zijn en laat Cc = dx, mn = dy en Mn = dz. Dan volgt met (3.1) dat

$$\frac{dy}{dz} = \frac{t}{c_1} \tag{3.2}$$

waarin t de snelheid in het punt M en c_1 een constante is.
Dan hebben we $c_1\, dy = t\, dz$ of

$$c_1^2\, dy^2 = t^2\, dz^2 = t^2\, dx^2 + t^2\, dy^2. \tag{3.3}$$



Figuur 4

Dat wil zeggen

$$(3.4) \qquad dy = \frac{t\ dx}{\sqrt{(c_1^2 - t^2)}} .$$

Met $t = c_2\sqrt{x}$ (hypothese van Galilei) volgt

$$(3.5) \qquad dy = \frac{c_2\ \sqrt{x}\ dx}{\sqrt{(c_1^2 - c_2^2 x)}}$$

of, met $c_1/c_2 = a$,

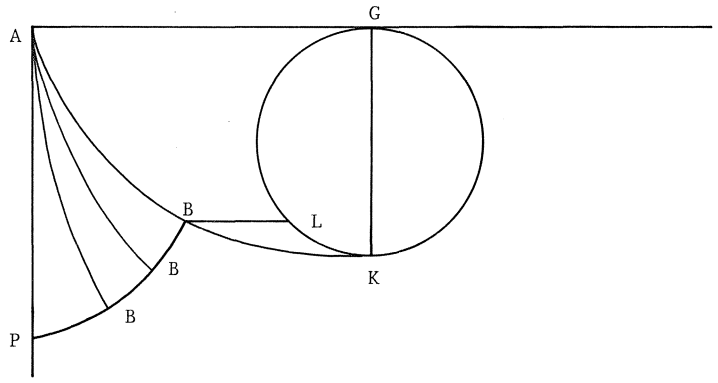
$$(3.6) \qquad dy = dx\ \sqrt{(\frac{x}{a - x})}.$$

Vervolgens toont Johann Bernoulli aan dat de door punt A gaande cycloide ontstaan door een cirkel met straat a op de y - as te laten rollen, aan deze differentiaal vergelijking voldoet. De lezer kan dit gemakkelijk zelf controleren.

Het artikel van Johann Bernoulli bevat meer dan alleen de oplossing van het probleem van de brachistochroon. Duidelijk is dat voor t een willekeurige functie $c_2 f(x)$ van x kan kiezen. De brachistochroon wordt dan door de volgende vergelijking beschreven

$$(3.7) \qquad dy = \frac{f(x)\ dx}{\sqrt{(a - (f(x))^2)}} .$$

Aan het eind van zijn artikel zegt Johann dat hem bij het schrijven van het voorafgaande nog een opmerkelijk resultaat is ingevallen. Men beschouwt (fig. 5) alle cycloiden met basis AG.

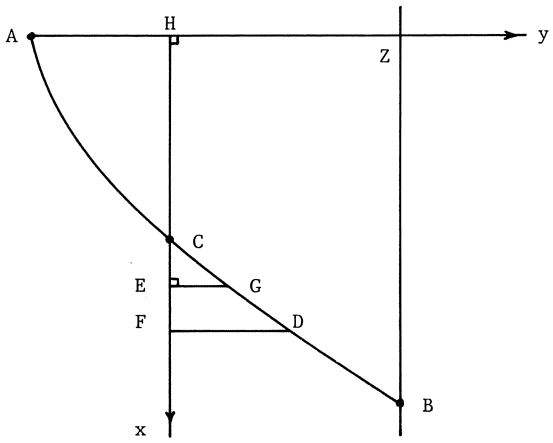


Figuur 5

De kromme PB (door Bernoulli *synchroon* genoemd) van alle punten, die door vanuit A langs die cycloiden vallende stoffelijke punten na een bepaalde tijd worden bereikt, laat zich puntsgewijs gemakkelijk "construeren". Als P en een middellijn GK gegeven zijn vindt men het snijpunt van de synchroon door P en de door de cirkel met middellijn GK bepaalde cycloide als volgt. Bepaal L zodanig (fig. 5) dat boog GL = $\sqrt{}$(AP.GK). Dan snijdt de horizontale lijn door L de cycloide in kwestie in het gezochte punt B. Bernoulli vermeldde dit resultaat zonder bewijs.

4. JAKOB BERNOULLI (1654 - 1705), ISOPERIMETRISCHE PROBLEMEN

De oplossing van Johann Bernoulli van het probleem van de brachistochroon is zeer elegant, maar heeft tegen de achtergrond van de verdere ontwikkelingen enigszins een ad hoc karakter. De oplossing van zijn broer Jakob is voor die verdere ontwikkelingen belangrijker geweest. Die oplossing verloopt als volgt. Laat ACDB de gezochte kromme zijn met C en D oneindig dicht bij elkaar (fig. 6). We beschouwen nu uitsluitend het stukje CD van de kromme (fig. 7) en vergelijken de tot de oplossingskromme behorende route CGD met een alternatieve route CLD, waarbij GL oneindig klein is t.o.v. EG (d.w.z. oneindig klein van de 2^{e} orde (N.B. CE = EF)).



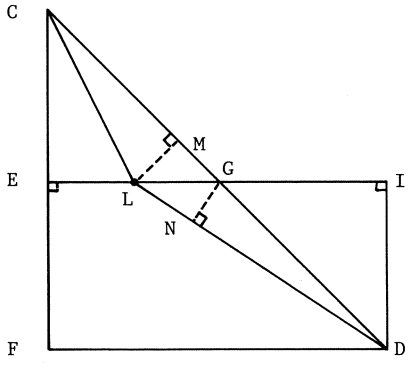
Figuur 6

Dan moet voor de tijden t_{CL}, t_{LD}, t_{CG} en t_{GD} die nodig zijn om resp. CL, LD, CG en GD te doorlopen, gelden dat

$$t_{CL} + t_{LD} = t_{CG} + t_{GD}$$

of anders geschreven,

(4.1) $$t_{CG} - t_{CL} = t_{LD} - t_{GD} .$$



Figuur 7

Vergelijken we de routes CE, CL en CG dan levert dat (op grond van de hypothese van Galilei)

$$\frac{CE}{CG} = \frac{t_{CE}}{t_{CG}} \quad \text{en} \quad \frac{CE}{CL} = \frac{t_{CE}}{t_{CL}}$$

of, anders geschreven,

$$\frac{CE}{CG - CL} = \frac{t_{CE}}{t_{CG} - t_{CL}} \cdot \tag{4.2}$$

Als nu LM ⊥ CG (fig. 7) dan is CL = CM (op 2^{e} orde termen na, die weggelaten kunnen worden) en Δ MLG ~ Δ ECG zodat

$$\frac{MG}{GL} = \frac{EG}{CG} \tag{4.3}$$

en dus volgt met MG = CG - CL uit (4.2) en (4. 3)

$$\frac{CE}{GL} = \frac{EG.t_{CE}}{CG.(t_{CG} - t_{CL})} \cdot \tag{4.4}$$

Op analoge wijze volgt door vergelijking van de routes EF, GD, LD dat

$$\frac{EF}{GL} = \frac{GJ.t_{EF}}{GD.(t_{LD} - t_{GD})} \cdot \tag{4.5}$$

Uit (4.1), (4.4) en (4.5) en CE = EF volgt

$$\frac{EG.t_{CE}}{GJ.t_{EF}} = \frac{CG}{GD} \cdot \tag{4.6}$$

Volgens de valwet geldt

$$\frac{t_{CE}}{t_{EF}} = \frac{\sqrt{HE}}{\sqrt{HC}} \tag{4.7}$$

en dus volgt uit (4.6) en (4.7) dat

$$\frac{EG}{\sqrt{HC}.CG} = \frac{GJ}{\sqrt{HE}.GD} \tag{4.8}$$

hetgeen we kunnen schrijven als

$$\frac{dx}{\sqrt{x}\ dz} = \frac{dx'}{\sqrt{x'}\ dz'} \tag{4.9}$$

hetgeen betekent dat tijdens de beweging langs de brachistochroon geldt dat

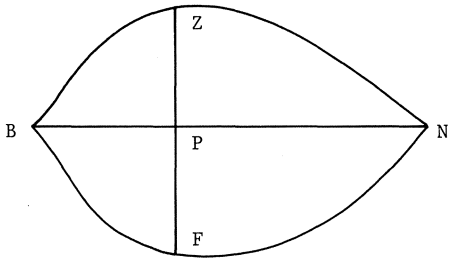
$$\frac{dx}{\sqrt{x}\,dz} = \text{constant}$$

waarmee Jakob precies dezelfde differentiaal-vergelijking had gevonden als zijn broer Johann. Ook Jakob besluit zijn oplossing met een bewijs dat de cycloide aan de differentiaal-vergelijking voldoet.

Het artikeltje waarin Jakob zijn oplossing van het brachistochroon-probleem levert, is getiteld: Oplossing van de opgaven van mijn broer, aan wie ik daarvoor andere voorleg.

De eerste opgave die Jakob aan Johann voorlegde is de volgende. Beschouw de verzameling van alle cycloiden (of ook cirkels, parabolen of andere krommen) die door A gaan en AH als basis hebben (zie fig. 6). De vraag is nu op welke van die krommen een uit A vertrekkend stoffelijk punt zo snel mogelijk de verticaal ZB bereikt.

Jakob schreef verder: In het bijzonder echter zou hij (dat is Johann), als hij wraak wil nemen, mogen proberen het volgende algemene probleem op te lossen. Onder alle isoperimetrische figuren op de gemeenschappelijke basis BN moet de kromme BFN bepaald worden, die weliswaar niet zelf de grootste oppervlakte omsluit, maar tot gevolg heeft dat een andere kromme BZN het doet, waarvan de ordinaat PZ evenredig is met een of andere macht of wortel van het lijnstuk PF of de boog BF (fig. 8).



Figuur 8

Isoperimetrische problemen vinden we al in de oudheid. *Virgilius* verhaalt in zijn *teneis* (gezang 1, 340 - 368) dat koningin *Dido* op de vlucht voor haar broer *Pygmalion* aan de Noord-Afrikaanse kust van de bewoners daar net zoveel land geschonken kreeg als zij met de huid van een stier kon omspannen. In het door *Justinus* geschreven uittreksel van *Pompeius Trogus*'

"Wereldgeschiedenis" (boek 18, §5.9) staat dat Dido beval om de huid in de fijnste delen te snijden om op die manier een groot stuk land in bezit te nemen. Zo zou Carthago zijn gesticht, 72 jaar vóór de stichting van Rome ([SZABÓ, 1977], p. 117).

Ook is er het verhaal van de heldendaad van *Horatius Cocles*, die in zijn eentje op de Etruskische oever van de Tiber de Etrusken tegenhield, totdat zijn metgezellen de brug achter hem hadden afgebroken. In volledige wapenrusting zwom hij vervolgens de Tiber over en werd aan Romeinse zijde beloond met zoveel land als waar hij op één dag met de ploeg omheen kon rijden. (Verhaald door *Titius Livius* in diens "Geschiedenis van Rome" ([SZABÓ, 1977], p. 118).)

Een tamelijk uitvoerige behandeling van de wiskundige problematiek op de achtergrond van deze verhalen vinden we in een verhandeling van *Zenodorus*, geschreven tussen ± 200 voor Christus en 90 voor Christus. Zo bewees hij de volgende drie stellingen:

(1) Van alle veelhoeken met hetzelfde aantal zijden en gelijke omtrek heeft de regelmatige veelhoek de grootste oppervlakte.

(2) Van alle regelmatige veelhoeken met gelijke omtrek heeft de veelhoek met de meeste hoeken de grootste oppervlakte.

(3) Een cirkel is groter dan elke regelmatige veelhoek met dezelfde omtrek ([HEATH, 1963], pp. 382 - 383).

Heath brengt de belangstelling van Griekse wiskundigen voor het isoperimetrische probleem in verband met foutieve opvattingen met betrekking tot oppervlakte en omtrek van figuren, waarover verschillende auteurs schrijven. Zo verhaalt Proclus volgens Heath, van leden van communes die grootmoedig stukken land met grote omtrek (en kleine oppervlakte) aan anderen lieten en zelf genoegen namen met een stuk land met kleine omtrek (en grote oppervlakte) en zodoende een reputatie van grote eerlijkheid wisten op te bouwen (Ibid.).

Jakob Bernoulli legde zijn broer, naast het al genoemde probleem, in feite twee gegeneraliseerde isoperimetrische problemen voor.

Wat moderner geformuleerd:

PROBLEEM 1. Gevraagd $y(x)$ met $a \leq x \leq b$, $y(a) = y(b) = 0$, zodanig dat

$$W = \int_a^b Z(y(x))\, dx$$

maximaal is, onder de nevenvoorwaarde dat

$$V = \int_a^b \sqrt{(1 + (y'(x))^2)}\, dx = \text{constant}.$$

PROBLEEM 2. Gevraagd $y(x)$ met $a \leq x \leq b$, $y(a) = 0$, zodanig dat

$$W = \int_a^b Z(s(x))\, dx$$

maximaal is, met $s(x) = \int_a^x \sqrt{(1 + (y'(x))^2)}\, dx$,

onder de nevenvoorwaarde dat

$$V = \int_a^b \sqrt{(1 + (y'(x))^2)}\, dx = \text{constant}.$$

Jakob beperkte zich in zijn vraagstelling tot functies Z met $Z(t) = t^n$, n geheel of gebroken.

Jakob beëindigde zijn artikel met de mededeling dat hij Johann drie maanden gaf om de uitdaging te aanvaarden en dat bij aanvaarding de oplossingen voor het eind van 1696 zouden moeten worden gepresenteerd. Jakob stelde Johann tevens een beloning van 50 dukaten in het vooruitzicht, die zouden worden betaald door een onbekende, een *non nemo*, waarvoor Jakob naar zijn zeggen borg stond.

5. HET FALEN VAN JOHANN BERNOULLI

In een brief aan Basnage, die in juni 1697 werd gepubliceerd in het *Journal des Savans* deelde Johann mede dat hij de door zijn broer gestelde problemen binnen drie minuten had opgelost. Het probleem betreffende de cycloiden kan worden opgelost met de synchronen, die hij tegelijk met zijn

oplossing van het brachistochroon-probleem had geïntroduceerd, zo deelde hij mede. In december van hetzelfde jaar verscheen in het Journal des Savans een brief van Johann aan Varignon, waarin hij, weliswaar zonder bewijs, zijn oplossingen gaf. In die brief gaf hij voor het isoperimetrische Probleem 1 als oplossing de functie gedefinieerd door

$$(5.1) \qquad x = \int \frac{y^n \, dy}{\sqrt{(a^{2n} - y^{2n})}}$$

voor het geval dat $Z(y) = y^n$ en voor het geval dat $Z(y) = f(y)$ willekeurig:

$$(5.2) \qquad x = \int \frac{b \, dy}{\sqrt{(a^2 - b^2)}}$$

met $b := \int \frac{f(y)}{y} \, dy$.

(Dit laatste zou hij later corrigeren tot $b := f(y)$.)

Het isoperimetrische Probleem 2 deed hij af met één zin, die er op neer komt dat het evident is hoe je in dat geval een differentiaal vergelijking kunt afleiden. In dezelfde brief aan Varignon deelde Johann mee dat hij zijn oplossingen aan Leibniz had gestuurd en dat Leibniz zich bereid had verklaard om de rol van arbiter op zich te nemen mits Jakob daarmee zou instemmen. Daarop ontstond de volgende situatie. Jakob aanvaardde klaarblijkelijk de arbitrage van Leibniz niet; hij wenste een publicatie van de volledige correcte oplossingen met bewijzen. Johann weigerde in feite zijn volledige oplossingen te publiceren. Begin 1701 zou hij ze door Varignon in een verzegelde envelop aan de Parijse Academie laten aanbieden met de bepaling dat het pak pas geopend mocht worden nadat Jakob diens oplossingen zou hebben vrijgegeven ([Joh. BERNOULLI, 1968 I], p. 424).

De verstandhouding tussen de beide broers, die toch al te wensen overliet bereikt nu weldra een dieptepunt. In februari 1698 verscheen in het Journal des Savans van Jakob een korte mededeling, waarin hij, reagerend op de brief van Johann aan Varignon, die in december 1697 was verschenen, mededeelde dat Johann's oplossing van het belangrijkste probleem, nl. dat betreffende de isoperimetrische figuren, niet helemaal correct is ([Joh. BERNOULLI, 1968 I], p. 214).

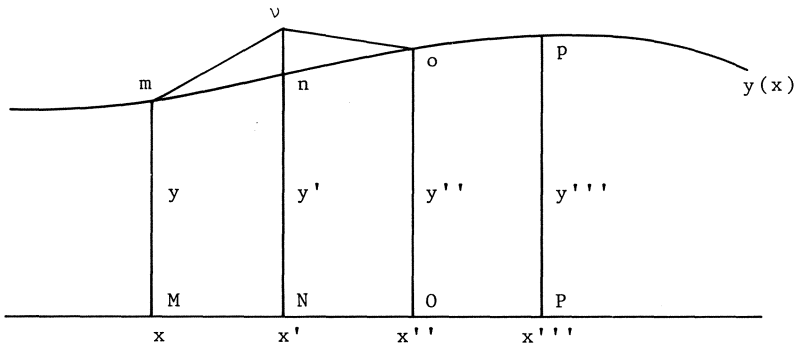
Johann reageerde in het Journal des Savans van 21 april 1698. Volgens hem had hij de problemen volstrekt bevredigend opgelost. Het probleem betreffende de cycloiden met behulp van zijn synchronen en de isoperimetrische problemen met behulp van zijn brachistochronen. Hij corrigeerde de naar zijn zeggen door haast ontstane fout in zijn oplossing van het isoperimetrische Probleem 1 (zie onder (5.2)) en merkte op dat hij eigenlijk alle problemen al had opgelost vóórdat zijn broer ze hem voorlegde. En inderdaad is de verzameling oplossingskrommen van isoperimetrisch Probleem 1 precies de verzameling brachistochronen die Johann had gevonden bij variatie van de manier waarop de snelheid van een vallend lichaam van de afgelegde hoogte afhangt (vergelijk (5.2) met (3.7)). Het is mogelijk dat Johann inmiddels in de gaten had gekregen dat hij zich had vergist in de moeilijkheid van isoperimetrisch Probleem 2. Hij was er heel kort over. Hij schreef zelfs dat de formulering van zijn broer zo gelezen kan worden dat slechts een oplossing van één van de twee isoperimetrische problemen werd gevraagd: de voegwoorden *vel*, *ve*, waarvan in de probleemstelling gebruik wordt gemaakt lijken van mij slechts de oplossing van het ene òf het andere probleem te verlangen ([Joh. BERNOULLI, 1968 I], p. 218).
De ruzie liep vervolgens en plein public hoog op. Jakob ging zelfs zover dat hij veronderstellingen ging uiten over de manier waarop Johann het isoperimetrisch Probleem 1 *zou* hebben opgelost en vervolgens fouten in de redenering aanwees. Johann zou zó hebben geredeneerd: Iedere mens is van steen; Iedere kei is mens; Dus iedere kei is van steen ([Joh. BERNOULLI, 1968 I], p. 227). M.a.w. Johann zou door middel van foute uitgangspunten tot een goed antwoord zijn gekomen. Jakob ging zover dat hij schreef: Ik heb nooit geloofd dat mijn broer de ware methode voor het isoperimetrische probleem bezat, maar nu twijfel ik er meer dan ooit aan. Johann reageerde met een open brief aan Jakob waarin hij zei: het uittreksel van de brief, die ik gisteren ontving, deden me inzien dat alle omzichtigheid en alle beleefdheden jegens hem niet hebben kunnen verhinderen [...] partij tegen mij te kiezen; en op een manier zo verhit en zo heftig dat er niemand meer is, die niet ziet dat het in plaats van de lofwaardige wedijver waarmee

ik mij vleide, slechts blinde jalouzie is die hem drijft [...] ; hersenschimmen, sterker en levendiger dan die van die zogenaamde heksen, die zich lichamelijk op de heksensabbat aanwezig waanden, hebben hem verleid; hij laat zich meeslepen door de stroom van zijn ijdele vermoedens; in één woord, hij lijkt niet meer voor rede vatbaar, zelfs niet meer in staat om alles wat ik hem over de kwestie zou kunnen zeggen te horen. Ik laat hem dus over aan zijn hartstochten en ik zal me ertoe beperken aan de lezer de absurditeit van zijn aanvallen te laten zien ([Joh. BERNOULLI, 1968 I], pp. 231 - 232).

Tegen de achtergrond van het voorafgaande is het niet vreemd dat, toen op 14 februari 1699 de beide broers tot buitenlands lid van de Parijse Academie werden benoemd, dat gebeurde "onder deze conditie dat men u beiden verzoekt om als er tussen u wetenschappelijke twistpunten zijn, fatsoenlijker te handelen dan u in het verleden hebt gedaan en dat beledigingen niet meer voor mogen komen" (geciteerd uit een brief, d.d. 16 februari 1699 van De L'Hopital aan Johann Bernoulli [Joh. BERNOULLI, 1955], p. 367).

Het conflict tussen de broers is nooit bijgelegd. De oplossingen van Jakob van de isoperimetrische Problemen 1 en 2 verschenen in twee artikelen in de Acta Eruditorum in 1700 en 1701 (zie e.g. [Joh. BERNOULLI, 1968 II], pp. 214 - 218 èn pp. 219 - 234). De oplossingen van Johann verschenen door onbekende omstandigheden pas in 1706 nà de dood van Jakob op 16 augustus 1705.
Uit die publicaties blijkt dat Jakob een methode bezat die tot een correcte oplossing van beide isoperimetrische problemen voert. Euler zou later voortbouwen op het centrale idee van die methode. Ook blijkt dat Johann géén bevredigende oplossing voor het isoperimetrische Probleem 2 had kunnen leveren.
Zowel Johann als Jakob pakten de isoperimetrische problemen aan in de geest van Jakob's oplossing van het brachistochroon-probleem: uitgangspunt is dat de maximum eigenschap ook locaal geldt, d.w.z. dat een locale variatie van de oplossingskromme een variatie 0 in de waarde van de te

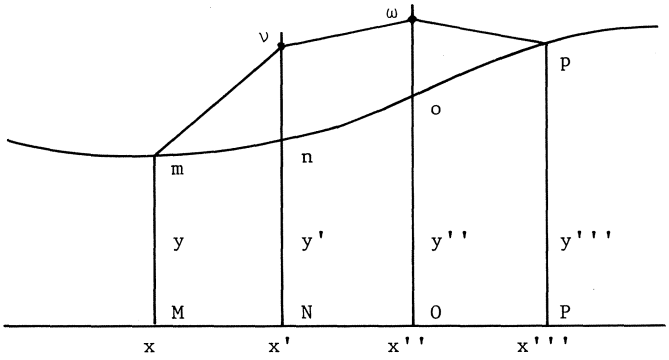
maximaliseren uitdrukking teweegbrengt en men tracht op die basis een differentiaal-vergelijking voor de oplossingskromme te vinden. Johann probeerde dat als volgt.



Figuur 9

Laat y(x) de gezochte kromme (fig. 9) zijn en x, x', x'', x''', etc. een *progressie* van infinitesimaal van elkaar verschillende waarden. Daarmee correspondeert dan een progressie y, y', y'', y''', etc.. Johann varieerde de kromme locaal door één punt te variëren: n wordt ν. De nevenconditie houdt dan in dat mn + no = mν + νo, m.a.w. dat ν op de ellips door n met brandpunten m en o moet liggen.

Jakob ging ook uit van de gezochte kromme y(x) (fig. 10).



Figuur 10

Hij varieerde de kromme echter locaal door *twee* punten verticaal te varieren: n wordt ν en o wordt ω. De nevenconditie houdt dan in dat mn + no + + op = mν + νω + ωp. Met deze aanpak slaagde Jakob er in om de beide isoperimetrische problemen op te lossen, terwijl Johann met diens methode

vastliep bij isoperimetrisch Probleem 2.

Omdat het werk van Jakob nogal wat op zichzelf bewonderenswaardig maar saai rekenwerk met zich meebrengt, zullen we dat hier niet bespreken. Evenmin bespreken we een artikel uit 1718 van zijn broer Johann waarin hij de ideeën van zijn 13 jaar daarvoor overleden broer Jakob in gestroomlijnde vorm behandelde. Voor een uitvoerige bespreking kan men [GOLDSTINE, 1980] raadplegen. De moeite van het lezen waard is ook [CARATHEODORY, 1945]. Caratheordory betoogt daarin dat ook bij isoperimetrisch Probleem 1, ondanks het goede antwoord, de redenering van Johann niet deugt.
In zijn artikel van 1718 gaf Johann Bernoulli in feite zijn ongelijk met betrekking tot isoperimetrisch Probleem 2 toe. Hij schreef dat hij vergeten had "om aandacht te schenken aan een zekere omstandigheid, die verhindert dat zij (i.e. zijn methode) zonder enige modificatie zou kunnen worden toegepast" ([Joh. BERNOULLI, 1968 II], p. 237) op isoperimetrisch Probleem 2.
In de volgende paragraaf zullen we zien welke vorm de methode van Jakob Bernoulli in de handen van Euler kreeg.

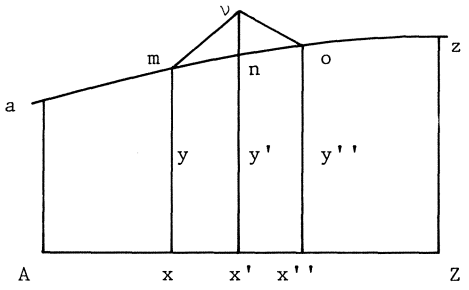
6. EULER's METHODUS

Zoals wij in de inleiding hebben opgemerkt hebben Euler's beschouwingen in de *Methodus* betrekking op vraagstukken van het volgende type: Gevraagd een kromme y(x) met $a \leq x \leq b$ waarvoor de integraal

$$W = \int_a^b Z \, dx \tag{6.1}$$

een extreme waarde heeft.
Indien er geen nevenvoorwaarden zijn dan hanteert Euler wat hij de *absolute methode van de maxima en de minima* noemt. We zullen laten zien wat Euler daaronder verstaat in het geval dat $Z = Z(x,y,p)$ met $dy = p\,dx$, m.a.w als Z alleen van x, y en $\frac{dy}{dx}$ afhangt.

Laat het interval [a,b] corresponderen met AZ (fig. 11). (N.B. We volgen, ook wat de notatie betreft, Euler op de voet.)



Figuur 11

Laat amnoz de gezochte kromme zijn en laat met de progressie x, x', x'' etc. de progressies y, y', y'' etc., p, p', p'' etc., Z, Z', Z'' etc. corresponderen. Omdat amnoz de gezochte kromme is, zal de waarde van W niet veranderen als we amnoz vervangen door de ermee *oneindig weinig* afwijkende kromme amνoz. Nu kunnen we W opvatten als de som

(6.2) $$W = \ldots + Z\,dx + Z'\,dx + Z''\,dx + \ldots .$$

De progressie x, x', x'' etc. is zodanig gekozen dat opeenvolgende waarden gelijke onderlinge afstand du hebben. Laten we nu de ordinaat Nn aangroeien met nν dan veranderen ..., x, x', x'' etc. niet; ..., y, y', y'' etc. veranderen in respectievelijk ..., y, y' + nν, y'' etc.; en ..., p, p', p'' etc. veranderen in respectievelijk ..., $p + \frac{n\nu}{dx}$, $p' - \frac{n\nu}{dx}$, p'' etc..

In (6.2) veranderen alleen Z dx en Z' dx en wel als volgt.

Gebruik makend van de totale differentiaal

(6.3) $$dZ = M\,dx + N\,dy + P\,dp$$

nemen Z dx en Z' dx toe met respectievelijk

$$dx\ P\ \frac{\nu n}{dx} \qquad \text{en} \qquad dx\ (N'.n\nu - P'\ \frac{n\nu}{dx}).$$

De aangroeiing van W is

(6.4) $$n\nu(N'\,dx - \frac{P' - P}{dx}\,dx)$$

ofwel

$$(6.5) \qquad \nu n(N - \frac{dP}{dx})\, dx$$

omdat N = N', afgezien van een kleine hogere orde term die kan worden weggelaten. Hieruit volgt onmiddellijk dat

$$(6.6) \qquad N - \frac{dP}{dx} = 0$$

of, modern opgeschreven,

$$(6.7) \qquad \frac{\partial Z}{\partial y} - \frac{d}{dx}(\frac{\partial Z}{\partial p}) = 0$$

een nodige voorwaarde is opdat y(x) met een extreem van W correspondeert. Vergelijking (6.6) vinden we in [EULER, 1744], p. 54, waar Euler de indruk wekt de voorwaarde ook als voldoende te beschouwen.
Het ligt voor de hand dat één van de eerste toepassingen die Euler van (6.7) geeft het brachistochroon-probleem betreft ([EULER, 1744], pp. 60 - 61). Het assenstelsel van fig. 4 hanterend, wordt y(x) gezocht, waarvoor

$$W = \int_0^{x_B} \frac{dx\ \sqrt{(1 + p^2)}}{\sqrt{x}}$$

minimaal is. (6.7) leidt onmiddellijk tot (2.7).

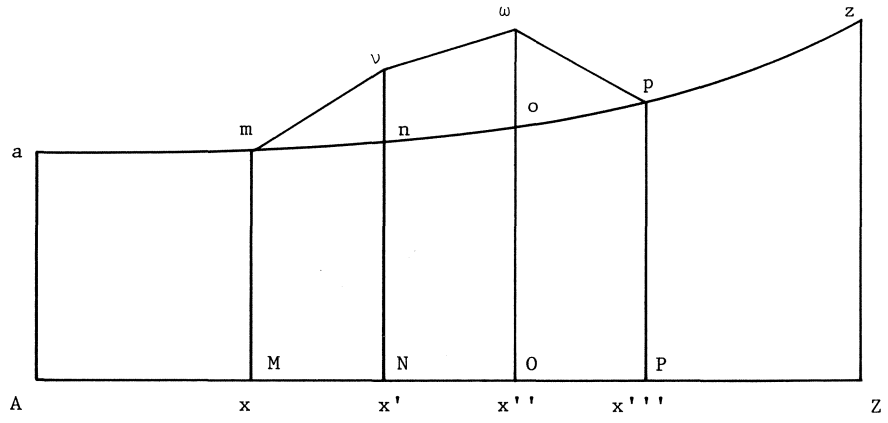
In het geval dat er nevenvoorwaarden zijn hanteert Euler de *relatieve methode van de maxima en minima*. Zij is toepasbaar bijv. op vragen van het type: Voor welke kromme y(x) met $a \leq x \leq b$ heeft

$$(6.8) \qquad W = \int_a^b Z\, dx$$

een extreme waarde onder de nevenvoorwaarde dat

$$(6.9) \qquad V = \int_a^b G\, dx = \text{constant}.$$

Euler redeneert als volgt. Laat amnopz de gezochte kromme zijn (fig. 12). De ordinaten Nn en Oo laten we nu aangroeien met resp. nν en oω.



Figuur 12

De aangroeiing nν alleen heeft tot gevolg dat W aangroeit met

(6.10) $n\nu.dA$

waarvoor geldt, als Z alleen afhankelijk is van x, y en p, dat, zoals we hebben gezien

(6.11) $dA = (N - \frac{dP}{dx})\, dx.$

De aangroeiing oω alleen heeft tot gevolg dat W aangroeit met

(6.12) $o\omega.dA'$

waarvoor geldt, als Z alleen afhankelijk is van x, y en p, dat

(6.13) $dA' = (N' - \frac{dP'}{dx})\, dx.$

De totale aangroeiing van W ten gevolge van de aangroeiingen nν en oω is dan

(6.14) $n\nu\, dA + o\omega\, dA'.$

Euler gaat hierop uitvoerig in ([EULER, 1744], pp. 95 - 98).
Voor de aangroeiing van V tengevolge van de aangroeiingen nν en oω vinden we een soortgelijke uitdrukking

(6.15) $n\nu\, dB + o\omega\, dB'.$

Opdat W extreem is onder de nevenvoorwaarde dat V = constant, moet gelden

$$(6.16)\qquad \begin{cases} n\nu\ dA + o\omega\ dA' = 0 \\ n\nu\ dB + o\omega\ dB' = 0 \end{cases}$$

waaruit volgt dat

$$(6.17)\qquad \frac{dA}{dB} = \text{constant}.$$

Euler leidt (6.17) op twee verschillende manieren uit (6.16) af ([EULER, 1744], pp. 99 - 100). Het snelst gaat het zo. Uit (6.16) volgt

$$\frac{dA'}{dB'} = \frac{dA}{dB}$$

en daaruit volgt onmiddellijk (6.17).

Indien Z en G beide slechts van x, y en p afhangen, kunnen we (6.17) wat moderner zo opschrijven

$$(6.18)\qquad \frac{\partial Z}{\partial y} - \frac{d}{dx}\left(\frac{\partial Z}{\partial p}\right) = c.\left\{\frac{\partial G}{\partial y} - \frac{d}{dx}\left(\frac{\partial G}{\partial p}\right)\right\}.$$

Een van de toepassingen die Euler van (6.18) geeft betreft isoperimetrisch Probleem 1: "Onder alle krommen az van dezelfde lengte is degene te bepalen, die bij draaiing om de as AZ het grootste volume teweegbrengt" ([EULER, 1744], p. 109). Duidelijk is dat gezocht wordt naar de kromme y(x) waarvoor

$$W = \int_a^b y^2\ dx$$

maximaal is onder de nevenvoorwaarde dat

$$V = \int_a^b \{1 + (y'(x))^2\}^{\frac{1}{2}}\ dx = \text{constant}.$$

(6.18) levert dan

$$(6.19)\qquad \frac{2y\ dx}{-\frac{d}{dx}\left(\frac{p}{\sqrt{(1 + p^2)}}\right) dx} = c_1$$

waaruit na kruislings vermenigvuldigen, vermenigvuldigen met p en integratie volgt

$$(6.20) \qquad y^2 + c_2 = \frac{c_1}{\sqrt{(1 + p^2)}} .$$

Oplossen naar p en integratie geeft

$$(6.21) \qquad x = \int \frac{y^2 + c_2}{\{c_1^2 - (y^2 + c_2)^2\}^{\frac{1}{2}}} \, dy$$

(zie [EULER, 1744], p. 110).

OPMERKING 1

Het is duidelijk dat ook het meer algemene isoperimetrische Probleem 1 op deze wijze kan worden opgelost. Als $W = \int_a^b f(y) \, dx$ dan vinden we de oplossing

$$(6.22) \qquad x = \int \frac{f(y) + c_2}{c_1^2 - (f(y) + c_2)^2} \, dy.$$

(Vergelijk (5.2).)

OPMERKING 2

Euler merkt op dat de kromtestraal van de oplossingskromme, in het algemeen gelijk aan

$$(6.23) \qquad \frac{dx}{d\frac{p}{\sqrt{(1 + p^2)}}} ,$$

hier gelijk is aan $\frac{c_1}{2y}$, m.a.w. omgekeerd evenredig is met de ordinaat y, waaruit volgt dat (6.21) de "curva elastica" is, d.w.z. de kromme die de vorm beschrijft van een onder invloed van een, op één uiteinde aangrijpende kracht, gebogen homogene elastische staaf.

In de *Methodus* lost Euler ook het isoperimetrische Probleem 2 op. Wij zullen die oplossing hier niet bespreken. We verwijzen de lezer naar de oorspronkelijke latijnse versie van de Methodus (L. Euler, Opera Omnia, Serie I, Vol. 24, pp. 198 - 200). In het eenvoudigste geval dat $W = \int_a^b s(x) \, dx$ vindt Euler dat de oplossingskromme de kettinglijn is.

7. HET PRINCIPE VAN DE KLEINSTE ACTIE

Zoals we in paragraaf 2 hebben gezien speelt het principe dat de natuur altijd de eenvoudigste weg kiest een rol in de voorgeschiedenis van de variatierekening. Echter, ook in de 18^{e} eeuw is dat idee invloedrijk. Dan wordt ook het probleem van de precieze formulering van het principe actueel.
De term "principe van de kleinste actie" ("le principe de la moindre quantité d'action") is afkomstig van *Pierre de Maupertuis* (1698 - 1759). Maupertuis gaf de volgende definitie: De natuur werkt zodanig dat m.v.s minimaal is (m is massa, v is snelheid en s de afgelegde weg).
Mach schreef: "Grootmoedig veranderde Euler de naam van het principe niet, liet de roem van de uitvinding aan Maupertuis, maar veranderde het in iets nieuws en werkelijk nuttigs" ([MACH, 1960], p. 550).
Euler's visie blijkt duidelijk uit de inleiding van "Additamentum I" van de Methodus. Hij schreef: "Omdat de constructie van het heelal zo volmaakt mogelijk is, zijnde het werk van een absoluut alwetende Schepper, zo geschiedt niets in de wereld, waarin geen maximum- of minimum-eigenschap blijkt". Het is volgens Euler niet eenvoudig die maximum- of minimum-eigenschappen a priori uit de principes van de metafysica te definiëren. De variatierekening stelde Euler echter in staat om a posteriori zulke eigenschappen te formuleren. We noemen er hier één. In "Additamentum II" van de Methodus komt Euler o.a. tot de conclusie dat, indien een stoffelijk met massa m zich beweegt in een "krachtveld" waarin de kracht uitsluitend van de plaats afhangt, de baan van het stoffelijk punt zodanig is dat de integraal

$$\int mv\, ds = \int mv^2\, dt \tag{7.1}$$

minimaal is. Aan Euler kan dan ook de ontdekking van het principe van kleinste actie voor vlakke beweging in een conservatief krachtveld worden toegeschreven.

LITERATUUR

Joh. BERNOULLI, (1955), *Der Briefwechsel von Johann Bernoulli*, Band I, Basel.

Joh. BERNOULLI, (1968 I), *Opera Omnia I*, Hildesheim.

Joh. BERNOULLI, (1968 II), *Opera Omnia II*, Hildesheim.

P. BRUNET, (1938), *Étude historique sur le principe de la moindre action*, Paris.

C. CARATHEODORY, (1945), *Basel und der Begin der Variationsrechnung*; in C. Caratheodory, Gesammelte Mathematische Schriften, Band II, München, pp. 108 - 128.

E.J. DIJKSTERHUIS & R.J. FORBES, (1961), *Overwinning door gehoorzaamheid II, van Newton tot Lorentz*, Zeist.

L. EULER, (1744), *Methode Curven zu finden denen eine Eigenschaft im höchsten oder geringsten Grade zukommt*, (gedeeltelijke vertaling van de Methodus), pp. 21 - 143 in [Stäckel, 1894].

G. GALILEI, (1974), *Two New Sciences*, Madison.

H.H. GOLDSTINE, (1980), *A History of the Calculus of Variations from the 17 - th through the 19 - th Century*, New York.

Th.L. HEATH, (1963), *Greek Mathematics*, New York (Dover Edition).

E. MACH, (1960), *The Science of Mechanics: a Critical and Historical Account of its Development*, Illinois.

D.J. STRUIK, (1969), *A Source Book in Mathematics, 1200 - 1800*, Massachusetts.

I. SZABÓ, (1977), *Geschichte der mechanischen Prinzipien*, Basel.

P. STÄCKEL (e.a.), (1894), *Abhandlungen über Variationsrechnung*, Erster Teil: Abhandlungen von Joh. Bernoulli (1696), Jac. Bernoulli (1697) und Leonhard Euler (1744), Leipzig. (Ostwald's Klassiker der Exakten Wissenschaften, Nr 46.)

HOOFDSTUK 2

ASPECTEN VAN VARIATIEREKENING

E.W.C. van GROESEN

1. INLEIDING

In deze bijdrage worden enkele van de basisideeën van de variatierekening behandeld. De keuze van onderwerpen is voor een deel bepaald door de noodzaak de behandeling zo beperkt mogelijk te houden. Anderzijds moesten enkele speciale, voor de volgende voordrachten voorbereidende, onderwerpen aan bod komen.
De tekst omvat meer dan in een uur behandeld kan worden.

Van oorsprong is de variatierekening ontstaan door de ontdekking dat veel problemen in de Mathematische Fysica een "variationeel" karakter hebben. Voor sommige van deze problemen betekent dit dat zij geformuleerd kunnen worden als een minimaliseringsprobleem waarvan de oplossingen overeenkomen met de oplossingen van het fysische probleem. Voor andere problemen betekent het dat de oplossingen overeenkomen met de oplossingen van een meer algemeen variatieprobleem, in casu met de zogenaamde stationaire punten van een functionaal.
Om dit te verduidelijken geven we eerst de algemene formulering, en karakteristieke probleemstelling, van mimimaliserings- en variatie-problemen; daarna geven we verschillende voorbeelden.

1.1. MINIMALISERINGSPROBLEMEN

De ingrediënten van een minimaliseringsprobleem (afgekort: min. pr.) zijn:

$$\begin{aligned} &V \text{ een lineaire ruimte,} \\ &M \subset V \text{ een verzameling toegestane elementen,} \\ &J: M \to \mathbb{R} \text{ een functionaal.} \end{aligned} \tag{1.1}$$

V is in de meeste toepassingen een oneindig-dimensionale ruimte van functies u; zo'n functie u zal dan de "toestand" van een fysisch systeem, of de evolutie daarvan, representeren. De deelverzameling M bestaat in die gevallen uit functies die nog aan extra *nevenvoorwaarden* voldoen, zoals bijvoorbeeld randvoorwaarden, positiviteit of zekere integraalvoorwaarden. De functionaal J is de te minimaliseren doelfunctie; J(u) heeft in veel gevallen een specifieke fysische betekenis, bijvoorbeeld de totale energie van

het systeem in toestand u.

Onder het *minimaliseringsprobleem voor* J *op* M, dat genoteerd wordt als

$$\inf_{u \in M} J(u) \quad \text{of als} \quad \inf \{J(u) \mid u \in M\} \tag{1.2}$$

verstaan we ruwweg het zoeken van oplossingen, dat is, per definitie, het zoeken van elementen $\hat{u} \in V$ waarvoor

$$\hat{u} \in M \text{ en} \tag{1.3}$$

$$J(\hat{u}) = \inf_{u \in M} J(u)$$

zodat dus $J(\hat{u}) \leq J(u)$ voor alle $u \in M$.

Meer precies geformuleerd behoren de volgende vragen tot het onderzoeksterrein:

(i) Is het infimum eindig ($> -\infty$) ?

(ii) Bestaat er een oplossing $\hat{u}$, en, zo ja, zijn er eventueel meerdere oplossingen ?

(iii) Bepaal (benaderingen voor) de oplossing(en).

Voor het onderzoek van deze vragen is het vaak nodig het globale gedrag van J op de hele verzameling M te beschouwen. Daartoe zijn functionaalanalytische en topologische methoden ontwikkeld; enkele van deze *globale variatiemethoden* worden in paragraaf 5 besproken.

Het is in het algemeen niet mogelijk expliciete oplossingen te vinden. Gebruikelijk is dan het probleem te *benaderen* door een eenvoudiger probleem dat wèl oplosbaar is, of waarvan tenminste meer te zeggen valt over de oplossingsverzameling. Benaderingen kunnen verkregen worden door J en/of M te benaderen. Bij gebruik van numerieke methoden, bijvoorbeeld, wordt de oneindig-dimensionale ruimte V benaderd door (zo geschikt mogelijk te kiezen) eindig-dimensionale ruimten (zie Hoofdstuk 5). Voorbeelden van benaderingen van J komen hierna (en i.h.b. in Hoofdstuk 4) aan de orde. Belangrijk, maar in de meeste gevallen erg moeilijk, is het onderzoek van de vraag hoe "goed" (in nader te preciseren zin) de oplossing van het benaderende probleem de exacte oplossing representeert.

1.2. LOCALE VARIATIEMETHODEN IN DE KLASSIEKE VARIATIEREKENING

Naast het globale onderzoek van J op M is een heel andere methode het zoeken naar nodige en/of voldoende voorwaarden waaraan $\hat{u}$ moet voldoen opdat het een oplossing is van het min. pr.. Door een *locaal* onderzoek van de functiewaarden $J(u)$, voor $u \in M$ in een omgeving van $\hat{u}$, kan in veel gevallen relatief eenvoudig een nodige voorwaarde afgeleid worden. Nodig daarvoor is dat J en M voldoende "glad" zijn in een omgeving van $\hat{u}$; in de klassieke variatierekening worden problemen onderzocht waarvoor aan deze *gladheidsvoorwaarden* wordt voldaan.
De nodige voorwaarde voor $\hat{u}$ die dan gevonden wordt kan ruwweg omschreven worden als de eis dat de afgeleide van J langs de verzameling M in het punt $\hat{u}$ nul moet zijn; laten we dit symbolisch noteren als

$$d_M J(\hat{u}) = 0. \tag{1.4}$$

Bijvoorbeeld, voor $M = \mathbf{R}^n$ en $J \in C'(\mathbf{R}^n)$ is $d_M J(\hat{u})$ = grad $J(\hat{u})$ met "grad", (verderop ∇), de gradiënt van de functie J. In dat geval is grad $J(\hat{u}) = 0$ het bekende resultaat dat elk minimaal element van J een stationair punt van J moet zijn. Naar analogie worden daarom elementen van M die voldoen aan $d_M J(u) = 0$ *stationaire* (ot *kritieke*) *punten van* J *op* M genoemd en wordt deze voorwaarde wel de *stationairiteits-voorwaarde* genoemd.
Het afleiden van de stat. v.w. voor algemene min. problemen wordt behandeld in de paragrafen 2 en 3.

In enkele gevallen zijn efficiënte methoden beschikbaar om oplossingen van de stat. v.w. te vinden; in die gevallen kan de stat. v.w. gebruikt worden om kandidaat-oplossingen van het min. pr. te vinden.
Stationairiteitsvoorwaarden zijn echter van veel fundamenteler belang dan alleen als nodige voorwaarde voor minimale elementen. Om dat te verduidelijken is het nodig op te merken dat deze voorwaarde voor $\hat{u}$ van een geheel ander karakter is dan het oorspronkelijke min. pr.. Bijvoorbeeld, voor het bovenstaande geval dat $M = \mathbf{R}^n$, is de voorwaarde grad $J(u) = 0$ een stelsel van n vergelijkingen (in het algemeen niet lineair) voor de n componenten

van de vector $u \in \mathbb{R}^n$. Voor problemen uit de mathematische fysica is de stat. v.w. in het algemeen een gewone of partiële differentiaalvergelijking voor de functie u.
Vanwege het meer algemene, en anderssoortige, karakter zullen we het onderzoek van de stat. v.w. als een apart probleem beschouwen.

1.3. VARIATIEPROBLEMEN

Gegeven V, M en J als in 1.1., waarbij J en M aan zekere gladheidsvoorwaarden voldoen zoals in 1.2.
Dan verstaan we onder het *variatieprobleem voor* J *op* M het onderzoek van de stationaire punten van J op M

$$u \in M \quad \text{met} \quad d_M J(u) = 0. \tag{1.5}$$

Dit variatieprobleem wordt ook wel genoteerd als

$$\operatorname*{stat}_{u \in M} J(u) \quad \text{of als} \quad \operatorname{stat} \{J(u) \mid u \in M\}. \tag{1.6}$$

Omdat de stat. v.w. afgeleid wordt als nodige voorwaarde (die i.h.a. niet voldoende is) voor oplossingen van het min. pr., is het variatieprobleem meer algemeen dan het min. pr..
De reden om variatieproblemen te bestuderen is dat verschillende klassen van problemen in de mathematische fysica juist door zo'n variatieprobleem worden beschreven en niet door het meer beperkte min. pr.. In het bijzonder kan het variatieprobleem zinvol zijn in gevallen waarin het infimum van J op M $-\infty$ is (denk aan de functie $x \to x^3$ op $\mathbb{R}$). Zelfs voor veel problemen die oorspronkelijk als min. pr. zijn geformuleerd, kunnen ook stationaire punten die niet corresponderen met minima van J op M van fysisch belang zijn. In het bijzonder geldt dit voor dynamische systemen.

Karakteristieke vragen die tot het onderzoek van een variatieprobleem horen, zijn

(i) Bestaan er stationaire punten; karakteriseer de stationaire punten al naar gelang het gedrag van de functionaal in een omgeving (bijv. locale minima, maxima, of *"zadelpunten"* [zie voor dit laatste type stationaire

punten Hoofdstuk 3]).

(ii) Vind benaderingen van tenminste enkele stationaire punten. In de gevallen waarin het corresponderende min. pr. zinvol is en oplossingen heeft kunnen de specifieke benaderingsmethoden die daarvoor bestaan gebruikt worden. Maar ook voor stationaire punten van zadelpunt-type worden steeds meer constructieve benaderingsmethoden ontwikkeld. Bijvoorbeeld, door het invoeren van zogenaamde *natuurlijke nevenvoorwaarden* $N \subset V$ kan een zadelpunt van J op M een globaal minimum geven van J op de extra beperkte verzameling $M \cap N$.

De laatst genoemde mogelijkheden om sommige stationaire punten op constructieve manier te vinden, is een van de redenen waarom het van belang is te weten dat een gegeven fysisch probleem ook een variatieprobleem is. Natuurlijk zijn niet alle problemen variatieproblemen: in $\mathbf{R}^n$, bijvoorbeeld, is niet elke afbeelding $F: \mathbf{R}^n \to \mathbf{R}^n$ te schrijven als de gradiënt van een functie; alleen als $F(x) = \text{grad } J(x)$ kan het probleem van het zoeken naar oplossingen van $F(x) = 0$ met variatiemethoden worden aangepakt.
Het zoeken naar voorwaarden waaronder een probleem een variatieprobleem is, en het in dat geval construeren van de functionaal J en de verzameling M, is bekend als het *"inverse-probleem van de variatierekening"*. Bijvoorbeeld, in $\mathbf{R}^n$ met $F(x) = Ax$, A een $n \times n$ - matrix, is nodig en voldoende dat A symmetrisch is; in dat geval is $J(x)$ gelijk aan $\frac{1}{2}x.Ax$. We gaan hier niet verder op in; voorbeelden van het "herkennen" van een specifiek probleem als een variatieprobleem komen in het vervolg nog aan de orde (zie ook de Hoofdstukken 5 en 7).
Na deze algemene inleiding zullen we enkele voorbeelden geven van problemen uit de Mathematische Fysica. We beperken ons hier voornamelijk tot problemen die zich laten formuleren als min. pr.. Het befaamde principe van stationaire actie, ter beschrijving van de beweging van discrete puntmassa's of continue media, zal besproken worden in paragraaf 4, na een precieze behandeling van stationairiteitsvoorwaarden.

1.4. GEODETISCHE MINIMALISERINGS-PROBLEMEN

Bij geodetische problemen denken we aan problemen die te maken hebben met het vinden van een weg van kleinste "afstand", een zogenaamde *geodeet*, waarbij het begrip "afstand" van geval tot geval verschillend kan zijn. We geven hiervan een aantal voorbeelden. Ook analoge meer-dimensionale problemen komen vaak voor (bijv. "minimale oppervlakken").

1.4.1. WEG VAN KLEINSTE AFSTAND

Met het gewone begrip van afstand formuleren we de volgende problemen als minimum problemen.

PROBLEEM 1. Gegeven twee verschillende punten A en B in $\mathbb{R}^3$. Vind de weg van kleinste afstand tussen A en B ("weg" = continue kromme).

PROBLEEM 2. Gegeven twee verschillende punten A en B op de eenheidssfeer S^2 in $\mathbb{R}^3$. Vind die weg op de sfeer van kleinste afstand tussen A en B.

De oplossingen van beide problemen zijn wel bekend. Merk op dat Probleem 2, afhankelijk van de ligging van A en B, één of oneindig veel oplossingen heeft.

De wiskundige symbolisering van Probleem 1 als min. pr. is mogelijk met:

$$\hat{M} = \{\gamma \mid \gamma \text{ is weg van A naar B}\}$$

$$J(\gamma) = \int_{\gamma} ds \qquad (= \text{lengte van } \gamma).$$

Een meer analytische formulering is mogelijk en wenselijk. Neem daartoe e een Cartesisch coördinatenstelsel (x,y,z) in de ruimte $\mathbb{R}^3$. Het is altijd mogelijk ervoor te zorgen dat A samenvalt met de oorsprong $\underline{0}$ en B = (b,0,0) = b $\underline{e}_1$ met b > 0 (eventueel ook b = 1 te kiezen). Een parameter-voorstelling van een "weg" in $\mathbb{R}^3$ is een continue afbeelding

$$t \to \underline{r}(t) = (x(t),y(t),z(t)).$$

We beperken ons in eerste instantie tot continue differentieerbare (C^1 -) krommen. Dan geldt, met s de booglengte, $\dot{\underline{r}} := \frac{d\underline{r}}{dt} = (\dot{x},\dot{y},\dot{z})$, en met $|\ |$ de Euclidische norm

$$ds = |\dot{\underline{r}}(t)|\, dt = \{\dot{x}^2(t)+\dot{y}^2(t)+\dot{z}^2(t)\}^{\frac{1}{2}}\, dt.$$

Laat M de verzameling C^1 - krommen zijn van A naar B:

$$M = \{\underline{r} \in C^1([0,1],\mathbb{R}^3) \mid \underline{r}(0) = \underline{0},\ \underline{r}(1) = b\,\underline{e}_1\}$$

(het parameter-interval, hier $[0,1]$, kan willekeurig gekozen worden).
De voorwaarden $\underline{r}(0) = 0$ en $\underline{r}(1) = b\,\underline{e}_1$ zijn voorbeelden van randvoorwaarden ("rand" van het parameter-interval). Voor $\underline{r} \in M$ is

$$J(\underline{r}) := \int_0^1 |\dot{\underline{r}}|\, dt = \int_0^1 \{\dot{x}^2 + \dot{y}^2 + \dot{z}^2\}^{\frac{1}{2}}\, dt$$

de lengte van de kromme beschreven door $\underline{r}$. (Deze uitdrukking verklaart de reden voor beperking tot C^1 - krommen.)
Het min. pr. voor J op M kan met standaard-methoden uit de variatierekening aangepakt worden. Omdat $M \subset \hat{M}$ en $M \neq \hat{M}$, is een extra argument nodig, en mogelijk, om te laten zien dat de oplossing van $\inf_M J$ ook de oplossing is van $\inf_{\hat{M}} J$.

Een vaak gebruikte andere aanpak is om alleen krommen te bekijken die geparametriseerd kunnen worden met de x-coördinaat. Onder verdere beperking tot krommen in een vlak door A en B, zeg (x,y,0)-vlak, wordt de formulering dan voor C^1 - krommen $y = y(x)$:

$$M = \{y \in C^1([0,b]) \mid y(0) = 0,\ y(b) = 0\}$$

en

$$J(y) = \int_0^b \{1+y_x^2\}^{\frac{1}{2}}\, dx \quad \text{met } y_x := \frac{dy}{dx}\ .$$

Zoals we weten is ook de oplossing van dit meer beperkte probleem de oplossing van het oorspronkelijke probleem.

Ons meteen beperkend tot C^1 - krommen, zijn er twee voor de hand liggende formuleringen voor Probleem 2. De eerste is met de functionaal $J(\underline{r})$ van hierboven met

$$M = \{\underline{r} \in C^1([0,1],\mathbb{R}^3) \mid \underline{r}(0) = \underline{a},\ \underline{r}(1) = \underline{b},\ |\underline{r}(t)| = 1 \text{ voor } t \in (0,1)\};$$

hierin zijn $\underline{a}$ en $\underline{b}$ de coördinaten van de punten A en B, en is $|\underline{r}(t)| = 1$ een nevenvoorwaarde die specificeert dat een toegestane kromme op S^2 moet liggen.

In een tweede formulering vermijden we zo'n nevenvoorwaarde door gebruik te maken van bolcoördinaten: elk punt $\underline{r}$ op S^2 wordt beschreven met een 2-tal bolhoeken $(\phi,\theta) \in [0,2\pi) \times [0,\pi]$ volgens

$$\underline{r} = (\cos\phi \sin\theta, \sin\phi \sin\theta, \cos\theta).$$

Dan is Probleem 2 voor C^1 - krommen te formuleren als min. pr. met ($\underline{\omega}_A$ en $\underline{\omega}_B$ uit $\mathbf{R}^2$ definiëren de punten A en B)

$$M = \{(\phi,\theta) \in C^1([0,1],\mathbf{R}^2) \mid (\phi(0),\theta(0)) = \underline{\omega}_A,\ (\phi(1),\theta(1)) = \underline{\omega}_B\}$$

en

$$J(\phi,\theta) = \int_0^1 \{\dot{\phi}^2 \sin^2\theta + \dot{\theta}^2\}^{\frac{1}{2}}\, dt.$$

1.4.2. WEG VAN KLEINSTE GEWOGEN AFSTAND

Als specifiek voorbeeld kunnen we hier denken aan de modificatie van Probleem 1 uit 1.4.1. die eruit bestaat tussen twee gegeven punten in een plat vlak die weg te vinden waarvan de aanlegkosten zo klein mogelijk zijn. Tengevolge van bodemgesteldheid, bijvoorbeeld, zal de prijs per lengte-eenheid afhangen van de plaats in het (x,y) vlak. Laat n(x,y) ds de prijs zijn voor een stukje weg ter lengte ds door het punt (x,y) (n is een gegeven, positieve functie op $\mathbf{R}^2$; de prijs hangt *niet* af van de richting van de weg). Ons beperkend tot C^1 - wegen zoeken we oplossingen van het min. pr. met

$$M = \{\underline{r} \in C^1([0,1], \mathbf{R}^2) \mid \underline{r}(0) = (0,0);\ \underline{r}(1) = (b,0)\}$$

en

$$J(\underline{r}) = \int_0^1 n(x(t),y(t))\, \{\dot{x}^2 + \dot{y}^2\}^{\frac{1}{2}}\, dx.$$

Een fysisch voorbeeld van dit probleem is de voortplanting van licht in een inhomogeen, isotroop optisch medium. "Inhomogeen" wil zeggen dat de voortplantingssnelheid c van het licht afhangt van de plaats, en "isotroop" dat c niet afhangt van de richting.

Dan is $dt = \frac{ds}{c(x,y)}$ de tijd (die het licht nodig heeft) om een stukje weg met lengte ds, door het punt (x,y) te doorlopen. Zij $n(x,y) := 1/c(x,y)$ de zogenaamde brekingsindex. Dan geeft de oplossing van bovenstaand min. pr. de fysische baan van het licht, en wel op grond van het befaamde

Principe van Fermat: Voor de werkelijke, fysische baan die licht volgt tussen twee punten in een optisch medium geldt dat de daarvoor benodigde totale tijd zo klein mogelijk is in vergelijking met de tijd voor een willekeurige andere baan door die punten.

1.4.3. GEODETEN OP VARIËTEITEN

Bovenstaande voorbeelden zijn speciale gevallen van de volgende situatie:

M is een n-dimensionale gladde variëteit,

g is een gladde metriek op M.

Dat wil zeggen, als $u = (u^1, \ldots, u^n)$ een coördinatenstelsel is op M, dan is $g(u)$, voor elke $u \in M$, een symmetrische $n \times n$ - matrix met elementen $g_{ij}(u)$. Verder is $u \to g(u)$ glad en voor de lengte van een lijnelement op M door u geldt

$$ds^2 = \sum_{i,j=1}^{n} g_{ij}(u)\, du^i\, du^j \equiv g(u)\, du.du.$$

Voor een C^1 - kromme op M wordt dan de lengte gegeven door

$$J(u) = \int_0^1 [g(u(t))\, \dot{u}(t).\dot{u}(t)]^{\frac{1}{2}}\, dt =$$

$$= \int_0^1 [\Sigma\, g_{ij}(u(t))\, \dot{u}^i\, \dot{u}^j]^{\frac{1}{2}}\, dt.$$

In de voorbeelden van 1.4.1 en 1.4.2 was steeds $g(u)$ een positief-definiete matrix (Riemann-metriek). In dat geval corresponderen geodeten met minima van J op de verzameling C^1 - krommen met dezelfde eindpunten. Een belangrijk grotere klasse van problemen wordt verkregen door ook metrieken g toe te staan die niet positief-definiet zijn. In dat geval corresponderen de fysisch relevante oplossingen vaak met stationaire punten van J in plaats

van met minimale elementen. In de algemene relativiteitstheorie wordt de 4-dimensionale ruimte-tijd als zo'n Riemann-variëteit gezien met g_{ij} de potentialen van het gravitatieveld. In het bijzonder, voor "vrije deeltjes" (geen gravitatie), is g de Minkowski-metriek:

$$g(u) = \text{diag}\ (1,-1,-1,-1).$$

1.4.4. OBSTAKEL-PROBLEMEN

Een extreem geval van de in 1.4.2 besproken situatie treedt op als er een open gebied $\Omega \subset \mathbf{R}^2$ is dat uitgesloten is voor wegaanleg (of voor lichtvoortplanting). Ω fungeert dan als een "obstakel". We kunnen dit op twee manieren behandelen. (Aangenomen wordt natuurlijk dat $A \notin \Omega$, $B \notin \Omega$.)
De eerste manier is om, formeel, toe te laten dat de functie n de waarde $+\infty$ heeft. Voor het geval dat $n(x,y) = 1$ buiten Ω, bijvoorbeeld, zetten we dan

$$n(x,y) = \begin{cases} \infty & \text{voor } (x,y) \in \Omega \\ 1 & \text{anders.} \end{cases}$$

Daardoor zal dan de functionaal J alleen dan eindig zijn als de baan $\underline{r}(t)$ de verzameling Ω niet doorsnijdt: i.h.b. zal een oplossing van het min. pr. deze, gewenste, eigenschap hebben.
De andere manier is om de verbodsbepaling in de definitie van M als nevenvoorwaarde op te nemen:

$$M = \{\underline{r} \in C^1([0,1], \mathbf{R}^2) \mid \underline{r}(0) = (0,0),\ \underline{r}(1) = (b,0);\ \underline{r}(t) \notin \Omega \text{ voor } t \in (0,1)\}.$$

Een speciaal geval hiervan, waarvoor Ω beschreven kan worden met een functie $\psi(x)$, met $\psi(0) < 0$ en $\psi(b) < 0$, leidt tot een obstakelprobleem voor functies $y(x)$:

$$M = \{y \in C^1([0,b]) \mid y(0) = 0,\ y(b) = 0;\ y(x) \geq \psi(x) \text{ voor } x \in (0,b)\}$$

en

$$J(y) = \int_0^b \{1 + y_x^2\}^{\frac{1}{2}}\, dx.$$

OPMERKING: Dit voorbeeld maakt duidelijk dat het simpel kan zijn functionalen met waarden in $\mathbf{R} \cup \{+\infty\}$ te bestuderen. Het zal ook duidelijk zijn

dat in veel van deze gevallen niet voldaan wordt aan de standaardgladheidsvoorwaarde van de klassieke variatierekening (zie ook Hoofdstuk 6 voor uitbreidingen van de methoden tot niet-gladde problemen).

1.5. PRINCIPE VAN MINIMALE POTENTIËLE ENERGIE

De statische (= tijdonafhankelijke) toestand van veel systemen uit de mathematische fysica kunnen worden beschreven als een minimaliseringsprobleem. In algemene termen geformuleerd luidt dit variatie-principe:

Principe van minimale potentiële energie (M.P.E.):
Laat M de (wiskundig) mogelijke toestanden van een systeem zijn en laat $P(S)$ de potentiële energie zijn van het systeem als het zich in de toestand $S \in M$ bevindt. Dan wordt de werkelijke, fysische, toestand van het systeem gegeven door die toestand $\hat{S} \in M$ die oplossing is van

$$\inf_{S \in M} P(S).$$

In het bovenstaande wordt het begrip "systeem" zonder nadere aanduiding gebruikt. Merk op dat de vertaling van het fysische probleem naar het wiskundige min. pr. volledig vastligt door specificatie van M en P; anders gezegd: M en P definiëren het systeem. We geven enkele voorbeelden.

1.5.1. KETTINGLIJN

Dit is het probleem de vorm te vinden van een homogene, onrekbare, volkomen buigzame, ketting met constante massadichtheid ρ en lengte L, die in het constante zwaartekrachtsveld is opgehangen in twee punten A en B.
Kies in het verticale vlak door A en B een coördinatenstelsel met y-as tegengesteld aan de richting van de gravitatiekracht. De geïdealiseerde beschrijving van de ketting als een één-dimensionaal continuum gebeurt door $\underline{r}(s) = (x(s),y(s))$, met $s \in [0,L]$ de booglengte. Als we de x-as als nul-niveau nemen voor de potentiële energie, dan is de potentiële energie van een stukje ketting ter lengte ds, ter plaatse s gelijk aan

$$\rho \, ds.g.y(s)$$

waarin g = gravitatieconstante.

Voor dit probleem is dan het MPE-principe toepasbaar met (ons beperkend tot C^1 - krommen)

$$M = \{\underline{r} \in C^1([0,L], \mathbb{R}^2) \mid \underline{r}(0) = A,\ \underline{r}(L) = B\}$$

en

$$P(\underline{r}) = \int_0^L \rho\, g\, y(s)\, ds.$$

Als we de krommen willen beschrijven met een andere parameter, bijv. met x, dan moet de constante-lengte-conditie expliciet worden meegenomen:

$$M = \{y \in C^1([\alpha,\beta]) \mid y(\alpha) = a,\ y(\beta) = b;\ \int_\alpha^\beta \{1 + y_x^2\}^{\frac{1}{2}}\, dx = L\}$$

en

$$P(y) = \int_\alpha^\beta \rho g y \{1 + y_x^2\}^{\frac{1}{2}}\, dx$$

(waarin $A = (\alpha,a)$, $B = (\beta,b)$, met $\alpha < \beta$ ondersteld).

1.5.2. SNAAR EN BALK

We bekijken nu de vorm die een snaar of een balk aanneemt onder invloed van een gegeven uitwendig krachtveld. Voor het gemak beschouwen we alleen configuraties in een plat vlak, beschreven als $\underline{r}(s) = (x(s),y(s))$, met s de booglengte. Als "nul-stand", i.e. de vorm waarin de potentiële energie nul is, nemen we de toestand waarin de snaar of balk langs de x-as ligt, $\underline{r}(s) = (s,0)$, $s \in [0,L]$. De potentiële erergie van de toestand $\underline{r}(s)$ tengevolge van een gegeven krachtveld $\underline{f}(\underline{r}) = (f_1(\underline{r}), f_2(\underline{r}))$ is

$$-\int_0^L \underline{f}(\underline{r}(s)).\underline{r}(s)\, ds.$$

Daarnaast is er een *inwendige potentiële energie* die voor balk en snaar verschillend zijn. Een snaar is, per definitie, volkomen buigzaam, maar voor uitrekking is energie nodig; een balk, daarentegen, wordt verondersteld onrekbaar te zijn, maar buiging vereist energie. (In werkelijkheid treden natuurlijk steeds beide effecten tezamen op.)
De afleiding van de uitdrukking voor deze inwendige energie is lastig; in veel leerboeken wordt een eenvoudige uitdrukking in veel gevallen als

vanzelfsprekend naar voren gebracht. Voor de balk geven we enkele details van een benaderingswijze die tot zo'n eenvoudig resultaat leidt.

Voor *de balk* veronderstellen we dat een eindpunt met de oorsprong samenvalt: zeg $\underline{r}(0) = \underline{0}$. Omdat de lengte L constant is, kan niet ook het andere eindpunt voorgeschreven worden. Wel nemen we steeds y(L) = 0, zodat dan $\underline{r}(L) = (\ell,0)$ met $\ell = x(L)$ afhankelijk van de configuratie.
De precieze materiaaleigenschappen van de balk worden beschreven door een functie $E(s,k)$, waarvan de betekenis is dat $E(s,k)$ ds de inwendige potentiële energie is van een stukje balk ter lengte ds op de plaats s en met kromming k. Voor een configuratie $\underline{r}(s)$ wordt k(s) op het teken na bepaald door

$$|k(s)| = |\underline{r}_{ss}(s)|.$$

Voor een homogene balk zal E een even functie zijn van k, en zal $E(s,0) = 0$ wegens de definitie van "nul-stand".
De inwendige potentiële energie wordt dan gegeven door

$$\int_0^L E(s, |\underline{r}_{ss}(s)|)\, ds$$

en het principe van M.P.E. kan toegepast worden met

$$M = \{\underline{r}(s) \in C^2([0,L]) \mid \underline{r}(0) = 0,\ y(L) = 0\}$$

$$P(\underline{r}) = \int_0^L E(s, |\underline{r}_{ss}|)\, ds - \int_0^L \underline{f}(\underline{r}(s)).\underline{r}(s)\, ds.$$

In plaats van deze "exacte" uitdrukkingen wordt vaak een *benadering* hiervan gegeven. De gezochte benadering is voor toestanden waarvoor de kromming in elk punt klein is, zeg $|k(s)| < \varepsilon$, voor alle s.
Deze eis op $|\underline{r}_{ss}|$, tezamen met de randvoorwaarden $\underline{r}(0) = 0$ en $y(L) = 0$, impliceert ook restricties op de functies $\underline{r}(s)$ en $\underline{r}_s(s)$.
Eenvoudige analyse leert dat

$$|y_s(s)| < 2L\varepsilon \quad \text{en} \quad 1 - 2\varepsilon^2L^2 < x_s \leq 1.$$

Dit betekent i.h.b. dat x_s tot in de $2^{\underline{e}}$ orde in ε benaderd wordt door 1, zodat $x(s) = s + \mathcal{O}(\varepsilon^2)$. In het bijzonder is dan $\ell = L + \mathcal{O}(\varepsilon^2)$.
Als E voldoende glad is, kan voor $|k| < \varepsilon$ de functie $k \to E(s,k)$ benaderd worden door

$$E(s,k) = \tfrac{1}{2}\sigma(s)\, k^2 + \mathcal{O}(|k|^3)$$

waarin $\sigma(s) = \dfrac{\partial^2 E}{\partial k^2}(s,0)$ een bekende functie is.
Door nu in de inwendige potentiële energie alleen termen te nemen van laagste orde in ε, i.e. ε^2, vinden we voor krommen beschreven door $x \to y(x)$

$$\int_0^L E(s, |\underline{r}_{ss}|)\, ds = \int_0^L \tfrac{1}{2}\sigma(s)\, |\underline{r}_{ss}|^2\, ds + \mathcal{O}(\varepsilon^2) =$$

$$= \varepsilon^2 \int_0^L \tfrac{1}{2}\sigma(x)\, y_{xx}^2\, dx + \mathcal{O}(\varepsilon^3).$$

(Merk op dat nu $x \in [0,L]$. Voor oplossingen met $y \not\equiv 0$ wordt nu dus niet meer exact aan de constante-lengte-conditie voldaan; wèl in de gewenste benadering: $\ell = L + \mathcal{O}(\varepsilon^2)$.)
Voor krachtvelden $\underline{f}$ van de orde ε geldt in dezelfde benadering

$$\int_0^L \underline{f}(\underline{r}(s)).\underline{r}(s)\, ds = \varepsilon \int_0^L f_2((x,0))\, y(x)\, dx + c_1 + \mathcal{O}(\varepsilon^3)$$

waarin $c_1 = \int_0^L f_1((x,0))\, x\, dx + \mathcal{O}(\varepsilon^3)$ een bekende constante is.
Stellen we $f_2((x,0)) = g(x)$, dan vinden we een benaderde beschrijving voor de balk uit het principe van M.P.E. met

$$M = \{y \in C^2([0,L]) \mid y(0) = 0,\ y(L) = 0\}$$

$$P(y) = \int_0^L \{\tfrac{1}{2}\sigma(x)\, y_{xx}^2 - g(x)\, y\}\, dx.$$

De hier opgelegde randvoorwaarden, $y(0) = y(L) = 0$, zijn een voorbeeld van homogene *Dirichlet-randvoorwaarden*, en beschrijven een zgn. "opgelegde balk". Voor dit probleem is het mogelijk ook nog de eerste afgeleiden in $x = 0$ en/of $x = L$ voor te schrijven (*Neumann*-randvoorwaarden); bijvoorbeeld $y_x(0) = y_x(L) = 0$ voor een in beide eindpunten "ingeklemde balk",

Voor *een snaar* kunnen de twee eindpunten vast gekozen worden, zeg $\underline{0}$ en (L,0), omdat de lengte variabel is. Een "exacte" beschrijving van de inwendige potentiële energie is ook nu mogelijk; omdat lengtevariaties deze energie bepalen, zijn in zo'n uitdrukking eerste-orde-afgeleiden de hoogst voorkomende.
De interpretatie van de exacte uitdrukking, zowel als de afleiding van een benadering daarvoor, is wat lastiger dan voor de balk, en wordt hier achterwege gelaten. We geven alleen het resultaat van een benadering die afgeleid is voor toestanden $x \to y(x)$ waarvoor $|y_x(x)|$, en bijgevolg $|y(x)|$, voor alle $x \in [0,L]$ klein is:

$$P(y) = \int_0^L \{\tfrac{1}{2}\sigma(x)\ y_x^2 - g(x)\ y\}\ dx$$

waarin $\sigma(x)$ een gegeven functie is die geïnterpreteerd kan worden als de spanning in de snaar in de nul-stand.

1.5.3. TROMMELVLIES

De uitwijking van een voorgespannen trommelvlies t.g.v. een gegeven kracht is een 2-dimensionaal analogon van de snaar.
Laat $\Omega \subset \mathbb{R}^2$ een gebied zijn dat correspondeert met de ruststand van het vlies. Laat $u(x,y)$ de verticale uitwijking zijn van het vlies op de plaats $(x,y) \in \Omega$. Op de rand $\partial\Omega$ van Ω, veronderstellen we het vlies vastgeklemd:

$$u(x,y) = 0 \quad \text{voor} \quad (x,y) \in \partial\Omega.$$

Dit is weer een voorbeeld van een (homogene) Dirichlet-randvoorwaarde.
Met de notatie $\nabla u = \text{grad } u = (\frac{\partial u}{\partial x}, \frac{\partial u}{\partial y})$ voor de gradiënt van u, wordt door het principe van M.P.E. een beschrijving gegeven van de uitwijking t.g.v. een verticale kracht $g(x,y)$ (voor configuraties met "kleine uitwijking"; meer precies, waarvoor $|\nabla u|$ klein is op Ω) met

$$M = \{u \in C^1(\Omega) \mid u(x,y) = 0 \text{ voor } (x,y) \in \partial\Omega$$

en

$$P(u) = \int_\Omega \{\tfrac{1}{2}\sigma\ |\nabla u|^2 + gu\}\ dx\ dy$$

waarin $\sigma = \sigma(x,y)$ een bekende functie is (spanning in het vlies in rust-stand).

1.5.4. ELECTROSTATICA

Het principe van M.P.E. is ook geldig voor problemen uit de electrostatica. Laat $\Omega \subset R^3$ een (enkelvoudig samenhangend) gebied zijn, met rand $\partial\Omega$, waarin zich ladingen bevinden met ladingsdichtheid $\rho(\underline{x})$ in $\underline{x} \in \Omega$. Tengevolge daarvan is er een electrisch veld $E(\underline{x})$. Een van de wetten van Maxwell leert dat in een tijdonafhankelijke situatie de rotatie van E nul is, zodat E beschreven kan worden met een electrostatische potentiaal ϕ volgens

$$E = - \nabla\phi .$$

Voor zo'n veld kan de uitdrukking

$$P(\phi) := \int_{\Omega} \{ \tfrac{1}{2}\varepsilon(\underline{x}) \; |\nabla\phi|^2 - \rho(\underline{x}) \; \phi \} \; d\underline{x}$$

geïnterpreteerd worden als de totale potentiële energie; hierin is de scalarfunctie ε, de diëlectrische "constante", een gegeven functie die het materiaal in Ω karakteriseert.
Als het materiaal een zgn. *isolator* is, zal het veld niet uit het gebied Ω kunnen ontsnappen; er geldt dan voor elke $\underline{x} \in \partial\Omega$ dat $E(\underline{x}).\underline{n}(\underline{x}) = 0$, met $\underline{n}(\underline{x})$ de (naar buiten wijzende) normaal op de (gladde) rand $\partial\Omega$ in het punt $\underline{x} \in \partial\Omega$. Met de notatie

$$\nabla\phi(\underline{x}).\underline{n}(\underline{x}) \equiv \frac{\partial\phi}{\partial n}(\underline{x}), \qquad \underline{x} \in \partial\Omega$$

is deze randvoorwaarde te schrijven als $\frac{\partial\phi}{\partial n}(\underline{x}) = 0$, $\underline{x} \in \partial\Omega$, en staat bekend als een Neumann-randvoorwaarde. Later (§ 3.3) zal blijken dat dit een natuurlijke randvoorwaarde is, en dat in dit geval voor M de hele ruimte $C^1(\Omega)$ genomen dient te worden.
Voor een *geleider* zal gelden $E(\underline{x}).\underline{\tau}(\underline{x}) = 0$, $\underline{x} \in \partial\Omega$, voor elke raakvector $\underline{\tau}(\underline{x})$ aan $\partial\Omega$, hetgeen impliceert dat ϕ constant moet zijn op $\partial\Omega$, zeg $\phi = 0$ op $\partial\Omega$ (homogene Dirichlet-randvoorwaarde). In dat geval is

$$M = \{ \phi \in C^1(\Omega) \mid \phi(\underline{x}) = 0 \text{ voor } \underline{x} \in \partial\Omega \}.$$

1.5.5. DIFFUSIE EN TEMPERATUURGELEIDING

Diffusie is het verschijnsel dat t.g.v. moleculaire bewegingen en plaatselijke concentratieverschillen een bepaalde stof, opgelost in een oplosmiddel, zich verspreidt. In het algemeen is dit een tijdafhankelijk proces in de zin dat de concentratie van die stof op een vaste plaats verandert als functie van de tijd. (Denk bijvoorbeeld aan suiker of inkt in water.) Indien er "bronnen" en "putten" aanwezig zijn, i.e. indien door oorzaken van buitenaf (bijv. door chemische reacties), plaatselijk stof toegevoegd of afgevoerd wordt, kan er een tijdonafhankelijke concentratieverdeling optreden. Als we de tijd-onafhankelijke concentratie ter plaatse $\underline{x} \in \Omega$ aangeven met $c(\underline{x})$, dan zegt de *wet van Fourier* dat de snelheid $\underline{v}$ van de stof t.g.v. concentratieverschillen evenredig is en tegengesteld van richting, met de concentratiegradiënt:

$$\underline{v}(\underline{x}) = - \kappa \nabla c(\underline{x});$$

de evenredigheidscoëfficiënt $\kappa > 0$ hangt i.h.a. van $\underline{x}$ en $c(\underline{x})$ af.
De "kinetische" energie van de massa in een volume-element is dan evenredig met $\frac{1}{2}c(\underline{x})\, d\underline{x}.|\kappa \nabla c(\underline{x})|^2$. Voor zo'n volume-element is de potentiële energie t.g.v. een brondichtheid $g(\underline{x})$ in Ω: $- g(\underline{x})c(\underline{x})\, d\underline{x}$.
Voor kleine concentratie-verschillen t.o.v. een uniforme concentratie, zeg $c(\underline{x}) = c_0 + \varepsilon u(\underline{x}) + \mathcal{O}(\varepsilon^2)$, vinden we dan in laagste orde van ε voor u als uitdrukking van de totale potentiële energie (op een constante na):

$$P(u) = \int_{\Omega} \{\tfrac{1}{2}D(\underline{x}) \, |\nabla u|^2 - g(\underline{x})u\} \, d\underline{x}$$

waarin $D(\underline{x})$ de zgn. diffusie-coëfficiënt is.

Een Dirichlet-randvoorwaarde voor u, bijvoorbeeld $u = \phi$ op $\partial\Omega$, met ϕ een gegeven functie gedefinieerd op $\partial\Omega$, betekent in dit geval dat de concentratie op de rand wordt voorgeschreven. Homogene Neumann-randvoorwaarden, $\frac{\partial u}{\partial n} = 0$ op $\partial\Omega$, betekenen dat geen stof door de rand $\partial\Omega$ het gebied Ω kan verlaten of binnenkomen.
Vanzelfsprekend kan ook een 1- of 2-dimensionaal model worden bekeken.

Thermische geleiding, bijvoorbeeld in een metaal, wordt op precies dezelfde manier beschreven met deze functionaal P; $u(\underline{x})$ is dan de temperatuur in $\underline{x}$, en g is een warmtebron.

2. ALGEMENE STATIONAIRITEITSVOORWAARDEN

In deze paragraaf zal duidelijk worden waaraan de variatierekening zijn naam te danken heeft. Variëren is namelijk, historisch gezien, de eerst gebruikte methode geweest om nodige en voldoende voorwaarden te vinden waaraan een element $\hat{u} \in M$ moet voldoen opdat $J(\hat{u})$ een *locaal minimum* is *van* J *op* M, i.e. opdat er een $\delta > 0$ is zodat

$$J(\hat{u}) \leq J(u) \quad \text{voor alle } u \in M \text{ met } \|u - \hat{u}\| < \delta.$$

(Elke oplossing van het min. pr. voor J op M geeft natuurlijk een locaal minimum.)

Het basisidee is eenvoudig: Als $\hat{u}$ zo'n locaal minimaal element is, wordt de waarde $J(\hat{u})$ vergeleken met de waarden van J in gevarieerde punten, i.e. in punten u die in een kleine omgeving van $\hat{u}$ liggen en die tot M behoren; $u-\hat{u}$ heet dan een "toegestane variatie" van $\hat{u}$. Het is duidelijk dat het onderzoek van $J(u) - J(\hat{u})$ afhangt van de functionaal J in een omgeving van $\hat{u}$, alsook van de "vorm" van de verzameling M in een omgeving van $\hat{u}$, i.e. van de toegestane variaties. Het resultaat dat gebaseerd is op lineaire benaderingen voor zowel J als M in de buurt van $\hat{u}$, i.e. het onderzoek van $J(u) - J(\hat{u})$ in eerste orde van $\|u - \hat{u}\|$, levert de stationairiteitsvoorwaarde en is bekend als de *theorie van eerste variatie*. Verdere nodige en voldoende voorwaarden worden bestudeerd in de theorie van tweede variatie, die hier geheel buiten beschouwing wordt gelaten.

In deze paragraaf geven wij kort de uitwerking van deze ideeën uitgaande van een algemene functionaal J en zekere deelverzamelingen M van een genormeerde ruimte V. In paragrafen 3 en 4 worden de hier gevonden resultaten gespecificeerd voor de in toepassingen meest voorkomende functionalen, nl. dichtheidsfunctionalen op ruimten van functies.

In het vervolg is:

V een genormeerde lineaire ruimte met norm $\|\ \|$;

J een functionaal op V: $J: V \to \mathbf{R}$.

2.1. GATEAUX- EN FRÉCHET-DIFFERENTIEERBAARHEID VAN FUNCTIONALEN

Het idee van richtingsafgeleide zoals bekend van functies op $\mathbf{R}^n$ is bruikbaar voor generalisatie tot functionalen op oneindig-dimensionale ruimten. Dit leidt tot het begrip "Gateaux-afgeleide". Zodra van het differentiequotient zekere uniformiteit in de verschillende richtingen wordt geëist, wordt Fréchet-differentieerbaarheid de generalisatie van differentieerbaarheid op $\mathbf{R}^n$.

2.1.1. DEFINITIE

Laat $\hat{u} \in V$. Indien zinvol, wordt de *eerste variatie* van J in $\hat{u}$ in de richting $\eta \in V$ gedefinieerd door

$$\delta J(\hat{u};\eta) := \frac{d}{d\varepsilon} J(\hat{u}+\varepsilon\eta)\Big|_{\varepsilon=0} = \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} [J(\hat{u}+\varepsilon\eta) - J(\hat{u})].$$

Indien $\delta J(\hat{u};\eta)$ bestaat voor elke $\eta \in V$, en de functionaal $\eta \to \delta J(\hat{u};\eta)$ is een continue lineaire functionaal op V, dan heet J *Gateaux-differentieerbaar* in $\hat{u}$, en wordt de *Gateaux-afgeleide* van J in $\hat{u}$ gedefinieerd als de continue lineaire functionaal $J'(\hat{u}): V \to \mathbf{R}$, dus $J'(\hat{u}) \in V^*$, door

$$\langle J'(\hat{u}),\eta\rangle = \delta J(\hat{u};\eta).$$

2.1.2. OPMERKINGEN

* In sommige boeken wordt de Gateaux-afgeleide gedefinieerd door de éénzijdige limiet voor $\varepsilon \downarrow 0$ van $\frac{1}{\varepsilon}[J(\hat{u}+\varepsilon\eta) - J(\hat{u})]$ te beschouwen. In het vervolg zou ook deze definitie gebruikt kunnen worden.
* De Gateaux-afgeleide in een punt, indien die bestaat, is uniek.
* Als J Gateaux-differentieerbaar is in $\hat{u}$, dan is J niet noodzakelijk continu in $\hat{u}$; bijvoorbeeld: $V = \mathbf{R}^2$, en $J(x,y) = 1$ voor $y = x^2$, $x \neq 0$, en $= 0$ anders, in $\hat{u} = (0,0)$.

* Als J Gateaux-differentieerbaar is in $\hat{u}$, dan geldt voor elke $\eta \in V$, de lineaire benadering

$$J(\hat{u} + \varepsilon\eta) = J(\hat{u}) + \varepsilon \langle J'(\hat{u}),\eta\rangle + o(\varepsilon), \qquad \varepsilon \to 0$$

maar de $o(\varepsilon)$ - term zal i.h.a. niet uniform zijn in $\eta \in V$.

2.1.3. DEFINITIE

Stel J is Gateaux-differentieerbaar in $\hat{u}$, met Gateaux-afgeleide $J'(\hat{u})$.
J heet *Fréchet-differentieerbaar* in $\hat{u}$ als

$$\lim_{\substack{\|v\| \to 0 \\ v \in V}} \frac{1}{\|v\|} [J(\hat{u} + v) - J(\hat{u}) - \langle J'(\hat{u}),v\rangle] = 0.$$

In dat geval noemen we $J'(\hat{u})$ de *Fréchet-afgeleide* in $\hat{u}$.

2.1.4. OPMERKINGEN

* Als $V = \mathbb{R}^n$ en $\langle\ ,\ \rangle$ wordt opgevat als het Euclidische inproduct, dan komt Fréchet-differentieerbaarheid overeen met de gebruikelijke definitie en is $J'(\hat{u})$ de gradiënt van J in $\hat{u}$: $J'(\hat{u}) = \nabla J(\hat{u})$.

* Als J Fréchet-differentieerbaar is in $\hat{u}$, dan is J continu in $\hat{u}$. Uit de definitie volgt dat nu zelfs de lineaire benadering voor $J(\hat{u} + v)$ uniform is in $v \in V$:

$$J(\hat{u} + v) = J(\hat{u}) + \langle J'(\hat{u}),v\rangle + o(\|v\|), \qquad \|v\| \to 0,\ v \in V.$$

* Er kan aangetoond worden: als J Gateaux-differentieerbaar is in een omgeving U van $\hat{u}$, en als dan de Gateaux-afgeleide $J': U \to V^*$ continu is, dan is J Fréchet-differentieerbaar in $\hat{u}$.

2.1.5. VOORBEELDEN

(i) Als $\ell \in V^*$, dan is ℓ Fréchet-differentieerbaar in elk punt $\hat{u} \in V$, en $\ell'(\hat{u}) = \ell$ (onafhankelijk van $\hat{u}$).

(ii) Laat $a(.,.): V \times V \to \mathbb{R}$ een continue, symmetrische bilineaire functionaal zijn. De *quadratische vorm* daarbij geven we ook aan met a, dus $a(u) = a(u,u)$. Dan is deze quadratische vorm $a: V \to \mathbb{R}$ Fréchet-differentieerbaar op V, met als Fréchet-afgeleide in $\hat{u} \in V$:

$$a'(\hat{u}) = 2a(\hat{u},.) \qquad (\text{dus } \langle a'(\hat{u}),\eta\rangle = 2a(\hat{u},\eta),\ \eta \in V).$$

2.1.6. ENKELVOUDIGE INTEGRALEN

Zij Ω een interval, zeg $\Omega = (0,1)$, en beschouw scalar- of m-vectorfuncties u op Ω: $u: (0,1) \to \mathbf{R}^m$, $m \geq 1$.
Een dichtheidsfunctionaal, of enkelvoudige integraal, is dan een functionaal van de vorm

$$J(u) = \int_0^1 F[u(x)]\, dx$$

waarin de "dichtheid" $F[u(x)]$ de afgekorte notatie is voor $F(x, u(x), u_x(x), u_{xx}(x), \ldots)$, waarin F een gegeven functie van zijn argumenten is, en geëvalueerd wordt in x, u(x), en afgeleiden van u in x tot zekere orde.
Een specifiek voorbeeld (waartoe we de algemene formules zullen beperken i.v.m. de toenemende complexiteit van de notatie) is

$$F \in C^1([0,1] \times \mathbf{R}^m \times \mathbf{R}^m)$$

$$J(u) = \int_0^1 F[u(x)]\, dx = \int_0^1 F(x,u(x),u_x(x))\, dx.$$

Noteer de partiële afgeleiden van de functie $(x,u,v) \to F(x,u,v)$ op standaard manier:

$$F_u = \frac{\partial F}{\partial u} = \left(\frac{\partial F}{\partial u_1}, \ldots, \frac{\partial F}{\partial u_m}\right), \qquad F_v \text{ analoog.}$$

(Om begrijpelijke redenen wordt F_v ook vaak genoteerd als F_{u_x}.)
Het is eenvoudig in te zien dat voor $u \in C^1([0,1])$ en $\eta \in C^1_x([0,1])$ de eerste variatie van J wordt gegeven door

$$\delta J(u;\eta) = \int_0^1 \{F_u[u(x)].\eta(x) + F_v[u(x)].\eta_x(x)\}\, dx$$

waarin $F_u.\eta = \sum_{i=1}^m F_{u_i}\, \eta_i$ etc.. Hieruit volgt meteen dat $\eta \to \delta J(u;\eta)$ een lineaire functionaal is op $C^1([0,1])$.
We kunnen differentieerbaarheid onderzoeken als de norm op $C^1([0,1])$ gespecificeerd is. Het blijkt dat J voor elke $u \in C^1([0,1])$ Gateaux-, en zelfs Fréchet-differentieerbaar is voor de, voor deze functionaal meest

voor de hand liggende, normen:

de C^1 - norm, gedefinieerd door

$$\|\eta\|_{C^1} = \|\eta\|_{C^0} + \|\eta_x\|_{C^0} \quad \text{waarin } \|\eta\|_{C^0} := \max_{x \in [0,1]} |\eta(x)|$$

en de H^1 - norm, een Hilbert-norm, gedefinieerd door

$$\|\eta\|_{H_1}^2 = \|\eta\|_0^2 \quad \|\eta_x\|_0^2 \quad \text{waarin } \|\eta\|_0^2 = \int_0^1 |\eta(x)|^2 \, dx$$

($\|\ \|_0$ is de gebruikelijke L_2 - norm). Immers, om Gateaux-differentieerbaarheid aan te tonen, merken we op dat

$$|\delta J(u;\eta)| \leq \|F_u\ u\|_{C^0}.\|\eta\|_{C^0} + \|F_v\ u\|_{C^0}.\|\eta_x\|_{C^0} \leq \alpha\|\eta\|_{C^1}$$

waarin $\alpha := \max\{\|F_u[u]\|_{C^0}, \|F_v[u]\|_{C^0}\}$ eindig is vanwege $F \in C^1$ en $u \in C^1$. Daaruit volgt dan dat $\eta \to \delta J(u;\eta)$ continu is. Dus is J Gateaux-differentieerbaar op C^1 met norm $\|\ \|_{C^1}$. Hetzelfde resultaat volgt voor de H^1 - norm: door gebruik te maken van de ongelijkheid van Cauchy-Schwarz

$$\left|\int_0^1 f(x).g(x)\, dx\right| \leq \|f\|_0.\|g\|_0$$

volgt namelijk

$$|\delta J(u;\eta)| \leq \|F_u[u]\|_0.\|\eta\|_0 + \|F_v[u]\|_0.\|\eta_x\|_0 \leq \sqrt{2}.\beta\|\eta\|_{H^1}$$

met $\beta = \max\{\|F_u[u]\|_0, \|F_v[u]\|_0\}$ eindig.

2.1.7. Voor de (benaderde) potentiële energie van snaar en balk (zie 1.5.2) geldt:

$$P(u) = \int_0^1 \{\tfrac{1}{2}\sigma(x)u_x^2 - f(x)u(x)\}\, dx$$

met

$$\delta P(u;\eta) = \int_0^1 \{\sigma(x)u_x.\eta_x - f(x)\eta(x)\}\, dx,$$

respectievelijk

$$P(u) = \int_0^1 \{\tfrac{1}{2}\sigma(x)u_{xx}^2 - f(x)u(x)\}\, dx$$

met

$$\delta P(u;\eta) = \int_0^1 \{\sigma(x)u_{xx}.\eta_{xx} - f(x)\eta(x)\}\, dx.$$

2.1.8. MEERVOUDIGE INTEGRALEN

Beschouw nu meer algemeen een gebied $\Omega \subset \mathbf{R}^n$ en scalarfuncties u op Ω: $u: \Omega \to \mathbf{R}$ (vectorfuncties kunnen analoog behandeld worden). In tegenstelling tot paragraaf 1 laten we voortaan streepjes onder vectoren achterwege. Een dichtheidsfunctionaal, of meervoudige integraal, is nu een functionaal van de vorm

$$J(u) = \int_\Omega F[u(x)]\, dx$$

waarin F[u(x)] de afgekorte notatie is voor een gegeven functie F geëvalueerd in de argumenten x, u(x), $u_{x_1}(x)$,..., $u_{x_n}(x)$, $u_{x_1x_1}(x)$, ...
Als specifiek voorbeeld nemen we weer dat alleen afgeleiden van eerste orde in F optreden:

$$F \in C^1(\overline{\Omega} \times \mathbf{R} \times \mathbf{R}^n)$$

en

$$J(u) = \int_\Omega F[u(x)]\, dx = \int_\Omega F(x,u(x),\nabla u(x))\, dx.$$

Dan vinden we voor de eerste variatie in $u \in C^1(\Omega)$, in de richting $\eta \in C^1(\Omega)$:

$$\delta J(u;\eta) = \int_\Omega \{F_u[u(x)]\eta(x) + F_v[u(x)].\nabla\eta(x)\}\, dx.$$

Met voor de hand liggende aanpassing van de definitie van C^1 en H^1 - norm voor functies op $\Omega \subset \mathbf{R}^n$, kan dan aangetoond worden, als in 2.1.6, dat J in elke $u \in C^1(\Omega)$ Gateaux- en Fréchet-differentieerbaar is m.b.t.elk van die twee normen.

2.1.9. Voor de potentiële energie van het diffusie-probleem (1.5.5) geldt

$$P(u) = \int_\Omega \{\tfrac{1}{2}D(x)\ |\nabla u|^2 - g(x)u\}\ dx$$

$$\delta P(u;\eta) = \int_\Omega \{D(x)\nabla u.\nabla\eta - g(x)\eta\}\ dx.$$

2.2. STATIONAIRITEITSVOORWAARDEN VOOR EEN AFFIENE VERZAMELING M

2.2.1. Het eenvoudigst te onderzoeken is het geval dat M een affiene ruimte is:

$$M = \{u_0\} + M_0 \quad \text{met}$$

$u_0 \in V$ en M_0 een gesloten lineaire deelruimte van V.

Voor elke $\hat{u} \in M$ is M dan ook te schrijven als $M = \{\hat{u}\} + M_0$. Daaraan zien we dat M_0 de verzameling van toegestane variaties is in het punt $\hat{u}$: voor elke $\eta \in M_0$ behoort $\hat{u} + \varepsilon\eta$ tot M voor elke $\varepsilon \in \mathbf{R}$.
Beschouw daarom voor willekeurige $\eta \in M_0$ de functie van één variabele:

$$\mathbf{R} \ni \varepsilon \to j(\varepsilon) := J(\hat{u} + \varepsilon\eta).$$

Triviaal, maar essentieel, is de opmerking dat als J locaal minimaal is in $\hat{u}$, dan is j locaal minimaal in $\varepsilon = 0$, zodat dan moet gelden $\frac{d}{d\varepsilon}j(0) = 0$.
Omdat

$$\frac{d}{d\varepsilon}j(\varepsilon)\Big|_{\varepsilon=0} = \delta J(\hat{u};\eta)$$

volgt hieruit het eerste resultaat.

2.2.2. STELLING
Zij M als in 2.2.1, en $\hat{u} \in M$, J Gateaux-differentieerbaar in $\hat{u}$, J locaal minimaal op M in $\hat{u}$.
Dan geldt dat $\hat{u}$ voldoet aan de volgende, zogenaamde, *stationairiteitsvoorwaarde*:

$$\langle J'(\hat{u}),\eta\rangle = 0 \quad \text{voor alle } \eta \in M_0.$$

2.2.3. DEFINITIE

Een element $\hat{u} \in V$ heet een *stationair punt* of *kritiek punt* van J op M, met M als in 2.2.1, als J Gateaux-differentieerbaar is in $\hat{u}$, $\hat{u} \in M$, en $\hat{u}$ voldoet aan de stationairiteitsvoorwaarde 2.2.2.

In veel toepassingen is M_0 een lineaire deelruimte van eindige codimensie in V.

2.2.4. In het speciale geval dat $M = V$ drukt de stationairiteitsvoorwaarde

$$\langle J'(\hat{u}), \eta \rangle = 0 \quad \text{voor alle } \eta \in V$$

uit dat $J'(\hat{u})$ de nul-functionaal is:

$$J'(\hat{u}) = 0 \quad \text{in } V^*.$$

2.2.5. Een andere veel voorkomende situatie is:

$$M = \{u \in V \mid \langle \ell_i, u \rangle = c_i,\ 1 \leq i \leq p\}$$

met $(c_1,\ldots,c_p) \in \mathbf{R}^p$ en $\ell_1,\ldots,\ell_p \in V^*$, lineair onafhankelijk.
Uit de lineaire onafhankelijkheid van $\ell_1,\ldots,\ell_p$ volgt dat

$$\{(\langle \ell_1, u \rangle, \langle \ell_2, u \rangle, \ldots, \langle \ell_p, u \rangle) \mid u \in V\} = \mathbf{R}^p.$$

Hieruit volgt:

(i) voor elke keuze $(c_1,\ldots,c_p) \in \mathbf{R}^p$ is de verzameling M niet leeg;

(ii) er bestaan elementen $e_1,\ldots,e_p \in V$ zodanig dat

$$\ell_i(e_j) = \delta_{ij} = \begin{cases} 0 & \text{als } i \neq j \\ 1 & \text{als } i = j \end{cases} \qquad \text{voor } 1 \leq i,j \leq p;$$

(iii) V is de directe som

$$V = [|e_1,\ldots,e_p|] + M_0$$

waarin M_0 de gesloten lineaire deelruimte van codimensie p is, gegeven door

$$M_0 = \{v \in V \mid \langle \ell_i, v \rangle = 0,\ 1 \leq i \leq p\};$$

anders gezegd: elke $u \in V$ kan op eenduidige manier geschreven worden als

$$u = \sum_{i=1}^{p} \alpha_i e_i + v, \text{ met } v \in M_0, \ (\alpha_1,\ldots,\alpha_p) \in \mathbf{R}^p$$

en α_i wordt in feite gegeven door

$$\alpha_i = \langle \ell_i, u \rangle, \ 1 \leq i \leq p.$$

De stationairiteitsvoorwaarde 2.2.2 impliceert in dit geval dat voor elke $u \in V$:

$$\langle J'(\hat{u}), u \rangle = \sum_{i=1}^{p} \langle J'(\hat{u}), e_i \rangle \langle \ell_i, u \rangle$$

m.a.w. dat $J'(\hat{u})$ een bepaalde lineaire combinatie is van $\ell_1,\ldots,\ell_p$.

Dit resultaat formuleren we expliciet.

2.2.6. LINEAIRE MULTIPLICATORREGEL

Zij M als in 2.2.5, $\hat{u}$ een stationair punt van J op M. Dan zijn er eenduidig bepaalde getallen $(\lambda_1,\ldots,\lambda_p) \in \mathbf{R}^p$ zodanig dat

$$J'(\hat{u}) = \lambda_1 \ell_1 + \ldots + \lambda_p \ell_p \quad \text{in } V^*.$$

Deze getallen $\lambda_1, \lambda_2, \ldots, \lambda_p$ worden *multiplicatoren* genoemd.

Een andere interpretatie van dit resultaat, die ook voor meer complexe situaties van belang is, wordt verkregen door 2.2.6 als volgt te herformuleren.

2.2.7. Voor M als in 2.2.5, definieer de zogenaamde *Lagrange-functionaal* L op $V \times \mathbf{R}^p$ door

$$L(u;\lambda_1,\ldots,\lambda_p) = J(u) - \lambda_1 \langle \ell_1, u \rangle - \ldots - \lambda_p \langle \ell_p, u \rangle.$$

Dan geldt: $\hat{u}$ is een stationair punt van J op M dan en slechts dan als er multiplicatoren $(\lambda_1,\ldots,\lambda_p) \in \mathbf{R}^p$ zijn zodanig dat $\hat{u} \in M$ een stationair punt is van $L(.;\lambda_1,\ldots,\lambda_p)$ op V.

2.2.8. De formulering 2.2.7 geeft een indicatie dat het in principe mogelijk is de *nevenvoorwaarden* (i.e. de restrictie van J tot M) te *elimineren* door introductie van een geschikt gemodificeerde functionaal $L(.,\lambda_1,\ldots,\lambda_p)$.

Er zij met nadruk op gewezen dat er zo'n verband, zonder verdere voorwaarden, alléén bestaat tussen *stationaire* punten, en dus, bijvoorbeeld, niet tussen (locaal) minimale elementen. Bijvoorbeeld:

$$V = \mathbb{R}^2,\ M = \{(x,y) \mid x = 0\},\ J(x,y) = x^3 + y^2,\ \hat{u} = (0,0).$$

2.2.9. EQUIVALENTIE VAN VARIATIEPROBLEEM EN MINIMALISERINGSPROBLEEM

In sectie 5.2 geven we voorwaarden waaronder de stationairiteitsvoorwaarde equivalent is met de formulering als min. pr.. We geven hier een voor toepassingen belangrijk voorbeeld daarvan.

Laat $Q: V \to \mathbb{R}$ een quadratische functionaal zijn, i.e. $Q(u) = a(u) - \ell(u)$, met $a: V \to \mathbb{R}$ een continue, quadratische vorm, en $\ell \in V^*$.

STELLING: Laat M zijn als in 2.2.1, en Q als boven. Veronderstel dat a niet negatief is op M_0, i.e. $a(\eta) \geq 0$ voor alle $\eta \in M_0$.
Dan geldt: $\hat{u}$ is een stationair punt van Q op M d.e.s.d.a. $\hat{u}$ een oplossing is van het min. pr. voor Q op M.

Bovendien, als a positief is op M_0, i.e. $a(\eta) > 0$ voor alle $\eta \in M_0$, $\eta \neq 0$, dan is een eventueel stationair punt (= minimaal element) *eenduidig*.

BEWIJS: Merk op dat voor elke $\hat{u} \in M$

$$Q(\hat{u} + \eta) = Q(\hat{u}) + \langle Q'(\hat{u}), \eta \rangle + a(\eta) \quad \text{voor alle} \quad \eta \in M_0.$$

Daaruit volgt dat voor een stationair punt $\hat{u}$ van Q op M geldt

$$Q(\hat{u} + \eta) \geq Q(\hat{u}) \quad \text{voor alle} \quad \eta \in M_0.$$

Omdat elk element $u \in M$ te schrijven is als $\hat{u} + \eta$ met $\eta \in M_0$ (n.l. $\eta = u - \hat{u}$) volgt hieruit dat $Q(\hat{u})$ minimaal is op M, en dat $Q(\hat{u} + \eta) > Q(\hat{u})$ voor $\eta \neq 0$ als $a(\eta) > 0$ voor $\eta \in M_0 \setminus \{0\}$.
Met Stelling 2.2.2 volgt dan het gestelde.

2.3. STATIONAIRITEITSVOORWAARDE VOOR EEN GLADDE VARIËTEIT M

In geval M geen affiene ruimte is, kunnen we in het algemeen geen lineaire deelruimte $M_0 \subset V$ vinden zodanig dat voor $\eta \in M_0$ geldt $\hat{u} + \varepsilon\eta \in M$ voor $\varepsilon \in \mathbf{R}$, $\varepsilon \neq 0$. In plaats daarvan zoeken we meer algemene één-parameterkrommen $\varepsilon \to u(\varepsilon)$ op M waarvoor $u(0) = \hat{u}$ en $u(\varepsilon) \in M$ voor alle $\varepsilon \in \mathbf{R}$, $|\varepsilon|$ voldoende klein. Als de raakrichting v aan zo'n kromme in $\varepsilon = 0$ goed gedefinieerd is, i.e. als $\frac{d}{d\varepsilon} u(\varepsilon)|_{\varepsilon=0} = v \in V$ bestaat, dan geldt dat $\varepsilon \to \hat{u} + \varepsilon v$ een raaklijn is aan M in $\hat{u}$, d.w.z. dat $\hat{u} + \varepsilon v$ dan (tot op hogere orde dan lineair in ε, $|\varepsilon| \to 0$) tot M behoort. De verzameling van al dit soort raakrichtingen definieert de raakruimte aan M in $\hat{u}$.

2.3.1. DEFINITIE

Voor $\hat{u} \in M$ wordt de *raakruimte* $TM_{\hat{u}}$ *aan* M *in* $\hat{u}$ gedefinieerd als de verzameling $v \in V$ waarvoor geldt: er is een $\varepsilon_0 > 0$ en een afbeelding $w(.,v)\colon (-\varepsilon_0,\varepsilon_0) \to V$ zodanig dat

(i) $\hat{u} + \varepsilon v + w(\varepsilon,v) \in M$ voor alle $\varepsilon \in (-\varepsilon_0,\varepsilon_0)$

(ii) $\lim_{\varepsilon\to 0} \frac{1}{|\varepsilon|} \|w(\varepsilon,v)\| = 0.$

2.3.2. STELLING

Zij $\hat{u} \in M$, en $TM_{\hat{u}}$ de raakruimte, J Fréchet-differentieerbaar in $\hat{u}$ en J locaal minimaal op M in $\hat{u}$.

Dan voldoet $\hat{u}$ aan de volgende *stationairiteitsvoorwaarde*:

$$\langle J'(\hat{u}),\eta\rangle = 0 \quad \text{voor alle} \quad \eta \in TM_{\hat{u}}.$$

BEWIJS: Neem $\eta \in TM_{\hat{u}}$ willekeurig en ε_0 en $w(.,\eta)$ als in 2.3.1. Beschouw dan voor $|\varepsilon| < \varepsilon_0$ de functie

$$j(\varepsilon) = J(\hat{u} + \varepsilon\eta + w(\varepsilon,\eta)).$$

Uit de differentieerbaarheid van J in $\hat{u}$ volgt m.b.v. 2.3.1. (ii) dat

$$J(\hat{u} + \varepsilon\eta + w(\varepsilon,\eta)) = J(\hat{u}) + \langle J'(\hat{u}),\varepsilon\eta + w\rangle + o(\|\varepsilon v + w\|) =$$

$$= J(\hat{u}) + \varepsilon\langle J'(\hat{u}),\eta\rangle + o(|\varepsilon|) \quad \text{als} \quad \varepsilon \to$$

Dit impliceert dat j differentieerbaar is in $\varepsilon = 0$ met $\frac{d}{d\varepsilon} j(\varepsilon)\big|_{\varepsilon=0} =$
$= \langle J'(\hat{u}), \eta \rangle$. Omdat j locaal minimaal is in $\varepsilon = 0$, is deze afgeleide gelijk aan nul en het resultaat volgt.

2.3.3. DEFINITIE

Een element $\hat{u} \in V$ heet een *stationair* (of *kritiek*) *punt* van J op M als J Fréchet-differentieerbaar is in $\hat{u} \in M$, en $\hat{u}$ aan de stationairiteitsvoorwaarde 2.3.2 voldoet.

2.3.4. We bepalen nu de raakruimte $TM_{\hat{u}}$ in het volgende geval:

$$M = \{u \in V \mid G_i(u) = c_i,\ 1 \le i \le p\} \text{ met gegeven}$$

$$(c_1,\ldots,c_p) \in \mathbb{R}^p \text{ zodanig dat } M \neq \emptyset \text{ en waarbij}$$

$$G_1,\ldots,G_p \text{ functionalen zijn op } V$$

en $\hat{u}$ is een zogenaamd *regulier punt* van M, i.e. $\hat{u} \in M$, G_i Fréchet-differentieerbaar in $\hat{u}$, voor $1 \le i \le p$, en de Fréchet-afgeleiden $G_1'(\hat{u}),\ldots,G_p'(\hat{u})$ in $\hat{u}$ lineair onafhankelijke functionalen in V^*.

LEMMA (Lyusternik)

Voor M en $\hat{u}$ als hierboven is $TM_{\hat{u}}$ de gesloten lineaire deelruimte van codimensie p in V gegeven door

$$TM_{\hat{u}} = \{v \in V \mid \langle G_i'(\hat{u}), v \rangle = 0,\ 1 \le i \le p\}.$$

BEWIJSSCHETS: Het moeilijkste deel van het bewijs is om bij gegeven $v \in V$ met $\langle G_i'(\hat{u}), v \rangle = 0$, $1 \le i \le p$, de afbeelding w te vinden die aan de eisen van 2.3.1 voldoet. We zoeken een w van de vorm $w = \sum_1^p \alpha_i e_i$, waarin de getallen $\alpha_1,\ldots,\alpha_p$ van ε (en v) zullen afhangen, en waarin $e_1,\ldots,e_p$ uit V gekozen worden als in 2.2.5 met $\ell_i = G_i'(\hat{u})$. Een vrij directe toepassing van de impliciete functiestelling (in de eindig-dimensionale ruimte $\mathbb{R}^p$) op de functie

$$(\varepsilon,\alpha_1,\ldots,\alpha_p) \to (G_1(\hat{u}+\varepsilon v+\sum_1^p \alpha_i e_i),\ldots,G_p(\hat{u}+\varepsilon v+\sum_1^p \alpha_i e_i))$$

levert dan het gewenste resultaat.

De verdere uitwerking hiervan laten wij achterwege.

2.3.5. MULTIPLICATORREGEL

Laat M en $\hat{u}$ zijn als in 2.3.4 en zij $\hat{u}$ een stationair punt van J op M.

Dan geldt: Er zijn multiplicatoren $\lambda_1,\ldots,\lambda_p$ zodanig dat in V^*

$$J'(\hat{u}) = \lambda_1 G_1'(\hat{u}) + \ldots + \lambda_p G_p'(\hat{u}).$$

Equivalent: Er zijn multiplicatoren $\lambda_1,\ldots,\lambda_p$ zodanig dat $\hat{u}$ een stationair punt is van de volgende *Lagrange-functionaal* $L(.;\lambda_1,\ldots,\lambda_p)$ op V

$$L(u;\lambda_1,\ldots,\lambda_p) := J(u) - \lambda_1 G_1(u) - \ldots - \lambda_p G_p(u).$$

BEWIJS: Dit volgt meteen uit 2.3.2, 2.3.4 en de lineaire multiplicatorregel 2.2.6.

2.4. STATIONAIRITEITSVOORWAARDEN VOOR EEN CONVEXE VERZAMELING M

2.4.1. Een verzameling M heet *convex* als voor alle paren punten u en v van M het gehele lijnstuk tussen u en v tot M behoort:

$$u,v \in M \Rightarrow \lambda u + (1 - \lambda)v \in M, \quad \forall \lambda \in (0,1).$$

Voor ons doel belangrijk, kan dit ook anders geïnterpreteerd worden:
voor $\hat{u} \in M$ en elke $v \in M$ geldt dat $\hat{u} + \varepsilon(v - \hat{u}) \in M$ voor elke $\varepsilon \in [0,1]$.
Met andere woorden: voor $\hat{u} \in M$ is $\{\varepsilon(v - \hat{u}) \mid v \in M, \varepsilon \in [0,1]\}$ de verzameling toegestane variaties.
Beschouw nu de functie $\varepsilon \to j(\varepsilon) := J(\hat{u} + \varepsilon(v - \hat{u}))$, gedefinieerd voor $\varepsilon \in [0,1]$. Als J op M locaal minimaal is in $\hat{u}$, dan is j locaal minimaal in het linker eindpunt $\varepsilon = 0$ van het interval $[0,1]$. Voor de rechter afgeleide van j in $\varepsilon = 0$ moet dan dus gelden $\frac{d}{d\varepsilon} j(\varepsilon)\big|_{\varepsilon=+0} \geq 0$.
Omdat $\frac{d}{d\varepsilon} j(\varepsilon)\big|_{\varepsilon=0} = \delta J(\hat{u};v - \hat{u})$, vinden we het volgende resultaat.

2.4.2. STELLING

Zij $M \subset V$ een convexe verzameling, $\hat{u} \in M$, J Gateaux-differentieerbaar in $\hat{u}$ en J locaal minimaal op M in $\hat{u}$.

Dan geldt dat $\hat{u}$ voldoet aan de volgende, zogenaamde, *variatie-*(of *variati-*

onele-) ongelijkheid:

$$\langle J'(\hat{u}), v - \hat{u} \rangle \geq 0, \quad \forall v \in M.$$

2.4.3. OPMERKING

In het eenvoudigste geval dat $M = [\alpha,\beta] \subset \mathbb{R}$, $J \in C^1([\alpha,\beta])$ en J locaal minimaal is in $\hat{u}$, volgen uit de variatie-ongelijkheid de bekende eigenschappen: $J'(\hat{u}) \geq 0$ als $\hat{u} = \alpha$, $J'(\hat{u}) \leq 0$ als $\hat{u} = \beta$ en $J'(\hat{u}) = 0$ als $\hat{u} \in (\alpha,\beta)$.
Meer algemeen, als $\hat{u} \in M$ een inwendig punt is dan volgt uit de variatie-ongelijkheid het bekende resultaat $J'(\hat{u}) = 0$ door op te merken dat voor elke $w \in M$ en elke ε, met $|\varepsilon|$ voldoende klein, $v = \hat{u} + \varepsilon w \in M$.

2.4.4. Een *kegel* K is een verzameling met de eigenschap: als $u \in K$ dan is $\lambda u \in K$ voor alle $\lambda > 0$.
Een kegel K heet een *puntkegel* als $0 \in K$.
LEMMA. Zij M een convexe puntkegel. Dan is de variatie-ongelijkheid equivalent met

$$\langle J'(\hat{u}), \hat{u} \rangle = 0 \quad \text{èn} \quad \langle J'(\hat{u}), v \rangle \geq 0, \quad \forall v \in M.$$

BEWIJS: Neem achtereenvolgens $v = 0$ en $v = 2\hat{u}$ in de variatie-ongelijkheid. Dat levert $\langle J'(\hat{u}), \hat{u} \rangle = 0$ en $\langle J'(\hat{u}), v - \hat{u} \rangle = \langle J'(\hat{u}), v \rangle \geq 0$, $\forall v \in M$.

Zie verder Hoofdstuk 7 voor vatiatie-ongelijkheden.

2.4.5. EQUIVALENTIE VAN VARIATIEONGELIJKHEID EN MINIMALISERINGSPROBLEEM

STELLING.
Zij $M \subset V$ een convexe verzameling, $Q(u) = a(u) - \ell(u)$ een quadratische functionaal als in 2.2.9. Veronderstel dat a niet-negatief is op V, i.e.

$$a(u) \geq 0, \quad \forall u \in V.$$

Dan geldt: $\hat{u}$ voldoet aan de variatie-ongelijkheid

$$2a(\hat{u}, v - \hat{u}) - \ell(v - \hat{u}) \geq 0, \quad \forall v \in M$$

d.e.s.d.a. $\hat{u}$ een oplossing is van het min. pr. voor Q op M.

Bovendien, als a positief is op V, i.e. $a(u) > 0$, $\forall u \in V$, $u \neq 0$, dan is een eventuele oplossing *eenduidig*.

BEWIJS: Dit is een eenvoudig gevolg van 2.4.2 en de observatie dat

$$Q(v) = Q(\hat{u} + (v - \hat{u})) = Q(\hat{u}) + (Q'(\hat{u}), v - \hat{u}) + a(v - \hat{u})$$

met $(Q'(\hat{u}), v - \hat{u}) = 2a(\hat{u}, v - \hat{u}) - \ell(v - \hat{u})$ voor elke $v \in M$.

3. STATIONAIRITEITSVOORWAARDEN VOOR DICHTHEIDSFUNCTIONALEN

In de voorgaande paragraaf is gebleken dat in veel gevallen de stat. v.w. voor een stationair punt $\hat{u} \in M$ te herleiden is tot de vorm

(3.1) $\quad \langle J'(\hat{u}), \eta \rangle = 0$ voor alle $\eta \in V$ zodat $J'(\hat{u}) = 0$ in V^*.

Immers, dit is precies de voorwaarde opdat $\hat{u}$ een stationair punt is van de functionaal J op $M = V$. Maar ook voor stationaire punten van J op zekere deelverzamelingen $M \subset V$ geldt zo'n voorwaarde als we voor $J'(\hat{u})$ van hierboven lezen $L'(\hat{u}, \lambda_1, \ldots, \lambda_p)$ met L de Lagrange-functionaal bij J en M, en $\lambda_1, \ldots, \lambda_p$ zekere multiplicatoren, zoals voor verschillende gevallen gedefinieerd in 2.2.7 en 2.3.5.
In deze paragraaf doen we niets anders dan de voorwaarde dat $J'(\hat{u})$ de nul-functionaal is van V^* concretiseren voor in de toepassingen meest voorkomende gevallen. Voor concrete functionalen wordt de notatie van een en ander al gauw wat onoverzichtelijk. Daarom schetsen we nu slechts, als leidraad, de te volgen procedure in algemene termen.

V zal een ruimte zijn van functies gedefinieerd op een (open) gebied $\Omega \subset \mathbf{R}^n$, waarvan de rand, $\partial\Omega$, voldoende glad wordt verondersteld.
V is steeds een deelverzameling van $L_2(\Omega)$, de verzameling van (equivalentieklassen van) quadratisch Lebesgue-integreerbare functies op Ω. Dit introduceert het L_2 - inproduct, aan te geven met $(\ ,\ )_0$ en bijbehorende norm $\|\ \|_0$:

$$(f,g)_0 := \int_\Omega f(x).g(x)\, dx, \qquad \|f\|_0 := (f,f)^{\frac{1}{2}}.$$

Voorts zal V steeds de ruimte $C_0^\infty(\Omega)$ van willekeurig vaak differentieerbare functies met compacte drager in Ω omvatten; in het bijzonder betekent dit dat de elementen van V, mogelijk met uitzondering van (homogene) randvoorwaarden niet aan enige nevenvoorwaarde voldoen. Dus V is een ruimte van functies op Ω met $C_0^\infty(\Omega) \subset V \subset L_2(\Omega)$.

De 18$^{\underline{e}}$ eeuwse ideeën volgend (die veel later geformaliseerd zijn in distributietheorieën) redeneren we als volgt.

Het komt nogal eens voor, i.h.a. slechts voor een beperkte klasse van functies $\hat{u} \in V$, dat er een element $\frac{\delta J}{\delta u}(\hat{u})$ van $L_2(\Omega)$, of zelfs van $C^0(\Omega)$, bestaat zodanig dat

$$\delta J(\hat{u};\eta) = \langle J'(\hat{u}),\eta\rangle = (\tfrac{\delta J}{\delta u}(\hat{u}),\eta)_0 \quad \text{voor alle} \quad \eta \in C_0^\infty(\Omega). \tag{3.3}$$

Omdat $C_0^\infty(\Omega)$ dicht ligt in $L_2(\Omega)$ wordt $\frac{\delta J}{\delta u}(\hat{u})$, als element van $L_2(\Omega)$, hierdoor volledig bepaald. In die gevallen wordt $\frac{\delta J}{\delta u}(\hat{u})$ de *variatie-afgeleide* van J in het punt $\hat{u}$ genoemd.

Als zo'n $\frac{\delta J}{\delta u}(\hat{u})$ bestaat, definieer dan een uitdrukking R door

$$\langle J'(\hat{u}),\eta\rangle = (\tfrac{\delta J}{\delta u}(\hat{u}),\eta)_0 + R(\hat{u};\eta) \quad \text{voor alle} \quad \eta \in V. \tag{3.4}$$

(In concrete gevallen vinden we de uitdrukking (3.4) door "partiële integratie".) Voor R geldt dan

$$R(\hat{u};\eta) = 0 \quad \text{voor alle} \quad \eta \in C_0^\infty(\Omega).$$

Als, zoals vaak het geval is, functies uit V een voldoend glad gedrag hebben in de buurt van en op de rand $\partial\Omega$, dan hangt $R(\hat{u};\eta)$ alleen af van de restrictie tot $\partial\Omega$ van $\hat{u}$ en η, en eventuele afgeleiden.

Als de variatieafgeleide bestaat, impliceert de stat. v.w. (vanwege $C_0^\infty(\Omega) \subset V$):

$$(\tfrac{\delta J}{\delta u}(\hat{u}),\eta)_0 = 0 \quad \text{voor alle} \quad \eta \in C_0^\infty(\Omega). \tag{3.5}$$

Een fundamenteel lemma van Lagrange herleidt deze voorwaarde in geval $\frac{\delta J}{\delta u}(\hat{u})$ een continue functie is op Ω tot een puntsgewijze voorwaarde op Ω:

$$\tfrac{\delta J}{\delta u}(\hat{u}) = 0 \quad \text{op} \quad \Omega. \tag{3.6}$$

Dit is dan de befaamde *Euler-Lagrange* (of Euler-, of Lagrange-) *vergelijkin*
Hiervan gebruik makend volgt als restant van de stat. v.w.:

(3.7) $\qquad R(\hat{u};\eta) = 0 \quad$ voor alle $\quad \eta \in V.$

Afhankelijk van eventuele randvoorwaarden die al in de definitie van V zijn opgenomen, zal blijken dat (3.7) nog extra voorwaarden kan geven voor $\hat{u}$ of afgeleiden daarvan op de rand $\partial\Omega$, de zogenaamde *natuurlijke randvoorwaarden*.

Deze algemene ideeën worden hierna uitgewerkt voor verschillende concrete voorbeelden. In deze paragraaf behandelen we elk van de coördinaten x_i van $x = (x_1,\ldots,x_n) \in \Omega$ op gelijke wijze, in tegenstelling tot de volgende paragraaf waarin één coördinaat (de tijd) onderscheiden wordt van de andere, ruimtelijke, coördinaten.

3.1. LEMMA'S VAN LAGRANGE EN DU BOIS-REYMOND

Een (reëel-waardige) *functie* f op Ω is per definitie een voorschrift dat aan elke $x \in \Omega$ een getal f(x) toevoegt. Een fundamenteel andere beschrijvingswijze van f zou zijn om i.p.v. deze puntsgewijze karakterisering, de waarden $\int_\Omega f(x)\eta(x)\,dx$ voor te schrijven voor een grote klasse van (test-) functies η. In hoeverre de functie f hierdoor bepaald is wordt in de volgende lemma's onderzocht. (Het fundamentele idee om functies op te vatten als (continue) lineaire functionalen op zekere ruimte van testfuncties heeft geleid tot een generalisatie van het functiebegrip en wordt bestudeerd in de distributietheorie.)

3.1.1. De ruimte $C_0^\infty(\Omega)$ van willekeurig vaak differentieerbare functies met compacte drager in Ω bevat meer dan alleen de nulfunctie. In het bijzonder behoort voor elke $x_0 \in \Omega$ en $\delta > 0$ voldoende klein de radiaalsymmetrische functie $h(.;x_0,\delta)$ tot $C_0^\infty(\Omega)$, waarbij

$$h(x;x_0,\delta) = \begin{cases} \exp\left(-\dfrac{1}{\delta^2 - |x - x_0|^2}\right) & \text{voor } x \text{ met } |x - x_0| < \delta, \\ 0 \quad \text{voor } x \in \Omega \text{ met } |x - x_0| \geq \delta. \end{cases}$$

3.1.2. LEMMA VAN LAGRANGE

Zij f een functie op Ω waarvoor geldt $f \in C^0(\Omega)$ en $\int_\Omega f(x)\eta(x)\,dx = 0$ voor alle $\eta \in C_0^\infty(\Omega)$. Dan geldt $f(x) = 0$ voor elke $x \in \Omega$.

BEWIJS: Als er een punt $x_0 \in \Omega$ zou zijn met $f(x_0) \neq 0$, zeg $\alpha = f(x_0) > 0$, dan is er vanwege $f \in C^0(\Omega)$ een $\delta > 0$ zodanig dat $\{x \mid |x - x_0| < \delta\} \subset \Omega$ en $f(x) > \frac{1}{2}\alpha$ voor x met $|x - x_0| < \delta$. Dan geldt

$$\int_\Omega f(x)h(x;x_0,\delta)\,dx > \tfrac{1}{2}\alpha \int_\Omega h(x;x_0,\delta)\,dx > 0$$

in strijd met het gegeven omdat $h(.;x_0,\delta) \in C_0^\infty(\Omega)$.

3.1.3. LEMMA VAN DU BOIS-REYMOND

Laat $n = 1$ en neem, bijvoorbeeld, $\Omega = (0,1) \subset \mathbb{R}$. Zij f een functie op (0,1) waarvoor geldt

$$f \in C^0((0,1)) \text{ en } \int_0^1 f(x)\eta_x(x)\,dx = 0 \text{ voor alle } \eta \in C_0^\infty(0,1).$$

Dan is de functie f constant op (0,1).

BEWIJS: Voer in de functie $\hat{f}(x) = f(x) - \int_0^1 f(x)\,dx$.

Dan geldt, voor elke $r \in C_0^\infty$ met $\bar{r} \equiv \int_0^1 r(x)\,dx$, dat

$$\int_0^1 \hat{f}(x)[r(x) - \bar{r}]\,dx = \int_0^1 f(x)[r(x) - \bar{r}]\,dx$$

en dit is gelijk aan nul volgens het gegeven omdat $r(x) - \bar{r} = \eta_x(x)$ met $\eta(x) = \int_0^x [r(s) - \bar{r}]\,ds \in C_0^\infty$. Uit het lemma van Lagrange volgt dan dat $\hat{f}(x) = 0$ voor $x \in (0,1)$, en daaruit het resultaat.

3.2. ENKELVOUDIGE INTEGRAAL-PROBLEMEN

We beschouwen enkelvoudige integralen zoals geïntroduceerd in 2.1.6, en, voor notationeel gemak, bekijken we (als algemeen geval) alleen dichtheden waarin de hoogst voorkomende afgeleide van orde 1 is. Bovendien eisen we wat meer gladheid van de dichtheidsfunctie:

$$(3.8)\qquad \begin{cases} F \in C^2([0,1] \times \mathbb{R}^m \times \mathbb{R}^m) \\ J(u) = \int_0^1 F[u(x)]\,dx = \int_0^1 F(x,u(x),u_x(x))\,dx. \end{cases}$$

Ter herinnering, met $V = C^1([0,1],\mathbb{R}^m)$, geldt voor $u \in V$, $\eta \in V$:

$$(3.9)\qquad \delta J(u;\eta) = \int_0^1 \{F_u[u].\eta + F_v[u].\eta_x\}\,dx.$$

We gebruiken eerst het lemma van Lagrange om tot de Euler-Lagrange vergelijking te komen. Dat is voor dit 1-dimensionale probleem niet de beste manier, maar geeft wel precies de methode aan waarmee meervoudige integralen moeten worden aangepakt. Daartoe wordt de laatste term in de integrand van $\delta J(u;\eta)$ partieel geïntegreerd, met als resultaat (vergelijk met (3.4)):

$$(3.10)\qquad \delta J(u;\eta) = \int_0^1 \{F_u[u] - \frac{d}{dx}F_v[u]\}.\eta\,dx + F_v[u(x)].\eta(x)\Big|_{x=0}^{x=1}.$$

3.2.1. LEMMA

Veronderstel dat $u \in C^2((0,1),\mathbb{R}^m)$. Dan bestaat de variatie-afgeleide $\frac{\delta J}{\delta u}(u)$ en deze wordt gegeven door

$$\frac{\delta J}{\delta u}(u) = F_u[u] - \frac{d}{dx}F_v[u]$$

en is een continue (m-vector) functie op (0,1). Als geldt dat $\delta J(u;\eta) = 0$ voor alle $\eta \in C_0^\infty((0,1),\mathbb{R}^m)$, dan voldoet u aan de m Euler-Lagrange vergelijkingen

$$(3.11)\qquad F_u[u(x)] - \frac{d}{dx}F_v[u(x)] = 0 \text{ voor alle } x \in (0,1).$$

(Uitgeschreven:

$$F_{u_i}(x,u(x),u_x(x)) - \frac{d}{dx}F_{v_i}(x,u(x),u_x(x)) = 0, \quad i=1,2,\ldots,m;$$

dit zijn m, gekoppelde, $2^{\underline{e}}$ orde gewone differentiaal-vergelijkingen voor de m componenten $u_1,\ldots,u_m$ van de functie u.)

BEWIJS: Vanwege de aannamen $F \in C^2$ en $u \in C^2$ is de partiële integratie die leidt tot (3.10) gerechtvaardigd en is $\frac{\delta J}{\delta u}(u)$ een continue functie op (0,1).

Door beperking tot $\eta \in C_0^\infty$ verdwijnen de stoktermen in (3.10) en het lemma van Lagrange leidt tot de Euler-Lagrange vergelijkingen (direct als m = 1, en als m > 1 door voor η te nemen $\eta(x) = \alpha(x)e_i$ met $\alpha \in C_0^\infty$ scalarfuncties en e_i, $1 \leq i \leq m$, de basisvectoren in $\mathbf{R}^m$).

Door gebruik te maken van het lemma van du Bois-Reymond kunnen we de aanname $u \in C^2$ laten vervallen in bovenstaand lemma en toch concluderen dat $\frac{\delta J}{\delta u}(u)$ een continue functie is *als* u een stationair punt is van J(de variationele eigenschap van u impliceert wat extra regulariteit !).

3.2.2. LEMMA

Veronderstel dat $u \in C^1([0,1],\mathbf{R}^m)$ en $\delta J(u;\eta) = 0$ voor alle $\eta \in C_0^\infty$.
Dan geldt dat $\frac{\delta J}{\delta u}(u)$ bestaat en een continue functie is op (0,1), en dat u voldoet aan de m Euler-Lagrange vergelijkingen (3.11) op (0,1).

BEWIJS: We integreren nu de eerste (!) term in (3.9) partieel, met als resultaat:

$$\int_0^1 -\{G(x) + F_v[u(x)]\}.\eta_x \, dx + G(x).\eta(x)\Big|_{x=0}^{x=1} = 0 \quad \forall \eta \in C_0^\infty$$

waarin $G(x) = \int_0^x F_u[u(\xi)]\, d\xi$. Merk op dat G een C^1 - functie is en dat de term tussen accoladen een C^0 - functie is.
Vanwege $\eta(0) = \eta(1) = 0$ verdwijnt de stokterm. Het lemma 3.1.3 (toegepast op elk van de componenten als m > 1 zoals in 3.2.1) geeft dan de zogenaamde geïntegreerde vorm van de Euler-Lagrange vergelijkingen:

$$-G(x) + F_v[u(x)] = c \quad \text{voor elke } x \in (0,1)$$

voor zekere constante vector c. Omdat $c + G(x)$ een C^1 - functie is volgt hieruit dat ook $F_v[u(x)]$ een C^1 - functie is. We kunnen dit resultaat dus differentiëren en het resultaat is dan (3.11).

Tot zover de Euler-Lagrange vergelijkingen. Een onderzoek van vergelijking (3.7) die in dit geval luidt

$$F_v[u(x)].\eta(x)\Big|_{x=0}^{x=1} = 0 \text{ voor alle } \eta \in V \tag{3.12}$$

kan nog eventuele natuurlijke randvoorwaarden opleveren, en wel

$$F_{v_k}[u(\xi)] = 0$$

als voor $\eta \in V$ de k^{e} component van η in het punt $x = \xi$, met $\xi = 0$ of $\xi = 1$, *niet* is voorgeschreven.

Als voorbeeld van het bovenstaande geven we nu een sterk resultaat dat de equivalentie uitdrukt tussen een oplossing van een min. pr. en de oplossing van een randwaarde-probleem (r.w.p.) voor de Euler-Lagrange vergelijking.

3.2.3. TOEPASSING: SNAARVERGELIJKING EN 1-DIMENSIONALE DIFFUSIE

Beschouw de quadratische functionaal (zie 1.5.2 en 1.5.5)

$$Q(u) = \int_0^1 \{\tfrac{1}{2}\sigma(x)u_x^2 - f(x)u\}\, dx$$

met $f \in C^0(0,1)$, $\sigma \in C^1(0,1)$ en $\sigma(x) > 0$ voor $x \in [0,1]$, op een van de volgende verzamelingen (u_0 en u_1 zijn gegeven getallen; in het geval van diffusie, de voorgeschreven waarden van de concentraties in $x = 0$ en in $x = 1$, respectievelijk):

$$M_1 = \{u \in C^1([0,1]) \mid u(0) = u_0,\ u(1) = u_1\}$$

of

$$M_2 = \{u \in C^1([0,1]) \mid u(0) = u_0\}.$$

STELLING: Er geldt de volgende equivalentie: de functie $\hat{u} \in C^1([0,1])$ is de eenduidige oplossing van het min. pr. voor Q op M_1, resp. voor Q op M_2, dan en slechts dan als: de functie $\hat{u}$ is regulier, i.e. $\hat{u} \in C^2(0,1)$, en is de oplossing van de Euler-Lagrange vergelijking

(3.13) $$-(\sigma(x)u_x)_x = f(x) \text{ voor } x \in (0,1)$$

die voldoet aan de randvoorwaarden:

in geval M_1: $u(0) = u_0$, $u(1) = u_1$

in geval M_2: $u(0) = u_0$, $u_x(1) = 0$.

BEWIJS: Laat $\hat{u} \in C^1$ een oplossing zijn van $\inf \{Q(u) \mid u \in M_i\}$, $i = 1$ of 2. Dan geldt $\delta Q(\hat{u};\eta) = 0$ voor alle $\eta \in V_i$ met

$$V_1 = \{\eta \in C^1([0,1]) \mid \eta(0) = \eta(1) = 0\}$$

en

$$V_2 = \{\eta \in C^1([0,1]) \mid \eta(0) = 0\}.$$

Vanwege $\sigma > 0$ op $[0,1]$ is volgens 2.2.9 $\hat{u}$ de enige oplossing van het min. pr.. Uit 3.2.2 volgt dat $\sigma(x)u_x \in C^1$, zodat $\hat{u} \in C^2$, èn dat $\hat{u}$ voldoet aan (3.13). Voor het geval M_1 moet $\hat{u}$ voldoen aan beide voorgeschreven randvoorwaarden. Voor het geval M_2 moet $\hat{u}$ voldoen aan $u(0) = u_0$; de waarde $\eta(1)$, voor $\eta \in V_2$, is nu niet voorgeschreven, zodat uit (3.12) volgt: $\sigma(1)\hat{u}_x(1) = 0$. Omdat $\sigma(1) > 0$, volgt daaruit $u_x(1) = 0$ als natuurlijke randvoorwaarde.

Omgekeerd, laat $\hat{u} \in C^2$ voldoen aan (3.13) en de randvoorwaarden. Dan geldt zeker dat $\hat{u} \in M_i$. Neem nu het L_2 - inproduct van (3.13) met een willekeurige functie $\eta \in V_i$. Een partiële integratie, gebruik makend van de randwaarden voor $\hat{u}$ en η, levert dan $\delta Q(\hat{u};\eta) = 0$ voor alle $\eta \in V_i$. Dit betekent dat $\hat{u}$ een stationair punt is van Q op M_i, en dus, wegens 2.2.9, dat $\hat{u}$ de eenduidige oplossing is van het min. pr. van Q op M_i.

3.2.4. Ook voor dichtheden die hogere orde afgeleiden bevatten is door generalisatie van het lemma van du Bois-Reymond aan te tonen dat stationaire punten een extra regulariteitseigenschap hebben. Als specifiek voorbeeld beschouwen we kort de

BALKVERGELIJKING.

Zij $\sigma \in C^2(0,1)$ met $\sigma(x) > 0$ op $[0,1]$, $f \in C^0(0,1)$ en

$$Q(u) = \int_0^1 \{\tfrac{1}{2}\sigma(x)u_{xx}^2 - f(x)u\}\, dx.$$

Zowel voor de ingeklemde balk:

$$M_1 = \{u \in C^2([0,1]) \mid u(0) = u(1) = 0 = u_x(0) = u_x(1)\}$$

als voor de opgelegde balk:

$$M_2 = \{u \in C^2([0,1]) \mid u(0) = u(1) = 0\}$$

volgt dan uit $\hat{u} \in C^2([0,1])$ en $\delta Q(\hat{u};\eta) = 0$ voor alle $\eta \in V_i$ met $V_i \equiv M_i$ dat $\hat{u}$ een C^4 - functie is op $(0,1)$ die moet voldoen aan de $4^{\underline{e}}$ orde Euler-Lagrange vergelijking:

$$(\sigma(x)u_{xx})_{xx} = f(x) \quad \text{op} \quad (0,1).$$

Er is dan weer een equivalentie tussen de oplossing van het min. pr. voor Q op M_i en de oplossing van de Euler-Lagrange vergelijking die voldoet aan de voorgeschreven randvoorwaarden en, in geval M_2, nog aan de extra natuurlijke randvoorwaarden:

$$u_{xx}(0) = u_{xx}(1) = 0.$$

3.3. MEERVOUDIGE INTEGRAAL-PROBLEMEN

Als in de voorgaande sectie geven we voor dichtheidsfunctionalen op een gebied $\Omega \subset \mathbf{R}^n$ (zie 2.1.8) alleen algemene formules voor het geval

$$J(u) = \int_\Omega F[u(x)]\,dx = \int_\Omega F(x,u(x),\nabla u(x))\,dx$$

waarin $F \in C^2(\overline{\Omega} \times \mathbf{R} \times \mathbf{R}^n)$ een gegeven functie is.
In de uitdrukking voor de eerste variatie:

$$\delta J(u;\eta) = \int_\Omega \{F_u[u]\eta + F_v[u].\nabla\eta\}\,dx \tag{3.14}$$

willen we nu de laatste term "partieel integreren" om het lemma van Lagrange te kunnen gebruiken. Daartoe memoreren we eerst enkele standaardresultaten.

3.3.1. Laat $\alpha \in C^1(\Omega)$ een scalarfunctie zijn en $a = (a_1,\ldots,a_n) \in C^1(\Omega)$ een n-vectorfunctie.

(i) De *divergentie* van a, div a, wordt gedefinieerd door

$$\text{div } a(x) = \frac{\partial a_1}{\partial x_1}(x) + \frac{\partial a_2}{\partial x_2}(x) + \ldots + \frac{\partial a_n}{\partial x_n}(x).$$

(ii) Er geldt

$$\text{div } (\alpha(x)a(x)) = \alpha(x)\ \text{div } a(x) + a(x).\nabla\alpha(x).$$

(iii) Laat $\partial\Omega$ de (stuksgewijs gladde) rand zijn van Ω en, voor $x \in \partial\Omega$, $n(x)$ de naar buiten wijzende normaal op $\partial\Omega$. Dan geldt de *stelling van Gauss*:

$$\int_{\Omega} \operatorname{div} a(x)\, dx = \int_{\partial\Omega} a(x).n(x)\, d\sigma.$$

3.3.2. Als we hiervan gebruik maken en we schrijven

$$F_v.\nabla\eta = \operatorname{div}\{\eta F_v\} - \eta \operatorname{div} F_v$$

dan volgt formeel (i.e. als $F_v \in C^1$) voor (3.14):

$$\delta J(u;\eta) = \int_{\Omega} \{F_u[u] - \operatorname{div} F_v[u]\}\eta(x)\, dx + \int_{\partial\Omega} \eta(x)F_v[u(x)].n(x)\, d\sigma.$$

Vergeleken met (3.4) levert dit formeel als *Euler-Lagrange vergelijking* de tweede orde partiële differentiaal-vergelijking:

(3.15) $$F_u[u(x)] - \operatorname{div} F_v[u(x)] = 0 \quad \text{voor} \quad x \in \Omega;$$

eventuele natuurlijke nevenvoorwaarden moeten volgen uit (3.7), i.e.

(3.16) $$\int_{\partial\Omega} \eta(x)F_v[u(x)].n(x)\, d\sigma = 0 \quad \text{voor alle} \quad \eta \in V.$$

Als concreet voorbeeld beschouwen we het meer-dimensionale analogon van 3.2.3.

3.3.3. TOEPASSING: DIFFUSIE-PROBLEEM MET DIRICHLET-RANDVOORWAARDEN

Zij $\sigma \in C^1(\Omega)$ met $\sigma > 0$ op $\overline{\Omega}$, $f \in C^0(\Omega)$,

$$Q(u) = \int_{\Omega} \{\tfrac{1}{2}\sigma(x)|\nabla u(x)|^2 - f(x)u(x)\}\, dx$$

en

$$\phi \in C^0(\partial\Omega),$$

$$M = \{u \in C^1(\overline{\Omega}) \mid u(x) = \phi(x) \quad \text{voor} \quad x \in \partial\Omega\}.$$

STELLING

Een functie $u \in M \cap C^2(\Omega)$ is (de eenduidige) oplossing van het min. pr. voor Q op M dan en slechts dan als $u \in C^2(\Omega)$ voldoet aan de Euler-Lagrange vergelijking:

$$(3.17) \qquad -\operatorname{div}(\sigma(x)\nabla u(x)) = f(x) \quad \text{voor} \quad x \in \Omega$$

en aan de randvoorwaarde

$$(3.18) \qquad u(x) = \phi(x) \quad \text{voor} \quad x \in \partial\Omega.$$

BEWIJS: Omdat nu gegeven is dat $u \in C^2(\Omega)$, zodat $\sigma(x)\nabla u(x) \in C^1$, is het lemma van Lagrange bruikbaar en loopt het bewijs verder analoog aan dat van het 1-dimensionale geval 3.2.3.

3.3.4. OPMERKING

In verschillende gevallen, zoals hierboven met wat extra condities op f, kan op een directe manier bewezen worden dat het min. pr. een oplossing $u \in M$ heeft (zie §5). In dat geval geldt dan zeker

$$(3.19) \qquad \int_\Omega \{\sigma(x)\nabla u(x).\nabla\eta(x) - f(x)\eta(x)\}\, dx = 0$$

voor alle $\eta \in \{\eta \in C^1(\Omega) \mid \eta(x) = 0 \text{ voor } x \in \partial\Omega\}$. Echter, als $u \notin C^2(\Omega)$, of als we dat nog niet weten, dan kan (3.16) (nog) niet opgeschreven worden. We zeggen daarom wel: u is een *variationele* (of zwakke) *oplossing van het randwaardeprobleem* (3.17), (3.18) als: $u \in M$ en u aan (3.19) voldoet. Daarmee vinden we dan: $u \in M$ is oplossing van het min. pr. voor Q op M d.e.s.d.a. $u \in M$ een variationele oplossing is van (3.17), (3.18). Voorts, onder de extra regulariteitsaanname: $u \in M \cap C^2(\Omega)$ i.e. variationele oplossing van (3.17), (3.18) d.e.s.d.a. $u \in C^2(\Omega)$ oplossing is van (3.17), (3.18).

3.3.5. We modificeren nu het diffusieprobleem 3.3.3 door de Dirichlet-voorwaarde slechts op een deel $\partial\Omega_1$ van de rand $\partial\Omega$ voor te schrijven, zeg

$$u(x) = \phi(x) \quad \text{voor} \quad x \in \partial\Omega_1.$$

Omdat dan $\eta(x) = 0$ op $\partial\Omega_1$, en $\eta(x)$ willekeurig is op $\partial\Omega_1^* \equiv \partial\Omega \setminus \partial\Omega_1$, volgt uit (3.16) dat

$$\int_{\partial\Omega_1^*} \eta(x) \nabla u(x).n(x)\, d\sigma = 0.$$

Met het lemma van Lagrange, onder de aanname dat

$$\frac{\partial u}{\partial n}(x) \equiv \nabla u(x).n(x) \in C^0(\partial\Omega_1^*)$$

volgt dan dat u op $\partial\Omega_1^*$ moet voldoen aan de homogene Neumann-randvoorwaarde

$$\frac{\partial u}{\partial n}(x) = 0 \quad \text{voor} \quad x \in \partial\Omega_1^*$$

als natuurlijke randvoorwaarde.

3.3.6. Een verdere modificatie bestaat eruit dat we op een deel $\partial\Omega_1$ van de rand Dirichlet-voorwaarden willen voorschrijven en op het resterende deel *inhomogene Neumann-randvoorwaarden*:

$$u(x) = \phi(x) \quad \text{voor} \quad x \in \partial\Omega_1$$

(3.20)

$$\frac{\partial u}{\partial n}(x) = \psi(x) \quad \text{voor} \quad x \in \partial\Omega_1^*.$$

In plaats van deze inhomogene Neumann-voorwaarden op te nemen in het definitiegebied van Q, kan hetzelfde doel bereikt worden door een andere functionaal te bekijken, nl. de functionaal J die de som is van de quadratische functionaal Q uit 3.3.3 en een lineaire *rand-functionaal* over $\partial\Omega_1^*$:

$$J(u) = Q(u) - \int_{\partial\Omega_1^*} \sigma(x)\psi(x)u(x)\, d\omega$$

(met $d\omega$ als oppervlakte-element i.p.v. $d\sigma$ om verwarring te voorkomen). Voor een stationair punt u van J op $M_1 = \{u \in C^1(\Omega) \mid u(x) = \phi(x) \text{ op } \partial\Omega_1\}$ geldt dan

$$\int_{\Omega} \{\sigma(x)\nabla u(x).\nabla\eta(x) - f(x)\eta(x)\}\, dx - \int_{\partial\Omega_1^*} \sigma(x)\psi(x)\eta(x)\, d\omega = 0$$

voor alle $\eta \in C^1(\Omega)$ met $\eta(x) = 0$ op $\partial\Omega_1$.
Weer onder de aanname dat dit stationaire punt regulier, C^2, is, vinden we dan

$$\int_{\Omega} \{- \operatorname{div} (\sigma(x)\nabla u(x)) - f(x)\}\eta(x)\, dx + \int_{\partial\Omega_1^*} \sigma(x)\{\nabla u(x).n(x)-\psi(x)\}\eta(x)\, d\omega=0.$$

Daaruit volgt dan eenvoudig de equivalentie tussen een eenduidige oplossing van het randwaardeprobleem (3.17), (3.20) en een eenduidige, reguliere, oplossing van het minimaliseringsprobleem voor J op M_1.

4. VARIATIONELE DYNAMISCHE SYSTEMEN

In deze paragraaf behandelen we voorbeelden van variatieprincipes die de evolutie beschrijven van veel dynamische systemen, zowel systemen uit de klassieke mechanica bestaande uit een eindig aantal puntmassa's (discrete systemen) alsook van systemen uit de continuum-mechanica bestaande uit een continuum van deeltjes (continue systemen).
De optredende functionalen zijn in wezen van dezelfde soort als die behandeld in de voorgaande paragraaf (i.e. dichtheidsfunctionalen), maar de speciale rol die één van de onafhankelijke variabelen, nl. de fysische tijd t, speelt geeft reden voor een aparte behandeling met vermelding van de specifieke naamgeving en de te bezigen notatie.

4.1. PRINCIPE VAN STATIONAIRE ACTIE

Beschouw een systeem waarvan de "toestand" op elk tijdstip t, gedurende zeker tijdsinterval (t_1,t_2), beschreven kan worden door een element z(t) uit een of andere toestandsruimte Z. (We gaan hier niet in op de vraag wat verstaan kan worden onder de toestand van een gegeven systeem. De beschrijving $t \to z(t)$ zal voor ons, per definitie, betekenen dat het systeem op elk tijdstip volledig gekarakteriseerd is.)
Van alle mogelijke evoluties van het systeem, i.e. van de familie van banen $t \to z(t)$ in de toestandsruimte Z, zijn we geïnteresseerd in die banen die het systeem in werkelijkheid, i.e. tengevolge van fysische wetten, zal kunnen doorlopen. Ter onderscheiding noemen we deze laatste banen de (werkelijke, of) fysische evolutie (beweging) van het systeem.

Variationele dynamische systemen worden gekarakteriseerd door het feit dat de fysische wet die de beweging van zo'n systeem bepaalt, geformuleerd kan worden m.b.v. een *actie-principe*. In algemene termen geformuleerd luidt dit variatieprincipe:

PRINCIPE VAN STATIONAIRE ACTIE: Er is een functionaal S, de zogenaamde *actie-functionaal*, die gedefinieerd is op de verzameling van banen in de toestandsruimte Z, zodanig dat stationaire punten van S (op nader te specificeren deelverzamelingen M van evoluties) corresponderen met de fysische evoluties van het systeem.

In sommige gevallen corresponderen de fysische evoluties met minimale elementen van S op M; in die gevallen kan men spreken van een principe van minimale actie. In het algemeen is de benaming stationaire actie meer gepast.

OPMERKING

Ter beschrijving van de fysische evolutie gedurende zeker tijdsinterval $[t_1,t_2]$ zal de verzameling M i.h.a. bestaan uit evoluties die, behalve aan eventuele andere nevenvoorwaarden, ten tijde t_1 en/of t_2 tenminste gedeeltelijk voorgeschreven eigenschappen hebben. In veel leerboeken worden deze randvoorwaarden niet expliciet beschouwd. Het min of meer expliciet beschreven actie-principe komt er dan op neer dat, bij afwezigheid van extra nevenvoorwaarden, een fysische evolutie correspondeert met een functie $\hat{z}(t)$ die voldoet aan

$$\frac{\delta S}{\delta z}(\hat{z}(t)) = 0$$

(de formele variatie-afgeleide), i.h.a. een gewone of partiële differentiaal-vergelijking: de "evolutie-vergelijking".

De reden voor de vaagheid hierover is misschien begrijpelijk: de aard van de evolutie-vergelijkingen is namelijk zodanig dat het *rand*-waardeprobleem een niet goed gesteld probleem is: het willekeurig voorschrijven van randvoorwaarden in t_1 en t_2 zal i.h.a. betekenen dat er geen oplossing bestaat ! (Dit in tegenstelling tot de algemene situatie bij *begin*-voorwaarden: dan worden op één tijdstip, zeg t_1, voorwaarden gegeven en meestal bestaat

er dan een oplossing op zeker tijdsinterval $[t_1,t_2]$ mits $t_2 - t_1$ voldoende klein is.)

De remedie is dan ook om het actie-principe niet te beschouwen als een formulering waarmee de *existentie* van een fysische evolutie wordt uitgesproken. Een deel van de formulering dient als volgt gelezen te worden: als $\hat{z}(t)$ een fysische evolutie is, waarvan, behalve de existentie, niets anders bekend is dan zekere eigenschappen in t_1 en $t_2 > t_1$, dan wordt $\hat{z}$ gekarakteriseerd (i.h.b. voor $t \in (t_1,t_2)$) door de eigenschap dat $\hat{z}$ een stationair punt van S is op de verzameling evoluties die diezelfde eigenschappen hebben op t_1 en t_2. (Het andere deel van de formulering geeft de omkering: als er een stationair punt is van S bij zekere voorwaarden in t_1 en t_2 dan is dit een fysische evolutie.)

We beschrijven nu de twee bekendste klassen van variationele dynamische systemen, nl. *Lagrangiaanse* en *Hamiltonse systemen*. Dit zijn geen elkaar uitsluitende klassen van systemen; integendeel, veel systemen kunnen, afhankelijk van wat gekozen wordt als de toestand van het systeem, zowel als een Lagrangiaans alsook als een Hamiltons systeem worden opgevat. In die gevallen spreken we dan over de *Lagrange-* resp. *Hamilton-formulering* van zo'n systeem. Het verband tussen beide beschrijvingswijzen is in die gevallen gebaseerd op convexiteitseigenschappen en wordt gegeven door een Legendre transformatie.

4.2. LAGRANGIAANSE SYSTEMEN

4.2.1. In de eenvoudigste gevallen is de toestand van een Lagrangiaans systeem, de *positie* van dat systeem, beschreven in zeker coördinatenstelsel. In plaats van Z en z(t) is de gebruikelijke notatie Q en q(t), waarin Q de zogenaamde *configuratie-ruimte* is. De actiefunctionaal is van de vorm

$$S(q) = \int_{t_1}^{t_2} L[q(t)]\, dt$$

waarin L[q(t)], de *Langrangiaan* van het systeem, een bekende functie is van zijn argumenten t, q(t) en afgeleiden van q naar t: $\dot{q}(t) = \frac{dq}{dt}(t)$, $\ddot{q}$, ... Wij zullen steeds te maken hebben met systemen waarvoor $\dot{q}$ de hoogst

voorkomende afgeleide is:

$$L[q(t)] = L(t,q(t),\dot{q}(t))$$

met $L \in C^2([t_1,t_2] \times TQ)$, waarin TQ de raakruimte van Q is:

$$TQ = \{(q,v) \mid q \in Q,\ v \in T_qQ\}.$$

Het is voorts gebruikelijk te veronderstellen dat de functie $v \to L(t,q,v)$ niet (positief) homogeen van de graad 1 is, maar dat, in het bijzonder, $L_{vv}(t,q,v) \neq 0$, voor alle voorkomende (t,q,v).

Voor een Lagrangiaans systeem met actiefunctionaal S luidt het *actieprincipe*: Laat q_1 en q_2 twee gegeven toestanden in Q zijn en $t_1 < t_2$. Beschouw alle mogelijke C^1 - banen in de configuratieruimte van q_1 (ten tijde t_1) naar q_2 (ten tijde t_2) geparametriseerd met t:

$$M = \{q \in C^1([t_1,t_2],Q) \mid q(t_1) = q_1,\ q(t_2) = q_2\}.$$

Dan geldt: $\hat{q}$ is een fysische evolutie van het systeem waarvoor $\hat{q} \in M$, dan en slechts dan als $\hat{q}$ een stationair punt is van S op M.

4.2.2. DISCRETE SYSTEMEN

Voor discrete systemen is de configuratieruimte Q een deelverzameling van $\mathbf{R}^N$ voor zekere $N \geq 1$. Als $Q = \mathbf{R}^N$, is N het aantal *vrijheidsgraden*, i.e. het minimale aantal positie-coördinaten die nodig zijn om de posities van alle deeltjes van het systeem te beschrijven.
Bijvoorbeeld, voor een systeem bestaande uit n deeltjes waarvan elk deeltje, onafhankelijk van de andere, elke positie in de ruimte ($\mathbf{R}^3$) kan innemen, is $N = 3n$ en $q(t) = (\underline{x}_1(t),\ldots,\underline{x}_n(t)) \in \mathbf{R}^{3n}$ geeft de positie van het totale systeem als $\underline{x}_i(t) \in \mathbf{R}^3$ de positie is van het i-de deeltje.
In dat geval is de Lagrangiaan een functie op $[t_1,t_2] \times \mathbf{R}^N \times \mathbf{R}^N$:

$$(t,q,v) \to L(t,q,v).$$

Uit het actieprincipe volgt dat de fysische evolutie van het systeem beschreven wordt door de Euler-Lagrange vergelijking bij S, i.e. door de N $2^{\underline{e}}$ orde gewone differentiaal-vergelijkingen voor q(t):

$$- \frac{d}{dt}\left(\frac{\partial L}{\partial v}[q(t)]\right) + \frac{\partial L}{\partial q}[q(t)] = 0$$

waarin weer $[q(t)] = (t,q(t),\dot{q}(t))$.

4.2.3. Voor veel Lagrangiaanse systemen, de zgn. *natuurlijke systemen*, is de Lagrangiaan het verschil tussen kinetische energie en potentiële energie. Neem als voorbeeld de n puntmassa's bewegend in $\mathbf{R}^3$ waarvan de posities $\underline{x}_i$ beschreven worden met Cartesische coördinaten. Als m_i (> 0) de massa is van het i-de deeltje, dan is de totale kinetische energie van het systeem

$$\tfrac{1}{2}\dot{q}(t).Mq(t) = \sum_{i=1}^{n} \tfrac{1}{2}m_i|\dot{\underline{x}}_i|^2$$

waarin M, de massa-matrix, de 3n × 3n diagonaal matrix is die gegeven wordt door $M = \text{diag}\,(m_1,m_1,m_1,m_2,\ldots,m_n,m_n,m_n)$.
Veronderstel dat de potentiële energie $V = V(t,q)$ een gegeven functie van de tijd en de positie is: $V \in C^1([t_1,t_2] \times \mathbf{R}^N)$.
Dan wordt de actiefunctionaal voor dit systeem gegeven door

$$S(q) = \int_{t_1}^{t_2} \{\tfrac{1}{2}\dot{q}(t).M\dot{q}(t) - V(t,q(t))\}\,dt$$

en de bewegingsvergelijkingen zijn de Euler-Lagrange vergelijkingen hiervan:

$$- M\ddot{q}(t) = \frac{\partial V}{\partial q}(t,q(t)).$$

Zonodig na het herschrijven van deze 3n vergelijkingen als een n-tal vergelijkingen voor de posities $\underline{x}_i$:

$$- m_i\ddot{\underline{x}}_i(t) = \frac{\partial V}{\partial \underline{x}_i}(t,\underline{x}_1,\underline{x}_2,\ldots,\underline{x}_n), \qquad i = 1,2,\ldots,n$$

herkennen we deze bewegingsvergelijkingen als de *wet van Newton* voor de beweging van de massapunten onder invloed van de kracht $F = -\frac{\partial V}{\partial q}$.

4.2.4. SLINGERVERGELIJKING

Een speciaal geval van een natuurlijk systeem is de slingerbeweging van een massapunt m bewegend in een vlak, onder invloed van de zwaartekracht en verbonden aan een vast ophangpunt met een massaloos koord van constante

lengte ℓ. Dit voorbeeld dient ook om te illustreren dat de toestand niet noodzakelijk met Cartesische coördinaten beschreven hoeft te worden. Laat nl. $\phi(t)$ de hoek van uitwijking zijn t.o.v. de verticale ruststand. De kinetische energie is dan $\frac{1}{2}m(\ell\dot{\phi})^2$ en de potentiële energie (met de arbitraire keuze $V(\phi) = 0$ voor $\phi = 0$) wordt gegeven door $mg\ell(1 - \cos\phi)$. De actiefunctionaal

$$S(\phi) = \int_{t_1}^{t_2} \{\tfrac{1}{2}m\ell^2\dot{\phi}^2 - mg\ell(1 - \cos\phi)\}\, dt$$

geeft dan via het actieprincipe als bewegingsvergelijking de bekende slingervergelijking:

$$\ddot{\phi}(t) + \omega^2 \sin\phi(t) = 0 \quad \text{met} \quad \omega^2 = \frac{g}{\ell}.$$

OPMERKING

Voor kleine uitwijkingen ($|\phi(t)| << 1$ voor alle t) wordt deze vergelijking vaak benaderd (ingegeven door $\sin\phi \approx \phi$ voor $|\phi|$ klein) door:

$$\ddot{\phi} + \omega^2\phi = 0.$$

Merk op dat deze linearisering van de vergelijking ook verkregen kan worden door "quadratisering" van de actiefunctionaal, nl. door benadering van $(1 - \cos\phi)$ door $\frac{1}{2}\phi^2$.

4.2.5. CONTINUE SYSTEMEN

Voor continue systemen is de configuratieruimte Q zelf een (deelverzameling van een) functieruimte: voor elke t wordt de toestand (positie) van het systeem beschreven door een functie van zekere ruimte-coördinaten x, zeg $x \in \Omega$ met $\Omega \subset \mathbf{R}^m$. ($x \in \Omega$ speelt de rol van continue index als we dat vergelijken met de discrete index $i \in \{1,2,\ldots,N\}$ voor q_i van een discreet systeem van N vrijheidsgraden.) Als we de positie beschrijven met u(x,t), dan is de Lagrangiaan $L[u(.,t)]$ nu zelf een functionaal die afhangt van t, $u(.,t)$ en $u_t(.,t)$ (en eventuele hogere orde afgeleiden naar t). In de gevallen die wij bekijken is L een dichtheidsfunctionaal over Ω:

$$L[u(.,t)] = L(t,u(.,t),u_t(.,t)) = \int_\Omega \mathcal{L}[u(x,t)]\, dx$$

waarin $L[u]$ de *Lagrange-dichtheid* is, een uitdrukking in t, x, u(x,t) en partiële afgeleiden van u en u_t naar x: u_x, u_t, u_{xx}, u_{tx},... .
De Euler-Lagrange vergelijking kan in dat geval geschreven worden als

$$-\frac{\partial}{\partial t}\left(\frac{\delta L}{\delta v}[u(.,t)]\right) + \frac{\delta L}{\delta u}[u(.,t)] = 0$$

waarin we voor $L = L(t,u(.,t),v(.,t))$ de variatie-afgeleiden naar u en naar v gebruiken die in paragraaf 3 zijn gedefinieerd.
Voor het speciale geval dat

$$L = L(t,x,u,u_x,u_t) = L[u(x,t)]$$

worden deze variatie-afgeleiden gegeven door (met voor de hand liggende, maar wat slordige, notatie):

$$\frac{\delta L}{\delta v} = \frac{\partial L}{\partial u_t}, \qquad \frac{\delta L}{\delta u} = \frac{\partial L}{\partial u} - \frac{\partial}{\partial x}\left(\frac{\partial L}{\partial u_x}\right)$$

zodat de Euler-Lagrange vergelijking luidt:

$$-\frac{\partial}{\partial t}\left(\frac{\partial L}{\partial u_t}[u(x,t)]\right) - \frac{\partial}{\partial x}\left(\frac{\partial L}{\partial u_x}[u(x,t)]\right) + \frac{\partial L}{\partial u}[u(x,t)] = 0.$$

(Merk op dat dit resultaat analoog is aan (3.15) als we de set (t,x) opvatten als de onafhankelijke variabelen met bijbehorende $\nabla = \left(\frac{\partial}{\partial t}, \frac{\partial}{\partial x}\right)$.)

4.2.6. DYNAMICA VAN SNAREN

Zij u(x,t) de uitwijking in een plat vlak, loodrecht op de x-as, van een langs de x-as gespannen snaar ter plaatse x, zeg $x \in (0,\ell)$, ten tijde t. Als de snaar vastgebonden is aan de x-as in x = 0 en $x = \ell$, dus $u(0,t) = u(\ell,t) = 0$ voor alle t, dan is $Q = \{y \in C^1([0,\ell]) \mid y(0) = y(\ell) = 0\}$ en de Lagrangiaan van dit natuurlijke systeem wordt gegeven (in de benadering voor kleine uitwijkingen) door

$$L[u(.,t)] = \int_0^\ell \{\tfrac{1}{2}\rho(x)u_t^2 - \tfrac{1}{2}\sigma(x)u_x^2\}\, dx.$$

Hierin is

1) $\rho(x)$ de massadichtheid van de snaar ter plaatse x;
2) $\int_0^\ell \tfrac{1}{2}\rho(x)u_t^2\, dx$ de totale kinetische energie;

3) $\sigma(x)$ (> 0) een maat voor de spankracht in de snaar en

4) $\int_0^\ell \frac{1}{2}\sigma(x)u_x^2\,dx$ de totale potentiële energie.

De bewegingsvergelijking die volgt uit het actieprincipe luidt:

$$\rho(x)u_{tt}(x,t) = (\sigma(x)u_x(x,t))_x, \quad x \in (0,\ell).$$

Deze partiële differentiaal-vergelijking staat bekend als de *golfvergelijking*. (In het speciale geval dat σ en ρ constant zijn, en $c^2 := \frac{\sigma}{\rho}$, is

$$u(x,t) = f(x - ct) + g(x + ct)$$

voor willekeurige functies f en g een oplossing van deze vergelijking. De functie $f(x - ct)$, resp. $g(x + ct)$, is te interpreteren als een zich met constante snelheid c, onvervormd naar rechts, resp. links, voortplantende "golf".)

4.2.7. DYNAMICA VAN BALKEN

De beweging van een balk wordt voor dezelfde configuratie als voor de snaar hierboven, gevonden uit de Lagrangiaan

$$L[u(.,t)] = \int_0^\ell \{\tfrac{1}{2}\rho(x)u_t^2 - \tfrac{1}{2}\sigma(x)u_{xx}^2\}\,dx.$$

De bewegingsvergelijking luidt nu

$$\rho(x)u_{tt}(x,t) = -(\sigma(x)u_{xx})_{xx}, \quad x \in (0,\ell);$$

als alleen de randvoorwaarden $u(0,t) = u(\ell,t) = 0$ worden voorgeschreven (opgelegde balk), volgen uit de variationele formulering bovendien nog de natuurlijke randvoorwaarden

$$u_{xx}(0,t) = u_{xx}(\ell,t) = 0.$$

4.3. HAMILTONSE SYSTEMEN

4.3.1. Voor een Hamiltons systeem wordt de toestand veelal beschreven door een paar variabelen: $z(t) = (q(t),p(t)) \in Z$; Z heet nu de *faseruimte*. In de eenvoudigste gevallen is q(t) te interpreteren als de *positie* (zoals bij een Lagrangiaans systeem) en p(t) als de *impuls* van het systeem ten tijde t.

In die gevallen is Z de coraakruimte van een configuratie-ruimte Q: $Z=T^*Q$, en heten $q \in Q$ en $p \in T^*_qQ$ een paar *kanonieke* (of kanoniek geconjugeerde) *variabelen*. Voor elke gladde baan $t \to (q(t),p(t))$ in de faseruimte is dan de uitdrukking $\langle p(t),\dot{q}(t)\rangle$ goed gedefinieerd voor elke t.
De (*kanonieke*) *actiefunctionaal* is van de vorm:

$$S(q,p) = \int_{t_1}^{t_2} \{\langle p(t),\dot{q}(t)\rangle - H[q(t),p(t)]\}\, dt$$

waarin H[q(t),p(t)], de *Hamiltoniaan* van het systeem, een bekende functie is van zijn argumenten t, q(t) en p(t). (Merk op dat in H geen afgeleiden van q en/of p naar t voorkomen.) Voor zo'n Hamiltons systeem luidt het (*kanonieke*) *actieprincipe*:

> Laat q_1 en q_2 twee gegeven toestanden zijn in Q, en $t_1 < t_2$. Beschouw in de faseruimte de volgende verzameling banen:
>
> $$M = \{(q,p) \in C^1([t_1,t_2],T^*Q) \mid q(t_1) = q_1,\ q(t_2) = q_2\}.$$
>
> Dan geldt: $(\hat{q},\hat{p})$ is een fysische evolutie van het systeem waarvoor $(\hat{q},\hat{p}) \in M$, d.e.s.d.a. $(\hat{q},\hat{p})$ een stationair punt is van S op M.

4.3.2. DISCRETE SYSTEMEN

Voor een discreet systeem met N vrijheidsgraden is de faseruimte $\mathbf{R}^{2N}$ 2N-dimensionaal, en de Hamiltoniaan is een gewone functie van t, q en p: $H \in C^1([t_1,t_2] \times \mathbf{R}^N \times \mathbf{R}^N)$. De actiefunctionaal luidt in dat geval

$$S(q,p) = \int_{t_1}^{t_2} \{p(t).\dot{q}(t) - H(t,q(t),p(t))\}\, dt$$

Door gebruik te maken van het feit dat met het lemma van du Bois-Reymond extra regulariteit voor stationaire punten bewezen kan worden, kunnen we in dit geval de verzameling M iets groter laten zijn door alleen $p \in C^0$ te eisen:

$$M = \{(q,p) \in C^1([t_1,t_2],\mathbf{R}^N) \times C^0([t_1,t_2],\mathbf{R}^N) \mid q(t_1)=q_1, q(t_2)=q_2\}.$$

Het actieprincipe leidt tot de volgende set van 2N eerste orde gewone differentiaal-vergelijkingen voor de functies q(t) en p(t), de zgn. *Hamilton-vergelijkingen*:

$$\dot{q}(t) = \frac{\partial H}{\partial p}(t,q(t),p(t))$$

$$-\dot{p}(t) = \frac{\partial H}{\partial q}(t,q(t),p(t)).$$

VOORBEELDEN

Laat

$$H(t,q,p) = \tfrac{1}{2}p.M^{-1}p + V(t,q)$$

waarin M^{-1} de inverse is van de massamatrix en V de potentiële energie als in 4.2.3. Dan zijn de Hamilton-vergelijkingen:

$$\dot{q}(t) = M^{-1}p(t)$$

$$-\dot{p}(t) = \frac{\partial V}{\partial q}(t,q(t)).$$

Merk op dat na eliminatie van de functie p(t) uit deze vergelijkingen, dezelfde bewegingsvergelijking voor q(t) resulteert als in 4.2.3.
Op analoge manier kan de slingerbeweging uit 4.2.4 beschreven worden m.b.v. de Hamiltoniaan

$$H(q,p) = \frac{1}{2m\ell^2}p^2 + mg\ell(1 - \cos q)$$

door eliminatie van p, en te zetten $q(t) = \phi(t)$, in de Hamilton vergelijkingen:

$$\dot{q}(t) = \frac{1}{m\ell^2}p(t)$$

$$-\dot{p}(t) = mg\ell \sin q(t).$$

4.3.3. CONTINUE SYSTEMEN

De faseruimte is nu zelf weer een (deel van een) functieruimte, en de kanonieke variabelen q en p zijn nu (scalar-, voor het gemak) functies van t en van de ruimtelijke variabelen $x \in \Omega$: $q = q(x,t)$, $p = p(x,t)$.

De Hamiltoniaan $H[q(.,t),p(.,t)]$ is dan een functionaal, in veel gevallen een dichtheidsfunctionaal van de vorm

$$H[q(.,t),p(.,t)] = \int_{\Omega} \mathcal{H}[q(x,t),p(x,t)]\, dx$$

waarin $\mathcal{H}[q,p]$, de *Hamilton-dichtheid*, een functie is van t, x, q(x,t), p(x,t en partiële afgeleiden van q en p naar x: q_x, p_x, q_{xx}, ...
De actiefunctionaal luidt

$$S(q,p) = \int_{t_1}^{t_2} dt \{\int_{\Omega} p(x,t)q_t(x,t)\, dx - H[q(.,t),p(.,t)]\} =$$

$$= \int_{t_1}^{t_2} dt \int_{\Omega} dx \{p(x,t)q_t(x,t) - \mathcal{H}[q(x,t),p(x,t)]\}.$$

Gebruik makend van variatie-afgeleiden, leidt het kanoniek actieprincipe tot de Hamilton-vergelijkingen:

$$q_t(.,t) = \frac{\delta H}{\delta p}[q(.,t),p(.,t)]$$

$$-p_t(.,t) = \frac{\delta H}{\delta q}[q(.,t),p(.,t)].$$

Bijvoorbeeld, als $\mathcal{H} = \mathcal{H}(t,x,q,p,q_x)$, dan is

$$\frac{\delta H}{\delta p} = \frac{\partial \mathcal{H}}{\partial p} \quad \text{en} \quad \frac{\delta H}{\delta q} = \frac{\partial \mathcal{H}}{\partial q} - \frac{\partial}{\partial x}\left(\frac{\partial \mathcal{H}}{\partial q_x}\right).$$

VOORBEELDEN

De beweging van een snaar (4.2.6) en van een balk (4.2.7) is ook te beschrijven als een Hamilton-systeem, zoals blijkt door te zetten u(x,t) = q(x,t) en door eliminatie van de functie p(x,t) uit de volgende Hamilton-vergelijkingen:

voor snaar:

met: $\mathcal{H} = \frac{1}{2\rho(x)}\, p^2 + \frac{1}{2}\sigma(x)q_x^2$:

$$q_t(x,t) = \frac{1}{\rho(x)}\, p(x,t)$$

$$-p_t(x,t) = -(\sigma(x)q_x)_x$$

voor balk:

met: $H = \frac{1}{2\rho(x)}\, p^2 + \frac{1}{2}\sigma(x) q_{xx}^2$:

$$q_t(x,t) = \frac{1}{\rho(x)}\, p(x,t)$$

$$-p_t(x,t) = (\sigma(x) q_{xx})_{xx}.$$

4.4. EQUIVALENTIE VAN LAGRANGE- EN HAMILTON-FORMULERING

We laten in deze sectie zien dat als de Lagrangiaan van een Lagrangiaans systeem aan zekere convexiteitseigenschappen voldoet, dit systeem dan ook een Hamiltons systeem is. De Hamiltoniaan volgt dan uit de Lagrangiaan middels een Legendre-transformatie. Het omgekeerde, dat een Hamiltons systeem waarvan de Hamiltoniaan zekere convexiteitseigenschappen heeft, ook een Lagrangiaans systeem is, kan op analoge manier bewezen worden. Voorbeelden hiervan zijn al gegeven in 4.3.2 en 4.3.3.
Voor de eenvoud van presentatie beperken we ons tot discrete systemen met configuratieruimte $\mathbf{R}^N$. Er kan dan gebruik gemaakt worden van de resultaten over Fenchel- en Legendre-transformatie zoals beschreven in 5.2.6.

4.4.1. Beschouw de actiefunctionaal van een discreet Lagrangiaans systeem met N vrijheidsgraden:

$$S(q) = \int_{t_1}^{t_2} L(t,q(t),\dot{q}(t))\, dt$$

$$M = \{q \in C^1([t_1,t_2],\mathbf{R}^N) \mid q(t_1) = q_1,\ q(t_2) = q_2\}.$$

We veronderstellen dat de Lagrangiaan voldoet aan:

(a) $L \in C^2([t_1,t_2] \times \mathbf{R}^N \times \mathbf{R}^N)$

(b) voor elke $(t,q) \in [t_1,t_2] \times \mathbf{R}^N$ geldt dat de functie

$\mathbf{R}^N \ni v \to L(t,q,v)$

(i) strikt convex is op $\mathbf{R}^N$

(ii) superlineair is in het oneindige, i.e.

$$\frac{L(t,q,v)}{|v|} \to \infty \quad \text{voor} \quad |v| \to \infty.$$

Definieer dan de functie H: $[t_1,t_2] \times \mathbb{R}^N \times \mathbb{R}^N$ door

$$H(t,q,p) := \max_{v\in\mathbb{R}^N} [p.v - L(t,q,v)].$$

Volgens 5.2.6 is H, de Fenchel-transformatie van L m.b.t. de variabele v, eindig gedefinieerd, en wordt ook gegeven als de Legendre-transformatie:

$$H(t,q,p) = p.v - L(t,q,v) \text{ waarin } v \text{ zo dat } p = L_v(t,q,v).$$

Voorts geldt:

$$L(t,q,v) = \max_{p\in\mathbb{R}^N} [p.v - H(t,q,p)] =$$

$$= p.v - H(t,q,p) \quad \text{waarin } p \text{ zo dat } v = H_p(t,q,p).$$

4.4.2. BEWERING

Het Hamilton-systeem met Hamiltoniaan H zoals hierboven gedefinieerd, is equivalent met het oorspronkelijk gegeven Lagrangiaanse systeem met Lagrangiaan L, i.e. de bewegingsvergelijkingen:

$$-\frac{\partial}{\partial t}(L_v(t,q(t),\dot{q}(t))) + L_q(t,q(t),\dot{q}(t)) = 0$$

en

$$\dot{q}(t) = \frac{\partial H}{\partial p}(t,q(t),p(t))$$

$$-\dot{p}(t) = \frac{\partial H}{\partial q}(t,q(t),p(t))$$

zijn equivalent via de transformatie

$$p(t) = L_v(t,q(t),\dot{q}(t)) \quad \text{of} \quad \dot{q}(t) = H_p(t,q(t),p(t)).$$

OPMERKING

Het standaardbewijs van dit resultaat is een directe verificatie dat de ene set bewegingsvergelijkingen volgt uit de andere via de gegeven transformatie tussen p(t) en $\dot{q}(t)$. In plaats hiervan zullen we een direct bewijs geven via de respectievelijke actieprincipes door gebruik te maken van de Fenchel-karakteriseringen van L en H.

4.4.3. BEWIJS: We zullen aantonen dat voor elke $q \in C^1([t_1,t_2], \mathbb{R}^N)$ geldt:

$$S(q) = \sup_{p\in C^0(t_1,t_2)} \tilde{S}(q,p) = \operatorname{stat}_{p\in C^0(t_1,t_2)} \tilde{S}(q,p)$$

waarin $\tilde{S}(q,p)$ de kanonieke actiefunctionaal is bij H:

$$\tilde{S}(q,p) = \int_{t_1}^{t_2} \{p(t).\dot{q}(t) - H(t,q(t),p(t))\}\, dt.$$

Daaruit volgt dan dat stationaire punten van S op M overeenkomen met de stationaire punten van $\tilde{S}(q,p)$ op

$$\tilde{M} = \{(q,p) \in C^1([t_1,t_2],\mathbb{R}^N) \times C^0((t_1,t_2),\mathbb{R}^N) \mid q(t_1)=q_1, q(t_2)=q_2\}.$$

Omdat stationaire punten van S op M de Euler-Lagrange vergelijkingen leveren, en stationaire punten van $\tilde{S}$ op $\tilde{M}$ de Hamilton-vergelijkingen, zijn dit dan twee equivalente formuleringen van hetzelfde systeem.
Met de extremaal-karakterisering van L in termen van H kunnen we schrijven:

$$S(q) = \int_{t_1}^{t_2} \max_{p\in\mathbb{R}^N} [p.\dot{q}(t) - H(t,q(t),p)]\, dt.$$

Voor elke $t \in (t_1,t_2)$ wordt het maximum in de integrand aangenomen voor zekere $p(t) \in \mathbb{R}^N$ waarvoor $p(t) = L_v(t,q(t),\dot{q}(t))$, en dit is tevens het enige stationaire punt. Omdat $q \in C^1$, zien we dat $t \to p(t)$ een C^0 - functie is, zodat

$$S(q) = \sup_{p\in C^0} \int_{t_1}^{t_2} [p(t).\dot{q}(t) - H(t,q(t),p(t))]\, dt$$

en sup kan hier vervangen worden door stat. Daarmee volgt het gestelde.
$p\in C^0$ $p\in C^0$

5. GLOBALE VARIATIEMETHODEN

In deze paragraaf beschrijven we enkele aspecten van het globale onderzoek van een functionaal J op een verzameling M.
Het eerste aspect betreft het onderzoek naar de *existentie* van een globaal minimaal element van J op M. Het blijkt dat in veel gevallen generalisaties van de stelling van Weierstrass gebruikt kunnen worden voor zo'n existentie-

onderzoek. Daarvoor moet gebruik gemaakt worden van daartoe ontwikkelde functionaal-analytische methoden.
Een ander aspect is het onderzoek hoe *convexiteit* de oplossingsverzameling van een min. pr. beperkt en hoe in die gevallen stationaire punten en globale minimale elementen overeenkomen (equivalentie tussen het min. pr. en het variatieprobleem).
Zoals i.h.b. in Hoofdstuk 6 tot uitdrukking zal komen, kunnen voor convexe minimaliseringsproblemen veel van de gladheidsvoorwaarden die meestal geeist worden in de klassieke variatierekening verzwakt worden. In Sectie 5.2 zullen echter wel (onnodig sterke) gladheidsvoorwaarden opgelegd worden bij het introduceren van *Fenchel-transformaties* van zekere klasse van functies. Daardoor wordt het verband, of beter, de equivalentie, met de veel oudere *Legendre-transformatie* eenvoudig beschreven.

5.1. GENERALISATIE VAN DE STELLING VAN WEIERSTRASS

De bekende stelling van Weierstrass zegt dat een continue functie op een compacte verzameling begrensd is en z'n maximale en minimale waarde aanneemt. Voor het minimaliseringsprobleem voor J op M betekent dit dat er een oplossing bestaat als J continu is en M compact.
In veel toepassingen voldoet de verzameling M niet aan deze sterke compactheidsvoorwaarden. Nader onderzoek leert dat het enige dat nodig is om de existentie van een minimaal element te bewijzen, is dat er een minimale rij gevonden kan worden waarvan de convergentie (in zekere zin) bewezen kan worden. Ook voor onbegrensde verzamelingen M kan dit (in feite voor deelrijen van *elke* minimale rij) bewezen worden mits de functionaal J aan zekere extra groeicondities (coërciviteit) voldoet. Een verdere, voor oneindigdimensionale problemen essentiële, verzwakking van de eisen is dat onderhalf continuïteit van J volstaat.

5.1.1. DEFINITIES. V is een genormeerde lineaire ruimte, $M \subset V$ en $J: V \to \mathbb{R}$.
(i) J heet *coërcief* op M als voor elke rij $\{u_n\} \subset M$, met $\|u_n\| \to \infty$ geldt dat $J(u_n) \to \infty$

(ii) J heet *onderhalf continu* (o.h.c.) in het punt $\hat{u} \in V$ als voor elke rij $\{u_n\} \subset V$ die convergeert naar $\hat{u}$ geldt:

$$J(\hat{u}) \leq \liminf_{n \to \infty} J(u_n).$$

VOORBEELDEN

Als M begrensd is dan is elke functionaal coërcief op M.

De functie $x \to e^x$ is niet coërcief op $\mathbf{R}$; de functie $(x,y) \to x^2-y^2+x^2y^2+y^4$ is wel coërcief op $\mathbf{R}^2$.

Als J continu is in $\hat{u}$, dan is J o.h.c. in $\hat{u}$.

De functie

$$x \to \begin{cases} \alpha \text{ als } x > 0 \\ \beta \text{ als } x \leq 0 \end{cases}$$

is o.h.c. op $\mathbf{R}$ dan en slechts dan als $\alpha \geq \beta$.

5.1.2. Laat $V = \mathbf{R}^N$, $N \geq 1$. Dan geldt

STELLING: Veronderstel dat J en M voldoen aan:

(i) M is gesloten in V

(ii) J is coërcief op M

(iii) J is o.h.c. in elk punt van M.

Dan geldt: het min. pr.

$$\mu = \inf_{u \in M} J(u)$$

heeft een eindige waarde μ, en er bestaat tenminste één oplossing.

BEWIJS: Neem een minimale rij $\{u_n\}$, i.e. een rij $\{u_n\}$ met $u_n \in M$ voor alle n, en $J(u_n) \to \mu$. Op grond van (ii) is dan deze rij begrensd: er is een $M > 0$ zodat $||u_n|| \leq M$ voor alle n. Vanwege de stelling van Bolzano-Weierstrass is er een deelrij, zeg $\{u_{n'}\}$ die convergeert naar een zeker element, zeg $\hat{u} \in V$: $u_{n'} \to \hat{u}$ in V. Vanwege (i) geldt dan $\hat{u} \in M$. Uit (iii) volgt voorts $J(\hat{u}) \leq \liminf J(u_{n'}) = \mu$, dus $J(\hat{u}) \leq \mu$. omdat J eindig gedefinieerd is in alle punten van M, volgt hieruit, ten eerste, dat μ eindig is, en, ten tweede, dat $J(\hat{u}) = \mu$ op grond van de definitie van μ. Dit voltooit het bewijs.

5.1.3. Zoals duidelijk blijkt uit het bewijs, berust dit resultaat op de volgende twee essentiële eigenschappen van $V = \mathbb{R}^N$:

(a) V is volledig: elke Cauchy rij in V heeft een limiet die tot V behoort;

(b) de eenheidsbol is compact: elke begrensde rij heeft een deelrij die convergeert (Bolzano-Weierstrass).

Omdat (b) een definiërende eigenschap is van eindig-dimensionale ruimten, geldt deze eigenschap zeker niet voor oneindig-dimensionale ruimten.
Aan (b) kan, in veel gevallen, tegemoet gekomen worden door introductie van het begrip zwakke convergentie. Met V^* de duale van V (de verzameling van alle continue lineaire functionalen ℓ op V) wordt gedefinieerd:
een rij $\{u_n\} \subset V$ is *zwak convergent* naar $\hat{u} \in V$ als $\langle \ell, u_n \rangle \to \langle \ell, \hat{u} \rangle$ voor elke $\ell \in V^*$.

Dan kan aangetoond worden, of als definitie genomen worden, dat een Banachruimte $\mathcal{B}$ voldoet aan

$(b)_{zw}$: de eenheidsbol is zwak compact: elke (norm-) begrensde rij heeft een deelrij die zwak convergent is,

dan en slechts dan als $\mathcal{B}$ een *reflexieve* Banach-ruimte is.
Alle Hilbertruimten blijken reflexieve Banachruimten te zijn.
Het is duidelijk dat begrippen als geslotenheid van verzamelingen, en (onderhalf-) continuïteit van functionalen ook gedefinieerd kunnen worden m.b.t. deze zwakke convergentie. (Merk op dat de norm-functionaal *niet* continu, maar wel onderhalf continu is voor zwakke convergentie; dit illustreert het belang van o.h.c. in oneindig-dimensionale ruimten).
Exact hetzelfde bewijs als dat van Stelling 5.1.2, nu steeds gebruik makend van de zwakke convergentie, en van eigenschap $(b)_{zw}$, levert dan de volgende oneindig-dimensionale generalisatie van 5.1.2.

5.1.4. STELLING

Laat V een reflexieve Banachruimte zijn. Veronderstel

(i) M is gesloten in V m.b.t. zwakke convergentie;

(ii) J is coërcief op M;

(iii) J is o.h.c. voor zwakke convergentie in elk punt van M.

Dan geldt hetzelfde resultaat als in 5.1.2.

Als, abstracte, toepassing van 5.1.4 bewijzen we de stelling van Riesz die zegt dat de duale van een Hilbertruimte H geïdentificeerd kan worden met de ruimte H zelf.

5.1.5. STELLING (RIESZ)

Zij H een Hilbertruimte, met $(\, , \,)_H$ als inproduct. Laat ℓ een continue lineaire functionaal zijn op H. Dan bestaat er precies één element $u_\ell \in H$ zódat

$$<\ell, v> = (u_\ell, v)_H \quad \text{voor alle} \quad v \in H.$$

BEWIJS: Beschouw de volgende quadratische functionaal op H:

$$Q(u) := \tfrac{1}{2} \| u \|_H^2 - <\ell, u>, \quad u \in H$$

waarin $\| u \|_H^2 = (u,u)_H$. Op grond van Stelling 5.1.4 is Q naar beneden begrensd en heeft een minimaal element $\hat{u} \in H$. Immers, ($M = H$ in dit geval) Q is de som van een zwak onderhalf continue en een zwak continue functionaal, dus zwak o.h.c., en Q is coërcief:

$$Q(u) = \| u \|_H \{ \tfrac{1}{2} \| u \|_H - \frac{<\ell, u>}{\| u \|_H} \} \geq$$

$$\geq \| u \|_H . \{ \| u \|_H - \| \ell \| \} \to \infty, \text{ voor } \| u \|_H \to \infty.$$

De stationairiteitsvoorwaarde voor $\hat{u}$ luidt in dit geval

$$(\hat{u}, v)_H = <\ell, v> \quad \text{voor alle} \quad v \in H$$

en volgens 2.2.9 is dit element $\hat{u}$ eenduidig. Met $\hat{u} \equiv u_\ell$ volgt het gestelde.

5.1.6. TOEPASSEN VAN HILBERTRUIMTE-METHODEN

Bovenbeschreven ideeën, zoals het herkennen van het belang van volledige ruimten (Banach- en Hilbert-ruimten) en het introduceren en bestuderen van zwakke topologieën zijn sinds het begin van deze eeuw ontwikkeld. Ontstaan door onderzoek van vragen uit de variatierekening is het tot een zelfstandig vakgebied uitgegroeid, de *Functionaalanalyse*.

Het toepassen van deze zogenaamde Hilbertruimte-methoden op concrete

problemen, zoals de functionalen die we tot nu toe zijn tegengekomen, bestaat er uit dat geprobeerd wordt een formulering van het probleem te vinden waarvoor (een variant van) Stelling 5.1.4 toepasbaar is. In de praktijk betekent dit dat een Hilbert- (of reflexieve Banach-) ruimte gevonden moet worden zodat aan de eisen van de stelling wordt voldaan. Afgezien van de eis op de verzameling M, kan aan de coërciviteitseis van J in concrete gevallen alleen voldaan worden als de topologie zwak genoeg is, terwijl aan de continuïteitseis voor J juist eerder door keuze van een sterkere topologie voldaan kan worden. Voor dichtheidsfunctionalen zal voorts, vanwege de coërciviteitseis, steeds een *integraal-norm* vereist zijn in plaats van de bekende puntsgewijze C^0, C^1, ... - normen.

Daarom zijn speciale Hilbertruimten geconstrueerd die in veel toepassingen gebruikt kunnen worden. Een eenvoudige klasse daarvan wordt gevormd door de zgn. *Sobolev-ruimten* $H^m(\Omega)$, $m = 0,1,2, \ldots$ voor (scalar-) functies op een gebied $\Omega \subset \mathbf{R}^N$. $H^0(\Omega)$ is niets anders dan de bekende $L_2(\Omega)$; $H^1(\Omega)$, bijv is de completering van C^∞ - functies onder de norm $\| \ \|_{H_1}$

$$\|u\|^2_{H_1} = \int_\Omega (u_x^2 + u^2)\, dx.$$

De bekende methode om een niet-volledige ruimte te *completeren* tot een volledige, in feite door te werken met equivalentieklassen van Cauchy-rijen die een nulrij verschillen, vereist in praktische situaties nog een aanvulling. Om nl. een gevoel te hebben voor deze nieuwe objecten (equivalentieklassen) zoekt men graag naar *representaties* van zo'n completering, zodat de elementen van de completering behoren tot een bekende ruimte en/of geïnterpreteerd kunnen worden als gegeneraliseerde functies. Bijvoorbeeld, alle elementen van $H^1(\Omega)$ kunnen opgevat worden als $L_2(\Omega)$ - functies, en de inbedding die dan ontstaat $H_1(\Omega) \subset L_2(\Omega)$ is continu en zelfs compact. (Voo functies van één variabele, zeg $\Omega = (0,1)$, geldt zelfs dat $H^1(0,1) \subset C^0(0,$

Na herformulering van een concreet probleem tot een probleem waarin bovenbeschreven Hilbertruimte-methoden toepasbaar zijn en tot existentie van een minimaal element leiden, dient i.h.a. nog een *regulariteitsonderzoek* plaats te vinden, d.w.z. een onderzoek hoe "glad" (regulier) zo'n minimaal

element in feite is. Het blijkt namelijk dat in veel gevallen de extremaliteits- (of stationairiteits-) eigenschap extra regulariteit impliceert in vergelijking met de regulariteit van willekeurige elementen van de gecompleteerde ruimte. (Een voorbeeld van extra regulariteit voor stationaire punten hebben we in een eenvoudig geval gezien bij het lemma van du Bois-Reymond, i.h.b. in de toepassing 3.2.3.)
Op de concrete uitwerking van deze methoden kunnen we hier niet verder ingaan.

5.2. CONVEXITEIT

De eenvoudigste eindig-dimensionale voorbeelden laten al zien dat er i.h.a. meerdere oplossingen van het minimaliseringsprobleem voor J op M kunnen zijn. Eenduidigheid van de oplossing is verzekerd als de functionaal J en de verzameling M aan zekere convexiteitseigenschappen voldoen.

5.2.1. DEFINITIE: Laat $M \subset V$ een convexe verzameling zijn.
De functionaal J heet *convex* op M als

$$J(\lambda u + (1 - \lambda)v) \leq \lambda J(u) + (1 - \lambda)J(v)$$

voor elk paar u,v in M, en $\lambda \in (0,1)$.
J heet *strikt convex* als gelijkheid alleen optreedt als $u = v$.

OPMERKING: Door introductie van de *epigraaf* van J:

$$\text{epi}(J) := \{(u,\mu) \in M \times \mathbf{R} \mid \mu \geq J(u)\} \subset V \times \mathbf{R}$$

is convexiteit van de functie J op M equivalent met convexiteit van de deelverzameling epi(J) van $V \times \mathbf{R}$.

5.2.2. STELLING
Zij $M \subset V$ een convexe verzameling en $J: M \to \mathbf{R}$ strikt convex op M.
Dan is een eventuele oplossing van het minimaliseringsprobleem voor J op M *eenduidig*.

BEWIJS: Stel $\hat{u}$ en $\hat{v}$ zijn beide oplossing van $\mu = \inf \{J(u) \mid u \in M\}$.
Dan geldt $J(\hat{u}) = J(\hat{v}) = \mu$. Als $\hat{u} \neq \hat{v}$ dar volgt uit de strikte convexiteit:

$$J(\lambda\hat{u} + (1 - \lambda)\hat{v}) < \mu \quad \text{voor elke} \quad \lambda \in (0,1)$$

hetgeen, omdat $\lambda\hat{u} + (1 - \lambda)\hat{v} \in M$, in strijd is met de definitie van μ.

5.2.3. VOORBEELDEN

De eenduidigheidsresultaten in 2.2.9 en 2.4.5 (en de toepassingen daarvan: 3.2.3 en 3.3.6) zijn een speciaal geval van 5.2.2.
Beschouw (de volgende modificatie van) het *obstakel-probleem* 1.4.4:
Met $\psi \in C^0([0,1])$ gegeven, en $\psi(0) < 0$, $\psi(1) < 0$, laat

$$M = \{u \in C^1([0,1]) \mid u(0) = u(1) = 0;\ u(x) \geq \psi(x) \text{ voor } x \in (0,1)$$

$$J_1(u) = \int_0^1 u_x^2 \, dx \quad \text{en} \quad J_2(u) = \int_0^1 (1 + u_x^2)^{\frac{1}{2}} \, dx.$$

De verzameling M is een convexe verzameling in C^1, en zowel de functionaal J_1 als J_2 is strikt convex. Bijgevolg: het min. pr. voor J_i op M, i=1,2, heeft ten hoogste één oplossing in $C^1([0,1])$.

5.2.4. Analoog als voor functies van één variabele geldt:
LEMMA: Laat J convex en Gateaux-differentieerbaar zijn op M. Dan geldt:

(i) $J(v) - J(u) \geq \langle J'(u), v-u \rangle$ voor alle $u,v \in M$

(ii) $\langle J'(u) - J'(v), u-v \rangle \geq 0$ voor alle $u,v \in M$.

BEWIJS: (i) Uit de convexiteit van J volgt meteen:

$$J(u + \lambda(v - u)) - J(u) \leq \lambda[J(v) - J(u)] \quad \text{voor alle} \quad \lambda \in (0,1$$

Delen door λ en dan limiet $\lambda \downarrow 0$ geeft het resultaat.
(ii) Door optelling van

$$J(v) - J(u) \geq \langle J'(u), v-u \rangle$$

en

$$J(u) - J(v) \geq \langle J'(v), u-v \rangle.$$

5.2.5. Uit 5.2.4 (i) volgt meteen de *equivalentie van het min. pr. en het variatieprobleem* voor convexe problemen:
als M een lineaire variëteit is en J convex en Gateaux-differentieerbaar, dan zijn de enig mogelijke stationaire punten de globale minimale elementen

van J op M;

als M convex is, dan zijn de enig mogelijke oplossingen van de variatie-ongelijkheid $\langle J'(u), v-u\rangle \geq 0$, $\forall v \in M$ de globale minimale elementen van J op M.

5.2.6. LEGENDRE- EN FENCHEL-TRANSFORMATIE

We willen tenslotte een idee geven van de begrippen Legendre- en Fenchel-transformatie van een functie. Fenchel-transformaties liggen, al dan niet expliciet, ten grondslag aan bijna alle dualiteitsmethoden in de convexe analyse (en uitbreidingen daarvan tot niet-convexe problemen). We zullen zien dat de Fenchel-transform van een voldoend gladde, convexe functie de globale formulering is van de veel oudere, en in de klassieke mechanica veel gebruikte Legendre-transformatie. Voor het gemak beperken we ons tot functies op $\mathbf{R}^n$ (een directe generalisatie tot functionalen op (reflexieve-) Banachruimten is mogelijk).

Voor een gegeven functie $f: \mathbf{R}^n \to \mathbf{R}$ wordt de *Fenchel-transformatie* gedefinieerd als de functie g met

$$g(y) := \sup_{x \in \mathbf{R}^n} [x.y - f(x)], \quad y \in \mathbf{R}^n. \tag{5.1}$$

Merk op dat g(y) i.h.a. niet eindig hoeft te zijn, en dat g(y) ook gegeven wordt door

$$g(y) = \inf \{\alpha \mid f(x) \geq x.y - \alpha \text{ voor alle } x \in \mathbf{R}^n\}.$$

Dit laatste maakt duidelijk dat (als g(y) eindig is), van alle lineaire functies $x \to x.y - \alpha$ waarvan de grafiek geheel ligt onder de grafiek van f, de functie $x \to x.y - g(y)$ de maximale is; de grafiek van deze lineaire functie is een stuthypervlak aan de epigraaf van f.

In het vervolg beschouwen we transformaties van de volgende klasse van functies F: $f \in F$ als

(i) $f \in C^1(\mathbf{R}^n)$

(ii) f is strikt convex op $\mathbf{R}^n$

(iii) f is *superlineair* in het oneindige, in de volgende zin:

(5.2) $\frac{f(x)}{|x|} \to \infty$ voor $|x| \to \infty$.

STELLING

Zij $f \in F$, en beschouw de Fencheltransformatie g gegeven door (5.1).

Dan geldt:

(i) g is eindig gedefinieerd op $\mathbb{R}^n$ en het sup in (5.1) wordt voor elke $y \in \mathbb{R}^n$ aangenomen in een uniek punt x, nl. het punt x waarvoor y=f'(x
De afbeelding $f': \mathbb{R}^n \to \mathbb{R}^n$ is bijectief.

(ii) g kan op equivalente manier gedefinieerd worden als de zgn. *Legendretransformatie* van f:

(5.3) $g(y) = x.y - f(x)$ waarin x zo dat $y = f'(x)$

(iii) Voor de functie f geldt

(5.4) $f(x) = \sup_{y \in \mathbb{R}^n} [x.y - g(y)], \quad x \in \mathbb{R}^n$

en het sup wordt voor elke x aangenomen in het unieke punt $y = f'(x)$

$f(x) = x.y - g(y)$ met $y = f'(x)$.

(iv) De functie g is strikt convex en zelfs differentieerbaar; $g': \mathbb{R}^n \to \mathbb{R}^n$ is de inverse van f':

(5.5) $y = f'(x) \Longleftrightarrow x = g'(y)$.

OPMERKING

Uit (iii) volgt dat $f = g^*$, zodat $f = (f^*)^*$ voor $f \in F$.

BEWIJS: Voor $y \in \mathbb{R}^n$ is de functie $x \to f(x) - x.y$ strikt convex, en coërcie vanwege (5.2). Met Stelling 5.1.2 en 5.2.2 volgt dat $\sup_{x \in \mathbb{R}^n}[x.y - f(x)]$ precies één oplossing x heeft, waarvoor dan $y = f'(x)$. Daaruit volgt dat f' bijectief is en (5.3).

Als g gegeven wordt door (5.3), dan is met 5.2.5 en 5.2.2 direct in te zie dat g ook gekarakteriseerd wordt door (5.1). Uit (5.1) volgt

$$g(y) \geq x.y - f(x), \quad \forall x, \forall y$$

zodat

$$f(x) \geq x.y - g(y), \quad \forall x, \forall y$$

en ook

$$f(x) \geq \sup_{y \in \mathbf{R}^n} [x.y - g(y)].$$

Uit (5.3) volgt dat voor $y = f'(x)$ dit sup wordt aangenomen en dan de waarde $f(x)$ heeft. Daaruit volgt (5.4).
De functie g is gedefinieerd in (5.1) als het sup van een familie convexe (want lineaire) functies $y \to x.y - f(x)$, en is dus convex. Strikte convexiteit volgt vrij eenvoudig met de karakterisering (5.3).
Het wat technischer bewijs dat g in feite C^1 is (als gevolg van het feit dat f strikt convex is) laten we achterwege. Uit de stationairiteitsvoorwaarde voor het maximale element $y(= f'(x))$ van (5.4): $x = g'(y)$, volgt dan (5.5).

5.2.7. VOORBEELD

Laat $\alpha > 1$, en $f(x) := \frac{1}{\alpha} |x|^{\alpha}$, $x \in \mathbf{R}^n$.
Dan geldt

$$f^*(y) = \frac{1}{\beta}|y|^{\beta}, \quad y \in \mathbf{R}^n$$

waarin β zódat $\frac{1}{\alpha} + \frac{1}{\beta} = 1$.

Tot slot wil ik prof. dr. A. van ROOIJ van harte bedanken voor zijn prompte commentaar en bruikbare opmerkingen bij een eerdere versie. De invloed van zijn getoonde belangstelling is in deze tekst, hoewel onzichtbaar, duidelijk aanwezig.

LITERATUURAANWIJZINGEN

Introductie tot variatierekening met toepassingen uit Mathematische Fysica

1. R. COURANT & D. HILBERT, *Method of Mathematical Physics*, Vol. I, Interscience Publ., New York, 1953. (Chapters IV-VI, veel voorbeelden

2. I.M. GELFAND & S.V. FOMIN, *Calculus of Variations*, Prentice-Hall Inc., New Jersey, 1963. (beknopt, veel informatie, fysisch georiënteerd)

3. H.A. LAUWERIER, *Calculus of Variations in Mathematical Physics*, M.C. Tract, Amsterdam, 1966. (elementair)

4. V.I. ARNOLD, *Mathematical Methods of Classical Mechanics*, Springer-Verlag, New York, 1978. (fundamenteel, wiskundig georiënteerd)

5. A.D. IOFFE & V.M. TIHOMIROV, *Theory of extremal problems*, North-Holland, Amsterdam, 1979. (fundamenteel, wiskundig georiënteerd)

6. J.L. TROUTMAN, *Variational calculus with elementary convexity*, Springer-Verlag, New York, 1983. (elementair en uitvoerig)

7. E.W.C. van GROESEN, *Optimaliseren*, Mathematisch Instituut, Katholieke Universiteit Nijmegen, 1985. (standaard-dictaat)

Introductie tot globale methoden met functionaal-analyse

8. M.M. VAINBERG, *Variational methods for the study of Nonlinear Operators*, Holden-Day Inc., San Francisco, 1964. (fundamenteel, uitvoerig)

---, *Variational method and method of Monotone Operators in the theory of Nonlinear Equations*, Wiley, New York, 1973. (vervolg)

9. D.G. LUENBERGER, *Optimization by vector space methods*, Wiley, New York, 1969. (uitvoerig, elementaire functionaal-analyse)

10. A.F. MONNA, *Dirichlet's principle. A mathematical comedy of errors and its influence on the development of analysis*, Oosthoek, Utrecht, 197! (historisch)

11. K. REKTORYS, *Variational Methods in Mathematics, Science and Engineering*, D. Reidel, Boston, 1975. (zeer uitvoerig, veel toepassingen)

12. E.W.C. van GROESEN, *Variational methods for nonlinear operator equations*; in Nonlinear Analysis, Vol. 2, N.M. Temme (ed.), M.C. Syllabus 26.2, Amsterdam, 1976. (elementaire introductie)

Introductie tot convexiteitstheorie (zie ook literatuur bij Hoofdstuk 6)

13. R.T. ROCKAFELLAR, *Convex Analysis*, Princeton Univ. Press, New Jersey, 1970. (zeer uitvoerig)

14. M.R. HESTENES, *Optimization Theory, the finite dimensional case*, Wiley, New York, 1975. (fundamentele ideeën gedemonstreerd in $\mathbb{R}^n$)

15. I. EKELAND & R. TEMAM, *Convex Analysis and Variational Problems*, North-Holland, Amsterdam, 1977. (fundamenteel, toepassingen uit Mathematische Fysica)

16. J. van TIEL, *Convexe Analyse*, M.C. Syllabus 40, Amsterdam, 1979. (elementair)

17. D. KINDERLEHRER & G. STAMPACCHIA, *An Introduction to Variational Inequalities and their Applications*, Academic Press, New York, 1980. (fundamenteel, variatie-ongelijkheden uit de Mathematische Fysica)

18. F.H. CLARKE, *Optimization and Nonsmooth Analysis*, Wiley, New York, 1983. (fundamenteel, gegeneraliseerde afgeleiden)

HOOFDSTUK 3

MINIMAX METHODEN

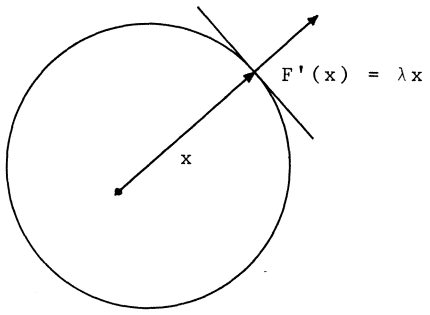
P.P.J.E. CLÉMENT

0. INLEIDING

Gedurende de laatste jaren zijn er aanzienlijke vorderingen gemaakt op het gebied van de variatierekening. Zonder overdrijving kan men zeggen dat het bewijs van P.H. RABINOWITZ [12] van een stelling van A. WEINSTEIN [13] over het bestaan van periodieke oplossingen met gegeven energie voor Hamiltonsystemen een belangrijke stap betekent in de geschiedenis van de variatierekening. Dit bewijs heeft een nieuwe impuls gegeven aan de directe methoden die gebruik maken van karakterisering van kritieke waarden van een functionaal met behulp van "minimax methoden". Het spreekt vanzelf dat heel wat andere -al dan niet hiermee verwante- methoden (dualiteitsmethoden, voortgekomen uit de convexe analyse, theorie van Morse, etc.) uitermate actueel zijn, maar in deze voordracht zullen we ons beperken tot - wat men noemt - de "minimax methoden". Deze methoden vinden hun oorsprong in the "theorie van algemene eigenwaarden" van LYUSTERNIK & SCHNIRELMANN [4],[5]. Bij wijze van voorbeeld van deze theorie beschouwen we de volgende situatie: Laat $F : \mathbb{R}^N \to \mathbb{R}$ $(n > 1)$ een functie met continue afgeleide zijn (notatie $F \in C^1(\mathbb{R}^n, \mathbb{R})$ en laat S^{N-1} de eenheidssfeer zijn in $\mathbb{R}^N$:

$$S^{N-1} := \{x \in \mathbb{R}^N \mid \|x\| = 1\}.$$

Wij interesseren ons nu voor de kritieke punten van de beperking van F tot S^{N-1} (notatie voor deze beperking: $F|_{S^{N-1}}$), d.w.z. voor de punten x van S^{N-1} met de eigenschap dat de gradiënt van $F|_{S^{N-1}}$ daar nul wordt of, met andere woorden, voor de punten $x \in S^{N-1}$ met de eigenschap dat de gradiënt van F in x loodrecht staat op het raakvlak in x aan S^{N-1}, of, nog weer anders, voor de punten x van S^{N-1} waarvoor er een $\lambda \in \mathbb{R}$ bestaat (multiplicator van Lagrange) zodanig dat $F'(x) = \lambda x$.



Figuur 1

Indien $\bar{x}$ een *kritiek punt* van $F|_{S^{N-1}}$ is, dan zegt men dat $F(\bar{x})$ een *kritieke waarde* is van $F|_{S^{N-1}}$. Daar F continu is op S^{N-1} en daar S^{N-1} compact is (d.w.z. begrensd en gesloten), bezit F een maximum en een minimum op S^{N-1}. Men kan nu bewijzen dat, indien $F|_{S^{N-1}}$ zijn *maximum* (resp. *minimum*) bereikt in $\bar{x}$, dit punt $\bar{x}$ tevens een *kritiek punt* is van $F|_{S^{N-1}}$ en dus dat $a := \max_{x \in S^{N-1}} F(x)$ en $b := \min_{x \in S^{N-1}} F(x)$ kritieke waarden zijn van $F|_{S^{N-1}}$. Het kan voorkomen dat deze a en b de enige kritieke waarden zijn van $F|_{S^{N-1}}$, bijv. indien $F(x) = x_1$, waarbij geldt:

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}.$$

LYUSTERNIK [4] heeft evenwel bewezen dat, indien $F|_{S^{N-1}}$ even is, d.w.z. indien $F(x) = F(-x)$ voor $x \in S^{N-1}$, de restrictie $F|_{S^{N-1}}$ van F tot S^{N-1} minstens N paren kritieke punten bezit. Het getal N is optimaal, zoals uit het volgende voorbeeld blijkt: Laat $F(x) := x^T A x$, waarbij A een symmetrische N×N matrix is met N verschillende eigenwaarden. Inderdaad is in dit geval een kritiek punt $\bar{x}$ van $F|_{S^{N-1}}$ een eigenvector van A en zelfs genormaliseerd omdat $\| \bar{x} \| = 1$ en men weet dat A precies N paren van zulke eigenvectoren bezit. Men ziet ook dat, indien $A\bar{x} = \bar{\lambda}\bar{x}$, tevens geldt:

$$F(\bar{x}) = \bar{x}^T A \bar{x} = \bar{\lambda} \bar{x}^T \bar{x} = \bar{\lambda}$$

en dus dat $\bar{\lambda}$ (de bijbehorende eigenwaarde) een kritieke waarde is van $F|_{S^{N-1}}$. Indien men de eigenwaarden van A naar opklimmende grootte ordent:

$$\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_N$$

dan weet men dat $\lambda_1 = \min F|_{S^{N-1}}$ en $\lambda_N = \max F|_{S^{N-1}}$. Anderzijds heeft FISCHER [3, 1905] de volgende karakterisering gegeven van de eigenwaarden:

$$\lambda_k = \inf_{D \in \Gamma_k} \max_{x \in D} F(x)$$

waarbij $\Gamma_k = \{D \subset S^{N-1} \mid D = S^{N-1} \cap E_k\}$. Hierin is E_k een deelruimte van $\mathbb{R}^N$ met de dimensie minstens k. Men ziet dat, indien $\lambda_k = \lambda_{k+1}$ voor een zekere waarde van k, $F|_{S^{N-1}}$ een oneindig aantal kritieke punten bezit. In het geval dat F even is, maar niet noodzakelijker wijze van de tweede graad, bestaat de methode van Schnirelmann daarin dat men klassen Γ_k van gesloten deelverzamelingen van S^{N-1} kiest die de eigenschap hebben dat de waarden c_k, gedefinieerd door:

$$c_k = \inf_{D \in \Gamma_k} \max_{x \in D} F(x)$$

kritieke waarden zijn en wel zodanig dat, indien voor zekere k geldt:

$$c_k = c_{k+1}$$

de functie F een oneindig aantal kritieke punten bezit. Het bewijs van de existentie van dergelijke klassen vereist resultaten van topologische aard (stelling van Borsuk) die wij hier niet zullen introduceren. Het doel van deze voordracht is aan te tonen dat men in het geval van een functie $F : \mathbb{R}^N \to \mathbb{R}$, kritieke waarden van F kan definiëren die corresponderen met kritieke punten van F (zonder nadere voorwaarden) die niet noodzakelijk ook extrema van F zijn. Aan de andere kant zullen wij ook "kritieke waarden", zeg c, van F definiëren, waarvoor er niet noodzakelijk een $\bar{x} \in \mathbb{R}^N$ bestaat zodanig dat $F(\bar{x}) = c$ en $F'(\bar{x}) = 0$ maar waarvoor er wel een rij (x_n) met $x_n \in \mathbb{R}^N$ bestaat zodanig dat

$$\lim_{n \to \infty} F(x_n) = c \quad \text{en} \quad \lim_{n \to \infty} F'(x_n) = 0.$$

1. MINIMA EN DE "PALAIS-SMALE" CONDITIE

Zij $F : \mathbb{R}^N \to \mathbb{R}$ een C^1 funktie. Definieer voor $c \in \mathbb{R}$ de verzameling

$$K_c := \{x \in \mathbb{R}^N \mid F(x) = c \text{ en } F'(x) = 0\}.$$

Als $K_c \neq \emptyset$ is, dan heet elke $x \in K_c$ een *kritiek punt* van F en c een *kritieke waarde* van F. Als F naar beneden begrensd is, dan is $c := \inf_{x \in \mathbb{R}^N} F(x)$ een eerste kandidaat om een kritieke waarde te zijn. Immers als er een $\bar{x} \in \mathbb{R}^N$ bestaat zó dat $F(\bar{x}) = c$ dan is $F'(x) = 0$ en $K_c \neq \emptyset$.
Om te onderzoeken of er zo'n minimaal element $\bar{x}$ bestaat, merken we op dat, op grond van de definitie van c, er voor elk $n \in \mathbb{N}^+$ een element $x_n \in \mathbb{R}^N$ bestaat zodat

$$(1.1) \qquad c \leq F(x_n) \leq c + \frac{1}{n} .$$

Als deze rij (x_n) begrensd zou zijn, dan bestaat (x_{n_k}) zodanig dat $\bar{x} = \lim_{k \to \infty} x_{n_k}$. Wegens de continuiteit van F geldt dat $F(\bar{x}) = c$, waaruit volgt dat $\bar{x}$ zo'n minimaal element is. Begrensdheid van zo'n minimale rij (x_n) als boven kan volgen uit een bepaald globaal gedrag van de functie F. Definieer daartoe: de functie F heet *coercief* als geldt

$$F(x_n) \to +\infty \text{ voor elke rij } (x_n) \text{ met } \|x_n\| \to \infty.$$

Als F coercief is, dan zijn de verzamelingen $A_d := \{x \in \mathbb{R}^N \mid F(x) \leq d\}$ voor alle $d \in \mathbb{R}$ begrensd en gesloten, en bovendien is F naar beneden begrensd. Hieruit volgt dan dat $c = \inf_{x \in \mathbb{R}^N} F(x) = \min_{x \in \mathbb{R}^N} F(x)$ een kritieke waarde van F is.

VOORBEELD 1.

Zij A een N×N symmetrische matrix die positief definiet is (er is een $a > 0$ zo dat $x^T Ax \geq a\, x^T x$ voor alle $x \in \mathbb{R}^N$) en zij $G : \mathbb{R}^N \to \mathbb{R}$, $G \in C^1$, $G(0) = 0$ zodat $g(x) := G'(x)$ (we beschouwen $G'(x)$ als kolomvector) begrensd is, d.w.z. er is $M > 0$ zodat $\|g(x)\| \leq M$ voor alle $x \in \mathbb{R}^N$.

(B.v. $\quad G(x) = \sin\left(\sqrt{1 + \|x\|^2}\right)$, $\quad G'(x) = \dfrac{\cos\left(\sqrt{1 + \|x\|^2}\right)}{\sqrt{1 + \|x\|^2}} x$.

Dan is $F(x) := \frac{1}{2} x^T Ax - G(x)$ coercief. Immers, merk op dat $G(x) = G(x) - G(0) = \int_0^1 g(tx)^T x \, dt$ en $F(x) = \frac{1}{2}x^T Ax - G(x) \geq \frac{a}{2} x^T x - M \| x \|$. Dus

$$\lim_{\|x\| \to \infty} F(x) \geq \lim_{\|x\| \to \infty} \|x\| (\frac{a}{2} \|x\| - m) = + \infty.$$

F bezit dus een globaal minimum en de vergelijking

$$Ax = g(x), \quad x \in \mathbb{R}^N$$

heeft een oplossing.

De in de toepassingen meest voorkomende situatie is niet dat het definitiegebied van de functie F eindig - maar juist oneindig dimensionaal is. Om een idee te geven hoe de eindig-dimensionale methoden gegeneraliseerd kunnen worden tot problemen in oneindig-dimensionale ruimten, beschouwen we in de rest van deze paragraaf functies F gedefinieerd op een reële Hilbertruimte H. Het inproduct in H geven we aan met (,). Als $F \in C^1(H;\mathbb{R})$ naar beneden begrensd is en coercief is, dan volgt dat de rij (x_n) gedefinieerd door (1.1) uniform begrensd is. Echter, daaruit kan niet geconcludeerd worden dat er een deelrij (x_{n_k}) en een $\bar{x} \in H$ moeten bestaan zodat $\lim_{k \to \infty} x_{n_k} = \bar{x}$. Immers, begrensde en gesloten deelverzamelingen in een oneindig-dimensionale Hilbertruimte zijn niet noodzakelijk compact. Er zijn tenminste twee methoden om deze moeilijkheid te overwinnen. Een van deze berust op het feit dat er wél een $\bar{x} \in H$ en een deelrij (x_{n_k}) bestaan zodat $\lim_{k \to \infty} (x_{n_k}, y) = (\bar{x}, y)$ voor alle $y \in H$. Men spreekt dan van zwakke convergentie van (x_{n_k}) naar $\bar{x}$. Zelfs als F continu differentieerbaar is, geldt dan niet noodzakelijk

$$\lim_{k \to \infty} F(x_{n_k}) = F(\bar{x}).$$

[Denk aan $H = \ell^2$, $e_1 = (1,0,\ldots)$, $e_2 = (0,1,\ldots)$ enz.. Dan geldt voor $(x,y) = \sum_{i=1}^{\infty} x_i y_i$ dat $\lim_{i \to \infty} (e_i, y) = \lim y_i = 0$ voor alle $y \in \ell^2$ en toch $\lim_{i \to \infty} \|e_i\|^2 = \lim_{i \to \infty} (e_i, e_i) = 1$.]

Maar voor het bestaan van een minimum is het voldoende om aan te tonen dat $F(x) \leq \lim_{k \to \infty} F(x_{n_k})$. Dit geldt in het bijzonder als F convex is. Dit

soort redeneringen is het uitgangspunt van een belangrijk deel van de niet-lineaire functionaal analyse: de zogenaamde monotoniciteit, pseudo-monotoniciteit methoden. We zullen in deze lezing deze weg niet volgen. (Zie Hoofdstuk 2, §5)

Een tweede methode bestaat er uit een beter gebruik te maken van de differentieerbaarheid van de functie F. Laten we een eenvoudig, maar belangrijk, voorbeeld beschouwen. Zij A een zelfgeadjungeerde operator in H die positief definiet is (er is $a > 0$ zodat $(Ax,x) \geq a\|x\|^2$ voor alle $x \in H$) en zij $f \in H$. Definieer

$$F(x) := \tfrac{1}{2}(Ax,x) - (f,x).$$

Dan geldt: $F \in C^1(H;\mathbb{R})$ en F is naar beneden begrensd en coercief. Laat $c := \inf_{x\in H} F(x)$ en beschouw een rij $(x_n) \subset H$ zodat

$$c \leq F(x_n) \leq c + \frac{1}{n}.$$

Dan geldt voor alle $t \in \mathbb{R}$ en elke $h \in H$: $F(x_n) - \frac{1}{n} \leq c \leq F(x_n + th)$. Hieruit volgt

$$t^2 \|h\|^2 + 2t(Ax_n - f,h) + \frac{2}{n} \geq 0$$

voor alle $t \in \mathbb{R}$ en alle $h \in H$. Dus $|(Ax_n - f,h)|^2 \leq \frac{2}{n}\|h\|^2$ voor alle $h \in H$. Kies $h = Ax_n - f$. Dan geldt:

$$\|Ax_n - f\| \leq \sqrt{\frac{2}{n}}.$$

Merk op dat

$$F'(x_n) = Ax_n - f.$$

We hebben dus een rij (x_n) gevonden zodat

$$\lim_{n\to\infty} F(x_n) = c, \quad \lim_{n\to\infty} F'(x_n) = 0.$$

Deze informatie is voldoende om te garanderen dat (x_n) een convergente rij is. Immers

$$a\,\|x_m - x_n\|^2 \le (A(x_m - x_n),\, x_m - x_n) =$$

$$((Ax_m - f) - (Ax_n - f),\, x_m - x_n) \le$$

$$\le (\|Ax_m - f\| + \|Ax_n - f\|)\,\|x_m - x_n\| \le$$

$$\le (\sqrt{\tfrac{2}{m}} + \sqrt{\tfrac{2}{n}})\,\|x_m - x_n\| .$$

Hieruit volgt

$$\|x_m - x_n\| \le a^{-1}\,[\sqrt{\tfrac{2}{m}} + \sqrt{\tfrac{2}{n}}]$$

en dus dat (x_m) een Cauchyrij is in H. Omdat H volledig is, is er dan een element $\bar{x} \in H$ zodat $\lim_{m\to\infty} x_m = \bar{x}$, en we hebben dan $A\bar{x} = f$. Dit eenvoudige voorbeeld laat zien hoe belangrijk het is te weten dat

$$\lim_{n\to\infty} F'(x_n) = 0.$$

We besluiten deze paragraaf met een (zwakkere vorm van een) stelling van Ekeland waaruit volgt dat deze informatie altijd aanwezig is als F naar beneden begrensd is.

STELLING. (EKELAND [2])
Laat (V,d) een volledige metrische ruimte zijn en zij $F : V \to \mathbb{R} \cup \{+\infty\}$ onder-half continu, naar beneden begrensd en $\not\equiv +\infty$. Zij $\varepsilon > 0$ en $u \in V$ zodanig dat

$$F(u) \le \inf_{v \in V} F(v) + \varepsilon.$$

Dan bestaat er voor alle $\lambda > 0$ een $v \in V$ met de eigenschappen:

$$F(v) \le F(u),$$

$$d(u,v) \le \lambda.$$

Voor alle $w \neq v$, $F(w) > F(v) - \varepsilon\lambda^{-1}\, d(v,w)$.

GEVOLG. Laat V een Banachruimte zijn (b.v. $\mathbb{R}^N$), en laat $F \in C^1(V;\mathbb{R})$, met $c = \inf F(x) > -\infty$. Neem een rij $U_n \in V$, $n \in \mathbb{N}^+$, $x \in V$ zodanig dat

$$F(U_n) \le c + \frac{1}{n}.$$

Beschouw $\frac{1}{n}$ als de ε in bovenstaande stelling en neem $\lambda = 1/\sqrt{n}$. Dan bestaat er een element $x_n \in V$ zodanig dat

$$F(x_n) \leq F(U_n) \leq c + \frac{1}{n},$$

$$\|U_n - x_n\| \leq \sqrt{\frac{1}{n}},$$

en voor elke $t > 0$ en $h \in V \setminus \{0\}$ geldt:

$$F(x_n + th) - F(x_n) > -\sqrt{\frac{1}{n}}\,|t|\,\|h\|.$$

Als we de limiet voor $t \downarrow 0$ nemen, krijgen we

$$F'(x_n)h \geq -\sqrt{\frac{1}{n}}\,\|h\|$$

voor alle $h \in V$.

Vervang h door -h, dan volgt

$$|F'(x_n)\,h| \leq \sqrt{\frac{1}{n}}\,\|h\|$$

voor alle $h \in V$ en dus $\|F'(x_n)\| \leq \sqrt{\frac{1}{n}}$.

Dit resultaat motiveert de volgende definitie:

DEFINITIE. De functie $f \in C^1(\mathbb{R}^N;\mathbb{R})$ voldoet aan "P.S" in $c \in \mathbb{R}$ indien het bestaan van een rij (x_n) in $\mathbb{R}^N$ met de eigenschappen:

$$\lim_{n\to\infty} F(x_n) = c, \qquad \lim_{n\to\infty} F'(x_n) = 0$$

impliceert dat c een kritieke waarde van F is. Als F aan "P.S" voldoet voor alle $c \in \mathbb{R}$, dan zeggen we dat F aan "P.S" voldoet. De notatie "P.S" staat voor PALAIS & SMALE [8] die voor het eerst zo'n conditie hebben ingevoerd.

Voor de volgende paragraaf is het van belang op te merken dat, in de situatie van Voorbeeld 1, als de matrix A niet positief-definiet is, maar wel regulier is, de functie F aan "P.S" voldoet.

2. TWEE "MINIMAX" STELLINGEN

Beschouw het voorbeeld 1 met A noch positief definiet, noch negatief definiet maar wel regulier en, om te beginnen, g = 0(!). Dan is 0 een kritiek punt van F, en wel een zadelpunt in de zin dat iedere omgeving van 0 punten y en z bevat waarvoor geldt

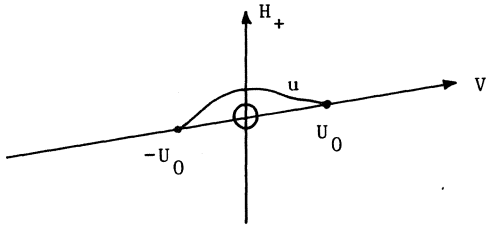
$$F(y) > F(0) > F(z).$$

Laat H_+ (resp. H_-) de deelruimte van $\mathbb{R}^N$ zijn die voortgebracht wordt door de eigenvectoren van A waarvoor de bijbehorende eigenwaarden positief

(resp. negatief) zijn. Dan is $F|_{H_+}$ naar beneden begrensd en $\inf_{x \in H_+} F(x) = 0$. Voor het gemak nemen we aan dat dim $H_- = 1$. Zij V een één-dimensionale deelruimte van $\mathbb{R}^N$ zodanig dat $\mathbb{R}^N = H_+ \oplus V$ (bijvoorbeeld $V = H_-$). Het is niet moeilijk in te zien dat er punten U_0 en $-U_0$ in V bestaan waarvoor $F(U_0), F(-U_0) < \inf_{x \in H_+} F(x)$. In dat geval kan de kritieke waarde 0 op de volgende manier gekarakteriseerd worden: zij

$$\Gamma := \{u : [-1,1] \to \mathbb{R}^N \mid u \text{ continu en } u(-1) = -U_0 \,, u(1) = U_0\}.$$

Voor $u \in \Gamma$ beschouw $\max_{t \in [-1,1]} F(u(t))$. Er is dan zeker een $\bar{t} \in (0,1)$ zodanig dat $u(\bar{t}) \in H_+$.



Figuur 2

Dus $\max_{t \in [-1,1]} F(u(t)) \geq F(u(\bar{t})) \geq \inf_{x \in H_+} F(x) = 0$.

Dan geldt dat $\inf_{u \in \Gamma} \max_{t \in [-1,1]} F(u(t)) = 0$ omdat $\bar{u}$ gedefinieerd door $\bar{u}(t) = tu_0$ een element van Γ is waarvoor

$$\max_{t \in [-1,1]} F(\bar{u}(t)) = 0$$

($t \to F(\bar{u}(t)) = \alpha t^2$ als u_0 een eigenvector is van A bij de negatieve eigenwaarde α). Zelfs als $g \neq 0$ (en $x \to g(x)$ begrensd), kan men een complement V van H_+ in $\mathbb{R}^N$ vinden zodat $F(-U_0), F(U_0) < \inf_{x \in H_+} F(x)$, met $U_0 \in V$.
Op dezelfde manier definiëren we dan

$$c = \inf_{u \in \Gamma} \max_{t \in [-1,1]} F(u(t))$$

met $\Gamma := \{u : [-1,1] \to \mathbb{R}^N \mid u \text{ is continu, } u(-1) = -U_0 \,, u(1) = U_0\}$. Weer geldt $c > \max(F(-U_0), F(U_0))$. Dat c een kritieke waarde van F is, is een gevolg van

STELLING A. (RABINOWITZ [11]])

Zij $F \in C^1(\mathbb{R}^N;\mathbb{R})$ een functie die aan "P.S" voldoet. Laat E_+ en E_- deelruimten van $\mathbb{R}^N$ zijn zo dat $\mathbb{R}^N = E_+ \oplus E_-$, en definieer voor $R > 0$ $B_R := \{x \in E_- \mid \|x\| \leq R\}$ en $\partial B_R := \{x \in E_- \mid \|x\| = R\}$. Veronderstel nu dat er een $R > 0$ bestaat zodanig dat

$$(2.1) \qquad \max_{x \in \partial B_R} F(x) < \inf_{x \in E_+} F(x).$$

Definieer de verzameling Γ door

$$\Gamma := \{u : B_R \to \mathbb{R}^N \mid u \text{ is continu en } u(x) = x \text{ voor } x \in \partial B_R\}.$$

Dan is

$$c := \inf_{u \in \Gamma} \max_{x \in B_R} F(u(x))$$

een kritieke waarde van F en er geldt

$$c > \max_{x \in \partial B_R} F(x).$$

We zijn nu in staat om het bestaan van een oplossing van (1.2) aan te tonen zelfs als $\dim H_- > 0$. Immers, als $0 < \dim H_- < N$, kies dan $E_+ = H_+$ en $E_- = H_-$ in stelling A en merk op dat uit de begrensdheid van g volgt dat aan (2.1) wordt voldaan voor elke R die voldoende groot is.

OPMERKINGEN

1. Het bestaan van een oplossing van (1.2) met A regulier en $G : \mathbb{R}^N \to \mathbb{R}^N$ continu, begrensd (niet noodzakelijk de gradient van een functie G) kan op een andere manier aangetoond worden. Immers, definieer T door $Tx := A^{-1}g(x)$ voor $x \in \mathbb{R}^N$. Dan is T continu en wegens de begrensdheid van G bestaat er een $R > 0$ zodanig dat T de bol met straal R in $\mathbb{R}^N$ in zichzelf afbeeldt ($\|Tx\| \leq R$ voor R groot genoeg). Uit een beroemde stelling van Brouwer volgt dat T een dekpunt $\bar{x}$ heeft, dus $\bar{x} = T\bar{x} = A^{-1}g(\bar{x})$, of wel $A\bar{x} = G(\bar{x})$.
2. Naar aanleiding van de discussie van §1 komt de volgende vraag aan de orde: als niet aan "P.S" voldaan wordt, bestaat er dan toch een rij (x_n) zodat $\lim_{n\to\infty} F(x_n) = c$ en $\lim_{n\to\infty} F'(x_n) = 0$? Het antwoord op deze vraag

wordt gegeven in het volgende lemma.

LEMMA

Laat $F \in C^1(\mathbf{R}^N;\mathbf{R})$, $K \subset \mathbb{R}^N$ begrensd en gesloten, en $K_0 \subset K$ niet leeg en gesloten. Zij $\psi \in C(K_0;\mathbf{R}^N)$ en definieer

$$\Gamma := \{u \in C(K;\mathbf{R}^N) \mid u(x) = \psi(x) \text{ voor alle } x \in K_0\}.$$

Laat $b = \max_{x \in K_0} F(\psi(x))$ en $c = \inf_{u \in \Gamma} \max_{x \in K} F(u(x))$.

Dan geldt: als $b < c$, dan bestaat er een rij $(x_n) \subset \mathbf{R}^N$ zodat

$$\lim_{n \to \infty} F(x_n) = c$$

$$\lim_{n \to \infty} F'(x_n) = 0.$$

We geven een schets van het bewijs in de Appendix.

OPMERKING

Het feit dat K_0 begrensd en gesloten is en dat ψ continu is garandeert dat $\max_{x \in K} F(\psi(x))$ bestaat. Γ bestaat uit alle continue voortzettingen van ψ tot K. Dat zulke voortzettingen bestaan is een gevolg van een stelling uit de algemene topologie (stelling van Tietze); dus $\Gamma \neq \emptyset$ en $b \leq c < \infty$.

BEWIJS VAN STELLING A.

Neem in het lemma $K_0 = \partial B_R$, $K = B_R$ en $\psi(x) = x$ op ∂B_R. Zij $P : \mathbb{R}^N \to \mathbb{R}^N$ de projectie op E_- "langs" E_+. Voor $u \in \Gamma$, definieer $v := p \circ u$. Dan geldt $v \in C(B_R;\mathbb{R}^N)$ en $v(x) = x$ voor alle $x \in \partial B_R$. Met behulp van bovengenoemde stelling van Brouwer is het mogelijk om aan te tonen dat v een nulpunt heeft, d.w.z. er is een $\bar{x} \in B_R$ zodanig dat $v(\bar{x}) = 0$. Voor $\underline{x}$ is dan $u(\bar{x}) = (I-P)u(\bar{x}) \in E_+$. Dan geldt

$$\max_{x \in B_R} F(u(x)) \geq F(u(\bar{x})) \geq \inf_{x \in E_+} F(x)$$

en

$$c = \inf_{u \in \Gamma} \max_{x \in B_R} F(u(x)) \geq \inf_{x \in E_+} F(x) > \max_{x \in \partial B_R} F(x) = b.$$

Aan de voorwaarden van het lemma wordt dus voldaan. De conclusie van het lemma en "P.S" geeft dan het resultaat van Stelling A.

We beëindigen deze paragraaf met de nu beroemde stelling van AMBROSETTI & RABINOWITZ [1].

STELLING B. HET "MOUNTAIN PASS LEMMA"

Zij $G \in C^1(\mathbb{R}^N;\mathbb{R})$. Neem aan dat er twee getallen d en R, $R > 0$, zijn zodanig dat

$$F(x) \geq d \text{ op de sfeer } S_R := \{x \in \mathbb{R}^N \mid \|x\| = R\},$$

$$F(0) \text{ en } F(e) < d \text{ voor een } e \in \mathbb{R}^N \text{ met } \|e\| > R.$$

Dan bestaat er een rij $(x_n) \subset \mathbb{R}^N$ met

$$\lim_{n\to\infty} F(x_n) = c, \quad \lim_{n\to\infty} F'(x_n) = 0$$

waarbij

$$c = \inf_{u\in\Gamma} \max_{t\in[0,1]} F(u(t)) \geq d$$

en

$$\Gamma := \{u : [0,1] \to \mathbb{R}^N \mid u \text{ is continu en } u(0) = 0,\ u(1) = e\}.$$

VOORBEELD 2.

We beschouwen een voorbeeld van een functie $F : \mathbb{R}^2 \to \mathbb{R}$ die als enige kritieke punten een lokaal minimum in (0,0) en een zadelpunt in (1,-1) bezit.

Zij

$$F(x,y) := x^2 + xy + \frac{y^2}{2} - \frac{x^3}{3}.$$

De kritieke punten van F voldoen aan:

$$\frac{\partial F}{\partial x}(x,y) = 2x + y - x^2 = 0$$

$$\frac{\partial F}{\partial y}(x,y) = x + y = 0.$$

Dus $y = -x$ en $x = x^2$, of wel $x = 0$, $y = 0$ en $x = 1$, $y = -1$.

De Hessiaan van F wordt gegeven door

$$\begin{bmatrix} \frac{\partial^2 F}{\partial x^2}(x,y) & \frac{\partial^2 F}{\partial x \partial y}(x,y) \\ \frac{\partial^2 F}{\partial x \partial y}(x,y) & \frac{\partial^2 F}{\partial y^2}(x,y) \end{bmatrix} = \begin{bmatrix} 2-2x & 1 \\ 1 & 1 \end{bmatrix}.$$

Voor $x = y = 0$ is $\begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}$ positief definiet, dus $(0,0)$ is een lokaal minimum.
Voor $x = 1$, $y = -1$ is $\det \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} = -1$, dus $(1,-1)$ is een zadelpunt. Omdat $F(0,0) = 0$, en $(0,0)$ een strikt locaal minimum is, zijn er getallen $R, d > 0$ zodat $F(x,y) \geq d$ voor $x^2 + y^2 = R^2$ (R voldoende klein). Kies $e = (3,0)$; dan is $F(3,0) = 0 < d$.

Voor iedere rij $(x_n, y_n) \in \mathbb{R}^2$ zodanig dat

$$\lim_{n\to\infty} F(x_n, y_n) \text{ bestaat en } \lim_{n\to\infty} F'(x_n, y_n) = 0$$

hebben we

a) $$\lim_{n\to\infty} \left(x_n^2 + x_n y_n + \frac{y_n^2}{2} - \frac{x_n^3}{3}\right) = \alpha \in \mathbb{R}$$

b) $$\lim_{n\to\infty} (2x_n + y_n - x_n^2) = 0$$

c) $$\lim_{n\to\infty} (x_n + y_n) = 0.$$

Uit b) en c) volgt: $\lim_{n\to\infty} x_n(1-x_n) = 0$. Dus

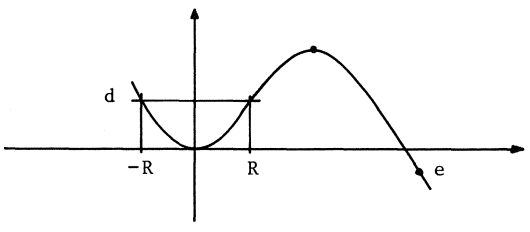
of $$\lim_{n\to\infty} x_n = 0 \text{ en } \lim_{n\to\infty} y_n = 0$$

of $$\lim_{n\to\infty} x_n = 1 \text{ en } \lim_{n\to\infty} y_n = -1$$

en aan "P.S." is voldaan.

Merk op dat $c = F(1,-1) = \frac{1}{6} > 0$. Omdat $(0,0)$ en $(1,-1)$ de enige kritieke punten zijn, is $c = \frac{1}{6}$ de kritieke waarde, met $K = \{(1,-1)\}$ die gekarakteriseerd wordt door het Mountain Pass Lemma.

Situatie voor $N = 1$



Figuur 3

BEWIJS VAN HET MOUNTAIN PASS LEMMA

We gebruiken weer het Lemma, nu met $K = [0,1]$, $K_0 = \{0,1\}$, $\psi(0) = 0$, $\psi(1) = e$.

Voor $u \in \Gamma$ bestaat er een $t_0 \in (0,1)$ zodanig dat $\| u(t_0) \| = R$ zodat

$$\max_{t\in[0,1]} F(u(t)) \geq F(u(t_0)) \geq d \text{ en } c = \inf_{u\in\Gamma} \max_{t\in[0,1]} F(u(t)) \geq d > b =$$

$\max (F(0), F(e))$. □

We beschouwen weer het voorbeeld 1 waar A positief definiet is, maar nu met g niet begrensd.

Als $g(0) = 0$ en $\lim_{x\to 0} \frac{g(x)}{\| x \|} = 0$ dan is 0 een lokaal minimum van F

(bijvoorbeeld $G(x) = \frac{1}{4} \sum_{i=1}^{n} x_i^4$, $g'(x) = \begin{bmatrix} x_1^3 \\ \vdots \\ x_n^3 \end{bmatrix}$).

Er zijn dan getallen $d > 0$ en $R > 0$ (R voldoende klein) waarvoor geldt $F(0) < d$ en $F(x) \geq d$ voor $\| x \| = R$. Als er bovendien een $y \in \mathbb{R}^n$ bestaat zodanig dat

$$\sup_{\lambda > 0} \frac{G(\lambda y)}{\lambda^2} \geq \frac{1}{2} y^T A y$$

dan geldt voor λ voldoende groot $F(\lambda y) < d$. Hieruit volgt dan dat F een ander kritiek punt (dan (0,0)) heeft indien aan "P.S." voldaan is. Bijvoorbeeld

$$Ax = \begin{bmatrix} x_1^3 \\ \vdots \\ x_n^3 \end{bmatrix}, \text{ (met A positief-definiet)}$$

bezit altijd tenminste één oplossing ongelijk aan de nuloplossing.

OPMERKING

In dit laatste voorbeeld is F even en dus heeft F een paar kritieke punten ongelijk aan nul. Bovendien geldt hiervoor dat $F(x) \leq 0 < d$ voor alle $x \in \mathbb{R}^n$ mits $\| x \| \geq R'$, met R' groot genoeg. Een resultaat analoog aan dat van de stelling van Lyusternik garandeert het bestaan van tenminste N paren verschillende kritieke punten ongelijk aan nul! [1].

Als laatste toepassing van het Mountain Pass Lemma vermelden we de volgende stelling [9]:

Zij $F \in C^1(\mathbb{R}^n;\mathbb{R})$ een functie die aan "P.S." voldoet en neem aan dat F twee verschillende lokale minima bezit. Dan bestaat er een derde kritieke punt.

APPENDIX

BEWIJS VAN HET LEMMA (voor $F \in C^2(\mathbb{R}^N;\mathbb{R})$).

Beschouw de differentiaalvergelijking

$$\text{(D.V.)} \qquad \begin{cases} \frac{d}{dt} u(t) + g(u(t))\, h(\| F'(u(t)) \|)\, F'(u(t)) = 0 \\ u(0) = x \in \mathbb{R}^N \end{cases}$$

met $$h(s) := \begin{cases} 1 & \text{voor } s \le 1 \\ s^{-1} & \text{voor } s > 1 \end{cases}$$

en $$g(y) := \frac{d(y,A)}{d(y,A) + d(y,B)}$$

waarbij $$A := \{x \in \mathbb{R}^N \mid F(x) \le b\}$$

$$B := \{x \in \mathbb{R}^N \mid F(x) \ge c\}$$

en $d(y,C) = \inf_{z \in C} \| y - z \|$ voor $C \subset \mathbb{R}^N$.

Omdat $b < c$, is $A \cap B = \emptyset$ en voldoet g aan:

i) $g(y) = 1$ als $F(y) \ge c$

ii) $g(y) = 0$ als $F(y) \le b$

iii) $0 \le g(y) \le 1$ voor alle $y \in \mathbb{R}^N$

iv) $g : \mathbb{R}^N \to \mathbb{R}$ is lokaal Lipschitz continu.

Omdat $y \to H(y) := g(y)\, h(\| F'(y) \|)\, F'(y)$ lokaal Lipschitz continu is met $0 \le \|H(y)\| \le 1$ voor alle $y \in \mathbb{R}^N$, bezit de D.V. voor elke $x \in \mathbb{R}^N$ één en slechts één oplossing $t \to u(t;x)$ en die bestaat voor alle $t \ge 0$. Bovendien is $x \to u(t;x)$ continu op $\mathbb{R}^N$ voor alle $t \ge 0$. Stel voor het gemak $T(t)x := u(t;x)$ voor alle $t \ge 0$, $x \in \mathbb{R}^N$. Dan is $T(t) : \mathbb{R}^N \to \mathbb{R}^N$ dus continu voor alle $t \ge 0$. Voor $x \in \mathbb{R}^N$ waarvoor $F(x) \le b$ geldt $g(x) = 0$, en dus $H(x) = 0$. Dan is $u(t,x) = x$, voor alle $t \ge 0$ een oplossing van (D.V.) en wegens de eenduidigheid van de oplossingen geldt $T(t)x = x$ voor alle

$t \geq 0$ en elke x met $F(x) \leq b$. Voor alle $x \in \mathbb{R}^N$ en $t > 0$ is

$$\frac{d}{dt} F(T(t)x) = F'(T(t)x)^T \frac{d}{dt} u(t) =$$

$$= -g(T(t)x)\, h(\| F'(T(t)x) \|) \| F'(T(t)x) \|^2 \leq 0.$$

Hieruit volgt dat voor alle $x \in \mathbb{R}^N$, de functie $t \to F(T(t)x)$ niet stijgend is. Kies $\varepsilon \in (0,1)$. Uit de definitie van c volgt dat er een $\phi_\varepsilon \in \Gamma$ bestaat zodanig dat $c \leq \max_{x \in K} F(\phi_\varepsilon(x)) \leq c + \varepsilon$.

Dan is $T(t) \circ \phi_\varepsilon : K \to \mathbb{R}^N$ continu voor alle $t \geq 0$

en $$T(t) \circ \phi_\varepsilon \mid K_0 = T(t) \circ \psi \text{ voor alle } t \geq 0.$$

Voor $x \in K_0$ geldt $F(\psi(x)) \leq b$ dus $T(t)\psi = \psi$ voor $t \geq 0$. Hieruit volgt dat $T(t) \circ \phi_\varepsilon \in \Gamma$ voor $t \geq 0$. Uit de definitie van c volgt dat

$$c \leq \max_{x \in K} F(T(t) \circ \phi_\varepsilon(x)) \leq \max_{x \in K} F(\phi_\varepsilon(x)) \leq c + \varepsilon.$$

Dit impliceert dat voor iedere $t > 0$ er een $x(t,\varepsilon) \in K$ bestaat zodanig dat

$$c \leq F(T(t) \circ \phi_\varepsilon(x(t,\varepsilon)) \leq c + \varepsilon.$$

Omdat K compact is, is er dan een element $x_\varepsilon \in K$ en een rij $(x(t_n,\varepsilon))$ zo dat $t_n \uparrow \infty$ en $x(t_n,\varepsilon) \to x_\varepsilon$ als $n \to \infty$. Wegens de continuiteit van ϕ_ε geldt $\lim_{n \to \infty} \phi_\varepsilon(x(t_n,\varepsilon)) = \phi_\varepsilon(x_\varepsilon)$. Neem nu $\bar{t} > 0$, en n groot genoeg zodat $t_n \geq \bar{t}$.
Dan geldt:

$$c \leq F(T(t_n) \circ \phi_\varepsilon(x(t_n,\varepsilon))) \leq F(T(\bar{t}) \circ \phi_\varepsilon(x(t_n,\varepsilon))) \leq$$

$$\leq F(\phi_\varepsilon(x(t_n,\varepsilon))) \leq c + \varepsilon.$$

Dus $$c \leq F(T(\bar{t}) \circ \phi_\varepsilon(x(t_n,\varepsilon))) \leq c + \varepsilon.$$

Als we de limiet voor $n \to \infty$ nemen krijgen we

$$c \leq F(T(\bar{t}) \circ \phi_\varepsilon(x_\varepsilon)) \leq c + \varepsilon.$$

Omdat $\bar{t} > 0$ willekeurig is, geldt $c \leq F(T(t)y_\varepsilon) \leq c + \varepsilon$ voor alle $t > 0$ met $y_\varepsilon := \phi(x_\varepsilon)$. Omdat $t \to F(T(t)y)$ niet stijgend, en naar beneden begrensd is bestaat er een $t_\varepsilon > 0$ zodanig dat $\left| \frac{d}{dt} F(T(t)y_\varepsilon) \right| < \varepsilon$.
Dan geldt:

$$g(T(t_\varepsilon)y_\varepsilon)h(\| F'(T(t_\varepsilon)y_\varepsilon)\|).\| F'(T(t_\varepsilon)y_\varepsilon)\|^2 < \varepsilon.$$

Merk op dat $F(T(t_\varepsilon)y_\varepsilon) \geq c$ en dus $g(T(t_\varepsilon)y_\varepsilon) = 1$. Omdat $\varepsilon \in (0,1)$, geldt $\| F'(T(t_\varepsilon)y_\varepsilon \| < \sqrt{\varepsilon}$. Neem nu, voor $n \in \mathbb{N}^+$, $\varepsilon = \frac{1}{n}$, en $z_n = T(t_\varepsilon)y_\varepsilon$; dan hebben we

(1) $$c \leq F(z_n) \leq c + \frac{1}{n}$$

(2) $$\| F'(z_n)\| \leq \sqrt{\frac{1}{n}}.$$

Daarmee is het lemma bewezen.

Ik wou graag mijn collega Prof.dr. A.W. Grootendorst bedanken voor de vertaling van de inleiding.

REFERENTIES

[1] A. AMBROSETTI & P. RABINOWITZ, *Dual variational methods in critical point theory and applications*, J. Funct. Anal., 14 (1973) pp. 349-381.

[2] I. EKELAND, *On the variational principle*, J. Math. Anal. Appl., 47 (2) (1974) pp. 324-353.

[3] E. FISCHER, *Über quadratische Formen mit reellen Koeffizienten*, Monatsh. f. Math. Phys., 16 (1905) pp. 234-249.

[4] L.A. LYUSTERNIK, *Topologische Grundlagen der allgemeinen Eigenwerttheorie*, Monatsh. f. Math. Phys., 37 (1930) pp. 125-130.

[5] L.A. LYUSTERNIK & L.G. SCHNIRELMANN, *Topological Methods in the Calculus of Variations*, Hermann, Paris (1934).

[6] L. NIRENBERG, *Variational and topological methods in nonlinear problems*, Bull. Amer. Math. Soc., 4 (1981) pp. 267-302.

[7] R.S. PALAIS, *Critical point theory and the minimax principle*, Proc. Sym. Pure Math., 15, A.M.S., Providence, R.I. (1970) pp. 185-212.

[8] R. PALAIS & S. SMALE, *A generalized Morse theory*, Bull. Amer. Math. Soc., 70 (1964) pp. 165-170.

[9] P. PUCCI & J. SERRIN, *Extensions of the Mountain Pass Theorem*, J. Funct. Anal., 59 (1984) pp. 185-210.

[10] P.H. RABINOWITZ, *Variational methods for nonlinear eigenvalue problems*; *Eigenvalues of Nonlinear Problems*, G. Prodi (editor), Edizione Cremonese, Roma (1974) pp. 141-195.

[11] P.H. RABINOWITZ, *A minimax principle and applications to elliptic partial differential equations;* in *Nonlinear Partial Differential Equations and Applications*, *Lecture Notes in Mathematics* 648, Springer-Verlag (1978) pp. 97-115.

[12] P.H. RABINOWITZ, *Periodic solutions of Hamiltonian systems*, Comm. Pure Appl. Math., 31 (1978) pp. 157-184.

[13] A. WEINSTEIN, *Periodic orbits for convex Hamiltonian systems*, Ann. Math., 108 (1978) pp. 507-518.

HOOFDSTUK 4

CONSISTENTE BENADERINGEN IN DE MATHEMATISCHE PHYSICA

L.J.F. BROER

1. INLEIDING

In de mathematische physica worden vergelijkingen die uit de theoretische physica volgen nader onderzocht en, zo mogelijk, opgelost. Dit is vaak heel moeilijk; er is dan behoefte aan betrouwbare vereenvoudigingen door geschikte benaderingen. Deze benaderingen dienen dan de voor het onderhavige probleem belangrijke eigenschappen van de vergelijkingen en hun oplossingen in hoofdzaak onveranderd te laten.

We beperken ons in het volgende tot conservatieve golfvergelijkingen. Ruwweg betekent dit dat er geen demping of wrijving is. Physisch komt het er als regel op neer dat de totale energie constant is. Dit soort vergelijkingen kan meestal afgeleid worden uit een variatieprincipe van de Lagrange of Hamilton soort. De kern van het volgende verhaal is te laten zien dat deze formulering van het probleem een belangrijk hulpmiddel kan zijn bij het opsporen van geschikte benaderingen. Hierbij zullen we de belangrijkste elementaire eigenschappen van het Lagrange- of Hamilton-(afgekort L - of H -) formalisme bekend veronderstellen. Een tweetal opmerkingen die in inleidende leerboeken meestal niet voorkomen voegen we hieraan toe.

De eerste betreft de H-formuleringen. Laten we indices (discreet systeem) of ruimtelijke coördinaten (continu systeem) weg dan kunnen de Hamiltonse vergelijkingen worden geschreven als

$$q_{,t} = \frac{\delta H}{\delta p}, \qquad p_{,t} = -\frac{\delta H}{\delta q} \tag{1.1}$$

waarin we gebruik maken van de notatie $\frac{\delta H}{\delta p}$, resp. $\frac{\delta H}{\delta q}$, met de betekenis van partiële afgeleide van de functie H naar p, resp. q, voor het geval van discrete systemen, en met de betekenis van variatieafgeleide van de functionaal H naar p, resp. q, voor continue systemen (zie Hoofdstuk 2, §4).

Beschouw nu q, p als componenten van een vector u. Dan kan (1.1) geschreven worden als

(1.2) $$u_t = S.\frac{\delta H}{\delta u} \quad \text{of} \quad S^{-1}u_t = \frac{\delta H}{\delta u}$$

waarin de operator S in matrixvorm is:

$$S = \begin{vmatrix} 0 & 1 \\ -1 & 0 \end{vmatrix} .$$

S is een symplectische operator, d.w.z. $SS^T = 1$ (unitair) en $S = -S^T$ (antisymmetrisch). De opmerking is nu dat voor de algemene theorie eigenlijk alleen het laatste van belang is. Vergelijkingen van het type:

(1.3) $$u_t = A\frac{\delta H}{\delta u}, \quad A^{-1}u_t = \frac{\delta H}{\delta u}$$

noemen we gegeneraliseerde Hamilton - vergelijkingen (afgekort g.H - vergelijkingen). Vergelijking (1.3) volgt uit een kanoniek actie - principe van de vorm

$$\int L(u,u_t)dt$$

waarin

$$L(u,u_t) = \tfrac{1}{2}(u,A^{-1}u_t) - H(u)$$

met (,) het Euclidische inproduct voor discrete systemen, en het L_2 - inproduct over de ruimtelijke variabele voor continue systemen. Wanneer A = S dan is dit de standaard formulering van het kanoniek actie - principe voor (1.2) (zie Hoofdstuk 2, §4).

Ofschoon we dat verder niet zullen gebruiken, merken we op dat het wel bekend is dat algemene eigenschappen van klassieke H- vergelijkingen geformuleerd kunnen worden met behulp van Poissonhaken.
Voor (1.2) kan de Poisson-haak van twee functies geschreven worden als een bilineaire vorm:

(1.4) $$\{F,G\} = \left(\frac{\delta F}{\delta u}, S\frac{\delta G}{\delta u}\right).$$

Het blijkt nu dat voor g.H - vergelijkingen de uitdrukkingen

(1.5) $$\{F,G\} = \left(\frac{\delta F}{\delta u}, A\frac{\delta G}{\delta u}\right)$$

dezelfde hoofdeigenschappen (i.h.b. de Jacobi-regel) hebben als (1.4).

In het bovenstaande is $A = -A^T$ een constante, i.e. niet van t afhangende, niet-singuliere operator. Het is mogelijk, maar voor ons doel niet nodig, de zaak nog verder te generaliseren voor een bepaalde klasse van niet-lineaire, van u afhankelijke, operatoren[1]).

De tweede opmerking betreft lineaire vergelijkingen. Een lineaire conservatieve vergelijking volgt middels het L-variatie principe uit een homogeen kwadratische L-functionaal die de tijd niet expliciet bevat. In het algemeen is het dan mogelijk om andere, essentieel verschillende L-formuleringen voor hetzelfde probleem te vinden. Uit het principe van Noether volgt dat er (voor elke t-onafhankelijke L) een kwadratische behoudswet is. Slechts één hiervan kan de energie zijn. Welke dit is, is een physisch probleem; het volgt uit de physische betekenis van de gebruikte variabelen. Laat nu de te onderzoeken benadering van een gecompliceerd probleem bestaan uit een linearisering. Neem aan dat voor deze lineaire vergelijkingen een kwadratische Langrangiaan te vinden is. Er is dan geen enkele zekerheid dat deze iets te maken heeft met een "kwadratisering" van de L van het complete probleem. Dit probleem heeft n.l. in het algemeen maar één L, het lineaire probleem heeft er vele. Overeenkomst zou dus een toevalstreffer zijn. Een specifiek voorbeeld hiervan komt aan het eind van de volgende paragraaf aan de orde.

2. KORTE GOLVEN IN EEN INHOMOGEEN MEDIUM

In deze paragraaf bekijken we golfvergelijkingen van het type

$$\frac{1}{c^2(x)}\, u_{tt} = u_{xx} \tag{2.1}$$

b.v. geluidsgolven in een inhomogeen medium. De golfsnelheid c(x) is overal positief. We nemen verder aan dat c in het oneindige naar vaste, eventueel links en rechts verschillende, asymptotische waarden gaat en dat er zoveel eindige en continue afgeleiden bestaan als nodig.

Vergelijking (2.1) volgt uit het variatieprincipe voor $\int L(u,u_t)dt$ met

$$L = \int dx \left[\frac{u_t^2}{2c^2} - \frac{u_x^2}{2} \right]. \tag{2.2}$$

Er is een behoudswet:

(2.3) $$E_t + F_x = 0$$

met

(2.4) $$E = \frac{u_t^2}{2c^2} + \frac{u_x^2}{2}, \qquad F = - u_x u_t .$$

(2.3) bestaat, volgens Noether, omdat L de tijd niet expliciet bevat. Een algemene analytische oplossing van (2.1) voor een willekeurige functie c(x) is niet bekend. We gaan nu zoeken naar een benadering voor korte golven. Dit betekent ruwweg dat de relatieve verandering van c, $\frac{\Delta c}{c}$, over een golflengte klein is t.o.v. 1. De grootte orde van deze gradiënt van c wordt verder aangegeven door de parameter ε. Een eerste stap is op te merken dat c(x) de karakteristieke snelheid in (2.1) is. D.w.z. kleine discontinuïteiten in u of u_x lopen volgens $dx = \pm\, c\,dt$. Het ligt dus voor de hand een nieuwe variabele y in te voeren volgens

(2.5) $$c\,dy = dx, \qquad y = \int^{x} \frac{dx'}{c(x')} .$$

Om nu de resterende factor c in (2.2) weg te werken voeren we een nieuwe afhankelijke variabele in:

(2.6) $$u = c^{\frac{1}{2}}\, w.$$

Substitutie van (2.5) en (2.6) in (2.2) levert

(2.7) $$L = \int dy\ \tfrac{1}{2}[w_t^2 - w_y^2 - w^2(\beta' + \beta^2)].$$

Hierin is $\beta = - \frac{1}{2c} \cdot \frac{dc}{dy}$, $\beta' = \frac{d\beta}{dy}$. De getransformeerde vergelijking is dan:

(2.8) $$w_{tt} - w_{yy} + w(\beta' + \beta^2) = 0.$$

De laatste termen in (2.7) en (2.8) zijn $O(\varepsilon^2)$. Een benadering voor korte golven wordt dus verkregen door deze termen eenvoudig weg te laten. Dit levert:

(2.9) $$w_{tt} - w_{yy} = 0.$$

Oplossingen zijn:

(2.10) $$w(q,t) = r_0(t-y) + s_0(t+y)$$

voor willekeurige functies r_0 en s_0, en dus

(2.11) $$u(x,t) = c^{\frac{1}{2}}(x)\,[r_0(t - \int^x \frac{dx'}{c(x')}) + s_0(t+ \int^x \frac{dx'}{c(x')})].$$

De oplossing (2.11) wordt wel de W.K.B.-benadering genoemd. Historisch juister is de naam Liouville-Green benadering[2]). Het is lokaal een goede benadering. Een tekortkoming is echter dat er geen reflectie in voorkomt. De uitgangsvergelijking (2.1) heeft, als c(x) niet constant is, altijd reflectie. Dit is b.v. in te zien door de volgende combinatie van wiskundige en fysische argumenten. Omdat c de karakteristieke snelheid is zou voor een naar rechts lopende golfberg, zeg $U(x,t) = \phi(x-c(x,t)$ in laagste orde, met ϕ bijvoorbeeld een functie met compacte drager, uit (2.3) volgen dat in laagste orde:

$$F - c\,E = 0.$$

Uit (2.4) volgt echter

(2.12) $$F - cE = -\frac{1}{c}(u_t + cu_x)^2 .$$

Het is niet moeilijk in te zien dat oplossingen van (2.1) waarvoor $u_t + cu_x$ overal nul is alleen mogelijk zijn voor constante c. Reflectie is dus essentieel voor niet-lokale oplossingen. Dit reflectieprobleem kan als volgt worden aangepakt.

Ga uit van de vergelijkingen:

(2.13) $$\begin{aligned} r_t + r_y &= \beta s \\ s_t - s_y &= -\beta r. \end{aligned}$$

Differentiatie van beide vergelijkingen naar t en naar y en een geschikte combinatie van de resultaten leert dan dat w = r+s voldoet aan (2.8). Om de reflectie van een naar rechts lopende golf te onderzoeken nemen we als randvoorwaarden:

$$r = r_0(y-t), \quad s = 0 \quad \text{als} \quad y \to +\infty.$$

Omdat β van de orde r is schrijven we:

$$r = r_0 + \varepsilon^2 r_2 + \ldots; \quad s = \varepsilon s_1 + \varepsilon^3 s_3 + \ldots .$$

Dus $$r_{2n,t} + r_{2n,y} = \beta\, s_{2n+1}, \quad s_{2n+1,t} - s_{2n+1,y} = -\beta\, r_{2n}.$$

De eerste orde bijdrage tot de reflectie volgt dus uit

$$s_{1,t} - s_{1,y} = -\beta r_0 (y-t)$$

met $s_1 = 0$ als $y \to +\infty$. Voor monochromatische golven komt dit neer op de methode van Bremmer [3].

We vermelden nog dat uit (2.13) volgt:

$$(r^2 + s^2)_t + (r^2 - s^2)_y = 0$$

waaruit nog eens blijkt dat de splitising $w = r+s$ neerkomt op een splitsing in naar rechts en naar links lopende golven. We kunnen nu (2.13) schrijven in de vorm

$$\binom{r}{s}_t + \begin{vmatrix} \partial_y - \beta & \\ & \beta - \partial_y \end{vmatrix} \begin{pmatrix} \frac{\partial G}{\partial r} \\ \frac{\partial G}{\partial s} \end{pmatrix} = 0 \tag{2.14}$$

waarin $G = \frac{1}{2}(r^2 + s^2)$ dus een bewegingsconstante is. De vorm (2.14) is een voorbeeld van een g.H-vergelijking in de vorm (1.3).

Tenslotte merken we op dat andere transformaties van (2.1) van nut kunnen zijn [4]. Een voorbeeld is de substitutie

$$v = u_x . \tag{2.15}$$

Dit levert dan de golfvergelijking:

$$v_{tt} = (c^2 v_x)_x . \tag{2.16}$$

Deze vergelijking heeft de Langrangiaan

$$L' = \int dx \left[\frac{v_t^2}{2} - \frac{c^2}{2} v_x^2 \right]. \tag{2.17}$$

Volgens Noether is er weer een behoudswet die samenhangt met de invariantie van (2.17) voor tijdverschuiving. Deze is nu

$$(\tfrac{1}{2} v_t^2 + c^2 v_x^2)_t - (c^2 v_x v_t)_x = 0. \tag{2.18}$$

Het is duidelijk dat (2.17) en (2.18) niet overeenkomen met (2.2) en (2.4). De vraag of (2.3), (2.18) of wellicht nog een andere relatie, de energiebalans is kan slechts beantwoord worden wanneer de fysische betekenis van u of v bekend is. We geven deze transformatie hier slechts als voorbeeld van een in § 1 vermelde eigenschap van lineaire conservatieve vergelijkingen.

3. LANGE GOLVEN IN DISPERSIEVE HOMOGENE MEDIA

In deze paragraaf bekijken we vergelijkingen van het type:

$$u_{tt} + \int_{-\infty}^{\infty} dy.\, R(x-y)\, u(y) = 0 \tag{3.1}$$

waarin R(z) een reële, symmetrische en overigens "nette" functie is. Een formele oplossing van (3.1) kan verkregen worden door een Fourier transformatie. De hierbij gebruikte basisoplossingen zijn:

$$u_k(x,t) = \exp(-\, i[\omega t \pm kx]).$$

Dit is, met complexe schrijfwijze, een functie die een golf voorstelt, met golflengte $\frac{2\pi}{k}$, frequentie ω (dus $\frac{2\pi}{\omega}$ als periode) en constante voortplantingssnelheid $\frac{\omega}{k}$ (de zgn. fase-snelheid). Door substitutie van deze uitdrukking in (3.1) volgt dat dit een oplossing is mits ω en k aan een zekere relatie voldoen. Deze zogenaamde dispersie-relatie luidt in dit geval:

$$\omega^2 = \hat{R}(k) \tag{3.2}$$

waarin $\hat{R}$, symmetrisch in k, de F-getransformeerde van R is. We interesseren ons nu voor de dispersie-effecten voor golven die t.a.v. een of ander criterium lang zijn. Het ligt dan voor de hand om $\hat{R}$ naar k te ontwikkelen

$$\hat{R} = c^2k^2 + \alpha\, k^4 + \ldots \tag{3.3}$$

waarbij aangenomen is dat $\hat{R}(0) = 0$. c is de snelheid van zeer lange golven (i.e. nu $k \to 0$) en α een dispersieparameter. Een benadering voor de relatie (3.3) tot en met 4^e orde in k, $k \to 0$, leidt tot de golfvergelijking

(3.4) $$u_{tt} = c^2 u_{xx} - \alpha u_{xxxx} .$$

Deze vergelijking kan afgeleid worden uit de Lagrangriaan

$$L = \int dx \, [\tfrac{1}{2} u_t^2 - \frac{c^2}{2} u_x^2 - \frac{\alpha}{2} u_{xx}^2].$$

Dit gaat allemaal probleemloos zolang $\alpha > 0$, d.w.z. wanneer de kortere golven wat sneller zijn dan de lange. In de praktijk, b.v. bij watergolven, is het omgekeerde vaak het geval. Dit kan tot moeilijkheden leiden. Als voorbeeld nemen we de vergelijking

(3.5) $$u_{tt} = u_{xx} + u_{xxxx}$$

die als dispersierelatie heeft

(3.6) $$\omega^2 = k^2(1-k^2).$$

Blijkbaar is het systeem instabiel voor $k^2 > 1$: Wanneer het spectrum van de beginwaarden daar niet exact nul is kan de korte-golf bijdrage exponentieel, als functie van t, stijgen. Neem b.v. als begintoestand een symmetrische golfberg in rust:

$$u(x,0) = U(x) = U(-x), \quad u_t(x,0) = 0.$$

De oplossing is dan formeel

$$u(x,t) = \int_0^1 dk \, \hat{U}(k) \cos kx . \cos k(1-k^2)^{\frac{1}{2}} t + \int_1^\infty dk \, \hat{U}(k) \cos kx . \cosh k(1-k^2)^{\frac{1}{2}} t.$$

De tweede term "ontploft" voor $t \to \infty$.

Er is dus nu een betere benadering voor (3.1) nodig. Voor lineaire problemen kan dit natuurlijk op allerlei manieren. We willen echter graag iets doen waardoor de vergelijking niet veel ingewikkelder wordt om analoge methoden ook op meer gecompliceerde niet-lineaire vergelijkingen toe te passen zijn.

De vergelijking (3.5) heeft als Lagrangiaan:

(3.7) $$L = \int dx \, [\tfrac{1}{2} u_t^2 - \tfrac{1}{2} u_x^2 + \tfrac{1}{2} u_{xx}^2].$$

Substitueer nu

(3.8) $$u = v + v_x \ .$$

Dan vinden we

$$L = \int dx \left[\tfrac{1}{2} v_t^2 + \tfrac{1}{2} v_{tx}^2 - \tfrac{1}{2} v_x^2 + \tfrac{1}{2} v_{xxx}^2 \right] .$$

De laatste term is van de zesde orde en kan worden weggelaten. Dit is immers ook al gebeurd met de zesde orde termen bij de overgang van (3.1) naar (3.5). Op deze manier vinden we de benadering

(3.9) $$v_{tt} - v_{ttxx} - v_{xx} = 0$$

met de dispersierelatie:

(3.10) $$\frac{\omega^2}{k^2} = \frac{1}{1+k^2} \ .$$

Voor kleine k^2 is dit ongeveer hetzelfde als (3.6). Voor grote k^2 zijn (3.6) en (3.10) verschillend en beide i.h.a. niet juist. Het essentiële punt is echter dat dit in (3.10) onschadelijk is omdat ω reëel blijft. De oplossing van het golfbergprobleem is

$$u(x,t) = \int_0^\infty dk \ \hat{U}(k) \cos kx. \cos \frac{kt}{(1+k^2)^{\frac{1}{2}}} \ .$$

Wanneer U(k) klein genoeg is voor grote k is dit een bruikbare benadering voor een oplossing van (3.1). Voor het analoge resultaat bij (3.5) is dit i.h.a. niet het geval.

4. WATERGOLVEN

De beweging van vloeistoffen is een buitengewoon ingewikkelde zaak; benaderingen zijn hiervoor dus belangrijk en noodzakelijk. Deze benaderingen kunnen, naar de aard van het probleem, van de meest uiteenlopende soort zijn. In deze paragraaf kijken we naar golven die onder invloed van de zwaartekracht ontstaan in een horizontale laag water met een vrij oppervlak. Het eenvoudigste is dit wanneer de bodem horizontaal is en er slechts één horizontale coördinaat meegenomen wordt. De wrijving (viscositeit) verwaarlozen we ook nog, evenals de compressibiliteit. Het overblijvende probleem laat een, reeds lang bekende, fraaie analytische

formulering toe. Exacte oplossingen hiervan zijn echter slechts in speciale gevallen (periodieke en solitaire golven) te vinden.[5]) Ook de computer biedt niet altijd uitkomst: verdere benaderingen zijn daarom zeer gewenst. Bij de formulering hiervan zijn twee dimensieloze parameters van belang. Deze zijn:

$$\frac{a}{h} = \frac{\text{amplitude}}{\text{waterdiepte}} \quad \text{en} \quad h^2k^2 \sim \left(\frac{\text{waterdiepte}}{\text{golflengte}}\right)^2 .$$

De eerste heeft te maken met de niet-lineariteit, de tweede met de dispersie. Een belangrijk en veel bestudeerd geval is nu dat beide parameters van dezelfde orde en vrij klein zijn. Ze worden dan vaak slechts in de eerste orde meegenomen. Ook met deze beperking zijn er nog vele benaderingen mogelijk. Een van de oudste wordt gevormd door de Boussinesq vergelijkingen. Na geschikte schaaltransformaties komen deze neer op:

$$\eta_t + [(1+\eta)u + u_{xx}]_x = 0 \tag{4.1}$$

$$u_t + (\tfrac{1}{2}u^2 + \eta)_x = 0. \tag{4.2}$$

Hierin is u (een maat voor) de horizontale snelheid aan het oppervlak, η het hoogteverschil tussen oppervlak in beweging en in evenwicht. We merken op dat (4.1) en (4.2) een g.H-stelsel vormen met

$$A = \begin{vmatrix} 0 & -\partial_x \\ -\partial_x & 0 \end{vmatrix}, \quad H = \int [\tfrac{1}{2}(1+\eta)u^2 + \tfrac{1}{2}\eta^2 - \tfrac{1}{2}u_x^2]dx .$$

In een aantal gevallen hebben deze vergelijkingen redelijke oplossingen. Echter, het teken van de laatste term in H bedreigt de stabiliteit van het systeem (4.1), (4.2). Linearisering levert een vergelijking als (3.1): het systeem is dus toch verdacht. Tientallen jaren heeft men zich dit niet gerealiseerd; duidelijke gevallen van situaties waarin (4.1) en (4.2) geen aanvaardbare oplossingen hebben zijn pas later gevonden met behulp van een computer. Een voorbeeld hiervan is weer het golfbergprobleem:

$$\eta(x,0) = f(x) , \qquad u(x,0) = 0$$

wanneer $\hat{f}(k) \neq 0$ voor $k^2 > 1$.

Verbetering kan bereikt worden op dezelfde manier als in §3. We voeren een nieuwe variabele in door:

$$u = v + v_x = Dv.$$

Hiervoor geldt:

$$u_t = Dv_t , \qquad D^T \frac{\delta H}{\delta u} = \frac{\delta H}{\delta v}$$

waarin D^T de formeel geadjungeerde is van de operator D, dus $D^T v = v - v_x$. De nieuwe vergelijkingen in g.H-vorm zijn dus

$$D^T \eta_T = - \left(\frac{\delta H}{\delta v}\right)_x$$

en

$$Dv_t = - \left(\frac{\delta H}{\delta \eta}\right)_x .$$

De getransformeerde H is:

$$H = \int dx \left[\tfrac{1}{2}(1+\eta)\, v^2 + \tfrac{1}{2}\eta^2\right] \tag{4.3}$$

wanneer termen van een hogere orde dan 3 worden weggelaten. Op deze manier vinden we de vergelijkingen:

$$\eta_t - \eta_{tx} + [(1+\eta).v]_x = 0 \tag{4.4}$$

$$v_t + v_{tx} + [\tfrac{1}{2}v^2 + \eta]_x = 0. \tag{4.5}$$

Computerresultaten voor (4.4) en (4.5) zijn mij niet bekend. De "gekortwiekte" Hamiltoniaan (4.3) lijkt echter de stabiliteit in een voldoend ruime omgeving van het evenwicht $v = \eta = 0$ te garanderen.
Er is nog een andere manier om een voldoend stabiele benadering te verkrijgen, en wel in de oorspronkelijke variabelen.[6]) Dit gaat met behulp van een Green-operator of Greense functie. Laat G de inverse zijn van de, positieve, operator $1 - \partial_x^2$. We nemen dan:

$$H = \int dx. \left[\tfrac{1}{2}\, u.Gu + \tfrac{1}{2}\, \eta u^2 + \tfrac{1}{2}\, \eta^2\right] \tag{4.6}$$

en dus:

$$\eta_t + [Gu + \eta u]_x = 0 \tag{4.7}$$

$$u_t + [\tfrac{1}{2}u^2 + \eta]_x = 0. \tag{4.8}$$

In het gelineariseerde probleem komt dit neer op het vervangen van de gevaarlijke factor $(1-k^2)$ in de dispersierelatie door de tamme factor $(1+k^2)^{-1}$. In (4.7) is G een integraaloperator; voor computerwerk lijkt dit geen probleem. Het stabiliteitsgebied van (4.6) lijkt weer groot genoeg voor de praktische toepassingen van dit soort vergelijkingen.

5. EEN PROCES-VERGELIJKING

In de praktijk van de theoretische mathematische fysica is de laatste jaren bijzondere aandacht gegeven aan een soort niet-lineaire partiële differentiaalvergelijkingen die als "volledig integreerbaar" bekend staan. De meest markante eigenschappen hiervan zijn dat het, niet-lineaire, conservatieve vergelijkingen zijn met een oneindige rij van bewegingsconstanten of behoudswetten en dat zij een zekere relatie hebben met een lineaire vergelijking waardoor zij in principe opgelost kunnen worden.[7]) Het bekendste voorbeeld hiervan is de Korteweg-de Vries vergelijking. Nadere studie hiervan begon ongeveer twintig jaar geleden; dit werk was het begin van de bovengenoemde trend. Deze K.d.V vergelijking was reeds sinds 1894 bekend. Hij was uit (4.1) en (4.2) afgeleid als een benadering voor lange golven die in hoofdzaak naar rechts lopen. Voor deze golven zijn dan u en η weinig verschillend. Invoeren van de variabelen $r = u + \eta$, $s = u - \eta$ en verwaarlozen van s leidt dan tot:

$$r_t + (r + \frac{3}{4}r^2 + \tfrac{1}{2}r_{xx})_x = 0. \tag{5.1}$$

De bewegingsconstanten waar het hier om gaat zijn polynomen in r en de afgeleiden hiervan. Voor elke positieve gehele k kan er een geconstrueerd worden die een term r^k bevat.[8]) Het is inmiddels gebleken dat de vergelijkingen (4.1) en (4.2) analoge eigenschappen hebben. De bewegingsconstanten beginnen hier met een term ηv^k.[9])

De fundering en verdere uitwerking hiervan kan onmogelijk in een stuk van deze omvang gegeven worden. In plaats daarvan zullen we een minder boeiend stel vergelijkingen behandelen waarbij het allemaal veel eenvoudiger verloopt. Deze vergelijkingen zijn

$$u_t + u_x = uv$$
5.2)
$$v_t - v_x = - uv .$$

en kan u en v opvatten als de dichtheden van twee populaties die zich met e snelheden +1 resp. -1 door elkaar heen bewegen waarbij het u-volk zich en koste van het v-volk uitbreidt.

et is niet direct duidelijk dat (5.2) een conservatief systeem is. We gaan aarom eerst op zoek naar een variatieprincipe. Optellen van de vergelijingen levert

$$(u+v)_t + (u-v)_x = 0.$$

us

$$u + v = 2\psi_x$$
5.3)
$$u - v = -2\psi_t$$

f

$$u = \psi_x - \psi_t$$
5.4)
$$v = \psi_x + \psi_t .$$

nvullen van (5.4) in (5.2) levert dan voor de potentiaal de vergelijking:

5.5) $$\psi_{tt} + \psi_t^2 = \psi_{xx} - \psi_x^2 .$$

eze vergelijking kan gelineariseerd worden door de transformatie:

5.6) $$\phi = \exp(-\psi), \quad \psi = - \ln \phi.$$

it levert:

$$\phi_t = - \psi_t . \exp(-\psi), \quad \phi_{tt} = (-\psi_{tt} + \psi_t^2) \exp(-\psi)$$

n dus

5.7) $$\phi_{tt} = \phi_{xx}$$

n dit is een lineaire vergelijking waarvan "alles" bekend is.

e algemene oplossing is:

$$\phi = r(x-t) + s(x+t)$$

et r en s willekeurige functies van één variabele.

et behulp van (5.4) en (5.6) vinden we dan

(5.8) $$u = \frac{-2r'}{r+s} \quad , \quad v = \frac{-2s'}{r+s} \quad .$$

Hiermee is in principe alles rond. We maken nog enkele opmerkingen die de structuur van deze transformatie van (5.5) naar (5.7) nader kunnen toelichten. Dit omdat in de meer ingewikkelde gevallen sommige bijverschijnselen gemakkelijker op te sporen zijn dan de lineariserende transformatie zelf. Het is duidelijk dat (5.7) volgt uit het variatieprincipe met Lagrangiaan

(5.9) $$L = \int dx \, [\tfrac{1}{2}\phi_t^2 - \tfrac{1}{2}\phi_x^2] .$$

Met behulp van (5.6) kan (5.9) getransformeerd worden in

(5.10) $$L = \int dx \, [\exp(-2\psi) . \, (\frac{\psi_t^2}{2} - \frac{\psi_x^2}{2})] .$$

Het is gemakkelijk na te gaan dat variatie van ψ in (5.10) juist (5.5) oplevert. Behoudswetten zijn ook direct door transformatie te vinden. Bijvoorbeeld:

$$\frac{d}{dt} \int dx \, (\frac{\phi_t^2 + \phi_x^2}{2}) = 0$$

als ϕ aan (5.7) voldoet. Transformatie geeft:

$$\frac{d}{dt} \int dx . \, \exp(-2\phi) . \, \frac{\phi_t^2 + \phi_x^2}{2} = 0 .$$

Met behulp van (5.3) is dit te schrijven als

(5.11) $$\frac{d}{dt} \int_{-\infty}^{\infty} dx \, \frac{u^2+v^2}{4} . \, \exp(- \int_{-\infty}^{x} (u+v)) = 0 .$$

De exponentiële factoren in (5.10) en (5.11) maken duidelijk waarom het niet direct te zien is dat (5.2) een conservatief systeem is. Andere bewegingsconstanten van (5.7) kunnen op analoge wijze worden ombebouwd. Tenslotte schrijven we de transformatie nog in een iets andere vorm. De bedoeling is een gelineariseerd analogon van (5.2) te krijgen zoals (5.7) dat van (5.5) is. Laat:

(5.12) $$f = \phi_t - \phi_x \quad (= -2r')$$
$$b = -\phi_t - \phi_x \quad (= -2s').$$

Uit (5.7) en (5.12) volgen dan de bewegingsvergelijkingen in de vorm

$$(5.13)\qquad \begin{aligned} f_t + f_x &= 0 \\ b_t - b_x &= 0. \end{aligned}$$

Deze vergelijkingen beschrijven hetzelfde systeem als (5.2). De relatie kan gevonden worden d.m.v. (5.3), (5.6) en (5.12). Er komt:

$$f = \phi_t - \phi_x = (\psi_x - \psi_t)\exp(-\psi) = u\exp(-\psi) = u\exp\left(-\int^x \frac{u+v}{2}\right),$$

$$b = v\exp\left(-\int^x \frac{u+v}{2}\right)$$

of

$$(5.14)\qquad \begin{cases} \dfrac{f_x}{f} = \dfrac{u_x}{u} - \dfrac{u+v}{2} \\[2ex] \dfrac{b_x}{b} = \dfrac{v_x}{v} - \dfrac{u+v}{2}\,. \end{cases}$$

Willekeurige functies van f, f_x,... zijn bewegingsconstanten van (5.13), evenzo functies van b, b_x,... . Er geldt

$$\frac{\partial}{\partial t}\left[F(f,f_x,\ldots) + B(b,b_x,\ldots)\right] + \frac{\partial}{\partial x}\left[F - B\right] = 0.$$

Met behulp van (5.14) kunnen nu bewegingsconstanten van (5.2) gevonden worden door geschikte keuze van F en B. Wanneer men ook exponentiële termen in de bewegingsconstanten toelaat dan kunnen F en B willekeurig gekozen worden.

NOTEN

1) Deze kwestie wordt uitvoerig besproken in "Canonical and non-canonical symmetries for Hamiltonian systems", H.M.M. ten Eikelder, Proefschrift THE, 1984. Het basis-resultaat is dat voor de H-structuur alleen vereist is dat er een variatieprincipe is dat linear is in de tijdsafgeleide. Dus:

$$\delta \int dt \int dx \left[u_t \cdot a\{u\} + h\{u\}\right] = 0$$

waarin $f\{u\} = f(u,u_x,u_{xx},\ldots)$.

2) Benaderde oplossingen van het type (2.11) voor de vergelijking (2.1)

lijken voor het eerst aangegeven te zijn in:
G. Green, Trans. Cambridge Phil. Soc. 6 (1837) p. 457, J. Liouville, Journal de Mathématiques 2 (1837) p. 16.
Blijkbaar heeft dit lange tijd niet veel aandacht getrokken. Daardoor is te begrijpen dat (2.11) vaak de W.K.B.-benadering wordt genoemd. In feite slaat dit op werk van de physici Wentzel, Kramers en Brillouin die, onafhankelijk van elkaar, in 1926 lieten zien hoe een monochromatische versie van (2.11) gebruikt kan worden om hogere eigenwaarden van quantum-mechanische problemen te benaderen.

3) H. Bremmer, "Handelingen Natuur- en Geneeskundig Congres, Nijmegen", Ruigrok, Haarlem (1939) p. 808, Physica 15 (1949) p. 593; Commun. pure en applied math. IV (1955) p. 105.
Het verband tussen de Bremmer- en de Liouvillebenadering is o.a. bestudeerd in L.J.F. Broer en J.B. van Vroonhoven, Physica 52 (1971) p. 441 en in Th. Verheggen, "Series Solutions of $\phi_{xx} - c^{-2}(x)\,\phi_{tt} = 0$ based on the W.K.B approximations", Proefschrift T.H.E. (1974).

4) Zie b.v. L.J.F. Broer, Radio Science 14 (1979) p. 245.

5) Klassiek zijn op dit gebied: T. Levi-Civita, Math. Ann.43 (1925) p. 264; D. Struik, Math. Ann. 45 (1926) p. 595.

6) L.J.F. Broer, Appl.Sci.Res. 29 (1974) p. 430; 31 (1975) p. 377; L.J.F. Broer, E.W.C. van Groesen & J.M.W. Timmers, Appl.Sci.Res. 32 (1976) p. 619.

7) De literatuur hierover is zeer uitgebreid en nogal onoverzichtelijk. Een zeer globale inleiding en enkele belangrijke publicaties zijn te vinden in: L.J.F. Broer, Ned. Tijdschr. Nat. A49 (1983) p. 43.

8) R.M. Miura c.s., J. Math. Phys. 9 (1968) p. 1204.

9) Dit volgt door toepassing van methoden te vinden bij ten Eikelder[1]) op een recent resultaat van Ito: M. Ito, Physics Letters 104A (1984) p. 248.

HOOFDSTUK 5

VARIATIEREKENING EN NUMERIEKE ANALYSE: DE EINDIGE ELEMENTEN METHODE

C. CUVELIER

INLEIDING

In de mathematische fysica worden stationaire problemen vaak geformuleerd in termen van een elliptische (partiële) differentiaalvergelijking. Deze differentiaalvergelijkingen zijn in het algemeen niet analytisch op te lossen. Om echter toch iets te kunnen zeggen over het kwantitatieve gedrag van de oplossing, zullen we zoeken naar een benadering u_N van de oplossing u van de diff. vgl.. Voor het zoeken van zo'n benadering zullen we gebruik maken van de variatierekening. In Hoofdstuk 2 hebben we gezien dat een (part.) diff. vgl. onder bepaalde voorwaarden geschreven kan worden als een minimaliseringsprobleem. De werkwijze voor het berekenen van een benadering u_N is nu als volgt. De (part.) diff. vgl. wordt (equivalent) beschreven als een variatieprobleem met dezelfde oplossing u. Van dit variatieprobleem kunnen we op systematische wijze de oplossing benaderen. De methode die we hier zullen toepassen is de methode van Ritz. Deze methode gaat uit van de volgende gedachte. De oplossing u van het variatieprobleem is gedefinieerd in een functieruimte met oneindig veel vrijheidsgraden, of anders gezegd, u is een lineaire combinatie van oneindig veel basisfuncties van de betreffende functieruimte. De benadering u_N wordt nu gedefinieerd door rekening te houden met slechts een eindig aantal, zeg N, basisfuncties. Voor de keuze en de definitie van de basisfuncties gebruiken we de Eindige Elementen Methode (EEM). Een voor de hand liggende eis die we moeten stellen is natuurlijk dat bij toename van het eindig aantal basisfuncties de daarbij behorende benadering u_N een nauwkeurige benadering wordt van de oplossing u.
De geschetste werkwijze zal worden toegepast op een warmtegeleidingsprobleem in een 1-dimensionale staaf met twee typen randvoorwaarden. Tevens zal een uitbreiding naar twee dimensies worden gegeven. Met een aantal rekenvoorbeelden zal het geheel worden geïllustreerd, waarbij we aandacht zullen schenken aan een schatting van de fout $|u - u_N|$ als functie van het aantal basis functies N.

De equivalente formulering van een (part.) diff. vgl. als minimaliseringsprobleem is slechts mogelijk voor zelfgeadjungeerde differentiaaloperatoren

(zie Hoofdstuk 2). Voor niet-zelfgeadjungeerde differentiaaloperatoren kunnen we de methode van Ritz niet toepassen. Voor deze klasse van problemen kunnen we toch een benadering u_N definiëren, waarbij we gebruik maken van de zwakke (of variationele) formulering van de diff. vgl.. Met behulp van de EEM kunnen dan weer basisfuncties gedefinieerd worden. Het definiëren van een benadering u_N met behulp van een zwakke formulering staat bekend als de methode van Galerkin. Deze methode is ook toepasbaar op zelfgeadjungeerde problemen en geeft, voor dat geval, dezelfde benadering u_N als de methode van Ritz. We kunnen dan ook stellen dat de methode van Galerkin een generalisatie is van de methode van Ritz. Ritz is toepasbaar in het zelf-geadjungeerde geval, Galerkin is ook toepasbaar in het niet-zelfgeadjungeerde geval.

1. DE WARMTEGELEIDINGSVERGELIJKING

In deze paragraaf beschouwen we de vgl. die de warmtegeleiding beschrijft in een n-dimensionaal lichaam Ω. Voor n = 1 denken we daarbij aan een staaf $\Omega = (0,1)$. Het geval n = 2 correspondeert met een "oneindig dunne" plaat, en voor n = 3 kunnen we de warmtegeleiding in een aardappelvormig gebied in gedachte hebben (zie fig. 1.a,b,c).

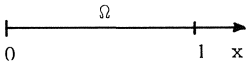


Fig. 1.a $\Omega \subset \mathbb{R}^1$

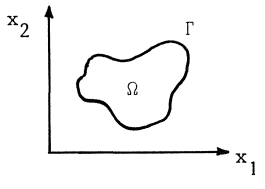


Fig. 1.b $\Omega \subset \mathbb{R}^2$

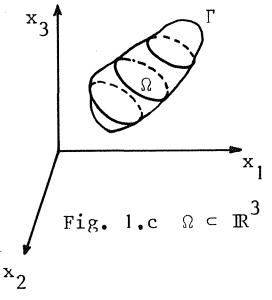


Fig. 1.c $\Omega \subset \mathbb{R}^3$

In een isotroop medium wordt de warmtegeleiding beschreven door de wet van Fourier, die zegt dat de warmtestroom $\underline{q}$ per oppervlakte-eenheid rechtevenredig is met de temperatuurgradiënt, maar tegengesteld gericht. In formule wordt dit

$$\underline{q} = - k \text{ grad } u \qquad (1.1)$$

waarbij $k > 0$ de (niet noodzakelijk konstante) thermische conductiviteit van het lichaam is (een materiaalkonstante) en waarbij u de temperatuur voorstelt. In het 1-dimensionale geval geldt dan

$$\underline{q} = - k \frac{du}{dx} . \tag{1.2}$$

Voor $n = 2$ kan (1.1) geschreven worden als

$$\underline{q} = \begin{pmatrix} q_{x_1} \\ q_{x_2} \end{pmatrix} = - k \begin{pmatrix} \frac{\partial u}{\partial x_1} \\ \frac{\partial u}{\partial x_2} \end{pmatrix}. \tag{1.3}$$

Veronderstellen we het lichaam Ω in rust, dan geldt de volgende wet van behoud van energie

$$\operatorname{div} \underline{q} = f \tag{1.4}$$

waarbij f de warmteproductie per volume eenheid voorstelt.
Combinatie van de formules (1.1) en (1.4) geeft de volgende warmtegeleidingsvergelijking voor u:

$$- \operatorname{div} (k \operatorname{grad} u) = f \quad \text{in } \Omega. \tag{1.5}$$

In het algemeen zal de thermische conductiviteit van de plaats afhangen: $k = k(\underline{x})$, $\underline{x} = \{x_1,\ldots,x_n\}$. Als k echter konstant is in het hele lichaam Ω dan reduceert (1.5) tot

$$- k \operatorname{div} \operatorname{grad} u = f \quad \text{in } \Omega \tag{1.6}$$

oftewel

$$- k \Delta_n u = f \quad \text{in } \Omega \tag{1.7}$$

waarbij Δ_n de n-dimensionale Laplace operator voorstelt:

$$\Delta_n = \frac{\partial^2}{\partial x_1^2} + \ldots + \frac{\partial^2}{\partial x_n^2} \; ; \quad \Delta_1 = \frac{d^2}{dx^2} \; ; \quad \Delta_2 = \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2} . \tag{1.8}$$

Om de oplossing u van (1.5) eenduidig vast te leggen zijn randvoorwaarden nodig. Dit zijn voorwaarden waaraan u moet voldoen op de rand Γ van het ge-

bied Ω. In het 1-dimensionale geval bestaat Γ uit de punten $x = 0$ en $x = 1$ (zie Fig. 1a). Voor $n = 2$ en 3 is Γ aangegeven in de Figuren 1b en c. Twee mogelijke randvoorwaarden zijn:

(1.9) Dirichlet-randvoorwaarden. De temperatuur is voorgeschreven op Γ:

$$u(\underline{x}) = g_0(\underline{x}) , \quad \underline{x} \in \Gamma.$$

(1.10) Neumann-randvoorwaarden. De warmtestroom in de richting van de uitwendige eenheidsnormaal ν op Γ is voorgeschreven

$$\underline{q}.\underline{\nu}(\underline{x}) \equiv (-k \text{ grad } u \,.\, \underline{\nu})(\underline{x}) \equiv -k \frac{\partial u}{\partial \nu}(\underline{x}) = g_1(\underline{x}), \quad \underline{x} \in \Gamma.$$

In ieder punt van de rand Γ moet één van de twee randvoorwaarden zijn gegeven.

(1.11) VOORBEELD

$n = 1$, $\Omega = (0,1)$, $k = 1$, $f = -e^x$,

$$\begin{cases} -\dfrac{d^2u}{dx^2} = -e^x \quad \text{op } (0,1) \\ u(0) = 1 \quad \text{(temperatuur in } x = 0 \text{ is 1)} \\ -\dfrac{du}{dx}(1) = -e \quad \text{(warmtestroom in } x = 1 \text{ is } -e). \end{cases}$$

Dit probleem kan analytisch opgelost worden en heeft als oplossing (zie Fig. 2)

$$u(x) = e^x.$$

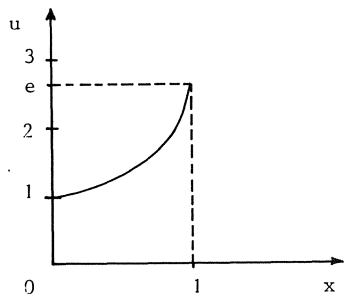


Fig. 2. $u(x) = e^x$

Het tweede voorbeeld dat we zullen bekijken is de warmtegeleiding in een 1-dimensionale staaf met niet konstante conductiviteit k. Beschouw een staaf [0,1] waarbij

$$(1.12) \qquad k(x) = \begin{cases} k_1 & \text{als} \quad x \in [0, \tfrac{1}{2}) \\ k_2 & \text{als} \quad x \in (\tfrac{1}{2}, 1] \end{cases}$$

met k_1 en k_2 positieve konstanten. Op $(0,\frac{1}{2})$ geldt de vgl. (1.7):

$$(1.13) \qquad -k_1 \frac{d^2u}{dx^2} = f.$$

Op $(\frac{1}{2},1)$ geldt

$$(1.14) \qquad -k_2 \frac{d^2u}{dx^2} = f.$$

Als randvoorwaarden in x = 0 en x = 1 kiezen we

$$(1.15) \qquad u(0) = g_0 , \qquad -\frac{du}{dx}(1) = g_1 .$$

Het probleem (1.13), (1.14), (1.15) is nog niet eenduidig op te lossen. Uit de thermodynamica van het probleem volgt dat we de volgende twee eisen moeten opleggen. Ten eerste verloopt de temperatuur continu in de staaf, d.w.z. u is continu in $x = \frac{1}{2}$. Ten tweede geldt de continuïteit van de warmtestroom in $x = \frac{1}{2}$, hetgeen geformuleerd kan worden als:

$$(1.16) \qquad -k_1 \frac{du}{dx}\Big|_{x=\frac{1}{2}} = -k_2 \frac{du}{dx}\Big|_{x=\frac{1}{2}+} .$$

(1.17) VOORBEELD

$n = 1$, $\Omega = (0,1)$, $k_1 = 2$, $k_2 = 1$, $f = -\frac{2}{(\frac{1}{2}+x)^2}$ op $(0,\frac{1}{2})$, $f = 0$ op $(\frac{1}{2},1)$.

$$-2\frac{d^2u}{dx^2} = -\frac{2}{(\frac{1}{2}+x)^2} \qquad \text{op } (0,\tfrac{1}{2})$$

$$-\frac{d^2u}{dx^2} = 0 \qquad \text{op } (\tfrac{1}{2},1)$$

$$u(0) = 1 + \ell n\, 2$$

$$-\frac{du}{dx}(1) = 2$$

u continu in $x = \frac{1}{2}$

$$-2 \frac{du}{dx}\Big|_{x = \frac{1}{2}-} = - \frac{du}{dx}\Big|_{x = \frac{1}{2}+} .$$

Dit probleem heeft de volgende oplossing:

$$u(x) = \begin{cases} 1 - \ell n(\frac{1}{2} + x) & \text{als } x \in [0, \frac{1}{2}] \\ 2 - 2x & \text{als } x \in [\frac{1}{2}, 1] \end{cases}$$

die geschetst is in Fig. 3.

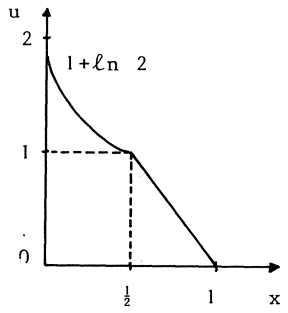


Fig. 3. $u(x)$

Het derde voorbeeld is een 2-dimensionaal warmtegeleidingsprobleem.

(1.18) VOORBEELD

$n = 2$, $\Omega = (0,1) \times (0,1)$, $k = 1$, $f = 8x_1$,

$$\begin{cases} -\left(\frac{\partial^2 u}{dx_1^2} + \frac{\partial^2 u}{\partial x_2^2}\right) = 8x_1 & \text{in } \Omega \\ u = 0 \quad \text{op } \Gamma_0 \\ \frac{\partial u}{\partial \nu}(\underline{x}) = \frac{\partial u}{\partial x_1}(\underline{x}) = 4x_2(1-x_2) & \text{voor } \underline{x} = \{x_1, x_2\} \in \Gamma_1 . \end{cases}$$

De situatie is geschetst in Fig. 4

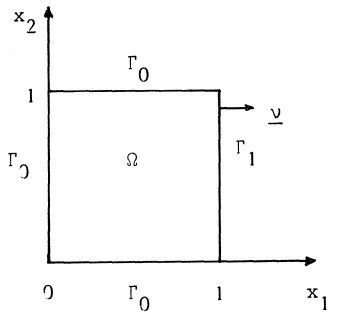


Fig 4. Ω met rand $\Gamma_0 \cup \Gamma_1$

De analytische oplossing van dit probleem is

$$u(x_1,x_2) = 4x_1x_2(1-x_2).$$

Voor deze drie voorbeelden was het mogelijk een analytische oplossing te geven. In het algemeen is dat niet mogelijk en moet een benadering van de oplossing u berekend worden.

2. HET EQUIVALENTE MINIMALISERINGSPROBLEEM

In Hoofdstuk 2 is aangetoond dat de (part.) diff. vgln van §1 equivalent geformuleerd kunnen worden als minimaliseringsprobleem. Bij het volgende probleem

$$(2.1) \qquad \begin{cases} -\operatorname{div}(k \operatorname{grad} u) = f \quad \text{in } \Omega, \quad k = k(\underline{x}) \text{ continu} \\ u = g_0 \quad \text{op } \Gamma_0, \quad g_0 = g_0(\underline{x}) \\ -k \dfrac{\partial u}{\partial \nu} = g_1 \text{ op } \Gamma_1, \quad g_1 = g_1(\underline{x}) \end{cases}$$

behoort het volgende minimaliseringsprobleem:

$$(2.2) \qquad \begin{cases} \text{Zoek een functie } u : \overline{\Omega} \to \mathbb{R} \text{ met } u|_{\Gamma_0} = g_0 \text{ zodanig dat} \\ J(u) = \inf_{\substack{v:\overline{\Omega}\to\mathbb{R} \\ v|_{\Gamma_0} = g_0}} J(v) \end{cases}$$

waarbij

$$(2.3) \qquad J(v) = \tfrac{1}{2} \int_\Omega k\,|\operatorname{grad} v|^2 \, d\Omega - \int_\Omega f v \, d\Omega + \int_{\Gamma_1} g_1 v \, d\Gamma.$$

Passen we dit toe op Voorbeeld (1.11) dan vinden we

$$
(2.4)\qquad \left\{\begin{array}{l} \text{Zoek } u : [0,1] \to \mathbb{R} \text{ met } u(0) = 1 \text{ zodanig dat} \\ J(u) = \inf_{v:[0,1]\to\mathbb{R},\ v(0)=1} J(v) \\ \text{met} \\ J(v) = \frac{1}{2}\int_0^1 \left|\frac{dv}{dx}\right|^2 dx + \int_0^1 e^x v\,dx - e\,v(1). \end{array}\right.
$$

Voor Voorbeeld (1.18) vinden we

$$
(2.5)\qquad \left\{\begin{array}{l} \text{Zoek } u : \bar{\Omega} \to \mathbb{R} \text{ met } u = 0 \text{ op } \Gamma_0 \text{ zodanig dat} \\ J(u) = \inf_{v:\bar{\Omega}\to\mathbb{R},\ w=0 \text{ op } \Gamma_0} J(v) \\ \text{met} \\ J(v) = \frac{1}{2}\int_\Omega \left\{\left(\frac{\partial v}{\partial x_1}\right)^2 + \left(\frac{\partial v}{\partial x_2}\right)^2\right\} d\Omega - 8\int_\Omega x_1\,v\,d\Omega - 4\int_{\Gamma_1} x_2(1-x_2)\,v\,d\Gamma. \end{array}\right.
$$

In Voorbeeld (1.17) is de coëfficient k discontinu, zodat de equivalentie met (2.2) niet triviaal is. Er geldt echter dat (1.17) equivalent is met

$$
(2.6)\qquad \left\{\begin{array}{l} \text{Zoek } u : [0,1] \to \mathbb{R} \text{ met } u(0) = 1 + \ell n\ 2 \text{ zodanig dat} \\ J(u) = \inf_{v:[0,1]\to\mathbb{R},\ v(0)=1+\ell n\ 2} J(v) \\ \text{met} \\ J(v) = \frac{1}{2}\int_0^1 k\left|\frac{dv}{dx}\right|^2 dx + 2\int_0^{\frac{1}{2}} \frac{v}{(\frac{1}{2}+x)^2}\,dx + 2\,v(1) \\ \text{waarbij} \\ k(x) = \begin{cases} 2 & \text{als } x \in (0,\frac{1}{2}) \\ 1 & \text{als } x \in (\frac{1}{2},1). \end{cases} \end{array}\right.
$$

We zullen bewijzen dat de oplossing u van (2.6) ook voldoet aan het probleem van Voorbeeld (1.17). Zij u een oplossing van (2.6), dan moet aan de volgende noodzakelijke voorwaarde voldaan zijn (zie Hoofdstuk 2):

$$\frac{d}{d\lambda} J(u+\lambda w)\Big|_{\lambda = 0} = 0 \text{ voor alle } w : [0,1] \to \mathbb{R} \text{ met } w(0) = 0.$$

Dit impliceert dat

$$(2.7) \qquad \int_0^1 k \frac{du}{dx} \frac{dw}{dx}\, dx + 2 \int_0^{\frac{1}{2}} \frac{w\, dx}{(\frac{1}{2}+x)^2} + 2\, w(1) = 0.$$

Relatie (2.7) geldt voor alle $w: [0,1] \to \mathbb{R}$ met $w(0) = 0$, dus zeker voor alle $w : [0,1] \to \mathbb{R}$ met $w = 0$ op $[\frac{1}{2},1]$ en $w(0) = 0$, zodat

$$0 = \int_0^{\frac{1}{2}} k \frac{du}{dx} \frac{dw}{dx}\, dx + 2 \int_0^{\frac{1}{2}} \frac{w\, dx}{(\frac{1}{2}+x)^2} = \int_0^{\frac{1}{2}} 2 \frac{du}{dx} \frac{dw}{dx}\, dx + 2 \int_0^{\frac{1}{2}} \frac{w\, dx}{(\frac{1}{2}+x)^2} =$$

$$= \int_0^{\frac{1}{2}} \left\{ -2 \frac{d^2u}{dx^2} + \frac{2}{(\frac{1}{2}+x)^2} \right\} w\, dx$$

waaruit volgt

$$(2.8) \qquad -2 \frac{d^2u}{dx^2} = -\frac{2}{(\frac{1}{2}+x)^2} \qquad \text{op } (0,\tfrac{1}{2}).$$

Uitdrukking (2.7) geldt voor alle $w : [0,1] \to \mathbb{R}$ met $w(0) = 0$, dus zeker voor alle $w : [0,1] \to \mathbb{R}$ met $w = 0$ op $[0,\frac{1}{2}]$ en $w(1) = 0$. Dit geeft

$$0 = \int_{\frac{1}{2}}^1 k \frac{du}{dx} \frac{dw}{dx}\, dx = \int_{\frac{1}{2}}^1 \frac{du}{dx} \frac{dw}{dx}\, dx = \int_{\frac{1}{2}}^1 \left\{ -\frac{d^2u}{dx^2} \right\} w\, dx$$

waaruit volgt

$$(2.9) \qquad -\frac{d^2u}{dx^2} = 0 \qquad \text{op } (\tfrac{1}{2},1).$$

Relatie (2.7) geldt voor alle $w: [0,1] \to \mathbb{R}$ met $w(0) = 0$, dus zeker voor alle $w : [0,1] \to \mathbb{R}$ met $w = 0$ op $[0,\frac{1}{2}]$, zodat

$$0 = \int_{\frac{1}{2}}^{1} k \frac{du}{dx} \frac{dw}{dx} dx + 2 w(1) =$$

$$= \int_{\frac{1}{2}}^{1} \frac{du}{dx} \frac{dw}{dx} dx + 2 w(1) =$$

$$= \int_{\frac{1}{2}}^{1} \left\{ - \frac{d^2u}{dx^2} \right\} w \, dx + \frac{du}{dx} w \Big|_{\frac{1}{2}}^{1} + 2 w(1) =$$

$$= \left\{ \frac{du}{dx} (1) + 2 \right\} w(1)$$

waaruit volgt dat

$$(2.10) \qquad - \frac{du}{dx} (1) = 2.$$

Zij w vervolgens een willekeurige functie op [0,1] met w(0) = 0. We vermenigvuldigen (2.8) met w, integreren over $(0,\frac{1}{2})$ en passen partiële integratie toe. Dit resulteert in

$$(2.11) \qquad \int_{0}^{\frac{1}{2}} 2 \frac{du}{dx} \frac{dw}{dx} dx - 2 \frac{du}{dx} (\tfrac{1}{2}-) \cdot w(\tfrac{1}{2}-) + \int_{0}^{\frac{1}{2}} \frac{2w \, dx}{(\frac{1}{2}+x)^2} = 0.$$

Vervolgens vermenigvuldigen we (2.9) met w, integreren over $(\frac{1}{2},1)$, passen partiële integratie toe en substitueren (2.10). Dit geeft

$$(2.12) \qquad \int_{\frac{1}{2}}^{1} \frac{du}{dx} \frac{dw}{dx} dx + 2 w(1) + \frac{du}{dx} (\tfrac{1}{2}+) \cdot w(\tfrac{1}{2}+) = 0.$$

Tellen we (2.11) en (2.12) bij elkaar op, dan geldt:

$$(2.13) \qquad \int_{0}^{1} k \frac{du}{dx} \frac{dw}{dx} dx + 2 \int_{0}^{\frac{1}{2}} \frac{w \, dx}{(\frac{1}{2}+x)^2} + 2 w(1) + \left\{ \frac{du}{dx} (\tfrac{1}{2}+) - 2 \frac{du}{dx} (\tfrac{1}{2}-) \right\} w(\tfrac{1}{2}) = 0.$$

Vergelijken we (2.13) met (2.7) dan volgt, daar $w(\frac{1}{2})$ willekeurig is, dat

(2.14) $$-2\,\frac{du}{dx}\,(\tfrac{1}{2}-) = -\,\frac{du}{dx}\,(\tfrac{1}{2}+)$$

waarmee is "bewezen" dat een oplossing u van (2.6) voldoet aan probleem (1.17).

De minimaliserings-formulering van het probleem heeft, vanuit numeriek analytisch oogpunt bekeken, voordelen ten opzichte van de formulering in termen van de (part.) diff. vgl.. Ten eerste bevat de functionaal J afgeleiden van lagere orde dan de diff. vgl.. De oplossing kan dan ook in een grotere klasse van functies gezocht worden. Een tweede voordeel is, zoals blijkt uit de voorbeelden, dat bij de variatieformulering niet met alle randvoorwaarden rekening gehouden behoeft te worden. Slechts de Dirichlet-randvoorwaarden moeten opgelegd worden aan de oplossing; aan de Neumann-randvoorwaarden is automatisch voldaan bij juiste keuze van de functionaal. De Dirichlet-randvoorwaarden worden wel essentiële en de Neumann-randvoorwaarden natuurlijke randvoorwaarden genoemd. Vanuit de numerieke analyse bezien is dat een groot voordeel van de variatie-formulering, daar juist de Neumann-randvoorwaarden, bij discretisatie, onnauwkeurigheden veroorzaken.

3. DE METHODE VAN RITZ

In deze paragraaf zullen we een methode bespreken voor het minimaliseringsprobleem

(3.1) $$\text{Zoek } u \in V \quad \text{zodanig dat } J(u) = \inf_{v \in V} J(v)$$

waarbij V een klasse van functies is (zie de voorbeelden), die aan de (essentiële) Dirichlet-randvoorwaarden voldoen. Zij u_0 een willekeurig element uit V; dan kan ieder element $v \in V$ geschreven worden als

(3.2) $$v = u_0 + w \qquad \text{voor zekere } w \in V_0$$

waarbij V_0 de klasse van functies is, die aan de homogene Dirichlet-randvoorwaarden voldoen.

We veronderstellen dat de functieruimte V_0 een aftelbare basis $\{\phi_j\}_{j=1,2,\ldots}$ heeft. Dat wil zeggen dat ieder element $w \in V_0$, in zekere zin, geschreven kan worden als een oneindige lineaire combinatie van de basisfuncties ϕ_j :

$$w = \sum_{j=1}^{\infty} \alpha_j \phi_j \tag{3.3}$$

waarbij α_j de coëfficienten zijn. Voor de oplossing u geldt dan

$$u = u_0 + \sum_{j=1}^{\infty} \alpha_j \phi_j \tag{3.4}$$

met nog te bepalen coëfficienten α_j.

Een benadering u_N van u zullen we nu zoeken in de klasse V_N van functies die te schrijven zijn als de som van u_0 en een lineaire combinatie van de eerste N basisfuncties ϕ_j. Een element v_N uit V_N is dan te schrijven als

$$v_N = u_0 + \sum_{j=1}^{N} \alpha_j \phi_j \tag{3.5}$$

en de benadering u_N van u wordt gedefinieerd als

$$\begin{aligned} & u_N \in V_n \\ & J(u_N) = \inf_{v_N \in V_n} J(v_N) \ . \end{aligned} \tag{3.6}$$

Daar $J(u_N)$ slechts afhangt van de coëfficienten $\alpha_1,\ldots,\alpha_N$ kan het minimaliseringsprobeem (3.6) ook geschreven worden als

$$\left\{ \begin{array}{l} \text{Zoek coëfficienten } \alpha_1,\ldots,\alpha_N \text{ zodanig dat} \\ J(\alpha_1,\ldots,\alpha_N) \text{ zijn minimum waarde aanneemt.} \end{array} \right. \tag{3.7}$$

Een nodige voorwaarde voor het minimaal zijn is dat

$$\frac{\partial J}{\partial \alpha_i} = 0 , \qquad i = 1,2,\ldots,N. \tag{3.8}$$

Het minimaliseringsprobleem (3.1) is nu gereduceerd tot een stelsel van N vgln. met N onbekenden $\alpha_1,\ldots,\alpha_N$.

We zullen deze werkwijze nu gaan toepassen op het probleem van Voorbeeld

(1.11). De functie $u_0 = 1$ voldoet aan de Dirichlet-voorwaarde in $x = 0$. Het is een klassiek resultaat dat in de functie-ruimte, bestaande uit continue functies op $[0,1]$ die gelijk nul zijn in $x = 0$, de volgende basis gekozen kan worden

$$\phi_j(x) = x^j , \quad j = 1,2,\ldots . \tag{3.9}$$

De oplossing u heeft dan de vorm

$$u = 1 + \sum_{j=1}^{\infty} \alpha_j x^j . \tag{3.10}$$

Een benadering u_N van u wordt nu geschreven als

$$u_N = 1 + \sum_{j=1}^{N} \alpha_j x^j \tag{3.11}$$

waarbij de coëfficienten $\alpha_1,\ldots,\alpha_N$ zodanig zijn dat de functionaal

$$J(u_N) = J\left(1 + \sum_{j=1}^{N} \alpha_j x^j\right) = J(\alpha_1,\ldots,\alpha_N) \tag{3.12}$$

minimaal is. Een nodige voorwaarde is dat

$$\frac{\partial J(\alpha_1,\ldots,\alpha_N)}{\partial \alpha_i} = 0 , \qquad i = 1,2,\ldots,N. \tag{3.13}$$

In geval van Voorbeeld (1.11) is $J(\alpha_1,\ldots,\alpha_N)$ gegeven door (zie (2.4)):

$$J(\alpha_1,\ldots,\alpha_N) = \tfrac{1}{2}\int_0^1 \left|\frac{d}{dx}\left(1+\sum_{j=1}^{N}\alpha_j x^j\right)\right|^2 dx + \int_0^1 e^x\left(1+\sum_{j=1}^{N}\alpha_j x^j\right)dx - e\left(1+\sum_{j=1}^{N}\alpha_j\right) = \tag{3.14}$$

$$= \tfrac{1}{2}\int_0^1 \left|\sum_{j=1}^{N} v_j \alpha_j x^{j-1}\right|^2 dx + \int_0^1 e^x\left(1+\sum_{j=1}^{N}\alpha_j x^j\right)dx - e\left(1+\sum_{j=1}^{N}\alpha_j\right).$$

De i-de vgl. (3.13) wordt nu

$$\int_0^1 \left(\sum_{j=1}^{N} j\alpha_j x^{j-1}\right)(ix^{i-1})dx + \int_0^1 e^x x^i dx - e = 0$$

oftewel

$$\sum_{j=1}^{N} ij\left(\int_0^1 x^{i+j-2}dx\right)\alpha_j = e - \int_0^1 e^x x^i dx$$

oftewel

$$\sum_{j=1}^{N} \frac{ij}{i+j-1}\,\alpha_j = e - \int_0^1 e^x x^i dx\,, \qquad i = 1,2,\ldots,N\,. \tag{3.15}$$

In matrix-notatie wordt (3.15)

$$\underline{A}\ \underline{\alpha} = \underline{F} \tag{3.16}$$

met

$$\underline{A} = \begin{bmatrix} 1 & 1 & 1 & 1 & \cdots & 1 \\ 1 & 4/3 & 6/4 & 8/5 & \cdots & \frac{2N}{N+1} \\ 1 & 6/4 & 9/5 & 13/6 & \cdots & \frac{3N}{N+2} \\ 1 & 8/5 & 12/6 & 16/7 & \cdots & \frac{4N}{N+3} \\ \cdot & \cdot & \cdot & \cdot & & \\ \cdot & \cdot & \cdot & \cdot & & \\ \cdot & \cdot & \cdot & \cdot & & \\ 1 & \cdot & \cdot & \cdot & \cdots & \frac{N^2}{2N-1} \end{bmatrix}, \underline{\alpha} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \alpha_N \end{bmatrix}, \underline{F} = \begin{bmatrix} -\int_0^1 e^x x\,dx \\ \int_0^1 e^x x^2\,dx \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ -\int_0^1 e^x x^N\,dx \end{bmatrix}.$$

De matrix $\underline{A}$, die bekend staat als de Hilbert-matrix, is positief-definiet, zodat het stelsel lineaire vgln. (3.16) een éénduidig bepaalde oplossing $\underline{\alpha}$ heeft. Vanuit praktisch oogpunt bezien, dient opgemerkt te worden dat de matrix $\underline{A}$ vol is, hetgeen wil zeggen dat er weinig (of geen) nullen in voorkomen. Bovendien heeft deze Hilbert-matrix een slechte conditie en dit betekent dat bij toenemende N de afrondfouten een rol gaan spelen en de nauwkeurigheid van de oplossing verstoren.

In het algemeen kunnen we stellen dat de methode van Ritz bestaat uit het schrijven van de benadering u_N in termen van een eindig aantal basisfuncties ϕ_j, $j = 1,\ldots,N$, met onbekende parameters α_j, $j = 1,\ldots,N$. De benadering wordt vervolgens gesubstitueerd in de functionaal. Een systeem van lineaire vgln voor de α_j wordt vervolgens afgeleid door de functionaal te differentiëren naar iedere α_i en de afgeleide gelijk nul te stellen. Dit leidt dan tot een stelsel van de vorm

$$\underline{A}\ \underline{\alpha} = \underline{F} \tag{3.17}$$

$$\underline{A} \text{ een } N \times N \text{-matrix}, \quad A_{ij} = \int_0^1 k \frac{d\phi_j}{dx} \frac{d\phi_i}{dx} dx$$

$$\underline{\alpha} \text{ een } N \text{-vector}, \quad (\underline{\alpha})_j = \alpha_j$$

$$\underline{F} \text{ een } N \text{-vector}, \quad (\underline{F})_j = F_j = \int_0^1 f \phi_j dx .$$

In de nu volgende paragraaf zullen we zien hoe door een speciale keuze van de basisfuncties de matrix $\underline{A}$ ijl wordt, d.w.z. veel nullen bevat. Dit zal dan de conditie van de matrix $\underline{A}$, de snelheid en de nauwkeurigheid waarmee het stelsel (3.17) opgelost wordt gunstig beïnvloeden. Een systematische werkwijze voor de constructie van dit soort speciale basisfuncties is de Eindige Elementen Methode.

4. DE EINDIGE ELEMENTEN METHODE

De EEM is een systematische methode voor het genereren van basisfuncties, die een ijle matrix $\underline{A}$ tot gevolg heeft. Uit hetgeen in de vorige paragraaf is besproken volgt dat de matrix $\underline{A}$ ijl is als de basisfuncties een "kleine" drager hebben. Onder de drager van een functie verstaan we de afsluiting van de deelverzameling van Ω waarop de functie ongelijk is aan nul. Het is duidelijk dat het matrix-element A_{ij} gelijk nul is als de dragers van ϕ_i en ϕ_j een lege doorsnede hebben.
Het basis-idee van de EEM is om het gebied $\Omega \subset \mathbb{R}^n$ op te delen in deelgebieden. In het 1-dimensionale geval wordt het interval $[0,1]$ opgedeeld in subintervallen. In 2 dimensies kunnen we het gebied Ω opdelen in driehoekjes (triangulatie). Aan deze deelgebieden stellen we de volgende eisen:

(4.1)

(i) Er is een eindig aantal deelgebieden e_k, $k = 1,\ldots,K$

(ii) Als e_k en e_ℓ twee deelgebieden zijn, dan geldt ôf $e_k = e_\ell$ ôf $e_k \wedge e_\ell = \emptyset$ ôf

(ii)' e_k en e_ℓ hebben een punt gemeen voor $n = 1$

(ii)" e_k en e_ℓ hebben een punt of een zijde gemeen voor $n = 2$

(iii) De vereniging van alle deelgebieden is precies $\overline{\Omega} = \Omega \cup \Gamma$.

In Fig. 5 is een 1-dimensionaal voorbeeld gegeven voor een opdeling van $\Omega = (0,1)$ in deelgebieden .

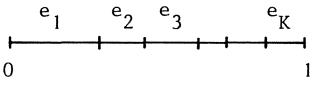


Fig. 5. $\Omega = (0,1)$, n = 1

Fig. 6a geeft een 2-dimensionaal voorbeeld. Het voorbeeld van Fig. 6b voldoet niet aan eis (4.1) (ii)''.

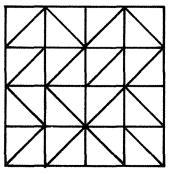


Fig. 6a. $\Omega = (0,1)^2$, n = 2

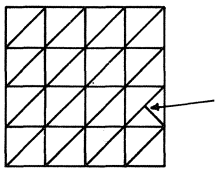


Fig. 6b. Triangulatie voldoet niet aan eis (4.1)(ii)''

Vervolgens kiezen we in ieder deelgebied een eindig aantal punten. Deze punten worden knooppunten genoemd. Enige voorbeelden zijn gegeven in Fig. 7 (n = 1) en Fig. 8 (n = 2).

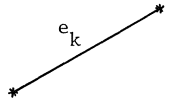


Fig. 7a. 2 knooppunten

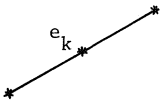


Fig. 7b. 3 knooppunten

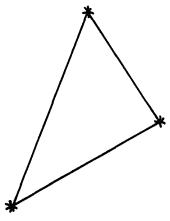


Fig. 8a. 3 knooppunten

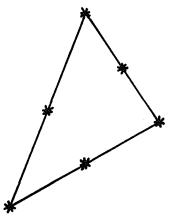


Fig. 8b. 6 knooppunten

De knooppunten worden aangegeven door $\underline{x}^j$ met coördinaten

$$x^j (n=1) \quad \text{en} \quad \{x_1^j, x_2^j\} \ (n=2).$$

Ten slotte definiëren we een basisfunctie $\phi_i : \overline{\Omega} \to \mathbb{R}$ voor ieder knooppunt $\underline{x}^i$ in $\overline{\Omega}$. Deze functie ϕ_i moet voldoen aan de volgende eisen:

(i) ϕ_i heeft de waarde 1 in $\underline{x}^i$ en is gelijk nul in alle andere knooppunten:

$$\phi_i(\underline{x}^j) = \delta_{ij} = \begin{cases} 1 \text{ als } i = j \\ 0 \text{ als } i \neq j \end{cases}$$

(ii) ϕ_i heeft een voorgeschreven gedrag op ieder deelgebied. Bijvoorbeeld affien of kwadratisch, afhankelijk van het aantal knooppunten op ieder deelgebied.

(iii) ϕ_i is continu op $\overline{\Omega}$.

We zullen nu enige voorbeelden geven van basisfuncties ϕ_i.

(4.2) VOORBEELD

$n = 1$, $\Omega = (0,1)$, 2 knooppunten per deelgebied, affiene basisfuncties op subinterval. De knooppunten zijn de randpunten van ieder subinterval: $\underline{x}^0, \underline{x}^1, \ldots, \underline{x}^N$

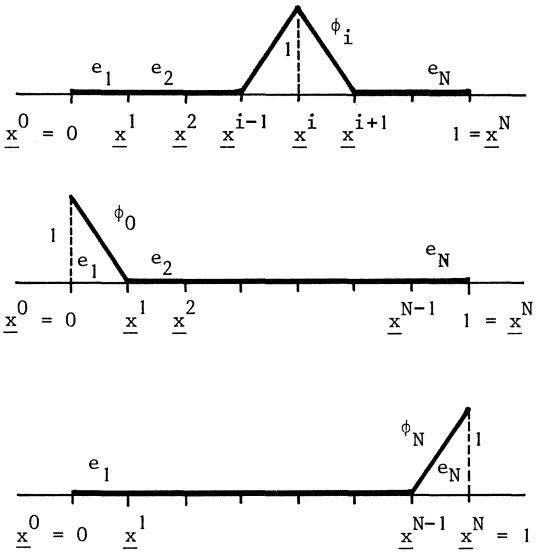


Fig. 9. Basisfuncties ϕ_i, ϕ_0 en ϕ_N behorend bij $\underline{x}^i$, $\underline{x}^0$ en $\underline{x}^N$

(4.3) VOORBEELD

n = 1 , Ω = (0,1) , 3 knooppunten per deelgebied, kwadratische basisfunctie op subinterval. De knooppunten zijn de randpunten $\underline{x}^{i-1}$, $\underline{x}^{i}$ en het midden $\underline{x}^{i-1i}$ van ieder subinterval e_i.

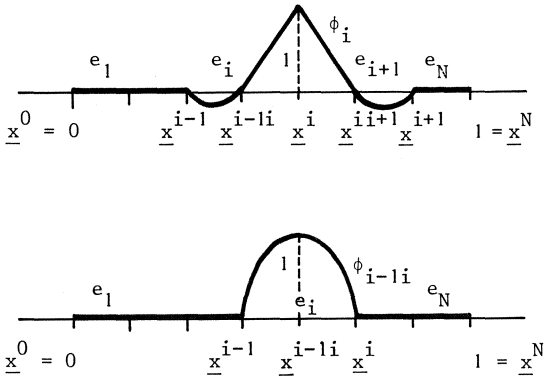


Fig. 10. Basisfuncties ϕ_i en ϕ_{i-1i} behorend bij $\underline{x}^i$ en $\underline{x}^{i-1i}$

(4.4) VOORBEELD

n = 2 , 3 knooppunten per driehoek (de hoekpunten), basisfuncties affien per driehoek.

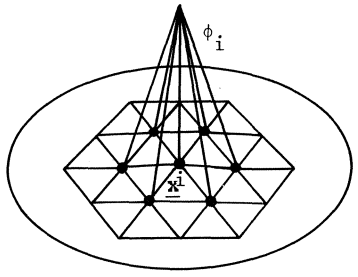


Fig. 11. Basisfunctie ϕ_i behorend bij $\underline{x}^i$

Met behulp van deze basisfuncties ϕ_j kan een benadering u_N geconstrueerd worden. De functie

$$u_N(x) = \sum_{j=1}^{N} u_{Nj}\, \phi_j(\underline{x}) \qquad (\text{of } \sum_{j=0}^{N} \ldots) \tag{4.5}$$

is een lineaire combinatie van basisfuncties ϕ_j , continu op $\overline{\Omega}$ met een voorgeschreven gedrag per deelgebied. De parameter U_{Ni} in (4.5) is precies de waarde van u_N in het punt $\underline{x}^i$:

$$u_N(\underline{x}^i) = \sum_{j=1}^{N} u_{Nj}\ \phi_j(\underline{x}^i) = \sum_{j=1}^{N} u_{Nj}\ \delta_{ij} = u_{Ni}\ .$$

De onbekende parameters u_{Nj} , $j = 1,2,\ldots,N$, kunnen nu bepaald worden met behulp van de methode van Ritz. Dat wil zeggen, we substitueren u_N in de functionaal, die daardoor slechts afhangt van $u_{N1},\ldots,u_{NN}$. Stellen we de afgeleide van de functionaal naar iedere parameter u_{Ni} gelijk aan nul, dan ontstaat een stelsel lineaire vgln. voor de onbekenden $u_{N1},\ldots,u_{NN}$. Onder een Eindig Element wordt verstaan: (i) een deelgebied e_k , (ii) de knooppunten op e_k , (iii) de algemene gedaante van de basisfuncties op e_k . De hier beschreven combinatie van de methode van Ritz en de EEM zullen we in de volgende paragrafen toepassen op de voorbeelden (1.11), (1.17) en (1.18).

5. RITZ EN EEM TOEGEPAST OP VOORBEELD (1.11)

Beschouw voorbeeld (1.11) en de minimaliserings-formulering (2.4). We volgen de eindige elementen procedure. Het interval [0,1] wordt opgedeeld in een eindig aantal, zeg N, subintervallen van willekeurige lengte. We veronderstellen dat aan de voorwaarde (4.1) is voldaan. Als knooppunten $\underline{x}^i$, $i = 0,1,\ldots,N$, kiezen we de randpunten van de subintervallen (zie Fig. 7.a). De basisfuncties zijn gedefinieerd zoals in Voorbeeld (4.2). De benaderende oplossing u_N stellen we van de vorm

$$u_N(x) = \sum_{j=0}^{N} u_{Nj}\ \phi_j(x) \tag{5.1}$$

waarbij u_{Nj} de (nog) onbekende waarde is van u_N in het punt $\underline{x}^j$. Omdat $u(0) = 1$, eisen we dat $u_N(0) = 1$, hetgeen impliceert dat

$$1 = u_N(0) = \sum_{j=0}^{N} u_{Nj}\ \phi_j(0) = \sum_{j=0}^{N} u_{Nj}\ \phi_j(\underline{x}^0) = \sum_{j=0}^{N} u_{Nj}\ \delta_{j0} = u_{N0}.$$

De vorm van u_N is dus

$$u_N(x) = \phi_0(x) + \sum_{j=1}^{N} u_{Nj}\ \phi_j(x). \tag{5.2}$$

Uitdrukking (5.2) voor u_N substitueren we vervolgens in de functionaal. Het minimaliseringsprobleem wordt:

Minimaliseer

$$J(u_N) = \tfrac{1}{2}\int_0^1 \left| \frac{d}{dx}\left(\phi_0(x) + \sum_{j=1}^{N} u_{Nj}\,\phi_j(x)\right)\right|^2 dx + \int_0^1 e^x\left(\phi_0(x) + \sum_{j=1}^{N} u_{Nj}\,\phi_j(x)\right) dx +$$

$$(5.3) \qquad - e\left(\phi_0(1) + \sum_{j=1}^{N} u_{Nj}\,\phi_j(1)\right)$$

met betrekking tot $u_{N1},\ldots,u_{NN}$.

Daar $\phi_j(1) = \phi_j(\underline{x}^N) = \delta_{jN}$ kan J geschreven worden als

$$(5.4) \quad J(u_N) = \tfrac{1}{2}\int_0^1 \left| \frac{d\phi_0}{dx} + \sum_{j=1}^{N} u_{Nj}\,\frac{d\phi_j}{dx}\right|^2 dx + \int_0^1 e^x(\phi_0 + \sum_{j=1}^{N} u_{N_j}\,\phi_j)\,dx - e\,u_{NN} =$$

$$= J(u_{N1},\ldots,u_{NN}) .$$

Differentiatie met betrekking tot $u_{N1},\ldots,u_{NN}$ leidt tot het volgende stelsel vergelijkingen:

$$\frac{\partial J(u_{N1},\ldots,u_{NN})}{\partial u_{Ni}} = 0 , \quad i = 1,2,\ldots,N$$

oftewel

$$\int_0^1 \left(\frac{d\phi_0}{dx} + \sum_{j=1}^{N} u_{Nj}\,\frac{d\phi_j}{dx}\right)\frac{d\phi_i}{dx}\,dx = -\int_0^1 e^x\,\phi_i\,dx + e\,\delta_{iN}$$

oftewel

$$(5.5) \qquad \sum_{j=1}^{N} u_{Nj}\left(\int_0^1 \frac{d\phi_j}{dx}\,\frac{d\phi_i}{dx}\,dx\right) = -\int_0^1 \frac{d\phi_0}{dx}\,\frac{d\phi_i}{dx}\,dx - \int_0^1 e^x\,\phi_i\,dx + e\,\delta_{iN} .$$

Dit is een stelsel van N lineaire vgln voor de N onbekenden $u_{N1},\ldots,u_{NN}$.
In matrix-notatie hebben we

$$(5.6) \qquad \underline{A}\,\underline{u}_N = \underline{F}$$

met

$$A_{ij} = \int_0^1 \frac{d\phi_j}{dx}\,\frac{d\phi_i}{dx}\,dx , \quad (\underline{u}_N)_i = u_{Ni}$$

$$(\underline{F})_i = -\int_0^1 \frac{d\phi_0}{dx}\,\frac{d\phi_i}{dx}\,dx - \int_0^1 e^x\,\phi_i\,dx + e\,\delta_{iN} .$$

Dit stelsel zullen we nu gaan uitwerken voor het geval van een konstante subintervallengte $h = \frac{1}{N}$. Door de speciale keuze van de basisfuncties geldt dat het matrixelement $A_{ij} = 0$ als $|i-j| > 1$. Want in dat geval hebben de dragers van ϕ_i en ϕ_j een lege doorsnede. Verder geldt

$$\phi_i(x) = \frac{x-\underline{x}^{i-1}}{h} \quad , \quad \frac{d\phi_i}{dx}(x) = \frac{1}{h} \quad , \quad \text{als} \quad \underline{x}^{i-1} < x < \underline{x}^{i}$$

$$\phi_i(x) = \frac{\underline{x}^{i+1}-x}{h} \quad , \quad \frac{d\phi_i}{dx}(x) = -\frac{1}{h} \quad \text{als} \quad \underline{x}^{i} < x < \underline{x}^{i+1}$$

$$\phi_i(x) = 0 \quad , \quad \frac{d\phi_i}{dx}(x) = 0 \quad \text{als} \quad x < \underline{x}^{i-1} \text{ of } x > \underline{x}^{i+1} .$$

De matrix-elementen A_{ij} kunnen nu berekend worden. Als $i \neq N$ dan geldt

$$j = i: \quad A_{i,i} = \int_0^1 \frac{d\phi_j}{dx} \frac{d\phi_i}{dx} dx = \int_{\underline{x}^{i-1}}^{\underline{x}^{i+1}} \left(\frac{d\phi_i}{dx}\right)^2 dx = 2h \frac{1}{h}\frac{1}{h} = \frac{2}{h}$$

$$j = i+1: \quad A_{i,i+1} = \int_0^1 \frac{d\phi_j}{dx} \frac{d\phi_i}{dx} dx = \int_{\underline{x}^{i}}^{\underline{x}^{i+1}} \frac{d\phi_{i+1}}{dx} \frac{d\phi_i}{dx} dx = h \frac{1}{h}\left(-\frac{1}{h}\right) = -\frac{1}{h}$$

$$j = i-1: \quad A_{i,i-1} = \int_0^1 \frac{d\phi_j}{dx} \frac{d\phi_i}{dx} dx = \int_{\underline{x}^{i-1}}^{\underline{x}^{i}} \frac{d\phi_{i-1}}{dx} \frac{d\phi_i}{dx} dx = h\left(-\frac{1}{h}\right)\frac{1}{h} = -\frac{1}{h} .$$

Dus voor $i \neq N$: $A_{i,i-1} = -\frac{1}{h}$, $A_{ii} = \frac{2}{h}$, $A_{ii+1} = -\frac{1}{h}$, $A_{ij} = 0$ als $|i-j|>1$.

Voor $i = N$ vinden we

$$A_{N,N} = \int_{\underline{x}^{N-1}}^{\underline{x}^{N}} \left(\frac{d\phi_N}{dx}\right)^2 dx = h \frac{1}{h^2} = \frac{1}{h}$$

$$A_{N,N-1} = \int_{\underline{x}^{N-1}}^{\underline{x}^{N}} \frac{d\phi_{N-1}}{dx} \frac{d\phi_N}{dx} dx = h\left(-\frac{1}{h}\right)\frac{1}{h} = -\frac{1}{h} .$$

De matrix $\underline{A}$ van het stelsel (5.6) heeft dus de volgende vorm

$$\underline{A} = \frac{1}{h}\begin{pmatrix} 2 & -1 & & & & 0 \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & & \cdots\cdots\cdots & & \\ & 0 & & & -1 & 2 & -1 \\ & & & & & -1 & 1 \end{pmatrix}.$$

De matrix $\underline{A}$ is ijl (bevat veel nullen) en tri-diagonaal. Bovendien kunnen we aantonen dat $\underline{A}$ positief-definiet is.

Het rechterlid van (5.6) zullen we gedeeltelijk niet exact berekenen, maar met behulp van de volgende numerieke integratie formule (trapezium regel):

$$\int_a^b f(x)\,dx \approx \frac{(b-a)}{2}\,(f(b) + f(a)) .$$

Er geldt:

$$F_1 = -\int_0^1 \frac{d\phi_0}{dx}\,\frac{d\phi_1}{dx}\,dx - \int_0^1 e^x \phi_1\,dx = \frac{1}{h} - \int_0^{\underline{x}^1} e^x\,\frac{x}{h}\,dx - \int_{\underline{x}^1}^{\underline{x}^2} \frac{e^x(\underline{x}^2 - x)}{h}\,dx =$$

$$= \frac{1}{h} - \frac{h}{2}\left(\frac{e^h h}{h} + 0\right) - \frac{h}{2}\,\frac{0 + e^h h}{h} = \frac{1}{h} - h\,e^h$$

$$F_i = -\int_0^1 e^x \phi_i\,dx = -\int_{\underline{x}^{i-1}}^{\underline{x}^i} e^x\,\frac{x - \underline{x}^{i-1}}{h}\,dx - \int_{\underline{x}^i}^{\underline{x}^{i+1}} e^x\,\frac{\underline{x}^{i+1} - x}{h}\,dx =$$

$$= -\frac{h}{2}\left(e^{\underline{x}^i}\,\frac{h}{h} + 0\right) - \frac{h}{2}\left(0 + e^{\underline{x}^i}\,\frac{h}{h}\right) = -h e^{\underline{x}^i} = -h e^{ih}, \quad i = 2,3,\ldots,N-1$$

$$F_N = -\int_0^1 e^x \phi_N\,dx + e\,\delta_{NN} = -\int_{\underline{x}^{N-1}}^{\underline{x}^N} e^x\,\frac{x - \underline{x}^{N-1}}{h}\,dx + e = -\frac{h}{2}\left(e^{\underline{x}^N}\,\frac{h}{h} + 0\right) + e =$$

$$= -\frac{h}{2}\,e + e = e\left(1 - \frac{h}{2}\right).$$

De rechterlid-vector van stelsel (5.6) heeft dus de volgende gedaante

$$\underline{F} = \begin{pmatrix} \frac{1}{h} - h e^h \\ -h e^{2h} \\ \vdots \\ -h e^{(N-1)h} \\ e\left(1 - \frac{h}{2}\right) \end{pmatrix}.$$

Het stelsel (5.6) hebben we voor verschillende waarden van N opgelost. In de volgende tabel staan naast elkaar de waarde van N en het maximale absolute verschil tussen de exacte oplossing $u(x) = e^x$ en de benadering $u_N(x)$.

TABEL 1

N	$\max \lvert u - u_N \rvert$	$N^2 \max \lvert u - u_N \rvert$
4	2.31×10^{-2}	3.70×10^{-1}
8	5.78×10^{-3}	3.70×10^{-1}
16	1.44×10^{-3}	3.69×10^{-1}
32	3.61×10^{-4}	3.70×10^{-1}
64	9.03×10^{-5}	3.70×10^{-1}
128	2.26×10^{-5}	3.70×10^{-1}
256	5.64×10^{-6}	3.70×10^{-1}
512	1.41×10^{-6}	3.70×10^{-1}
1024	3.50×10^{-7}	3.67×10^{-1}

Uit de tabel blijkt dat bij toenemende waarde van het aantal punten N (d.w.z. een steeds kleiner wordende sub-intervallengte h) het verschil tussen u en u_N afneemt. We komen hier later nog op terug, maar merken nu al vast op dat $\max \lvert u - u_N \rvert$ rechtevenredig is met $\frac{1}{N^2}$, d.w.z. met h^2.

6. RITZ EN EEM TOEGEPAST OP VOORBEELD (1.17)

Beschouw Voorbeeld (1.17) en het bijhorende minimaliseringsprobleem (2.6). We kiezen voor dezelfde eindige elementen procedure als in paragraaf 5. De benaderende oplossing u_N heeft de gedaante

$$u_N(x) = \sum_{j=0}^{N} u_{Nj}\, \phi_j(x). \tag{6.1}$$

Daar $u(0) = 1$, geldt $u_N(0) = 1$, waaruit volgt $u_{N0} = 1$:

$$u_N(x) = \phi_0(x) + \sum_{j=1}^{N} u_{Nj}\, \phi_j(x). \tag{6.2}$$

Substitutie van (6.2) in de functionaal (2.6) en de relatie $\phi_j(1) = \delta_{jN}$ leidt tot:

Minimaliseer

$$(6.3)\qquad J(u_N) = \tfrac{1}{2}\int_0^1 k\left|\frac{d}{dx}(\phi_0 + \sum_{j=1}^{N} u_{Nj}\phi_j)\right|^2 dx + 2\int_0^{\frac{1}{2}} \frac{1}{(\frac{1}{2}+x)^2}(\phi_0 + \sum_{j=1}^{N} u_{Nj}\phi_j)\,dx + 2u_{NN}$$

met betrekking tot $u_{N1},\ldots,u_{NN}$.

Differentiatie van $J(u_N)$ naar $u_{N1},\ldots,u_{NN}$ en nul stellen geeft:

$$(6.4)\qquad \int_0^1 k\left(\frac{d\phi_0}{dx} + \sum_{j=1}^{N} u_{Nj}\frac{d\phi_j}{dx}\right)\frac{d\phi_i}{dx}\,dx + 2\int_0^{\frac{1}{2}} \frac{\phi_i}{(\frac{1}{2}+x)^2}\,dx + 2\delta_{iN} = 0$$

oftewel

$$(6.5)\qquad \sum_{j=1}^{N} u_{Nj}\left(\int_0^1 k\frac{d\phi_j}{dx}\frac{d\phi_i}{dx}\,dx\right) = -\int_0^1 k\frac{d\phi_0}{dx}\frac{d\phi_i}{dx}\,dx - 2\int_0^{\frac{1}{2}} \frac{\phi_i}{(\frac{1}{2}+x)^2}\,dx - 2\delta_{iN}.$$

Dit is weer een stelsel van N lineaire vergelijkingen met N onbekenden $u_{N1},\ldots,u_{NN}$. In matrix-notatie staat er

$$(6.6)\qquad \underline{A}\,\underline{u}_N = \underline{F}$$

met

$$A_{ij} = \int_0^1 k\,\frac{d\phi_j}{dx}\frac{d\phi_i}{dx}\,dx\,, \qquad (\underline{u}_N)_i = u_{Ni}$$

$$F_i = -\int_0^1 k\,\frac{d\phi_0}{dx}\frac{d\phi_i}{dx}\,dx - 2\int_0^{\frac{1}{2}} \frac{\phi_i}{(\frac{1}{2}+x)^2}\,dx - 2\delta_{iN}.$$

De berekening van de matrix elementen gaat op exact dezelfde wijze als in paragraaf 5. Er dient echter onderscheid gemaakt te worden tussen het gebied $(0,\frac{1}{2})$, waar $k = k_1 = 2$, en het gebied $(\frac{1}{2},1)$, waar $k = k_2 = 1$.

Kiezen we de subintervallen met konstante lengte $h = \frac{1}{N}$ met $N = 2^M$, $M = 2,3,4,\ldots$ dan valt knooppunt $\underline{x}^{2^{M-1}}$ samen met $\frac{1}{2}$.

Er geldt:

$$A_{ii} = k_1 \frac{2}{h} = \frac{4}{h}, \qquad i = 1,2,\ldots,2^{M-1}-1$$

$$A_{ii} = (k_1+k_2)\frac{1}{h} = \frac{3}{h}, \qquad i = 2^{M-1}$$

$$A_{ii} = k_2 \frac{2}{h} = \frac{2}{h}, \qquad i = 2^{M-1}+1,\ldots,N-1$$

$$A_{NN} = k_2 \frac{1}{h} = \frac{1}{h},$$

$$A_{ii+1} = -k_1 \frac{1}{h} = -\frac{2}{h}, \qquad i = 1,2,\ldots,2^{M-1}-1$$

$$A_{ii+1} = -k_2 \frac{1}{h} = -\frac{1}{h}, \qquad i = 2^{M-1},\ldots,N-1$$

$$A_{ii-1} = -k_1 \frac{1}{h} = -\frac{2}{h}, \qquad i = 1,2,\ldots,2^{M-1}$$

$$A_{ii-1} = -k_2 \frac{1}{h} = -\frac{1}{h}, \qquad i = 2^{M-1}+1,\ldots,N$$

oftewel

$$\underline{A} = \frac{1}{h}\begin{bmatrix} 2k_1 & -k_1 & & & & & & \\ -k_1 & 2k_2 & -k_1 & & & & & \\ & \cdot & \cdot & \cdot & & 0 & & \\ & -k_1 & 2k_1 & k_1 & & & & \\ & & -k_1 & k_1+k_2 & -k_2 & & & \\ & & & -k_2 & 2k_2 & -k_2 & & \\ & 0 & & & \cdot & \cdot & \cdot & \\ & & & & -k_2 & 2k_2 & -k_2 & \\ & & & & & -k_2 & k_2 & \end{bmatrix}, \quad \begin{matrix} k_1 = 2 \\ k_2 = 1 \end{matrix}.$$

$\underline{A}$ is, evenals in paragraaf 5, ijl, tri-diagonaal en positief-definiet.

Voor de rechterlid-vector $\underline{F}$ vinden we

$$F_i = -\int_0^1 k \frac{d\phi_0}{dx}\frac{d\phi_i}{dx}\,dx - 2\int_0^{\frac{1}{2}} \frac{\phi_i}{(\frac{1}{2}+x)^2}\,dx - 2\delta_{iN} =$$

$$= 2\frac{1}{h}\delta_{i1} - 2\int_{\underline{x}^{i-1}}^{\underline{x}^i} \frac{1}{(\frac{1}{2}+x)^2}\frac{x-\underline{x}^{i-1}}{h}\,dx - 2\int_{\underline{x}^i}^{\underline{x}^{i+1}} \frac{1}{(\frac{1}{2}+x)^2}\frac{\underline{x}^{i+1}-x}{h}\,dx - 2\delta_{iN} =$$

$$= \frac{2}{h}\delta_{i1} - 2.\frac{h}{2}\frac{1}{(\frac{1}{2}+\underline{x}^i)^2} - 2.\frac{h}{2}\frac{1}{(\frac{1}{2}+\underline{x}^i)^2} - 2\delta_{iN} =$$

$$= \frac{2}{h}\,\delta_{i1} - 2\delta_{iN} - \frac{2h}{(\frac{1}{2}+ih)^2} =$$

$$= \frac{2}{h}\,\delta_{i1} - \frac{2h}{(\frac{1}{2}+ih)^2} \qquad \text{voor } i = 1,2,\ldots,N-1$$

en

$$F_N = -2\int_{\underline{x}^{N-1}}^{\underline{x}^N} \frac{\phi_N}{(\frac{1}{2}+x)^2}\,dx - 2 = -h\,\frac{1}{(\frac{1}{2}+1)^2} - 2 = -\frac{4h}{9} - 2.$$

Het stelsel (6.6) is opgelost voor verschillende waarden van N, corresponderend met M = 2,3,...,10. In Tabel 2 zijn opgenomen de waarde van N, het maximale absolute verschil $\max|u-u_N|$ en $N^2 \max|u-u_N|$.

N	$\max\|u-u_N\|$	$N^2 \max\|u-u_N\|$
4	1.95×10^{-2}	3.12×10^{-1}
8	5.12×10^{-3}	3.27×10^{-1}
16	1.30×10^{-3}	3.33×10^{-1}
32	3.25×10^{-4}	3.33×10^{-1}
64	8.14×10^{-5}	3.33×10^{-1}
128	2.03×10^{-5}	3.33×10^{-1}
256	5.09×10^{-6}	3.34×10^{-1}
512	1.27×10^{-6}	3.33×10^{-1}
1024	3.22×10^{-7}	3.38×10^{-1}

TABEL 2

7. RITZ EN EEM TOEGEPAST OP VOORBEELD (1.18)

We beschouwen voorbeeld (1.18) en de minimaliseringsformulering (2.5). De EEM is als volgt. Het gebied Ω wordt opgedeeld (getrianguleerd) in een eindig aantal driehoekjes e_k, k = 1, ..., K die voldoen aan voorwaarde (4.1), zie fig. 6a. In iedere driehoek kiezen we de hoekpunten als knooppunten. De basisfunctie ϕ_i, behorend bij knooppunt $\underline{x}^i$, wordt

gedefinieerd zoals in voorbeeld (4.4). De basisfunctie ϕ_i is gelijk aan 1 in $\underline{x}^i$ en gelijk aan nul in ieder ander knooppunt. Verder is ϕ_i affien op iedere driehoek en continu op $\overline{\Omega}$ (zie fig. 11). Een benaderende oplossing u_N kan geschreven worden als

$$u_N(x) = \sum_{j=1}^{N} u_{Nj}\, \phi_j(x) \tag{7.1}$$

waarbij N het aantal knooppunten $\underline{x}^1$, $\underline{x}^2$,...,$\underline{x}^N$ is in $\overline{\Omega}$ die niet tot Γ_0 behoren; immers op Γ_0 moet gelden $u_N = 0$. De functie u_N is gedefinieerd op $\overline{\Omega}$, affien per driehoek, continu op $\overline{\Omega}$, gelijk aan nul op Γ_0 en heeft de waarde u_{Nj} in het knooppunt $\underline{x}^j$, $j = 1,2,\ldots,N$. De algemene gedaante (7.1) van u_N wordt gesubstitueerd in de functionaal J en we formuleren het benaderende probleem als volgt

$$\begin{cases} \text{Minimaliseer } J(u_N) \\ \text{met betrekking tot } u_{N1},\ldots,u_{NN}. \end{cases} \tag{7.2}$$

Nodige voorwaarde voor het minimaal zijn is dat

$$\frac{\partial J(u_N)}{\partial u_{Ni}} = 0 , \quad i = 1,2,\ldots,N$$

hetgeen impliceert dat

$$\int_\Omega \text{grad } u_N \cdot \text{grad } \phi_i \, d\Omega - 8 \int_\Omega x_1 \phi_i \, d\Omega - 4 \int_{\Gamma_1} x_2(1-x_2)\, \phi_i \, d\Gamma = 0$$

oftewel

$$\sum_{j=1}^{N} u_{Nj} \left(\int_\Omega \text{grad } \phi_j \cdot \text{grad } \phi_i \, d\Omega \right) = 8 \int_\Omega x_1 \phi_i \, d\Omega + 4 \int_{\Gamma_1} x_2(1-x_2)\phi_i \, d\Gamma, \quad i = 1,2,\ldots,N. \tag{7.3}$$

Hier staat een stelsel van N lineaire vgln in $u_{N1},\ldots,u_{NN}$

$$\underline{A}\, \underline{u}_N = \underline{F} \tag{7.4}$$

met

$$\begin{cases} A_{ij} = \int_\Omega \text{grad } \phi_j \cdot \text{grad } \phi_i \, d\Omega , \quad (\underline{u}_N) = u_{Ni} \\ F_i = 8 \int_\Omega x_1 \phi_i \, d\Omega - 4 \int_{\Gamma_1} x_2(1-x_2)\, \phi_i \, d\Gamma. \end{cases}$$

We zien reeds dat de matrix $\underline{A}$ ijl is, daar $A_{ij} = 0$ als de dragers van ϕ_j en ϕ_i een lege doorsnede hebben. Dit is het geval als de knooppunten $\underline{x}^i$ en $\underline{x}^j$ niet tot een zelfde driehoek behoren.

Door de opbouw (assemblage) van de matrix $\underline{A}$ en de rechterlid-vector $\underline{F}$ worden in (7.3) de integralen over Ω opgesplitst in een som van integralen over alle driehoeken e_k, $k = 1,\ldots,K$. De integraal over Γ_1 wordt geschreven als een som van integralen over zijden ℓ_p, $p = 1,\ldots,P$, van driehoeken die op Γ_1 gelegen zijn (zie fig. 12).

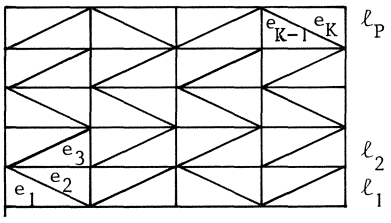


Fig. 12. Driehoeken e_k, zijden ℓ_p

Schrijven we $\int_\Omega \ldots d\Omega = \sum_{k=1}^{K} \int_{e_k} \ldots d\Omega$ en $\int_{\Gamma_1} \ldots d\Gamma = \sum_{p=1}^{P} \int_{\ell_p} \ldots d\Gamma$ dan gaat (7.3) over in

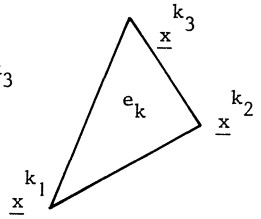
$$(7.5) \qquad \sum_{k=1}^{K} \sum_{j=1}^{N} u_{hj} \left(\int_{e_k} \text{grad } \phi_j . \text{grad } \phi_i \, d\Omega \right) = \sum_{k=1}^{K} \int_{e_k} 8x_1 \phi_i \, d\Omega + \sum_{p=1}^{P} \int_{\ell_p} 4x_2(1-x_2)\phi_i \, d\Gamma.$$

We zien dus dat de assemblage van de matrix $\underline{A}$ en de vector $\underline{F}$ neerkomt op het berekenen van de integralen

$$(7.6) \qquad \int_{e_k} \text{grad } \phi_j . \text{grad } \phi_i \, d\Omega \,, \quad \int_{e_k} 8x_1 \phi_i \, d\Omega \text{ en } \int_{\ell_p} 4x_2(1-x_2)\phi_i \, d\Gamma.$$

Wat betreft de eerste integraal van (7.6) merken we op dat op de driehoek e_k slechts drie basisfuncties ongelijk nul zijn, nl. die basisfuncties die behoren bij de hoekpunten $\underline{x}^{k_1}$, $\underline{x}^{k_2}$, $\underline{x}^{k_3}$ van driehoek e_k (zie fig. 13):

Fig. 13. Driehoek e_k met knooppunten $\underline{x}^{k_1}$, $\underline{x}^{k_2}$ en $\underline{x}^{k_3}$



De integraal $I_{ij}^{e_k} \equiv \int_{e_k} \text{grad } \phi_j . \text{grad } \phi_i \, d\Omega$ is dus ongelijk nul voor $i = k_1, k_2, k_3$ en $j = k_1, k_2, k_3$. Dit levert 9 combinaties {i,j} op, die we in een matrix $\underline{A}^{e_k}$ plaatsen:

$$(7.7) \qquad \underline{A}^{e_k} = \begin{bmatrix} I_{k_1k_1}^{e_k} & I_{k_1k_2}^{e_k} & I_{k_1k_3}^{e_k} \\ I_{k_2k_1}^{e_k} & I_{k_2k_2}^{e_k} & I_{k_2k_3}^{e_k} \\ I_{k_3k_1}^{e_k} & I_{k_3k_2}^{e_k} & I_{k_3k_3}^{e_k} \end{bmatrix} .$$

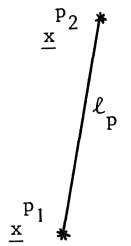
Deze matrix heet de element-matrix.

De tweede integraal van (7.6), $I_i^{e_k} = \int_{e_k} 8x_1 \, \phi_i \, d\Omega$ is niet nul als $i = k_1, k_2, k_3$. Deze drie mogelijkheden plaatsen we in de element-vector $\underline{F}^{e_k}$:

$$(7.3) \qquad \underline{F}^{e_k} = \begin{bmatrix} I_{k_1}^{e_k} \\ I_{k_2}^{e_k} \\ I_{k_3}^{e_k} \end{bmatrix} .$$

De derde integraal in (7.6), $I_i^{\ell_p} = \int_{\ell_p} 4x_2(1-x_2)\phi_i \, d\Gamma$ is ongelijk nul als het knooppunt $\underline{x}^i$ tot de zijde ℓ_p behoort. Dit levert twee mogelijkheden op voor i, nl. $i = p_1, p_2$ (zie fig. 14):

Fig. 14. Zijde ℓ_p met knooppunten $\underline{x}^{p_1}$, $\underline{x}^{p_2}$



De twee grootheden $I_i^{\ell_p}$, $i = p_1, p_2$, plaatsen we in de lijnelement-vector $\underline{F}^{\ell_p}$

$$\underline{F}^{\ell_p} = \begin{bmatrix} I_{p_1}^{\ell_p} \\ I_{p_2}^{\ell_p} \end{bmatrix} . \tag{7.9}$$

De algemene procedure voor de opbouw van de matrix $\underline{A}$ en de vector $\underline{F}$ is als volgt. Begin met een matrix $\underline{A}$ bestaande uit N×N nullen, en een vector $\underline{F}$ bestaande uit N nullen, waarbij N het aantal onbekenden is. Bereken vervolgens voor iedere driehoek e_k de elementmatrix $\underline{A}^{e_k}$ en de elementvector $\underline{F}^{e_k}$ en bereken voor iedere op Γ_1 gelegen zijde van een driehoek de lijnelement-vector $\underline{F}^{\ell_p}$. Deze matrices en vectoren worden nu opgeteld bij de matrix $\underline{A}$, in die zin, dat het matrixelement $I^{e_k}_{k_i k_j}$ van $\underline{A}^{e_k}$ wordt opgeteld bij matrixelement $A_{k_i k_j}$; het vectorelement $I^{e_k}_{k_i}$ wordt opgeteld bij F_{k_i} en het vectorelement $I^{\ell_p}_{p_i}$ wordt opgeteld bij F_{p_i} . Na assemblage van $\underline{A}$ en $\underline{F}$ kan het stelsel vgln worden opgelost. We hebben dit uitgevoerd voor de volgende triangulatie van Ω (zie fig. 15):

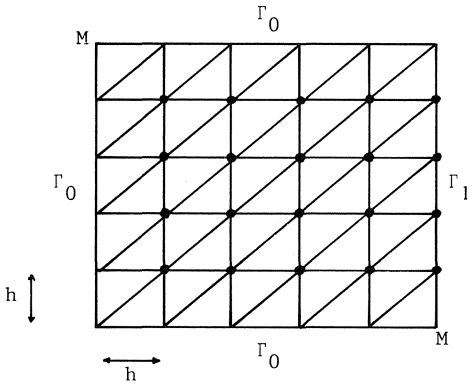


Fig. 15. Triangulatie van Ω.
N = M(M-1) onbekenden,
M^2 driehoeken, $h = \frac{1}{M}$,
M zijden op Γ_1

In Tabel 3 zijn opgenomen het maximale absolute verschil tussen u_N en u: $\max_{x \in \overline{\Omega}} |u - u_N|$ en de maaswijdte (= grootte van de driehoeken) h.

h	max $\lvert u - u_N \rvert$	$\frac{1}{h^2}$ max$\lvert u - u_N \rvert$
2^{-3}	1.31×10^{-2}	8.38×10^{-1}
2^{-4}	3.01×10^{-3}	7.70×10^{-1}
$2^{-4.5}$	1.48×10^{-3}	7.58×10^{-1}

TABEL 3

8. FOUTSCHATTING VAN DE EEM

Bij een foutschatting van de EEM vragen we ons af hoe groot het verschil is tussen de exacte oplossing u (die we meestal niet kunnen berekenen) en de benadering u_N. De fout wordt gedefinieerd als

(8.1) $$\varepsilon(\underline{x}) = u(\underline{x}) - u_N(\underline{x}).$$

Ondanks het feit dat u onbekend is, is het toch mogelijk iets over het gedrag van ε te zeggen. Informatie over het gedrag van ε is van groot belang voor praktische berekeningen want het zegt iets over de vraag of de benaderende oplossing acceptabel is of dat nog een verfijning van de deelgebieden uitgevoerd moet worden.

In de theorie van de EEM wordt een schatting gemaakt van een norm van ε. Gebruikelijke normen zijn bijvoorbeeld

$$\|\varepsilon\|_0 = \{\int_\Omega |\varepsilon^2(\underline{x})| \, d\Omega\}^{\frac{1}{2}}, \qquad (L_2\text{-norm})$$

$$\|\varepsilon\|_1 = \{\|\varepsilon\|_0^2 + \|\text{grad } \varepsilon\|_0^2\}^{\frac{1}{2}}, \qquad (H^1\text{-norm})$$

$$\|\varepsilon\|_\infty = \max_{\underline{x}\in\Omega} |\varepsilon(\underline{x})|, \qquad (\infty\text{-norm}).$$

Deze laatste grootheid is steeds getabelleerd in de Tabellen 1, 2 en 3. Voor de hier beschouwde model-problemen en voor affiene basisfuncties kunnen we bewijzen dat

(8.2) $$\|\varepsilon\|_\infty \leq c.h^2$$

waarbij h een maat is voor de grootte van de deelgebieden (bijv. $h = \frac{1}{N}$ in het 1-dimensionale geval) en waarbij c een constante is die van de gegevens

van het probleem afhangt maar niet van h. In de derde kolom van de Tabellen 1, 2 en 3 is te zien dat aan de foutschatting (8.2) is voldaan: bij halvering van h (d.w.z. een verdubbeling van N) neemt de fout met een factor $\frac{1}{4}$ af.

9. DE METHODE VAN GALERKIN

Bij de methode van Ritz hebben we essentieel gebruik gemaakt van de minimaliserings-formulering van de (part.) diff. vgl.. De methode van Ritz is dan ook slechts toe te passen op die (part.) diff. vgln waarvoor een equivalent minimaliseringsprobleem bestaat. De zelf-geadjungeerde diff. vgln voldoen aan deze voorwaarde. Voor het niet zelf-geadjungeerde geval kunnen we de methode van Ritz niet toepassen. Een voorbeeld van een niet zelf-geadjungeerd geval is het voorbeeld (1.11) waaraan een eerste orde term wordt toegevoegd:

$$(9.1)\qquad \begin{cases} -\dfrac{d^2u}{dx^2} + \dfrac{du}{dx} = -e^{-x} & \text{op } (0,1) \\ u(0) = 1 & \\ -\dfrac{du}{dx}(1) = -e. & \end{cases}$$

Voor dit probleem bestaat geen equivalent minimaliseringsprobleem. We zullen nu een generalisatie van de methode van Ritz bespreken, die ook op niet zelf-geadjungeerde problemen toepasbaar is. Deze methode heet de methode van Galerkin. We zullen de methode van Galerkin toepassen op het 1-dimensionale voorbeeld (1.11) en op het 2-dimensionale voorbeeld (1.18), natuurlijk zonder gebruik te maken van de minimaliseringsformulering.

De algemene werkwijze van de methode van Galerkin is als volgt:

(i) Vermenigvuldig de (part.) diff. vgl. met een zgn. testfunctie w, die voldoet aan de homogene Dirichlet-randvoorwaarden van het probleem.

(ii) Integreer over het gebied Ω.

(iii) Pas de formule van Green toe (dit is partiële integratie in het 1-dimensionale geval) om de tweede orde afgeleide te verlagen tot een eerste orde afgeleide.

(iv) Substitueer de randvoorwaarden.

(v) Ga over op eindig-dimensionale functieruimten met behulp van de EEM.

Beschouw de diff. vgl. van voorbeeld (1.11). Zij $w : [0,1] \to \mathbb{R}$ een willekeurige (test-) functie die voldoet aan de homogene Dirichlet-randvoorwaarde van het probleem: $w(0) = 0$. Vermenigvuldig de diff. vgl. met w en integreer over $\Omega = (0,1)$:

$$-\int_0^1 \frac{d^2u}{dx^2}\, w\, dx = -\int_0^1 e^x\, w\, dx\ .$$

Partiële integratie geeft

$$\int_0^1 \frac{du}{dx}\,\frac{dw}{dx}\, dx - \left[\frac{du}{dx}\, w\right]_0^1 = -\int_0^1 e^x w\, dx$$

en wegens $w(0) = 0$ wordt dit

$$\int_0^1 \frac{du}{dx}\,\frac{dw}{dx}\, dx - \frac{du}{dx}(1)\, w(1) = -\int_0^1 e^x\, w\, dx\,.$$

Substitutie van de randvoorwaarde $-\frac{du}{dx}(1) = -e$ levert op

$$\int_0^1 \frac{du}{dx}\,\frac{dw}{dx}\, dx = -\int_0^1 e^x w\, dx + e\, w(1).$$

We introduceren nu de zwakke formulering van (1.11) als volgt

(9.2)
$$\begin{cases} \text{Zoek } u : [0,1] \to \mathbb{R} \text{ met } u(0) = 1 \text{ zodanig dat} \\ \displaystyle\int_0^1 \frac{du}{dx}\,\frac{dw}{dx}\, dx = -\int_0^1 e^x w\, dx + e\, w(1) \\ \text{voor alle } w : [0,1] \to \mathbb{R} \text{ met } w(0) = 0. \end{cases}$$

De klassen van functies waarin u en w zijn gedefinieerd worden vervolgens met de EEM gediscretiseerd. Dat wil zeggen: we stellen u_N van de vorm (5.2) en w_N van de vorm

(9.3)
$$w_N(x) = \sum_{i=1}^{N} w_{Ni}\, \phi_i(x).$$

Het gediscretiseerde probleem wordt nu gedefinieerd als:

(9.4)
$$\left|\begin{array}{l} \text{Zoek } u_N : [0,1] \to \mathbb{R} \text{ van de vorm (5.2) zodanig dat} \\ \int_0^1 \frac{du_N}{dx}\,\frac{dw_N}{dx}\,dx = -\int_0^1 e^x\, w_N\, dx + e\, w_N(1) \\ \text{voor alle } w_N \text{ van de vorm (9.3).} \end{array}\right.$$

Wegens de lineariteit van (9.4) in w_N, is (9.4) equivalent met:

(9.5)
$$\left|\begin{array}{l} \text{Zoek } u_N : [0,1] \to \mathbb{R} \text{ van de vorm (5.2) zodanig dat} \\ \int_0^1 \frac{du_N}{dx}\,\frac{d\phi_i}{dx}\,dx = -\int_0^1 e^x\,\phi_i\,dx + e\,\phi_i(1), \\ i = 1,2,\ldots,N \end{array}\right.$$

oftewel

(9.6)
$$\left|\begin{array}{l} \text{Zoek } u_{N1},\ldots,u_{NN} \text{ zodanig dat} \\ \sum_{j=1}^{N} u_{Nj}\left(\int_0^1 \frac{d\phi_j}{dx}\,\frac{d\phi_0}{dx}\,dx\right) = -\int_0^1 \frac{d\phi_0}{dx}\,\frac{d\phi_i}{dx}\,dx - \int_0^1 e^x\,\phi_i\,dx + e\,\phi_i(1), \\ i = 1,2,\ldots,N \end{array}\right.$$

en dit is precies hetzelfde probleem als (5.5). We hebben dus voor voorbeeld (1.11) hetzelfde stelsel lineaire vgln (5.6) afgeleid, maar nu zonder gebruik te maken van de minimaliseringsformulering.

Beschouw vervolgens de part. diff. vgl. van voorbeeld (1.18). Zij $w : \overline{\Omega} \to \mathbb{R}$ een willekeurige (test-)functie die voldoet aan $w = 0$ op Γ_0. We vermenigvuldigen de part. diff. vgl. met w en integreren over Ω:

$$-\int_\Omega \left(\frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2}\right) w\, d\Omega = 8\int_\Omega x_1 w\, d\Omega.$$

Toepassen van de formule van Green geeft

(9.7)
$$\int_\Omega \text{grad } u.\text{grad } w\, d\Omega - \int_\Gamma \frac{\partial u}{\partial \underline{\nu}}\, w\, d\Gamma = 8\int_\Omega x_1 w\, d\Omega.$$

Omdat $w = 0$ op Γ_0, reduceert (9.7) zich tot

$$\int_\Omega \text{grad } u.\ \text{grad } w\, d\Omega - \int_{\Gamma_1} \frac{\partial u}{\partial \underline{\nu}}\, w\, d\Gamma = 8\int_\Omega x_1 w\, d\Omega.$$

Substitutie van de randvoorwaarde $\frac{\partial u}{\partial \underline{\nu}}(1,x_2) = 4x_2(1-x_2)$ leidt tot de volgende zwakke formulering:

(9.8)
$$\left|\begin{array}{l} \text{Zoek } u : \bar{\Omega} \to \mathbb{R} \quad \text{met } u = 0 \text{ op } \Gamma_0 \quad \text{zodanig dat} \\ \int_\Omega \text{grad } u.\ \text{grad } w\ d\Omega = 8 \int_\Omega x_1 w\ d\Omega = 4 \int_{\Gamma_1} x_2(1-x_2)\, w\, d\Gamma \\ \text{voor alle } w : \bar{\Omega} \to \mathbb{R} \text{ met } w = 0 \text{ op } \Gamma_0. \end{array}\right.$$

Discretisatie van u en w met de EEM geeft:

(9.9)
$$\left|\begin{array}{l} \text{Zoek } u_N : \bar{\Omega} \to \mathbb{R} \text{ van de vorm (7.1)} \quad \text{zodanig dat} \\ \int_\Omega \text{grad } u_N.\ \text{grad } w_N\, d\Omega = 8 \int_\Omega x_1 w_N\, d\Omega + 4 \int_{\Gamma_1} x_2(1-x_2) w\, d\Gamma \\ \text{voor alle } w_N \text{ van de vorm } w_N = \sum_{i=1}^{N} w_{Ni}\, \phi_i. \end{array}\right.$$

Op grond van de lineariteit in w_N, is (9.9) equivalent met

(9.10)
$$\left|\begin{array}{l} \text{Zoek } u_N : \bar{\Omega} \to \mathbb{R} \text{ van de vorm (7.1)} \quad \text{zodanig dat} \\ \int_\Omega \text{grad } u_N.\text{grad } \phi_i\, d\Omega = 8 \int_\Omega x_1 \phi_i\, d\Omega + 4 \int_{\Gamma_1} x_2(1-x_2)\phi_i\, d\Gamma, \\ i = 1,2,\ldots,N \end{array}\right.$$

oftewel

(9.11)
$$\left|\begin{array}{l} \text{Zoek } u_{N1},\ldots,u_{NN} \quad \text{zodanig dat} \\ \sum_{j=1}^{N} u_{Nj} \left(\int_\Omega \text{grad } \phi_j.\ \text{grad } \phi_i\, d\Omega\right) = 8 \int_\Omega x_1 \phi_i\, d\Omega + 4 \int_{\Gamma_1} x_2(1-x_2)\phi_i\, d\Gamma, \\ i = 1,2,\ldots,N \end{array}\right.$$

hetgeen identiek is met stelsel (7.3).

Algemeen kunnen we stellen dat de methode van Galerkin een generalisatie is van de methode van Ritz, in die zin dat Galerkin is toe te passen op een grotere klasse van (part.) diff. vgln dan Ritz, en dat voor problemen waarop zowel Ritz als Galerkin toepasbaar zijn, beide methoden resulteren in hetzelfde stelsel lineaire vgln.

LITERATUUR

1. E.B. BECKER, G.F. CAREY, J.T. ODEN, *Finite Elements, an Introduction*, Vol. I, Prentice-Hall, New Jersey, 1981. (Inleiding tot EEM.)

2. P.G. CIARLET, *The FEM for Elliptic Problems*, North-Holland, 1978. (Geavanceerd, functionaal-analytische benadering van EEM.)

3. C. CUVELIER, A. SEGAL, A.A. v. STEENHOVEN, *EEM in de stromingsleer*, 1983. (PATO-cursus Eindhoven - Delft.)

4. C. CUVELIER, A. SEGAL, A.A. v. STEENHOVEN, *FEMs and the Navier-Stokes Equations*, Reidel-Publishing Company, verschijnt in 1985. (Eerste gedeelte: inleidend; tweede gedeelte: toepassing in stromingsleer; derde gedeelte: functionaal-analytische benadering.)

5. A.J. DAVIES, *The FEM, A First Approach*, Oxford, Clarendon Press, 1980. (Inleiding tot EEM.)

6. E. HINTON, D.R.J. OWEN, *Finite Element Programming*, Academic Press, New York, 1977.

7. A.R. MITCHELL, R. WAIT, *The FEM in Partial Differential Equations*, John Wiley & Sons, London, 1977. (Inleidend.)

8. J.T. ODEN, J.N. REDDY, *An Introduction to the Mathematical Theory of Elements*, John Wiley, New York, 1976. (Geavanceerd.)

9. P.A. RAVIART, J.M. THOMAS, *Introduction à l'analyse numérique des équations aux dérivées partielles*, Masson, Paris, 1983. (Geavanceerd.)

10. G. STRANG, G.J. FIX, *An Analysis of the FEM*, Prentice-Hall, New Jersey, 1973. (Geavanceerd.)

HOOFDSTUK 6

DUALITEIT IN DE OPTIMALISERING

J. PONSTEIN

. INLEIDENDE OPMERKINGEN EN DEFINITIES

'anzelfsprekend zijn er twee typen optimaliseringsproblemen:

$$\inf_x \{f(x) : x \in G\} \text{ en } \sup_x \{f(x) : x \in G\}.$$

:n beide gevallen is x de *beslissingsvariabele*. Deze moet element zijn van :en of andere *ruimte* X. In principe mag X een willekeurige verzameling :ijn, maar vrijwel altijd is het een *lineaire ruimte*, waar men vanzelf bin- ıen blijft als men zich beperkt tot optelling en skalaire vermenigvuldiging. /erder is f(x) een reëel getal en is G een deelverzameling van X. f is de *doelstelling*(*sfunktie*) en G is het *toelaatbare gebied*. Bekend zijn X, f en ;; gevraagd wordt het inf of het sup van f(x) te bepalen onder de voorwaar- de dat x in G ligt, en liefst, als er één is, een $x^0 \in G$, zdd $f(x^0) = \inf$ of $f(x^0) = \sup$, in welk geval x^0 een *optimale oplossing*, of gewoon *oplossing* is van het gegeven probleem. Als $x \in G$, dan heet x *toelaatbaar*; als $x \in X$, maar $x \notin G$, dan heet x *ontoelaatbaar*. Als $G \neq \emptyset$ resp. $G = \emptyset$, dan heet het gegeven probleem *toelaatbaar* resp. *ontoelaatbaar*. Dit laatste lijkt een nut- teloze definitie, immers G is bekend ondersteld. Met G bekend te onderstel- len bedoelen we echter alleen dat het (meestal eenvoudig) mogelijk is om vast te stellen of $x \in G$ dan wel $x \notin G$.

VOORBEELD

$X = \mathbb{R}^n$, A is een m bij n matrix, $b \in \mathbb{R}^m$, $G = \{x : Ax \geq b\}$. De ongelijkheid hier betekent dat $(Ax)_i \geq b_i$ voor alle i.

Inderdaad is gemakkelijk vast te stellen of $Ax \geq b$, dan wel $Ax \ngeq b$. Maar dit wil nog niet zeggen dat we gemakkelijk alles over G te weten kunnen komen. De vraag of G al of niet leeg is kan lastig zijn, alsook de vraag naar de hoekpunten van G als die er zijn. Dus met "G is bekend" is alleen bedoeld: G is volledig omschreven.

Omdat $\sup_x \{f(x) : x \in G\} = -\inf_x \{-f(x) : x \in G\}$ kunnen we ons beperken tot inf-problemen (voor de straks in te voeren duale problemen houdt dit de be- perking in tot sup-problemen).

Zoals gezegd is f(x) een element van $\mathbb{R}$. Dit is echter niet het meest algemene geval, want ook als bijvoorbeeld $f(x) \in \mathbb{R}^3$ kan men daar een optimaliserings-

probleem bij maken. Dat is er dan een met *meervoudige doelstelling*, waarvoor de begrippen infimum en supremum op gepaste wijze dienen te worden gegeneraliseerd. We zullen dit algemenere geval buiten beschouwing laten. Datzelfde geldt voor het meeste dat te maken heeft met *geheeltalligheid* en *kombinatoriek*.

Hoewel het voor de praktijk weinig interessant en voor de numerieke oplossing van optimaliseringsproblemen weinig handig is om met $+\infty$ en met $-\infty$ te werken, is dit voor de theorie zeer plezierig. En wel omdat het ogenschijnlijk principiële verschil in karakter van doelstellingsfunktie en beperking ermee kan worden opgeheven! Definieer namelijk, uitgaande van het inf-probleem

$$f^0(x) = f(x) \text{ als } x \in G \quad \text{en } f^0(x) = +\infty \text{ als } x \notin G$$

dan kunnen we het inf-probleem simpelweg schrijven als

$$\inf_x f^0(x)$$

schijnbaar een probleem zonder beperkingen. Als omgekeerd, $f^0(x) \in \mathbb{R} \cup \{+\infty\}$, dan kunnen we stellen

$$G = \{x : f^0(x) < +\infty\},\ f(x) = f^0(x) \text{ als } x \in G,\ f(x) = \text{willekeurig als } x \notin G$$

en krijgen we de uitgangsformulering weer terug.

Een eerste voordeel van deze manipulatie valt ons toe bij de definitie van konvexiteit. We noemen het inf-probleem *konvex* indien $f^0(x)$ *konvex* is, en dit is per definitie het geval indien de verzameling

$$\{(x,\mu) : \mu \in \mathbb{R},\ \mu \geq f^0(x)\}$$

konvex is (dit vereist dat X een lineaire ruimte is). Gemakkelijk is aan te tonen dat het inf-probleem konvex is precies dan als f(x) konvex is in de gewone zin van het woord, en G konvex is. Dus de konvexiteit van f^0 impliceert die van f en van G, en omgekeerd. De zojuist ingevoerde verzameling heet *epigraaf* van f^0, met als notatie

$$\text{epi } f^0.$$

2. LAGRANGE MULTIPLIKATOREN EN DUALITEIT

Een bijzonder eenvoudige manier om door te dringen tot de kern van de Lagrange multiplikator en van dualiteit, is in het gegeven inf-probleem een willekeurig aantal willekeurige *parameters* te selekteren, en na te gaan wat er met het infimum gebeurt indien deze parameters worden gevarieerd. Laten we alle geselekteerde parameters representeren door één enkele vektor y, dan kunnen we het infimum in afhankelijkheid van y voorstellen door

$$p(y).$$

Laten we verder aannemen dat $y = 0$ overeenkomt met de waarden der parameters in het gegeven probleem.

VOORBEELD

$\inf_x \{f(x) : Ax \geq b\}$. Kies als parameters de komponenten van b. Dan is bijvoorbeeld $p(y) = \inf_x \{f(x) : Ax + y \geq b\}$, maar -y of 3y i.p.v. y mag ook.

Nemen we $y \neq 0$ dan kunnen we ook zeggen dat het gegeven probleem is *gestoord*. De te selekteren parameters mogen zowel in f als in G thuis horen, en daarom is het handig weer gebruik te maken van de f^0 van §1. Kies namelijk $F(x,y)$ zodanig dat $F(x,0) = f^0(x)$.

VOORBEELD (vervolg).

$F(x,y) = f(x)$ als $Ax + y \geq b$.

Bij elke keus voor f, G en de te variëren parameters is zo een F te konstrueren (doorgaans is F niet eenduidig). Omgekeerd kunnen we uitgaan van een willekeurige $F(x,y) \in \mathbb{R} \cup \{+\infty\}$, en stellen $f^0(x) = F(x,0)$. Het werken met f^0 en F geeft veel minder omhaal, omdat we anders apart een storing y_1 voor f, en een storing y_2 voor G zouden moeten invoeren, tenminste als we volledige flexibiliteit wensen te behouden. Het nadeel is natuurlijk dat we met oneindige waarden moeten werken, maar die zullen we bij het dualiseren toch niet kunnen ontlopen!

VOORBEELD

$$F(x,y) = \begin{cases} f(x) + g(Ax+y) & \text{als } x \in C \text{ en } Ax + y \in D \\ +\infty & \text{zo niet.} \end{cases}$$

Kennelijk is $p(y) = \inf_x \{f(x)+g(Ax+y) : x \in C, Ax + y \in D\}$

zodat $p(0) = \inf_x \{f(x)+g(Ax) : x \in C, Ax \in D\}$.

Voorlopig nemen we aan dat $y = (y_1,\ldots,y_m) \in \mathbb{R}^m$ voor zekere m. In het algemeen zal uiteraard $p(y) \neq p(0)$ indien $y \neq 0$, dus indien het gegeven probleem wordt gestoord. Laten we trachten het effekt van de storing te neutraliseren m.b.v. een *lineaire term*

$$\lambda y = \lambda_1 y_1 + \ldots + \lambda_m y_m$$

waarin de λ_i de Lagrange multiplikatoren zijn (of λ de Lagrange multiplikator is).

Neem nu eens aan (met excuus omdat deze veel gebezigde uitdrukking in wiskundige teksten ook hier camoufleert dat er iets uit de lucht komt vallen), dat er een λ^0 was zodanig dat

$$p(y) + \lambda^0 y \geq p(0) \text{ voor alle } y \in \mathbb{R}^m \text{ (groot of klein)}.$$

Anders gezegd, neem eens aan dat we het introduceren van de storing y vergezeld laten gaan van het *bijtellen* van de term $\lambda^0 y$, en dat dit tot resultaat heeft dat het netto effekt van de storing *niet* van voordeel is. Deze simpele aanname voert ons direkt tot de kern van de zaak. Want de ongelijkheid is ekwievalent met

$$\inf_y \{p(y)+\lambda^0 y\} = p(0) = \inf_x F(x,0)$$

en als we een willekeurige λ nemen, dan volgt dat

$$\inf_y \{p(y)+\lambda y\} \leq p(0) + \lambda.0 = p(0) = \inf_y \{p(y)+\lambda^0 y\}$$

en dus dat

$$\sup_y \{\inf_y \{p(y)+\lambda y\}\} = \inf \{p(y)+\lambda^0 y\}.$$

Bijna vanzelf hebben we een nieuw optimaliseringsprobleem gemaakt, met λ als beslissingsvariabele, en met λ^0 als optimale oplossing. Kortom,

als voor alle y *geldt dat* $p(y) + \lambda^0 y \geq p(0)$ *dan is* sup = max = inf.

Omgekeerd,

als sup = inf *en als het supremum wordt aangenomen voor* λ^0 *dan geldt dat* $p(y) + \lambda^0 y \geq p(0)$ *voor alle* y,

zoals gemakkelijk is na te gaan. Het nieuwe probleem is het *duale probleem*, het oude is het *primaire probleem*. Uiteraard is het duale probleem ook gedefinieerd indien er geen λ^0 bestaat die aan de uitgangsongelijkheid voldoet, en we hebben

$$\sup_\lambda \{\inf_y \{p(y)+\lambda y\}\} = \sup_\lambda \{\inf_{x,y} \{F(x,y) + \lambda y\}\} =$$
$$= \sup_\lambda \{\inf_x \{\inf_y \{F(x,y) + \lambda y\}\}\}$$

uitdrukkingen die er niet aantrekkelijk uitzien. In veel gevallen is er echter een aanzienlijke reduktie mogelijk.

VOORBEELD (LINEAIR PROGRAMMEREN)

$x \in \mathbb{R}^n$, $c \in \mathbb{R}^n$, A $m \times n$, $b \in \mathbb{R}^m$, $y \in \mathbb{R}^m$,

$$F(x,y) = \begin{cases} cx & \text{als } Ax + y \geq b,\ x \geq 0 \\ +\infty & \text{zo niet} \end{cases}$$

dus $p(0) = \inf_x \{cx : Ax \geq b,\ x \geq 0\}$ en $p(y) = \inf_x \{cx : Ax + y \geq b,\ x \geq 0\}$. Gemakkelijk is in te zien dat als $\lambda \not\geq 0$, dan is $\inf_y \{p(y) + \lambda y\} = -\infty$, een uitkomst die voor het duale probleem weinig interessant is. Dus kunnen we aannemen dat $\lambda \geq 0$, maar dan volgt dat $\inf_y \{F(x,y) + \lambda y\} = cx + \lambda(b-Ax)$. Maar het infimum over x hiervan is weer $-\infty$ indien $c - \lambda A \not\geq 0$, en dus kunnen we bovendien aannemen dat $c - \lambda A \geq 0$, en dan blijft er van dit infimum nog λb over. Dus het duale probleem is $\sup_\lambda \{\lambda b : \lambda A \leq c,\ \lambda \geq 0\}$, en dit is opnieuw lineair programmeren. Nogmaals dualiseren levert het uitgangsprobleem weer op.

VOORBEELD (NIET-LINEAIR PROGRAMMEREN)

Gewoonlijk bedoelt men hiermee voor $x \in \mathbb{R}^n$, $f(x) \in \mathbb{R}$, $g(x) \in \mathbb{R}^m$, $y \in \mathbb{R}^m$, dat

$$F(x,y) = \begin{cases} f(x) & \text{als } g(x) \leq y \\ +\infty & \text{zo niet} \end{cases}$$

zodat $p(0) = \inf_x \{f(x) : g(x) \leq 0\}$ en $p(y) = \inf_x \{f(x) : g(x) \leq y\}$. Het duale probleem is $\sup_{\lambda \geq 0} \{\inf_x \{f(x) + \lambda g(x)\}\}$, zoals analoog aan het vorige voorbeeld is na te gaan. Veelal neemt men aan dat f en g differentieerbaar zijn, zodat, als λ^0 bestaat, $f'(x) + \lambda^0 g'(x) = 0$, en dit is de klassieke vorm van het toepassen van Lagrange multiplikatoren. Het duale probleem nogmaals

dualiseren levert in het algemeen niet het primaire probleem op.

VOORBEELD (VARIANT OP NIET-LINEAIR PROGRAMMEREN, FENCHEL DUALITEIT)

$x \in \mathbb{R}^n$, $f(x) \in \mathbb{R}$, A $m \times n$, $G \subset \mathbb{R}^m$, $y \in \mathbb{R}^m$,

$$F(x,y) = \begin{cases} f(x) & \text{als } Ax + y \in G \\ +\infty & \text{zo niet} \end{cases}$$

dus $p(0) = \inf_x \{f(x) : Ax \in G\}$ en $p(y) = \inf_x \{f(x) : Ax + y \in G\}$. Het duale probleem is $\sup_\lambda \{\inf_{x,y} \{f(x) + \lambda y : Ax + y \in G\}\} = \sup_\lambda \{\inf_{x,z} \{f(x) - \lambda Ax + \lambda z : z \in G\}\} = \sup_\lambda \{\inf_x \{f(x) - \lambda Ax\} + \inf_z \{\lambda z : z \in G\}\}$. Als inf = sup = max spreekt men van *Fenchel dualiteit*.

VOORBEELD (EVENEENS FENCHEL DUALITEIT)

$F(x,y) = f(x) + g(Ax+y)$, met $f(x) \in \mathbb{R}$ of $f(x) = +\infty$ en $g(y) \in \mathbb{R}$ of $g(y) = +\infty$. Het primaire probleem is $\inf_x \{f(x) + g(Ax)\}$ en het duale is $\sup_\lambda \{\inf_x \{f(x) - \lambda Ax\} + \inf_z \{g(z) + \lambda z\}\}$. Met

$$g(z) = \begin{cases} 0 & \text{als } z \in G \\ +\infty & \text{zo niet} \end{cases}$$

gaat dit voorbeeld over in het vorige; dan is g de *indikator funktie* van G (elders in de wiskunde is de definitie daarvan anders).

De laatste drie voorbeelden maken duidelijk dat dualiseren niet eenduidig is.

VOORBEELD

$$p_1(y) = \inf_{x=(x_1,x_2)} \{(x_1-3)^2 + (x_2-4)^2 : x_1^2 + x_2^2 \leq 1 + y\}$$

$$p_2(y) = \inf_{x=(x_1,x_2)} \{(x_1-3)^2 + (x_2-4)^2 : (x_1+y_1)^2 + (x_2+y_2)^2 \leq 1\}$$

$$p_3(y) = \inf_{x=(x_1,x_2)} \{(x_1+y_1-3)^2 + (x_2-y_2-4)^2 : x_1^2 + x_2^2 \leq 1\}.$$

Steeds is het primaire probleem $\inf_x \{(x_1-3)^2 + (x_2-4)^2 : x_1^2 + x_2^2 \leq 1\}$, maar $\lambda \in \mathbb{R}$, $\lambda \in \mathbb{R}^2$ en $\lambda \in \mathbb{R}^2$, respektievelijk. Aan $p_3(y)$ is te zien dat het niet nodig is beperkingen te storen, maar dat dit ook mag geschieden met de doelstellingsfunktie (of dat in dit geval handig is, is een tweede).

VOORBEELD (KONVEX PROGRAMMEREN)

$F(x,y)$ is konvex in (x,y), $p(0) = \inf_x F(x,0)$, $p(y) = \inf_x F(x,y)$;

het duale probleem is

$\sup_\lambda \{\inf_{x,y} \{F(x,y) + \lambda y\}\}$. Speciale gevallen:

a) lineair programmeren

b) niet-lineair programmeren met f en g konvex

c) Fenchel dualiteit met f, G en g konvex, enz..

Onder enkele extra (redelijke) voorwaarden geldt dat het duale van het duale probleem weer het primaire probleem is. Zie een volgende paragraaf.

3. DE LAGRANGE FUNKTIE EN EEN SIMPEL RESULTAAT

De klassieke Lagrange funktie is $f(x) + \lambda g(x)$. De generalisatie daarvan is

$$L(x,\lambda) = \inf_y \{F(x,y) + \lambda y\}.$$

VOORBEELDEN

a) lineair programmeren

$$L(x,\lambda) = \begin{cases} cx + \lambda(b-Ax) & \text{als } x \geq 0 \text{ en } \lambda \geq 0 \\ +\infty & \text{als } x \ngeq 0 \\ -\infty & \text{als } x \geq 0 \text{ en } \lambda \ngeq 0. \end{cases}$$

b) niet-lineair programmeren

$$L(x,\lambda) = \begin{cases} f(x) + \lambda g(x) & \text{als } \lambda \geq 0 \\ -\infty & \text{zo niet} \end{cases}$$

c) Fenchel dualiteit, $L(x,\lambda) = f(x) - \lambda Ax + \inf_z \{\lambda z : z \in G\}$.

Aan de Lagrange funktie kan men goed zien wat er gebeurt bij de introduktie van de multiplikatoren. Bij a) in het laatste voorbeeld *verdwijnen* de beperkingen $Ax \geq b$, en komen de ingrediënten daarvan terecht in de doelstellingsfunktie. Bij b) gebeurt iets dergelijks. Bij c) verdwijnt de beperking weliswaar niet, maar er vindt een *ontkoppeling* plaats tussen doelstelling en beperking, want in het gedeelte $f(x) - \lambda Ax$ speelt G geen rol, en in het gedeelte $\inf_z \{\lambda z : z \in G\}$ speelt f geen rol. Voor het oplossen van optimaliseringsproblemen zijn deze aspekten van groot belang. Afhankelijk van het probleem zal men kiezen voor het "verdwijn"-effekt of het "ontkoppelings"-effekt, of een ander effekt, want er zijn vele variaties op het thema "dualiseren".

We komen nu weer terug op inf = sup = max. Laat $f^d(\lambda)$ de duale doelstellingsfunktie zijn, zodat

$$f^d(\lambda) = \inf_{x,y} \{F(x,y) + \lambda y\} = \inf_x L(x,\lambda)$$

waaruit volgt dat $f^d(\lambda) \in \mathbb{R}$ of $f^d(\lambda) = -\infty$ (en er dus toch niet te ontsnappen valt aan oneindige funktiewaarden; zie een eerdere opmerking).

STELLING 1.

a) $f^d(\lambda) \leq F(x,0)$ *voor alle* (x,λ) (*d.i. zwakke dualiteit*)

b) *als voor zekere* x^0 en λ^0 *geldt dat* $\inf_x L(x,\lambda^0) = F(x^0,0) \in \mathbb{R}$, *dan zijn* x^0 *en* λ^0 *beide optimaal, is* inf = min = sup = max (*d.i. sterke dualiteit*) *en is* $\inf_x L(x,\lambda^0) = L(x^0,\lambda^0)$.

BEWIJS

a) $f^d(\lambda) = \inf_{x,y} \{F(x,y) + \lambda y\} \leq \inf_x \{F(x,0) + \lambda.0\} \leq F(x,0)$.

b) Er volgt $f_d(\lambda^0) = F(x^0,0) \in \mathbb{R}$, dus x^0 en λ^0 zijn beide optimaal en min = max.

Daar $\inf_x L(x,\lambda^0) \leq L(x^0,\lambda^0) \leq F(x^0,0)$ geldt ook de laatste bewering.

OPMERKINGEN

1) Ook inf = sup of inf = sup = max wordt met sterke dualiteit aangegeven.

2) Kontinuïteit, konvexiteit en differentieerbaarheid spelen geen rol.

3) Er volgt niet dat als de onderstelling onder b) geldt dat dan x' optimaal is indien $\inf_x L(x,\lambda^0) = L(x',\lambda^0)$. Tegenvoorbeeld: lineair programmeren, $\lambda^0 \geq 0$, $\lambda^0 A \leq c$, zodat $L(x,\lambda^0) = cx + \lambda^0(b-Ax)$ als $x \geq 0$. Dan is $\inf_x L(x,\lambda^0) = L(0,\lambda^0)$, maar x' = 0 hoeft niet een optimale oplossing te zijn en kan zelfs ontoelaatbaar zijn.

4) Voor het praktisch oplossen stelle men dus:

$$\inf_x L(x,\lambda^0) = L(x^0,\lambda^0) = F(x^0,0) \in \mathbb{R}.$$

Lukt het dit stelsel op te lossen, dan zijn zowel het primaire als het duale probleem opgelost. Het stelsel zal echter veelal onoplosbaar zijn!

VOORBEELD

$p_1(y)$ als in §2, dus $\lambda^0 \geq 0$ en $L(x,\lambda^0) = (x_1-3)^2 + (x_2-4)^2 + \lambda^0(x_1^2 + x_2^2 - 1)$.

We krijgen $x_1^0 - 3 + \lambda^0 x_1^0 = 0$, $x_2^0 - 4 + \lambda^0 x_2^0 = 0$ en

$\lambda^0(\frac{25}{(1+\lambda^0)^2} - 1) = 0$, dus $\lambda^0 = 0$ of $\lambda^0 = 4$. Als $\lambda^0 = 0$ dan is $x^0 = (3;4)$, maar $F(x^0,0) = +\infty$ omdat $3^2 + 4^2 > 1$. Als $\lambda^0 = 4$ dan is $x^0 = (3/5;4/5)$ en $F(x^0,0) \in \mathbb{R}$. Dus $\lambda^0 = 4$ en $x^0 = (3/5;4/5)$ zijn optimaal.
Omdat $L(x,\lambda^0)$ strikt konvex is, is Opmerking 3) niet van toepassing, en is x' optimaal als $\inf_x L(x,\lambda^0) = L(x',\lambda^0)$, uiteraard als λ^0 optimaal is!
Strikte konvexiteit kan ook voor numerieke methoden een prettige eigenschap zijn. Reden waarom men soms lineair programmeren vervangt door "kwadratisch" programmeren, door (kleine) kwadratische termen in de doelstellingsfunktie toe te voegen.

4. GENERALISATIES

Eigenlijk zijn we al veel algemener bezig dan in de klassieke theorie van de Lagrange multiplikatoren, maar de tot zover gevolgde weg nodigt tot verdere verkenningen. Bovendien heeft Jacobi gezegd: "man muss immer generalisieren".
De eerste beperking die we willen laten vallen is dat $y \in \mathbb{R}^m$ en dat $\lambda y = \lambda_1 y_1 + \ldots + \lambda_m y_m$. We vervangen $\mathbb{R}^m$ door een willekeurige *topologische vektorruimte* Y, en nemen voor λ een *kontinue, lineaire funktionaal* op Y, zodat als Y^* de met Y gekonjugeerde ruimte is, $\lambda \in Y^*$. Aan het formalisme m.b.v. F(x,y) verandert dan niets en ook Stelling 1 blijft van kracht.

VOORBEELD

$\inf_x \{\int_0^1 c(t)x(t)dt : A(t)x(t) \geq b(t),\ 0 \leq t \leq 1\}$, met $x(t) \in \mathbb{R}^n$, $c(t) \in \mathbb{R}^n$, A(t) m bij n, en $b(t) \in \mathbb{R}^m$. De beslissingsvariabele is nu $x : t \mapsto x(t)$.
Uiteraard moeten x en $c : t \mapsto c(t)$ zodanig zijn dat de integraal bestaat.
Stel nu $y : t \mapsto y(t)$ en $\lambda : t \mapsto \lambda(t)$, en kies voor Y bijvoorbeeld $L_m^2([0,1])$, zodat ook $Y^* = L_m^2([0,1])$. Dan is $\lambda y = \int_0^1 \lambda(t)y(t)dt$.

VOORBEELD (OPTIMALE BESTURING)

$\inf_{x,u} \{\int_0^1 f(x(t),u(t),t)dt : \dot{x}(t) = g(x(t),u(t),t),\ x(0) = \xi,\ u(t) \in G,\ 0 \leq t \leq 1\}$. Hierin is niet x, maar (x,u) de beslissingsvariabele.
u is de besturing, x(t) de toestand ten tijde t, die wordt bepaald door de besturing ten tijde t en onmiddellijk daaraan voorafgaande tijdstippen.
Ook nu is het formalisme van toepassing bij geschikte keuze van X en Y.

VOORBEELD (ONEINDIGE HORIZON, DISKRETE TIJD)

Een minder vergaand voorbeeld is het volgende, dat in de wiskundige ekonomie van toepassing is:

$$F(x,y) = \begin{cases} \sum_{t=1}^{\infty} f_t(x_t) \text{ als } (x_{t-1}, x_t + y_t) \in G_t,\ t=1,2,\ldots, x_0 = \xi, \text{ reeks konvergent} \\ +\infty \text{ zo niet.} \end{cases}$$

Hierin is $x = (x_1, x_2, \ldots)$, $y = (y_1, y_2, \ldots)$. Voor Y zou men kunnen nemen de ruimte van alle absoluut konvergente rijen, l_1 dus. Dan is $Y^* = l_\infty$.

Voor een anderssoortige generalisatie keren we terug naar $Y = \mathbb{R}^m$ en naar

$$p(y) + \lambda^0 y \geq p(0).$$

De term $\lambda^0 y$ viel op te vatten als een korrektie-term, die diende om het effekt van het introduceren van $y \neq 0$ te neutraliseren. Waarom alleen lineaire korrektietermen gebruiken? Laten we voor λ^0 een willekeurige funktionaal kiezen, zodat we $\lambda^0 y$ moeten vervangen door $\lambda^0(y)$, waarbij we eisen dat $\lambda^0(0) = 0$. Ook dan, als we ook λy vervangen door $\lambda(y)$, verandert er vrijwel niets aan het formalisme, en blijft Stelling 1 van kracht. Er zit echter een bedenkelijke kant aan deze generalisatie. Stel eens dat onder de toegelaten $\lambda(y)$ ook $p(0) - p(y)$ voorkomt.

Dan is op triviale wijze voldaan aan de bovenstaande ongelijkheid en volgt dus inf = sup = max. Dat lijkt een mooi resultaat; het vergt echter dat we $\lambda(y) = p(0) - p(y)$ kunnen bepalen, en dat is in de praktijk niet eenvoudig. Bovendien zou met de bepaling van $p(0) - p(y)$ het probleem eigenlijk al zijn opgelost. Het grote voordeel van $\lambda_1 y_1 + \ldots + \lambda_m y_m$ is juist dat deze uitdrukking gemakkelijk is te bepalen, en slechts m parameters bevat. Hoe minder parameters hoe beter. Er is dus terughoudendheid geboden bij de introduktie van niet-lineaire $\lambda(y)$. Anderzijds geldt dat voor de meeste problemen niet kan worden voldaan aan de aanname onder b) van Stelling 1, en in elk geval het onder Opmerking 4) van §3 genoemde stelsel onoplosbaar is. Verruiming van de mogelijke $\lambda(y)$ kan dan aantrekkelijk zijn.

OORBEELD (GEHEELTALLIG PROGRAMMEREN)

eschouw $\sup_x \{2x_1 + 3x_2 : x_1 + 2x_2 \leq 4,\ 2x_1 + x_2 \leq 4,\ x \geq 0,\ x \text{ geheel}\}$,
dus, voor de verandering sup in plaats van inf). Het storen van de beide 4-en,
oor resp. $4 - y_1$ en $4 - y_2$ levert als duaal probleem
$\inf_\lambda \{4\lambda_1 + 4\lambda_2 : \lambda_1 + 2\lambda_2 \geq 2,\ 2\lambda_1 + \lambda_2 \geq 3,\ \lambda \geq 0\}$. Het sup is gelijk aan 6,
et inf is gelijk aan 20/3, en dus is sup < inf. We kunnen dit dualiteitsgat
lichten door niet te werken met $\lambda_1 y_1 + \lambda_2 y_2$, maar met $\lambda_1 y_1 + \lambda_2 y_2 + \lambda_3 \times$
$\lambda_4 y_1 + \lambda_5 y_5\rfloor$, waarin $\lfloor . \rfloor$ afronden naar beneden voorstelt.
let duale probleem wordt dan

$$\inf_{\lambda_1,\ldots,\lambda_5 \geq 0}\{\sup_x \{2x_1 + 3x_2 + \lambda_1(4-x_1-2x_2) + \lambda_2(4-2x_1-x_2) + \lambda_3\lfloor\lambda_4(4-x_1-2x_2) + \lambda_5(4-2x_1-x_2)\rfloor : x \geq 0,\ x \text{ geheel}\}\}.$$

olgens de zwakke dualiteit geldt sup ≤ inf. Zelfs geldt sup = inf, want neem
naar $\lambda_1 = 1$, $\lambda_2 = 0$, $\lambda_3 = 1$, $\lambda_4 = \lambda_5 = 1/3$, dan is de waarde van de duale doel-
stellingsfunktie gelijk aan $\sup_x \{6 : x \geq 0,\ x \text{ geheel}\} = 6$. De waarden der λ's
zijn min of meer met proberen bepaald, en het is niet duidelijk hoe men in het
algemeen met deze wijze van storen moet werken, al was het alleen maar omdat
we dan de waarde van het primaire supremum niet kennen. Doen we enig water in
de wijn dan blijkt het wel te lukken en komen we terecht bij de zogenoemde
fraktionele snede-methode van Gomory. Die blijkt meer λ's te gebruiken dan
strikt nodig is.

VOORBEELD (PENALTY METHODE)

Neem voor $\inf_x \{-x^4 : x^2 \leq 1\}$, $\lambda(y) = \lambda_1 y + \lambda_2 y^2$, en $p(y) = \inf_x \{-x^4 :$
$x^2 \leq 1 + y\}$. Dan wordt het duale probleem

$$\sup_\lambda \{\inf_{x,y} \{-x^4 + \lambda_1 y + \lambda_2 y^2 : x^2 \leq 1 + y\}\}.$$

Als $\lambda_2 \leq 0$ dan is het supremum gelijk aan $-\infty$. Dus neem $\lambda_2 > 0$. Het infimum
over y wordt aangenomen voor $y = \max(x^2-1;\ -\lambda_1/(2\lambda_2))$. Dus het duale probleem
wordt:

$$\sup_{\lambda_1,\lambda_2>0} \{\min [\inf_x \{-x^4 + \lambda_1(x^2-1) + \lambda_2(x^2-1)^4 : x^2 \geq 1 - \lambda_1/(2\lambda_2)\};\ \inf_x \{-x^4 - \lambda_1^2/(4\lambda_2) : x^2 \leq 1 - \lambda_1/(2\lambda_2)\}]\}.$$

Met behulp van numerieke methoden kan men de λ's bepalen. Voor $\lambda_1 = 2$
en $\lambda_2 = 1$ wordt min [...] = -1 en dat is juist het primaire infimum. De term

$\lambda_2 y^2$ is een penalty term.

In dit voorbeeld is de doelstellingsfunktie niet konvex. De penalty methode kan ook zinrijk zijn als dit wel het geval is en ook het toelaatbare gebied van het primaire probleem konvex is. Door het optreden van $\lambda_2 > 0$, kan men niet bij voorbaat zeggen dat $\lambda_1 \geq 0$.

VOORBEELD (BARRIER METHODE)
Beschouw hetzelfde probleem als bij de penalty methode. Maar laat $\lambda(y) = -1/(\lambda y)$, voor $\lambda > 0$. Het duale probleem wordt dan $\sup_{\lambda>0} \{\inf_x \{-x^4 - \lambda^{-1}(x^2-1)^{-1}\}\}$. Helaas wordt het supremum niet voor een eindige waarde van λ aangenomen. Uitwerken levert dat $x^2 = 1 + \delta$, met $\delta^2 = (2\lambda)^{-1} + \ldots$ en voor λ naderend naar $+\infty$ komt er $x^2 = 1$, zodat het supremu wel gelijk is aan het primaire minimum. Bij de numerieke uitwerking moet men λ geleidelijk naar $+\infty$ laten naderen. Iets dergelijks kan nodig zijn bij penal ty methoden, maar er zijn gelukkig ook zogenoemde "exakte" penalty methoden, waarbij het supremum voor een eindige waarde van de "penalty parameter" wordt aangenomen.

Zeer vele methoden kan men samen nemen door met $\lambda(y)$ in plaats van met λy te werken. Dat geldt in het bijzonder voor de meeste, zo niet alle, methoden met behulp van *gemodificeerde Lagrange funkties* (modified Lagrangeans, ook augmented Lagrangeans genoemd).

In het vervolg zullen we ons echter beperken tot lineaire $\lambda(y)$. Wel zullen we ook in het vervolg aandacht schenken aan oneindig dimensionale X en Y.

5. WANNEER IS inf = sup = max?

Als Stelling 1 van toepassing is, is er weinig meer te wensen over want dan hebben we blijkbaar zowel een optimale oplossing x^0 als een optimale oplossin λ^0 in handen. In de praktijk zal "meestal" niet aan de voorwaarden van Stelli 1 zijn voldaan, zodat er dus geen (x^0,λ^0) is te vinden zodanig dat

$$\inf_x L(x,\lambda^0) = F(x^0,0) \in \mathbb{R}.$$

Mogelijk blijkt het niet bestaan van (x^0,λ^0) na veel rekenwerk, werk dat achteraf bezien voor niets is geweest. Het ware gewenst inzicht te krijgen

in de omstandigheden waaronder we bij *voorbaat* zeker weten dat (x^0,λ^0) wèl bestaat. Dit voert tot een vrij omvangrijke theorie, waarvan we enkele aspekten in deze syllabus zullen toelichten.

We beginnen met een *meetkundig* beeld van zwakke en van sterke dualiteit, in beide gevallen onder de aanname dat sup = max, dus dat λ^0 bestaat. Neem eens aan dat

$$\inf_{x,y} \{F(x,y) + \lambda^0 y\} = \inf_y \{\inf_x F(x,y) + \lambda^0 y\} = \inf_y \{p(y) + \lambda^0 y\} = p(y^0) + \lambda^0 y^0$$

voor zekere y^0. Dan is voor alle (x,y)

$$F(x,y) + \lambda^0 y \geq p(y) + \lambda^0 y \geq p(y^0) + \lambda^0 y^0$$

en dus ligt elk punt $(y,\mu) \in Y \times \mathbb{R}$ niet "onder" het hypervlak

$$H = \{(y,\mu) : y \in Y,\ \mu \in \mathbb{R},\ \lambda^0 y + 1\cdot\mu = \lambda^0 y^0 + p(y^0)\}$$

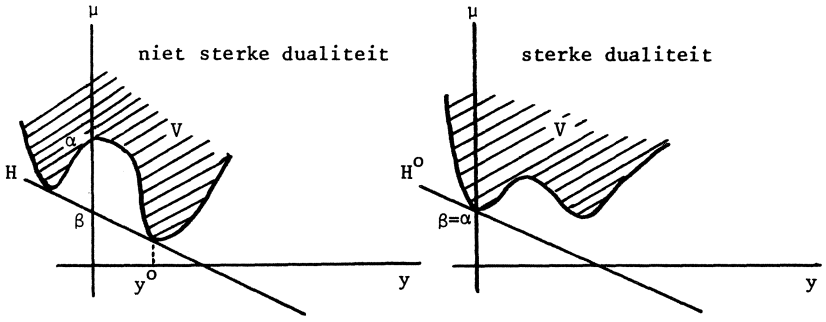
indien $\mu \geq F(x,y)$ voor een of andere x. Dit suggereert de volgende definitie:

$$V = \{(y,\mu) : y \in Y,\ \mu \in \mathbb{R},\ \mu \geq F(x,y) \quad \text{voor een } x \in X\}.$$

Dus: V ligt niet "onder" H. Merk op dat de koëfficiënt van μ in H gelijk is aan 1, en dus $\neq 0$, zodat H niet "vertikaal" staat! Sterke dualiteit wil zeggen dat $\inf_{x,y} \{F(x,y) + \lambda^0 y\} = p(0) = \alpha$, en dus dat er zo'n y^0 is, namelijk $y^0 = 0$. Daar hoort het hypervlak

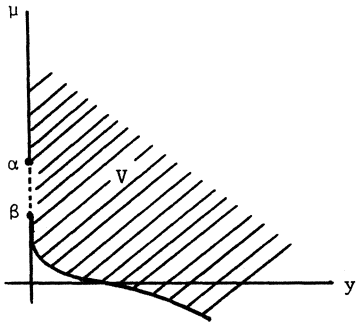
$$H^0 = \{(y,\mu) : y \in Y,\ \mu \in \mathbb{R}, \lambda^0 y + \mu = \alpha\}, \quad \text{waarin } \alpha = p(0) = \inf_x F(x,0)$$

bij, en V ligt niet "onder" H^0. Bovendien hebben V en H^0 een punt gemeen, en wel $(y,\mu) = (0,\alpha)$. Dit laatste geldt trouwens ook voor V en H, maar wel op grond van de onderstelling dat y^0 bestaat.



In de bovenstaande figuren is niet alleen $\alpha = p(0) = \inf_x F(x,0)$, maar ook $\beta = f^d(\lambda^0) = \inf_{x,y} \{F(x,y) + \lambda^0 y\}$ aangegeven. Bij niet-sterke dualiteit is $\beta < \alpha$, en bij sterke dualiteit is $\beta = \alpha$.

De vraag is nu: wanneer kunnen we rekenen op sterke dualiteit? De figuren suggereren het volgende antwoord: als er in elk rand-punt van V een steunend hypervlak is te bepalen, dat wil zeggen een hypervlak zodanig dat V geheel aan één kant ervan ligt en zodanig dat het door dat punt van V gaat. Uiteraard is deze voorwaarde niet noodzakelijk, want in de tweede figuur is er niet aan vol daan en toch heerst daar sterke dualiteit. Maar het lijkt moeilijk een algemene karaktereigenschap voor V aan te geven die alleen geldt voor het punt $(y,\mu) = (0,\alpha)$. Dus als je iets zegt over het ene punt, dan moet je tegelijk iets zeggen over alle andere punten van de rand van V. Deze voorwaarde is ekwievalent met de konvexiteit van V. Wanneer is V konvex? Als F als funktie van (x,y) konvex is! En daarmee is de belangstelling verklaard voor konvexe programmering (zie §2). Toch hebben we een foutje gemaakt, want het hypervlak H^0 mag niet-vertikaal staan. Staat H^0 namelijk vertikaal dan is er hoogstens sprake van sterke dualiteit in de zin van inf = sup, maar niet in de zin van inf = sup = max. In onderstaande figuur behoort het gestippelde lijnstuk niet tot V!



ok het punt (β,0) hoort niet tot V. De in de figuur geschetste situatie kan ich ook voordoen als V konvex is. Maar al is α = β dan nog bestaat λ^0 niet.

'OORBEELD

$(y) = \inf_x \{x : x^2 \leq y\}$. Gemakkelijk volgt dat $f^d(\lambda) = -1/(4\lambda)$ als $\lambda > 0$, en $^d(\lambda) = -\infty$ als $\lambda \leq 0$, zodat sup = 0 = inf, maar max bestaat niet.

let is vooral de mogelijkheid van het bestaan van vertikale steun-hypervlakken .n (0,α) die verantwoordelijk is voor veel theorie, want behalve konvexiteit ioet er nog iets extra worden ondersteld om de existentie van vertikale steun-ıypervlakken uit te sluiten. Eigenlijk doen we dan trouwens weer te veel, vant het gaat er niet zozeer om het bestaan van vertikale steun-hypervlakken ıit te sluiten, als wel om het bestaan van niet-vertikale steun-hypervlakken in (0,α) te garanderen. Want het is mogelijk dat beide typen tegelijk mogelijk zijn!

VOORBEELD

$p(y) = \inf_x \{x : x \leq y, x \geq 0\}$. Dan is V het niet-negatieve kwadrant in $Y \times \mathbb{R}$. Dit voorbeeld betreft het eenvoudigste van alle optimaliseringsproblemen onder nevenvoorwaarden: lineair programmeren!

Wil men, althans voor konvex programmeren, het onderste uit de kan, en dus een noodzakelijke en voldoende voorwaarde voor het bestaan van een niet-vertikaal hypervlak in (0,α), waarbij we natuurlijk ook nog aannemen dat α eindig is

(maar dat is geen ernstige beperking) dan komt men te staan voor een stelling van de volgende soort.

STELLING 2.

Als

a) Y *een lokaal-konvexe topologische vektorruimte is*

b) $\alpha = \inf_X F(x,0)$ *eindig is*

c) $V = \{(y,\mu) : (y,\mu) \in Y \times \mathbb{R},\ \mu \geq F(x,y)$ *voor een* $x \in X\}$ *konvex is*

d) T *is gedefinieerd door* $T = \{(y,0) : (y,0) \in Y \times \mathbb{R}\}$

e) K *de kegel is die door* V *wordt gegenereerd in het punt* $(0,\alpha)$, *dat wil zeggen dat*

$$K = \{k : k = \lambda v + (1-\lambda)(0,\alpha) \text{ voor een } \lambda \geq 0 \text{ en een } v \in V\},$$

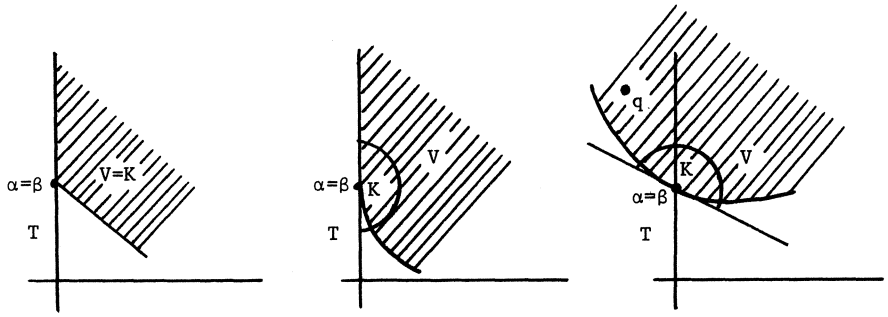
dan geldt $\alpha = \inf = \sup = \max = \beta$ *precies dan als* $\operatorname{cl} K \not\ni T$.

Het bewijs van deze stelling berust op de stelling van Hahn-Banach, waarvan het bewijs niet-konstruktief van aard is. Voor (eindig-dimensionaal) lineair programmeren en andere speciale gevallen zijn er alternatieve bewijzen, die wel konstruktief van aard zijn. In feite houdt de simplex-methode voor lineai programmeren zo'n bewijs in.

Voor de praktijk is de voorwaarde $\operatorname{cl} K \not\ni T$ meestal lastig direkt te verifiëren. Vandaar het bestaan van vele iets sterkere voorwaarden, bijvoorbeeld dat er een punt (y,μ) van V "links" van T (de μ-as) ligt. Preciezer betekent dit het volgende. Veronderstel dat er in Y een zogenaamde *niet-negatieve kegel* P is gedefinieerd, die niet-negativiteit in Y vastlegt, en wel als volg

$$y \geq 0 \text{ precies dan als } y \in P.$$

P dient een gesloten, konvexe kegel te zijn, met $0 \in Y$ als hoekpunt. Bovendie moet P "gepunt" zijn, dat wil zeggen $P \cup (-P) = \{0\}$. We kunnen nu aangeven wat bedoeld wordt met "links" van T: (y,μ) ligt "links" van T als $y \in \operatorname{int}(-P)$.



We kunnen de voorwaarden van Stelling 2 als volgt klassificeren:

a) eindigheid van $\inf_x F(x,0)$

b) konvexiteit van V

c) regulariteitsvoorwaarde cl K $\not\ni$ T.

Voorwaarde a) is redelijk, want $\inf_x F(x,0) = +\infty$ betekent dat het gegeven primaire probleem ontoelaatbaar is (beschouw bijvoorbeeld F(x,y) = f(x) als $g(x) \leq y$, $F(x,y) = +\infty$, zo niet), en $\inf_x F(x,0) = -\infty$ betekent dat er toelaatbare x zijn met een willekeurig negatief-grote waarde van F(x,0), en dat komt in de praktijk niet voor. Aan voorwaarde c) is veelal voldaan. Het lijkt er zelfs op alsof alleen "pathologische" gevallen er niet aan voldoen, zoals bijvoorbeeld $p(y) = \inf_x \{x : x^2 \leq y\}$.

Bij lineair programmeren met eindig veel beperkingen en eindig veel variabelen, dus met een eindige matrix A, is vanzelf aan c) voldaan.

Over blijft de voorwaarde b). Dit is de voorwaarde waarmee in de praktijk rekening dient te worden gehouden. Veelal is er niet aan voldaan, en is dat er de oorzaak van dat inf > sup of dat inf = sup maar dat het supremum niet als maximum wordt aangenomen.

VOORBEELD (NIET-LINEAIR PROGRAMMEREN)

$p(y) = \inf_x \{f(x) : g(x) \leq y, x \in C\}$. (De voorwaarde $x \in C$ hebben we in een eerder voorbeeld niet opgenomen.) Veronderstel dat f en g konvexe funkties zijn, en dat C een konvexe verzameling is. Dan is V konvex. Veronderstel dat Y een niet-negatieve kegel P heeft met niet-leeg inwendige (int P), en dat er

een $\hat{x}$ is met $g(\hat{x}) \in \text{int } (-P)$, $\hat{x} \in C$. Veronderstel tenslotte dat $p(0)$ eindig is. Dan is Stelling 2 van toepassing, en weten we dus bij voorbaat dat er een $\lambda^0 \in Y^*$ is, zdd $p(0) = \alpha = \inf_x \{f(x) : g(x) \leq 0, x \in C\} = \inf_{x,y} \{f(x) + \lambda^0 y : g(x) \leq y, x \in C\}$.

De niet-negatieve kegel P in Y induceert een niet-negatieve kegel in Y^*

$$P^* = \{\lambda : \lambda y \geq 0 \text{ voor alle } y \in P\}.$$

Met behulp van deze "duale" kegel kunnen we het resultaat van het zojuist gegeven voorbeeld vereenvoudigen, want het blijkt dat als $\lambda^0 \neq 0$ (dus als $\lambda^0 \notin P^*$) dat dan $f^d(\lambda) = -\infty$, zodat we de voorwaarde $\lambda^0 \geq 0$ kunnen opleggen. (Zoiets hebben we al eerder gedaan!) En dan komt er $p(0) = \inf_x \{f(x) + \lambda^0 g(x) : x \in C\}$, waarmee we de "vertrouwde" Lagrange-funktie $f(x) + \lambda^0 g(x)$ weer terug hebben.

Het is toegestaan P te laten ontaarden tot $P = \{0\}$. Dan betekent $y \geq 0$ in feite niets anders dan $y = 0$. Met andere woorden we kunnen gelijkheidsbeperkingen ook via een niet-negatieve kegel beschrijven. Sterker, we kunnen elke kombinatie van ongelijkheids- en gelijkheidsbeperkingen beschrijven met één enkele P. Wel lopen we daarbij het risiko dat int P leeg zal zijn (maar dan nog kan aan cl $K \not\supset T$ zijn voldaan).

VOORBEELD

Stel de voorwaarden zijn $2x_1^2 + 3x_2 \leq 4$ en $x_2 + x_3 = 6$. Definieer P in $\mathbb{R}^2$ als volgt. $P = \{(y_1, y_2) : y_1 \geq 0, y_2 = 0\}$. Definieer verder $g(x) = (2x_1^2 + 3x_2 - 4; x_2 + x_3 - 6)$. Dan is $g(x) \in -P$ precies dan als aan de voorwaarden is voldaan.

OPMERKING.

Uit de bovenstaande diskussie volgt dat ongelijkheden fundamenteler zijn dan gelijkheden, want de laatste verschijnen als bijzondere gevallen van de eerste Ook voor de wiskunde in het algemeen is het werken met kegels van groot belang. Beschouw bijvoorbeeld een lineaire, begrensde afbeelding A van X in Y, en beschouw de multifunktie T gedefinieerd door $T(x) = Ax + P$. Dan geldt onder bepaalde voorwaarden dat een open verzameling S in X onder T overgaat in

:en open verzameling in Y (generalisatie van de "open mapping" stelling).
leer aandacht voor deze onderwerpen in de wiskunde-opleiding lijkt gewenst.

EVOLG VAN STELLING 2 (EN OOK VAN STELLING 1) VOOR NIET-LINEAIR PROGRAMMEREN

ndien voor het probleem in het laatste voorbeeld inf = max, en indien boven-ien inf = min, zodat er x^0 en λ^0 zijn met $f(x^0) = f^d(\lambda^0) = \min_x$ $f(x) + \lambda^0 g(x) : x \in C\}$, dan volgt enerzijds, omdat $g(x^0) \leq 0$ en $\lambda^0 \geq 0$, dat $^0g(x^0) \leq 0$, en anderzijds dat $f(x^0) \leq f(x^0) + \lambda^0 g(x^0)$, want $x^0 \in C$, en dus .at $\lambda^0 g(x^0) \geq 0$. Ergo $\lambda^0 g(x^0) = 0$. Voor praktische berekeningen is dit een iiterst handige vergelijking, die als $Y = \mathbb{R}^m$ in feite bestaat uit m verge-.ijkingen $\lambda_i^0 g_i(x^0) = 0$, i = 1,...,m. Bij lineair programmeren worden dit de :ogenoemde *komplementaire spelingsrelaties*, bij optimale besturing de *trans-)ersaliteitsvoorwaarden*.

). TOEPASSINGEN VAN STELLING 2

1. KLEINSTE KWADRATEN ONDER NIET-NEGATIVITEITSVOORWAARDEN

)e variabele is nu niet x, maar (x,z); x stelt de helling voor van de te be-)alen rechte lijn en z stelt de konstante term voor. Gegeven zijn n meetpun-:en (a_i,b_i).

)PMERKING

In de statistiek noteert men dit geheel anders! De meetpunten zijn (x_i,y_i) en (x,z) = (α,β), waarin α en β de te schatten parameters zijn.
$p(y) = \min_{x,z} \{\Sigma_i (xa_i+z-b_i)^2 : x + y_1 \geq 0, z + y_2 \geq 0\}$. Toepassing van Stelling 2 geeft

$$2\Sigma a_i(x^0a_i+z^0-b_i) = \lambda_1^0, \quad 2\Sigma(x^0a_i+z^0-b_i) = \lambda_2^0,$$
$$(x^0,z^0,\lambda^0) \geq 0, \quad \lambda_1^0 x^0 = \lambda_2^0 z^0 = 0.$$

In de statistiek laat men de voorwaarde (x,z) ≥ 0 meestal weg, ook als die gewenst is! Men rekent dan op $(x^0,z^0) > 0$, in welk geval $\lambda^0 = 0$ en er voor (x^0,z^0) twee lineaire vergelijkingen overblijven. Hierdoor kan echter een strijdigheid ontstaan. In feite zijn er 4 mogelijkheden: $(x^0,z^0) = 0$, $\lambda^0 = 0$, $(x^0,\lambda_2^0) = 0$, $(z^0,\lambda_1^0) = 0$. Deze vier mogelijkheden dienen te worden uitgeprobeerd, of men dient een techniek voor "kwadratisch programmeren" toe te passen.

Afhankelijk van de ligging van de puntenwolk komt één van de 4 mogelijkheden voor de dag (afgezien van een speciaal geval. Welk?). Uiteraard zijn er op dit voorbeeld vele variaties. Allerlei andere beperkingen zijn denkbaar. Ook niet-lineariteit is denkbaar, voor zover het de beperkingen betreft. De doelstellingsfunktie die we hier hebben gekozen is het kwadraat van een Euclidisc norm, maar ook andere normen komen in aanmerking.

B. SIMPEL STOCHASTISCH PRODUKTIE-VOORRAAD PROBLEEM

x stelt de produktie van een artikel voor, $0 \leq x \leq m$. De beginvoorraad van da artikel is 0, de vraag naar dat artikel is verdeeld volgens een kontinue verdeling over $[0,\infty)$, met ϕ als dichtheidsfunktie en F als verdelingsfunktie. De produktie vindt plaats voordat de vraag zich realiseert. Produktiekosten zijn c per eenheid, overschotkosten (overschot = produktie-vraag indien ≥ 0) zijn h per eenheid, tekortkosten (tekort = vraag-produktie indien ≥ 0) zijn p per eenheid, $c > 0$, $h > 0$, $p > c$.

$p(y) = \inf_x \{cx + h\int_0^x (x-v)\phi(v)dv + p\int_x^\infty (v-x)\phi(v)dv : 0 \leq x + y_1,\ x \leq m + y_2\}$.

Dit is een konvex probleem, want de tweede afgeleide van de doelstellingsfunktie is $F'(x)$. Ook kan men de doelstellingsfunktie beschouwen als de integraal (som) van konvexe funkties, te weten $h[x-v]^+$ en $p[v-x]^+$, waarin $[z]^+ = \max(0$

Toepassing van Stelling 2 levert

$$c - p + (p+h)F(x^0) - \lambda_1^0 + \lambda_2^0 = 0,\ \lambda_1^0 x^0 = 0,\ \lambda_2^0(x^0-m) = 0,\ \lambda^0 \geq 0,\ 0 \leq x^0 \leq m.$$

$x^0 = 0$ geeft $p \leq c$, wat niet kan, dus $x^0 > 0$ en $\lambda_1^0 = 0$. Als $\lambda_2^0 = 0$, dan $F(x^0) = (p-c)/(p+h)$, en dus $x^0 = F^{-1}((p-c)/(p+h))$, maar dan moet $x^0 \leq m$. Klop dit niet, dan is $x^0 = m$.

VARIATIES: beginvoorraad ongelijk aan nul; meer dan één "periode", met in elk periode een produktie aan het begin van de periode en stochastische vraag lat in die periode; oneindig veel perioden. Andere methoden dan het direkt toepassen van Stelling 2 komen dan echter ook in aanmerking (met name dynamisch programmeren) en kunnen de voorkeur verdienen.

C. SIMPEL OPTIMAAL BESTURINGSPROBLEEM

Eigenlijk is dit meer een toepassing van Stelling 1, want het verifiëren van

een regulariteitsvoorwaarde zullen we achterwege laten. Ook gaan we niet in op de aard van de ruimte Y en zijn gekonjugeerde. De variabele is niet x maar (x,u), waarin x: $t \mapsto x(t) \in \mathbb{R}$, u: $t \mapsto u(t) \in \mathbb{R}$. Dus (x,u) is een 2-dimensionale vektorfunktie over [0,1]. x(t) is de hoek waarover de as van een motor draait, u(t) is de aan de motor toegevoerde electrische stroom, en $\ddot{x}(t)+\dot{x}(t) = u(t)$. In termen van optimale besturing is x de toestand, u de besturing. Het probleem is $\inf_{x,u} \{\int_0^1 u^2(t)dt : x(0) = \dot{x}(0) = 0, x(1) = 1, \dot{x}(1) = 0, \ddot{x}(t) + \dot{x}(t) = u(t), 0 \le t \le 1\}$. Ten tijde t = 0 staat de motor dus stil en is de hoek gelijk aan 0, ten tijde t = 1 moet de motor weer stil staan, maar moet de hoek gelijk zijn aan 1. Dit moet zo gebeuren dat de toegevoerde energie minimaal is. Kies als Lagrange-funktie:

$$L(x,u,\lambda) = \lambda_1 x(0) + \lambda_2 \dot{x}(0) + \lambda_3(x(1)-1) + \lambda_4 \dot{x}(1) + \int_0^1 \lambda_5(t)(\ddot{x}(t) + \dot{x}(t) - u(t))dt + \int_0^1 u^2(t)dt.$$

Dus $\lambda = (\lambda_1,\lambda_2,\lambda_3,\lambda_4,\lambda_5)$, waarin $(\lambda_1,\lambda_2,\lambda_3,\lambda_4) \in \mathbb{R}^4$ en $\lambda_5 : t \mapsto \lambda_5(t) \in \mathbb{R}$. Hierin moet λ_5 twee maal kontinu differentieerbaar zijn. Na partiële integratie waardoor $\ddot{x}(t)$ en $\dot{x}(t)$ dienen over te gaan in x(t) en (Fréchet) afgeleide nul stellen komt er:

$$2u(t) - \lambda_5(t) = 0$$

$$\ddot{\lambda}_5(t) - \dot{\lambda}_5(t) = 0$$

$$\lambda_1 - \lambda_5(0) + \dot{\lambda}_5(0) = 0$$

$$\lambda_2 - \lambda_5(0) = 0$$

$$\lambda_3 - \dot{\lambda}_5(1) + \lambda_5(1) = 0$$

$$\lambda_4 + \lambda_5(1) = 0.$$

Dus $\lambda_5(t) = a + be^t$, $u(t) = \frac{1}{2}at + \frac{1}{2}be^t + c$, $x(t) = \frac{1}{2}a^t - \frac{1}{2}a + \frac{1}{4}be^t + c + de^{-t}$ en begin- en eindvoorwaarden toepassen levert: $a = -2(e+1)/(e-3)$, $b = 4/(e-3)$, $c = -2/(e-3)$, $d = -e/(e-3)$, zodat $x(t)(3-e) = 1 - e + (e+1)t - e^t + e^{1-t}$ en $u(t)(3-e) = e + 1 - 2e^t$, dus $u(0) > 0$ en $u(1) < 0$.

(Dit voorbeeld is een variatie op een voorbeeld in Luenberger.)

7. HET WERKEN MET GEKONJUGEERDE FUNKTIES

In plaats van te werken met behulp van de verzamelingen V en T (zie boven), dus met behulp van een meer meetkundige aanpak, kunnen we de theorie ook presenteren met behulp van gekonjugeerde funkties. Als X een lokaal-konvexe topologische vectorruimte met X^* als gekonjugeerde is, en $f(x) \in \mathbb{R}$ of $f(x) = +\infty$, dan is

$$f^*(\zeta) = \sup_x \{\zeta x - f(x)\} \quad \text{voor } \zeta \in X^*.$$

OPMERKING

Vergelijk dit met transformaties zoals de Laplace-transformatie.

f^* heet de gekonjugeerde van f. Aangenomen dat $X^{**} = X$, is

$$f^{**}(x) = \sup_\zeta \{\zeta x - f^*(\zeta)\}.$$

Altijd geldt dat f^* konvex is, en dus dat f^{**} konvex is. Indien ook f konvex is, is dan $f^{**} = f$? Niet altijd! Het geldt wel indien f naar beneden begrensd is en als zijn epigraaf gesloten is. Algemener geldt het als f gesloten is, dat wil zeggen

a) $a(x) = \inf_\mu \{\mu : \mu \in \mathbb{R}, (x,\mu) \in \text{cl epi } f\} > -\infty$, en $f(x) = a(x)$ voor alle $x \in X$, óf

b) $f(x) = -\infty$ voor alle $x \in X$ en $a(x) = -\infty$ voor tenminste één $x \in X$.

Gemakkelijk volgt dat $\sup_\lambda \inf_{x,y} \{F(x,y) + \lambda y\} = p^{**}(0)$, zodat inf = sup precies dan als $p(0) = p^{**}(0)$. Dus, is p konvex en gesloten, dan is inf = sup.

Met behulp van gekonjugeerde funkties kan men een fraaie symmetrie in de dualiteitstheorie inbouwen. Laat namelijk

$$F^d(\lambda,\zeta) = \inf_{x,y} \{F(x,y) + \lambda y - \zeta x\}, \quad \lambda \in Y^*, \ \zeta \in X^*$$

(waarin het minus-teken ervoor dient om in speciale gevallen, en dat zijn de lineaire programmeringsproblemen, mooie resultaten te verkrijgen). Laat het gestoorde duale probleem zijn:

$$\sup_\lambda F^d(\lambda,\zeta) = p^d(\zeta).$$

Zij verder

$$F^{dd}(x,y) = \sup_{\lambda,\zeta} \{F^d(\lambda,\zeta) - \lambda y + \zeta x\}$$

(met andere tekens dan in F^d, omdat het duale probleem een konkaaf probleem is!). Gemakkelijk volgt nu dat

$$F^{dd} = F^{**}$$

waarbij de konjugatie moet geschieden met betrekking tot (x,y) en (λ,ζ). Neem daarbij aan niet alleen dat $X^{**} = X$, maar ook dat $Y^{**} = Y$. Dus, *het duale probleem van het duale probleem is de dubbel-gekonjugeerde van het primaire probleem.*

We laten het bij deze definities en opmerkingen. Zie o.a. Rockafellar.

8. DIFFERENTIEERBAARHEID

Hoewel het niet strikt nodig is, beperken we ons in deze paragraaf tot "Lagrange-dualiteit" oftewel "rechterlid-dualiteit":

$$p(y) = \inf_x \{f(x) : g(x) \leq y,\ x \in C\}.$$

Voorts nemen we $C = X$, zodat als inf = max, er een $\lambda^0 \geq 0$ is met

$$\alpha = \inf_x \{f(x) : g(x) \leq 0\} = \inf_x \{f(x) + \lambda^0 g(x)\}.$$

Zijn f en g differentieerbaar, en is x^0 een optimale oplossing, dan geldt

$$f'(x^0) + \lambda^0 g'(x^0) = 0$$

en samen met $f(x^0) = \alpha$, $\lambda^0 g(x^0) = 0$, $\lambda^0 \geq 0$, $g(x^0) \leq 0$ zijn dit de zogenoemde Kuhn-Tucker voorwaarden van het probleem. Zijn f en g konvex dan levert de oplossing van de KT-voorwaarden zowel een optimale oplossing x^0 voor het primaire als een optimale oplossing λ^0 voor het duale probleem.

Echter, ook al zijn f en g konvex, dan hoeven ze nog niet differentieerbaar te zijn en de vraag is: wat komt er dan in de plaats van $f'(x^0) + \lambda^0 g'(x^0) = 0$? Om deze vraag te kunnen beantwoorden, moeten we eerst het begrip differentieerbaarheid generaliseren. Zij X een lokaal-konvexe topologische vektorruimte, zij $f(x) \in \mathbb{R}$ of $f(x) = +\infty$, en zij $f(x')$ eindig. Dan is $\zeta \in X^*$ een subgradiënt van f in x' indien

$$f(x) - f(x') \geq \zeta(x-x') \text{ voor alle } x \in X.$$

Een subgradiënt hoeft niet uniek te zijn.

VOORBEELD

$f(x) = |x|$, $x' = 0$, dan $-1 \leq \zeta \leq 1$.

De verzameling van alle subgradiënten van f in x heet subdifferentiaal van f in x, en wordt aangegeven met $\partial f(x)$. Er geldt

$$x^0 \text{ is minimum van f (f konvex) precies dan als } 0 \in \partial f(x^0).$$

Deze inklusie komt dus in de plaats van het "afgeleide nul stellen". Dus voo f en g konvex, en $\lambda^0 \geq 0$, levert $\alpha = \inf_x \{f(x) + \lambda^0 g(x)\}$, dat

$$0 \in \partial(f(x^0) + \lambda^0 g(x^0)).$$

Helaas geldt dat $\partial(f_1+f_2)(x) \supset \partial f_1(x) + \partial f_2(x)$, maar onder niet al te veeleisende voorwaarden (zie Rockafellar) geldt dat $\partial(f_1+f_2)(x) = \partial f_1(x) + \partial f_2(x)$ en kunnen we dus verwachten dat

$$0 \in \partial f(x^0) + \lambda^0 \partial g(x^0).$$

Is $f(x) = |x|$, waarin $|x|$ een of andere norm is van $x \in X$, en is $|\zeta| = \sup_x \{|\zeta x| : |x| \leq 1\}$, $\zeta \in X^*$ dan geldt

$$\partial|x| = \{\zeta : |\zeta| \leq 1, \zeta x = |x|\}$$

en hiermee kunnen we "minimum-norm"-problemen aanpakken.

VOORBEELD

$\inf_x \{|x| : x = (x_1,x_2), 2x_1 + x_2 = 3\}$, $|x| = \max(|x_1|,|x_2|)$, zodat $|\zeta| = |\zeta_1| + |\zeta_2|$. Uitwerken geeft $0 = \zeta_1^0 + 2\lambda^0$, $0 = \zeta_2^0 + \lambda^0$, en dus $|\zeta_1^0| + |\zeta_2^0| \leq 1$, $\zeta_1^0 x_1^0 + \zeta_2^0 x_2^0 = \max(|x_1^0|,|x_2^0|)$, zodat $\max(|x_1^0|,|x_2^0|) \leq 1$ en $2x_1^0 + x_2^0 = 3$, en dat (natuurlijk) $x_1^0 = x_2^0 = 1$.

In een volgende paragraaf komen we terug op het generaliseren van differentieerbaarheid, en wel als we niet-konvexe problemen beschouwen, die niet in de klassieke zin differentieerbaar zijn.

9. NIET-KONVEXE PROBLEMEN DIE DIFFERENTIEERBAAR ZIJN IN KLASSIEKE ZIN

Als uitgangspunt kiezen we weer

$$p(y) = \inf_x \{f(x) : g(x) \leq y\}$$

waarin we f en g differentieerbaar onderstellen (in de klassieke zin), maar

niet konvex. Teneinde toch het toepassen van de dualiteitstheorie van de konvexe programmering mogelijk te maken, passen we linearisatie toe. Veronderstel dat x^0 een optimale oplossing is. Ontwikkel f en g als volgt

$$f(x) = f(x^0) + f'(x^0)(x-x^0) + \ldots \text{ en } g(x) = g(x^0) + g'(x^0)(x-x^0) + \ldots$$

en beschouw

$$p(y) = \inf_x \{f'(x^0)x : g(x^0) + g'(x^0)(x-x^0) \leq y\}$$

(zodat de eerstgenoemde p(y) vervalt!). Plezierig zou het zijn indien x^0 ook een optimale oplossing was van het gelineariseerde probleem. Om voorwaarden op te sporen waaronder dit het geval is, nemen we aan dat het niet het geval is en dat er een x' is met $g(x^0) + g'(x^0)(x'-x^0) \leq 0$ en $f'(x^0)(x') < f'(x^0)x^0$. De eerste ongelijkheid impliceert, daar $g(x^0) \leq 0$, dat, voor $0 \leq \tau \leq 1$, $g(x^0) + g'(x^0)\tau(x'-x^0) \leq 0$. Op grond van een zogenoemde approximatiestelling (verwant aan de bekende impliciete funktiestellingen) geldt nu dat $g(x^0+\tau(x'-x^0) + \mathcal{O}(\tau)) \leq 0$ voor een geschikte $\mathcal{O}(\tau)$. Deze approximatiestelling verlangt dat g' kontinu is in x^0, dat de niet-negatieve kegel P in Y gesloten is en dat $g'(x^0)X + P = Y$. Er volgt dan dus dat $x^0 + \tau(x'-x^0) + \mathcal{O}(\tau)$ een toelaatbare oplossing van het gegeven probleem is, en derhalve dat $f(x^0+\tau(x'-x^0) + \mathcal{O}(\tau)) \geq f(x^0)$, maar dit komt in konflikt met $f'(x^0)(x'-x^0) < 0$. Met andere woorden als g' kontinu is in x^0 en als $g'(x^0)X + P = Y$, met P gesloten, dan is x^0 ook een optimale oplossing van het gelineariseerde probleem. Daar het gelineariseerde probleem konvex is, kunnen we er Stelling 2 op toepassen, en onder de nodige onderstellingen geldt dan dat er een λ^0 moet zijn, zdd $\lambda^0 \geq 0$, en $f'(x^0) + \lambda^0 g'(x^0) = 0$. Voorts volgt weer dat $\lambda^0 g(x^0) = 0$.

OPMERKINGEN

a. Het bewijs van de approximatiestelling berust op nieuwere analytische resultaten, die verband houden met wat werd vermeld in de laatste opmerking van §5. Zie Robinson.

b. Met extra onderstellingen kan ook het probleem worden behandeld waarbij naast $g(x) \leq 0$, er nog een beperking is van het type $x \in C$; zie onder.

c. Wil men een probleem met gelijkheidsbeperkingen zo onderzoeken, dan hoeft men slechts P te laten ontaarden tot {0}, zodat $g'(x^0)X + P = Y$ overgaat in $g'(x^0)X = Y$, een bekende aanname in de minder recente theorie, die zegt dat $g'(x^0)$ een afbeelding is van X *op* Y.

Indien we een beperking van het type $x \in C$ opnemen, is het noodzakelijk dat ook C op een of andere wijze wordt "gelineariseerd". Het resultaat van een "gelineariseerde" C is een kegel! En wel een konvexe kegel. De top van de kegel ligt in het punt van C van waaruit de approximatie wordt uitgevoerd. Door de jaren heen zijn er allerlei soorten approximerende kegels geïntroduceerd. De meest belovende lijkt een kegel te zijn die door Clarke werd ingevoerd. Dit is de tangentiaal kegel van C in het punt $x \in C$, aangeduid door het symbool $T_C(x)$:

$$T_C(x) = \{v : v \in X, \forall x_i \in C, t_i > 0, x_i \to x, t_i \to 0, i = 1,2,\ldots,$$
$$\exists v_i \to v, \text{ zdd } x_i + t_i v_i \in C, \ i = 1,2,\ldots\}.$$

Het verrassende van deze definitie is o.a. dat $T_C(x)$ automatisch gesloten en konvex is. Het resultaat van de linearisatie van $\inf_x \{f(x) : g(x) \leq 0, x \in C\}$ wordt

$$\inf_x \{f'(x^0)x : g(x^0) + g'(x^0)(x-x^0) \leq 0, \ x \in T_C(x^0)\}$$

waarop we Stelling 2 kunnen toepassen. Details blijven echter achterwege.

OPMERKING

Opnieuw funktioneert het begrip kegel!

Voor de rest van dit verhaal is nog de normaalkegel van C in x van belang, aangeduid met $N_C(x)$:

$$N_C(x) = \{\zeta : \zeta \in X^*, \ \zeta v \leq 0 \text{ voor alle } v \in T_C(x)\} \text{ (let op: } \leq 0, \text{ niet } \geq 0).$$

Dus, terwijl $T_C(x)$ een deelverzameling is van X, is $N_C(x)$ een deelverzameling van X^*.

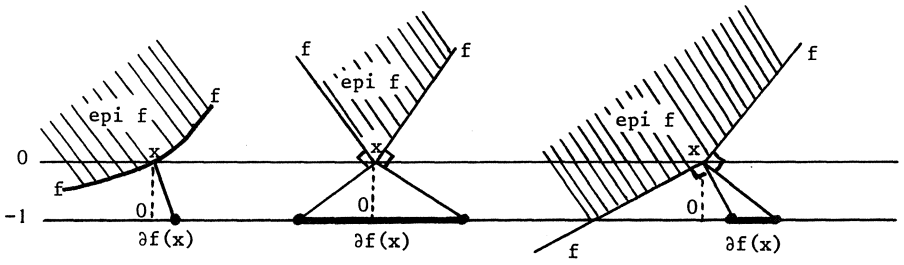
0. NIET-KONVEXE PROBLEMEN DIE NIET DIFFERENTIEERBAAR ZIJN IN KLASSIEKE ZIN

Wat valt er nog te doen indien zowel konvexiteit als differentieerbaarheid in de klassieke zin worden opgegeven? Een belangrijke observatie in dezen is dat als $f(x)$, $x \in \mathbb{R}$, gewoon differentieerbaar is, het punt $(-1,\zeta)$ ligt op de normaal op de kromme van f in x (naar beneden gericht), indien $\zeta = f'(x)$. Dit is een welbekend feit, waar als volgt door Clarke en Rockafellar gebruik van is gemaakt voor de volgende generalisatie van differentieerbaarheid (deze sluit de reeds behandelde generalisatie van differentieerbaarheid voor konvexe funkties in). Zij f gegeven, zodanig dat $f(x) \in \mathbb{R}$ of $f(x) = +\infty$. Kies een punt x waar f eindig is. Bepaal de normaalkegel in (x,μ) met $\mu = f(x)$, van de epigraaf van f, dus bepaal

$$N_{\text{epi } f}(x,f(x))$$

en definieer als gegeneraliseerde subdifferentiaal van f in x, de verzameling

$$\partial f(x) = \{\zeta : (-1,\zeta) \in N_{\text{epi } f}(x,f(x))\}.$$



De zwaar aangezette punten of lijnstukken geven $\partial f(x)$ aan. In het tweede geval is dit een interval links en rechts van 0, in het derde geval een interval rechts van 0. Merk op dat in het derde geval f konkaaf is.

Onder betrekkelijk weinig veeleisende voorwaarden (zie Rockafellar en Clarke), geldt nu dat, als x^0 een optimale oplossing is van $\inf_x \{f(x) : x \in C\}$,

$$0 \in \partial f(x^0) + N_C(x^0).$$

Met Lagrange-multiplikatoren heeft dit niet direkt te maken, maar beschouw nu

$$\inf_x \{f(x) : g(x) \leq y,\ x \in C\}$$

waarin de beperking $g(x) \leq 0$ wel, maar $x \in C$ niet is gestoord. Dan volgt, als x^0 optimaal is, en als geschikte voorwaarden gelden, dat er een $\lambda^0 \geq 0$ is zodanig dat

$$0 \in \partial f(x^0) + \lambda^0 \partial g(x^0) + N_C(x^0).$$

11. GEMENGDE PROGRAMMERING/MINIMUM PRINCIPES

Tot zover hebben we de nadruk gelegd òf op de konvexiteit van het gegeven probleem, òf op het lineariseren van dat probleem. In het eerste geval ging het om *globale* optimalisatie, dus het zoeken naar een x^0 zodanig dat $f(x^0) \leq f(x)$ voor elke toelaatbare x, al of niet dicht gelegen bij x^0, die natuurlijk zelf ook toelaatbaar moet zijn. In het tweede geval moesten we genoegen nemen met *lokale* optima, dat wil zeggen met x^0 zodanig dat $f(x^0) \leq f(x)$ voor alle x dicht genoeg gelegen bij x^0, want we hadden immers approximaties rondom x^0 beschouwd. In deze laatste paragraaf kombineren we beide aspekten. De beslissingsvariabele zal (x,u) zijn, niet x. T.o.v. u zullen we naar globaliteit streven, t.o.v. x zullen we genoegen nemen met lokaliteit. Dit houdt in dat we differentieerbaarheid t.o.v. x zullen moeten aannemen. Voor de eenvoud nemen we aan dat dit de gewone (Fréchet) differentieerbaarheid is. Generalisaties tot gegeneraliseerde differentieerbaarheid, zoals genoemd in §10 liggen dan voor de hand.

Laten we het volgende probleem beschouwen:

$$\inf_{x,u} \{f(x,u) : g(x,u) \leq 0,\ x \in C,\ u \in D\}.$$

Neem aan dat f en g (Fréchet) differentieerbaar zijn met betrekking tot x, en beschouw het volgende gelineariseerde probleem:

$$\inf_{x,u} \{f(x^0,u) + f'_x(x^0,u^0)(x-x^0) : g(x^0,u) + g'_x(x^0,u^0)(x-x^0) \leq 0,$$
$$x \in T_C(x^0),\ u \in D\}.$$

Zouden we u fixeren dan stond hier een konvex optimaliseringsprobleem, waarop we de boven besproken theorie konden toepassen. Om voor variërende u voldoende konvexiteit in het gelineariseerde probleem te brengen nemen we aan da

$$S = \{s : s = (s_1,s_2),\ s_1 \in \mathbb{R},\ s_2 \in Y,\ s_1 \geq f(x^0,u),\ s_2 \geq g(x^0,u) \text{ voor een } u \in D\}$$

een konvexe verzameling is. Deze S is geheel analoog aan de eerder ingevoerde V, waarvan we konvexiteit hebben geëist om inf = sup te kunnen garanderen. Onder enkele aanvullende regulariteitsvoorwaarden (o.a. dat $g'_x(x^0,u^0)X + P = Y$, vgl. §9), geldt dat als (x^0,u^0) een optimale oplossing is van het gegeven probleem, er een $\lambda^0 \geq 0$ is zodanig dat $\lambda^0 g(x^0,u^0) = 0$ en

$$f'_x(x^0,u^0)(x-x^0) + \lambda^0 g'_x(x^0,u^0)(x-x^0) \geq 0 \quad \text{als } x \in T_C(x^0),$$

$$f(x^0,u) + \lambda^0 g(x^0,u) \geq f(x^0,u^0) \qquad \text{als } u \in D.$$

De eerste ongelijkheid is een lokale voorwaarde, want x moet in de buurt van x^0 liggen (het feit dat $T_C(x^0)$ een kegel is helpt natuurlijk niet veel). De tweede ongelijkheid is echter een globale voorwaarde, omdat die moet gelden voor elke $u \in D$. Aan enkele voorbeelden zullen we nu laten zien dat we deze twee ongelijkheden samen als een abstrakt *minimum principe* mogen opvatten.

VOORBEELD

$\inf_{x,u} \{2x_1^2 - x_2 : x_1 = \int_0^1 u(t)dt,\ x_2 = \int_0^1 u^2(t)dt,\ 0 \leq u(t) \leq 1,\ 0 \leq t \leq 1\}$.

De verzameling S wordt

$$S = \{(s_1,s_2,s_3) : s_1 \geq 2x_1^2 - x_2,\ s_2 = x_1 - \int_0^1 u(t)dt,$$

$$s_3 = x_2 - \int_0^1 u^2(t)dt,\ \ 0 \leq u(t) \leq 1\}.$$

Aan te tonen is dat cl S konvex is. Voorts blijkt dat het voldoende is om de konvexiteit van cl S aan te nemen, in plaats van de konvexiteit van S zelf. Het minimum principe geeft:

$$4x_1^0 + \lambda_1^0 = 0, \quad -1 + \lambda_2^0 = 0, \quad \text{en}$$

$$2(x_1^0)^2 - x_2^0 + \lambda_1^0(x_1^0 - \int_0^1 u(t)dt) + \lambda_2^0(x_2^0 - \int_0^1 u^2(t)dt) \geq 2(x_1^0)^2 - x_2^0, \text{ en dus}$$

$$\int_0^1 (4x_1^0 u(t) - u^2(t))dt \geq \int_0^1 (4x_1^0 u^0(t) + (u^0(t))^2)dt \text{ voor alle } u,$$

$$0 \leq u(t) \leq 1.$$

Hieruit volgt dat $4x_1^0 u(t) - u^2(t) \geq 4x_1^0 u^0(t) + (u^0(t))^2$ voor alle t, $0 \leq t \leq 1$, en dit is een ongelijkheid die typisch is voor het minimum principe van Pontryagin. Omdat het linkerlid voor vaste t konkaaf is in u(t), volgt dat $u^0(t) = 0$ of $u^0(t) = 1$. Als ν de maat is van de t's waarvoor

$u^0(t) = 1$, dan volgt $x_1^0 = x_2^0 = \nu$, en $\inf_\nu \{2\nu^2 - \nu\}$ geeft $\nu = \frac{1}{4}$, zodat het gevraagde minimum gelijk is aan $-1/8$, hetgeen te realiseren is door bijvoorbeeld $u^0(t) = 1$ voor $0 \le t \le \frac{1}{4}$ en $u^0(t) = 0$ voor $\frac{1}{4} < t \le 1$ te nemen. De optimale oplossing voor u^0 is derhalve "bang-bang".

OPMERKING

Vervangen we dit voorbeeld door $\inf_{x,u} \{2x_1^2 - x_2 : x_1 = u,\ x_2 = u^2\}$, dan is S niet konvex en is het minimum principe niet van toepassing!

VOORBEELD

Optimale besturing. Zij $f(x,u) = f_1(x,u) + f_2(x,u)$, met

$$f_1(x,u) = \int_0^1 \phi_1(x(t),u(t),t)dt \text{ en } f_2(x,u) = \phi_2(x(1),u(1)).$$

Zij verder $g(x,u) = (g_1(x,u);\ g_2(x,u)) = 0$ met

$$g_1(x,u) = x(0) - \xi^0 \text{ en } g_2(x,u)(t) = -\dot{x}(t) + \psi(x(t),u(t),t).$$

Onder bepaalde voorwaarden geldt dat we λy als volgt kunnen kiezen:

$$\lambda y = \lambda_1 y_1 + \lambda_2 y_2 = \lambda_1 y_1 + \int_0^1 \lambda_2(t)y(t)dt$$

zodat

$$\begin{aligned}\lambda g(x,u) &= \lambda_1(x(0)-\xi^0) - \int_0^1 \lambda_2(t)\{\dot{x}(t) - \psi(x(t),u(t),t)\}dt = \\ &= \lambda_1(x(0)-\xi^0) - \lambda_2(1)x(1) + \lambda_2(0)x(0) + \int_0^1 \dot{\lambda}_2(t)x(t)dt + \\ &+ \int_0^1 \lambda_2(t)\psi(x(t),u(t),t)dt.\end{aligned}$$

De eerste ongelijkheid van het minimum principe geeft dan:

$$\int_0^1 \phi_1'(x^0(t),u^0(t),t)h(t)dt + \phi_2'(x^0(1),u^0(1))h(1) + \lambda_1^0 h(0) - \lambda_2^0(1)h(1) + \lambda_2^0(0)h(0) +$$

$$+ \int_0^1 \{\dot{\lambda}_2^0(t) + \lambda_2^0(t)\psi'(x^0(t),u^0(t),t)\}h(t)\,dt = 0 \text{ voor alle } h, \text{ zodat}$$

$$\lambda_1^0 + \lambda_2^0(0) = 0,\quad \lambda_2^0(1) = \phi_2'(x^0(1),u^0(1)) \text{ en}$$

$$\phi_1'(x^0(t),u^0(t),t) + \lambda_2^0(t)\psi'(x^0(t),u^0(t),t) + \dot{\lambda}_2^0(t) = 0$$

(differentiaties alle naar x), of als we de Hamilton funktie H invoeren

$$H(\xi,\mu,t,\omega) = \phi_1(\xi,\mu,t) + \omega\psi(\xi,\mu,t),$$

$$H'(x^0(t),u^0(t),\lambda_2^0(t)) + \dot{\lambda}_2^0(t) = 0$$

hetgeen een differentiaalvergelijking is voor de geadjungeerde variabele λ_2^0 met $\lambda_2^0(1) = \phi_2'(x^0(1),u^0(1))$ als eindvoorwaarde.
De tweede ongelijkheid van het minimum principe geeft

$$\int_0^1 H(x^0(t),u(t),t,\lambda_2^0(t))dt + \phi_2(x^0(1),u(1)) \geq$$
$$\int_0^1 H(x^0(t),u^0(t),t,\lambda_2^0(t))dt + \phi_2(x^0(1),u^0(1))$$

waaruit volgt dat

$$H(x^0(t),u(t),t,\lambda_2^0(t)) \geq H(x^0(t),u^0(t),t,\lambda_2^0(t)) \text{ voor alle } t$$

hetgeen weer Pontryagin's minimum principe is.

We zien uit deze voorbeelden dat het abstrakte minimum principe overeenkomt met een geïntegreerde vorm van het minimum principe van Pontryagin, althans als integraties een rol spelen. Om te laten zien dat het abstrakte minimum principe van de gemengde optimalisering algemener is dan Pontryagin's minimum principe, tenslotte nog een voorbeeld.

VOORBEELD

Zogenoemd gegeneraliseerd lineair programmeren. De variabele is niet (x,u), maar (x,u,v); het probleem is

$$\inf_{x,u,v} \{\Sigma u_i x_i : \Sigma v_i x_i = b,\ x_i \geq 0,\ (u_i,v_i) \in D_i,\ i = 1,\ldots,n\}.$$

Hierin zijn $x_i \in \mathbb{R}$, $u_i \in \mathbb{R}$, $v_i \in \mathbb{R}^m$, $b \in \mathbb{R}^m$ en (u,v) komt in de plaats van u. Als alle D_i konvex zijn, dan volgt dat er een λ^0 is zodanig dat $u_i^0 - \lambda^0 v_i^0 \geq 0$ en $\Sigma u_i x_i^0 + \lambda^0(b-\Sigma v_i x_i^0) \geq \Sigma u_i^0 x_i^0$ voor alle (u,v), zodanig dat $(u_i,v_i) \in D_i$. Als bijvoorbeeld alle D_i polyhedra zijn, dan zijn er speciale technieken om dit stelsel op te lossen, zie bijvoorbeeld Lasdon.

LITERATUUR

1. Clarke, F.H., Optimization and non-smooth analysis, Wiley, 1983.
2. Lasdon, L.S., Optimization theory for large systems, Macmillan, 1970.
3. Luenberger, D.G., Optimization by vector space methods, Wiley, 1969.
4. Ponstein, J., Approaches to the theory of optimization, Cambridge University Press, 1980.
5. Rockafellar, R.T., Convex analysis, Princeton University Press, 1970.
6. Rockafellar, R.T., The theory of subgradients and its applications to problems of optimization; Convex and nonconvex functions, Heldermann, 1981.
7. Robinson, S.M., Stability theory for systems of inequalities, Part II: differentiable non-linear systems, SIAM J. Num. Anal. 13(1976) pp. 497-513.
8. Robinson, S.M., Normed convex processes, Trans. Am. Math. Soc. 174(1972) pp. 127-140.
9. Robinson, S.M., An inverse function theorem for a class of multivalued functions, Proc. Am. Math. Soc. 41(1973) pp. 211-218.
10. Ioffe, A.D. & V.M. Tihomirov, Theory of extremal problems, North-Holland, 1979.

HOOFDSTUK 7

VARIATIONELE ONGELIJKHEDEN MET TOEPASSINGEN OP HET OBSTAKEL- EN MEMBRAAN PROBLEEM

C. CUVELIER

INLEIDING

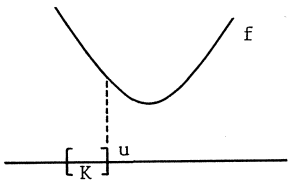
In dit hoofdstuk zullen we twee fysische problemen beschouwen, die mathematisch geformuleerd kunnen worden als het minimaliseren van een functionaal met nevenvoorwaarden. De twee fysische problemen zijn het obstakelprobleem en het membraanprobleem. Bij het obstakel probleem wordt, onder invloed van eventuele externe krachten, een snaar gespannen tussen twee punten, waartussen zich een obstakel bevindt. De vraagstelling is: wat is de uitwijking van de snaar. Bij het tweede probleem, het membraan probleem, probeert men in een staaf de temperatuurverdeling te bepalen, waarbij de warmtestroom wel van buiten naar binnen, maar niet van binnen naar buiten gericht kan zijn. De wand van het gebied heet semi-permeabel. We zullen eerst een oplossingsmethode geven voor het bepalen van het minimum van een convexe functie van $\mathbb{R}$ naar $\mathbb{R}$ met nevenvoorwaarden. De geprojecteerde gradiënt methode zal geïntroduceerd worden als een numerieke methode voor het daadwerkelijk berekenen van dit minimum. Vervolgens zullen we het obstakel- en het membraan-probleem formuleren in termen van een te optimaliseren functionaal met nevenvoorwaarden. Voor deze optimaliseringsproblemen zullen we equivalente formuleringen geven, die dan geïnterpreteerd kunnen worden als zwakke formuleringen (zie Hoofdstuk 5) van (part.) diff. vgln met randvoorwaarden. Uit de theorie van Hoofdstuk 6 volgt dat dit soort problemen equivalent is met het bepalen van het zadelpunt van een Lagrangiaan. Deze laatste formulering geeft aanleiding tot een geprojecteerde gradiënt methode, dat wil zeggen een methode voor het iteratief bepalen van het zadelpunt van de Lagrangiaan. Tenslotte zullen we met enige numerieke (eindige elementen) berekeningen de gevolgde werkwijze illustreren.

1. MINIMALISEREN VAN EEN CONVEXE FUNCTIE MET NEVENVOORWAARDEN

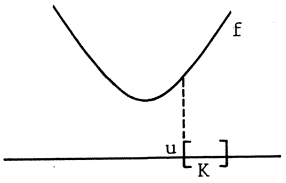
Zij $f : \mathbb{R} \to \mathbb{R}$ een convexe differentieerbare functie van $\mathbb{R}$ naar $\mathbb{R}$ en zij K een gesloten interval op $\mathbb{R}$. Beschouw het volgende probleem: Bepaal het minimum van f op K. Dat wil zeggen:

$$(1.1) \qquad \begin{cases} \text{Zoek } u \in K & \text{zodanig dat} \\ f(u) \leq f(v) & \text{voor alle } v \in K. \end{cases}$$

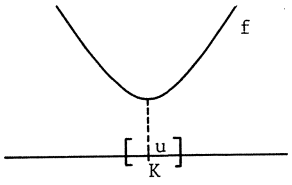
Er kunnen zich drie gevallen voordoen die zijn weergegeven in Fig. 1.



Figuur 1.a.



Figuur 1.b.



Figuur 1.c.

In Figuur 1a wordt het minimum van f op K aangenomen in het punt u. Verder geldt dat $f'(u) < 0$ en dat $v - u \leq 0$ voor alle $v \in K$, zodat

$$(1.2) \qquad u \in K, \quad f'(u)\cdot(v-u) \geq 0 \quad \text{voor alle} \quad v \in K.$$

In Figuur 1b geldt voor de oplossing u: $f'(u) > 0$ en $v - u \geq 0$ voor alle $v \in K$, hetgeen ook leidt tot (1.2).
In Figuur 1c is $f'(u) = 0$ zodat ook hier (1.2) geldig is. We kunnen dus zeggen dat de oplossing u van (1.1) gekarakteriseerd wordt door het volgende probleem:

$$(1.3) \qquad \begin{cases} \text{Zoek} \quad u \in K \quad \text{zodanig dat} \\ f'(u).(v-u) \geq 0 \quad \text{voor alle } v \in K. \end{cases}$$

Probleem (1.3) heet een variationele ongelijkheid (VO). Laten we eens een paar speciale gevallen bekijken.

(1.4) VOORBEELD: $K = \mathbb{R}$.

In dit geval reduceert het probleem zich tot het minimaliseren zonder nevenvoorwaarden. De karakterisering (1.3) wordt geschreven als:

$$(1.5) \qquad \begin{cases} \text{Zoek} \quad u \in \mathbb{R} \quad \text{zodanig dat} \\ f'(u).(v-u) \geq 0 \quad \text{voor alle} \quad v \in \mathbb{R}. \end{cases}$$

Kiezen we $v = u + w$ dan geldt

(1.6) $$\begin{cases} \text{Zoek } u \in \mathbb{R} \text{ zodanig dat} \\ f'(u).w \geq 0 \text{ voor alle } w \in \mathbb{R}. \end{cases}$$

Daar w zowel positief als negatief kan zijn volgt onmiddellijk dat $f'(u) = 0$, hetgeen ook inderdaad de oplossing karakteriseert voor $K = \mathbb{R}$.

(1.7) VOORBEELD: $K = \{v \in \mathbb{R} \mid v \geq a\}$.

De karakterisering van de oplossing (1.3) is:

(1.8) $$\begin{cases} \text{Zoek } u \geq a \text{ zodanig dat} \\ f'(u).(v-u) \geq 0 \text{ voor alle } v \geq a. \end{cases}$$

Kiezen we achtereenvolgens $v = a \in K$ en $v = 2u - a \in K$ dan vinden we dat tegelijk moet gelden

$$f'(u).(a-u) \geq 0 \quad \text{en} \quad f'(u).(a-u) \leq 0$$

zodat

(1.9) $$f'(u).(a-u) = 0.$$

Kiezen we $v = u + w$ met $w \geq 0$ (dan geldt zeker $v \geq a$) dan krijgen we

$$f'(u).w \geq 0 \quad \text{voor alle } w \geq 0$$

waaruit volgt

(1.10) $$f'(u) \geq 0.$$

Als de oplossing u strikt groter is dan a: $u > a$, kies dan een willekeurige v met $a \leq v \leq 2u + a$, dan neemt $v-u$ alle waarden aan tussen $a-u < 0$ en $a + u > 0$. Uit (1.8) volgt dan dat $f'(u) = 0$.

Resumerend kunnen we stellen dat

(1.11) $$\begin{cases} u \geq a \quad \text{dan } f'(u) \geq 0. \\ \text{Als } u > a \quad \text{dan geldt } f'(u) = 0. \end{cases}$$

Een iteratieve methode voor het bepalen van de oplossing u van (1.1) is de gradiënt methode met projectie. Bekijk eerst eens het geval (1.4) met $K = \mathbb{R}$. De oplossing u is dan de limiet van de rij $\{u_n\}_n$ die bepaald wordt door

(1.12) $$u_{n+1} = u_n - \rho f'(u_n), \quad n \geq 0$$

waarbij ρ een positieve parameter is en u_0 een startwaarde. Zij I een omgeving van u, dan kan bewezen worden dat de rij $\{u_n\}_n$ voor iedere startwaarde $u_0 \in I$ convergeert naar u als $0 < \rho \leq \frac{2}{M_f}$ met $M_f = \max_{v \in I} f''(v)$.

In het geval dat K = [a,b], met eventueel $a = -\infty$, $b = \infty$, kan de oplossing u bepaald worden met de geprojecteerde gradiënt methode. De oplossing u is dan de limiet van de rij $\{u_n\}_n$ bepaald door

$$(1.13) \qquad \begin{cases} u_{n+\frac{1}{2}} = u_n - \rho\, f'(u_n) \\ u_{n+1} = \mathrm{Proj}_K\,[u_{n+\frac{1}{2}}]. \end{cases}$$

In twee stappen wordt, uitgaande van u_n, de nieuwe benadering u_{n+1} bepaald. In stap 1 wordt de gradiënt methode toegepast; in stap 2 wordt $u_{n+\frac{1}{2}}$ geprojecteerd op K. Dit houdt in dat

$$u_{n+1} = \begin{cases} a & \text{als } u_{n+\frac{1}{2}} < a \\ u_{n+\frac{1}{2}} & \text{als } a \leq u_{n+\frac{1}{2}} \leq b \\ b & \text{als } u_{n+\frac{1}{2}} > b. \end{cases}$$

Er kan bewezen worden dat (1.13) convergeert als $0 < \rho \leq \frac{2}{M_f}$.

2. TWEE EQUIVALENTE FORMULERINGEN

In deze paragraaf beschouwen we de algemene gedaante van een minimaliseringsprobleem met nevenvoorwaarden. Zonder bewijs geven we een resultaat betreffende existentie en eenduidigheid van een oplossing. De oplossing van het probleem kan gekarakteriseerd worden met behulp van een VO. We zullen bewijzen dat, onder zekere voorwaarden, het oorspronkelijke minimaliseringsprobleem equivalent is met de VO. We zullen ons beperken tot één-dimensionale problemen.

Zij V een functieruimte bestaande uit functies $v : [0,1] \to \mathbb{R}$ en zij K een niet-lege, gesloten, convexe deelverzameling in V. Zij vervolgens $J : V \to \mathbb{R}$ een functionaal op V. We veronderstellen dat J de volgende gedaante heeft:

$$(2.1) \qquad J(v) = \tfrac{1}{2} \int_0^1 \left(\frac{dv}{dx}\right)^2 dx - \int_0^1 fv\, dx$$

met $f: \mathbb{R} \to \mathbb{R}$. We formuleren het probleem als volgt:

$$(2.2)\qquad \begin{cases} \text{Zoek } u \in K \text{ zodanig dat} \\ J(u) = \inf_{v\in K} J(v). \end{cases}$$

Zoals we in de voorbeelden zullen zien, kunnen de nevenvoorwaarden opgenomen worden in de definitie van de verzameling K. We kunnen bewijzen dat probleem (2.2) een eenduidige oplossing u heeft. Deze oplossing u wordt gekarakteriseerd door het volgende probleem:

$$(2.3)\qquad \begin{cases} \text{Zoek } u \in K \text{ zodanig dat} \\ \int_0^1 \frac{du}{dx}\ \frac{d(v-u)}{dx}\,dx \geq \int_0^1 f(v-u)\,dx \quad \text{voor alle } v \in K. \end{cases}$$

In de volgende stelling bewijzen we

(2.4) STELLING. De problemen (2.2) en (2.3) zijn equivalent.

BEWIJS.

(i) Zij u een oplossing van (2.2) dan geldt voor alle $v \in K$ en voor alle $\lambda \in (0,1)$ dat

$$J(u) \leq J((1-\lambda)u + \lambda v)$$

oftewel

$$\frac{1}{\lambda}\,[J(u + \lambda(v-u)) - J(u)] \geq 0;$$

immers, het minimum van J wordt aangenomen in u. Substitueren we (2.1) in deze ongelijkheid dan volgt:

$$\frac{1}{\lambda}\left[\tfrac{1}{2}\int_0^1 \left(\frac{d(u+\lambda(v-u))}{dx}\right)^2 dx - \int_0^1 f(u+\lambda(v-u))\,dx - \tfrac{1}{2}\int_0^1 \left(\frac{du}{dx}\right)^2 dx + \int_0^1 f\,u\,dx\right] \geq 0.$$

Laten we $\lambda \downarrow 0$ dan krijgen we

$$\int_0^1 \frac{du}{dx}\ \frac{d(v-u)}{dx}\,dx \geq \int_0^1 f(v-u)\,dx \quad \text{voor alle } v \in K.$$

Dus u is ook oplossing van (2.3).

(ii) Zij u een oplossing van (2.3). We merken op dat de afbeelding $v \to J(v)$ convex is, hetgeen inhoudt dat

$$J((1-\lambda)w+\lambda v) \leq (1-\lambda)J(w) + \lambda J(v) \quad \text{voor alle } v,w \in K,\ \lambda \in (0,1)$$

oftewel

$$(2.4) \qquad J(v) - J(u) \geq \frac{1}{\lambda} \left[J((1-\lambda) w + \lambda v) - J(w) \right] .$$

Substitueren we de uitdrukking (2.1) in het rechterlid van (2.4) en laten we $\lambda \downarrow 0$ gaan, dan geldt

$$J(v) - J(w) \geq \int_0^1 \frac{dw}{dx} \frac{d(v-w)}{dx} dx - \int_0^1 f(v-w) dx \quad \text{voor alle} \quad v,w \in K.$$

Kiezen we $w = u \in K$, dan geldt, omdat u voldoet aan (2.3),

$$J(v) - J(u) \geq \int_0^1 \frac{du}{dx} \frac{d(v-u)}{dx} dx - \int_0^1 f(v-u) dx \geq 0 \text{ voor alle } v \in K$$

waarmee we bewezen hebben dat u ook oplossing is van (2.2). □

Ongelijkheid (2.3) heet een variationele ongelijkheid (VO).

De uitdrukking

$$(2.5) \qquad \lim_{\lambda \downarrow 0} \frac{1}{\lambda} \left[J(w + \lambda (v-w)) - J(w) \right]$$

is de Gateaux - afgeleide van de functionaal J in het punt w in de richting v-w en wordt genoteerd als

$$(2.6) \qquad J'(w) [v-w].$$

De VO (2.3) kan nu geschreven worden als:

$$(2.7) \qquad \begin{cases} \text{Zoek } u \in K \text{ zodanig dat} \\ J'(u) [v-u] \geq 0. \end{cases}$$

De analogie met het probleem uit paragraaf 1 is nu duidelijk. Probleem (1.1) komt overeen met (2.2) en formulering (1.3) komt overeen met (2.7). In de toepassingen van de paragrafen 3 en 4 zullen we steeds, uitgaande van een minimaliseringsprobleem met nevenvoorwaarden, de VO afleiden. Vervolgens zullen we deze VO interpreteren als de zwakke formulering van een diff. vgl. met randvoorwaarden.

Het zou voor de hand liggen om, voor het oplossen van het minimaliseringsprobleem met nevenvoorwaarden, de geprojecteerde gradiënt methode toe te passen op de VO (2.7). Dit zou leiden tot een iteratieve methode van het volgende type:

$$(2.8)\qquad \begin{cases} \text{Kies } u_0 \in K \text{ . Bepaal } \{u_n\}_n \text{ met} \\ u_{n+\frac{1}{2}} = u_n - \rho\, J'(u_n)\,, \quad u_{n+\frac{1}{2}} \in V \\ u_{n+1} = \mathrm{Proj}_K\,[u_{n+\frac{1}{2}}]. \end{cases}$$

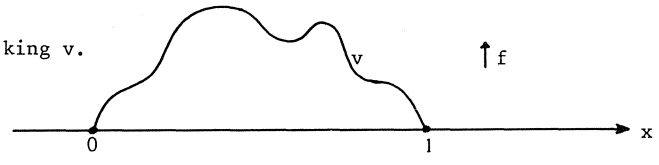
De $u_{n+\frac{1}{2}}$ kan bepaald worden, eventueel met behulp van de EEM (zie Hoofdstuk 5). Echter, de tweede stap, dus het bepalen van de projectie, is een gecompliceerde operatie omdat hier de afgeleiden van de functies in de ruimte V een rol spelen. Het zou ons te ver voeren hier nader op in te gaan; er wordt verwezen naar de literatuur. We zullen in de paragrafen 5 en 7 streven naar een herformulering van het probleem waarop de geprojecteerde gradiënt methode toepasbaar is en waarbij geprojecteerd wordt in een functieruimte waarin de afgeleiden geen rol spelen.

3. OBSTAKEL PROBLEEEM

Bij dit probleem spannen we een elastische snaar tussen twee punten, zeg x = 0 en x = 1. De positie die de snaar zal innemen is die welke behoort bij een minimale potentiële energie van de snaar. Stel dat $f:[0,1] \to \mathbb{R}$ een externe kracht (bijvoorbeeld zwaartekracht) is die loodrecht op de ongestoorde toestand van de snaar werkt, dan heeft een willekeurige uitwijking $v : [0,1] \to \mathbb{R}$ van de snaar (zie Fig. 2) een potentiële energie

$$(3.1)\qquad J(v) = \tfrac{1}{2} \int_0^1 \left(\frac{dv}{dx}\right)^2 dx - \int_0^1 fv\, dx.$$

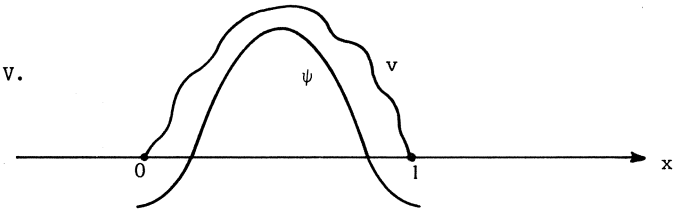
Figuur 2. Willekeurige uitwijking v.



De snaar is ingeklemd in x = 0 en x = 1 en dus geldt v(0) = v(1) = 0. De functieruimte V is nu gedefinieerd als de ruimte van functies v, waarvoor J(v) eindig is en waarvoor geldt v(0) = v(1) = 0. De positie die de snaar wil innemen wordt nu gehinderd door een obstakel $\psi : [0,1] \to \mathbb{R}$,

dat voldoet aan $\psi(0) < 0$, $\psi(1) < 0$ (zie Fig. 3).

Figuur 3. Obstakel ψ en $v \in V$.



Aan de uitwijking v leggen we nu de nevenvoorwaarde op dat

$$v \geq \psi \quad \text{op} \quad [0,1].$$

We zoeken de oplossing dus in de deelverzameling K van V die als volgt gedefinieerd is:

$$K = \{ v : [0,1] \to \mathbb{R} \mid v(0) = v(1) = 0,\ v \geq \psi \ \text{ op } [0,1]\}. \tag{3.2}$$

Het obstakel-probleem wordt nu geformuleerd als:

$$\begin{cases} \text{Zoek } u \in K \quad \text{zodanig dat} \\ J(u) = \inf\limits_{v \in K} J(v). \end{cases} \tag{3.3}$$

Volgens stelling (2.4) kan dit probleem equivalent geformuleerd worden als:

$$\begin{cases} \text{Zoek } u \in K \quad \text{zodanig dat} \\ \int\limits_0^1 \frac{du}{dx}\ \frac{d(v-u)}{dx}\, dx \geq \int\limits_0^1 f(v-u)\, dx \quad \text{voor alle } v \in K. \end{cases} \tag{3.4}$$

De interpretatie van dit probleem is gecompliceerd; we kunnen het dan ook niet streng behandelen. Formeel kunnen we echter als volgt te werk gaan. Het interval $\Omega = (0,1)$ splitsen we op in twee stukken, nl. Ω^+ en Ω^0, die als volgt gedefinieerd zijn:

$$\begin{cases} \Omega^0 = \{x \in (0,1) \mid u(x) = \psi(x)\} \\ \Omega^+ = \{x \in (0,1) \mid u(x) > \psi(x)\}. \end{cases} \tag{3.5}$$

Dat wil dus zeggen dat op Ω^0 de snaar aan het obstakel raakt en dat op Ω^+ de snaar zich strikt boven het obstakel bevindt. Kies nu de functie $v \in K$ zodanig dat

(3.6) $\qquad v = u + \varepsilon\phi \quad$ met $\quad \varepsilon > 0 \quad$ en $\phi = 0 \quad$ op Ω^0.

Daar op Ω^+ geldt $u > \psi$, kan bij iedere functie ϕ een $\varepsilon > 0$ gekozen worden zódat $v \in K$. Substitutie van (3.6) in (3.4) geeft

$$\int_{\Omega^+} \frac{du}{dx} \frac{d(\varepsilon\phi)}{dx}\, dx \geq \int_{\Omega^+} f\varepsilon\phi\, dx \quad \text{voor alle} \quad \phi \in V \text{ met } \phi = 0 \text{ op } \partial\Omega^+.$$

Daar Ω willekeurig is en ε weggedeeld kan worden volgt

$$\int_{\Omega^+} \frac{du}{dx} \frac{d\phi}{dx}\, dx = \int_{\Omega^+} f\phi\, dx \quad \text{voor alle} \quad \phi \in V \text{ met } \phi = 0 \text{ op } \partial\Omega^+$$

en dit is de zwakke formulering van de volgende diff. vgl.

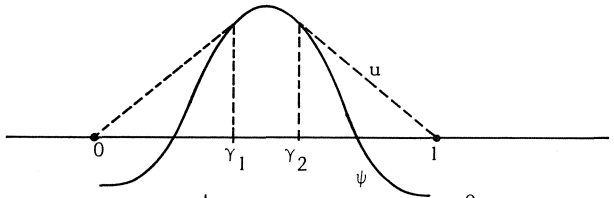
(3.7) $\qquad -\frac{d^2u}{dx^2} = f \quad$ op Ω^+.

Het obstakel-probleem (3.3) kan dus als volgt worden geïnterpreteerd:

(3.8)
$$\begin{cases} \text{Zoek } u : [0,1] \to \mathbb{R} \quad \text{zodanig dat} \\ u = \psi \quad \text{op} \quad \Omega^0 \\ -\frac{d^2u}{dx^2} = f \quad \text{op} \quad \Omega^+ \\ u(0) = u(1) = 0. \end{cases}$$

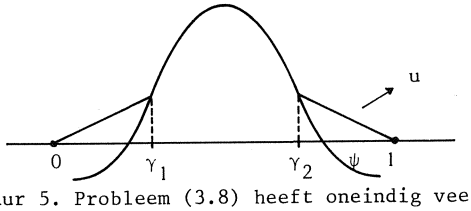
Laten we ons wel realiseren dat Ω^0 en Ω^+ onbekend zijn. Om dit tot uiting te laten komen herformuleren we (3.8) als volgt (zie Fig. 4).

(3.9)
$$\begin{cases} \text{Zoek } u : [0,1] \to \mathbb{R}\,, \ \gamma_1 \text{ en } \gamma_2 \quad \text{zodanig dat} \\ u = \psi \quad \text{op} \quad [\gamma_1, \gamma_2] \\ -\frac{d^2u}{dx^2} = f \quad \text{op} \quad (0,\gamma_1) \text{ en } (\gamma_2, 1) \\ u(0) = u(1) = 0 \\ u \quad \text{continu in } \gamma_1 \text{ en } \gamma_2. \end{cases}$$



Figuur 4. Oplossing u. $\Omega^+ = (0,\gamma_1) \cup (\gamma_2,1)$, $\Omega^0 = (\gamma_1,\gamma_2)$.

Formulering (3.8) (of (3.9)) is echter niet voldoende om de oplossing u eenduidig te karakteriseren. In feite zijn er oneindig veel oplossingen van (3.8). Eén ervan is aangegeven in Fig. 5.

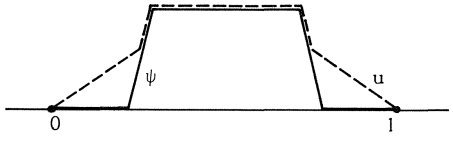


Figuur 5. Probleem (3.8) heeft oneindig veel oplossingen.

We zullen dus nog een extra conditie aan probleem (3.8) moeten toevoegen om het probleem eenduidig oplosbaar te maken. Is het obstakel ψ voldoende glad, bijvoorbeeld één keer continu differentieerbaar, dan maakt de extra conditie

(3.10) $\qquad \frac{du}{dx}$ continu in γ_1 en γ_2

probleem (3.8) eenduidig oplosbaar. Is de functie ψ niet glad en stellen we toch de eis (3.10) dan kan het probleem geen oplossing hebben zoals blijkt uit Fig. 6.a.



Figuur 6.a. ψ niet glad. Probleem (3.8) heeft oneindig veel oplossingen. Met conditie (3.10) bestaat er geen oplossing.

De conclusie is dat de formulering in termen van een diff. vgl. slechts gegeven kan worden voor gladde obstakels ψ. De variationele formulering ((3.3), (3.4)) kan ook gegeven worden van niet gladde, zelfs discontinue functies ψ.

4. MEMBRAAN PROBLEEM (SEMI-PERMEABELE WAND)

Beschouw het volgende probleem:

$$(4.1) \qquad \begin{cases} \text{Zoek } u \in K \text{ zodanig dat} \\ J(u) = \inf_{v \in K} J(v) \end{cases}$$

met $$J(v) = \tfrac{1}{2} \int_0^1 \left(\frac{dv}{dx} \right)^2 dx - \int_0^1 fv \, dx$$

$$K = \{v : [0,1] \to \mathbb{R} \mid v(0) = 0 \, , \, v(1) \geq 0\} \subset V$$

$$V = \{v : [0,1] \to \mathbb{R} \mid v(0) = 0\}.$$

Uit Stelling (2.4) volgt dat (4.1) equivalent is met de volgende VO:

$$(4.2) \qquad \begin{cases} \text{Zoek } u \in K \text{ zodanig dat} \\ \int_0^1 \frac{du}{dx} \, \frac{d(v-u)}{dx} \, dx \geq \int_0^1 f(v-u) \, dx \quad \text{voor alle } v \in K. \end{cases}$$

Dit probleem zullen we nu weer interpreteren als de zwakke formulering van een diff. vgl. met randvoorwaarden. Kiezen we achtereenvolgens $v = 0$ en $v = 2u$ dan geldt:

$$(4.3) \qquad \int_0^1 \left(\frac{du}{dx} \right)^2 dx = \int_0^1 fu \, dx.$$

Trekken we (4.3) van (4.2) af dan volgt

$$(4.4) \qquad \int_0^1 \frac{du}{dx} \, \frac{dv}{dx} \, dx \geq \int_0^1 fv \, dx \quad \text{voor alle } v \in K.$$

Kies vervolgens $v = \pm \phi$ met ϕ een willekeurige functie van $[0,1] \to \mathbb{R}$ met $\phi(0) = \phi(1) = 0$; dan geldt $v \in K$ en uit (4.4) volgt

$$(4.5) \qquad \int_0^1 \frac{du}{dx} \, \frac{d\phi}{dx} \, dx = \int_0^1 f\phi \, dx \quad \text{voor alle } \phi \text{ met } \phi(0) = \phi(1) = 0.$$

Dit is de zwakke formulering van de volgende diff. vgl.

$$(4.6) \qquad -\frac{d^2u}{dx^2} = f \quad \text{op} \quad (0,1).$$

Vermenigvuldigen we (4.6) met $v-u$, waarbij $v \in K$ willekeurig is, integreren we over $(0,1)$ en passen we de partiële integratie regel toe dan volgt:

$$(4.7) \qquad \int_0^1 \frac{du}{dx}\,\frac{d(v-u)}{dx}\,dx - \left[\frac{du}{dx}\,(v-u)\right]_0^1 = \int_0^1 f(v-u)\,dx.$$

Daar $v(0) = u(0) = 0$, reduceert (4.7) tot

$$(4.8) \qquad \int_0^1 \frac{du}{dx}\,\frac{d(v-u)}{dx}\,dx - \frac{du}{dx}(1).(v-u)(1) = \int_0^1 f(v-u)\,dx.$$

Vergelijken we nu (4.8) met (4.2) dan moet kennelijk gelden dat

$$(4.9) \qquad \frac{du}{dx}\,(1).(v-u)(1) \geq 0 \quad \text{voor alle } v \in K.$$

Kiezen we achtereenvolgens $v = 0$ en $v = 2u$ dan geldt

$$(4.10) \qquad \frac{du}{dx}\,(1).u(1) = 0.$$

Aftrekken van (4.9) geeft

$$(4.11) \qquad \frac{du}{dx}\,(1).v(1) \geq 0 \quad \text{voor alle} \quad v \in K.$$

Daar $v \in K$, geldt $v(1) \geq 0$, zodat

$$(4.12) \qquad \frac{du}{dx}\,(1) \geq 0.$$

De conclusie is dat (4.2) de zwakke formulering is van

$$(4.13) \qquad \begin{cases} -\dfrac{d^2u}{dx^2} = f \quad \text{op } (0,1) \\ u(0) = 0 \\ u(1) \geq 0 \\ \dfrac{du}{dx}\,(1) \geq 0 \\ \dfrac{du}{dx}\,(1).u(1) = 0. \end{cases}$$

In het punt $x = 1$ geldt dus

(4.14)
$$\text{ôf} \quad u = 0 \quad \text{en} \quad \frac{du}{dx} \geq 0$$
$$\text{ôf} \quad u > 0 \quad \text{en} \quad \frac{du}{dx} = 0.$$

Fysisch kan dit probleem gezien worden als een model voor warmtegeleiding in een staaf (0,1), waarbij de temperatuur u is voorgeschreven in $x = 0$: $u(0) = 0$. In $x = 1$ geldt dat er geen warmte van binnen naar buiten kan stromen; wel van buiten naar binnen. Dus voor de warmtestroom q in $x = 1$ geld

$$q(1) = -\frac{du}{dx}(1) \leq 0 \Rightarrow \frac{du}{dx}(1) \geq 0.$$

Stel de buiten-temperatuur op 0 : $u = 0$ voor $x > 1$. Dan kunnen zich de volgende gevallen voordoen.

(i) $u(1) > 0$. Er zou dan warmte naar buiten stromen, maar dit kan niet. Dus $\frac{du}{dx}(1) = 0$. (4.14b).

(ii) Stel $u(1) \leq 0$; dan zal er warmte van buiten naar binnen stromen. Dit is toegestaan. Nu kan er zich echter geen sprong in de temperatuur voordoen, zodat $u(1) = 0$ = buiten-temperatuur. (4.14a).

Tenslotte laten we nog zien dat (4.2) afgeleid kan worden uit (4.13). Zij $v : [0,1] \to \mathbb{R}$ zodanig dat $v(0) = 0$. Vermenigvuldig (4.13a) met (v-u), integreer over (0,1), pas partiële integratie toe:

$$(4.15) \quad \int_0^1 \frac{du}{dx} \frac{d(v-u)}{dx}\, dx + \frac{du}{dx}(0).(v-u)(0) - \frac{du}{dx}(1).(v-u)(1) = \int_0^1 f(v-u)\, dx.$$

In $x = 0$ geldt $(v-u)(0) = 0$. In $x = 1$ geldt $\frac{du}{dx}(1)\, u(1) = 0$. Verder geldt $\frac{du}{dx}(1) \geq 0$. Kiezen we nu functies v met $v(1) \geq 0$, dan geldt

$$\frac{du}{dx}(1).v(1) \geq 0$$

zodat uit (4.15) volgt dat

$$(4.16) \quad \int_0^1 \frac{du}{dx} \frac{d(v-u)}{dx}\, dx \geq \int_0^1 f(v-u)\, dx \quad \text{voor alle } v \in K$$

waarmee is aangetoond dat (4.2) uit (4.13) volgt.

5. ZADELPUNT FORMULERING

Zoals we reeds in paragraaf 2 hebben opgemerkt is de geprojecteerde gradiënt methode niet direct toepasbaar op de VO van de paragrafen 3 en 4. We zullen de VO nu gaan schrijven als inf sup problemen (zie Hoofdstuk 6). Anders gezegd: de oplossing van de VO wordt nu gezocht via het bepalen van een zadelpunt van een Lagrangiaan. De inf sup formulering heet het primaire probleem. Onder zekere voorwaarden (zie Hoofdstuk 6) kunnen we hiervoor schrijven: sup inf, het duale probleem. Op dit duale probleem gaan we dan in paragraaf 6 de geprojecteerde gradiënt methode toepassen.

We bekijken nog eens de formulering (3.3) van het obstakel-probleem. De nevenvoorwaarden $v \geq \psi$ zijn opgenomen in de verzameling K. We gaan nu deze nevenvoorwaarden verwerken in de uitdrukking voor de functionaal J, zodanig dat de oplossing u gezocht kan worden in de hele functieruimte V en dat toch, uiteindelijk, u aan de nevenvoorwaarden voldoet. We bekijken de uitdrukking

$$(5.1) \qquad \int_0^1 q(x).(v-\psi)(x)\, dx$$

waarbij de functie $q : [0,1] \to \mathbb{R}$ niet-positief is:

$$q \leq 0.$$

Voor de integraal geldt dat

$$(5.2) \qquad \sup_{q \leq 0} \int_0^1 q(v-\psi)\, dx = \begin{cases} 0 & \text{als} \quad v \geq \psi \\ \infty & \text{als} \quad v \not\geq \psi. \end{cases}$$

Immers, als $v \geq \psi$ dan is $v - \psi \geq 0$ en $q(v-\psi) \leq 0$. Het supremum wordt dan aangenomen voor $q = 0$. Als echter $v(x_0) < \psi(x_0)$ dan geldt op grond van continuïteitseigenschappen dat $v(x) < \psi(x)$ voor x behorend tot een omgeving B_{x_0}. Kiezen we q nu zodanig dat $q = 0$ op $[0,1] \setminus B_{x_0}$ en "oneindig negatief" op een interval $\tilde{B}_{x_0} \subset B_{x_0}$, dan geldt

$$\sup_{q \leq 0} \int_0^1 q(v-\psi)\, dx = \int_{\tilde{B}_{x_0}} \overset{<0}{\uparrow} q \overset{<0}{\uparrow} (v-\psi)\, dx \to \infty.$$

We kunnen dus stellen dat

$$\inf_{v\in K} J(v) = \inf_{v\in V} [J(v) + \sup_{q\le 0} \int_0^1 q(v-\psi)\, dx]$$

want als v niet groter dan of gelijk is aan ψ, is het rechterlid gelijk aan oneindig. Daarentegen is voor $v \ge \psi$ de waarde eindig. Dus bij het zoeken van het infimum over de hele ruimte V, komen we automatisch in de klasse K terecht. Dus voor (3.3) kunnen we schrijven:

$$(5.3) \qquad \begin{cases} \text{Zoek } u \in V \text{ en } p \le 0 \text{ zodanig dat} \\ L(u,p) = \inf_{v\in V} \sup_{q\le 0} L(v,q) \end{cases}$$

met

$$(5.4) \qquad L(v,q) = J(v) + \int_0^1 q(v-\psi)\, dx$$

de Lagrangiaan. Merk op dat

$$\sup_{q\le 0} L(v,q) = \sup_{q\le 0} \{J(v) + \int_0^1 q(v-\psi)\, dx\} = \begin{cases} J(v) & \text{als } v \ge \psi \\ \infty & \text{als } v \not\ge \psi \end{cases}$$

zodat

$$\inf_{v\in V} \sup_{q\le 0} L(v,q) = \inf_{v\in V} \left\{ \begin{cases} J(v) & \text{als } v \ge \psi \\ \infty & \text{als } v \not\ge \psi \end{cases} \right\} = \inf_{v\ge\psi} J(v).$$

Voor het linkerlid van (5.3) geldt dan

$$L(u,p) \equiv J(u) + \int_0^1 p(u-\psi)\, dx = J(u)$$

zodat $\int_0^1 p(u-\psi)\, dx = 0$. De relaties

$$(5.5) \qquad \begin{cases} \int_0^1 p(u-\psi)\, dx = 0 \\ p \le 0 \\ u - \psi \ge 0 \end{cases}$$

worden de Kuhn-Tucker relaties van het zadelpunt-probleem (5.3) genoemd.

De oplossing $\{u,p\}$ van (5.3) heet een zadelpunt en voldoet aan

$$(5.6) \qquad L(u,q) \le L(u,p) \le L(v,p) \quad \text{voor alle } v \in V \text{ en voor alle } q \le 0.$$

Voor het probleem met de semi-permeabele wand van paragraaf 4 bekijken we de uitdrukking

(5.7) $\quad q.v(1)$

met q een niet-positief reëel getal : $q \leq 0$. Analoog aan (5.2) geldt

$$\sup_{q\leq 0} \; q.\, v(1) = \begin{cases} 0 & \text{als} \quad v(1) \geq 0 \\ \infty & \text{als} \quad v(1) < 0. \end{cases}$$

Er geldt dus voor probleem (4.1) dat

$$\inf_{v\in K} J(v) = \inf_{v\in V} [J(v) + \sup_{q\leq 0} q.v(1)].$$

De zadelpunt-formulering van (4.1) is dan:

(5.8)
$$\begin{cases} \text{Zoek } u \in V \quad \text{en} \quad p \leq 0 \quad \text{zodanig dat} \\ L(u,p) = \inf_{v\in V} \sup_{q\leq 0} L(v,q) \end{cases}$$

waarbij

(5.9) $\quad L(v,q) = J(v) + \; q.v(1).$

We merken weer op dat

$$\sup_{q\leq 0} L(v,q) = \sup_{q\leq 0} [J(v) + q.v(1)] = \begin{cases} J(v) & \text{als } v(1) \geq 0 \\ \infty & \text{als } v(1) < 0 \end{cases}$$

zodat

$$\inf_{v\in V} \sup_{q\leq 0} L(v,q) = \inf_{v\in V} \left\{ \begin{matrix} J(v) & \text{als } v(1) \geq 0 \\ \infty & \text{als } v(1) < 0 \end{matrix} \right\} = \inf_{v(1)\geq 0} J(v).$$

Voor het linkerlid van de gelijkheid in (5.8) moet dan gelden

$$L(u,p) \equiv J(u) + p.u(1) = \inf_{v(1)\geq 0} J(v)$$

waaruit volgt dat $p.u(1) = 0$. De relaties

(5.10)
$$\begin{cases} p.u(1) = 0 \\ p \leq 0 \\ u(1) \geq 0 \end{cases}$$

heten de Kuhn-Tucker relaties van dit zadelpunt-probleem (zie Hoofdstuk 6).

6. OUDE PROBLEMEN EN GEPROJECTEERDE GRADIËNT METHODE

Op de primaire problemen (5.3) en (5.8) passen we nu de theorie van Hoofdstuk 6 toe. We definiëren de (equivalente) duale problemen door verwisseling van inf en sup. Voor het obstakelprobleem wordt dit:

(6.1)
$$\left|\begin{array}{l} \text{Zoek } u : [0,1] \to \mathbb{R} \text{ in } V \text{ en } p : [0,1] \to \mathbb{R} \text{ met } p \le 0 \text{ zodanig dat} \\ L(u,p) = \sup_{q \le 0} \inf_{v \in V} L(v,q) \end{array}\right.$$

met

$$\left|\begin{array}{l} V = \{v : [0,1] \to \mathbb{R} \mid v(0) = v(1) = 0\} \\ L(v,1) = J(v) + \int_0^1 q(v-\psi)\, dx \\ J(v) = \tfrac{1}{2} \int_0^1 \left(\frac{dv}{dx}\right)^2 dx - \int_0^1 f\, v\, dx. \end{array}\right.$$

Het membraanprobleem heeft de volgende duale formulering

(6.2)
$$\left|\begin{array}{l} \text{Zoek } u : [0,1] \to \mathbb{R} \text{ in } V \text{ en een reële } p \le 0 \text{ zodanig dat} \\ L(u,p) = \sup_{q \le 0} \inf_{v \in V} L(v,q) \end{array}\right.$$

met

$$\left|\begin{array}{l} V = \{v : [0,1] \to \mathbb{R} \mid v(0) = 0\} \\ L(v,q) = J(v) + q.v(1) \\ J(v) = \tfrac{1}{2} \int_0^1 \left(\frac{dv}{dx}\right)^2 dx - \int_0^1 f\, v\, dx . \end{array}\right.$$

De algemene "setting" van beide problemen is dus:

(6.3)
$$\left|\begin{array}{l} \text{Zoek } \{u,p\} \in V \times P \text{ zodanig dat} \\ L(u,p) = \sup_{q \in P} \inf_{v \in V} L(v,q) \end{array}\right.$$

met

$$\left|\begin{array}{l} L(v,q) = J(v) + \langle q, \phi(v) \rangle \\ P = \{q : [0,1] \to \mathbb{R} \mid q \le 0\}\ ,\ \langle q, \phi(v) \rangle = \int_0^1 q(v-\psi)\, dx \text{ (obstakel probleem)} \\ P = \{q \in \mathbb{R} \mid q \le 0\}\ ,\ \langle q, \phi(v) \rangle = q.v(1) \text{ (membraan-probleem).} \end{array}\right.$$

De afbeelding ϕ is affien, d.w.z. kan geschreven worden als $\phi(v) = \phi_0 + \phi_\ell(v)$ met ϕ_0 onafhankelijk van v en ϕ_ℓ lineair in v.

Probleem (6.3) gaan we nu oplossen door de geprojecteerde gradiënt methode toe te passen op:

$$(6.4)\qquad \begin{cases} \text{Zoek } p \in P \text{ zodanig dat} \\ I(p) = \sup\limits_{q \in P} I(q) \end{cases}$$

met

$$(6.5)\qquad I(q) = \inf_{v \in V} L(v,q).$$

(6.4) is equivalent met:

$$(6.6)\qquad \begin{cases} \text{Zoek } p \in P \text{ zodanig dat} \\ -I(p) = \inf\limits_{q \in P} \{-I(q)\}. \end{cases}$$

Eerst berekenen we $I(q) = \inf\limits_{v \in V} L(v,q)$. Passen we Stelling (2.4) toe en bedenken we dat het inf gezocht moet worden over de hele functieruimte V, dus zonder nevenvoorwaarden, dan kan bewezen worden dat

$$(6.7)\qquad I(q) = \tfrac{1}{2} \int_0^1 \left(\frac{du_q}{dx}\right)^2 dx - \int_0^1 f u_q \, dx + \langle q, \phi(u_q) \rangle$$

waarbij u_q het eenduidig bepaalde element in V is dat voldoet aan

$$(6.8)\qquad \int_0^1 \frac{du_q}{dx} \frac{dv}{dx} \, dx - \int_0^1 fv \, dx + \langle q, \phi_\ell(v) \rangle = 0 \text{ voor alle } v \in V$$

waarbij $\phi_\ell(v)$ het lineaire deel is van $\phi(v)$; dus

$$\langle q , \phi_\ell(v) \rangle = \int_0^1 q\, v \, dx \qquad \text{(obstakel-probleem)}$$

$$\langle q , \phi_\ell(v) \rangle = q.v(1) \qquad \text{(membraan-probleem).}$$

Formulering (6.8) kan geïnterpreteerd worden als de zwakke formulering van de volgende diff. vgl. met randvoorwaarden.

Obstakel-probleem:

$$(6.9)\qquad \begin{cases} -\dfrac{d^2 u_q}{dx^2} = f - q \quad \text{op} \quad (0,1) \\ u_q(0) = u_q(1) = 0. \end{cases}$$

Membraan-probleem:

$$(6.10)\qquad \begin{cases} -\dfrac{d^2u_q}{dx^2} = f \\ u_q(0) = 0 \\ \dfrac{du_q}{dx}(1) = -q. \end{cases}$$

Voor het toepassen van de geprojecteerde gradiënt methode op (6.6) dient de Gateaux-afgeleide van $-I(q)$ berekend te worden. Deze bepalen we uit (6.7) en maken gebruik van (6.8). Na enig rekenwerk vinden we

$$(6.11)\qquad I'(p)\,[q] = \lim_{\lambda \downarrow 0} \frac{1}{\lambda}\{I(p+\lambda q) - I(p)\} = < q, \phi(u_p) >.$$

Dus voor het obstakel-probleem vinden we

$$(6.12)\qquad I'(p) = u_p - \psi$$

en voor het membraan-probleem

$$(6.13)\qquad I'(p) = u_p(1).$$

Met de geprojecteerde gradiënt methode kunnen we nu het inf bepalen over de niet-lege convexe gesloten verzameling P in (6.6):

$$(6.14)\qquad \begin{cases} \text{Zij } p_0 \in P \text{ willekeurig. We definiëren } \{p_n\}_n,\ n \geq 1, \text{ door} \\ p_{n+\frac{1}{2}} = p_n - \rho(-I'(p_n)) \\ p_{n+1} = \mathrm{Proj}_P\,[p_{n+\frac{1}{2}}]. \end{cases}$$

De projectie op P kan eenvoudig worden uitgevoerd daar P een klasse van functies is, waarin afgeleiden geen rol spelen.

De iteratieve methode voor het oplossen van het obstakel-probleem, gebaseerd op de geprojecteerde gradiënt methode toegepast op het duale probleem (Uzawa algorithme) is nu als volgt:

(6.15) We starten met $p_0 \leq 0$, een willekeurige functie. Zijn $\{u_1,p_1\},\ldots,\{u_n,p_n\}$ bekend, dan wordt $\{u_{n+1},p_{n+1}\}$ als volgt gedefinieerd

$$(6.16') \qquad \begin{cases} -\dfrac{d^2 u_{n+1}}{dx^2} = f - p_n \quad \text{op } (0,1) \\ u_{n+1}(0) = u_{n+1}(1) = 0 \end{cases}$$

en

$$(6.16'') \qquad \begin{cases} p_{n+\frac{1}{2}} = p_n + \rho(u_{n+1} - \psi) \\ p_{n+1} = \text{Proj}_{\leq 0}\,[p_{n+\frac{1}{2}}] \end{cases}$$

waarbij geldt

$$p_{n+1}(x) = \min\{0,\ p_{n+\frac{1}{2}}(x)\}.$$

Het Uzawa algorithme voor het membraan-probleem is als volgt:

(6.17) Zij $p_0 \leq 0$ een willekeurig reëel getal. Zijn $\{u_1,p_1\},\ldots,\{u_n,p_n\}$ bekend, dan wordt $\{u_{n+1},p_{n+1}\}$ als volgt gedefinieerd

$$(6.18') \qquad \begin{cases} -\dfrac{d^2 u_{n+1}}{dx^2} = f \quad \text{op} \quad (0,1) \\ u_{n+1}(0) = 0 \\ \dfrac{du_{n+1}}{dx}(1) = -p_n \end{cases}$$

$$(6.18'') \qquad \begin{cases} p_{n+\frac{1}{2}} = p_n + \rho\, u_{n+1}(1) \\ p_{n+1} = \text{Proj}_{\leq 0}\,[p_{n+\frac{1}{2}}] = \min\{0,\ p_{n+\frac{1}{2}}\}. \end{cases}$$

Betreffende de convergentie van het Uzawa algorithme geldt de volgende stelling.

(6.19) STELLING.

Als $0 < \rho < \rho_0$ ($\rho_0 = 2\pi^2$ voor het obstakel-probleem, $\rho_0 = 2$ voor het membraan-probleem) dan geldt voor $n \to \infty$

$$\int_0^1 |u_n - u|^2\, dx \to 0$$

en

$$\int_0^1 \left| \frac{du_n}{dx} - \frac{du}{dx} \right|^2 dx \to 0.$$

7. NUMERIEKE RESULTATEN

We zullen in deze paragraaf enige numerieke resultaten laten zien.

OBSTAKEL PROBLEEM

(7.1) Ten eerste zullen we het geval bekijken dat $f \equiv 0$ en ψ gegeven is door

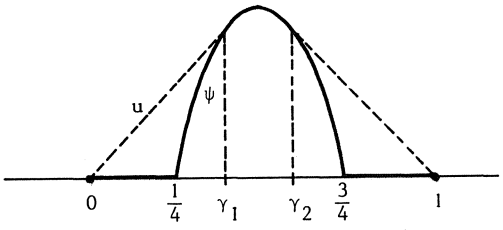
$$\psi(x) = \begin{cases} 0, & 0 \le x < \frac{1}{4} \\ -8(x-\frac{1}{4})(x-\frac{3}{4}), & \frac{1}{4} \le x \le \frac{3}{4} \\ 0, & \frac{3}{4} < x \le 1. \end{cases}$$

Dit probleem kan analytisch worden opgelost. We vinden

$$\begin{cases} \gamma_1 = \frac{1}{4}\sqrt{3} \approx 0.433, & \gamma_2 = 1 - \frac{1}{4}\sqrt{3} \approx 0.567 \\ u(x) = (8 - 4\sqrt{3})\,x & \text{op } [0,\gamma_1] \\ u(x) = \psi(x) & \text{op } (\gamma_1,\gamma_2) \\ u \text{ symmetrisch} & \text{t.o.v. } x = \frac{1}{2}. \end{cases}$$

Het obstakel ψ en de oplossing u zijn afgebeeld in Fig. 6.b.

Figuur 6.b. Obstakel ψ en oplossing u.



We passen het Uzawa-algorithme toe, waarbij de oplossingen u_{n+1} van (6.16') numeriek berekend worden met behulp van de EEM (zie Hoofdstuk 5). Het interval [0,1] wordt opgedeeld in 32 equidistante intervallen, dus $h = \frac{1}{32}$. We starten het algorithme met $p_0(x) = 0$ en stoppen het algorithme als $\max_{x\in[0,1]} |u_{n+1}(x) - u_n(x)| < 10^{-3}$. In Tabel 1 staan, voor verschillende waarden van ρ, het aantal benodigde iteraties it en het verschil tussen $u_{it}(x)$ en de exacte oplossing u(x): $\max_{x\in[0,1]} |u_{it}(x) - u(x)|$

ρ	aantal iteraties it	$\max_{x\in[0,1]} \lvert u_{it}(x) - u(x) \rvert$
1	85	0.195
3	75	0.106
5	58	0.087
7	52	0.072
9	46	0.066
11	47	0.053
13	41	0.052
15	37	0.050
17	35	0.048
19	35	0.044
21	40	0.032
23	32	0.035
25	55	0.016
27	>100	-

TABEL 1

(7.2) In het tweede voorbeeld van het obstakelprobleem kiezen we $f \equiv 0$ en

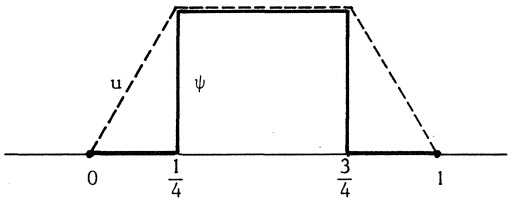
$$\psi(x) = \begin{cases} 0 & 0 \le x < \frac{1}{4} \\ 1 & \frac{1}{4} \le x \le \frac{3}{4} \\ 0 & \frac{3}{4} < x \le 1. \end{cases}$$

De analytische oplossing wordt gegeven door

$$u(x) = \begin{cases} 4x, & 0 \le x < \frac{1}{4} \\ 1, & \frac{1}{4} \le x \le \frac{3}{4} \\ 4 - 4x, & \frac{3}{4} < x \le 1. \end{cases}$$

De functies ψ en u zijn weergegeven in Fig. 7.

Figuur 7. Obstakel ψ en oplossing u.



Uzawa wordt weer toegepast met de EEM (zie Voorbeeld 7.1). De resultaten zijn vermeld in Tabel 2.

ρ	aantal iteraties it	$\max_{x\in[0,1]} \lvert u_{it}(x) - u(x) \rvert$
1	57	0.047
3	26	0.035
5	18	0.033
7	19	0.019
9	15	0.018
11	13	0.018
13	11	0.018
15	10	0.018
17	8	0.018
19	9	0.017
21	10	0.015
23	13	0.007
25	13	0.006
27	17	0.005
29	38	0.002
31	>100	

TABEL 2

MEMBRAANPROBLEEM

Hier kiezen we $f(x) = \lambda x - 1$, met λ een parameter. De analytische oplossing wordt als volgt geconstrueerd. Er geldt

$$-\frac{d^2u}{dx^2} = f$$

zodat

$$u(x) = -\frac{1}{6}\lambda x^3 + \frac{1}{2}x^2 + c_1 x + c_2$$

met c_1 en c_2 constanten. Uit $u(0) = 0$ volgt $c_2 = 0$. Verder geldt

$$\begin{cases} u(1) = -\frac{1}{6}\lambda + \frac{1}{2} + c_1 \\ \frac{du}{dx}(1) = -\frac{1}{2}\lambda + 1 + c_1 . \end{cases}$$

Er moet dus gelden

$$\begin{cases} -\frac{1}{6}\lambda + \frac{1}{2} + c_1 \geq 0 \Rightarrow c_1 \geq \frac{\lambda}{6} - \frac{1}{2} \\ -\frac{1}{2}\lambda + 1 + c_1 \geq 0 \Rightarrow c_1 \geq \frac{\lambda}{2} - 1 \\ (-\frac{1}{6}\lambda + \frac{1}{2} + c_1)(-\frac{1}{2}\lambda + 1 + c_1) = 0. \end{cases}$$

We merken nog op dat $\frac{\lambda}{6} - \frac{1}{2} = \frac{\lambda}{2} - 1$ voor $\lambda = \frac{3}{2}$, zodat

$$\begin{cases} c_1 = \frac{\lambda}{2} - 1, & \text{dus } u(1) > 0, \ \frac{du}{dx}(1) = 0, & \text{als } \lambda > \frac{3}{2} \\ c_1 = \frac{\lambda}{6} - \frac{1}{2}, & \text{dus } u(1) = 0, \ \frac{du}{dx}(1) \geq 0, & \text{als } \lambda \leq \frac{3}{2} . \end{cases}$$

Voor $p = \lim_{n \to \infty} p_n$ vinden we

$$p = \begin{cases} 0 \text{ als } \lambda > \frac{3}{2} \\ -(-\frac{\lambda}{2} + 1) - (\frac{\lambda}{6} - \frac{1}{2}) = \frac{\lambda}{3} - \frac{1}{2} \quad \text{als } \lambda \leq \frac{3}{2} . \end{cases}$$

Het Uzawa algorithme met $p_0 = 1$ wordt weer toegepast met de EEM; het aantal subintervallen is 32, $h = \frac{1}{32}$. De resultaten voor verschillende waarden van λ en ρ staan vermeld in Tabel 3.

ρ	$\lambda = 0.5$		$\lambda = 1.0$		$\lambda = 1.5$		$\lambda = 2.0$		$\lambda = 2.5$	
0.33	15	0.0017	13	0.0019	3	0.0002	3	0.0003	3	0.0004
0.67	8	0.0005	8	0.0002	3	0.0002	3	0.0003	3	0.0004
1.00	3	0.0000	3	0.0000	3	0.0002	3	0.0003	3	0.0004
1.33	9	0.0002	9	0.0002	4	0.0002	4	0.0003	3	0.0004
1.67	20	0.0003	18	0.0004	4	0.0002	4	0.0003	3	0.0004
2.00	>100	-	>100	-	4	0.0002	4	0.0003	4	0.0004

TABEL 3. Voor iedere λ is vermeld: het aantal iteraties it en $\max_{x \in [0,1]} |u_{it}(x) - u(x)|$.

Zoals uit de Tabellen 1, 2 en 3 blijkt wordt de convergentiesnelheid van het Uzawa algorithme bepaald door de grootte van ρ. Voor te grote waarden van ρ is het algorithme divergent; voor te kleine waarden is de convergentiesnelheid laag.

LITERATUUR

1. C. CUVELIER, *Introduction to the Numerical Analysis of Variational Inequalities*, Delft University of Technology, Report NA-22,1978.

2. C. CUVELIER, *Functional and Numerical Analysis of Partial Differential Equations* (Boek in voorbereiding).

3. G. DUVAUT & J.L. LIONS, *Les Inéquations en Mécanique et Physique*, Dunod, Paris, 1972 (Engelse vertaling: Springer, Heidelberg).

4. R. GLOWINSKI, J.L. LIONS & R. TREMOLIERES, *Analyse Numérique des Inéquations Variationnelles*, Dunod, Paris, 1976 (Engelse vertaling: North-Holland, Amsterdam).

APPENDIX

OVER BERGPASSEN TOT EEN GODSBEWIJS ?

(Uit: W & N bulletin, Faculteit Wiskunde en Natuurwetenschappen, Katholieke Universiteit te Nijmegen.)

E.W.C. van GROESEN

Zonder dat U er zich van bewust bent, bent U op dit moment bezig met zowel het formuleren als het oplossen van een optimaliserings probleem. Hoewel het volgende verhaal zal gaan over dit soort problemen, en meer in het bijzonder over variatierekening, zal de inhoud ervan alleen op een indirecte manier van invloed zijn op de oplossing van Uw probleem.

VARIATIEREKENING is een van die onderdelen van de wiskunde waarvoor geldt dat de wiskundige ontwikkeling ervan belangrijke impulsen heeft ontvangen vanuit een bepaald toepassingsgebied en waarvoor, omgekeerd, de behaalde wiskundige resultaten van groot belang zijn geweest voor dit toepassingsgebied. We zullen dat hieronder duidelijk maken.

Ruim geïnterpreteerd bestudeert variatierekening het gedrag van functies (meestal functies van oneindig veel variabelen), en meer speciaal van die punten in het definitiegebied waarvoor de functie een zogenaamde stationaire waarde heeft, bijvoorbeeld de punten waar de functie een maximale of minimale (in één term: extremale) waarde heeft. Als we het definitiegebied van de functie aangeven met M, een of andere gegeven verzameling "elementen" ofwel "punten", en J een reëelwaardige functie is op M, dan is de bestudering van het *minimaliseringsprobleem* van J op M een voorbeeld van een probleem uit de variatierekening. Een minimaal punt $\hat{x} \in M$ voldoet dan per definitie aan

$$(1) \qquad J(\hat{x}) = \min_{x \in M} J(x)$$

dus $\qquad J(\hat{x}) \leq J(x)$ voor alle $x \in M$.

Een klasse van problemen die hieronder vallen en tegenwoordig alom onderzocht worden zijn *"optimale handelings-problemen"*. Voor een specifiek voorbeeld daarvan kunt U denken aan M als de verzameling van alle mogelijke productie processen voor een eenheid van een bepaald product (waarbij de processen meestal aan allerlei beperkende voorwaarden moeten voldoen) en aan J(x) als de productiekosten van een eenheid bij keuze van x als fabrikageproces. Een oplossing van het minimaliseringsprobleem (1)

correspondeert dan met een "optimaal" proces waarvoor de kosten zo laag mogelijk zijn.

Minimaliseringsproblemen als (1) komen zelfs in de mythologie al voor: volgens Vergilius kocht koningin Dido een stuk land, waarop het latere Carthago gebouwd zou worden, dat omgeven moest kunnen worden door de huid van een rund. Haar verwachting dat de maximale oppervlakte van het omsloten land bereikt wordt door de in repen gescheurde huid in een halve cirkel te leggen, met de eindpunten aan de "rechte" Middellandse Zeekust, is wiskundig juist, hoewel een verificatie daarvan pas in de 18e eeuw gegeven kon worden.

Een serieuze bestudering van minimaliseringsproblemen van deze aard (meer precies, van problemen zoals (1) waarin M een deelverzameling is van een oneindig dimensionale ruimte) begon namelijk in de 17e en 18e eeuw, en werd geïnitieerd door vragen uit de Natuurkunde, i.h.b. de Klassieke Mechanica en Mathematische Fysica. Rond die tijd werd men zich bewust van het feit dat veel basisvergelijkingen van de natuurkunde geformuleerd kunnen worden als minimaliseringsprobleem. Het eerste omschreven, en ook meest eenvoudig te formuleren, van deze zogenaamde *variatieprincipes*, is het befaamde *principe van Fermat* (1662) over de voortplanting van licht. Dit principe luidt dat de voortplanting van een lichtstraal in een optisch medium tussen twee punten plaatsvindt langs die baan waarvoor geldt dat de benodigde tijd zo klein mogelijk is in vergelijking met de tijd die nodig is langs enig andere baan tussen de twee gegeven punten.

Deze omschrijving verklaart ook de naam *variatie*-principe: variëren van de optimale baan, blijvend binnen de restrictieve verzameling M (= banen door de gegeven punten), geeft een baan die in het algemeen met een grotere waarde van J (= tijd langs baan) correspondeert. Metafysische argumenten lagen vaak ten grondslag aan het geloven in, en het zoeken naar, variatieprincipes. EULER, 1707-1783, bijvoorbeeld, schreef: "....je suis convaincu que par tout *la* nature agit selon quelque principe d'un maximum ou minimum ...".

MAUPERTUIS gebruikte de gevonden extremaliteitseigenschappen van de natuur zelfs ter ondersteuning van een bijzonder existentiebewijs: ".....des loix du mouvement où l'action est toujours employée avec la plus *grande économie démontreront l'existence* de l'Etre supreme;...", in : "Examen *philosophique* de la preuve de l'existence de Dieu", 1757.

Hoe dit ook zij, feit is dat sinds die tijd variatieprincipes een essentiële rol gespeeld hebben in de mathematische fysica: het principe van "minimale" actie, waarmee dynamische problemen uit de klassieke en continuum mechanica worden beschreven, en een variatieprincipe voor de Maxwell vergelijkingen van het electro-magnetisme, zijn voorbeelden waarvoor geldt dat het verschijnsel al eerder op een andere dan variationele manier beschreven kon worden. Een in dit opzicht meer essentiële rol spelen variatieprincipes tegenwoordig bij het opsporen van de natuurwetten die een (universele) beschrijving (pogen te) geven van alle fundamentele krachten van de natuur; te denken valt aan de relativiteits-theorie en aan moderne ijkvelden-theoriën zoals die van Yang-Mills.
Het wiskundig onderzoek van minimaliseringsproblemen heeft zich na Euler tot op de dag van vandaag voortgezet en heeft veel ander onderzoek geïnitieerd of beinvloed. Een voorbeeld is *Functionaalanalyse*, nu een zelfstandig vakgebied, dat ontstaan is door bestudering van de vraag naar het bestaan van een minimaal punt voor het geval M een oneindig-dimensionale ruimte is.

Hiervoor is steeds gesproken over problemen waarvoor een minimale of maximale waarde wordt gezocht. Echter, aan het eind van de vorige eeuw (POINCARÉ,1897) was al bekend dat meer punten dan alleen die waarvoor de functie extremaal is van belang zijn voor de beschrijving van de wetten van de natuur. Meer algemeen dan extreme punten blijken zogenaamde *stationaire punten* van belang te zijn. (Het principe van minimale actie wordt dan ook sindsdien meer correct het principe van stationaire actie genoemd.)

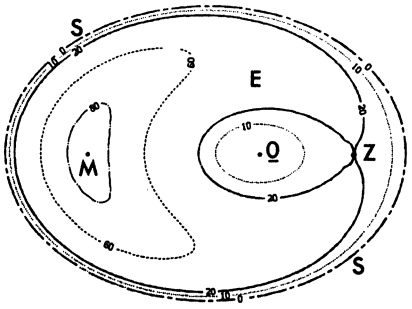
Hieronder illustreren we de betekenis van dit begrip en geven een indicatie van de huidige ontwikkeling in het wiskundig onderzoek daarvan. Met een

functie van slechts twee variabelen kunnen al enkele essentiële ideeën uitgelegd worden. Laat daarom J een differentieerbare functie zijn van twee variabelen x_1 en x_2; schrijf $\underline{x} = (x_1,x_2)$. De grafiek van J, de functiewaarde $J(\underline{x})$ loodrecht uitgezet boven het punt $\underline{x}$ in het x_1, x_2-vlak, kan voorgesteld worden als een berglandschap. Een stationair punt $\hat{\underline{x}}$ wordt dan gedefinieerd als een punt waarvoor het raakvlak aan dit berglandschap horizontaal is met het x_1,x_2-vlak. Een andere manier om dit voor te stellen is door op te merken dat een "testballetje" op de grafiek van J boven de plaats $\underline{x}$ onder invloed van de zwaartekracht alleen dan niet zal gaan rollen als $\underline{x}$ een stationair punt is. [Analytisch kan dit naar analogie van functies van één variabele uitgedrukt worden door te zeggen dat de afgeleiden van J in elk van de twee richtingen x_1 en x_2 nul zijn in $\hat{\underline{x}}$.]
Punten waarvoor J een extremale waarde heeft zijn stationaire punten: in dat geval ligt de grafiek geheel boven of onder het horizontale raakvlak. Het punt (0,0) van de eenvoudige functie $J(x_1,x_2) = (x_1)^2 - (x_2)^2$ is een voorbeeld van een stationair punt dat géén extremaal punt is. Het is een zogenaamd *zadelpunt*, i.e. een stationair punt $\hat{\underline{x}}$ waarvoor in elke omgeving punten $\underline{y}$ en $\underline{z}$ te vinden zijn met

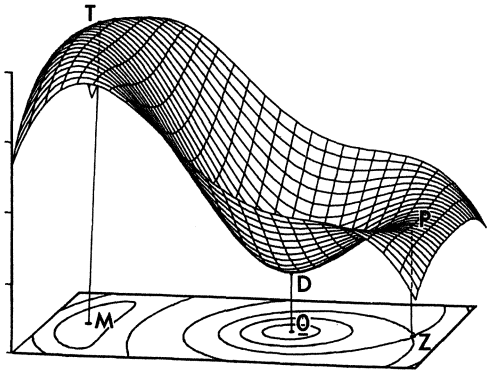
$$J(\underline{y}) < J(\hat{\underline{x}}) < J(\underline{z})$$

(in elke omgeving van $\hat{\underline{x}}$ zijn er zowel punten waarvoor de grafiek boven als onder het horizontale raakvlak ligt).

Een voorbeeld van een functie met verschillende soorten stationaire punten is te verkrijgen door voor J de hoogte (boven zeeniveau) te nemen van een berglandschap op een eiland E. Veronderstel i.h.b. dat elk punt van het eiland boven zeeniveau ligt, m.u.v. één punt dat we als oorsprong van het x_1,x_2-vlak nemen, dus $J(0) = 0$. Dit punt $\underline{0}$ denken we dan omgeven door een bergketen die het punt $\underline{0}$ scheidt van het "strand" S van het eiland (S is de rand van E; voor elk punt $\underline{x}$ van S geldt ook $J(\underline{x}) = 0$). Het is duidelijk dat $\underline{0}$ een stationair punt is, want het is een (locaal) minimaal punt voor



Hoogtekaart van het eiland E; 0, M en Z zijn respectivelijk het minimale-, maximale- en zadelpunt.



Een deel van het berglandschap op E, met T de top, D het dal en P de bergpas.

Daarnaast is er tenminste nog één ander extremaal punt dat correspondeert met het hoogste punt van het eiland. Het is niet moeilijk een voorbeeld te verzinnen van een hoogtefunctie J met voorgaande eigenschappen die geen andere stationaire punten heeft dan 0 en (in dat geval een continuum van) punten waarvoor J maximaal is. Echter, in het algemeen zal zo'n functie ook een stationair punt hebben dat een zadelpunt is. Ter illustratie zijn enkele niveaukrommem van zo'n functie getekend en geeft de andere figuur een 3-dimensionale voorstelling van een deel van de grafiek van die functie. Het punt Z in deze figuren is een zadelpunt. Het erbij horende punt P op de grafiek zal zeker corresponderen met het juiste gevoel van een zadelpunt in het berglandschap, dat in het dagelijks spraakgebruik een *bergpas* heet: het is het hoogste punt van een weg door het landschap dat tevens het

laagste punt is in vergelijking met punten in een richting "loodrecht" op die weg.

Het karakteriseren (of zelfs alleen al het bewijzen van het bestaan) van een stationair punt dat niet noodzakelijk een extremaal punt is, is i.h.a. moeilijker dan het onderzoek van een minimaliseringsprobleem zoals (1). Hoewel het idee wat ouder is, is pas in 1973 een wiskundig bevredigende manier daarvoor aangegeven: de *mountainpass stelling* van AMBROSETTI en RABINOWITZ. Het idee van deze stelling is eenvoudig te beschrijven voor het berglandschap op het eiland, en is gebaseerd op de *methode van het minimale maximum*. Laat γ een continue weg zijn in het x_1, x_2 -vlak van het punt $\underline{0}$ van het eiland naar het strand S en laat Γ de collectie zijn van al dit soort wegen. Neem voor een weg γ het punt ervan van grootste hoogte:

$$(2) \qquad \max_{\underline{x} \in \gamma} J(\underline{x})$$

Zoek dan dié weg $\hat{\gamma}$ waarvoor die grootste hoogte zo klein mogelijk is, zeg $\hat{c}$:

$$\hat{c} = \min_{\gamma \in \Gamma} \max_{\underline{x} \in \gamma} J(\underline{x}).$$

De stelling zegt dan dat er minstens één stationair punt $\hat{\underline{x}}$ is met $J(\hat{\underline{x}}) = \hat{c}$. Dit stationaire punt kan een extremaal punt zijn van J, maar is in het algemeen een zadelpunt.

(Het bewijs van deze stelling berust ruwweg op het idee dat als geen enkel punt van de optimale weg een stationair punt zou zijn, deze weg in elk punt $\underline{x}$ ervan "gedeformeerd" kan worden (blijvend binnen Γ) door in elk punt van die weg te deformeren in de richting van snelste afname van J (i.e. in de richting waaring een testballetje zou gaan rollen); daardoor zou dan op de gedeformeerde weg de maximale hoogte kleiner zijn dan $\hat{c}$, in strijd met de definitie van $\hat{c}$).

Ofschoon dit, zonder extra eisen op te leggen aan het gedrag van J, de tot nu toe eenvoudigst bekende manier is om een stationair punt, anders dan een extremaal punt, te vinden, heeft deze formulering (2) een ernstig bezwaar.

De verzameling Γ is namelijk oneindig-dimensionaal, en dat lijkt voor ons eenvoudige voorbeeld van een functie van maar twee variabelen, wat veel van het goede. Zeer recent onderzoek, o.a. hier in Nijmegen, richt zich erop een eenvoudiger karakterisering voor zadelpunten te vinden. In het bijzonder, voor berglandschappen met de eigenschap dat, lopend vanuit $\underline{0}$ in een vaste richting er precies één punt is met maximale hoogte, geldt dat (2) essentieel vereenvoudigd kan worden. In feite kan dan zelfs het mini-max probleem teruggebracht worden tot een echt minimaliseringsprobleem van de vorm (1), nl. voor de functie J op de verzameling M bestaande uit de punten van maximale hoogte in de verschillende richtingen. Dit is dan een minimaliseringsprobleem voor een functie van effectief maar één variabele, een aanzienlijke verbetering in vergelijking met (2).

Aan de extra aanname die nodig is om een mini-max probleem als (2) te kunnen reduceren tot een minimaliseringsprobleem van de vorm (1), blijkt door veel problemen uit de mathematische fysica voldaan te worden. Zonder verder de conclusies te bespreken die in de geest van Maupertuis uit zo'n resultaat getrokken kunnen worden, merken we op dat zo'n vereenvoudiging het mogelijk maakt een zadelpunt met numerieke methoden te benaderen, en dat de extra informatie uit het reductieproces in veel gevallen direct relevant is voor het eermee corresponderende probleem uit de mathematische fysica.

Uit de enkele hier beschreven recente ontwikkelingen moge duidelijk zijn dat de variatierekening met zijn roemruchte historie voor de wiskunde alsook voor z'n toepassingsgebieden nog springlevend is en nog steeds in betekenis toeneemt.
Tenslotte, als U niet tot hier bent gekomen met het lezen van dit stuk, hebt U ondertussen Uw optimaliseringsprobleem dat in de eerste regel ter sprake kwam al opgelost; als U wel tot hier bent gekomen wens ik U succes met de verdere oplossing ervan.

ADRESSEN (van de sprekers)

L.J.F. Broer
Sint Jorisstraat 24
5091 SE Middelbeers

P.P.J.E. Clément
Technische Hogeschool te Delft
Onderafdeling der Wiskunde en Informatica
Vakgroep Algemene Wiskunde
Julianalaan 132
2628 BL Delft

C. Cuvelier
Technische Hogeschool te Delft
Onderafdeling der Wiskunde en Informatica
Vakgroep Algemene Wiskunde
Julianalaan 132
2628 BL Delft

E.W.C. van Groesen
Katholieke Universiteit te Nijmegen
Mathematisch Instituut
Vakgroep Toegepaste Analyse
Toernooiveld
6525 ED Nijmegen
(toekomstig adres, vanaf 1 september 1985:
Technische Hogeschool Twente
Onderafdeling der Toegepaste Wiskunde
Postbus 217
7500 AE Enschede)

T. Koetsier
Vrije Universiteit te Amsterdam
Wiskundig Seminarium
Vakgroep Didactiek-Geschiedenis van de Wiskunde en
Relaties tussen Wiskunde en Samenleving
De Boelelaan 1081
1081 HV Amsterdam

J. Ponstein
Rijksuniversiteit Groningen
Subfaculteit Wiskunde
Interfaculteit Econometrie
Hoogbouw WSN, Universiteitscomplex Paddepoel
Postbus 800
9700 AV Groningen

MC SYLLABI

1.1 F. Göbel, J. van de Lune. *Leergang besliskunde, deel 1: wiskundige basiskennis.* 1965.

1.2 J. Hemelrijk, J. Kriens. *Leergang besliskunde, deel 2: kansberekening.* 1965.

1.3 J. Hemelrijk, J. Kriens. *Leergang besliskunde, deel 3: statistiek.* 1966.

1.4 G. de Leve, W. Molenaar. *Leergang besliskunde, deel 4: Markovketens en wachttijden.* 1966.

1.5 J. Kriens, G. de Leve. *Leergang besliskunde, deel 5: inleiding tot de mathematische besliskunde.* 1966.

1.6a B. Dorhout, J. Kriens. *Leergang besliskunde, deel 6a: wiskundige programmering 1.* 1968.

1.6b B. Dorhout, J. Kriens, J.Th. van Lieshout. *Leergang besliskunde, deel 6b: wiskundige programmering 2.* 1977.

1.7a G. de Leve. *Leergang besliskunde, deel 7a: dynamische programmering 1.* 1968.

1.7b G. de Leve, H.C. Tijms. *Leergang besliskunde, deel 7b: dynamische programmering 2.* 1970.

1.7c G. de Leve, H.C. Tijms. *Leergang besliskunde, deel 7c: dynamische programmering 3.* 1971.

1.8 J. Kriens, F. Göbel, W. Molenaar. *Leergang besliskunde, deel 8: minimaxmethode, netwerkplanning, simulatie.* 1968.

2.1 G.J.R. Förch, P.J. van der Houwen, R.P. van de Riet. *Colloquium stabiliteit van differentieschema's, deel 1.* 1967.

2.2 L. Dekker, T.J. Dekker, P.J. van der Houwen, M.N. Spijker. *Colloquium stabiliteit van differentieschema's, deel 2.* 1968.

3.1 H.A. Lauwerier. *Randwaardeproblemen, deel 1.* 1967.

3.2 H.A. Lauwerier. *Randwaardeproblemen, deel 2.* 1968.

3.3 H.A. Lauwerier. *Randwaardeproblemen, deel 3.* 1968.

4 H.A. Lauwerier. *Representaties van groepen.* 1968.

5 J.H. van Lint, J.J. Seidel, P.C. Baayen. *Colloquium discrete wiskunde.* 1968.

6 K.K. Koksma. *Cursus ALGOL 60.* 1969.

7.1 *Colloquium moderne rekenmachines, deel 1.* 1969.

7.2 *Colloquium moderne rekenmachines, deel 2.* 1969.

8 H. Bavinck, J. Grasman. *Relaxatietrillingen.* 1969.

9.1 T.M.T. Coolen, G.J.R. Förch, E.M. de Jager, H.G.J. Pijls. *Colloquium elliptische differentiaalvergelijkingen, deel 1.* 1970.

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.

10 J. Fabius, W.R. van Zwet. *Grondbegrippen van de waarschijnlijkheidsrekening.* 1970.

11 H. Bart, M.A. Kaashoek. H.G.J. Pijls, W.J. de Schipper, J. de Vries. *Colloquium halfalgebra's en positieve operatoren.* 1971.

12 T.J. Dekker. *Numerieke algebra.* 1971.

13 F.E.J. Kruseman Aretz. *Programmeren voor rekenautomaten; de MC ALGOL 60 vertaler voor de EL X8.* 1971.

14 H. Bavinck, W. Gautschi, G.M. Willems. *Colloquium approximatietheorie.* 1971.

15.1 T.J. Dekker, P.W. Hemker, P.J. van der Houwen. *Colloquium stijve differentiaalvergelijkingen, deel 1.* 1972.

15.2 P.A. Beentjes, K. Dekker, H.C. Hemker, S.P.N. van Kampen, G.M. Willems. *Colloquium stijve differentiaalvergelijkingen, deel 2.* 1973.

15.3 P.A. Beentjes, K. Dekker, P.W. Hemker, M. van Veldhuizen. *Colloquium stijve differentiaalvergelijkingen, deel 3.* 1975.

16.1 L. Geurts. *Cursus programmeren, deel 1: de elementen van het programmeren.* 1973.

16.2 L. Geurts. *Cursus programmeren, deel 2: de programmeertaal ALGOL 60.* 1973.

17.1 P.S. Stobbe. *Lineaire algebra, deel 1.* 1973.

17.2 P.S. Stobbe. *Lineaire algebra, deel 2.* 1973.

17.3 N.M. Temme. *Lineaire algebra, deel 3.* 1976.

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.* 1973.

19 A. Hordijk, R. Potharst, J.Th. Runnenburg. *Optimaal stoppen van Markovketens.* 1973.

20 T.M.T. Coolen, P.W. Hemker, P.J. van der Houwen, E. Slagt. *ALGOL 60 procedures voor begin- en randwaardeproblemen.* 1976.

21 J.W. de Bakker (red.). *Colloquium programmacorrectheid.* 1975.

22 R. Helmers, J. Oosterhoff, F.H. Ruymgaart, M.C.A. van Zuylen. *Asymptotische methoden in de toetsingstheorie; toepassingen van naburigheid.* 1976.

23.1 J.W. de Roever (red.). *Colloquium onderwerpen uit de biomathematica, deel 1.* 1976.

23.2 J.W. de Roever (red.). *Colloquium onderwerpen uit de biomathematica, deel 2.* 1977.

24.1 P.J. van der Houwen. *Numerieke integratie van differentiaalvergelijkingen, deel 1: eenstapsmethoden.* 1974.

25 *Colloquium structuur van programmeertalen.* 1976.

26.1 N.M. Temme (ed.). *Nonlinear analysis, volume 1.* 1976.

26.2 N.M. Temme (ed.). *Nonlinear analysis, volume 2.* 1976.

27 M. Bakker, P.W. Hemker, P.J. van der Houwen, S.J. Polak, M. van Veldhuizen. *Colloquium discretiseringsmethoden.* 1976.

28 O. Diekmann, N.M. Temme (eds.). *Nonlinear diffusion problems.* 1976.

29.1 J.C.P. Bus (red.). *Colloquium numerieke programmatuur, deel 1A, deel 1B.* 1976.

29.2 H.J.J. te Riele (red.). *Colloquium numerieke programmatuur, deel 2.* 1977.

30 J. Heering, P. Klint (red.). *Colloquium programmeeromgevingen.* 1983.

31 J.H. van Lint (red.). *Inleiding in de coderingstheorie.* 1976.

32 L. Geurts (red.). *Colloquium bedrijfsystemen.* 1976.

33 P.J. van der Houwen. *Berekening van waterstanden in zeeën en rivieren.* 1977.

34 J. Hemelrijk. *Oriënterende cursus mathematische statistiek.* 1977.

35 P.J.W. ten Hagen (red.). *Colloquium computer graphics.* 1978.

36 J.M. Aarts, J. de Vries. *Colloquium topologische dynamische systemen.* 1977.

37 J.C. van Vliet (red.). *Colloquium capita datastructuren.* 1978.

38.1 T.H. Koornwinder (ed.). *Representations of locally compact groups with applications, part I.* 1979.

38.2 T.H. Koornwinder (ed.). *Representations of locally compact groups with applications, part II.* 1979.

39 O.J. Vrieze, G.L. Wanrooy. *Colloquium stochastische spelen.* 1978.

40 J. van Tiel. *Convexe analyse.* 1979.

41 H.J.J. te Riele (ed.). *Colloquium numerical treatment of integral equations.* 1979.

42 J.C. van Vliet (red.). *Colloquium capita implementatie van programmeertalen.* 1980.

43 A.M. Cohen, H.A. Wilbrink. *Eindige groepen (een inleidende cursus).* 1980.

44 J.G. Verwer (ed.). *Colloquium numerical solution of partial differential equations.* 1980.

45 P. Klint (red.). *Colloquium hogere programmeertalen en computerarchitectuur.* 1980.

46.1 P.M.G. Apers (red.). *Colloquium databankorganisatie, deel 1.* 1981.

46.2 P.G.M. Apers (red.). *Colloquium databankorganisatie, deel 2.* 1981.

47.1 P.W. Hemker (ed.). *NUMAL, numerical procedures in ALGOL 60: general information and indices.* 1981.

47.2 P.W. Hemker (ed.). *NUMAL, numerical procedures in ALGOL 60, vol. 1: elementary procedures; vol. 2: algebraic evaluations.* 1981.

47.3 P.W. Hemker (ed.). *NUMAL, numerical procedures in ALGOL 60, vol. 3A: linear algebra, part I.* 1981.

47.4 P.W. Hemker (ed.). *NUMAL, numerical procedures in ALGOL 60, vol. 3B: linear algebra, part II.* 1981.

47.5 P.W. Hemker (ed.). *NUMAL, numerical procedures in ALGOL 60, vol. 4: analytical evaluations; vol. 5A: analytical problems, part I.* 1981.

47.6 P.W. Hemker (ed.). *NUMAL, numerical procedures in ALGOL 60, vol. 5B: analytical problems, part II.* 1981.

47.7 P.W. Hemker (ed.). *NUMAL, numerical procedures in ALGOL 60, vol. 6: special functions and constants; vol. 7: interpolation and approximation.* 1981.

48.1 P.M.B. Vitányi, J. van Leeuwen, P. van Emde Boas (red.). *Colloquium complexiteit en algoritmen, deel 1.* 1982.

48.2 P.M.B. Vitányi, J. van Leeuwen, P. van Emde Boas (red.). *Colloquium complexiteit en algoritmen, deel 2.* 1982.

49 T.H. Koornwinder (ed.). *The structure of real semisimple Lie groups.* 1982.

50 H. Nijmeijer. *Inleiding systeemtheorie.* 1982.

51 P.J. Hoogendoorn (red.). *Cursus cryptografie.* 1983.

CWI SYLLABI

1 Vacantiecursus 1984 *Hewet - plus wiskunde.* 1984.

2 E.M. de Jager, H.G.J. Pijls (eds.). *Proceedings Seminar 1981-1982. Mathematical structures in field theories.* 1984.

3 W.C.M. Kallenberg, et.al. *Testing statistical hypotheses: worked solutions.* 1984.

4 J.G. Verwer (ed.). *Colloquium topics in applied numerical analysis, volume 1.* 1984.

5 J.G. Verwer (ed.). *Colloquium topics in applied numerical analysis, volume 2.* 1984.

6 P.J.M. Bongaarts, J.N. Buur, E.A. de Kerf, R. Martini, H.G.J. Pijls, J.W. de Roever. *Proceedings Seminar 1982-1983. Mathematical structures in field theories.* 1985.

7 Vacantiecursus 1985 *Variatierekening.* 1985.