CWI Syllabi

CWI
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands
Telephone + 31 - 20 592 9333
Telefax + 31 - 20 592 4199
WWW page http://www.cwi.nl/publications_bibl/

CWI is the nationally funded Dutch institute for research in Mathematics and Computer Science.

Vakantiecursus 2000
Is wiskunde nog wel mensenwerk?

Centrum voor Wiskunde en Informatica
48 CWI SYLLABUS

De Vakantiecursus Wiskunde voor leraren in de exacte vakken in VWO, HAVO en HBO en andere belangstellenden is een initiatief van de Nederlandse Vereniging van Wiskundeleraren. De cursus wordt sinds 1946 jaarlijks gegeven op het Centrum voor Wiskunde en Informatica en aan de Technische Universiteit Eindhoven.

ISBN 90 6196 491 1
NUGI-code: 811

Inhoud

Ten Geleide

Is wiskunde nog wel mensenwerk? Menigeen die de stormachtige ontwikkelingen van de informatietechnologie in de laatste decennia van de afgelopen eeuw gevolgd heeft, zal zich dat wel eens hebben afgevraagd. En de man of vrouw op straat, voor wie wiskunde toch al hetzelfde is als rekenen, maar dan op een moeilijke manier, heeft al lang het idee dat de computer het rekenen, en dus ook alle wiskunde, van de mens overgenomen heeft. Maar ook wiskundeleraren en wiskundigen die in andere sectoren van de maatschappij werkzaam zijn, zullen zich de laatste tijd steeds vaker zijn gaan afvragen welk deel van de wiskunde nog mensenwerk blijft, nu steeds meer van het traditionele handwerk door computers overgenomen wordt.

De vraag ligt daarbij voor de hand of er nog wel werk voor wiskundigen overblijft, met name in het onderwijs. Symbolisch differentiëren en integreren kan de computer immers veel beter dan wij, grafieken tekenen en functieonderzoek plegen is met de computer een fluitje van een cent. Het moeizaam benaderen van kansverdelingen door een normale verdeling en die vervolgens op standaardvorm herleiden, is dankzij de computer een volstrekt overbodige vaardigheid geworden. Voer een willekeurige stelling uit traditionele vlakke euclidische meetkunde in symbolische vorm in het computeralgebrapakket MAPLE in, en een slim deelprogramma[1] antwoordt onmiddellijk 'TRUE' of 'FALSE'.

Is wiskunde nog wel mensenwerk? en: *Wat voor werk is er nog voor wiskundigen in de komende eeuw?* – Dat zijn de centrale vragen van de Vacantiecursus 2000. De sprekers laten zien dat er voor het hierboven geschetste pessimisme geen enkele reden bestaat. De wiskunde is nog steeds springlevend – je zou zonder enige overdrijving kunnen zeggen: springlevender dan ooit, als dat niet zulk lelijk taalgebruik was. Want weliswaar neemt de computer ons steeds meer werk uit handen, maar tegelijkertijd zorgt diezelfde computer voor tal van nieuwe mogelijkheden en uitdagingen, niet alleen in de wetenschap, de techniek en de industrie, maar ook in de klas. Het programma van de Vakantiecursus en de in deze Syllabus verzamelde teksten van de verschillende voordrachten tonen overduidelijk aan dat de wiskunde ons in de nabije toekomst voor een overvloed aan fascinerende, nieuwe opgaven stelt. De nieuwe informatietechnologie zal daarbij belangrijke gereedschappen leveren, maar uiteindelijk blijft wiskunde wel degelijk mensenwerk – ook in de eenentwintigste eeuw!

Gaarne wil ik deze inleiding besluiten met een woord van dank aan allen die hebben bijgedragen aan het welslagen van de cursus. In de eerste plaats natuur-

[1] zie het artikel 'Plane Geometry: An Elementary School Textbook (ca. 2050 AD)' van Shalosh B.Ekhad XIV (pseudoniem van Doron Zeilberger), The Mathematical Intelligencer, 21 (3), 1999, pp. 64-70.

lijk de sprekers, die naast hun lezing ook een schriftelijke neerslag ervan voor deze Syllabus hebben aangeleverd. Het Centrum voor Wiskunde en Informatica te Amsterdam en de Technische Universiteit Eindhoven stelden zaalruimte beschikbaar, de administratieve en praktische organisatie was in handen van mw. Wilmy van Ojik en dr. Miente Bakker, die ook de inhoudelijke coördinatie van de Syllabus verzorgde.

Allen hartelijk dank!

Jan van de Craats

Vakantiecursus 2000

pp. 3 – 20

Interactief onderwijs in de algebra

Hans Sterk
Technische Universiteit Eindhoven

1. Computers: van rekenen naar (re)presenteren

Met de komst van computers voorzien van programma's en interfaces die steeds meer mogelijkheden aan de gebruikers bieden, rijst de vraag wat deze nieuwe technologie kan betekenen voor het doen van wiskunde op de computer in de breedste zin des woords. Het verleden heeft overtuigend laten zien hoe nuttig de *rekenkracht* van computers voor de wiskunde is, maar de rol van computers ten behoeve van *presentatie* van wiskunde (hoe laat je wiskunde met of op de computer zien?)

$$\int_0^\infty e^{-x^2}\,dx$$

Figuur 1. Mooie presentatie.

en *representatie* van wiskunde (hoe sla je wiskunde op de computer op ten behoeve van allerlei gebruik?) komt pas de laatste jaren van de grond.

```
x*y+1
```

Figuur 2. Softwarepakket Maple 'herkent' deze wiskunde-expressie als $xy + 1$.

Het gaat hier natuurlijk niet om gescheiden grootheden: moderne presentatie van wiskunde beoogt onder meer gebruikers integraal diverse rekenmogelijkheden aan te bieden en daarbij speelt impliciet de representatie een grote rol. Centraal staat nu de vraag welke manier van (re)presenteren van wiskunde een meerwaarde heeft boven klassieke (re)presentatievormen, gegeven een aantal technologische innovaties, zoals een toenemend reken- en manipulatievermogen, en niet te vergeten internet. Met andere woorden, hoe structureren we met moderne middelen wiskundedocumenten? Meerwaarde moeten we natuurlijk in het licht zien van een doel, zoals het vlot kunnen communiceren van wiskundeberekeningen, of van een doelgroep, zoals middelbare scholieren, ingenieurs, wiskundigen enz. De vraag is te veelomvattend om hier volledig te bespreken. In plaats daarvan schets ik in deze bijdrage hoe een en ander in de groep waarin ik werk gestalte krijgt en wat we voor de nabije toekomst voorzien.

Bij de leerstoel DAM (Discrete Algebra en Meetkunde) van de Technische Universiteit Eindhoven is in 1999 het cursusmateriaal *Algebra Interactive!* (zie

[3]) in gebruik genomen, bestaande uit een CD-rom en een boek, voor het onderwijs in de algebra aan eerste– en tweedejaars studenten van de ingenieursopleidingen Technische Wiskunde en Technische Informatica. De CD-rom is het hart van *Algebra Interactive!* en maakt gebruik van nieuwe electronische middelen, waarin de bovengenoemde zaken een rol spelen. *Algebra Interactive!* is ontstaan uit een zogenaamd 'studeerbaarheidsproject' van het Ministerie OC & W, maar is gaandeweg een project geworden dat op meerdere (internationale) pilaren steunt en is feitelijk een momentopname van lopende ontwikkelingen. Juist algebra hebben we gekozen, omdat studenten dit vak zo vaak als theoretisch en saai zien. Onze hoop is dat we met *Algebra Interactive!* het onderwijs in de algebra nieuw leven inblazen, dat we met *Algebra Interactive!* de studenten een stimulerende werkomgeving bieden die aansluit bij het multimediale tijdperk waarin zij opgroeien en dat gebruikers overtuigd raken van de fundamentele, maar niet altijd zichtbare rol van de algebra.

In deze bijdrage gaan we in op de verschuiving in de richting van electronische middelen, op veranderende visies ten aanzien van de inhoud van te presenteren wiskunde, op de realisatie bij *Algebra Interactive!* en op toekomstige ontwikkelingen, met name de representatie van wiskunde op de computer. Concreet illustreren we diverse aspecten tijdens de Vakantiecursus met de keuzes die gemaakt zijn bij *Algebra Interactive!* en zijn opvolger(s) in wording.

Mijn mede-auteurs, Hans Cuypers en Arjeh Cohen, wil ik graag bedanken voor hun constructieve opmerkingen bij eerdere versies.

2. Van boeken naar multimedia

Wiskundeboeken en tijdschriftartikelen vormen het klassieke medium om wiskunde vast te leggen en over te dragen. Vorm en inhoud leken daarbij een langdurig verbond te zijn aangegaan: ieder van ons is vroeg of laat wel eens geconfronteerd met de strakke lijn definitie–stelling–bewijs die in oude (leer)boeken jarenlang heeft overheerst.

Definitie. Een geheel getal $n > 1$ is een *priemgetal* als het geen andere positieve delers heeft dan 1 en zichzelf.

Voorbeeld. 3 is een priemgetal want de enige positieve delers zijn 1 en 3.

Stelling. *Er zijn oneindig veel priemgetallen.*

Bewijs. ...

Figuur 3. Klassieke opbouw

In de loop der jaren is de stijl van wiskundeboeken op allerlei manieren aangepast aan veranderende inzichten op het gebied van kennisoverdracht, vaak parallel aan veranderende mogelijkheden die druktechnieken bieden. Denk bij aanpassingen van de stijl die de leesbaarheid kunnen vergroten bijvoorbeeld

aan regelafstanden, lettergrootten en lettertypen, illustraties (tabellen, tekeningen, foto's), omkaderen van tekst, gebruik van kleuren, gebruik van marges, het toevoegen van toepassingen waar men wiskunde aan het werk ziet, het onderbreken van de tekst met vragen en/of opgaven 'om de lezer bij de les te houden', het opnemen van becommentariërende teksten enz. In het licht van de beschikbare moderne middelen komen totaal andere varianten in beeld om wiskundemateriaal te structureren. Met de huidige gigantische productiesnelheid van software is het componeren van een wiskundedocument dat optimaal gebruik maakt van deze software een complexe en uitdagende taak geworden, waarin inhoudelijke en didactische aspecten opnieuw aandacht vragen.

Laten we de ontwikkelingen, voor zover relevant voor de wiskunde, even in grote lijnen doorlopen. Die ontwikkelingen bestrijken het terrein van tekstverwerkers, internet en op wiskundige berekeningen gerichte softwarepakketten.

2.1. Tekstverwerkers voor wiskunde

Met de komst van professionele tekstverwerkers en geavanceerde printers is het aanmaken van een wiskundetekst met allerlei toeters en bellen binnen het bereik van auteurs zelf gekomen (deze tekst is bijvoorbeeld door de auteur in LaTeXgeschreven). Het tekstverwerkingsprogramma LaTeX(en de wat minder uitgebreide versie TeX), zie [8], [9], heeft een ware revolutie in de wiskundewereld teweeggebracht in de zin dat wiskundigen hun artikelen en boeken hiermee veelal zelf haast drukklaar bij een drukker kunnen aanleveren[2]. Zowel de bovengenoemde als vele andere mogelijkheden om teksten qua lay-out aan te passen, zijn aanwezig in dit programma, of zijn met enig programmeerwerk zelf aan deze programma's toe te voegen. Het enorme scala aan mogelijkheden dat nu direct in handen van auteurs is, heeft het oude verbond tussen vorm en inhoud definitief opengebroken. De vraag dringt zich op hoe je deze middelen effectief inzet bij de presentatie van wiskunde. Zo is het verleidelijk teksten te zeer op te smukken met allerlei franje, die bij nader inzien de kwaliteit wellicht schaadt. Functioneel gebruik van het arsenaal aan middelen blijkt een (nieuw) ambacht.

2.2. Internet/World Wide Web

Stond de ontwikkeling van tekstverwerkers nog niet zo prominent in de schijnwerpers, de ontwikkeling van internet gedurende de afgelopen vijf à zes jaar daarentegen krijgt alle mogelijke aandacht van de media en grijpt zichtbaar diep in onze samenleving in. Het World Wide Web of internet (er is wel een

[2] Deze tekstverwerkers zijn in het alledaagse leven niet zo bekend; daar overheersen programma's, die niet primair voor de aanmaak van wiskundedocumenten bedoeld zijn; overigens biedt het programma Word intussen al wel opmaakmogelijkheden voor wiskunde. Curieus aspect van TeX en LaTeX is dat teksten gecodeerd worden ingevoerd, ze behoren daarmee niet tot de wereld van de 'WYSIWYG' (what you see is what you get). In dit opzicht is er een overeenkomst met HTML (zie verderop) dat gebruikt wordt voor de aanmaak van internetdocumenten, al bestaan hiervoor inmiddels editors die het WYSIWYG-aspect hebben teruggebracht.

verschil maar dat is hier niet relevant, we gebruiken beide termen door elkaar, zie [1] voor een indruk van de historie) heeft andere karakteristieken en behelst een arsenaal aan nieuwe mogelijkheden: verwijzingsmogelijkheden ('links' naar andere digitale informatiebronnen), het oproepen van nieuwe vensters, de stuurbaarheid (door de bezoeker van een site), de minder tekstuele oriëntatie van de documenten, de mogelijkheid informatiebronnen te 'downloaden' en de dynamiek in algemene zin: denk aan bewegende figuren, over het scherm wandelende teksten, plaatjes waar je als gebruiker soms zelf iets mee kunt doen, en in- en uitvoervelden waarmee je bijvoorbeeld een database kunt raadplegen of jezelf bij een instantie kunt aanmelden.

2.3. Rekenen op de computer

Ten slotte: rekenkracht op de computer is uitgemond in redelijk gebruikersvriendelijke pakketten, variërend van 'general purpose' computeralgebrapakketten en numerieke paketten als Maple ([13]), Mathematica ([10]) en Matlab ([12]), tot gespecialiseerde pakketten op numeriek en symbolisch terrein. Een voorbeeld van dat laatste is GAP ([6]), bedoeld voor specialistische berekeningen in de groepentheorie.

```
Int(x/(x^2+1), x=0..1): \% = value(\%);
```

$$\int_0^1 \frac{x}{x^2+1}\,dx = \frac{1}{2}\ln(2)$$

FIGUUR 4. Een berekening in Maple

Sommige pakketten pogen al diverse moderne technologieën in zich te verenigen. Maple en Mathematica bijvoorbeeld bieden naast reken- en programmeerfaciliteiten ook de mogelijkheid om sessies als volwaardige documenten aan te bieden en bijvoorbeeld om te zetten in LaTeX. Nadeel is wel dat in alle gevallen pakketafhankelijke kennis vereist is om ermee te kunnen omgaan: een geoefend gebruiker van Maple kan niet zonder meer overweg met Mathematica.

Tekstverwerkers, internet, rekenen op de computer, de uitdaging bestaat eruit in deze veelheid aan mogelijkheden wegen uit te stippelen die tot coherente, wiskundig zinvolle producten leiden. Wat zijn nu de ingrediënten van zo'n nieuw product? Wat willen we met wiskunde doen?

In de eerste plaats willen we natuurlijk wiskundige berekeningen van allerlei soort doen en die berekeningen en resultaten op een deugdelijke manier archiveren en op een begrijpelijke en inspirerende manier aan anderen tonen en met anderen delen. Berekeningen willen we graag door computers laten uitvoeren. Met het World Wide Web dient zich de mogelijkheid aan berekeningen op machines elders te laten uitvoeren. Archivering hoeft ook niet per se lokaal te gebeuren, maar kan aan gespecialiseerde instanties uitbesteed worden, die

digitaal toegankelijk zijn. Communiceren van wiskunde is via email al een stuk versneld ten opzichte van enkele decennia geleden, maar meestal bedient de wiskundige zich dan van een soort LaTeX-achtige pseudotaal om de betekenis over te brengen.

```
Heb je nog naar de vergelijking x^n + y^n = z^n gekeken?
```

FIGUUR 5. Wiskundecommunicatie per email

Mooier is het als je wiskunde-expressies met betekenis en al makkelijk kunt oversturen, niet alleen naar personen, maar ook naar machines, bijvoorbeeld om een berekening te laten uitvoeren. Kortom: hoe kunnen we wiskundige zaken tot hanteerbare objecten op een computer ombouwen (het aspect representatie)?

En van de didactische kant: oude teksten boden wellicht een efficiënte leidraad door een stuk wiskunde, maar lieten impliciet veel aan de lezer over. Ze vormden een toegangsdeur tot een wereld van verwante zaken (voorbeelden, berekeningen, variaties op resultaten, toepassingen), maar konden die wereld hooguit summier betreden, al was het alleen maar om de lijn van het verhaal niet te zeer te onderbreken. De computer nodigt uit allerlei illustraties in te bouwen (meer het aspect presentatie), van statisch tot dynamisch, zonder de voortgang te verstoren, omdat nieuwe technologie het mogelijk maakt materiaal anders aan te bieden, bijvoorbeeld in een meer gelaagde structuur.

In de volgende paragraaf gaan we in op een inhoudelijke verschuiving bij ons materiaal die is ingegeven door bovenvermelde ontwikkelingen en door onze belangrijkste doelgroep: toekomstige ingenieurs. Het betreft een verschuiving naar een meer algoritmische, op concrete berekeningen gerichte aanpak. Ingenieurs zullen beroepshalve veel met door computers uit te voeren berekeningen te maken krijgen en dienen vertrouwd te zijn met de wiskundige achtergronden van de relevante software.

In de paragraaf daarna komt de opbouw van *Algebra Interactive!* aan bod.

3. WISKUNDE: VAN BEWIJZEN NAAR ALGORITMEN

Er is geen koninklijke weg naar de wiskunde antwoordde Euclides aan de koning van Egypte, die zich zo graag wat sneller dan gewone stervelingen de wiskunde eigen wilde maken.[3] Wiskunde leren is, zo luidde de boodschap, voor iedereen een intensieve bezigheid, maar, in een variatie op Euclides, niet een eenduidig vastgelegde activiteit: er zijn vele wegen om je iets eigen te maken. Die wegen hebben gemeen dat je wiskunde moet *doen*. Met de huidige computeralgebrapakketten voorzien van programmeerfaciliteiten krijgen we op een dienblad de mogelijkheid aangereikt de wiskunde te benaderen vanuit een meer algoritmisch standpunt. Inderdaad gaan achter resultaten uit de wiskunde regelmatig berekeningswijzen schuil die niet altijd zo duidelijk uit de formulering blijken.

[3] Deze uitspraak is te vinden in het werk van de Romeinse auteur Proclus (410–485), zie bijvoorbeeld [4], vol. 1, p. 1.

In *Algebra Interactive!* hebben we ervoor gekozen om de wiskunde met nadruk op deze algoritmische kant te presenteren. Dit maakt de beschreven algebra meteen wat operationeler. Bovendien geven we hiermee een concrete invulling aan het ingenieurskarakter van de opleidingen waar we *Algebra Interactive!* gebruiken.

3.1. Delen met rest

Denk maar eens aan het elementaire 'delen met rest' om de gedachten te bepalen. Deling met rest formuleren we in eerste instantie tamelijk abstract als volgt:

Gegeven zijn de gehele getallen a en $b \neq 0$. Dan bestaan er unieke gehele getallen q en r zó dat $a = qb + r$ en $0 \leq r < b$.

Het bewijs van deze stelling kan op diverse abstractieniveaus gegeven worden, maar een concrete variant richt zich tevens op een methode ter bepaling van quotiënt en rest; het bewijs van de uniciteit laten we hierbij buiten beschouwing. Kortom, we willen de bekende staartdeling, zoals bijvoorbeeld de deling

$$\begin{array}{rl} 24/ & 2371 \quad \backslash 98 \\ & \underline{216} \\ & 211 \\ & \underline{192} \\ & 19 \end{array}$$

die leidt tot de conclusie $2371 = 98 \cdot 24 + 19$, dat wil zeggen, $q = 98$ en $r = 19$, tot een (door een machine uitvoerbaar) algoritme ombouwen. Laten we ons gemakshalve tot niet-negatieve getallen a en b beperken. Een constructie ter bepaling van quotiënt q en rest r verloopt stapsgewijs als volgt. Bij het begin geven we q (voorlopig) de waarde 0. We onderscheiden twee stappen.

- Als $a < b$ dan zijn we klaar omdat $q = 0$ en $r = a$ in dit geval voldoen.
- Als $a \geq b$, dan vervangen we a door $a - b$, zetten q voorlopig op 1 en herhalen het procédé met $a - b$ en b. Dat wil zeggen, als $a - b < b$, dan zijn we klaar, en wel met $q = 1$ en $r = a - b$. Als $a - b \geq b$, dan herhalen we ons procédé met $a - 2b$ en b en verhogen q tot 2, etc.

In de optiek van *Algebra Interactive!* gaat het bij deze stelling om een algoritme ter bepaling van het quotiënt en van de rest. Daarop ligt de nadruk, en bij de presentatie speelt dat algoritme een prominente rol. De twee items zijn het voorstadium van een 'echt' algoritme (efficiëntie laten we hier buiten beschouwing), dat er in pseudo-taal zo uit zou kunnen zien:

Input: $a, b \geq 0$
Output: quotiënt q en rest r
$i = 1$;
while $a - ib \geq 0$ **do** $i := i + 1$;
$q := i - 1$, $r = a - qb$

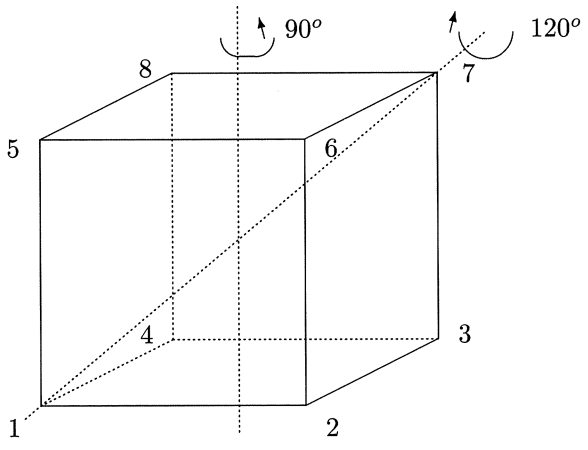
Het existentiegedeelte van de stelling is nu gegoten in de vorm van een algoritme, en het bewijs (dat we hier niet verder toelichten) zou zich kunnen richten op de correctheid van het algoritme. Natuurlijk is de bespreking hier summier en verdienen allerlei facetten in het onderwijsmateriaal nadere toelichting.

Opgave. Hoe moet dit algoritme aangepast worden zodat het werkt voor alle gehele getallen a, b met $b \neq 0$?

Het algoritmisch aspect verlegt de aandacht enigszins naar de vraag: hoe laat ik wiskunde werken, hoe reken ik iets uit, hoe kan ik dat automatiseren en hoe zit het met de correctheid? Iets ambitieuzer nog: welke onderdelen van dit proces kan ik uitbesteden aan de computer? Kun je bijvoorbeeld (correctheids)bewijzen of andere redeneringen in enigerlei vorm automatiseren? Zie ook paragraaf 5.

3.2. Symmetrieën

Het onderwerp groepentheorie uit de algebra leent zich ook uitstekend voor een meer algoritmische aanpak. Dit laat zich goed illustreren aan de hand van het onderzoek naar symmetrieën van meetkundige objecten. *Groepen* vormen het natuurlijke kader om symmetrieën te beschrijven. Groepentheorie tref je vaak in een abstracte vorm aan, maar wanneer je je aanpak richt op algoritmen en berekeningen, dan verandert dit abstracte perspectief zonder verlies van diepgang. Om de gedachten te bepalen, kijken we naar de symmetrieën van een kubus. Overigens zijn in dit geval de berekeningen nog wel met de hand uit te voeren, maar wordt al wel duidelijk dat de hoeveelheid (reken)werk in dergelijke situaties makkelijk flink groot kan zijn.



FIGUUR 6. Symmetrieën van de kubus

De eerste fase bestaat eruit symmetrieën op een voor berekeningen en voor de computer handige wijze te beschrijven. We nummeren de hoekpunten van een

kubus met de getallen 1 tot en met 8 als in de figuur. Draaien we de kubus bijvoorbeeld om de as door de centra van boven- en benedenvlak over 90^o in de aangegeven richting, dan kunnen we deze draaiing vastleggen door te vertellen wat er met de hoekpunten is gebeurd:

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 2 & 3 & 4 & 1 & 6 & 7 & 8 & 5 \end{pmatrix},$$

waarin elke kolom onderin het beeld van het getal bovenin heeft staan. Een draaiing om de as door de hoekpunten 1 en 7 over 120^o geven we weer met

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 1 & 5 & 6 & 2 & 4 & 8 & 7 & 3 \end{pmatrix}.$$

Efficiënter is het om deze *permutaties* in (disjuncte) *cykelnotatie* weer te geven: de eerste draaiing wordt in deze notatie $(1,2,3,4)(5,6,7,8)$ en de tweede wordt $(2,5,4)(3,6,8)$. De permutatie $(2,5,4)(3,6,8)$ lezen we als volgt: de eerste kring, $(2,5,4)$, vertelt dat 2 naar 5 gaat, 5 naar 4, en 4 weer terug naar 2; de tweede kring dat 3 naar 6 gaat, 6 naar 8, en 8 naar 3. Merk op dat hier – weliswaar op papier – de vraag naar de representatie van een permutatie is binnengeslopen! Hoe stel je een permutatie eigenlijk voor?

Opgave. Beschrijf nog enkele symmetrieën met behulp van permutaties. Geef ook hun cykelnotatie. Waarom is

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 2 & 3 & 4 & 1 & 5 & 6 & 7 & 8 \end{pmatrix}$$

geen symmetrie van de kubus? Als σ en τ permutaties zijn (van dezelfde getallen) dan betekent $\sigma\tau$: voer eerst τ uit en dan σ (de samenstelling van σ en τ). Welke permutatie is dan bijvoorbeeld

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 2 & 3 & 4 & 1 & 6 & 7 & 8 & 5 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 1 & 5 & 6 & 2 & 4 & 8 & 7 & 3 \end{pmatrix}?$$

Beschrijf een algoritmische aanpak om een permutatie in cykelnotatie weer te geven. Hoe bepaal je (algoritmisch) de inverse van een permutatie?

Nu we symmetrieën met behulp van (rijtjes) getallen vast hebben gelegd, staat de weg open om deze met een computer te verwerken. We kunnen ons bijvoorbeeld afvragen hoeveel symmetrieën van de kubus er zijn. Omdat er 8! permutaties van de 8 hoekpunten zijn, vormt dit getal een bovengrens. Structuurloos opsommen van de mogelijkheden is natuurlijk een zwakke strategie, een wiskundige niet waardig.

Een fraaie theoretische, maar tamelijk ad hoc, aanpak om dit aantal te bepalen, is de volgende (de preciese details vergen iets meer moeite, het gaat hier even om de grote lijn). De kubus bevat vier lichaamsdiagonalen; in termen van hoekpunten gaat het om de puntenparen $\{1,7\}$, $\{2,8\}$, $\{3,5\}$, $\{4,6\}$. Zo'n puntenpaar bestaat uit twee punten die op maximale afstand van elkaar liggen,

en wel afstand 3 als we de kubus even als graaf opvatten. Omdat deze vier paren alle paren punten op maximale afstand zijn, permuteert elke symmetrie van de kubus deze vier diagonalen. Het is binnen handbereik te laten zien dat alle 24 mogelijke permutaties voorkomen. Nu levert een puntspiegeling in het centrum van de kubus de symmetrie $(1,7)(2,8)(3,5)(4,6)$ op die de lichaamsdiagonalen vasthouden. Door combinatie van deze spiegeling met de eerder genoemde 24 permutaties vind je alle 48 symmetrieën van de kubus.

Een algoritmische aanpak probeert lijn te brengen in zulke telexercities en leidt tot een andere strategie, die we nu bespreken. Merk eerst op dat als we symmetrieën samenstellen (na elkaar uitvoeren), we weer een symmetrie vinden en dat de inverse van een symmetrie uiteraard ook weer een symmetrie is. Dit weerspiegelt de aanwezige groepsstructuur. We zeggen dat de symmetrieën een *groep* vormen, zeg G. Om structuur in onze tellerij aan te brengen, maken we gebruik van het hieronder vermelde resultaat. Het is gebaseerd op de overweging dat we de groep als volgt kunnen opdelen in acht disjuncte deelverzamelingen, corresponderend met de 8 hoekpunten: in het i–de deel zitten die symmetrieën van de kubus die het hoekpunt 1 in het hoekpunt i overvoeren. Het bijzondere van deze opsplitsing is dat elk van deze acht delen even groot is.

Laat G de groep van alle symmetrieën van de kubus zijn en laat G_1 de deelverzameling van permutaties uit G zijn die het hoekpunt 1 vastlaten. Dan geldt voor het aantal elementen $|G|$ van G en $|G_1|$ van G_1:

$$|G| = 8 \cdot |G_1|.$$

Bewijs. In de eerste plaats is het eenvoudig om vast te stellen dat elk van de acht delen niet leeg is: door een geschikte combinatie van de permutaties $(1,2,3,4)(5,6,7,8)$ en $(2,5,4)(3,6,8)$ te gebruiken, kunnen we hoekpunt 1 overvoeren in elk van de acht hoekpunten (de relevantie van deze opmerking blijkt straks). We zeggen wel dat $\{1,2,3,4,5,6,7,8\}$ de *baan* van 1 is. Nu bewijzen we dat elk van de acht delen even groot is als G_1 door een bijectief verband aan te geven. Kies een index $i \in \{1,2,3,4,5,6,7,8\}$ en een permutatie g_i die 1 overvoert in i. Is $g \in G_1$ dan is de samenstelling $g_i \circ g$ een permutatie die 1 in i overvoert: $(g_i \circ g)(1) = g_i(g(1)) = g_i(1) = i$. Dus $g_i \circ g$ behoort tot de permutaties die 1 in i overvoeren. Is omgekeerd h een permutatie die 1 in i overvoert, dan is de samenstelling $g_i^{-1} \circ h$ een permutatie die 1 in 1 overvoert, dus tot G_1 behoort. □

De deelverzameling G_1 is zelf weer een groep: samenstellingen van permutaties die 1 op 1 afbeelden, hebben zelf die eigenschap ook weer, evenals inversen van dergelijke permutaties. De andere zeven deelverzamelingen vormen geen groep; de situatie is vergelijkbaar met de opsplitsing van de gehele getallen $\mathbb{Z}$ in de drievouden enerzijds en de drievouden plus 1 en de drievouden plus 2 anderzijds. We zeggen dat G_1 een *ondergroep* is van G. Ook heet G_1 wel de *stabilisator* van 1 onder G. De permutaties van G_1 permuteren eigenlijk alleen nog maar de getallen $2, \ldots, 8$. We passen nu dezelfde strategie als daarnet toe

op G_1: we delen G_1 op in 7 disjuncte stukken: voor $i = 2, \ldots, 7$ bestaat het i-de stuk uit die permutaties in G_1 die hoekpunt 2 naar hoekpunt i sturen. In dit geval treedt er een complicatie op. Omdat 2 vastzit aan hoekpunt 1, kan hoekpunt 2 hooguit naar de hoekpunten 4, 5 en 2 zelf gaan. Er blijken nu 3 niet-lege stukken te zijn, maar die zijn wel alledrie even groot (de baan van 2 onder G_1 is $\{2, 4, 5\}$); dit volgt met een redenering als boven. Geven we met $G_{1,2}$ de ondergroep van G_1 aan bestaande uit die permutaties die zowel 1 als 2 vastlaten, dan geldt dus:

$$|G_1| = 3 \cdot |G_{1,2}|.$$

In de volgende stap vinden we op soortgelijke wijze $|G_{1,2}| = 2 \cdot |G_{1,2,3}|$.

Opgave. Naar welke hoekpunten kan het hoekpunt 3 nog gaan als 1 en 2 vast zijn? Kunt u permutaties aangeven die dit realiseren?

Bij $G_{1,2,3}$ stopt het proces omdat er maar één symmetrie is van de kubus die zowel 1, 2 als 3 vastlaat: dat is de permutatie die alle hoekpunten vastlaat. Alles bij elkaar vinden we dus 48 symmetrieën:

$$|G| = 8 \cdot |G_1| = 8 \cdot 3 \cdot |G_{1,2}| = 8 \cdot 3 \cdot 2 \cdot |G_{1,2,3}| = 48.$$

Bovenstaande berekeningswijze is algemeen toepasbaar bij het analyseren van symmetrieën, niet alleen om het aantal symmetrieën te bepalen, maar ook om de 'fijnere' structuur van een groep bloot te leggen.

Welke algoritmische aspecten zitten verborgen in deze tellerij? Boven wezen we er al op dat samenstellingen en inversen van symmetrieën weer symmetrieën zijn. We starten nu met enkele permutaties waarvan we denken dat we daarmee behoorlijk veel, zo niet alle, symmetrieën van de kubus door samenstelling kunnen opbouwen[4]. Het zijn de permutaties

$$a = (2,5,4)(3,6,8), \quad b = (1,2,3,4)(5,6,7,8), \quad c = (1,4)(2,3)(5,8)(6,7).$$

De groep van permutaties die hieruit door samenstelling opgebouwd kan worden, noemen we Γ (de permutaties a, b, c heten dan *voortbrengers* van Γ) en we schrijven hiervoor $\Gamma = \langle a, b, c \rangle$.

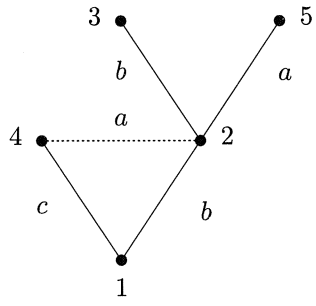
Gegeven enkele symmetrieën van een object zoals de kubus, weergegeven als permutaties, kan de vraag naar het totale aantal symmetrieën dat daarmee door samenstelling te verkrijgen is, als volgt algoritmisch benaderd worden. Er moeten twee typen berekeningen gedaan worden:

- berekeningen van *banen* (de baan van ω bestaat uit de verzameling beelden van ω onder alle elementen van de groep),
- berekeningen van *stabilisatoren.*

[4] De reden dat we naast a en b nog een derde element gebruiken, is dat a en b beide rotaties voorstellen. Samenstellingen van rotaties zijn weer rotaties, dus daarmee kunnen we geen spiegelingen construeren. Het element c is een spiegeling.

Om de baan van een element ω (denk aan het hoekpunt 1 bij de kubus) te bepalen, bereken je het beeld van ω onder elke voortbrenger van G. Dit levert mogelijk een aantal nieuwe elementen op. Vervolgens bereken je het beeld van elk van deze elementen onder iedere voortbrenger van G, enz. Dit is het grondidee van het algoritme. Met een handige boekhouding kunnen banen nog wat efficiënter bepaald worden.

Om stabilisatoren te bepalen, leggen we de informatie zoals bij het berekenen van banen verkregen vast in een graaf (een *Schreier–boom*). Om de gedachten te bepalen illustreren we dit in het geval van de kubus en groep $\Gamma = \langle a, b, c \rangle$. Maak een graaf (in feite een boom) met wortel 1. De beelden $\neq 1$ van 1 onder de voortbrengers a, b en c van Γ worden nieuwe hoekpunten van de graaf verbonden met 1 (komt een beeld meerdere malen voor, dan wordt alleen de eerste keer geadministreerd). De kanten labellen we met een a, een b of een c, de corresponderende voortbrenger. Deze gegevens laten zich administreren middels een algoritme. Uit deze graaf lezen we voortbrengers af van de stabilisator Γ_1 van 1 op de volgende manier (zie Figuur 7).



FIGUUR 7. Stukje Schreier–boom met wortel 1 (stippellijn hoort niet bij de boom)

Bij de constructie van de Schreier–boom administreerden we alleen de eerste keer dat een beeld voorkomt in de graaf. De andere keren leiden tot 'bruggen', waarvan er één gestippeld is in Figuur 7. Uit de brug halen we het element $b^{-1}ac$ dat 1 vastlaat (start in de wortel, ga met c naar 4, dan met de brug a naar 2 en vervolgens met b^{-1} terug naar de wortel). Ter controle:

$$b^{-1}ac(1) = b^{-1}a(4) = b^{-1}(2) = 1.$$

Het is een stelling dat je met behulp van dergelijke bruggen een voortbrengend stelsel permutaties krijgt van de stabilisator, in dit geval Γ_1. Dergelijke bruggen zijn ook weer op een algoritmische manier uit de data van de Schreier–boom te halen. Het voert te ver daarop in te gaan (evenals op de complexiteit).

Algoritmen voor het type berekeningen als hier besproken, zijn bijvoorbeeld geïmplementeerd in het pakket GAP, speciaal ontworpen voor berekeningen met groepen, maar niet in pakketten als Maple en mathematica. Zoals ik

boven al enigszins heb aangegeven, zijn deze algoritmen gebouwd op een stevig fundament van wiskunderesultaten.

Overigens kun je de hier geschetste technieken ook gebruiken om 'Rubik's cube' en verwante spelletjes te analyseren.

Opgave. Laat zien dat Γ uit alle symmetrieën van de kubus bestaat in de volgende stappen: (a) de baan van 1 bestaat uit alle acht hoekpunten; (b) de baan van 2 onder Γ_1 bestaat uit drie hoekpunten; (c) de baan van 3 onder de stabilisator $\Gamma_{1,2}$ van 1 en 2 bestaat uit twee elementen. [Hint: gebruik het element *abc*.]

Opgave. Bij de bespreking van symmetrieën van de kubus zijn we voortvarend over een subtiel aspect heengelopen: bedoelen we met de kubus een graaf, in dit geval een verzameling van 8 hoekpunten met kanten als in de figuur aangegeven? Een symmetrie is dan een bijectie van de hoekpunten met de eigenschap dat als twee hoekpunten verbonden zijn hun beelden ook verbonden zijn. Of bedoelen we met de kubus de gehele verzameling bestaande uit hoekpunten, zijden, zijvlakken, inwendige? In dat geval bedoelen we met symmetrieën lineaire transformaties van de ruimte waarvan de beperking tot de kubus een bijectie oplevert van de kubus op zichzelf. Ga na dat beide standpunten op hetzelfde neerkomen: elke symmetrie van de ene soort hoort bij een unieke symmetrie van de andere soort.

Opgave. Bereken het aantal symmetrieën van de hyperkubus in $\mathbb{R}^4$ op de bovenbeschreven manier. Deze hyperkubus heeft 2^4 hoekpunten, bijvoorbeeld beschreven door $(\pm 1, \pm 1, \pm 1, \pm 1)$.

4. De opbouw van *Algebra Interactive!*

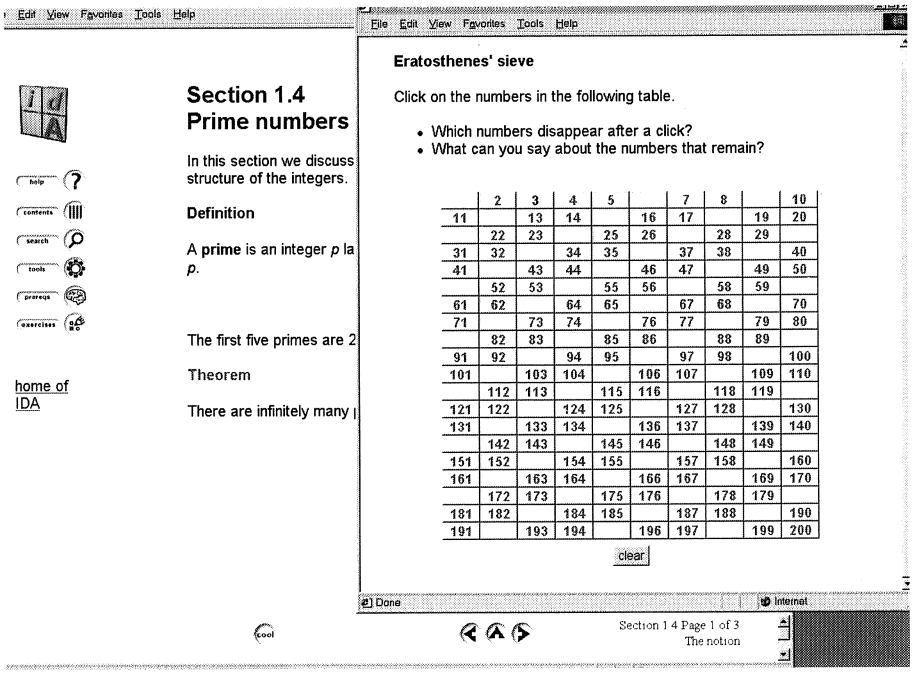
Met de one-liner 'Wiskunde is overal' in het hoofd is het geen gekke gedachte gebruik te gaan maken van een technologische ontwikkeling die informatie op talloze plaatsen op de wereld met elkaar verbindt: het World Wide Web. Deze mogelijkheden beperken zich niet alleen tot statische informatie, maar strekken zich, in principe, ook uit tot dynamische activiteiten als het uitbesteden van een berekening aan een machine elders of het binnenhalen van programmatuur. Bovendien is het World Wide Web zo langzamerhand een door iedereen gebruikt medium. Dat is voor ons de reden geweest in deze ontwikkeling te participeren en uit te gaan van een webdocument.

4.1. Textuele aspecten

Net als andere documenten op internet, bekijk je *Algebra Interactive!* met een *webbrowser*, bijvoorbeeld een versie van Netscape of van MicroSoft Internet Explorer[5]. Webbrowsers verenigen diverse mogelijkheden in zich: een browser

[5] De ontwikkeling van webbrowsers in de laatste 5 à 6 jaar vormt een schitterend voorbeeld van de manier waarop informatietechnologie (IT) en geld in Silicon Valley hand in hand gaan. Netscape maakte de ontwikkelaars, Jim Clarck en de door hem van de universiteit geplukte Marc Andreessen, in een mum van tijd schatrijk (zie [5] voor een relaas van

is een venster op de internetwereld van documenten die overal op de wereld op machines aanwezig zijn; een browser biedt editor-faciliteiten, zodat je er documenten geschikt voor het internet mee kunt aanmaken; een browser biedt zoekmogelijkheden, biedt mogelijkheden om documenten van het net te downloaden voor eigen gebruik en nog veel meer. De basisstandaard waarin de documenten op het net geschreven zijn is (Dynamic) HTML (HyperText Markup Language). Het is geen programmeertaal, maar enkel een manier om gegevens betreffende de opmaak van teksten vast te leggen; browsers zijn in staat deze gegevens op de bedoelde manier weer te geven. Naast de mogelijkheid platte tekst te produceren, levert HTML ook een scala aan verwijsmogelijkheden, de zogenaamde *hyperlinks* of eenvoudigweg links, naar plekken in en buiten het eigen document. Tegenover de mogelijkheden die browsers bieden staat ook een in het oog springend nadeel: het weergeven van wiskunde-expressies in HTML moet meestal gebeuren met een vervelende kunstgreep: wiskundesymbolen geef je dan weer door kleine plaatjes die relatief veel geheugen gebruiken. In nieuwere varianten (XML, MathML) van HTML worden nadelen als deze gedeeltelijk ondervangen. Het zal wel niet lang duren voor de opmaak van wiskunde-expressies geen probleem meer is.



FIGUUR 1. Een snapshot van een 'bladzijde' en een applet uit *Algebra Interactive!*[3]

dichtbij). Gestimuleerd door dit succes begaf de concurrent MicroSoft zich op dezelfde markt met Internet Explorer.

In *Algebra Interactive!* zijn door het gebruik van vensters hoofdlijnen en detailzaken in een gelaagde structuur ondergebracht: hoofdzaken kom je het eerst tegen bij bladeren, terwijl bewijzen, voorbeelden en andere illustraties een niveau dieper zijn gelegen. Naast de gebruikelijke navigatiemogelijkheden zijn verschillende andere, op het document toegesneden, manieren van navigeren opgenomen. Rekenaspecten worden direct geïllustreerd met *gapplets* (zie onder) die tot experimenteren met een bepaald type berekening aanzetten. Illustraties van concepten zijn toegevoegd in de vorm van *Java applets* (zie onder), die soms het karakter van een spel hebben. Ook in de vorm van applets zijn opgenomen rekenmachines voor elke samenhangende groep rekentechnieken (rekenen met getallen, rekenen met polynomen, rekenen met permutaties). Definities en resultaten gaan vergezeld van een multiple choice vraag die fungeert als eerste begripstest. Naast een collectie open opgaven is verder een database van multiple choice vragen opgenomen waaruit op twee manieren toetsen gegenereerd kunnen worden: toetsen over een hoofdstuk naar keuze of toetsen over de stof tot en met een hoofdstuk naar keuze. Deze toetsen worden random gegenereerd en online geëvalueerd.

4.2. Dynamiek: applets

De dynamiek in webpagina's is grotendeels te danken aan de door het computerbedrijf SUN ontwikkelde programmeertaal *Java* [7]. Deze taal stelt je in staat programma's in te bouwen in internetdocumenten, zodat deze tezamen met een webpagina via het internet veilig binnengehaald en uitgevoerd kunnen worden zonder complicaties bij het gebruik van verscheidene platforms (de eerlijkheid gebiedt te zeggen dat de realiteit wat weerbarstiger is dan de claims van SUN suggereren). Zulke programma's heten *applets*. De programmeerfaciliteiten van Java zijn uitermate geschikt voor visuele aspecten: denk aan bewegende plaatjes. Eenvoudiger, aan Java verwant, is *JavaScript*, dat veel toegankelijker is voor de beginnende gebruiker. In *Algebra Interactive!* hebben we deze middelen gebruikt om online dynamische illustraties van concepten te tonen, niet alleen dynamisch in de zin dat er beweging in zit, maar vooral ook in de zin dat je als gebruiker mede stuurder bent in het geïllustreerde proces; dit maakt de lezer tot speler, experimentator en ontdekker. Deze mogelijkheden waren mede bepalend bij het besluit op HTML-basis te werken. Een andere vergelijkbare aanpak zou de schier onmogelijke taak van het ontwikkelen van relevante software bij de auteurs hebben gelegd (bij de oorspronkelijke oriëntatie op de mogelijkheden is een dergelijke poging binnen onze groep wel gedaan).

4.3. Berekeningen: gapplets

Ten slotte berekeningen en algoritmen. Voor vrijwel elk type berekening, vaak gekoppeld aan een algoritme, is er een *gapplet* toegevoegd. De gebruiker heeft

alleen met een in- en uitvoerveld te maken en met wat becommentariërende tekst. Een aantal berekeningen berust op serieuze wiskunde en kan daarom niet eenvoudig in Java geprogrammeerd worden. We hebben hiervoor het computeralgebrapakket GAP ingeschakeld (gapplet is een samentrekking van GAP en applet). Vanuit de webbrowser wordt een verbinding gelegd met dit computeralgebrapakket. Bij de opzet van de gapplets wilden we bereiken dat de gebruiker geen specifieke kennis van GAP nodig heeft; GAP fungeert als zogenaamde *backengine*. De verbinding met GAP vanuit een browser was een hobbel bij de ontwikkeling omdat hierbij invoer in een webdocument moet worden omgezet in invoer voor GAP (en omgekeerd uitvoer van GAP in HTML omgezet moet worden). Een dergelijke verbinding, die voor communicatie van wiskunde binnen een systeem zorgt, is het prototype van de problemen die zich bij communicatie van wiskunde met de computer voordoen.

5. Wiskunde op de computer

Op het starship Enterprise uit de bekende tv-serie worden allerlei gecompliceerde technische handelingen vanachter beeldschermen of vanuit iets nog geavanceerders in gang gezet. Blijkbaar vindt er tussen allerlei instanties moeiteloos communicatie plaats om berekeningen te doen, om (onderdelen van) het ruimtevaartuig in beweging te zetten, om databanken te raadplegen enz.

Anno 2000 zijn we in het stadium aanbeland waarin we ons daadwerkelijk kunnen bezighouden met het integreren van diverse nieuwe middelen tot een flexibele wiskunde–omgeving. Wat zijn de mogelijkheden waar we aan denken en aan werken? Hier volgt een summiere indruk.

5.1. Computeralgebragebruik

Berekeningen en algoritmen besteden we bij voorkeur uit aan bijvoorbeeld een computeralgebrasysteem. Tot op heden vereist dit kennis van het gebruikte pakket, zelfs voor de meest elementaire zaken, en de aanwezigheid van een dergelijk pakket. Dit kan vervelend zijn omdat het ene pakket wel routines voor een type berekening heeft en een ander pakket niet. Denk maar aan de symmetrieberekeningen waarvoor GAP wel routines kent, maar Maple, Mathematica en MatLab niet. Ook vervelend als je misschien voor de zekerheid de correctheid van een antwoord van het ene pakket wilt testen met een ander pakket. De aspecten wiskunde construeren, wiskunde testen, wiskundige experimenten uitvoeren, wiskundige informatiebronnen raadplegen, zijn erbij gebaat als je gebruik kunt maken van verschillende pakketten, wellicht op diverse locaties, zonder al die pakket-specifieke kennis. De interface in *Algebra Interactive!* met GAP is een eerste zeer minieme stap in de richting computeralgebrapakketten makkelijker te kunnen benaderen.

5.2. OpenMath

Binnen het project OpenMath ([14], zie ook [2]) werken we aan het betekenis geven aan wiskundige expressies zodat ze hanteerbaar worden voor gebruik op

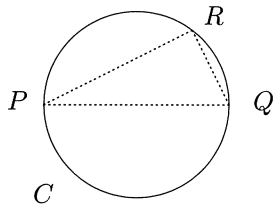
de computer en in het bijzonder uitgewisseld kunnen worden tussen allerlei software tools. (Het al eerder genoemde MathML ([11]) houdt zich voornamelijk bezig met presentatie van wiskundige objecten.) Je wilt bijvoorbeeld dat de expressie $\exp(x)$ herkend wordt als de exponentiële functie losgelaten op x, terwijl een andere interpretatie zou kunnen zijn dat het gaat om de expressie x. Of denk aan 7/4: bedoel je de breuk of heb je quotiënt 1 en rest 3 voor ogen? Het vastleggen van de betekenis is een zorgvuldige (en tijdrovende) bezigheid die in het project OpenMath gestalte krijgt. Idealiter kan een wiskundig object uit een of ander softwarepakket omgezet worden in een OpenMath object (de kernwoorden bij deze vertaalslag zijn *Phrasebooks* en *Content Dictionaries*, waarbij een Phrasebook een vertaling uitvoert en daarbij Content Dictionaries consulteert) en vervolgens opgepikt worden door een ander softwarepakket. In OpenMath gecodeerde objecten kunnen niet alleen uitgewisseld worden tussen softwarepakketten, maar kunnen ook door browsers getoond worden, kunnen getest worden op wiskundige deugdelijkheid en gebruikt worden bij het aanmaken van interactieve documenten.

De bestaande versie van *Algebra Interactive!* is nog zonder OpenMath gemaakt, maar in de nieuwe editie hopen we het concrete gebruik van OpenMath in een wiskundedocument te kunnen demonstreren.

5.3. Automatische bewijzen in de meetkunde

Van 'bewijzen verzinnen' en 'bewijzen verifiëren' is de eerste categorie ongetwijfeld het lastigst. In beide gevallen kan men zich afvragen of er automatiseringsmogelijkheden zijn. Dit is een behoorlijk ambitieus onderwerp. Ik wil me dan ook niet wagen aan een bespreking, maar me beperken tot een voorbeeld waarin algebra op de computer een rol kan spelen bij het vinden van bewijzen voor stellingen uit de vlakke meetkunde. De automatiseringsslag in de meetkunde waar we op doelen, bestaat eruit meetkundige beweringen en constructies, zo mogelijk, om te zetten in polynomiale vergelijkingen en vervolgens de (computer)algebra het werk te laten doen. Dit laat zich het best illustreren met een eenvoudig voorbeeld.

Teken een cirkel C en trek een middellijn PQ. Kies een derde punt R op de omtrek van de cirkel. De bewering is dat PR en QR loodrecht op elkaar staan.



FIGUUR 9. PR staat loodrecht op QR

Een cirkel met straal r en middelpunt in de oorsprong beschrijven we met de vergelijking $x^2 + y^2 - r^2 = 0$. De antipodale punten $P = (p, q)$ en $Q = (r, s)$ voldoen aan

$$p^2 + q^2 - r^2 = 0, \quad p = -s, \quad q = -t.$$

Omdat $R = (u, v)$ ook op de cirkel ligt, levert dit de vergelijking $u^2+v^2-r^2 = 0$. De claim is dat PR en RQ loodrecht op elkaar staan. In coördinaten vertaalt dit in

$$(u - p)(u - s) + (v - q)(v - t) = 0.$$

Het komt er nu op aan te laten zien dat deze vergelijking uit de vier hypothesevergelijkingen volgt. In de algebra is er een handige techniek om zo'n bewering na te gaan. In één variant daarvan komt het erop neer te laten zien dat de uitdrukking $(u - p)(u - s) + (v - q)(v - t)$ 'polynomiaal' is op te bouwen uit de vier uitdrukkingen $p^2 + q^2 - r^2, p + s, q + t, u^2 + v^2 - r^2$. Anders gezegd, de uitdrukking $(u - p)(u - s) + (v - q)(v - t)$ moet met behulp van deze vier uitdrukkingen tot 0 af te breken zijn. In elk van de volgende stappen breken we de uitdrukking een stukje verder af:

$$\begin{aligned} u^2 - u(p + s) + v^2 - v(q + t) + ps + qt &\rightarrow r^2 + ps + qt \\ &\rightarrow r^2 + p(p + s) - p^2 + \\ &\quad\ q(q + t) - q^2 \\ &\rightarrow r^2 - p^2 - q^2 \\ &\rightarrow 0. \end{aligned}$$

Bijvoorbeeld, bij de eerste overgang hebben we $u^2 + v^2$ door r^2 vervangen en $p+s$, $q+t$ beide door 0. Ofschoon dat uit deze berekening niet blijkt, kan deze berekening geheel geautomatiseerd worden. De gedachte achter het leveren van bewijzen in deze context op een geautomatiseerde manier is de volgende. Uit een database haal je vergelijkingen die bepaalde standaardsituaties beschrijven, zoals 'rechte door twee punten', 'lijnstukken zijn even lang', 'punt ligt op gegeven cirkel'. Vervolgens giet je de te bewijzen bewering in de vorm van een vergelijking (als dat kan), en ten slotte zet je geautomatiseerde algebraïsche technieken in om de bewering te bewijzen of te weerleggen. Varianten hiervan kunnen in principe ook gebruikt worden om stellingen te ontdekken. Het zal u niet ontgaan zijn dat met wat meetkundig inzicht vereenvoudigingen zijn aan te brengen in bovenstaand voorbeeld, maar in complexere situaties is dat wellicht niet meer mogelijk.

Ideaal zou het zijn als je meetkundige configuraties op de computer kunt tekenen en uit deze plaatjes volautomatisch vergelijkingen in een of ander coördinatensysteem kunt genereren. Zulke ontwikkelingen zijn wel gaande.

Overigens vergt het wel enige studie om de wereld van de algebra van polynomen te betreden; de technieken achter het bovenstaande voorbeeld komen dan ook pas in een vervolgcursus algebra aan de orde en niet in *Algebra Interactive!*. Daarin is echter wel een inleiding in het werken met polynomen te vinden.

6. Tot slot

De ontwikkelingen gaan snel: naast universitaire onderzoekers begeven namelijk industriële spelers met flinke investeringen zich op diverse terreinen die we boven hebben aangeroerd. Natuurlijk richt de industrie zich primair op die aspecten waar ze een reusachtige markt vermoeden. Dat laat voor wiskundigen de taak de nieuw geproduceerde middelen naar hun eigen inzichten te combineren. Onvermijdelijk vergt dat links en rechts toch nog ontwikkelingswerk omdat de industrie natuurlijk niet in eerste instantie aan wiskundige gebruikers denkt. Overigens volgen de technologische innovaties elkaar zo snel op dat ongetwijfeld binnen enkele maanden na verschijnen van deze bijdrage diverse onderwerpen in een ander licht zijn komen te staan. Het illustreert ook, zoals we uit ervaring weten, dat wiskunde op de computer een zo omvangrijk terrein is dat enkel in samenwerkingsverbanden vooruitgang is te boeken. Wiskundigen wacht een boeiende taak bij het realiseren van een flexibele wiskundige werkomgeving.

Referenties

1. T. Berners-Lee (2000). *De wereld van het World Wide Web.* Uitgeverij Nieuwezijds, Amsterdam (vertaling van *Weaving the Web - The original design and ultimate destiny of the World Wide Web by its inventor*, uitgegeven in 1999 bij Orion Business Press, London (UK) en HarperSanFrancisco (US))
2. A.M. Cohen. Communicating Mathematics across the Web. Te verschijnen in: Wilfried Schmid, Björn Engquist (eds.). *Mathematics Unlimited – 2001 and Beyond.* Springer-Verlag, Berlin, New York, etc.
3. A.M. Cohen, H. Cuypers, H. Sterk (1999). *Algebra Interactive.* Springer Verlag, Berlin, New York, etc. Voor een demo-versie, zie `http://www.win.tue.nl/~ida/inhoud.html`
4. T.L. Heath (1956). *The thirteen books of Euclid's elements.* Dover, New York.
5. M. Lewis (2000). *Het nieuwste van het nieuwste.* Uitgeverij Balans, Amsterdam (vertaling van *The new new thing*, uitgegeven door W.W. Norton & Company, New York).
6. GAP: `http://www-groups.dcs.st-and.ac.uk/~gap`, 2000
7. Java: `http://www.javasoft.com/`
8. D.E. Knuth (1984). *The Texbook.* Addison-Wesley, Reading, Mass.
9. L. Lamport (1986). *LATEX: a document preparation system.* Addison-Wesley. Reading, Mass.
10. Mathematica: `http://www.wolfram.com/products/mathematica`, 2000
11. MathML: `http://www.w3.org/Math/`
12. Matlab: `http://www.mathworks.com/products/matlab/`, 2000
13. Maple: `http://www.maplesoft.com/products`, 2000
14. OpenMath: `http://www.nag.co.uk/projects/openmath/omsoc`
15. Vakantiecursus 1994: *Computeralgebra.* CWI-syllabus 36.

Vakantiecursus 2000 pp. 21 – 40

Rekenen aan beelden: is een plaatje duizend woorden waard?

Henk J.A.M. Heijmans
Centrum voor Wiskunde en Informatica
Kruislaan 413, 1098 SJ Amsterdam
email: henkh@cwi.nl

1. Inleiding en motivatie

Informatie komt steeds vaker tot ons in de vorm van beelden. Dit is het geval in de medische diagnostiek waar artsen over steeds betere beeldvormingstechnieken beschikken, o.a. dankzij steeds geavanceerdere MRI en CT-scanners. In een groot aantal produktie-processen worden tussen- en eindprodukten vaak aan geautomatiseerde visuele inspectie onderworpen. Op het militaire vlak probeert men steeds vaker wapensystemen uit te rusten met computers welke een zekere mate van 'visuele intelligentie' bezitten. Een steeds groter aantal satellieten zendt een almaar toenemende hoeveelheid visuele data naar de aarde, die vervolgens geanalyseerd en opgeslagen moet worden.

Maar ook in het dagelijks leven speelt beeldinformatie, en dan met name in digitale vorm, een steeds belangrijkere rol. Daarbij kan men natuurlijk in de eerste plaats aan internet denken, maar ook digitale fotografie en video maken hun opmars en digitale televisie staat op het punt onze huiskamers te veroveren.

Deze en andere toepassingen hebben geleid tot een breed scala van onderzoeksvragen op het gebied van de digitale beeldverwerking, de computervisie en de patroonherkenning. Om er een paar te noemen: hoe kunnen we de kwaliteit van een beeld verbeteren? Hierbij kan de kwaliteit om allerlei redenen te wensen overlaten; het contrast kan te laag zijn, er kan bewegingsonscherpte voorkomen, het beeld kan door ruis verstoord zijn, enz. Welke informatie bevat een beeld en hoe kan men deze informatie aan een beeld onttrekken? Hoe houden we opslagruimte en transmissietijden van beelddata binnen de perken?

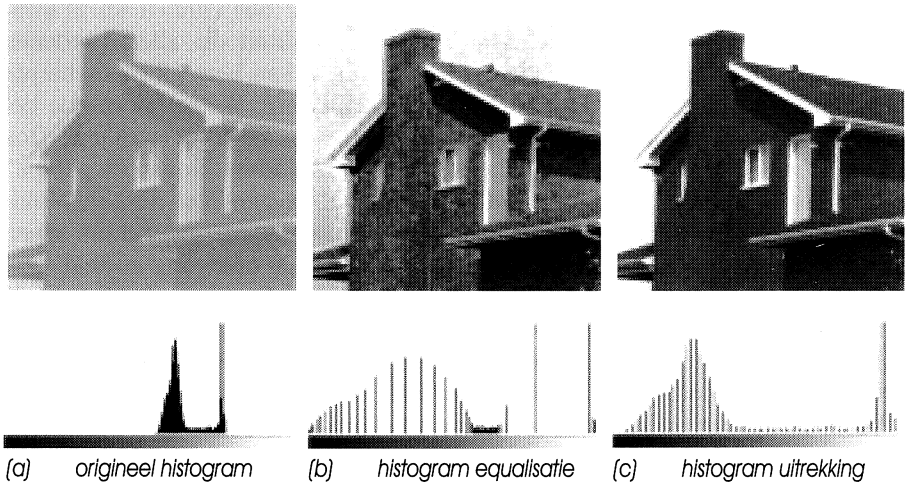
Bij het aanpakken van deze en talloze andere problemen in de digitale beeldverwerking speelt wiskunde van oudsher een belangrijke rol, een rol die alleen maar belangrijker wordt naarmate de eisen die aan de oplossing gesteld worden steeds hoger worden. In deze bijdrage zullen we een paar problemen uit de digitale beeldverwerking de revue laten passeren. Daarnaast zullen we kort ingaan op één specifieke beeldverwerkingsmethodiek, de mathematische morfologie, een methodiek die sterk meetkundig georiënteerd is en derhalve bij uitstek geschikt voor het extraheren van geometrische informatie uit beelden. Het spreekt voor zich dat we in zo'n kort tijdsbestek nergens echt diep op in kunnen gaan, maar niettemin hopen we dat we er in zullen slagen de lezer een globale indruk te geven van de onderliggende principes.

Een bekend Engels spreekwoord zegt: 'A picture is worth a thousand words'. Dat klinkt wellicht voor de hand liggend, maar een 'intelligente' computer zou daar onmiddelijk tegen inbrengen dat een 'picture' veel meer ruimte inneemt dan 'thousand words', en dat de informatie voor een computer nou niet onmiddelijk voor het oprapen ligt. Ook beeldverwerking, zo zal blijken, is voor een belangrijk deel mensenwerk ....

2. Elementaire beeldverwerking

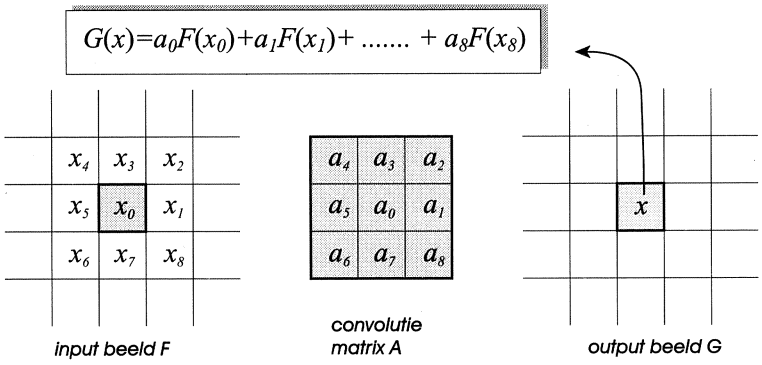
Het uiteindelijke doel van een groot aantal beeldverwerkingstechnieken is het (helpen) interpreteren van de inhoud van een beeld. Bij dit alles speelt het proces van visuele perceptie een belangrijke rol. Alvorens een beeld op een computerscherm te tonen is het van belang te weten hoe het menselijk visueel systeem omgaat met verschillen in helderheid. Door het eenvoudigweg variëren van de helderheids-representatie van een beeld kan men grote visuele verschillen bewerkstelligen. We lichten dit aspect toe aan de hand van een voorbeeld.

We zullen echter eerst een wiskundige definitie van een beeld geven. We beperken ons voor het gemak tot beelden op een discreet domein, bijvoorbeeld een 256×256 vierkant in $\mathbb{Z}^2$. Punten in dit domein worden *pixels* genoemd, een afkorting van 'picture elements'. Een beeld is nu een functie van het domein in een grijswaarden-verzameling; we beperken ons hier tot een eindige één-dimensionale grijswaarden-verzameling, zeg $\{0, 1, \ldots, n\}$. Het *histogram* van een digitaal beeld is de functie die aan elke grijswaarde $t \in \{0, 1, \ldots, n\}$ de waarde $p(t) = N_t/N$ toekent. Hierin is N_t het aantal pixels met grijswaarde t en N het totaal aantal pixels. We kunnen de functie p dus interpreteren als een soort kansverdeling over de verschillende grijswaarden. In Figuur 1(a) zien we een afbeelding met een laag contrast. Alle grijswaarden bevinden zich



Figuur 1. Een tweetal beeldtransformaties gebaseerd op het histogram.

in een relatief klein deelinterval van $[0, 255]$, namelijk $[132, 200]$; zie ook het bijbehorende histogram. *Histogram equalisatie* is een eenvoudige transformatie welke het histogram zo homogeen mogelijk maakt, d.w.z. de verdeling p zoveel mogelijk transformeert naar een uniforme verdeling. Het effect daarvan zien we in Figuur 1(b). Allerlei details welke in de eerste afbeelding niet zichtbaar waren, zijn nu duidelijk te zien. Een eenvoudige variant hierop is de *histogram uitrekking*. Deze transformattie doet niets anders dan het bereik van het histogram, in ons geval $[132, 200]$, uitrekken naar $[0, 255]$. Het effect is te zien in Figuur 1(c). De histogram-gebaseerde operatoren werken op pixel-niveau: de uitkomst in een gegeven pixel hangt alleen af van de grijswaarde in dat pixel. Een meer geavanceerde familie van operatoren wordt verkregen als de uitkomst in een pixel afhangt van de inputwaarden in een specifieke omgeving, *venster* genaamd, van de betreffende pixel. Vaak betreft het hier een lineaire combinatie zoals geillustreerd in Figuur 2. Hierbij wordt voor elk punt x in het input beeld de output waarde $G(x) = \sum_{i=0}^{8} a_i F(x_i)$ berekend. In Figuur 2 hebben we voor het gemak aangenomen dat de afmeting van het venster 3×3 is (dus 9 pixels bevat), maar dit is niet noodzakelijk. De operator geillustreerd in Figuur 2 heet wel een *convolutie-operator* en de bijbehorende matrix A de *convolutie-matrix*.



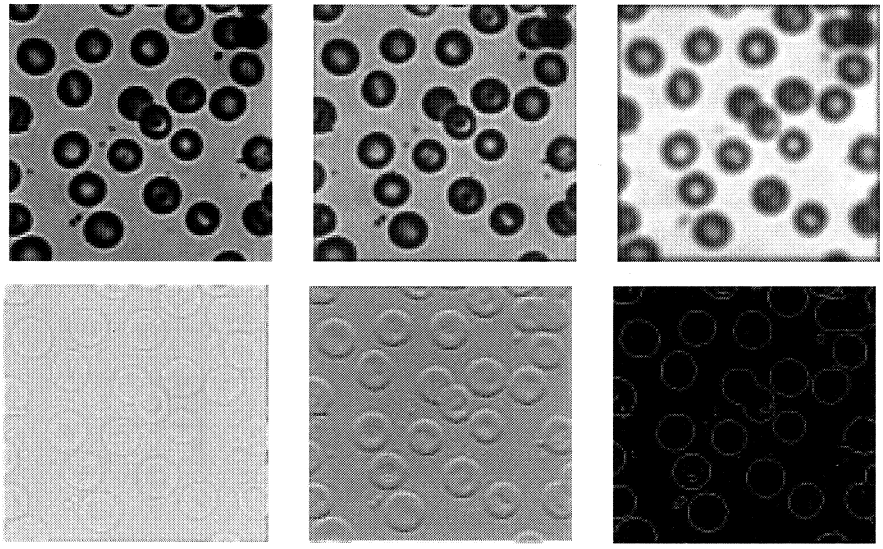
FIGUUR 2. Een lineaire operator.

Een aanzienlijk deel van de klassieke, maar ook moderne, beeldverwerkingstechnieken (bijvoorbeeld de wavelet-transformaties elders in deze bundel) zijn gebaseerd op lineaire convolutie-operatoren. Dit is niet zo verrassend als men zich realiseert dat door variatie van de grootte van het venster en de gewichten in A men de beschikking heeft over een groot aantal operatoren, middels welke men in staat is om verschillende aspecten van een beeld te bestuderen. In Figuur 3 laten we dit aan de hand van een paar voorbeelden zien. In al deze voorbeelden, behalve (c), is de venstergrootte 3×3. Voor de convolutie-matrices

kiezen we

$$A_M = \frac{1}{9}\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}, \quad A_L = \begin{pmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \quad A_S = \begin{pmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{pmatrix},$$

welke respectievelijk corresponderen met het berekenen van een gemiddelde, een som van tweede afgeleiden (ook wel Laplace filter genoemd), en een verticale gradient (Sobel filter geheten). In Figuur 3 worden achtereenvolgens getoond: het originele beeld zijnde een opname van bloedcellen, de convolutie met A_M, het resultaat na een viervoudige convolutie met A_M (feitelijk een middeling over een 9×9 venster), de convolutie met A_L, de convolutie met A_S, resulterend in horizontale randen (edges), en een z.g.n. 'edge detector' gebaseerd op convoluties met horizontale en verticale Sobel matrices.



FIGUUR 3. Van links naar rechts en onder naar boven: een input beeld, een middeling over een 3×3 en 9×9 venster, de convolutie met de Laplace matrix A_L en de Sobolev matrix A_S, en als laatste een binair beeld verkregen m.b.v. een zogenaamde 'Sobolev edge detector'.

Deze sectie beoogde slechts een paar elementaire beeldverwerkings-operaties de revue te laten passeren en bevat slechts een fractie van wat er in de literatuur allemaal over dit onderwerp te vinden is. [6] De lezer die meer wil weten over dit onderwerp verwijzen we graag naar het boek van Gonzalez and Woods [1], één van de standaardwerken op dit gebied.

[6] Merk op dat er alleen al op het gebied van de digitale beeldverwerking zo'n 7 of 8 belangrijke tijdschriften bestaan, gezamenlijk goed voor een paar duizend pagina's per jaar!

3. Principes van de mathematische morfologie

Het woord "morfologie" is te herleiden tot de Griekse woorden $\mu o \rho \phi \eta$ en $\lambda o \gamma o \varsigma$ en betekent zoveel als "vormenleer". Het woord "morfologie" komt in verschillende takken van de wetenschap voor: in de biologie duidt het de richting aan welke zich bezighoudt met vorm en bouw van organismen; in de geografie betreft het een methodiek voor de beschrijving van de vorm van aardoppervlakken; in de taalkunde tenslotte, is morfologie het vakgebied dat zich bezighoudt met de vraag hoe in een taal woorden en verbuigingen daarvan worden gevormd. Met "mathematische morfologie" daarentegen, wordt een specifieke methodologie uit de digitale beeldverwerking bedoeld. Ook hier zijn de trefwoorden "vorm" en "structuur", maar nu in betrekking tot informatie, expliciet of impliciet, bevat in digitale beelden.

De mathematische morfologie vindt zijn oorsprong in het werk van Matheron [3] en Serra [4] van de Ecole Nationale Supérieure des Mines de Paris in Fontainebleau. Hun werk is in belangrijke mate geinspireerd door zeer praktische problemen afkomstig uit de mijnbouw, zoals de beschrijving van de porositeit van een bepaald materiaal (gesteente, legering) aan de hand van visuele kenmerken, veelal gegeven in de vorm van een opname van een dwarsdoorsnede van het betreffende materiaal.

De technieken welke door Matheron en Serra werden ontwikkeld, berusten op elementaire operaties uit de verzamelingstheorie, zoals vereniging, doorsnijding, en verzamelingscomplement, en daarnaast op geometrische transformaties zoals translatie, rotatie en schaling. Ondanks (of misschien dankzij) hun eenvoud bleken deze technieken uitermate geschikt voor een groot aantal toepassingen, en in de afgelopen decennia heeft de mathematische morfologie zich ontwikkeld tot één van de belangrijkste methodieken in de beeldverwerking. Wat de mathematische morfologie voor de wiskundige zo interessant maakt, is dat deze discipline berust op solide theoretische fundamenten, samengesteld uit verschillende gebieden van de wiskunde zoals verzamelingstheorie, algebra, meetkunde, topologie en waarschijnlijkheidsrekening, om de belangrijkste te noemen.

In deze uiteenzetting zullen we ons noodgedwongen beperken tot een aantal elementaire operaties uit de mathematische morfologie. Bovendien kiezen we er voor ons grotendeels te beperken tot binaire (i.e., zwart-wit) beelden. Deze laatstgenoemde keuze moge weliswaar erg restrictief zijn, het biedt ons wel de mogelijkheid om in een zeer kort ruimtebestek de lezer een redelijk inzicht te geven in de uitgangspunten van het vakgebied. Overigens zullen we in de laatste sectie alsnog aandacht besteden aan uitbreidingen naar grijswaarden-beelden. Een tweetal werken op het gebied van de mathematische morfologie van vrij recente datum zijn [2, 5].

Binaire beelden kunnen wiskundig worden gerepresenteerd middels een verzameling. Als we de onderliggende discrete ruimte aangeven met $\mathbb{Z}^2$, i.e., de paren (i, j) met $i, j \in \mathbb{Z}$, dan kunnen we een binair beeld representeren als een deelverzameling $X \subseteq \mathbb{Z}^2$. De ruimte van al zulke deelverzamelingen geven we aan met $\mathcal{P}(\mathbb{Z}^2)$, ook wel de machtsverzameling van $\mathbb{Z}^2$ genoemd. Onder een *morfologische operator* verstaan we een afbeelding ψ welke $\mathcal{P}(\mathbb{Z}^2)$ in $\mathcal{P}(\mathbb{Z}^2)$

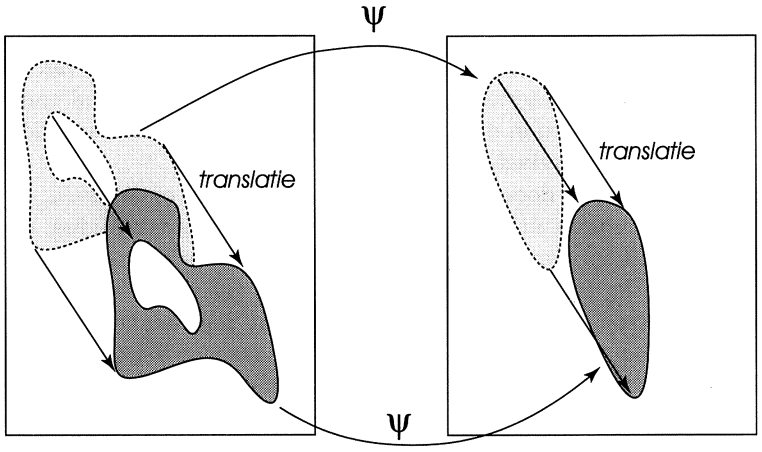
afbeeldt, en welke vaak een aantal additionele eigenschappen heeft; deze zullen in de loop van dit betoog aan de orde komen. Een belangrijke eigenschap in dit verband is *translatie-invariantie*:

$$\psi(X_h) = (\psi(X))_h, \quad X \in \mathcal{P}(\mathbb{Z}^2),\ h \in \mathbb{Z}^2 .$$

Hierin is X_h de translatie van X over de vector h:

$$X_h = \{x + h \mid x \in X\} .$$

Translatie-invariantie van een operator ψ betekent dus dat ψ commuteert met de translatie-operator $X \mapsto X_h$; zie Figuur 4 voor een illustratie. Alle operatoren waar we hier mee te maken krijgen, hebben deze eigenschap. We besluiten



FIGUUR 4. Translatie-invariantie van een operator.

deze sectie met een paar afspraken betreffende notatie. Als X, Y verzamelingen zijn, dan geven we met $X \cap Y$ en $X \cup Y$ de *doorsnede* resp. de *vereniging* van X en Y aan. Met X^c bedoelen we het *complement* van X. Morfologische operatoren duiden we aan met Griekse letters $\psi, \delta, \varepsilon$, etc.

4. DILATIE EN EROSIE

In deze sectie bespreken we twee morfologische operatoren, dilatie en erosie, welke de bouwstenen vormen voor een groot aantal andere operatoren. De dilatie δ_A is als volgt gedefinieerd:

$$\delta_A(X) = X \oplus A, \quad X \in \mathcal{P}(\mathbb{Z}^2) .$$

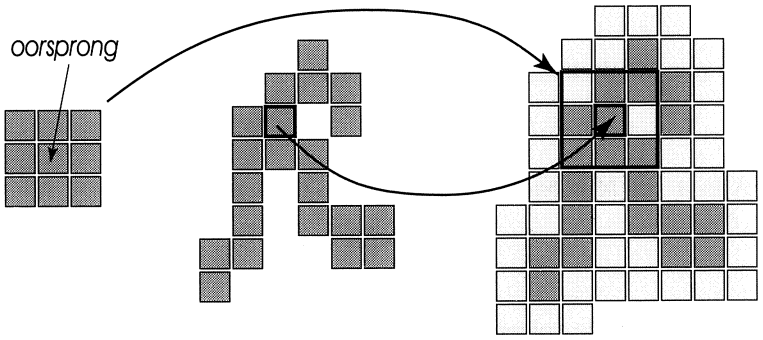
Hierin is $A \in \mathcal{P}(\mathbb{Z}^2)$ gegeven en is $X \oplus A$ de *Minkowski som* van X en A:

$$X \oplus A = \{x + a \mid x \in X \text{ en } a \in A\} .$$

Het is eenvoudig in te zien dat $X \oplus A$ kan worden herschreven als

$$X \oplus A = \bigcup_{a \in A} X_a = \bigcup_{x \in X} A_x \,. \tag{4.1}$$

Het eerste gelijk-teken betekent dat $X \oplus A$ de vereniging van alle getransleerde verzamelingen X_a is, waarbij a een willekeurige vector uit A is. Het tweede gelijk-teken heeft een analoge interpretatie. In de mathematische morfologie is X het input beeld en heeft A de interpretatie van een zogeheten *structurerend element*. Dit is een verzameling $A \subseteq \mathbb{Z}^2$ welke naar eigen inzicht kan worden gekozen. In de praktijk is deze keuze sterk afhankelijk van de specifieke toepassing. Een veelgemaakte keuze voor A is een 3×3 vierkant met de oorsprong als middelpunt. In Figuur 5 geven we een illustratie van de dilatie.



FIGUUR 5. Van links naar rechts: het structurerend element A, de input verzameling X (donkere pixels) en de gedilateerde $X \oplus A$ (lichte en donkere pixels).

De dilatie heeft een aantal interessante eigenschappen. De eerste eigenschap is translatie-invariantie:

$$X_h \oplus A = (X \oplus A)_h, \quad X \in \mathcal{P}(\mathbb{Z}^2)\,,\ h \in \mathbb{Z}^2\,. \tag{4.2}$$

Een tweede eigenschap, karakteristiek voor dilaties, is:

$$(X \cup Y) \oplus A = (X \oplus A) \cup (Y \oplus A)\,, \tag{4.3}$$

voor alle $X, Y \in \mathcal{P}(\mathbb{Z}^2)$. D.w.z., als we de dilatie van een verzameling, welke de vereniging van een aantal kleinere componenten is, willen berekenen, dan kunnen we ook eerst de dilaties van al deze componenten berekenen, en vervolgens de vereniging van alle afzonderlijke uitkomsten nemen.

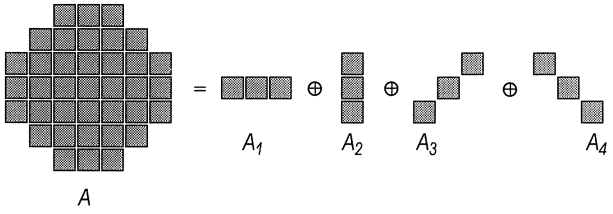
Een onmiddelijk gevolg van (4.3) is:

$$X \subseteq Y \text{ impliceert } X \oplus A \subseteq Y \oplus A\,. \tag{4.4}$$

We zeggen wel dat dilatie een *(monotoon) stijgende* operator is. Ten slotte geldt dat

$$(X \oplus A) \oplus B = X \oplus (A \oplus B)\,. \tag{4.5}$$

Dit betekent dat dilatie met A gevolgd door een tweede dilatie met B hetzelfde resultaat geeft als dilatie met $A \oplus B$. Deze laatste eigenschap betekent dat dilatie met een "groot" structurerend element A ontbonden kan worden in termen van dilaties met de componenten $A_1, A_2, \ldots, A_n$ welke zo gekozen zijn dat $A_1 \oplus A_2 \oplus \cdots \oplus A_n = A$. Bijvoorbeeld:



De tweede morfologische operator is de *erosie* ε_A gegeven door

$$\varepsilon_A(X) = X \ominus A, \quad X \in \mathcal{P}(\mathbb{Z}^2),$$

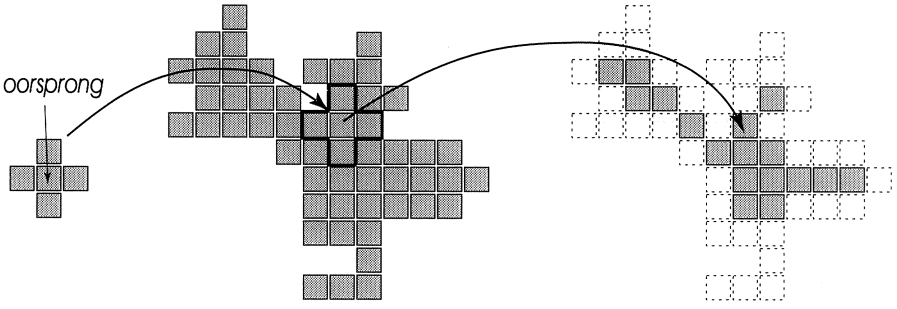
waarin $X \ominus A$ het *Minkowski verschil* van X en A is:

$$X \ominus A = \bigcap_{a \in A} X_{-a}. \tag{4.6}$$

Deze operator heeft de volgende belangrijke geometrische interpretatie:

$$X \ominus A = \{h \in \mathbb{Z}^2 \mid A_h \subseteq X\}, \tag{4.7}$$

m.a.w., $X \ominus A$ bestaat uit alle punten zodat A gecentreerd in dit punt geheel binnen X ligt. Een illustratie vindt men in Figuur 6. Ook de erosie is translatie-



FIGUUR 6. Van links naar rechts: het structurerend element A, de input verzameling X en de geërodeerde verzameling $X \ominus A$ (donkere pixels).

invariant en stijgend (c.f. (4.2) en (4.4)), maar i.p.v. (4.3) vinden we:

$$(X \cap Y) \ominus A = (X \ominus A) \cap (Y \ominus A). \tag{4.8}$$

Het analogon van de decompositie in (4.5) is

$$(X \ominus A) \ominus B = X \ominus (A \oplus B)\,. \tag{4.9}$$

Dilatie en erosie zijn aan elkaar gerelateerd middels de volgende relatie

$$X \oplus A \subseteq Y \iff X \subseteq Y \ominus A, \quad X, Y \in \mathcal{P}(\mathbb{Z}^2)\,, \tag{4.10}$$

welke de *adjunctie-relatie* wordt genoemd. Zonder overdrijving kan men stellen dat de abstracte versie van deze relatie ten grondslag ligt aan de mathematische morfologie. Allerlei uitbreidingen van de theorie, bijvoorbeeld naar grijswaarden- en kleurenbeelden zijn op deze adjunctie-relatie gebaseerd [2].

Ook geldt de volgende dualiteits-relatie:

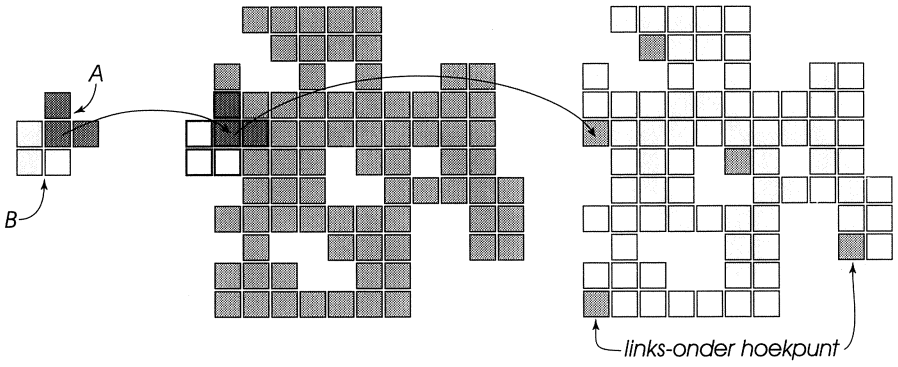
$$X^{\mathrm{c}} \ominus A = (X \oplus \check{A})^{\mathrm{c}}\,,$$

waarin $\check{A} = \{-a \mid a \in A\}$ de gespiegelde van A t.o.v. de oorsprong is. Feitelijk zegt deze relatie dat erosie van de achtergrond overeenkomt met dilatie van de voorgrond.

Een belangrijke morfologische operator, welke met behulp van de erosie kan worden gedefiniëerd, is de *hit-or-miss operator*. Een belangrijk verschil met de operatoren welke we hierboven besproken hebben, is dat het structurerend element hier uit twee componenten bestaat, een voorgrond component A en een achtergrond component B. Gegeven een binair beeld X zeggen we dat (A, B) op positie h met X overeenkomt als A_h binnen de voorgrond X en B_h binnen de achtegrond X^{c} ligt:

$$X \circledast (A, B) = \{h \in \mathbb{Z}^2 \mid A_h \subseteq X \text{ en } B_h \subseteq X^{\mathrm{c}}\}\,.$$

Deze operator is het best te illustreren aan de hand van een concreet voor-



FIGUUR 7. Door de specifieke keuze van A en B (links) waarin A de donkere en B de lichte pixels bevat, levert de hit-or-miss operator van een verzameling (midden) als output alle linker-beneden hoekpunten (rechts).

beeld. Dit doen we in Figuur 7 waar we laten zien hoe, middels een eenvoudige keuze van het paar (A, B), de hit-or-miss operator als output de linker-beneden hoekpunten in een beeld oplevert. Het is eenvoudig in te zien dat de hit-or-miss operator, welke ook kan worden geschreven als

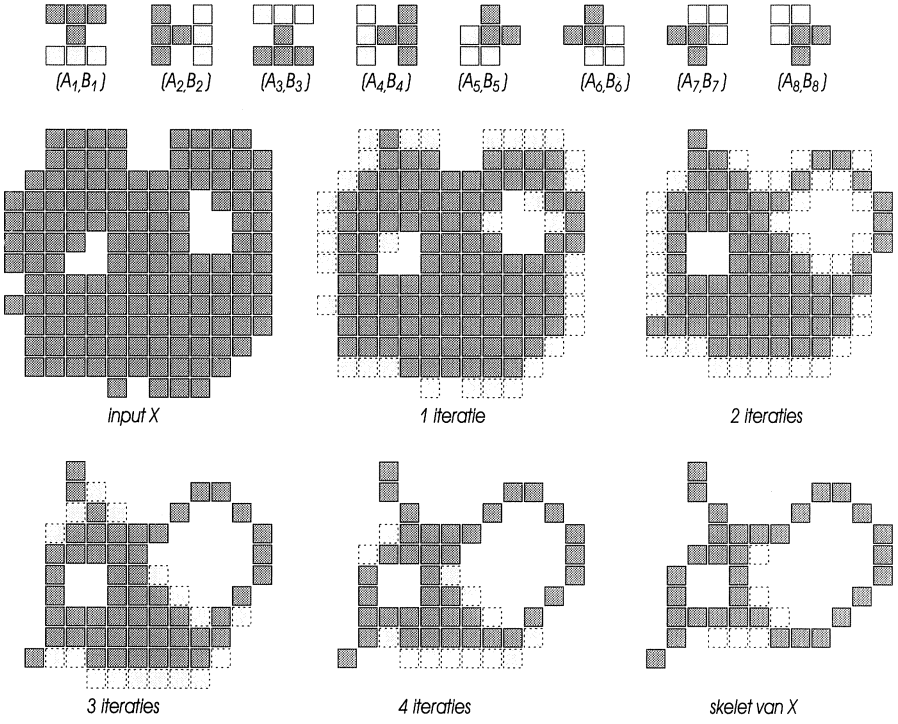
$$X \circledast (A, B) = (X \ominus A) \cap (X^c \ominus B)\,,$$

translatie-invariant is. De operator is echter in het algemeen niet stijgend.

Een ander operator, welke is afgeleid van de hit-or-miss operator, is de *thinning* (Nederlands: verdunning!). Deze wordt verkregen door de output van de hit-or-miss operator uit het oorspronkelijke beeld X weg te laten:

$$X \circledcirc (A, B) = X \setminus X \circledast (A, B)\,;$$

hierbij wordt altijd aangenomen dat $0 \notin B$. Er bestaan talloze algoritmen in de morfologie welke op de thinning gebaseerd zijn. De bekendste is ongetwijfeld het *skelet*. Het idee hierachter is erg simpel: bereken de thinning met (A_1, B_1),



FIGUUR 8. Het skelet van input beeld X wordt verkregen na 5 iteraties, waarbij elke iteratie uit 8 thinnings is opgebouwd.

dan met (A_2, B_2), enzovoorts tot aan (A_8, B_8) (hierin zijn (A_i, B_i) de acht rotaties over veelvouden van 45° van een gegeven (A_1, B_1)) en herhaal deze

reeks van acht thinnings tot er niets meer verandert. Een illustratie is gegeven in Figuur 8. Het skelet is een belangrijk concept in de digitale beeldverwerking en de patroonherkenning. Het wordt o.a. gebruikt in software voor karakterherkenning (OCR).

5. Opening en sluiting

Op het eerste gezicht lijkt het alsof dilatie en erosie elkaars inverse zijn, maar het is eenvoudig in te zien dat dit niet juist is. Als X kleiner is dan A in de zin dat geen enkele translatie van A binnen X past, dan is $X \ominus A$ de lege verzameling. Dit betekent met name dat de erosie niet inverteerbaar is. Niettemin heeft de operator welke wordt verkregen als erosie gevolgd door dilatie, i.e.,

$$X \circ A = (X \ominus A) \oplus A$$

een aantal interessante eigenschappen. Behalve dat deze operator wederom translatie-invariant en stijgend is, geldt ook dat

$$X \circ A \subseteq X \tag{5.1}$$

$$(X \circ A) \circ A = X \circ A, \tag{5.2}$$

voor alle $X \in \mathcal{P}(\mathbb{Z}^2)$, ongeacht de keuze van het structurerend element A. De eerste eigenschap, aangeduid met *anti-extensiviteit*, betekent dat het beeld $X \circ A$ altijd bevat is in X. De tweede eigenschap is feitelijk de meest interessante en zegt dat de operator $X \mapsto X \circ A$ *idempotent* is. Herhaalde toepassing van een idempotente operator heeft per definitie geen zin. Idempotentie is een wenselijke eigenschap waar het om het filteren van ruis in beelden gaat. We zullen daar in Sectie 7 een voorbeeld van zien.

De operator $\alpha_A(X) = X \circ A$ heet de *opening van X met A*. Deze bewerking heeft de volgende geometrische interpretatie:

$$X \circ A = \bigcup \{A_h \mid h \in \mathbb{Z}^2 \text{ en } A_h \subseteq X\}, \tag{5.3}$$

m.a.w., $X \circ A$ is de vereniging van alle translaties van het structurerend element A welke binnen X liggen. In Figuur 9 geven we een illustratie. Als we eerst de dilatie van X met A berekenen en daar vervolgens de erosie op loslaten, dan vinden we

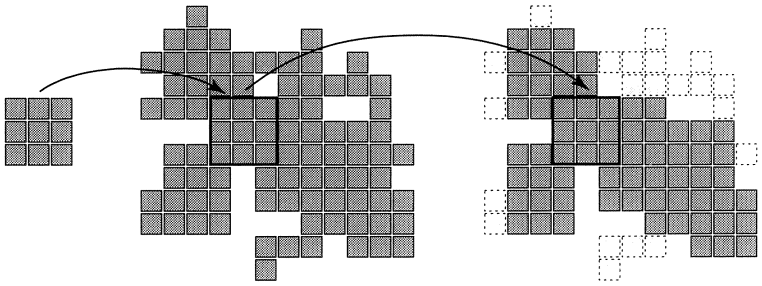
$$X \bullet A = (X \oplus A) \ominus A,$$

de *sluiting van X met A*. De operator $\beta_A(X) = X \bullet A$ is translatie-invariant, stijgend, extensief, d.w.z.

$$X \subseteq X \bullet A, \quad X \in \mathcal{P}(\mathbb{Z}^2), \tag{5.4}$$

en, evenals de opening, idempotent:

$$(X \bullet A) \bullet A = X \bullet A. \tag{5.5}$$

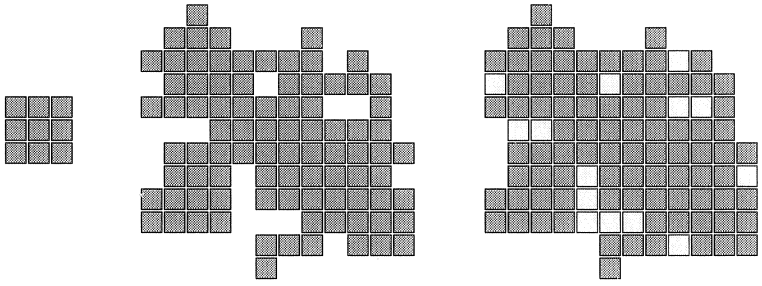


FIGUUR 9. Van links naar rechts: het structurerend element A, de input verzameling X en de opening $X \circ A$.

Verder geldt de dualiteits-relatie

$$X^c \circ A = (X \bullet \check{A})^c . \tag{5.6}$$

Dit betekent zoveel als dat de opening van de voorgrond overeenkomt met de sluiting van de achtergrond (waarbij het structurerend element, in het niet-symmetrische geval, gespiegeld dient te worden). Een illustratie vindt men in Figuur 10. Merk op dat in het geval van dilatie en erosie de positie van het



FIGUUR 10. Van links naar rechts: het structurerend element A, de input verzameling X en de sluiting $X \bullet A$.

structurerend element t.o.v. de oorsprong belangrijk is; in het geval van de opening en de sluiting maakt dit niet uit.

6. PATROON-SPECTRUM

In deze sectie zullen we aan de hand van een paar eenvoudige voorbeelden laten zien hoe een eenvoudige morfologische operator als de opening kan worden gebruikt om uiterst bruikbare geometrische informatie uit een beeld te halen. Zoals we tot nu toe steeds hebben gedaan, zullen we ons ook hier weer beperken tot het binaire geval, wederom met de kanttekening dat alles uitbreidbaar is naar het geval van grijswaarden-beelden.

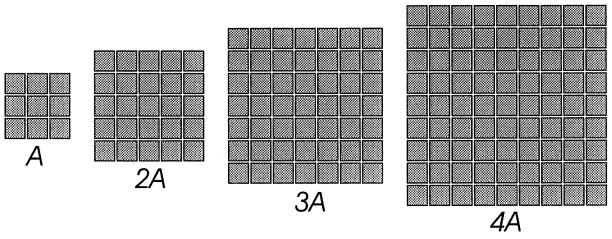
Het principe welk ten grondslag ligt aan de definitie van het patroonspectrum is het volgende: als A, B twee structurerende elementen zijn, zodanig dat B open is m.b.t. A, d.w.z., $B \circ A = B$ (en hieraan is voldaan dan en slechts dan als er een C is zodat $B = A \oplus C$), dan geldt dat

$$X \circ B \subseteq X \circ A \text{ voor alle } X \in \mathcal{P}(\mathbb{Z}^2)\,.$$

Dit is eenvoudig in te zien: $X \circ B$ is een vereniging van translaties van B, maar B kan weer worden geschreven als een vereniging van translaties van A omdat $B \circ A = B$. Kies nu een rij structurerend elementen $A_1, A_2, \ldots$ zodat

$$A_{n+1} \circ A_n = A_{n+1},\ n \geq 1\,. \tag{6.1}$$

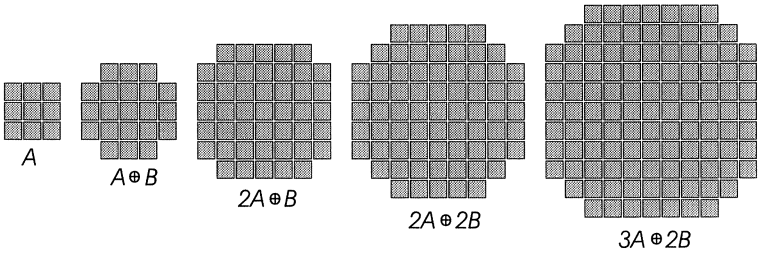
Aan deze relatie is eenvoudig voldaan als $A_n = nA := A \oplus A \oplus \cdots \oplus A$ (n termen), zoals in Maar we kunnen zo'n rij ook als volgt construeren:



$$A_1 = A,\ A_2 = A \oplus B,\ A_3 = 2A \oplus B,\ A_4 = 2A \oplus 2B,\ \text{ etc.}$$

Hierin kunnen we de structurerend elementen A en B willekeurig kiezen. Zo levert de keuze

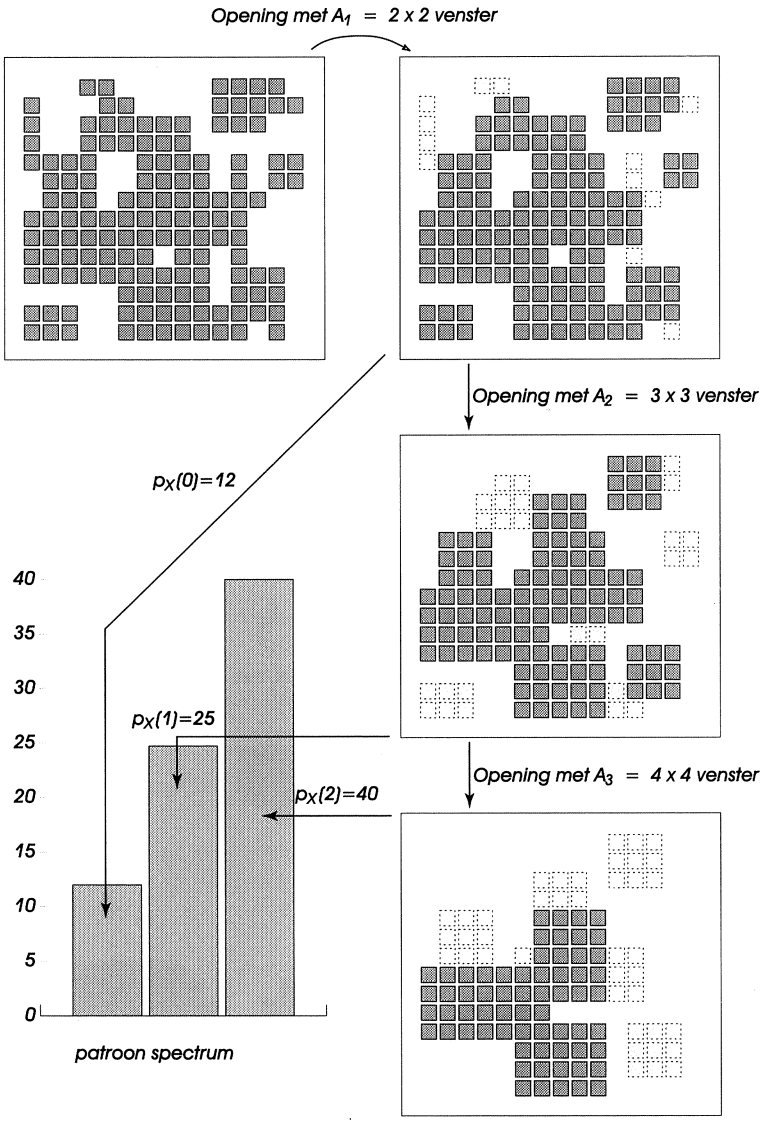
de volgende rij op:



Als aan (6.1) is voldaan, dan geldt, op grond van bovenstaande, voor willekeurige input verzameling X:

$$X \circ A_{n+1} \subseteq X \circ A_n, \; n \geq 1 \,. \tag{6.2}$$

Bij elke stap n, waarbij we de resterende verzameling openen met A_n, verdwij-

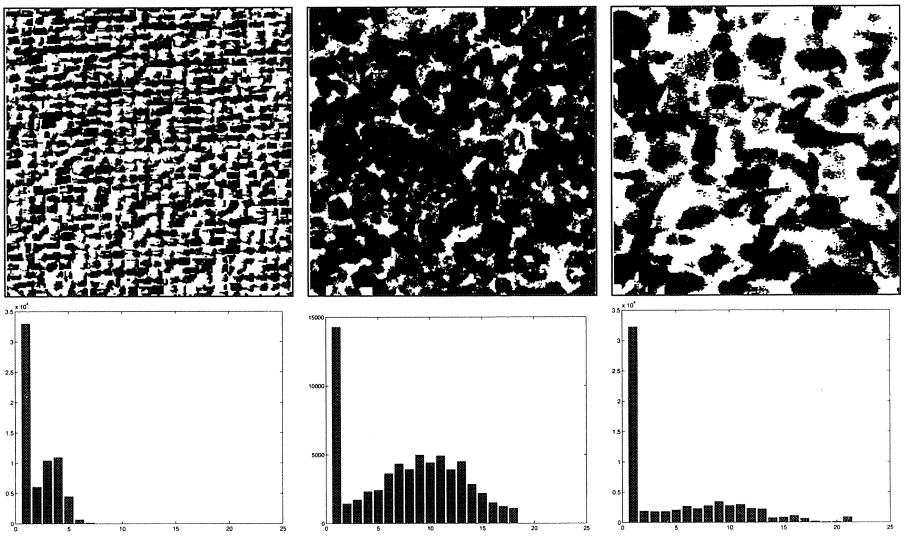


FIGUUR 11. De berekening van het patroon-spectrum.

nen er een aantal punten, en wel

$$p_X(n) = |X \circ A_{n-1} \setminus X \circ A_n|\,, \tag{6.3}$$

waarbij $|Y|$ het aantal punten in Y is (we nemen voor het gemak aan dat we met eindige verzamelingen werken), en waarbij $A_0 = \{0\}$ en dus $X \circ A_0 = X$. De functie p_X wordt in de literatuur wel het *patroon-spectrum* van X genoemd, omdat deze functie informatie bevat over de afmetingen en de vormen van objecten in X. In Figuur 11 laten we aan de hand van een eenvoudig voorbeeld zien hoe deze functie wordt berekend. In de praktijk wordt het patroon-spectrum (soms in een iets andere hoedanigheid ook wel *granulometrie* genoemd) gebruikt om texturen te analyseren, of om objecten aan de hand van hun vorm van elkaar te onderscheiden. We geven een illustratie in Figuur 12.



FIGUUR 12. Het patroon-spectrum van drie verschillende binaire textuurbeelden.

7. MORFOLOGIE VOOR GRIJSWAARDEN-BEELDEN

We hebben ons, wat de morfologie betreft, tot nu toe beperkt tot binaire beelden. In deze sectie laten we aan de hand van concrete voorbeelden zien hoe binaire operatoren uitgebreid kunnen worden naar grijswaarden-beelden. Verder zullen we ook een morfologische operator construeren welke uiterst bruikbaar is voor het filteren van ruis in beelden.

We hebben binaire beelden gemodelleerd met $\mathcal{P}(\mathbb{Z}^2)$, de machtsverzameling van $\mathbb{Z}^2$. We zullen in het nu volgende grijswaarden-beelden representeren middels de ruimte $\mathrm{Fun}(\mathbb{Z}^2)$ bestaande uit alle functies welke $\mathbb{Z}^2$ in een grijswaardenverzameling $\mathcal{T}$ afbeelden. We zullen hier voor het gemak $\mathcal{T} = \{0, 1, \ldots, n\}$ kiezen, waarbij 0 zwart en n wit representeert (vaak is $n = 255$). Merk op

dat voor RGB-kleurenbeelden $\mathcal{T}$ drie-dimensionaal is. Zoals we verzamelingen partiëel kunnen ordenen middels de inclusie-relatie '$X \subseteq Y$', zo kunnen we dat ook voor functies m.b.v. de relatie $F \leq G$, waarmee bedoeld wordt dat in elk punt $x \in \mathbb{Z}^2$ de grijswaarde $F(x)$ niet groter is dan de grijswaarde $G(x)$. We hebben ook een analogon van de intersectie en de vereniging, en wel het minimum en maximum: $F \wedge G$ en $F \vee G$ zijn de functies gegeven door:

$$(F \wedge G)(x) = \min(F(x), G(x)), \quad (F \vee G)(x) = \max(F(x), G(x)),$$

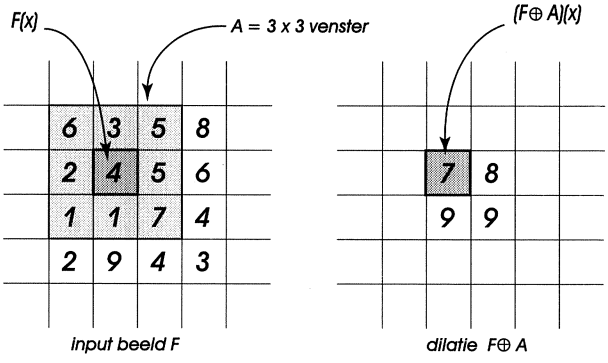
voor alle $x \in \mathbb{Z}^2$. Uitgaande van (4.1) kunnen we nu de dilatie van een functie $F \in \mathrm{Fun}(\mathbb{Z}^2)$ met een structurerend element $A \subseteq \mathbb{Z}^2$ definiëren als:

$$F \oplus A = \bigvee \{F_a \mid a \in A\},$$

waarbij F_a de translatie van F over a is: $F_a(x) = F(x - a)$. We kunnen bovenstaande uitdrukking ook schrijven als

$$(F \oplus A)(x) = \max\{F(x - a) \mid a \in A\}.$$

Met andere woorden, $(F \oplus A)(x)$ is de grootste grijswaarde van F binnen het venster A gepositioneerd in punt x; zie Figuur 13. Evenzo is de erosie gegeven



FIGUUR 13. De dilatie $(F \oplus A)(x)$ is de grootste grijswaarde $F(y)$ waarbij $y \in A_x$, het venster A gecentreerd in x.

door

$$(F \ominus A)(x) = \min\{F(x + a) \mid a \in A\}.$$

Allerlei eigenschappen welke in Sectie 4 zijn afgeleid voor het binaire geval blijven gelden, bijvoorbeeld translatie-invariantie, het analogon van (4.3), i.e.,

$$(F \vee G) \oplus A = (F \oplus A) \vee (G \oplus A),$$

enz. De belangrijkste eigenschap is echter de adjunctie-relatie (4.10), welke voor functies als volgt luidt:

$$F \oplus A \leq G \iff F \leq G \ominus A\,, \quad F, G \in \mathrm{Fun}(\mathbb{Z}^2)\,.$$

We kunnen door dilatie en erosie samen te stellen, openingen en sluitingen definiëren voor grijswaarden-beelden, en wederom blijven alle resultaten bewezen in Sectie 5 overeind. We zullen hier nu niet verder op ingaan. In plaats daarvan introduceren we een specifieke klasse van operatoren, de zogenaamde *rang-operatoren*. Het idee is simpel. Kies eerst een (eindig) structurerend element; we zullen hier voor het gemak een 3×3 venster gecentreerd in de oorsprong nemen, maar de lezer zal geen moeite hebben om in te zien dat dit gemakkelijk gegeneraliseerd kan worden. Leg A neer in een willekeurig punt $x = x_0$ van het input beeld F (zie Figuur 2) en orden de grijswaarden $F(x_0), F(x_1), \ldots, F(x_8)$ in aflopende grootte; dit levert een rij $t_1 \geq t_2 \geq \cdots \geq t_9$. We definiëren voor $k = 1, 2, \ldots, 9$ de *rang-operator* ρ_k middels

$$\rho_k(F)(x) = t_k\,.$$

Uit deze definitie volgt onmiddelijk dat

$$\rho_1(F) \geq \rho_2(F) \geq \cdots \geq \rho_9(F)\,, \quad F \in \mathrm{Fun}(\mathbb{Z}^2)\,. \tag{7.1}$$

Het is verder eenvoudig in te zien dat elk van de operatoren ρ_k translatie-invariant en stijgend is. Uit de definitie volgt dat $\rho_1(F)(x)$ het maximum van F in een 3×3 venster om x is, met andere woorden,

$$\rho_1(F) = F \oplus A\,.$$

Evenzo is

$$\rho_9(F) = F \ominus A\,.$$

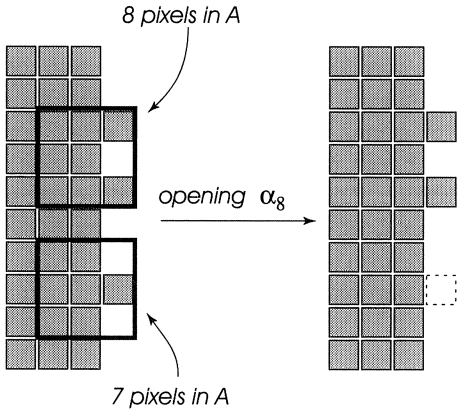
Van bijzonder belang is verder ρ_5 aangezien $\rho_5(F)(x)$ de mediaan van de waarden $F(x_0), F(x_1), \ldots, F(x_8)$ is, en we noemen ρ_5 dan ook de *mediaan operator*. Deze operator wordt in de praktijk vaak gebruikt om ruis te filteren, en we zullen daar verderop in deze sectie een voorbeeld van zien.

We construeren nu m.b.v. deze rang-operatoren een nieuwe familie van operatoren, en wel de zogeheten *RAS-filters* (hierin is 'RAS' de afkorting van 'rank alternating sequential'). In de morfologie is een *filter* een stijgende operator ψ welke *idempotent* is, d.w.z. $\psi^2 = \psi$, oftewel $\psi(\psi(F)) = \psi(F)$, hetgeen zoveel betekent als dat tweemaal toepassen van de operator niet meer effect heeft als slechts één keer toepassen, een eigenschap welke buitengewoon interessant is als de operator gebruikt wordt om ruis te filteren. We hebben eerder al gezien dat de opening en de sluiting deze eigenschappen bezitten. Definiëer, voor $k = 1, 2, \ldots, 9$, de operator

$$\alpha_k(F) = F \wedge (\rho_k(F) \oplus A)\,,$$

d.w.z., bereken de rang-getransformeerde $\rho_k(F)$, dilateer met A, en neem het minimum van de uitkomst en het oorspronkelijke input beeld. De operator α_k

heeft alle eigenschappen welke de opening $F \mapsto F \circ A$ ook heeft: ze is stijgend, translatie-invariant, anti-extensief (d.w.z. $\alpha_k(F) \leq F$) en idempotent. In het binaire geval heeft α_k een eenvoudige geometrische interpretatie: neem een getransleerde A_h van het structurerend element A zodat tenminste k punten uit A_h in X liggen, dan is $A_h \cap X$ bevat in $\alpha_k(X)$; zie Figuur 14. Het is



FIGUUR 14. Illustratie van de opening α_8.

duidelijk dat α_9 de oorspronkelijke opening $F \mapsto F \circ A$ is. De opening α_1 is domweg de identiteits-operator (welke elke functie op zichzelf afbeeldt). De familie openingen α_k voldoet als gevolg van (7.1) aan

$$\alpha_1(F) \geq \alpha_2(F) \geq \cdots \geq \alpha_9(F), \quad F \in \mathrm{Fun}(\mathbb{Z}^2).$$

Gebruikmakend van het dualiteitsprincipe dat men overal in de morfologie tegenkomt, definiëren we een tweede familie operatoren

$$\beta_k(F) = F \vee (\rho_{10-k}(F) \ominus A).$$

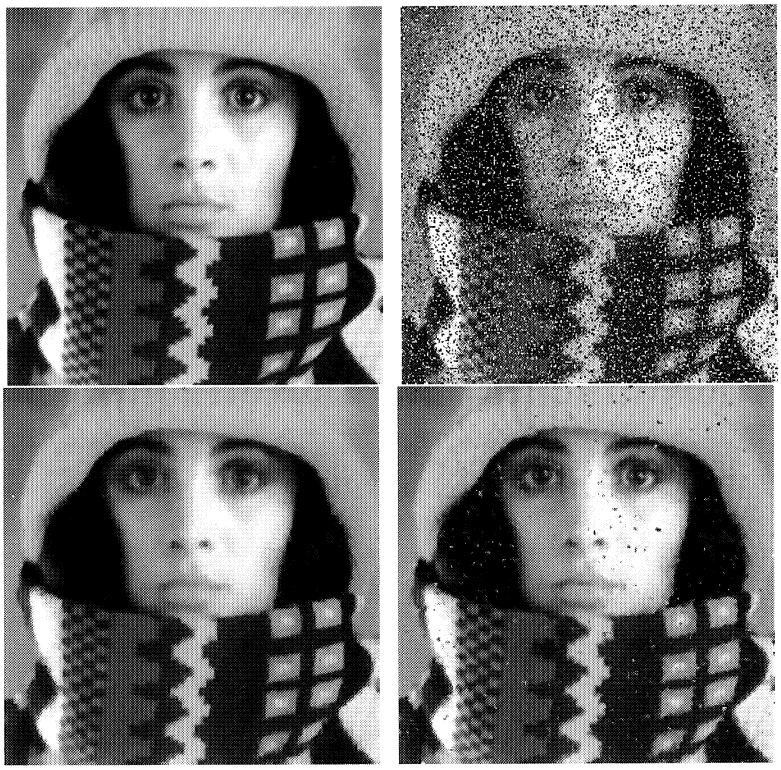
Het zal niemand verbazen dat ieder β_k een sluiting is, dat β_1 de identiteits-operator is, en dat β_9 de oorspronkelijke sluiting $F \mapsto F \bullet A$ is. Verder geldt

$$\beta_1(F) \leq \beta_2(F) \leq \cdots \leq \beta_9(F), \quad F \in \mathrm{Fun}(\mathbb{Z}^2).$$

We hebben nu alle gereedschappen om het RAS-filter ψ_k te definiëren: voor $k = 1, 2, \ldots, 9$ is

$$\psi_k = \beta_k \alpha_k \beta_{k-1} \alpha_{k-1} \cdots \beta_1 \alpha_1 .$$

D.w.z., we passen eerst een opening gevolgd door een sluiting toe, beiden met een zwakke werking, waarvan alleen kleine ruisdeeltjes effect ondervinden. (Merk op dat we in bovenstaande uitdrukking β_1 en α_1 weg kunnen laten aangezien dit beide identiteits-operatoren zijn.) Vervolgens nemen we een iets krachtiger opening en sluiting, en we gaan zo door tot aan niveau k. In Figuur 15 laten we het effect van het RAS-filter ψ_9 zien. Het is interessant dit filter te vergelijken met de mediaan-operator, omdat deze laatste operator een uitstekende reputatie heeft waar het verwijdering van ruis betreft.



FIGUUR 15. Boven een grijswaarden-beeld voor- en nadat er ruis aan is toegevoegd; ongeveer 40% van de pixels is door de ruis aangetast. Onder ziet men het effect van twee verschillende ruisfilters: links het RAS-filter ψ_9 en rechts de mediaan operator ρ_5. Merk op dat de mediaan operator geen filter is omdat niet aan de idempotentie-eigenschap is voldaan.

8. TOT SLOT ....

We herhalen ten overvloede nog eens dat deze bijdrage slechts de pretentie heeft om als smaakmaker te fungeren. We hopen dat we er in geslaagd zijn de lezer te overtuigen dat het mogelijk is om met elementaire geometrische operaties interessante "beeld-operatoren" te bouwen. Zulke operatoren kunnen worden gebruikt om een gegeven input beeld in een ander beeld over te voeren, bijvoorbeeld met het oogmerk om de kwaliteit te verbeteren zoals in het geval van het RAS-filter. Ze kunnen ook worden gebruikt om kwantitatieve geometrische informatie aan een beeld te onttrekken. Een voorbeeld daarvan is het patroonspectrum besproken in Sectie 6. Natuurlijk gaat de beeldverwerkings-literatuur in het algemeen, en de mathematische morfologie in het bijzonder, veel verder. In ieder geval ver genoeg om te bereiken dat de meeste plaatjes inderdaad veel meer waard zijn dan duizend woorden. Maar dat is dan wel voor een belangrijk deel aan de wiskunde te danken.

REFERENTIES

1. GONZALEZ, R. C., AND WOODS, R. E. *Digital Image Processing.* Addison-Wesley, Reading, 1992.
2. HEIJMANS, H. J. A. M. *Morphological Image Operators.* Academic Press, Boston, 1994.
3. MATHERON, G. *Random Sets and Integral Geometry.* John Wiley & Sons, New York, 1975.
4. SERRA, J. *Image Analysis and Mathematical Morphology.* Academic Press, London, 1982.
5. SOILLE, P. *Morphological Image Analysis.* Springer-Verlag, Berlin, 1999.

Vakantiecursus 2000

pp. 41 – 63

Wavelets in beeld en geluid

Hennie ter Morsche
Technische Universiteit Eindhoven

1. Inleiding

De wiskunde kent een aantal methoden dat bij het analyseren van geluid en beeld of meer in het algemeen bij het analyseren van signalen goed kan worden gebruikt. Klassiek is de Fourier-transformatie, die bij de frequentie-analyse een belangrijke rol speelt. De noodzaak om signalen simultaan te beschrijven in zowel tijd/plaats als frequentie heeft aanleiding gegeven tot de introductie van de zogenaamde Short Time Fourier-transformatie. Bekend in deze is ook de Wigner transformatie ofwel de Wigner distributie. Al deze transformaties hebben het gemeenschappelijke kenmerk dat ze bepaalde informatie die een signaal bevat op een gestructureerde manier proberen te achterhalen en wel zodanig dat "het werk" door een computer kan worden uitgevoerd. Zo kan de Fourier-transformatie worden ingezet om frequenties in een geluidssignaal op te sporen en met behulp van bijvoorbeeld de Wigner transformatie kunnen we ook nog een idee hebben op welk tijdsinterval zich een bepaalde frequentie manifesteert.

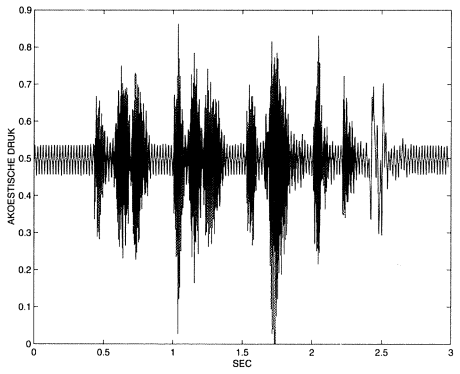
In dit rijtje van transformaties past nu ook de Wavelet-transformatie (er bestaan overigens verschillende versies). Met deze transformatie is men in staat om, ondersteund door snelle algoritmen voor het nodige rekenwerk, een signaal lokaal op verschillende detail niveau's te onderzoeken. Dit maakt de Wavelet-transformatie erg aantrekkelijk voor verschillende toepassingen. Belangrijke daarbij zijn het comprimeren en restaureren van data bij beeld en geluid, het onderdrukken van ruis en het opsporen van speciale objecten of saillante punten in signalen.

Deze cursus gaat over de Wavelet-transformatie, de wiskundige theorie en enkele toepassingen in relatie met beeld en geluid. De wiskundige theorie in deze cursus is vooral beschrijvend en weinig bewijzend, desondanks zullen we bij de presentatie de wiskundige context van functieruimten gebruiken. Paragraaf 3 geeft een beknopt overzicht van de begrippen, die we uit de theorie van functieruimten zullen gebruiken. Fourier-analyse is onderwerp van Paragraaf 4. Deze paragraaf is bedoeld om enerzijds vertrouwd te raken met functieruimten en anderzijds als opstap te dienen voor de wavelettheorie. In de nu volgende paragraaf worden geluid en beeld als wiskundige objecten geïntroduceerd. De wavelets worden behandeld in Paragraaf 5. De cursus wordt afgesloten met het laten zien, vooral in het hoorgedeelte van de cursus, van enkele toepassingen.

De wavelettheorie kent een erg interessante geschiedenis. Degene die zich hiervoor interesseren zij verwezen naar een artikel van Ingrid Daubechies in een proceedings van IEEE (cf [2]).

2. Beeld en geluid zijn functies

Het meest eenvoudige mathematisch model van een geluidssignaal is een functie, zeg f, gedefinieerd op een deel van de reële getallenrechte $\mathbb{R}$ waarvan de functiewaarden $f(t)$ reële getallen zijn. De getallenrechte $\mathbb{R}$ staat model voor de tijd-as en de functiewaarde $f(t)$ op tijdstip t stelt de door een geluidsbron veroorzaakte akoestische druk op dat tijdstip voor. Een geluidssignaal is dus een functie van één variabele. We zeggen dat geluid een 1D-signaal (één dimensionaal) is.
In Figuur 1 is een grafiek van een geluidsignaal getekend. Dit signal is een spraaksignaal met de tekst " Ik zie ik zie wat jij niet ziet".



Figuur 1. Grafiek van een spraaksignaal

Ook het mathematisch model van een beeld is een functie f, maar dan een functie van twee variabelen, zeg x en y. Het paar (x, y) zijn de cartesische coördinaten (t.o.v. van een rechthoekig assenstelsel) van een punt van het beeld en de functiewaarde $f(x, y)$ stelt de grijswaarde in dat punt voor. Daarbij zijn we er gemakshalve van uitgegaan dat we te maken hebben met een zwart/wit beeld. In onze terminologie is dus een beeld een 2D-signaal.

Het bewerken van signalen betekent dan ook het bewerken van functies. Tegenwoordig wordt een signaal bewerkt met computers, die alleen met digitale signalen kunnen omgaan. Het gevolg is dat een signaal eerst bemonsterd (gesampled) moet worden voordat er een digitale bewerking kan worden uitgevoerd. Voor een functie betekent dit dat we op een regelmatig rooster in het definitiegebied punten kiezen en de functie alleen in die punten beschouwen.

Bij het bemonsteren van signalen moet men in het algemeen verwachten dat er informatieverlies zal optreden. Hoge tonen in een geluidssignaal gaan verloren als de zogenaamde bemonsteringsfrequentie te laag is. Bij geluid is de bemonsteringsfrequentie het aantal samples per seconde, waarbij dan aangenomen wordt dat de samples op de tijdas op onderling gelijke afstand liggen. De samples van een beeld worden meestal op een uniform rechthoekig rooster gekozen, zodat er sprake is van een bemonsteringsfrequentie in de horizontale

richting en een bemonsteringsfrequentie in de verticale richting. Details in een beeld kunnen verloren gaan, wanneer een van deze twee bemonsteringsfrequenties te laag is. Maar wat betekent in een dergelijke situatie dan te laag? Het zogenaamde bemonsteringstheorema van Shannon (reference) zegt dat er bij een geluidssignaal geen informatieverlies zal optreden indien de bemonsteringsfrequentie minstens tweemaal zo hoog is als de hoogste frequentie die in het signaal voorkomt. In een enigszins gemodificeerde vorm bestaat een dergelijke theorema ook voor 2D- signalen. De Fourier-analyse leert ons hoe een signaal kan worden gesplitst in verschillende frequenties. Een 1D-signaal met precies de frequentie f_0 is het cosinussignaal

$$A\cos(2\pi f_0 t + \varphi_0) \qquad (t \in \mathbb{R}).$$

Het getal $A \geq 0$ heet amplitude en φ_0 de beginfase. Voor 2D-signalen is het enigszins gecompliceerder, omdat we dan met twee richtingen te maken hebben. Bijvoorbeeld een 2D-signaal $f(x, y)$ met de frequentie f_0 in de x-richting en de frequentie f_1 in de y-richting zou men kunnen aangeven met

$$f(x, y) = A\cos(2\pi f_0 x + \varphi_0)\cos(2\pi f_1 y + \varphi_1)$$

3. Vectorruimten van functies

In de inleiding is al gezegd dat geluid is opgebouwd uit signalen met verschillende frequenties. In feite staat hier dat een functie is opgebouwd uit een set van specifieke functies (cosinussignalen), die elk een bepaalde frequentie vertegenwoordigen. De wiskundige context om dit goed te beschrijven is die van vectorruimten en bases in vectorruimten.

In een (eindig dimensionale) vectorruimte is een basis een onafhankelijk stelsel $\mathbf{a_1}, \ldots, \mathbf{a_n}$ van vectoren, zodat iedere vector $\mathbf{x}$ in deze ruimte kan worden geschreven als een lineaire combinatie van de vectoren $\mathbf{a_i}$ uit dit stelsel, i.e. $\mathbf{x} = \xi_1 \mathbf{a_1} + \xi_2 \mathbf{a_2} + \cdots + \xi_n \mathbf{a_n}$ met eenduidig bepaalde coëfficiënten ξ_i. Deze coëfficiënten identificeren de vector x volledig, zodat in principe alle informatie over $\mathbf{x}$ is opgeslagen in de coëfficiëntenrij ξ_i. We zeggen dat de vector $\mathbf{x}$ is öpgebouwd" uit de basiselementen $\mathbf{a_i}$.

Omdat we weten hoe we functies kunnen optellen en hoe een functie met een getal kan worden vermenigvuldigd, kunnen we functies interpreteren als vectoren in een vectorruimte. Een vectorruimte van functies noemen we een *functieruimte*. Een voorbeeld is de ruimte van polynomen van de graad hooguit 3. Deze ruimte is een eindig dimensionale ruimte van dimensie vier. Immers elk cubisch polynoom is een lineaire combinatie van de vier basispolynomen $p_0(t) = 1$, $p_1(t) = t$,$p_2(t) = t^2$ en $p_3(t) = t^3$. Vaak zijn functieruimten oneindig dimensionaal en hebben we een oneindige set van basisfuncties nodig om de functies te representeren.

In een vectorruimte zijn natuurlijk verschillende keuzes van bases mogelijk. Zelfs als we in onze bekende ruimte $\mathbb{R}^n$ een zogenaamde orthonormale basis willen, dit is een basis waarvan de elementen onderling loodrecht zijn en lengte 1

hebben, dan zijn er nog oneindig veel mogelijkheden. Hoeken tussen vectoren en lengte van vectoren zijn ook zinvolle begrippen in functieruimten. De reden is dat deze begrippen afkomstig zijn van een *inproduct*. In $\mathbb{R}^n$ wordt het inproduct $(\mathbf{x}, \mathbf{y})$ van twee vectoren $\mathbf{x}$ en $\mathbf{y}$ gedefinieerd door

$$(\mathbf{x}, \mathbf{y}) = \sum_{i=0}^{n} x_i \, y_i,$$

waarbij $x_1, \ldots, x_n$ de kentallen zijn van $\mathbf{x}$ en $y_1, \ldots, y_n$ die van $\mathbf{y}$. Dit inproduct voldoet aan de volgende eigenschappen

$$\begin{aligned} (\alpha \, \mathbf{x}, \mathbf{y}) &= \alpha \, (\mathbf{x}, \mathbf{y}), \\ (\mathbf{x} + \mathbf{y}, \mathbf{z}) &= (\mathbf{x}, \mathbf{z}) + (\mathbf{y}, \mathbf{z}), \\ (\mathbf{x}, \mathbf{y}) &= (\mathbf{y}, \mathbf{x}). \end{aligned} \tag{1}$$

Bovenstaande eigenschappen gelden voor alle reële getallen α en alle vectoren $\mathbf{x}$,$\mathbf{y}$ en $\mathbf{z}$ in de betreffende vectorruimte. Maar in een ruimte van functies, bijvoorbeeld functies gedefinieerd op het interval $[a, b]$ kan men een inproduct (f, g) van twee functies f en g, definieren door

$$(f, g) = \int_a^b f(t) \, g(t) \, dt,$$

en dit inproduct voldoet aan dezelfde eigenschappen als vermeld in (1, waarbij de vectoren $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}$ zijn vervangen door functies zeg f, g en gh.

In de ruimte $\mathbb{R}^n$ weten we dat de lengte $\|\mathbf{x}\|$ van een vector $\mathbf{x}$ gelijk is aan

$$\|x\| = \sqrt{(\mathbf{x}, \mathbf{x})} = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}.$$

Men kan nu ook de lengte $\|f\|$ van een functie f overeenkomstig definieren.

$$\|f\| = (f, f)^{\frac{1}{2}}.$$

De stelling van Pythagoras voor functies ziet er nu als volgt uit:

Geldt voor de functies $f_1, f_2, \ldots, f_n$ dat $(f_i, f_j) = 0$ voor $i \neq j$ (we zeggen dat de functies onderling loodrecht zijn) en $f = f_1 + f_2 + \cdots + f_n$, dan geldt

$$\|f\|^2 = \|f_1\|^2 + \|f_2\| + \cdots + \|f_n\|^2. \tag{2}$$

Een belangrijke eigenschap van een inproduct staat bekend als de ongelijkheid van Cauchy-Schwartz. In een functieruimte zegt deze ongelijkheid dat voor alle f en alle g in die ruimte geldt dat

$$|(f, g)| \leq \|f\| \, \|g\|, \tag{3}$$

waarbij het gelijkteken alleen dan optreedt, indien $f = \alpha \, g$ voor een zeker getal α. Hier is gemakshalve aangenomen dat f niet de nulfunctie is.

Vanwege de ongelijkheid van Cauchy-Schwartz, speelt het inproduct een voorname rol bij het vergelijken van signalen. We zullen zien dat Fourier-coëfficiënten en ook wavelet-coëfficiënten ook inproducten zijn

De ruimte $\mathbb{R}^n$ is een eindig dimensionale ruimte van dimensie n. Een voorbeeld van een oneindig dimensionale ruimte zijn de oneindige rijtjes $\mathbf{x} = (x_1, x_2, \dots)$ van reële getallen waarvoor geldt

$$\sum_{i=1}^{\infty} x_i^2 < \infty.$$

In deze ruimte wordt het inproduct $(\mathbf{x}, \mathbf{y})$ van twee vectoren $\mathbf{x}$ en $\mathbf{y}$ gegeven door de oneindige som $\sum_{i=1}^{\infty} x_i\, y_i$ en natuurlijk geldt ook hiervoor de ongelijkheid van Cauchy -Schwartz (cf 3).

De natuurlijke basis $\mathbf{e_1}, \mathbf{e_2}, \dots$ ($\mathbf{e_i}$ is het rijtje met op de i-de plaats 1 en elders overal nullen) is een orthonormale basis en het is duidelijk dat

$$\mathbf{x} = \sum_{i=1}^{\infty} (\mathbf{x}, \mathbf{e_i})\, \mathbf{e_i}.$$

Een belangrijke oneindig dimensionale functieruimte, is de ruimte $L_2(\mathbb{R})$ waarvan de vectoren functies f zijn gedefinieerd op de reële getallen rechte $\mathbb{R}$ met de eigenschap

$$\int_{-\infty}^{\infty} f^2(t)\, dt < \infty$$

In de signaaltheorie wordt deze integraal de energie-inhoud van het signaal f genoemd. Omdat een signaal een functie is, zal er in deze cursus geen onderscheid worden gemaakt in een functie en een signaal. De ruimte $L_2(\mathbb{R})$ kent ook een inproduct, gegeven door

$$(f, g) = \int_{-\infty}^{\infty} f(t)\, g(t)\, dt.$$

Deze ruimte kent diverse fraaie orthonormale bases. Voor ons zijn straks orthonormale bases van wavelets van belang.

Een functie f die gedefinieerd is op het interval $[0, T]$ voor zekere $T > 0$ en waarvoor $\int_0^T f^2(t)\, dt < \infty$ kan op een natuurlijke manier worden uitgebreid tot een functie behorende tot $L_2(\mathbb{R})$. Men stelt eenvoudig $f(t) = 0$ voor alle t buiten het interval $[0, T]$. Dit heet *zeropadding*. Desondanks heeft het zijn voordeel, met name voor de frequentie-analyse van signalen met eindige tijdsduur, om deze verzameling van functies als een aparte vectorruimte te beschouwen met als inproduct

$$(f, g) = \frac{1}{T} \int_0^T f(t)\, g(t)\, dt.$$

Deze ruimte wordt aangegeven met het symbool $L_2([0,T])$.

We besluiten deze paragraaf met een korte verhandeling over deelruimten en orthogonale sommen van deelruimten. We weten dat in de $\mathbb{R}^3$ een rechte door de oorsprong of een vlak door de oorsprong op zich weer als vectorruimten kunnen worden beschouwd van dimensie 1 respectievelijk 2. De rechte en het vlak zijn voorbeelden van lineaire deelruimten. Een lineaire deelruimte van een vectorruimte is een deel van die ruimte dat met de geldende optelling en vermenigvuldiging ook weer een vectorruimte is. Deelruimten kunnen worden opgeteld. Een som van deelruimten is ook weer een deelruimte. Als L_0 en DL_1 twee lineaire deelruimten zijn van een vectorruimte V, dan is $L_0 + L_1$ de deelruimte van V die alle vectoren $\mathbf{x}$ van V bevat die geschreven kunnen worden als een som $\mathbf{x} = \mathbf{dl_1} + \mathbf{l_2}$ van een vector $\mathbf{l_1}$ in L_1 en een vector $\mathbf{l_2}$ in L_2. Als in de $\mathbb{R}^3$ bijvoorbeeld L_1 een rechte lijn is door de oorsprong en ook L_2 en deze rechte lijnen vallen niet samen, dan is de som het vlak door deze twee rechten. Vallen deze lijnen wel samen dan is de som gelijk aan die lijn.

Twee deelruimten L_1 en DL_2 zijn onderling loodrecht indien voor alle $\mathbf{dl_1}$ in L_1 en alle $\mathbf{l_2}$ in L_2 geldt dat $(\mathbf{l_1}, \mathbf{l_2}) = 0$. Wij zullen vooral sommen van onderling loodrechte deelruimten tegenkomen. Als in een vectorruimte V, de vectoren $\mathbf{a_1}, \mathbf{a_2}, \ldots$ een orthonormale basis vormen en L_k, $k = 0, 1, \ldots$ is de twee-dimensionale deelruimte met als basis $\mathbf{a_{2\,k+1}}, \mathbf{a_{2\,k+2}}$ (we zeggen ook wel D_k wordt opgespannen door $\mathbf{a_{2\,k+1}}, \mathbf{a_{2\,k+2}}$)). Dan geldt dat de ruimten onderling loodrecht zijn en dat

$$V = L_0 + L_1 + \ldots .$$

De ruimten L_i vormen een *orthogonale decompositie* van de ruimte V.

4. Fourier-analyse

Met de kennis van de voorafgaande paragraaf, zijn we in staat om de zinsnede "geluid is opgebouwd uit frequenties" beter te begrijpen. We stellen de periode waarin we een geluidssignaal waarnemen gelijk aan het tijdsinterval $[0,T]$. De Fourier-theorie leert ons dat dit signaal kan worden opgebouwd uit signalen waarvan de frequenties gehele veelvouden zijn van de zogenaamde grondfrequentie $\omega_0 = 2\,\pi/T$. De frequentie 0 correspondeert met het constante signaal $(f(t) \equiv c)$ en een signaal met frequentie $k\,\omega_0$ $(k = 1, 2, \ldots)$ is een lineaire combinatie van $c_k(t) = \cos(k\,\omega_0\,t)$ en $s_k(t) = \sin(k\,\omega_0\,t)$. We weten dat een lineaire combinatie $a_k\,\cos(k\,\omega_0\,t) + b_k\,\sin(k\,\omega_0\,t)$ van deze cosinus- en sinus functie geschreven kan worden in de vorm

$$A_k\,\cos(k\,\omega_0\,t + \varphi_k).$$

In deze representatie heet $A_k = \sqrt{a_k^2 + b_k^2}$ de amplitude van het signaal en φ_k $(\cos\varphi_k = a_k/A_k,\ \sin\varphi_k = -b_k/A_k)$ de beginfase.

De signalen $1, \cos(\omega_0\,t), \sin(\omega_0\,t), \cos(2\,\omega_0\,t), \sin(2\,\omega + 0\,t), \ldots$ zijn functies die we beschouwen als functies in de ruimte $L_2([0,T])$. Het fraaie is nu dat deze functies in $L_2([0,T])$ een basis vormen en onderling loodrecht zijn. Door

ze te "hernormeren" kunnen we ervoor zorgen dat ze allemaal lengte 1 hebben. Dit "hernormeren" levert ons dan de volgende functies op:

$$\begin{aligned} c_0(t) &= 1, \\ c_k(t) &= \sqrt{\tfrac{1}{2}}\cos(k\,\omega_0\,t), \\ s_k(t) &= \sqrt{\tfrac{1}{2}}\sin(k\,\omega_0\,t), \end{aligned}$$

Deze functies vormen een orthonornale basis van de ruimte $L_2([0,T])$, zodat iedere functie $f \in L_2([0,T])$ als volgt kan worden geschreven.

$$f = (f, c_0) + \sum_{k=1}^{\infty}((f, c_k)\,c_k + (f, s_k)\,s_k)$$

We merken op dat de signalen die een lineaire combinatie zijn van $\cos(k\,\omega_0\,t)$ en $\sin(k\,\omega_0\,t)$ een twee dimensionale deelruimte vormen van $L_2([0,T])$. We geven deze ruimte aan met het symbool U_k. Definieren we nu U_0 als de één dimensionale deelruimte van de constante functies met als basis de constante functie 1, dan hebben we een orthogonale decompositie van de ruimte $L_2([0,T])$ in termen van de ruimten P_k. Elke deelruimte past bij één bepaalde frequentie. Met de Fourier-analyse is het mogelijk om te onderzoeken welke frequenties er in een signaal voorkomen. Het nadeel is wel dat het heel moeilijk valt te achterhalen waar die frequenties voorkomen. Neem als voorbeeld het volgende signaal f dat bestaat uit twee aan elkaar geknoopte sinussignalen van verschillende frequenties op het interval $[0,T]$ met $T = 1$.

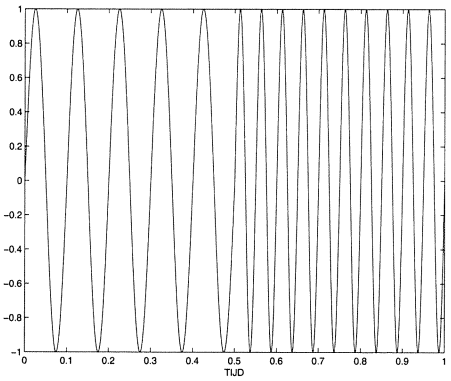
$$f(t) = \begin{cases} \sin(20\,\pi\,t) & 0 \leq t \leq 0.5, \\ \sin(40\,\pi\,t) & 0.5 < t \leq 1. \end{cases}$$

In deze cursus noemen we dit signaal de "frequency break". De grafiek van dit signaal is getekend in Figuur 2 en het amplitude- en fasespectrum in Figuur 3,

We zien duidelijk dat er twee frequenties significant zijn, echter het is erg lastig om te kunnen zien dat de hoge frequentie zich in de tweede helft van het tijdsinterval manifesteert en de lage frequentie in de eerste helft. Om lokale informatie te verkrijgen over de verschillende frequenties of variaties van een signaal hebben we een andere basisset van signalen nodig. Deze basisset moet dan kunnen weergeven welke variaties in een signaal lokaal kunnen voorkomen. Een dergelijke basisset leveren ons de wavelets.

5. Orthogonale wavelets

Een 1D-signaal beschouwen we nu als een functie in de ruimte $L_2(\mathbb{R})$. In plaats deze ruimte op te splitsen in orthogonale deelruimten horende bij bepaalde frequenties, gaan we $L_2(\mathbb{R})$ opsplitsen in ruimten die passen bij verschillende



FIGUUR 2. Grafiek van de "frequency break" $f(t)$

detail niveaus. In de wavelettheorie heet een dergelijke opsplitsing een *Multiresolutie analyse.* Wat een detail niveau is, zal uit de context duidelijk moeten worden. We geven geen formele definitie.

Bij die opsplitsing of decompositie spelen twee functies een belangrijke rol. Een zogenaamde *moederwavelet* $\psi(t)$ en een *schalingsfunctie* $\varphi(t)$. Van deze twee functies worden nieuwe functies $\varphi_{n,j}$ en $\psi_{n,j}$ $(n, j = 0, \pm 1, \pm 2, \dots)$ afgeleid op de volgende manier.

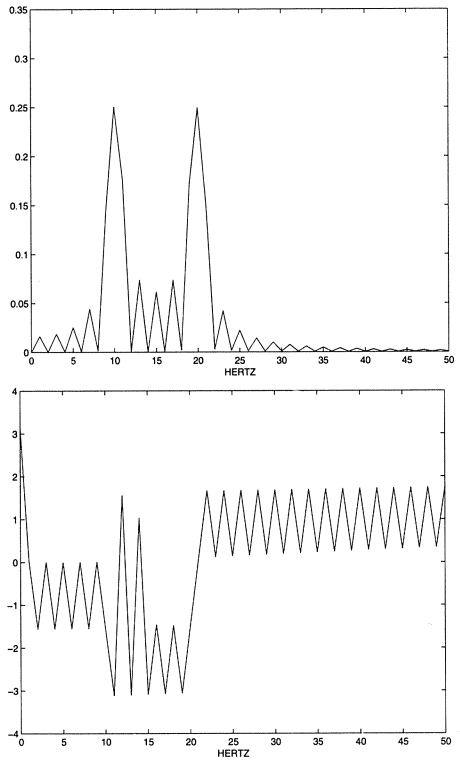
$$\varphi_{n,j}(t) = 2^{-j/2}\varphi(2^{-j}\,t - n),$$
$$\psi_{n,j}(t) = 2^{-j/2}\,\psi(2^{-j}\,t - n).$$

De factor $2^{-j/2}$ zorgt er voor dat de lengte of norm van deze functies dezelfde waarde hebben als die van φ respectievelijk ψ. We zien dat voor $j \to -\infty$ de functies $\varphi_{n,j}$ en $\psi_{n,k}$ smaller en smaller worden en dat voor $j \to \infty$ het tegenovergestelde gebeurt. Het ligt dan ook voor de hand om de parameter j te associëren met een bepaald detail niveau. De parameter n duidt een translatiestap aan. Omdat $\psi(2^j\,t - n) = \psi(2^j(t - n\,2^{-j}))$, zijn de translatiestappen klein bij kleine j en groot bij grote j. De translatiestappen passen zich dus aan bij het detail niveau.

Bij vaste j definieren we de detailruimte W_j als de deelruimte van $L_2(\mathbb{R})$, die opgespannen wordt door de functies $\psi_{n,j}$ $(n = 0, \pm 1, \pm 2, \dots)$ In deze cursus beschouwen we alleen maar die functies $\psi_{n,j}$ die onderling loodrecht zijn en lengte 1 hebben. Dat betekent dat ook de ruimten W_j onderling loodrecht zijn. Bovendien willen we dat de totale som van deze ruimten gelijk is aan de gehele $L_2(\mathbb{R})$, dus

$$L_2(\mathbb{R}) = \cdots + W_{-1} + W_0 + W_1 + \cdots \tag{5.1}$$

Dit betekent dat de ruimten W_j een orthogonale decompositie vormen van



FIGUUR 3. Het amplitude- en fasespectrum van $f(t)$

$L_2(\mathbb{R})$. Een moederwavelet ψ die hiervoor zorgt heet een *orthogonale wavelet.* In deze cursus zullen we verschillende orthogonale wavelets tegenkomen.

Het gevolg van het voorafgaande is dat elke functie f in $L_2(\mathbb{R})$ geschreven kan worden in de vorm

$$f(t) = \sum_{j=-\infty}^{\infty} D_j(t),$$

met

$$D_j(t) = \sum_{n=-\infty}^{\infty} d_{n,j}\, \psi_{n,j}(t), \tag{5}$$

$$d_{n,j} = (f, \psi_{n,j}). \tag{6}$$

Men zou kunnen zeggen dat f_j de functie f vertegenwoordigt op detail niveau j. De functie f_j heet de detailfunctie van f op detail niveau j. De coëfficiënten $d_{n,j}$ worden *wavelet-coëfficiënten* genoemd.

Uit (6) zou men kunnen concluderen dat het berekenen van wavelet-coëfficiënte het berekenen van inproducten betekent. Echter als er sprake is van een scha-

lingsfunctie is het berekenen van inproducten overbodig en zijn er snelle algoritmen beschikbaar. Daarom zullen we nog een een verdere eis stellen aan een de orthogonale wavelet. Bij deze eis speelt de schalingsfunctie φ een rol.

Het is duidelijk dat een functie f, waarvan alle waveletcoëfficiënten $d_{n,j}$ gelijk zijn aan 0 voor alle n en voor $j = 0, -1, -2, \ldots$ behoort tot de ruimte $V_0 := W_1 + W_2 + \cdots$. Onze voorwaarde is dat deze ruimte een speciale orthonormale basis heeft, die bestaat uit alle translaties $\varphi(t-n); (n = 0, \pm 1, \pm 2, \ldots)$ van de schalingsfunctie φ. Een gevolg hiervan is dat bij vaste k de set $\varphi_{n,k}$ $(n = 0, \pm 1, \pm 2, \ldots$ een orthonormale basis is van

$$V_k = W_{k+1} + W_{k+2} + \cdots .$$

In deze cursus gaan we niet in op de constructie van paren φ en ψ die het bovenstaande realiseren. Het bekende boek "Ten lectures on Wavelets" van Ingrid Daubechies (cf [1]) geeft hierover uitgebreide informatie. Verder zijn er heel wat leerboeken op dit gebied verschenen. In de referentielijst staan er enkele vermeld (cf [3],[4], [5] en [6]). Paragraaf 5.2 gaat over een van de meest eenvoudige voorbeelden van een orthogonale moederwavelet, de zogenaamde *Haar-wavelet*. We zullen deze wavelet met de bijbehorende schalingsfunctie hier alvast introduceren.

$$\psi(t) = \begin{cases} 1 & 0 < t < 0.5, \\ -1 & 0.5 < t < 1, \\ 0 & \text{elders}, \end{cases} \tag{7}$$

$$\varphi(t) = \begin{cases} 1 & 0 < t < 1, \\ 0 & \text{elders}. \end{cases} \tag{8}$$

In Paragraaf 5.2 in Figuur (5) zijn enkele detailfuncties (D_j (cf. 5) weergegeven van de "frequency break", die met de Haar-wavelet zijn bepaald.

We hebben al opgemerkt dat het berekenen van wavelet-coëfficiënten op een efficiënte manier kan worden uitgevoerd via de zogenaamde discrete wavelet-transformatie. Dit betekent dat de wavelet-coëfficiënten worden berekend via filterbewerkingen. In Paragraaf 5.3 wordt hieraan aandacht besteed. In Paragraaf 5.2 komt het berekenen van wavelet-coëfficiënten eveneens aan de orde, maar dan speciaal voor de Haar-wavelet.

De filtercoëfficiënten bij de discrete wavelet-transformatie komen voort uit een aantal belangrijke betrekkingen die er voor de paren φ en ψ bestaan. Fundamenteel voor onze orthogonale wavelet is de volgende betrekking, waarin alleen de schalingsfunctie voorkomt.

$$\phi(t) = \sum_{k=0}^{M-1} p_k \, \varphi(2\,t - k), \tag{9}$$

voor zekere getallen $p_0, p_1, \ldots, p_{M-1}$. Deze betrekking heeft als gevolg dat de schalingsfunctie $\varphi(t)$ identiek nul is buiten het interval $[0, M-1]$. We noemen

$[0, M-1]$ de *drager* van φ. Een geschikte orthogonale moederwavelet kan nu uit de schalingsfunctie worden gemaakt met behulp van de betrekking

$$\psi(t) = \sum_{k=0}^{N-1} (-1)^k \, q_k \, \varphi(2\,t - k), \tag{10}$$

met

$$q_k = (-1)^k \, p_{N-k-1}.$$

Net als de schalingsfunctie φ heeft nu ook de moederwavelet ψ als drager het interval $[0, M-1]$.

De volgende betrekking blijkt bij de decompositie van functies in detail-functies van belang te zijn.

$$\phi(2\,t - n) = \tfrac{1}{2} \sum_{k=-\infty}^{\infty} (p_{n-2\,k} \, \varphi(t-k) + q_{n-2\,k} \, \psi(t-k)). \tag{11}$$

5.1. Voorbeelden van orthogonale wavelets

De meest eenvoudige wavelet, de Haar-wavelet, is reeds in de vorige paragraaf geïntroduceerd. De Haar-wavelet is in feite een exemplaar uit een hele familie van wavelets, de zogenaamde Daubechies wavelets, die vernoemd zijn naar de Belgische wiskundige Ingrid Daubechies. We weten dat de getallen p_l in Formule (9) in feite de orthogonale wavelet ψ bepalen. We stellen in deze formule $M = 2\,N$ met $N = 1, 2, 3, \ldots$. Ingrid Daubechies heeft voor elke N een rij p_l gemaakt, die een geschikte schalingsfunctie φ levert en een moederwavelet ψ. Deze moederwavelet heet dan de Daubechies-wavelet van de orde N. De Haar-wavelet is een Daubechies wavelet van de orde 1. Deze wavelet is een discontinue functie. Bij toenemende N worden de schalingsfuncties en dus ook de moederwavelet gelukkig gladder. Dit gaat overigens wel ten koste van de lengte van de drager van φ en ψ en dus ook van de lengte van de filters die bij het berekenen van wavelet-coëfficiënten worden gebruikt (zie paragraaf 5.3). Een voordeel is wel dat het aantal "vanishing moments" bij toenemende N toeneemt. Dit aantal is van belang als we de wavelets willen gebruiken voor het comprimeren van data. Men kan afleiden dat voor elke wavelet ψ de gemiddelde waarde gelijk is aan nul, i.e.

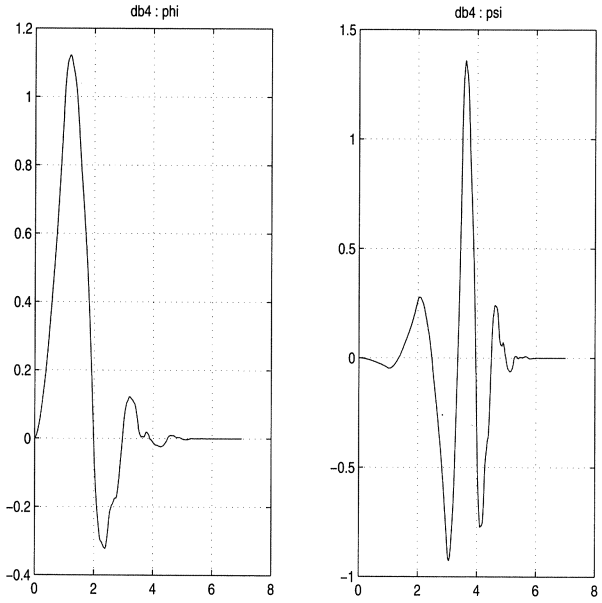
$$\int_{-\infty}^{\infty} \psi(t)\, dt = 0.$$

Het μ-de moment $\mu(\psi)$ ($\mu = 0, 1, 2, \ldots$) van een wavelet wordt gedefinieerd door

$$\mu(\psi) = \int_{-\infty}^{\infty} t^{\mu} \, \psi(t)\, dt. \tag{12}$$

Nu geldt voor de Daubechies wavelet van de orde N dat alle momenten tot en met $N-1$ gelijk zijn nul. De Daubechies wavelets , de Haar-wavelet

uitgesloten, kan men niet door middel van formules met elementaire functies worden weergegeven. Voor het rekenen met wavelets zal men moeten volstaan met de formules (9) en (10). Dit is overigens geen bezwaar. De Daubechies wavelet van de orde 4 met de bijbehorende schalingsfunctie zijn getekend in Figuur (4)



FIGUUR 4. Daubechies wavelet van de orde 4 met schalingsfunctie

In de Wavelet-Toolbox van MATLAB zijn verschillende orthogonale wavelet families opgenomen, waaronder de Symlet wavelet, de Coiflet wavelet etc. Deze wavelets onderscheiden zich in gladheid, in het aantal vanishing moments en eigenschappen van de bijbehorende filters. We gaan hier verder niet op in.

5.2. Decompositie met de Haar-wavelet

Een heel bekend voorbeeld van een moederwavelet is de zogenaamde Haar-wavelet, die we reeds eerder hebben ingevoerd.

De formules (9), (10) zien er voor de Haar-wavelet er als volgt uit.

$$\phi(t) = \phi(2\,t) + \phi(2\,t-1),$$
$$\psi(t) = \varphi(2\,t) - \varphi(2\,t-1).$$

Formule (11) kan worden afgeleid uit de voor de Haar-wavelet gemakkelijk

te verifieren eigenschappen

$$\phi(2\,t) = \tfrac{1}{2}\phi(t) + \tfrac{1}{2}\psi(t),$$
$$\phi(2\,t-1) = \tfrac{1}{2}\phi(t) - \tfrac{1}{2}\psi(t),$$

In de praktijk worden wavelet-coëfficiënten berekend via een discrete wavelet-transformatie. Dit betekent dat met eenvoudige filterbewerkingen de coëfficiënten worden bepaald. We zullen dit voor de Haar-wavelet laten zien. Hierbij wordt wel verondersteld dat van het gegeven signaal $f(t)$ alleen een bemonstering beschikbaar is op tijdstippen, die op de tijd-as onderling gelijke afstand hebben. Gemakshalve stellen we deze afstand gelijk aan 1 en noteren $a_n = f(n)$ voor alle gehele n. Met het verschalen van de tijd-as kan dit worden bereikt. In de punten n, waar f niet is gedefinieerd, stellen we $a_n = 0$ (zero padding). Men dient zich te realiseren dat door het bemonsteren informatie op fijne schaal verloren is gegaan. Het heeft dan ook geen zin om wavelet-coëfficiënten $d_{n,j}$ uit te rekenen met $j \geq 0$. We weten dat een functie waarvoor $d_{n,j} = 0$ voor alle $j \leq 0$ en voor alle gehele n te schrijven is als een lineaire combinatie van $\varphi(t-n)$. Voor de schalingsfunctie bij de Haar-wavelet is deze lineare combinatie een stuksgewijs constante functie met eventuele sprongpunten in de gehele getallen. Het ligt dan ook voor de hand om $f(t)$ te vervangen door

$$f_0(t) = \sum_{n=-\infty}^{\infty} a_n\,\varphi(t-n).$$

De wavelet-coëffciënten $d_{n,j}$, gaan we schaal voor schaal äfpellen" van f_0. De manier waarop volgt uit het volgende rekenwerk. In dit rekenwerk is $a_{n,0} = a_n$ gesteld.

$$f_0(t) = \sum_{n=-\infty}^{\infty} a_{n,0}\,\varphi(t-n) =$$
$$\sum_{l=-\infty}^{\infty} a_{2\,l,0}\,\varphi(t-2\,l) + \sum_{l=-\infty}^{\infty} a_{2\,l+1,0}\,\varphi(t-2\,l-1) =$$
$$\sum_{l=-\infty}^{\infty} \tfrac{1}{2}\,a_{2\,l,0}\varphi(t/2-l) + \sum_{l=-\infty}^{\infty} \tfrac{1}{2}\,a_{2\,l,0}\psi(t/2-l) +$$
$$\sum_{l=-\infty}^{\infty} \tfrac{1}{2}\,a_{2\,l+1,0}\varphi(t/2-l) - \sum_{l=-\infty}^{\infty} \tfrac{1}{2}\,a_{2\,l+1,0}\psi(t/2-l) =$$
$$\tfrac{1}{\sqrt{2}} \sum_{l=-\infty}^{\infty} a_{l,1}\,\varphi_{l,1}(t) + \tfrac{1}{\sqrt{2}} \sum_{l=-\infty}^{\infty} d_{l,1}\,\psi_{l,1}(t).$$

Hierin is

$$a_{l,1} = \frac{1}{\sqrt{2}} a_{2\,l,0} + \frac{1}{\sqrt{2}} a_{2\,l+1,0},$$
$$d_{l,1} = \frac{1}{\sqrt{2}} a_{2\,l,0} - \frac{1}{\sqrt{2}} a_{2\,l+1,0}. \tag{13}$$

We hebben het signaal $f_0(t)$ gesplitst in een approximatiesignaal $f_1(t)$ en een detailsignaal $D_1(t)$. Dus

$$f_0(t) = f_1(t) + D_1(t)$$

met

$$f_1(t) = \sum_{l=-\infty}^{\infty} a_{l,1}\, \varphi_{l,1}(t),$$

en

$$D_1(t) = \sum_{-\infty}^{\infty} d_{l,1}\, \psi_{l,1}(t).$$

Ditzelfde rekenwerk kan ook worden uitgevoerd op het approximatiesignaal $f_1(t)$. dan krijgen we $f_1(t) = f_2(t) + D_2(t)$ met

$$f_2(t) = \sum_{l=-\infty}^{\infty} a_{l,2}\, \varphi_{l,2}(t),$$
$$D_2(t) = \sum_{-\infty}^{\infty} d_{l,2}\, \psi_{l,2}(t).$$

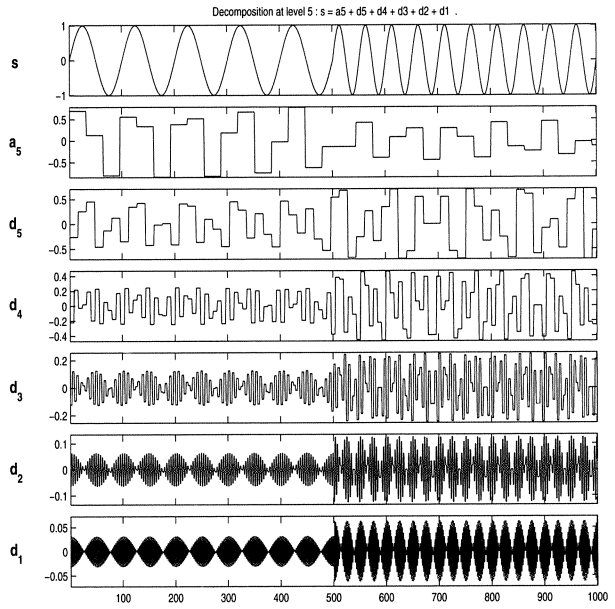
De nieuwe coëfficiënten voldoen weer aan dezelfde betrekkingen als gegeven in (13), dus

$$a_{l,2} = \frac{1}{\sqrt{2}} a_{2\,l,1} + \frac{1}{\sqrt{2}} a_{2\,l+1,1},$$
$$d_{l,2} = \frac{1}{\sqrt{2}} a_{2\,l,1} - \frac{1}{\sqrt{2}} a_{2\,l+1,1}.$$

De decompositie kan worden voortgezet. Elke keer opnieuw pellen we wavelet coëfficiënten af. Doen we dit zeg k keer dan hebben we uiteindelijk één approximatierij $a_{l,k}$ en k detailrijen $d_{l,1}$ tot en met $d_{l,k}$ met wavelet-coëfficiënten ter beschikking. Bij de approximatierij past het signaal $f_k(t) = \sum_{l=-\infty}^{\infty} a_{l,k}\, \varphi_{l,k}(t)$ en bij de rij $d_{l,j}$ het signaal $D_j(t) = \sum_{l=-\infty}^{\infty} d_{l,j}\, \psi_{l,j}(t)$. Bovendien geldt

$$f_0(t) = f_k(t) + \sum_{j=1}^{k} D_j(t).$$

Met de wavelet Toolbox van MATLAB zijn voor $k = 5$ deze zes signalen getekend. We zien dat bij steeds grovere schaal de hoekigheid van de schalingsfunctie en van de Haar-wavelet terug. Voor de analyse van een frequentie break zijn de Haar-wavelets niet zo geschikt. We gebruikken hiervoor liever gladdere wavelets.



FIGUUR 5. Decompositie van de "frequency break" tot en met detailniveau 5

5.3. Wavelets en filters

De manier waarop uit $a_{l,0}$ de nieuwe approximatiecoëfficiënten $a_{l,1}$ en de waveletcoëfficiënten $d_{l,1}$ worden gemaakt, kan worden gezien als een filterproces gevolgd door een zogenaamde downsampling. Dit is echter niet voorbehouden voor alleen de Haar-wavelet. Gebruiken we een willekeurige andere orthogonale wavelet om een signaal $f(t)$ at te pellen, dan gaan we ook $f(t)$ vervangen door

$$f_0(t) = \sum_{n=-\infty}^{\infty} a_n \, \varphi(t),$$

met a_n de waarden van de samples van f en φ de bijbehorende schalingsfunctie. De fout die dan wordt gemaakt, wordt gemakshalve verwaarloosd. Bij een hoge bemonsteringsfrequentie en een φ met en erelatief smalle drager is dit ook het geval, anders is men aangewezen op geavanceerde methoden om een goede set van a_n te maken.

Door gebruik te maken van de Formule (11), kan men net als bij de Haar-

wavelet aantonen dat $f_0(t) = f_1(t) + D_1(t)$ met

$$f_1(t) = \sum_{-\infty}^{\infty} a_{l,1}\, \varphi_{l,1}(t),$$

$$f_2(t) = \sum_{-\infty}^{\infty} d_{l,1}\, \psi_{l,1},$$

waarin

$$a_{l,1} = \frac{1}{\sqrt{2}} \sum_{k=-\infty}^{\infty} a_{k,0}\, p_{k-2\,l}, \tag{14}$$

$$d_{l,1} = \frac{1}{\sqrt{2}} \sum_{k=-\infty}^{\infty} a_{k,0}\, q_{k-2\,l}. \tag{15}$$

Op welke wijze deze bovenstaande sommen gerelateerd zijn met lineaire filtering gevolgd door downloading laten we nu zien.

Als een rij getallen y_l wordt geconstrueerd uit een gegeven rij getallen x_l via een formule van de vorm

$$y_l = \sum_{k=-\infty}^{\infty} c_k\, x_{l-k} \qquad (l = 0, \pm 1, \pm 2, \dots), \tag{16}$$

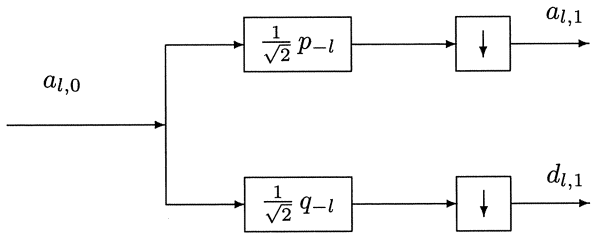
dan zeggen we dat de rij y_l is ontstaan uit x_l door middel van een lineaire filtering met filtercoëfficiënten c_l. Nu kunnen we in formule (14) $a_{l,1}$ ook als een dergelijke som schrijven. Merk op dat $a_{l,1}$ wordt verkregen ddor de som

$$\tfrac{1}{\sqrt{2}} \sum_{k=-\infty}^{\infty} a_{k,0}\, p_{k-l}$$

te berekenen en daarna l te vervangen door $2\,l$. Deze laatste stap wordt nu downsampling genoemd. De som is een lineaire filtering met filtercoëfficiënten $\frac{1}{\sqrt{2}}\, p_{-l}$, waarvan er gelukkig slechts eindig veel ongelijk zijn aan nul. Ook de rij $d_{l,1}$ wordt door filtering en downloading verkregen uit $a_{l,0}$. Van dit filter bevat de rij $\frac{1}{\sqrt{2}}\, q_{-l}$ de filtercoëfficiënten. In onderstaand figuur hebben we de filters met downloading schematisch weergegeven.

Natuurlijk kan men deze filters vervolgens toepassen op de rij $a_{l,1}$ om de wavelet-coëfficiënten $d_{l,2}$ te berekenen. Voor onze "frequency break" is dit op 10 niveaus gebeurd. Het resultaat is als een grijswaarde plaatje weergegeven in Figuur (7).

De gebruikte wavelet is een zogenaamde Daubechies wavelet van de orde 3. In de afbeelding correspondeert een donkere grijstint met een (in absolute waarde) relatief grote waveletcoëfficiënt. In de verticale richting staan de schaalwaarden $a = 2^j, j = 0, 1, 2, \dots, 8$ van fijn naar grof en in de horizontale



FIGUUR 6. Decompositie

richting de translatiestappen, die afgestemd zijn op het aantal samples (in ons geval 1000) op het tijdsinterval $[0, 1]$. Het nu al duidelijk dat rond $t = 0.5$ een overgang van grof naar fijn detail optreedt.

Het bewerken van signalen met behulp van wavelets komt neer op het in eerste instantie het splitsen van een signaal in wavelets om vervolgens sommige wavelet-coëfficiënten te veranderen of in geval ze erg klein zijn weg te laten. Zo ontstaan er een nieuwe detailsignalen die we met behulp van het overgebleven approximatiesignaal weer worden gereconstrueerd naar een signaal op het oorspronkelijk niveau.

Stel we hebben de beschikking over de approximatie-coëfficiënten $a_{l,1}$ en over de wavelet-coëfficiënten $d_{l,-1}$ en de vraag is hoe de oorspronkelijke coëfficiënte $a_l = a_{l,0}$ kunnen worden teruggevonden. Hiervoor zijn de formules (9) en (10) belangrijk. Uit deze formules volgt namelijk dat

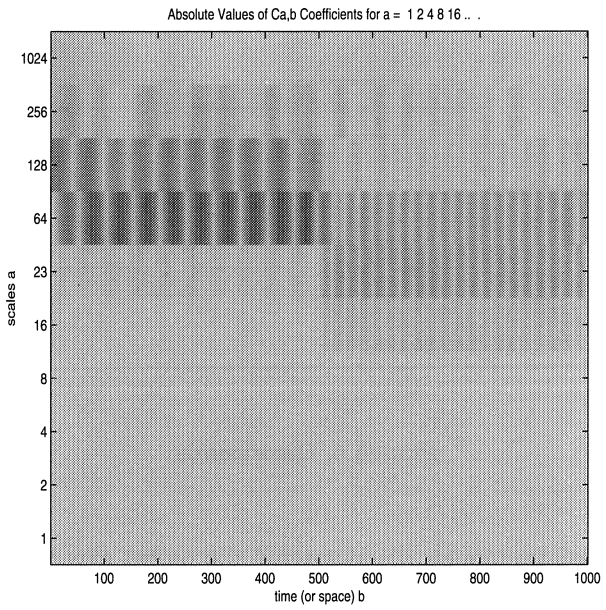
$$a_{l,0} = u_{l,0} + v_{l,0},$$

met

$$u_{l,0} = \frac{1}{\sqrt{2}} \sum_{k=-\infty}^{\infty} a_{k,1}\, p_{l-2\,k},$$

$$u_{l,0} = \frac{1}{\sqrt{2}} \sum_{k=-\infty}^{\infty} a_{k,1}\, p_{l-2\,k}.$$

Net als bij de decompositie van een signaal, kunnen deze sommen geassociëerd worden met een lineair filterproces, maar in plaats van downsampling na het filteren, wordt nu van te voren een upsampling op de te filteren rij toegepast. Upsampling wil zeggen dat we steeds tussen twee opeenvolgende elementen van een rij een nul plaatsen. Passen we dit toe op de rij x_l dan ontstaat er een rij y_l met $y_{2\,l} = x_l$ en $y_{2\,l+1} = 0$. Upsampling en daarna filteren met filtercoëfficiënten c_l geeft dus voor een invoerrij x_l het resultaat $\sum_{k=-\infty}^{\infty} x_k\, c_{l-2\,k}$.



FIGUUR 7. Grijswaardenplaatje van waveletcoëfficiënten

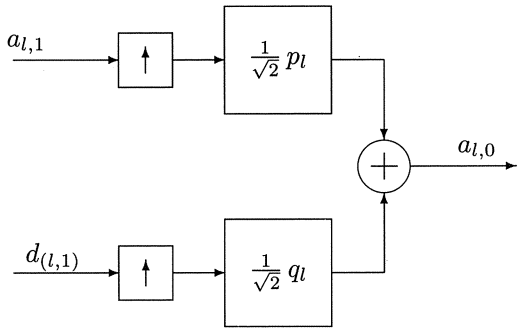
Het filter bij de berekening van $u_{l,0}$ kent dus als filtercoëfficiënten de rij $\frac{1}{\sqrt{2}}\, p_l$. De rij $\frac{1}{\sqrt{2}}\, q_l$ bevat de filtercoëfficiënten voor de berekening van $v_{l,0}$. De reconstructie is schematisch weergegeven in Figuur (8)

5.4. Wavelets in twee dimensies

De theorie voor orthogonale wavelets in meer dimensies kent dezelfde opbouw als die in één dimensie. Er is wel een belangrijk verschil. Bijvoorbeeld bij functies gedefinieerd op het x-y vlak hebben we in plaats van één moederwavelet drie moederwavelets nodig. Om dit in te zien gaan we uit van een functie an het type $f(x)\, g(y)$ en proberen deze te splitsen met behulp van één-dimensionale wavelets met een schalingsfunctie $\varphi_1(x)$ en moederwavelet $\psi_1(x)$ voor de x-richting en het paar $\varphi_2(y)$, $\psi_2(y)$ voor de y-richting. Zoals we dat in de een-dimensies gewend zijn, gaan we de functie $f(x)$ vervangen door $f_0(x)$ en $g(y)$ door $g_0(y)$ en vervolgens worden zowel $f_0(x)$ als $g_0(y)$ gesplitst naar detailniveau 1. De detailsignalen hierbij noemen we $D_1^1(x)$ respectievelijk $D_1^2(y)$. Dus

$$f_0(x)\, g_0(y) = (f_1(x) + D_1^1(x))\, (g_1(y) + D_1^2(y)) =$$
$$f_1(x)\, g_1(y) + f_1(x)\, D_1^2(y) + D_1^1(x)\, g_1(y) + D_1^1(x)\, D_1^2(y).$$

We onderscheiden hierin een nieuw approximatie signaal $f_1(x)\, g_1(y)$ en drie detailsignalen. Het signaal $f_1(x)\, D_1^2(y)$ met alleen details in de verticale richting (y-richting); het signaal $D_1^1(x)\, g_1(y)$ met alleen details in de horizontale



FIGUUR 8. Reconstructie

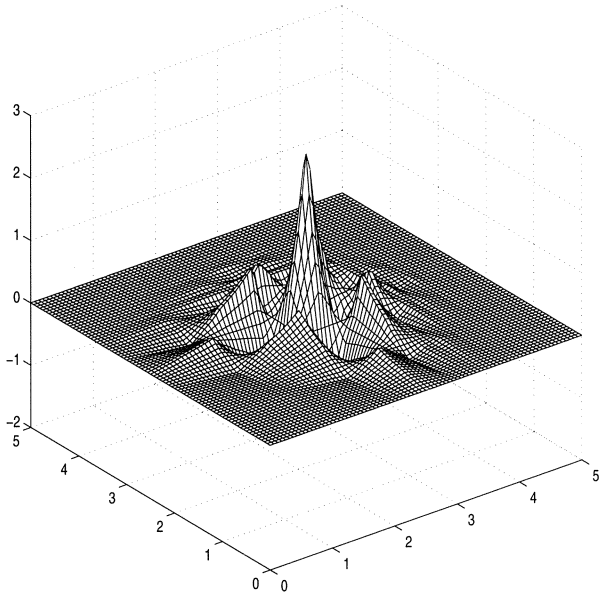
(x-richting) en het detailsignaal $D_1^1(x)\, D_2^(y)$ met details zowel in de x-richting als in de y-richting. Deze laatste wordt in de wavelettheorie ook wel het detailsignaal in de diagonale richting genoemd. Het is ook duidelijk dat de twee-dimensionale schalingsfunctie gelijk is aan het product $\phi(x,y) = \phi_1(x)\,\phi_2(y)$ van de een-dimensionale schalingsfuncties en dat er sprake is van drie moederwavelets te weten:

$$\begin{aligned}
\psi^1(x,y) &= \varphi_1(x)\,\psi_2(y),\\
\psi^2(x,y) &= \psi_1(x)\,\varphi_2(y),\\
\psi^3(x,y) &= \psi_1(x)\,\psi_2(y).
\end{aligned}$$

Meestal worden bij de twee-dimensionale voor de x-richting en y-richting dezelfde wavelets en schalingsfuncties gebruikt. In Figuur 9 is de wavelet getekend voor de diagonale details met behulp van de Daubechies wavelet van de orde 3.

Men kan ook twee-dimensionale wavelets construeren die niet van een product $\varphi_1(x)\,\varphi_2(y)$ afkomstig zijn, maar van een " echte" twee-dimensionale schalingsfunctie $\varphi(x,y)$. Dan ook dienen er drie moederwavelets te worden gereconstrueerd. We gaan hier verder niet op in.

Het berekenen van wavelet-coëfficiënten in twee dimensies kan ook weer efficiënt worden uitgevoerd met filters en downloading. Ook hier zullen we ons niet verder uitwijden. In de volgende paragraaf laten we een voorbeeld zien van een toepassing van een waveletdecompositie op beelden.



FIGUUR 9. Een bivariate Daubechies wavelet, orde 3 en diagonale detail

6. TOEPASSINGEN EN VERDERE ONTWIKKELINGEN

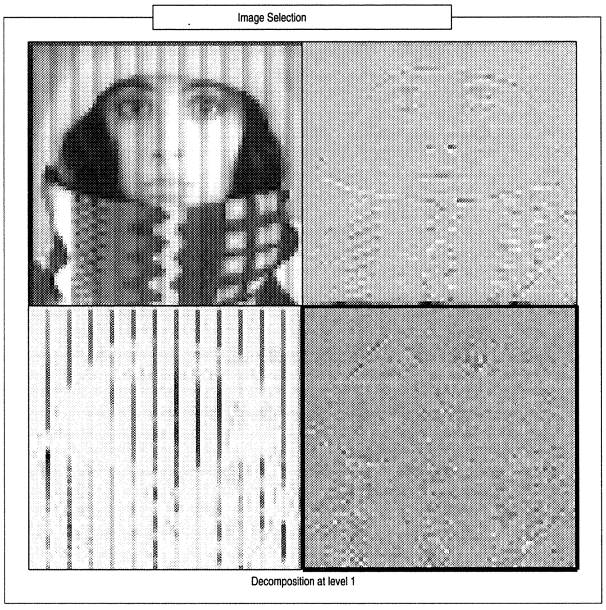
In de afbeelding hierna is een jonge dame achter tralies geplaatst.

Deze dame is naar detailniveau 1 gesplitst in vier beelden met behulp van de een-dimensionale Daubechies wavelets van de orde 3.

Met de klok mee, vanaf linksboven, zien we het nieuwe approximatiebeeld, een detailbeeld in de horizontale richting, de diagonale richting en de verticale richting. Het detailbeeld in de verticale richting bevat blijkbaar de tralies. Door de corresponderende wavelet coëficiënten op nul te zetten kan de jonge dame worden bevrijd.

6.1. Compressie van data

Omdat met behulp van wavelets een beeld kan worden gesplitst op diverse detailniveau's, zijn we in staat om in elk niveau wavelet-coëfficiënten die in absolute waarde klein zijn weg te laten. Hiervoor zijn in de regel wavelets geschikt die op z'n minst 2, liever 3, opeenvolgende vanishing moments hebben. Op delen van het beeld met weinig detail kan het percentage kleine wavelet-coëfficiënten behoorlijk groot, zodat een grote data-reductie kan worden bereikt. Tijdens het hoorgedeelte van deze cursus zal met de Wavelet Toolbox van MATLAB (cf. [7])enkele demo's worden getoond.



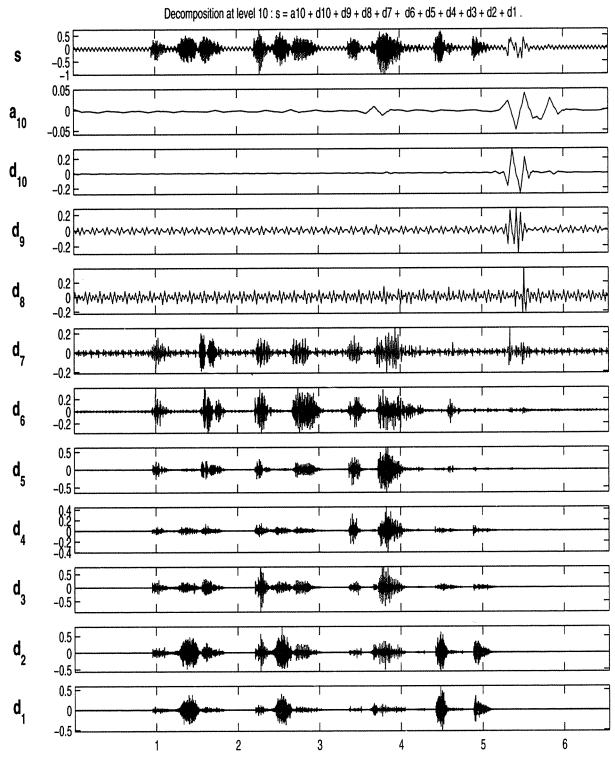
FIGUUR 10. Een jonge dame achter tralies

6.2. Onderdrukken van ruis in audio signalen

Wavelets kunnen ook goed gebruikt worden voor het onderdrukken van ruis. Het boekje van James Walker (cf. [10]) geeft hiervan enkele mooie voorbeelden. In feite wordt hier dezelfde techniek gebruikt als bij data-compressie. Op elk detailniveau wordt geprobeerd de ruis te herkennen in de wavelet-coëfficiënten. Deze coëfficiënten worden op nul gezet. Op nul zetten gebeurt via een keuze van een treshold waarde, die in feite per detailniveau mag verschillen. De Wavelet Toolbox van MATLAB kent enkele commando's die kunnen helpen bij een goede keuze van treshold waarden. In het spraaksignaal " ik zie ik zie wat jij niet ziet" in Paragraaf 2 zit een stoorsignaal die afkomstig is van het elektrisch netwerk tijdens de opname van het signaal. Het signaal is gesplitst met behulp van de Daubeschies wavelet van de orde 3 tot en met niveau 10 (cf Figuur 11). In de de detailsignalen vanaf $D_7(t)$ zien we de storing terug. De decompostie is gemaakt met de Wavelet Toolbox van MATLAB ([7]).

6.3. Opsporen van features in een signaal

Een breed toepassingsgebied van wavelets is het opsporen van speciale kenmerken (features) in een signaal. Bijvoorbeeld bij het lokaliseren van het epicentrum van een aardbeving in een meetstation, denk bijvoorbeeld aan het KNMI, is het belangrijk om te weten wat het tijdsverschil is van binnenkomst van de zogenaamde P-golf en S-golf die in de binnengekomen trilling zijn opgeborgen. Bij het bepalen van dit tijdsverschil spelen wavelets een nuttige rol (zie [8]).



FIGUUR 11. Decompositie van het spraaksignaal "Ik zie ... "

6.4. Een nieuwe generatie wavelets

Een van de beperkingen van de wavelets die door verschaling en translatie zijn gemaakt uit één moederwavelet is dat er per detailniveau steeds sprake is van één staplengte en dat de signalen op een of andere manier moeten worden voortgezet buiten het interval, waar ze zijn gegeven. Dit veroorzaakt allerlei randstoringen. Bijvoorbeeld zeropadding geeft meestal een dicontinuiteit aan de rand en dit werkt door als een artifact in de wavlet-coëfficiënten. Dit soort problemen hebben aanleiding gegeven tot een nieuwe generatie wavelets, waar nog steeds het splitsen van een signaal in een approximatiesignaal en een detail-signaal een centrale rol speelt. Er is geen sprake meer van een enkele wavelet. Met de nieuwe generatie wavelets zijn het aantal toepassingsmogelijkheden zelfs nog groter geworden. De geïnteresseerde lezer zij verwezen naar het artikel van Wim Sweldens (http://www.sweldens.com). Wim mag beschouwd worden als een van de geestelijke vaders van de nieuwe generatie wavelets. Hij is ook de redacteur van het bekende elektronisch magazine op het gebied van wavelets, The Wavelet Digest., met boordevol actuele informatie over toepassingen en nieuwe ontwikkelingen.

REFERENTIES

1. I. DAUBECHIES, *Ten lectures to wavelets*, CBMS-NSF Regional Conference Series in Applied Mathematics, SIAM, Philadelphia 1992.
2. I. DAUBECHIES, "Where do wavelets come from?—A personal point of view", *Proceedings of the IEEE Special Issue on Wavelets* **84** (no. 4), 510–513, April 1996.
3. C. SIDNEY e.a., *Introduction to Wavelets and Wavelet transforms - A primer*, Prentice Hall, New Jersey,1998
4. C.R.TRAAS, H.G.TER MORSCHE, R.M.J. VAN DAMME, *Splines en Wavelets*,Epsilon uitgaven Utrecht, 2000
5. G.STRANG, TRUONG NGUYEN, *Wavelets and Filter banks*, Wellesley Cambridge Press, 1996
6. C.K. CHUI, *An introduction to wavelets*, Academic Press, New York, 1992
7. MICHEL MISITI e.s., *Wavelet Toolbox, for use with MATLAB*, The Mathworks Inc, Mass.,1996
8. P.J. OONINCX, *Automatic phase dedection in seismic data using the discrete wavelet transform*, CWI-report PNA-R9811,1998
9. W. SWELDENS, The Lifting Scheme: A New Philosophy in Biorthogonal Wavelet Constructions, *Wavelet Applications in Signal and Image Processing III*, A. F. LAINE, M. UNSER, Proc. SPIE 2569, 1995, 68–79
10. J.S. WALKER, *Primer on Wavelets and their Scientific Applications*, Chapman&Hall/CRC, Boca Raton, London, 1999.

CWI Vakantiecursus 2000

pp. 65 – 90

Een computerwerkplaats voor wiskunde

André Heck
Universiteit van Amsterdam

1. ICT IN DE EXACTE VAKKEN

In onderwijs op VWO/HAVO wordt een steeds actievere rol aan de leerling toegekend. Belangrijk argument is dat leerlingen meer leren door zelf actief bezig te zijn. Het in de eind jaren tachtig opgekomen constructivisme bepleit een onderwijsaanpak waarin leerlingen op basis van reeds aanwezige kennis hun eigen ideeën vormen, deze uitwerken en toetsen, en door na te denken over eigen handelen nieuwe kennis vergaren. In deze visie past het uitvoeren van praktische opdrachten en het doen van eigen onderzoeken. Om leerlingen beter voor te bereiden op een vervolgopleiding en latere beroepspraktijk worden nog voorwaarden gesteld zoals het zelfstandig en in teamverband planmatig kunnen werken en het functioneel gebruiken van informatie- en communicatietechnologie (ICT). De moderne leerkracht houdt zich bij dit alles meer bezig met het ontwerpen, voorbereiden en begeleiden van leeractiviteiten dan met de traditionele doceertaak. Op zichzelf allemaal mooie ideeën, maar het onderwijs moet dit dan wel mogelijk maken!

Een effectieve en efficiënte omgeving is nodig die leerlingen aanzet tot actief leren en ze in staat stelt om taken — in het bijzonder praktische opdrachten en het profielwerkstuk — goed uit te voeren. Zo'n omgeving moet erg leerlinggestuurd zijn, maar ook de docent de mogelijkheid bieden om het leerproces te volgen, waar nodig bij te sturen en tenslotte het werk te beoordelen. ICT is hier één van de hulpmiddelen voor. Wat het er echter niet gemakkelijker op maakt is dat elk schoolvak een eigen accent op de ICT-gereedschappen legt. Bij de natuurwetenschappelijke vakken zijn dit middelen om metingen te doen, gegevens te verwerken, en om deze dan te vergelijken met resultaten uit computersimulaties. Bij techniek gaat het om het combineren van meten van gegevens en aansturen van apparaten. Wiskunde heeft behoefte aan rekenfaciliteiten — exact, numeriek en grafisch — om gemakkelijk berekeningen uit te kunnen voeren en om gegevens te verwerken. Ook is er behoefte aan het kunnen werken met wiskundige modellen op de computer, in het bijzonder daar waar exacte oplossingen niet mogelijk zijn of de nodige wiskundekennis bij leerlingen ontbreekt. In alle vakken wordt van leerlingen verwacht dat zij kennis en vaardigheden kunnen toepassen in situaties van het dagelijkse leven. Manieren om deze situaties in de klas te brengen zijn het door leerlingen zelf laten uitvoeren van experimenten, hen laten werken met reële data uit statistisch onderzoek en met informatie gevonden op Internet. Daarnaast bieden digitale videobronnen waaraan gemeten kan worden een goede mogelijkheid om met echte, zelf gekozen data te werken.

Hoe divers het ICT-gebruik in exacte vakken ook is, toch is er is een gemeenschappelijke rol van ICT in exacte vakken te herkennen. Deze rol hangt samen met een didactische en onderwijskundige vernieuwing die de laatste jaren in gang is gezet. In het verleden stond de leerstof centraal en was de activiteit van leerlingen bij exacte vakken vooral gericht op het oplossen van standaardvraagstukken. Impliciet werd verondersteld dat het begrip dan wel vanzelf zou volgen. Een doorgaande hervorming van het onderwijs in bètavakken is om de concepten voldoende aandacht te geven en het leerproces van de leerlingen een belangrijke plaats toe te bedelen. De natuurwetenschappelijke vakken lopen hiermee voorop, maar ook bij wiskunde maakt probleem-oplossen meer en meer plaats voor wiskundig modelleren, redeneren en presenteren. De rol van ICT is om deze nieuwe activiteiten te accommoderen: toepassen van ICT in exacte vakken is dan ook vooral omgaan met ICT-gereedschappen. Computeralgebra-systemen nemen het vervelende wiskundige handwerk over en bieden toegang tot het brede spectrum van wiskundige methoden en technieken. Modelleerprogramma's maken het mogelijk om de concepten achter een onderwerp te bestuderen zonder dat hierbij de nadruk gelegd hoeft te worden op oplossingsmethoden. Simulatiepakketten maken het mogelijk om parameters in een model te variëren en het effect hiervan in eindresultaten te onderzoeken. Softwarepakketten voor tekstverwerking en vormgeving motiveren leerlingen om op professionele wijze werkzaamheden af te sluiten en niet eerder te rusten voor een goede versie van de verslaglegging af is.

Welke ICT-gereedschappen worden er momenteel op Nederlandse scholen in de exacte vakken ingezet? Bij alle practica in natuurkunde, scheikunde, biologie, ANW en techniek wordt vrijwel overal de aan de Universiteit van Amsterdam ontwikkelde software- en hardwareomgeving Coach gebruikt [6, 14]. Deze leeromgeving biedt ook diverse wiskundige faciliteiten, om de simpele reden dat experiment en waarneming meestal gepaard gaan met wiskundige presentatie en analyse van gegevens. Maar hiaten in faciliteiten voor wiskunde moeten nog opgevuld worden voordat er van een bètabreed inzetbare omgeving gesproken kan worden. Voor wiskundeonderwijs is de situatie compleet anders en bestaat er niet een met Coach te vergelijken softwarematige werkomgeving, die bij alle wiskundige activiteiten en op diverse leerlingenniveaus inzetbaar is. Integendeel, er is een versnipperd gebruik van ICT-gereeedschappen: naast de landelijk ingevoerde grafische rekenmachine wemelt het van domeinspecifieke computerprogramma's. We noemen er een paar: Cabri Geometry, Geometer's Sketchpad en Cinderella voor meetkunde, VU-dynamo, Dynasis en Stella voor dynamische systemen, VU-stat voor statistiek, symbolische rekenmachines en Derive voor het werken met formules, StudyWorks en TI-InterActive voor wiskundige werkbladen en Excel voor werken met spreadsheets. Hoe prachtig en weldoordacht al deze programma's ook zijn, ze vormen in feite een grabbelton van onderling slecht samenwerkende pakketten.

Wat echter het meest vervelend is: er worden momenteel bij wiskunde heel andere ICT-gereedschappen gebruikt dan bij de overige exacte vakken. Dit probleem komt meer naar de oppervlakte nu met de invoering van de profielen, met name van de profielen Natuur & Gezondheid en Natuur & Techniek,

aan de wenselijkheid om bètavakken inhoudelijk beter op elkaar af te stemmen en de systeemscheiding tussen de vakken op te heffen gevolg gegeven kan worden. Deze samenhang zou je dan ook graag terug zien bij de hulpmiddelen voor leerlingen en docenten. De behoefte aan een gemeenschappelijke ICT-leeromgeving voor alle bètavakken neemt toe. Tijdsbesparing en grotere effectiviteit mag verwacht worden wanneer leerlingen en docenten met slechts één omgeving werken, maar ook aspecten als kostenbesparing spelen een rol.

Op basis van groeiende inzichten in ICT-gebruik bij wiskunde en van ervaringen opgedaan met Coach bij de andere exacte vakken, wordt in de komende jaren aan het AMSTEL Instituut van de UvA een bètabreed inzetbare leeromgeving gemaakt, kortweg met β-tool aangeduid. Maar waarom wachten op de eerste versie van de β-tool? Misschien is de huidige versie van Coach, die op veel scholen aanwezig is, al inzetbaar. In [11] kunt u een overzicht vinden van de huidige faciliteiten van Coach en lezen hoe dit pakket met al zijn beperkingen en mogelijkheden inderdaad binnen wiskundeonderwijs bruikbaar is.

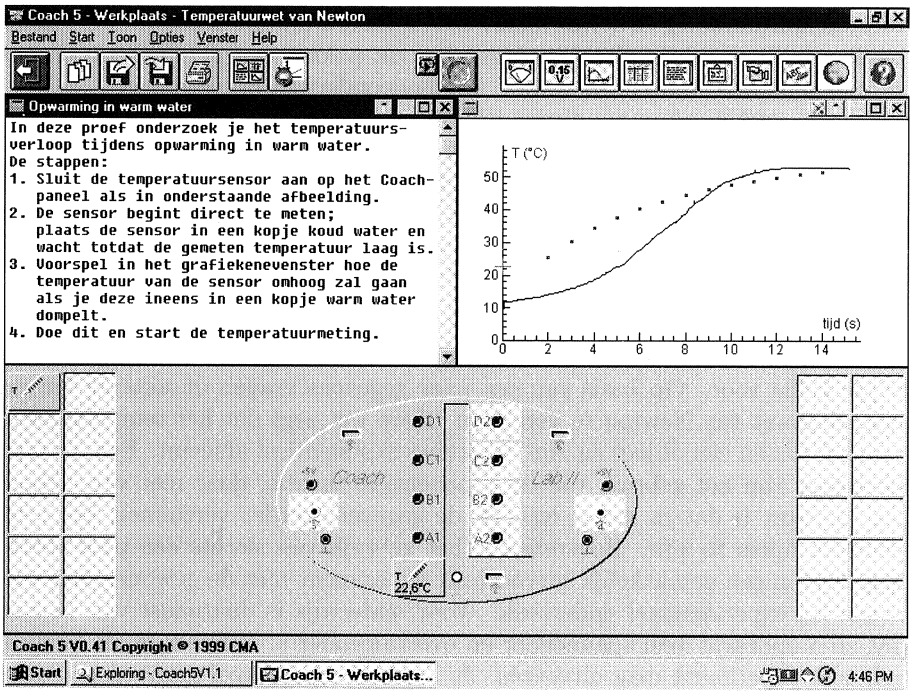
In dit artikel zullen we dieper ingaan op de mogelijkheden van Coach bij praktische opdrachten en het profielwerkstuk. Onderwerpen zijn: temperatuursverandering, vrije worp bij basketbal, kettinglijn, tumorgroei en lengtegroei bij mensen. Hiermee hopen we u enig zicht te geven op enkele ingrediënten van de β-tool in spe en op de eisen die wiskunde aan dit pakket stelt.

2. Herontdek de temperatuurwet van Newton

In elk schoolboek voor wiskunde komt wel het voorbeeld van afkoeling van een kopje koffie voor. Op basis van een paar gegevens vragen de auteurs om de afkoelingswet van Newton te verifiëren. Deze wet zegt dat het temperatuursverschil tussen een object en zijn omgeving exponentieel afneemt. Een aardige illustratie van het gebruik van de exponentiële functie, daar niet van. Maar wat irriteert is dat in menig lesboek de gegevens eerder verzonnen dan echt gemeten lijken te zijn. Bovendien is het gebruik van slechts een paar meetgegevens in een gemakkelijk uitvoerbaar experiment niet de gewone gang van zaken bij experimenteel onderzoek. Voor onderwijs is daarnaast van belang dat zo'n opdracht veel spannender en overtuigender is wanneer je als leerling zelf zo'n experiment mag uitvoeren. Je kunt dan gelijk veel meer vragen over temperatuursverandering onderzoeken: hoe hangt het temperatuursverloop af van begin- en eindtemperatuur? Maakt het verschil of je een object laat afkoelen in water of lucht? Speelt de vorm of het materiaal van het object een rol? Verloopt opwarmen van een object net zo? Bij de beantwoording van deze vragen gaan natuurkunde en wiskunde hand in hand.

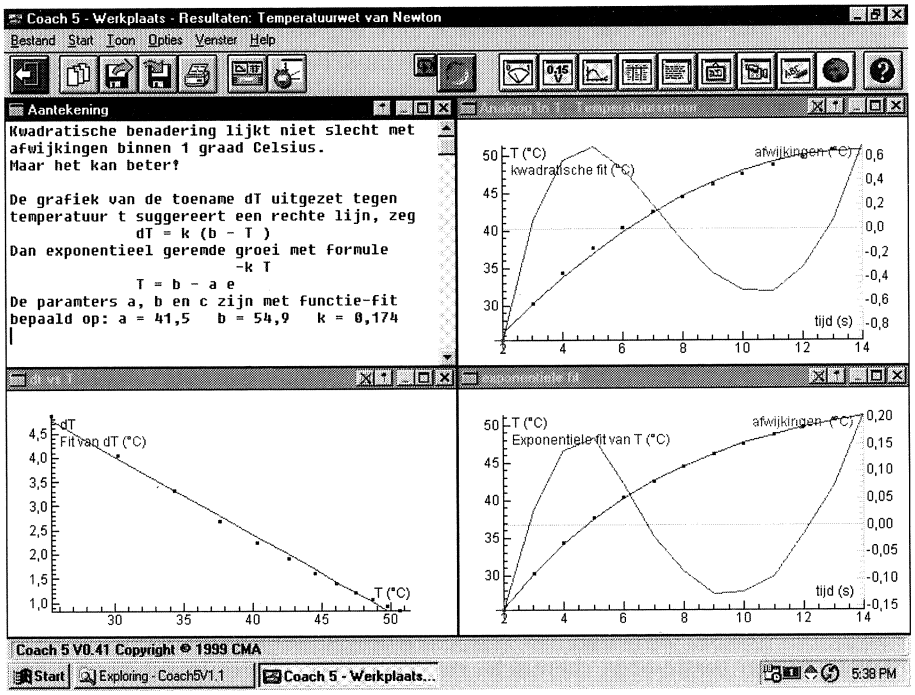
Om een indruk te geven van hoe je met Coach in het echt gegevens verzamelt zullen we een eenvoudig experiment beschrijven om de opwarming van een object te onderzoeken. De proef is als volgt: koel een temperatuursensor eerst in een kopje koud water af, dompel deze dan ineens in een bakje warm water en laat de computer elke seconde de temperatuur van de sensor registreren. De beschrijving van de proef en de meetopstelling kan in begeleidend lesmateriaal staan, maar kan ook binnen het computerpakket getoond worden, zoals in onderstaande schermafbeelding (figuur 1) van Coach in werking te zien is.

Onderaan in figuur 1 staat de afbeelding van het Coachlab-paneel waarop de temperatuursensor is aangesloten. Linksboven staan de opdrachten en rechtsboven is een diagram waarin punten getekend zijn die de gemeten temperatuur op zeker moment aangeven. Desgewenst kan de docent deze schermopbouw vooraf instellen: zo kan de software zelf herkennen welke sensor op welke plaats op het meetpaneel wordt aangesloten; de meetinstellingen kunnen van tevoren vastgelegd worden; de grootheden temperatuur T en *tijd* kunnen al in een diagramvenster klaargezet zijn. Voorafgaand aan de meting kan gevraagd worden over mogelijke uitkomsten na te denken en in het diagramvenster te voorspellen hoe de opwarming verloopt. De vorm van de getekende grafiek doet vermoeden dat hier aan een logistische kromme gedacht is.



FIGUUR 1. Temperatuurmeting met Coach.

Zodra de meting klaar is kan met de verwerking van de meetgegevens en met de speurtocht naar een wiskundig model van opwarming begonnen worden. De voorspelde grafiek was fout en wordt snel gewist. Met de menu-optie `functie-fit` kun je eenvoudig nagaan of een bekend regressiemodel goed past bij de gemeten grafiek. Het venster rechtsboven in figuur 2 laat zien dat de kleinste kwadratenmethode met een bergparabool al een regressiekromme oplevert die maximaal 1°C afwijkt. Merk op dat de afwijkingen in het diagramvenster uitgezet zijn met een eigen verticale coördinaat-as om de grafiek beter in beeld te krijgen.



FIGUUR 2. Wiskundige analyse van opwarming.

Rest de vraag hoe goed deze regressiekromme is. Op de eerste plaats is de som van de kwadraten van de afwijkingen een aanwijzing voor de nauwkeurigheid: hoe kleiner deze som, hoe beter. Ook is het verstandig om het voorspellende karakter van de kromme onder de loep nemen. Dan voldoet de bergparabool al veel minder: na 30 seconden wordt een temperatuur van 8°C voorspeld, wat minder is dan de berekende begintemperatuur van 17,6°C. Het diagramvenster rechtsonder in figuur 2 illustreert dat een exponentieel regressiemodel een geschiktere formule oplevert, met afwijkingen binnen 0,2°C . Voor dit ene experiment is deze formule een mooi resultaat, maar als wiskundig model van opwarming is het weinig bevredigend. Dan wil je ook weten wat de wiskundige en fysische grondslag voor de formule is, zodat je niet alleen iets weet over de opwarming van deze ene sensor, maar over de temperatuursverandering in een algemener geval.

In dit experiment ligt het voor de hand om, behalve naar de temperatuur, ook naar de temperatuurtoename tijdens de proef te kijken. In het venster linksonder in figuur 2 is de grootheid dT voor de toename in temperatuur per seconde ingevoerd en is de grafiek getekend van dT als functie van temperatuur T: een rechte lijn verschijnt op het scherm. Kennelijk is er een lineair verband, zeg van de vorm $dT = k(b - T)$. Door aflezen in de grafiek of door het uitvoeren van een lineair regressiemodel via de menu-optie `functie-fit`

kunnen de waarden van de parameters k en b bepaald worden. Deze waarden komen later goed van pas wanneer je met de modelleeromgeving van Coach een simulatie van het opwarmingsproces wilt uitvoeren.

Een docent mag van een leerling verlangen dat hij of zij over de betekenis van parameters in een model nadenkt: het moet bijvoorbeeld duidelijk zijn dat b de maximale temperatuur bij opwarming voorstelt. Verder is er natuurlijk de hoop en verwachting dat een leerling door naar het verband tussen temperatuurtoename en temperatuur te kijken zelf op een idee komt voor een geschikt wiskundig model van het natuurkundige verschijnsel. Het werken met zelf gemeten data vergroot het gevoel zelf een ontdekking gedaan te hebben. Daarnaast zal menig scholier het doen van experimenten een aantrekkelijke werkvorm vinden. Een heleboel pluspunten van werken met reële data.

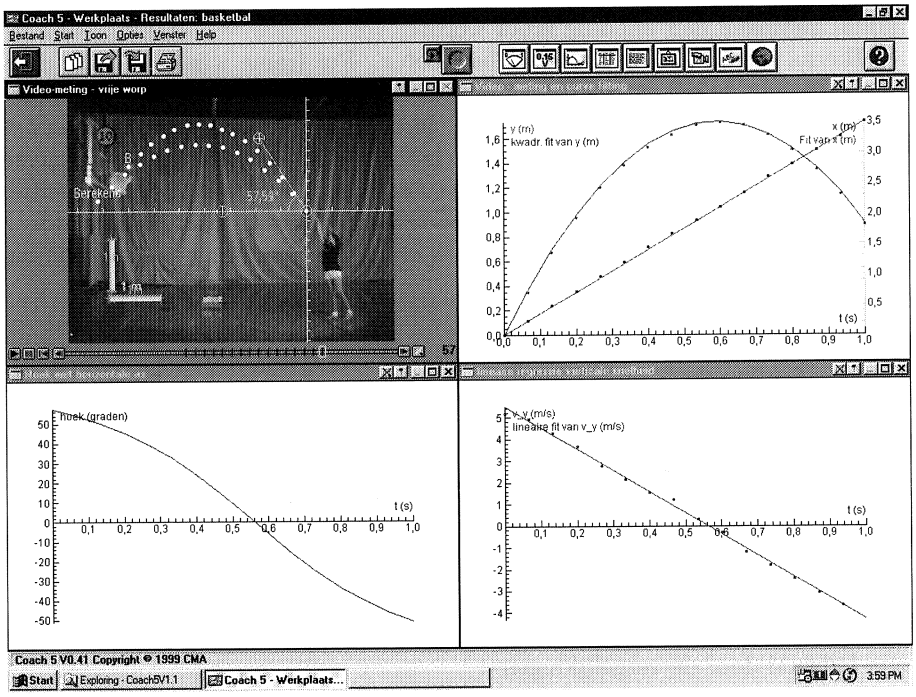
In deze ene leerlingactiviteit passeren al veel onderdelen uit de wiskunde de revue: toenamendiagram, differentiequotiënt, lineair en exponentieel verband, exponentiële functie, kleinste kwadraten methode, etc. Maar het belangrijkste is misschien wel dat een leerling hier ervaart dat dit geen gekunstelde, maar concrete wiskundige begrippen zijn die hun diensten bewijzen in een wiskundig model van een natuurkundig verschijnsel. Bovendien kan dit model toegepast worden op verschijnselen van verschillende aard: laat leerlingen maar experimenteren met de ontlading van een condensator of de concentratie van een reagens in een eerste-orde chemische reactie volgen. Later in dit artikel zullen we zien hoe dit model ook onderdeel vormt van een wiskundig model voor lengtegroei van jongens en meisjes.

3. Analyseer een vrije worp bij basketbal

Een onderwerp dat dicht bij het gewone schoolleven staat is basketbal in de gymnastiekles. Een interessante onderzoeksvraag hierover is hoe je het beste een vrije worp kunt nemen: bovenhands of onderhands? hoog of laag? Om deze vraag te kunnen beantwoorden heb je wel gegevens nodig en deze kunnen jammer genoeg niet of slechts met veel moeite via sensoren gemeten worden. Wat veel beter werkt is het opnemen van een aantal vrije worpen met een digitale videocamera om daarna gegevens uit de videoclips te halen. In videometing klik je met de muis in de videoclip op punten waarvan je de coördinaten wilt weten en meet je afstanden en hoeken. Deze gegevens kun je vervolgens met dezelfde middelen als bij een echt experiment verwerken en analyseren. Voordelen van videometing zijn:

- Je hoeft zelf geen proefopstelling op te bouwen.
- Processen die zich minder goed lenen voor directe metingen kun je toch bestuderen.
- Je hoeft niet van tevoren tot in elk detail te bedenken wat precies gemeten gaat worden.
- Je kunt metingen gemakkelijk en snel doen, achteraf nog eens verifiëren en indien nodig corrigeren.

Figuur 3 is een schermafbeelding van een videometing en analyse van een bovenhandse vrije worp van een scholiere in de sportzaal. Hierin gaat de aandacht uit naar de baan die de bal volgt, maar deze video is even goed te gebruiken om de beweging van de scholiere tijdens het gooien van de bal te analyseren.



FIGUUR 3. Meting en analyse van een vrije worp bij basketbal.

Hoe voer je bovenstaande videometing en analyse uit? Als kant en klare films niet voorhanden zijn moet je eerst zelf opnamen maken en deze in geschikte digitale videoclips omzetten. Heb je eenmaal een videoclip in een Coach-activiteit geladen, dan kun je eerst nog eens het filmpje afspelen om een idee te krijgen van wat er gebeurt. Zo kun je ook beter beslissen wat te gaan onderzoeken en hoe.

Stel dat je de baan van de geworpen bal wilt bepalen, dan moet je de posities van de bal op verschillende tijdstippen meten. Hiervoor is nodig dat je horizontale en verticale lengtematen instelt en een coördinatenstelsel kiest. Op het getoonde filmpje zijn horizontale en verticale maatstokken aangebracht, maar anders had je de lengtematen nog uit de voorgeschreven hoogte van de basketring en de voorgeschreven afstand van de vrijeworp-lijn kunnen halen. Om de wiskunde gemakkelijker te maken kun je de oorsprong van het coördinatenstelsel het beste neerzetten op het punt waar het meisje de bal los laat. Wanneer je de positieve x- en $y-$as respectievelijk van rechts naar links en van onder naar boven laat open, komen deze variabelen overeen met de

grootheden 'horizontale afstand' en 'hoogte', gezien vanuit de werpster.

Je beëindigt het voorwerk met het kiezen van de individuele beeldjes (frames) waarin je metingen wilt doen. In dit voorbeeld ligt voor de hand om alleen frames te kiezen in de periode tussen het weggooien van de bal en het belanden van de bal in het netje. Ook kun je met regelmaat frames overslaan zodat het aantal metingen niet onnodig groot wordt.

Hierna kun je aan de slag met het meten. Per frame wijs je met de muis de positie van de bal aan en Coach registreert de (x, y) coördinaten samen met het bijpassende tijdstip in de videoclip. Je kunt de gemeten punten ook in het videovenster tonen: in bovenstaande schermafbeelding markeren de bovenste punten in de videoclip de gemeten positie van de geworpen bal. Behalve posities kun je in Coach ook afstanden en hoeken meten: in bovenstaande schermafdruk (figuur 3) is met de gradenboog een hoek van 57,6° gemeten waaronder de bal wordt weggegooid.

Tijdens of na het meten kun je de gegevens in een tabel of diagram op het scherm zetten: zie de punten in het diagramvenster rechtsboven in figuur 3. Deze grafiek brengt je op het idee van een bergparabool voor de hoogte van de bal. Met lineaire regressie vind je onmiddellijk de beste formule. En dit klopt heel goed: niet alleen wiskundig gezien, maar ook als je de mechanicawetten van Newton toepast. Als luchtweerstand namelijk verwaarloosd mag worden, dan werkt er in horizontale richting geen kracht op de bal en heeft de bal in deze richting een eenparige, rechtlijnige beweging. In verticale richting werkt alleen de zwaartekracht op de bal, met als gevolg een eenparig versnelde, rechtlijnige beweging. In formuletaal: de horizontale afstand x en hoogte y van een bal, die op tijdstip $t = 0$ met beginsnelheid v onder een hoek α geworpen wordt, zijn bepaald door de vergelijkingen

$$x = (v \cos \alpha)\, t,$$
$$y = (v \sin \alpha)\, t - \frac{1}{2} g\, t^2.$$

Hierin is g de valversnelling. De natuurkunde verklaart dus waarom je een rechte lijn voor de horizontale positie van de bal uitgezet tegen de tijd krijgt en waarom de verticale positie een bergparabool is. Uit de tweede vergelijking volgt eenvoudig de verticale component v_y van de snelheid van de bal:

$$v_y = v \sin \alpha - gt.$$

Rechtsonder in figuur 3 zijn de resultaten van numeriek differentiëren van y in een diagramvenster getekend: de berekende punten liggen inderdaad op een bijna rechte lijn. Door lineaire regressie, maar nog eenvoudiger door de helling van de lijn met het hiervoor bestemde gereedschap in Coach te bepalen, vind je een waarde voor de valversnelling g en beginsnelheid v. De waarden $g = 9{,}8\,\mathrm{m/s^2}$ en $v = 6{,}5\,\mathrm{m/s}$ stemmen mooi overeen met literatuurgegevens.

Door eliminatie van t in de vergelijkingen van x en y krijg je het volgende kwadratische verband tussen horizontale afstand en hoogte:

$$y = x \tan \alpha - \frac{g}{2v^2 \cos^2 \alpha} x^2 .$$

Differentiëren levert de helling in elk punt van de baan op:

$$\frac{dy}{dx} = \tan\alpha - \frac{g}{v^2\cos^2\alpha}x\,.$$

Voor de hoek van inval α_i bij de basketring op afstand L geldt dus:

$$\tan\alpha_i = \tan\alpha - \frac{gL}{v^2\cos^2\alpha}.$$

Meet L op in de video en je kunt de hoek van inval bij de basketring berekenen: deze blijkt in deze worp gelijk te zijn aan $-53°$. Dit kun je weer onmiddellijk controleren in de videoclip of in de grafiek van de hoek die de snelheidsvector van de bal tijdens de vlucht maakt met de horizontale richting.

Met de formules kun je ook de baan van de bal uitrekenen als de scholiere de bal met dezelfde snelheid als in het filmpje weggooit, maar met een beginhoek van $50°$ i.p.v. $57,6°$. Berekende punten van deze baan van de bal kun je in de videoclip laten tonen: de onderste gemarkeerde punten in het videovenster van figuur 3 geven de berekende baan weer. Ook in dit geval lijkt de bal door de basketring te gaan. Met andere woorden, bij de gegeven startsnelheid is er best wel een ruime marge voor de hoek waaronder de bal weggegooid kan worden om toch nog een punt op te leveren. Omgekeerd zou je ook kunnen onderzoeken of bij gegeven hoek van weggooien van de bal de snelheid voldoende mag variëren om toch nog een punt voor je team te scoren. Met de wiskundige formules kun je ook de starthoek bepalen waarvoor de benodigde snelheid van weggooien van de bal minimaal is en de controle over de worp dientengevolge maximaal is.

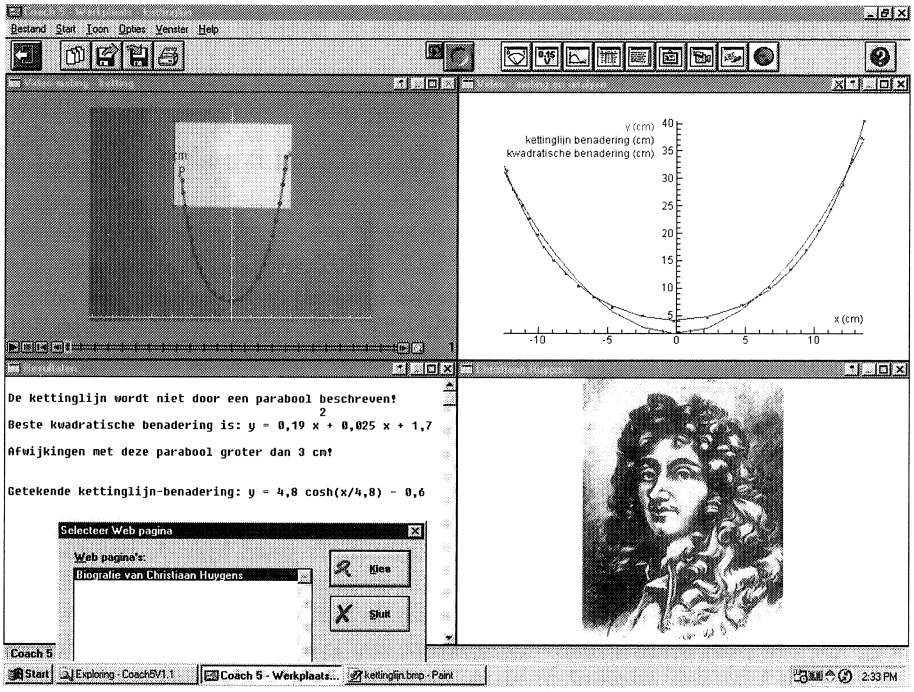
Wat in het tweede voorbeeld opnieuw opvalt is dat de wiskundige formules, functies en grafieken geen abstracte noties zijn, maar juist concrete en direct inzetbare begrippen zijn. Je kunt formules en grafieken direct op hun waarde schatten door terugkoppeling naar de echte situatie. De wiskunde gaat ergens over en is vormgegeven in uitdagende opdrachten.

Wie geïnteresseerd is in de combinatie van sport, wiskunde en natuurkunde en wie inspiratie op wil doen voor praktische opdrachten of profielwerkstuk in deze richting raden we de boeken "Mathematics in Sport" [15] en "The Physics of Sports" [1] aan.

4. Vind de formule van de kettinglijn

Bij videometing denk je in eerste instantie aan bewegende beelden. Maar ook aan stilstaande beelden kun je meten en zinvolle wiskunde doen. Hang bijvoorbeeld een halsketting aan de uiteinden op, maak er een digitale foto van en laad deze in een Coach-activiteit als ware het een videoclip. Nu kun je posities van punten op de ketting meten. In onderstaande schermafbeelding (figuur 4) wordt de vorm van een hangende halsketting opgemeten (met dank aan Olga Zika).

Als je in Coach handmatig een parabool als benadering voor de kettinglijn uitprobeert, dan kom je er snel achter dat de gemeten punten niet goed op zo'n kromme passen. In het diagramvenster rechtsboven in figuur 4 is de beste



FIGUUR 4. Meten en rekenen aan de kettinglijn.

parabool met automatische regressie bepaald: een afwijking van meer dan 3 cm! De kettinglijn is dus geen parabool; iets wat Christiaan Huygens al in 1646 op ingenieuze wijze aantoonde. Uit eerbetoon hebben we een portret van Internet opgehaald en in ons werkstuk geplaatst. De hyperlink naar de website waar het plaatje vandaan komt en waar de biografie van Huygens te lezen is kan een docent van tevoren klaarzetten, zodat leerlingen snel aan de slag kunnen gaan en hun werk niet hoeft te verzanden in speurtochten op Internet. Tussen twee haakjes, het maken van lesmateriaal of het verslag doen van een experiment met Coach is technisch gezien zo eenvoudig, dat men zich volledig op de inhoud kan concentreren. Tekstuitleg, een videoclip, een grafiek en elk ander onderdeel is in een venster te plaatsen door eerst een bijpassend icoon in de taakbalk aan te klikken, dan een dialoogvenster in te vullen (bijvoorbeeld door de naam van een hyperlink of een bestand aan te klikken), om tenslotte het icoon naar een venster te slepen en los te laten.

We keren terug naar de wiskunde van de kettinglijn. Jan van de Craats heeft in [5] uitvoerig beschreven hoe de ketting wel hangt en hoe dit met leerlingen uit te zoeken is. Laat $y(x)$ een functie zijn waarvan de grafiek een 'idealisatie' van de hangende ketting is. Deze functie voldoet aan de volgende

differentiaalvergelijking:

$$\frac{d^2y}{dx^2} = \frac{1}{k}\sqrt{1+(\frac{dy}{dx})^2},$$

voor zekere $k > 0$. De algemene oplossing is:

$$y(x) = k\cosh(x/k + b) + c\,,$$

voor zekere constantes b en c. Als het coördinatenstelsel zodanig gekozen wordt dat het laagste punt van de ketting bij $x = 0$ past, dan is $b = 0$ en is de oplossing in termen van de exponentiële functie te schrijven als

$$y(x) = k\frac{e^{x/k} + e^{-x/k}}{2} + c\,.$$

Met een trial-and-error methode kun je geschikte waarden voor k en c vinden. In het diagramvenster rechtsboven in figuur 4 zijn de gemeten punten te zien, samen met de grafiek van een geschikte 'wiskundige' kettinglijn. Een mooi resultaat, niet waar?

Deze wiskundige benadering van de kettinglijn kan weer gebruikt worden om vragen als "Wat is de lengte van de ketting tussen twee punten?" te beantwoorden. Laat s de functie zijn waarvoor $s(x)$ de lengte van de 'wiskundige' ketting tussen het ophangpunt aan de linkerkant en het punt $(x, y(x))$ is. Dan geldt:

$$s(x) = \int_{x_l}^{x} \sqrt{1+(\frac{dy}{d\xi})^2}\, d\xi\,,$$

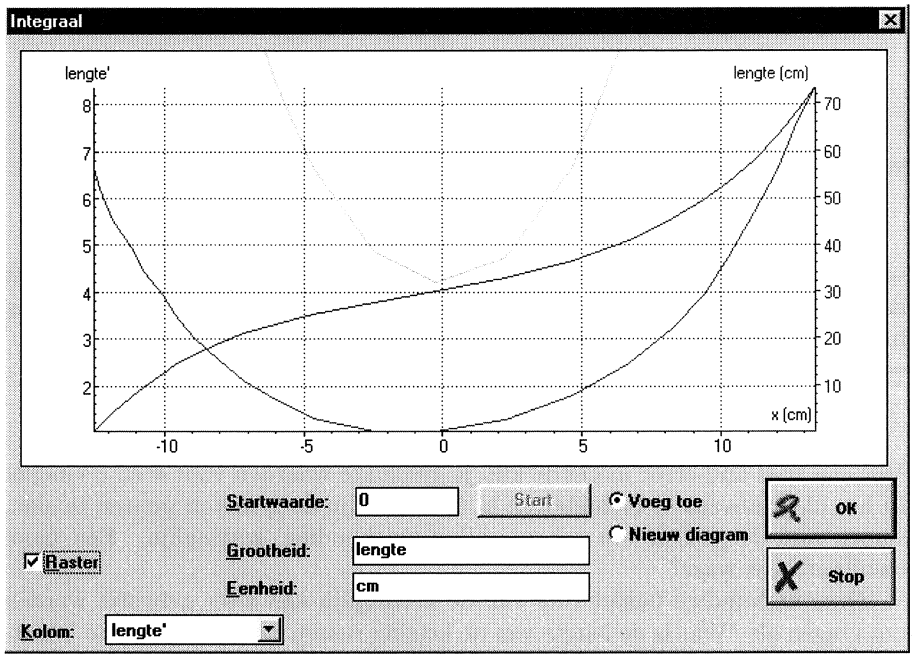
waarbij x_l de x-positie van het ophangpunt aan de linkerkant is.

Coach kan numerieke differentiëren en integreren: in figuur 5 staat de grafiek van $s(x)$ die het gereedschap voor integratie oplevert. De totale lengte van de ketting zou volgens deze grafiek 73,5 cm zijn en dit blijkt bij meting van de echte ketting aardig te kloppen. In Coach kun je ook de oppervlakte onder een grafiek op een willekeurig gekozen segment bepalen. Hiermee kun je gemakkelijk Coach de lengte van de kettinglijn tussen elk tweetal punten laten uitrekenen.

Aangespoord door dit succes kun je andere situaties met kettingen gaan onderzoeken: Hoe hangt een ketting waar in het midden een hangertje aan hangt? Hoe is de vorm van een ketting die gedeeltelijk in het water hangt? Als je op een hangende ketting weer twee ophangpunten kiest waaraan een tweede ketting gehangen wordt, wat kun je dan zeggen over de vorm van de twee kettingen? Allemaal onderzoeksvragen waar meting aan digitaal beeldmateriaal en het beschikbaar hebben van een groot arsenaal van wiskundige faciliteiten helpen om antwoorden te vinden.

5. Maak een wiskundig model voor tumorgroei

Wiskundige modellen bieden een uitkomst als metingen om praktische, technische of ethische redenen niet uitvoerbaar zijn. Het lukt niet altijd om een



FIGUUR 5. Lengte van kettinglijn tussen twee punten.

model te onderbouwen: onderstaand model van Gompertz wordt gebruikt om de groei van sommige tumoren wiskundig te beschrijven, maar een afdoende biologische of medische verklaring waarom en onder welke voorwaarden het wiskundig model werkt is er nog niet (zie [10]).

Er bestaan grofweg twee manieren waarop wiskundig modellen tot stand komen: empirische afleiding via statistische analyse, bijvoorbeeld regressie, en een meer theoretische benadering, gevolgd door een computersimulatie. We concentreren ons hier op de tweede manier van werken. Het proces van modelleren kent dan vier fasen:

1. Analyse van het systeem en bepaling van de basiscomponenten nodig voor het model.
2. Definitie van de belangrijkste variabelen om het systeem te beschrijven.
3. Afleiding van de wiskundige vergelijkingen.
4. Computerimplementatie, schatting van parameterwaarden en simulatie.

We kijken hier alleen naar de laatste fase. Vooraf noemen we nog enkele voordelen van een wiskundig model en van een computersimulatie:

– Een conceptueel model van een verschijnsel wordt op deze manier gekwantificeerd, hetgeen weer kan leiden tot een dieper inzicht in het gemodelleerde fenomeen.

- Hypothesen kunnen op wetenschappelijke wijze onderzocht worden zonder daarvoor tijdrovende metingen te doen.
- Simulaties nodigen als het ware uit tot het stellen van 'wat als'-vragen.
- Eenzelfde wiskundig model kan in bijna identieke vorm in meerdere toepassingsgebieden toegepast worden.

Om het laatste voordeel kracht bij te zetten merken we op dat het groeimodel van Gompertz ook toegepast wordt bij bestudering van bladgroei van planten [4] en bij studies naar gewichtstoename van varkens [9].

Laten we beginnen met het groeimodel van Gompertz wiskundig te beschrijven. Exponentiële groei komt in elk wiskunde-schoolboek voor: een grootheid, zeg g, groeit of neemt exponentieel af als de snelheid waarmee de waarde van de grootheid verandert evenredig is met de waarde op dat moment. Anders geformuleerd: de relatieve groeisnelheid is constant. In de taal van differentiaalvergelijkingen:

$$\frac{dg}{dt} = c\, g(t)\,,$$

voor zekere constante c. De algemene oplossing is:

$$g(t) = g(0)e^{ct}$$

Voegen we aan het rechterlid een kwadratische term toe, dan krijgen we het Verhulst model voor een proces van geremde groei:

$$\frac{dg}{dt} = c\, g(t)\,\big(k - g(t)\big)\,,$$

voor zekere positieve constanten c en k. De algemene oplossing van wat ook als logistische groei bekend staat is

$$g(t) = \frac{kg(0)}{g(0) + (k - g(0))e^{-ckt}}\,,$$

ofwel

$$g(t) = \frac{k}{1 + e^{d-ckt}},$$

voor zekere positieve constante d. In dit model nadert g tot k als $t \longrightarrow \infty$.

Er bestaat een andere aanpassing van de differentiaalvergelijking voor exponentiële groei die geremde groei oplevert: neem geen constante relatieve groeisnelheid, maar laat deze volgens een exponentieel proces in de tijd afnemen. In formuletaal:

$$\frac{dc}{dt} = -a\, c(t)\,,$$
$$\frac{dg}{dt} = c(t)\, g(t)\,,$$

voor zekere positieve constante a. Dit heet het groeimodel van Gompertz. Het stelsel van differentiaalvergelijkingen voor g en c is te herschrijven tot één vergelijking:

$$\frac{dg}{dt} = a\, g(t) \left(b - \ln\big(g(t)\big)\right),$$

voor zekere positieve constante b. In deze vorm komen gelijkenis en verschil met het Verhulst model goed tot uitdrukking. Ook het model van Gompertz voor geremde groei kan exact opgelost worden:

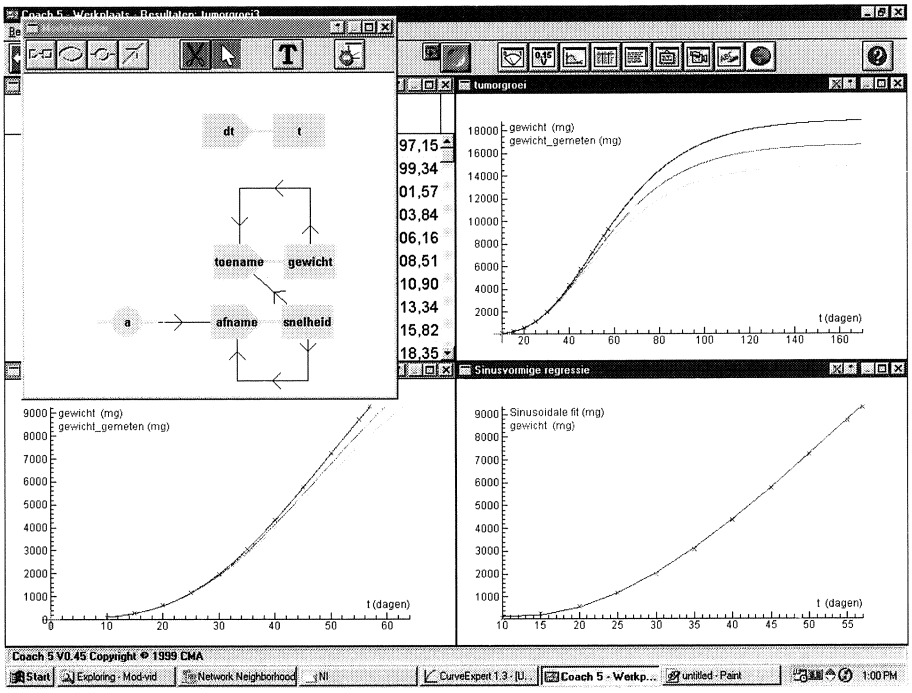
$$g(t) = \exp\left(b - e^{k-at}\right),$$

voor zekere positieve constante k. Voor groei volgens het Gompertz model geldt dat g nadert tot e^b als $t \longrightarrow \infty$.

Een leerling hoeft dan wel niet zelf de exacte oplossingen van bovenstaande differentiaalvergelijkingen te kunnen bepalen, controle op correctheid van de gegeven oplossingen en interpretatie van de parameters hoor wel tot de eisen die te stellen zijn. En laten we ook eerlijk zijn: de meeste differentiaalvergelijkingen die je in praktijk tegen komt zijn niet analytisch oplosbaar, maar kunnen slechts numeriek aangepakt worden. Als zo'n numerieke aanpak in de wiskundige software beschikbaar is, dan breidt dit de mogelijkheden voor wiskundig onderzoek door scholieren enorm uit. De techniek van het oplossen van differentiaalvergelijkingen hoeven ze niet eerst meester te zijn voordat ze met interessante toepassingen kunnen beginnen.

We passen het model van Gompertz toe op de groei van de C3H tumor. De groeigegevens ontlenen we aan [2] en referenties hierin. Maar voor we dit gaan doen, gebruiken we de gegevens om het verschil tussen empirisch en theoretisch modelleren toe te lichten. In het diagramvenster rechtsonder in figuur 6 is de grafiek van de beste sinusvormige benadering (bepaald met de `functie-fit` faciliteit van Coach). Het resultaat is mooi, maar elke kankerpatiënt zou wensen dat het in overeenstemming met de werkelijkheid is. Wat ontbreekt aan deze regressiekromme is de motivatie voor de keuze van formule. We hadden net zo goed een 4e-graads veelterm kunnen nemen en een nog mooiere overeenstemming gekregen hebben. Bij regressie geven alleen de som van kwadraten van afwijkingen, de correlatiecoefficiënt en de spreiding van residuen een aanwijzing voor toepasbaarheid van een gekozen model. Bij theoretisch modelleren is juist de onderbouwing van een gekozen model startpunt van het verdere werk.

Het modelleren op de computer kent twee fasen in de implementatie van het wiskundig model: het specificeren van het wiskundige model en het onderzoeken van het model. Om met het eerste te beginnen: er bestaat een grafisch interface om het model kwalitatief te beschrijven (linksboven in figuur 6). Hierin geef je op welke grootheden in het wiskundige model een rol spelen (met onderscheid tussen parameters en toestandsvariabelen), hoe ze van elkaar afhangen, welke formules voor grootheden precies gebruikt worden en welke waarden de constanten hebben.



FIGUUR 6. Modelleeromgeving ingezet bij het Gompertz model.

Het grafische model wordt automatisch vertaald naar een stel vergelijkingen die gebruikt worden in een computersimulatie van het model. Ook kun je vergelijkingen, startwaarden van toestandsgrootheden en waarden van constanten via een vergelijkingsgeoriënteerd interface intoetsen. Het ingevoerde model kan vervolgens worden doorgerekend. Resultaten kun je in een tabel of diagram weergeven en direct vergelijken met de echte data. In bovenstaande schermafbeelding zijn grafieken getekend voor verschillende waarden van a in het model van Gompertz, toegepast op de tumorgroei. Op deze manier kan het effect van wijzigen van een parameterwaarde eenvoudig onderzocht worden en zijn geschikte waarden van parameters proefondervindelijk op te sporen.

Wiskundige modellen maak je ook om de invloed van wijzigingen in condities op het verloop van een proces te bestuderen. Een voorbeeld: stel dat met chemotherapie begonnen wordt of een deel van het gezwel verwijderd wordt, wat is dan het effect op de groei van de tumor? Zo'n verandering van condities kan ook een verandering of uitbreiding van het model tot gevolg hebben. Bijvoorbeeld, chemotherapie leidt tot de volgende verandering van het rechterlid van de differentiaalvergelijking van Gompertz:

$$\frac{dg}{dt} = a\,g(t)\left(b - \ln\big(g(t)\big)\right) - r\,g(t)\,C(t),$$

waarbij r een positieve constante is en $C(t)$ de concentratie van het chemotherapeuticum op tijdstip t is (zie [2]). Ook een dergelijke verandering in het wiskundige model is in Coach gemakkelijk aan te brengen en uit te werken.

6. MAAK EEN FORMULE VOOR DE GEMIDDELDE LENGTE VAN JONGENS

Lengtegroei van jongens en meisjes kom je op verschillende plaatsen in een wiskundeboek tegen. Bij de behandeling van

- veranderingsbegrippen zoals toenamendiagram, differentiequotiënt en helling.
- statistische begrippen als normale verdeling, gemiddelde, mediaan en percentielen.
- discrete en dynamische modellen van groei.

Bij nadere beschouwing van de vragen en opdrachten krom je echter regelmatig je tenen. Zo kun je een vraagstuk tegenkomen waarin een 15-jarige jongen van 175 cm lengte klein van stuk genoemd wordt, terwijl dit volgens de jongste Nederlandse groeicijfers boven het landelijke gemiddelde ligt. Een andere tekst voert een jongen ten tonele die op twaalfjarige leeftijd 115 cm lang is en een eindlengte van 191 cm bereikt. Kennelijk heeft deze jongen stiekem een behandeling met groeihormonen ondergaan! Immers, op jonge leeftijd was zijn lengte extreem klein (ruim 40 cm onder de gemiddelde lengte van leeftijdgenoten) en ver onder indicatie voor verwijzing naar een kinderarts, maar zijn eindlengte komt ruim boven het landelijke gemiddelde van 184 cm uit. Een schoolvoorbeeld van wat je vaker tegenkomt in wiskundeteksten: een verzonnen context, voorgeschoteld als reële context, maar alleen bedoeld als omlijsting of als 'ideale' illustratie van een wiskundig begrip.

Het veel gehoorde argument dat echte gegevens te weerbarstig zijn om succesvol mee te kunnen werken gaat kennelijk niet meer op in een praktische opdracht om de lengtegroei van jongens en meisjes te onderzoeken. Deze opdracht staat ook letterlijk in hedendaagse tekstboeken. Verwachten de opstellers van zo'n opdracht dat leerlingen dan ineens wel met weerbarstigheid van echte data overweg kunnen? Of gaan ze op voorhand uit van weinig wiskundige diepgang bij onderzoek door leerlingen? Dit laatste is amper voor te stellen.

In onderstaand voorbeeld hopen we aan te tonen dat een doordachte aanpak, waarin de docent de leerling tijdens de onderzoeksopdracht begeleidt, en voldoende inzet van ICT-hulpmiddelen het werken met echte gegevens heel goed mogelijk maakt en tot interessante resultaten kan leiden. Met name aan de rol van ICT in het doen van wiskunde zullen we aandacht besteden in onze studie van lengtegroei van Nederlandse jongens.

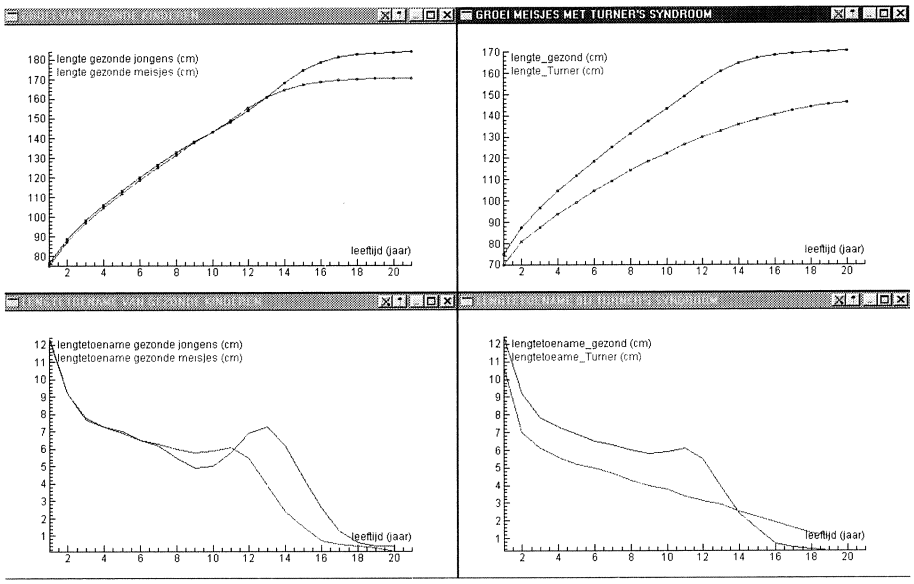
ICT wordt om te beginnen gezien als hulp bij het verkrijgen van informatie, bijvoorbeeld op Internet en CD-rom. Klopt, maar als het gaat om het vinden van recente Nederlandse groeicijfers, dan kom je van de koude kermis thuis. Deze cijfers zijn uit commerciële motieven niet on-line beschikbaar. Je komt dus als leerling niet eens toe aan de vraag of gevonden gegevens wel geschikt

zijn voor je onderzoek. En dit is nu juist de enige wiskundige waarde van zoeken naar informatie op Internet. Amerikaanse groeicijfers zijn daarentegen wel snel te vinden op Internet, maar deze zochten we hoogstens ter vergelijking met de Nederlandse situatie. De Nederlandse overheid zal het beschikbaar stellen van cijfermateriaal moeten stimuleren wil Internet de rol van informatiebron in Tweede Fase-onderwijs goed kunnen waarmaken.

Wat je overigens wel op Internet kunt vinden is een uitgebreid artikel uit de wetenschapsbijlage van NRC Handelsblad van 4 maart 2000 dat gaat over de meest recente landelijke groeistudie. Het is een mooie tekst als introductie op het onderwerp. Deze tekst en het wetenschappelijke artikel [8] dat de aanleiding tot het kranteartikel vormde en waarin precieze groeicijfers staan kunnen eigenlijk net zo goed in papiervorm aan de leerling gegeven worden. U kijkt er misschien van op, maar het is helemaal niet zo raar om het wetenschappelijke artikel in handen van scholieren te geven. Bij lezing blijkt het qua wiskunde helemaal niet zo ver weg te staan van wat er op school aan theorie behandeld wordt. Op deze manier krijgt een leerling ook een beeld van hedendaags onderzoek en verslaglegging.

Een tweede rol van ICT in exacte vakken, nl. verwerking van meetgegevens, komt in de volgende fase van het project aan bod: als je de groeicijfers eenmaal in handen hebt, dan moet je deze gemakkelijk in een computerprogramma kunnen invoeren voor grafische presentatie en verdere wiskundige behandeling. Importeren van gegevens in database- of spreadsheet-formaat is nuttig voor snelle verzameling van bestaande gegevens. Om verschillen in gemiddelde lengtegroei tussen jongens en meisjes te kunnen ontdekken moet je beide grafieken in een voldoende groot beeldscherm in beeld kunnen brengen samen met toenamendiagrammen.

De grafieken aan de linkerkant in figuur 7 hebben betrekking op de lengtegroei bij doorsnee-kinderen. Wat in de toenamendiagrammen onmiddellijk opvalt is de groeispurt in de pubertijd en dat deze bij jongens later optreedt dan bij meisjes. Maar het is ook leerzaam te kijken naar groeidiagrammen van kinderen met groeistoornissen. In figuur 7 staan aan de rechterkant de groeidiagrammen en toenamendiagrammen getekend van gezonde meisjes en van meisjes met het syndroom van Turner. Twee symptomen van Turner's syndroom zijn in de grafieken terug te vinden: trage groei en het ontbreken van de pubertaire groeispurt. Niet-wiskundige informatie over het syndroom van Turner en andere groeistoornissen kun je vinden op de website van de Belangenvereniging Van Kleine Mensen (`www.bvkm.nl`). Bij meisjes met het syndroom van Turner neemt na het vierde levensjaar de lengtetoename nagenoeg lineair in de tijd af. Anders gezegd, vanaf deze leeftijd kun je de gemiddelde lengte wiskundig beschrijven met een parabool. Hier merk je dat enige kennis van de relatie tussen de vorm van de afgeleide en de vorm van de oorspronkelijke functie van nut is. Bij lineaire regressie met een parabool krijg je afwijkingen van minder dan 1 mm tussen de regressiekromme en de echte gegevens voor meisjes met het syndroom van Turner. Wie durft nog te zeggen dat het werken met echte gegevens zo lastig is binnen wiskunde? Het lijkt meer een kwestie van geschikte toepassingen zoeken.



FIGUUR 7. Gemiddelde lengte en lengtetoename van jongens en meisjes.

In het vervolg kijken we alleen nog maar naar de gemiddelde lengtegroei van Nederlandse jongens. Het zelf bedenken van een wiskundig model gaat veel te ver en is waarschijnlijk tot mislukken gedoemd. Maar het uitproberen van een algemeen aanvaard en veelgebruikt model in de kindergeneeskunde hoort wel tot de mogelijkheden. We behandelen hier het zogenaamde KKP-model (zie [13]). In dit wiskundige model worden drie componenten gebruikt die elk met een groeifase geassocieerd zijn:

1. *Kleutertijd* (0-3 jaar): geremde groei, waarbij lengtetoename vanaf de geboorte exponentieel afneemt. De bijpassende formule is:

$$L_1 = a_1 - b_1 e^{-c_1 t} .$$

2. *Kindertijd*: lengtetoename neemt lineair af (denk terug aan de groei van meisjes met Turner's syndroom) en leidt tot de volgende formule:

$$L_2 = a_2 t^2 + b_2 t + c_2 .$$

3. *Pubertijd*: logistische groei voor de bijdrage van de pubertijdspurt aan de lengte, met als formule:

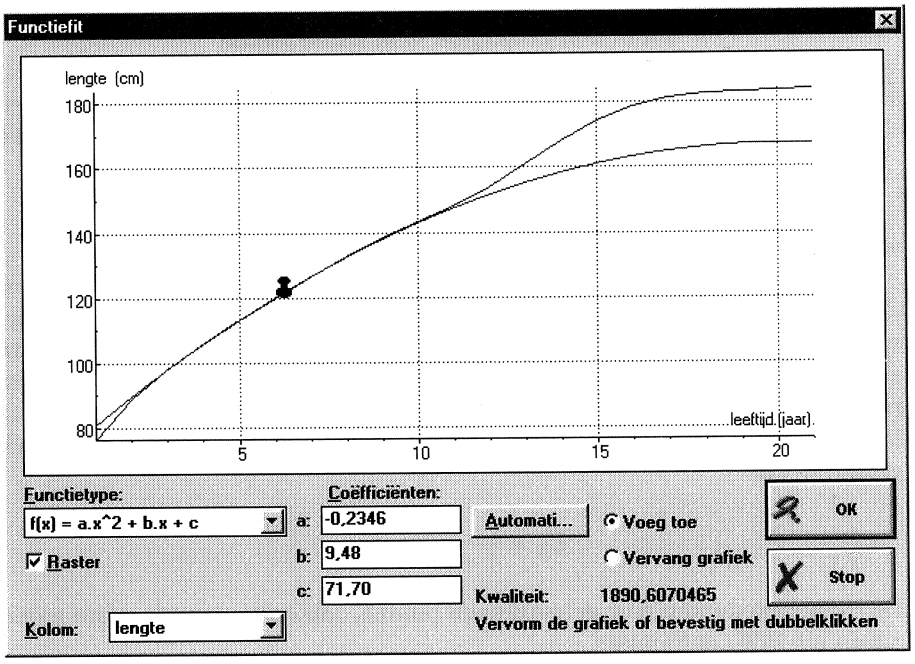
$$L_3 = \frac{a_3}{1 + e^{c_3 - b_3 t}} .$$

Hierbij zijn a_1, b_1, c_1, a_2, b_2, c_2, a_3, b_3 en c_3 parameters met positieve waarden, die op basis van de groeicijfers bepaald worden. De gemiddelde lengte L wordt op elke leeftijd gegeven door de som $L_1 + L_2 + L_3$.

Hoe gaan we met dit model aan de slag? Omdat de component voor de kindertijd het enige onderdeel met niet-geremde groei is en we toch een realistische formule willen vinden voor lengtegroei van 0 tot 21 jaar is het verstandig hiermee te beginnen. We zoeken een bergparabool die enerzijds de groei tussen het 3e en 10e levensjaar aardig beschrijft en anderzijds zijn maximum bereikt rondom de leeftijd van 20 jaar. De kleinste kwadratenmethode werkt in dit geval niet; we selecteren dan maar handmatig en op het oog een geschikte bergparabool. In onderstaande schermafdruk (figuur 8) zie je onze keuze van

$$lengte = -0{,}235\, leeftijd^2 + 9{,}5\, leeftijd + 71{,}7\,.$$

De punaise in de schermafdruk geeft aan dat we op die plaats de benadering vastgepind hebben. Door een ander punt op de parabool met de muis te verslepen is een andere tweedegraads kromme te maken. Als je de punaise door dubbelklikken losmaakt, kun je de parabool transleren. Op deze manier kun je in Coach op eenvoudige wijze met de muis in de hand een kromme van een voorgeschreven vorm construeren.

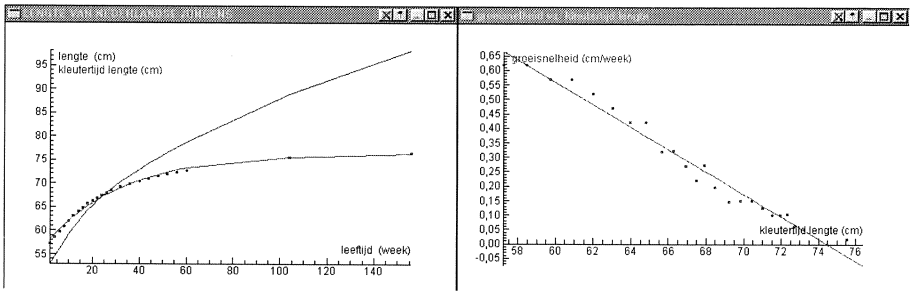


FIGUUR 8. Handmatige bepaling van de kromme voor groei in kinderperiode

We verschuiven deze parabool in verticale richting zodanig dat een positieve bijdrage aan de totale lichaamslengte optreedt vanaf de leeftijd van 6 maanden. Kortom, we nemen als formule:

$$L_2 = -0{,}235t^2 + 9{,}5t - 4{,}7\,.$$

We trekken vervolgens deze bijdrage af van de groeicijfers voor lengtegroei in de eerste drie levensjaren. We krijgen zo aangepaste cijfers voor de lengte in de kleuterperiode. We veronderstellen geremde groei met een lineair afnemende groeisnelheid. Hoe goed of slecht dit model is merk je als je de groeisnelheid uitzet tegen de lengte. In het diagramvenster rechts in figuur 9 is de beste rechte lijn volgens de kleinste kwadratenmethode bij de punten getekend. Links



FIGUUR 9. Lengtegroei in de kleuterperiode

staan de grafieken van de oorspronkelijke groeicijfers in de kleuterperiode en de aangepaste cijfers, samen met de handmatig bepaalde kromme met formule

$$L_1 = 76{,}4 - 19{,}4e^{-0{,}03t},$$

waarbij de leeftijd t in weken gegeven is.

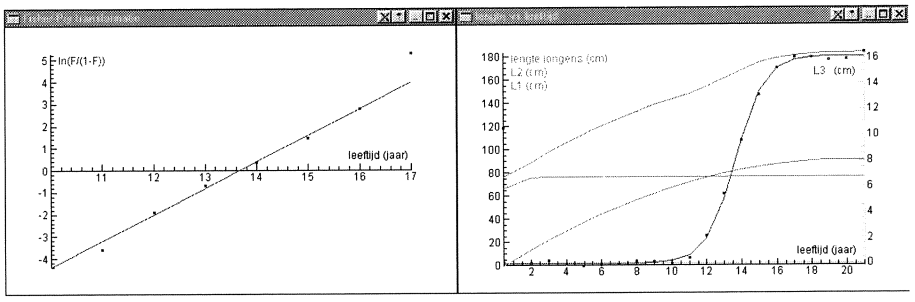
Komen we tenslotte bij de bijdrage van de pubertijdspurt. We trekken eerst de bijdragen L_1 en L_2 af van de gegeven groeicijfers en krijgen zo de bijdrage van de pubertijdspurt aan de lengte, genoteerd met L_3. Zoals in figuur 10 te zien is, lijkt de grafiek van L_3 inderdaad op een logistische kromme. Met een trial-and-error methode kun je geschikte waarden vinden voor de parameters in de formule van een logistische kromme. Rechts in figuur 10 is de grafiek getekend van

$$L_3 = \frac{16{,}1}{1 + e^{16{,}4 - 1{,}2t}}.$$

De overeenstemming met de getekende punten, afkomstig van de echte groeicijfers uit 1997, is frappant.

Bij een wiskundig model moet je jezelf altijd afvragen hoe goed of slecht het functioneert. Hoe goed het logistisch model voor de derde component van het KKP-model werkt is in het linkerdeel van figuur 10 te zien. Hier is de Fisher-Pry transformatie van de gegevens uitgezet tegen de leeftijd in de periode van 10 tot 17 jaar. De Fisher-Pry transformatie van gegevens die gemodelleerd worden via een logistische model met formule

$$L(t) = \frac{a}{1 + e^{c - bt}}$$



FIGUUR 10. Bijdrage van de pubertijdspurt aan de lengte

is gedefinieerd als

$$\ln\Big(\frac{F}{1-F}\Big),$$

met $F = L/a$. Er geldt nu een lineair verband:

$$\ln\Big(\frac{F}{1-F}\Big) = b\,t - c.$$

Dit is ook de basisgedachte achter de Fisher-Pry transformatie: als het logistische groeimodel goed functioneert, dan moet een lineair verband herkenbaar zijn. De parameters b en c zijn dan via lineaire regressie te bepalen. Je ziet het omzetten van gegevens naar een andere vorm, waarin modelparameters zich gemakkelijker laten schatten en waarin de deugdelijkheid van een model beter aan het licht komt, vaak terug bij wiskundig modelleren. In het diagramvenster links in figuur 10 hebben we als waarde voor a gekozen 16,1 en de gegevens na de Fisher-Pry transformatie getekend samen de rechte lijn horende bij de formule $16{,}4 - 1{,}2t$. De gevonden parameterwaarden geven aan dat de maximale lengtetoename tijdens de pubertaire groeispurt van jongens gemiddeld op de leeftijd van 13 jaar en 8 maanden optreedt (16,4/1,2).

We hebben hier een trial-and-error methode gebruikt om de logistische groeiparameters te schatten. De waarden van de parameters hadden ook m.b.v. de modelleeromgeving van Coach bepaald kunnen worden. Elk van deze methoden heeft zijn charme; de β-tool moet dan ook beide technieken beschikbaar hebben.

De opdracht wordt natuurlijk interessanter als je ook de gemiddelde lengte van Nederlandse meisjes met het KKP-model bestudeert. Je vindt dan een pubertaire groeispurt met maximale lengtetoename gemiddeld op de leeftijd van 11 jaar en 4 maanden en een bijdrage aan de volwassen lengte van 8,7 cm (bijna de helft van de 16,1 cm bij jongens). Je vindt zo een getalsmatige onderbouwing van het gegeven dat meisjes eerder in de pubertijd geraken en ook eerder de bijpassende groeispurt doormaken.

Denk niet dat het KKP-model het enige succesvolle wiskundige model voor lengtegroei van jongens en meisjes is. In de literatuur (zie [7]) zijn diverse

andere modellen te vinden. Twee gangbare modellen met respectievelijk 9 en 7 parameters zijn:

- Het trilogistische model van Bock en Thissen [3], dat de volgende formule voor lengtegroei L hanteert:

$$L(t) = \frac{\theta_1}{1+\exp(\frac{\theta_2-t}{\theta_3})} + \theta_9\Big(\frac{1-\theta_8}{1+\exp(\frac{\theta_4-t}{\theta_5})} + \frac{\theta_8}{1+\exp(\frac{\theta_6-t}{\theta_7})}\Big) .$$

- Het JPPS-model [12], met als formule:

$$L(t) = \theta_1 \Big(1 - \frac{1}{1+(t/\theta_2)^{\theta_3}+(t/\theta_4)^{\theta_5}+(t/\theta_6)^{\theta_7}}\Big) .$$

Wil je deze modellen kunnen toepassen dan moet je wel de beschikking hebben over een rekenprogramma dat niet-lineaire regressie toestaat en kennis of ervaring hebben in het schatten van beginwaarden van parameters. Het trilogistische model, toegepast op de gemiddelde lengte van Nederlandse jongens, levert als benadering op:

$$L(t) = \frac{48{,}1}{1+e^{-2{,}29(t+0{,}09)}} + \frac{101{,}5}{1+e^{-0{,}33(t-3{,}23)}} + \frac{34{,}9}{1+e^{-0{,}71(t-13{,}32)}}$$

Afwijkingen met de echte groeicijfers zijn tussen 0 en 21 jaar minder dan 1 cm. Belangrijkste oorzaak van de tamelijk grote afwijkingen is dat de groei in het eerste levensjaar er niet zo goed mee beschreven wordt. Laat je deze periode weg en pas je het model toe op de groei tussen 1 en 21 jaar dan benaderen de resultaten de echte groeicijfers tot op een millimeter. Hetzelfde geldt voor het JPPS-model: de beste formule voor lengte tussen 1 en 21 jaar is

$$L(t) = 184{,}0 \Big(1 - \frac{1}{1+(t/12{,}45)^{13{,}18}+(t/7{,}43)^{2{,}27}+(t/2{,}91)^{0{,}33}}\Big)$$

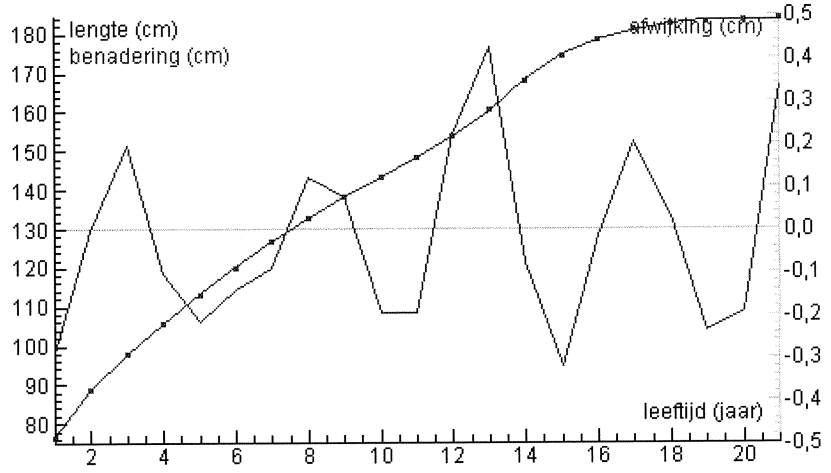
en wijkt minder dan 2 mm af van de echte groeigegevens.

Wij kunnen tevreden zijn met het resultaat van het KKP-model. In figuur 11 zijn de bijdragen aan de lengte van de drie componenten bij elkaar opgeteld en samen met de echte groeigegevens getekend. De afwijking tussen de echte en berekende lengtes zijn minder dan een halve centimeter over de hele periode van geboorte tot volwassen lengte. En dit met de betrekkelijk eenvoudige formule

$$L(t) = -19{,}4\, e^{-1{,}56t} - 0{,}235\, t^2 + 9{,}5\, t + 71{,}7 + \frac{16{,}1}{1+e^{16{,}4-1{,}2t}} ,$$

die helemaal opgebouwd is uit drie op school behandelde wiskundige modellen.

We besluiten met nog wat verdere onderzoeksvragen die je in een project over menselijke groei aan de orde kunt stellen:



FIGUUR 11. KKP-model voor gemiddelde lengte van Nederlandse jongens

- Hoe verloopt de toename in gewicht bij jongens en meisjes vanaf de geboorte tot vroegvolwassen leeftijd? Kun je dit met een eenvoudig wiskundig model beschrijven?
- Welke veranderingen in de ontwikkeling van kinderen van de ene generatie naar de andere kun je halen uit de groeicijfers van de landelijke groeistudies van 1980 en 1997? Klopt het dat Nederlanders gemiddeld langer en dikker worden?
- Welke verschillen zijn er in lengte, gewicht en Quetelet index tussen Amerikaanse en Nederlandse kinderen?

Antwoorden op bovenstaande onderzoeksvragen heeft de auteur zelf met Coach achterhaald.

7. CRITERIA VOOR DE β-TOOL

Mede op basis van ervaringen opgedaan met het gebruik van Coach in zetten we een aantal criteria waaraan de β-tool dient te voldoen op een rijtje. Dit lijstje maakt duidelijk hoe complex het maken van een bètabreed inzetbare leeromgeving is en dat de β-tool niet in één nacht gebouwd kan worden.

Gebruiksvriendelijkheid is een voor de hand liggend criterium van hulpmiddelen in en buiten onderwijs, maar omdat een bètabreed pakket heel veel faciliteiten moet bieden, zijn een heldere opzet en goede hulpfaciliteiten nog belangrijker. Het organiseren van werk en veranderen van representaties moet soepel gaan. Naar hartelust moet je met verschillende, al dan niet gekoppelde, representaties kunnen werken en deze tegelijkertijd in beeld kunnen brengen.

Het *instrumentele en activerende karakter* van de β-tool is essentieel. Dit betekent dat het in eerste instantie een toegankelijk, krachtig en 'neutraal' stuk gereedschap vormt, dat vrij is van didactische contexten of opvattingen. Dit laatste betekent niet dat de software niet speciaal ontworpen is om binnen een onderwijssituatie dienst te doen of dat het constructivistische principe van actief en zelfstandig leren niet ondersteund wordt.

Profielbreed gebruik van de software is uitgangspunt. Dit vereist dat de computer zowel voor vergaren van gegevens (via een echt experiment of videometing) als voor het wiskundig verwerken van data en modelleren gebruikt wordt. En dit alles binnen één leeromgeving.

Studiehuis-geschiktheid van de leeromgeving impliceert dat deze de leerlingen ondersteunt in het maken van grotere opdrachten en werkstukken. Daarnaast biedt de β-tool de docent de mogelijkheid om een lesonderwerp toe te lichten, uit te leggen of van een demonstratieve proef te voorzien en zijn er faciliteiten om leerlingen (op afstand) te begeleiden. Dit pleit ook voor een softwarematige leeromgeving, omdat werken door leerling en docent niet meer gekoppeld is aan een bepaalde werkruimte. Werken en leren kan plaatsvinden in vak- en computerlokaal, laboratorium, bibliotheek of thuis.

Het *open karakter* van de leeromgeving komt nog meer naar voren als er open problemen mee bestudeerd worden. Er is geen voorkeursmethode of strategie op de achtergrond aanwezig. De gebruiker bepaalt in hoge mate zelf wat er gedaan wordt, waarom en hoe.

Aanpasbaarheid van de leeromgeving staat ook hoog in het vaandel. Het is mogelijk om met verschillende leerbronnen binnen de omgeving te werken. Deze zijn door docenten en auteurs van leermiddelen van tevoren klaar te zetten en zijn op het niveau van leerlingen aan te passen. Deze bronnen moeten ruim worden opgevat: het kunnen bestaande proeven zijn, teksten, video's, opdrachten en vragen, enzovoort.

De behoefte aan een *gemeenschappelijke werkplaats* voor leraren en leerlingen is met de veranderende visie over de rol van de docent groter geworden. Niet alleen het ontwerpen van leersituaties, maar ook het (op afstand) kunnen volgen en aansturen van het leerproces en het kunnen beoordelen van leerlingenwerk dient ondersteund te zijn.

Koppeling met andere software is en blijft nodig, al is het maar om resultaten te transporteren naar verslagen, die opgemaakt worden met een favoriete tekstverwerker, om uitwisseling met presentatie-software te hebben of om gegevens uit bestaande databestanden te importeren.

Aantrekkelijkheid van de leeromgeving wordt mede bepaald door de mogelijkheden die deze biedt om in exacte vakken leerlingen realistische problemen uit hun eigen leefwereld te laten bestuderen. Wiskundige gereedschappen zoals de regressie-tool en de modelleeromgeving zorgen ervoor dat een leerling hiervoor niet eerst de algoritmische vaardigheden hoeft te verwerven. Hierdoor

wordt de wiskunde aantrekkelijk, heel concreet, direct verifieerbaar en illustreerbaar: mooie grafieken en krommen, interessante resultaten uit simulaties, enzovoort.

8. Tot slot

Het AMSTEL instituut wil de opvolger van Coach, de β-tool, net als voorheen ontwikkelen in samenwerking met allen die zich betrokken voelen bij onderwijs in exacte vakken. In het bijzonder wiskundeleraren worden uitdrukkelijk uitgenodigd mee te denken en mee te bouwen aan de nieuwe leeromgeving plus lesmaterialen.

Wat het laatste betreft: hopelijk put u inspiratie uit de voorbeelden in dit artikel en wordt u erdoor aangemoedigd om zelf de mogelijkheden van Coach bij wiskunde te gaan bekijken. Belangrijk hierbij is te weten dat de software reeds op de meeste scholen aanwezig is en dat het pakket op een PC met bescheiden geheugen en schijfruimte gebruikt kan worden. U kunt dus direct aan de slag. Meer informatie over software, hardware, toe- passingen en lesmateriaal kunt u vinden op de website `www.cma.science.uva.nl`.

Referenties

1. Armenti, Jr. A. (red) *The Physics of Sports.* Springer-Verlag, New York, 1992.
2. Aroesty, J. *et al.* Tumor growth and chemotherapy: Mathematical models, computer simulations, and experimental foundations. *Math. Biosciences*, **17**: 243-300, 1973.
3. Bock, R.D. & Thissen, D.M. Fitting multicomponent models for growth in stature. In *Proceeedings of the Ninth International Biometric Conference, Boston, August 22-27, 1976*, pp. 431-442, The Biometric Society, Raleigh, North Carolina.
4. Causton, D.R. & Venus, J.C. *The Biometry of Plant Growth.* Edward Arnold, London, 1981.
5. Craats van de, J. Hoe hangt een ketting? *Nieuwe Wiskrant* **19**(1): 32-36, 1999.
6. Ellermeijer, T. & Mulder, C. IP-Coach: een succesvolle leeromgeving voor de natuurwetenschappelijke vakken en techniek. *TINFON* **7**(4): 131-134, 1998.
7. Falkner, F. & Tanner, J.M. (red) *Human growth: a comprehensive treatise. Vol. 3.* Plenum Press, New York, 2nd edition, 1986.
8. Fredriks, A.M., *et al.* Continuing Positive Secular Growth Change in the Netherlands 1955-1997. *Pediatric Research*, **47** (3): 316-323, 2000.
9. Greef de, K.M. *Prediction of production: nutriton induced tissue partitioning in growing pigs.* Proefschrift, Universiteit Wageningen, 1992.
10. Gyllenberg, M. & Webb, G.F. Quiescence as an Explanation of Gompertzian Tumor Growth. *Growth, Development and Aging,* **53**: 25-33, 1989.
11. Heck, A. Coach: β-tool in spe en nu al inzetbaar bij wiskunde. *Nieuwe Wiskrant*, **19**(4): 36-42, 2000.

12. JOLICOEUR, P., PONTIER, J., PERNIN, M.-O., SEMPÉ, M. A lifetime asymptotic growth curve for human height. *Biometrics*, **44**: 995-1003, 1988.
13. KARLBERG, J. *et al.* Linear growth retardation in relation to the three phases of growth. In *Causes and Mechanisms of Linear Growth Retardation.* WATERLOW, J.C. & SCHÜRCH, B. (red), Proceedings I/D/E/G/C Workshop, London, U.K., 1993. Op 14 juni 2000 electronisch beschikbaar op URL `www.unu.edu/unupress/food2/uid06e/uid06e00.htm`
14. MULDER, C. *Computer-based Investigation in Physics.* Proefschrift, Universiteit van Amsterdam (in voorbereiding), 2000.
15. TOWNEND, M.S. *Mathematics in Sport.* Ellis Horwood Limited, Chichester, 1984.
16. WIT, J.M (red) *De vierde landelijke groeistudie*, Boerhave Commissie, Leiden, 1998.

Vakantiecursus 2000

pp. 91 – 102

Wiskunde werkt!

Jaap Molenaar
Technische Universiteit Eindhoven

1. Aard en functie van de wiskunde

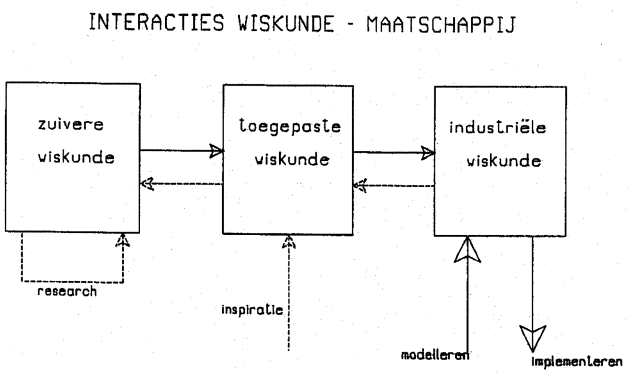
Wiskunde wordt gekarakteriseerd door een strenge opbouw: uitgaande van een aantal basisgegevens, de axioma's, worden er beweringen, de stellingen, gedaan en bewezen en de aldus verkregen inzichten vormen tezamen een solide bouwwerk, met een fraaie structuur en zonder tegenstrijdigheden oftewel inconsistenties. In principe kan dit bouwwerk geheel abstract zijn. Als de gekozen basisgegevens ontsproten zijn aan onze fantasie en geen relatie met de ons omringende werkelijkheid hebben, zal het bouwwerk deze relatie in eerste instantie ook niet hebben. In de praktijk echter komt volledige abstractie weinig voor. Meestal is er een inspiratiebron die heel concreet is. De meetkunde bijvoorbeeld heeft relatie met landmeetkunde, de differentiaalvergelijkingen komen voort uit de behoefte mechanische systemen te beschrijven. En zelfs gedeelten van de wiskunde die begonnen zijn op een volledig abstracte basis, blijken later toch een relatie te hebben met onverwachte toepassingen in andere vakken. Er is weliswaar een fundamenteel verschil tussen wiskunde en de andere wetenschappen, maar in de praktijk is er een sterke band. Wiskunde is inductief, kan zijn eigen problemen stellen en bouwt eigen formalismen los van concrete vragen, terwijl de andere vakken hun vraagstellingen krijgen aangereikt vanuit het betreffende vakgebied - de dode natuur (natuurkunde, scheikunde), de levende natuur (de biologie), de economie, interactie tussen mensen (sociologie), enz. - en dus een deductief karakter hebben. Hoewel deze onderscheiding volledig juist is, leidt het niet tot een tegenstelling. De onderwerpen die de niet-wiskundevakken beschrijven, blijken vaak wiskundig van aard te zijn, d.w.z. ze zijn te beschrijven met behulp van de taal van de wiskunde. Zeker voor de natuurwetenschappen geldt dat de schepping functioneert volgens wetten die in wiskundige termen zijn te formuleren. De wiskunde is in dit opzicht "unreasonably effective" [1]. Stukken wiskunde die een bepaald verschijnsel of proces beschrijven noemen we een "wiskundig model". De optredende grootheden en vergelijkingen hebben dan een interpretatie. Een wiskundig model kan goed of slecht zijn. Als een model slecht is betekent dat niet dat de gebruikte wiskunde incorrect is, maar dat de resultaten van het model niet goed overeenstemmen met de experimentele gegevens. Zo'n model dient aangepast te worden.

Daarmee belanden we op de mogelijkheid de wiskunde te classificeren aan de hand van de relatie tot de 'buitenwereld'. Eigenlijk classificeren we dan niet de wiskunde zelf - de aard daarvan is éénduidig - maar de manier waarop deze functioneert. Een groffe indeling is: zuivere wiskunde, toegepaste wiskunde, industriële wiskunde (zie Figuur 1) met als kenmerken:

- Zuivere wiskunde: De inspiratie en de axioma's komen voort uit de creativiteit van de wiskundige en de resultaten hebben niet een directe toepassing elders. Een voorbeeld vormen bepaalde gedeelten van de getaltheorie, zoals het recente bewijs van de laatste stelling van Fermat door Wiles.

- Toegepaste wiskunde: De inspiratie komt voort uit concrete probleemstellingen. Daarvoor wordt de benodigde wiskunde ontwikkeld en zo ver mogelijk gegeneraliseerd, Het gebruik ervan wordt overgelaten aan niet-wiskundigen. Een voorbeeld is het vier-kleurenprobleem. Iedere kaartenmaker weet uit de praktijk dat de gebieden op een kaart met hoogstens 4 kleuren aangegeven kunnen worden, ook als je eist dat buurlanden altijd een verschillende kleur hebben. In de grafentheorie is dit bewezen. Voor het specifieke probleem is dat niet van belang, de algemene resultaten echter van dit vakgebied zijn zeer breed toepasbaar.

- Industriële wiskunde: De motivatie wordt gevormd door concrete vragen uit de industrie of andere delen van de maatschappij. Mathematische modellering is een belangrijk onderdeel van het werk. De benodigde wiskunde kan meestal geput worden uit het bestaande wiskundereservoir, hoewel soms nieuwe resultaten moeten worden afgeleid. De industrieel wiskundige is intensief betrokken bij het gebruik van zijn resultaten bij het oplossen van de oorspronkelijke vraag. Voorbeelden worden verderop in deze bijdrage gegeven.



FIGUUR 1.

Bovenstaande indeling is nuttig maar kan misleidend zijn. De grenzen zijn namelijk niet scherp. De wiskundige als persoon kan in principe actief zijn in alle genoemde categorieën tegelijk. Opvallend is dat wiskundige kennis die als zuivere wiskunde gegenereerd is, soms (veel) later een onverwachte toepassing krijgt bij industriële vragen. En omgekeerd leiden heel concrete vragen en observaties soms tot het ontwikkelen van nieuwe wiskundige vakgebieden.

Gegevensbescherming bijvoorbeeld is een zeer actuele vraag die inspireert tot veel wiskundig onderzoek.

2. INDUSTRIËLE WISKUNDE

Industriële wiskunde wordt meestal projectmatig beoefend. De stappen in zo'n project kunnen heel schematisch als volgt weergegeven worden:

a) Inventariseren

b) Mathematisch modelleren

c) Dimensieloos formuleren

d) Reduceren

e) Analyseren

f) Interpreteren en verifiëren

g) Optimaliseren door itereren

h) Implementeren

Over iedere stap is veel te zeggen; u zij daarvoor verwezen naar [2] en de vele boeken die de afgelopen decennia verschenen zijn en "mathematical modelling" in de titel hebben. Niet altijd zijn alle stappen even duidelijk aanwezig. In ieder geval wel altijd de stappen a), b), e) en h). Bij a) en b) wordt nagegaan welke bestaande wiskundige kennis toepasbaar is op de situatie. Voor sommige verschijnselen zijn de regels zo fundamenteel dat ze 'wetten' genoemd worden. Voor fysische processen gaat het dan vaak om het formuleren van de relevante behoudswetten. Voorbeelden hiervan zijn de wetten van Newton in de mechanica, de wetten van Maxwell voor electro-magnetische verschijnselen en de Navier-Stokes vergelijkingen in de vloeistofmechanica. In de meeste andere gevallen zijn de regels veel minder goed bekend en dient men ze zelf te ontwerpen. Deze regels, die de verbanden aangeven tussen de relevante grootheden, worden in wiskundetaal weergegeven en vormen samen een *model*. De stappen c) en d) leiden tot vereenvoudigingen in het model. Het is vaak mogelijk aan te geven welke effecten dominant zijn en dus zeker in het model opgenomen moeten worden en welke in eerste instantie verwaarloosd kunnen worden. Nadat de oplossingen van het mathematische model bepaald zijn - hetzij analytisch, hetzij numeriek - dient er nagegaan te worden of deze zinnig zijn: stap f). Liefst moeten ze gekontroleerd worden aan de hand van enkele proefnemingen. Als blijkt dat de afwijkingen tussen berekende en gemeten waarden te groot zijn, moet nagegaan worden welke aanpassingen gedaan kunnen worden om het model beter de realiteit te laten beschrijven. Het kan bijvoorbeeld zijn dat er ten onrechte effecten verwaarloosd zijn.
De laatste stap, het implementeren, is van groot praktisch belang. Deze fase kan beslissend zijn voor het succes van het project als geheel. Het ontwikkelde model kan nog zo goed zijn, als het niet in een vorm gebracht wordt waarmee

de vraagsteller uit de voeten kan, is de kans groot dat het niet gebruikt wordt. De voorgestelde oplossingen moeten binnen de technische mogelijkheden van het betreffende bedrijf liggen, de geleverde software moet gebruikersvriendelijk zijn, het model moet niet toegepast worden op situaties waarvoor het niet ontworpen is: er zijn altijd vele randcondities waaraan voldaan moet worden en die meestal weinig met de wiskunde zelf te maken hebben.

3. Studiegroep Wiskunde met de Industrie

De stappen a) - e) genoemd in paragraaf 2 komen heel expliciet aan de orde tijdens de Studiegroep Wiskunde met de Industrie (SWI). Dit fenomeen is zo'n dertig jaar geleden gestart in Oxford, o.a. door A. Tayler. Men kreeg voortdurend vragen vanuit de industrie en kwam tot de conclusie dat de beantwoording ervan vraagt om speciale expertise in het mathematisch modelleren van industriële processen. Men besloot een aantal vragen op te sparen en die te behandelen tijdens een daarvoor gereserveerde week. Dit initiatief is sindsdien ieder jaar herhaald. De locatie wisselt de laatste jaren. Het idee is overgenomen op verschillende andere plaatsen in de wereld: Australië, Canada, Mexico, Denemarken en sinds 1998 ook in Nederland. Het programma is overal hetzelfde. Op maandag presenteren bedrijfsvertegenwoordigers de problemen. Het zijn in de regel moeilijke vragen waar men al enige tijd mee tobt of die men van groot belang acht voor toekomstige technologische doorbraken. De problemen worden gepresenteerd zoals ze zich in de praktijk voordoen, dus zonder wiskundige termen. De aanwezige wiskundigen - de meesten afkomstig van universiteiten, sommigen uit het bedrijfsleven - gaan vervolgens in groepen aan de problemen werken. De samenstelling van de groepen kan wisselen; iedereen die een goed idee heeft is welkom in iedere groep. Omdat de meeste deelnemers weinig weten van de context van de vragen, wordt de eerste tijd vooral besteed aan vragen stellen, inventariseren en brainstormen. Van maandagmiddag tot en met donderdagavond wordt er gewoonlijk met groot enthousiasme en inzet gewerkt aan het opstellen van modellen en - indien de tijd het toelaat - het analyseren en oplossen ervan. Het is vaak verbazend wat er in zo'n korte tijd bereikt wordt. Op vrijdag wordt aan degenen die de vragen inbrachten gepresenteerd wat de stand van zaken is. Soms is er al een compleet model met oplossingen beschikbaar. Meestal zijn er verschillende modellen en technieken onderzocht en kan geadviseerd worden hoe de betreffende industrie verder zou moeten gaan. Na de week wordt er door de deelnemers een rapport geschreven over de resultaten van het werk. Vaak blijken er tijdens het schrijven nog een aantal goede ideeën naar voren te komen, zodat het verslag rijker is dan de resultaten van de week zelf. Tijdens een SWI is de interactie tussen wiskundigen en het bedrijfsleven heel intens: men ziet er als het ware industriële wiskunde in bedrijf. Voor nadere inlichtingen over deze Studiegroepen: zie [3,4,5]. Onderwerpen die tijdens SWI'98 en SWI'99 aan de orde kwamen waren o.a.:

- Betrouwbaarheidsintervallen bij kleine steekproeven (Ned. Meetinstituut)
- Optimalisatie van data-overdracht over verbindingen met ruis (KPN-research)

- Het boren van koelgaten m.b.v. lasers (Eldim)
- Modellering van het persen van gebogen tafelbladen (Trespa International)
- Detectie van metastases op longfoto's (Daniel den Hoedkliniek)
- Modellering van het verwijderen van modder uit boorschachten (Schlumberger)
- Schatten van verkeersdichtheden uit telefonische steekproeven (Ericsson Telecommunicatie)
- Efficiënt gebruik van World Wide Web Caches (KPN-research)
- Het mengen van kleuren (AKZO)
- Positiebepaling in een windtunnel (NLR)
- Compacte modellering van een transisteronderdeel (Philips)

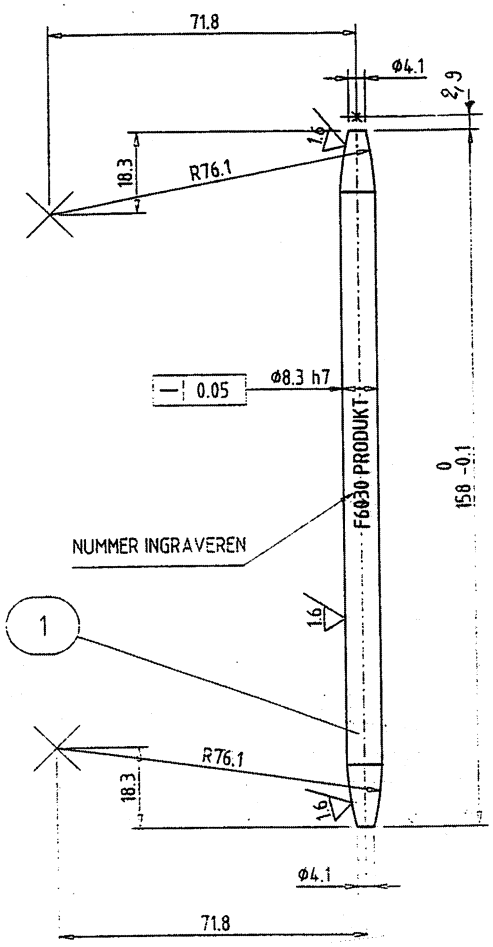
4. Voorbeeldproject: Sigaren rollen

Lareka Machines B.V. te Valkenswaard ontwerpt en produceert innovatieve machines, o.a. voor de sigarenindustrie. Bestaande machines om sigaren te rollen zijn traag en het kost erg veel tijd ze "om te stellen", d.w.z. van het ene type sigaar op het andere over te gaan. Voor de besturing van het rolproces klopte men aan bij wiskundigen. Het probleem blijkt oplosbaar met behulp van reeds lang bekende inzichten uit de differentiaalgeometrie. Voor een uitgebreide beschrijving van dit project voor de sigarenindustrie, zie [6].

4.1. De vraag

Een sigaar bestaat uit opvulsel van kruimeltabak met daarom heen het dekblad. De kwaliteit van het dekblad is bepalend voor de kwaliteit van de sigaar. De vorm van de sigaar zonder dekblad - de zogenaamde "klos" - is zeer nauwkeurig voorgeschreven. Het is een omwentelingsfiguur; de om te wentelen curve is gespecificeerd door middel van lijnstukken en cirkelbogen. Zie Figuur 2 voor een voorbeeld.

Het dekblad is een strook gestansd uit een sigarenblad. Bij iedere klos wordt de vorm van de strook dekblad - de zogenaamde "stansvorm" - met grote zorg ontworpen. Voor voorbeelden: zie Figuur 3. Bij het rollen wordt de stansvorm met een pin aan het uiteinde van de klos bevestigd onder een nauwkeurig voorgeschreven starthoek tussen de lengteas van de klos en de *backbone* van de stansvorm. Deze *backbone* is de rechte lijn aangegeven in Figuur 3. De klos gaat draaien en de stansvorm wikkelt zich eromheen. De stansvorm die in eerste instantie rond een cilinder is bevestigd rolt daarbij van deze cilinder af. Om ervoor te zorgen dat de stansvorm soepel om de klos windt en niet gaat kreukelen, moet deze cilinder bewegen langs de klos. Tevens moet hij draaien om zich aan te passen aan de varierende hoek tussen lengteas klos en *backbone* van de stansvorm. De vragen die door Lareka gesteld werden zijn:

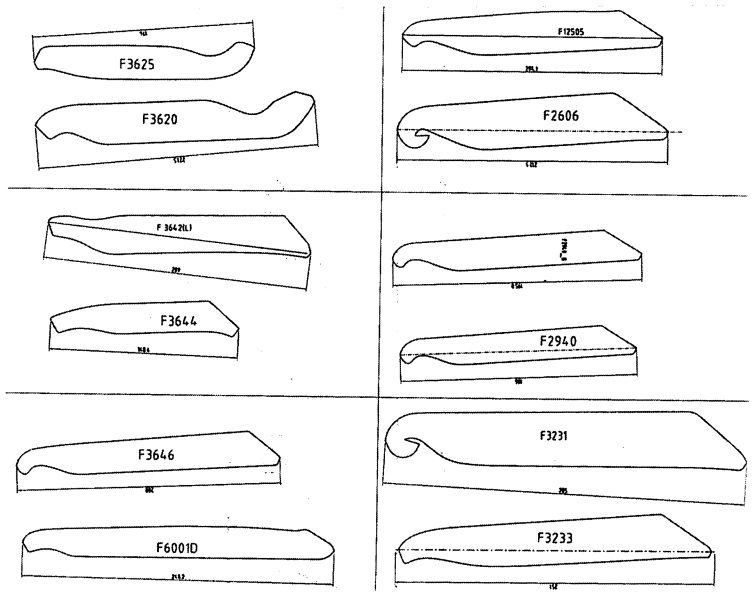


FIGUUR 2.

a) Gegeven de klosvorm, bereken de baan die de *backbone* rond de klos volgt als functie van de starthoek en bereken daaruit de gegevens voor de besturing van cilinder.

b) Als a) opgelost is voor zekere klosvorm, bereken dan voor gegeven stansvorm de overlap van opeenvolgende wikkelingen van de stansvorm.

c) Gegeven de klosvorm, ontwerp voor gegeven starthoek de stansvorm zodanig dat aan zekere eisen op de overlap is voldaan.



FIGUUR 3.

4.2. De modellering

Het is duidelijk dat bovengenoemde vragen liggen op het terrein van de theorie van krommen en oppervlakken, een tamelijk klassiek wiskundig vakgebied, dat in de industriële praktijk overigens opvallend relevant is.

Centraal staat in dit project de vraag wat de baan van de *backbone* is op het sigaaroppervlak. In eerste instantie kan de precieze vorm van de stans verwaarloosd worden. Wel heeft het feit dat de *backbone* onderdeel is van een strook grote consequenties. Een 1-dimensionaal touwtje kan op een oneindig aantal manieren rond de klos gevonden worden. Voor de stansvorm geldt echter dat, indien de beginpositie en de starthoek gegeven zijn, de baan volledig vastligt. Dit is een constatering die iedereen kan maken na enig 'spelen' met het materiaal. Wiskundig is dit ook voortreffelijk te beschrijven. Maar eerst moet er een "hobbel" overdacht worden.

4.3. Mathematisch model als benadering

Als we een strook papier rond een bol wikkelen, raakt het papier de bol slechts langs een lijn, de *backbone*. Als we daarentegen een strook rond een cilinder of kegel wikkelen, raakt de gehele strook de cilinder of kegel. We kunnen dus bij de sigaar verwachten dat de stansvorm in het midden, waar de sigaar vrijwel cilindervormig is, de klosvorm goed volgt, maar dat dekblad en klos niet goed "passen" aan de uiteinden. Wiskundig gezien is hier een probleem, maar in de praktijk valt dit reuze mee omdat het dekblad - natuurlijk materiaal - vrij

rekbaar is en zich tamelijk makkelijk rond de klosuiteinden plooit. Er zal dus een discrepantie tussen mathematisch model en de reële situatie optreden. Deze discrepantie zal klein zijn als de breedte van de stansvorm klein is daar waar de kromming van het klosoppervlak groot is. Deze discrepantie is typisch voor mathematisch modelleren. De werkelijkheid en het model verschillen altijd en het is noodzakelijk voor ieder model de condities te specificeren waaronder het toepasbaar is.

4.4. Krommen en oppervlakken

Een goede samenvatting van de theorie van krommen en oppervlakken is te vinden in [7,8].

De sigaar vormt een omwentelingslichaam. De vorm is gegeven door een functie $f(x)$, $0 \leq x \leq L$, met de L de sigaarlengte. We kiezen Cartesische coördinaten met als x-as de lengteas van de sigaar, en als oorsprong het begin van de sigaar. Een punt $\mathbf{x}$ op de sigaar wordt dan gegeven door

$$\mathbf{x}(x,\varphi) = (x, f(x)\cos\varphi,\ f(x)\sin\varphi) \tag{1}$$

en ligt dus vast als x en de omwentelingshoek φ (met $0 \leq \varphi < 2\pi$) gespecificeerd zijn. In ieder punt kunnen we twee tangentvectoren vinden, die raken aan het oppervlak. Het zijn

$$\frac{\partial \mathbf{x}}{\partial x} \equiv \mathbf{x}_x \text{ en } \frac{\partial \mathbf{x}}{\partial \varphi} \equiv \mathbf{x}_\varphi \tag{2}$$

De vectoren $\mathbf{x}_x$ en $\mathbf{x}_\varphi$ spannen het tangentvlak in het betreffende punt op. De normaalvector $\mathbf{N}$ in dat punt staat loodrecht op het tangentvlak, dus (zonder normalisatie)

$$\mathbf{N} = \mathbf{x}_u \times \mathbf{x}_\varphi \tag{3}$$

Veel eigenschappen van het oppervlak worden bepaald door de zogenaamde eerste-fundamentaalcoëfficiënten, gedefinieerd als:

$$\begin{aligned} E &= \mathbf{x}_x \cdot \mathbf{x}_x \\ F &= \mathbf{x}_x \cdot \mathbf{x}_\varphi \\ G &= \mathbf{x}_\varphi \cdot \mathbf{x}_\varphi \end{aligned} \tag{4}$$

Uit deze definitie blijkt dat E en G gegeven worden door (het kwadraat van) de lengten van de twee afzonderlijke tangentvectoren, terwijl F bepaald wordt door de hoek ertussen. Voor het omwentelingsoppervlak gegeven door (1) geldt:

$$\mathbf{x}_x = \begin{pmatrix} 1 \\ f'\cos\varphi \\ f'\sin\varphi \end{pmatrix},\ \mathbf{x}_\varphi = \begin{pmatrix} 0 \\ -f\sin\varphi \\ f\cos\varphi \end{pmatrix}, \tag{5}$$

en dus

$$E = 1 + (f')^2$$

(6) $$F = 0$$

$$G = f^2$$

Het is typerend voor een omwentelingsoppervlak dat E, F, G alleen van x afhangen en niet van φ.

Laten we de baan van de *backbone* op het sigaaroppervlak aangeven met $\mathbf{y}(x(s), \varphi(s))$. Deze baan kunnen we zien als een afbeelding van een baan $(x(s), \varphi(s))$ in het (x, φ) vlak naar de sigaar. Hierbij is s de booglengte van de baan in het (x, φ) vlak.
Voor een willekeurige kromme $\mathbf{y}(s)$ definiëren we de tangentvector $\mathbf{t}$ en de normaalvector $\mathbf{n}$ in een punt door

(7) $$\mathbf{t} = \frac{d\mathbf{y}}{ds}, \; \mathbf{n} = \frac{d\mathbf{t}}{ds}$$

We kunnen hier eenheidsvectoren van maken door te delen door de lengtes, maar dat is hier niet relevant.

4.5. Sigaargeometrie

We komen nu tot een inzicht dat essentieel is voor de modellering. Merk op dat langs de baan van *backbone* op het sigarenoppervlak de tangentvlakken van het sigarenoppervlak en de stansvorm samenvallen. Of met andere woorden, de normaalvectoren $\mathbf{N}$ van het oppervlak en $\mathbf{n}$ van de *backbone* zijn evenwijdig. Dit impliceert dat de baan van de *backbone* een *geodeet* is. In een vlak zijn de rechte lijnen de geodeten. Door ieder punt van het vlak gaat een schaar geodeten waarvan de leden onder verschillende hoeken vertrekken. Op een bol zijn de geodeten de grote cirkels. Door ieder punt gaat zo'n familie grote cirkels. Analoog geldt op een willekeurig (glad) oppervlak dat door ieder punt een familie geodeten gaat. Door de starthoek vast te leggen kiezen we dus op het sigaaroppervlak aan het begin een geodeet en de *backbone* volgt deze uniek bepaalde baan.

Nu dit inzicht verkregen is, wordt de rest een kwestie van techniek. Er is van alles bekend van geodeten. In het (x, φ) vlak zijn het oplossingen van het volgende stelsel gewone differentiaalvergelijkingen:

(8)
$$\frac{d^2x}{dx^2} + \Gamma^1_{11}\Big(\frac{dx}{ds}\Big)^2 + 2\Gamma^1_{12}\Big(\frac{dx}{ds}\Big)\Big(\frac{d\varphi}{ds}\Big) + \Gamma^1_{22}\Big(\frac{d\varphi}{ds}\Big)^2 = 0$$

$$\frac{d^2\varphi}{ds^2} + \Gamma^2_{11}\Big(\frac{dx}{ds}\Big)^2 + 2\Gamma^2_{12}\Big(\frac{dx}{ds}\Big)\Big(\frac{d\varphi}{ds}\Big) + \Gamma^2_{22}\Big(\frac{d\varphi}{ds}\Big)^2 = 0$$

De Christoffelsymbolen Γ^{ij}_k zijn functies van E, F, G. Voor omwentelingslichamen, waarvoor (6) geldt, vinden we

(9) $$\Gamma^2_{11} = \Gamma^2_{22} = 0; \;\; \Gamma^2_{12} = \frac{f'(x)}{f(x)}$$

Daarmee vereenvoudigt (8b) sterk, namelijk tot

$$\frac{d^2\varphi}{ds^2} + 2\frac{f'(x)}{f(x)}\left(\frac{dx}{ds}\right)\left(\frac{d\varphi}{ds}\right) = 0 \tag{10}$$

Als we dit met $f^2(x)$ $(\neq 0)$ vermenigvuldigen, vinden we

$$\frac{d}{ds}(f^2(x)\frac{d\varphi}{ds}) = 0 \tag{11}$$

Dus

$$f^2(x)\frac{d\varphi}{ds} = c_0 \tag{12}$$

voor zekere constante c_0. Deze 'behoudswet' biedt bijzonder veel informatie. We leggen dit uit aan de hand van Figuur 4. In deze figuur is de baan $(x(s), \varphi(s))$ getekend voor het geval van een perfecte cilinder, waarvoor f niet van x afhangt. In deze situatie is de hoek α tussen de tangent van de baan en de lengteas van de sigaar constant. In het algemeen geldt op het sigaaroppervlak

$$f\frac{d\varphi}{ds} = \sin\alpha\ , \tag{13}$$

zodat (12) geschreven kan worden als

$$f(x)\sin\alpha(x) = c_0 \tag{14}$$

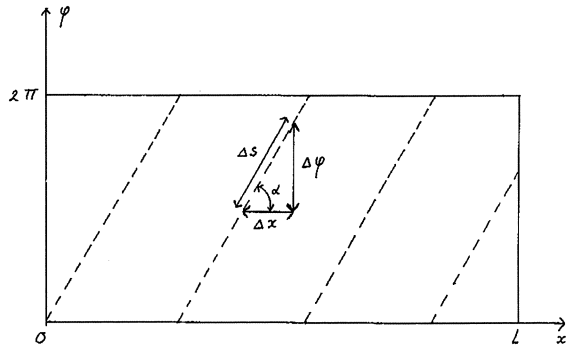
Deze relatie is zeer inzichtelijk te maken. De waarde van c_0 wordt bepaald door de waarden van f en starthoek α in de oorsprong. Als $f(x)$ voor $x \geq 0$ constant is, geldt dat ook voor α. Dat is het geval van een perfecte cilinder. Als $f(x)$ stijgt, moet $\sin\alpha(x)$ en dus $\alpha(x)$ kleiner worden en omgekeerd. Bijvoorbeeld, als we bij een kegel beginnen te wikkelen vlak bij de top in de richting van toenemende f, dan zal $\alpha(x)$ steeds kleine worden en zal de baan van de *backbone* een steeds kleiner hoek met de kegelas maken. Kortom de spoed - dat is de voortbeweging langs de x-as per omwenteling - wordt steeds groter. Voor $x \to \infty$ gaat de baan evenwijdig de as lopen en de spoed wordt dan oneindig groot. Als we daarentegen op een kegel wikkelen in de richting van afnemende f, zal de spoed afnemen. Bij de start zal altijd gelden

$$\sin\alpha(0) = \frac{c_0}{f(0)} \leq 1, \text{ dus } f(0) \geq c_0 \tag{15}$$

Een speciale situatie doet zich voor als $f(x)$ daalt en gelijk wordt aan c_0. De spoed wordt dan nul en het dekblad zal niet verder wikkelen maar in omgekeerde richting gaan wikkelen! Bij het uiteinde van een sigaar neemt $f(x)$ inderdaad af en moet men dus hierop bedacht zijn.

5. Besluit

Bovenstaande inzichten bevatten genoeg informatie om het project uit te voeren. We zullen hier niet verder ingaan op de verdere stappen in het project. Het construeren van de geodeten op de klos is relatief eenvoudig gevonden, omdat, gegeven $f(x)$ en c_0 uit de begincondities, de hoek α volgt uit



FIGUUR 4.

$$\alpha(x) = \arcsin\left(\frac{c_0}{f(x)}\right) \tag{16}$$

Een andere aanpak (zie [7]) leidt tot het inzicht dat voor een geodeet op een omwentelingsoppervlak de hoek φ als functie van x gegeven wordt door de integraal

$$\varphi(x) = \varphi(0) + c_0 \int_0^x \frac{\sqrt{1 + f'(t))^2}}{f(t)\sqrt{f^2(t) - c_0^2}} dt \tag{17}$$

Hoewel de punten a) en b) uit de vraagstelling van Lareka hiermee vrij rechttoe rechtaan te beantwoorden zijn, geldt dit niet voor c), het vinden van een geschikte stansvorm bij een gegeven klos. Er is op dit punt veel vrijheid en in gesprek met de opdrachtgever zullen de randcondities geformuleerd moeten worden voor een optimalisatieprobleem. Daarbij zullen geheel andere wiskundige technieken een rol gaan spelen. Tenslotte kwam de opdrachtgever nog met het verzoek of ook de rek in het dekblad in rekening gebracht zou kunnen worden. Dat gooit bovenstaande analyse flink door de war. Typerend voor industriële wiskunde: correcte antwoorden voor geïdealiseerde situaties zijn zelden toereikend. Aan de andere kant: men is zelden geïnteresseerd in het wiskundig gezien optimale antwoord. Een model dat antwoord geeft op de vragen die leven is het enige dat telt. Daardoor heeft de industrieel wiskundige meestal grote vrijheid om zijn creativiteit te ontplooien.

REFERENTIES

[1] E.P. WIGNER, The unreasonable effectiveness of mathematics in the natural sciences, *Comm. on pure and applied math.* **VIII**, 1–14, 1960.

[2] R.M.M. MATTHEIJ, J. MOLENAAR, *ODE in theory and practice*, Chapter XII, John Wiley & Sons, 1997, ISBN 0-471-95674-0.

[3] P.W. HEMKER (ED.), *Proceedings of the 33^{th} European Study Group with Industry*, CWI Syllabus 46, 1999, ISBN 90-6196-486-5.

[4] J. MOLENAAR (ED.), *Proceedings of SWI'99*, TUE Report00-WSK-01, ISSN 0167-9708.

[5] Zie de wegpage: `http://www.itw.nl/` en klik op studiegroep. Aldaar zijn ook links naar vele andere studiegroepen te vinden.

[6] F. MEZEI, G. SCHMITZ, *Cigar rolling*, report of a project for Lareka Machines B.V., June 2000, Mathematics for Industry Programme, Eindhoven University of Technology.

[7] M. LIPSCHITZ, *Differential Geometry*, Schaum's Outline Series, McGraw Hill, 1969.

[8] E. KREYSZIG, *Differential Geometry*, Dower Publications, New York, 1991, ISBN 0-486-66721-9.

CWI Vakantiecursus 2000

pp. 103 – 120

Computers: ook voor de wiskunde zelf

N.G. de Bruijn
Faculteit Wiskunde en Informatica
Technische Universiteit Eindhoven
email: n.g.d.bruijn@tue.nl

1. Inleiding

Iedereen gebruikt tegenwoordig computers, dus wiskundigen doen dat ook. Ze gebruiken computers voor het samenstellen van teksten, het zoeken in teksten, zoeken naar teksten en naar allerlei gegevens in het WorldWideWeb. Ze gebruiken ze voor hun correspondentie, meestal via email, en voor het archiveren daarvan. Dat alles verschilt niet zoveel van wat andere mensen er voor hun werk mee doen.

Alleen moet gezegd worden dat de wiskundige aan een tekstverwerker andere eisen stelt dan de doorsnee gebruiker. Het wiskundige formulewerk, voorheen uitgevoerd door gespecializeerde handzetters, wordt nu geproduceerd door grote computerprogramma's, waarvan de uit TEX afgeleide tegenwoordig de belangrijkste zijn. De gebruiker ziet ze als tweetraps tekstverwerkers. In de eerste trap wordt met een gewone tekstverwerker een brontekst gemaakt waarin de tekst gelardeerd is met wat TEX idioom. Het TEX programma transformeert die brontekst in drukwerk dat op de monitor te lezen is en op papier kan worden afgedrukt, in een kwaliteit die vroeger alleen door de allerbeste drukkers kon worden geleverd. Soms gaat het verder. Een deel van dit artikeltje is gemaakt door drietraps tekstverwerking. Voor al het vlagvertoon is er nl. een hulpprogramma dat een brontekst met alle gegevens omzet in een brontekst voor TEX. Zo kan iedereen voor speciale behoeften nog zelf allerlei dingen aan TEX vastbouwen.

Dat schrijven in systemen als TEX gaat zó gemakkelijk, dat veel wiskundigen het overal gebruiken waar maar even iets moet worden vastgelegd. Zonder geknoei en met groot gemak kunnen allerlei wijzigingen worden aangebracht. Deze systemen zijn zelfs te gebruiken voor dingen die vroeger alleen op kladpapier gebeurden. Een belangrijke reden om veel meer dan vroeger electronisch te werken schuilt natuurlijk ook in de electronische communicatie met anderen.

Dit alles is computerwerk ten bate van de wiskundige, maar het is nog geen wiskundig werk. Toch kunnen computers dat laatste ook, want daar is het immers allemaal mee begonnen. Oorspronkelijk waren het rekenmachines (het woord "computer" betekende niet de machine maar de menselijke rekenaar) die voor rekenwerk werd gebruikt door wat men in ruime zin toegepast wiskundigen zou kunnen noemen.

De zuiver wiskundige had er niet zo veel mee te maken. Die had altijd succes geboekt door erin te slagen het rekenen te vermijden, zoals het schooljongetje K.F. Gauss al deed toen hij de getallen van 1 tot 100 bij elkaar moest tellen. De wiskundige geniet daarvan. Met het uitvoeren van 'mechanische' procedures is nooit veel eer te behalen.

Na de de opkomst van de programmeerbare rekenmachine kreeg het simpele rekenwerk vleugels, waardoor men zich met zaken kon gaan bezighouden waar vroeger niet over viel te dromen. De hele wereld profiteerde ervan, maar de wiskunde zelf eigenlijk niet zo erg veel. Het grootste profijt is misschien dat de wiskunde uitgroeit en uitbloeit zodra er wat van wordt toegepast. We zien dat bijna overal, ook in wat hier in deze cursus wordt aangeboden. Maar die wiskundige productie, bedoeld als toelevering voor toepassingen die (meestal met behulp van computers) aan de buitenwereld ten goede komen, wordt zelf doorgaans weer geheel op traditionele wijze zonder computers bedreven.

2. Wiskunde ontwikkelen met de computer

Echt computerwerk (dus afgezien van tekstverwerking) ten bate van de wiskunde zelf treedt wat minder in het daglicht. Soms levert de computer resultaten die direct worden gebruikt maar meestal produceert hij materiaal op grond waarvan vermoedens kunnen worden opgesteld die daarna met traditionele methoden worden bewezen. Soms doet de wiskundige dit met kant-en-klaar programmapakketten in hapklare brokken, soms zal de wiskundige eigen programmatuur ontwikkelen aan de hand van de vragen die in het onderzoek opkomen. Soms gaat het om belangrijke dingen die de voorpagina's van de kranten halen, meestal verdwijnt het werk in de prullemand, omdat het dezelfde rol heeft gespeeld als ons kladpapier.

Het klinkt allemaal nogal onsystematisch, en zo hoort het eigenlijk ook. We moeten niet vergeten dat veel wiskundige ontdekkingen tot stand zijn gekomen door te *spelen*. Zowel de combinatoriek als de waarschijnlijkheidsrekening zijn uit spelen voortgekomen! En voor vele onderwerpen is de computer een prettige speelkameraad.

Ik speel zelf al sinds 1962 met computers, niet alleen om spelletjes en puzzles te programmeren maar ook om materiaal te vergaren waaruit wiskundige ideeën kunnen voortkomen. Aan dat materiaal kunnen weer nieuwe vragen worden gesteld, totdat er eindelijk een structuur wordt ontdekt. Meestal denk ik dan: hoe kon ik zo stom zijn om het zonder computer niet direct te zien? Maar zo gaat dat nu eenmaal, ook als je gewoon op kladpapier werkt en zo maar wat zit te proberen.

Een computertoepassing die erg voor de hand ligt is de hulp bij het zoeken naar het asymptotisch gedrag van bijv. een functie. Numeriek werk leidt daar nooit tot iets definitiefs, maar heel vaak tot de goede vermoedens. Dat is belangrijk: een goed vermoeden kan ons vertellen welke reken- en bewijsmethoden we het beste kunnen toepassen.

De eis om een probleem goed te programmeren leidt heel vaak tot zuivere probleemstellingen. Ook ontwikkel je vaak wiskundig inzicht doordat je pro-

beert het programma efficient te maken om het ook in zware gevallen goed te laten werken. Zo was het succes dat ik in 1980 had met het ontdekken van een algebraische structuur achter de vlakvullingen met Penrose tegeltjes een gevolg van langdurige pogingen om er met behulp van een computer wat over te ontdekken. Achteraf gezien was het allemaal geklungel, maar de theorie kwam tot stand.

Een recenter voorbeeld. In 1987 deed V. Strehl een merkwaardige ontdekking over een soort halskettingen. Men neemt een natuurlijk getal n en bekijkt alle 2^n rijtjes van n symbolen 0 of 1. Twee van deze rijtjes worden equivalent genoemd als ze door cyclische verwisseling in elkaar overgaan, maar ook wanneer ze in elkaar overgaan als nullen in enen worden veranderd en omgekeerd, en dan eventueel nog cyclisch verwisseld. Zo zijn 1110100, 1001110 en 1000101 onderling equivalent, maar 1110100 is niet equivalent met 1001011. Het aantal equivalentieklassen wordt gegeven door de formule

$$\frac{1}{2n}\sum_{d|n}\phi(d)\ 2^{n/d}\ (d,2)$$

waarin ϕ de indicator van Euler is en $(d,2)$ de grootste gemene deler van d en 2. De sommatieindex d doorloopt de verzameling van alle delers van n.

In elke equivalentieklasse nemen we de kleinste representant (d.w.z. degene die gelezen als binair getal de kleinste is). In het gegeven voorbeeld is dat 0001011. In het geval $n = 7$ zijn er 10 equivalentieklassen. We ordenen de kleinste representanten naar opklimmende grootte:

0000001
0000011
0000101
0000111
0001001
0001011
0001101
0010011
0010101
1111111

Er is één uitzondering gemaakt. De geheel uit nullen bestaande rij is vervangen door de ermee equivalente geheel uit enen bestaande, en onderaan gezet.

De laatste kolom bestaat geheel uit enen, wat gemakkelijk te begrijpen is. Maar Strehl zag nog wat anders: in de voorlaatste kolom wisselen de nullen en enen elkaar precies af. En hij zag dat dit zo bleef bij hogere waarden van n, tot ongeveer 20. Ik probeerde het ook, en het lukte me (op een nu als voorwereldlijk te beschouwen PC) de waarheid vast te stellen tot en met $n = 25$. Het bleef kloppen, in al die millioenen gevallen, en de oorzaak was mysterieus.

Jaren later begon ik er nog eens aan, en na veel geploeter lukte het me te bewijzen dat die keurige afwisseling van nullen en enen voor *elke* n geldt. Tijdens de bewijspogingen kwamen steeds weer nieuwe vragen op, en de computer

verschafte daarbij waarnemingsmateriaal.

De kunst van het bedrijven van wiskunde met de hulp van een computer is niet zozeer de kunst om goede antwoorden te krijgen, maar de kunt om goede vragen te stellen.

3. Computers kunnen ons ook bedriegen

Soms is experimentele wiskunde bedriegelijk. Kijk eens naar de rij waarvan de eerste 320 elementen zijn:

0 0 2 4 2 6 8 6 2 4 6 6 8 4 0 0 8 2 8 4 2 4 8 2 0 0 0 4 8 6 2 6
6 2 0 0 4 6 8 2 4 6 8 4 0 0 2 2 2 4 8 2 8 0 4 6 6 4 2 6 8 4 4 2
6 8 6 6 0 0 0 8 8 2 4 8 2 0 0 8 6 4 6 8 2 4 0 4 2 6 8 4 6 8 0 0
0 0 2 4 2 6 8 6 2 4 6 6 8 4 0 0 8 2 8 4 2 4 8 2 0 0 0 4 8 6 2 6
6 2 0 0 4 6 8 2 4 6 8 4 0 0 2 2 2 4 8 2 8 0 4 6 6 4 2 6 8 4 4 2
6 8 6 6 0 0 0 8 8 2 4 8 2 0 0 8 6 4 6 8 2 4 0 4 2 6 8 4 6 8 0 0
0 0 2 4 2 6 8 6 2 4 6 6 8 4 0 0 8 2 8 4 2 4 8 2 0 0 0 4 8 6 2 6
6 2 0 0 4 6 8 2 4 6 8 4 0 0 2 2 2 4 8 2 8 0 4 6 6 4 2 6 8 4 4 2
6 8 6 6 0 0 0 8 8 2 4 8 2 0 0 8 6 4 6 8 2 4 0 4 2 6 8 4 6 8 0 0
0 0 2 4 2 6 8 6 2 4 6 6 8 4 0 0 8 2 8 4 2 4 8 2 0 0 0 4 8 6 2 6

De rij is verkregen door een eenvoudig procedé dat in één zin is uit te drukken. Kan iemand dat procedé raden? Het lijkt niet erg waarschijnlijk, want er is weinig anders op te merken dan dat er alleen maar cijfers 0, 2, 4, 6 en 8 in staan, met ongeveer dezelfde frequentie. Regelmatigheden zijn er niet te zien. Statistisch is er nog iets anders: als we naar de paren van twee opeenvolgende kijken, zien we dat bijv. 6 6 bijna twee keer zoveel voorkomt als 4 4. En ook over andere paren valt statistisch wat waar te nemen. Dat lijkt een sleutel te zijn, maar het helpt ons niet op weg. Toch is het recept heel simpel. Van de kwadraten van de opvolgende priemgetallen schrijft men (in het tientallig stelsel) het op één na laatste cijfer op! De kwadraten zijn 4, 9, 25, 49, 121, 169, 289, 361, 529, 841, ..., en de voorlaatste cijfers zijn 0, 0, 2, 4, 2, 6, 8, 6, 2, 4, ... Het is niet moeilijk te zien dat er alleen maar even cijfers komen. Er kan ook worden bewezen dat de cijfers 0, 2, 4, 6, 8 alle dezelfde frequentie hebben. Maar veel meer is er niet. Met voor de hand liggende (maar misschien niet gemakkelijk te bewijzen) vermoedens over priemgetallenstatistiek kan men aannemelijk maken dat gedurende heel lange tijd (milliarden keer) een paar als 6 6 veel vaker voorkomt dan 4 4, maar ook dat dit *niet* tot in het oneindige geldt. De statistiek die men aan een beginstuk waarneemt is bedriegelijk.

4. Meetkunde met computer

Wat jammer dat de belangstelling voor meetkunde zo is achteruitgegaan, want met onze moderne computer kan je er zulke leuke dingen mee doen. Misschien kan de computer helpen om de meetkunde weer een beetje terug te laten komen.

De computer kan tekenen, rekenen en redeneren, en die facetten met elkaar combineren. Het tekenen helpt het inzicht, zeker als het over meer dan twee dimensies gaat. Het is een betere en snellere versie van wat we op papier gewend waren, tenminste als we over de beginmoeilijkheden heen zijn. Er zijn schitterende programma's voor, en de toekomst zal nog veel meer te bieden hebben. Vooral voor het onderwijs is er veel van te verwachten. Misschien komt er ook weer wat belangstelling voor niet-euclidische meetkunde, want op dat gebied is onze eigen tekenvaardigheid problematisch terwijl het voor de computer niet zoveel verschil maakt.

Het rekenen ligt natuurlijk in het domein van de analytische meetkunde. Het kan op twee manieren gebeuren: door computeralgebra en door numerieke benaderingen. Computeralgebra heeft bewijskracht, numeriek rekenen kan wel met zeer grote, maar nooit met oneindige precisie worden gedaan. Het levert geen bewijzen op en kan dus worden vergeleken met het meetkundige tekenen. De computer doet het natuurlijk allebei tegelijk, maar waar het ons om gaat is de vraag of we alles in alle tekeningen met de ogen moeten bekijken dan wel dat we een aantal vragen formuleren en aan de computer de antwoorden laten vinden. Op de laatstgenoemde wijze kan men een groot aantal verschillende situaties doorrekenen en de computer laten melden of er wat interessants gevonden is.

Onlangs heb ik met driehoeksmeetkunde gespeeld. Bij elke driehoek kan men nieuwe driehoeken bouwen, zoals de voetpuntsdriehoek, de driehoek gevormd door de middens van de zijden, en nog zo wat. Stel de operaties die van de oorspronkelijke driehoek naar die nieuwe leiden voor door α, β, ... Nu kunnen we ook producten bekijken, zoals $\alpha\beta\alpha$. We laten een flink aantal van die producten op de oorspronkelijke driehoek los en vragen dan of er onder de ontstane driehoeken soms gelijke voorkomen, en of er soms paren zijn die perspectivisch liggen. Iets dergelijks was op grote schaal door Kimberling [2] gedaan voor bijzondere punten en lijnen in een driehoek. Ook hij werkte met nauwkeurige numerieke benaderingen in plaats van met computer algebra, omdat het veel meer vrijheid geeft.

Door zulk onderzoek ontdek je natuurlijk een aantal trivialiteiten, maar ook wel eens iets interessants, iets onverwachts. Wanneer er dan wat gevonden wordt dat de moeite waard is, kunnen altijd nog de tekeningen vertoond worden. Daarna bedenk je weer nieuwe vragen, enz.

De uitdaging bij al deze dingen is natuurlijk niet het vaststellen van waarheden, maar het leveren van leuke *opgaven* die we met de aloude meetkundige redeneringen willen oplossen. De gevonden waarheden zijn op zichzelf beschouwd erg onbelangrijk.

Het meetkundige *redeneren* staat natuurlijk in principe los van rekenen of tekenen. We zijn er nog niet echt goed aan toe om het met een computer te doen. Een van de redenen is dat de grondslagen van de meetkunde zo moeilijk in perfecte vorm te geven zijn. De belangrijkste vraag is hier niet of we het zouden kunnen maar of we het eigenlijk wel zouden willen. Ook hier geldt de opmerking dat ons doel met het leren van meetkunde ligt in de methodieken,

niet in de resultaten. Maar natuurlijk is meetkunde een prachtig gebied om het automatizeren van redeneringen te oefenen.

5. Logica

Op het gebied van de logica zijn er veel dingen waarbij computers behulpzaam kunnen zijn. Een van de belangrijkste dingen is de kunstmatige intelligentie, waarmee voor opgegeven uitspraken bewijzen worden gezocht. Maar met wat minder ambities kan men denken aan een *bewijsassistent*, die de wiskundige of de logicus helpt met het in elkaar zetten van bewijzen. De assistent controleert alles tot in de puntjes, verzorgt de boekhouding en vult alle kleine gaatjes die de wiskundige te flauw vindt om zich er druk over te maken. Soms is de bewijsassistent in staat knopen te ontwarren als de wiskundige er zelf niet uitkomt. Verschillende van deze systemen zijn nu al aardig op dreef en zullen in de toekomst bij de steeds verbeterende computertechnologie hoe langer hoe bruikbaarder worden, ook voor de gewone man. Om er een paar te noemen: Coq, Isabelle, Newprl, Lego. Een basis voor zulke systemen werd al gelegd in het Automath project dat een zekere bruikbaarheid toonde door een geheel wiskundeboek te verifiëren met behulp van de computertechnologie van omstreeks 1970, toen computers zeker wel duizend keer langzamer waren dan nu en veel minder geheugen hadden.

We zullen hier iets bespreken dat er een heel klein beetje op lijkt: *natuurlijke deductie* gepresenteerd met vlaggen en vlaggenstokken, beperkt tot propositiecalculus. Hoe eenvoudig ook, het geeft een inzicht in wat een bewijs is, in het bijzonder in de geneste structuur van het invoeren en dechargeren van onderstellingen. Ook wordt er er iets zichtbaar van de strategieën waarmee met tot een bewijs komt.

Het heeft ook wel met computers te maken. Er zijn computerprogramma's mogelijk die veel van de sommen oplossen die we aan leerlingen opgeven, en daarbij de verantwoording (de toegepaste regels) leveren die we ook van de leerlingen verwachten. Het leuke is dat zulke programma's ook hun beperkingen hebben, zodat er een eerlijke concurrentie is tussen mens en computer. Voor de niet-klassieke calculus ML$(\rightarrow, \wedge, \vee)$ die nog ter sprake zal komen is er geen programma bekend dat alles oplost wat er op te lossen valt.

Een redelijk programma werd in 1987 gemaakt door twee Eindhovense studenten, Paul Rambags en Maurice Schekkerman, in het kader van een stage onder leiding van R.P. Nederpelt. In een wat bijgeschaafde vorm kan het programma aan de deelnemers van deze cursus ter beschikking worden gesteld.

6. De minimale calculus

We zullen niet beginnen met een formele beschrijving van een systeem van natuurlijke deductie, maar proberen het een en ander duidelijk te maken aan de hand van een voorbeeld.

(1) $\boxed{((a \to b) \to b) \to a}$
(2) $\quad\boxed{b}$
(3) $\quad\quad\boxed{a \to b}$
(4) $\quad\quad b$
(5) $\quad(a \to b) \to b$
(6) $\quad a$
(7) $b \to a$
(8) $\quad\boxed{(a \to b) \to a}$
(9) $\quad\quad\boxed{a \to b}$
(10) $\quad\quad a$
(11) $\quad\quad b$
(12) $\quad(a \to b) \to b$
(13) $\quad a$
(14) $((a \to b) \to a) \to a$
(15) $(b \to a) \wedge (((a \to b) \to a) \to a)$
(16) $(((a \to b) \to b) \to a) \to ((b \to a) \wedge (((a \to b) \to a) \to a))$

Dit voorbeeld geeft de *afleiding* van de formule (16). Het begint in (1) met de *onderstelling* $((a \to b) \to b) \to a$. Met deze onderstelling wordt van alles en nog wat afgeleid, en het eindigt in (15) met het resultaat $(b \to a) \wedge ((a \to b) \to a) \to a$. Een van de afleidingsregels zegt: wanneer in een onderstelling p een resultaat q is afgeleid, dan beshouwen we dat als een afleiding van de *implicatie* $p \to q$ *zonder* die onderstelling p.

De *vlag* bij (1) geeft aan dat er een onderstelling is opgevoerd, en de *vlaggestok* laat zien gedurende welk deel van de tekst de onderstelling werkzaam is. In de stap van (15) naar (16) wordt de onderstelling (1) *gedechargeerd*. Het toepassen van zulke decharge, waarbij een implicatie is afgeleid, zullen we IN$\to$ noemen. Aan de tekst op een vlag wordt geen andere eis gesteld dan dat het syntactisch goed in elkaar zit, dat het leesbaar is.

Bij elke tekstregel hebben we wat we noemen een *context*. Het is het rijtje veronderstellingen dat daar geldig is. Zo is de context van (11) gevormd door de onderstellingen bij (1), (8) en (9), wat we kunnen zien door de vlaggen op te zoeken die worden gedragen door de stokken vooraan in (11). In (16) staan geen stokken; we zeggen dat (16) in de *lege* context staat.

De letters a, b, c stellen *proposities* voor. We noemen ze ook wel *propositievariabelen*. De hele afleiding kan worden beschouwd als een *blauwdruk*: wanneer we hem herhalen met vervanging (overal) van a, b, c door andere proposities, bijv. resp. door $p \to q$, $q \to p$, q, dan ontstaat weer een geldige afleiding.

Zulke substituties zijn trouwens wezenlijk wanneer we een stukje logica toepassen op de wiskunde. Dan kunnen a, b, c bijv. worden vervangen door resp. $x > 5$, $x + y = z$, $z < 0$. En het zou ook gemengd kunnen gebeuren, door zowel wiskundige proposities als propositievariabelen toe te laten.

De formules in de afleiding (1)-(16) zijn opgebouwd uit de a, b, c met

behulp van haakjes en *connectieven*: hier komen alleen $\rightarrow$ en $\wedge$ voor. Die connectieven zijn infix geschreven (d.w.z. dat bijv. in $a \rightarrow b$ de pijl tussen de beide argumenten a en b staat). Door geschikte afspraken kan er op het aantal haakjes worden bezuinigd, maar voor beginnelingen is het beter dat achterwege te laten.

We kijken nog even verder in de afleiding. In (4) mocht de geldigheid van b worden vastgesteld op grond van het feit dat die in een van de daar nog geldige onderstellingen staat: één van de stokken vooraan in (4) wappert met de vlag bij (2). Dit soort herhaling van een nog geldige onderstelling wordt later de regel HERHAAL genoemd.

In (5) wordt weer IN$\rightarrow$ toegepast: het resultaat (4) is gedechargeerd van de onderstelling (3). In (6) gebeurt wat nieuws: daar wordt de afleidingsregel *modus ponens* toegepast, die we ook wel EL $\rightarrow$ zullen noemen. Laten we even $(a \rightarrow b) \rightarrow b$ door p voorstellen. In de context van (6) zijn zowel $p \rightarrow a$ als p geldig, nl. wegens (1) en (5). Modus ponens zegt dat nu ook het rechterlid van de implicatie, nl. a, geldig is (nog steeds in diezelfde context).

Misschien niet ten overvloede nog even iets over het woord *geldig*, dat altijd wordt gebruikt relatief niet alleen ten opzichte van een bepaalde context, maar ook ten opzichte van een bepaalde *plaats*. Om dit aan een voorbeeld te laten zien: in (11) zijn alle eerder gemelde formules, voor zover het geen onderstellinegn zijn, geldig wanneer hun context dezelfde is als de context van (11) of althans een beginstuk ervan. Bovendien zijn de onderstellingen geldig die op vlaggen staan waarvan de stok bij (11) langskomt. In (11) zijn dus de formules (7),(10) en de onderstellingen (1), (8), (9) geldig. Bovendien worden in elke context dingen geldig verklaard op grond van de aangenomen afleidingsregels, zoals de reeds besproken IN$\rightarrow$ en EL$\rightarrow$. Tenslotte, wat ergens geldig is, mag worden opgeschreven. En wat eenmaal opgeschreven is, heeft invloed op wat er verderop weer geldig is.

We kijken nog even verder. In (7) is IN$\rightarrow$ toegepast ((6) wordt gedechargeerd van (2)). In (8) wordt (in de context (1)) een nieuwe onderstelling opgevoerd, evenzo in (9). In (10) is a geldig op grond van EL$\rightarrow$ (modus ponens), toegepast op de implicatie (8) en het linkerlid ervan dat in (9) staat; beide zijn hier geldig. In (11) is b geldig op grond van EL$\rightarrow$, toegepast op (9) en (10). In (12) is (11) gedechargeerd van (9), en (13) past modus ponens toe op (1) en (12), beide daar geldig. In (14) is (13) gedechargeerd van (8).

In (15) treedt een nieuw connectief $\wedge$ op en er wordt een afleidingsregel IN$\wedge$ toegepast, die zegt dat wanneer p en q ergens geldig zijn, er ook de *conjunctie* $p \wedge q$ geldt.

De afleiding eindigt met (16) waarin, zoals eerder gemeld, (15) is gedechargeerd van (1).

Bij eerste kennismaking lijkt het op toverij, maar de afleiding wordt niet opgebouwd op de manier waarop we die hier hebben gelezen. Het komt als volgt tot stand. We hebben ons een *doel* gesteld: formule (16). Alleen kunnen we het nummer er nog niet bijzetten, want we weten nog niet hoe lang de afleiding worden zal, maar de formulenummers worden hier toch maar even

gebruikt voor onze bespreking. Het doel (16) is een implicatie, en die valt te bewijzen door decharge. We schrijven dus (1) en (15) op, met vlag en al. Tussen (1) en (15) is een gat dat we nog moeten vullen. Nu is (15) ons enige doel. Deze (15) heeft de vorm van een conjunctie, en die kunnen we afleiden door de afzonderlijke delen af te leiden. We stellen ons dus, in plaats van (15), twee nieuwe subdoelen, (7) en (14), in dezelfde context als (15). Er zijn nu *twee* gaten te vullen. Wat het eerste gat betreft hebben we (7) af te leiden. Dat heeft de vorm van een implicatie, dus we gaan weer aan de gang met een onderstelling (2) en een nieuw subdoel (6). Deze (6) heeft niet de vorm van een implicatie of een conjunctie, dus we moeten wat anders bedenken. Boven (6) staan alleen nog pas (1) en (2). We zien dat de a voorkomt aan de rechterkant van de implicatie (1). Daarom gaan we de modus ponens proberen, waarbij we het linkerlid van de implicatie (1) als nieuw subdoel (5) gaan stellen. Dit is weer een implicatie, zodat we (3) en (4) opschrijven met de bijbehorende vlag. Nu is (4) ons nieuwe subdoel. Maar (4) is daar geldig op grond van (2) en het gat is gedicht.

Nu het andere gat. In (14) staat weer een implicatie, zodat we (8) en (13) opschrijven met bijbehorende vlag. In (13) staat a en is dus te vergelijken met (6), maar de context is anders. Bij (13) zijn geldig: (1), (7) en (8). Zowel (1) als (8) eindigen op a; met beide zouden we weer modus ponens kunnen proberen. Met (8) leidt het tot niets, omdat er nergens iets bruikbaars op b eindigt. We gaan dus weer met (1) aan de gang, (12) wordt ons nieuwe subdoel, dat laat ons (9) en (11) opschrijven. Ons nieuwe subdoel is nu wèl een rechterlid van een bruikbare implicatie, nl. (9). Daarom stellen we het linkerlid als nieuw subdoel (10). Nu is het gat gedicht, want (8) en (9) zijn hier beide geldig, zodat nog eens modus ponens kan worden toegepast.

Een formule die in de lege context is afgeleid heet een *tautologie*. Formule (16) is een voorbeeld ervan.

Er valt nog een simpele afleidingsregel te melden die nog niet aan de orde gekomen was. Het is de EL$\wedge$, die zegt dat wanneer ergens de conjunctie $p \wedge q$ geldt, ook de p en q daar geldig zijn.

In het volgende overzicht staan de afgesproken afleidingsregels met de daaraan gehechte namen (IN is een afkorting voor "introductie", en EL voor "eliminatie"):

$$\frac{\begin{array}{c}a\\ \hline b\end{array}}{a \to b}\ (\text{IN}\to) \qquad \frac{a \to b \quad a}{b}\ (\text{EL}\to)$$

$$\frac{a \quad b}{a \wedge b}\ (\text{IN}\wedge) \qquad \frac{a \wedge b}{a}\ (\text{EL1}\wedge) \qquad \frac{a \wedge b}{b}\ (\text{EL2}\wedge)$$

De hier ontwikkelde calculus zullen we ML($\to, \wedge$) noemen (ML staat voor "minimale logica": ongeveer het minste wat we kunnen doen).

Evenals in de wiskunde kunnen afleidingen worden verkort door hier en daar *hulpstellingen* af te leiden waar later een beroep op kan worden gedaan. Dat betekent eigenlijk dat een hulpstelling een blauwdruk is waarvan in andere afleidingen een kopie kan worden gemaakt, eventueel na aanpassing van de in de blauwdruk voorkomende variabelen door middel van substitutie.

Ook kunnen er nieuwe connectieven worden ingevoerd door middel van afkortingen . Gemakkelijk is bijv. de *equivalentie* waarvoor we het teken $\Leftrightarrow$ gebruiken. De definitie van $a \Leftrightarrow b$ is $(a \rightarrow b) \wedge (b \rightarrow a)$. Als hulpstelling kan men nu o.a. nemen de afleiding van $a \Leftrightarrow b$ uit de onderstelling $a \wedge b$.

Terloops zij hier iets vermeld waarbij het gebruik van een computer onmisbaar is. Laat $S(\rightarrow, \wedge, a, b, c)$ de (oneindige) collectie van alle formules zijn die gebouwd kunnen worden met de connectieven $\rightarrow$, $\wedge$ en de propositievariabelen a, b en c. Twee formules P en Q uit deze collectie heten equivalent als $P \Leftrightarrow Q$ in ML$(\rightarrow, \wedge)$ kan worden afgeleid. Het aantal equivalentieklassen is eindig, nl. 623662965552330. Er staat een computerprogramma ter beschikking waarmee van elke formule uit $S(\rightarrow, \wedge, a, b, c)$ snel (rekentijd evenredig met de lengte van de formule) de afleidbaarheid kan worden vastgesteld of verworpen. Dat programma werkt met een partieel geordende verzameling van slechts 61 elementen, en de equivalentieklassen corresponderen met zekere deelverzamelingen daarvan.

Overigens kan worden vermeld dat, wanneer we maar twee letters a, b toelaten in plaats van drie, er nog maar 18 equivalentieklassen zijn (en in plaats van 61 punten krijgen we er 5), en als we maar één letter toelaten zijn er maar twee equivalentieklassen: elke formule uit $S(\rightarrow, \wedge, a)$ is òf equivalent met a òf met $a \rightarrow a$.

Wat hier gezegd werd is *metatheorie*: het bestaat niet uit afleidingen *in* de calculus ML$(\rightarrow, \wedge)$ maar uit wiskundige beschouwingen *over* die calculus.

7. Negatie

We gaan de calculus ML$(\rightarrow, \wedge)$ uitbreiden door bij iedere propositie p ook de *negatie* ervan toe te laten, en die met $\neg p$ te noteren.

Eerst een waarschuwing. Als we zeggen dat ergens p niet geldig is, dan is dat wat anders dan dat $\neg p$ geldig is. Het eerste betekent dat p niet kan worden afgeleid, het tweede dat $\neg p$ wèl kan worden afgeleid. Sinds Gödel weten we dat er in de wiskunde proposities p zijn waarvoor noch p noch $\neg p$ kan worden afgeleid.

We voeren de negatie in door een propositie*constante* F toe te laten die *falsum* of *contradictie* wordt genoemd. Vervolgens wordt $\neg p$ gedefinieerd als de implicatie $p \rightarrow F$ ("p leidt tot een contradictie").

Erg veel kan men met deze negatie nog niet doen. Men kan bijv. wel de tautologie $(a \rightarrow b) \rightarrow ((\neg b) \rightarrow (\neg a))$ afleiden maar niet omgekeerd $((\neg b) \rightarrow (\neg a)) \rightarrow (a \rightarrow b)$ (de onafleidbaarheid is na te gaan met die methode der 61 punten).

Evenzo kunnen we nagaan dat $\neg\neg a$ uit a kan worden afgeleid, maar niet omgekeerd. Hier zijn de afleidingen voor de tautologieën $(a \rightarrow b) \rightarrow ((\neg b) \rightarrow (\neg a))$ en $a \rightarrow (\neg\neg a)$:

(1)	$\boxed{a \to b}$	
(2)	$\quad\boxed{\neg b}$	
(3)	$\quad\quad\boxed{a}$	
(4)	$\quad\quad b$	EL $\to$ (1, 3)
(5)	$\quad\quad F$	EL $\to$ (2, 4)
(6)	$\quad\neg a$	IN $\to$ (5)
(7)	$(\neg b) \to (\neg a)$	IN $\to$ (6)
(8)	$(a \to b) \to ((\neg b) \to (\neg a))$	IN $\to$ (7)
(9)	$\boxed{a}$	
(10)	$\quad\boxed{\neg a}$	
(11)	$\quad F$	EL $\to$ (10, 9)
(12)	$\quad\neg\neg a$	IN $\to$ (11)
(13)	$a \to (\neg\neg a)$	IN $\to$ (12)

Nog een ander voorbeeldje van een redenering waarin negaties optreden. Er blijkt het volgende uit. Waneer we een *negatieve* propositie p hebben die zowel uit a als uit $\neg a$ kan worden afgeleid, dan is p ook zonder onderstellingen geldig. Dit kan worden beschouwd als een zwakke vorm van de (hierna te bespreken) regel van de uitgesloten derde.

(1)	$\boxed{a \to (\neg c)}$	
(2)	$\quad\boxed{(\neg a) \to (\neg c)}$	
(3)	$\quad\quad\boxed{c}$	
(4)	$\quad\quad\quad\boxed{a}$	
(5)	$\quad\quad\quad\neg c$	EL $\to$ (1, 4)
(6)	$\quad\quad\quad F$	EL $\to$ (5, 3)
(7)	$\quad\quad\neg a$	IN $\to$ (6)
(8)	$\quad\quad\neg c$	EL $\to$ (2, 7)
(9)	$\quad\quad F$	EL $\to$ (8, 3)
(10)	$\quad\neg c$	IN $\to$ (9)

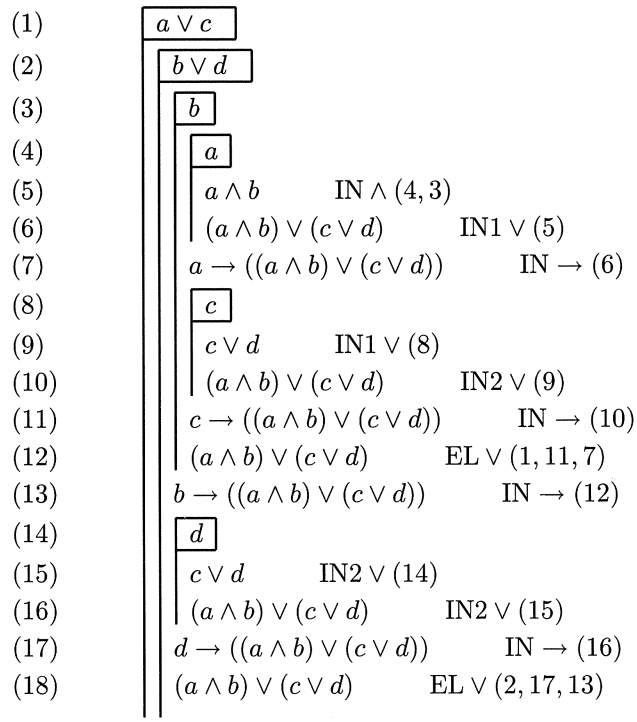
8. DISJUNCTIE

We keren terug naar de mininale logica ML($\to, \wedge$) en gaan een nieuw connectief $\vee$ toevoegen: $a \vee b$ (uitgesproken als "a of b") heet de *disjunctie* van a en b. De regels voor introductie en eliminatie zijn

$$\frac{a}{a \vee b}\ (\text{IN1}\vee) \qquad \frac{b}{a \vee b}\ (\text{IN2}\vee) \qquad \frac{a \vee b \quad a \to c \quad b \to c}{c}\ (\text{EL}\vee)$$

Het toevoegen van deze disjunctie maakt het niet eenvoudiger! Eerder werd vermeld dat ML($\to, \wedge$) in de collectie $S(\to, \wedge, a, b)$ 18 equivalentieklassen heeft,

maar bij ML($\rightarrow, \wedge, \vee$) heeft de collectie $S(\rightarrow, \wedge, \vee, a, b)$ er oneindig veel. Computerprogramma's hebben het met het vinden van afleidingen in ML($\rightarrow, \wedge, \vee$) ook aanzienlijk moeilijker dan in ML($\rightarrow, \wedge$). Hier is een voorbeeld van een afleiding in ML($\rightarrow, \wedge, \vee$):



9. Klassieke logica

We formuleren twee nieuwe afleidingsregels die ons spel versterken. Als eerste de *falsumregel*:

$$\frac{F}{a} \quad \text{(FALSR)}$$

die zegt dat uit een contradictie alles volgt wat men maar zou willen. Als tweede de regel voor de uitgesloten derde:

$$a \vee (\neg a) \quad \text{(UITGD)}.$$

Evenals bij de andere afleidingsregels is het weer toegestaan de variabele a door een willekeurige propositie te vervangen. Uit deze twee regels samen kan de regel voor de dubbele negatie worden afgeleid:

$$\frac{\neg\neg a}{a} \quad \text{(DBNG)}.$$

Hier is een afleiding:

(1)	$\boxed{\neg\neg a}$	
(2)	$a \vee (\neg a)$	UITGD
(3)	$\boxed{a}$	
(4)	a	HERHAAL (3)
(5)	$a \to a$	IN $\to$ (4)
(6)	$\boxed{\neg a}$	
(7)	F	EL $\to$ (1, 6)
(8)	a	FALSR (7)
(9)	$(\neg a) \to a$	IN $\to$ (8)
(10)	a	EL $\vee$ (2, 5, 9)

De falsumregel wordt door de intuitionisten nog wèl aangenomen, de regel van uitgesloten derde niet. Die is, samen met de falsumregel, de basis voor de *klassieke logica.*

Overigens kan de falsumregel onmiddellijk uit DBNG worden afgeleid:

(1)	$((a \to F) \to F) \to a$	DBNG
(2)	$\boxed{F}$	
(3)	$\boxed{a \to F}$	
(4)	F	HERHAAL (2)
(5)	$(a \to F) \to F$	IN $\to$ (4)
(6)	a	EL $\to$ (1, 5)

De regel van de dubbele negatie maakt *indirect bewijzen* mogelijk. Dat gaat als volgt. We willen a bewijzen, nemen $\neg a$ aan, en leiden een contradictie af. Met IN$\to$ komen we tot $\neg\neg a$ en met DBNG tot a.

We hebben nu de gehele klassieke propositielogica in handen. Als voorbeeld bewijzen we de tautologie $((\neg b) \to (\neg a)) \to (a \to b)$ (waarvan het omgekeerde al zonder DBNG was aangetoond).

(1)	$\boxed{(\neg b) \to (\neg a)}$	
(2)	$\boxed{a}$	
(3)	$\boxed{\neg b}$	
(4)	$\neg a$	EL $\to$ (1, 3)
(5)	F	EL $\to$ (4, 2)
(6)	$(\neg b) \to F$	IN $\to$ (5)
(7)	b	DBNG (6)
(8)	$a \to b$	IN $\to$ (7)

Met klassieke logica leiden we ook de formule van Peirce af. Het is $((a \to b) \to a) \to a$, bevat geen spoor van falsum of negatie, maar is toch in de calculus

ML($\rightarrow, \wedge$) niet afleidbaar (zoals valt aan te tonen met de eerder genoemde methode der 61 punten). Met klassieke logica gaat het wèl:

(1) $\boxed{(a \rightarrow b) \rightarrow a}$
(2) $\quad\boxed{\neg a}$
(3) $\quad\quad\boxed{a}$
(4) $\quad\quad F \qquad \text{EL} \rightarrow (2,3)$
(5) $\quad\quad b \quad \text{FALSR}(4)$
(6) $\quad a \rightarrow b \qquad \text{IN} \rightarrow (5)$
(7) $\quad a \qquad \text{IN} \rightarrow (1,6)$
(8) $\quad F \qquad \text{EL} \rightarrow (2,7)$
(9) $(\neg a) \rightarrow F \qquad \text{IN} \rightarrow (8)$
(10) $a \qquad \text{DBNG}\ (9)$
(11) $((a \rightarrow b) \rightarrow a) \rightarrow a \qquad \text{IN} \rightarrow (10)$

In de klassieke logica geldt als tautologie $(a \vee b) \Leftrightarrow ((\neg a) \rightarrow b)$. Eerst bewijzen we de implicatie van links naar rechts:

(1) $\boxed{a \vee b}$
(2) $\quad\boxed{\neg a}$
(3) $\quad\quad\boxed{a}$
(4) $\quad\quad F \qquad \text{EL} \rightarrow (2,3)$
(5) $\quad\quad b \quad \text{FALSR}(4)$
(6) $\quad a \rightarrow b \qquad \text{IN} \rightarrow (5)$
(7) $\quad\quad\boxed{b}$
(8) $\quad\quad b \qquad \text{HERHAAL}\ (7)$
(9) $\quad b \rightarrow b \qquad \text{IN} \rightarrow (8)$
(10) $\quad b \qquad \text{EL} \vee (1,6,9)$
(11) $(\neg a) \rightarrow b \qquad \text{IN} \rightarrow (10)$

En dit is de afleiding van de implicatie van rechts naar links:

(1) $\boxed{(\neg a) \rightarrow b}$
(2) $\quad\boxed{\neg(a \vee b)}$
(3) $\quad\quad\boxed{a}$
(4) $\quad\quad a \vee b \qquad \text{IN1} \vee (3)$
(5) $\quad\quad F \qquad \text{EL} \rightarrow (2,4)$
(6) $\quad \neg a \qquad \text{IN} \rightarrow (5)$
(7) $\quad b \qquad \text{EL} \rightarrow (2,6)$

(8)	$a \vee b$	IN2 $\vee$ (7)
(9)	F	EL $\rightarrow$ (2, 8)
(10)	$\neg\neg(a \vee b)$	IN $\rightarrow$ (9)
(11)	$a \vee b$	DBNG (10)

Dit betekent dat we ons in de klassieke logica kunnen permitteren de disjunctie $a \vee b$ te *definiëren*, bijv. door $(\neg a) \rightarrow b$, maar het zou ook kunnen door $\neg((\neg a) \wedge (\neg b))$. De regels voor introductie en eliminatie kunnen dan als hulpstelling worden afgeleid.

De metatheorie van de klassieke propositielogica is eenvoudig. Van iedere propositie kan de afleidbaarheid worden vastgesteld dan wel verworpen met de *waarderingsmethode*. Als de formule n verschillende propositievariabelen $a_1, \ldots, a_n$ telt zijn er 2^n manieren om aan elk der letters één der waarden 0 of 1 toe te voegen. Aan de eventueel voorkomende F wordt steeds de waarde 0 toegekend. Bij elk van de 2^n mogelijkheden wordt vervolgens met de zo bekende waarheidstafels de waarde van de gehele formule bepaald. Als in al die gevallen die waarde op 1 uitkomt, is de formule in de klassieke logica afleidbaar, en anders niet.

10. GEBRUIK VAN SYMMETRIE

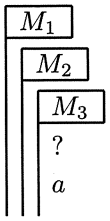
Uit de letters a, b en c vormen we de proposities

$$(a \Leftrightarrow b) \rightarrow (a \wedge (b \wedge c))$$

$$(b \Leftrightarrow c) \rightarrow (b \wedge (c \wedge a))$$

$$(c \Leftrightarrow a) \rightarrow (c \wedge (a \wedge b))$$

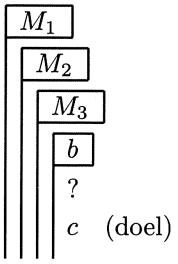
die we respectievelijk door M_1,M_2,M_3 voorstellen. Gevraagd wordt om, wanneer M_1,M_2,M_3 ondersteld zijn, de geldigheid van a, b en c af te leiden. Wegens de symmetrie is het voldoende alleen a af te leiden. We stellen ons dus de opgave



en willen het door het vraagteken aangeduide gat vullen.

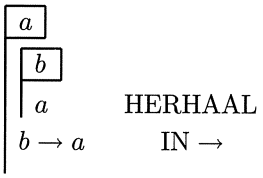
Om tot a te besluiten kunnen we met één van de drie equivalenties $a \Leftrightarrow b$, $b \Leftrightarrow c$, $c \Leftrightarrow a$ volstaan (en ons dan beroepen op de bijpassende uit M_1,M_2,M_3). Wegens de symmetrie doet het er niet toe welke we nemen. We kiezen als doel $b \Leftrightarrow c$. Dit is $(b \rightarrow c) \wedge (c \rightarrow b)$. Volgens de regel IN$\wedge$ kunnen we ons nu als doel stellen om zowel $b \rightarrow c$ als $c \rightarrow b$ te bewijzen. Als het bewijs voor

de eerste gelukt is, krijgen we dat voor de tweede eenvoudig door herhaling met lettterverwisseling. We behoeven dus alleen $b \rightarrow c$ als doel te stellen. Dit kunnen we bewijzen door b aan te nemen, daaruit c te bewijzen, en dan een beroep te doen op de regel IN$\rightarrow$. We moeten dus het gat nog vullen in de volgende redenering:

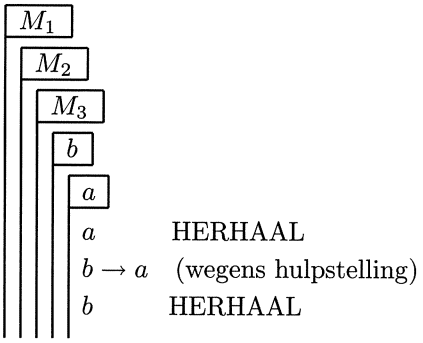


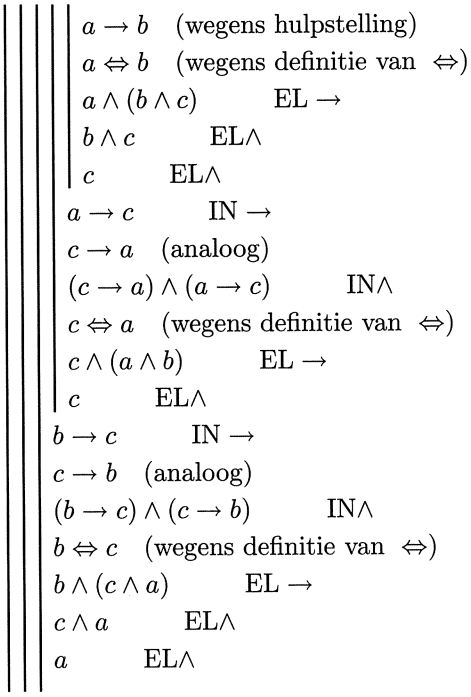
Hoe kunnen we tot c komen? Natuurlijk door één van de M_1,M_2,M_3 te gebruiken en ons op de regel EL$\rightarrow$ te beroepen. We moeten daartoe in deze context een van de drie equivalenties $a \Leftrightarrow b$, $b \Leftrightarrow c$, $c \Leftrightarrow a$ afleiden. Het aantrekkelijkste is $c \Leftrightarrow a$, want dat vraagt om een bewijs van $c \rightarrow a$ en $a \rightarrow c$, en wegens de symmetrie die op deze plaats nog tussen a en c heerst, hoeven we ons maar over één van de twee zorgen te maken, want de andere is nagenoeg hetzelfde.

We gaan nu op de plaats van het vraagteken een afleiding van $a \rightarrow c$ maken. Dat doen we door a te onderstellen en c af te leiden. Die c is weer te bereiken door één van de $a \Leftrightarrow b$, $b \Leftrightarrow c$, $c \Leftrightarrow a$ af te leiden. Daar op deze plaats nu zowel a als b ter beschikking staan, bereiken we $a \Leftrightarrow b$ het gemakkelijkst. Laten we even een *hulpstelling* formuleren: uit a volgt $b \rightarrow a$. Bewijs als volgt:



Nu vullen we het gat:





De lezers kunnen hun krachten beproeven op een analoog geval waarbij ook weer het gebruik van negaties verboden is: men leide onder de onderstellingen $a \vee b$, $b \vee c$, $c \vee a$ af dat $(a \wedge b) \vee ((b \wedge c) \vee (c \wedge a))$.

11. Slotbeschouwing

Er is veel geklaagd dat de schoolmeetkunde niet meer functioneert als leerschool voor het logische denken en voor het leren geven van waterdichte bewijzen. Dat komt niet alleen doordat er minder tijd voor de meetkunde wordt ingeruimd of doordat de sociale en de onderwijsorganisatorische omgeving minder geschikt zijn dan vroeger. Het ligt ook aan de meetkunde zèlf. De basis van de schoolmeetkunde is te slap om er op een eerlijke manier streng op verder te bouwen. Wanneer men die basis op orde brengt, zoals Hilbert deed omstreeks 1900, wordt het voor behandeling op school te moeilijk. In de meetkunde leeft strengheid nu eenmaal op gespannen voet met meetkundige inzichtelijkheid.

De hier voorgestelde behandeling van natuurlijke deductie zou de logisch opvoedende taak van de meetkunde heel goed kunnen overnemen, met minder moeite en met groter duidelijkheid. Redeneerschema's leert men spelenderwijs door ze helder bloot te leggen. Ook is er een spelelement: we krijgen puzzeltjes voor het kiezen van geschikte strategie. Voorts ligt er opvoedende waarde in het feit dat er volledige verantwoording kan worden verlangd van de bij de oplossing van een probleem toegepaste regels. Aantrekkelijk is nog dat vele leerlingen in staat zullen zijn opgaven voor zichzelf te bedenken, en zelf hulpstellingen in te voeren die verder werk kunnen versnellen. Allemaal facetten die wijzen op geschiktheid voor het onderwijs. Bovendien zijn er relaties met computers.

Een hardnekkige traditie praat ons aan dat logica een spel is van nullen en enen, en dat de waarheidstafels de grondslag van ons denken zijn. Deze traditie heeft veel schade berokkend. Wiskundigen hebben zich in het algemeen van de logica afgekeerd. Ze vonden het een abstract spel dat wel in staat is om snel tautologieën vast te stellen (tenminste wanneer het aantal propositievariabelen niet te groot is), maar weinig laat zien van het wiskundige redeneren. Waar wordt trouwens in de wiskunde een beroep op zulke tautologieën gedaan?

De natuurlijke deductie, zeker in de hier gepresenteerde stijl met de vlaggen, sluit wèl aan bij het echte redeneren. Dat blijkt ook duidelijk wanneer aan de propositicalculus de *predikatencalculus* wordt toegevoegd. Er komen dan wat regels bij voor introductie en eliminatie van kwantoren, maar de opzet blijft dezelfde. We kunnen zo de gehele wiskunde in één schema onderbrengen met twee verschillende soorten vlaggen: rechthoekige vlaggen voor het invoeren van onderstellingen, puntige vlaggen voor het opvoeren van (getypeerde variabelen).

In Einhoven zijn gedurende een kleine 20 jaar gunstige ervaringen opgedaan met het onderwijzen van logica aan eerstejaarsstudenten in de wiskunde en de informatica volgens dit systeem. Verwezen kan worden naar Eindhovense collegesyllabi, en (in een groter verband) naar Nederpelt's boek [1].

LITERATUUR

1. R.P. NEDERPELT, *De taal van de wiskunde*, Versluys, Almere, 1987.
2. CLARK KIMBERLING, Central points and central lines in the plane of a triangle. *Mathematics Magazine* **67** (1994), 163–187.

CWI *Vakantiecursus 2000*

pp. 121 – 141

Elliptische krommen en cryptografie

M.J. Coster
W.W.J. Hulsbergen
Ministerie van Defensie

Deel I: Grondbeginselen

1. Inleiding

1.1. Public Key Cryptografie (PKC)

Tot 1976 bestond alle cryptografie uit Secret Key Cryptografie (SKC). Bij SKC bezitten de zender en ontvanger dezelfde sleutel. Bij Public Key Cryptografie (ook wel asymmetrische cryptografie genoemd) daarentegen bezitten zender en ontvanger verschillende sleutels. In het algemeen bezit de zender (**A**ns) de publieke sleutel, terwijl de ontvanger (**B**en) de geheime sleutel bezit. In principe kan iedereen die de publieke sleutel bezit **B**en een boodschap sturen.

Het bekendste PKC–systeem is RSA, genoemd naar de auteurs Rivest, Shamir en Adleman [7] . Bij RSA kiest **B**en twee (verschillende) priemgetallen p en q en een random getal d zó dat $\gcd(d,(p-1)(q-1)) = 1$. **B**en berekent $n = pq$ (de **modulus**) en e zodanig dat $de \equiv 1 \bmod (p-1)(q-1)$. Vervolgens publiceert hij de getallen e, n en houdt hij p, q en d geheim. Als **A**ns een boodschap M, herschreven als getal m $(0 < m < n)$, wil vercijferen, dan berekent zij $c \equiv m^e \bmod n$. In plaats van m verstuurt ze de versleutelde boodschap c. Uit c kan **B**en dan de oorspronkelijke boodschap m berekenen, want $c^d \equiv (m^e)^d \equiv m^{de} \equiv m \bmod n$. (Dit kan worden geverifiëerd door de bovenstaande congruenties modulo p en modulo q na te gaan.) Indien **C**hantal het bericht c onderschept en hieruit de oorspronkelijke boodschap m wil bepalen dan zal haar dat niet meezitten. Als zij de geheime sleutel (p,q,d) niet kent, dan is er tot op heden nog geen methode bekend om m te berekenen in 'beperkte tijd'. Het beste wat **C**hantal kan doen, is te trachten n te factoriseren. In 1999 hebben medewerkers van het CWI voor het eerst een 512–bits RSA–getal gefactoriseerd. Als n echter een 1024–bits RSA–getal is (of groter), dan wordt dit momenteel gezien als onmogelijk.

Een ander veel gebruikt PKC–systeem is de **discrete logaritme** (DL). In 1984 kwam T. ElGamal met een protocol dat hierop is gebaseerd. Dit protocol won snel populariteit. Het werkt als volgt: **B**en kiest een priem p (de

[7] Voor een heldere uiteenzetting over RSA verwijzen we naar Pythagoras 37-ste jaargang, **nr. 5**, juni 1998.

modulus), een voortbrenger g van de groep $(\mathbb{Z}/p\mathbb{Z})^\times$, en een geheime sleutel $k_{\mathbf{B}}$. Hij berekent $B \equiv g^{k_{\mathbf{B}}} \bmod p$, publiceert (p, g, B) en houdt $k_{\mathbf{B}}$ geheim. Als **A**ns een boodschap m, $0 < m < p$, wil vercijferen voor **B**en, dan bepaalt zij een random getal k en berekent $C = mB^k$ en $K = g^k$, en verstuurt deze twee getallen. **B**en berekent $K^{k_{\mathbf{B}}}$ en aangezien $K^{k_{\mathbf{B}}} \equiv B^k \equiv g^{k_{\mathbf{B}}k} \bmod p$ volgt $m \equiv C(K^{k_{\mathbf{B}}})^{-1} \bmod p$. Op dit moment lijkt het ondoenlijk voor **C**hantal, als zij p, g, B, K en C kent, om daaruit de boodschap m te distilleren, en zeker als men ervoor zorgt dat $p - 1$ niet in 'kleine' priemfactoren uiteenvalt, maar een grote priemfactor p_1 bevat met $p_1 > 10^{100}$, dan is het praktisch onmogelijk voor haar om m te bepalen. Men veronderstelt dat (ongeveer) gelijke moduli bij RSA en bij de discrete logaritme gelijke veiligheid garanderen.

In 1985 stelden V. Miller en, onafhankelijk van hem, N. Koblitz een cryptografisch protocol (ECC, Elliptic Curve Cryptography) voor dat op een equivalent van de discrete logaritme, die op elliptische krommen (gedefinieerd over een voldoende groot eindig lichaam), is gebaseerd. Elliptische krommen zijn al van oudsher zeer populair in de zuivere wiskunde vanwege hun intrigerende eigenschappen, zowel in de analyse als in de getallentheorie. Een voordeel van het gebruik van elliptische krommen is dat er vele voorhanden zijn bij een gegeven grondlichaam (ook al zijn ze niet allemaal geschikt), en nog belangrijker voor cryptografische doeleinden, men kan met (veel) kleinere lichamen werken dan bij de gewone discrete logaritme om dezelfde veiligheid te bewerkstelligen. Dit leidt tot eenvoudiger implementatie en grotere efficiency. Dit is ook een voordeel t.o.v. RSA. Om een idee te krijgen van de grootte van de modulus bij RSA en van het grondlichaam bij ECC waarbij een vergelijkbare veiligheid wordt geboden volgen hier enige waarden (in bits):

- RSA/DL: 512 1024 2048 4096
- ECC: 128 174 234 313.

Er zijn in de afgelopen 25 jaar diverse PKC–systemen ontwikkeld. Sommige daarvan zijn inmiddels ook weer gekraakt. Echter RSA, de Discrete Logaritme en de Discrete Logaritme voor Elliptische Krommen (ECDL) zijn tot op heden sterke cryptografische systemen gebleken.

In deze voordracht zal aandacht worden besteed aan elliptische krommen en hun toepassing in de cryptografie aan de hand van het ElGamal protocol. In Hoofdstuk 2 worden de begrippen abelse groepen en eindige lichamen in herinnering geroepen. Deze begrippen spelen een belangrijke rol in de theorie (en praktijk) van elliptische krommen. In Hoofdstuk 3 zetten we een en ander over elliptische krommen uiteen. In Hoofdstuk 4 gaan we kort in op de discrete logaritme. In hoofdstuk 5 bespreken we twee typen van elliptische krommen die ongeschikt blijken voor cryptografische doeleinden. We beëindigen de lezing met enkele voorbeelden. Het manuscript is verdeeld in twee delen, waarvan het eerste de nodige begrippen in de simpelste gevallen introduceert. In het tweede deel wordt iets dieper ingegaan op wiskundige aspecten die op de achtergrond een belangrijke rol spelen bij de toepassing van elliptische krommen in de cryptografie. Hieruit moge de indruk naar voren komen dat er in de hedendaagse cryptografie nog een belangrijke rol voor mensenwerk in de vorm

van zuivere wiskunde, aangevuld door listige uitbuiting van de rekencapaciteit van de computer, is weggelegd.

2. Abelse groepen en lichamen

2.1. Abelse groepen

Omdat de begrippen (abelse) groep en lichaam aan de basis staan van PKC brengen we hier de definities nog eens in herinnering.

Definitie. Een **abelse groep** G is een (niet lege) verzameling met een operatie $\star : G \times G \to G$, zó dat

1. $(a \star b) \star c = a \star (b \star c)$, $\forall a, b, c \in G$ (**associativiteit** van $\star$);
2. $\exists\, e \in G$, het **neutrale element**, zó dat $e \star a = a \star e = a$, $\forall a \in G$;
3. $\forall a \in G$, $\exists\, a^{-1} \in G$, de **inverse** van a, zó dat $a \star a^{-1} = a^{-1} \star a = e$;
4. $a \star b = b \star a$, $\forall a, b \in G$ (**commutativiteit** van $\star$).

Men schrijft $(G, \star)$ voor G als men de operatie $\star$ expliciet wil vermelden. Als het aantal elementen in de groep G eindig is, heet dit aantal de **orde** van de groep G, notatie: $|G|$ of $\#(G)$. In een eindige groep $(G, \star)$ is er voor elk element $g \in G$ een kleinste natuurlijk getal n zó dat $g \star g \star \cdots \star g = e$ (n keer g). Deze n heet de **orde** van g, en n is een deler van $\#(G)$. Als elk element van een group $(G, \star)$ eenduidig kan worden geschreven als macht van a voor zekere $a \in G$, dan heet G **cyclisch** met voortbrenger a.

Voorbeelden. *(i)* $(\mathbb{Z}, +)$, $(\mathbb{Q}, +)$, $(\mathbb{R}, +)$, $(\mathbb{C}, +)$ met de gewone optelling $+$ en het neutrale element $e = 0$ in de vier gevallen, hier is $\mathbb{Z}$ cyclisch met voortbrenger 1 (of -1).

(ii) $(\mathbb{Q}^\times, \cdot)$, $(\mathbb{R}^\times, \cdot)$, $(\mathbb{C}^\times, \cdot)$ met de gewone vermenigvuldiging $\cdot$ (meestal niet genoteerd) en het neutrale element $e = 1$ in de drie gevallen.

(iii) $(\mathbb{Z}/m\mathbb{Z}, +)$ (ook genoteerd $\mathbb{Z}/m$), de gehele getallen modulo m met de optelling $+$ mod m, met $\#(\mathbb{Z}/m\mathbb{Z}) = m$. Bijvoorbeeld, voor $m = 12$ heeft men $\mathbb{Z}/12\mathbb{Z} = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11\}$ en $3 + 5 = 8$, $5 + 8 = 1$, etc.

(iv) $((\mathbb{Z}/m\mathbb{Z})^\times, \cdot)$, de gehele getallen mod$m$ die onderling priem zijn met m, met vermenigvuldiging (veelal niet genoteerd) $\cdot$ mod m, bv. $(\mathbb{Z}/12\mathbb{Z})^\times = \{1, 5, 7, 11\}$ met $5 \cdot 11 = 7$, etc. Hier geldt: $\#((\mathbb{Z}/m\mathbb{Z})^\times) = \varphi(m)$, de totiëntfunktie van Euler, bv. $\varphi(12) = 4$. De elementen 5, 7 en 11 hebben orde 2.

Er geldt de volgende

Hoofdstelling voor eindige abelse groepen. *Laat G een eindige abelse groep zijn. Dan is G te schrijven als*

$$G = \mathbb{Z}/p_1^{s_1} \oplus \mathbb{Z}/p_2^{s_2} \oplus \cdots \oplus \mathbb{Z}/p_k^{s_k}, \quad \textit{met } p_1 \leq p_2 \leq \cdots \leq p_k,$$

waarin de priemen p_i's (en hun machten s_i) eenduidig bepaald zijn.

2.2. Commutatieve lichamen

DEFINITIE. Een commutatief **lichaam** k is een (niet lege) verzameling met twee operaties $+ : k \times k \to k$ en $\cdot : k \times k \to k$, zó dat k een abelse groep (met neutraal element genoteerd 0) is onder de 'optelling' $+$, en zó dat $k^\times \times k^\times \to k^\times$, met $k^\times := k \backslash \{0\}$, een abelse groep is (met neutraal element genoteerd $1 \in k^\times$) onder de 'vermenigvuldiging' $\cdot$. De operaties $+$ en $\cdot$ voldoen aan $(a+b)\cdot c = a\cdot c + b\cdot c$. De inverse van a ten aanzien van $+$ wordt genoteerd als $-a$, en die ten aanzien van $\cdot$ als a^{-1}, soms $1/a$.

Kort gezegd: in een groep kan men optellen en aftrekken, in een lichaam kan men optellen, aftrekken, vermenigvuldigen en delen (behalve door 0).

OEFENING. Toon aan dat in een lichaam k geldt: (i) $0 \cdot a = 0, \forall a \in k$;
(ii) $-1 \cdot a = -a, \forall a \in k$.

VOORBEELDEN. (i) $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$ met $+$ en $\cdot$ als boven.
(ii) Voor een priemgetal p, $\mathbb{Z}/p\mathbb{Z}$ met $+$ en $\cdot$ modulo p. Dit lichaam wordt meestal geschreven als $GF(p)$ of $\mathbb{F}_p$. Het heeft p elementen. De multiplicatieve groep $\mathbb{F}_p^\times$ is cyclisch van orde $p-1$. Een openstaand probleem is het bepalen van een voortbrenger.
(iii) In Deel II zullen we ook voorbeelden tegenkomen van andere eindige lichamen $\mathbb{F}_q$, waarbij $q = p^n$ met $n \geq 2$.

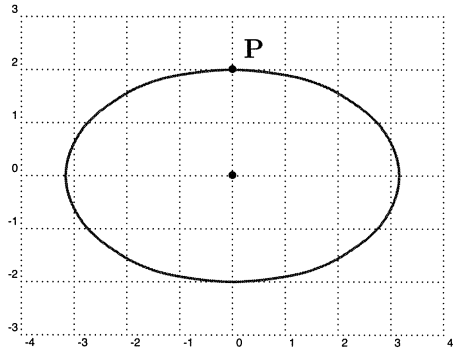
OEFENINGEN. (i) Bewijs de 'kleine stelling' van Fermat: Voor elk positief natuurlijk getal a en elk priemgetal p geldt: als $p \not| a$, dan $a^{p-1} \equiv 1 \bmod p$.
(ii) Bewijs dat geldt: $(a+b)^p = a^p + b^p$, $\forall a, b \in \mathbb{F}_p$.

3. ELLIPTISCHE KROMMEN

3.1. Weierstraßvorm voor elliptische krommen

We beginnen met een eenvoudig voorbeeld van een reële algebraïsche kromme die geheel op het papier past, anders gezegd, waarvan de grafiek elk reëel punt te zien geeft:

de *ellips* $\mathcal{E} : 2x^2 + 5y^2 = 20$

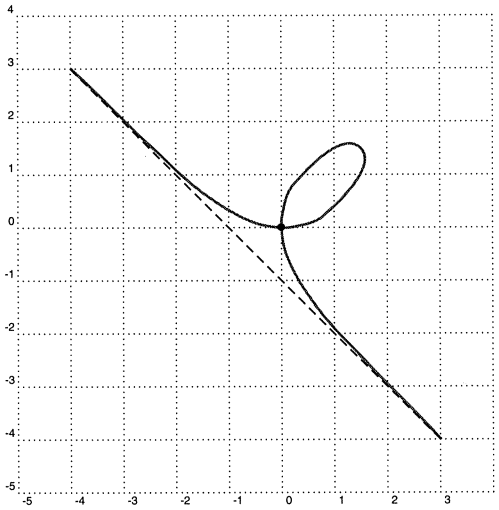


De ellips is een voorbeeld van een kegelsnede, evenals de parabool en de hyperbool. In tegenstelling tot de ellips verdwijnen de parabool en de hyperbool

'ergens in het oneindige'. Kegelsneden worden gegeven door tweedegraadsvergelijkingen. Kegelsneden zijn sinds de Grieken vanuit velerlei gezichtspunten zeer uitvoerig bestudeerd.

Als tweede voorbeeld bekijken we het *folium van Descartes* gegeven door de vergelijking:

$$\mathcal{F} : x^3 + y^3 - 3xy = 0$$



Hier doen zich onmiddellijk twee vragen voor: (i) wat gebeurt er in oneindig? (ii) hoe zit het in de buurt van de oorsprong $(0,0)$? Het antwoord op vraag (i) is dat men eigenlijk naar het projectieve model van de kromme moet kijken, d.w.z. men voert homogene coördinaten $(x : y : z)$ in, en werkt modulo de equivalentierelatie

$$(x : y : z) \sim (\lambda x : \lambda y : \lambda z),$$

$\lambda \neq 0$. Men maakt de vergelijking van de kromme homogeen:

$$\mathcal{F} : x^3 + y^3 - 3xyz = 0.$$

De vertrouwde affiene punten corresponderen dan met z-waarden $\neq 0$. De punten 'in het oneindige' zijn van de vorm $(x : y : 0)$, waarbij x en y niet beide 0 zijn. Hier dus alleen $(-1 : 1 : 0)$.

Het antwoord op vraag (ii) is dat de oorsprong een singulier punt, een zgn. dubbelpunt, is van de kromme. Over singuliere punten kan nog veel gezegd worden, maar we volstaan met het geven van een criterium voor het singulier zijn van een punt $P = (x_0 : y_0 : z_0)$ (projektief!) op een kromme gegeven door de homogene vergelijking $F(x, y, z) = 0$:

$$F(P) = F(x_0, y_0, z_0) = 0 \quad \wedge \quad \frac{\partial F}{\partial x}(P) = \frac{\partial F}{\partial y}(P) = \frac{\partial F}{\partial z}(P) = 0.$$

OEFENING. Wat zijn de punten 'in het oneindige' van de ellips, de parabool $2x - 5y^2 = 20$, de hyperbool $2x^2 - 5y^2 = 20$?

Na al deze voorbereidingen kunnen we nu een nette definitie van een elliptische kromme geven.

DEFINITIE. *Een elliptische kromme E, gedefinieerd over een willekeurig lichaam k, is een niet-singuliere vlakke projektieve derdegraadskromme over k met een k-rationaal punt O (d.w.z. met coördinaten in k) op E.*

Zo'n kromme kan worden gegeven door de zgn. **Weierstraßvorm**, in homogene coördinaten x, y, z,

$$\boxed{E : y^2z + a_1xyz + a_3yz^2 = x^3 + a_2x^2z + a_4xz^2 + a_6z^3}$$

met $a_1, \cdots, a_6 \in k$, zodanig dat de **discriminant** $\Delta \neq 0$. Deze discriminant [8] is een polynomiale uitdrukking in de coëfficiënten $a_1, \ldots, a_6$. De voorwaarde $\Delta \neq 0$ is nodig en voldoende opdat E niet-singulier is. De kromme E heeft precies één k-rationaal punt in 'oneindig', d.w.z. als $z = 0$, nl. het punt $(0 : 1 : 0)$. Dit punt speelt de rol van O. Wil men expliciet aangeven dat E gedefinieerd is over het lichaam k, d.w.z. gegeven wordt door een vergelijking met coëfficiënten in k, dan schrijft men veelal $E/k : y^2 + \cdots = x^3 + \cdots$ etc. Men kan zich beperken tot het affiene deel ($z \neq 0$) van E:

$$E : y^2 + a_1xy + a_3y = x^3 + a_2x^2 + a_4x + a_6.$$

Voor lichamen k zoals $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$ of $\mathbb{F}_p$ met $p > 3$ kan men de Weierstraß gedaante voor de affiene kromme door de coördinatentransformatie

$$X = x + \frac{a_1^2 + 4a_2}{12} \quad \text{en} \quad Y = y + \frac{a_1}{2}x + \frac{a_3}{2}$$

omvormen tot een kromme van de vorm

$$\boxed{E/k : y^2 = x^3 + Ax + B}$$

met $A, B \in k$ zodanig dat de discriminant $\Delta = 4A^3 + 27B^2 \neq 0$. Deze gedaante zal in het vervolg steeds gebruikt worden.

OPMERKING. Voor praktische toepassingen zijn eindige lichamen van de vorm $k = GF(2^n) = \mathbb{F}_{2^n}$ van groot belang. Voor elliptische krommen over zulke lichamen dient men de theorie zoals we die hier bespreken, aan te passen.

3.2. Punten op elliptische krommen

Notatie. Als k het grondlichaam is waarover de kromme E is gedefinieerd, en K is een lichaam dat k bevat [9] , dan noteren we de verzameling van punten van E met coördinaten in K als $E(K)$. Aan de hand van een voorbeeld laten we zien dat het lichaam K het aanzien van de kromme bepaalt. In het geval van een kromme die gedefinieerd is door een vergelijking met rationale coëfficiënten

[8] Vgl. de discriminant $\Delta = b^2 - 4ac$ van de kwadratische uitdrukking $ax^2 + bx + c$.
[9] Voor meer over lichaamsuitbreidingen zie Deel II

is dit het duidelijkst. Voor krommen over eindige lichamen moet men toch ook aan iets dergelijks denken.

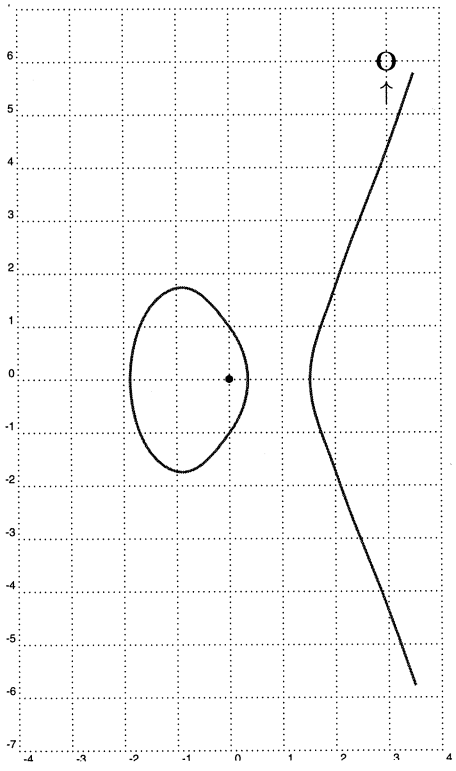
VOORBEELD. Bekijk de kromme

$$E/\mathbb{Q} : y^2 = x^3 - 3x + 1$$

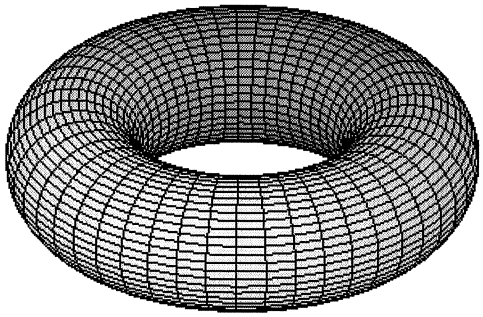
• $E(\mathbb{Q})$ wordt gegeven door de verzameling van $\mathbb{Q}$-waardige oplossingen van de **diophantische vergelijking** $y^2 = x^3 - 3x + 1$, bv. $(0, \pm 1)$, $(\frac{9}{4}, \pm\frac{19}{8})$, en een iets minder triviaal paar punten:

$$\left(-\frac{926527914284812196336}{499758792156170515809}, \pm\frac{4864488622791473479562279219161}{1117225050627356845186949580737 7}\right), \text{etc.}$$

• Daarnaast wordt $E(\mathbb{R})$ gegeven door de volgende figuur:



• Voor de complexe punten $E(\mathbb{C})$ vindt men iets dat geïnterpreteerd kan worden als een torus in $\mathbb{R}^3$:



Hoe dit precies zit wordt in leerboeken over complexe analyse waarin de $\wp$-functie van Weierstraß wordt behandeld, uit de doeken gedaan.

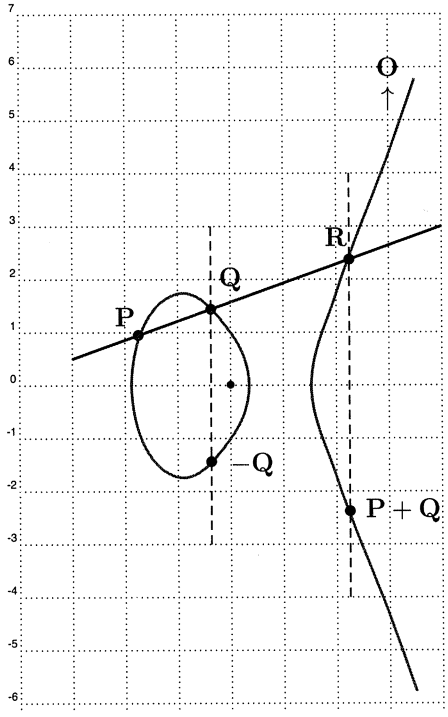
• Beschouw nu de kromme $E/\mathbb{F}_p : y^2 = x^3 - 3x + 1$, dus over een eindig lichaam (met $p > 3$), bv. neem $p = 11$. Men kan zich afvragen wat nu de verzameling van punten $E(\mathbb{F}_{11})$ is. Door achtereenvolgens $x = 0, 1, \ldots, 10$ in de vergelijking van E in te vullen kan men uitproberen of bij een x-waarde een y (dan ook $-y$) voldoet. In dit voorbeeld vindt men de volgende punten: (0,1), (0,10), (2,5), (2,6), (4,3), (4,8), (5,1), (5,10), (6,1), (6,10), (7,2), (7,9), (8,4), (8,7), (10,5), (10,6). Voor punten op E met coördinaten in een groter lichaam van de vorm $\mathbb{F}_{11^n}$ verwijzen we naar Deel II. Dat het uitproberen door invullen niet de geëigende methode is om de punten op E, of zelfs maar het *aantal* $\mathbb{F}_p$-punten op E, zelfs m.b.v. computers, te bepalen als p groter [10] wordt, is aanzet geweest tot veel onderzoek naar betere methodes. De laatste jaren zijn enige spectaculaire resultaten bereikt voor de berekening van het aantal $\mathbb{F}_p$-punten op een elliptische kromme over $\mathbb{F}_p$. De recente methodes berusten op werk van de Nederlandse wiskundige René Schoof uit 1984 en verfijningen van A.O.L. Atkin en N. Elkies uit de jaren 90 van de vorige eeuw. Geavanceerde methodes uit de arithmetische algebraïsche meetkunde (modulaire vormen) worden hierbij niet geschuwd. Dit heeft geleid tot het zgn. **SEA-algoritme** (**S**choof-**E**lkies-**A**tkin) voor de berekening van het aantal punten op een elliptische kromme over $\mathbb{F}_p$ voor grote waarden van p. Het record ligt bij een priem p van 500 cijfers, een resultaat dat in $\pm$9000 uur rekentijd op een netwerk van DEC-alpha's door Francois Morain e.a. aan de École Polytechnique te Parijs werd behaald in januari 1995.
Het is gebruikelijk, als men spreekt over het aantal punten op een elliptische kromme, ook het punt $O = (0 : 1 : 0)$ van het projektieve model van de kromme mee te nemen. Dus in bovenstaand voorbeeld heeft men $\#(E(\mathbb{F}_{11})) = 17$ (en niet 16). Voor het aantal punten $E(\mathbb{F}_p)$ heeft men de ongelijkheden van **Hasse-Weil**:

[10] Voor cryptografische toepassingen moet p minstens 160 bits lang zijn.

M.a.w. men kan schrijven $\#(E(\mathbb{F}_p)) = p + 1 - t$, waarbij $|t| \leq 2\sqrt{p}$. Het SEA-algoritme is er nu op gericht deze t te bepalen.

3.3. De groepstruktuur

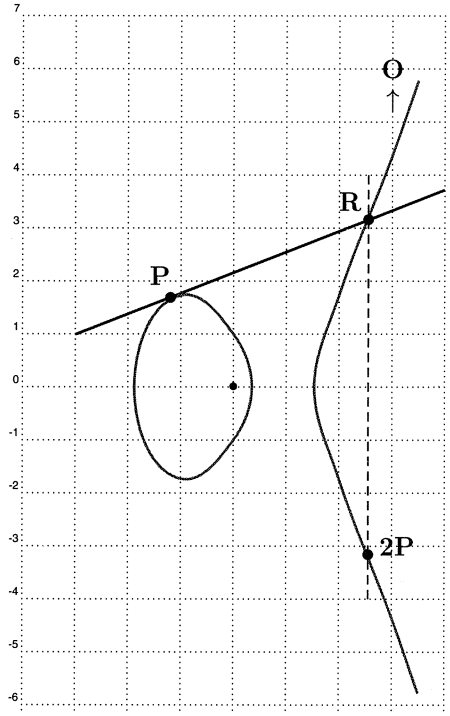
Wat maakt elliptische krommen tot zulke interessante objecten van studie in de algebraïsche meetkunde en tevens tot mogelijke gebruiksvoorwerpen in de cryptografie? Welnu, als de kromme E gedefinieerd is over een lichaam k, dan vormen de punten $E(k)$ een **abelse groep** [11] waarbij de optelling van twee punten P en Q op de kromme meetkundig het duidelijkst wordt voorgesteld aan de hand van de figuur van $E(\mathbb{R})$ (als we even aannemen dat $k = \mathbb{R}$):



Om het punt $2P = P + P$ te construeren neemt men de raaklijn in P aan de

[11] Dit is ook waar voor de punten $E(K)$ in een uitbreiding K van k.

kromme, enz.:



Om bovenstaande constructie in formulevorm om te zetten definieert men

$$\lambda = \frac{y_Q - y_P}{x_Q - x_P} \text{ als } P \neq Q, \quad \text{en} \quad \lambda = \frac{3x_P^2 + A}{2y_P} \text{ als } P = Q,$$

m.a.w. λ is de richtingscoëfficiënt van de lijn door P en Q, resp. van de raaklijn in P mits $y_P \neq 0$. Als $y_P = 0$ dan geldt: $2P = O$. Een eenvoudige berekening geeft nu (uiteraard met de geschikte λ's):

$$\boxed{x_{P+Q} = \lambda^2 - (x_P + x_Q)} \quad \text{en} \quad \boxed{y_{P+Q} = -y_Q - \lambda(x_{P+Q} - x_Q)}$$

en

$$\boxed{x_{2P} = \lambda^2 - 2x_p} \quad \text{en} \quad \boxed{y_{2P} = -\lambda^3 - y_P + 3\lambda x_P}$$

OEFENING. Ga deze formules na.

Men ziet aan deze formules dat ze niet meer bewerkingen bevatten dan optelling, aftrekking, vermenigvuldiging en deling in het lichaam waarin de coördinaten van P en Q zitten, dus de punten $P + Q$ en $2P$ hebben ook weer coördinaten in dat lichaam. Men gaat na dat de optelling van twee punten aan de axioma's van een abelse groep voldoet [12] met als neutraal element het punt in het oneindige $O = (0 : 1 : 0)$. Voor de inverse $-P$ van het punt $P = (x, y)$ vindt men

[12] De associativiteit is het moeilijkst te bewijzen.

eenvoudig: $-P = (x, -y)$. De som van drie punten op de kromme die op een lijn liggen is juist O. De formules zijn geldig in een willekeurig lichaam, i.h.b. in $\mathbb{F}_p$ en eventuele uitbreidingen.

OEFENING. (*i*) Bereken $2 \cdot (0,1)$ op $E(\mathbb{Q})$ voor $E/\mathbb{Q} : y^2 = x^3 - 3x + 1$.

(*ii*) Bereken $2 \cdot (0,1)$ op $E(\mathbb{F}_{11})$ voor $E/\mathbb{F}_{11} : y^2 = x^3 - 3x + 1$.

(*iii*) (Voor de onversaagden!) Zelfde vragen voor $9 \cdot (0,1)$. Hint: $9 = 2 \cdot 2 \cdot 2 + 1$.

VOORBEELD. Het feit dat $\#(E(\mathbb{F}_{11})) = 17$ voor bovenstaande kromme $E/\mathbb{F}_{11}$ betekent dat elk punt ($\neq O$) orde 17 heeft en dat $E(\mathbb{F}_{11})$ cyclisch is. Dit soort informatie is van nut voor cryptografische toepassingen.

4. CRYPTOGRAFIE

4.1. Inbedding van tekst op de kromme

Met het idee van de gelijkmatige verdeling van rationale punten op een elliptische kromme in het achterhoofd, heeft N. Koblitz een eenvoudige manier voorgesteld om tekst van een gekozen alfabet om te zetten in punten op een (geschikte) elliptische kromme, en omgekeerd, om punten van de kromme terug te vertalen naar leesbare tekst. Om hierbij moeilijkheden te voorkomen kiest men van te voren een voldoende groot natuurlijk getal κ, bv. $\kappa = 20$ of $30, \ldots$. Definieer een alfabet $\mathcal{A}$ bestaande uit de gewenste letters, cijfers en symbolen ter grootte $M = |\mathcal{A}|$. We geven de letters, cijfers en symbolen van $\mathcal{A}$ weer door natuurlijke getallen [13] m, $0 \leq m < M$. Zorg er voor dat $M\kappa < q$, waarin q de orde van het eindige lichaam $\mathbb{F}_q$ is waarover onze elliptische kromme gedefinieerd is. Voor het gemak nemen we steeds voor q een priemgetal p, maar omwille van de algemeenheid bespreken we de methode zoals Koblitz voorstelt. Dan is elk element μ van het alfabet te schrijven als $m\kappa + j$, $0 \leq j < \kappa$, waarbij voor gegeven m, j achtereenvolgens $0, 1, 2, \ldots, \kappa - 1$ genomen wordt totdat men een punt $(x, y) = (m\kappa + j, y) \in E(\mathbb{F}_q)$ te pakken heeft. Het voorschrift $\mu \mapsto P_\mu := (m\kappa + j, y) \in E(\mathbb{F}_q)$ aldus beschreven, geeft een inbedding van het alfabet in de $\mathbb{F}_q$-punten van E. Een boodschap van s letters, cijfers en symbolen van het alfabet wordt nu omgezet in een geordende reeks van s $\mathbb{F}_q$-rationale punten op E. De j dient ervoor om een oplossing van de vergelijking $y^2 = x^3 + Ax + B$ in $\mathbb{F}_q$ te vinden. Voor een willekeurige j-waarde zal de kans op geluk ongeveer $\frac{1}{2}$ bedragen, dus om κ mislukkingen te moeten incasseren, zal de kans $2^{-\kappa}$ bedragen. De grootte van κ bepaalt aldus het risico op mislukking.

VOORBEELD. We nemen het alfabet bestaande uit de spatie ␣ , de hoofdletters A, B, C, ... , Z, en ! (en eventueel nog wat symbolen). We geven deze de respectieve numerieke waarden ␣=0, A=1, B=2 , ... , Z=26, !=27, etc. Neem $\kappa = 30$ en het grondlichaam $\mathbb{F}_p$ met $p = 997$. Voor de elliptische kromme E kiezen we

$$E/\mathbb{F}_{997} : y^2 = x^3 + 2x + 13$$

Met bovenstaande afspraken wil **A**ns de boodschap

[13] In de praktijk gebruikt men blokken van letters, cijfers en symbolen van lengte n, heden ten dage n minstens 20, en $M = |\mathcal{A}|^n$.

'DAG BEN!' (d.w.z. DAG␣BEN!)

als een reeks punten P_D, P_A, P_G, $P_␣$, P_B, P_E, P_N, $P_!$ op E aan **B**en sturen. Dus, bv. $D = 4$, en de bijbehorende x-coördinaat van P_D is $4\kappa+j = 4*30+j = 120+j$ met $j = 0, 1, 2, \ldots, \kappa - 1$ totdat de vergelijking

$$y^2 = (120+j)^3 + 2*(120+j) + 13 \pmod{997}$$

een oplossing (dan ook twee) heeft. Het blijkt dat $j = 1$ al goed is voor de oplossing $y = 407$, dus $P_D = (121, 407) \in E(\mathbb{F}_{997})$, etc. Uiteindelijk vindt men de volgende serie punten:$(121, 407)$, $(32, 464)$, $(210, 558)$, $(0, 451)$, $(62, 440)$, $(150, 285)$, $(424, 327)$, $(811, 438)$.

OEFENING. Ga dit na.

Het antwoord van **B**en is: $(241, 110)$, $(451, 144)$, $(274, 401)$, $(0, 546)$, $(32, 464)$, $(424, 327)$, $(572, 220)$, $(811, 559)$.

OEFENING. Wat zegt dit antwoord?

4.2. De discrete logaritme (ECDL)

In een (abelse) groep G (multiplicatief geschreven) kan men de vergelijking $y = x^n$, $x, y \in G$, $n \in \mathbb{Z}$, bekijken. Als x en n zijn gegeven kan men doorgaans eenvoudig y berekenen. Omgekeerd, als x en y zijn gegeven, is het veelal veel moeilijker om n te achterhalen. Naar analogie met de vertrouwde reële logaritme noemt men hier n de **discrete logaritme** van y bij het grondtal x.

VOORBEELDEN. *(i)* $G = (\mathbb{F}_p, +)$, de discrete logaritme komt neer op het inverteren van n in de vergelijking $y = nx$, hetgeen triviaal is;
(ii) $G = (\mathbb{F}_p^\times, \cdot)$, dit is de discrete logaritme zoals deze in de 'klassieke' ElGamal voorkomt: gegeven $x, y \in \mathbb{F}_p$, zó dat $y = x^n$ voor zekere $n \in \mathbb{Z}$, vind n;
(iii) (ECDLP) $G = E(\mathbb{F}_p)$, $P, Q \in E(\mathbb{F}_p)$, zó dat $Q = nP$ voor zekere $n \in \mathbb{Z}$, vind n. Dit lijkt tot op heden een moeilijk probleem te zijn. Eind 1998 is het gelukt om het ECDL-probleem succesvol aan te vallen voor een grondlichaam $\mathbb{F}_p$ waarbij p 98 bits lang is.

4.3. Het protocol van ElGamal

Als voorbeeld van een cryptografisch protocol bekijken we het Public Key encryptie schema zoals dat door T. ElGamal in 1984 is voorgesteld, weliswaar nog voordat het idee om elliptische krommen in PKC te gebruiken, was gelanceerd, maar direkt te vertolken in termen van het ECDL-probleem. We nemen weer het eenvoudige voorbeeld van **A**ns die **B**en een boodschap wil sturen, maar nu versleuteld volgens het ElGamal protocol. Daartoe nemen we maar weer de elliptische kromme (over een priemlichaam voor het gemak)

$$E/\mathbb{F}_{997} : y^2 = x^3 + 2x + 13.$$

Tevens nemen we een *basispunt* $B \in E(\mathbb{F}_{997})$, bv. $B = (100, 333)$. Het punt B is openbaar. Men vindt $\#(E(\mathbb{F}_p)) = 1051$, maar we hebben dit hier niet

expliciet nodig. **A**ns en **B**en kiezen ieder een **geheime** sleutel $k_{\mathbf{A}} \in \mathbb{N}$, resp. $k_{\mathbf{B}} \in \mathbb{N}$. Ze maken evenwel de punten $k_{\mathbf{A}}B$, resp. $k_{\mathbf{B}}B$ bekend. Als **A**ns weer de boodschap 'DAG BEN!', maar nu versleuteld, aan **B**en wil sturen, neemt ze voor elk te versleutelen symbool een willekeurig getal $k \in \mathbb{N}$ dat ze geheim houdt, berekent $C = P_\mu + k(k_{\mathbf{B}}B)$, en stuurt de serie puntenparen

$$\boxed{(kB, C),\ \mu = D, \ldots, N}$$

aan **B**en. Deze laatste ontcijfert zo'n puntenpaar via

$$\boxed{C - k_{\mathbf{B}}(kB) = P_\mu + k(k_{\mathbf{B}}B) - k_{\mathbf{B}}(kB) = P_\mu}$$

VOORBEELD. Laten de geheime sleutels $k_{\mathbf{A}} = 13$ en $k_{\mathbf{B}} = 150$ zijn, en neem aan dat **A**ns eerst $k = 73$ heeft gekozen. Men vindt: $k_{\mathbf{B}}B = 150 \cdot (100, 333) = (252, 598)$, $kB = 73 \cdot (100, 333) = (24, 202)$ en $k(k_{\mathbf{B}}B) = 73 \cdot (252, 598) = (203, 824)$. Het eerste versleutelde puntenpaar wordt: $((24, 202), (121, 407) + (203, 824)) = ((24, 202), (38, 392))$. Als **A**ns vervolgens voor k de waarden 11, 65, 136, 84, 765, 23, 673 neemt, zal **B**en de volgende reeks puntenparen ontvangen [14] : $((87, 872), (720, 628))$, $((514, 642), (687, 194))$, $((435, 976), (54, 230))$, $((558, 6), (912, 846))$, $((553, 615), (944, 885))$, $((990, 211), (203, 824))$, $((818, 481), (568, 637))$.

OEFENING. Ga dit na en ontcijfer vervolgens de boodschap.

5. WELKE ELLIPTISCHE KROMMEN WEL EN WELKE NIET?

Zoals uit het bovenstaande valt af te lezen, valt of staat het ElGamal protocol (evenals andere, die we hier niet bespreken) met de oplossing van het ECDLP. Men kent de punten B, kB, $k_{\mathbf{A}}B$ en $k_{\mathbf{B}}B$, maar niet k, $k_{\mathbf{A}}$ en $k_{\mathbf{B}}$. Wil het ECDLP onhanteerbaar zijn, dan zal men er moeite voor doen om te zorgen dat de cyclische groep voortgebracht door B groot is, anders gezegd, $E(\mathbb{F}_q)$ moet een grote cyclische ondergroep bevatten, of zelf cyclisch zijn. Dit maakt het wenselijk $\#(E(\mathbb{F}_q))$ te kennen. Maar kennis van $\#(E(\mathbb{F}_q))$ is slechts een eerste stap. De vraag naar geschikte krommen zal in de loop van de tijd mogelijk een veranderend antwoord krijgen. In ieder geval zijn er op dit moment twee [15] klassen van elliptische krommen die voor cryptografische toepassingen niet geschikt zijn: de supersinguliere en de abnormale elliptische krommen.

5.1. Supersinguliere elliptische krommen

Een elliptische kromme $E/\mathbb{F}_p$ heet **supersingulier** als $\#(E(\mathbb{F}_p)) = p + 1$. In 1993 publiceerden A. Menezes, T. Okamoto en S. Vanstone een verrassende constructie om het ECDLP te reduceren tot het DLP in een uitbreiding van $\mathbb{F}_p$. Deze constructie berust op geavanceerde algebraïsche meetkunde voor elliptische krommen. Voor supersinguliere krommen blijkt dat de graad van de

[14] Volledige kennis van de y-coördinaat is overbodig. Het is voldoende het 'teken' van y te weten. Dit levert een voordeel voor de snelheid van het protocol.

[15] Voor een derde zie Deel II.

uitbreiding hoogstens 6 is. Het ECDLP zou opgelost kunnen worden in subexponentiële tijd. Dit betekent dat men onevenredig grote lichamen moet hanteren om voldoende veiligheid te garanderen bij gebruik van deze krommen in de cryptografie, en dit gaat dan weer ten koste van de snelheid.

5.2. Abnormale elliptische krommen

Men heeft enige tijd geloofd dat elliptische krommen $E/\mathbb{F}_p$ die juist p $\mathbb{F}_p$-punten hebben, zeer geschikt zijn voor cryptografische doeleinden. Aan dit idee kwam aan het eind van de vorige eeuw een einde toen uit werk van G. Frey & H.-G. Rück, van I. Semaev, van T. Satoh & K. Araki en van N. Smart bleek dat men eenvoudig een isomorphisme tussen $E(\mathbb{F}_p)$ en de additieve groep van $\mathbb{F}_p$ kan aangeven. In deze laatste groep is het discrete logaritme probleem gemakkelijk op te lossen. Elliptische krommen $E/\mathbb{F}_p$ met $\#(E(\mathbb{F}_p)) = p$ heten **abnormaal**. Voor gegeven p zijn deze krommen (gelukkig) dungezaaid.

Deel II: Meer wiskundige achtergronden

6. Lichaamsuitbreidingen en Frobenius

6.1. Lichaamsuitbreidingen

Men kan een gegeven lichaam k uitbreiden tot een groter lichaam door er op geschikte manier 'nieuwe elementen aan toe te voegen'. Zo is de verzameling van getallen van de vorm $a+b\sqrt{5}$ met $a, b \in \mathbb{Q}$ en $\sqrt{5}$ oplossing van de vergelijking $x^2 = 5$ (dus $\sqrt{5} \cdot \sqrt{5} = 5 \in \mathbb{Q}$), met de bekende optelling en vermenigvuldiging, een lichaamsuitbreiding van $\mathbb{Q}$, genoteerd $\mathbb{Q}(\sqrt{5})$. Evenzo zijn $\mathbb{R}$ en $\mathbb{C}$ lichaamsuitbreidingen van $\mathbb{Q}$, zij het van andere aard. Men zegt dat lichaamsuitbreidingen van $\mathbb{Q}$ **karakteristiek nul** hebben. Men kan ook uitbreidingen van $\mathbb{F}_p$ maken, bv. door toevoeging van wortels van geschikte veeltermen met coëfficiënten in $\mathbb{F}_p$. Ingeval dit leidt tot nieuwe eindige lichamen, dan hebben deze laatste altijd een macht (> 1) van p als aantal elementen. Traditiegetrouw wordt een eindig lichaam als $\mathbb{F}_q$ genoteerd, waarbij $q = p^m$ voor zekere $m \geq 1$.

Voorbeeld. Neem als grondlichaam $\mathbb{F}_{11} = \{0, 1, \ldots, 10\}$ en het polynoom $F(x) = x^2+x+1$. Door 'formeel' [16] een wortel α van de vergelijking $x^2+x+1 = 0$ aan $\mathbb{F}_{11}$ toe te voegen, en de verzameling van elementen [17] van de vorm $a + b\alpha$, $a, b \in \mathbb{F}_{11}$, met optelling $(a + b\alpha) + (c + d\alpha) := (a + c) + (b + d)\alpha$ en vermenigvuldiging $(a+b\alpha)(c+d\alpha) := ac+(ad+bc)\alpha+bd\alpha^2 = ac+(ad+bc)\alpha+bd(-1-\alpha) = (ac-bd)+(ad+bc-bd)\alpha$, te bekijken, en op voor de hand liggende wijze inversen te definiëren, heeft men een uitbreiding van graad twee van $\mathbb{F}_{11}$ gemaakt. Men noteert deze uitbreiding als $\mathbb{F}_{11^2}$. Deze hangt dus van de keuze van het polynoom F af. Neemt men evenwel een ander tweedegraadspolynoom dat evenmin wortels in $\mathbb{F}_{11}$ heeft, dan is de aldus gecreëerde uitbreiding van $\mathbb{F}_{11}$ isomorf met de voorgaande $\mathbb{F}_{11^2}$.

Oefeningen. (*i*) Toon dit laatste aan.
(*ii*) Wat is de multiplicatieve inverse van $3 + 7\alpha$ in $\mathbb{F}_{11^2}$?
(*iii*) Maak een uitbreiding van graad drie van $\mathbb{F}_{11}$, enz.

Men zegt dat lichamen die uitbreidingen van $\mathbb{F}_p$ zijn, **karakteristiek p** hebben. $\mathbb{Q}$, resp. $\mathbb{F}_p$ zelf zijn op te vatten als de 'kleinste' lichamen van karakteristiek nul, resp. p. Door de wortels van alle polynomen met coëfficiënten in een lichaam k aan k toe te voegen krijgt men de **algebraïsche afsluiting** $\bar{k}$ van k, i.h.b. $\bar{\mathbb{F}}_p$. Dit is een lichaam met oneindig veel elementen.

6.2. Het Frobeniusendomorfisme

We keren terug naar de elliptische krommen, bv. $E/\mathbb{F}_{11} : y^2 = x^3 - 3x + 1$ en proberen de punten van $E(\mathbb{F}_{11^2})$ te bepalen. Men vindt bv. O, $(5 + 2\alpha, 4 + \alpha)$, $(5 + 2\alpha, 7 + 10\alpha)$, $(7 + 5\alpha, 1 + 10\alpha)$, $(7 + 5\alpha, 1 + 10\alpha)$, ..., en nog 114 andere.

[16] Voor insiders: men bekijkt het quotiëntlichaam van de ring $\mathbb{F}_{11}[x]/(x^2 + x + 1)$.
[17] Vergelijk de constructie van complexe getallen uitgaande van de reële getallen.

In het algemeen is het ondoenlijk de punten van $E(\mathbb{F}_q)$ te bepalen. Voor cryptografische toepassingen is dit ook niet nodig. Men is evenwel geïnteresseerd in het *aantal* punten met coördinaten in een uitbreidingslichaam. Het blijkt dat men dit aantal kan bepalen zodra men het aantal punten met coördinaten in het grondlichaam, d.w.z. het lichaam waarover de kromme is gedefinieerd, weet.

Voor een elliptische kromme E gedefinieerd over het lichaam $\mathbb{F}_q$ geeft een algemenere vorm van Hasse-Weil de volgende afschattingen:

$$\boxed{q+1-2\sqrt{q} \leq \#(E(\mathbb{F}_q)) \leq q+1+2\sqrt{q}}$$

d.w.z.

$$\boxed{|\#(E(\mathbb{F}_q)) - (q+1)| \leq 2\sqrt{q}}$$

Voor zo'n kromme $E/\mathbb{F}_q$ definieert men $t \in \mathbb{Z}$ door $\#(E(\mathbb{F}_q)) = q+1-t$, dus $|t| \leq 2\sqrt{q}$. Deze t (veelal als $\mathrm{Tr}(F_q)$ genoteerd) heet het **spoor** van het **Frobenius endomorfisme**

$$F_q : E(\bar{\mathbb{F}}_q) \to E(\bar{\mathbb{F}}_q),$$

met $F_q : (x,y) \mapsto (x^q, y^q)$ en $F_q : O \mapsto O$. Voor twee punten $P, Q \in E(\bar{\mathbb{F}}_q)$ geldt $F_q(P+Q) = F_q(P) + F_q(Q)$. Men ziet zo dat $E(\mathbb{F}_q)$ juist uit die punten bestaat die invariant zijn onder F_q, m.a.w. voor $P \in E(\bar{\mathbb{F}}_q)$ geldt dat $P \in E(\mathbb{F}_q) \iff F_q(P) = P$.

Het Frobenius endomorfisme $F_q : E(\bar{\mathbb{F}}_q) \to E(\bar{\mathbb{F}}_q)$ voldoet aan de karakteristieke vergelijking

$$\boxed{F_q^2 - tF_q + q = 0}$$

die men moet interpreteren in de zin van operatoren die werken op de groep van punten van E:

$$(x^{q^2}, y^{q^2}) - t \cdot (x^q, y^q) + q \cdot (x,y) = O, \quad \forall\, (x,y) \in E(\bar{\mathbb{F}}_q).$$

Als α en β de eigenwaarden van F_q zijn, geldt:

$$t = \alpha + \beta, \quad \text{en} \quad \alpha\beta = q,$$

dus $\#(E(\mathbb{F}_q)) = q+1-(\alpha+\beta)$. Nu geldt $\mathrm{Tr}(F_q^n) = \alpha^n + \beta^n$, maar de punten van E die invariant zijn onder F_q^n vormen juist de groep $E(\mathbb{F}_{q^n})$, dus

$$\boxed{\#(E(\mathbb{F}_{q^n})) = q^n + 1 - (\alpha^n + \beta^n)}$$

VOORBEELD. We nemen $q = p = 11$ en komen terug op de kromme $E/\mathbb{F}_{11}$: $y^2 = x^3 - 3x + 1$. We weten dat $\#(E(\mathbb{F}_{11})) = 11+1-(\alpha+\beta) = 17$ (vergeet het punt O niet), dus $\#(E(\mathbb{F}_{11^2})) = 11^2+1-(\alpha^2+\beta^2) = 122-((\alpha+\beta)^2-2\alpha\beta) = 122-((-5)^2-2\cdot 11) = 119$.

Via $\alpha^n + \beta^n = (\alpha+\beta)(\alpha^{n-1}+\beta^{n-1}) - q(\alpha^{n-2}+\beta^{n-2})$ berekent men recursief $\#(E(\mathbb{F}_{q^n}))$ voor $n \geq 3$.

Het zal nu duidelijk zijn dat als men het spoor $\mathrm{Tr}(F_q)$ van Frobenius kan bepalen, men alle informatie over het aantal punten van E over een willekeurige eindige uitbreiding van het grondlichaam $\mathbb{F}_q$ achterhaalt. De berekening van het spoor van Frobenius is de kern van het SEA-algoritme.

7. Het spoor van Frobenius en de groepstruktuur

Als men $\#(E(\mathbb{F}_q))$ en de factorisatie ervan eenmaal weet, is er een (probabilistisch) algoritme van V. Miller om de groepstructuur van $E(\mathbb{F}_q)$ in ongeveer $(ln\, ln\, q)^2$ stappen te bepalen, tenminste als q en $\#(E(\mathbb{F}_q))$ onderling priem zijn, hetgeen bijna altijd het geval zal zijn.

Werk van W. Waterhouse (1969), R. Schoof (1985), H.-G. Rück (1987) en J. Voloch (1988) geeft een volledige classificatie van de mogelijkheden voor t en $E(\mathbb{F}_q)$:

Classificatiestelling. *Laat $q = p^m$, dan bestaat er een elliptische kromme $E/\mathbb{F}_q$ met $\#(E(\mathbb{F}_q)) = q + 1 - t$ en bijvermelde groepstructuur dan en slechts dan als*

1. *$t = 0$, m oneven of $p \not\equiv 1 \bmod 4$, dan is $E(\mathbb{F}_q) = \mathbb{Z}/2 \oplus \mathbb{Z}/(q+1)/2$ of cyclisch als $q \equiv 3 \bmod 4$, anders cyclisch;*
2. *$t = \pm\sqrt{q}$, m even of $p \not\equiv 1 \bmod 3$, dan is $E(\mathbb{F}_q)$ cyclisch;*
3. *$t = \pm\sqrt{2q}$, m oneven en $p = 2$, dan is $E(\mathbb{F}_q)$ cyclisch;*
4. *$t = \pm\sqrt{3q}$, m oneven en $p = 3$, dan is $E(\mathbb{F}_q)$ cyclisch;*
5. *$t = \pm 2\sqrt{q}$, m even, dan is $E(\mathbb{F}_q) = \mathbb{Z}/(q^{1/2} \mp 1) \oplus \mathbb{Z}/(q^{1/2} \mp 1)$;*
6. *$ggd(t, q) = 1$, dan is $E(\mathbb{F}_q)$ van de vorm*

$$E(\mathbb{F}_q) = \mathbb{Z}/p^{v_p(N)} \oplus \bigoplus_{\ell \neq p} \mathbb{Z}/\ell^{r_\ell} \oplus \mathbb{Z}/\ell^{s_\ell},$$

met $N = \#(E(\mathbb{F}_q))$, $r_\ell + s_\ell = v_\ell(N)$ en $0 \leq \min(r_\ell, s_\ell) \leq v_\ell(q-1)$, waarbij, voor een priemgetal ℓ, $v_\ell(k)$ ($k \in \mathbb{N}_{>0}$) het grootste gehele getal is zodat $\ell^{v_\ell(k)} | k$.

Gevolg. *$E(\mathbb{F}_q)$ is een abelse groep van rang 1 of 2, en kan geschreven worden als*

$$E(\mathbb{F}_q) = \mathbb{Z}/n_1 \oplus \mathbb{Z}/n_2,$$

waarbij $n_2 | n_1$ en bovendien $n_2 | (q-1)$.

7.1. Spoor **0***: Supersinguliere elliptische krommen*

Een elliptische kromme $E/\mathbb{F}_q$, $q = p^m$, heet **supersingulier** als $\mathrm{Tr}(F_q)$ deelbaar is door p, i.h.b. als $q = p$, dan $\mathrm{Tr}(F_p) = 0$ en $\#(E(\mathbb{F}_p)) = p + 1$.

Laat $E/\mathbb{F}_q$, $q = p^m$, een elliptische kromme zijn met

$$E(\mathbb{F}_q) \cong \mathbb{Z}/n_1 \oplus \mathbb{Z}/n_2,$$

met $n_2|n_1$ en $n_2|(q-1)$. Voor $n \in \mathbb{N}_{\geq 2}$ schrijft men gewoonlijk $E[n] := E(\bar{\mathbb{F}}_q)[n]$ voor de groep van de n**-torsie punten** van $E(\bar{\mathbb{F}}_q)$, d.w.z. de punten $P \in E(\bar{\mathbb{F}}_q)$ met $nP = O$. De struktuur van de groep $E[n]$ is bekend:

$$E[n] \cong \mathbb{Z}/n \oplus \mathbb{Z}/n, \quad \text{mits } \mathrm{ggd}(n, q) = 1.$$

Neem nu een punt $P \in E(\mathbb{F}_q)$ van orde n, waarbij $n|n_1$. De cruciale stap in de constructie van Menezes, Okamoto en Vanstone berust op het bepalen van de minimale $k \in \mathbb{N}$ zó dat $E[n_1] \subset E(\mathbb{F}_{q^k})$, dus dan zeker ook $E[n] \subset E(\mathbb{F}_{q^k})$.

Uit bovenstaande classificatiestelling leest men af dat er slechts 6 gevallen overblijven voor de groepstruktuur van supersinguliere elliptische krommen:

- $t = 0$, $n_1 = \#(E(\mathbb{F}_q)$, $n_2 = 1$, en $E(\mathbb{F}_q) \cong \mathbb{Z}/(q+1)$ cyclisch;
- $t = 0$, $q \equiv 3 \pmod 4$, $n_1 = (q+1)/2$, $n_2 = 2$, en
$$E(\mathbb{F}_q) \cong \mathbb{Z}/(q+1)/2 \oplus \mathbb{Z}/2;$$
- $t = \pm\sqrt{q}$, m even, $n_1 = q + 1 \mp \sqrt{q}$, $n_2 = 1$, en $E(\mathbb{F}_q) \cong \mathbb{Z}/n_1$ cyclisch;
- $t = \pm\sqrt{2q}$, $p = 2$, m oneven, $n_1 = q + 1 \mp \sqrt{2q}$, $n_2 = 1$, en $E(\mathbb{F}_q) \cong \mathbb{Z}/n_1$ cyclisch;
- $t = \pm\sqrt{3q}$, $p = 3$, m oneven, $n_1 = q + 1 \mp \sqrt{3q}$, $n_2 = 1$, en $E(\mathbb{F}_q) \cong \mathbb{Z}/n_1$ cyclisch;
- $t = \pm 2\sqrt{q}$, m even, $n_1 = n_2 = \sqrt{q} \mp 1$, en
$$E(\mathbb{F}_q) \cong \mathbb{Z}/(\sqrt{q} \mp 1) \oplus \mathbb{Z}/(\sqrt{q} \mp 1).$$

De bijbehorende uitbreidingsgraden k zijn achtereenvolgens 2, 2, 3, 4, 6, 1.

Om dit in te zien, bekijken we bijvoorbeeld het geval $t = \sqrt{q}$. We hebben $\alpha + \beta = \sqrt{q}$ en $\alpha\beta = q$. Dit geeft $\alpha^2 + \beta^2 = (\alpha + \beta)^2 - 2\alpha\beta = -q$ en $\alpha^3 + \beta^3 = (\alpha+\beta)^3 - 3\alpha\beta(\alpha+\beta) = -2q\sqrt{q}$. We zoeken de kleinste $k \in \mathbb{N}$ zó dat $E[n_1] \hookrightarrow E(\mathbb{F}_{q^k})$. Er moet zeker gelden dat n_1^2 een deler is van $\#(E(\mathbb{F}_{q^k}))$. Voor $k = 1$ geldt $\#(E(\mathbb{F}_q)) = q + 1 - (\alpha + \beta) = q + 1 - \sqrt{q} = n_1$. Vervolgens vinden we voor $k = 2$, $\#(E(\mathbb{F}_{q^2})) = q^2 + 1 - (\alpha^2 + \beta^2) = q^2 + 1 + q = n_1(q + 1 + \sqrt{q})$. Tenslotte $\#(E(\mathbb{F}_{q^3})) = q^3 + 1 - (\alpha^3 + \beta^3) = q^3 + 1 + 2\sqrt{q^3}$, dus volgens 5. van de classificatiestelling (met q vervangen door q^3) geldt $E(\mathbb{F}_{q^3}) = \mathbb{Z}/(q\sqrt{q}+1) \oplus \mathbb{Z}/(q\sqrt{q} + 1)$, maar $q\sqrt{q} + 1 = (q + 1 - \sqrt{q})(\sqrt{q} + 1)$ en $E[n_1] \hookrightarrow E(\mathbb{F}_{q^3})$.

OEFENING. Verifieer de overige gevallen zelf.

Het ECDLP heeft nu als input een punt $P \in E(\mathbb{F}_q)$ van orde n en een punt $Q \in \langle P \rangle$, de cyclische groep voortgebracht door P. De output is een geheel getal ℓ zó dat $Q = \ell P$. Menezes e.a. hebben aangetoond dat dit probleem voor supersinguliere krommen in probabilistisch polynomiale tijd (in ln q) is terug te voeren tot het DLP in $\mathbb{F}_{q^k}$, en hierin kan men het DLP in probabilistisch subexponentiële tijd proberen aan te pakken.

7.2. Spoor **1***: Abnormale elliptische krommen*

Elliptische krommen $E/\mathbb{F}_p$ met $\mathrm{Tr}(F_p) = 1$ heten **abnormaal**. Het blijkt dat men gemakkelijk een expliciet isomorfisme $\phi : E(\mathbb{F}_p) \xrightarrow{\sim} \mathbb{Z}/p\mathbb{Z} = \mathbb{F}_p^+$ kan geven.

Eerst voeren we een begrip in: het **Legendre symbool** $\left(\frac{a}{p}\right)$. Dit is als volgt gedefinieerd:

$$\left(\frac{a}{p}\right) = \begin{cases} 0 & \text{als } a \equiv 0 \bmod p \\ 1 & \text{als } a \text{ een kwadratisch residu mod } p \text{ is,} \\ -1 & \text{als } a \text{ een kwadratisch niet-residu mod } p \text{ is.} \end{cases}$$

Hierbij zegt men dat a een kwadratisch (niet-)residu mod p is als de vergelijking $x^2 = a \bmod p$ een (geen) oplossing heeft. Het Legendre symbool heeft vele mooie eigenschappen, maar voor ons is de volgende voldoende:

LEMMA. $a^{\frac{p-1}{2}} \equiv \left(\frac{a}{p}\right) \; mod p.$

Bewijs. Als $p|a$ dan is er niets te bewijzen, dus neem aan dat $a \not\equiv 0 \bmod p$. Dan geldt volgens de 'kleine Fermat': $a^{\frac{p-1}{2}} = \pm 1 \bmod p$. Laat g een voortbrenger van $\mathbb{F}_p^\times$ zijn. Dus de kleinste macht k van g zó dat $g^k = 1 \bmod p$ is $p-1$. Als $a = g^i$ voor zekere i, dan hebben we dus dat $\left(\frac{a}{p}\right) = 1$ dan en slechts dan als i even is. Maar tevens $a^{\frac{p-1}{2}} = g^{\frac{i(p-1)}{2}} = 1 \bmod p$ d.e.s.d. als $\frac{i(p-1)}{2}$ een heeltallig veelvoud is van $p-1$, d.w.z. als i even is. ∎

We zien nu dat het aantal $\mathbb{F}_p$-punten op de kromme

$$E/\mathbb{F}_p : y^2 = x^3 + Ax + B$$

wordt gegeven door de uitdrukking [18]

$$\#(E(\mathbb{F}_p)) = 1 + \sum_{x \in \mathbb{F}_p} \left(1 + \left(\frac{x^3 + Ax + B}{p}\right)\right) = 1 + p + \sum_{x \in \mathbb{F}_p} \left(\frac{x^3 + Ax + B}{p}\right)$$

dus voor het spoor van Frobenius vinden we hier:

$$t = \mathrm{Tr}(F_p) = -\sum_{x \in \mathbb{F}_p} \left(\frac{x^3 + Ax + B}{p}\right).$$

[18] Vergeet het punt O niet!

Bekijk de uitdrukking $y^{p-1} = (x^3 + Ax + B)^{\frac{p-1}{2}}$ en sommeer over $\mathbb{F}_p$. Men vindt

$$\sum_{x\in\mathbb{F}_p} y^{p-1} = \sum_{x\in\mathbb{F}_p} (x^3 + Ax + B)^{\frac{p-1}{2}} \equiv \sum_{x\in\mathbb{F}_p} \left(\frac{x^3 + Ax + B}{p}\right) = -t \pmod{p}.$$

Anderzijds, schrijf y^{p-1} als

$$y^{p-1} = (x^3 + Ax + B)^{\frac{p-1}{2}} = U(x) + Hx^{p-1} + x^p V(x),$$

waarin $U, V \in \mathbb{F}_p[x]$ en U van graad $\leq p-2$ is. De coëfficiënt $H \in \mathbb{F}_p$ heet de **Hasse-invariant** van E.

LEMMA. $\sum_{x\in\mathbb{F}_p} x^k = \begin{cases} p-1 & \text{als } (p-1)|k, \\ 0 & \text{anders.} \end{cases}$

Bewijs. $S = \sum_{x\in\mathbb{F}_p} x^k = \sum_{x\in\mathbb{F}_p^\times} x^k = \sum_{x\in\mathbb{F}_p^\times} (zx)^k = z^k S$ voor willekeurige $z \in \mathbb{F}_p^\times$, dus $(1 - z^k)S = 0$ en de bewering volgt.

■

Door dit lemma toe te passen op $y^{p-1} = U(x) + Hx^{p-1} + x^p V(x)$ vindt men nu dat

$$\boxed{H \equiv t \pmod{p}}$$

In het geval van een abnormale kromme geldt dus $H = 1$. Zonder bewijs [19] geven we het volgende resultaat:

STELLING. *Zij E een abnormale elliptische kromme over $\mathbb{F}_p$. Dan is de afbeelding*

$$\phi : E(\mathbb{F}_p) \longrightarrow \mathbb{F}_p^+,$$

gegeven door $\phi : (x, y) \mapsto yV(x)$ en $\phi(O) = 0$ een isomorfisme.

We lichten dit toe aan de hand van een eenvoudig voorbeeld.

VOORBEELD. Neem de kromme $E/\mathbb{F}_{13} : y^2 = x^3 + 7x + 3$. Deze kromme is abnormaal met

$$E(\mathbb{F}_{13}) = \{O, (0, \pm 4), (2, \pm 5), (3, \pm 5), (4, \pm 2), (6, \pm 1), (8, \pm 5)\}.$$

Bereken $y^{12} = (x^3 + 7x + 3)^6$ en schrijf dit als

$$y^{12} = U(x) + x^{12} + x^{13} V(x),$$

met U een polynoom in x met coëfficiënten in $\mathbb{F}_{13}$ en van graad ≤ 11. Voor $V(x)$ vindt men hier

$$V(x) = x^5 + 3x^3 + 5x^2 + 7x + 6.$$

[19] Hier is weer diepgaande algebraïsche meetkunde nodig.

Het isomorfisme $\phi : E(\mathbb{F}_{13}) \xrightarrow{\sim} \mathbb{Z}/13\mathbb{Z} = \mathbb{F}_{13}^{+}$ wordt gedefinieerd door

$$\phi : (x, y) \mapsto yV(x) \quad \text{en} \quad \phi(O) = 0.$$

In dit voorbeeld vindt men $\phi(0, \pm 4) = \pm 11$, $\phi(2, \pm 5) = \pm 12$, $\phi(3, \pm 5) = \pm 4$, $\phi(4, \pm 2) = \pm 8$, $\phi(6, \pm 1) = \pm 7$, $\phi(8, \pm 5) = \pm 10$. De discrete logaritme wordt nu die in $(\mathbb{F}_{13}, +)$, b.v. gegeven $P = (8, 5)$ en $Q = (0, 9)$, waarbij men $Q = \ell P$ schrijft en ℓ wil bepalen. Men vindt $\ell = \frac{-11}{10} \pmod{13} = \frac{1}{5} \bmod 13) = 8 \pmod{13}$, dus $(0, 9) = 8 \cdot (8, 5)$.
Het zal de lezer snel duidelijk worden dat bovenstaande methode ter berekening van het isomorfisme ϕ bij grotere waarden van p ondoenlijk wordt, immers de graad van $V(x)$ is $(p-3)/2$. Er bestaan evenwel algoritmen om ϕ voor grotere waarden van p te berekenen. Dit valt echter buiten het bestek van dit verslag.

7.3. Spoor **2***: Meer zuivere wiskunde*

Zeer recent werk (1998) van G. Frey, M. Müller en H.-G. Rück, gebaseerd op vroeger werk van J. Tate en van S. Lichtenbaum, leidt tot een interessante toepassing op het ECDLP. Men kan dit zien als een uitbreiding van en een aanvulling op het resultaat van Menezes, Okamoto en Vanstone voor supersinguliere krommen. I.h.b. als p een priem is zo dat ook $m = \dfrac{p-1}{2}$ priem is, en $E/\mathbb{F}_p$ is een elliptische kromme met $\mathrm{Tr}(F_p) = 2$, d.w.z. $\#(E(\mathbb{F}_p)) = p - 1$, dan kan het ECDLP direct worden opgelost of worden teruggebracht tot het DLP in $\mathbb{F}_p$ in polynomiale $(\log m)$ tijd [20] . De benodigde begrippen en technieken berusten weer op geavanceerde algebraïsche meetkunde (mensenwerk!) en gaan de theorie zoals beschreven in deze voordracht ver te boven.

LITERATUUR EN SOFTWARE

1. N. KOBLITZ, *A Course in Number Theory and Cryptography*, GTM 114, Springer-Verlag.
2. A. MENEZES, *Elliptic Curve Public Key Cryptosystems*, Kluwer Academic Publishers.
3. J. SILVERMAN, *The Arithmetic of Elliptic Curves*, GTM 106 Springer-Verlag.
4. M.J. COSTER, `www.coster.demon.nl/elliptic/index.html`

[20] Het resultaat van genoemde auteurs is veel algemener.

Medewerkers

Prof.dr. N.G. de Bruijn (TUE)
Eikenlaan 2, 5671 AB Nuenen, n.g.d.bruijn@tue.nl

Dr. M. Coster (Ministerie van Defensie)
Wielingenstraat 24r, 1078 KL Amsterdam, matthijs@coster.demon.nl

Prof.dr. J. van de Craats
KMA, Postbus 90154, 4800 RG Breda, tel. 076-5273816, jcr@euronet.nl

Drs. A. Heck (UvA)
Amstel Instituut, Kruislaan 404, 1098 SM Amsterdam, heck@wins.uva.nl

Dr.ir. H.J.A.M. Heijmans (CWI)
Centrum voor Wiskunde en Infmatica, Kruislaan 413, 1098 SJ Amsterdam, henkh@cwi.nl

Dr. W.W.J. Hulsbergen (Ministerie van Defensie)
Engelandlaan 408, 2034 NN Haarlem, hulsw@wxs.nl

Prof.dr. J. Molenaar (TUE, UT)
Technische Universiteit Eindhoven, Faculteit Wiskunde en Informatica IWDE, Postbus 513, 5600 MB Eindhoven, jaapm@win.tue.nl

Dr. H.G. ter Morsche (TUE)
Technische Universiteit Eindhoven, Faculteit Wiskunde en Informatica, Postbus 513, 5600 MB Eindhoven, morscheh@win.tue.nl

Dr. H.J.M. Sterk (TUE)
Technische Universiteit Eindhoven, Faculteit Wiskunde en Informatica, Postbus 513, 5600 MB Eindhoven, sterk@win.tue.nl

Contacten Centrum voor Wiskunde en Informatica

Dr. M. Bakker
Centrum voor Wiskunde en Informatica, Kruislaan 413, Postbus 94079,
1090 GB Amsterdam, 020 592 4172, e-mail: Miente.Bakker@cwi.nl

Wilmy van Ojik
Centrum voor Wiskunde en Informatica, Kruislaan 413, Postbus 94079,
1090 GB Amsterdam, 020 592 4200, e-mail: Wilmy.van.Ojik@cwi.nl

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 Vacantiecurus 1985: *Variatierekening.* 1985.

8 G.M. Tuynman. *Proceedings Seminar 1983–1985. Mathematical structures in field theories, Vol.1 Geometric quantization.* 1985.

9 J. van Leeuwen, J.K. Lenstra (eds.). *Parallel computers and computations.* 1985.

10 Vacantiecursus 1986: *Matrices.* 1986.

11 P.W.H. Lemmens. *Discrete wiskunde: tellen, grafen, spelen en codes.* 1986.

12 J. van de Lune. *An introduction to Tauberian theory: from Tauber to Wiener.* 1986.

13 G.M. Tuynman, M.J. Bergvelt, A.P.E. ten Kroode. *Proceedings Seminar 1983–1985. Mathematical structures in field theories, Vol.2.* 1987.

14 Vacantiecursus 1987: *De personal computer en de wiskunde op school.* 1987.

15 Vacantiecursus 1983: *Complexe getallen.* 1987.

16 P.J.M. Bongaarts, E.A. de Kerf, P.H.M. Kersten. *Proceedings Seminar 1984–1986. Mathematical structures in field theories, Vol.1.* 1988.

17 F. den Hollander, H. Maassen (eds.). *Mark Kac seminar on probability and physics. Syllabus 1985–1987.* 1988.

18 Vacantiecursus 1988. *Differentierekening.* 1988.

19 R. de Bruin, C.G. van der Laan, J. Luyten, H.F. Vogt. *Publiceren met LATEX.* 1988.

20 R. van der Horst, R.D. Gill (eds.). *STATAL: statistical procedures in Algol 60, part 1.* 1988.

21 R. van der Horst, R.D. Gill (eds.). *STATAL: statistical procedures in Algol 60, part 2.* 1988.

22 R. van der Horst, R.D. Gill (eds.). *STATAL: statistical procedures in Algol 60, part 3.* 1988.

23 J. van Mill, G.Y. Nieuwland (eds.). *Proceedings van het symposium wiskunde en de computer.* 1989.

24 P.W.H. Lemmens (red.). *Bewijzen in de wiskunde.* 1989.

25 Vacantiecursus 1989: *Wiskunde in de Gouden Eeuw.* 1989.

26 G.G.A. Bäuerle et al. *Proceedings Seminar 1986–1987. Mathematical structures in field theories.* 1990.

27 Vacantiecursus 1990: *Getallentheorie en haar toepassingen.* 1990.

28 Vacantiecursus 1991: *Meetkundige structuren.* 1991.

29 A.G. van Asch, F. van der Blij. *Hoeken en hun Maat.* 1992.

30 M.J. Bergvelt, A.P.E. ten Kroode. *Proceedings seminar 1986–1987. Lectures on Kac-Moody algebras.* 1992.

31 Vacantiecursus 1992: *Systeemtheorie.* 1992.

32 F. den Hollander, H. Maassen (eds.). *Mark Kac seminar on probability and physics. Syllabus 1987–1992.* 1992.

33 P.W.H. Lemmens (ed.). *Meetkunde van kunst tot kunde, vroeger en nu.* 1993.

34 J.H. Kruizinga. *Toegepaste wiskunde op een PC.* 1992.

35 Vacantiecursus 1993: *Het reële getal.* 1993.

36 Vacantiecursus 1994: *Computeralgebra.* 1994.

37 G. Alberts. *Wiskunde en praktijk in historisch perspectief. Syllabus.* 1994.

38 G. Alberts, J. Schut (eds.). *Wiskunde en praktijk in historisch perspectief. Reader.* 1994.

39 E.A. de Kerf, H.G.J. Pijls (eds.). *Proceedings Seminar 1989–1990. Mathematical structures in field theory.* 1996.

40 Vacantiecursus 1995: *Kegelsneden en kwadratische vormen.* 1995.

41 Vacantiecursus 1996: *Chaos.* 1996.

42 H.C. Doets. *Wijzer in Wiskunde.* 1996.

43 Vacantiecursus 1997: *Rekenen op het Toeval.* 1997.

44 Vacantiecursus 1998: *Meetkunde, Oud en Nieuw.* 1998.

45 Vacantiecursus 1999: *Onbewezen Vermoedens.* 1999.

46 P.W. Hemker, B.W. van de Fliert (eds.). *Proceedings of the 33^{rd} European Study Group with Industry.* 1999.

47 K.O. Dzhaparidze. *Introduction to Option Pricing in a Securities Market.* 2000.

48 Vacantiecursus 2000: *Is wiskunde nog wel mensenwerk?* 2000.

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.