CWI Syllabi

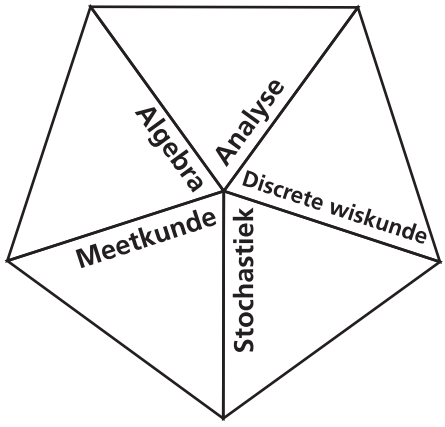
Centrum voor Wiskunde en Informatica

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

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

Vakantiecursus 2005

De schijf van vijf



Centrum voor Wiskunde en Informatica
CWI SYLLABUS 54

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.

Deze cursus is mede mogelijk gemaakt door een subsidie van de Nederlandse Organisatie voor Wetenschappelijk Onderzoek.

Ontwerp omslag: Tobias Baanders naar een illustratie uit de Œuvres Complètes van Christaan Huygens.

ISBN 90 6196 531 4
NUGI-code: 811

Inhoud

Docenten

Dr. A. Bolck, Dr. M. Sjerps
Nederlands Forensisch Instituut
Laan van Ypenburg 6, 2497 GB Den Haag, 070-8886666
{a.bolck,m.sjerps}@nfi.minjus.nl

Dr. J. Brinkhuis
Erasmus Universiteit Rotterdam, Econometrisch Instituut
Postbus 1738, 3000 DR Rotterdam, 010-4081271/81277
brinkhuis@few.eur.nl

Dr. E. Coplakova
Technische Universiteit Delft, Elektrotechniek Wiskunde en Informatica
Mekelweg 4, 2628 CD Delft, 015 2785800
e.coplakova@ewi.tudelft.nl

Prof.dr. J. van de Craats
Marinus de Jongstraat 12, 4904 PL Oosterhout, 0162-457364
jcr@euronet.nl

Dr. R.H. Jeurissen
rhjeuris@wanadoo.nl

Dr. R.H. Kaenders
Radboud Universiteit Nijmegen, Instituut voor Leraar en School
Postbus 38250, 6503 AG Nijmegen, 024 35 30090
R.Kaenders@ils.ru.nl

Dr. N. Litvak
Universiteit Twente, Elektrotechniek Wiskunde & Informatica
Postbus 217, 7500 AE Enschede, 053-4893388
n.litvak@math.utwente.nl

Dr. R.H. Vermij
Universiteit Utrecht, Instituut voor de Geschiedenis en de Grondslagen van de Natuurwetenschappen
Postbus 80000 3508 TA Utrecht 030-2533173
R.H.Vermij@phys.uu.nl

Ten geleide

J. van de Craats
Open Universiteit
Universiteit van Amsterdam
e-mail: jcr@euronet.nl

Al jarenlang is de *schijf van vijf* een symbool voor gezonde voeding: wie gezond wil leven, moet in zijn menukeuze voor voldoende variatie zorgen: de schijf van vijf is daarbij een leidraad. Ook de leraar die in zijn vak bij wil blijven, moet van tijd tot tijd geestelijk voedsel tot zich nemen. En ook daarbij is eenzijdigheid uit den boze. De organisatoren van de CWI-Vakantiecursus 2005 bieden een gevarieerd menu, samengesteld uit een wiskundige *'schijf van vijf'*, bestaande uit meetkunde, algebra, analyse, discrete wiskunde en stochastiek.

Op het programma van deze cursus staan een verhaal over Christiaan Huygens en diens meetkundige aanpak van de mechanica, een bijdrage over statistiek in de misdaadbestrijding, een lezing over zoekmachines op het internet en een discreet-wiskundige bijdrage over foutenherstellende codes. De tweede cursusdag staat voor een groot deel in het kader van de analyse en de algebra, maar ook daar komt soms weer meetkunde aan te pas. Er is een voordracht over kranen en lemniscaten, een lezing over het oplossen van algebraïsche vergelijkingen, een inleiding in de Fouriertheorie en een bijdrage over continue optimaliseringsmethoden in de economie.

Gaarne wil ik hier allen bedanken die in 2005 opnieuw een Vakantiecursus mogelijk hebben gemaakt. In de eerste plaats natuurlijk de sprekers, die naast hun lezing ook een tekst voor deze Syllabus hebben geleverd. Het Centrum voor Wiskunde en Informatica te Amsterdam en de Technische Universiteit Eindhoven stelden zaalruimte beschikbaar, de administratieve en praktische organisatie van de cursus was in handen van Wilmy van Ojik en dr. Miente Bakker, die ook samen met Minnie Middelberg de inhoudelijke coördinatie van deze Syllabus verzorgden.

Allen hartelijk dank!

CWI Vakantiecursus 2005

pp. 3 – 18

Christiaan Huygens: de wiskunde en de werkelijkheid van de zeventiende eeuw

R.H. Vermij
Universiteit Utrecht
e-mail: R.H.Vermij@phys.uu.nl

Deze bijdrage aan de syllabus bestaat uit twee gedeelten:

1. Een verhandeling over Huygens' publicatie *De motu corporum ex percussione*
2. Een herdruk van de voordracht *Huygens and mathematics*, die Henk Bos in april 2004 in Noordwijk heeft gehouden tijdens de ESA-conferentie 'Titan – From Discovery to Encounter' met dank aan de ESA

Deel 1 – De motu corporum ex percussione

1. Toelichting

Huygens begon zich voor de botsingswetten te interesseren in 1652, toen hij 22 jaar was. Aanleiding waren de regels die Descartes had gegeven voor de botsingen van harde lichamen. Huygens stelde vast dat Descartes' regels met elkaar in strijd waren. Hij correspondeerde hierover met zijn leermeester Frans van Schooten. Rond 1656 schreef hij een verhandeling waarin hij de botsingsregels wiskundig afleidde. In de jaren daarop werd hij evenwel door tal van zaken in beslag genomen, eerst door de optica en de vervaardiging van telescopen, vervolgens door het slingeruurwerk. Mede om die reden bleef de verhandeling over de botsingsregels liggen. In de latere jaren van zijn leven keerde Huygens verschillende keren tot het onderwerp terug. In zijn nalatenschap bevinden zich verschillende versies van de verhandeling over botsingen, geschreven op verschillende tijdstippen. Bij zijn leven kwam het echter nooit tot een publicatie. Alleen publiceerde hij in 1669 in het Journal des Scavans een kort stukje met de belangrijkste conclusies. De verhandeling zelf werd pas na zijn dood gepubliceerd in 1703, in zijn Opuscula postuma, onder de titel 'De motu corporum ex percussione' [over de beweging van lichamen ten gevolge van botsingen].

Deze versie van 1703 is het uitgangspunt van de onderstaande vertaling. De vertaling betreft niet de hele verhandeling, maar alleen de stellingen, lemmata en uitgangspunten. Met andere woorden, de hele bewijsvoering blijft achterwege, aangezien dat wat ver zou voeren. De structuur van het stuk is op

Figuur 1. *Christiaan Huygens*

deze manier echter goed zichtbaar. Het biedt een goed voorbeeld van Huygens' gebruik van wiskunde en van zijn redeneertrant.

Twee opmerkingen tot slot. Waar Huygens over snelheid spreekt, moet bedacht worden dat het vectorbegrip in deze tijd nog onbekend is en dat het steeds alleen over de grootte van de snelheid gaat. En de middenevenredige tussen twee grootheden a en b is de grootheid c waarvoor geldt: $a : c = c : b$.

2. Tekst

Hypothese 1 *Wanneer een lichaam eenmaal in beweging is zal het, als het niet gehinderd wordt, altijd doorgaan te bewegen met dezelfde snelheid en volgens een rechte lijn.*

Hypothese 2 *Wat ook de oorzaak is van het terugkaatsen van harde lichamen bij onderling contact wanneer ze op elkaar worden gestoten: wij stellen dat wanneer twee gelijke lichamen met gelijke snelheid van weerszijden rechtstreeks op elkaar lopen, ze allebei teruggekaatst worden met dezelfde snelheid waarmee ze kwamen aanlopen.*

Hypothese 3 *De beweging van lichamen, en de gelijke dan wel ongelijke snelheden, moeten worden verstaan betrekking te hebben op andere lichamen die als in rust worden beschouwd, hoewel mogelijk zowel deze als gene deelnemen aan een andere, gemeenschappelijke, beweging. Wanneer dus twee lichamen op elkaar stoten zullen ze, ook als ze beide gelijkelijk aan een gemeenschappelijke beweging onderworpen zijn, elkaar niet anders terugstoten ten opzichte van [een lichaam] dat eveneens deze gemeenschappelijke beweging uitvoert, dan als die beweging geheel afwezig zou zijn.*

Stelling 1 *Als een lichaam botst met een even groot lichaam dat in rust is, zal het na het contact in rust verkeren. Het lichaam dat eerst in rust was zal de snelheid hebben verkregen van het lichaam dat de botsing veroorzaakte.*

Stelling 2 *Wanneer twee even grote lichamen op elkaar botsen met ongelijke snelheden, zullen zij na het contact bewegen met onderling uitgewisselde snelheden.*

Hypothese 4 *Wanneer een groter lichaam botst tegen een kleiner dat in rust is, geeft hij hem een zekere beweging, en derhalve verliest hij daar zelf van.*

Stelling 3 *Hoe groot een lichaam ook is, als het door een ander lichaam, hoe klein en met wat voor beweging ook, wordt aangestoten, zal het worden bewogen.*

Hypothese 5 *Wanneer bij een botsing van twee harde lichamen de uitkomst is dat de een al zijn beweging heeft behouden, dan zal ook de ander geen beweging erbij hebben gekregen of zijn kwijtgeraakt.*

Stelling 4 *Altijd als twee lichamen met elkaar botsen zal de snelheid waarmee zij zich ten opzichte van elkaar verwijderen hetzelfde zijn als de snelheid waarmee zij elkaar naderden.*

Stelling 5 *Als twee lichamen op elkaar botsen met dezelfde snelheden als waarmee zij zich, ieder voor zich, na een eerdere botsing van elkaar verwijderden, dan zal elk na deze tweede botsing dezelfde snelheid krijgen als waarmee hij bij de eerdere botsing op het andere af ging.*

Stelling 6 *Wanneer twee lichamen op elkaar botsen zal na de botsing de hoeveelheid beweging van het systeem als geheel niet altijd hetzelfde blijven als hij voor de botsing was. Hij kan meer of minder zijn geworden.*

Stelling 7 *Als een groter lichaam botst op een kleiner lichaam dat in rust is, geeft hij hem een snelheid die kleiner is dan het dubbele van zijn eigen snelheid.*

Stelling 8 *Als twee lichamen op elkaar botsen waarvan de snelheden omgekeerd evenredig zijn aan hun grootte, zal elk worden teruggekaatst met dezelfde snelheid als waarmee hij aan kwam lopen.*

Stelling 9 *Gegeven zijn twee lichamen van ongelijke grootte die rechtstreeks op elkaar botsen en die of beide bewegen, of waarvan slechts een beweegt. Gegeven is verder de snelheid van beide, of van de ene als de ander in rust is. Gevraagd worden de snelheden waarmee beide na de botsing bewegen.*

Stelling 10 *De snelheid die een groter lichaam geeft aan een kleiner dat in rust is verhoudt zich tot de snelheid die dat kleinere lichaam geeft aan het grotere als het met dezelfde snelheid botst en het grotere in rust is, als de onderlinge groottes van de lichamen.*

Stelling 11 *Wanneer twee lichamen op elkaar botsen zullen de produkten van hun groottes met het kwadraat van hun snelheden, bij elkaar opgeteld, voor en na de botsing gelijk worden gevonden. Namelijk als de verhoudingen van de groottes en van de snelheden in getallen of in lijnstukken worden gesteld.*

Lemma 1 *Laat de rechte AB verdeeld zijn in C en D zo dat het deel AC kleiner is dan CD en CD kleiner is dan BD. Dan zeg ik dat de rechthoek met zijden AD en CB kleiner is dan het dubbele van de som van de rechthoeken ACD en CDB.*

Lemma 2 *Laat AB, AC en AD drie rechte lijnen zijn in evenredige verhouding [d.w.z. $BA : AC = CA : AD$], waarvan AB de grootste is. Voeg aan elk van hen dezelfde lengte AE toe. Dan zeg ik dat de rechthoek met zijden BE en DE groter is dan het vierkant met zijde CE.*

Stelling 12 *Als enig lichaam op een groter of kleiner lichaam af beweegt dat in rust verkeert, zal hij dit een grotere snelheid geven door tussenplaatsing van een eveneens rustend lichaam, met een grootte die het midden houdt tussen die van de beide andere lichamen, dan wanneer hij er zonder zo'n middelaar op botst. Het zal het andere lichaam de grootste snelheid geven, wanneer [de grootte van] het tussengeplaatste lichaam van de beide uitersten de middenevenredige is.*

Stelling 13 *Naarmate er tussen twee ongelijke lichamen, waarvan het ene stilstaat en het andere [naar het eerste toe] beweegt, meer lichamen worden geplaatst, zal er aan het rustende lichaam een grotere beweging kunnen worden overgedragen. Bij een gegeven hoeveelheid tussengeplaatste lichamen zal de grootste beweging worden overgedragen, als de [groottes van de] tussengeplaatste lichamen samen met [die van] de beide uiterste een evenredige reeks uitmaken.*

Deel 2 – Huygens and mathematics

H.J.M. Bos
Universiteit Utrecht
e-mail: Bos@math.uu.nl

1. Drawings

The drawing in the first figure is by Christiaan Huygens. You may still find some spots quite like it not far from ESTEC at Noordwijk. As you see, Huygens was a creditable amateur draftsman. He was also a professional draftsman in as far as his professional work involved drawing many mathematical figures.

Drawings, especially those appearing in early notes and drafts of arguments, have a special status in the process of mathematical research: they often are the first materializations of the thoughts in the brain of the mathematician.

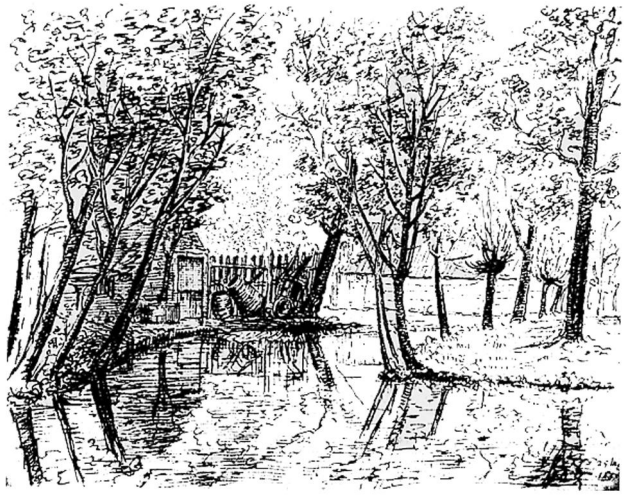


Figure 1. *Drawing by Christiaan Huygens, 1657 (O.C.*[1] *Vol 22, pp. 78–79)*

And even if they are redrawn later, and finally printed, these drawings retain a nearness to mathematical thought which written words and formulas often lack.

With this in mind, I decided to deal with my subject, Huygens and Mathematics, via Huygens' mathematical drawings, and I begin with a very brief, even somewhat hasty tour through the gallery of these drawings and figures.

2. A tour of the gallery

In Figure 2 we have Huygens thinking about rolling. In the middle a hexagon is rolling along a line. He draws a rather bumpy approximation of the process of a circle rolling smoothly: a series of successive turns of the hexagon around a corner. Below a pentagon is rolling, above again a hexagon, now rolling along a curve. Huygens used these sketches to understand the rolling process. Obviously there is a limit process involved: regular polygons with more and more sides are less and less bumpy; real rolling is when the polygons transform into a circle. The drawings in Figure 3 illustrate a similar approach. They are from the beginning of Huygens' career, when he studied the catenary, the form of a free hanging cord or chain.

Again he uses an approximation. He considers a weightless cord, with equal weights hanging at equal distances. What happens along the successive weights can be exactly determined by statics; the drawing suggest to extrapolate this knowledge to the continuous case where the weights are as it were spread out all along the chain or the cord. Again, a limit process. In 1646 Huygens managed

[1] O.C. = Œuvres Complètes de Christiaan Huygens, see Acknowledgements at page 18

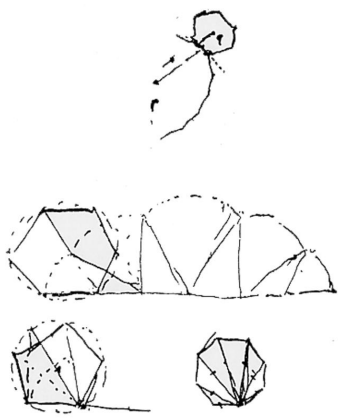


Figure 2. *Sketches of rolling figures, 1678 (O.C. Vol 18, pp. 402)*

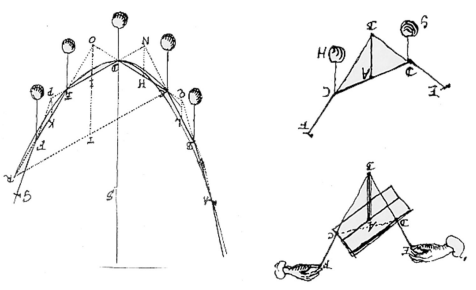


Figure 3. *Approximating the caternary, 1646 (O.C. Vol 11 pp. 37–40)*

to prove by such an extrapolation that the catenary could not be a parabola (as Galileo had suggested), but only much later he was able to determine the true form of the curve. Then another drawing (Figure 4), from October 27th, 1657, and marked (in Greek) *Heureka*, so Huygens had found something. What that was I'll tell later. For now we'll just look at the elements of the drawing. There are curves and axes. Along the curve to the right we see a sequence of tangents. Near the point where they touch the curve they almost coincide with it. The curve is approximated by a polygon of tangent pieces along it.

In the middle there is another curve. Over an area between this curve and the vertical axis narrow strips are drawn; together they form a rectilineal area

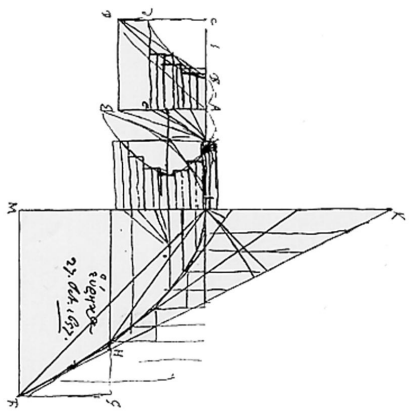


Figure 4. *Curves: tangents and areas, 1657 (O.C. Vol 14 pp. 234)*

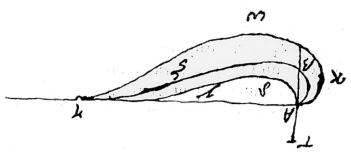


Figure 5. *Curves: the 'paracentric isochrone'*

approximating the area to the right of the curve.

Small strips under a curve and small straight tangent segments along a curve; they are perpetually recurring themes in Huygens' drawings; we will see more of them. Huygens saw them as very small, or becoming ever smaller, or infinitely small; I will use the term infinitesimals for these elements. And of course you sense their relation to what we know as differentiation and integration.

Another recurring theme in the drawings are curves. Figure 5 shows an example taken from a letter Huygens wrote in 1694. Huygens called the three curves in the figure 'paracentric isochrones'; they had to do with a complicated problem, actually at the very edge of research at the time, about motion in a vertical plane along curved trajectories.

Two other isochronic curves drawn by Huygens is in Figure 6. I show them mainly because I like the spiralling effect. Figure 7 shows a curve whose nature is more easily explained. It concerns what was at the time called an 'inverse tangent problem.' The usual tangent problem was: given a curve, determine its tangents. The inverse one was: given a property of tangents, determine

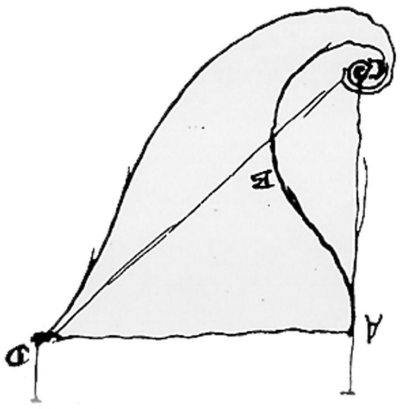


Figure 6. *Curves: a spiralling isochrone, 1694 (O.C. Vol 10 pp. 668)*

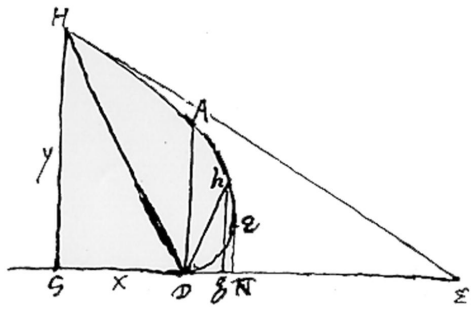


Figure 7. *Curves: solution of an 'inverse tangent problem', 1694 (O.C. Vol 10 pp. 475*

a curve whose tangents have that property. Here the property is that at any point H on the curve, the subtangent, i.e. the segment along the axis below the tangent, should be equal to the sum of the coordinates x and y (Huygenbs takes x and y positive):

$$GE = x + y.$$

You will realize that such problems are equivalent to differential equations. In the case of figure 7 the corresponding differential equation is:

$$\frac{dy}{dx} = \frac{y}{(x+y)} \ .$$

These inverse tangent problems were difficult, indeed often very difficult.

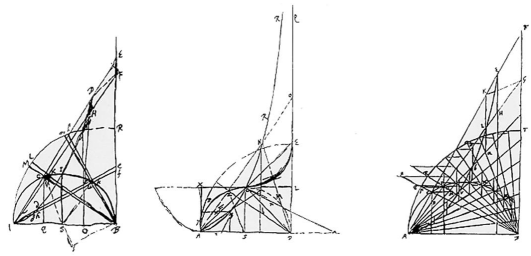


***Figure 8**. At work on the conchoid, 1657 (O.C. Vol 14 pp. 309–311)*

I noted that the seventeenth-century infinitesimals involved in tangents and areas of curves relate to what soon after became differentiation and integration. Similarly, curves in the seventeenth century had the role which was later taken over by the concept of function. Actually, that transition came later, roughly by the middle of the eighteenth century. For Huygens curves, not functions, were the natural means to represent mathematical relationships.

Finally three drawings (Figure 8) showing Huygens at work on a curve called the conchoid; it is the one from A to D in the left-hand drawing, in which Huygens first roughly sketched the curve. You note the infinitesimals he was interested in here: they are the small triangular strips. In the middle drawing he added some details and apparently decided that the drawing was still too sketchy for clarity about the infinitesimals, so for the right-hand drawing Huygens turned to tools of the trade, ruler and compass, to get a better result.

3. Seeing through the Drawings

So far some glimpses from the gallery of Huygens' mathematical drawings. How did they function in Huygens' research?

Obviously they helped him, first of all to order complex spatial information. But they also showed him something which is not on them. He could, as it were, see through the drawings to what cannot be represented in a static drawing, notably motion and the infinitely small. He could see motion of objects along curves, and he could see limits when rolling polygons turned into a rolling circle and when curves temporarily took the form of a polygon of tangent lines.

I shall now turn to a few examples in which Huygens used his drawings in this way to represent the unrepresentable. I divide them according to the following three themes: infinitesimals and limits, motion, and the modeling of processes of movement and change.

3.1. Infinitesimals and Limits

For the infinitesimals and limits I return to a drawing shown earlier (Figure 9), the one with the 'heureka,' which I used as an illustration of infinitesimals, the small tangent parts along a curve and the small strips approximating the area

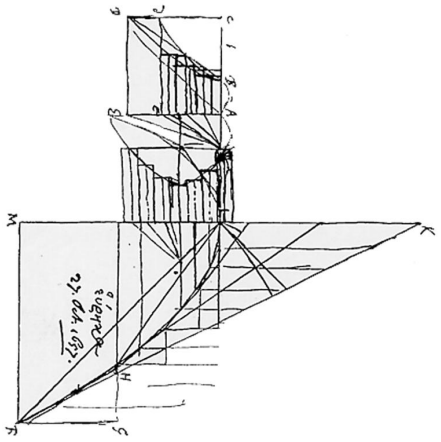


Figure 9. *Arc lengths and areas, 1657 (O.C. Vol 14 pp. 234)*

under a curve. The curve to the right in the drawing is a parabola; the one in the middle is a hyperbola.

What Huygens found — "heureka!" — was a relation between two problems that were famously difficult at the time. The one was to determine the arclength of a parabola between two given points on it; the so-called 'rectification of the parabola.' The other problem was to determine the area under a hyperbola between two given ordinates; the 'quadrature of the hyperbola.' About 1657, when Huygens made the drawing, a few mathematicians had seen that the quadrature of the hyperbola depended on logarithms.

Huygens noticed that the small tangent pieces along the parabola are equal to the corresponding strips under the hyperbola. To see that requires an intimate familiarity with the properties of both curves. Huygens concluded that the sum of all the tangent pieces along the parabola is equal to the sum of the strips under the hyperbola. In the limit, when the corresponding pieces and strips are 'infinitely small,' the sums become equal to the arclength of the parabola and the area under the hyperbola, respectively. Hence the two problems were strictly related: if the quadrature of the hyperbola was found, then the rectification of the parabola was found as well, and *vice versa*. And thus Huygens had found that for determining the lengths of parabolic arcs one needed logarithms in the same way as for the quadrature of the hyperbola.

The drawing, then, illustrates how Huygens used a sketch of curves and infinitesimals to see and understand the limit processes involved in measuring curvilinear lengths and areas.

It is instructive to compare this visual understanding of rectification with the

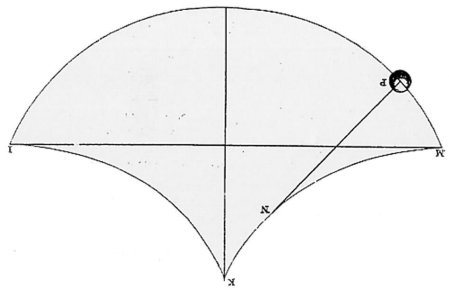


***Figure 10**. Unrolling a curve and the radius of curvature, 1673 (O.C. Vol 18 pp. 105)*

modern, analytic, standard formula for the arclength of a curve with equation $y = f(x)$:

$$s = \int \sqrt{1 + (\frac{dy}{dx})^2}\, dx.$$

Huygens' drawing, as it were, carries the proof that this formula indeed provides the arclengths, as well as the fact that in the case of the parabola the function to be integrated is a hyperbola. Both the proof and the fact are implied in the formula but they are much less visible than in the drawing.

3.2. Motion

Infitesimals, such as in the previous example, occur in Huygens' work especially in connection with motion and dynamics. My second example is about a special kind of motion, namely the unrolling or 'evolution' of curves. Figure 10, taken from Huygens' book on pendulum motion from 1673, illustrates the process. The pendulum consists of a weight P, connected via a thread to a fixed point K in a vertical plane in which two curved strips (of metal for instance) KM and KI are fixed. At rest, the weight hangs vertically under the point K and the thread is straight. If the weight is moved to the left, the thread will wind up along the curve MK; when the weight is in P, as drawn in the figure, the thread is partly straight (the part PN) and partly wound up along the curve (the part NK). If the weight is released from position M it will swing down, pass the lowest point, and move up towards I, and then return along the same path to M, then back down again, and so on. During this motion the thread first unwinds from the curve MK and then winds up along KI, and then winds off KI again and so on. Huygens was fascinated by this process of threads winding, or rolling up or from curves. In the case illustrated in the figure the two curves KM and KI are symmetrically placed halves of a special curve called the 'cycloid'; in that case the path MPI of the weight turns out to be a full cycloid. This phenomenon was crucial in Huygens' theory of oscillation. But the process of unrolling can be generalised to apply for any

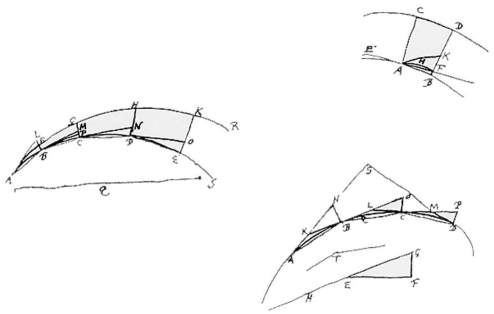


Figure 11. *Evolutes and second-order infinitesimals, 1659 (O.C. Vol 14 pp. 400–402)*

curve MK producing 'evolutes' of KM such as the curve described by P. Huygens derived various properties of curves and their evolutes, such as the fact that the curvature of the evolute at P is equal to the curvature of a circle with center N and radius NP. This length is therefore called the 'radius of curvature' of the evolute at P.

Figure 11 shows some of the drawings through which Huygens 'saw' the process of unrolling along curves with varying curvatures, a process involving infinitely small line segments along the curve and even doubly infinitely small ones perpendicular to the curve. In the drawing to the left (the other two are variants or details of it) we recognize the tangent pieces AL, BM, CN, DO, etc., touching the curve AS. They are infinitesimals in the sense that in the limit, when the arc AS is divided in more and more (infinitely many) pieces, their number becomes (is) infinite, and the sum of their lengths becomes (is) equal to the total length of the arc AS. Now consider the small sides BL, CM, DN, EO, etc. of the triangles ABL, BCM CDB, DEO etc. They are perpendicular to the curve. In the limit process these perpendiculars will of course become zero, but the drawing suggests that they will also become very (infinitely) small with respect to AB, BC, etc along the curve, which themselves also become infinitely small. Huygens made precise what this meant: unlike the 'first order'infinitesimals AB, BC, etc, which become zero but whose sum becomes equal to a finite value (namely the length of the curve), these perpendiculars are 'second order' infinitesimals; they will become zero and their sum will become zero as well. Huygens even provided an explicit proof of this phenomenon, which formed the basis of his further theory of the evolutes of curves.

Again it is instructive to compare Huygens' infinitesimal geometric arguments based on drawings with a modern formula for one of his results. Let ρ be the radius of curvature of a curve $y = f(x)$. Then

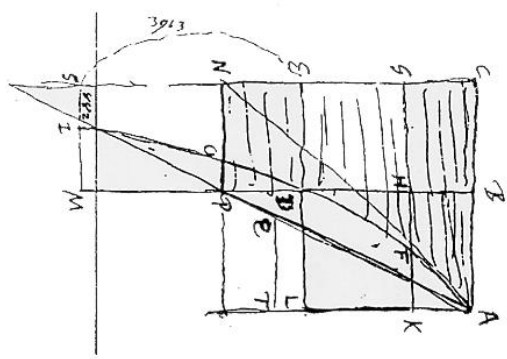


***Figure 12**. Huygens' geometrical model for fall in a medium with resistance proportional to velocity, 1668 (O.C. Vol 19 pp. 102)*

$$\rho = \frac{d^2y/dx^2}{(1+(dy/dx)^2)\sqrt{1+(dy/dx)^2}}.$$

One notes that the formula implies the same ingredients as Huygens' drawings: the tangents to the given curve (the derivative dy/dx), and the second order infinitesimals (the second order derivative d^2y/dx^2).

3.3. Models

Before I turn to my third and last example of Huygens' use of drawings I owe the reader a remark about the mathematical technicalities in my discussion of Huygens' drawings (the example below has even more). I am aware that I may well lose some readers for the good reason of lack of time for, or affinity with, the details of the material. I hope however that the text can still be used as a guideline in taking some time to look at the drawings, note their charm and esthetics, and imagine Huygens making them and pondering natural phenomena by means of the art of scientific drawing. These aspects, I feel, are in fact more important that the technical mathematical details. —

The remaining example concerns the motion of a body, falling under the influence of gravity through a medium with resistance proportional to the velocity of the moving body. Figure 12 shows Huygens' drawing in which he incorporated the four variables involved in the process, velocity, acceleration, time and resistance, as well as their mutual relations. I will use the letters v, a, t, and r respectively for these, but note that Huygens did not use these letters in his drawing. His drawing served the function of a 'mathematical model,' be it that at present we expect such a model to consist of a set of formulas giving the equations and/or differential equations, which describe the process. Huygens' model was a geometrical one.

In Figure 13 I have indicated the elements of his drawing corresponding to

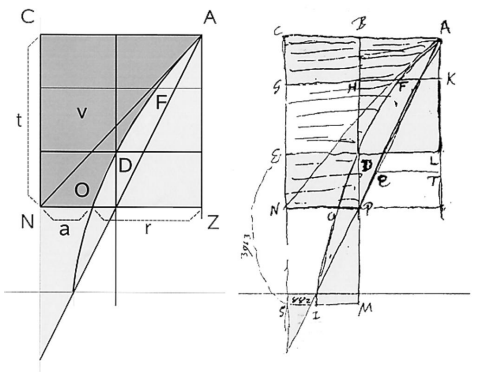


Figure 13. *Fall in a medium with resistance, the variables redrawn*

the four variables mentioned (I have added the letter Z for a point which in Huygens' drawing was not lettered):

- The time t is represented along a vertical axis AZ (or equivalently along CN),
- the velocity v by an area under an as yet unknown curve $AFDO\ldots$ with respect to the axes CA and CN,
- the acceleration a by the ordinate NO of the unknown curve (whereby the relation $a = dv/dt$ is incorporated in the drawing),
- and the resistance r turns out to be represented by the segment OZ.

The problem, then, is to determine the nature of the curve $AFDO\ldots$ from the given that the resistance r is proportional to the velocity v.

Here is how Huygens argued on the basis of his drawing: If there were no resistance the velocity would be proportional to the time, according to Galileo's law of fall. Thus the area v would be proportional to the time t, which implies that the curve from A coincides with the axis AZ. We conclude that, because there is resistance, the unknown curve must extend from A to the left of the axis, and that CA represents the acceleration if resistance is absent, that is the gravitational acceleration (modern: g). Moreover the curve cannot extend to the left of the axis CS because then the acceleration would be negative and the body would rise again. Thus the geometrical model directly provides a global insight in the process of fall with resistance.

Then Huygens incorporates the given that the resistance is proportional to the velocity. NO represents the acceleration of the body, which is the sum of the gravitational acceleration represented by CA and the (negative) acceleration caused by the resistance. Thus the resistance is represented by $CA - NO$, that is, by OZ. Hence the curve has the property that the difference $CA - NO$ between any of its ordinates and the first ordinate CA is proportional to the area between these two ordinates. Note that the argument until now

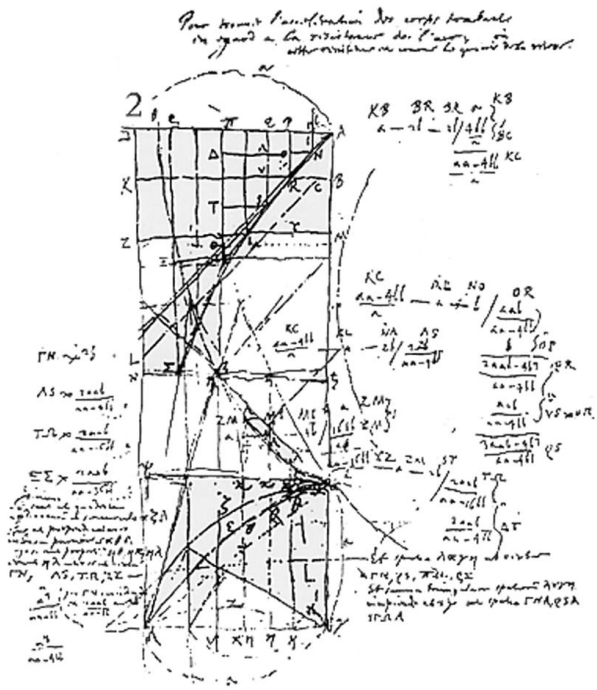


***Figure 14**. Fall in medium with resistance proportional to the square of the velocity, 1668 (O.C. Vol 19 pp. 159)*

corresponds to the derivation of the differential equation $dv/dt = g - \beta v$ from the Newtonian law $F = m \times a$ and the given proportionality $r \propto \beta v$. (The correspondence, however, is less straightforward than it may seem because the drawing models proportionalities rather than equalities.)

But a differential equation is no solution of the problem it describes; it has to be solved. Similarly Huygens' result about the unknown curve is not the answer to which curve it is. He did determine the curve however, because in earlier studies he had encountered a curve with the same property, namely the 'Logarithmica,' which was the seventeenth-century name of what now is called the exponential curve with equation $y = e^x$. Huygens' solution corresponds to the solution $v(t) = \frac{g}{\beta}(1 - e^{-\beta t})$ of the differential equation above. Finally, Figure 14 illustrates how Huygens could adapt his geometrical model with the four variable quantities involved in fall with resistance, to other assumptions about the relation between the resistance and the velocity. In this case he assumed the resistance to be proportional to the square of the velocity, and succeeded in determining the required curve. Thus the drawing illustrates the power of geometrical modeling in the hands of the master who pioneered this approach.

4. Conclusion

After this brief survey, how to characterize Huygens' mathematics? It was geometrical infinitesimal analysis of curves and of motion. As to inspiration and imagery it was inseparable from mechanics; in style it was pure mathematics. It was geometrical because it was essentially dependent on drawings for handling infinitesimals, limit processes, and motion.

Huygens brought this kind of mathematics to great heights. But this mathematics passed. The next generations changed the style: the drawings were replaced by formulas; the infinitesimal lines and strips were replaced by differential quotients dy/dx and integrals $\int ydx$; drawing figures was replaced by manipulation of formulas.

Newton and Leibniz set this transformation in motion. Huygens was the grand master of the previous style. In the long run, this style could not compete with the new, formula based, differential calculus in solving the problems that confronted mathematicians and mechanicists.

So Huygens was no longer the solution, and, as the saying goes, if you're not part of the solution, you're part of the problem. Something like this has indeed happened to him. Historians of science and modern scientists often experience Huygens' mathematics as problematical and they sometimes see the stylistic aspect of his mathematics as a deplorable detour from how it should have been. This is understandable, because his mathematics is indeed difficult; it takes time, and lack of time is a valid excuse for a historian to take a short-cut in the telling. But the idea that Huygens took a detour is nonsense. Geometrical analysis and physics was an essential and necessary phase in the development of mathematics.

Acknowledgements

This article was first published in

> K. Fletcher (ed.)(2004). *Proceedings of the International Conference 'Titan – From Discovery to Encounter'* Noordwijk, The Netherlands (ESA SP-1278).

and is used in this syllabus with kind permission of ESA.

Images from the manuscripts of Christiaan Huygens are taken from the *Codices Hugeniani* and have been reproduced with kind permission of the Universiteitsbibliotheek Leiden (UBL).

Images taken from *Œuvres Complètes de Christiaan Huygens* (22 volumes) are referenced by O.C. +volume number + page number. These images are reproduced with kind permission of the Hollandse Maatschappij der Wetenschappen, the publisher of the O.C.

CWI Vakantiecursus 2005

pp. 19 – 34

Hoeveel is voldoende?

A. Bolck
Nederlands Forensisch Instituut
e-mail: abolck@nfi.minjus.nl

Het Nederlands Forensisch Instituut (NFI) is een onderdeel van het ministerie van Justitie en doet technisch en natuurwetenschappelijk onderzoek ten behoeve van strafzaken. Zo worden bij het NFI onder andere DNA profielen bepaald, schoensporen onderzocht, computers gekraakt en maaginhouden van overleden slachtoffers geanalyseerd. Bij veel onderzoek wordt ook wiskunde gebruikt. Van de schijf van vijf betreft dit voornamelijk statistiek en kansrekening.

De kansrekening is met name belangrijk bij het rapporteren van de (altijd onzekere) resultaten. Zoals de kans dat een bepaald schoenspoor van een willekeurig persoon komt. Statistiek is vooral belangrijk in het wetenschappelijk onderzoek ter verbetering van methoden en technieken. Zoals de methoden bij het vergelijken van papiersoorten van dreigbrieven. Een ander aspect waarbij statistiek en kansrekening een rol spelen is het bepalen van het aantal te analyseren monsters. In dit stuk zal ik ingaan op hoe het aantal te analyseren eenheden uit een partij discrete eenheden bepaald kan worden, en hoe dit in de forensische praktijk (met name op het gebied van illegale drugs) gebeurt.

Laten we ons verplaatsen naar een inval in een woonhuis. De politie stuit hierbij op een grote partij verdachte eenheden (bijvoorbeeld pillen of CD's). Zij kunnen niet ter plekke met redelijke zekerheid bepalen of en om welk illegaal materiaal het hier gaat. En al helemaal niet in welke hoeveelheid. Hooguit kunnen ze op basis van hun ervaring of de aanwezigheid van andere materialen vermoeden dat het om illegaal materiaal gaat. Soms kunnen ze ook een paar simpele testen doen om hun vermoedens te versterken. Zo kunnen CD's in een eventueel meegebracht notebook snel worden bekeken. Een zogenaamde kleurentest kan snel een indicatie geven of er sprake kan zijn van illegale middelen volgens de opiumwet. Om definitief vast te stellen of eenheden illegaal materiaal bevatten, om welk materiaal het gaat en eventueel in welke hoeveelheid zal de vondst door experts onderzocht moeten worden in laboratoria zoals het Nederlands Forensisch Instituut.

Het is vaak onmogelijk of onpraktisch om alle eenheden naar laboratoria te sturen dus de politie moet monsters nemen. Behalve dat dit representatief moet gebeuren moet er een beslissing genomen worden over hoeveel monsters er genomen moeten worden. In het vervolg ga ik er voor het gemak even van uit dat de partijen homogeen zijn en steekproeven a select worden getrokken.

Figuur 1. *Een verdachte partij pillen*

1. Arbitraire methoden

Door de jaren heen heeft ieder land en ieder vakgebied eigen methoden ontwikkelt voor het bepalen van het aantal te nemen monsters. Zo ontstond in de jaren 20, uit de behoefte van Amerikaanse landbouwinspecteurs aan een eenvoudige goed te onthouden regel, de zogenaamde *wortelregel* (Izenman 2001, Colon e.a. 1992). Volgens deze regel moeten de inspecteurs een aantal monsters nemen dat (afgerond) gelijk is aan de wortel uit het totaal aantal eenheden. Dit betekent dat van een partij van 150 verdachte eenheden er ongeveer 12 daadwerkelijk onderzocht moeten worden. Deze regel is tegenwoordig ook buiten de landbouw erg populair in veel landen inclusief Nederland.

Een andere bekende regel is de vijf procentregel (of tien procentregel) waarbij de steekproefgrootte n op 0,05N (of 0,1N) wordt gesteld. Hierbij moet het aantal monsters dus gelijk zijn aan 5% (of 10%) van het totaal aantal eenheden.

Ook varianten en andere regels komen voor (Izenman 2001, Colon e.a.1992). De United Nations International Drug Control Programme (UNDCP 1998) adviseert o.a.:

- In het geval van partijen met 10 of minder eenheden: Onderzoek de hele partij
- In het geval van partijen met 10 tot 100 eenheden: Onderzoek 10 eenheden
- In het geval van partijen met meer dan 100 eenheden: Onderzoek $\sqrt{N}$ eenheden.

Een nadeel van al deze methoden is dat het aantal te nemen monsters erg groot wordt als de partij uit zeer veel eenheden bestaat. Bij een partij van 10.000 eenheden (wat bij drugs niet zeldzaam is) zouden volgens de wortelmethode 100 eenheden onderzocht moeten worden en volgens de Partijen van 100.000 eenheden of meer leveren dan helemaal problemen op. Dergelijke grote steekproeven zijn ook helemaal niet nodig, zoals later in dit stuk zal blijken. Daarnaast hebben alle bovenstaande regels als nadeel dat ze niet wetenschappelijk onderbouwd zijn. De monsters geven een bepaald beeld van de partij en achteraf, na de monstername, valt wel een uitspraak te doen over hoe goed

de monsters de partij weergeven en met welke foutenmarge, maar het aantal monsters is niet bepaald op van tevoren vastgestelde criteria over hoe goed de monsters met welke betrouwbaarheid de partij moeten weergeven. Het zijn slechts vuistregels die in de praktijk zijn ontstaan omdat ze makkelijk in het gebruik zijn en eenvoudig te onthouden.

2. Statistisch onderbouwde methoden voor het bepalen van de monstergrootte

Statistisch onderbouwde methoden voor het bepalen van de steekproefgrootte laten het aantal te analyseren monsters afhangen van een minimaal vereist percentage illegale eenheden dat met een bepaalde gewenste betrouwbaarheid kan worden gegarandeerd in de partij indien alle monsters illegaal zijn. Er worden bijvoorbeeld zoveel monsters genomen dat met 95% betrouwbaarheid gegarandeerd kan worden dat tenminste 90% van de eenheden illegaal materiaal bevat, als inderdaad alle monsters illegaal blijken te zijn.

Bij arbitraire methoden kun je achteraf ook een betrouwbaarheid uitrekenen. Het verschil tussen arbitraire en statistisch gefundeerde methoden is echter dat bij de laatste de steekproefgrootte gebaseerd is op een gekozen betrouwbaarheid dat de partij tenminste een bepaald van tevoren vastgesteld percentage illegale eenheden bevat, terwijl de betrouwbaarheid van de arbitraire methoden fluctueert en nergens op gebaseerd is.

Er zijn twee typen statistisch onderbouwde methoden gangbaar, frequentistische methoden en Bayesiaanse methoden. Ze komen beide aan bod.

3. Frequentistische methoden

3.1. De hypergeometrische verdeling

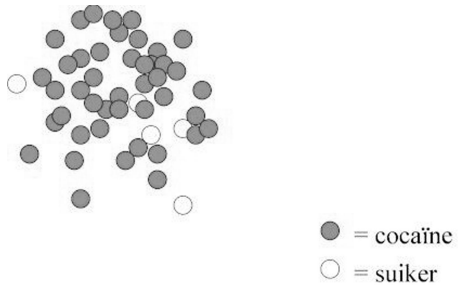
De hypergeometrische verdeling is een discrete kansverdeling. Deze verdeling geeft bij trekken zonder terugleggen de kans op X positieven in een steekproef die getrokken wordt uit een populatie van N eenheden (bv. pillen), met N_1 positieven (bv. pillen die drugs bevatten) en de rest ($= N - N_1$) negatieven (pillen die geen drugs bevatten):

$$P(X = x|N_1, N, n) = \frac{\binom{N_1}{x}\binom{N - N_1}{n - x}}{\binom{N}{n}}$$

In een zak met 50 pillen, waarvan er 45 cocaïne bevatten en 5 alleen suiker is de kans dat een steekroef van 5 precies 4 pillen met cocaïne bevat dus

$$P(X = 4|N_1 = 45, N = 50, n = 5) = \frac{\binom{45}{4}\binom{50 - 45}{5 - 4}}{\binom{50}{5}} = 0,3516$$

De kans dat alle 5 pillen in de steekproef drugs bevatten is gelijk aan 0,5766 oftewel 57,66%. En de kans dat tenminste 4 van de 5 pillen illegaal zijn 0,3516 + 0,5766 = 0,9282. Omgekeerd kun je uitrekenen dat de kans op maximaal 1 illegale pil slechts 0,01% is.



***Figuur 2**. 50 pillen, waarvan 45 pillen cocaïne bevatten en 5 alleen suiker*

In de praktijk van de inbeslagname van verdachte eenheden is echter niet bekend wat de proportie illegale eenheden (N_1) in de populatie is. Op basis van een steekproef wil men hier juist een uitspraak over doen. Of beter gezegd men wil weten hoe groot de steekproef moet zijn om een betrouwbare uitspraak te doen over het percentage illegale eenheden in een partij. De vraag daarbij is hoe groot de steekproef moet zijn om dit te kunnen garanderen.

Hiervoor moet de hypergeometrische verdeling op een iets andere manier gebruikt worden dan hierboven. Om mensen te kunnen veroordelen moet men kunnen aantonen dat het aantal illegale eenheden in de populatie tenminste gelijk is aan een van te voren bepaald aantal K, oftewel, het is gewenst dat $N_1 \geq K$. Daarom wordt het aantal te onderzoeken monsters berekend door het toetsten van de nul hypothese dat het aantal illegale eenheden in de populatie ter grootte van N kleiner is dan K. De alternatieve hypothese is dat het aantal illegale eenheden groter of gelijk is aan K;

$$H_0 \ : \ N_1 < K;$$

$$H_1 \ : \ N_1 \geq K.$$

Dit betekent dat er bewijs gevonden moet worden om de nul hypothese te verwerpen, zodanig dat de kans (α) dat de nul hypothese ten onrechte wordt verworpen klein is. Zoals gebruikelijk bij toetsen wordt voor deze fout 5% genomen, of in sommige gevallens 1%. De betrouwbaarheid is dan respectievelijk 95% of 99%, en in het algemeen $(1-\alpha)$ 100%.

De hypotheses worden getoetst met het aantal illegale eenheden dat gevonden wordt in de steekproef (X) als toetsgrootheid. De nul hypothese wordt verworpen als dit aantal X groter is dan een bepaald (verwacht)kritisch aantal x. Dan moet er voor de steekproefgrootte n het laagste aantal gekozen worden waarvoor geldt dat

$$P(X \geq x | N_1 < K) \leq \alpha.$$

De hypergeometrische verdeling is een functie die daalt als N_1 afneemt, daarom zullen alle afzonderlijk kansen met waarden voor N_1 die kleiner zijn dan K maximaal gelijk zijn aan de kans met de hoogste waarde kleiner dan K, dat is $K-1$. Oftewel

$$P(X \geq x | N_1 < K) \leq P(X \geq x | N_1 = K-1) \Rightarrow$$
$$P(X \geq x | N_1 < K) \leq \sum_{i=x}^{n} \frac{\binom{K-1}{i}\binom{N-K+1}{n-1}}{\binom{N}{n}} \leq \alpha.$$

Als men verwacht dat alle onderzochte eenheden in de steekproef illegaal zijn (oftewel $x = n$) dan geldt:

$$\frac{\binom{K-1}{n}\binom{N-K+1}{0}}{\binom{N}{n}} \leq \alpha$$
$$\Rightarrow$$
$$P_0 = \frac{(K-1)!(N-n)!}{(K-n-1)!N!} = \frac{(K-1)(k-2)\cdots(K-n)}{N(N-1)\cdots(N-n+1)} \leq \alpha.$$

In het geval van 1 negatieve (= 1 legale eenheid) in de steekproef leidt dit tot:

$$P_0 \left[1 + \frac{n(N-K+1)}{(K-n)}\right] \leq \alpha$$

en in het geval van 2 negatieven leidt dit tot:

$$P_0 \left[1 + \frac{n(N-K+1)}{(K-n)} \left\{1 + \frac{(n-1)(N-K)}{2(K-n+1)}\right\}\right] \leq \alpha,$$

enzovoort.

Voor de steekproefgrootte bepaald kan worden moet dus een inschatting gemaakt worden over de hoeveelheid positieven, oftewel illegale eenheden, die men verwacht in de steekproef. Met andere woorden de hypergeometrische verdeling zit zo in elkaar dat daarmee het aantal te analyseren monsters n te berekenen is zodanig dat met bijvoorbeeld 95% betrouwbaarheid gegarandeerd kan worden dat bijvoorbeeld 90% van de eenheden illegaal is, gegeven dat x van de n monsters illegaal zijn.

De praktijk heeft geleerd dat een partij meestal helemaal illegaal is of helemaal niet. Het komt niet vaak voor dat slechts een gedeelte illegaal is. Dus als er verdachte omstandigheden zijn waardoor men illegale goederen vermoedt, verwacht men meestal dat een hele partij illegaal is en, als gevolg daarvan, dat alle monsters illegaal zijn. Bijvoorbeeld in het geval van een bolletjesslikker is dit heel aannemelijk. Het is niet waarschijnlijk dat een bolletjesslikker naast bolletjes heroïne ook bolletjes zetmeel heeft geslikt. Of dit inderdaad zo is zal blijken na de analyse van de steekproef. Maar eerst is dan de hypergeometrische verdeling gebruikt om de steekproefgrootte te bepalen onder de aanname dat alle monsters illegaal zijn. Het aantal monsters wordt gekozen zodanig dat als ze allemaal illegaal blijken te zijn er met een betrouwbaarheid van $(1-\alpha)100\%$ (meestal 95%) gezegd kan worden dat tenminste $k100\%$ (bijvoorbeeld 50% of 90%, dan is $k = 0,5$ of 0,9) in de partij illegaal is. Als dan toch maar een bepaald gedeelte (of geen enkele) van de monsters in de steekproef illegaal blijken te zijn dan zal een grotere steekproef genomen moeten worden of moet de

betrouwbaarheid of het aangetoonde percentage illegale eenheden in de partij worden aangepast. Dit zal ik later laten zien.

In Singapore werd in juli 1996 een Nederlander opgepakt die 2238 pillen bij zich had (*zie onder andere de NRC van 12–7–96, 16–7–96, 30–8–96, 2–9–96, 3-9–96, 4–9–96 en 13–9–96*). Bij een visuele check bleken alle pillen er hetzelfde uit te zien. In totaal zijn 212 van de 2238 pillen chemisch getest in een laboratorium. Dit bleken allemaal XTC pillen te zijn. De Nederlander moest voorkomen op verdenking van de smokkel van 2238 XTC pillen, maar de verdediging vond dat de aanklacht veranderd moest worden in smokkel van 212 pillen, omdat slechts van 212 pillen daadwerkelijk was aangetoond dat het om XTC ging. Op die manier zou de Nederlander een veel lichtere straf krijgen dan wanneer hij voor alle 2238 pillen veroordeeld zou worden.

Hieruit blijkt weinig begrip voor de statistiek. Inderdaad weet je maar van 212 tabletten 100% zeker dat het om XTC ging, maar de betrouwbaarheid is zeer hoog dat minstens de overgrote meerderheid van de tabletten ook XTC is. Met behulp van de hypergeometrische verdeling valt uit te rekenen dat bij een steekproef van 212 waarbij alle 212 pillen XTC blijken te bevatten met een betrouwbaarheid van 99% gezegd kan worden dat tenminste 2193 pillen (=98%) XTC bevatten. Daarnaast is de kans dat uit een partij van 2238 pillen er 212 worden getrokken die als enige allemaal XTC bevatten $9,742 \times 10^{-304}$. Dit is onvoorstelbaar klein.

Tabel 1 op bladzijde 25 geeft bij betrouwbaarheden van zowel 95% als 99% en drie verschillende percentages minimaal te garanderen illegaal materiaal aan wat de steekproefgrootte moet zijn op basis van de hypergeometrische verdeling bij verschillende populatiegroottes als wordt aangenomen dat alle monsters illegaal zullen zijn (oftewel dat er geen negatieve monsters zullen zijn).

Een paar jaar geleden zijn op Schiphol bankbiljetten in beslag genomen waarvan men vermoedde dat daar een meer dan gebruikelijke hoeveelheid cocaïne opzat. Bankbiljetten bevatten altijd sporen van cocaïne omdat ze door veel handen, en dus ook criminele handen, gaan. Soms worden de biljetten echter ook gebruikt voor cocaïne smokkel. Stel er zijn 1000 verdachte bankbiljetten. Dat betekent dat (zie Tabel 1) 28 bankbiljetten onderzocht moeten worden om met 95% betrouwbaarheid te garanderen dat tenminste 90% van de bankbiljetten (dat is tenminste 900) een meer dan gebruikelijke hoeveelheid cocaïne bevatten als de 28 monsters allemaal een meer dan gebruikelijke hoeveelheid cocaïne blijken te bevatten.

Stel er is met betrekking tot XTC besloten dat men het voor partijen boven de 100 pillen voldoende vindt om met 95% betrouwbaarheid de garantie te krijgen dat tenminste de helft van een partij pillen XTC bevat. In dat geval is voor partijen van enkele honderden of duizenden mogelijke XTC pillen, een steekproef van 5 genoeg, als men er vanuit gaat dat alle 5 XTC zullen bevatten. Dit is het getal wat de ENFSI adviseert en wat ook in Tabel 1 is terug te vinden voor partijen groter dan 100 eenheden.

In het zeldzame geval dat er een vermoeden is dat niet alle eenheden in de partij illegaal zijn, bijvoorbeeld omdat de politie dit in een simpele test heeft geconstateerd, kan het aantal verwachte negatieven (niet illegale eenheden) in

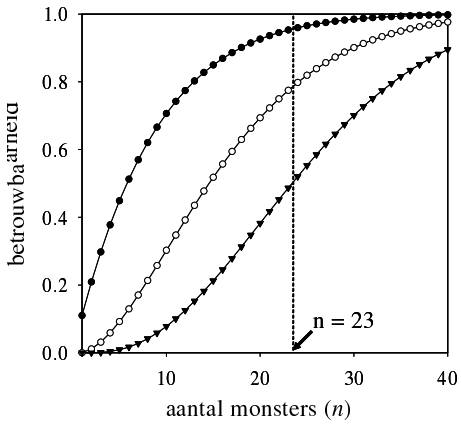
populatie-	95% betrouwbaarheid			99% betrouwbaarheid		
grootte N	$k = 0,5$	$k = 0,8$	$k = 0,9$	$k = 0,5$	$k = 0,8$	$k = 0,9$
10	3	6	8	4	7	9
20	4	9	12	5	11	15
30	4	10	15	6	13	20
40	4	11	18	6	15	23
50	4	11	19	6	16	26
60	4	12	20	6	17	28
70	5	2	21	7	17	30
80	5	12	22	7	18	31
90	5	12	23	7	18	32
100	5	12	23	7	18	33
200	5	13	26	7	20	38
300	5	13	27	7	20	40
400	5	14	27	7	20	41
500	5	14	28	7	20	41
600	5	14	28	7	21	42
700	5	14	28	7	21	42
800	5	14	28	7	21	42
900	5	14	28	7	21	43
1000	5	14	28	7	21	43
5000	5	14	29	7	21	44
10000	5	14	29	7	21	44

Tabel 1. *Hypergeometrische verdeling. Steekproefgrootte om met 95% of 99% betrouwbaarheid te garanderen dat $k100\%$ in de partij illegaal is, als de hele steekproef illegaal is (0 negatieven)*

de steekproef op 1 of 2 (of welk willekeurig aantal dan ook) gesteld worden. Ook hier zijn natuurlijk weer tabellen of rekenprogramma's (excel sheets) voor te maken.

Bij bankbiljetten is het misschien niet vreemd een aantal negatieven te verwachten. Als men uitgaat van 2 negatieven in de steekproef dan valt uit te rekenen dat 51 monsters van de 1000 bankbiljetten geanalyseerd moeten worden om nog steeds met 95% betrouwbaarheid te garanderen dat tenminste 90% van de bankbiljetten een meer dan normale hoeveelheid cocaïne bevatten. Dan moeten ook inderdaad precies 2 van de 51 monsters negatief zijn. Zijn er meer of minder negatieven dan zal de steekproefgrootte of de betrouwbaarheid of het minimale percentage moeten worden bijgesteld (zie Figuren 3 en 4).

Figuur 3 geeft de relatie tussen de betrouwbaarheid en het aantal gekozen monsters, bij een populatie grootte van 100 en een minimaal te garanderen proportie illegale eenheden van 90%, in het geval dat 0, 1 of 2 negatieven worden verwacht in de steekproef. Hieruit valt te lezen dat voor een betrouwbaarheid van 95% 23 monsters nodig zijn, indien er geen negatieven in de steekproef worden verwacht (bovenste curve). Als achteraf blijkt dat toch 1 van de 23



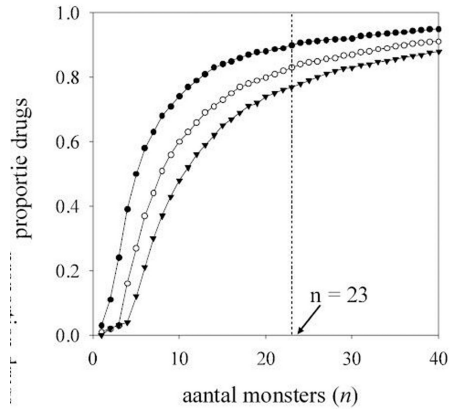
Figuur 3. *Betrouwbaarheid van een steekproef afgezet tegen de steekproefgrootte* (n) *in een populatie te grootte van* $N = 100$ *waarbij minimaal 90% illegaal materiaal moet zijn gegarandeerd bij respectievelijk 0 negatieven* $(-\bullet)$, *1 negatief* $(-\circ-)$ *en 2 negatieven* $(-\blacktriangledown-)$

monsters niet illegaal is, dus negatief is, levert de middelste curve de werkelijke betrouwbaarheid bij de steekproefgrootte van 23. Deze betrouwbaarheid is ongeveer 78%. Als 2 van de 23 monsters negatief blijkt te zijn daalt de betrouwbaarheid naar 48% om nog steeds een percentage van 90% illegalen in de partij te garanderen.

Het is ook mogelijk het percentage te garanderen illegale eenheden te verminderen en de betrouwbaarheid constant te houden. Dit valt te zien in Figuur 4. Bij het constateren van 1 negatief onder 23 monsters daalt het percentage minimaal te garanderen illegale eenheden van 90% naar 83% als de betrouwbaarheid 95% blijft. Bij 2 negatieven daalt het percentage zelfs naar 77%.

Daarnaast is het natuurlijk ook mogelijk zowel de betrouwbaarheid als het te garanderen percentage illegale eenheden aan te passen als het aantal illegale eenheden in de monsters anders blijkt dan verwacht.

In 1997 ontving het NFI (toen nog het Gerechtelijk Laboratorium) van de technische recherche Zaanstreek-Waterland 403 verpakkingen met poeders en brokjes. Op basis van uiterlijke kenmerken werden 20 groepen onderscheiden waaruit uiteindelijk in totaal 40 monsters werden onderzocht. Deze bleken alle cocaïne te bevatten. Uit eerder politieonderzoek was echter gebleken dat zich ook een aantal negatieven in de partij zouden bevinden. Deze werden in de steekproef niet gevonden. Er valt voor iedere samenstelling van de partij met de hypergeometrische verdeling uit te rekenen wat de kans is dat geen negatieven in de steekproef van 40 worden gevonden. Een aantal voorbeelden staan in Tabel 2. Hieruit valt op te maken dat als de partij voor de helft uit negatieven bestaat, dat dan de kans dat alle 40 monsters positief zijn nagenoeg nihil is. Ook als 100 van de 403 in de partij negatief zijn is deze kans klein. Pas



Figuur 4. *Proportie minimaal te garanderen illegale eenheden* $(K\%)$ *afgezet tegen de steekproefgrootte* (n)*, bij een populatiegrootte van* $N = 100$*, en een betrouwbaarheid van 95% bij respectievelijk 0 negatieven* $(-\bullet)$*, 1 negatief* $(-\circ-)$ *en 2 negatieven* $(-\blacktriangledown-)$

Samenstelling		
Cocaïne	negatieve monsters	kans op geen negatieve in de steekproef
202	201	0,0000000000118%
303	100	0,000556%
353	50	0,372%
360	43	0,855%
363	40	1,21%
373	30	3,83%
383	20	11,71%
393	10	34,72%

Tabel 2.

als er hoogstens 42 negatieven in de partij van 403 zitten stijgt de kans dat de steekproef geen negatieven bevat boven de 1woorden met een betrouwbaarheid van 99% kan gesteld worden dat de partij dan ten hoogste 42 negatieven bevat. Dus als in een steekproef van 40 geen negatieven worden gevonden dan kun je stellen met een betrouwbaarheid van 99% dat maximaal ongeveer 10% van de partij negatief is.

Bij de afdeling verdovende middelen van het Nederlands Forensisch Instituut wordt voor grote partijen de richtlijn gebruikt dat een steekproef groot genoeg moet zijn om met 99% betrouwbaarheid te garanderen dat tenminste 80% van de eenheden illegaal is. Men heeft gekozen voor 80% omdat het vaak belangrijker is dat iets drugs bevat, dan dat alles drugs bevat. Bovendien blijkt uit ervaring dat als enkele eenheden drugs bevatten meestal alle eenheden drugs

bevatten. Daarbij is 80% dus nog aan de voorzichtige kant gekozen, een meerderheid (50%) zou volgens deze redenering ook voldoende moeten zijn. Wel vond men het heel belangrijk dat de resultaten betrouwbaar zijn. Vandaar dat men voor 99% in plaats van de ook heel gebruikelijke 95% heeft gekozen. In de praktijk komt het erop neer dat in principe 20 monsters worden genomen (en vaak ook nog eens 20 monsters als reserve worden achtergehouden). Dit wordt door het NFI ook aan de politie geadviseerd op cursussen over het nemen van monsters. In Tabel 1 valt te zien waar dit getal 20 vandaan komt. Bij $N = 80$ tot en met $N = 100$ zijn 18 monsters nodig om met 99Bij 200-500 eenheden zijn dit 20 monsters en daarboven 21 monsters. Deze getallen worden voor het gebruiksgemak afgerond tot 20. Zelfs als de partij dan uit 10.000 eenheden bestaat is de betrouwbaarheid misschien niet exact 99%, maar toch wel 98,8% dat op basis van 20 positieve monsters gesteld kan worden dan ten minste 80% van de partij drugs bevat.

Dit getal 20 werkt in de praktijk erg goed. Het is nog makkelijker hiermee te werken dan bijvoorbeeld met de wortelregel, waar toch eerst nog de wortel van het aantal verdachte eenheden berekend moet worden. Bovendien zijn bij bijvoorbeeld 10.000 eenheden nog steeds maar 20 monsters nodig terwijl bij de wortelmethode dan 100 monsters vereist zijn. Maar het belangrijkste is dat het getal 20 gebaseerd is op een statistisch model. Het is niet zomaar gekozen, maar gekozen op basis van een van tevoren gekozen betrouwbaarheid die een van tevoren gewenst minimaal aantal illegale eenheden garandeert volgens de hypergeometrische verdeling. In de meeste gevallen zullen alle 20 eenheden illegale bestanddelen volgens de opiumwet bevatten. Er wordt dan gerapporteerd dat hier sprake is van een partij illegale eenheden, met de bijbehorende betrouwbaarheid en proportie. Indien er ook negatieven worden gevonden wordt er daarnaast een schatting gerapporteerd van het aantal illegale en niet illegale eenheden.

In 2003 verscheen uit een samenwerking tussen verschillende Europese forensische laboratoria een rapport met 'Guidelines on representative Drugsampling' (ENFSI 2004). Hierin wordt het advies gegeven in standaard situaties 5 monsters te nemen. Ook dit is een eenvoudig te onthouden getal. Daarbij is er van uit gegaan dat de garantie dat met 95% betrouwbaarheid tenminste de helft van de partij illegaal is voldoende is. Landen die strenger willen zijn of situaties die garantie op een groter percentage vereisen (bijvoorbeeld zeer grote vondsten met honderdduizenden pillen) kunnen dan alsnog met behulp van de hypergeometrische berekening en de gewenste minimaal te garanderen proportie illegale eenheden de benodigde steekproefgrootte berekenen. Ook als men negatieven vermoedt zal de steekproef groter moeten zijn dan 5. Het ENFSI rapport adviseert hier ook over.

4. De binomiale verdeling

De hypergeometrische verdeling gaat uit van trekken zonder terugleggen. Als een eenheid eenmaal is geanalyseerd wordt deze niet weer teruggelegd in de populatie om eventueel opnieuw getrokken en geanalyseerd te worden. Zou dat wel het geval zijn dan is er sprake van trekken met terugleggen en moet de

binomiale verdeling gebruikt worden.

Net als bij de hypergeometrische verdeling is de binomiale verdeling in eerste instantie bedoeld om de kans op het aantal eenheden met een bepaalde eigenschap (positieven) in een steekproef ter grootte van n te bepalen, gegeven dat de proportie positieven in de populatie $\theta = \frac{N_1}{N}$ is:

$$P(X = x|\theta, n) = \binom{n}{x}\theta^x(1-\theta)^{n-x}.$$

De binomiale verdeling kan echter ook gebruikt worden om de steekproefgrootte n te berekenen zodanig dat met een betrouwbaarheid van $(1-\alpha)100\%$ een populatie proportie van tenminste k100% gegarandeerd kan worden.

Ook hier gebeurt dat met behulp van het statistische toetsen van twee hypotheses:

$$H_0 \ : \ \theta < k;$$
$$H_1 \ : \ \theta \geq k$$

en wederom is de steekproefgrootte n het kleinste getal waarvoor $P(X \geq x|\theta < k) \leq \alpha$. In dit geval betekent dat n zodanig gekozen moet worden dat:

$$P(X \geq x|\theta < k) = \sum_{i=x}^{n} \binom{n}{x}\theta^x(1-\theta)^{n-x} \leq \alpha.$$

In het geval dat $x = n$ betekent dat n het kleinste getal is zodanig dat

$$\theta^n \leq \alpha.$$

Als je dit herschrijft krijg je:

$$n \leq \frac{\log \alpha}{\log \theta}.$$

De waarde van n die hier aan voldoet geldt voor alle N.

De binomiale verdeling kan gebruikt worden als benadering van de hypergeometrische verdeling in het geval van grote populaties, waarbij de steekproef relatief klein is ten opzichte van de populatie. In de praktijk wordt deze benadering al gebruikt bij populatie groter dan 50 eenheden. De binomiale verdeling wordt dan gebruikt omdat het een eenvoudigere verdeling is die bovendien steekproefgroottes oplevert die onafhankelijk zijn van de populatiegrootte. Echter het met deze benadering gevonden gewenste aantal te analyseren monsters is altijd minimaal even groot als dat wordt gevonden met de hypergeometrische verdeling.

Om met een betrouwbaarheid van 95% te garanderen dat ten minste 90 van de 100 bolletjes uit een bolletjesslikker cocaïne bevat zijn 29 monsters nodig (Tabel 3) in plaats van de 23 die met de hypergeometrische verdeling gevonden wordt. In Tabel 3 zie je ook meteen dat ongeacht de populatiegrootte een steekproef van 21 voldoende is om met 99% betrouwbaarheid tenminste 80% illegaal materiaal in de partij te garanderen als alle onderzochte monsters illegaal blijken te zijn. Dit is dus een goede benadering van het getal 20 dat de afdeling verdovende middelen van het NFI gebruikt.

	95% betrouwbaarheid			99% betrouwbaarheid		
	$k = 0,5$	$k = 0,8$	$k = 0,9$	$k = 0,5$	$k = 0,8$	$k = 0,9$
0 negatief	5	14	29	7	21	44
1 negatief	7	22	46	11	31	64
2 negatief	11	30	61	14	39	81

Tabel 3. *Bionomiale verdeling. Steekproefgrootte om met 95% of 99% betrouwbaarheid te garanderen dat een proportie k in de partij illegaal is, bij 0, 1 of 2 negatieven.*

5. De Bayesiaanse benadering

Bayesiaanse methoden (Aitkin 1997, Aitkin 1999, Aitkin 2000) om de steekproefgrootte te bepalen hebben net als frequentistische methoden een statistische basis. De steekproefgrootte wordt zo bepaald dat met een zekere kans gegarandeerd kan worden dat de populatie tenminste een gewenste proportie illegaal materiaal bevat indien de steekproef een bepaalde verwacht aantal illegalen bevat. Er kan nu van kans worden gesproken in plaats van betrouwbaarheid omdat de Bayesiaanse methoden ervan uitgaan dat de partijsamenstelling (θ) en niet de steekproefresultaten (x) een statistische verdeling volgt.

Naast een andere benadering van steekproeven en populaties verschillen Bayesiaanse methoden van frequentistische methoden door het gebruik van voorkennis. Als alle illegale eenheden er hetzelfde uitzien, dezelfde geur hebben, hetzelfde wegen e.d., dan is de kans dat de partij allemaal hetzelfde materiaal bevat groter dan wanneer alle eenheden er verschillend uitzien, geuren, wegen etc. Dergelijke informatie wordt meegenomen in de berekening van een kans op een bepaalde samenstelling van de partij.

De voorkennis die men veronderstelt bij Bayesiaanse methoden kan gegoten worden in de vorm van een prior verdeling voor de onbekende proportie illegale eenheden θ in de partij $f\theta$. Ook als er helemaal geen voorkennis is of beschouwd mag worden, kan dit in een (neutrale) prior worden verwerkt. De prior verdeling wordt gecombineerd met de likelihood functie, die informatie over de steekproef bevat (x), tot een posterior kansverdeling voor θ:

$$P(\theta|x) \propto L(\theta|x)f(\theta)$$

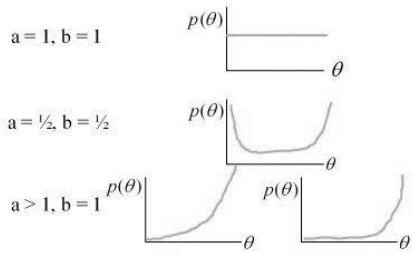
Voor de prior wordt meestal een bèta verdeling verondersteld:

$$f(\theta|a,b) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)},$$

Hierbij is $B(a,b)$ de bèta functie $\int_0^1 y^{a-1}(1-y)^{b-1}dy$.

In het geval dat men geen voorinformatie veronderstelt, kunnen a als b gelijk aan 1 worden genomen. In dat geval is de bèta verdeling namelijk gelijk aan de uniforme verdeling. Als men meer informatie heeft, bijvoorbeeld dat alle eenheden er hetzelfde uitzien (zelfde planten of zelfde pillen e.d.) dan kan men andere waardes voor a en b nemen. Als bijvoorbeeld alle pillen er

hetzelfde uitzien is de kans groot dat alle pillen drugs bevatten of dat geen enkele pil drugs bevat. In dat geval kan men het beste zowel a als b gelijk aan $1/2$ nemen. Als er een gefundeerd vermoeden is dat men een partij illegaal materiaal te pakken heeft, bijvoorbeeld door de vindplaats (vaten die op een verdachte plek zijn gedumpt), is de kans groot dat θ hoog is. In dat geval kan men het beste b op 1 vast stellen en voor a 3 kiezen of zelfs 10 of nog hoger, al naar gelang men zekerder is over de illegaliteit



Figuur 5. *De bèta verdeling bij verschillende waarden van a en b*

De likelihood functie bevat informatie uit de steekproef. Het is dezelfde kansverdeling die voor het aantal illegale eenheden in de steekproef wordt gebruikt in de frequentistische benadering als de populatie groot is (de binomiale verdeling), behalve dat nu de proportie illegaal materiaal in de populatie (θ) variabel wordt verondersteld gegeven het aantal illegale eenheden in de steekproef (x). Dit laatste aantal wordt juist constant veronderstelt.

De likelihood functie combineert samen met de prior verdeling tot de posterior verdeling van de proportie illegaal materiaal in de populatiegegeven de steekproefresultaten x.

$$f(\theta|x,n,a,b) = Be(x+a, n-x+b) = \frac{\theta^{x+a-1}(1-\theta)^{n-x+b-1}}{B(x+a,n-x+b)}.$$

Dit is een bèta verdeling (Be) met parameters $x+a$ en $n-x+b$.

De kans dat de populatieproportie tenminste $k(=K/N)$ is, kan nu berekend worden met. En dat kan weer gebruikt worden om de steekproefgrootte n te bepalen waarvoor deze kans tenminste $(1-\alpha)$ is.

Als we er van uitgaan dat alle geanalyseerde monsters illegaal materiaal bevatten dan geldt:

$$f(\theta|n,n,a,b) = Be(n+a,b) = \frac{\theta^{n+a-1}(1-\theta)^{b-1}}{B(n+a,b)}.$$

Om in dit geval de steekproefgrootte n te bepalen zodat de kans $(1-\alpha)100\%$ is dat tenminste $k100\%$ van alle eenheden in de partij illegaal is, moet n zo gekozen worden dat:

$a=1$	95% kans			99% kans		
$b=1$	$k=0,5$	$k=0,8$	$k=0,9$	$k=0,5$	$k=0,8$	$k=0,9$
0 negatief	4	13	28	6	20	28
1 negatief	7	21	45	9	30	63
2 negatief	10	29	60	13	38	80

$a=3$	95% kans			99% kans		
$b=1$	$k=0,5$	$k=0,8$	$k=0,9$	$k=0,5$	$k=0,8$	$k=0,9$
0 negatief	2	11	26	4	18	41
1 negatief	5	19	43	8	28	61
2 negatief	8	27	58	11	36	78

$a=0,5$	95% kans			99% kans		
$b=0,5$	$k=0,5$	$k=0,8$	$k=0,9$	$k=0,5$	$k=0,8$	$k=0,9$
0 negatief	3	9	18	5	15	32
1 negatief	6	18	38	9	26	55
2 negatief	9	27	54	12	35	73

Tabel 4. *Bèta verdeling (met parameters $x+a$ en $n-xb$). Steekproefgrootte om met 95% of 99% kans te garanderen dat een proportie k in de partij illegaal is, bij $0,1$, or 2 negatieven in de steekproef.* ($N>50$). *Gebruik $a=1$ en $b=1$ als er geen voorkennis aangenomen wordt, $a=0,5$ en $b=0,5$ als redelijkerwijs aangenomen kan worden dat óf alles legaal is óf alles illegaal en $a=3$ en $b=1$ (of extremere waarden voor a) als er een reden is om aan te nemen dat het merendeel van de partij illegaal is.*

$$P(\theta > k|n,n,a,b) = \int_k^1 \theta^{n+a-1}(1-\theta)^{b-1}d\theta/B(n+a,b) = (1-\alpha).$$

In het geval van kleine partijen zijn de berekeningen allemaal iets anders. Er wordt dan een bèta-binomiale verdeling ipv een bèta verdeling gebruikt. Het idee is verder hetzelfde. Ik zal hier niet verder over uitweiden.

Bij een bolletjesslikker wil men zoveel bolletjes analyseren dat men met 95% kans kan garanderen dat tenminste 90% van de bolletjes cocaïne bevat. Zonder enige voorkennis over de bolletjes zal men in de Bayesiaanse benadering 28 bolletjes moeten analyseren, tenminste als men er wel van uitgaat dat de hele steekproef illegaal zal zijn (Tabel 4). Dit is meer dan met de hypergeometrische verdeling vereist is (23) maar net eentje minder dan met de binomiale verdeling vereist is (29). Het komt echter vrijwel nooit voor dat als 1 bolletje cocaïne bevat dat een ander bolletje geen cocaı"ıne bevat. Neem je deze voorkennis mee, dan kan men een bèta verdeling als prior nemen met $b=1$ en a een hoge waarde bv. 3 of misschien wel 10. Dan zal de gewenste steekproefgrootte flink dalen. In de praktijk analyseert men vaak slechts 1 bolletje. Met behulp van de Bayesiaanse theorie en de keuze van een voldoende grote a valt dit te

verdedigen.

De keuze van de grootte van a en b is een subjectieve beslissing. Hier wordt onderzoek naar gedaan, maar er zijn op het moment (nog) geen richtlijnen. In Nederland wordt in de praktijk nog geen gebruik gemaakt van de Bayesiaanse methode en in bijvoorbeeld Groot-Brittannië gebeurt dit alleen op ad hoc basis. Dat wil niet zeggen dat de Bayesiaanse aanpak daarom maar helemaal genegeerd moet worden. In veel praktijk situaties is het niet onrealistisch om aan te nemen dat naar alle waarschijnlijkheid de hele partij illegaal is, alleen al op basis van het feit dat dit meestal het geval is. Binnen het Bayesiaanse perspectief kan het gebruik van deze informatie betekenen dat efficiënter steekproeven getrokken kunnen worden (oftewel minder monsters nodig zijn).

Figuur 6. *Een hennepplantage*

Een hennepplantage (zie Figuur 6) is door iedere (ervaren) opsporingsambtenaar als zodanig te herkennen. Als de opsporingsambtenaar een plantage goed heeft bekeken, en er van alle kanten er omheen is gelopen en dan van mening is dat alle planten er hetzelfde uitzien is het zeer waarschijnlijk dat als één plant inderdaad hennep blijkt te zijn dat alle planten dit dan zijn. Dit is een geval waar voorinformatie (namelijk dat alle planten er hetzelfde en als hennep er uitzien) sterk zou kunnen meewegen. Dit vertaalt zich in het gebruik van een bèta-verdeling als priorverdeling met $b = 1$ en a een hoge waarde. Als de waarde van a hoog genoeg wordt genomen dan kan de analyse van 1 plant al voldoende zijn om met 95% betrouwbaarheid te garanderen dat tenminste de helft hennep is.

6. Slot

Sommige statistici vinden dat er altijd met de Bayesiaanse methode gewerkt moet worden omdat dit een logische benadering is. Daarnaast kan eventuele extra informatie meegenomen, maar het hoeft niet. In het laatste geval gebruik je een neutrale prior. In het eerste geval kan dit leiden tot een flinke reductie in het aantal te analyseren monsters.

Tegenstanders komen vooral uit de praktijk en redeneren dat de Bayesiaanse aanpak statistisch misschien wel logischer is, maar ingewikkelder in de prakti-

sche toepassing. Bovendien bevat ze een subjectief element waar men moeilijk mee om kan gaan. De meeste mensen zijn het er echter wel over eens dat statistisch onderbouwde methodes beter zijn en meestal tot minder te analyseren monsters leiden.

Literatuur

1. C.C.G. Aitken, J. Bring, T. Leonard, O. Papasouliotis, *Estimation of quantities of drugs handled and the burden of proof.* Statist. Soc., 1997, **160**(2), 333–350.
2. C.C.G. Aitken, *Interpretation of Evidence and Sample Size Determination.* Statistical Science in the Courtroom, (ed.) Joseph L. Gastwirth, Springer Verlag, 2000, 1–24.
3. C.C.G. Aitken, *Sampling - How big a sample ?* Journal of Forensic Sciences, JFSCA, 1999, **44**(4), 750–760.
4. C.C.G. Aitkin, *Estimation of the Quantity of a Drug in a Consignment from Measurements on a Sample.* Journal Forensic Sci., Sept. 2002, **47**(5), 968–975
5. M. Azoury, D. Grader-Sageev, S. Avraham, *Evaluation of Sampling Procedure for Heroin Street Doses.* Journal of Forensic Sciences, JFSCA, 1998, **43**(6), 1203–1207.
6. A.B. Clark, C.B. Clark, *Sampling of Multi-unit Drug Exhibitsi.* Journal of Forensic Sciences, JFSCA, 1990, **35**(3), 713–719.
7. W.G. Cochran, *Sampling Techniquesi*, Wiley and Sons, New York, 1977.
8. M. Colon, G. Rodriguez, R.O. Diaz, *Representative Sampling of 'street' Drug Exhibits.* Journal of Forensic Sciences, JFSCA, 1993, **38**(3), 641–648.
9. S.A. Coulson, A. Coxon, J.S. Buckleton, *How many Samples from a drug Seizure Need to be analysed.* Journal of Forensic Sciences, JFSCA, 2001, **46**(6), 1456–1461.
10. ENFSI drugs working group, *Guidelines on representative drug sampling.* Opmeer, Den Haag, 2003, te vinden op www.enfsi.org
11. R.S. Frank, S.W. Hinkley, C.G. Hoffman, *Representative Sampling of Drug Seizures in Multiple Containers.* Journal of Forensic Sciences, JFSCA, 1991, **36**(2), 350–357.
12. A.J. Izenman, *Statistical and Legal Aspects of the Forensic Study of Illicit Drugs.* Statistical Science, 2001, **16**(1), 35–57.
13. S.K. Thompson, *Sampling.* John Wiley and Sons, New York, 1992.
14. D. Tzidony, M. Ravreby, *Statistical Approach to Drug Sampling: A case Study.* Journal of Forensic Sciences, JFSCA, 1992, **37**(6), 1541–1549.

CWI Vakantiecursus 2005

pp. 35 – 47

Mathematical aspects of the World Wide Web and search engines

N. Litvak
University of Twente
e-mail: n.litvak@math.utwente.nl

In business and every-day life, it is hard to overestimate the role of the World Wide Web – a giant virtual network binding together several billions of hyperlinked pages. In this note, we attempt to highlight some structural properties of the Web and show how they can be modeled mathematically. Further, we discuss the general scheme of a Web search engine and explain ideas behind some search engine techniques such as Google PageRank.

1. Introduction

In our informational society, the World Wide Web has quickly become one of the most if not *the* most important media. Given a gigantic size of the Web and its uncontrollable random expansion, the structure of the Web may seem completely chaotic, and the high performance of modern search engines looks almost like a magic. This note has a two-fold purpose. First, we attempt to highlight some well-known structural properties of the World Wide Web and show how they can be modeled mathematically. Second, we shall explain the principal scheme of a Web search engine and discuss important ranking algorithms used for listing the search results in an appropriate order.

The common viewpoint in the literature is to present the Web as a *graph*, a mathematical object composed of a number of vertices and edges between them. Thus, the pages are viewed as vertices and the links – as *directed* edges. This simplified representation suffices to answer many important questions such as: What is a typical number of in- and out-going links? Does the Web consist of one giant knot of pages and links (the graph is connected) or is it more like several separate 'islands'? What is the average path length between two connected pages? These extremely important questions have been partly answered in the famous paper by Broder et al. [7] that we shall discuss in the next section.

Several typical properties of the Web can also be observed in other complex stochastic networks such as networks of collaborations, networks of airline

routes, biological networks, scientific citations, children friendships, and many others [12]. This suggests that a network structure builds up in a certain way, which is similar for various large systems. Understanding of how this structure appears enables one to predict the developments in a highly dynamic environment such as the World Wide Web. Currently, *growing network* models with *preferential attachment* are widely accepted as a possible mathematical explanation for many empirically discovered properties of complex networks. The main idea in these models is that the observed structure is a result of a network growth driven by 'rich get richer' mechanism. That is, a newly created node is more likely to link to the nodes that are already well-connected. We shall address the growing network models in more detail in Section 3.

For a user, the practical availability of enormous amounts of information offered by the Web depends greatly on efficiency of search engines. In Section 4 we shall briefly explain how a search engine works and focus on the hyperlink-based techniques used for listing the search results in a user-friendly order. In particular, we shall explain the relatively simple mathematical model behind the Google PageRank.

2. Graph structure in the Web

As mentioned in the introduction, we view the Web as a set of vertices (pages, nodes), and directed edges (links) between them. We say that there is an edge from page i to page j if i has a hypelink to j, i.e. a user can go from i to j by just one click.

Given a spontaneous chaotic development of the Web, one can hardly expect any regularity in its structure. From the first sight, it looks like the Web graph may have any shape you please. However, the fundamental research by Broder et al. [7] revealed several robust structural properties of the Web. The experiments in [7] were carried out on two large crawls each containing about 200 million pages and 1.5 billion links. To indicate the importance of this paper, we just mention that it was cited in about 500 articles! Strikingly, most typical traits discovered in the World Wide Web have also been observed in other complex networks such as social networks, networks of scientific citations, biological networks, etc. Below we discuss the properties of the Web graph presented in [7] and other articles that continued empirical studies of the Web graph.

2.1. Bow-tie structure

In Figure 1, we present the structure of the Web as discovered in [7]. We see that the majority of the pages are united in one connected component, which has a shape of a 'bow tie'. For any two pages i and j in the 'bow tie', there is a hyperlink path either from i to j or from j to i. In the middle, there is a *Strongly Connected Component (SCC)* containing more than one quarter of all pages. The term for the SCC comes from the graph theory and stands for a set of nodes where each node can be reached from any other node by traversing directed edges. For the Web, it means that in the SCC, each page

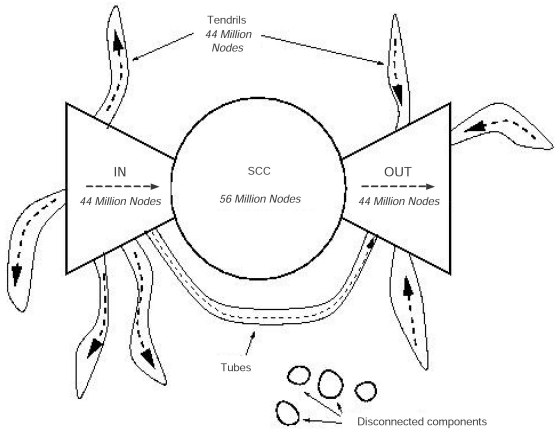


Figure 1. *Graph structure in the Web (from [7])*

can be reached from any other page by clicking on hyperlinks. Next, there are large IN and OUT components. The pages in IN (OUT) have a path to (from) the SCC, but not back. There are also smaller groups such as *Tendrils* branching from IN or leading directly to OUT, and *Tubes* offering a path from IN to OUT. The little 'islands' represent the *Disconnected components*, which amount to less than 10% of the Web.

From the above, one may have an impression that the Web is greatly connected, and for two random pages i and j, a hyperlink path from i to j is likely to exist. However, a closer look suggests that it is not so. Roughly speaking, a hyperlink path exists only if page i belongs to IN+SCC, and page j is in SCC+OUT. As both IN and OUT contain slightly less than 1/4 of all pages, the probability that the path exists is (only!) about 24%.

2.2. Small-world effect

Assuming that there is a path from one page to another, what is the average path length? Despite the enormous size of the Web, the average path turns out to be relatively short. Experiments in [7] report about 16 clicks only! Moreover, if links can be traversed in both ways, the average path length reduces to the value about 7.

This phenomenon – a short average distance between the nodes in large networks – is known for a long time as a *small-world effect*. One of the most famous experiments in this respect was carried out by Stanley Milgram in the 60s in a context of social networks. The participants were asked to pass a letter to their close acquaintances so that it would finally reach the assigned target individual. About 1/4 of the letters reached the target passing, on average, through the hands of only about 6 people! [12]

While looking quite astonishing, the small-world effect actually has a simple mathematical explanation. Here is a greatly simplified argument, which is far from being rigorous but it helps to grasp the main idea. For a given node, assume that the number of nodes within a distance r is roughly a^r, where $a > 0$ is a constant. This assumption is true for many real-life networks. Now, let l be the maximal distance from one node to another. Then it follows from geometric series that the total number of nodes N is

$$N = 1 + a + a^2 + \cdots + a^l = (a^{l+1} - 1)/(a - 1) \approx a^{l+1}/(a - 1).$$

Hence, for large N, the value l turns out to be of the order $\log(N)$. For example, in the network of one billion nodes, this number is of the order 10, which is exactly the small-world effect observed in experiments. For more detail on the small-world effect, we refer to the brilliant survey by Mark Newman [12] and references therein.

2.3. Power laws

Let us consider the number of in- and out-going links, called, respectively, *in-* and *out-degree*, of a Web page. The question is, for instance, what is the fraction p_k of pages whose in-degree is exactly k? The experiments carried in different time on different crawl sizes agree that p_k is approximately proportional to $k^{-2.1}$. On the left plot in Figure 2, we present the experimental results on the in-degree distribution [7]. Plotted in the log-log scale, for each in-degree k, the

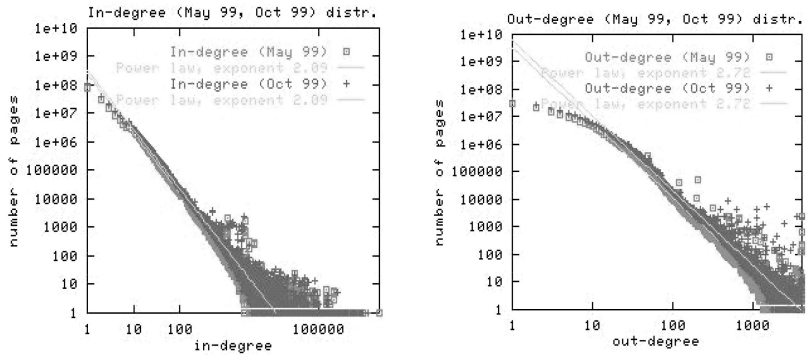


***Figure 2**. (from [7]) In- and out-degree (May99, Oct99) in the log-log scale*

values of p_k concentrate around a straight line: $\log(p_k) = -2.09\log(k) + const$, which signals the power law: $p_k \approx const \cdot k^{-2.09}$.

Put in words, the power law means that the majority of the pages have a relatively small in-degree but there is noticeable group of pages whose in-degree is high. To see that, let us evaluate the number of pages with in-degree 1000. According to the power law distribution, the fraction of such pages is of the order $1000^{-2.1} \approx 10^{-6}$. Hence, for a net of one billion pages, the number of pages whose in-degree is one thousand is of the order $10^9 * 10^{-6} = 1000$. A group

of this magnitude cannot be neglected in any reasonable network analysis. To demonstrate a difference with, for instance, exponential law, assume that p_k is of the order 10^{-k}. Then for $k = 1000$, the proportion 10^{-1000} results in a negligible group of pages, even in a truly giant network. As we know that *there are* well-connected nodes (think for instance of the homepage of Google or CNN), the power law seems to be a more realistic model for the in-degree distribution. As we see from on the right plot in Figure 2, the out-degree also obeys a power law but it has an exponent about -2.7. Thus, for large enough k, the probability of in-degree k is greater than the probability of out-degree k.

2.4. Self-similarity

Clearly, the Web can be subdivided into large logically united components, for instance, by domain or by topic. Surprisingly, it turns out that such large components have similar structure as the Web as a whole, which is a result of many essentially independent stochastic processes evolving in the Web at various scales. This phenomenon – called *self-similarity* – was observed in a number of experiments on different crawl sizes, and analyzed in detail in [8]. It was shown in [8] that a Web-like structure is present in so-called *thematically unified clusters* (TUC), i.e. sets of pages that share some common feature, for instance, content, domain, or geographical location. The authors also note that in a purely random set of pages the structure will be lost. Indeed, assume that a sample of one million pages out of possibly one billion is chosen at random. Then the probability that both ends of some edge belong to the chosen sample is $(10^6/10^9) \cdot (10^6/10^9) = 10^{-6}$. Since the average number of links per page is just about 8, we get on average $8 \cdot 10^9 \cdot 10^{-6} = 8000$ links in the random collection of one million nodes. With such a small amount of links one can hardly expect any interesting graph-theoretic structure.

3. Mathematical models of the Web: preferential attachment

Currently, growing network models with preferential attachment are widely accepted as a possible mathematical explanation of many empirically discovered properties of the Web. In these models, a newly created page is more likely to link to the pages that are already well-connected. The most famous model of this sort was suggested in 1999 by Barabasi and Albert [3], and many modifications appeared since then. We shall closely follow Newman [12] in explaining how the model works and why it leads to the power law in-degree distribution.

In [3], a networks starts with one node. When a new node appears, it has $m \geq 1$ *undirected*, or, equivalently, *bi-directed* links to distribute among the existing nodes. In doing so, a node follows the 'rich get richer' strategy. That is, the probability that some node v gets a new link is proportional to the current in-degree of v. Thus, if the fraction of nodes with in-degree k is p_k, then a probability that a new link goes to this group is

$$\frac{kp_k}{1 \cdot p_1 + 2 \cdot p_2 + 3 \cdot p_3 + \cdots + kp_k + \cdots} = \frac{kp_k}{2m}, \quad k \geq 1.$$

Here the denominator in the left-hand side is first defined such that the sum of the probabilities equals 1, and then we notice that it equals the average number of links per page, which is $2m$ since each node brings m bi-directed edges.

Now, with each new node, the group of in-degree k receives on average $m \cdot kp_k/(2m) = (1/2)kp_k$ links, which is independent of m. Thus, the number of vertices with in-degree k *decreases* by this amount since these nodes join the group of in-degree $k+1$. On the other hand, on average $(1/2)(k-1)p_{k-1}$ nodes of in-degree $k-1$ will also receive a new link, so the number of vertices with in-degree k will *increase* by this number. If the total number of nodes is very large then the proportion of nodes with in-degree k almost does not change (in fact, this proportion converges to a constant when the number of nodes goes to infinity). So, when the nth new node is added and n is large enough, then the number of nodes with in-degree k changes approximately by $np_k - (n-1)p_k = p_k$. Equating the increments in the number of nodes with in-degree k, we can write so-called master equations, which hold when the number of vertices in the graph goes to infinity:

$$p_k = \begin{cases} \frac{1}{2}(k-1)p_{k-1} - \frac{1}{2}kp_k, & \text{for } k > m; \\ 1 - \frac{1}{2}mp_m, & \text{for } k = m. \end{cases}$$

Here the last equation reflects that there is always one new node with exactly m links, and at the same time the group of such nodes becomes smaller by $(1/2)mp_m$, as it happens for any other value of k. Writing the equation for p_m we get $p_m = 2/(m+2)$, and for other values of k we obtain $p_k = p_{k-1}(k-1)/(k+2)$. Solving recursively, we arrive at

$$p_k = \frac{(k-1)(k-2)\cdots m}{(k+2)(k+1)\cdots(m+3)}\, p_m = \frac{2m}{(k+2)(k+1)k}.$$

Thus, for large k, we have $p_k \sim k^{-3}$, which is a power law with exponent 3. We note that the present model deviates from the experimental results that suggest the exponent 2.1. However, this was fixed in later generalizations by other authors.

The significance of the Barabasi and Albert model is that besides modeling the growing random graph that exhibits the power law in-degree distribution, it also aims to explain *why* such distribution appears. We note that this model is in the spirit of the earlier model developed in 1965 by Derek de Solla Price in his work on scientific citations. Even before that in 1950s, Herbert Simon showed that the power law distributions arise from the 'rich get richer' mechanism, also called *Matthew effect* in sociology [12]. As an example, we note that something like the power law can be observed in a group of school children: a few boys and girls are very popular while others have only one or two friends. Is it not natural to explain this by the tendency of the children to make friends with popular, or 'well-connected' classmates?

The models with preferential attachment have received a great attention in the network literature. There is a lot of research on generalizing these models in such a way that they better reflect complicated features of the Web

such as directed links, hierarchical structure, appearing and disappearing of links and even willingness of users to link to highly ranked pages [16]. The other research direction is a rigorous mathematical analysis of (generalized) preferential attachment models, based on the theory of random graphs. In particular, the power law distributions were rigorously derived, and it was also shown that the models with preferential attachment exhibit a small-world effect [4, 5]. The 'rich get richer' mechanism appears to be responsible for many typical developments in complex networks.

4. Search engines

The study and modeling of the Web structure is one of the main challenges in the Web search engines [9], which are of extreme importance in navigating the Web. In this section, we shall explain in broad lines how a search engine works, and discuss two prominent hyperlink-based ranking techniques (in particular, the Google PageRank) for selecting important and interesting Web pages.

4.1. General scheme of a search engine

The general scheme of a crawler-based search engine is presented in Figure 3, which resembles a figure found in Google images. Below we highlight essential

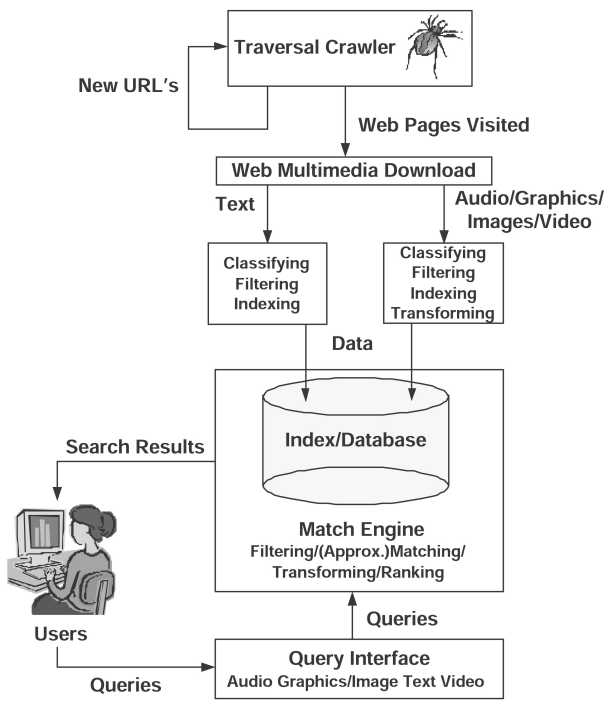


Figure 3. *General scheme of a search engine*

components of the Web search, using a nice article by Gallianno Cosme published in Search Engine Guide in May 2005 (http://www.searchengineguide.com). According to this article, a crawler-based search engine consists of three main parts: the spider (also known as crawler or robot), the index, and the software.

The *spider* is a program that visits pages and follows hyperlinks to move from one page to another. The main goal is to obtain the most recent copy of each page but, as we see later, it may be also important to record the hyperlink structure. The spiders start their journey from the pages that are already in the search engine database. The most active spiders on the Web are Googlebot (Google), Slurp (Yahoo!) and MSNBot (MSN Search).

The copies of crawled pages are stored in *index* which is essentially a giant catalogue or database where the pages are classified, filtered, indexed, transformed (if needed) and grouped according to some rules, for instance, according to the topic. The size of the index is truly enormous. Just to give an idea, the last figure revealed by Google is 8 billion pages! In practice, both crawling and indexing require significant time and capacity.

When a user types a query in a query interface (like a famous white page of Google), the search engine finds the relevant pages in its database using a software based on sophisticated algorithms and state-of-art information retrieval techniques. In Figure 3 this software is denoted as Match Engine. It is beyond the scope of this note to discuss the methods for retrieving relevant pages from the database. We mention only that, in one or another way, the query is compared with the key-words related to a Web page. Naturally, these are mainly the keywords and the text included in the page itself. However, curiously enough, the text of the hyperlinks connecting to a page is also taken into account, at least by Google. That is why the query *'miserable failure'* returns a biography of G.W. Bush although these words have never been mentioned in this document! The relevant pages are then ranked in some order that the search engine finds most appropriate, and the results are presented to the user. The ranking algorithm is a well-kept secret, and it depends on many factors, for instance, on geographical location of the user and maybe even on his/her last searches.

We would like to emphasis that the main structural feature of a search engine is that the hard and slow job such as crawling and indexing happens without involving the user, who 'only' needs to consult a database and receive the results from the *index* rather than from the Web itself. Search engines are equipped with modern software and powerful computers that scan the index extremely fast so that the search results appear on a screen almost instantaneously. As a minor drawback of this scheme, the user may access only the pages listed in the index, which is not complete and not entirely up-to-date. For instance, new pages that have been crawled but have not yet been added to the index, will *not* be available to those searching with the search engine. This is the reason why the Web site owners have to make sure that their sites are timely indexed and highly ranked. In a business world, there is a whole branch of marketing called Search Engine Optimization that developes techniques for increasing the Web site's ranking performance.

At the end, the main goal of any search engine is to satisfy the user, so we may trust that the vastly expanding index is frequently updated, and matching/ranking algorithms are steadily improving to provide us with the desirable results. Note, by the way, that the index and the matching/ranking mechanisms are different for different search engines, and therefore it is not uncommon to use several search engines for the same query.

4.2. Node ranking based on the hyperlink structure

We started this section claiming that the knowledge of the hyperlink structure is important for search engines. Obviously, such knowledge helps to optimize the spider's crawl, and it can be used for matching as well. However, some search engines, and in particular Google, also use hyperlinks for ranking the Web pages according to their importance.

Suppose, a search engine has received a query and found relevant pages in its index. Then another problem arises. Namely, there can be thousands of pages matching the query. How to define which page is the *most important* and has to be placed on top of the list? At the beginning, this problem was solved solely by finding pages with best-matching text. However, with fast expansion of the Web, such methods soon became inefficient. Two innovative path breaking approaches were presented in 1998: one belongs to a well-known academician Jon Kleinberg [11], and another came from two PhD students from Stanford, Sergey Brin and Larry Page [6] known as 'founding fathers' of Google. Although the two approaches are different, the main idea is similar: the page should be ranked high and listed high if many other good pages have a hyperlink to this page, and thus the page is recognized by the Web community as an important source of information. In contrast to the previously used methods, these novel ranking techniques are based not on the content of the pages but on the most fundamental feature of the Web – the hyperlink structure. Naturally, both methods require the knowledge of who is linking to whom. This information is recorded in the *adjacency matrix* A defined as follows:

$$A_{ij} = \begin{cases} 1, & \text{there is a link from } i \text{ to } j, \\ 0, & \text{otherwise.} \end{cases}$$

Such matrix can be obtained by the spider while crawling the Web.

In his work [11], Kleinberg considers two sorts of pages: hubs an authorities. A hub serves as a reference giving many links *to* important authorities. The authorities, on the other hand, contain important information and thus receive many links *from* the hubs. Formally, let a_i and h_i be, respectively, the authority and the hub score of page $i = 1, \ldots, n$. Then

$$a_i = \sum_j A_{ji} h_j, \qquad h_i = \sum_j A_{ij} a_j.$$

The HITS algorithm suggested by Kleinberg, is as follows:

– Retrieve a set of relevant pages from the database.

– Extend this set by adding all pages that has links to and from the selected pages.

– For the extended set, compute the hub and authority scores.

– Since the user is mostly interested in authoritative sources, list the search results according to the authority score.

If we want to include this ranking algorithm into the scheme in Figure 3, then we should add an Authority Score Computation block between the Database and the User (this option is depicted in Figure 4 with dashed arrows). Fortunately,

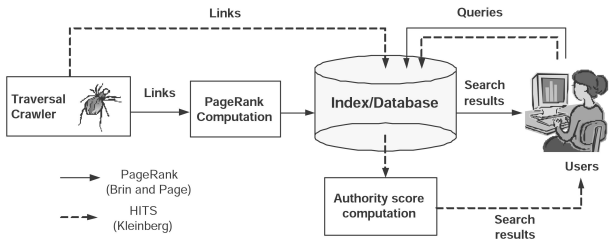


Figure 4. *Ranking computation: HITS vs PageRank*

the authority scores can be computed very fast because the number of pages involved in computations is not very large, which results in a well solvable linear algebra problem.

The approach of Brin and Page is different. In their work [6], they introduce a universal popularity measure – the PageRank. The PageRank $PR(i)$ of page i depends on how many other pages link to i and how important these pages are. The original formula is as follows:

$$PR(i) = c \sum_j \frac{A_{ji}}{d_j} PR(j) + (1 - c), \quad i = 1, \ldots, n, \tag{1}$$

where d_j is the number of out-going links from page j, n is the number of pages in the Web, and c is a constant between zero and one (Google originally used $c = 0.85$). The algorithm works as follows.

– Right after crawling the Web, retrieve and store the matrix A.

– Compute the PageRank score for each page and store the PageRank vector.

– For each query, list the matching pages according to their PageRank.

In order to reflect this procedure in Figure 3, we have to add a chain that is depicted in Figure 4 by solid arrows: there is a large computation block right after crawling but there is no computation involved after consulting the database, which in general helps to deliver the search results faster.

Let us now take a closer look at the famous PageRank formula (1). We see that two factors are taken into account: the quality and the quantity of

incoming links. The idea is that if we view a link as a vote, then pages with many links deserve attention. Plus, if a page has only a few links but these links come from important sites, then this page is also worth browsing.

PageRank has an insightful probabilistic interpretation. The $PR(i)$'s in (1) can be normalized so that they sum up to one. We denote the normalized PageRank values by π_i's:

$$\pi_i = PR(i)/[PR(1) + PR(2) + \cdots + PR(n)], \quad i = 1, \ldots, n. \tag{2}$$

The vector $\pi = (\pi_1, \pi_2, \ldots, \pi_n)$ is a probability distribution that can be interpreted via the so-called *easily bored surfer* model. Consider a random surfer that starts navigating the Web from a random page. At each page, with probability c, the surfer follows a randomly chosen hyperlink, and with probability $1 - c$ he/she gets 'bored' and jumps to a random page. To keep the model equivalent to (1), we have to make a natural assumption that the surfer always jumps to a random page when reaching a page which does not have out-going links. Such pages, also called *dangling nodes*, should not influence the ranking. The described surfing process can be modeled as a *Markov chain*, a random process whose state at time $t+1$ depends *only* on the state at time t and does not depend on the past states. Our Markov chain is *irreducible* [10] by definition, since there is a possibility to make a random jump, and thus, any two pages (states) can be reached from each other. Hence, it follows from the theory of Markov chains [10] that π_i is nothing else but the long-run probability, or long-run fraction of time, that a random surfer spends on page i. Moreover, this probability is uniquely defined for all $i = 1, \ldots, n$. Naturally, higher the probability, more popular is the page. It follows from the description of the surfing process that π_i's satisfy

$$\pi_i = c \sum_j \frac{A_{ji}}{d_j} \pi_j + c \frac{1}{n} \sum_{j \in D} \pi_j + \frac{1-c}{n}, \; i = 1, \ldots, n; \quad \sum_i \pi_i = 1,$$

where D is a set of dangling nodes. The last equation is equivalent to (1) and (2).

The 'dumping factor' $c < 1$ is needed in particular because the random jump option guarantees that the unique distribution π exists. With $c = 1$, it is quite likely that some pages cannot be reached from each other (see also Section 2), and then according to the Markov chain theory, the PageRank vector is not well defined.

The PageRank citation ranking technique is very efficient and *is* actually used by Google, although maybe not in its original form. The disadvantage of this method is however obvious. Equation (1) must hold for each $i = 1, \ldots, n$, where n is the number of pages in the index, so we have a huge linear system with n equations and n variables, where n is of the order of billions. Solving such linear system directly is practically unfeasible. Google originally proposed to use a *power iteration method* that works as follows. First, put $\pi^{(0)} = (1/n, \ldots, 1/n)$. Then for each $k \geq 1$ compute

$$\pi^{(k)} = c \sum_j \frac{A_{ji}}{d_j} \pi_j^{(k-1)} + c \frac{1}{n} \sum_{j \in D} \pi_j^{(k-1)} + \frac{1-c}{n}.$$

The algorithm stops when $\pi^{(k)}$ and $\pi^{(k-1)}$ are close enough. In their first work, Brin and Page reported convergence in 50–100 iterations.

It can be shown using Perron-Frobenius theory in linear algebra [17] that the difference between the approximation $\pi^{(k)}$ and the real PageRank value π is of the order c^k (see e.g. [13]). Thus, the power iterations converge exponentially fast, and smaller values of c ensure a faster convergence, which is a valid reason to keep c not too close to 1. On the other hand, in (1), the term that depends on hyperlinks decreases with c, so small c results in almost uniform PageRank. Hence, a reasonable compromise has to be found, and Google's original choice was $c = 0.85$. We refer to the interesting and extremely well written survey [13] for more detail on this respect.

The present value of c and the actual algorithm used by Google nowadays is not known but nevertheless, the PageRank distribution still plays an important role in defining the order of search results. Moreover, according to the publicly available information, power iterations are still used for the PageRank computation. There are a lot of intelligent techniques developed for making the power method more efficient, such as parallel computing, block iteration methods, rearranging, two-stage methods, and many others.

There are also alternative algorithms that allow to compute the PageRank on-line while crawling the Web. One of the methods that works surprisingly well is a *Monte Carlo* algorithm [2]. In a nutshell, this algorithm runs a random surfing process from each page. If a random jump has to be made, the simulation stops and then starts from the next page. At the end, the PageRank of page i is computed as the number of visits to this page divided by the total number of steps performed. Amazingly, it is sufficient to run such simulation only *once* from each page to obtain a reasonable estimate of the PageRank.

Another intelligent on-line method is proposed in [1]. Initially, each page receives an equal amount of cash, and whenever a page is crawled, it distributes all its cash among its outgoing links. After several crawls, the importance of a given page is evaluated as a fraction of cash spent by this page compared to the total amount spent by all pages together. The algorithm converges very fast, does not require any storage of the hyperlink matrix, and quickly adopts to the changes in the Web. Besides, analytical studies of this algorithm gives rise to many interesting mathematical problems.

Although the PageRank is not directly related to the number of incoming links there is an intimate connection between these two measures of page popularity. For instance, it turns out that the fraction of pages whose PageRank is about k/n is roughly proportional to $k^{-2.1}$ [16], exactly as the fraction of pages with k incoming links! Since the models with preferential attachment explain the power law phenomenon for the in-degree, it is interesting to study the PageRank and its evolution in these models. This leads to a whole class of challenging research problems in the novel exciting area of complex stochastic networks.

Literatuur

1. S. Abiteboul, M. Preda, G. Cobena (2003). Adaptive on-line page im-

portance computation. *The Twelfth International World Wide Web Conference WWW2003.*

2. K. Avrachenkov, N. Litvak, D. Nemirovsky, N. Osipova (2005). Monte Carlo methods in PageRank computation: When one iteration is sufficient, Technical Report 1754, University of Twente.
3. A.-L. Barabási, R. Albert (1999). Emergence of scaling in random networks. *Science* **286**, 509–512.
4. B. Bollobás, O. Riordan, J. Spencer, G.E. Tusnády (2001). The degree sequence of a scale-free random graph process. *Random Struct. Algorithms* **18**(3), 279–290.
5. B. Bollobás, O. Riordan (2004). The diameter of a scale-free random graph. *Combinatorica* **4**, 5–34.
6. S. Brin, L. Page, R. Motwami, T. Winograd (1998). The PageRank citation ranking: bringing order to the web. Stanford University Technical Report.
7. A.Z. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, J.L. Wiener (2000). Graph structure in the Web. *Computer Networks* **33**, 309–320.
8. S. Dill, R. Kumar, K. McCurley, S. Rajagopalan, D. Sivakumar, A. Tomkins (2002). Self-similarity in the Web. *ACM Trans. Internet Technology* **2**(3), 205–223.
9. M.R. Henzinger (2003). Algorithmic challenges in Web search engines. *Internet Mathematics* **1**(1), 115–126.
10. S. Karlin, H.M. Taylor (1998). *An Introduction to Stochastic Modeling.* Academic Press, San Diego.
11. J. Kleinberg (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM* **46**, 604–632.
12. M.E.J. Newman (2003). The structure and function of complex networks. *SIAM Rev.* **45**(2), 167–256.
13. A.N. Langville, C.D. Meyer (2005). Deeper inside PageRank. *Internet Mathematics* **1**(3), 335–380.
14. A.N. Langville, C.D. Meyer (2004). A reordering for the PageRank problem. NCSU CRSC Technical Report CRSC-TR04-16.
15. C.P.-C. Lee, G.H. Golub, S.A. Zenios (2004). A fast two-stage algorithm for computing PageRank. Stanford University Technical Report.
16. G. Pandurangan, P. Raghavan, E. Upfal (2002). Using PageRank to characterize Web structure. *8th Annual International Computing and Combinatorics Conference (COCOON).*
17. W.J. Stewart (1994). *An Introduction to the Numerical Solution of Markov Chains.* Princeton University Press.

Vakantiecursus 2005

pp. 49 – 69

Coderingstheorie

R.H. Jeurissen
e-mail: rhjeuris@wanadoo.nl

Data (Latijn: gegevens) worden tegenwoordig bijna altijd digitaal bewaard en verzonden. Tekst, spraak, foto's, muziek en waarnemingen zijn dan opgeslagen als rijen van nullen en enen (bits). Er is ook al digitale radio en televisie. Bij opslag of transport van die data kunnen fouten optreden (fout of krasje op een CD, ruis, atmosferische storing, zwakke zender van een ruimtesonde): een 0 wordt een 1 of andersom. Door toevoegen van extra bits (het coderen) kan worden bereikt dat na ontvangst fouten (mits niet teveel) kunnen worden ontdekt en zelfs verbeterd. In tekst en spraak is 'van nature' al zulke extra ('redundant') informatie aanwezig (een boek met veel drukfouten is wel te reconstrueren, een slordige spreker nog wel te volgen). Aan de hand van voorbeelden wordt een indruk gegeven van hoe coderen in zijn werk gaat en hoe wiskunde daarbij een beslissende rol speelt.

1. Wat is coderen?

De zender kan de ontvanger bereiken via een **binair kanaal**: hij kan alleen 0'en en 1'en versturen. Stel hij wil vier verschillende boodschappen kunnen verzenden. Voor die boodschappen slechts twee bits nodig, de ontvanger weet wat hij bedoelt met respectievelijk $00, 01, 10$ en 11. Neem aan dat de kans dat een bit fout overkomt 1% is. De kans dat de boodschap fout overkomt is dan 2% (eigenlijk $1,99\%$, maar vooruit). We nemen aan dat fouten onafhankelijk van elkaar optreden; bij ruis is dat niet altijd waar, maar dat is weer een ander probleem, buiten ons bestek.

Een voor de hand liggend idee is: herhaal de boodschap en zend 0101 in plaats van 01. Nu is de kans op een foute boodschap 4%. Ook als bij ontvangst van 0111 wordt aangenomen dat er maar één fout in zit (veel waarschijnlijker dan drie fouten, en twee kan niet) en met een kans van 50% goed gegokt wordt op 01 (en niet 11) komt nog de boodschap in 2% van de gevallen fout door. We schieten er weinig mee op, de verzending kost dubbel zoveel tijd (jammer bij een te pletter vallende ruimtesonde, lastig als de CD-speler de muziek niet kan bijhouden). Het enige wat we bereiken is dat we een fout ontdekken als er één of drie zijn gemaakt. Erg efficiënt is het evenmin: we gebruiken 4 bits om een boodschap van 2 bits te verzenden. De **(efficiency) rate** is $\frac{2}{4}$.

Dat kan echter eenvoudiger door een pariteitscontrole ('**parity check**') te gebruiken. We voegen één extra bit toe, zó, dat ons bericht een even aantal

1'en bevat: $00, 01, 10$ en 11 worden respectievelijk $000, 011, 101$ en 110. Ook nu wordt één fout (kans 3%) door de ontvanger opgemerkt. (En dan gokken maakt daar weer 2% van). De rate is nu $\frac{2}{3}$.

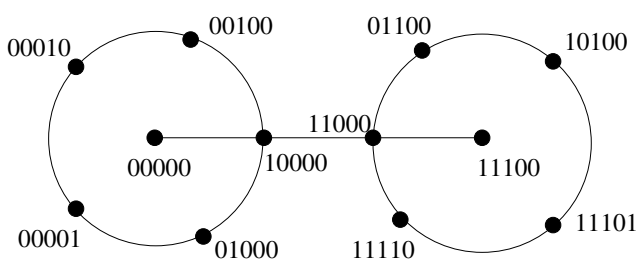
De vier **codewoorden** $000, 011, 101$ en 110 van **lengte** 3 vormen onze **code**. De andere vier 0,1-rijtjes heten **woorden**. Dat we één fout kunnen ontdekken komt doordat een codewoord niet door één fout bit in een ander codewoord kan overgaan, de codewoorden verschillen onderling op twee plaatsen. We zeggen dat ze onderling **afstand** 2 hebben. De code is **1-error-detecting**. Het aantal 1'en in een woord is zijn **gewicht**, het is zijn afstand tot het 'all-zero'-woord 000.

Zijn er maar twee boodschappen, 0 en 1, dan kunnen we als codewoorden 000 en 111 gebruiken. Niet erg efficiënt, maar de codewoorden hebben nu afstand 3. Bij ontvangst van 010 kunnen we tamelijk veilig aannemen dat er 000 is verzonden, de kans op één fout is veel groter dan die op twee fouten. De code is nu 1-fout-verbeterend (**1-error-correcting**). Omdat de afstand 3 is, ligt 010 (op afstand 1 van 000) toch nog op afstand 2 van 111.

Zouden we voor onze vier boodschappen $00, 01, 10, 11$ ook een 1-fout-verbeterende code willen maken, dan moeten we dus vier codewoorden zoeken met onderlinge afstand ≥ 3. Dat lukt alleen als we de codewoorden langer maken. We zoeken uit hoe lang.

2. Bollen

We proberen het met codewoorden van lengte 5. De afstanden tussen codewoorden veranderen niet als we in alle codewoorden het eerste bit veranderen (0 wordt 1 en 1 wordt 0). Idem voor elk ander bit. We mogen dus rustig aannemen dat een van de codewoorden 00000 is. Er zijn 5 woorden op afstand 1 ervan. Samen vormen zij een 'bolletje' met 'straal' 1 waarin $1 + 5 = 6$ woorden zitten. Net zo is er een bolletje met 6 woorden om elk ander codewoord. Deze bolletjes moeten disjunct zijn (anders zouden er twee codewoorden met afstand < 3 zijn). We zoeken vier zulke bolletjes. Die hebben samen $4 \times (1 + 5) = 24$ verschillende woorden. Er zijn $2^5 = 32$ woorden. Dat bewijst nog niet dat het kan, maar in dit geval lukt het, bijvoorbeeld met $00000, 11100, 10011, 01111$. We hebben een code met **minimum afstand** 3 (al zijn er ook paren met afstand 4). Lengte 4 was zeker onmogelijk, omdat $4 \times (1 + 4) > 2^4$.



Figuur 1. *Disjuncte bolletjes met straal 1*

Het zal duidelijk zijn dat bij een 2-fout-verbeterende code gezocht moet worden naar codewoorden met onderlinge afstand ≥ 5, dus naar disjuncte bolletjes met straal 2.

Voor de eigenschappen van een code is het minder relevant welk van die vier codewoorden bij welk van de vier boodschap-woorden hoort. Voor de praktijk is een handige manier natuurlijk wel gewenst, coderen en decoderen is computerwerk (zie het kader in §3) en vereist een systematische methode.

Concentreren we ons op de code zelf, dan zijn van belang:

1. de *lengte* van de codewoorden (graag klein voor efficiëncy)
2. de *afstand* tussen de codewoorden (graag groot voor foutcorrectie)
3. het *aantal* codewoorden (graag veel boodschappen mogelijk)

Legt men de lengte en de gewenste minimum afstand vast, dan zal er een bovengrens aan het aantal codewoorden zijn. Legt men het aantal codewoorden en de gewenste minimum afstand vast, dan zal er een ondergrens aan de lengte zijn. Legt men de lengte en het gewenste aantal codewoorden vast, dan zal er een bovengrens aan de minimum afstand zijn.

De informatie-theorie kent aan een binair kanaal een capaciteit toe. Bij een foutkans van p per bit is die capaciteit $1 + p^{\,2} \log p + (1-p)^{\,2} \log(1-p)$; voor $p = 0,001$ is dat ongeveer $0,986$. De Stelling van Shannon zegt dat met zo'n kanaal voor elke te kiezen rate, mits kleiner dan de capaciteit, codes bestaan (die dan grote lengte moeten hebben) met willekeurig kleine foutkans (in principe kan dus bijna foutloos worden gecommuniceerd).

In de volgende paragraaf zien we dat we niet bij voorbaat aan het aantal boodschappen, dus het aantal codewoorden, zijn gebonden. Vier boodschappen van 2 bits kunnen, door ze twee aan twee te koppelen, behandeld worden als zestien boodschappen van 4 bits. Omgekeerd kunnen de 1024 boodschappen van 10 bits gezien worden als paren van boodschappen van 5 bits.

3. Een perfecte code

We besluiten onze vier boodschappen $00, 01, 10, 11$ twee aan twee te koppelen. We hebben nu dus 16 mogelijke boodschappen:

$$0000, 0001, 0100, 0010, \ldots\ldots, 0111, 1011, 1101, 1110, 1111.$$

We willen een code maken met 16 codewoorden die één fout kan verbeteren. De minimum afstand moet dus 3 zijn. We beginnen als volgt

$$\begin{array}{ccccccc} 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 \end{array}$$

We hebben hier woorden van lengte 7 met onderlinge afstand zelfs 4 (controle is eenvoudig door het cyclische karakter; als het eerste en het vierde woord afstand 4 hebben, geldt dat vanzelf ook voor het tweede en het vijfde woord, enzovoort). We kunnen nog het 'all-one' woord 1111111 toevoegen met behoud van afstand 4 en hebben er dan al acht.

De acht complementaire woorden (verander elke 1 in 0 en elke 0 in 1) hebben dan ook onderling afstand 4. Het is een eenvoudige oefening te laten zien dat nu elke twee van die 16 woorden een afstand ≥ 3 hebben. We hebben een 1-fout-verbeterende code van lengte 7. De rate is $\frac{4}{7}$.

Had het efficiënter gekund, met lengte 6? Nee, we zouden dan 16 bolletjes met inhoud $1 + 6$ moeten hebben, samen dus $16 \times 7 = 116$ woorden, terwijl $2^6 = 64$. Dus 7 is de minimale lengte waarvoor 16 woorden met onderlinge afstand 3 mogelijk zijn. Het rekenwerk daarbij: $16 \times (1 + 7) = 128$, terwijl ook $2^7 = 128$. Er is iets heel bijzonders aan de hand: de bolletjes bevatten samen alle woorden, we zouden dus ook niet nog een zeventiende codewoord erbij kunnen maken. Zo'n code heet **perfect**.

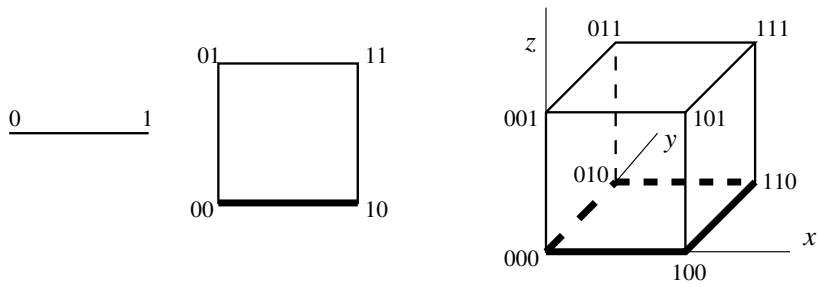
In 1969 werd bij de verzending van Mars-foto's door de Mariners 6 en 7 een codering gebruikt die boodschapblokken van lengte 92 omzette in codewoorden van lengte 127 met onderlinge afstand ≥ 11. Dat maakt het mogelijk 5 fouten in een ontvangen boodschap nog te corrigeren: hij blijft daarbij op afstand ≥ 6 van een andere. Voor zo'n foto moesten $15 \cdot 10^7$ bits worden verzonden, met een snelheid van 16.200 per seconde. Dat duurde 3 uur. Het z.g. decoderen (de fouten in het ontvangen blok van 127 bits verbeteren en het terugvertalen naar het verzonden blok van 92 bits) is natuurlijk computer-werk. Daar zit heel wat algebra achter, want zoeken in een tabel is onmogelijk (de tabel zou $2^{92} \approx 5000000000000000000000000000$ codewoorden bevatten).

4. Kubussen

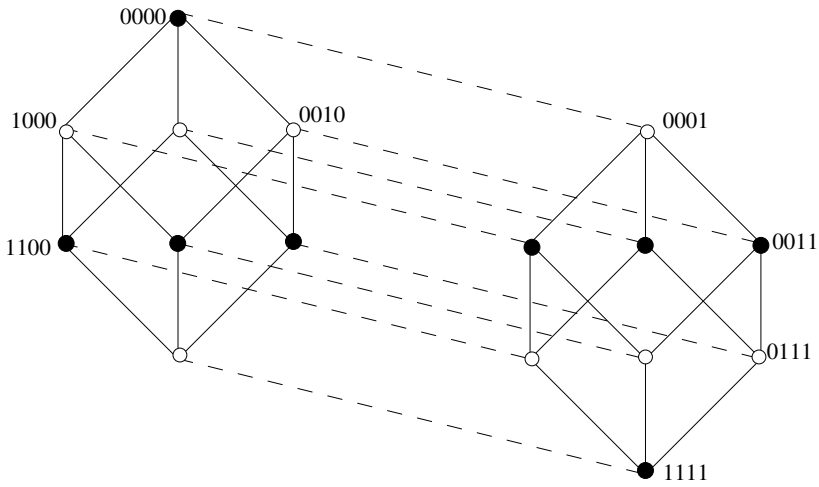
Vanuit een 1-kubus kun je een 2-kubus maken, vanuit een 2-kubus een 3-kubus en vanuit een 3-kubus een 4-kubus. (Maak twee copiën van de 3-kubus, voeg bij de ene een 0 toe aan de coördinaten, bij de andere een 1, en verbind de twee copiën van eenzelfde hoekpunt). Zie figuren 2 en 3.

In figuur 3 en figuur 4 pogingen om een 4- en een 5-kubus voor te stellen. Ga in figuur 3 na dat er geen 4 codewoorden van lengte 4 zijn met minimum afstand 3. De woorden bij de zwarte punten zijn die met even gewicht.

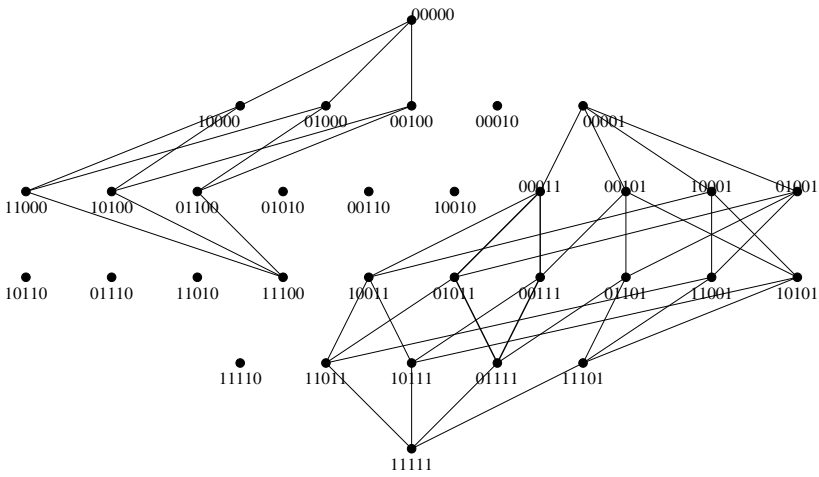
Zo doorgaande bedenk je een 7-kubus. De hoekpunten zijn de 0,1-rijtjes van lengte 7. De woorden van onze perfecte code kun je zien als zekere hoekpunten van de 7-kubus. Tussen twee hoekpunten loopt een ribbe als ze in precies één coördinaat verschillen. Dat twee woorden afstand drie hebben betekent dat je op de kubus van het ene punt in het andere kunt komen door drie stappen langs een ribbe. De 'gewone' (Euclidische) afstand tussen de punten is dan $\sqrt{3}$.



Figuur 2. *1-kubus, 2-kubus en 3-kubus*



Figuur 3. *4-kubus*



Figuur 4. *5-kubus, met 42 van de 80 ribben*

Nog een manier om (code)woorden te interpreteren krijg je door de 0'en en 1'en op te vatten als elementen van $\mathbb{Z}_2$, en de woorden te zien als de vectoren in de lineaire ruimte $\mathbb{Z}_2^7$. Zo kun je er algebra mee bedrijven. Reken dan dus steeds met $1 + 1 = 0$.

5. Algebra

Onze perfecte code kwam nogal uit de lucht vallen. We pakken het nu wat systematischer aan. De berekening met bolletjes leert dat we ten minste lengte 7 nodig hebben als we een code met 16 codewoorden willen maken die één fout kan verbeteren. We proberen het met parity-checks op stukjes van de boodschap (a_1, a_2, a_3, a_4). Zo voegen we een vijfde coördinaat a_5 toe als 'check' op a_1, a_2, a_3, m.a.w. zó, dat $a_1 + a_2 + a_3 + a_5 \equiv 0 \pmod 2$, ofwel $a_5 \equiv a_1+a_2+a_3 \pmod 2$. Op een dergelijke manier voegen we nog twee coördinaten toe. We coderen de boodschap als volgt:

$$(a_1, a_2, a_3, a_4) \mapsto (a_1, a_2, a_3, a_4, a_1 + a_2 + a_3, a_1 + a_2 + a_4, a_1 + a_3 + a_4).$$

Coderen is hier een lineaire afbeelding van $\mathbb{Z}_2^4$ naar $\mathbb{Z}_2^7$, dus te beschrijven met een matrix:

$$\begin{pmatrix} a_1 & a_2 & a_3 & a_4 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 \end{pmatrix}.$$

De codewoorden zijn de lineaire combinaties van de rijen van die matrix, de **generator-matrix** van de code.

$(a_1, a_2, a_3, a_4, a_5, a_6, a_7)$ is een codewoord als voldaan is aan de vergelijkingen (steeds in $\mathbb{Z}_2$):

$$\begin{array}{lllllllllll} a_5 = a_1 + a_2 + a_3 & \text{ofwel} & a_1+ & a_2+ & a_3+ & & a_5 & & & = 0 \\ a_6 = a_1 + a_2 + a_4 & \text{ofwel} & a_1+ & a_2+ & & a_4+ & & a_6 & & = 0 \\ a_7 = a_1 + a_3 + a_4 & \text{ofwel} & a_1+ & & a_3+ & a_4+ & & & a_7 & = 0 \end{array}$$

We hebben nu 16 codewoorden gemaakt. Het zijn de oplossingen van het beschreven stelsel van drie lineaire vergelijkingen. Maar vormen die oplossingen inderdaad een code die één fout kan verbeteren?

De ontvanger test het ontvangen woord $(b_1, b_2, b_3, b_4, b_5, b_6, b_7)$ door te kijken of het aan de vergelijkingen voldoet:

$$\begin{pmatrix} 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \\ b_6 \\ b_7 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}?$$

Hij kijkt dus of het ontvangen woord in de kern van de matrix zit. Stel er is een fout gemaakt, $b_3 = 0$ terwijl het als 1 was verzonden, of $b_3 = 1$ terwijl het als 0 was verzonden. Hij ontvangt nu *codewoord* + $(0,0,1,0,0,0,0)$ (modulo 2!), Bij het codewoord was de **foutvector** $(0,0,1,0,0,0,0)$ opgeteld.

De matrix daarop loslaten levert op $\begin{pmatrix}0\\0\\0\end{pmatrix}+\begin{pmatrix}1\\0\\1\end{pmatrix}=\begin{pmatrix}1\\0\\1\end{pmatrix}$, de derde kolom van de matrix. Dit resultaat heet het **syndroom**. Hij constateert een fout, en aannemend dat het er maar één is weet hij dat die op de derde plaats moet zitten. Dat weet hij omdat alle kolommen verschillend zijn. Hij **corrigeert** door bij het ontvangen woord $(0,0,1,0,0,0,0)$ op te tellen dat maakt van de foute 0 (of 1) weer een 1 (of 0), en **decodeert** dan door de eerste vier bits als de gezonden boodschap te lezen. De matrix wordt de **parity-check-matrix** genoemd.

Decoderen is hier eenvoudig, omdat de boodschap zelf een deel is van het codewoord (**systematic encoding**), maar dat is niet altijd het geval.

Interessant is dat de code, als kern van een matrix, een lineaire deelruimte van $\mathbb{Z}_2^7$ is. De som van twee codewoorden (modulo 2) is dus weer een codewoord.

Net als de eerder gemaakte perfecte code is deze 1-fout-corrigerend, van lengte 7, en met 16 woorden. Ook nu vullen de bolletjes met straal 1 dus de ruimte precies op. We hebben 16 hoekpunten van de 7-kubus, elk op afstand ≥ 3 van de 15 andere.

Je zou ook als code de kern kunnen gebruiken van

$$\begin{pmatrix}1&0&0&1&0&1&1\\0&1&0&1&1&1&0\\0&0&1&0&1&1&1\end{pmatrix}$$

Dat geeft, op verwisselen van coördinaten na, dezelfde codewoorden. Deze code is die welke in §3 is gemaakt.

Een aardige vorm is ook $\begin{pmatrix}0&0&0&1&1&1&1\\0&1&1&0&0&1&1\\1&0&1&0&1&0&1\end{pmatrix}$. Het syndroom, bijvoorbeeld $\begin{pmatrix}1\\1\\0\end{pmatrix}$ bij een fout op positie 6, geeft direct de plaats van de fout (110_{bin} is de schrijfwijze voor 6 in het tweetallig stelsel)

Deze codes heten Hamming-codes H_3, naar hun bedenker ($\pm$1950).

6. Meer Hamming-codes

Je kunt blijkbaar codes maken door de kern te nemen van een matrix over $\mathbb{Z}_2$. Eén fout in een codewoord resulteert in een kolom van de matrix. Hij kan worden *ontdekt* omdat geen kolom $\begin{pmatrix}0\\0\\0\end{pmatrix}$ voorkomt; de plaats ervan kan worden gevonden en hij kan dus worden *verbeterd* omdat er geen twee kolommen gelijk

zijn. Merk wel op dat twee fouten tot misverstand leiden: een fout in het zesde bit samen met een in het zevende bit geeft als syndroom (bij de laatste matrix) $\begin{pmatrix}0\\0\\1\end{pmatrix}$; de ontvanger denkt dat er één fout zit en wel op positie 1.

Dezelfde redenering leert dat de kern van een matrix met 4 rijen ook een 1-fout-verbeterende code is mits de kolommen verschillend en niet 0 zijn. Je kunt $2^4 - 1 = 15$ verschillende kolommen maken; de rang is dan 4 en de kern is een lineaire deelruimte van $\mathbb{Z}_2^{15}$ van dimensie $15 - 4 = 11$ We hebben dus 2^{11} codewoorden, genoeg om boodschappen van lengte 11 te coderen met codewoorden van lengte 15, vectoren in $\mathbb{Z}_2^{15}$.

We hebben nu 2^{11} bolletjes met inhoud $1+15 = 2^4$, en omdat $2^{11} \times 2^4 = 2^{15}$ hebben we weer een perfecte code, de Hamming code H_4.

Merk op: is M_3 een parity-check-matrix voor H_3, dan is de volgende matrix er een voor H_4

$$\begin{pmatrix} 0 & \cdots\cdots & 0 & 1 & \cdots\cdots & 1 \\ & & & 0 & & \\ & M_3 & & \vdots & M_3 & \\ & & & 0 & & \end{pmatrix}$$

Zo kun je doorgaan. Een matrix met m rijen en het maximale aantal van $2^m - 1$ verschillende kolommen niet 0 levert als kern een code met 2^{2^m-1-m} codewoorden in $\mathbb{Z}_2^{2^m-1}$ en disjuncte bolletjes met inhoud 2^m die samen de ruimte vullen ($2^{2^m-1-m} \times 2^m = 2^{2^m-1}$), weer een perfecte code.

De rate van zo'n code is $\frac{2^m-1-m}{2^m-1}$, voor grote m dus bijna 1. Bedenk wel: hoe langer de codewoorden, hoe groter de kans op meer dan één fout, en verkeerde decodering. Bij langere codewoorden is betere fout-correctie nodig.

7. Niet-binaire Hamming-codes

Coderingstheorie is niet beperkt is tot binaire codes, die over het lichaam $\mathbb{F}_2$ dus. In deze paragraaf wordt dat kort geïllustreerd. Daarna zullen we ons (behalve in §14) eenvoudigheidshalve alleen met binaire codes bezighouden, al spelen ook daarbij andere eindige lichamen een grote rol.

In plaats van uit woorden van 0'en en 1'en kunnen codewoorden ook bestaan uit woorden van elementen uit een ander priemlichaam dan $\mathbb{Z}_2$, bijvoorbeeld $\mathbb{Z}_7$ met de elementen $\overline{0}, \overline{1}, \overline{2}, \overline{3}, \overline{4}, \overline{5}, \overline{6}$. Die zijn via een binair kanaal te zenden door elk van die zeven elementen voor te stellen door een ander rijtje van drie bits (zend bijvoorbeeld $\overline{3}$ als 011, $\overline{4}$ als 100, etc.). Ook nu geldt als afstand tussen twee woorden het aantal posities waarin ze verschillen.

We proberen weer een 1-fout-verbeterende code te maken met behulp van een parity-check matrix, zeg met 4 rijen. Nu is het niet genoeg om de positie van een fout te kennen; als een $\overline{2}$ op positie 3 fout is, wil je ook nog weten welk van de zes andere elementen het moest zijn. Stel het moest $\overline{5}$ zijn. Dat betekent dat bij het codewoord de **foutvector** $(\overline{0}, \overline{0}, \overline{4}, \overline{0}, \ldots, \overline{0})$ is opgeteld ($5 + 4 \equiv 2 \pmod 7$). De matrix levert dan als syndroom niet de nulvector maar zijn derde kolom, vermenigvuldigd met $\overline{4}$. Mits geen enkele andere kolom van de matrix

een veelvoud is van de derde kolom, weet de ontvanger dat de fout in positie 3 zit en dat daar $\overline{4}$ moet worden afgetrokken. Geen enkele kolom moet dus een veelvoud zijn van een andere.

Hoeveel kolommen kunnen er maximaal zijn in het algemene geval? Neem m rijen en werk met $\mathbb{Z}_p$, p een priemgetal. Kolommen zijn vectoren $\neq 0$ in $\mathbb{Z}_p^m$. Daarvan zijn er $p^m - 1$, verdeeld in groepjes van $p-1$ die veelvoud zijn van elkaar. We kunnen er uit elk groepje één gebruiken, dus het maximaal aantal kolommen is $L := \frac{p^m-1}{p-1}$. Dat is tevens de lengte van de codewoorden.

De rang van zo'n matrix is m, dus de codewoorden vormen een deelruimte van dimensie $L - m$ in $\mathbb{F}_p^L$. Er zijn er dus p^{L-m}. Op elk van de L plaatsen in een woord kan het op $p-1$ manieren fout gaan, dus de inhoud van een bolletje met straal 1 is $1 + (p-1) \cdot L = p^m$. Samen overdekken de bolletjes dus $p^{L-m} \times p^m = p^L$ punten, dus de gehele ruimte. Weer een perfecte, nu p-aire, code.

p moet een priemgetal zijn, anders is $\mathbb{Z}_p$ geen lichaam en kunnen we geen goede algebra bedrijven. Het gaat ook goed bij lichamen $\mathbb{F}_q$, q een priemmacht.

Perfecte codes zijn zeldzaam. Bewezen is (door van Lint en Tietäväinen): behalve de Hamming-codes (en sommige niet-lineaire codes met dezelfde lengte, hetzelfde aantal codewoorden en dezelfde minimumafstand 3 als de Hamming-codes) zijn er nog maar twee, de Golay-codes.

8. Lineaire codes

$\mathbb{F}_2$ heeft twee elementen, $\overline{0}$ en $\overline{1}$. In het vervolg laten we die streepjes gemakshalve vaak weg. We schrijven dan $1 + 1 = 0$ als we bedoelen $\overline{1} + \overline{1} = \overline{0}$.

In de vectorruimte $\mathbb{F}_2^n$ definiëer je op de gebruikelijke wijze een inproduct: als $x = (a_1, a_2, \ldots, a_n)$ en $y = (b_1, b_2, \ldots, b_n)$, is $(x, y) = a_1b_1 + a_2b_2 + \cdots + a_nb_n$. Het is een lineaire functie van x en van y. Je zegt dat x loodrecht staat op y ($x \perp y$) als $(x, y) = 0$. Een vector kan loodrecht op zichzelf staan ($(x, x) = 0$ als x even gewicht heeft). Een vector staat loodrecht op alle vectoren van een lineaire deelruimte T van dimensie r als hij loodrecht staat op de r vectoren van een basis van T. Zulke vectoren vormen een lineaire deelruimte $T^\perp$ van dimensie $n - r$, een kwestie van lineaire vergelijkingen. $(T^\perp)^\perp = T$. Is $T = \{(0,0,0,0), (1,0,0,0), (0,1,0,0), (1,1,0,0)\}$ in $\mathbb{F}_2^4$, dan is $T^\perp = \{(0,0,0,0), (0,0,1,0), (0,0,0,1), (0,0,1,1)\}$. Maar is $T = \{(0,0,0,0), (1,1,0,0), (0,0,1,1), (1,1,1,1)\}$, dan is $T^\perp = T$.

Een code van lengte n heet **lineair** als de codewoorden een lineaire deelruimte vormen van F_2^n (in ons binaire geval wil dat alleen maar zeggen dat de som van twee codewoorden weer een codewoord is). De kern van een parity-check-matrix is zo'n lineaire code.

Een lineaire code C heeft een dimensie, zeg r. Er is dan een basis van r codewoorden. Gebruik je die als rijen van een $r \times n$-matrix, dan is elk codewoord

een lineair combinatie van die rijen (met coëfficiënten 0 en 1, dus de som van een aantal van die rijen). Er zijn dus 2^r codewoorden. Die matrix is een generatormatix van de code.

Bij een lineaire code C kun je de **duale** code $C^\perp$ maken. Die bestaat uit de woorden die loodrecht staan op alle woorden van C; het is de kern van de generatormatrix. De generatormatrix van $C^\perp$ heeft $n-r$ rijen en zijn kern is weer $(C^\perp)^\perp = C$. Het is dus een parity-check-matrix van C.

In het begin van §5 zagen we een generatormatrix en de bijbehorende parity-check-matrix van H_3.

$$G = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 \end{pmatrix}$$

$$K = \begin{pmatrix} 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 \end{pmatrix}$$

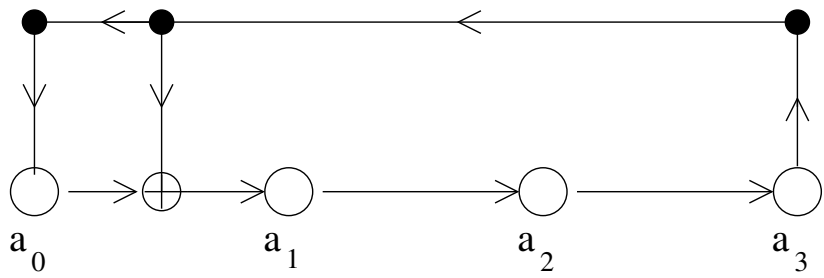
De loodrechtheid brengt mee dat GK^t en KG^t nulmatrices zijn.

9. Lichaamsuitbreiding

Een lichaam is een verzameling met twee operaties, optelling en vermenigvuldiging, waarvoor de 'gewone' rekenregels gelden: beide operaties zijn associatief en commutatief, hebben een neutraal element (0 resp. 1), elk element x heeft een tegengestelde $-x$ en (mits $x \neq 0$) een inverse x^{-1} en de distributieve wet geldt $(x(y+z) = xy + xz)$. De bekende voorbeelden zijn $\mathbb{Q}$, $\mathbb{R}$ en $\mathbb{C}$. Eenvoudige voorbeelden van eindige lichamen zijn de priemlichamen $\mathbb{F}_p = \mathbb{Z} \bmod p$, waarin p een priemgetal is.

Een lichaam kan worden uitgebreid door **adjungeren** van een nulpunt van een irreducibel (niet ontbindbaar) polynoom. Zo'n polynoom heeft geen nulpunt in het gegeven lichaam. Als voorbeeld (zonder bewijs) maken we uit $\mathbb{Z}_2$ (elementen $\overline{0}$ en $\overline{1}$, de streepjes laten we verder maar weg) een lichaam $\mathbb{F}_8$ met 8 elementen. (Bedenk: de **karakteristiek** is 2, d.w.z. $2 \cdot x = 0$, dus $x = -x$ en $(x+y)^2 = x^2 + y^2$). Als polynoom nemen we $x^3 + x + 1$. De elementen 0 en 1 van $\mathbb{Z}_2$ zijn geen nulpunt ervan. Introduceer een nulpunt α. Daarvoor geldt $\alpha^3 = \alpha + 1$. Het lichaam bestaat nu uit 0, 1, α, $1+\alpha$, α^2, $1+\alpha^2$, $\alpha+\alpha^2$, $1+\alpha+\alpha^2$. Elk element is een lineaire combinatie, met coëfficiënten 0 en 1, van de basiselementen $1, \alpha, \alpha^2$. De som van twee zulke elementen is weer zo'n element.

De elementen $\neq 0$ zijn ook te schrijven als machten van α (α is een **primitief** element). Zie $\alpha^0 = 1$, $\alpha^1 = \alpha$, α^2, $\alpha^3 = \alpha + 1$, $\alpha^4 = \alpha^2 + \alpha$, $\alpha^5 = \alpha^3 + \alpha^2 = \alpha^2 + \alpha + 1$, $\alpha^6 = \alpha^3 + \alpha^2 + \alpha = \alpha^2 + 1$, $\alpha^7 = \alpha^3 + \alpha = 1$. Daarmee zie je dat het product van twee van die elementen ook zo'n element is, bijvoorbeeld



Figuur 5. *Vermenigvuldigen met* α

$\alpha^6 \cdot \alpha^5 = \alpha^{11} = \alpha^7 \cdot \alpha^4 = \alpha^4$. Net zo voor inversen, bijvoorbeeld $\alpha^{5^{-1}} = \alpha^2$ want $\alpha^5 \cdot \alpha^2 = 1$.

Voor elke priemmacht p^n is zo'n uitbreiding van $\mathbb{Z}_p$ tot een lichaam met p^n elementen mogelijk, met een irreducibel polynoom van graad n.

Een lichaam met 16 elementen kun je maken met, bijvoorbeeld, de irreducibele veelterm $x^4 + x + 1$. Een nulpunt α is primitief, $\alpha^4 = \alpha + 1$ en $1, \alpha, \alpha^2, \alpha^3$ vormen een basis. Rekenwerk in zo'n lichaam kan met speciale electronica worden uitgevoerd ('wiring'). Figuur 5 toont een schakeling die de vermenigvuldiging met α uitvoert

$$\alpha \times (a_0 + a_1\alpha + a_2\alpha^2 + a_3\alpha^3) = a_3 + (a_0 + a_3)\alpha + a_1\alpha^2 + a_2\alpha^3$$

Op commando van een klok wordt de inhoud (0 of 1) van de vier flip-flops volgens de pijlen verstuurd. $\oplus$ is een 'binary adder'. De coördinaten (a_0, a_1, a_2, a_3) worden omgezet in $(a_3, a_0 + a_3, a_1, a_2)$.

N.B. $x^4+x^3+x^2+x+1$ is ook irreducibel. Ook adjungeren van een nulpunt β van déze veelterm levert een lichaam van 16 elementen, met basis $1, \beta, \beta^2, \beta^3$ en met $\beta^4 = \beta^3 + \beta^2 + \beta + 1$. Maar β is niet primitief, want $\beta^5 = 1$. Overigens is er, op isomorfie na, voor elke priemmacht p^n precies één lichaam met p^n elementen.

10. Meer fouten verbeteren

We beperken ons verder tot binaire codes. Voor een code die twee fouten kan verbeteren moet de minimum afstand 5 zijn. Een parity-check matrix voor zo'n code moet de eigenschap hebben dat elke foutvector met één of twee 1'en erin een ander syndroom $\neq 0$ oplevert. Zo'n syndroom is een kolom van de matrix of de som van twee kolommen. Een (te) eenvoudig voorbeeld is $\begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$. De drie kolommen en de drie sommen van twee kolommen zijn alle zes verschillend. De code (de kern) bestaat helaas alleen uit 000. Bij ontvangst van 110 is het syndroom $(1, 1, 0, 0)$, de som van de eerste twee kolommen. De snuggere ontvanger besluit dat 000 is verzonden (door bij het ontvangen woord $(1, 1, 0)$ de foutvector $(1, 1, 0)$ op te tellen.

De matrix is ontstaan door verlengen van de kolommen van de parity-checkmatrix $\begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix}$ van de Hamming-code H_2. Toevoegen van rijen heeft de rang verhoogd naar 3, de kern $(0,0,0),(1,1,1)$ is daardoor verkleind tot $(0,0,0)$.

We willen zoiets proberen met de Hamming-code H_3. Het resultaat zal nog steeds teleurstellend zijn, maar dit voorbeeld is eenvoudig en representatief voor het principe.

De parity-checkmatrix van H_3 bezien we eerst eens met andere ogen, door het primitieve element α van $\mathbb{F}_8$ uit §9 te gebruiken.

$$\begin{array}{c} \\ 1 \\ \alpha \\ \alpha^2 \end{array} \begin{array}{c} \begin{array}{ccccccc} 1 & \alpha & \alpha^2 & \alpha^3 & \alpha^4 & \alpha^5 & \alpha^6 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \end{pmatrix} \end{array}$$

De kolommen zijn de machten van α, uitgedrukt in de basiselementen $1, \alpha, \alpha^2$. In de kolom onder α^4 lees je dat $\alpha^4 = \overline{0} \cdot 1 + \overline{1} \cdot \alpha + \overline{1} \cdot \alpha^2$.

Het codewoord 1000110 associeerden we al met de vector $(1,0,0,0,1,1,0)$ in de kern van de matrix, alles over $\mathbb{Z}_2$. De matrix toepassen op die vector betekent de som nemen van de kolommen onder 1, α^4 en α^5. Dat geeft $\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$, voorstellende $\overline{0} \cdot 1 + \overline{0} \cdot \alpha + \overline{0} \cdot \alpha^2$, het nulelement van $\mathbb{F}_8$. Dat betekent dat (in $\mathbb{F}_8$) geldt: $1+\alpha^4+\alpha^5 = 0$, dus dat α een nulpunt is van de veelterm $1+x^4+x^5$.

Algemeen: $a_0a_1a_2a_3a_4a_5a_6$ is een codewoord precies dan als α nulpunt is van de veelterm $a_0 + a_1x + a_2x^2 + a_3x^3 + a_4x^4 + a_5x^5 + a_6x^6$.

We zien dat we de codewoorden ook kunnen associëren met de veeltermen van graad ≤ 6 met coëfficiënten in $\mathbb{Z}_2$ die α als nulpunt hebben. Het idee (van Bose/Chauduri en Hocquenghem, $\pm$1960) is nu een grotere minimum afstand te bereiken door de code 'uit te dunnen' **expurgation**) en wel door van de veeltermen te eisen dat ze meer nulpunten hebben dan alleen α.

Nu is α^2 geen geschikte keuze, want als α nulpunt is, is α^2 het vanzelf ook al: is (bijvoorbeeld) $1+\alpha^4+\alpha^5 = 0$, dan is ook $1+(\alpha^2)^4+(\alpha^2)^5 = (1+\alpha^4+\alpha^5)^2 = 0$ (denk aan $(a+b)^2 = a^2 + b^2$).

Probeer dan α^3. We eisen dus dat een codewoord ook in de kern zit van de matrix waarvan de kolommen voorstellen $1, \alpha^3, \alpha^6, \alpha^9, \alpha^{12}, \alpha^{15}, \alpha^{18}$, d.w.z. van de matrix

$$\begin{array}{c} \\ 1 \\ \alpha \\ \alpha^2 \end{array} \begin{array}{c} \begin{array}{ccccccc} 1 & \alpha^3 & \alpha^6 & \alpha^2 & \alpha^5 & \alpha & \alpha^4 \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 0 & 1 \end{pmatrix} \end{array}$$

De nieuwe parity-checkmatrix wordt daarmee (combineer de matrices)

$$\begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 0 & 1 \end{pmatrix}$$

Waarom is de bijbehorende code nu inderdaad 2-fout-verbeterend (minimum afstand ≥ 5)?

De 7 mogelijke fout-vectoren van gewicht 1 zullen syndromen van de vorm $\begin{pmatrix} \alpha^k \\ \alpha^{3k} \end{pmatrix}$ opleveren, allemaal verschillend dus. De 21 mogelijke fout-vectoren van gewicht 2 zullen syndromen van de vorm $\begin{pmatrix} \beta+\gamma \\ \beta^3+\gamma^3 \end{pmatrix}$ opleveren, met β en γ verschillende machten van α. Omdat dan $(\beta+\gamma)^3 = \beta^3+\gamma^3+\beta\gamma(\beta+\gamma) \neq \beta^3+\gamma^3$ kan het syndroom van een foutvector van gewicht 2 niet hetzelfde zijn als dat van een foutvector van gewicht 1. Bovendien zijn uit dat syndroom β en γ, dus de plaatsen van de twee fouten, terug te vinden: uit $\beta+\gamma$ en $\beta^3+\gamma^3$ bereken je $\beta\gamma(\beta+\gamma)$ (dat is $(\beta+\gamma)^3+\beta^3+\gamma^3$) en daaruit $\beta\gamma$. Uit $\beta+\gamma$ en $\beta\gamma$ bereken je β en γ (met de goede oude vierkantsvergelijking).

Qua bruikbaarheid is ons voorbeeld niet indrukwekkend. De matrix heeft rang 6, de kern dus dimensie 1, de code bestaat uit het 'all-zero' woord en het 'all-one' woord van lengte 7. De twee bollen van straal 2 om de codewoorden bevatten elk $1+7+\binom{7}{2} = 29$ van de $2^7 = 128$ woorden in $\mathbb{Z}_2^7$.

11. BCH-codes

Het principe van de vorige paragraaf laat zich op grote schaal toepassen. Langere en/of meer codewoorden kun je maken door grotere lichamen te gebruiken, correctie van meer fouten kun je bereiken door nog meer nulpunten voor het polynoom te eisen.

Een voorbeeld. Met behulp van een irreducibel polynoom van graad 7 (bv. x^7+x+1 of $x^7+x^5+x^3+x+1$) breid je $\mathbb{Z}_2$ uit tot een lichaam met $2^7 = 128$ elementen. Neem een primitief element δ. Dan zijn de elementen van het lichaam $1, \delta, \delta^2, \ldots\ldots \delta^{125}, \delta^{126}$, en $\delta^{127} = 1$. Ze zijn ook te schrijven als lineaire combinaties van de basiselementen $1, \delta, \delta^2, \delta^3, \delta^4, \delta^5, \delta^6$, met coëfficiënten uit $\mathbb{Z}_2$.

Gebruik deze machten, uitgedrukt in de basiselementen, als kolommen van een 7×127 matrix. We hebben dan een parity-checkmatrix van H_7.

Net zo maken we een matrix bij de opvolgende machten van δ^3 (kolommen met $1, \delta^3, \delta^6, \ldots\ldots, \delta^{375} = \delta^{121}, \delta^{378} = \delta^{124}$.

Doe hetzelfde nog eens, nu met machten van δ^5, met die van δ^7 en met die van δ^9.

Als we de vijf matrices onder elkaar zetten hebben we een 35×127 parity-checkmatrix.

De codewoorden zijn de coëfficiënten-rijtjes van de veeltermen van graad < 127 die δ, δ^3, δ^5, δ^7 en δ^9 als nulpunt hebben en dus vanzelf ook δ^2, δ^4, δ^6, δ^8 en δ^{10}; die vijf matrices hoefden er dus niet ook nog bij.

Er kan bewezen worden (met Vandermonde matrices) dat de minimum afstand van deze code ≥ 11 is, hij kan dus vijf fouten verbeteren. Omdat de matrix 35 rijen heeft is de rang ≤ 35. In feite is hij 35 en heeft de kern dus dimensie $127 - 35 = 92$. Er zijn dus 2^{92} codewoorden. Dit is de code uit het kader op pagina 52.

Bij de binaire BCH-code van lengte $2^m - 1$ bestaande uit de veeltermen van graad $< 2^m - 1$ die $\delta, \delta^2, \ldots, \delta^k$ als nulpunt hebben (δ primitief in $\mathbb{F}_{2^m}$) heet $k + 1$ de **ontwerpafstand**. Bewezen kan worden dat de minimumafstand ten minste $k+1$ is. De rang van de parity-check-matrix is $\leq km$; de code heeft dus dimensie $\geq (2^m - 1) - km$, de rate is dus $\geq \frac{(2^m-1)-km}{2^m-1}$. Voor grote m komt dit dicht bij 1.

Ook over andere eindige lichamen dan $\mathbb{F}_2$ kunnen, op analoge manier, BCH codes worden gedefinieerd.

De muziek op een CD is ook gecodeerd. In (zeer) grote lijnen gaat dat als volgt. Bij de opname wordt er van het linker- en van het rechterkanaal 44100 keer per seconde een 'sample' genomen die in 2 bytes wordt opgeslagen (dat zijn 1411200 bits per seconde). Telkens 24 bytes worden gegroepeerd tot een 'frame'. Zo'n frame wordt gecodeerd (met een Reed-Solomon code, een speciaal soort BCH-code), waarbij 4 bytes worden toegevoegd, en daarna wordt op de 28 bytes nog eens een Reed-Solomon codering toegepast, weer 4 bytes erbij. Dan wordt ook nog elke byte vertaald in een woord van 14 bits. Intussen worden de bytes van een frame gemengd met die van andere frames, zodat ze op de schijf niet te dicht op elkaar staan (een krasje tast zo niet een heel frame aan) en er komen nummers bij en wat bits voor informatie en synchronisatie. Uiteindelijk leidt elk frame van $24 \times 8 = 192$ bits tot 588 bits op de schijf; zo'n bit wordt weergegeven doordat het spoor op de schijf wel of niet overgaat van een 'putje' in een 'dammetje' of andersom. De afspeler leest per seconde ruim 4 miljoen bits af van 125 cm spoor (totale lengte van het spoor ruim 5 km met 20000 windingen), waarvoor het toerental van de schijf moet variëren van 3.5 (aan de buitenrand) tot 8, terwijl de leeskop nauwkeurig boven het juiste spoor moet blijven (speling $< 0.1\mu$m). Foutcorrectie en decodering geschieden door het afspeelapparaat, dat tenslotte het digitale signaal omzet in een analoog signaal voor versterker en luidsprekers.

12. Cyclische codes. Meer algebra

Sommige codes hebben de eigenschap dat bij elk codewoord $(a_1, a_2, \ldots, a_n)$ ook $(a_n, a_1, a_2, \ldots, a_{n-1})$ een codewoord is, m.a.w. een 'cyclic shift' toegepast op een codewoord levert weer een codewoord op.

Ook BCH-codes hebben die eigenschap. Dat zie je het gemakkelijkst door codewoorden weer als veeltermen op te vatten. Een voorbeeld van de redenering moge volstaan. We gebruiken de code van lengte 127 uit §11. Voor het primitieve element δ in $\mathbb{F}_{128}$ gold $\delta^{127} = 1$.

Dat $(a_0, a_1, \ldots\ldots, a_{126})$ een codewoord is (de a_i in $\mathbb{Z}_2$) betekent dat de veelterm

$$a_0 + a_1x + a_2x^2 + \cdots\cdots + a_{126}x^{126}$$

(o.a.) δ, δ^3, δ^5, δ^7 en δ^9 als nulpunten heeft.

Dan is, voor bijvoorbeeld δ^5,

$$a_0 + a_1\delta^5 + a_2\delta^{10} + \cdots\cdots + a_{126}(\delta^5)^{126} = 0. \tag{1}$$

Bedenk nu dat $\delta^5 \cdot (\delta^5)^{126} = (\delta^5)^{127} = (\delta^{127})^5 = 1$ en vermenigvuldig (1) met δ^5. Er komt

$$a_{126} + a_0\delta^5 + a_1\delta^{10} + \cdots\cdots + a_{125}\delta^{5\cdot 126} = 0 \tag{2}$$

Net zo voor de andere nulpunten. Dus de veelterm

$$a_{126} + a_0x + a_1x^2 \cdots\cdots + a_{125}x^{126}$$

heeft ook die nulpunten, en $(a_{126}, a_0, a_1, \ldots\ldots, a_{125})$ is dus een codewoord.

Een code heet **cyclisch** als hij die bewuste eigenschap heeft en bovendien lineair is. BCH-codes zijn lineair als kernen van matrices. (Voor binaire codes betekent 'lineair' slechts dat de som van twee codewoorden weer een codewoord is).

Vatten we de codewoorden van een cyclische code C van lengte n weer op als veeltermen van graad $< n$, dan vinden we:

$$c_0+c_1x+\cdots+c_{n-2}x^{n-2}+c_{n-1}x^{n-1} \in C \Rightarrow c_{n-1}+c_0x+c_1x^2+\cdots+c_{n-2}x^{n-1} \in C.$$

Blijkbaar is de verzameling van die (code)veeltermen 'bestand' tegen vermenigvuldigen met x, mits we daarbij met de veeltermen rekenen alsof $x^n = 1$ (ofwel modulo $(x^n - 1)$). Maar dan ook tegen vermenigvuldigen met x^2, x^3, enzovoort (steeds gebruikend $x^n \equiv 1$, $x^{n+1} \equiv x$, $x^{n+2} \equiv x^2$, etc.).

(Formeler: werk in de veeltermring $\mathbb{F}_2[x]$ modulo het hoofdideaal $(x^n - 1)$; de codeveeltermen worden gezien als representanten van zekere elementen van die quotiëntring $\mathbb{F}_2[x]/(x^n - 1)$).

Laat nu $g(x)$ een codeveelterm $\neq 0$ zijn van laagste graad in een cyclische code, zeg die graad is s. De hoogste term van $g(x)$ is x^s (de enig mogelijke coëfficient $\neq 0$ is immers 1). Er is geen andere codeveelterm van graad s, anders zou hun som (wat hetzelfde is als hun verschil) een codeveelterm $\neq 0$ van lagere graad dan s zijn. Zij $f(x)$ een codeveelterm van graad r, $r > s$. Dan is $x^{r-s}g(x)$ en dus ook $f(x) - x^{r-s}g(x)$, een codeveelterm. Die heeft graad $< r$. Doorgaande met aftrekken van geschikte veeltermen van de vorm $x^jg(x)$ moeten we tenslotte een veelterm krijgen van graad $< s$, en dat kan alleen 0 zijn.

We zien: $f(x)$ is een som van veeltermen van de vorm $x^j g(x)$, dus $f(x) = t(x)g(x)$, met $t(x)$ een veelterm (van graad $r - s < n - s$). Omgekeerd is elke veelterm van de vorm $t(x)g(x)$, met $t(x)$ van graad $< n - s$, een som van veeltermen van de vorm $x^j g(x)$, dus een codeveelterm.

De codeveeltermen zijn dus de lineaire combinaties (met coëfficiënten 0 en 1) van $g(x), xg(x), x^2g(x), \ldots, x^{n-s-1}g(x)$. Deze vormen een basis van de code; ze zijn onafhankelijk omdat ze respectievelijk graad $s, s+1, \ldots, n-1$ hebben.

Nu is $x^{n-s}g(x)$, na weer vervangen van x^n door 1, ook een codeveelterm, dat is immers een cyclic shift. Idem voor hogere machten van x, en dus ook voor sommen van dergelijke machten, Dat betekent dat ook voor veeltermen $t(x)$ van graad $\geq n - s$ geldt dat $t(x)g(x)$, na reduceren mod $(x^n - 1)$, een codeveelterm is (maar geen nieuwe, want we hadden ze al allemaal).

Deel nu $x^n - 1$ door $g(x)$ ('staartdeling'): $x^n - 1 = d(x)g(x) + r(x)$. De rest $r(x)$ is een veelterm vam graad $< s$. Reduceer beide leden mod $(x^n - 1)$; het linkerlid wordt 0, het rechterlid wordt een codeveelterm$+r(x)$. Dus $r(x)$ is ook een codeveelterm. Omdat zijn graad $< s$ is, moet $r(x) = 0$ zijn: $g(x)$ is een deler van $x^n - 1$.

Er is dus een veelterm $k(x)$ van graad $n - s$ zo, dat $x^n - 1 = k(x)g(x)$. De veeltermen $g(x)$ en $k(x)$ zijn het **generator-polynoom** en het **parity-check-polynoom** van de cyclische code. Het waarom staat in de volgende paragraaf.

13. Generator-polynoom en parity-check-polynoom

Bij elk paar polynomen $g(x)$ en $k(x)$ met $k(x)g(x) = x^n - 1$ kunnen we een cyclische code maken met woorden van lengte n. Heeft $g(x)$ graad s, dan vormen de veeltermen $g(x), xg(x), \ldots, x^{n-s-1}g(x)$ een basis van de code, die dus dimensie $n - s$ heeft, dus 2^{n-s} boodschappen kan coderen. (Die code is cyclisch: $k(x) = x^{n-s} + t(x)$, $t(x)$ van graad $< n - s$, dus $x^{n-s}g(x) = (k(x) - t(x))g(x) = k(x)g(x) + t(x)g(x) \equiv t(x)g(x)$). Alleen bij speciale n (van de vorm $2^m - 1$) en $g(x)$ (met speciale nulpunten) is het een BCH-code.

Een veelterm $c(x)$ van graad $< n$ kun je delen door $g(x)$ met een rest: $c(x) = t(x)g(x) + r(x)$, waarbij $r(x)$ graad $< s$ heeft en 0 is precies dan als $c(x)$ in de code zit. Dan is

$$\begin{aligned} k(x)c(x) &= t(x)k(x)g(x) + k(x)r(x) = t(x)(x^n - 1) + k(x)r(x) \equiv \\ &\quad k(x)r(x) \bmod (x^n - 1). \end{aligned}$$

Is $c(x)$ een codewoord, dan is dus $k(x)c(x) \equiv 0 \bmod (x^n - 1)$, en anders niet, want de graad van $k(x)r(x)$ is $< n$.

Testen of een woord een codewoord is kan dus door zijn veelterm te vermenigvuldigen met $k(x)$ en te kijken of het resultaat $\equiv 0 \bmod (x^n - 1)$ is. In het volgende zien we het verband tussen $k(x)$ en een parity-check-matrix.

Stel $g(x) = g_0 + g_1 x + \cdots + g_s x^s$ en $k(x) = k_0 + k_1 x + \cdots + k_{n-s} x^{n-s}$. De generatormatrix bestaat uit de $n - s$ coëfficiëntenrijen van de veeltermen

$x^j g(x)$, $j = 1 \ldots n-s$:

$$G = \begin{pmatrix} g_0 & g_1 & \cdots & \cdots & \cdots & g_{s-1} & g_s & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & g_0 & g_1 & \cdots & \cdots & \cdots & g_{s-1} & g_s & \ddots & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & & & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & & & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & \ddots & \ddots & & & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & \cdots & 0 & g_0 & g_1 & \cdots & \cdots & \cdots & g_{s-1} & g_s \end{pmatrix}$$

Zo maken we ook een matrix bij $k(x)$, maar schrijven de s rijen op *in omgekeerde volgorde*.

$$K = \begin{pmatrix} 0 & \cdots & \cdots & \cdots & 0 & k_{n-s} & \cdots & \cdots & \cdots & \cdots & k_1 & k_0 \\ \vdots & & & \iddots & k_{n-s} & \cdots & \cdots & \cdots & \cdots & k_1 & k_0 & 0 \\ \vdots & & \iddots & \iddots & & & & & \iddots & \iddots & \iddots & \vdots \\ \vdots & \iddots & \iddots & & & & & \iddots & \iddots & \iddots & & \vdots \\ 0 & \iddots & & & & & \iddots & \iddots & \iddots & & & \vdots \\ k_{n-s} & \cdots & \cdots & \cdots & \cdots & k_1 & k_0 & 0 & \cdots & \cdots & \cdots & 0 \end{pmatrix}$$

Het inproduct van de eerste rij van G met de eerste rij van K is $g_{s-1}k_{n-s} + g_s k_{n-s-1}$; dat is de coëfficiënt van x^{n-1} in $g(x)k(x)$. De eerste rij van G met de tweede van K (of de tweede van G met de eerste van K) geeft $g_{s-2}k_{n-s} + g_{s-1}k_{n-s-1} + g_s k_{n-s-2}$; dat is de coëfficiënt van x^{n-2} in $g(x)k(x)$. Zo doorgaande vinden we bij de twee laatste rijen $g_0k_1 + g_1k_0$, de coëfficiënt van x^1. Al die coëfficiënten zijn echter 0 want $g(x)k(x) = x^n - 1$.

De $n-s$ onafhankelijke rijen van G staan dus loodrecht op de s onafhankelijke rijen van K. Ze spannen dus de kern van K op. K is blijkbaar een parity-check-matrix van de code en de generator-matrix van de duale code, die dus als generator-polynoom $k_{n-s} + k_{n-s-1}x + \cdots + k_1 x^{n-s-1} + k_0 x^{n-s} = x^{n-s}k(x^{-1})$ heeft.

We gebruiken als voorbeeld (toch maar) een BCH-code van lengte 15 die twee fouten kan verbeteren.

We maken het lichaam $\mathbb{F}_{16}$ door aan $\mathbb{F}_2$ een nulpunt α van $x^4 + x + 1$ te adjungeren; α is primitief. Daarmee vinden we een parity-check-matrix voor de Hamming-code H_4:

$$\begin{array}{c} \\ 1 \\ \alpha \\ \alpha^2 \\ \alpha^3 \end{array} \begin{array}{c} \begin{array}{cccccccccccccc} 1 & \alpha & \alpha^2 & \alpha^3 & \alpha^4 & \alpha^5 & \alpha^6 & \alpha^7 & \alpha^8 & \alpha^9 & \alpha^{10} & \alpha^{11} & \alpha^{12} & \alpha^{13} & \alpha^{14} \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 1 & 1 \end{pmatrix} \end{array}$$

De kern bestaat uit de coëfficiënten-rijtjes van de veeltermen in $\mathbb{Z}_2[X]$ die nulpunt α hebben. Door die veeltermen ook nog het nulpunt α^3 'op te dringen'

komen er in de parity-check-matrix nog vier rijen bij. We weten dan dat de rang ≤ 8 is, dus de code heeft dimensie $\geq 15 - 8 = 7$.

De veelterm van laagste graad die α als nulpunt heeft is $1 + x + x^4$; de veelterm van laagste graad die α^3 als nulpunt heeft is $1 + x + x^2 + x^3 + x^4$. De veelterm van laagste graad die α en α^3 als nulpunten heeft is hun product $g(x) = 1 + x^4 + x^6 + x^7 + x^8$. Dat is dus het generator-polynoom van de code. Het codewoord daarbij is 100010111000000. Samen met de codewoorden bij $xg(x), x^2g(x), \ldots, x^6g(x)$ vormt het een basis voor de code. Ze staan in de rijen van de 7×15 generator-matrix:

$$G = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \end{pmatrix}$$

De dimensie van de code is blijkbaar precies 7. Er zijn 2^7 codewoorden van lengte 15. Je kunt het codewoord bij een boodschap $b = (b_0, b_1, \ldots, b_6)$ maken door een lineaire combinatie van de rijen te maken met als coëfficiënten $b_0, b_1, \ldots, b_6$. Dat komt neer op het (matrix-)product bG. In veeltermvorm: je maakt $b_0g(x) + b_1x(g(x) + \cdots + b_6x^6g(x)$, dus je vermenigvuldigt $b_0 + b_1x + \cdots + b_6x^6$ met het generator-polynoom $g(x)$.

Omdat $x^{15} - 1 = (1 + x^4 + x^6 + x^7 + x^8)(1 + x^4 + x^6 + x^7)$ is $k(x) = 1 + x^4 + x^6 + x^7$ het parity-check-polynoom.

Om een ontvangen woord $w = w_0w_1 \ldots w_{14}$ te testen moet zijn veelterm worden vermenigvuldigd met $1 + x^4 + x^6 + x^7$.

De parity-check-matrix laat zich uit $k(x)$ bepalen:

$$K = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

De code kan twee fouten verbeteren, maar heeft nog een goede eigenschap die uit het cyclisch karakter volgt. Omdat het generator-polynoom graad 8 heeft kan geen codewoord bij een veelterm van graad 7 horen, dus al zijn 1'en in de eerste 8 posities hebben, ofwel eindigen op 7 nullen. Evenmin kan het dan elders 7 in cyclische zin op elkaar volgende nullen hebben. Een foutvector waarvan de fouten in 8 opeenvolgende posities voorkomen geeft dus een syndroom $\neq 0$. Zo'n 'burst error' wordt dus ontdekt. Algemeen: een cyclische code van lengte n met een generator-polynoom van graad s kan burst errors van lengte $\leq s$ ontdekken.

Bij BCH-codes kennen we de ontwerp-afstand, bij andere cyclische codes is het bepalen van de minimum-afstand minder eenvoudig.

14. De Golay-codes

Als voorbeeld van een binaire cyclische code die geen BCH-code is nemen we de Golay-code G_{23} (zie slot van §7)

In $\mathbb{F}_2[X]$ is $x^{23}-1=(x+1)g(x)k(x)$, waarbij

$$g(x)=x^{11}+x^{10}+x^6+x^5+x^4+x^2+1 \qquad k(x)=x^{11}+x^9+x^7+x^6+x^5+x+1$$

Beide veeltermen zijn irreducibel, dus we hebben de volledige ontbinding van $x^{23}-1$ (MAPLE doet dat voor je met '`Factor(x^(23)-1) mod 2`'). $g(x)$ is het generator-polynoom van een code van lengte 23 en dimensie $23-11=12$, dus met $2^{12}=4096$ codewoorden.

De minimum afstand is 7 (geen bewijs), dus de code is 3-fout-verbeterend. De bolletjes met straal 3 om een codewoord bevatten $1+23+\binom{23}{2}+\binom{23}{3}=2048=2^{11}$ woorden; ze bevatten samen dus $2^{12}\cdot 2^{11}=2^{23}$ woorden en overdekken de ruimte precies. Een perfecte code.

Hetzelfde geldt natuurlijk voor de code met generator-polynoom $k(x)$. Maar omdat $x^{11}k(x^{-1})=g(x)$ is dat dezelfde code met omgekeerde bit-volgorde. Merk op dat $(x+1)g(x)$ ook een cyclische code voortbrengt. Omdat daarbij de veeltermen ook nog nulpunt 1 moeten hebben, dus een even aantal termen, bestaat die uit de 2048 woorden van G_{23} die even gewicht hebben.

De andere Golay-code is een ternaire code en berust op de ontbinding

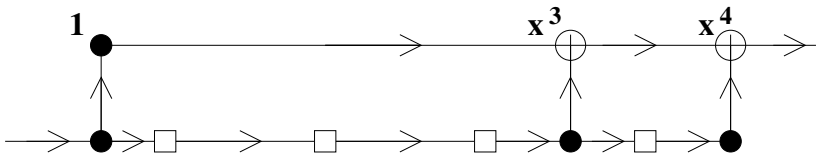
$$x^{11}-1=(x-1)(x^5+x^4-x^3+x^2-1)(x^5-x^3+x^2-x-1) \text{ in } \mathbb{F}_3[X].$$

Op soortgelijke manier als bij binaire cyclische codes kun je uit de veeltermen een generator-matrix en een parity-check-matrix maken. Ook nu geven de twee polynomen van graad 5 dezelfde codes, op de bitvolgorde na. De lengte is 11, de dimensie 6 en er zijn $3^6=729$ codewoorden. De minimum afstand is 5, de bolletjes van straal 2 bevatten elk $1+2\cdot 11+4\cdot\binom{11}{2}=3^5$ woorden, en $3^6\cdot 3^5=3^{11}$, een perfecte code.

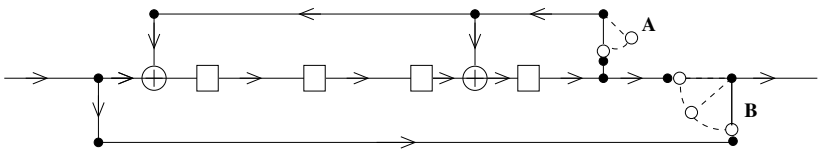
15. Coderen en decoderen

Om een boodschap te coderen met de generator-matrix moet eerst het hele boodschap-woord worden ingevoerd voordat deze met de matrix kan worden vermenigvuldigd. Op basis van een generator-polynoom is een snellere methode beschikbaar. Als voorbeeld nemen we de Hamming-code H_4 van lengte 15 met dimensie 11 en generator-polynoom $g(x)=x^4+x^3+1$ (in §13 gebruikten we x^4+x+1).

Voor de boodschap $b(x)=b_0+b_1x+\cdots+b_{10}x_{10}$ gebruiken we het codewoord $b(x)g(x)$. Dat levert een lineaire combinatie van basis-codeveeltermen: $b_0g(x)+b_1xg(x)+b_2x^2g(x)+\cdots+b_{10}x^{10}g(x)$. (De coëfficiënten van deze veelterm komen ook tevoorschijn als we de rij $b_0,b_1,\ldots,b_{10}$ rechts vermenigvuldigen met de generator-matrix die bij $g(x)$ hoort; de rijen daarvan corresponderen immers



Figuur 6. *Coderen*



Figuur 7. *Coderen*

met de $x^j g(x)$). De coëfficiënt van x^r in $b(x)g(x)$ is $b_r + b_{r-3} + b_{r-4}$ (lees $b_m = 0$ als $m < 0$ of $m > 10$).

Daarvoor wordt de volgende electronica gebruikt (zie Figuur 6).

Aanvankelijk staan de vier 'flip-flops' van het register op 0. Achtereenvolgens worden $b_0, b_1, \ldots, b_{10}$ ingevoerd waarbij telkens de inhoud van de flip-flops wordt doorgeschoven. Na invoer van b_0, b_1 en b_2 zijn deze ook uitgevoerd (bovenlangs). Bij de invoer van b_3 wordt $b_3 + b_0$ uitgevoerd en daarna staat in het register b_3, b_2, b_1, b_0. Bij de invoer van b_4 is de uitvoer $b_4 + b_1 + b_0$, bij de invoer van b_{10} is de uitvoer $b_{10} + b_7 + b_6$ en het register bevat b_{10}, b_9, b_8, b_7.

Nu worden vier 0'en ingevoerd; de uitvoer is $b_7 + b_8, b_8 + b_9, b_9 + b_{10}, b_{10}$, de coëfficiënten van $x^{11}, x^{12}, x^{13}, x^{14}$. Het register staat nu op 0 en is klaar voor de invoer van de volgende boodschap.

Voor de liefhebbers van de ouderwetse staartdeling is ook de schakeling in Figuur 7 interessant. De boodschap $b_0, b_1, \ldots, b_{10}$ wordt nu opgevat als veelterm $B(x) = b_{10} + b_9 x + \cdots + b_1 x^{13} + b_0 x^{14}$. De invoer van $b_0, b_1, \ldots, b_{10}$ wordt ongewijzigd doorgegeven (in de figuur onderlangs). Intussen vindt in het register een bewerking plaats. Na invoer van b_3 staat daar b_3, b_2, b_1, b_0. Zie dat als het hoogste deel van B(x): $b_3 x^{11} + b_2 x^{12} + b_1 x^{13} + b_0 x^{14}$. Nu wordt b_4 in-(en door-)gevoerd en het register bevat daarna $b_4 + b_0, b_3, b_2, b_1 + b_0$. Zie dit als $(b_4 - b_0)x^{10} + b_3 x^{11} + b_2 x^{12} + (b_1 - b_0)x^{13}$. Dat is het hoogste deel van $B(x)$ nadat daar $b_0 x^{10} g(x) = b_0 x^{14} + b_0 x^{13} + b_0 x^{10}$ van is afgetrokken, zoals bij een staartdeling. De graad is nu ≤ 13. Dat gaat zo door. Na invoer van b_{10} houden we in het register de coëfficiënten over van een veelterm van graad 3 die de rest is van $B(x)$ bij deling door $x^4 + x^3 + 1$. Nu gaat schakelaar A open, B klapt naar boven en er worden vier 0'en ingevoerd waardoor de rest ongewijzigd wordt uitgevoerd.

De codeveelterm bij $B(x)$ is dus $B(x) + r(x)$, $r(x)$ de rest van $B(x)$ na deling door $x^4 + x^3 + 1$. Deze codering is systematisch, de boodschap staat zelf in het codewoord. Dezelfde schakeling kan gebruikt worden om fouten te ontdekken, bij een codewoord $B(x) + r(x)$ moet de bepaling van de rest immers

$r(x) + r(x) = 0$ opleveren.

In het algemeen is het niet moeilijk om te constateren dat een ontvangen woord geen codewoord is, het syndroom is niet 0. Het vinden van de fouten is veel lastiger, al zijn daar voor BCH-codes goede algoritmen voor. Ze hebben te maken met het zoeken van nulpunten van veeltermen; begrijpelijk dat dat niet eenvoudig is, en hier niet verder zal worden onthuld.

Literatuur

Klassiekers

1. Elwyn R. Berlekamp. Algebraic Coding Theory (McGraw-Hill, 1968).
2. J.H. van Lint. Coding Theory (Springer, 1971).
3. F.J. MacWilliams, N.J.A. Sloane. The Theory of Error-Correcting Codes (North Holland, 1977).

Lichtere kost

J.H. van Lint (red.). Inleiding in de Coderingstheorie (MC Syllabus 31, 1976).

Ook op het internet is veel te vinden (probeer bv. AMS, NASA, History site van St. Andrews).

Misschien kan ik op aanvraag (jeuris@math.ru.nl) een collegedictaat via E-mail toesturen (format .ps of .pdf).

CWI Vakantiecursus 2005

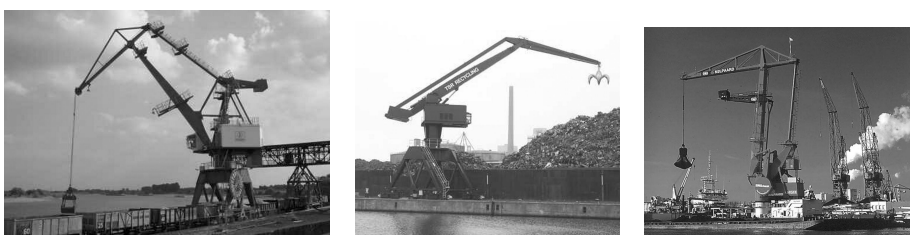
pp. 71 – 92

Kranen en lemniscaten

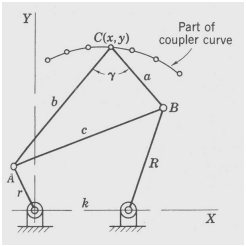
R.H. Kaenders
Radboud Universiteit Nijmegen
e-mail: R.H.Kaenders@ils.ru.nl

Floating lemniscate cranes (*intrekdraaikranen met lemniscaatsturing*), de zogenaamde *double boom cranes* of *balanskranen* zijn moderne hijskranen die allemaal berusten op een eenvoudig principe: het zogenaamde *Doppellenkerwippprinzip met lemniscaatsturing* dat in de jaren dertig door de *ARDELT Werke* in het Duitse Eberswalde (tegenwoordig *KE Kranbau Eberswalde*) werd ingevoerd bij de bouw van hijskranen (zie [17] en [18]). Het is de bedoeling van een dergelijke constructie om hijskranen met een enigszins horizontale lastweg te bouwen, d.w.z. waarbij het uiteinde van de kabel nagenoeg op één hoogte blijft ('level luffing', [13], p.1367). Hierdoor hoeft geen extra kracht te worden ingezet om de last in opwaartse richting te verplaatsen. Er zijn verschillende constructies bekend waarmee dit effect van (bijna) horizontale lastverplaatsing bereikt kan worden en we zullen er een paar van behandelen. Bijvoorbeeld voor een intrekdraaikraan met ellipssturing is het elementair meetkundig te bewijzen dat hij de last langs een horizontale lijn vervoert. Het basismechanisme bij veel van deze hijskranen treft men ook aan in talloze andere constructies: een scharnierende stangenvierhoek in het platte vlak met een vaste basisstang. We zullen zien dat zulke stangenvierhoeken in twee grote klassen kunnen worden ingedeeld die door de zogenaamde *doorslaande stangenvierhoeken* van elkaar zijn gescheiden. Verder zullen we kijken naar de diagonalen van zulke stangenvierhoeken.

Vervolgens beschouwen wij de tegenover de vaste stang liggende stang, *koppelstang* geheten, als de basis AB van een driehoek $\triangle ABC$ waarvan het toppunt C bij beweging van de driehoek een kromme beschrijft: een zogenaamde



Figuur 1. *Lemniscate-, balance- en floating level-luffing crane*

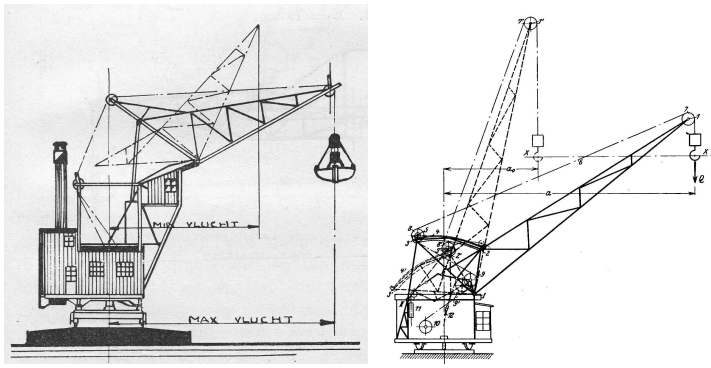


Figuur 2. *Constructie van een koppelkromme (tekening uit [4])*

koppelkromme (zie Figuur 2). Een dergelijke koppelkromme hangt af van vijf lengtes en daardoor kan een grote variatie van krommen ontstaan. Hieronder zijn beroemde voorbeelden zoals de Watt-krommen en de klassieke lemniscaat van Bernoulli. Van de laatste laten we twee constructies als koppelkromme zien. De bewijzen zijn enkel gebaseerd op gelijkvormigheid. Uitdrukkelijk sluiten wij niet de koppelkrommen met ontaarde driehoeken $\triangle ABC$ uit, waarbij het toppunt van de driehoek op de koppelstang of in zijn verlengde ligt. Daarmee beschrijft de haak waar de last aanhangt bij intrekdraaikranen altijd een koppelkromme – soms met dubbelpunt en soms zonder. Door te kijken naar de zogenaamde *pivotcirkel* zullen wij een handig criterium voor het bestaan en het vinden van dubbelpunten van koppelkrommen formuleren.

Ten slotte zullen wij ingaan op een speciale vraag over stangenvierhoeken. *Welke stangenvierhoeken met een extra vijfde stang zijn nog beweegbaar?* In het speciale geval van een parallellogram wordt deze vraag in de cursus *Gelijkvormigheid* (tweede klas) in de Ratio-internetmethode [19] behandeld. Wij geven hier een algemeen antwoord met een zelf gevonden bewijs dat gebruikmaakt van de zogenaamde *afbeelding van Darboux*. De wiskundige Gaston Darboux (1842–1917) heeft in het jaar 1879 een verrassend eenvoudige en ingenieuze manier bedacht om aan een stangenvierhoek een derdegraads kromme in het complexe projectieve vlak te verbinden (de afbeelding van Darboux). De eigenschappen van deze kromme weerspiegelen de eigenschappen van de stangenvierhoek. De afbeelding van Darboux geeft inzicht in mogelijke bewegingspatronen van stangenvierhoeken en daarmee ook van hijskranen. We zullen uitleggen hoe dat gaat en deze techniek toepassen en zo een definitief antwoord geven op de bovengestelde vraag.

Dit onderwerp is rijk aan de prachtige meetkunde waarvan wij hier alleen een tipje van de sluier kunnen oplichten. Wij laten bijvoorbeeld alle differentiaalmeetkundige en theoretisch-mechanische aspecten ervan achterwege. Dynamische meetkundeprogramma's zoals CABRI geven schitterende mogelijkheden voor de visualisering van stangen- en kraanconstructies. Dit zal wél in de voordracht te zien zijn maar in de syllabustekst moeten we het doen met starre afbeeldingen. Nochtans is dit een uitnodiging om, wellicht met leerlingen, de



***Figuur 3**. Wipkraan (uit [11]) en intrekdraaikraan met krommenbaan (uit [3])*

constructies uit de tekst met zulke programma's na te bouwen[1] en verder over het onderwerp na te denken.

Collega-docent en levenslang Meccano-bouwer Alfons Gijselhart bedank ik omdat hij mij attent maakte op artikelen als [12] en de ontdekking dat er prachtige wiskunde bij hijskranen te ontdekken is. Verder dank aan Leon van de Broek, Jozef Steenbrink, Nellie Verhoef en Edith Verbeet voor het kritisch doorlezen van deze bijdrage.

Dit verhaal over vernuftige ingenieurskunst heb ik geschreven in liefdevolle herinnering aan mijn vader, Dipl. Ing. Josef Kaenders.

1. Hijskranen met horizontale lastweg

Reeds in 1911 heeft het Haarlemse bedrijf FIGEE ([17]) (zie ook [11]) voor een 'pakhuis in Rotterdam' wipkranen gebouwd. Over het algemeen zijn *wipkranen* hijskranen waar de kraanboom of kraanarm rond een vast punt scharniert. Zij vormen een grote klasse van hijskranen waarvan wij enkele speciale constructies zullen bekijken. De hijskraan links in Figuur 3 geeft een voorbeeld van een wipkraan die recht doet aan zijn naam.

In [11] vat ir. De Vries een aantal criteria samen waarmee bij het ontwerp van wipkranen rekening moet worden gehouden.

> "Ter vermindering van de energie voor het wippen, wordt er naar gestreefd: 1^e. de lastweg horizontaal te maken; 2^e de massakrachten binnen matige grenzen te houden; 3^e de wrijvingsverliezen tot een minimum te beperken en 4^e het aandrijfmechanisme zoodanig uit te voeren, dat de massa's langzaam versneld worden, terwijl de snelheid tegen het einde weer afneemt. Het is vanzelfsprekend, dat de bewegende ijzerconstructie gebalanceerd wordt en dat de resultante der krachten, die op den laadboom werken, door het draaipunt gaat of er dichtbij valt."

[1] Bijvoorbeeld in het kader van profielwerkstukken, zoals [15], [16] en [14].

In deze syllabustekst beperken we ons alleen tot 1. Dit neemt niet weg dat ook de andere aspecten van belang zijn en eveneens aanleiding geven tot interessante wiskundige beschouwingen over de statica en de kinematica van kraanconstructies. Als wij kijken naar horizontale lastwegen, dan doen we dat hier voornamelijk wiskundig. Een ingenieur kijkt er veel praktischer tegenaan en moet bij de bouw van een kraan nog heel wat andere aspecten in de gaten houden. "Veelal wordt geeischt een zoveel mogelijk horizontale lastweg, wat echter alleen te prefereeren is, waneer dit zonder complicaties te bereiken is. In den regel is de weg niet mathematisch zuiver recht, doch een goede praktische benadering. Meer hoeft men niet te eischen, vooral als dit met een eenvoudige constructie bereikt wordt." schrijft ir. De Vries in [11] (blz. 418).

De boeken van [11] en [3] geven een prachtig overzicht van kraanconstructies die in de jaren twintig in Nederland en Duitsland zijn ontwikkeld. Het lijkt erop dat er in die tijd een bloei in de constructie van nieuwe hijskranen heeft plaatsgevonden. Speciaal voor kraanconstructies met horizontale lastweg zijn meer voorbeelden te vinden in [8], [9] of [10]. De constructies van deze hijskranen werden onder meer uitgevoerd door bedrijven als FIGEE en Stork-Hijsch te Haarlem, M.A.N. te Augsburg, A.E.G. te Berlijn en Amsterdam, ARDELT te Eberswalde of DEMAG in Duisburg. In het decennium ervoor werden er in Engeland bij Babcock & Wilcox en bij Toplis nieuwe hijskranen ontwikkeld.

1.1. Toren- en portaal- en andere kranen

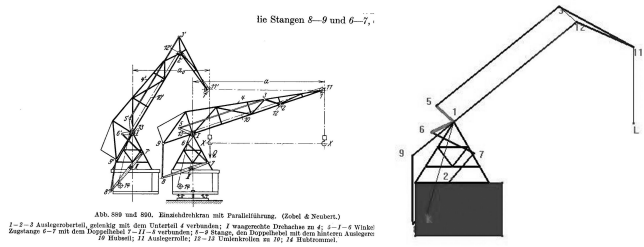
Er zijn veel verschillende strategieën om een horizontale lastweg te bereiken. Bij portaal- en torenkranen is de baan van de 'loopkat', de rol die langs de kraanarm schuift, door de constructie al horizontaal. De horizontale verplaatsing van de last wordt dan bereikt door een kleine maar vernuftige truc. De kabel loopt niet alleen vanaf een vast beginpunt op de kraanarm over de loopkat naar de last maar ook via een rol terug van de last langs de loopkat naar het andere einde van de kraanarm. Op die manier hangt de last aan een lus die altijd even lang blijft zolang men niet de kabel verlengt of inkort. Zelfs met schuine kraanbomen zijn vergelijkbare constructies mogelijk die de niet horizontale maar wel lineaire baan van de last compenseren en horizontaal maken.

Een beetje kort door de bocht maar wel gerealiseerd zijn ook constructies waar de last pas een horizontale baan beschrijft als de kabelrol langs een kromme beweegt die juist de kromme van de last vereffent zoals rechts in Figuur 3 te zien.

1.2. Parallel- en ellipskranen

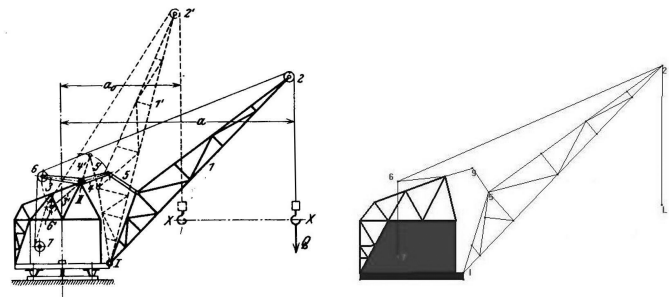
Naast deze meer simpele hijskranen zijn er veel kraanconstructies die berusten op stangenconstructies waaarbij de stangeneinden óf scharnieren óf langs andere stangen schuiven. Wij geven hier drie voorbeelden. In [8] treft men meer dan vijftig verschillende van dergelijke constructies aan. De constructie van deze hijskranen in CABRI maakt het thans mogelijk hen te visualiseren en te onderzoeken ten opzichte van een mogelijkerwijs horizontale lastweg.

In Figuur 5 zien wij een wipkraan waarbij de kabel met een schommelhendel



Figuur 4. *Intrekdraaikraan met parallelsturing; tekening uit [3] en CABRI-constructie uit [15]*

meer of minder ver wordt omgeleid met het effect dat de lastweg enigszins recht wordt. Het uiteinde van de schommelhendel beweegt langs een koppelkromme.



Figuur 5. *Intrekdraaikraan met schommelhendel; uit [3], CABRI-constructie uit [15]*

De *intrekdraaikraan met ellipssturing* is een kraan waarbij het uiteinde van de kraanarm een ellips beschrijft. Hier is de lastweg recht – zelfs wiskundig gezien. Het bewijs daarvoor is elementair en wij geven het hieronder als opgave. Op de webpagina [20] is er een applet van de ellipskraan te vinden.

Opgave

Met de notatie uit Figuur 6 is de afstand tussen 1 en 3 vier keer zo groot als de afstand tussen I en 2 die op zijn beurt weer gelijk is aan de afstand tussen 2 en 3. Punten I en 2 zijn scharnierpunten terwijl punt 3 op en neer schuift.

(i) Laat zien dat het middelpunt van de kraanarm langs de horizontale lijn door I loopt.

(ii) Er zijn twee verschillende kabels: de hijskabel (10, langs 13-12-5[11]-1) waarmee de last wordt opgehesen en de intrekkabel (4, langs 9-7-8-3) waar-

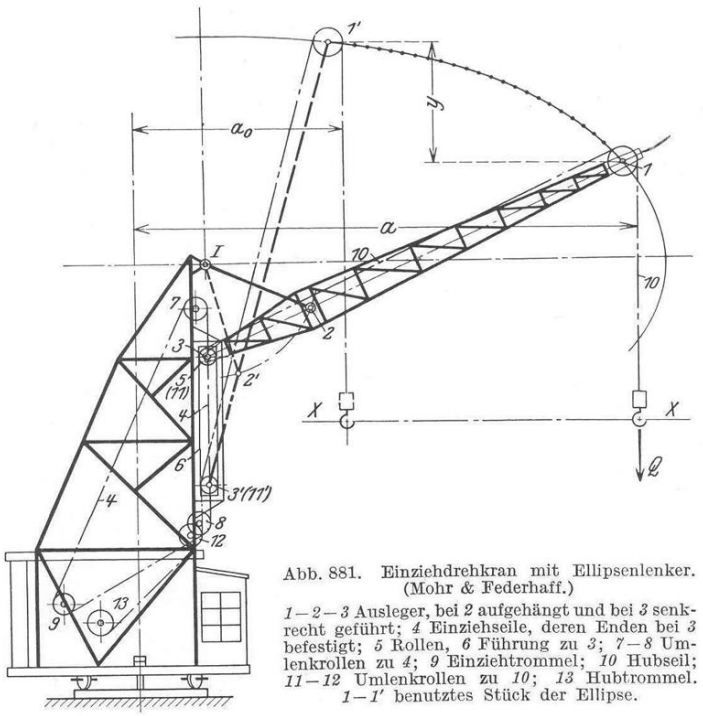


Abb. 881. Einziehdrehkran mit Ellipsenlenker. (Mohr & Federhaff.)

1–2–3 Ausleger, bei *2* aufgehängt und bei *3* senkrecht geführt; *4* Einziehseile, deren Enden bei *3* befestigt; *5* Rollen, *6* Führung zu *3*; *7–8* Umlenkrollen zu *4*; *9* Einziehtrommel; *10* Hubseil; *11–12* Umlenkrollen zu *10*; *13* Hubtrommel. *1–1'* benutztes Stück der Ellipse.

Figuur 6. *Intrekdraaikraan met ellipssturing, uit [3]*

mee de kraanarm kan worden ingetrokken. Laat zien dat de last precies op één hoogte beweegt.

(iii) Toon aan dat het uiteinde 1 van de kraanarm een ellips beschrijft.

1.3. Hijskranen met lemniscaatsturing

De hijskranen met lemniscaatsturing berusten allemaal op het *Doppellenker-wippprinzip*: een scharnierende stangenvierhoek waar de koppelstang, dat is de stang tegenover de vaste stang, tot kraanarm is verlengd. Dit mechanisme werd reeds in de jaren dertig door de *ARDELT Werke* in Eberswalde (tegenwoordig *KE Kranbau Eberswalde*) ingezet (zie [18]).

In het volgende wiskundige gedeelte proberen wij de bewegingspatronen van dergelijke hijskranen met lemniscaatsturing te begrijpen. In 3.3 zullen wij zien dat ook de klassieke lemniscaat van Bernoulli als lastkromme ontstaat. Allereerst zullen wij hiervoor de bewegingspatronen van stangenvierhoeken onderzoeken.

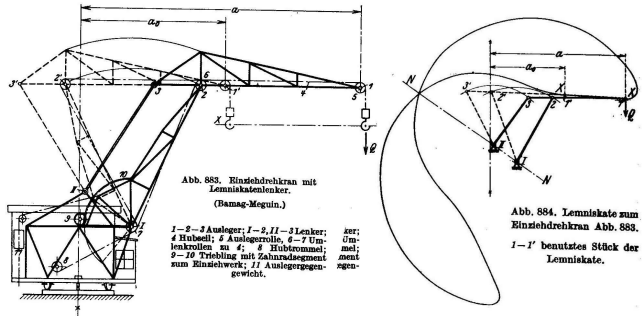


Abb. 883. Einziehdrehkran mit Lemniskatenlenker. (Bamag-Meguin.)

1—2—3 Ausleger; *I—2, II—3* Lenker; *4* Hubseil; *5* Auslegerrolle, *6—7* Umlenkrollen zu *4*; *8* Hubtrommel; *9—10* Triebling mit Zahnradsegment zum Einziehwerk; *11* Auslegergegengewicht.

Abb. 884. Lemniskate zum Einziehdrehkran Abb. 883.

1—1' benutztes Stück der Lemniskate.

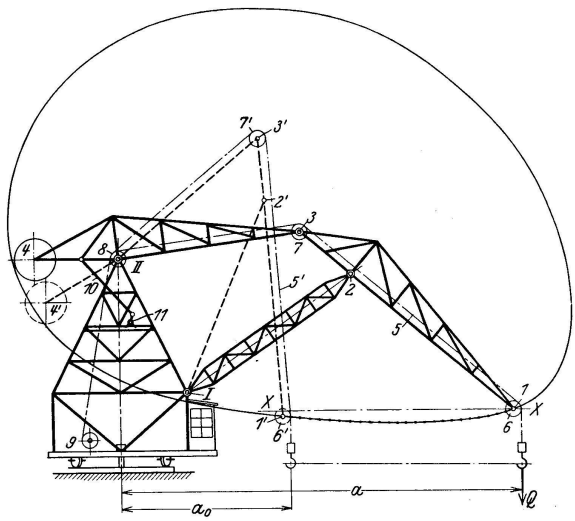


Abb. 882. Einziehdrehkran mit Lemniskatenlenker. (Demag.)

1—2—3 Ausleger; *I—2, II—3* Lenker; *4* Auslegergegengewicht; *5* Hubseil; *6* Auslegerrolle; *7—8* Umlenkrollen zu *5*; *9* Hubtrommel; *10* Zugstangen, den angetriebenen Lenker mit den Kurbeln *11* des Einziehwerks verbindend; *1—1'* benutzes Stück der Lemniskate.

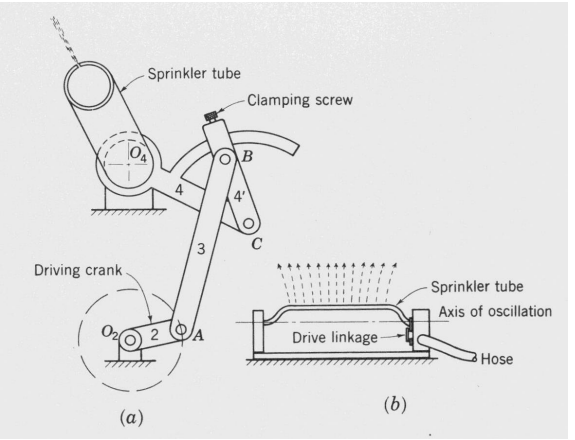
Figuur 7. *Intrekdraaikranen met lemniscaatsturing, uit [3]*

2. Stangenvierhoeken

Drie wederzijds aan elkaar verbonden stangen bewegen niet. De eerste scharnierende stangenveelhoek die voor beweeglijke mechanische constructies in aanmerking komt, wordt gevormd door vier stangen. Het is daarom ook niet toevallig dat stangenvierhoeken in alle mogelijke mechanismen terug te vinden zijn, zoals bij een gazonsproeier (zie Figuur 8), een busdeur, een vorkheftruck en vele andere apparaten en toestellen. Hoewel stangenvierhoeken voor de hand liggende eenvoudige objecten vormen is hun kinematica spannend en niet-triviaal.

Wij gaan uit van een intuitïtief begrip van alle stangenconstructies als wis-

kundige objecten. Deze intuïtieve voorstelling leidt niet tot misverstanden en daarom proberen wij hier ook niet haar in een formele wiskundige definitie te gieten.



Figuur 8. *Stangenvierhoek in een gazonsproeier, uit [4]*

2.1. Eigenschappen van stangenvierhoeken

Voor een dieper begrip van de bewegingspatronen van koppelkrommen is de eerste stap een beter begrip van de bewegingspatronen van stangenvierhoeken. Zulke stangenvierhoeken zijn op een natuurlijke manier in twee klassen op te delen. Wij volgen hier [4].

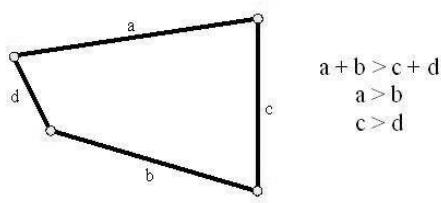
klasse I: Een stangenvierhoek behoort tot deze klasse als de kortste stang een volle draaiing ten opzichte van alle andere stangen kan maken. De andere drie stangen kunnen alleen maar heen en weer schommelen ten opzichte van elkaar.

klasse II: Geen stangenvierhoek uit deze klasse kan een volle draaiing maken ten opzichte van een van de andere stangen.

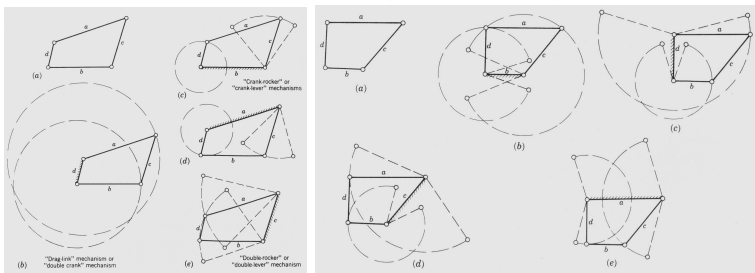
Altijd kunnen we de notatie van de lengtes van een stangenvierhoek in een van de twee klassen zo kiezen dat voor tegenover elkaar liggende zijden a en b geldt (zie ook 9):

$$a + b > c + d \quad \text{en} \quad a > b \quad \text{en} \quad c > d.$$

In de volgende beschouwingen gebruiken wij dezelfde letters als namen voor de stangen alsook voor hun lengte.



Figuur 9. *Notatie voor de classificatie van stangenvierhoeken*



Figuur 10. *Mechanismes van stangenvierhoeken: klasse I (links) en klasse II (rechts)*

Opgave

(i) Laat zien dat de stangenvierhoeken in de verschillende klassen door de volgende twee extra eisen worden gekenmerkt.

klasse I: $a - b < c - d$,

klasse II: $a - b > c - d$.

(ii) Laat zien dat bij alle stangenvierhoeken die niet passen in de klassen I en II de som van twee van de zijden gelijk is aan de som van de twee overige zijden of dat één zijde even lang is als de andere drie bij elkaar.

(iii) Ga na dat er bij de stangenvierhoeken die niet in klasse I of II zitten minstens twee stangen bestaan die een volle draaiing ten opzichte van een van de andere stangen kunnen maken.

Definitie 1 Een stangenvierhoek die niet behoort tot één van de klassen I of II heet een *doorslaande stangenvierhoek*.

Bij een stangenvierhoek uit klasse I of II kunnen vier verschillende mechanismes ontstaan, afhankelijk van welke stang wordt vastgehouden (zie Figuur 10).

Wij besluiten deze paragraaf met de volgende eenvoudige opmerking over de diagonalen van stangenvierhoeken.

STELLING 14 *Het inproduct van de diagonaalvectoren bij een stangenvierhoek is constant.*

Opmerking. Het inproduct van twee vectoren v en w is gelijk aan

$$< v,\, w >= |v||w| \cos \angle (v, w)\,.$$

In het bijzonder betekent dit: als bij één positie van de stangenvierhoek de diagonalen loodrecht op elkaar staan, dan is dat voor elke positie het geval.

Als het inproduct positief (negatief) is heeft dit tot gevolg dat tegenoverliggende scharnierpunten niet bij elkaar kunnen komen en de diagonalen in elke positie een scherpe (stompe) hoek met elkaar blijven vormen.

Bewijs: Wij stellen ons de vier stangen in de vierhoek voor als vier vectoren v_1, v_2, v_3, en v_4 waarlangs men de vierhoek in deze volgorde kan doorlopen. Er geldt

$$v_1 + v_2 + v_3 + v_4 = 0. \tag{1}$$

Wij willen bewijzen dat

$$< v_1 + v_2,\, v_3 + v_2 >=< v_1,\, v_3 > + < v_1,\, v_2 > + < v_2,\, v_3 > + |v_2|^2$$

constant is en doen dat door op te merken dat uit (1) volgt:

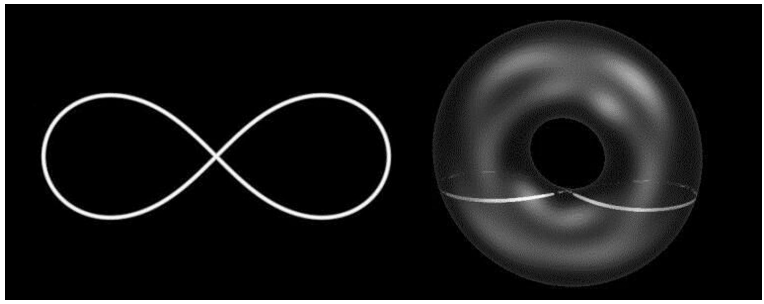
$$|v_4|^2 = |v_1|^2 + |v_2|^2 + |v_3|^2 + 2\,(< v_1,\, v_2 > + < v_1,\, v_3 > + < v_2,\, v_3 >)\,.$$

□

Figuur 11. *Heidense afbeelding op een doopsteen uit het begin van de* 13^e *eeuw (uit het dorpje Niel, Niederrhein)*

3. LEMNISCATEN

De lemniscaat is het symbool voor oneindig. Afbeeldingen van draken in een achtvorm die zichzelf verslinden, staan van oudsher voor het eeuwig terugkerende en zetten zich daarmee af tegen het christelijke A et Ω, waar de wereld een begin en een einde kent.



***Figuur 12**. De lemniscaat treedt ook op als doorsnijding met de torus waarbij de gr ote straal twee keer zo groot is als de kleine*

De lemniscaatvorm als symbool oefende altijd al een grote fascinatie uit op mensen. Naast de bovengenoemde associatie bestaan er nog veel meer. Wij beperken ons hier tot de wiskunde van de lemniscaat.

3.1. De lemniscaat van Jacob Bernoulli
Gegeven twee punten in het affine vlak F_1 en F_2 met een vaste afstand van $2a$ ertussen. De *lemniscaat* is de meetkundige plaats van alle punten P in het vlak waarvoor geldt

$$|PF_1| \cdot |PF_2| = a^2.$$

Hoewel deze kromme reeds door John Wallis (1616–1703) in zijn boek *Arithmetica Infinitorium* werd ingevoerd, wordt zij meestal verbonden met de naam Jacob Bernoulli (1654–1705), die er in het jaar 1694 een vergelijking voor heeft afgeleid.

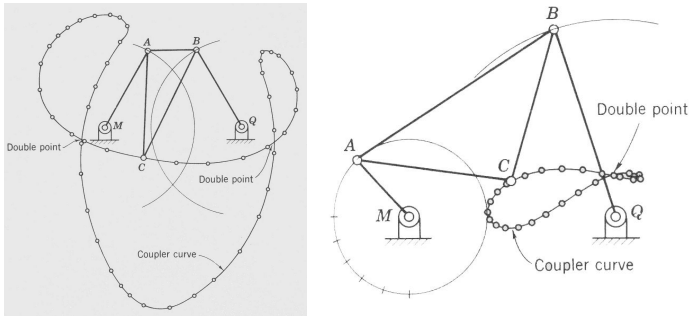
Opgave

(i) Wij leggen een poolcoördinatensysteem in het vlak zodat de oorsprong midden tussen F_1 en F_2 ligt en de straal door F_2 loopt. Laat zien dat de lemniscaat in poolcoördinaten (r, φ) wordt beschreven door de vergelijking:

$$r^4 = r^2 2a^2 \cos(2\varphi).$$

(ii) Nu leggen wij een cartesisch coördinatenstelsel in het vlak zodanig dat de brandpunten de coördinaten hebben: $F_1(-a, 0)$ en $F_2(a, 0)$. Laat zien dat de lemniscaat in (x, y) wordt beschreven door:

$$\left(x^2 + y^2\right)^2 = 2a^2(x^2 - y^2).$$



Figuur 13. *Voorbeelden van Koppelkrommen (uit [4])*

(iii) De coördinaten $\mathbb{R}^2$ zijn ook op te vatten als complex getallenvlak $\mathbb{C}$. Laat zien dat hier de lemniscaat wordt gegeven door de $z \in \mathbb{C}$ met:

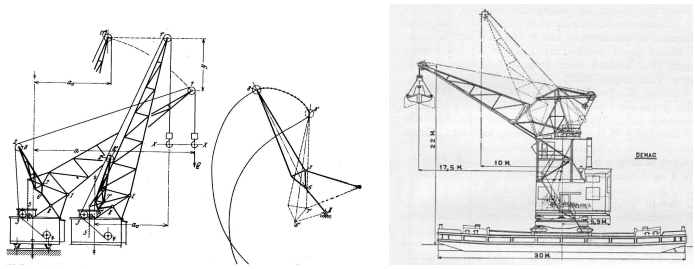
$$z^2\bar{z}^2 = a^2(z^2 + \bar{z}^2).$$

De *orthogonale hyperbool* is de meetkundige plaats van alle punten P waarvoor geldt: $\|PF_1| - \|PF_2\| = \sqrt{2}a$. De lemniscaat en de orthogonale hyperbool zijn nauw met elkaar verbonden. Bijvoorbeeld geldt:

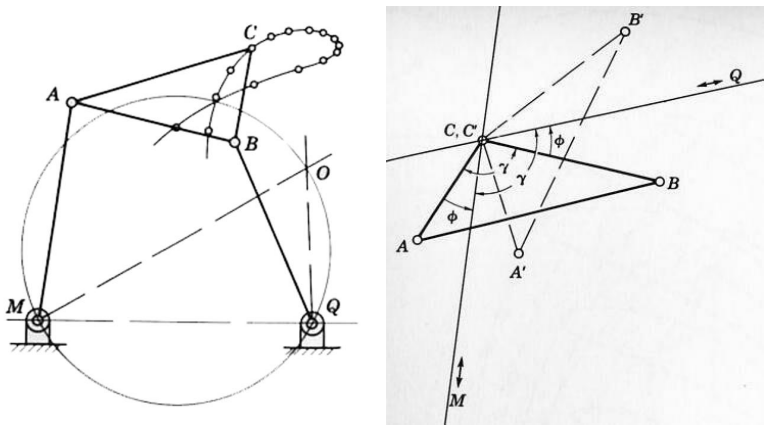
- Gegeven een willekeurige cirkel rond de oorsprong. De lemniscaat gespiegeld aan deze cirkel is een orthogonale hyperbool en andersom.
- De lemniscaat is de *voetpuntskromme* van de orthogonale hyperbool, d.w.z. gegeven een raaklijn aan de orthogonale hyperbool, dan is het voetpunt van de loodlijn vanuit de oorsprong (de projectie) op die raaklijn een punt van een lemniscaat waarvan precies alle punten op die manier ontstaan.
- De lemniscaat is de omhullende kromme die ontstaat als men alle cirkels door de oorsprong tekent met middelpunt op de orthogonale hyperbool.

Opgave
Bewijs de boven gemaakte beweringen – desnoods analytisch.

3.2. Koppelkrommen
Elk punt in het vlak met een vaste positie ten opzichte van de koppelstang van een stangenvierhoek waarbij de basisstang vast is, beschrijft een koppelkromme. Er zijn legio voorbeelden van mechanische constructies waar koppelkrommen een belangrijke rol spelen. Voorbeelden van koppelkrommen zijn de voor de constructie van stoommachines door James Watt verzonnen *Watt-krommen* en constructies voor het tekenen van redelijk rechte lijnen zoals wij er een paar hebben laten zien. In het boek van Kempe [7] treft men er nog meer fascinerende voorbeelden van aan.



Figuur 14. *Intrekdraaikranen; de kabel gaat langs een rol die een koppelkromme beschrijft*



Figuur 15. *Definitie van een pivotcirkel en bewijsschets bij Stelling 15, uit [4]*

De verscheidenheid van deze krommen is enorm. In de jaren vijftig vervaardigden Hrones en Nelson [6] een indrukwekkende atlas met ongeveer 10.000 koppelkrommen: een waar monnikenwerk. Dit betreft alleen koppelkrommen die ontstaan als meetkundige plaats van een punt op de koppelstang zelf.

Uit de rijke theorie van koppelkrommen geven wij hier een verrassend eenvoudig resultaat dat ons inzicht geeft over welke koppelkromme een dubbelpunt kan hebben, waar dit ligt, en welke koppelkromme geen dubbelpunt heeft. Ook hier volgen wij [4] maar bewijzen een iets sterkere stelling.

DEFINITIE 2 Gegeven een koppelmechanisme met notatie zoals is afgebeeld in Figuur 13, waarbij A, B en C niet op een lijn liggen. De *pivotcirkel* van dit koppelmechanisme is dan de cirkel door de punten M, Q en O waarbij O zo is gekozen dat $\triangle MQO$ gelijkvormig is met $\triangle ABC$. Als A, B en C wel op een lijn liggen noemen we deze lijn de *pivotlijn*.[2]

[2] Het Franse woord 'pivot' betekent zoiets als draaipunt. In een koppelmechanisme zoals in Figuur 15 zijn er de punten M en Q mee bedoeld.

Met behulp van deze definitie kunnen wij de volgende verrassend eenvoudige stelling formuleren en bewijzen. Zie Figuur 15.

STELLING 15 *Gegeven een koppelmechanisme met pivotcirkel. De dubbelpunten van de bijbehorende koppelkromme zijn precies de punten waar de kromme de pivotcirkel snijdt.*

Bewijs: Stel dat er een dubbelpunt op de pivotcirkel ligt. Dan zijn er twee posities van de driehoek aan de koppelstang: $\triangle ABC$ en $\triangle A'B'C'$ waarbij $C = C'$. Het vaste punt M moet nu ergens op de deellijn liggen van $\angle ACA'$. Evenzo moet Q op de deellijn liggen van $\angle BCB'$. Hieruit volgt dat $\angle MCQ$ óf gelijk is aan $\angle ACB$ óf aan het supplement $180° - \angle ACB$. Als wij nu bij gegeven M en Q kijken naar de meetkundige plaats van alle punten S waarvoor $\angle MCQ$ óf gelijk is aan $\angle ACB$ óf aan het supplement van $\angle ACB$, dan vinden wij de pivotcirkel. Deze constructie kan moeiteloos worden omgedraaid om te bewijzen dat bij een punt van de koppelkromme op de pivotcirkel twee verschillende driehoeken en daarmee een dubbelpunt behoort. □

Een ander spectaculair inzicht over koppelkrommen is de zogenaamde *stangenconstructie van Robert* waarmee men elke koppelkromme op drie verschillende manieren kan construeren. Voor dit en andere inzichten over koppelkrommen zie [4] en [1].

Opgave

Nu gaan we ervan uit dat A, B en C wel op een lijn liggen, de pivotlijn. Toon aan:

(i) Als de koppelkromme een dubbelpunt heeft dan ligt dat op de pivotlijn.

(ii) Als de stangenvierhoek niet doorslaand is, geldt ook de omkering: als de koppelkromme de pivotlijn snijdt dan heeft zij in dat snijpunt een dubbelpunt.

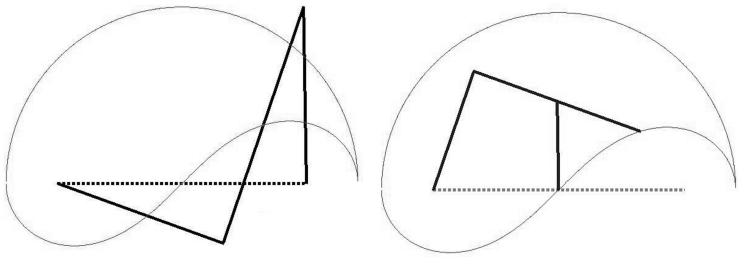
(iii) Onderzoek de koppelkrommen bij de afgebeelde hijskranen met lemniscaatsturing op hun dubbelpunten.

(iv) Onderzoek wanneer een koppelkromme met doorslaande stangenvierhoek die de pivotlijn snijdt een dubbelpunt in dat snijpunt heeft.

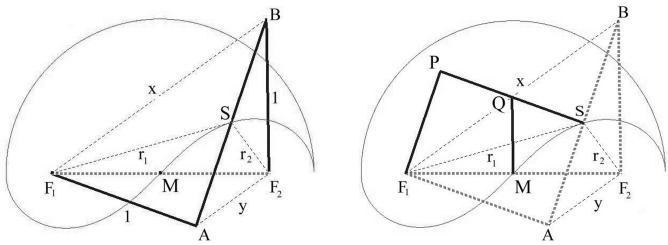
3.3. De lemniscaat als koppelkromme

De klassieke lemniscaat ontstaat op twee verschillende manieren als koppelkromme. In Figuur 16 schetsen wij deze constructies. De hoekpunten tussen de stangen zijn scharnierpunten.

Hierbij zijn de langere stangen $\sqrt{2}$ maal zo lang als de korte stangen waarvan wij de lengte a noemen. De horizontale gestippelde lijnen zijn bij beide constructies vast. Links wordt de lemniscaat beschreven als meetkundige plaats van het middelpunt van de andere lange zijde. In de rechter constructie is de bovenste stang twee keer zo lang (lengte $2a$) als de kortste stang (lengte a) en de lemniscaat is de meetkundige plaats van het losse eindpunt van deze stang.



Figuur 16. *Twee constructies van de lemniscaat als koppelkromme*



Figuur 17. *Notatie bij de constructies van de lemniscaat als koppelkromme*

STELLING 16 *Bij beide boven beschreven constructies is de koppelkromme gelijk aan de klassieke lemniscaat van Bernoulli.*

Wij bewijzen dat de constructie links de lemniscaat als koppelkromme levert. De constructie rechts laten we over aan de lezer.

Bewijs: Als wij de notatie uit het linker Figuur in 17 overnemen, dan is te bewijzen dat geldt: $r_1 r_2 = \frac{1}{2}$. Voor het gemak zetten wij hier $a = 1$.

Eerst merken wij op dat geldt: $\triangle F_1AS \cong \triangle BAF_1$ want de verhoudingen $\frac{|F_1A|}{|AB|}$ en $\frac{|AS|}{|F_1A|}$ zijn gelijk. Om een zelfde reden is driehoek $\triangle ABF_2$ gelijkvormig met driehoek $\triangle SBF_2$ en daarom:

$$\frac{x}{\sqrt{2}} = \frac{r_1}{1} \quad \text{en} \quad \frac{y}{\sqrt{2}} = \frac{r_2}{1}.$$

Bovendien uit $\triangle F_1AS \cong \triangle BAF_1$ volgt ook dat $\angle ABF_1 = \angle BF_1F_2$ gelijk is aan $\angle AF_1S$. Als wij bij beide hoeken, $\angle BF_1F_2$ en $\angle AF_1S$, de hoek $\angle SF_1F_2$ weghalen vinden wij $\angle F_2F_1A = \angle BF_1S$. Hieruit volgt $\triangle AF_2F_1 \cong \triangle SBF_1$ en daarmee $\frac{x}{\left(\frac{1}{\sqrt{2}}\right)} = \frac{\sqrt{2}}{y}$. Samen:

$$r_1 r_2 = \frac{1}{2}xy = \frac{1}{2}.$$

□

Opgave

(i) Bewijs meetkundig met behulp van de rechter tekening in Figuur 17 dat de koppelkromme van deze constructie de klassieke lemniscaat is. Maak hierbij gebruik van de linker constructie.

(ii) Geef met behulp van de vergelijking van de lemniscaat in poolcoördinaten op blz. 81 een meer analytisch bewijs voor deze constructie van de lemniscaat als koppelkromme.

4. De afbeelding van Darboux

De Franse wiskundige Gaston Darboux [2] publiceerde in het jaar 1879 een methode om aan een stangenvierhoek een projectieve derdegraads kromme over de complexe getallen te verbinden (zie ook [1]). De eigenschappen van deze kromme weerspiegelen de eigenschappen van de bijbehorende stangenvierhoek. Wij lichten hieronder zijn ingenieuze en verrassend eenvoudige constructie toe. Darboux gebruikte de toen nog jonge theorie over elliptische functies van Jacobi om er een meetkundige observatie over vierhoeken uit af te leiden.

Voor een complex getal $z \in \mathbb{C}$ met $|z| = 1$ geldt: $\bar{z} = \frac{1}{z}$. Wij representeren de stangen van een stangenvierhoek door vier vectoren in $\mathbb{R}^2$ met lengtes a, b, c, d die wij opvatten als vier complexe getallen az_1, bz_2, cz_3, $dz_4 \in \mathbb{C}$ waarbij $|z_1| = |z_2| = |z_3| = |z_4| = 1$ en $a, b, c, d \in \mathbb{R}^{>0}$. Het feit dat een stangenvierhoek gesloten is vertaalt zich naar de eis $az_1 + bz_2 + cz_3 + dz_4 = 0$. Samen met de geconjugeerde vergelijking vinden we:

$$az_1 + bz_2 + cz_3 + dz_4 = 0 \quad \text{en} \quad \frac{a}{z_1} + \frac{b}{z_2} + \frac{c}{z_3} + \frac{d}{z_4} = 0. \tag{2}$$

Als men de tweede vergelijking vermenigvuldigt met $z_1 z_2 z_3 z_4$ wordt duidelijk dat dit een homogene derdegraads vergelijking in vier variabelen is. Als je wilt kun je de eerste vergelijking gebruiken om een variabele te elimineren en dan blijft er nog maar een enkele homogene derdegraads vergelijking over de complexe getallen in drie variabelen over. Kortom, de twee vergelijkingen (2) beschrijven een complexe vlakke projectieve kromme. Wij noemen deze kromme de *Darbouxkromme $D = D_{a,b,c,d}$ bij de stangenvierhoek met lengtes a, b, c, d.*

De posities van de stangenvierhoek komen nu dus overeen met projectieve punten van de vorm $(z_1 : z_2 : z_3 : z_4) \in \mathbb{P}^3$ die voldoen aan de vergelijkingen (2) en aan de eis:

$$|z_1| = |z_2| = |z_3| = |z_4| = 1. \tag{3}$$

De rol van de vier variabelen is dezelfde. Deze kromme hoort bij alle stangenvierhoeken met zijden van lengte a, b, c, d, ongeacht de volgorde. Er zijn dus zes mogelijke volgordes van lengtes bij een stangenvierhoek. In [2] bespreekt Darboux de ambiguïteiten die hierdoor kunnen ontstaan. Dit komt neer op het onderscheid tussen de in 2.1 behandelde klassen I en II van stangenvierhoeken.

Opgave
Laat zien dat er over het algemeen drie verschillende stangenvierhoeken zijn waarbij de verzameling van de lengtes gelijk is aan $\{a, b, c, d\}$.

Er zijn nu verschillende eigenschappen van de derdegraads kromme die corresponderen met eigenschappen van de bijbehorende stangenvierhoeken. Wij noemen er hier maar één.

STELLING 17 *Gegeven een stangenvierhoek S met lengtes a, b, c, d en Darbouxkromme D. Dan is S doorslaand dan en slechts dan als D een singulariteit heeft.*

Opmerking. Vlakke complex-projectieve kubische krommen met een singulariteit kunnen als volgt worden geclassificeerd. Zij bezitten oftewel een rationale parametrisering (d.w.z. zij zijn het beeld van een algebraïsche afbeelding vanuit een projectieve lijn naar het projectieve vlak) of zij vallen uiteen in een gladde kegelsnede en een lijn of zij bestaan uit drie lijnen.

Bewijs: Eerst onderzoeken wij wat het voor de vergelijkingen (2) betekent als D een singulariteit in een punt $Q := (w_1 : w_2 : w_3 : w_4)$ bezit. Hiervoor definiëren wij de rationale functies: $L(z_1, z_2, z_3, z_4) := az_1 + bz_2 + cz_3 + dz_4$ en

$$F(z_1, z_2, z_3, z_4) := \frac{a}{z_1} + \frac{b}{z_2} + \frac{c}{z_3} + \frac{d}{z_4}.$$

Als wij F beperken tot de nulpuntsverzameling van L, dan verdwijnt de afgeleide van F precies daar waar de gradienten van F en L een veelvoud van elkaar zijn.

Dus er zal een complex getal λ bestaan met:

$$\frac{\partial F}{\partial z_i}(Q) = \lambda \frac{\partial L}{\partial z_i}(Q) \quad \text{voor} \quad i = 1, 2, 3, 4.$$

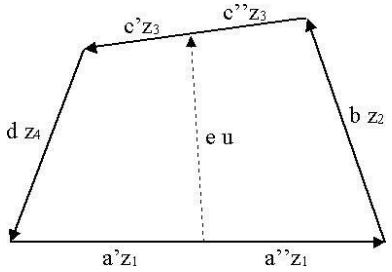
Als wij de afgeleides berekenen vinden wij de volgende vergelijkingen:

$$-\frac{a}{w_1^2} = \lambda a, \quad -\frac{b}{w_2^2} = \lambda b, \quad -\frac{c}{w_3^2} = \lambda c, \quad -\frac{d}{w_4^2} = \lambda d.$$

Hieruit volgt $w_1^2 = w_2^2 = w_3^2 = w_4^2$. Daarnaast geldt $L(Q) = 0$ en wij zien dat geldt $a \pm b \pm c \pm d = 0$ voor ene of andere keuze van tekens '+' of '−'. Op blz. 79 hebben wij gezien dat dit het criterium is voor een doorslaande stangenvierhoek. Verder blijkt dat het singuliere punt in $\mathbb{P}^4$ de gedaante $(\pm 1 : \pm 1 : \pm 1 : \pm 1)$ moet hebben.

Andersom impliceert een relatie van de vorm $a \pm b \pm c \pm d = 0$ dat D een singulariteit in het bijbehorende punt $(\pm 1 : \pm 1 : \pm 1 : \pm 1)$ heeft. □

Opmerking. Het bewijs laat bovendien zien dat de enige mogelijke punten waar D een singulariteit kan hebben van de volgende vorm zijn: $(\pm 1 : \pm 1 : \pm 1 : \pm 1) \in \mathbb{P}^4$.



Figuur 18. *Notatie bij stangenvierhoek met een vijfde stang*

4.1. Stangenvierhoeken met een vijfde stang
Wij beëindigen de bijdrage met de volgende vraag:

> Welke stangenvierhoeken met een extra vijfde stang zijn nog beweegbaar? Is er bijvoorbeeld een beweegbaar 'stangentrapezium' met een vijfde stang? Bestaat er überhaupt nog een andere stangenvierhoek met een vijfde stang behalve het stangenparallellogram met een vijfde stang die aan twee zijden parallel is?

Opmerking. De bovengenoemde vraag werd mij twee jaar geleden gesteld door Leon van den Broek en Dolf van den Hombergh, twee auteurs van de Ratio internetmethode [19]. In het onderzoeksgedeelte van het hoofdstuk *Gelijkvormigheid* voor de tweede klas behandelen zij deze vraag voor het speciale geval van stangenparallellogrammen. Op de algemene vraag konden wij geen antwoord in de literatuur vinden en een elementair bewijs van de stelling hieronder kennen wij tot nu toe niet. Het hier gegeven bewijs maakt gebruik van de afbeelding van Darboux, is verrassend eenvoudig en laat de kracht van abstractie zien.

Opgave
Ga naar de webpagina van Ratio, [19], lesmateriaal, hoofdstuk 14, *Gelijkvormigheid*, paragraaf 14.4, *Onderzoek* en maak opgave 3.

STELLING 18 *Gegeven een stangenvierhoek met een vijfde stang die twee punten van de stangenvierhoek verbindt. Dan is deze stangenconstructie beweegbaar dan en slechts dan als de stangenvierhoek een stangenparallellogram is waarbij de vijfde stang evenwijdig loopt aan twee van de vier stangen.*

Voordat wij met het bewijs beginnen noteren wij de stangen en lengtes in *de vierstangenvierhoek met een vijfde stang* zoals dat in Figuur 18 is te zien.

De constructie van Darboux levert ons voor elk van de twee kleine vierhoeken een projectieve kubische kromme. Wij vinden de vier vergelijkingen:

$$a'z_1 + eu + c'z_3 + dz_4 = 0, \tag{4}$$

$$\frac{a}{z_1} + \frac{e}{u} + \frac{c'}{z_3} + \frac{d}{z_4} = 0, \tag{5}$$

$$a''z_1 + bz_2 + c''z_3 - eu = 0, \tag{6}$$

$$\frac{a''}{z_1} + \frac{b}{z_2} + \frac{c''}{z_3} - \frac{e}{u} = 0. \tag{7}$$

De posities van de stangenconstructie komen overeen met

$$(z_1 : z_2 : u : z_3 : z_4) \in \mathbb{P}^4,$$

waarvoor geldt: $|z_1| = |z_2| = |u| = |z_3| = |z_4| = 1$.

Twee verschillende lijnen in het projectieve vlak snijden elkaar altijd precies een keer. Dat is het mooie aan projectieve meetkunde – zo is het projectieve vlak juist gemaakt. Ook andere krommen in het projectieve vlak zonder gemeenschappelijke component hebben altijd minstens één en nooit meer dan eindig veel snijpunten. Als wij de snijpunten met multipliciteiten tellen dan is het aantal snijpunten van twee krommen precies gelijk aan het product van de graden van hun vergelijkingen. Deze stelling staat bekend als *de stelling van Bezout* en is in ieder inleidend boek over algebraïsche krommen te vinden.

Bewijs van 18: Het is duidelijk dat een stangenparallellogram met een evenwijdige vijfde stang beweegbaar is. Het gaat dus om de omkering.

Nu gaan wij uit van een beweegbare stangenvierhoek met een extra verbindingsstang. De lineaire condities (4) en (6) snijden een projectief vlak V uit een vierdimensionale projectieve ruimte en de derdegraads vergelijkingen (5) en (7) definiëren twee kubische krommen D' en D'' in dit projectieve vlak V.

Als de stangenconstructie beweegbaar is dan zijn er oneindig veel punten (posities van de stangenconstructie) in $\mathbb{P}^4$ die aan alle vier vergelijkingen, (4) tot (7), voldoen. De stelling van Bezout echter zegt dat twee kubische krommen hoogstens eindig veel snijpunten (met multipliciteit geteld, negen) kunnen hebben tenzij zij een component gemeen hebben. Bij D_1 en D_2 moet dat dus het geval zijn en we noemen de gemeenschappelijke component C. De kromme C is óf de hele kromme óf een lijn óf een kegelsnede.

In alle drie gevallen weten wij eveneens door de stelling van Bezout dat C elke lijn in V minstens een keer snijdt. In het bijzonder moet C een snijpunt hebben met de 'lijnen in oneindig' op V: $z_1 = 0$, $z_2 = 0$, $u = 0$, $z_3 = 0$ of $z_4 = 0$.[3] Deze snijpunten gan wij nu onderzoeken.

$u = 0$: De vergelijkingen (5), (7) vermenigvuldigd met $z_1z_2z_3$ samen met 4 en 6 worden dan:

$$a'z_1 + c'z_3 + dz_4 = 0, \qquad ez_1z_3z_4 = 0,$$

$$a''z_1 + bz_2 + c''z_3 = 0, \qquad -ez_1z_2z_3 = 0.$$

Omdat wij te maken hebben met projectieve coördinaten kunnen nooit alle vijf coördinaten $(z_1 : z_2 : u : z_3 : z_4) \in \mathbb{P}^4$ tegelijkertijd verdwijnen. Nu zijn er vier disjuncte gevallen te onderscheiden.

[3] Let op de 'ironie van abstractie': punten op de lijnen in oneindig hebben geen voor de hand liggende betekenis als posities van de stangenconstructie meer.

$z_1 = 0$: Dan blijft $c'z_3 + dz_4 = 0$ en $bz_2 + c''z_3 = 0$. Er blijft dus nog maar één projectief punt over $P := (0 : -c''d : 0 : bd : -c'b)$.

$z_2 = 0$: Hier blijft $a'z_1 + c'z_3 + dz_4 = 0$ en $ez_1z_3z_4 = 0$ als ook de vergelijking $a''z_1 + c''z_3 = 0$.

Als hier een snijpunt is dan moet gelden:

$$R := (-c' : 0 : 0 : a' : 0) = (-c'' : 0 : 0 : a'' : 0).$$

$z_3 = 0$: Hier is het snijpunt gelijk aan $S := (-db : da'' : 0 : 0 : a'b)$.

$z_4 = 0$: Als hier een snijpunt is dan is dit snijpunt noodzakelijkerwijs gelijk aan $R := (-c' : 0 : 0 : a' : 0) = (-c'' : 0 : 0 : a'' : 0)$.

De snijpunten met de andere lijnen in oneindig worden op dezelfde manier gevonden.

$z_1 = 0$: Hier zijn er twee mogelijke snijpunten: $P(0 : -c''d : 0 : bd : -c'b)$ of $Q := (0 : ed : bd : 0 : -eb)$.

$z_2 = 0$: $R = (-c' : 0 : 0 : a' : 0) = (-c'' : 0 : 0 : a'' : 0)$ is het enig mogelijke snijpunt in dit geval. Dit snijpunt R heeft multipliciteit 1, want de lijn $z_2 = 0$ wordt geparametriseerd door $(z_1 : z_3) \in \mathbb{P}^1$: gegeven z_1 en z_3 kun je de variabelen u en z_4 uitrekenen. Gereduceerd tot $z_2 = 0$ is de kubische vergelijking van D' nu $z_1z_3(a'z_1 + c'z_3) = 0$. Daarmee hebben C en de lijn $z_2 = 0$ een snijpunt van multipliciteit 1 in R. Met de stelling van Bezout concluderen wij: C **is een lijn!**

$z_3 = 0$: $S = (-db : da'' : 0 : 0 : a'b)$ of $Q := (0 : ed : bd : 0 : -eb)$ zijn mogelijke snijpunten in dit geval.

$z_4 = 0$: $R = (-c' : 0 : 0 : a' : 0) = (-c'' : 0 : 0 : a'' : 0)$ levert het snijpunt en een conditie.

Samengevat vinden wij de volgende mogelijke snijpunten met de lijnen in oneindig:

$z_1 = 0$	$z_2 = 0$	$u = 0$	$z_3 = 0$	$z_4 = 0$
P, Q	R	P, R, S	Q, S	R

.

Omdat wij weten dat C een lijn moet zijn kan deze configuratie alleen worden bereikt als C de lijn is die R en Q verbindt. Daarom moet dan elk punt op de lijn

$$(t_0 : t_1) \mapsto (-t_0c' : t_1ed : t_1bd : t_0a' : -t_1eb)$$

voldoen aan (5) en (7). Invullen levert samen met de eis over het bestaan van R de gewenste identiteiten: $e = b = d$, $c' = a'$, $c'' = a''$. □

Opgave

Laat zien dat de snijpunten van C met de lijnen in oneindig zo zijn als in het bewijs wordt beweerd.

Literatuur

1. W. Blaschke, H.R. Müller (1956). *Ebene Kinematik*. Verlag von R. Oldenbourg, München.
2. M.G. Darboux (1879). *De lémploi des fonctions hyperelliptiques dans la théorie de quadrilatère plan*, Bulletin des sciences mathématiques et astronomiques, Gauthiers-Villars, Paris, 109–128.
3. R. Hänchen (1932). *Winden und Krane, Aufbau, Berechnung und Konstruktion*. Verlag Julius Springer, Berlin.
4. A.S. Hall (1961). *Kinematics an Linkage Design*, Prentice-Hall, Inc., Englewood Cliffs, N.J.
5. K.P. Hart (1999). *De lemniscaat*, Pythagoras, oktober 1999.
6. J.A. Hrones, G.L. Nelson (1951). *Analysis of the Four-Bar Linkage*, I and II, The Technology Press of M.I.T. and John Wiley & Sons. **3**, 31–59.
7. A.B. Kempe (1849). *How to draw a straight line*, National Council of Teachers in Mathematics, Classics in mathematics education, 6, 1977.
8. G. Niemann(1928). *Über Wippkrane mit waagerechtem Lastweg*, Dissertation Technische Hochschule Berlin.
9. K. Rüdiger (1930). *Einziehkrane mit waagerechter Lastbahn*, Fördertechnik und Frachtverkehr, XXIII (9), 169–172.
10. Selter (1927). *Wippauslegerkrane mit horizontal bewegter Last*. Fördertechnische Rundschau, **2**, 28–31.
11. J.E. de Vries (1929). *Hijschwerktuigen*. Firma Ruijgrok & Co., Haarlem.
12. B. Vrugt (1998). *Van wagenkraan tot lemniscaat*, deel I en II. Meccano Nieuws 16.1 en 16.2.
13. F. Westendorp (1927). *Handboek voor werktuigkundigen*. Zevende druk van Bernoulli's vademecum, Van Holkema & Warendorf's Uitg. Mij., Amsterdam.

Profielwerkstukken

14. D. van Gemert, T. Wijnen (2003). *De Ellipskraan*, vwo-profielwerkstuk, Pleincollege Eckart, Eindhoven, Cursus *Wiskundig denken*, Ratio, RU Nijmegen.
15. M. Hercules, J. van der Velde (2005). *Hijskranen met een horizontale belastingsweg*, havo-profielwerkstuk, Canisius College Nijmegen.
16. J. Hoek, R. Balsem(2004). *Lemniscaat*, vwo-profielwerkstuk, Canisius College Nijmegen.

Webpagina's

17. FIGEE crane building, Haarlem, http://www.figee.com/.
18. KE Kranbau Eberswalde, http://www.kranbau-eberswalde.de/ (voormalig ARDELT Werke).
19. Ratio Instituut en Internetmethode, http://www.ratio.ru.nl/,

Radboud Universiteit Nijmegen.
20. Wiskunde-B-dag (2004). *Dansende Stangen*, http://www.fi.uu.nl/wisbdag, Freudenthal Instituut, Universiteit Utrecht.

CWI Vakantiecursus 2005

pp. 93 – 108

De '*abc*-formule' voor hogere-graadsvergelijkingen

E. Coplakova
Faculteit EWI
Technische Universiteit Delft
e-mail: e.coplakova@ewi.tudelft.nl

1. Inleiding

Een van de oudste problemen in de wiskunde was oplossingen te vinden van vergelijkingen van de vorm

$$a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 = 0,$$

waarbij n een positief natuurlijk getal is, $a_n \neq 0$ en $a_n, a_{n-1}, \ldots, a_1, a_0$ reële getallen zijn. Zo'n vergelijking heet een *n-de graadsvergelijking* en de uitdrukking $p(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$ is een *reëel n-de graads-polynoom*. Men zocht oplossingen die uit te drukken waren met behulp van eindig veel toepassingen van de operaties optellen, aftrekken, vermenigvuldigen, delen en k-de machtsworteltrekken, dat wil zeggen, oplossingen in *radikalen*.

De eenvoudigste, lineaire, vergelijkingen (als $n = 1$) zijn makkelijk op te lossen: $ax + b = 0$ heeft precies één oplossing $x = -b/a$. We kennen ook allemaal de *abc*-formule voor het oplossen van een kwadratische vergelijking (als $n = 2$)

$$ax^2 + bx + c = 0. \tag{1}$$

Deze vergelijking heeft alleen dan oplossingen als de discriminant $D = b^2 - 4ac$ niet-negatief is en de oplossingen zijn

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}. \tag{2}$$

Het is niet moeilijk om de formule af te leiden. Omdat $a \neq 0$ kunnen we de vergelijking (1) door a delen (de oplossingen veranderen, natuurlijk, niet):

$$x^2 + \frac{b}{a}x + \frac{c}{a} = 0. \tag{3}$$

Door kwadraat af te splitsen elimineren we de lineaire term in (3):

$$x^2 + \frac{b}{a}x + \frac{c}{a} = \left(x + \frac{b}{2a}\right)^2 - \frac{b^2 - 4ac}{4a^2}. \tag{4}$$

Hiermee kan (3) geschreven worden als

$$\left(x + \frac{b}{2a}\right)^2 = \frac{b^2 - 4ac}{4a^2}.$$

Als de rechterkant niet negatief is vinden we

$$x + \frac{b}{2a} = \pm\frac{\sqrt{b^2 - 4ac}}{2a}.$$

De *abc*-formule volgt hieruit onmiddelijk.

Ook voor derde-graadsvergelijkingen had men al in de zestiende eeuw dergelijke formules; voor vergelijkingen van de vorm $x^3 + px + q = 0$ levert de volgende stelling een formule voor een oplossing (we zullen de formule in de volgende paragraaf afleiden):

Stelling 19 (Formules van Cardano) *De derde-graadsvergelijking*

$$x^3 + px + q = 0 \tag{5}$$

heeft een oplossing van de vorm

$$x = \sqrt[3]{-\frac{q}{2} + \frac{\sqrt{-3D}}{18}} + \sqrt[3]{-\frac{q}{2} - \frac{\sqrt{-3D}}{18}} \qquad \text{met} \qquad D = -4p^3 - 27q^2. \tag{6}$$

Het getal D heet de *discriminant* van de derde-graadsvergelijking (5). In tegenstelling tot de *abc*-formule vertelt deze formule ons niet of de vergelijking een oplossing heeft of niet zoals het volgende voorbeeld laat zien.

Voorbeeld 1 Bekijk de vergelijking

$$x^3 - 15x - 4 = 0.$$

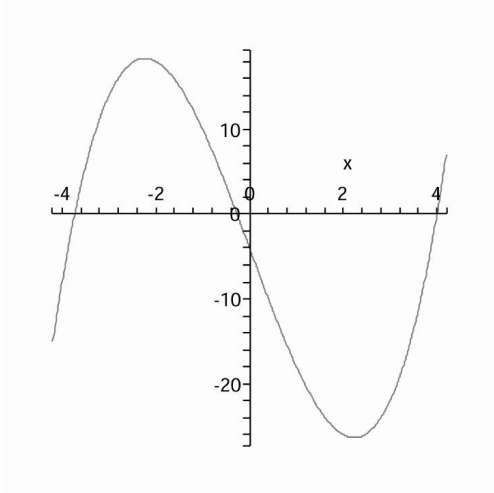
De discriminant D is gelijk aan $D = -4 \cdot (-15^3) - 27 \cdot (-4)^2 = 13068$ en dus is het getal $-3D$ negatief. De vergelijking lijkt dus onoplosbaar: we kunnen geen wortel van een negatief getal nemen. Invullen leert echter dat $x = 4$ wel een oplossing is en als we de grafiek van de functie $f(x) = x^3 - 15x - 4$ bekijken (zie Figuur 1) zien we dat er eigenlijk nog twee andere oplossingen zijn. Men zat dus met een probleem: de vergelijking heeft een reële oplossing die niet uit de formule te halen was.

De complexe getallen werden ingevoerd (nog niet met die naam) om toch enige zin aan zulke formules te geven. Invoering van complexe getallen heeft ook andere belangrijke 'neveneffecten' wat betreft oplossingen van vergelijkingen: we zullen zien dat elke algebraïsche vergelijking een (complexe) oplossing heeft, in het bijzonder ook kwadratische vergelijkingen met een negatieve discriminant.

2. Complexe getallen

De complexe getallen kunnen we als volgt definiëren: We voeren een nieuw getal, de *imaginaire eenheid* i, in met de eigenschap dat

$$i^2 = -1.$$



Figuur 1. *De grafiek van de functie* $f(x) = x^3 - 15x - 4$

De *complexe getallen* zijn dan de getallen van de vorm $z = a + bi$ met a en b reële getallen; het getal a heet het *reële deel* en b het *imaginaire deel van* z. De oude, reële, getallen kunnen we ook als complexe getallen beschouwen door het imaginaire deel gelijk aan 0 te stellen.

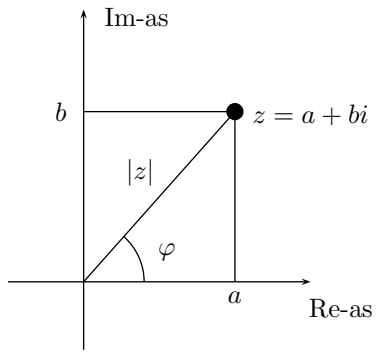
De nieuwe getallen kunnen we zonder problemen bij elkaar optellen en met elkaar vermenigvuldigen[1]; telkens als we i^2 tegenkomen vervangen we dat door -1. Ook delen gaat zonder problemen: om z/w te berekenen kunnen we de teller en de noemer met de complexe toegevoegde $\overline{w} = c - di$ van het getal $w = c + di$ vermenigvudigen.

Met behulp van complexe getallen is nu elke kwadratische vergelijking op te lossen: als de discriminant $b^2 - 4ac$ van (1) negatief is dan is $4ac - b^2 > 0$ en omdat $b^2 - 4ac = -1 \cdot (4ac - b^2) = i^2(4ac - b^2)$ krijgen we

$$x = \frac{-b \pm i\sqrt{4ac - b^2}}{2a}.$$

Net zoals we van de reële getallen een plaatje in ons hoofd hebben — de getallen-rechte met de negatieve getallen links en de positieve getallen rechts van het getal 0 — kunnen we ook de complexe getallen visualiseren. We hebben daarvoor het platte vlak nodig: het complex getal $z = a + bi$ correspondeert met het punt (a, b), zie Figuur 2.

[1] Om helemaal precies te zijn zouden we ook moeten nagaan dat de rekenregels voor het optellen en de vermenigvuldiging van reële getallen uit te breiden zijn voor complexe getallen.



***Figuur 2**. Het getal $z = a + bi$ in het complexe vlak*

De afstand $|z|$ van het punt (a, b) tot de oorsprong heet de *modulus* van z en de kleinste positieve hoek φ die het lijnstuk met eindpunten $(0, 0)$ en (a, b) met de positieve Re-as maakt wordt het *argument* van z genoemd. Uit de figuur is makkelijk af te lezen dat voor de modulus en het argument van z geldt

$$|z| = \sqrt{a^2 + b^2}, \qquad a = |z| \cos\varphi \qquad \text{en} \qquad b = |z| \sin\varphi.$$

We zien dus dat het complexe getal z ook eenduidig bepaald wordt door zijn modulus en argument:

$$z = |z| \left(\cos\varphi + i \sin\varphi\right).$$

Met behulp van de modulus en het argument kunnen we mooie formules opstellen voor het product en quotiënt van twee complexe getallen. Neem maar twee complexe getallen en schrijf ze in modulus-argument-vorm:

$$z = |z|(\cos\varphi + i \sin\varphi) \qquad \text{en} \qquad w = |w|(\cos\psi + i \sin\psi).$$

We werken het product uit en gebruiken een paar gonioformules:

$$\begin{aligned} zw &= |z|(\cos\varphi + i\sin\varphi)|w|(\cos\psi + i\sin\psi) \\ &= |z||w|\big((\cos\varphi\cos\psi - \sin\varphi\sin\psi) + i(\sin\varphi\cos\psi + \cos\varphi\sin\psi)\big) \\ &= |z||w|\big(\cos(\varphi+\psi) + i\sin(\varphi+\psi)\big). \end{aligned}$$

We zien dat we bij vermenigvuldiging van twee complexe getallen de modulussen met elkaar moeten vermenigvuldigen en de argumenten bij elkaar optellen (we moeten echter modulo 2π rekenen om het argument tussen 0 en 2π radialen te houden). Analoog kunnen we afleiden dat we bij het delen de modulussen door elkaar moeten delen en de argumenten van elkaar af moeten trekken.

Deze methode kunnen we generaliseren ook om producten van meer dan twee complexe getallen te berekenen. We krijgen de volgende stelling[2]:

STELLING 20 *Zij* $z = |z|(\cos\varphi + i\sin\varphi)$ *een complex getal en* n *een natuurlijk getal. Dan geldt*

$$z^n = |z|^n(\cos n\varphi + i\sin n\varphi).$$

We kunnen Stelling 20 gebruiken om n-de machtswortels van de complexe getallen te trekken. Laat w een complex getal zijn met modulus $|w|$ en argument ψ. Zij $n > 1$ een natuurlijk getal. We zoeken alle complexe getallen z waarvoor geldt $z^n = w$.

Welnu, schrijf z ook in modulus-argument-vorm: $z = |z|(\cos\theta + i\sin\theta)$ en bereken zijn n-de macht:

$$z^n = |z|^n(\cos n\theta + i\sin n\theta).$$

Nu moet dus gelden

$$|z|^n = |w| \qquad \text{en} \qquad \cos n\theta + i\sin n\theta = \cos\psi + i\sin\psi.$$

De eerste vergelijking geeft meteen $|z| = \sqrt[n]{|w|}$ (dit omdat de modulussen $|w|$ en $|z|$ niet-negatieve reële getallen zijn). De tweede vergelijking geeft $n\theta = \psi + 2k\pi$ voor een of ander geheel getal k. We kunnen dit ook schrijven als

$$\theta = \frac{\psi}{n} + \frac{2k\pi}{n}.$$

We merken nu op dat voor een k en een l die n verschillen de bijbehorende θ's 2π verschillen en dus dezelfde z opleveren. We zien dat er precies n verschillende n-de machtswortels[3] van w bestaan:

$$z_k = \sqrt[n]{|w|}\left(\cos\left(\frac{\psi + 2k\pi}{n}\right) + i\sin\left(\frac{\psi + 2k\pi}{n}\right)\right),$$

waarbij $k = 0, 1, \ldots, n-1$. De n-de machtswortels van w corresponderen met n punten op de cirkel met straal $\sqrt[n]{|w|}$ en middelpunt $(0,0)$ die een regelmatige n-hoek maken.

VOORBEELD 2 We bepalen alle vierde-machtswortels van -3. Het argument van -3 is π en de modulus is 3. Elke vierde-machtswortel heeft dus modulus $\sqrt[4]{3}$. Volgens de bovenstaande redenering vinden we vier hoeken, namelijk $\frac{\pi}{4} + k\frac{\pi}{2}$ met $k = 0, 1, 2, 3$. De bijbehorende complexe getallen zijn:

[2] In het speciale geval wanneer $|z| = 1$ krijgen we de *Formule van de Moivre*:

$$(\cos\varphi + i\sin\varphi)^n = \cos n\varphi + i\sin n\varphi.$$

[3] Het is niet mogelijk om n-de machtswortels van negatieve of complexe getallen die niet reëel zijn eenduidig te bepalen; we kunnen dus ook niet zo maar gewone rekenregels die voor wortels van reële getallen gelden gebruiken zoals de volgende 'redenering' laat zien:

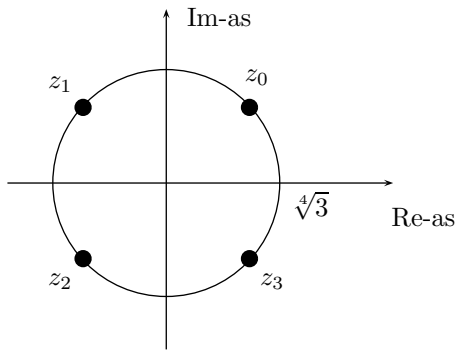
Omdat $i^2 = -1$ volgt er $\sqrt{-1} = i$. Maar dan zou moeten gelden (als we dus de bekende rekenregels voor de wortels van de reële getallen toepassen)

$$-1 = i^2 = i \cdot i = \sqrt{-1}\cdot\sqrt{-1} = \sqrt{(-1)(-1)} = \sqrt{1} = 1.$$

We kregen $-1 = 1$ wat natuurlijk onzin is.

$$
\begin{aligned}
z_0 &= \sqrt[4]{3}\left(\sqrt{2}/2 + i\sqrt{2}/2\right)\\
z_1 &= \sqrt[4]{3}\left(-\sqrt{2}/2 + i\sqrt{2}/2\right)\\
z_2 &= \sqrt[4]{3}\left(-\sqrt{2}/2 - i\sqrt{2}/2\right)\\
z_3 &= \sqrt[4]{3}\left(\sqrt{2}/2 - i\sqrt{2}/2\right).
\end{aligned}
$$

De oplossingen zijn getekend in Figuur 3.



Figuur 3. *De vierde machtswortels uit* $z = -3$

Opgave 1 Laat zien:

(a) Als z een reëel getal is dan $\overline{z} = z$.
(b) Voor elk complex getal z en elk natuurlijk getal n geldt $\overline{z^n} = \overline{z}^n$.

Opgave 2 Vind alle derde-machtswortels van 1.

Opgave 3 Vind alle complexe oplossingen van $z^4 = -1 + i$.

3. Derde- en vierdegraadsvergelijkingen

Al in de zestiende eeuw kon men vergelijkingen van de vorm $x^3 + px + q = 0$ oplossen; een van de eerste wiskundigen die al in 1515 een oplossing vond was Scipione del Ferro. Hij heeft helaas zijn uitvinding nooit gepubliceerd.

Veel meer bekend werd de oplossing gevonden door Tartaglia in 1535 die hij in 1539 aan Cardano heeft gegeven en die deze in zijn boek Ars Magna publiceerde, zie [3]. De beschrijving van die oplossing was kort en op rijm gesteld (zie [8], p.22):

> Quando che'l cubo con le cose appresso
> Se agguaglia a qualche numero discreto:
> Trovan dui altri, differenti in esso.

Dapoi terrai, questo per consueto,
Che'l lor produtto, sempre sia equale
Al terzo cubo delle cose neto;
El residuo poi suo generale,
Delli lor lati cubi, bene sottratti
Varra la tua cosa principale.

Cardano gaf in zijn boek niet alleen de formules voor de oplossing van derde-graadsvergelijkingen maar ook hun afleiding. In Ars Magna is ook te lezen hoe vierde-graadsvergelijkingen op te lossen zijn. We zullen — in moderne notatie — de afleidingen van Cardano uitleggen.

We beschouwen een algemene derde-graadsvergelijking

$$ax^3 + bx^2 + cx + d = 0, \tag{7}$$

waarbij a, b, c en d reële getallen zijn en $a \neq 0$. We delen eerst beide kanten van (7) door a:

$$x^3 + \frac{b}{a}x^2 + \frac{c}{a}x + \frac{d}{a} = 0. \tag{8}$$

Hierdoor veranderen de oplossingen niet: elke oplossing van (7) is een oplossing van (8) en omgekeerd. We hebben al gezien dat de *abc*-formule voor kwadratische vergelijkingen makkelijk af te leiden was zodra we de lineaire term hebben geëlimineerd. In het geval van derde-graadsvergelijkingen kunnen we bewijzen dat de kwadratische term in (8) ook te elimeneren is, zodat na een geschikte substitutie elke derde-graadsvergelijking in de vorm $y^3 + py + q = 0$ geschreven kan worden.

Het is niet moeilijk om na te gaan dat de volgende formule geldt:

$$(x+m)^3 = x^3 + 3mx^2 + 3m^2x + m^3.$$

Neem $m = \frac{b}{3a}$ dan geldt

$$\left(x + \frac{b}{3a}\right)^3 = x^3 + \frac{b}{a}x^2 + \frac{b^2}{3a^2}x + \frac{b^3}{27a^3}.$$

Hieruit volgt dat (8) te schrijven is als

$$\left(x + \frac{b}{3a}\right)^3 + \left(\frac{c}{a} - \frac{b^2}{3a^2}\right)x + \frac{d}{a} - \frac{b^3}{27a^3} = 0. \tag{9}$$

Substitueer $y = x + \frac{b}{3a}$ in de linkerkant van (9); we krijgen

$$\begin{aligned} y^3 + \left(\frac{c}{a} - \frac{b^2}{3a^2}\right)\left(y - \frac{b}{3a}\right) + \frac{d}{a} - \frac{b^3}{27a^3} = \\ = \quad y^3 + \left(\frac{c}{a} - \frac{b^2}{3a^2}\right)y + \frac{d}{a} - \frac{bc}{3a^2} + \frac{2b^3}{27a^3}. \end{aligned}$$

Noem nu $p = \frac{c}{a} - \frac{b^2}{3a^2}$ en $q = \frac{d}{a} - \frac{bc}{3a^2} + \frac{2b^3}{27a^3}$. De vergelijking (9) is dan van de vorm $y^3 + py + q = 0$. We hebben bewezen:

STELLING 21 *Elke derde-graadsvergelijking* $ax^3 + bx^2 + cx + d = 0$ *is na de substitutie* $y = x + \frac{b}{3a}$ *te schrijven in de vorm*

$$y^3 + py + q = 0 \tag{10}$$

met p *en* q *reële getallen.*

Als we de vergelijking (10) kunnen oplossen kunnen we ook de algemene derde-graadsvergelijking oplossen: voor elke oplossing y van (10) is $y - \frac{b}{3a}$ een oplossing van (7) en omgekeerd. Vanaf nu zullen we dus alleen derde-graadsvergelijkingen van de vorm (10) beschouwen.

Om de oplossing van (10) te vinden passen we weer een nieuwe substitutie toe: schrijf $y = u + v$. Dan geldt

$$y^3 = u^3 + 3u^2v + 3uv^2 + v^3 = u^3 + v^3 + 3uv(u + v),$$

de vergelijking (10) is dan te schrijven als

$$u^3 + v^3 + (3uv + p)(u + v) + q = 0.$$

Om de linkerkant gelijk aan 0 te krijgen proberen we u en v zó te vinden dat

$$u^3 + v^3 = -q \qquad \text{en} \qquad 3uv = -p.$$

Vul $u = -\frac{p}{v}$ in de vergelijking $u^3 + v^3 = -q$ in:

$$-\frac{p^3}{27v^3} + v^3 = -q.$$

Breng nu $-q$ naar de linkerkant en vermenigvuldig de hele vergelijking met v^3:

$$v^6 + qv^3 - \frac{p^3}{27} = 0.$$

Maar door $v^3 = V$ te stellen krijgen we dan een kwadratische vergelijking

$$V^2 + qV - \frac{p^3}{27} = 0$$

die we kunnen oplossen:

$$V = \frac{-q \pm \sqrt{q^2 + \frac{4p^3}{27}}}{2} = -\frac{q}{2} \pm \frac{\sqrt{3(27q^2 + 4p^3)}}{18}.$$

Bedenk nu dat $v^3 = V$ en $u^3 = -q - v^3$; we krijgen

$$u^3 = -\frac{q}{2} \mp \frac{\sqrt{3(27q^2 + 4p^3)}}{18}.$$

We hebben dus u en v gevonden:

$$u = \sqrt[3]{-\frac{q}{2} \mp \frac{\sqrt{3(27q^2 + 4p^3)}}{18}} \qquad \text{en} \qquad v = \sqrt[3]{-\frac{q}{2} \pm \frac{\sqrt{3(27q^2 + 4p^3)}}{18}}.$$

Het is niet moeilijk om in te zien dat $y = u + v = v + u$ de oplossing in (6) is. We hebben Stelling 19 bewezen.

VOORBEELD 3 We bekijken de vergelijking $x^3 - 15x - 4 = 0$ uit Voorbeeld (1) nog een keer. De discriminant D is gelijk aan 13068; in plaats van $\sqrt{-3D}$ schrijven we $i\sqrt{3D}$ in de formule (6) van Cardano. Als we die uitwerken krijgen we

$$x = \sqrt[3]{2 + 11i} + \sqrt[3]{2 - 11i}.$$

Dit ziet er complex uit en de vraag is hoe we de oplossing $x = 4$ uit deze formule kunnen krijgen. In 1572 merkte Rafaele Bombelli op dat $(2 \pm i)^3 = 2 \pm 11i$; hiermee is $\sqrt[3]{2 + 11i} + \sqrt[3]{2 - 11i}$ te vereenvoudigen tot $(2 + i) + (2 - i) = 4$.

Er is nog een vraag die we moeten beantwoorden: de *abc*-formule geeft alle twee oplossingen van de kwadratische vergelijking door middel van de $\pm$. Hoe kunnen we met behulp van de formule van Cardano de *drie* oplossingen van $x^3 - 15x - 4 = 0$ maken?

Hiertoe moeten we ons realiseren dat $\sqrt[3]{2 + 11i}$ eigenlijk *drie* betekenissen heeft: we kunnen elk van de drie oplossingen van $z^3 = 2 + 11i$ gebruiken. Die kunnen we in dit geval maken uit $2 + i$ en de oplossingen van $z^3 = 1$. Die oplossingen van $z^3 = 1$ zijn 1, $\omega = -\frac{1}{2} + \frac{1}{2}\sqrt{3}\,i$ en $\omega^2 = -\frac{1}{2} - \frac{1}{2}\sqrt{3}\,i$, zie Opgave 2. De derde-machtswortels van $2+11i$ zijn dan $u_0 = 2+i$, $u_1 = (2+i)\omega$ en $u_2 = (2 + i)\omega^2$.

Evenzo krijgen we de derde-machtswortels van $2 - 11i$: $v_0 = 2 - i$, $v_1 = (2 - i)\omega$ en $v_2 = (2 - i)\omega^2$. Met de eis $3uv = 15$ zien we dat van de negen mogelijke combinaties $x = u_j + v_k$ met $j, k = 0, 1, 2$ er drie zijn die een oplossing opleveren: $x_0 = u_0 + v_0$, $x_1 = u_1 + v_2$ en $x_2 = u_2 + v_1$, immers, er moet gelden

$$3u_j v_k = 15$$

en ook

$$3u_j v_k = 3(2 + i)(2 - i)\omega^{j+k} = 15\omega^{j+k}.$$

We kunnen dus alleen de paren met $j + k = 0$ of $j + k = 3$ gebruiken.

OPMERKING 1 De vergelijking $x^3 - 15x - 4 = 0$ was natuurlijk zó in elkaar gestoken dat $x = 4$ een makkelijk te vinden oplossing is. In het algemeen, als de discriminant D positief is en als we dus met $\sqrt[3]{\frac{q}{2} \pm \frac{i\sqrt{3D}}{18}}$ te maken krijgen, is het niet zo eenvoudig de oplossingen in zuiver reële vorm te krijgen. Er geldt namelijk het volgende: voor een vergelijking $x^3 + px + q = 0$ met positieve discriminant zijn er twee mogelijkheden:

1) er is een oplossing die *zonder* wortels al in p en q uit te drukken is, zoals $x = 4$ in Voorbeeld 3; of

2) géén van de oplossingen is met behulp van alléén reële getallen in radikalen uit te drukken, terwijl er wel een reële oplossing is (zie Stelling 22).

Dit geval heet daarom *Casus Irreducibilis*: de formule $\sqrt[3]{\frac{q}{2} + \frac{i\sqrt{3D}}{18}} + \sqrt[3]{\frac{q}{2} - \frac{i\sqrt{3D}}{18}}$ is niet verder te vereenvoudigen, met behulp van wortels alleen.

We hebben gezien, op pagina 97, dat elk complex getal derde-machtswortels heeft, deze vonden we door modulus en argument te rekenen. Dit levert in het irreducibele geval een formule voor de reële oplossing met behulp van de cosinus-functie, zoals het volgende voorbeeld laat zien.

VOORBEELD 4 Toepassing van de formule van Cardano op de vergelijking

$$x^3 - 3x - 1 = 0$$

geeft $D = 81$ en dus $\sqrt{-3D} = 9\sqrt{3}i$. Na uitwerking krijgen we

$$x = \sqrt[3]{\frac{1}{2} + \frac{1}{2}\sqrt{3}\,i} + \sqrt[3]{\frac{1}{2} - \frac{1}{2}\sqrt{3}\,i}$$

als oplossing. Nu geldt $\frac{1}{2} + \frac{1}{2}\sqrt{3}\,i = \cos\frac{\pi}{3} + i\sin\frac{\pi}{3}$ en dus kunnen we $\cos\frac{\pi}{3} + i\sin\frac{\pi}{3}$ als waarde van $\sqrt[3]{\frac{1}{2} + \frac{1}{2}\sqrt{3}\,i}$ nemen, en ook $\cos\frac{\pi}{3} - i\sin\frac{\pi}{3}$ als waarde van $\sqrt[3]{\frac{1}{2} - \frac{1}{2}\sqrt{3}\,i}$ (zie pagina 97). Daarmee vinden we $2\cos\frac{\pi}{9}$ als oplossing van $x^3 - 3x - 1 = 0$. Merk op dat deze oplossing niet in radikalen uitgedrukt is.

Analoog als bij tweede- en derde-graadsvergelijkingen kunnen we in elke vierde-graadsvergelijking door een geschikte substitutie de derde-machtsterm elimineren (zie Opgave (5)) en na enig werk een formule voor de oplossing afleiden. We laten slechts het idee zien; het is een goede oefening de details zelf uit te werken. De methode is afkomstig van Descartes (hij vond hem in 1637), zie [1]. Laat (na het elimineren van de derde-machtsvorm)

$$x^4 + px^2 + qx + r = 0$$

een vierde-machtsvergelijking zijn. Door invullen en alles uitschrijven kunnen we eerst laten zien dat de linkerkant te schrijven is als

$$x^4 + px^2 + qx + r = (x^2 + ux + v)(x^2 - ux + w),$$

waarbij u, v en w aan de volgende eisen voldoen:

$$\begin{aligned} v + w - u^2 &= p \\ u(w - v) &= q \\ vw &= r \end{aligned}$$

We elimineren nu v en w om een zesde-graadsvergelijking van de vorm

$$u^6 + su^2 + t = 0$$

te krijgen. Na de substitutie $u^2 = U$ hebben we een derde-graadsvergelijking $U^3 + sU + t = 0$ die we met behulp van de formule van Cardano kunnen oplossen.

OPGAVE 4 Los de volgende derde-graadsvergelijkingen op.

(a) $x^3 + 3x = 10$;
(b) $x^3 - 9x^2 + 21x - 5 = 0$;
(c) $x^3 - 7x - 7 = 0$.

Opgave 5 Laat zien dat elke vierde-graads vergelijking

$$ax^4 + bx^3 + cx^2 + dx + e = 0$$

na de substitutie $y = x + \frac{b}{4a}$ te schrijven is in de vorm

$$y^4 + py^2 + qy + r = 0. \tag{11}$$

Opgave 6 Los de volgende vierde-graadsvergelijkingen op.

(a) $x^4 + x^2 + 4x - 3 = 0$;
(b) $x^4 - 2x^2 + 8x - 3 = 0$.

4. Hoofdstelling van de Algebra

We hebben al gezien dat door het introduceren van complexe getallen elke kwadratische vergelijking een complexe oplossing heeft. De formules van Cardano laten ook zien dat elke derde-graadsvergelijking met behulp van complexe getallen op te lossen is. Hoe is dat met hogere-graadsvergelijkingen? Moeten we weer nieuwe, niet complexe, getallen introduceren om te garanderen dat, bijvoorbeeld, elke achtste-graadsvergelijking een oplossing heeft? Het antwoord is verrassend simpel:

Stelling 22 (Hoofdstelling van de Algebra) *Elke n-de graadsvergelijking heeft altijd een complexe oplossing*

We geven het idee van een bewijs dat te vinden is in [5] en dat toegeschreven wordt aan Argand. In [5] wordt ook een aantal andere bewijzen besproken, gelardeerd met historische argumenten. De stelling geldt universeel voor vergelijkingen met complexe coëfficiënten. Het bewijs volgt uit de volgende twee stellingen die we zonder bewijs geven. Laten we even herinneren dat de functie

$$p(z) = a_n z^n + a_{n-1} z^{n-1} + \cdots + a_1 z + a_0$$

waarbij de coëfficiënten $a_n, a_{n-1}, \ldots, a_1, a_0$ complexe getallen zijn en $a_n \neq 0$ een *complex polynoom van graad n* heet. Als we $z = x + iy$ schrijven dan is $p(z)$ te schrijven als

$$\begin{aligned} p(z) &= a_n z^n + a_{n-1} z^{n-1} + \cdots + a_1 z + a_0 \\ &= a_n (x+iy)^n + a_{n-1} (x+iy)^{n-1} + \cdots + a_1 (x+iy) + a_0 \end{aligned}$$

We werken de haakjes uit; we krijgen dat $p(z)$ van de vorm

$$p(z) = p_1(x, y) + i p_2(x, y)$$

is, waarbij $p_1(x,y)$ het reële deel en $p_2(x,y)$ het imaginaire deel van $p(z)$ is. Het modulus van $p(z)$ is dan gelijk aan

$$\big|p(z)\big| = \sqrt{p_1(x,y)^2 + p_2(x,y)^2}.$$

Merk op dat $\big|p(z)\big|$ een continue reële functie is van twee variabelen.

Stelling 23 (Minimumstelling van Cauchy) *Zij $p(z)$ een complex polynoom van graad n. Dan is er een complex getal c waarin de functie $\big|p(z)\big|$ een minimum aanneemt, dat wil zeggen, voor elk complex getal z geldt*

$$\big|p(c)\big| \leq \big|p(z)\big|.$$

Stelling 24 (Ongelijkheid van Argand) *Zij $p(z)$ een complex polynoom van een positieve graad n. Voor elk complex getal c met $p(c) \neq 0$ is er een getal c' te vinden zó dat*

$$\big|p(c')\big| < \big|p(c)\big|.$$

Nu volgt de Hoofdstelling bijna onmiddelijk: Zij

$$a_n z^n + a_{n-1} z^{n-1} + \cdots + a_1 z + a_0 = 0 \tag{12}$$

een n-de graadsvergelijking en zij $p(z)$ het polynoom gelijk aan de linkerkant van de vergelijking. Volgens Stelling (23) is er een complex getal c waarin de functie $\big|p(z)\big|$ een minimum aanneemt. Als $p(c) \neq 0$ dan is volgens de Stelling (24) een c' te vinden met $\big|p(c')\big| < \big|p(c)\big|$. Maar dat kan niet omdat $p(z)$ in c een minimum heeft. Dus er moet gelden dat $p(c) = 0$. We hebben bewezen dat c een oplossing van (12) is.

We kunnen eveneens makkelijk een gevolg van de Hoofdstelling afleiden. Volgens de Hoofdstelling van de Algebra heeft de vergelijking (12) een complexe oplossing c. Merk op dat het polynoom $p(z)$ aan de linkerkant van (12) door $z - c$ deelbaar moet zijn, anders is de rest bij het staartdelen ongelijk aan 0 en dus $p(z)$ is te schrijven als

$$p(z) = (z - c) \cdot q(z) + r,$$

waarbij $r \neq 0$. Maar dit is onmogelijk: als $z = c$ krijgen we $p(c) = 0$ maar $(c - c) \cdot q(c) + r = r \neq 0$. Hieruit volgt dat $p(z) = (z - c) \cdot q(z)$ waarbij $q(z)$ een polynoom van graad $n - 1$ is. Als $n - 1 > 0$ kunnen we de Hoofdstelling op $q(z) = 0$ toepassen om weer een oplossing vinden. Zo kunnen we n stappen lang doorgaan. We hebben bewezen:

Stelling 25 *Elke n-de graadsvergelijking heeft precies n complexe oplossingen, waarbij elke k-voudige oplossing*[4] *k keer geteld moet worden.*

[4] We zeggen dat c een *k-voudige oplossing* van

$$a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0 = 0$$

is als het polynoom $p(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$ deelbaar is door $(x - c)^k$ maar niet door $(x - c)^{k+1}$.

Hoewel vergelijkingen van een even graad geen reële oplossingen hoeven te hebben — denk aan $x^2+1=0$ — heeft elke vergelijking van oneven graad wel een reële oplossing. Dit feit volgt uit de volgende stelling:

STELLING 26 *Neem aan dat c een complexe oplossing is van een n-de graadsvergelijking*

$$a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 = 0$$

met $a_n, a_{n-1}, \ldots, a_1, a_0$ reële getallen. Dan is de complex toegevoegde $\overline{c}$ van het getal c ook een oplossing van de vergelijking.

De stelling kunnen we als volgt aantonen. Neem aan dat c een complex getal is met

$$a_0 + a_1 c + \cdots + a_n c^n = 0.$$

Er volgt

$$\overline{a_0 + a_1 c + \cdots + a_n c^n} = \overline{0}.$$

Volgens Opgave 1 geldt dat $\overline{a_k} = a_k$ en $\overline{0} = 0$ dus

$$\overline{a_0 + a_1 c + \cdots + a_n c^n} = a_0 + a_1\overline{c} + \cdots + a_n\overline{c}^n.$$

Hieruit volgt dat $\overline{c}$ ook een oplossing is.

We zien dat complexe, niet reële, oplossingen in 'paren' komen. Er zijn dus altijd even veel complexe, niet reële, oplossingen (het getal 0 is ook even). Als n een oneven getal is blijven er, omdat er volgens Stelling (25) precies n oplossingen zijn, oneven veel reële oplossingen over. We hebben bewezen:

GEVOLG 1 *Elke vergelijking van oneven graad heeft tenminste één reële oplossing.*

VOORBEELD 5 We zoeken een (reële) derde-graadsvergelijking die 1 en $1-2i$ als oplossingen heeft. Volgens Stelling 26 is ook het getal $\overline{1-2i} = 1+2i$ een oplossing. Hieruit volgt dat een gezochte vergelijking te schrijven is als:

$$(x-1)\big(x-(1-2i)\big)\big(x-(1+2i)\big) = 0$$

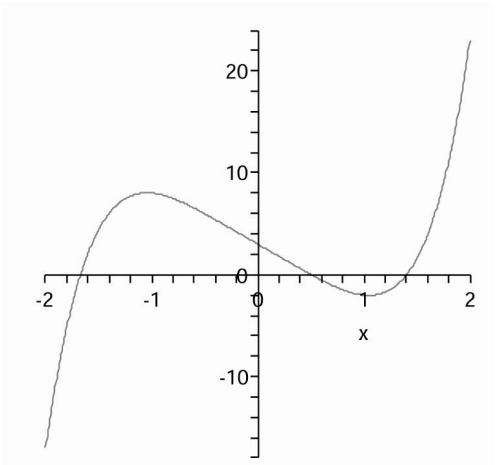
ofwel (na uitwerking van de linkerkant)

$$x^3 - 3x^2 + 7x - 5 = 0.$$

OPGAVE 7 Laat zien dat $x^{11}+21x^5+13x^3-156=0$ *precies* één reële oplossing heeft.

5. OPLOSSEN VAN n-DE MACHTSVERGELIJKINGEN IN RADIKALEN VOOR $n \geq 5$

We hebben al gezien dat er formules in radikalen bestaan voor oplossingen van eerste-, tweede-, derde- en vierde-graadsvergelijkingen. Als we een algemene vijfde-graadsvergelijking proberen op te lossen door slechts gebruik te maken



Figuur 4. *De grafiek van de functie* $f(x) = x^5 - 6x + 3$

van eindig vaak optellen, aftrekken, vermenigvuldigen, delen en k-de machtsworteltrekken is elke moeite tevergeefs: De jonge Noorse wiskundige Abel bewees in 1824 (hij was toen 21 jaar oud) dat er geen algemene formule voor oplossingen van vijfde-gaadsvergelijkingen bestaat die uitgedrukt kan worden in radikalen.

Het bewijs van de Stelling van Abel valt buiten het bestek van dit artikel maar een globaal idee ervan is wel te geven. In het bewijs wordt gebruik gemaakt van symmetrie in de formules. Die symmetrie is zichtbaar in de *abc*-formule: de $\pm$ geeft twee oplossingen. In de formule van Cardano zit drievoudige symmetrie dankzij de keuze-mogelijkheden 1, ω en ω^2. We hebben gezien dat door de eis $3uv = -p$ van de negen mogelijkheden er drie overblijven.

Wat Abel bewees is dat een potentiële algemene formule voor een vijfde-graadsvergelijking zoveel symmetrie moet bevatten dat deze 120 verschillende oplossingen van $a_5x^5 + \cdots + a_1x + a_0 = 0$ op *moet* leveren en dat is veel meer dan de vijf die Stelling (25) oplevert.

VOORBEELD 6 Een vijfde-graadsvergelijking die niet in radikalen op te lossen is is bijvoorbeeld $x^5 - 6x + 3 = 0$. Voor een bewijs verwijzen we naar [7]. Volgens de Hoofdstelling van de Algebra is er echter een reële oplossing en zoals uit de grafiek van de functie $f(x) = x^5 - 6x + 3$ blijkt (zie Figuur 6) zijn er in feite drie reële oplossingen, maar deze oplossingen zijn niet in radikalen uit te drukken.

Er zijn natuurlijk vijfde-graadsvergelijkinge die wel in radikalen kunnen worden opgelost: $x^5 + 2 = 0$ heeft $x = \sqrt[5]{2}$ als oplossing. De Franse wiskundige Évariste Galois heeft op heel jonge leeftijd (hij heeft zijn leven verloren bij een duel toen hij 21 jaar oud was) een antwoord gegeven op de vraag welke vergelijkingen wel en welke niet in radikalen kunnen worden opgelost. Ook zijn

resultaat valt buiten het bestek van dit artikel; er is een hele theorie (*Galois-theorie*) ontwikkeld om deze vraag te bestuderen. De sleutel is weer de notie van symmetrie. Hierbij bekijkt men bepaalde permutaties van de oplossingen van de vergelijking, namelijk die welke de rekenregels respecteren. Alleen zulke permutaties worden toegelaten; we leggen het idee uit aan de hand van het volgende voorbeeld.

Niels Henrik Abel (1802–1829)

Evariste Galois (1811–1832)

VOORBEELD 7 De oplossingen van de vergelijking $x^5 = 1$ zijn 1, α, α^2, α^3 en α^4, waarbij $\alpha = \cos\frac{2\pi}{5} + i\sin\frac{2\pi}{5}$, zie pagina 97. Een permutatie σ van $\{1, \alpha^2, \alpha^2, \alpha^3, \alpha^4\}$ is toegelaten als $\sigma(\alpha^j \cdot \alpha^k) = \sigma(\alpha^j) \cdot \sigma(\alpha^k)$ voor alle j en k met $j, k = 0, 1, 2, 3, 4$. Zo'n permutatie ligt vast zodra $\sigma(\alpha)$ bekend is. De vier mogelijkheden zijn dan $\sigma(\alpha) = \alpha$, $\sigma(\alpha) = \alpha^2$, $\sigma(\alpha) = \alpha^3$ en $\sigma(\alpha) = \alpha^4$. Dus slechts vier van de 120 mogelijke permutaties zijn toegelaten.

De structuur van de verzameling van toegelaten permutaties bepaalt of de oplossingen in radikalen uit te drukken zijn. Voor de vierde- en lagere graadsvergelijkingen is de structuur altijd van dien aard dat een oplossing in radikalen mogelijk is; er zijn vijfde-graadsvergelijkingen waarbij deze structuur te ingewikkeld is.

Achteraf kan men Abel's bewijs herinterpreteren: Voor de algemene vijfde-graadsvergelijkingen is *elke* permutatie van de oplossingen toegelaten. De structuur van de verzameling van alle permutaties verhinderd het bestaan van een algemene oplossingsformule.

Galois noemde zijn permutatieverzamelingen *groepen* en stond hiermee aan de wieg van de groepentheorie.

Literatuur

1. E.J. Barbeau, *Polynomials*, Springer-Verlag, New York, 1989.
2. G. Birkhoff and S. Mac Lane, *A Survey of Modern Algebra*, A K Peters, Massachusetts, 1997.
3. G. Cardano, *Ars Magna or the Rules of Algebra*, Dover, New York, 1993.
4. L. Childs, *A Concrete Introduction to Higher Algebra*, Springer-Verlag, New York, 1979.
5. H.-D. Ebbinghaus, H. Hermes, F. Hirzenbruch, M. Koecher, K. Mainzer, J. Neukirch, A. Prestel, R. Remmert, *Zahlen*, Springer-Verlag, Berlin, 1988.
6. J.-P. Escofier, *Galois Theory*, Springer Verlag, New York, 2001.
7. I. Stewart, *Galois Theory*, Chapman & Hall Mathematics, London, 2004.
8. J.-P. Tignol, *Galois' theory of algebraic equations*, Longman Scientific & Technical, 1988.

Vakantiecursus 2005

pp. 109 – 120

Complexe getallen en Fourier-theorie

J. van de Craats
Open Universiteit
Universiteit van Amsterdam
e-mail: jcr@euronet.nl

1. Introductie

Toen Napoleon Bonaparte zich op 19 mei 1798 met een leger van veertigduizend man te Toulon inscheepte voor een grote expeditie naar Egypte, liet hij zich vergezellen door prominente kunstenaars en wetenschappers, onder wie de wiskundigen Gaspard Monge (1746–1818) en Jean-Baptiste Joseph Fourier (1768–1830). Na een snelle verovering van Egypte op de Engelsen benoemde Bonaparte Fourier, die getoond had niet alleen over wetenschappelijke maar ook over bestuurlijke kwaliteiten te beschikken, tot gouverneur van het zuidelijke deel van dat land. Napoleons onderneming kreeg echter met tegenslagen te kampen: op 1 augustus 1798 vernietigde de Engelse admiraal Nelson de Franse vloot op de rede van Aboekir zodat de Fransen in Egypte opgesloten zaten. In 1799 keerde Napoleon met een groep getrouwen naar Frankrijk terug met achterlating van een bezettingsleger, dat echter in 1801 moest capituleren voor een gezamenlijke strijdmacht van Engelsen en Turken.

***Figuur 1**. Jean-Baptiste Joseph Fourier (1768–1830)*

Ook Fourier repatrieerde naar Frankrijk waar hij in 1802 tot prefect van het district Grenoble benoemd werd, een functie die hij tot 1815 zou blijven vervullen. In 1808 kreeg hij de titel van baron en in 1816 werd hij lid van de Académie des Sciences. Daarna wijdde hij zich geheel aan de wetenschap.

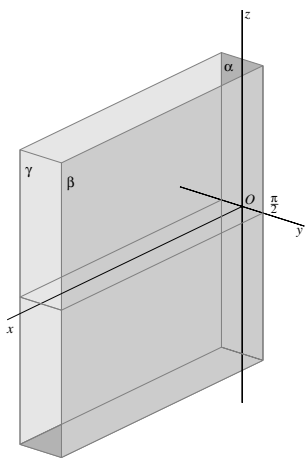
De publicatie van zijn hoofdwerk *Théorie analytique de la chaleur* in 1822 was echter lang tegengehouden door invloedrijke wiskundigen die vonden dat zijn baanbrekende ideeën niet exact genoeg geformuleerd, laat staan bewezen waren. Toch zou dit boek een revolutie teweegbrengen in de wiskunde van de negentiende en de twintigste eeuw en haar toepassingen. Fourieranalyse is een van de basistechnieken in de theorie van signalen en systemen geworden, en iedereen die een tv-toestel, een cd-speler of een mobiele telefoon bezit, maakt indirect gebruik van resultaten waarvoor Fourier de grondslag heeft gelegd.

2. De *Théorie de la chaleur*

Fourier hield zich echter nauwelijks met signaaltheorie bezig, maar veeleer met het vraagstuk hoe warmte door vaste lichamen stroomt. Daarnaar was al eerder onderzoek gedaan, en bekend was dat de warmtestroom door een homogeen isotroop lichaam gehoorzaamt aan de *diffusievergelijking*

$$\frac{\partial v}{\partial t} = C\left(\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} + \frac{\partial^2 v}{\partial z^2}\right)$$

voor een zekere constante C. Hierin is $v = v(x, y, z, t)$ de temperatuur van het lichaam in het punt (x, y, z) op tijdstip t. Een van de voorbeelden die Fourier



***Figuur 2**. De door Fourier bestudeerde oneindige plaat met een temperatuur van honderd graden in vlak α en nul graden in de vlakken β en γ*

bekeek, was een geïdealiseerd lichaam dat de vorm heeft van een oneindige plaat die in $\mathbb{R}^3$ gegeven wordt door

$$\{(x, y, z) \in \mathbb{R}^3 \mid x \geq 0, -\frac{\pi}{2} \leq y \leq \frac{\pi}{2}, -\infty < z < \infty\}$$

In het bijzonder vroeg hij zich af wat de uiteindelijke stationaire temperatuurverdeling in de plaat zou zijn wanneer het vlakdeel $\alpha = \{x = 0, -\frac{\pi}{2} \leq y \leq \frac{\pi}{2}\}$

door kokend water steeds op honderd graden Celsius gehouden werd, terwijl de temperatuur van de halfvlakken $\beta = \{x \geq 0, y = -\frac{\pi}{2}\}$ en $\gamma = \{x \geq 0, y = \frac{\pi}{2}\}$ met behulp van ijsblokjes op nul graden Celsius wordt gefixeerd. Er zal zich dan op den duur een stationaire warmtestroom van α naar β en γ instellen waarbij de temperatuur $v(x, y, z, t)$ niet meer van t en van z afhangt. Door v zo te normeren dat $v = 0$ met nul graden overeenkomt en $v = 1$ met honderd graden verkreeg Fourier een tweedimensionaal randwaardenprobleem voor de differentiaalvergelijking

$$\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} = 0 \tag{1}$$

op het gebied

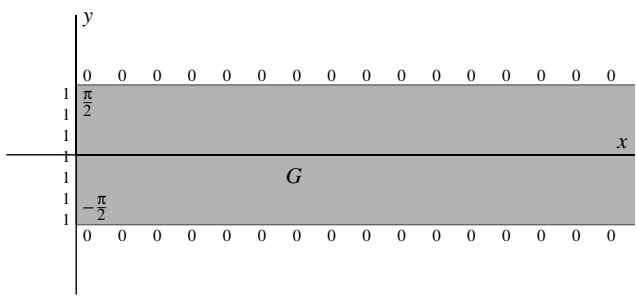
$$G = \{(x, y) \in \mathbb{R}^2 \mid x \geq 0, -\frac{\pi}{2} \leq y \leq \frac{\pi}{2}\}$$

met de volgende randvoorwaarden:

$$v(0, y) = 1 \quad \text{als } -\frac{\pi}{2} < y < \frac{\pi}{2}; \tag{2}$$

$$v(x, -\frac{\pi}{2}) = v(x, \frac{\pi}{2}) = 0 \quad \text{als } x \geq 0; \tag{3}$$

$$\lim_{x \to \infty} v(x, y) = 0 \quad \text{als } -\frac{\pi}{2} < y < \frac{\pi}{2}. \tag{4}$$



Figuur 3. *Het tweedimensionale randwaardenprobleem*

Fourier beredeneerde dat er oplossingen van de vorm $v(x, y) = F(x)f(y)$ moeten zijn, en daarvoor levert differentiaalvergelijking (1) na substitutie en herleiding

$$\frac{1}{F(x)} \frac{\partial^2 F(x)}{\partial x^2} = -\frac{1}{f(y)} \frac{\partial^2 f(y)}{\partial y^2},$$

waarin het linkerlid onafhankelijk van y, en het rechterlid onafhankelijk van x is. Wegens de gelijkheid moeten beide leden dus constant zijn, zeg m, dus

$$\frac{\partial^2 F}{\partial x^2} - mF(x) = 0 \quad \text{en} \quad \frac{\partial^2 f}{\partial y^2} + mf(y) = 0.$$

De reële oplossingen hiervan zijn e-machten of combinaties van sinussen en cosinussen, afhankelijk van het teken van m. Uit randvoorwaarde (4) kun je gemakkelijk afleiden dat $m > 0$ moet zijn, dus $F(x) = c\,\mathrm{e}^{-\sqrt{m}x}$ voor een zekere constante c. Fourier merkte vervolgens op dat de keuze $m = (2k+1)^2$ voor $k = 0, 1, 2, 3, \ldots$ speciale oplossingen $f(y)$ geeft van de vorm

$$f(y) = \cos\sqrt{m}y = \cos(2k+1)y,$$

waarvoor geldt dat $v(x,y) = F(x)f(y) = c\,\mathrm{e}^{-(2k+1)x}\cos(2k+1)y$ niet alleen aan randvoorwaarde (4), maar ook aan (3) voldoet. Tot op dit moment van de afleiding volgde Fourier nog gebaande paden. Maar het probleem is randvoorwaarde (2), waar geen van de gevonden oplossingen aan voldoet. Fourier pakte het dan als volgt aan. Hij schrijft: *'Nu is het echter gemakkelijk een nog algemenere functie voor v op te stellen. Want omdat $2k+1$ een willekeurig oneven positief getal mag zijn, en de differentiaalvergelijking lineair en homogeen is, zal ook*

$$\begin{aligned} v(x,y) &= \sum_{k=0}^{\infty} a_{2k+1}\mathrm{e}^{-(2k+1)x}\cos(2k+1)y \\ &= a_1\mathrm{e}^{-x}\cos y + a_3\mathrm{e}^{-3x}\cos 3y + a_5\mathrm{e}^{-5x}\cos 5y + \cdots \end{aligned}$$

een functie zijn die aan (1), (3) en (4) voldoet. Om bovendien nog aan (2) te voldoen, moeten de constanten a_1, a_3, a_5, ... zo bepaald worden dat voor alle $-\frac{\pi}{2} < y < \frac{\pi}{2}$ aan de vergelijking

$$1 = \sum_{k=0}^{\infty} a_{2k+1}\cos(2k+1)y = a_1\cos y + a_3\cos 3y + a_5\cos 5y + \cdots$$

voldaan is.' Daarmee was de eerste Fourierreeks geboren!

Het berekenen van die coëfficiënten a_k was voor Fourier echter geen sinecure. Na een bladzijdenlange rekenpartij kwam hij tot de formule $a_k = (-1)^k\frac{4}{\pi(2k+1)}$ dus tot

$$1 = \frac{4}{\pi}\left(\cos y - \frac{1}{3}\cos 3y + \frac{1}{5}\cos 5y - \frac{1}{7}\cos 7y \cdots\right) \quad \text{als} \quad -\frac{\pi}{2} < y < \frac{\pi}{2}. \tag{5}$$

Dat is een verrassend resultaat. De geldigheid ervan wordt echter ondersteund doordat een aantal substituties leiden tot bekende formules. Voor $y = 0$ krijg je bijvoorbeeld de bekende alternerende harmonische reeks van Leibniz, namelijk

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots$$

en voor $y = \frac{\pi}{4}$ vind je een formule die ook al door Euler op een heel andere wijze was afgeleid, namelijk

$$\pi = 2\sqrt{2}\left(1 + \frac{1}{3} - \frac{1}{5} - \frac{1}{7} + \frac{1}{9} + \frac{1}{11} - \frac{1}{13} - \frac{1}{15} + \cdots\right).$$

Voor het volgende, ook door Fourier genoemde substitutieresultaat citeer ik Fourier vrijwel letterlijk: *'Vermenigvuldigt men beide leden van vergelijking (5) met $\frac{\pi}{4}dy$ en integreert men dit vervolgens van $y = 0$ tot $y = y$, dan ontstaat*

$$\frac{\pi}{4}y = \sin y - \frac{1}{3^2}\sin 3y + \frac{1}{5^2}\sin 5y - \frac{1}{7^2}\sin 7y + \cdots$$

Wanneer men hierin $y = \frac{\pi}{2}$ substitueert, ontstaat een andere beroemde formule van Euler, namelijk

$$\frac{\pi^2}{8} = 1 + \frac{1}{3^2} + \frac{1}{5^2} + \frac{1}{7^2} + \cdots \tag{6}$$

Ik voeg daar zelf nog aan toe dat je hieruit direct de waarde van een nog beroemdere Euler-som kan afleiden, namelijk de waarde van de zètafunctie in het punt 2

$$\zeta(2) = \sum_{k=0}^{\infty} \frac{1}{k^2} = 1 + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \cdots$$

Euler had die zètafunctie als volgt gedefinieerd:

$$\zeta(s) = \sum_{k=0}^{\infty} \frac{1}{k^s} = 1 + \frac{1}{2^s} + \frac{1}{3^s} + \frac{1}{4^s} + \cdots \quad \text{voor } s > 1.$$

Met het integraalkenmerk kun je direct verifiëren dat die reeks voor $s > 1$ convergeert. De exacte berekening van de waarde van $\zeta(2)$ was een van de vele grote wapenfeiten van Euler geweest. In 1859 zou Riemann de zètafunctie uitbreiden tot een analytische functie op het gehele complexe vlak met uitzondering van $s = 1$. Zijn vermoeden dat de niettriviale nulpunten van deze functie allemaal voldoen aan $\mathrm{Re}(s) = \frac{1}{2}$ is inmiddels een van de belangrijkste open problemen in de wiskunde geworden.

De welbekende afleiding van $\zeta(2)$ uit formule (6) gaat als volgt

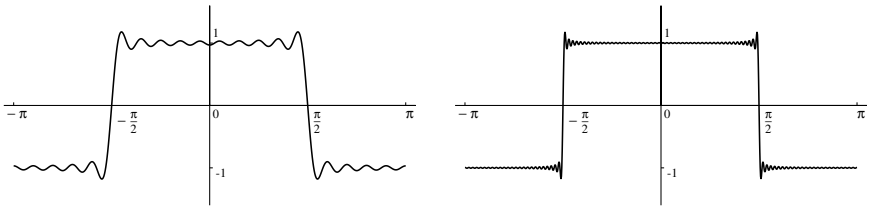
$$\frac{1}{2^2} + \frac{1}{4^2} + \frac{1}{6^2} + \frac{1}{8^2} + \cdots = \frac{1}{4}\left(1 + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \cdots\right) = \frac{1}{4}\zeta(2)$$

en dus is

$$\frac{\pi^2}{8} = 1 + \frac{1}{3^2} + \frac{1}{5^2} + \frac{1}{7^2} + \cdots = \zeta(2) - \frac{1}{4}\zeta(2) = \frac{3}{4}\zeta(2),$$

waaruit volgt dat $\zeta(2) = \dfrac{\pi^2}{6}$.

Het lijdt geen twijfel dat dit soort resultaten Fourier sterkten in zijn overtuiging dat hij met zijn methode op het juiste spoor zat. Zou hij al over computeralgebra beschikt hebben, dan was die overtuiging zeker nog verder vergroot door plotjes van de partiële sommen van zijn reeks. Hieronder geven we de grafiek van $S_n(y) = \sum_{k=0}^{n}(-1)^k \frac{4}{\pi(2k+1)}\cos(2k+1)y$ voor $n = 10$ en $n = 50$. Omdat alle termen van de reeks periodiek zijn met een periode 2π, zijn de partiële sommen dat ook. Op het interval $-\frac{\pi}{2} < y < \frac{\pi}{2}$ lijken



Figuur 4. *De partiële sommen $S_{10}(y)$ en $S_{50}(y)$ van de Fourierreeks*

de partiële sommen inderdaad naar 1 te convergeren. Hoewel? Wat zijn die rare bergpuntjes vlak boven $-\frac{\pi}{2}$ en vlak onder $\frac{\pi}{2}$? Misschien is het maar goed dat Fourier ze niet gezien heeft; misschien hadden ze zijn zelfvertrouwen ondermijnd. Dat 'doorschietverschijnsel' werd in 1848 voor het eerst opgemerkt door Wilbraham, wiens werk echter in de vergetelheid raakte. In 1898 verscheen in *Nature* echter een artikel van de fysicus Michaelson waarin hij er aandacht voor vroeg. In 1899 verklaarde Gibbs het verschijnsel door te laten zien dat het altijd optreedt bij een Fourierreeks die een periodieke functie representeert met sprongdiscontinuïteiten. De 'doorschiethoogte' nadert op den duur tot een vaste fractie van ongeveer 9 procent van de spronggrootte, onafhankelijk van de aard van de functie. Wel schuiven die doorschiettoppen steeds dichter naar het discontinuïteitspunt toe waardoor de puntsgewijze convergentie van de reeks in alle punten waar de functie continu is – hier het interval $(-\frac{\pi}{2}, \frac{\pi}{2})$ – niet in gevaar komt. Maar van uniforme convergentie op dat interval is dus geen sprake!

De Fourierreeks van een willekeurige functie

Aan de hand van de tekst van Fourier zelf zijn we het terrein van de trigonometrische reeksen binnengeleid. Fourier bepaalde een cosinusreeks voor de functie die 1 is op het interval $(-\frac{\pi}{2}, \frac{\pi}{2})$ en 0 in de beide eindpunten. Maar buiten dat interval stelt die reeks ook een functie voor, en wel een periodieke functie met periode 2π. Het is een 'blokfunctie' die nul is in alle punten van de vorm $\frac{\pi}{2} + k\pi$, 1 op alle intervallen van de vorm $(-\frac{\pi}{2} + 2k\pi, \frac{\pi}{2} + 2k\pi)$ en -1 op alle intervallen van de vorm $(\frac{\pi}{2} + 2k\pi, \frac{3\pi}{2} + 2k\pi)$ (k geheel).

Nadat hij dit voorbeeld behandeld had, wierp Fourier de vraag op of men ook een willekeurige periodieke functie $f(t)$ – stel voor de eenvoud maar weer dat de periode van die functie 2π is – als een trigonometrische reeks kan schrijven. Daniel Bernoulli had dit al beweerd in het kader van zijn onderzoek naar trillende snaren. Fourier is het met hem eens, en denkt het ook te kunnen bewijzen. Hij beweert dus dat er bij zo'n periodieke functie $f(t)$ altijd een reeks van de vorm

$$A_0 + \sum_{n=1}^{\infty} (a_n \cos nt + b_n \sin nt)$$

bestaat die $f(t)$ representeert. Tegenwoordig noemen we zo'n reeks een Fourierreeks. Fourier staaft zijn bewering door expliciete formules te geven voor de coëfficiënten A_0, a_n en b_n van de reeks, namelijk

$$A_0 = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(t)\,dt; \tag{7}$$

$$a_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)\cos nt\,dt; \tag{8}$$

$$b_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)\sin nt\,dt. \tag{9}$$

Het is een aardige opgave om te verifiëren dat deze formules in het hierboven behandelde voorbeeld van een blokfunctie met periode 2π inderdaad de gevonden reeks opleveren, dus dat in dat geval geldt dat $A_0 = 0$, $b_n = 0$ voor alle n en dat $a_{2k} = 0$ en $a_{2k+1} = \frac{4}{\pi(2k+1)}$ voor alle k. Bij de berekening van a_{2k+1} kun je wegens de periodiciteit van f het integratie-interval naar believen verschuiven, zolang de lengte ervan maar 2π blijft. Hier is $[-\frac{\pi}{2}, \frac{3\pi}{2}]$ een handige keuze.

We zullen nu niet laten zien hoe Fourier die integraalformules gevonden heeft. Wel merken we op dat ze direct een aantal vragen oproepen. Bijvoorbeeld: bestaan die integralen wel voor een willekeurige periodieke functie? Zo ja, convergeert de ermee geconstrueerde Fourierreeks dan voor alle t? En zo ja, stelt de som van die reeks dan ook echt in alle punten de oorspronkelijke functie $f(t)$ voor? Tot ver in de twintigste eeuw hebben dit soort vragen wiskundigen beziggehouden. Eerst onder andere Dirichlet en Riemann (de Riemann-integraal en de Riemann-sommen zijn geïntroduceerd in een artikel van Riemann over Fourierreeksen), later Cantor, Weierstrass, Lebesgue en vele anderen.

Voor nette periodieke functies, bijvoorbeeld functies die continu of stuksgewijs continu zijn, lukte het al vrij snel om de zaken tot klaarheid te brengen, maar het is niet al te moeilijk 'pathologische' functies te bedenken waar het helemaal misloopt. Wij zullen deze problematiek hier verder laten rusten en ons, net als de meeste toepassers, beperken tot nette functies. Ons interesseert hier vooral de vraag hoe je zo'n Fourierreeks vindt, en wat ervan de belangrijkste eigenschappen zijn. Daarbij blijkt, zoals zo vaak in de wiskunde, dat de eenvoudigste en overzichtelijkste weg naar reële resultaten door het complexe vlak loopt.

De complexe Fourierreeks

Tot nu toe hadden we voor de eenvoud de periode op 2π gesteld. Het algemene geval van een functie $f(t)$ met periode T kan hierop gemakkelijk worden teruggebracht via de substitutie $t = (2\pi/T)t'$. Men noemt $\omega = 2\pi/T$ dan de *hoekfrequentie* en de formules (7), (8) en (9) voor de Fouriercoëfficiënten worden (met weglating van de accenten)

$$A_0 = \frac{1}{T}\int_{T} f(t)\,dt; \tag{10}$$

$$a_n = \frac{2}{T}\int_T f(t)\cos n\omega t\,dt; \tag{11}$$

$$b_n = \frac{2}{T}\int_T f(t)\sin n\omega t\,dt, \tag{12}$$

waarbij de T onder het integraalteken aangeeft dat er over een 'volle periode', dat wil zeggen een interval van lengte T, geïntegreerd moet worden; waar dat interval begint, maakt vanwege de periodiciteit van $f(t)$ niets uit. De Fourierreeks zelf krijgt nu de vorm

$$A_0 + \sum_{n=1}^{\infty}(a_n\cos n\omega t + b_n\sin n\omega t).$$

Met behulp van de bekende relaties van Euler

$$\begin{aligned} \mathrm{e}^{i\varphi} &= \cos\varphi + i\sin\varphi;\\ \cos\varphi &= \frac{\mathrm{e}^{i\varphi}+\mathrm{e}^{-i\varphi}}{2};\\ \sin\varphi &= \frac{\mathrm{e}^{i\varphi}-\mathrm{e}^{-i\varphi}}{2i} \end{aligned}$$

kunnen we de Fourierreeks herschrijven als

$$\sum_{n=-\infty}^{\infty}\alpha_n\,\mathrm{e}^{in\omega t} = \cdots+\alpha_{-2}\mathrm{e}^{-2i\omega t}+\alpha_{-1}\mathrm{e}^{-i\omega t}+\alpha_0+\alpha_1\mathrm{e}^{i\omega t}+\alpha_2\mathrm{e}^{2i\omega t}+\cdots, \tag{13}$$

waarin

$$\begin{aligned} \alpha_0 &= A_0;\\ \alpha_n &= \frac{1}{2}(a_n - ib_n) \quad (n\geq 1);\\ \alpha_{-n} &= \frac{1}{2}(a_n + ib_n) \;=\; \overline{\alpha_n} \quad (n\geq 1). \end{aligned}$$

Wanneer we aannemen dat de Fourierreeks (13) inderdaad de functie $f(t)$ voorstelt, dat de hieronder beschreven integralen bestaan en dat we sommatie en integratie mogen verwisselen (dat zijn allemaal voorwaarden die vallen onder de vage veronderstelling dat $f(t)$ een 'nette' functie is), dan geldt voor elk geheel getal k dat

$$\begin{aligned} \int_T f(t)\,\mathrm{e}^{-ik\omega t}\,dt &= \int_T\left(\sum_{n=-\infty}^{\infty}\alpha_n\,\mathrm{e}^{in\omega t}\right)\mathrm{e}^{-ik\omega t}\,dt\\ &= \sum_{n=-\infty}^{\infty}\alpha_n\left(\int_T \mathrm{e}^{i(n-k)\omega t}\,dt\right). \end{aligned}$$

De cruciale opmerking is nu dat $\int_T \mathrm{e}^{i(n-k)\omega t}\,dt = 0$ voor alle n behalve $n=k$. Immers, op grond van Eulers relaties is

$$\begin{aligned} \int_T \mathrm{e}^{i(n-k)\omega t}\,dt &= \int_T \cos i(n-k)\omega t\,dt + i\int_T \sin i(n-k)\omega t\,dt\\ &= \int_T \cos i(n-k)\frac{2\pi}{T}t\,dt + i\int_T \sin i(n-k)\frac{2\pi}{T}t\,dt \end{aligned}$$

en omdat $n-k$ een geheel getal is en we over een interval van lengte T (dat wil zeggen $n-k$ volle periodes) integreren, zijn beide integralen nul. De enige uitzondering is het geval dat $n=k$, want dan is $\mathrm{e}^{i(n-k)\omega t}=\mathrm{e}^0=1$, en dus is dan $\int_T \mathrm{e}^{i(n-k)\omega t}\,dt=\int_T 1\,dt=T$. We concluderen dat

$$\alpha_k=\frac{1}{T}\int_T f(t)\,\mathrm{e}^{-ik\omega t}\,dt, \tag{14}$$

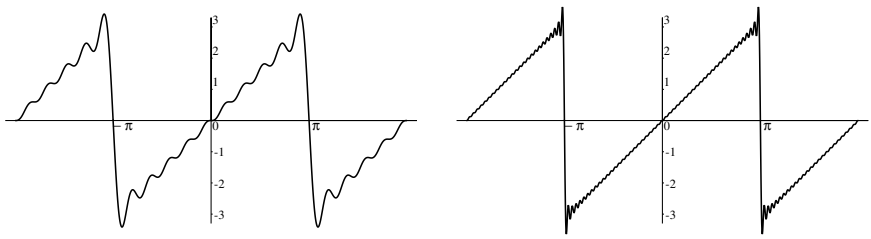
waarmee we een eenvoudige en overzichtelijke formule gevonden hebben voor de Fouriercoëfficiënten α_k. Dat zijn echter in veel gevallen wel *complexe* getallen. De som van de bijbehorende complexe Fourierreeks is echter een reële functie van t, mits natuurlijk $f(t)$ zelf een nette, reële functie is. Men kan bewijzen dat voor stuksgewijs continue functies de Fourierreeks naar $f(t)$ convergeert in alle continuïteitspunten van f, en naar het gemiddelde van de linker- en de rechterlimiet van $f(t)$ in alle sprongpunten.

We berekenen als voorbeeld de complexe Fourierreeks van de functie $f(t)$ met periode $T=2\pi$ (dus $\omega=1$) die op het interval $(-\pi,\pi)$ gegeven wordt door $f(t)=t$. Dit is een soort 'zaagtand-functie'. Daarvoor is $\alpha_0=0$ en voor $k\neq 0$ geeft partiële integratie

$$\begin{aligned}\alpha_k &= \frac{1}{2\pi}\int_{-\pi}^{\pi} t\,\mathrm{e}^{-ikt}dt=\frac{-1}{2\pi ik}\int_{-\pi}^{\pi} t\,d(\mathrm{e}^{-ikt})\\ &= \left[\frac{-1}{2\pi ik}t\,\mathrm{e}^{-ikt}\right]_{-\pi}^{\pi}+\frac{1}{2\pi ik}\int_{-\pi}^{\pi}\mathrm{e}^{-ikt}\,dt\\ &= \frac{-1}{2ik}(\mathrm{e}^{-ik\pi}+\mathrm{e}^{ik\pi})=\frac{i}{k}\cos k\pi=\frac{i(-1)^k}{k}.\end{aligned}$$

De reële Fourierreeks is dus een zuivere sinusreeks, en wel

$$f(t)=\sum_{k=1}^{\infty}\frac{2}{k}(-1)^{k+1}\sin kt=2\left(\sin t-\frac{1}{2}\sin 2t+\frac{1}{3}\sin 3t-\frac{1}{4}\sin 4t+\cdots\right)$$



Figuur 5. *De partiële sommen $S_{10}(y)$ en $S_{50}(y)$ van de Fourierreeks van de zaagtandfunctie. Let weer op het 'doorschietverschijnsel' rond de sprongpunten*

Het spectrum

In het algemeen is, zoals we al gezien hebben, α_k een complex getal. We schrijven het in de *polaire vorm*

$$\alpha_k = |\alpha_k|\, \mathrm{e}^{i\varphi_k}$$

Wegens $\alpha_{-k} = \overline{\alpha_k}$ geldt dat $|\alpha_k| = |\alpha_{-k}|$ en $\varphi_{-k} = -\varphi_k (\mathrm{mod}\, 2\pi)$. Vullen we dit in de complexe Fourierreeks in, dan kunnen we die schrijven als

$$\begin{aligned}
\sum_{n=-\infty}^{\infty} \alpha_n\, \mathrm{e}^{in\omega t} &= \sum_{n=-\infty}^{\infty} |\alpha_n|\, \mathrm{e}^{i(n\omega t+\varphi_n)} \\
&= \alpha_0 + \sum_{n=1}^{\infty} |\alpha_n| \left(\mathrm{e}^{i(n\omega t+\varphi_n)} + \mathrm{e}^{-i(n\omega t+\varphi_n)}\right) \\
&= \alpha_0 + \sum_{n=1}^{\infty} 2|\alpha_n| \cos(n\omega t + \varphi_n) \\
&= \alpha_0 + \sum_{n=1}^{\infty} A_n \cos(n\omega t + \varphi_n).
\end{aligned}$$

We hebben hier $A_n = 2|\alpha_n|$ $(n \geq 1)$ gesteld. Wanneer $f(t)$ een periodiek geluidssignaal voorstelt (dat wil zeggen een muzikale toon), is α_0 het gemiddelde niveau (nulniveau) van het geluidssignaal, $A_1 \cos(\omega t+\varphi_1)$ de grondtoon van het signaal en $A_n \cos(n\omega t + \varphi_n)$ de n-de boventoon. De frequentie is $\nu = \frac{1}{T} = \frac{\omega}{2\pi}$ Hertz. Dit is ook de frequentie van de grondtoon. De n-de boventoon heeft frequentie $n\nu$. Men noemt $A_n = 2|\alpha_n|$ de amplitude, en φ_n de fase(hoek) van de n-de boventoon. In het algemeen heet de rij $\{\alpha_n\}_{n=-\infty}^{\infty}$ het (complexe) *spectrum* van $f(t)$. Voor nette periodiek functies geldt dat het spectrum de functie via de Fourierreeks volledig bepaalt: zo'n functie ligt volledig vast als men zijn periode en zijn spectrum geeft. De rijen $\{A_n\}_{n=0}^{\infty}$ en $\{\varphi_n\}_{n=0}^{\infty}$ noemt men respectievelijk het *amplitudespectrum* en het *fasespectrum*.

Fourierintegralen

Kun je het bij Fourierreeksen met enige moeite nog wel zonder complexe getallen stellen, bij de *Fourier-integralen* is dat praktisch uitgesloten. Fourierintegralen vormen het analogon van Fourierreeksen bij niet-periodieke functies. Je kunt ze via een limietovergang intuïtief uit Fourierreeksen afleiden, maar daarvoor ontbreekt ons hier de tijd en de ruimte. In plaats daarvan laten we de definitie gewoon uit de lucht vallen.

Bij een gegeven 'nette' functie $f(t)$ (we laten weer in het midden wat men precies onder 'net' mag verstaan) wordt de *Fourier-getransformeerde* $\hat{f}(\omega)$ gedefinieerd door

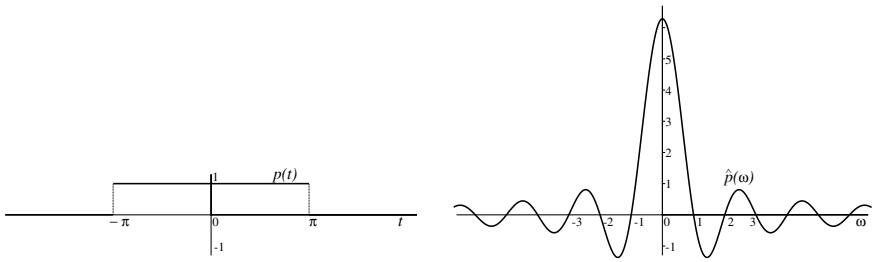
$$\hat{f}(\omega) = \int_{-\infty}^{\infty} f(t)\, \mathrm{e}^{-i\omega t} dt.$$

Deze functie speelt een rol die vergelijkbaar is met die van de Fourier-coëfficiënten α_k van de Fourierreeks van een periodieke functie. Men noemt $\hat{f}(\omega)$ ook wel de *spectrale dichtheid* van $f(t)$.

In zekere zin is het gebruik van ω in deze notatie ongelukkig, want de variabele ω speelt hier een andere rol dan de ω bij de Fourierreeksen. Daar was $\omega = \frac{2\pi}{T}$, maar hier is ω een variabele die de gehele $\mathbb{R}$ doorloopt, net als de variabele t bij de functie $f(t)$. In de toepassingen spreekt men vaak over het t-domein (of het tijddomein) en het ω-domein (of het frequentiedomein). De Fouriertransformatie zet dan een functie $f(t)$ in het tijddomein over in een functie $\hat{f}(\omega)$ in het frequentiedomein. Het verrassende is dat er bij deze overzetting geen informatie verloren gaat, althans wanneer de functies zich netjes gedragen. We hebben al gezien hoe een functie van het t-domein naar het ω-domein wordt getransformeerd. Er is ook een inverse transformatie die functies uit het ω-domein weer naar het t-domein terughaalt, en de formule waarmee dit gebeurt lijkt erg op die van de gewone Fourier-transformatie:

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{f}(\omega)\, \mathrm{e}^{i\omega t} d\omega.$$

Dat is in zekere zin het analogon van de Fourierreeks voor periodieke functies, die immers een functie schrijft als een oneindige som van Fouriercoëfficiënten en complexe e-machten. Ook hierover is veel meer te vertellen dan hier mogelijk is. Ik volsta ermee op te merken dat de Fourier-integralen, meer nog dan de Fourierreeksen, voor de toepassingen van eminent belang zijn.



Figuur 6. *De functie $p(t)$ in het tijddomein en de Fouriergetransformeerde $\hat{p}(\omega)$ in het frequentiedomein*

Als voorbeeld berekenen we de spectrale dichtheid $\hat{p}(\omega)$ van het signaal $p(t)$ dat gegeven wordt door

$$p(t) = \begin{cases} 1 & \text{als } |t| \leq \pi \ ; \\ 0 & \text{anders .} \end{cases}$$

In dat geval is

$$\begin{aligned} \hat{p}(\omega) &= \int_{-\infty}^{\infty} p(t)\, \mathrm{e}^{-i\omega t} dt = \int_{-\pi}^{\pi} \mathrm{e}^{-i\omega t} dt \\ &= \left[\frac{-1}{i\omega} \mathrm{e}^{-i\omega t} \right]_{t=-\pi}^{\pi} = \frac{-1}{i\omega} \left(\mathrm{e}^{-i\omega\pi} - \mathrm{e}^{i\omega\pi} \right) = \frac{2 \sin \pi\omega}{\omega}. \end{aligned}$$

De omkeerformule geeft nu

$$p(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \hat{p}(\omega)\,\mathrm{e}^{i\omega t}d\omega$$

en in het bijzonder is (substitueer $t = 0$)

$$p(0) = 1 = \frac{1}{2\pi}\int_{-\infty}^{\infty} \hat{p}(\omega)\,\mathrm{e}^{0}d\omega = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{2\sin\pi\omega}{\omega}d\omega$$

en hieruit volgt (stel $x = \pi\omega$ en merk op dat $\frac{\sin x}{x}$ een even functie is) het beroemde resultaat

$$\int_0^{\infty} \frac{\sin x}{x}dx = \frac{\pi}{2}.$$

Het belang van de Fouriertransformatie is alleen maar toegenomen door de komst van de computer. Die maakte het in principe mogelijk de integralen waarmee de Fouriertransformatie gedefinieerd is ook numeriek te berekenen via een zogenaamde *Discrete Fouriertransformatie* (DFT), maar dit werd pas echt doenlijk toen de *Fast Fourier Transform* (FFT) ten tonele verscheen, een buitengewoon efficiënt algoritme om van het discrete tijddomein naar het discrete frequentiedomein over te stappen en omgekeerd. Al deze ontwikkelingen hebben het mogelijk gemaakt zowel continue als discrete signalen in de beide domeinen te analyseren en te bewerken, met schier onbegrensde toepassingsmogelijkheden. En bij al die toepassingen zijn complexe getallen een onontbeerlijk hulpmiddel gebleken.

Verder lezen

Het aantal elementaire inleidingen in de Fourieranalyse is legio. In het Nederlands kan ik voor een eerste kennismaking deel 4, *Fourier-theorie en systeemtheorie* van de serie *Voortgezette wiskunde* van A. Kaldewaij en J. van Tiel aanbevelen (Bohn, Scheltema & Holkema, Utrecht/Antwerpen 1983, ISBN 90-3130575-8). Dit boek geeft een goede inleiding in de toepassingen in onder andere de electrotechniek op hts-niveau.

Zeer gedegen, met ook aandacht voor alle wiskundige finesses, is het boek dat oorspronkelijk als cursusboek voor de Open Universiteit werd geschreven, en dat later ook afzonderlijk in de handel is gebracht: R.J. Beerends, H.G. ter Morsche, J.C. van den Berg, E.M. van der Vrie: *Fourier- en Laplacetransformaties* (Educaboek, Culemborg, 1993, ISBN 90-11-021096).

CWI Vakantiecursus 2005

pp. 121 – 136

Optimalisatie in financiering, economie en wiskunde: welke toepassingen zijn overtuigend?

J. Brinkhuis
Econometrisch Instituut
Erasmus Universiteit Rotterdam
e-mail: brinkhuis@few.eur.nl

1. Toepassingen die door de mand vallen

Provocerende tentamenopgave

Ik heb gisteren mijn laatste college niet-lineaire optimalisering voor tweedejaars econometriestudenten gegeven. Eindelijk tijd om aan mijn tekst voor de zomercursus te beginnen. De titel van mijn lezing is bijna identiek aan die van een opgave uit het tentamen na afloop van mijn vorige cursus. Mijn college gisteren stond uiteraard in het teken van het komende tentamen en ik heb de studenten beloofd dezelfde vraag weer te zullen stellen. De precieze vraag is: welke toepassing van optimalisering vind je het meest overtuigend en welke het minst, en waarom? Dat ik wilde weten welke toepassingen het minst overtuigde, trok vooral de aandacht. 'Dat is zeker voor de tweede druk van uw boek[1]; dan gaat u die toepassingen zeker weglaten.' Daar zat wel wat in. Bij het vorige tentamen hadden een paar toepassingen het zwaar te verduren en daar heb ik toen consequenties uit getrokken.

Het huwelijksfeest van de sultan

Zo dacht ik bijvoorbeeld een mooie toepassing te hebben over het huwelijksfeest van de sultan. Deze wil het aantal dagen dat het feest zal duren en de dagelijkse hoeveelheden vaten wijn, die voor de onlesbare gasten uit de wijnkelder worden gehaald, zo kiezen dat het succes van het feest zo groot mogelijk is. In die wijnkelder bevinden zich bij het begin van het feest honderd vaten, en de sultan neemt aan dat het product van de gekozen dagelijkse aantallen wijnvaten een goede maat is voor het succes van het feest. Uiteraard is hier de bedoeling dat hoe groter dit product is, hoe beter het feest. Dus een feest van vier dagen met tien, twintig, dertig en veertig vaten is beter dan een feest van twee dagen met iedere dag vijftig vaten, want het product $10 \times 20 \times 30 \times 40$ is groter dan het product 50×50. Is het misschien zo dat hoe langer het feest duurt, hoe beter het is? Zo eenvoudig is het niet want een feest van honderd dagen met op elke dag een vat leidt tot het magere product een. Iedereen die houdt van een

[1] J. Brinkhuis, V. Tikhomirov, *Optimization: Insights and Applications*, Princeton University Press, 2005

eenvoudige breinbreker op zijn tijd, zal plezier kunnen beleven aan het vinden van de optimale afweging tussen een lang feest en veel wijn per dag.

Kritiek op 'Het huwelijksfeest van de sultan'
En nu de pijlen die de studenten bij het vorige tentamen op deze opgave hebben afgeschoten. Het meest pijnlijke schot betrof mijn culturele blunder wat betreft de combinatie wijn en sultan. Ik heb snel een kroonprins van de sultan gemaakt. Een andere pijl die doel trof was het vraagteken dat je kunt plaatsen bij de gekozen maat voor het succes van het feest: het product van de dagelijkse hoeveelheden wijnvaten. In economische termen is dat een keuze van een *nutsfunctie*. Deze nutsfunctie heeft wat dubieuze eigenschappen. Een dag zonder wijn verknoeit blijkbaar het hele feest, en een extra dag met maar een vat geeft geen extra plezier. Nog drie succesvolle pijlen:

(i) hoe kun je de uitkomst van dit probleem in de praktijk gebruiken ('pragmatisch nut'),
(ii) wat voor economisch of ander nieuw inzicht illustreert het ('wetmatigheid'),
(iii) leidt het tot de essentie van een wiskundig verschijnsel ('wiskunde')?

Een leuke puzzel is nog geen overtuigende toepassing
Een moment dacht ik nog dat in ieder geval de laatste pijl zou afketsen. Want als je de franje weglaat, blijft er dan niet een leuke wiskundige puzzel over? Dat is zo, maar diepgaande discussies met mijn coauteur Prof. V. Tikhomirov uit Moskou, hebben mij tot het inzicht gebracht dat zulke puzzels, hoe leuk en uitdagend ook, niet aan de hoogste eisen van echt overtuigende toepassingen voldoen. De oplossing leidt niet tot het doordringen tot de essentie van wat dan ook. Dat neemt overigens niet weg dat het oplossen van puzzels een boeiende bezigheid kan zijn.

Vergelijking van e^π en π^e
Hier is nog een aardige puzzel: welk getal is groter, e^π of π^e? Met een rekenapparaat zie je dit zo – het scheelt overigens maar weinig, die getallen zijn bijna gelijk – maar de uitdaging is om dit te bewijzen en het blijkt dat je dit het eenvoudigst kunt doen met optimaliseringsmethoden, hoewel de puzzel niets met optimalisering te maken lijkt te hebben. De clou is dat je het probleem kunt vervangen door het probleem welk getal groter is, $e^{e^{-1}}$ of $\pi^{\pi^{-1}}$; dit brengt je dan op het idee om de functie $f(x) = x^{x^{-1}}$ te maximaliseren. Overigens gebruikt Matlab deze puzzel om grafische mogelijkheden te demonstreren: aan de grafiek van de functie $x^y - y^x$ kun je duidelijk zien wat de oplossing is van deze puzzel. Maar een echt bewijs is dat natuurlijk niet.

Gearrangeerd huwelijk
Een andere toepassing die het bij het vorige tentamen moest ontgelden betrof het verschijnsel gearrangeerd huwelijk. Het ging hier om een situatie met een

beperkt vetorecht: de vraag is wat het optimale gebruik van dit vetorecht is. Hier zijn de details.

In grote delen van Afrika, Azië en het Midden Oosten is een aanzienlijk deel van alle huwelijken gearrangeerd. Het algemene idee achter dit gebruik is dat jonge mensen op plezier uit zijn en dat je er niet op kunt vertrouwen dat zij zelf een geschikte partner kunnen vinden. Daarom nemen ouders, vrienden en bemiddelaars de taak op zich om een geschikte bruid te vinden. We maken nu de volgende aannamen. De jongeman heeft een vetorecht: hij heeft de mogelijkheid om 'nee' te zeggen tegen een huwelijksvoorstel. Als hij dat doet, wordt er verder gezocht naar een nog betere kandidaat. Dus het weigeren van een voorstel is aantrekkelijk: het volgende voorstel zal beter zijn. Maar er is een probleem: altijd blijven weigeren is ook niet aantrekkelijk. Laten we eens kijken naar het geval van een jongeman die wil trouwen voordat hij de leeftijd T bereikt. Hij besluit om alle voorstellen voor een poosje te weigeren, zeg tot moment w, en dan ja te zeggen tegen het eerste voorstel dat hij daarna krijgt. Welke w zou hij dan het best kunnen kiezen? Bij keuze van een kleine w is de kans hoog dat er na het voorstel waartegen hij ja zegt, nog een ander voorstel gedaan zou worden voordat hij de leeftijd T bereikt. Dan zou hij dus niet de best mogelijke kandidaat gekregen hebben. Aan de andere kant, als hij w groot kiest, dan is de kans hoog dat hij de leeftijd T bereikt zonder te trouwen. We nemen aan dat de huwelijksvoorstellen die hem gedaan worden een continue Poissonverdeling volgen met gegeven gemiddelde. Verder nemen we aan dat zijn doel is om de kans op succes te maximaliseren. Deze kans is gelijk aan de kans dat hij precies één voorstel zal krijgen in het tijdsinterval $[w, T]$. Als hij in dit interval nul voorstellen zou ontvangen, dan zou hij ongetrouwd blijven en als hij in dit interval meer dan een voorstel zou krijgen, dan zou hij trouwen met de verkeerde kandidaat. Daarom is hij alleen succesvol als hij precies een voorstel krijgt.

Kritiek op 'Gearrangeerd huwelijk'

Er zijn zelfs wetenschappelijke publicaties waarin geprobeerd wordt om het afwegingsprobleem tussen te snel accepteren en te lang wachten te modelleren en op te lossen. Leuke modellen, daar niet van, maar geen overtuigende toepassingen, want de uitkomsten hebben geen praktisch nut en leiden niet tot waardevolle nieuwe inzichten of wetmatigheden. Overigens heb ik gemerkt dat het onderwerp gearrangeerde huwelijken de emoties hoog kan doen oplaaien... Ik heb daarom nog overwogen een dating show van het gearrangeerde huwelijk te maken. Maar eerlijk gezegd zijn de aannamen zelfs niet erg realistisch voor een datingshow, dus, alles bij elkaar, weg met dit voorbeeld; gelukkig waren de drukproeven voor de eerste druk nog niet de deur uit.

Wie heeft gelijk, de econoom of de econometrist?

Een toepassing die ik voorlopig ook maar heb geschrapt, heette: 'Wie heeft gelijk, de econoom of de econometrist?' Het behandelt een interessante kwestie, met een knipoog die duidelijk was, dacht ik, en aan de hand van een verhaal

waar de lezer aan het eind op het verkeerde been wordt gezet.

Een econoom onderzoekt een kruispunt in New York City. Het is een druk kruispunt en iedere week gebeuren er meestal wel een paar auto-ongelukken. De econoom wil dit modelleren en hij wil een schatting geven van de kans dat er in een bepaalde week geen ongeluk zal zijn. Hij begint met het verzamelen van informatie en dan maakt hij een lijst van het aantal ongelukken in elk van de afgelopen honderd weken. Hij neemt aan dat het aantal ongelukken in een week verdeeld is als een discrete Poissonverdeling. Daarom schat de econoom het gemiddelde λ van die verdeling—die zoals gezegd de verdeling bepaalt—door gewoon het gemiddelde te nemen van de honderd getallen op zijn lijst. Dan vult hij deze waarde in in de bekende formule voor de Poissonverdeling om de gezochte modellering te krijgen en daarna vult hij 0 in om de kans dat er geen ongeluk is te schatten. Een paar dagen later komt de econoom een kennis tegen die econometrie studeert. De econoom vertelt over zijn oplossing voor het auto-ongelukken-probleem, maar de econometrist zegt meteen dat de zaak niet zo eenvoudig ligt. "Je mag niet zomaar het gemiddelde aantal wekelijkse ongelukken schatten door het gemiddelde van de aantallen van de laatste honderd weken te nemen. Je moet eerst een schattingsmethode kiezen, bijvoorbeeld de Maximum Likelihood methode." Zij gaat aan de slag met die methode, en de schatting die zij dan vindt is gewoon het gemiddelde dat de econoom genomen heeft.

Kritiek op 'Wie heeft gelijk, de econoom of de econometrist?'

Het verhaal werd niet afgesloten met een moraal, de conclusie werd met opzet aan de lezer overgelaten. Dit leidde tot zeer diverse, soms felle, reacties bij mijn onderwijs aan econometriestudenten en bij het onderwijs aan PhD-studenten economie. Zo werd hier door sommigen uit geconcludeerd dat dit weer eens illustreert dat 'al die statistische moeilijkdoenerij' nergens goed voor is, anderen wierpen mij voor de voeten dat het een vergissing is om hier een voorbeeld te kiezen waar een eenvoudige formule achteraf de beste blijkt te zijn. Zelf had ik eigenlijk verwacht dat de boodschap dat bewezen is dat een eenvoudige schattingsformule ook de beste is, in goede aarde zou vallen. Overigens, een ander voorbeeld van dit verschijnsel is de zeer populaire kleinste kwadraten methode; het is van belang om er bij het onderwijs op te wijzen dat deze schattingsmethode niet alleen de nu eenmaal gangbare is, maar ook de *beste* schatting geeft. Omdat, tegen mijn verwachting in, de tekst tot misverstanden bleek te leiden, heb ik de toepassing voorlopig geschrapt, maar binnenkort hoop ik een versie te schrijven waarbij de aandacht niet nodeloos van de essentie wordt afgeleid.

Nieuwe pijlen?

Ik ben benieuwd wat het tentamen over twee weken voor nieuwe interessante reacties op de huidige collectie toepassingen zal opleveren.

De rode draad

Toepassingen van optimalisering met continue variabelen kunnen alleen dan als echt overtuigend worden beschouwd als ze bestand zijn tegen een van de volgende drie kritische vragen: dienen ze een pragmatisch doel (dit is bijvoorbeeld het criterium in de financiering), geven ze een verhelderend inzicht in een wetmatigheid (dit is bijvoorbeeld het criterium in de economie) of leiden ze tot het doorgronden van een diepzinnig verschijnsel (het criterium in de wiskunde).

Voorlopige samenvatting

De boven gegeven voorbeelden vallen allemaal door de mand.

(i) 'Het huwelijksfeest van de kroonprins' wekt door zijn vormgeving de indruk dat het een pragmatisch doel dient, een zo geslaagd mogelijk feest, maar het verhaal is onrealistisch en alleen bedoeld om een puzzel wat 'op te leuken'.

[Terzijde, alle wiskundeleerboeken voor het VWO staan vol van dit soort schijntoepassingen. De bedoeling is om de 'droge stof' aantrekkelijker te presenteren, maar de werkelijkheid is dat dit door bijna geen enkele leerling gewaardeerd wordt, en dat bestudering van deze schijntoepassingen bijdraagt aan een verkeerde en negatieve beeldvorming – wellicht voor het leven – bij middelbare scholieren over wat wiskunde is en wat je ermee kunt doen.]

Op het tweede gezicht lijkt het probleem een overtuigende wiskundige toepassing, maar het is alleen een puzzel, die met veel plezier kan worden opgelost, maar die verder nergens toe leidt.

(ii) 'Het gearrangeerde huwelijk' is bedoeld om inzicht in een aspect van de werkelijkheid om ons heen te krijgen, maar slaagt daar niet in: het leidt tot geen enkel verhelderend inzicht. Dit soort problemen heeft wel zijn waarde, het zijn leuke vingeroefeningen met behulp waarvan je kunt proberen te leren om economische verschijnselen in een simpel model te vangen.

(iii) 'Wie heeft gelijk de econometrist of de econoom?' bevat impliciet verschillende boodschappen, maar de ervaring heeft mij geleerd dat een boodschap tegelijk tot een beter effect leidt.

Verder hebben we gezien dat ook problemen die op het eerste gezicht niets met optimalisering te maken hebben, soms toch met optimaliseringsmethoden kunnen worden aangepakt. Bijvoorbeeld het is een verrassing dat de puzzel 'Welk getal is groter, e^π of π^e?' kan worden opgelost door een functie van een variabele te maximaliseren. Verder realiseren niet alle gebruikers van populaire methoden (zoals de kleinste kwadraten methode) zich dat dit eigenlijk oplossingen van optimaliseringsproblemen zijn.

2. De drie gouden klassen van overtuigende toepassingen: nut, inzicht en diepte

Mijn belangstelling voor wiskunde is voor een deel gewekt door het plezier dat ik op de middelbare school had in puzzels; dit werd verder gestimuleerd door

wiskundeolympiades. Ik vind zulke puzzels nog steeds boeiend. Ik was bijvoorbeeld heel enthousiast over de toepassing van optimalisering op het probleem om e^π en π^e te vergelijken dat ik hierboven heb genoemd. Ik besloot het een prominente plaats te geven in het boek over optimalisering dat ik aan het schrijven was samen met Prof. Tikhomirov. Maar hier stuitte ik onverwacht op groot verzet bij mijn coauteur. Na lange en interessante discussies zijn wij tot de volgende lijst van drie gouden klassen van serieuze en overtuigende toepassingen gekomen. Allebei kwamen we verrijkt uit die discussie. Ik kreeg een helder inzicht in het onderscheid tussen puzzels en wiskundige toepassingen. Mijn coauteur was aangenaam verrast hoeveel overtuigende toepassingen er zijn van optimaliseringsmethoden op het gebied van economie en financiering. Dit zijn de drie gouden klassen die volgens ons alle overtuigende toepassingen bevatten. Sommige overtuigende toepassingen behoren tot meer dan een klasse.

(i) *Pragmatische toepassingen ('nut').* Pragmatische overwegingen stimuleren een mens om het beste te doen dat hij kan, gegeven bestaande beperkingen. Het gaat hier vaak om een afweging tussen tegengestelde effecten. Bijvoorbeeld, hoe meer jaren onderwijs je volgt, hoe beter je baan zal zijn. Tegenover die baten staan ook kosten, je moet voor je opleiding betalen en in de tijd dat je studeert verdien je niets. Er zijn talloze voorbeelden van economische pragmatische problemen, bijvoorbeeld problemen van minimale kosten, maximale winst of maximaal sociaal welzijn.

Deze klasse wordt hieronder geïllustreerd met een toepassing uit de financiering die een enorme impact heeft gehad: het probleem van het prijzen van opties. Het werk van Black en Scholes heeft dit probleem opgelost en dit heeft geleid tot meer stabiliteit in de financiële wereld.

(ii) *Wetmatigheden ('inzicht').* De meeste – of alle – natuurwetten kunnen worden gezien alsof de natuur optimaliseert. Bijvoorbeeld, licht gedraagt zich in wezen alsof het de snelste weg kiest. In de economie zijn er soortgelijke voorbeelden. Een ervan betreft het marktprincipe, een van de fundamenten van de economie. Dit laat zien dat markten in de volgende zin tot een optimale uitkomst leiden: de prijs waarvoor vraag gelijk is aan aanbod is de prijs waarvoor het sociale welzijn maximaal is. Veel van die economische toepassingen zijn belangrijk voor het bepalen van economisch beleid.

Deze klasse illustreren we hieronder met behulp van het werk van Kydland en Prescott, winnaars van de Nobelprijs Economie 2004: aan hun ontdekking heeft de wereld een lange periode van lage inflatie te danken.

(iii) *Wiskundige toepassingen ('diepte').* De rechtvaardiging voor wiskundige toepassingen is minder recht toe recht aan. Het is het nobele streven om tot de essentie van alles door te dringen: wetenschappelijke nieuwsgierigheid drijft sommige mensen ertoe om kwesties tot op de bodem uit te zoeken. Bijvoorbeeld, als je al een boven- of benedengrens hebt die voldoet voor praktische doeleinden, dan kan je toch doorgaan om de scherpste grens te vinden.

Deze klasse zullen we illustreren met de hoofdstelling van de algebra, een veelgebruikt resultaat dat in principe bekend is aan iedere middelbare scholier: ieder polynoom van een variabele is volledig te ontbinden in lineaire en kwadratische factoren. Voor praktische doeleinden is het niet nodig om te begrijpen waarom dit zo is. Als je dit toch wilt weten, merk je dat dit niet zo eenvoudig is: bekende bewijzen vereisen of kennis van Galois-theorie of kennis van de theorie van de analytische functies. Hier bieden we een bewijs met behulp van optimaliseringsmethoden aan. Hiermee dring je door tot de essentie van de zaak, en het bewijs kan de vergelijking met de bestaande bewijzen goed doorstaan.

3. Financiering: de juiste prijs voor opties

De formule van Black en Scholes

Nog niet zo heel lang geleden was de handel in opties van weinig betekenis, omdat de juiste prijs voor een optie niet bepaald kon worden. Dankzij het baanbrekende werk van Black en Scholes kan dit nu wel. Dit werk heeft voor een revolutie gezorgd in de *praktijk* van de financiering. Opties spelen een essentiële rol in de economie omdat zij bedrijven in staat stellen om zich te beschermen tegen risico's die zij zelf niet kunnen dragen. Opties bestaan al duizenden jaren maar het leek altijd een kwestie van smaak hoe je ze zou moeten prijzen. Dit leek namelijk af te hangen van je eigen voorspelling van de toekomstige waarde van het onderliggende aandeel, en bovendien van je bereidheid om risico's te nemen. Een verrassend inzicht van Black en Scholes is dat de waarde van een optie voor iedereen precies dezelfde is. De spectaculaire ontwikkeling van de optiemarkten zou niet mogelijk zijn geweest zonder het werk van Black en Scholes.

Opties in het nieuws

Bijna iedereen heeft wel eens gehoord over opties, ze zelf misschien wel eens gekocht hebben, of iemand kennen die dat gedaan heeft. Het onderwerp heeft de laatste jaren vaak de kranten gehaald in verband met de beloning van topmanagers in de vorm van opties. Het idee hierachter is dat zulke beloningen een bedrijf weinig kosten, dat ze aantrekkelijk zijn voor de topmanager mede omdat de eventuele winst bij verkoop niet belast is, en dat het de topmanager zou aanzetten tot een betere 'performance' omdat de waarde van zijn optiepakket stijgt als het beter gaat met het bedrijf.

De pech van Michael

Mijn eerste ervaring met opties was via een studentassistent, laten we hem Michael noemen. Op een dag kwam Michael naar mij toe met een geheimzinnig gezicht, hij had van iemand van de afdeling Financiering en Belegging een gouden tip gekregen: de jaarcijfers van Akzo die binnenkort bekend gemaakt zouden worden, zouden veel beter zijn dan verwacht. Dus had hij onmiddellijk voor een relatief klein bedrag opties gekocht, die het recht gaven om voor de huidige koers aandelen te kopen na bekendmaking van de jaarcijfers. Dan zou

de koers naar verwachting omhoog gaan en dan zou hij het koersverschil kunnen incasseren. Een enorme winst voor Michael zou het gevolg zijn. Verder dacht hij geen risico's te lopen, want hij had de bank opdracht gegeven om de opties onmiddellijk te verkopen als de prijs van de aandelen onverhoopt omlaag zou gaan. Spannende weken braken aan, op de grote dag zag ik in het nieuws dat de jaarcijfers inderdaad hoger waren dan verwacht. Ik hoopte al op een feest, maar daarna verscheen Michael een poos niet op de universiteit. Later hoorde ik van hem wat er was gebeurd: de halve financiële wereld had de gouden tip ook gekregen, en deze kennis was al verwerkt in de prijs van de aandelen op het moment dat Michael zijn opties kocht. Er was zelfs een overenthousiasme en daardoor zakte na het bekend maken van de goede jaarcijfers de prijs van de aandelen enigszins. Maar als dat gebeurt zijn de opties ineens niets meer waard, en dus kelderde de prijs van de opties binnen een paar minuten, lang voordat de bank de opties van Michael probeerde te verkopen.

Waar zijn opties (niet) goed voor?
Deze anekdote illustreert een elementair feit: als individu kan je voor weinig geld met opties speculeren en zo soms een grote slag slaan, maar je kunt niet systematisch geld verdienen met opties. Dat is ook niet de bedoeling van opties; hun nuttige rol in de financiële wereld is dat ze door bedrijven gebruikt kunnen worden om bepaalde risico's af te dekken; zo leiden ze tot grotere stabiliteit.

Uitleg van de formule van Black en Scholes
De oorspronkelijke methode van Black en Scholes is niet zo eenvoudig uit te leggen. Die methode maakt gebruik van het oplossen van een partiële differentiaalvergelijking. Inmiddels is er een eenvoudig alternatief beschikbaar voor het volgende karaktistieke speciale geval (de formule voor het algemene geval kan hieruit worden afgeleid 'door het nemen van een limiet'): drie assets–een aandeel, een staatsobligatie en een optie–en twee scenario's–het aandeel gaat omlaag of het aandeel gaat omhoog, en een korte periode.

- Aandeel. Een aandeel heeft huidige prijs p; in het eerste scenario gaat de waarde naar beneden naar $v^{(1)}$, in het tweede scenario gaat de waarde naar boven naar $v^{(2)}$.
- Staatsobligatie. Bovendien is er een staatsobligatie, dat risicoloos is; we zetten de huidige prijs op 1, en we kunnen bereiken, door een normalisatie, dat de waarde 1 blijft. Dit schaadt de algemeenheid niet en vereenvoudigt de formules. In financiële termen komt die normalisatie neer op de aanname dat de rente gelijk is aan 0. Daardoor verandert de prijs van de obligatie niet, en maakt het niets uit dat je de optiepayoff pas na afloop van de periode krijgt terwijl je er nu al voor betaalt.
- Optie. Verder is er een Europese calloptie op het aandeel, met uitoefenprijs w. Dit is het recht – maar niet de plicht – om het aandeel te kopen na afloop van de investeringsperiode voor de prijs w.

– Formule van Black en Scholes. De prijs van de optie blijkt precies vast te liggen onder een plausibele aanname, de afwezigheid van *arbitragemogelijkheden*, waarover later meer, en wordt gegeven door de volgende formule :

$$p_0 = (v^{(2)} - v^{(1)})^{-1}(p - v^{(1)})(v^{(2)} - w). \qquad (*)$$

Deze optieprijsformule is de formule van Black en Scholes in ons speciale geval. We illustreren de formule met een numeriek voorbeeld. Daarna beginnen we met de afleiding van de formule.

Voorbeeld
We bekijken een aandeel en een staatsobligatie die allebei nu 100 euro waard zijn. De waarde van het aandeel na een maand zal of 50 euro of 200 euro zijn; de kansen van deze scenario's zijn niet noodzakelijk gelijk, in feite zijn die kansen niet bekend. De waarde van de staatsobligatie na een maand zal nog steeds 100 euro zijn. Nu bekijken we een optie die het recht geeft om het aandeel na een maand te kopen voor 110 euro, wat zijn waarde op dat moment ook is.
Welke prijs moet de bank voor deze optie vragen?
De juiste prijs in euros kan worden berekend met de formule van Black en Scholes:

$$(200 - 50)^{-1}(100 - 50)(200 - 110) = 30.$$

We zien dus dat we aan het begin van de investeringsperiode met de formule van Black en Scholes en met de beschikbare informatie de prijs van een optie kunnen bepalen.

Kansinterpretatie van de formule van Black en Scholes
We geven nu nog een mooie interpretatie van de formule van Black en Scholes (*) in termen van kansen.
Er is een unieke kansverdeling voor de twee mogelijke scenario's waarvoor de prijs van het aandeel en de staatsobligatie gelijk zijn aan hun verwachte waarde na de investerings periode. De juiste prijs voor de optie is de verwachte waarde na afloop van de investeringsperiode ten opzichte van deze kansverdeling.
Die unieke kansverdeling is: kans $\frac{v^{(2)}-p}{v^{(2)}-v^{(1)}}$ voor scenario 1 en kans $\frac{p-v^{(1)}}{v^{(2)}-v^{(1)}}$ voor scenario 2.

Voor alle duidelijkheid: die kansverdeling heeft niets te maken met de werkelijke kansverdeling voor de scenario's. Die is onbekend, en het mooie van de formule van Black en Scholes is onder andere dat deze laat zien dat de correcte prijs voor een optie helemaal niet afhangt van deze onbekende kansverdeling.

Het doel van deze paragraaf is om deze formule te begrijpen met behulp van optimaliseringsmethoden. We zullen zelfs meer doen en een algemene schattingsmethode voor derivaten – een algemene term voor optie-achtige financiële

producten – presenteren die de formule van Black en Scholes als een bijzonder geval bevat.

- Waarde optie na de periode. In de eerste plaats, wat heb je eigenlijk aan het bezit van een optie? Laten we eerst eens kijken naar het tweede scenario. In het tweede scenario kan de eigenaar van de optie het aandeel kopen voor het bedrag w en het daarna onmiddellijk doorverkopen voor de marktwaarde $v^{(2)}$, die hoger is. Zo maakt hij een winst van $v^{(2)} - w$. Die winst kan worden gezien als de waarde van de optie in het tweede scenario. In het eerste scenario heeft de optie geen waarde, want de optie geeft het recht om het aandeel te kopen voor de prijs w, maar dat recht is in het tweede scenario niets waard, want w is meer dan de prijs $v^{(1)}$ waarvoor iedereen het aandeel op dat moment zou kunnen kopen.

- Hoe waarschijnlijk is een scenario? Bij het begin van de investeringsperiode is er niets bekend over de waarschijnlijkheid van de scenario's.

- Het probleem van de juiste prijs voor opties. Het probleem is voor welke prijs p_0 de optie nu verkocht moet worden.
  Je kunt onmiddellijk een boven- en een ondergrens geven voor de waarde van een optie nu (= aan het begin van de investeringsperiode): die waarde is niet meer dan $v^{(2)} - w$, want dit is de waarde na afloop van de investeringsperiode in het tweede scenario, terwijl de waarde in het andere scenario nul is; aan de andere kant is de optie uiteraard wel *iets* waard. We gaan nu het begrip 'arbitragemogelijkheid' invoeren. Dit leidt tot een spectaculaire verfijning van de methode om boven- en ondergrenzen voor de prijs van een optie te geven: boven- en ondergrenzen blijken dan samen te vallen, en de waarde van een optie nu ligt dus vast.

- Econometriestudenten ontdekken in Polen een arbitragemogelijkheid. Een aantal jaren geleden ging de jaarlijkse studiereis van het Econometrisch Dispuut naar Polen. De geldmarkt was nog niet gereguleerd en ieder wisselkantoortje had zijn eigen wisselkoersen. Sommige studenten ontdekten dat ze zomaar geld konden verdienen aan die verschillen in wisselkoersen door een rondje te maken langs de wisselkantoren en op de juiste manier valuta te wisselen. Dit is een voorbeeld van een arbitragemogelijkheid. In principe hadden de studenten schatrijk naar Nederland terug kunnen komen, door steeds maar zulke lucratieve rondjes te maken. In werkelijkheid was dit natuurlijk nooit gelukt, want hun frequente bezoeken hadden er zeker toe geleid dat de wisselkoersen zo aangepast zouden worden dat de gevonden arbitragemogelijkheid verdwenen zou zijn.

- Definitie van het begrip arbitragemogelijkheid. Deze anekdote illustreert het begrip arbitragemogelijkheid en het feit dat zo'n mogelijkheid nooit lang kan bestaan. Het is in feite een goede benadering van de werkelijkheid om aan te nemen dat arbitragemogelijkheden helemaal niet voorkomen. Deze aanname wordt meestal gemaakt bij analyses van financiële markten.

De arbitragemogelijkheden die we hier bekijken kunnen ruwweg worden gedefinieerd als investeringsstrategiëen waarbij we nu geld krijgen maar nooit iets hoeven te betalen. De precieze definitie is dat een arbitragemogelijkheid een vector (x_1, x_2, x_3) is waarvoor de volgende ongelijkheden gelden:

$$\begin{array}{lcl} px_1 + x_2 + p_0 x_3 & < & 0, \\ v^{(1)} x_1 + x_2 & \geq & 0, \\ v^{(2)} x_1 + x_2 + (v^{(2)} - w) x_3 & \geq & 0. \end{array}$$

Hier is x_1 het aantal gekochte aandelen, x_2 het aantal gekochte staatsobligaties, en x_3 het aantal gekochte opties. De getallen x_1, x_2, x_3 kunnen positief, nul of negatief zijn ('going short'). 'Negatief zijn' heeft de voor de hand liggende interpretatie, en 'going short' is vaak toegestaan.

- Betekenis van de drie ongelijkheden: arbitrage. Wat betekenen die drie ongelijkheden boven nu precies? De eerste betekent dat je geld krijgt als je bereid bent om deze investering in bezit te hebben (met de eventuele hierbij behorende betalingsverplichtingen aan het einde van de investeringsperiode). De tweede en de derde ongelijkheid betekenen dat je helemaal geen geld hoeft te betalen (dat wil zeggen minstens evenveel krijgt als je moet betalen) aan het eind van de investeringsperiode in het eerste en tweede scenario, respectievelijk. Dit maakt duidelijk dat een oplossing van die drie ongelijkheden de naam 'arbitragemogelijkheid' verdient.

Voor een volledig begrip bekijken we vanaf nu een meer algemene context: n assets en m scenario's. Na afloop zullen we dit toepassen op drie assets – aandeel, staatsobligatie, optie – en twee scenario's – 'aandeel gaat omlaag' en 'aandeel gaat omhoog'.
We bekijken de volgende situatie:

- m assets en een investeringsperiode,
- $p_1, \ldots, p_m$, de prijzen bij het begin van de investeringsperiode (die willen we bepalen of we willen er op zijn minst boven- en ondergrenzen voor vinden),
- $x_1, \ldots, x_m$, de grootte van de investeringen aan het begin van de periode (hier is short-selling, dat wil zeggen, negatieve x_i, toegestaan),
- de kosten van de investeringen $p^T \cdot x$.

Nu modelleren we ook het risico:

- we bekijken n scenario's,
- $v_1^{(j)}, \ldots, v_m^{(j)}$, zijn de waarden in scenario j (we nemen aan dat de waarde van $v_i^{(j)}$ precies bekend is voor ieder asset i en ieder scenario j),
- de uiteindelijke waarde van de investeringen in scenario j is de j-de coördinaat van de vector $V^T x$, waarbij V de $m \times n$-matrix $(v_i^{(j)})_{ij}$ is en $V^T x$ het matrix product van de matrix V^T en de vector x is.

We nemen wel aan dat de $v_i^{(j)}$ bekend zijn, maar niet dat de waarschijnlijkheidsverdeling van de scenario's dat is. Het is misschien verrassend op het

eerste gezicht dat de prijzen $p_1, \ldots, p_m$ bij het begin van de investeringsperiode niet willekeurig zijn. De reden dat dit niet zo is, is dat het redelijk is om aan te nemen dat er geen arbitragemogelijkheden zijn. Arbitragemogelijkheden kunnen ruwweg gedefinieerd worden als investeringsstrategieën, waarbij we nu geld krijgen maar nooit iets hoeven te betalen. Dit idee kan worden omgezet in een precieze definitie. Een arbitragemogelijkheid is een vector x waarvoor

$$p^T \cdot x < 0, \quad V^T x \geq 0.$$

STELLING *Als er geen arbitragemogelijkheden zijn, dan is er een vector* $y \geq 0$ *waarvoor*

$$p = Vy.$$

BEWIJS

(i) Het optimaliseringsprobleem $f(x) = p^T \cdot x \to \min,\ V^T x \geq 0$ is oplosbaar wegens de aanname dat er geen arbitragemogelijkheden zijn.

(ii) Toepassing van de standaard noodzakelijke voorwaarden voor convexe optimaliseringsproblemen van Karush-Kuhn-Tucker geeft: er is een vector $y \geq 0$ waarvoor $p = Vy$.

Q.E.D.

Kansinterpretatie

Een risicoloos asset kan worden gedefinieerd als een waarvoor de waarde in ieder scenario gelijk is aan de prijs aan het begin van de investeringsperiode. Het bekendste voorbeeld hiervan is een staatsobligatie. Als de collectie van de n assets een risicoloos asset bevat, dan kan y geïnterpreteerd worden als een kansverdeling op de collectie van scenario's: inderdaad geldt dan

$$y_1 + \cdots + y_m = 1,\ y_i \geq 0,\ 1 \leq i \leq m.$$

Dan hebben de overige vergelijkingen van het systeem $p = Vy$ de volgende interpretatie in termen van kansen: voor ieder asset i is de prijs p_i aan het begin van de periode gelijk aan de verwachting van de waarde aan het einde van de periode ten opzichte van de kansverdeling y:

$$p_i = y_1 v_i^{(1)} + \cdots + y_m v_i^{(m)}.$$

Relatie met optimalisering

Bijvoorbeeld, als de prijzen aan het begin van de periode van alle assets op een na bekend zijn, dan geeft de bovenstaande stelling informatie over de prijs bij het begin van de periode van het overgebleven asset: die prijs ligt tussen de extreme waarden van het volgende optimalisatie probleem met gegeven $m \times n$-matrix V en met gegeven prijzen $p_1, \ldots, p_{m-1}$:

$$f(p, y) = p_m \to \text{extr},\ p = Vy,\ y \geq 0.$$

In het bijzonder, als de extreme waarden van dit probleem samenvallen, dan is de prijs van het overgebleven asset bepaald. Dit is zo in het hierboven gepresenteerde speciale geval (een aandeel, een staatsobligatie, een optie; twee scenario's: 'het aandeel gaat omlaag' en 'het aandeel gaat omhoog'). Dan krijgen we een stelsel van drie lineaire vergelijkingen in drie onbekenden y_1, y_2 en p_0:

$$\begin{aligned} p &= y_1 v^{(1)} + y_2 v^{(2)}, \\ 1 &= y_1 + y_2, \\ p_0 &= y_2(v^{(2)} - w). \end{aligned}$$

Zonder veel moeite kan worden vastgesteld dat dit stelsel precies een oplossing heeft, en dat voor die oplossing y_1 en y_2 niet-negatief zijn. Dit stelsel legt dus in het bijzonder p_o, de prijs van de optie, vast. Oplossen van dit stelsel geeft

$$\begin{aligned} y_1 &= \tfrac{v^{(2)}-p}{v^{(2)}-v^{(1)}}, \\ y_2 &= \tfrac{p-v^{(1)}}{v^{(2)}-v^{(1)}}, \\ p_0 &= (v^{(2)} - v^{(1)})^{-1}(p - v^{(1)})(v^{(2)} - w). \end{aligned}$$

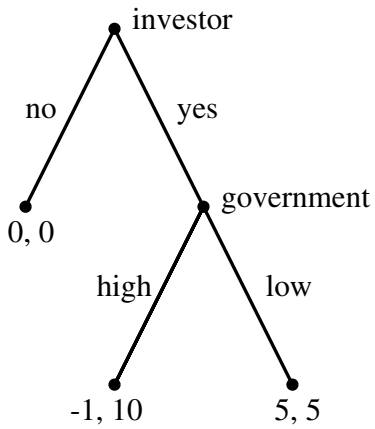
De laatste formule is de boven gegeven optieprijsformule $(*)$ van Black en Scholes; de eerste twee geven de unieke kansverdeling uit de boven gegeven kansinterpretatie van de optieprijsformule.

4. Economie: de waarde van commitment

De wereld economie heeft lang te kampen gehad met een hoge inflatie en de ongunstige effecten hiervan. Het beleid van centrale banken dat er op was gericht om de inflatie terug te dringen mislukte keer op keer. Het werk van Kydland en Prescott over de waarde van commitment – beloond met de Nobelprijs Economie in 2004 – heeft duidelijk gemaakt waarom. Dit *inzicht* heeft er toe geleid dat het nu al vele jaren gelukt is om de inflatie laag te houden.

Nu geven we een beschrijving in woorden van dit model. Op tijdstip $t = 0$ beslist een investeerder om te investeren in een land of niet. Een jaar later, op tijdstip $t = 1$ beslist de regering over het belastingtarief voor kapitaal. We nemen aan dat de payoffstructuur als volgt is. Als de investeerder niet investeert in het land, dan krijgen zowel de investeerder als de regering payoffs die we normaliseren tot nul. Als de investeerder investeert in het land en de regering tot een laag belastingtarief besluit, dan krijgen zowel de investeerder als de regering een payoff die gelijk is aan 5. Maar als de investeerder investeert en de regering op tijdstip $t = 1$ een hoog belastingtarief kiest dan verliest de investeerder en krijgt hij -1 terwijl de regering een hoge belastingopbrengst krijgt: een payoff gelijk aan 10.

Als de regering zich kan vastleggen ('commitment') op tijdstip $t = 0$ op de beslissing die zij neemt op $t = 1$ dan zou ze zich vastleggen op een laag belastingtarief. Dit kan door een wet te maken die bepaalt dat het belastingtarief laag is (dat wil zeggen, veranderen van de wet leidt tot kosten groter dan 5, wat switchen naar een hoger belastingtarief later niet rendabel maakt). Of



Figuur 1. *Payoff structure*

de regering kan proberen om een reputatie te krijgen voor lage belastingtarieven. In geval van een regering die probeert om inflatie te bestrijden, komt de commitment van het instellen van een centrale onafhankelijke bank.

Als de regering zich niet kan vastleggen, kiest zij een optimaal plan op elk moment dat zij een beslissing moet nemen. Dat impliceert dat zodra de investeerder geïnvesteerd heeft, de regering het hoge belastingtarief kiest omdat een payoff van 10 meer is dan een payoff van 5. De investeerder ziet dit van tevoren aankomen en kiest ervoor om niet te investeren omdat een payoff van 0 meer is dan een payoff van -1.

Het is duidelijk dat de commitment oplossing nooit slechter is dan de niet commitment oplossing en in het geval boven is het zelfs duidelijk beter. Dat wil zeggen, commitment heeft waarde. Een voorbeeld dat vaak gebruikt wordt is Odysseus die zich aan de mast heeft laten vastbinden om naar de Syrenen te luisteren. Op die manier had hij zich vastgelegd om niet naar ze toe te gaan terwijl hij toch kon luisteren.

We merken ook op dat in een deterministische wereld commitment altijd beter is dan geen commitment. In een wereld met onzekerheid, is er een afweging. Met commitment kan men niet reageren op onverwachte gebeurtenissen. In het voorbeeld boven, zou je je gecommitteerd kunnen hebben aan een laag belastingtarief, maar het land zou in een belastingcrisis terecht kunnen komen en dan kun je niet reageren door de belastingtarieven te verhogen.

Relatie met optimalisering

Een van de ideeën van Kydland en Prescott kan uitgelegd worden in termen van het verschil tussen oplossingen van dynamische optimaliseringsproblemen met de volgende twee methoden: Pontryagin's Maximum Principe en de Bellman vergelijking. De eerste geeft een *'open loop'-oplossing*: dit legt alle beslissingen vast die genomen moeten worden gedurende de planningperiode (in technische

termen: een stuurfunctie $u(t)$ van de tijd t). De laatste geeft een *'closed loop'* of *feedback* oplossing: deze oplossing geeft een strategie om beslissingen te nemen gedurende de planningsperiode afhankelijk van hoe de situatie in de toekomst zal zijn (in technische termen: een stuurfunctie $u(t, x)$ van de tijd t en de huidige toestand x). Op het eerste gezicht zou het kunnen lijken dat een 'closed loop' oplossing altijd de voorkeur verdient, wegens de grotere flexibiliteit. Maar dat is niet altijd het geval zoals het model van Kydland en Prescott dat boven gegeven is, laat zien. Als een beslisser het commitment kan maken om een Pontryagin-oplossing uit te voeren, dan is hij in een sterke positie. Een beslisser die een Bellman oplossing gebruikt, is kwetsbaar voor manipulaties.

Specifiek, in het model boven, komt de commitment oplossing ('om op $t = 1$ een laag belastingtarief te kiezen') overeen met de open loop oplossing of de Pontryagin oplossing van een dynamisch optimaliseringsprobleem; de geen-commitment oplossing ('om op $t = 1$ een hoog belastingtarief te kiezen') komt overeen met de Bellman oplossing of gesloten loop oplossing: op elk moment in de tijd kun je je optimale actie kiezen.

5. Wiskunde: fundamentaalstelling van de algebra

Iedere middelbare scholier leert een vorm van de fundamentaalstelling van de algebra: elk polynoom in een variabele is te ontbinden in lineaire en kwadratische factoren. Wie precies wil weten hoe de vork in de steel steekt ("het bewijs"), heeft een probleem. Ieder van de bekende bewijzen vereist gespecialiseerde voorkennis. Dit illustreert de *diepte* van deze stelling.

Fundamentaalstelling van de algebra
Iedere polynomiaalvergelijking

$$p(x) = a_0 + \cdots + a_n x^n = 0$$

in een complexe variabele x – van graad $n \geq 1$ heeft ten minste een oplossing.

Dit resultaat heeft op het eerste gezicht niets te maken met optimalisering.

Bewijs

(i) We bekijken het probleem

$$f(x) = |p(x)| \to \min,\ x \in \mathbb{C}.$$

Dit twee-variabelen-probleem ($x = x_1 + ix_2$ met $x_1, x_2 \in \Re$) heeft een oplossing $\widehat{x}$, omdat $x \to |p(x)|$ coercief is:

$$|p(x)| = |x|^n |a_n + \cdots + a_0 \frac{1}{x^n}|;$$

de eerste factor gaat naar $+\infty$ en de tweede naar $|a_n|$ voor $|x| \to +\infty$.

(ii) Laat $\tilde{x}$ een complex getal zijn waarvoor $p(\tilde{x}) \neq 0$. We drukken p uit als polynoom in de nieuwe variabele $y = x - \tilde{x}$:

$$p(x) = b_0 + b_k y^k + \cdots + b_n y^n$$

met $b_k \neq 0$. Merk op dat $b_0 = p(\tilde{x}) \neq 0$. Laat u een oplossing zijn van $b_0 + b_k y^k = 0$: dit is een van de k k-de wortels van het getal $-\frac{b_0}{b_k}$. Definieer de kromme lijn $y_\alpha = \alpha u,\ \alpha \geq 0$. Er geldt dat

$$f(\tilde{x} + y_\alpha) = |b_0 + (\alpha u)^k + o(\alpha^k)| = |(1 - \alpha^k)b_0 + o(\alpha^k)| < |b_0|$$

voor $\alpha > 0$ voldoende klein. Dit bewijst dat $\tilde{x}$ geen locaal minimum kan zijn, laat staan een globaal minimum.

(iii) Weglaten van de complexe getallen waarvoor we zojuist hebben uitgesloten dat ze minimaal zijn, leidt tot de conclusie dat alle locale oplossingen van het optimaliseringsprobleem wortels zijn van de vergelijking $p(x) = 0$.

(iv) De stelling is bewezen omdat $p(\widehat{x}) = 0$.

6. Conclusies lezing

- Vele problemen die op het eerste gezicht niets met optimalisatie te maken hebben, kunnen toch met optimaliseringsmethoden worden aangepakt.
- Er is een systematische methode beschikbaar om alle analytisch oplosbare optimaliseringsproblemen aan te pakken.
- Hoe overtuigend toepassingen van optimaliseringsmethoden in de financiering, economie en wiskunde zijn, hangt af van de impact op de *praktijk*, het toegevoegd *inzicht* in de economische werkelijkheid en de *diepte* van de verkregen resultaten, respectievelijk.

Dankbetuiging

Ik heb geprofiteerd van discussies over het materiaal in deze lezing met Jan Boone, Jan van de Craats, Marielle Non, Vladimir Protassov, Vladimir Tikhomirov, Shuzhong Zhang.

CWI SYLLABI

1 Vakantiecursus 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 Vakantiecursus 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 Vakantiecursus 1987: *De personal computer en de wiskunde op school.* 1987.

15 Vakantiecursus 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 Vakantiecursus 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 Vakantiecursus 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 Vakantiecursus 1990: *Getallentheorie en haar toepassingen.* 1990.

28 Vakantiecursus 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 Vakantiecursus 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 Vakantiecursus 1993: *Het reële getal.* 1993.

36 Vakantiecursus 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 Vakantiecursus 1995: *Kegelsneden en kwadratische vormen.* 1995.

41 Vakantiecursus 1996: *Chaos.* 1996.

42 H.C. Doets. *Wijzer in Wiskunde.* 1996.

43 Vakantiecursus 1997: *Rekenen op het Toeval.* 1997.

44 Vakantiecursus 1998: *Meetkunde, Oud en Nieuw.* 1998.

45 Vakantiecursus 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 Vakantiecursus 2000: *Is wiskunde nog wel mensenwerk?* 2000.

49 Vakantiecursus 2001: *Experimentele wiskunde.* 2001.

50 Vakantiecursus 2002: *Wiskunde en gezondheid.* 2002.

51 G.M. Hek (ed.). *Proceedings of the 42^{nd} European Study Group with Industry.* 2002.

52 Vakantiecursus 2003: *Wiskunde in het dagelijks leven.* 2003.

53 Vakantiecursus 2004: *Structuur in schoonheid.* 2004.

54 Vakantiecursus 2005: *De schijf van vijf – meetkunde, algebra, analyse, discrete wiskunde, stochastiek.* 2005.

MC SYLLABI

1.1 F. Göbel, J. van de Lune. Leergang besliskunde, deel l: 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. 1967.
1.6b B. Dorhout, J. Kriens, J.Th. van Lieshout. Leergang besliskunde deel 6b: wiskundige programmering. 1967
1.7a G. de Leve. Leergang besliskunde, deel 7a: dynamische programmering 1. 1969
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 7e: 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, RP. 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, EM. de Jager, N.M. Temme, R.J. de Vogelaere. Colloquium elliptische differentiaalvergelijkingen, deel 2. 1970.
10 l. 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 programma-correctheid. 1975.
22 R. Helmers, J. Oosterhoff, F.H. Ruymgaart, M.C.A. van Zuylen. Asymptotische methoden in de toe-tsingstheorie; toepassingen van naburigheid. 1976.
23.1 J.W. de Roever (red.). Colloquium onderwerpen uit de biomathematica, deel l. 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: eenstapsmethaden. 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, SJ. 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 1 B. 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 waerstanden 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 l. 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 procedures in ALGOL 60, vol. l: elementary procedures; vol. 2: algebraic evaluatons. 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, procedures in ALGOL 60, vol. 4: analytical evaluations; vol. 5A: analytical problems. part I. 1981
47.6 P.W. Hemker (ed.). NUMAL, procedures in ALGOL 60, vol 5B: analytical problems, part II. 1981
47.7 P.W. Hemker (ed.). NUMAL, 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 II. 1982.
49 T.H. Koomwinder (ed.) The structure of real semisimple Lie groups. 1982
50 H. Nijmeijer. Inleiding systeemtheorie. 1982.
51 P.J. Hoogendoorn (red.). Cursus cryptografie. 1983.