Papers dedicated to J.J. Seidel

TECHNISCHE HOGESCHOOL EINDHOVEN
NEDERLAND
ONDERAFDELING DER WISKUNDE
EN INFORMATICA

EINDHOVEN UNIVERSITY OF TECHNOLOGY
THE NETHERLANDS
DEPARTMENT OF MATHEMATICS AND
COMPUTING SCIENCE

PAPERS DEDICATED TO J.J. SEIDEL

edited by:

P.J. de Doelder, J. de Graaf and J.H. van Lint

EUT Report 84-WSK-03

ISSN 0167-9708

Coden: TEUEDE

Eindhoven

August 1984

Typewerk : Wetenschappelijk Secretariaat van de Onderafdeling der Wiskunde en Informatica: A.H.M. Brüssow-Hermens, J.G.W. Klooster-Derks en E.W. van Thiel-Niekoop.

Tekeningen: C.P. van Nieuwkasteele

ERRATA

Papers dedicated to J.J. Seidel

EUT Report 84-WSK-03

	er staat:	er moet staan:
blz. iv, r. 2	Jan Jacob Seidel	Johan Jacob Seidel
blz. 51, 3e kolom	.0217	0.014
	.0217	0.014
blz. 51, 4e kolom	.2483	.2583
	1.407	1.405
	1.406	1.405
	.081	.080
	.197	.195

INHOUDSOPGAVE

LEVENSLOOP VAN J.J. SEIDEL

Jan Jacob Seidel werd op 19 augustus 1919 geboren in Den Haag. Hij legde in 1937 het eindexamen Gymnasium β af aan het Stedelijk Gymnasium te Den Haag.

In datzelfde jaar begon hij zijn studie in de Wis- en Natuurkunde aan de Rijksuniversiteit te Leiden; in 1940 legde hij in die faculteit het candidaatsexamen A af. Wegens de sluiting van de Leidse Universiteit zette hij in 1941 de studie voort aan de Vrije Universiteit te Amsterdam, waar hij in de zomer van 1946 het doctoraal examen, met hoofdvak wiskunde, aflegde.

Vanaf maart 1946 tot juli 1946 was hij leraar aan de MTS voor bouwkunde te Amsterdam en van september 1946 tot september 1950 leraar aan het Vossiusgymnasium te Amsterdam. Ondertussen bereidde hij ook een proefschrift voor en hij promoveerde op 25 mei 1948 tot doctor in de Wis- en Natuurkunde aan de RU te Leiden, waarbij Prof.dr. J. Haantjes als promotor optrad. De titel van het proefschrift luidde: "De congruentie-orde van het elliptische vlak".

In 1950 werd hij benoemd als wetenschappelijk ambtenaar (instructeur) aan de TH Delft en in 1955 volgde zijn bevordering tot hoofdambtenaar. In het voorjaar van 1955 vertoefde hij met studieverlof in Rome aan de universiteit aldaar. In het cursusjaar 1955-1956 gaf hij een cursus Infinitesimaalrekening aan de RU te Leiden en vanaf 1 november 1955 nam hij de colleges Analytische Meetkunde II waar, die Haantjes wegens ziekte verhinderd was te geven.

Op 20 december 1955 werd hij voorgedragen tot adviseur van de commissie de Quay, die de oprichting van de THE voorbereidde. De benoeming vond plaats op 1 februari 1956. Bij Koninklijk Besluit no. 53 d.d. 5 november 1956 werd hij benoemd tot gewoon hoogleraar in de Wiskunde aan de THE. De ambtsaanvaarding vond plaats op 1 januari 1957. Zijn intreerede, getiteld "Wiskunde en Technisch Hoger Onderwijs" sprak hij uit op 25 februari 1958.

De werkzaamheden van Seidel in de jaren 1957-1984 kunnen worden onderscheiden naar Organisatorische en Bestuurlijke activiteiten, Onderwijsactiviteiten en Wetenschappelijke activiteiten.

Organisatorische en Bestuurlijke activiteiten

Behalve de vele werkzaamheden vanwege de opbouw van de Sectie Wiskunde, die in de beginjaren van de THE nodig waren, vervulde Seidel tientallen functies in bestuursorganen en in interne en externe commissies. De volgende opsomming is een kleine, vrij willekeurige greep daaruit:

1960-1966 Voorzitter van de Onderafdeling der Wiskunde.
Lid van het college van Rector Magnificus en Assessoren.

1970-1971 Secretaris van de senaat.

1963- Lid van de Sectie Wiskunde van de Academische raad. Later werd hij voorzitter van deze Sectie Wiskunde.

1969- Voorzitter van de Eindhovense commissie voor de wetenschappelijke begeleiding van het Moller instituut.
Lid van de adviescommissie exacte wetenschappen van de Nederlandse organisatie voor Zuiver Wetenschappelijk Onderzoek ZWO.

1972- Lid van de adviescommissie van het Mathematisch Centrum.
Lid van het bestuur van het Instituut voor Ontwikkeling van het wiskundeonderwijs IOWO.

1978- Curator van het Mathematisch Centrum.

1979- Lid van de adviescommissie Wiskunde van het Natuurkundig Laboratorium van NV Philips.

1982- Jurylid Prix Franqui-België.
Associated editor van "European Journal of Combinatorics", van "Combinatorica" en van "Linear Algebra and its applications".

Samen met Benders, de Bruijn en Veltkamp was Seidel wiskundig adviseur van de directie van Philips' Natuurkundig Laboratorium. Gesprekken met Teer, de Haan en anderen hebben ertoe geleid dat het aanstellingsbeleid van het Natuurkundig Laboratorium voor jonge wiskundigen gewijzigd is.

Onderwijsactiviteiten

Op het aan de TH Eindhoven gegeven onderwijs in de wiskunde heeft Seidel een duidelijk stempel gedrukt. In de begintijd van het bestaan van de TH heeft hij, eerst alleen en later samen met anderen, vorm gegeven aan de inhoud en de presentatie van het wiskundecurriculum ten behoeve van de ingenieursopleidingen. Gebruik makend van zijn Delftse ervaringen had hij daarop een visie ontwikkeld, die hij tot werkelijkheid heeft weten te maken. In de totstandkoming in 1960 van de opleiding tot wiskundig ingenieur in Eindhoven heeft hij een belangrijk aandeel gehad. De bloei en de reputatie van deze opleiding heeft hij mede bevorderd, doordat hij er in de tijd van zijn

voorzitterschap in geslaagd is om een voor de vervulling van deze taak adequate bemanning aan de Onderafdeling te verbinden. Een actieve rol heeft hij verder vervuld bij het tot stand brengen van een aan de eisen van de tijd aangepaste opleiding tot het verkrijgen van de onderwijsbevoegdheid in de wiskunde. Zo heeft hij met succes geijverd voor een evenwichtig pakket wiskunde-vakken voor de niet-wiskundige ingenieurs, die de bevoegdheid in de wiskunde wensten te behalen. Daarnaast heeft hij gezorgd voor de totstandkoming in de Onderafdeling van een groep didactiek, die hij voorts heeft georganiseerd en geleid. Ook buiten het verband van de TH heeft hij zich verdienstelijk gemaakt voor het voortgezet onderwijs. In de tijd voorafgaande aan de invoering van het nieuwe wiskundecurriculum in 1968, was hij lid van de Commissie Modernisering Leerplan Wiskunde en was hij betrokken bij de organisatie van de Heroriënteringscursussen voor leraren. Hij trad tevens als docent op bij deze cursussen. Uit zijn bestuurslidmaatschap van het IOWO bleek zijn belangstelling voor de modernisering van het wiskundecurriculum bij het voortgezet onderwijs.

Speciale vermelding verdient ook Seidel's voorzitterschap van de Commissie WIHBO (Wiskunde en Informatica bij het Hoger Beroeps Onderwijs). Deze commissie werd in 1979 opgericht en Seidel bekleedde het voorzitterschap tijdens de belangrijke beginjaren. Hij wist velen te mobiliseren om een bijdrage te leveren in de opzet en ondersteuning van het informatica-onderwijs in het HBO. (Leerplannen, begeleidingscommissie, bijscholingscursussen, enz.) Hij was een gepassioneerd leider van het geheel. Vergaderingen onder zijn voorzitterschap waren dikwijls echte "happenings".

Als meer specifiek resultaat van Seidel's onderwijsactiviteiten moet vermeld worden dat 6 studenten bij hem zijn afgestudeerd. Voorts is Seidel als eerste promotor opgetreden bij de promoties van W.H. Haemers op 30 oktober 1979 en A. Blokhuis op 30 september 1983.

Wetenschappelijke activiteiten

Het wetenschappelijke werk van Seidel begon met een proefschrift en een aantal artikelen die allemaal de niet-euclidische meetkunde betroffen. Dan volgt een periode, waarin hij alle energie en aandacht besteedde aan de opbouw van de Onderafdeling der Wiskunde van de THE en het opzetten van de wiskundecolleges. Enkele onderwijskundige publicaties getuigen van deze activiteiten.

Het belangrijkste deel van zijn werk betreft de theorie van sterk reguliere graphen, begonnen in 1966 na zijn aftreden als voorzitter van de Onderafdeling. In die tijd begon de zeer vruchtbare samenwerking met J.-M. Goethals (MBLE-Brussel), een samenwerking, waarbij vaak ook P. Delsarte (MBLE) en P.J. Cameron (Oxford) betrokken waren. De elf artikelen met Goethals als co-auteur moeten tot de belangrijkste van de lijst gerekend worden. De ontwikkeling van het idee van spherical designs leidde tot een terugkeer tot de oude liefde: de niet-euclidische meetkunde. Niet onvermeld mag blijven dat F.C. Bussemaker (THE) acht keer als co-auteur wordt vermeld.

De grote verdiensten van Seidel werden op 30 april 1975 gehonoreerd met het toekennen van een Koninklijke onderscheiding, t.w. Ridder in de Orde van de Nederlandsche Leeuw.

PUBLIKATIES VAN J.J. SEIDEL IN WETENSCHAPPELIJKE TIJDSCHRIFTEN E.D.

1. The congruence order of the elliptic plane
(with J. Haantjes).
Proc. Kon. Ned. Akad. Wet. Ser. A., 50 (1948),
892-894 (= Indag. Math. 9 (1947), 403-405).

2. De congruentie-orde van het elliptische vlak.
Universiteit van Leiden, 1948.
Thesis; iv + 71 pp.

3. Distance-geometric development of two-dimensional
euclidean, hyperbolical and spherical geometry, I, II.
Simon Stevin 29 (1952), 32-50; 65-76.

4. Angoli fra due sottospazi di uno spazio sferico od ellittico.
Rend. Accad. Naz. Lincei. (8) 16 (1954), 625-632.

5. An approach to n-dimenional euclidean and non-euclidean geometry.
Proc. of the Int. Math. Congress Amsterdam,
ed. J.C.H. Gerretsen and J. de Groot.
Vol. 2 (1954), 255.

6. Angles and distances in n-dimensional euclidean and non-euclidean
geometry, I, II, III.
Proc. Kon. Ned. Akad. Wet. Ser. A., 58
(= Indag Math. 17) (1955), 329-335; 336-340; 535-541.

7. De betekenis van het leerplan voor de toekomstige student.
Euclides 31 (1955), 245-256.

8. Afstandsmeetkunde.
Euclides 33 (1957), 161-165.

9. Wiskunde en Technisch Hoger Onderwijs.
Technische Hogeschool Eindhoven, 1958.
Inaugurale rede.

10. Wiskunde en Technisch Hoger Onderwijs.
Simon Stevin 32 (1958), 145-158.

11. On null vectors of certain semi-definite matrices
(with W. Peremans).
Simon Stevin 33 (1959), 101-105.

12. Mutually congruent conics in a net
(with J. van Vollenhoven).
Simon Stevin 37 (1963), 20-24.

13. Polytopes.
Math. Centrum, Amsterdam, 1966.
(Rapport Z.W.-055), 7 pp.

14. Remark concerning a theorem on eigenvectors of bounded linear operators.
Proc. Kon. Ned. Akad. Wet. Ser. A., 69
(= Indag. Math. 28) (1966), 358-359.

15. Equilateral point sets in elliptic geometry
(with J.H. van Lint).
Proc. Kon. Ned. Akad. Wet. Ser. A., 69
(= Indag. Math. 28) (1966), 335-348.

16. Strongly regular graphs of L_2 type and of triangular type.
Proc. Kon. Ned. Akad. Wet. Ser. A., 70
(= Indag. Math. 29) (1967), 188-196.

17. Orthogonal matrices with zero diagonal
(with J.M. Goethals).
Canad. J. Math. 19 (1967), 1001-1010.

18. Strongly regular graphs with (-1,1,0) adjacency matrix having eigenvalue 3.
Lin. Alg. and Appl. 1 (1968), 281-298.

19. Colloquium Discrete Wiskunde
(with P.C. Baayen and J.H. van Lint).
Math. Centrum, Amsterdam, 1968,
(Syllabus; 5); 108 pp.

20. Discrete Wiskunde.
Euclides 44 (1968/1969), 38-45.

21. Strongly regular graphs, (Waterloo).
Progress in Combinatorics; ed. W.T. Tutte.
Acad. Press Inc., New York, 1969, 185-197.

22. Quasiregular two-distance sets.
Proc. Kon. Ned. Akad. Wet. Ser. A., 72
(= Indag. Math. 31) (1969), 64-69.

23. Quasisymmetric block designs
(with J.M. Goethals).
Combinatorial Structures and their Applications,
ed. R. Guy.
Proc. Calgary Intern. Conference.
Gordon-Breach, New York 1970; 111-116.

24. A skew Hadamard matrix of order 36
(with J.M. Goethals).
J. Austr. Math. Soc. 11 (1970), 343-344.

25. Computerwiskunde.
Redacteur. Het Spectrum, Utrecht, 1969.
(Aulareeks; 407).

26. Strongly regular graphs derived from combinatorial designs
(with J.M. Goethals).
Canad. J. Math. 22 (1970), 597-614.

27. Symmetric Hadamard matrices of order 36
(with F.C. Bussemaker).
Int. Conference on Combinatorial Mathematics,
ed. A. Gewirtz and L.V. Quintas.
New York, Academy of Sciences, 1970, 66-79.
(Annals of the New York Academy of Sciences, 175).

28. A new family of nonlinear codes obtained from conference matrices
(with N,J, Sloane).
Int. Conference on Combinatorial Mathematics,
ed. A. Gewirtz and L.V. Quintas.
New York, Academy of Sciences, 1970.
(Annals of the New York Academy of Sciences, 175).

29. Symmetric Hadamard matrices of order 36
(with F.C. Bussemaker).
Technological University Eindhoven, 1970.
(Report 70-WSK-02).

30. A strongly regular graph derived from the perfect ternary Golay code
(with E.R. Berlekamp and J.H. van Lint).
A survey of combinatorial theory,
ed. J.N. Shrivastava.
Amsterdam, North-Holland, 1973; 25-30.

31. Orthogonal matrices with zero diagonal, part II
(with Ph. Delsarte and J.M. Goethals).
Canad. J. Math. 23 (1971), 816-832.

32. Equiangular lines
(with P.W.H. Lemmens).
J. of Algebra 24 (1973), 494-512.

33. Quadratic forms over GF(2)
(with P.J. Cameron).
Proc. Kon. Ned. Akad. Wet. Ser. A., 76
(= Indag. Math. 35) (1973), 1-8.

34. Equi-isoclinic subspaces of Euclidean spaces
(with P.W.H. Lemmens).
Proc. Kon. Ned. Akad. Wet. Ser. A., 76
(= Indag. Math. 35) (1973), 98-107.

35. Combinatorial designs.
Mathematical Recreations and Essays,
by W.W. Rouse Ball and H.S.M. Coxeter.
University of Toronto Press, 12th ed., 1974;
Chapter X; p. 271-311.

36. Van recreatie naar toepassing, van meetkunde naar codes, grafen en groepen.
Math. Centrum, Amsterdam, 1973.
(Syllabus 18), 31-37.

37. On two graphs and Shult's characterization of symplectic and orthogonal geometries over GF(2).
Technological University Eindhoven, 1973.
(Report 73-WSK-02).

38. The mathematical education of engineers, and the education of mathematical engineers in The Netherlands.
Bull. Inst. Math. Appl. 9 (1973), 305-307.

39. A survey of two-graphs.
Proc. Intern. Coll. Teorie Combinatorie, (Roma 1973).
Accad. Naz. Lincei, Roma, 1976; 481-511.

40. The regular two-graph on 276 vertices
(with J.M. Goethals).
Discrete Mathematics 12 (1975), 143-158.

41. Graphs and two-graphs.
5th Southeastern Conference on Combinatorics,
Graph Theory and Computing,
ed. F. Hoffmann.
Utilitas Math. Publ. Inc., Winnipeg, 1974; 125-143.

42. Prima introduzione alla matematica discreta e alla teoria dei codici,
Archimede (1973), 235-241.

43. Finite geometric configurations.
Foundation of geometry,
ed. P. Scherk.
University of Toronto Press, 1976; 215-250.

44. Bounds for systems of lines, and Jacobi polynomials
(with Ph. Delsarte and J.M. Goethals).
Philips Research Reports 30 (1975), 91-105.
Issue in honour of C.J. Bouwkamp.

45. Metric problems in elliptic geometry.
The geometry of metric spaces,
ed. L.M. Kelly.
Springer, Berlin, 1975. (Lecture Notes in
Mathematics 490); 32-43.

46. Line graphs, root systems, and elliptic geometry
(with P.J. Cameron, J.M. Goethals and E.E. Shult).
J. of Algebra 43 (1976), 305-327.

47. Spherical codes and designs
(with Ph. Delsarte and J.M. Goethals).
Geometriae Dedicata 6 (1977), 363-388.

48. Computer investigation of cubic graphs
(with F.C. Bussemaker, S. Cobeljić and D.M. Cvetković).
Technological University Eindhoven, 1976.
(Report 76-WSK-01).

49. Cubic graphs on ≤ 14 vertices
(with F.C. Bussemaker, S. Cobeljić and D.M. Cvetković).
J. Comb. Theory B, 23 (1977), 234-235.

50. Graphs related to exceptional root systems
(with F.C. Bussemaker and D.M. Cvetković).
Combinatorics,
ed. A. Hajnal and V.T. Sós.
Amsterdam, North-Holland, 1978; 185-191.
(Colloquia Mathematica Societatis János Bolyai; 18).

51. Graphs related to exceptional root systems
(with F.C. Bussemaker and D.M. Cvetković).
Technological University Eindhoven, 1976.
(Report 76-WSK-05).

52. Eutactic stars.
Combinatorics,
ed. A. Hajnal and V.T. Sós.
Amsterdam, North-Holland, 1978; 983-999.
(Colloquia Mathematica Societatis János Bolyai; 18).

53. On two-distance sets in Euclidean space
(with D.G. Larman and C.A. Rogers).
Bull. London Math. Soc. 9 (1977), 261-267.

54. Strongly regular graphs having strongly regular subconstituents
(with P.J. Cameron and J.M. Goethals).
J. of Algebra 55 (1978), 261-267.

55. The pentagon.
Proc. Bicentennial Congress Wisk. Genootschap, vol. I,
ed. P.C. Baayen, D. van Dulst, J. Oosterhoff.
Math. Centrum, Amsterdam, 1979, 81-96.
(Math. Centre Tracts; 100).
+
Intern. Conference on Combinatorial Mathematics,
ed. A. Gewirtz and L.V. Quintas.
New York Academy of Sciences, 1979, 497-507.
(Annals of the New York Academy of Sciences, 319).

56. The Krein condition, spherical designs, Norton algebras and permutation groups
(with P.J. Cameron and J.M. Goethals).
Proc. Kon. Ned. Akad. Wet. A., 81
(= Indag. Math. 40) (1978), 196-206.

57. Spherical designs
(with J.M. Goethals).
Relations between combinatorics and other parts of mathematics,
ed. D.K. Ray-Chaudhuri.
Amer. Math. Soc., Providence, 1979 (proc. Symp. Pure Math.; 34),
255-272.

58. Two-graphs; a second survey
(with D.E. Taylor).
Algebraic Methods in Graph Theory,
ed. L. Lovasz and V.T. Sós.
Amsterdam, North-Holland, 1981; 689-711.
(Colloquia Mathematica Societatis János Bolyai; 25).

59. Matematicko obrazovanje inzenjera i skolovanje matamatickih inzenjera u holandiji.
Diskretne matematicke strukture,
by D. Cvetković.
Univerzitet u Beogradu, 1978; 146-150.

60. Strongly regular graphs, an introduction.
Surveys in Combinatorics;
ed. B. Bollobás.
Cambridge UP, 1979.
(London Math. Soc. Lecture Note Series; 38); 157-180.

61. Cubature formulae, polytopes and spherical designs
(with J.M. Goethals).
The Geometric Vein;
ed. C. Davis, B. Grünbaum and F.A. Sherk.
Springer, Berlin, 1981; 203-218.

62. Tables of two graphs
(with F.C. Bussemaker and R. Mathon).
Technological University Eindhoven, 1979.
(Report 79-WSK-05).

63. The football
(with J.M. Goethals).
Nieuw Archief voor Wiskunde 29 (1981), 50-58.

64. Graphs and two-distance sets.
Combinatorial Mathematics VIII,
ed. K.L. McAvaney.
Springer, Berlin, 1981.
(Lecture Notes in Math.; 884), 90-98.

65. Discrete hyperbolic geometry
(with A. Neumaier).
Combinatorica, 3 (1983), 219-237.

66. Tables of two-graphs
(with F.C. Bussemaker and R. Mathon).
Combinatorics and Graph Theory,
ed. S.B. Rao.
Springer, Berlin, 1981.
(Lecture Notes in Math., 885), 70-112.

67. Regular non-Euclidean pentagons.
Nieuw Archief voor Wiskunde 30 (1982), 161-166.

68. The addition formula for hyperbolic space
(with E. Bannai, A. Blokhuis and Ph. Delsarte).
J. Comb. Theory A 36 (1984), 332-341.

69. Graphs and association schemes, algebra and geometry
(with A. Blokhuis and H.A. Wilbrink).
Eindhoven University of Technology, 1983.
(EUT-Report; 83-WSK-02).

70. Conference matrices from projective planes of order 9
(with C.A.J. Hurkens).
European J. Combinatorics.

71. Polytopes and non-Euclidean geometry.
Mitteil. Math. Semin. Giessen, 163 (1984), I, 1-17.
(Coxeter Festschrift).

72. Meetkunde van de ruimte
(with P.W.H. Lemmens).
Vakantiecursus C.W.I.
Zomer 1984.

THE BERNOULLI NUMERATORS

by

P.J. van Albada

Dedicated to J.J. Seidel on the occasion of his retirement.

INTRODUCTION

If written in the reduced form the Bernoulli numbers have a denominator which is a multiple of 6 and a divisor of $2(2^{2m}-1)$. If written with this latter number as denominator, the numerators themselves form a sequence of integers which can be defined independently of the Bernoulli numbers, but quite in an analogous way. Where numbers B_{2m} appear in the expansion of $\frac{u}{e^u-1}$, numbers C_{2m-1} appear in the expansion of $\frac{2}{e^u+1}$ with $C_{2m-1} = -2(2^{2m}-1)B_{2m}$. Where numbers B_m^* (we write the asterisk here to avoid confusion between both notations) are defined as $\frac{(2m)!}{2^{2m-1}\pi^{2m}} \sum_{n=1}^{\infty} n^{-2m}$, numbers C_m^* can be defined as $\frac{4(2m)!}{\pi^{2m}} \sum_{n=1}^{\infty} (2n-1)^{-2m}$.

These integers appear in Bernoulli-like polynomials too; in fact these polynomials can be defined as coefficients of $\frac{t^k}{k!}$ in the expansion of $\frac{e^{2nt}-1}{e^t+1}$. They can also be defined algebraically as polynomials in n equivalent to $\sum_{i=0}^{2n-1} (-1)^{i+1} i^k$.

1. We start with algebra and write $\varphi(k,n) = \sum_{i=0}^{2n-1} (-1)^{i+1} i^k$. We find directly

(a) $\varphi(k,1) = 1 \quad (k = 1,2,\dots)$ and

(b) $\varphi(k,n+1) - \varphi(k,n) = (2n+1)^k - (2n)^k$.

The difference equation (b) predicts that $\varphi(k,n)$ will come out as a polynomial in n of degree k which by (b) is determined apart from a constant. This constant can be obtained from (a).

The equations (a) and (b) also define $\varphi(k,n)$ for non-positive values of n. We find

(1) $\qquad \varphi(k,0) = 0,$

further $\varphi(k,-1) = -(-1)^k + (-2)^k$, and generally

(2) $\qquad \varphi(k,-n) = (-1)^{k-1}\varphi(k,n) + (-2n)^k$.

If we write $\varphi(k,n) = \sum_{i=0}^{k} a_i(k) n^{k-i}$ we see from (1) that $a_k(k) = 0$ and from (2) that $a_{2i}(k) = 0 \quad (i > 0$, while $a_0 = 2^{k-1})$.

From (b) we have

$$(3) \qquad \sum_{j=1}^{k} \binom{k}{j} (2n)^{k-j} = \sum_{i=0}^{k-1} a_i(k) \sum_{j=i+1}^{k} \binom{k-i}{j-i} n^{k-j} ,$$

whence

$$\binom{k}{j} 2^{k-j} = \sum_{i=0}^{j-1} a_i(k) \binom{k-i}{j-i}.$$

We write $a_i(k) = \dfrac{2^{k-i-1}}{i+1} \binom{k}{i} c_i(k)$.

From (4) we obtain

$$(5) \qquad 1 = \sum_{i=0}^{j-1} c_i(k)\binom{j}{i}\frac{2^{j-i-1}}{i+1} \quad .$$

We observe that the $c_i(k)$ do not depend on k. We will write them C_i in future. If then we write (5) in the form

$$(6) \qquad C_{j-1} = 1 - \sum_{i=0}^{j-2} C_i\binom{j}{i}\frac{2^{j-i-1}}{i+1} = 1 - \sum_{i=0}^{j-2} C_i\binom{j}{i+1}\frac{2^{j-i-1}}{j-i} \quad ,$$

we easily obtain in succession $C_0 = 1$, $C_1 = -1$, $C_3 = 1$, $C_5 = -3$, $C_7 = 17$, $C_9 = -155$, $C_{11} = 2073$, $C_{13} = -38227$, $C_{15} = 929569$, etc. As far as we go, the C_i remain integers.

However, even if for $i < 2N$ all C_i are integers (6) alone cannot guarantee that also C_{2N+1} is an integer.

Studying now

$$(7) \qquad \sum_{i=0}^{2n-1} (-1)^{i+1}(i+1)^k = \sum_{j=0}^{k}\binom{k}{j}\varphi(j,n) = (2n)^k - \varphi(k,n) + 1 \quad ,$$

we obtain, selecting the coefficients of n only

$$(8) \qquad \sum_{j=1}^{k-1}\binom{k}{j} C_{j-1} + 2C_{k-1} = 0.$$

In (6) if all C_i $(0 \le i \le j-2)$ were integrals C_{j-1} could be fractional because some of the cofactors $\binom{j}{i+1}\frac{2^{j-i-1}}{j-i}$ of the C_i are fractional. But in the quotient $\frac{2^{j-i-1}}{j-i}$ the numerator contains more factors 2 than does the denominator so C_{j-1}, if fractional, cannot have an even denominator.

Remembering this, we can conclude from (8) that C_{k-1} is an integer if all C_i $(i < k-2)$ are integral.

Once the $_i$ have been defined from (6) or from (8) starting with $C_0 = 1$, we can give the general formula for $\varphi(k,n)$:

$$(9) \qquad \sum_{i=0}^{2n-1} (-1)^{i+1} {}_i k = \sum_{0}^{k-1} C_i \binom{k}{i+1} \frac{2^{k-i-1}}{k-i} n^{k-i} ,$$

a polynomial with integral coefficients, as can be concluded from (6) and (9) together.

2. While

$$\sum_{k=1}^{\infty} \varphi(k,n) \frac{t^k}{k!} = \sum_{k=1}^{\infty} \frac{t^k}{k!} \sum_{i=1}^{2n-1} (-1)^{i+1} {}_i k = \sum_{i=0}^{2n-1} (-1)^{i+1} \sum_{k=1}^{\infty} \frac{(it)^k}{k!} =$$

$$= \sum_{i=0}^{2n-1} (-1)^{i+1} (e^{it} - 1) = \frac{e^{2nt} - 1}{e^t + 1} ,$$

the latter function can be used as the generating function for the polynomials $\varphi(k,n)$.

3. The generating function for the coefficients C_i is the polynomial $\frac{2}{e^t + 1}$; the coefficients of $\frac{t^k}{(k+1)!}$ in the expansion of this quotient satisfy equations (6) as can easily be checked.

4. If we compare $\varphi(k,n) = \sum_{i=0}^{2n-1} (-1)^{i+1} {}_i k$ with $f(k,n) = \sum_{i=0}^{n} (i)^k$ we see that

$$f(2m,2n) - 2^{2m+1} f(2m,n) = \varphi(2m,n) - (2n)^{2m} .$$

As known $f(2m,n)$ is a polynomial the last term of which is $B_{2m}n$. The polynomial

$$\varphi(2m,n) := \sum_{i=0}^{2m} a_i(2m)\, n^{2m-i} = \sum_{i=0}^{2m} \frac{2^{2m-i-1}}{i+1} \binom{2m}{i} C_i\, n^{2m-i}$$

contains for $i = 2m-1$ the term

$$\frac{2^0}{2m} \binom{2m}{2m-1} C_{2m-1}\, n = C_{2m-1}\, n.$$

We obtain the relation

$$B_{2m}(2 - 2^{2m+1}) = C_{2m-1} \quad . \tag{10}$$

5. In the complex plane the function $\frac{2}{e^z + 1}$ possesses simple poles on the imaginary axis in the points $\pi i \pm 2k\pi i$, each with residue -2. Hence

$$\frac{2}{e^z+1} = 1 + \sum_{m=1}^{\infty} \left(\frac{-2}{z-(2m-1)\pi i} - \frac{2}{(2m-1)\pi i} \right) + \left(\frac{-2}{z+(2m-i)\pi i} + \frac{2}{(2m-1)\pi i} \right) =$$

$$= 1 + \sum_{m=1}^{\infty} \frac{-4z}{z^2 + (2m-1)^2 \pi^2} = 1 + 4 \sum_{k=1}^{\infty} (-1)^k \frac{z^{2k-1}}{\pi^{2k}} \sum_{m=1}^{\infty} (2m-1)^{-2k},$$

for $|z| < \pi$. It follows that

$$C_{2m-1} = (-1)^m \frac{4(2m)!}{\pi^{2m}} \sum_{n=1}^{\infty} (2n-1)^{-2m} \quad . \tag{11}$$

Since $B_m^* = \frac{(2m)!}{2^{2m-1}\pi^{2m}} \sum_{n=1}^{\infty} n^{-2m}$ and $B_{2m} = (-1)^{m+1} B_m^*$ this leads again to the relation (10).

UNIQUENESS OF A ZARA GRAPH ON 126 POINTS AND NON-EXISTENCE OF A COMPLETELY REGULAR TWO-GRAPH ON 288 POINTS

by

A. Blokhuis and A.E. Brouwer

Dedicated to J.J. Seidel on the occasion of his retirement.

Abstract. There is a unique graph on 126 points satisfying the following three conditions:

(i) every maximal clique has six points;

(ii) for every maximal clique C and every point p not in C, there are exactly two neighbours of p in C;

(iii) no point is adjacent to all others.

Using this we show that there exists no completely regular two-graph on 288 points, cf. [4], and no (287,7,3)-Zara graph, cf. [1].

1. INTRODUCTION

A *Zara graph* with clique size K and nexus e is a graph satisfying:

(i) every maximal clique has size K;

(ii) every maximal clique has nexus e (i.e., any point not in the clique is adjacent to exactly e points in the clique).

For a list of examples, due to Zara, we refer to [1] and [6]. In this note we prove that there is only one Zara graph on 126 points with clique size 6

and nexus 2, which also has the property that no point is adjacent to all others. This graph, Z^*, is defined as follows:

Let W be a 6-dimensional vector space over GF(3), together with the bilinear form $\langle x|y\rangle = x_1y_1 + \ldots + x_6y_6$. Points of Z^* are the one-dimensional subspaces of W generated by a point x of norm 1, i.e., $\langle x|x\rangle = 1$. Two such subspaces are adjacent if they are orthogonal: $\langle x\rangle \sim \langle y\rangle$ iff $\langle x|y\rangle = 0$.
In the following section Z will denote any Zara graph on 126 points with $K = 6$ and $e = 2$.

2. BASIC PROPERTIES OF ZARA GRAPHS

A *singular subset* of a Zara graph is a set of points which is the intersection of a collection of maximal cliques. Let S denote the collection of singular subsets. From [1] we quote the main theorem for Zara graphs (a graph is called *coconnected* if its completement is connected):

THEOREM 1. Let G be a coconnected Zara graph. There exists a rank function $\rho : S \to \mathbb{N}$ such that

(i) $\rho(\emptyset) = 0$

(ii) If $\rho(x) = i$ and C is a maximal clique containing x while $p \in C\backslash x$, then $\exists y \in S$ with $\rho(y) = i+1$ and $x \cup \{p\} \subset y \subset C$.

(iii) $\exists r : \rho(c) = r$ for all maximal cliques C.

(iv) $\exists R_0, R_1, \ldots, R_r : \rho(x) = i \Rightarrow x$ is in R_i maximal cliques.

(v) $\exists K_0, K_1, \ldots, K_r : \rho(x) = i \Rightarrow |x| = K_i$.

(vi) The graph defined on the rank 1 sets by $x \sim y$ iff $\xi \sim \eta$ for all $\xi \in x$ and $\eta \in y$ is strongly regular. □

The number r is called the *rank* of the Zara graph. A coconnected rank 2 Zara graph with e = 1 is essentially a *generalized quadrangle*. In this case singular subsets are the empty set (rank 0) the points (rank 1) and the maximal cliques (rank 2). This graph is also denoted by $GQ(K-1,R_1-1)$. As an example we mention GQ(4,2). This is a graph on 45 points, maximal cliques have size 5, and each point is in three maximal cliques. This graph is unique [5] and has the following description:

Let W be a 4-dimensional vector space over GF(4) with hermitian form $\langle x|y\rangle = x_1\bar{y}_1 + \ldots x_4\bar{y}_4$, where $\bar{y}_i = y_i^2$. Points are the one-dimensional subspaces $\langle x\rangle$ with $\langle x|x\rangle = 0$ and $\langle x\rangle \sim \langle y\rangle$ if $\langle x|y\rangle = 0$ (and $\langle x\rangle \neq \langle y\rangle$). Another description of this graph is the following: Let W' be a 5-dimensional vector space over GF(3) with bilinear form $\langle x|y\rangle = x_1y_1 + \ldots + x_5y_5$. Points are the one-dimensional subspaces $\langle x\rangle$ with $\langle x|x\rangle = 1$ and $\langle x\rangle \sim \langle y\rangle$ if $\langle x|y\rangle = 0$.

From the main theorem on Zara graphs one can prove:

THEOREM 2. Z is a strongly regular graph, with $(v,k,\lambda,\mu) = (126,45,12,18)$. Each point is in 27 maximal cliques, each pair of adjacent points in 3. The induced graph on the neighbours of a given point is (isomorphic to) GQ(4,2). □

3. A FEW REMARKS ON GQ(4,2), Z^* AND FISCHER SPACES

The following facts can be checked directly from the description of GQ(4,2) and Z^* and the definition of Z. If x and y are points at distance two in the graph G then $\mu_G(x,y)$ (or just $\mu(x,y)$) denotes the induced graph on the set of common neighbours of x and y in G.

Fact 1. If $x \not\sim y$ in Z then $\mu(x,y)$ is a subgraph of GQ(4,2) on 18 points, regular with valency 3. If $x \not\sim y$ in Z^* then $\mu(x,y) \simeq 3 \times K_{3,3}$.

Fact 2. GQ(4,2) contains 40 subgraphs isomorphic to $3 \times K_{3,3}$. Through each 2-claw (i.e. $K_{1,2}$) in GQ(4,2) there is a unique $3 \times K_{3,3}$ subgraph, even a unique $K_{3,3}$.

Let $x \in Z^*$. Let $\Gamma(x)$ denote the induced graph on the neighbours of x, $\Delta(x)$ the induced graph on the non-neighbours, different from x. $\Gamma(x) \simeq GQ(4,2)$ and each point $y \in \Delta(x)$ determines the subgraph $K_y \simeq 3 \times K_{3,3}$ in $\Gamma(x)$, where $K_y = \mu(x,y)$

Fact 3. To each subgraph $K' \simeq 3 \times K_{3,3}$, of $\Gamma(x)$ there correspond exactly two points $y,y' \in \Delta(x)$, such that $K_y = K_{y'} = K'$. Note that $y \not\sim y'$.

This property can be used to show that Z^* is a *Fischer space*.

DEFINITION. A *Fischer space* is a linear space (E,L) such that

(i) All lines have size 2 or 3;

(ii) For any point x, the map $\sigma_x : E \to E$, fixing x and all lines through x, and interchanging the two points distinct from x on the lines of size 3 through x, is an automorphism.

THEOREM 3. There is a unique Fischer space on 126 points with 45 two-lines on each point.

The proof of this fact can be found in [2] p. 14. □

4. THE UNIQUENESS PROOF, PART I

Using a few lemmas, it will be shown that Z carries the structure of a Fischer space. By Theorem 3 then $Z \simeq Z^*$.

Notation: For a subset S of Z, we denote by $S^\perp$ the induced subgraph on the set of points adjacent to all of S.

LEMMA 1. Let {a,b,c} be a two-claw in Z: $a \sim b$, $a \sim c$, $b \not\sim c$. Then $\{a,b,c\}^\perp \simeq \bar{K}_3$ and there is a unique point $d \sim a$ such that $\{a,b,c,d\}^\perp = \{a,b,c\}^\perp$. Moreover, $d \not\sim b$, $d \not\sim c$.

Proof. Apply fact 2 to $\Gamma(a) \simeq GQ(4,2)$. □

LEMMA 2. Let $a \not\sim b$ in Z. Then $\mu(a,b) \simeq 3 \times K_{3,3}$.
This is the *main lemma*; the proof will be the subject of the next section. □

LEMMA 3. Let $a \not\sim b$ in Z. There is a unique point $c \in Z$ such that $\{a,b\}^\perp = \{a,b,c\}^\perp$. Moreover, $c \not\sim a$, $c \not\sim b$.

Proof. Consider a 2-claw {x,y,z} in $\mu(a,b)$. By Lemma 1 there is a point c in $\{x,y,z\}^\perp$ and $c \not\sim a$, $c \not\sim b$. By Lemma 2 $\mu(a,b) \simeq 3K_{3,3}$ and by fact 2 this subgraph of $\Gamma(a)$ is unique, hence $\mu(a,b) = \mu(a,c)$. □

THEOREM 4. Z carries the structure of a Fischer space with 126 points and 45 two-lines on each point.

Proof. Let the two-lines correspond to the edges of Z, the 3-lines to the triples {a,b,c} as in Lemma 3. This turns Z into a linear space with 45 two-lines on each point. It remains to be shown that σ_x is an automorphism for all $x \in Z$. Since $\sigma_x^2 = 1$ it suffices to show that $y \sim z$ implies $\sigma_x(y) \sim \sigma_x(z)$. The only non-trivial case is when $y,z \in \Delta(x)$. Let $Y = \Gamma(y) \cap \Delta(x)$, $Y' = \Gamma(\sigma_x(y)) \cap \Delta(x)$. Then $Y \cap Y' = \emptyset$ and $|Y| = |Y'| = 27$.

Since $|\{y,u,x\}^\perp| = 6$ for all $u \in Y$ (there are three maximal cliques passing through y, and x has two neighbours on each of them), and since $\mu(y,x) = \mu(\sigma_x(y),x)$, we also have $|\{y',u,x\}^\perp| = 6$ for $u \in Y$ and similarly $|\{y,u',x\}^\perp| = 6$ for $u' \in Y'$.

Counting edges between $\mu(x,y)$ and $\Delta(\infty)$ it follows that the average of $|\{y,u,x\}^\perp|$, with $u \in U = \Delta(\infty) \setminus (Y \cup Y' \cup \{y\} \cup \{\sigma_x(y)\})$ is 9. Consider an edge in $\mu(x,y) = \mu(x,y')$. There are three maximal cliques passing through that edge, containing x,y,y' respectively. Hence $\{y,u,x'\}$ is a coclique for $u \in U$, whence $|\{y,u,x\}^\perp| \leq 9$. Combining this yields $|\{y,u,z\}^\perp| = 9$ for all $u \in U$.

Next, consider a point z in Y. Since $\mu(x,z) = \mu(x,\sigma_x(z))$, we must have $\sigma_x(z) \in Y$ or $\sigma_x(z) \in Y'$. If $\sigma_x(z) \in Y$, then $y \sim z$ and $y \sim \sigma_x(z)$ but $y \sim x$, contradiction. Hence, $z \in Y'$, i.e., $\sigma_x(y) \sim \sigma_x(z)$. □

This finishes the uniqueness. It remains to prove Lemma 2.

5. THE UNIQUENESS PROOF, PART II: PROOF OF THE MAIN LEMMA

Main lemma. Let $\infty \not\sim \infty'$ in Z. Then $\mu(\infty,\infty') \simeq 3K_{3,3}$.

The proof will be split into a number of lemmas.

LEMMA 4. Let $S = \{a,b,c,d\}$ be a square in Z, i.e., $a \sim b \sim c \sim d \sim a$ and $a \not\sim c$, $b \not\sim d$. Then $|S^{\perp}| \in \{0,1,3\}$.

Proof. Clearly $S^{\perp}$ has at most three points, so it suffices to show that two points is impossible.

Let $\infty,\infty' \in S^{\perp}$. By Lemma 1 there is a point a' such that $\{d,a,b\}^{\perp} = \{a',\infty,\infty'\}$. Similarly there are points b',c',d'. If two of the points a',b',c',d' coincide, then we have found a third point adjacent to all of S. Hence, assume they are all different. There are three maximal cliques containing ab. One contains ∞, another ∞', whence the third one contains a' and b'. Hence $a' \sim b' \sim c' \sim d' \sim a'$. Considering again the clique $\{a,b,a',b'\}$, notice that $c' \not\sim a$, $c' \sim b$ and $c' \sim b'$. It follows that $c' \not\sim a'$ and similarly $b' \not\sim d'$. The situation is summarized in figure 1 where $A = \{a,a'\}$.

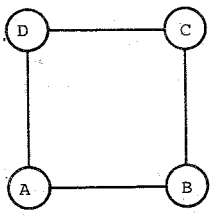


figure 1

Using the Zara graph property it follows that the picture can be completed to figure 2:

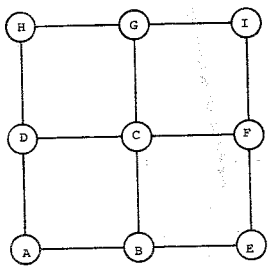


figure 2

where $E = \{e,e'\}$ etc.: Indeed, the clique $\{a,a',b,b'\}$ can be completed with points e,e'. Similarly DC can be completed and $\{e,e'\} \cap \{f,f'\} = \emptyset$. Having found E,F,G,H, complete the clique $\{e,e',f,f'\}$ using $\{i,i'\}$. Since i and i' have no neighbours in A,B,C,D, they must be adjacent to G and H.

Now ∞ and ∞' have one neighbour in each of A,B,C,D. It follows that both are adjacent to i,i'. However, there are three maximal

cliques through I, two of them already visible, whence ∞ and ∞' must be in the third clique. This is a contradiction since $\infty \not\sim \infty'$. The conclusion is that $a' = b' = c' = d'$ is the third point in $S^{\perp}$. □

LEMMA 5. If $\infty \not\sim \infty'$ in Z and $\mu(\infty,\infty')$ contains a square, then $\mu(\infty,\infty') \simeq 3K_{3,3}$, and there is a unique point ∞'' such that $\{\infty,\infty',\infty''\}^{\perp} = \mu(\infty,\infty')$.

Proof. Let $S = \{a,b,c,d\}$ be a square in $\mu(\infty,\infty')$. From the previous lemma it follows that there is a third point, e, adjacent to the square $\{\infty,a,\infty',c\}$. Similarly there is a point f adjacent to $\{\infty,b,\infty',d\}$, and $\{a,b,c,d,e,f\}$ is a $K_{3,3}$ in $\mu(\infty,\infty')$. Now $\mu(\infty,\infty')$ is a subgraph of $\Gamma(\infty) \simeq GQ(4,2)$ with 18 points and valency 3, containing a $K_{3,3}$. This is enough to guarantee that $\mu(\infty,\infty') \simeq 3K_{3,3}$. Let ∞'' be the third point adjacent to S. Since S is in a unique $K_{3,3}$ in $\Gamma(\infty)$ it follows that $\mu(\infty,\infty'') = \mu(\infty,\infty')$. □

LEMMA 6. Let $a,b,c \in Z$ with $|\{a,b,c\}^{\perp}| = 18$. Then $\{a,b,c\}^{\perp} \simeq 3K_{3,3}$.

Proof. First note that $\{a,b,c\}$ is a coclique. Let $M = \{a,b,c\}^{\perp}$ and $A = \Gamma(a)\backslash M$; B and C are defined similarly. Finally $R = Z\backslash(A \cup B \cup C \cup M \cup \{a\} \cup \{b\} \cup \{c\})$.

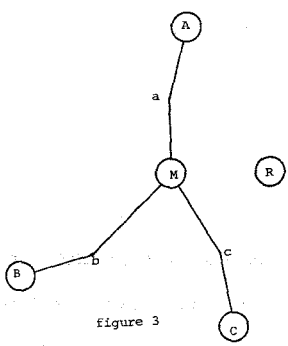


figure 3

$|A| = |B| = |C| = 27$

$|R| = 24, \quad |M| = 18$.

Two adjacent points in M have twelve common neighbours, three in A,B and C and none in R. It follows that the neighbours of a point $r \in R$ in M form a coclique. A point $m \in M$ has three neighbours in M, nine in

A,B and C (since Z is strongly regular with $\lambda = 12$). Hence m has twelve neighbours in R. Since the neighbours of $r \in R$ in M form a coclique, r has at most nine neighbours in M. But $9 \times 24 = 12 \times 18$, so it is exactly nine. If M is connected, there are at most two nine cocliques in M, whence at least twelve points of R are adjacent to the same 9-coclique. If there is an edge between two of the twelve we have a contradiction, if not also Hence M is disconnected. In this case however, one easily sees that M contains a square and hence $M \simeq 3K_{3,3}$. □

From now on we will identify $\Gamma(\infty) \simeq GQ(4,2)$ with the set of isotropic points in PG(3,4) w.r.t. a unitary form.

For $a \in \Delta(\infty)$ let $M_a = \mu(a,\infty)$. The graph M_a has 18 vertices and is regular of valency 3. By Lemma 5, if M_a contains a square, then $M_a \simeq 3K_{3,3}$. A computer search for all 18-point subgraphs of valency 3 and girth ≥ 5 of GQ(4,2) reveals that such a graph is necessarily (connected and) bipartite, i.e., a union of two ovoids. Now GQ(4,2) contains precisely two kinds of ovoids, plane ovoids and tripod ovoids (cf. [3]).

Let $x \in PG(3,4)\backslash U$, where U is the set of isotropic points.

A *plane ovoid* is a set of the form $x^{\perp} \cap U$.

A *tripod ovoid* (on x) is a set of the form

$$\bigcup_{i=1}^{3} xz_i \cap U ,$$

where $\{x,z_1,z_2,z_3\}$ is an orthonormal basis. On each non-isotropic point there are four tripod ovoids. Since two plane ovoids always meet, we find that each set M_a is one of the following (where T_x denotes some tripod ovoid on x):

I. $(x^{\perp} \cup T_x) \cap U$ $\quad (M_a \cong 3K_{3,3}$ in this case).

II. $(x^{\perp} \cup T_z) \cap U$ $\quad$ where $z \in x^{\perp}$ and $x \notin T_z$.

III. $(T_z \cup T_z') \cap U$, $\quad$ the union of two tripod ovoids on the same point.

(Note that $(T_x \cup T_z) \cap U$ for $z \in x^{\perp}$ and xz in T_z but not in T_x does contain squares, in fact $K_{2,3}$'s.)

If $a \sim b$, $a,b \in \Delta(\infty)$, then ∞ has two neighbours on each of the three 6-cliques on the edge ab, so that $M_a \cap M_b \cong 3K_2$.
By studying the intersections between sets of the three types, I,II,III we shall see that necessarily all sets M_a are of type I. Let us prepare this study by looking at the intersections of two ovoids in GQ(4,2).

A. $|x^{\perp} \cap y^{\perp} \cap U| = \begin{cases} 9 \text{ if } x = y ; \\ 3 \text{ if } x \perp y ; \\ 1 \text{ otherwise.} \end{cases}$

B. $|x^{\perp} \cap T_z \cap U| = \begin{cases} 0 \text{ if } x = z \text{ or } (z \in x^{\perp} \text{ and } x \notin T_z) ; \\ 6 \text{ if } z \in x^{\perp} \text{ and } x \in T_z ; \\ 2 \text{ otherwise.} \end{cases}$

C. $|T_x \cap T_z \cap U| = \begin{cases} 9 \text{ if } T_x = T_z ; \\ 3 \text{ if } z \in x^{\perp} \text{ and } xz \text{ occurs in both or none of } T_x, T_z ; \\ 0 \text{ if } (x = z \text{ and } T_x \neq T_z) \text{ or } (z \in x^{\perp} \text{ and } xz \text{ in one of } T_x, T_z); \\ 4 \text{ if } z \neq x \text{ and } z \notin x^{\perp} \text{ and } (xz)^{\perp} \text{ meets } T_x \cap T_z; \\ 1 \text{ otherwise.} \end{cases}$

Next let us determine which intersections of the sets of types I,II,III are of the form $3K_2$.

a) $(x^\perp \cup T_x) \cap (y^\perp \cup T_y) \cap U \simeq 3K_2$ iff $y \in x^\perp$, $x \notin T_y$ and $y \notin T_x$.

b) $(x^\perp \cup T_x) \cap (y^\perp \cup T_z) \cap U \simeq 3K_2$ iff either ($x \in \{y,z\}^\perp$ and $y \notin T_x$) or ($x \notin y^\perp \cup z^\perp$ and $T_x \ni w$ where $w \in \{y,z\}^\perp$).

c) $(x^\perp \cup T_x) \cap (T_z \cup T_z') \cap U \not\simeq 3K_2$.

d) $(x^\perp \cup T_y) \cap (z^\perp \cup T_w) \cap U \simeq 3K_2$ iff either ($x = w$ and $y = z$) or ($x = w$ and $y \in z^\perp$) or ($w \in x^\perp$ and $y = z$).

e) $(x^\perp \cup T_y) \cap (T_z \cup T_z') \cap U \not\simeq 3K_2$.

f) $(T_x \cup T_x') \cap (T_z \cup T_z') \cap U \not\simeq 3K_2$.

It follows immediately that no set M_a can be of type III, since no type is available for M_b when $b \sim a$. Each edge of $\Delta(\infty)$ is in three 6-cliques and these have two points each in $\Gamma(\infty)$, so that we find 4-cliques in $\Delta(\infty)$.

If some 4-clique $\{a, b, c, d\}$ has $M_a = (x^\perp \cup T_y) \cap U$ and $M_b = (y^\perp \cup T_z) \cap U$ (with $z \in x^\perp$), then M_c and M_d cannot both be of type I (for let $\{x,y,z,w\}$ be an orthonormal basis; if $M_c = (v^\perp \cap T_v) \cap U$ where $v \neq w$ then $v \in w^\perp$ and $w \in T_v$; now M_d cannot be $(w^\perp \cup T_w) \cap U$ so $M_d = (u^\perp \cup T_u) \cap U$ where $u \in \{v,w\}^\perp$ and $w \in T_u$, $v \notin T_u$, impossible by the definition of a tripod); so w.l.o.g. $M_c = (z^\perp \cup T_x) \cap U$.

Consequently the three 4-cliques on the edge ab each contain a point c with $M_c = (z^\perp \cup T_x) \cap U$, and by Lemma 6 these three sets are distinct, so we see that the three possibilities for T_x ($z \notin T_x$) all occur. Now fixing a and c

and repeating the argument we find three points b with $M_b = (y^\perp \cup T_z) \cap U$ and similarly three points a with $M_a = (x^\perp \cup T_y) \cap U$ and thus a subgraph $\simeq K_{3,3,3}$ in $\Delta(\infty)$. But that is impossible:

LEMMA 7. Let K be a subgraph of Γ with $K \simeq K_{3,3,3}$. Then any point x outside K is adjacent to precisely three points of K.

Proof. Standard counting arguments. □

Next: no 4-clique {a,b,c,d} has M_a of type II and M_b, M_c, M_d all of type I: Let $M_a = (x^\perp \cup T_y) \cap U$ and {x,y,z,w} be an orthonormal basis; let the three sets M_b, M_c and M_d be $(v_i^\perp \cup T_{v_i}) \cap U$ $(i = 1,2,3)$, then the points v_1,v_2,v_3 are pairwise orthogonal and each is in $\{w,z\} \cup w^\perp \cup z^\perp$; if $v_1 = w$, $v_2 = z$ then $v_3 \in \{x,y\}$, impossible; if $v_1 = w$, $v_2,v_3 \in w^\perp \backslash \{z\}$ then T_{v_2} must contain w and must not contain w, impossible; if $v_1,v_2,v_3 \in (w^\perp \cup z^\perp) \backslash \{w,z\}$ then we may suppose $v_2,v_3 \in w^\perp \backslash \{z\}$ and the same contradiction arises.

It follows that if a 4-clique {a,b,c,d} has M_a of type II, then there is precisely one other set of type II among M_b, M_c, M_d - if $M_a = (x^\perp \cup T_y) \cap U$ then $M_b = (y^\perp \cup T_x) \cap U$, but a is on 27 four-cliques and for each of the three possible b the edge ab is on only 3 four-cliques, a contradiction. This shows that sets of type II do not occur at all: the main lemma is proved. □

6. APPLICATIONS

THEOREM 4. There does not exist a rank 4 Zara graph G on 287 points with clique size 7 and nexus 3.

Proof. Using the main theorem for Zara graphs it is not difficult to show that G is a strongly regular graph with $(v,k,\lambda,\mu) = (287,126,45,63)$. Moreover, for each point $\infty \in G$, $\Gamma(\infty) \simeq Z^*$. To finish the proof we need two lemmas.

LEMMA 8. Let $a \not\sim b$ in G . Then $\mu(a,b)$ is a graph on 63 points, and for each $c \in \mu(a,b)$ we have $\Gamma_{\mu(a,b)}(c) \simeq 3K_{3,3}$.

Proof. Consider $\Gamma_G(c) \simeq Z^*$. In $\Gamma_G(c)$, $\mu(a,b) \simeq 3K_{3,3}$, but this just means that $\Gamma_{\mu(a,b)}(c) \simeq 3K_{3,3}$. □

LEMMA 9. Z^* does not contain a subgraph T on 63 points which is locally $3K_{3,3}$.

Proof. Let $\infty \in Z^*$ and suppose $\infty \in T$. Let $K \simeq 3K_{3,3}$ be the subgraph of $\Gamma_Z(\infty)$ also in T. Let figure 4 be one of the components of K, and consider $\Gamma_T(a)$. We see the points ∞,u,v,w.

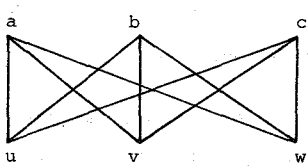


figure 4

Since $\Gamma_T(a) \simeq 3K_{3,3}$ there are points ∞' and ∞'' in T also adjacent to a,u,v,w. But we know these points, they are unique in Z. Hence: ∞,∞',∞'' have precisely the same neighbours in T. As a consequence the points of T can be divided into 21 groups of 3. Let T' be the graph defined on the 21 triples by $t_1 \sim t_2$ if $\tau_1 \sim \tau_2$ for all $\tau_1 \in t_1$, $\tau_2 \in t_2$. Then T' is a strongly regular graph on 21 points with $k = 6$, $\lambda = 1$ and $\mu = 1$ (this is a direct consequence of the structure of Z^*). Now such a graph does not exist, since it violates almost all known existence conditions for strongly regular graphs. This proves the lemma and the theorem. □

THEOREM 5. There does not exist a (non-trivial) completely regular two-graph on 288 points.

Proof. (For definitions and results about completely regular two-graphs see [4]).

A completely regular two-graph on 288 points, gives rise to at least one rank 4 Zara graph on 287 points with clique size 7 and nexus 3. But such a graph does not exist by the previous theorem. □

ACKNOWLEDGEMENTS.

We are very grateful to Henny Wilbrink for carefully reading the manuscript and pointing out a "few" mistakes.

REFERENCES

[1] Blokhuis, A., Few-distance sets, thesis , T.H. Eindhoven (1983).

[2] Brouwer, A.E., A.M. Cohen, H.A. Wilbrink, Near polygons with lines of size three and Fischer spaces. ZW 191/83 Math. Centre, Amsterdam.

[3] Brouwer, A.E., H.A. Wilbrink, Ovoids and fans in the generalized quadrangle GQ(4,2), report ZN 102/81 Math. Centre, Amsterdam.

[4] Neumaier, A., Completely regular two-graphs, Arch.Math. 38,(1982) 378-384.

[5] Seidel, J.J., Strongly regular graphs with (-1,1,0) adjacency matrix having eigenvalue 3, Linear Algebra and Applications 1 (1968), 281 - 298.

[6] Zara, F., Graphes liés aux espaces polaires, preprint.

THE SCHOOLGIRLS OF GRIJPSKERK

by

J.H. de Boer

(Catholic University, Nijmegen)

Dedicated to J.J. Seidel on the occasion of his 65th birthday.

In 1974 a new edition, the twelfth, of Rouse Ball's famous "Mathematical Recreations and Essays" appeared, the eleventh edition being of 1939. This new edition, completely revised by Coxeter, came as a pleasant surprise and a further pleasant surprise was the fact that the chapter on Kirkman's schoolgirls appeared to have been rewritten by Seidel.

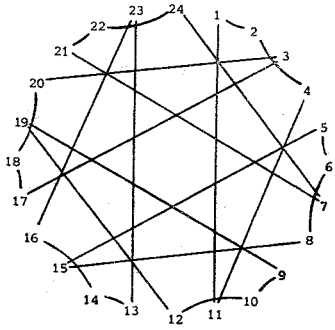
Though the title of the book allows for some deviation from frivolity, still the question may be asked: "Did Seidel write the chapter in office hours or in his own time?" I mean, is it concerned with *relevant* mathematics? The question is not a hard one. Combinatorial designs not only rank among, but have even moved up to the front rows of applicable mathematics. Inevitably, this has brought about a shift from nice isolated problems to theory, much the same as happens, in the case of chess, with some so-called endgame studies. Thus also Kirkman's schoolgirls have marched further into territories like finite geometries, Hadamard matrices, latin squares, etc. In the hands of Seidel, the chapter has become an introduction to combinatorial theory.

Gone, however, is the discussion of rotational solutions, where the schoolgirls are placed in a circle in order to allow a cyclic permutation without tiring them too much. One of the girls was thereby even placed in the centre and did not have to move at all.

It was precisely this idea of the turning wheel that helped me, back in 1951, to put this part of mathematics in the service of mankind. Mankind was represented by a club of cardplayers in the town of Grijpskerk, not far from the University of Groningen, where I had an assistant's position at the department of economics. In the past the club had consisted of 16 and later of 20 members and they had found and used schedules that allowed them to play, for a winter season of five saturday nights, in groups of four players ("tables" or "quadruples") in such a way that no two players would meet more than once (at a table) in the season. In 1951, however, as a contribution to the increase of the world's population, the club had acquired four more members and thus, henceforward, consisted of 24 players. I received a letter from its secretary in the morning of 4 October 1951 and it contained a cry for help. They wanted a new schedule to play again for five nights, the nature of the game still being such that two players who meet at a table, develop such a tremendous dislike for each other that they cannot meet again in the same season. Obviously, five nights is not enough to have all pairs of players meeting, so that is no requirement. But it is clear that otherwise the club's members wanted to continue to behave like schoolgirls.

The letter was a personal one, the secretary of the club being a cousin of mine. But as I was a consultant mathematician at that time, I confidently

took the problem with me to my office and worked on it during that day. After looking into "Mathematical Recreations and Essays" I managed to find a partly rotational solution:



My turning wheel has all 24 players on the circumference and no player at the centre. The fifth day, not shown in the figure, consists of the inscribed squares, the first day is as indicated and the other three days are obtained by rotating. So, that afternoon an official letter went out from the university, saying:

"Sir, we hope the following arrangement can help you.

1	2	4	11
5	6	8	15
9	10	12	19
13	14	16	23
17	18	20	3
21	22	24	7

2	3	5	12
6	7	9	16
10	11	13	20
14	15	17	24
18	19	21	4
22	23	1	8

3	4	6	13
7	8	10	17
11	12	14	21
15	16	18	1
19	20	22	5
23	24	2	9

4	5	7	14
8	9	11	18
12	13	15	22
16	17	19	2
20	21	23	6
24	1	3	10

1	7	13	19
2	8	14	20
3	9	15	21
4	10	16	22
5	11	17	23
6	12	18	24

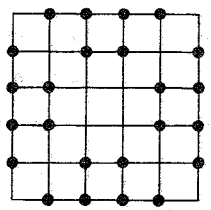
Sincerely yours".

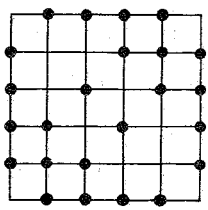
Are there other solutions? Or could the club even play one saturday more, i.e. for six days, without any pair of members meeting twice? I had already learned the principal rule for consulting mathematicians: Do not put your own question and as soon as you have answered a question, leave it at that and do not pursue the matter further. So I left it at that.

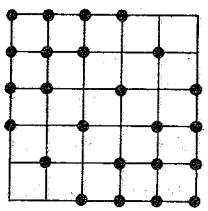
But now, with Seidel retiring, an event bringing about feelings of gratitude as well as some sadness and some work, I have to go a little further. Let me start by saying that I cannot even show that a schedule for seven days (and 24 players) is impossible. For that it does not suffice to ignore the parallelism, i.e., the grouping into the seven days: It is possible to form 42 tables of 4 players, where no pair of players sits together at more than one table, a "packing of pairs by quadruples". In fact, start with a so-called Steiner system with 25 players, $0,1,\ldots,24$, where each of the 300 pairs occurs in exactly one of 50 blocks of 4 players. Such designs exist ([1], [2]). Now simply leave out the 8 tables where player 0 participates and you end up with 42 tables. So to me certainly $N(6) = 0$ is only a conjecture. (Here $N(k)$ is the number of distinct solutions for k days.)

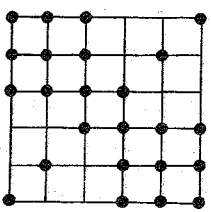
To do at least something, I looked for other solutions to the original request for a 5 days' schedule. I must say that it now took me more time to find (by hand) a second solution and simultaneously a third one. These have less symmetry, but still enough to describe a method of finding them. A fourth solution is suggested by [1]. So I will present three other solutions here.

How do I know the solutions are different, up to renumbering the players or the days or the tables? A first remark is that for a schedule for just two days (and 24 players) there are only 4 possibilities, up to renumbering. The intersection matrices for these 4 possibilities are A, B, C, D below: The tables of the first day are the rows, those of the second day are the columns.

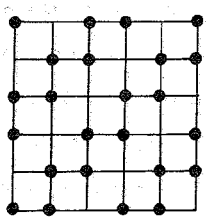
A = 

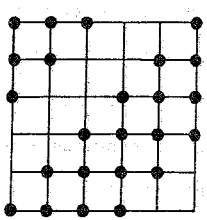
B = 

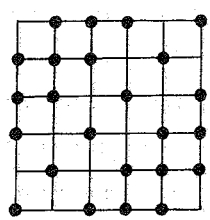
C = 

D = 

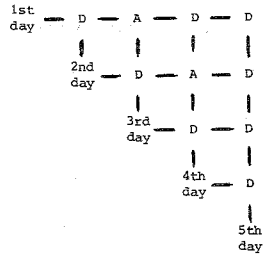
The rows or the columns may be permuted, so for instance A, C and D can also be drawn in the following form, indicating their cyclicity (B is not cyclic):

A = 

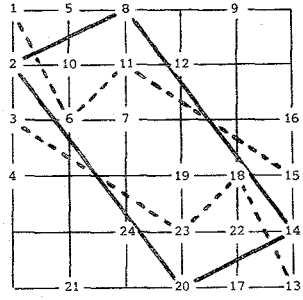
C = 

D = 

A solution of the Grijpskerk problem gives an intersection diagram in which for each of the 10 pairs of the 5 days it is indicated which of the possibilities A, B, C, D applies. For my first solution one obtains the following incidence matrices:



Is there a 5-day arrangement possible with incidence matrix B or C occurring? To find such a solution, one obviously could start with B or C for the first two days and try and supplement the other days. Silly enough, I started with a variation of incidence matrix D instead.



This matrix being symmetric with respect to its centre, one can speak of a pair of players who are opposite to each other, such as 6 and 18; the numbering is such that when placed in a circle the two players are also opposite.

I first drew in the third day, which is such that with each table also the four opposite players form a table. One such pair is drawn in the figure. Similarly for the fourth day. For the fifth day each table is its own opposite and the numbering of the players is further such that these tables are the squares in the circle. One such square is shown. Solutions with such a symmetry with respect to the centre of the D-matrix are easier to find and reasonably transparent.

Written out, my second solution is: 1st day the rows

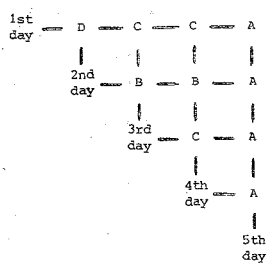
2nd day the columns

3rd day			
1	6	11	15
2	24	17	16
3	23	18	13
4	5	12	14
8	19	22	21
9	10	7	20

4th day			
1	10	18	14
2	6	22	13
3	8	12	17
4	7	23	21
5	24	20	15
9	11	19	16

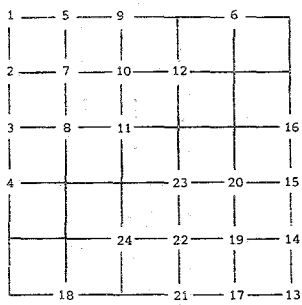
5th day			
1	7	13	19
2	8	14	20
3	9	15	21
4	10	16	22
5	11	17	23
6	12	18	24

Its incidence matrices are



A third solution, with the same first three days (but the players renumbered in order to have the fifth day again consisting of the squares in the circle)

is:

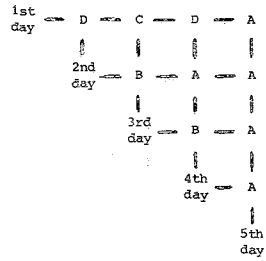


1st day the rows

2nd day the columns

	3rd day				4th day				5th day			
1	8	10	15	1	11	22	18	1	7	13	19
2	24	17	16	2	5	19	15	2	8	14	20
3	22	20	13	3	7	17	14	3	9	15	21
4	5	12	14	4	8	24	21	4	10	16	22
6	7	11	21	6	10	23	13	5	11	17	23
9	23	19	18	9	12	20	16	6	12	18	24

Its incidence matrices are



For a fourth solution I somewhat demolished the Steiner system with v = 25, k = 4 given in [1]. The players are now numbered $\{(i,j) ; 1 \leq i \leq 6 , 1 \leq j \leq 4\}$.

1st day			
11	41	12	42
13	43	14	44
21	51	23	53
22	52	24	54
32	62	33	63
31	61	34	64

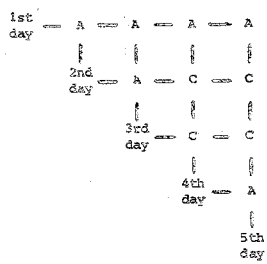
2nd day			
11	51	14	54
12	52	13	53
21	61	22	62
23	63	24	64
31	41	33	43
32	42	34	44

3rd day			
11	61	13	63
12	62	14	64
21	41	24	44
22	42	23	43
31	51	32	52
33	53	34	54

4th day			
11	22	33	44
21	32	43	64
31	42	53	24
41	52	63	14
51	62	13	34
61	12	23	54

5th day			
11	52	43	34
21	42	63	54
31	62	23	44
41	22	13	64
51	12	33	24
61	32	53	14

Its incidence matrices are



So, to end with a theorem, $N(5) \geq 4$.

REFERENCES

[1] Rokowska, B., A new construction of the block systems B(4,1,21) and B(4,1,25). Coll. Math. 38 (1977), 165 - 167.

[2] Bose, R.C., On the construction of balanced incomplete block designs. Annals of Eugenics 9 (1939), 353 - 399.
(Cf. Hanani, H., The existence and construction of balanced incomplete block designs. Annals of Math. Stat. 32 (1961), 361 - 386, p. 372.)

REPRODUCING INTEGRAL RELATIONS FOR SPHERICAL HARMONICS: ANSWER TO SOME QUESTIONS OF J.J. SEIDEL

by

J. Boersma

Dedicated to J.J. Seidel on the occasion of his retirement.

Preface to Jaap Seidel

Dear Jaap: Our first meeting dates back to December 4, 1962. On that day I was visiting Bouwkamp to discuss my thesis work, and on my arrival he told me that I was going to have lunch with you. I do remember our conversation quite well. As a chairman of the Mathematics Department you were very active in hiring new personnel, and so you offered me a job in Eindhoven after my graduation. Some two years later I accepted, but I still feel that I was actually hired over that lunch.

I have always admired your broad conception of mathematics as a unity. Your interest was certainly not confined to just your own research specialism. In addition, you love to talk and lecture about your research. Thus through the years I have got a fairly good idea of what your work has been about. In your work of recent years special functions, in particular orthogonal polynomials, play an important role. This subject which also has my interest, has brought us together mathematically. Ever since, you have occasionally employed me as a sort of free lance consultant on your problems in special functions. This paper reports on one such occasion, dated October 7, 1983, when you asked me a couple of questions on what I have called "reproducing" integral relations

for harmonic and homogeneous polynomials on the unit sphere. On the same day I sent you my final answers, however, I still owe you the details of my analysis which you may now find in Section 2 of this paper. In retrospect I realize that my results hardly differ from yours [6], except for a somewhat different approach; thus I might conclude that you could easily have answered the questions yourself!

While working on this paper I was reminded of another question which I asked your student Blokhuis on the occasion of his thesis defense. In his thesis [1] Blokhuis introduced an "algebraic" inner product for harmonic polynomials on the unit sphere, which seemed to be different from the usual inner product in terms of an integral over the unit sphere. My question concerned the precise relationship between the two inner products. Blokhuis' answer was not too explicit - because of uniqueness the inner products should coincide up to a scalar multiple. In Section 3 of this paper I present a more detailed answer which hopefully will interest you.

Finally, I have enjoyed working with you in the past years and I hope that our (modest) co-operation will continue after your retirement.

1. INTRODUCTION

We start from the (real) Euclidean space $\mathbb{R}^n$ of dimension $n \geq 2$, consisting of vectors $x = (x_1, x_2, \ldots, x_n)$, and provided with the standard inner product $(x,y) = \sum_{i=1}^{n} x_i y_i$. Let Ω denote the unit sphere in $\mathbb{R}^n$, i.e., $\Omega = \{x \mid (x,x) = 1\}$. For any integer $k \geq 0$, we introduce the following spaces of polynomials in n variables $x_1, x_2, \ldots, x_n$, restricted to Ω:

Pol(k) is the linear space of polynomials of total degree $\leq k$;

Hom(k) is the linear space of homogeneous polynomials of total degree k;

Harm(k) is the linear subspace of Hom(k) consisting of the harmonic polynomials, where a polynomial $f(x)$ is called harmonic if

$$\Delta f = f_{x_1 x_1} + f_{x_2 x_2} + \ldots + f_{x_n x_n} = 0 .$$

All spaces are provided with the positive definite inner product

$$(1) \qquad \langle f,g \rangle = \frac{1}{\omega_n} \int_{\Omega} f(x) g(x) d\omega(x)$$

in which $d\omega$ denotes the surface element of Ω and ω_n is the surface area of Ω, given by

$$\omega_n = \frac{2\pi^{n/2}}{\Gamma(n/2)} .$$

With respect to this inner product we have the following orthogonal decompositions of Hom(k) and Pol(k) (cf. [3, Thm. 3.1], [6, §1])

$$(2) \qquad \mathrm{Hom}(k) = \sum_{i=0}^{[k/2]} \mathrm{Harm}(k-2i) ,$$

$$\text{(3)} \qquad \mathrm{Pol}(k) = \mathrm{Hom}(k) \oplus \mathrm{Hom}(k-1) = \sum_{i=0}^{k} \mathrm{Harm}(i).$$

The elements of Harm(k) are also called spherical harmonics of degree k for the sphere Ω in $\mathbb{R}^n$. In Harm(k) we may choose an orthogonal basis with respect to the inner product (1), expressible in terms of Gegenbauer polynomials; see [4, Secs. 11.2, 11.3]. For later use we list some properties of these polynomials, quoted from [4, Sec. 10.9].

For $\lambda > -\frac{1}{2}$, $\lambda \neq 0$, the Gegenbauer polynomials $C_k^\lambda(t)$ may be defined by means of their generating function

$$(1 - 2tz + z^2)^{-\lambda} = \sum_{k=0}^{\infty} C_k^\lambda(t) z^k .$$

$C_k^\lambda(t)$ is a polynomial in t of degree k and the coefficient of t^k is given by

$$a_k = \frac{2^k \Gamma(\lambda + k)}{\Gamma(\lambda) k!} ;$$

furthermore

$$C_k^\lambda(1) = \frac{\Gamma(2\lambda + k)}{\Gamma(2\lambda) k!} .$$

Most important is the orthogonality property

$$\int_0^\pi C_k^\lambda(\cos\theta) C_\ell^\lambda(\cos\theta) (\sin\theta)^{2\lambda} d\theta = h_k \delta_{k\ell}$$

where

$$h_k = \frac{\pi^{\frac{1}{2}}\Gamma(\lambda+\frac{1}{2})\Gamma(2\lambda+k)}{(\lambda+k)\Gamma(\lambda)\Gamma(2\lambda)k!} .$$

2. ANSWER TO SEIDEL'S QUESTIONS

Seidel's first question dealt with the integral relation

$$(4) \qquad \frac{1}{\omega_n}\int_{\Omega} f(x)(x,y)^k d\omega(x) = Cf(y)$$

which he conjectured to be valid for all $f \in \mathrm{Harm}(k)$. Actually, his relation had the constant $C = 1$ which will turn out to be wrong. To verify (4), the function f is represented in terms of a suitable orthogonal basis of Harm(k). Let $n \geq 3$. Then we choose the basis to consist of the zonal spherical harmonic $C_k^{n/2-1}((x,y))$ with pole y, and a number of non-zonal spherical harmonics; cf. [4, Sec. 11.2, Thm. 1]. For f equal to a non-zonal harmonic, the relation (4) holds trivially because both the left- and right-hand sides vanish. Therefore it is sufficient to set $f(x) = C_k^{n/2-1}((x,y))$ in (4). Then by the substitution $(x,y) = \cos\theta$, $d\omega(x) = \omega_{n-1}(\sin\theta)^{n-2}d\theta$, the integral (4) reduces to

$$\frac{\omega_{n-1}}{\omega_n}\int_0^{\pi} C_k^{n/2-1}(\cos\theta)(\cos\theta)^k(\sin\theta)^{n-2}d\theta .$$

The latter integral is readily evaluated by use of the orthogonality property, after replacing $(\cos\theta)^k$ by $a_k^{-1}C_k^{n/2-1}(\cos\theta)$. Thus we find that the integral relation (4) is correct and the constant C is given by

$$(5) \qquad C = \frac{\omega_{n-1}h_k}{\omega_n a_k C_k^{n/2-1}(1)} = \frac{\Gamma(n/2)k!}{2^k\Gamma(n/2+k)} = \frac{k!}{n(n+2)\cdots(n+2k-2)},$$

valid for $n \geq 3$. The present result also holds for $n = 2$. In that case it is sufficient to set $f(x) = \cos(k\varphi)$ in (4), where $\cos \varphi = (x,y)$. Then the constant C is found to be

$$C = \frac{1}{2\pi} \int_0^{2\pi} \cos(k\varphi)(\cos \varphi)^k d\varphi = 2^{-k}$$

by means of [4, form. 1.5.1 (30)].

The integral relation (4) permits some obvious generalization, for example, the kernel $(x,y)^k$ may be replaced by a polynomial in (x,y) of degree k. Choosing the Gegenbauer polynomial $C_k^{n/2-1}$, with $n \geq 3$, as a kernel, we arrive at

$$(6) \qquad \frac{1}{\omega_n} \int_\Omega f(x) C_k^{n/2-1}((x,y)) d\omega(x) = a_k C f(y) = \frac{n-2}{n+2k-2} f(y) \ ,$$

valid for $f \in \mathrm{Harm}(k)$. Referring to [3, Remark 2.2], we now introduce the normalized Gegenbauer polynomial $Q_k(t)$ defined by

$$(7) \qquad Q_k(t) = \frac{n+2k-2}{n-2} C_k^{n/2-1}(t), \quad n \geq 3$$

where the dependence of Q_k on n has been suppressed in the notation. Then the integral relation (6) takes the simple form

$$(8) \qquad \frac{1}{\omega_n} \int_\Omega f(x) Q_k((x,y)) d\omega(x) = f(y) \ ,$$

valid for $f \in \mathrm{Harm}(k)$, $n \geq 3$. It can easily be shown that the relation (8) also holds for $n = 2$, if the polynomial $Q_k(t)$ is defined by

$$(9) \qquad Q_0(t) = 1, \quad Q_k(t) = 2T_k(t), \quad k \geq 1, \quad n = 2,$$

in accordance with [3, Remark 2.2]; here, $T_k(t)$ is the Tchebichef polynomial of the first kind, cf. [4, Sec. 10.11].

In terms of the inner product (1), the integral relation (8) can be shortly written as

$$(10) \qquad \langle f(x), Q_k((x,y)) \rangle = f(y), \qquad f \in \mathrm{Harm}(k)$$

which shows that the kernel Q_k has the "reproducing" property.

Seidel's next question was to extend the result to the spaces Hom(k) and Pol(k), i.e., to determine reproducing kernels C_k and R_k such that

$$(11) \qquad \langle f(x), C_k((x,y)) \rangle = f(y), \qquad f \in \mathrm{Hom}(k),$$

$$(12) \qquad \langle f(x), R_k((x,y)) \rangle = f(y), \qquad f \in \mathrm{Pol}(k).$$

The answer is straightforward. In view of the decompositions (2), (3), it is immediately seen that

$$(13) \qquad C_k(t) = \sum_{i=0}^{[k/2]} Q_{k-2i}(t), \quad R_k(t) = \sum_{i=0}^{k} Q_i(t) \ .$$

The notations C_k, R_k have been adopted from [3, Def. 2.3] where it is also pointed out that $C_k(t)$ and $R_k(t)$ can be expressed in terms of Gegenbauer and Jacobi polynomials. Indeed, by means of [4, form. 10.9 (36), 10.9 (4), 10.8(36)]

one has

$$C_k(t) = \sum_{i=0}^{[k/2]} \left\{ C_{k-2i}^{n/2}(t) - C_{k-2i-2}^{n/2}(t) \right\} = C_k^{n/2}(t) ,$$

$$R_k(t) = \sum_{i=0}^{k} \left\{ C_i^{n/2}(t) - C_{i-2}^{n/2}(t) \right\} = C_k^{n/2}(t) + C_{k-1}^{n/2}(t)$$

$$= \frac{2\Gamma(\frac{1}{2}n + \frac{1}{2})\,(n+k-2)!}{(n-1)!\,\Gamma(\frac{1}{2}n + k - \frac{1}{2})}\, p_k^{(\mu,\mu-1)}(t)$$

where $\mu = \frac{1}{2}(n-1)$.

It is interesting to reconsider these results from another viewpoint. To that purpose we start from the reproducing integral relation for continuous functions on Ω:

$$(14) \qquad \langle f(x), \omega_n \delta(x-y) \rangle = \int_\Omega f(x)\,\delta(x-y)\,d\omega(x) = f(y) ,$$

in which $\delta(x-y)$ is the Dirac distribution with pole y. Next we expand $\omega_n \delta(x-y)$ in a series of orthogonal polynomials $Q_i((x,y))$, then the expansion coefficients are equal to

$$\frac{\langle \omega_n \delta(x-y), Q_i((x,y)) \rangle}{\langle Q_i((x,y)), Q_i((x,y)) \rangle} = \frac{Q_i(1)}{Q_i(1)} = 1,$$

where the denominator has been evaluated by means of (10). Thus we have the formal expansion

$$(15) \qquad \omega_n \delta(x-y) = \sum_{i=0}^{\infty} Q_i((x,y)) .$$

The reproducing kernels Q_k, C_k, R_k, as determined before, are now immediately recognized as "projections" of $\omega_n \delta(x-y)$ onto Harm(k), Hom(k), Pol(k), respectively.

Except for a different approach, the results of this section are basically the same as those of Seidel [6]. In his approach the polynomial Q_k is in fact defined by the reproducing property (10). In [6, §2] Seidel starts from a linear functional on Harm(k) given by $f \mapsto f(y)$, with $y \in \Omega$ fixed. Then by the Riesz representation theorem there exists a unique element of Harm(k), denoted by $Q_k(x,y)$, such that $\langle f, Q_k(\cdot,y)\rangle = f(y)$ for all $f \in$ Harm(k). Next it is shown that Q_k depends on (x,y) only, thus leading to the reproducing property (10). Various properties of the polynomials Q_k are discussed in [6,§3], from which the relations (7), (9) to Gegenbauer polynomials readily follow. In our approach the reproducing kernel Q_k has been constructed by starting from a suitable orthogonal basis of Harm(k), expressed in terms of Gegenbauer polynomials.

3. RELATIONSHIP BETWEEN INNER PRODUCTS

In his thesis [1, §2.5] Blokhuis introduced an inner product for harmonic polynomials, different from (1), and defined by

$$\langle f,g\rangle_B = (f(\partial)g)(0), \qquad f,g \in \text{Harm}(k). \tag{16}$$

Here $f(\partial)$ is the differential operator obtained from $f(x) = f(x_1,x_2,\ldots,x_n)$ by replacing x_i by $\partial/\partial x_i$, $i = 1,2,\ldots,n$. In this section we shall establish the precise relationship between the inner products (1) and (16).

In [1, §2,6] Blokhuis determined the unique polynomial $\tilde{y} \in \mathrm{Harm}(k)$ with the reproducing property $\langle f,\tilde{y}\rangle_B = f(y)$ for all $f \in \mathrm{Harm}(k)$, and he found that

$$\tilde{y} = [(n-2)n \cdots (n+2k-4)]^{-1} C_k^{n/2-1}((x,y))$$

or equivalently, by (7),

$$\tilde{y} = [n(n+2) \cdots (n+2k-2)]^{-1} Q_k((x,y)).$$

A comparison of this result with (10) suggests the relationship

$$\langle f,g\rangle_B = n(n+2) \cdots (n+2k-2)\langle f,g\rangle, \qquad f,g \in \mathrm{Harm}(k) \tag{17}$$

which is proved below. Notice the similarity of (17) when written out, and the mean value theorem [2, p.277]

$$u(0) = \frac{1}{\omega_n} \int_{\Omega} u(x)\, d\omega(x)$$

for harmonic functions $u(x)$.

THEOREM. Let $f \in \mathrm{Hom}(k)$, $g \in \mathrm{Harm}(k)$. Then

$$\frac{1}{\omega_n} \int_{\Omega} f(x)g(x)\, d\omega(x) = \frac{(f(\partial)g)(0)}{n(n+2)\cdots(n+2k-2)} .$$

Proof. The result is trivial for $k = 0$. Next let $k \geq 1$, and let f be the monomial $f(x) = x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n}$ with $\sum_{i=1}^{n} \alpha_i = k$. Then there exists an exponent $\alpha_i \geq 1$ and we may set $f(x) = x_i p(x)$ with $p \in \mathrm{Hom}(k-1)$. Let N denote the outward unit normal to Ω with components $N_i = x_i$, $i = 1,2,\ldots,n$. Then by means

of Gauss' divergence theorem one has

$$\int_{\Omega} f(x)g(x)d\omega(x) = \int_{\Omega} p(x)g(x)N_i d\omega(x) = \int_{B} \frac{\partial}{\partial x_i}\{p(x)g(x)\}d\tau(x) ,$$

in which B is the unit ball in $\mathbb{R}^n$, given by $(x,x) \leq 1$, and $d\tau$ denotes the volume element of B. The integrand $(\partial/\partial x_i)\{p(x)g(x)\}$ is a homogeneous polynomial of degree $2k-2$, hence, by defining $(x,x)^{\frac{1}{2}} = r$, the volume integral over B can be reduced to

$$\int_0^1 r^{n+2k-3}dr \int_{\Omega} \frac{\partial}{\partial x_i}\{p(x)g(x)\}d\omega(x)$$

$$= \frac{1}{n+2k-2}\int_{\Omega} p_{x_i}(x)g(x)d\omega(x) + \frac{1}{n+2k-2}\int_{\Omega} p(x)g_{x_i}(x)d\omega(x) .$$

Here the first integral over Ω vanishes because of $p_{x_i} \in \mathrm{Hom}(k-2)$, $g \in \mathrm{Harm}(k)$, and the orthogonal decomposition (2). Thus by collecting the above results we have shown that

$$\int_{\Omega} f(x)g(x)d\omega(x) = \frac{1}{n+2k-2}\int_{\Omega} x_i^{-1}f(x)g_{x_i}(x)d\omega(x) .$$

By repeated application of the latter identity to the monomial f we find

$$\int_{\Omega} f(x)g(x)d\omega(x) = [n(n+2)\cdots(n+2k-2)]^{-1}\int_{\Omega} \prod_{i=1}^{k}\left(\frac{\partial}{\partial x_i}\right)^{\alpha_i} g(x)d\omega(x)$$

$$= [n(n+2)\cdots(n+2k-2)]^{-1}\int_{\Omega} f(\partial)g(x)d\omega(x) .$$

Finally, since $f(\partial)g(x)$ is a harmonic function, we have by the mean value theorem

$$\frac{1}{\omega_n}\int_{\Omega} f(\partial)g(x)\,d\omega(x) = (f(\partial)g)(0)$$

which completes the proof. □

The relationship (17) is now obvious. Moreover, our theorem implies that $(f(\partial)g)(0)=(g(\partial)f)(0)$ for $f,g \in \mathrm{Harm}(k)$, which shows again that Blokhuis' inner product is symmetric.

In [2, Ch.IV, §3,4] the mean value theorem is extended to other elliptic differential equations. Remembering that $\Delta^{k+1}(fg) = 0$ if $f,g \in \mathrm{Hom}(k)$, we may use the result in [2, p.289] to obtain

$$\frac{1}{\omega_n}\int_{\Omega} f(x)g(x)\,d\omega(x) = \Gamma(n/2)\sum_{m=0}^{k}\frac{(\Delta^m(fg))(0)}{2^{2m}m!\Gamma(m+n/2)}$$

$$= \frac{(\Delta^k(fg))(0)}{2^k k!\, n(n+2)\cdots(n+2k-2)}\,.$$

By combining this result with our theorem, we are led to the identity

$$(\Delta^k(fg))(0) = 2^k k!(f(\partial)g)(0)\,, \tag{18}$$

valid for $f \in \mathrm{Hom}(k)$, $g \in \mathrm{Harm}(k)$. Here is a short direct proof of (18), where it is assumed for simplicity that also $f \in \mathrm{Harm}(k)$. For $k = 0$ the identity holds trivially. Next, proceeding by induction, let the identity be true for some $k \geq 0$. Then for $f,g \in \mathrm{Harm}(k+1)$ we have

$$\Delta(fg) = 2 \sum_{i=1}^{n} f_{x_i} g_{x_i} ,$$

in which $f_{x_i}, g_{x_i} \in \mathrm{Harm}(k)$, and by use of the induction assumption we establish

$$\Delta^{k+1}(fg) = 2 \sum_{i=1}^{n} \Delta^{k}(f_{x_i} g_{x_i})$$

$$= 2^{k+1} k! \left(\left\{ \sum_{i=1}^{n} f_{x_i}(\partial) \frac{\partial}{\partial x_i} \right\} g \right)(0) = 2^{k+1}(k+1)!\,(f(\partial)g)(0) .$$

It is proper to end this paper with a (minor) suggestion to Seidel for his future research. Referring to [1, Ch. 2], we consider the vector space $\mathbb{R}^{p,q}$ provided with an indefinitie inner product of signature (p,q) with $p+q = n$. Then one may again introduce the space Harm(k) of harmonic polynomials f(x) which now satisfy the (ultra)hyperbolic equation

$$f_{x_1 x_1} + \ldots + f_{x_p x_p} - f_{x_{p+1} x_{p+1}} - \ldots - f_{x_n x_n} = 0 .$$

In Harm(k) one may define an inner product similar to (16), whereas there seems to be no analog of the inner product (1). As a suggestion to Seidel, one might possibly employ a suitable mean value theorem in order to reduce the inner product (16) to an expression in terms of integrals over unit spheres. Such a mean value theorem for solutions of ultrahyperbolic equations has been developed by Asgeirsson, see [2, Ch. VI, §16], [5, Ch. V].

REFERENCES

[1] Blokhuis, A., Few-distance sets, Thesis, Eindhoven University of Technology, Eindhoven 1983.

[2] Courant, R., and D. Hilbert, Methods of mathematical physics, Vol. II, Partial differential equations, Interscience, New York, 1962.

[3] Delsarte, Ph., J.M. Goethals and J.J. Seidel, Spherical codes and designs, Geom. Dedicata 6, 363 - 388 (1977).

[4] Erdélyi, A., W. Magnus, F. Oberhettinger and F.G. Tricomi, Higher transcendental functions, Vols. I, II, McGraw-Hill, New York, 1953.

[5] John, F., Plane waves and spherical means applied to partial differential equations, Interscience, New York, 1955.

[6] Seidel, J.J., Spherical harmonics and combinatorics, Memorandum 1981-07, Department of Mathematics, Eindhoven University of Technology, Eindhoven, June 1981.

EEN OUD PROBLEEM OPNIEUW "BENADERD"

door

A.J. Bosch

Opgedragen aan Prof. J.J. Seidel ter gelegenheid van zijn afscheid van de T.H. Eindhoven.

0. INLEIDING

Mijn eerste contact met Prof. Seidel dateert van precies 25 jaar geleden, toen ik een sollicitatiegesprek met hem had en met de Heer Hamaker (toen nog geen hoogleraar) in de stationsrestauratie te Eindhoven.... . Zo werd ik de eerste statistisch medewerker bij de onderafdeling der Wiskunde.

Het werkterrein van Prof. Seidel heeft zeker raakvlakken met de statistiek (ik denk o.a. aan correlatie, fractionele proeven, Grieks-Latijnse vierkanten e.d.). Doch mijn interesse lag op andere deelgebieden van de statistiek. Eén van mijn eerste statistiekpractica bestond uit het uitleggen van de Facit (de "koffiemolen"), een mechanisch eenvoudig rekentuig. Er werd in die tijd nog veel "met de hand" gerekend en men schrok niet terug voor het inverteren van 5 × 5-matrices met elementen in 4 significante cijfers.

Thans staat ons een grote en snelle computer ter beschikking, hetgeen uiteraard het onderwijs en onderzoek sterk heeft beïnvloed. Vele problemen die vroeger opgelost werden via ingewikkelde nomogrammen en tabellen, zijn nu met een computer eenvoudig op te lossen. Zo ook het volgende probleem.

Alweer geruime tijd geleden kwam een medewerker van de afdeling Bouwkunde bij mij met de vraag: "Hoe berekent men toch de coëfficiënt in $\bar{x}$ - ks bij een gegeven keuringsvoorschrift"?

Dit is een oud, bekend probleem, althans theoretisch. Een exacte oplossing kon ik echter niet geven en was mij niet bekend. Wel was mij een benadering bekend van Stange [3] uit ±1960.

Bij het oplossen van dit probleem stuit men nl. op een vergelijking van twee niet-centrale Studentverdelingen, waarbij bovendien de niet-centraliteitsparameter δ afhangt van het aantal vrijheidsgraden ν. Dit probleem leek onoplosbaar. Het toeval wilde echter dat juist die tijd het rekencentrum van de T.H.E. een computerprogramma ontwikkeld had voor de niet-centrale Studentverdeling. Toen was het probleem gauw geklaard. Bovendien kon ik nagaan hoe goed de benadering van Stange [3] wel was en tevens zoeken naar een betere benadering. Na het gereedkomen van dit artikel vernam ik dat Prof. Hamaker [1] voor dit probleem in 1979 een goede benadering had gevonden. Ook al heeft hierdoor dit artikel enigszins aan originaliteit ingeboet, toch leek het mij interessant de drie benaderingen eens met elkaar en met de computeroplossing te vergelijken. Temeer daar de methode van Stange [3], die het slechtst blijkt te zijn, nog steeds gedoceerd wordt bij de afdeling Bouwkunde van de T.H.E. (zie ook o.a. Kreijger [2]) en de benadering van Hamaker [1] daar niet bekend is.

1. PROBLEEMSTELLING

We beschouwen bij de keuring van een partij slechts één karakteristieke grootheid $\underline{x}$ (b.v. de druksterkte), waarvan we bovendien aannemen dat deze normaal

verdeeld is met onbekende μ en (bekende of onbekende) variantie σ^2, notatie $\underline{x} \sim N(\mu,\sigma^2)$.

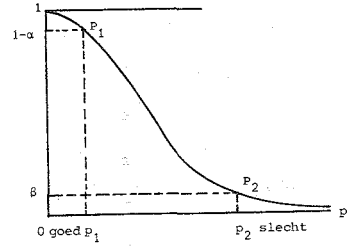
Een produkt heet defekt als $\underline{x} < L$ (b.v. L is de karakteristieke druksterkte).

Men wil een keuringsvoorschrift, dat voldoet aan de twee eisen:

a) de kans om een partij met foutenfractie p_1 af te keuren = α (producenten-risico).

b) de kans om een partij met foutenfractie p_2 goed te keuren = β (konsumenten-risico).

De keuringskarakteristiek ligt nu geheel vast door de punten $P_1(p_1,1-\alpha)$ en $P_2(p_2,\beta)$.



goedkeuringskans als functie van foutenfractie p in de partij

Als $P(\underline{x} < L) = p$, dan is p = foutenfractie in de partij.

Er geldt: $L = \mu - u_{1-p}\sigma$ (1.1)

Voor de definitie van u_{1-p} zie (2.3).

Om te beslissen of een partij goed- of afgekeurd moet worden, nemen we een aselecte steekproef ter grootte n: $x_1,\ldots,x_n$. De steekproefgrootte is vooralsnog onbepaald. Daar $L = \mu - u_{1-p}\sigma$, ligt het voor de hand als toetsingsgrootheid te nemen (er zijn ook andere mogelijkheden):

$$\underline{z} := \bar{\underline{x}} - k_\sigma\sigma \text{ als } \sigma \text{ bekend is of } \underline{z} := \bar{\underline{x}} - k_s\underline{s} \text{ als } \sigma \text{ onbekend is.} \quad (1.2)$$

We keuren de partij goed als $z \geq L$ en keuren deze af als $z < L$.

De vraag is nu:

Hoe groot moeten, bij gegeven α, β, p_1 en p_2, de steekproefgrootte n en de coëfficiënt k zijn, opdat aan de twee eisen van het keuringsvoorschrift is voldaan.

2. NOTATIES EN ENKELE EENVOUDIGE RESULTATEN

2.1. Toevalsvariabelen worden onderstreept zoals $\underline{x}$, $\underline{u}$, $\underline{s}$, $\underline{t}_\nu$ e.d.

2.2. $\underline{x} \simeq \underline{y}$ betekent: $\underline{x}$ en $\underline{y}$ hebben dezelfde kansverdeling.

2.3. $\underline{u} \sim N(0,1)$ d.w.z. $\underline{u}$ is normaal verdeeld met $\mu = 0$ en $\sigma = 1$.

u_α is gedefinieerd door $P(\underline{u} \leq u_\alpha) = \alpha$. Analoog $u_{1-\alpha}$, u_p, $t_{\nu,\gamma}$.

Wegens symmetrie geldt $u_\alpha = -u_{1-\alpha}$.

2.4. $\bar{x} := \Sigma x_i/n$ is het steekproefgemiddelde; $s^2 := \Sigma(x_i - \bar{x})^2/(n-1)$ de steekproefvariantie. Is σ bekend, dan noteren we de steekproefgrootte met n_σ. Is σ onbekend en geschat door s, dan noteren we deze met n_s.

2.5. $\underline{t}_\nu(\delta) \simeq \dfrac{\underline{u} + \delta}{\sqrt{\underline{\chi}^2_\nu/\nu}}$, waarbij teller en noemer onafhankelijk zijn, is een niet-centrale Studentstochast met ν vrijheidsgraden en niet-centraliteitsparameter δ.

Stel in het vervolg $\underline{x} \sim N(\mu,\sigma^2)$, dan geldt:

2.6. $$\frac{\underline{x} - \mu}{\sigma/\sqrt{n}} \simeq \underline{u} \sim N(0,1).$$

2.7. $\bar{\underline{x}}$ en $\underline{s}$ zijn onderling onafhankelijk; $\underline{s}/\sigma \simeq \sqrt{\underline{\chi}^2_\nu/\nu}$ met $\nu = n_s - 1$.

var $\underline{s} \approx \sigma^2/2\nu$. Daar $\mathcal{E}\underline{s}^2 = \sigma^2$ en var $\underline{s} = \mathcal{E}\underline{s}^2 - (\mathcal{E}\underline{s})^2$, geldt bij benadering:

$$\mathcal{E}\underline{s} \approx \sigma\sqrt{1 - \frac{1}{2\nu}} \approx \sigma\left(1 - \frac{1}{4\nu}\right) = \frac{4n_s - 5}{4n_s - 4}\sigma .$$

3. EXACTE OPLOSSING

De twee eisen waaraan het keuringsvoorschrift moet voldoen zijn:

a) $P(\underline{z} < L | p = p_1) = \alpha$ en b) $P(\underline{z} \geq L | p = p_2) = \beta$.

We beschouwen $P(\underline{z} < L | p = p) = \gamma$ en veronderstellen σ onbekend.

Met (1,2) en (1,1) wordt dit:

$$P(\bar{\underline{x}} - k\underline{s} < \mu + u_p\sigma) = \gamma = P(\bar{\underline{x}} - \mu - u_p\sigma < k\underline{s}) =$$

$$= P\left(\frac{\bar{\underline{x}} - \mu}{\sigma/\sqrt{n}} - u_p\sqrt{n} < k\underline{s}\sqrt{n}/\sigma\right) = \text{met (2.6) } P(\underline{u} + \delta_p < k\underline{s}\sqrt{n}/\sigma)$$

waarbij $\delta_p := u_{1-p}\sqrt{n}$

$$= P\left(\frac{\underline{u} + \delta_p}{\underline{s}/\sigma} < k\sqrt{n}\right). \text{ Met (2.5) en (2.7) geeft dit} \qquad (3.1)$$

$$P(\underline{t}_\nu(\delta_p) < k\sqrt{n}) = \gamma. \text{ Dus } k\sqrt{n} = t_{\nu,\gamma}(\delta_p) \qquad (3.2)$$

Substitutie van $p = p_1$, $\gamma = \alpha$ resp. $p = p_2$, $\gamma = 1 - \beta$ in (3.2) geeft:

$$\boxed{t_{\nu,\alpha}(\delta_{p_1}) = t_{\nu,1-\beta}(\delta_{p_2}) = k\sqrt{n}} \quad \text{met } \delta_p = u_{1-p}\sqrt{n} \text{ en } \nu = n-1 \qquad (3.3)$$

De oplossing van deze vergelijkingen geeft $n = n_s$ en $k = k_s$.

Opmerkingen

1. Daar n een natuurlijk getal is, is (3.3) i.h.a. niet exact op te lossen bij gegeven α, β, p_1 en p_2. Oplossen betekent hier dus: minimaliseren naar n van

$$\left| t_{n-1,\alpha}(u_{1-p_1}\sqrt{n}) - t_{n-1,1-\beta}(u_{1-p_2}\sqrt{n}) \right| \tag{3.4}$$

2. Is σ bekend, dan volgt uit (3.1) met $\underline{s} = \sigma$: $k\sqrt{n} - \delta_p = u_\gamma$.

Substitutie van $p = p_1$, $\gamma = \alpha$ resp. $p = p_2$, $\gamma = 1 - \beta$ geeft:

$$k\sqrt{n} = u_\alpha + u_{1-p_1}\sqrt{n} = u_{1-\beta} + u_{1-p_2}\sqrt{n} ,$$

oftewel

$$\boxed{n_\sigma = \left(\frac{u_{1-\alpha} + u_{1-\beta}}{u_{1-p_1} - u_{1-p_2}}\right)^2 \quad (3.5) \text{ en } \quad k_\sigma = \frac{u_{1-p_1}u_{1-\beta} + u_{1-p_2}u_{1-\alpha}}{u_{1-\alpha} + u_{1-\beta}}} \tag{3.6}$$

3. Voor $\alpha = \beta$ wordt $k_\sigma = \frac{1}{2}(u_{1-p_1} + u_{1-p_2})$ (3.7)

4. Uiteraard wordt n_σ op het dichtstbijzijnde natuurlijke getal afgerond.

4. BENADERDE OPLOSSING VOOR n_s EN k_s

Daar de vergelijkingen (3.3) moeilijk zonder computer op te lossen zijn, heeft men gezocht naar een benaderde oplossing.

Voor de hand ligt de benadering: $\underline{z} \sim N(\mu_z, \sigma_z^2)$

We beschouwen 3 benaderingen en stellen even $k_s = (1 + a_{n_s})k_\sigma$.

I. Stange [3] benadert μ_z door $\mathcal{C}(\underline{\bar{x}} - k\underline{s}) \approx \mu - k\sigma$ en $\text{var}\,\underline{z} = \text{var}(\underline{\bar{x}} - k\underline{s}) \approx \frac{\sigma^2}{n}(1 + k^2/2)$; $P(\underline{z} < L) = \gamma$ geeft $\frac{L - \mu_z}{\sigma_z} = u_\gamma$ dus

$$\frac{\mu + u_p\sigma - (\mu - k\sigma)}{\sigma/\sqrt{n}\ \sqrt{1+k^2/2}} = u_\gamma \quad \text{oftewel} \quad \frac{(k+u_p)\sqrt{n}}{\sqrt{1+k^2/2}} = u_\gamma \ . \tag{4.1}$$

Substitutie van $p = p_1$, $\gamma = \alpha$ resp. $p = p_2$, $\gamma = 1-\beta$ in (4.1) geeft:

$$\frac{(k+u_{p_1})\sqrt{n}}{\sqrt{1+k^2/2}} = u_\alpha \quad \text{en} \quad \frac{(k+u_{p_2})\sqrt{n}}{\sqrt{1+k^2/2}} = u_{1-\beta} \ ,$$

met als oplossing

$$k_s = \frac{u_{1-p_1}u_{1-\beta} + u_{1-p_2}u_{1-\alpha}}{u_{1-\alpha} + u_{1-\beta}} \tag{4.2}$$

en

$$n_s = \left(\frac{u_{1-\alpha} + u_{1-\beta}}{u_{1-p_1} - u_{1-p_2}}\right)^2 \left(1 + k_s^2/2\right) \tag{4.3}$$

Hier zien we, met (3.6) en (3.5) dat $k_s = k_\sigma$ dus $a_{n_s} = 0$ en

$$n_s = (1 + k_\sigma^2/2)n_\sigma \tag{4.4}$$

II. Hamaker [1] bepaalt n_s en k_s zó dat $\bar{x} - k_\sigma\sigma$ (met steekproefgrootte n_σ) en $\bar{x} - k_s s$ (met steekproefgrootte n_s) hetzelfde gemiddelde en dezelfde variantie hebben:

$$\mu - k_\sigma\sigma \approx \mu - k_s\left(\frac{4n_s - 5}{4n_s - 4}\right)\sigma \text{ zie (2.7) en } \frac{1}{n_\sigma} \approx \frac{1}{n_s} + \frac{k_s^2}{2(n_s - 1)} \ .$$

Hieruit volgt:

$$n_s - 1 \approx n_\sigma\left(1 + \frac{k_s^2}{2} - \frac{1}{n_s}\right) \approx n_\sigma\left(1 + k_\sigma^2/2\right) \ ,$$

dus

$$k_s = \left(1 + \frac{1}{4n_s - 5}\right)k_\sigma \quad (4.5) \quad \text{en} \quad n_s = \left(1 + k_\sigma^2/2\right)n_\sigma + 1 \quad (4.6)$$

Bij substitutie van k_σ en n_σ moet men de niet-afgeronde waarden nemen uit (3.6). n_s mag wel afgerond worden op natuurlijk getal; hier is $a_{n_s} = 1/(4n_s - 5)$.

We zien dat in (4.6) n_s 1 groter is dan n_s in (4.4); (4.6) blijkt correct te zijn.

III. Daar ik bij gegeven α, β, p_1 en p_2 via de computer meteen de exacte waarden (dwz.afgerond op 2 decimalen) van n_s en k_s kon berekenen, was ik in staat een kleine correctie aan te brengen op a_{n_s} onder II, en wel als volgt:

als $n_s < 10$ neem $a_{n_s} = 1/(4n_s - 1)$;

voor $n_s \geq 10$ neem $a_{n_s} = 1/(5{,}6n_s - 1)$.

In de volgende tabel kan men de resultaten van de 3 benaderingen vergelijken.

α $u_{1-\alpha}$	β $u_{1-\beta}$	p_1 u_{1-p_1}	p_2 u_{1-p_2}	n_s	$k_s=k_\sigma$ I	k_s II	k_s III	k_s computer
.10 u=1,282	.10	.01	.2853 u=.567	5	1.446 *	1.54	1.52	1.52
.10	.10	.02 u=2.05	.2483 u=.648	7	1.349 *	1.41	1.40	1.40
.05 u=1.645	.05	.0281 u=1.91	.306 u=.506	10	1.208 *	1.24	1.23	1.23
.05	.05	.01	.197 u=.858	12	1.592 *	1.63	1.62	1.62
.05	.05	.0217 u=2.20	.203 u=.831	13	1.516 *	1.55	1.54	1.54
.01 u=2.326	.01	.0217	.199 u=.846	26	1.523 *	1.54	1.53	1.53
.07 u=1.475	.10	.01	.080 u=1.407	25	1.834	1.85	1.85	1.85
.05	.01	.01	.081 u=1.406	55	1.945	1.95	1.95	1.95

Opmerkingen

1. (4.6) geeft de exacte waarde voor n_s; Stange [3] geeft steeds 1 te laag.

2. * dwz. $\alpha = \beta$ dus met (3.6) $k_\sigma = (u_{1-p_1} + u_{1-p_2})/2$.

3. $k_s = (1 + a_{n_s})k_\sigma$ met k_σ berekend uit (3.6) en voor I: $a_{n_s} = 0$;

voor II: $a_{n_s} = 1/(4n_s - 5)$; voor III: als $n_s < 10$ dan $a_{n_s} = 1/(4n_s - 1)$

als $n_s \geq 10$ dan $a_{n_s} = 1/(5{,}6n_s - 1)$.

REFERENTIES

[1] Hamaker, H.C., "Acceptance sampling for percent defective by variables and by attributes". The Journal of Quality Technology, vol. 11, no.3, July 1979.

[2] Kreijger, P.C., "Controle druksterkte van beton volgens ontwerp VB 1972", Cement XXIII, 1971, nr. 3.

[3] Stange, K., "Stichproben-pläne für messende Prüfung. Aufstellung und Handhabung mit Hilfe des doppelten Wahrscheinlichkeitsnetzes". (±1960), Universiteit van Berlijn.

PUZZEL(M)UURTJE

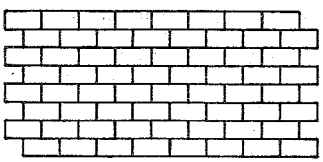
door

C.J. Bouwkamp en J.M.M. Verbakel

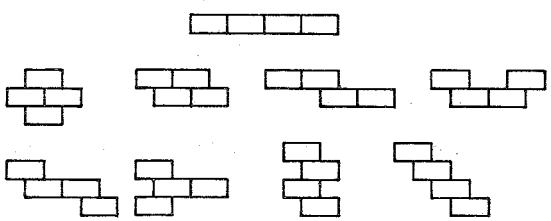
Opgedragen aan J.J. Seidel ter gelegenheid van zijn afscheid van de Onderafdeling der Wiskunde en Informatica.

Figuur 1 toont een stuk metselwerk van 64 stenen gebouwd in 8 horizontale lagen van ieder 8 stenen, hierna *muurtje* genoemd.

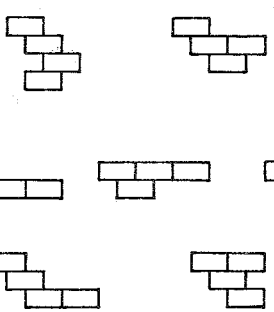
Figuur 1.



Figuur 2 toont de 16 verschillende (modulo translatie, draaiing en spiegeling) mogelijke combinaties van 4-steens *brokken* in metselverband.



Figuur 2.



De vraag is of het muurtje kan worden afgebroken in brokken zodanig dat elk brok van Figuur 2 één en slechts één keer in de sloop voorkomt.

We zullen ons iets positiever opstellen: geen sloop, maar bouw, en de vraag wordt, kunnen we het muurtje metselen met de 16 aangegeven brokken?

Het ligt voor de hand - wanneer je voor de eerste maal de puzzel onder ogen krijgt - te refereren aan pentomino-problemen, met hun vele, vaak zeer vele, oplossingen. Daarom maak je al snel kartonfiguren van de brokken en probeer je ze in te passen in Figuur 1. Het is niet moeilijk, 15 van de 16 brokken te plaatsen, ook zó dat een gat overblijft in de vorm van een brok dat net was gebruikt en dus niet meer beschikbaar is. In deze situatie is bij een pentomino-puzzel vaak een oplossing nabij: enkele stukjes verplaatsen en daar heb je een oplossing.

Ondanks vele pogingen van diverse puzzelaars, waaronder Oskar van Deventer, de aangever en (her?)uitvinder van de puzzel, werd geen enkele oplossing gevonden, ook niet door een computer-programma door Oskar enige tijd gedraaid.

Op zo'n ogenblik denk je aan het schaakbord met twee missende hoekvelden aan de uiteinden van een hoofddiagonaal, dat niet door dominostenen kan worden bedekt.

We zullen nu bewijzen dat het muurtje van Figuur 1 niet met de geprefabriceerde brokken van Figuur 2 kan worden opgebouwd. Zo'n bewijs gaat met *kleuring* en *pariteit*, net zoals in het geval van het schaakbord.

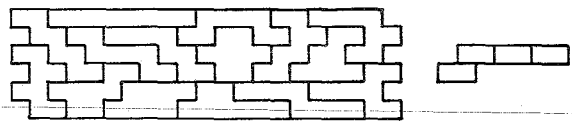
De 8 horizontale lagen van het muurtje kleuren we afwisselend zwart en wit. Vervolgens construeren we van elk der 16 brokken de pariteit. Dat gaat zo: zet het brok op Figuur 1, neem het absolute verschil van de aantallen zwarte en witte stenen (z,w) door het brok bedekt, en dat alles modulo 4:

$$\text{pariteit} = |z - w| \bmod 4 \ .$$

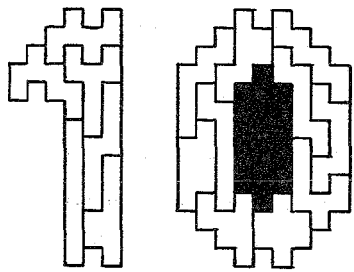
Het bovenste brok van Figuur 2 kan zowel 4 zwarte als 4 witte stenen bedekken. Zijn pariteit is dus 0. De volgende 10 brokken kunnen elk 2 zwarte en 2 witte stenen bedekken. Ook hun pariteit is dus 0. De laatste 5 brokken hebben elk pariteit 2, omdat ze elk 1 witte en 3 zwarte of 1 zwarte en 3 witte stenen kunnen bedekken. De totale pariteitsom van alle brokken tesamen is dus 2. Daarentegen is de pariteitsom van de muur als één geheel kennelijk 0. Ergo: het muurtje kan niet worden gemetseld uit de 16 brokken.

Gelukkig voor de puzzelaar kan je nog wel andere bouwsels bedenken. Bijvoorbeeld, leg een der brokken met pariteit 2 terzijde en probeer met de resterende 15 brokken een muur te construeren van 6 lagen elk van 10 stenen. Zo'n oplossing is aangegeven in Figuur 3.

Figuur 3.



Tot besluit, een "tien" (Figuur 4) voor de man die, meer dan een kwart eeuw geleden, de fundamenten metselde van de huidige Onderafdeling der Wiskunde en Informatica van de THE.



Figuur 4.

C.J. Bouwkamp
Technische Hogeschool Eindhoven
Onderafdeling der Wiskunde en
Informatica

J.M.M. Verbakel
Philips Research Laboratories
Eindhoven
The Netherlands

ASYMPTOTICS OF ITERATION

by

J.J.A.M. Brands and M.L.J. Hautus

Dedicated to J.J. Seidel on the occasion of his retirement.

Abstract. A uniform approach is presented to obtain the asymptotic behaviour of iteration sequences $x_{n+1} = F(n,x_n)$ for a large class of functions F. The treatment consists of two parts: First, a suitable transformation is given such that from the resulting equation the asymptotics can be derived in a straightforward and standard way; secondly, a theorem is given which yields the asymptotics via a comparison with solutions of an associated differential equation.

1. INTRODUCTION

The asymptotic behaviour of solutions of iteration problems has repeatedly been a topic of investigation at the Department of Mathematics in Eindhoven. To a great extent, interest in this matter was inspired by N.G. de Bruijn's book on Asymptotic Analysis [1], in which a whole chapter is devoted to this subject. As a result of these activities, a number of problems were published in the Problem Section of the "Nieuw Archief voor Wiskunde", of which [2], [3] and [4] can be handled by the methods of this paper. Also, J.J. Seidel has shown a keen interest in this topic. In a course on recurrence relations

given for students in industrial engineering, he taught the use of iterations in economic cobweb phenomena and growth models. He became so enthousiastic about the subject that he contributed to the lecture notes with an interesting new section on stability and bifurcation [5].

The variety of behaviours which solutions of iterations can have, is so immense that it seems impossible to develop something like a coherent general theory. Nevertheless, it turns out that a rather large number of problems can be handled in a unified way. The purpose of this paper is to demonstrate this fact for the class of iterations

$$(1) \qquad a_{n+1} = a_n + \psi(a_n) \ , \quad a_1 > 0,$$

where $\psi : (0,\infty) \to (0,\infty)$ has the property $\psi(x) = o(x) \quad (x \to \infty)$. Iterations of the kind

$$(2) \qquad x_{n+1} = x_n - \varphi(x_n) \ , \quad x_1 \in (a,b) \ ,$$

where φ is a function on (a,b) such that $0 < \varphi(x) < x-a \quad (a < x < b)$ and $\varphi(x) = o(x-a) \quad (x \downarrow a)$, are reduced to (1) by the transformation $a_n := (x_n - a)^{-1}$. Cobweb iteration can be handled by a transformation of the kind $a_n := (x_{2n+1} - a)^{-1}$.

In Section 2, a method will be presented which, for a large class of functions ψ, gives the asymptotic behaviuour of a_n. In many cases, the method provides a means to get as many terms of the asymptotics as we want.

In Section 3, we present a theorem that provides us with an easy way of obtaining the main term of the asymptotic behaviour of a_n for a large class

of functions . This class does not coincide with the class of functions to which the methods of Section 2 apply, but there is much overlap. The theorem even covers many cases of nonautonomous iteration

(3) $\qquad a_{n+1} + \psi(n, a_n)$.

Finally, in Section 4, the method of Section 2 is extended succesfully to various nonautonomous iteration problems. This becomes possible by the use of the asymptotic behaviour, obtained by Theorem (17) of Section 3.

2. THE METHOD

The ideas are the following: We try to find an invertable function h such that the transformation

(4) $\qquad h(a_n) =: b_n$

applied to (1) leads to an equation of the kind

(5) $\qquad b_{n+1} - b_n = u(b_n)$,

where u has the property

(6) $\qquad u(x) = 1 + o(1) \quad (x \to \infty)$.

If we succeed in determining such a function h then equation (5) enables us to find the asymptotic behaviour of b_n (with an obtainable accuracy determined by smoothness and monotonicity properties of u) by means of successive substitutions. Subsequently, the asymptotic behaviour of a_n can be determined via (4). Execution of transformation (4) leads to

(7) $$b_{n+1} - b_n = h(a_n + \psi(a_n)) - h(a_n) \ .$$

Heuristically, taking into account that we want $b_{n+1} - b_n \sim 1 \quad (n \to \infty)$, and approximating the right-hand side of (7) by $\psi(a_n)h'(a_n)$, we expect that a positive solution h of

(8) $$h'(x)\ \psi(x) = 1$$

will do. The positive solutions of (8) are

(9) $$h(x) = \int^{x} (\psi(s))^{-1} ds \ .$$

Denoting the inverse function of h by y we obtain for u in (5)

(10) $$u(x) = \int_{t}^{t+\psi(t)} (\psi(s))^{-1} ds \Bigg|_{t=y(x)} \ .$$

Clearly, condition (6) is fulfilled if

(11) $$\int_{t}^{t+\psi(t)} (\psi(s))^{-1} ds \to 1 \qquad (t \to \infty) \ .$$

About the function y we want to remark that it satisfies

(12) $$y' = \psi(y) \ .$$

This differential equation can be considered as the continuous analogue of the difference equation (1).

(13) *REMARK.* It is not necessary to use h given by (9). In many cases a good approximation of h will also be satisfactory.

(14) *EXAMPLE.* Consider the iteration

$$x_{n+1} = x_n - x_n^3 , \quad 0 < x_1 < 1 .$$

The substitution $a_n := x_n^{-1}$ leads to

$$a_{n+1} = a_n + (a_n^2 - 1)a_n , \quad a_1 > 1 .$$

A function satisfying (9) is $h(x) = \frac{1}{2}x^2 - \log x$. With (13) in mind and in order to achieve a simplification of the computation we try the substitution $b_n := \frac{1}{2}a_n^2$. We get

$$b_{n+1} - b_n = 1 + \frac{3}{2}(2b_n - 1)^{-1} + \frac{1}{2}(2b_n - 1)^{-2} .$$

Using successive substitution, we derive from this equation an asymptotic expansion for b_n and subsequently for $x_n = (2b_n)^{-\frac{1}{2}}$. ("$\approx$" signifies that the right-hand side is an asymptotic expansion.)

$$b_n \approx n + \frac{3}{4}\log n + c + \frac{3}{4}n^{-1}\log n + (\frac{1}{2} - c)n^{-1} + \ldots \qquad (n \to \infty) .$$

$$x_n \approx (2n)^{-\frac{1}{2}} - 3(8\sqrt{2})^{-1} n^{-3/2}\log n - c(2\sqrt{2})^{-1} n^{-3/2} + \ldots (n \to \infty) .$$

(15) *EXAMPLE.* In order to find the asymptotics of

$$a_{n+1} = a_n + e^{-a_n} ,$$

we substitute $b_n := e^{a_n}$. We get

$$b_{n+1} - b_n = (\exp(b_n^{-1}) - 1)b_n \approx 1 + \frac{1}{2}b_n^{-1} + \frac{1}{6}b_n^{-2} + \ldots \qquad (n \to \infty) ,$$

whence

$$b_n \approx n + \frac{1}{2}\log n + c + \frac{1}{2}n^{-1}\log n + (c - \frac{1}{6})n^{-1} + \ldots \qquad (n \to \infty)$$

and subsequently

$$a_n \approx \log n + \frac{1}{2}n^{-1}\log n + c\,n^{-1} - \frac{1}{4}n^{-2}(\log n)^2 + \ldots \qquad (n \to \infty) .$$

(16) *EXAMPLE*. We consider equation (1) with $\psi(x) = (x + \frac{1}{2}x \sin x^{\frac{1}{2}})^{-1}$. Clearly condition (11) is satisfied. With (13) in mind, we try $b_n := \frac{1}{2}a_n^2 - a_n^{3/2} \cos a_n^{\frac{1}{2}}$. We get

$$b_{n+1} - b_n = 1 + \frac{3}{2}a_n^{-\frac{1}{2}}(1 + \frac{1}{2}\sin a_n^{\frac{1}{2}})^{-1} \cos a_n^{\frac{1}{2}} + O(a_n^{-3/2}) \quad (n \to \infty).$$

From $a_n^2(1 - 2a_n^{-\frac{1}{2}} \cos a_n^{\frac{1}{2}}) = 2b_n$, we infer that $a_n = (2b_n)^{\frac{1}{2}} + O(b_n^{\frac{1}{4}})$ $(n \to \infty)$. It follows that $b_{n+1} - b_n = O(b_n^{-\frac{1}{4}})$ $(n \to \infty)$, whence $b_n = n + O(n^{\frac{3}{4}})$ $(n \to \infty)$. Hence

$$a_n = (2n)^{\frac{1}{2}} + O(n^{\frac{1}{4}}) \quad (n \to \infty).$$

The authors confess that they have not succeeded in finding sharper asymptotic results. Note that this example is not covered by Theorem (17).

3. THE MAIN TERM

In many cases the function ψ in equation (1) is nearly decreasing, i.e., ψ is a decreasing function plus a relatively small perturbation. We shall formulate a general result for nonautonomous iterations. First, we introduce some terminology. A function f is said to be decreasing if $x_1 > x_2$ implies $f(x_1) \leq f(x_2)$. (Strictly decreasing if $f(x_1) < f(x_2)$.) By a solution of a differential equation $y' = f(x,y)$ with a discontinuous function f, we mean a continuous function y satisfying an integral equation of the kind $y(x) = y(a) + \int_a^x f(s,y(s))ds$ $(x > a)$. We denote $[1,\infty)\times[\alpha,\infty)$ by D_α.

(17) *THEOREM.* Let $\psi_1 : D_\alpha \to (0,\infty)$ have the properties:

(i) For every $v \geq \alpha$ the function $\psi_1(\cdot,v)$ is decreasing, and for every $u \geq 1$ the function $\psi_1(u,\cdot)$ is decrasing,

(ii) $\psi_1(u,v) \to 0$ if $\min\{u,v\} \to \infty$.

(iii) $\sum_{n=1}^{\infty} \psi_1(n,v)$ diverges for every $v \geq \alpha$.

Let $\psi_2 : D_\alpha \to \mathbb{R}$ be such that $\psi := \psi_1 + \psi_2$ is positive on D_α and $\psi_2(u,v) = o(\psi_1(u,v))$ $(\min\{u,v\} \to \infty)$.

Let y be a positive solution of

$$y' = \psi_1(x,y) \ , \tag{18}$$

and (a_n) a sequence in $[\alpha,\infty)$ satifying

$$a_{n+1} - a_n = \psi(n,a_n) \qquad (n \in \mathbb{N}) \ . \tag{19}$$

Then we have

$$a_n = y(n)\ (1 + o(1)) \qquad (n \to \infty). \tag{20}$$

In the case that $\psi_2 = 0$ we have

$$a_n = y(n) + c + o(1) \qquad (n \to \infty). \tag{21}$$

If, in addition, equation (18) has the property that the difference of any two positive solutions tends to zero for $x \to \infty$, then the constant c in (21) is zero.

(22) *COROLLARY.* If $\psi : (0,\infty) \to (0,\infty)$ is decrasing, and (a_n) is a positive sequence satisfying (1), and y is a positive solution of $y' = \psi(y)$, then

$$a_n = y(n) + o(1) \qquad (n \to \infty).$$

The result of Corollary (22) has already been published as a problem [3].

PROOF OF THEOREM (17). The proof consists of two parts. In the first part, we prove Theorem (17) for the special case $\psi_2 \equiv 0$. In the second part we compare the solution sequence (a_n) of (19) with solution sequences of special cases of (19) with $\psi_2 \equiv 0$.

(23) *LEMMA*. Let ψ_1 and y be as in Theorem (17). Let (c_n) satisfy: $c_1 \geq \alpha, c_{n+1} = c_n + \psi_1(n,c_n)$ $(n \geq 1)$. Then there exists a constant c such that

(24) $$c_n = y(n) + c + o(1) \qquad (n \to \infty).$$

The constant c is zero if ψ_1 has the property mentioned after (21)

Proof. Let $k \in \mathbb{N}$. Let z be the solution of (20) with initial value $z(k) = c_k$. Then $z(n) \leq c_n$ $(n \geq k)$. For suppose that $m > k$ is the smallest integer such that $z(m) > c_m$. Then, clearly, there is a $T \in [m-1,m)$ such that $z(T) = c_{m-1}$. It follows that $z(m) = z(T) + \int_T^m \psi_1(s,z(s))ds \leq z(T) + (m-T)\psi_1(T,z(T))$ $\leq c_{m-1} + (m-T)\psi_1(m-1,c_{m-1}) \leq c_m$, contrary to the supposition.

Let $m > k$. Then $z(m) - z(m-1) = \int_{m-1}^m \psi_1(s,z(s))ds \geq \psi_1(m,z(m)) \geq \psi_1(m,c_m) =$ $= c_{m+1} - c_m$. Summing over m from $k+1$ to $n \geq k+1$ we find $z(n) \geq c_{n+1} - c_{k+1} + z_k$ $> c_n - \psi_1(k,c_k)$. So, by now we have

(25) $$0 \leq c_n - z(n) \leq \psi_1(k,c_k) \qquad (n \geq k+1).$$

Furthermore $c_n \to \infty$ $(n \to \infty)$. For c_n is increasing. Suppose $c_n \to \ell$. Then $\ell > c_n$, whence $c_{n+1} - c_k \geq \psi_1(n,\ell)$. Using condition (iii) in Theorem (17) we arrive at a contradiction. Now, on account of condition (ii), we can state that $\psi_1(n,c_n) \to 0$ if $n \to \infty$.

Now we proceed as follows. Let $\varepsilon > 0$. There exists a $k \in \mathbb{N}$ such that $\psi_1(k,c_k) < \frac{1}{2}\varepsilon$. Then z, defined as above, satisfies $0 \leq c_n - z(n) < \frac{1}{2}\varepsilon$ $(n \geq k+1)$ according to (25). Finally, we shall prove in due course that $\lim_{x\to\infty}(y(x) - z(x))$ exists. Once we have proved this, the conclusion (24) follows from the fact that $c_n - y(n)$ is a Cauchy sequence. Indeed,

$$|(c_n - y(n)) - (c_m - y(m))| \leq |c_n - z(n)| + |c_m - z(m)| +$$

$$|(z(n) - y(n)) - (z(m) - y(m))| < \varepsilon$$

for n and m sufficiently large. □

(26) *LEMMA*. Let y_1 and y_2 be two positive solutions of (18), defined on some interval $[x_0,\infty)$ (with $x_0 \geq \alpha$), and $y_1(x_0) > y_2(x_0)$. Then $y_1(x) > y_2(x)$ $(x \geq x_0)$, $y_1(x) - y_2(x)$ is decreasing, and converges.

Proof. From conditions (i) and (iii) of Theorem (17) we infer that every positive solution of (18) grows without limit. Hence, there is a $T > 0$ such that $y_2(x_0 + T) = y_1(x_0)$. We claim that $y_2(x + T) \leq y_1(x)$ $(x \geq x_0)$. For suppose to the contrary that $y_2(x_1 + T) > y(x_1)$ for some $x_1 > x_0$. Let $x_2 := \max\{x < x_1 | y_2(x+T) = y_1(x)\}$. Clearly $y_2(x+T) > y_1(x)$ on $(x_2, x_1]$. Then $y_2(x_1 + T) - y_1(x_1) = \int_{x_2}^{x_1} (\psi_1(s+T, y_2(s+T)) - \psi_1(s, y_1(s)))ds \leq 0$, a contradiction. So we have now $y_2(x) < y_2(x+T) \leq y_1(x)$ $(x \geq x_0)$.

Furthermore $y_1(t) - y_2(t) = y_1(s) - y_2(s) + \int_s^t (\psi_1(s,y_1(s)) - \psi(s,y_2(s)))ds$ $\leq y_1(s) - y_2(s)$ $(x_0 \leq s < t)$. Hence, $y_1(x) - y_2(x)$ tends decreasingly to a nonnegative limit if $x \to \infty$.

In the autonomous case there is a $T > 0$ such that $y_2(x+T) = y_1(x)$ $(x \geq x_0)$. It follows that $0 < y_1(x) - y_2(x) = \int_x^{x+T} \psi_1(y_2(s))ds \leq T\psi_1(y_2(x)) = o(1)$ $(x \to \infty)$. This proves also Corollary (22). □

(27) *REMARK*. We can replace the $o(1)$ term in (24) by a sharper result. Let z be as in the proof of Lemma (23), i.e., a solution of (20) with $z(k) = c_k$. We already know from Lemma (26) that $z(x) - y(x)$ tends monotonically to a limit, say $\ell(k)$. Then $|z(x) - y(x) - \ell(k)| =: u(x,k) \searrow 0$ $(x \to \infty)$. Combining this with (25) we have $\ell(k) - u(n,k) \leq c_n - y(n) = c_n - z(n) + z(n) - y(n) \leq \psi_1(k,c_k) + \ell(k) + u(n,k)$. By (24), letting $n \to \infty$, we find $\ell(k) \leq c \leq \psi_1(k,c_k) + \ell(k)$. It follows that $|c_n - y(n) - c| \leq \psi_1(k,c_k) + u(n,k)$ $(n > k)$. Now we can choose $k = k(n) < n$ in such a way that $k(n) \to \infty$ and $u(n,k(n)) \to 0$ if $n \to \infty$. An application will be shown in Example (31).

For the second part of the proof of Theorem (17) we need

(28) *LEMMA*. Let $\varphi : D_\gamma \to (0,\infty)$ satisfy the condition that for all $(u,v) \in D_\gamma$ $\varphi(u,\cdot)$ and $\varphi(\cdot,v)$ are decreasing. Let $n_0 \in \mathbb{N}$. Let (u_n) and (v_n) be sequences in $[\gamma,\infty)$ satisfying for all $n \in \mathbb{N}$: $u_{n+1} \geq u_n + \varphi(n+n_0,u_n)$, $v_{n+1} \leq v_n + \varphi(n+n_0,v_n)$. Then $v_n \leq u_n + \max\{v_1 - u_1, \varphi(n_0,\gamma)\}$ $(n \in \mathbb{N})$.

Proof. Let us assume that $v_k > u_k$ $(m \le k \le n)$, where $n \ge m \ge 1$. Then obviously $v_{k+1} - v_k \le u_{k+1} - u_k$ $(m \le k \le n)$, whence $v_\ell \le u_\ell + (v_m - u_m)$ $(m+1 \le \ell \le n+1)$. If $m > 1$ and $v_{m-1} \le u_{m-1}$ then $v_m - u_m \le v_{m-1} - u_{m-1} + \varphi(m-1, v_{m-1}) - \varphi(m-1, u_{m-1}) < \varphi(m-1, v_{m-1}) \le \varphi(n_0, \gamma)$. Clearly the conclusion of the lemma follows. □

Now we are in a position to complete the proof of Theorem (17). Let (a_n) and y satisfy the conditions of Theorem (17). Then conditions (i) and (iii) imply that a_n and $y(x)$ tend to infinity. Let $0 < \varepsilon < 1$. Choose k so large that $|\psi_2(n,a_n)|(\psi_1(n,a_n))^{-1} < \varepsilon$ $(n \ge k)$. Clearly (a_n) satisfies the inequalities $|a_{n+1} - a_n - \psi_1(n,a_n)| < \varepsilon\psi_1(n,a_n)$ $(n \ge k)$. We introduce two sequences (p_n) and (q_n) in $[\alpha,\infty)$ satisfying

$$p_{n+1} = p_n + (1+\varepsilon)\psi_1(n,p_n) \qquad (n \ge k)\ ,$$

$$q_{n+1} = q_n + (1-\varepsilon)\psi_1(n,q_n) \qquad (n \ge k)\ .$$

We apply Lemma (28) twice; first we apply it with $n_0 := k$, $\varphi := (1+\varepsilon)\psi_1$, $u_n := p_{n+k-1}$ and $v_n := a_{n+k-1}$; secondly we take $n_0 := k$, $\varphi := (1-\varepsilon)\psi_1$, $u_n := a_{n+k-1}$ and $v_n = q_{n+k-1}$. We get $q_n - \beta_2 \le a_n \le p_n + \beta_1$ $(n \ge k)$, where β_1 and β_2 are independent of n. Applying now Lemma (25) to the sequences (p_n) and (q_n) we get

$$p_n = y(n+\varepsilon n) + O(1) \quad (n \to \infty) \text{ and } q_n = y(n-\varepsilon n) + O(1) \quad (n \to \infty).$$

From

$$0 < y(n+\varepsilon n) - y(n) \le y(n) \quad (n-\varepsilon n) \le \varepsilon(1-\varepsilon)^{-1}(n-\varepsilon n)\psi_1(n-\varepsilon n, y(n-\varepsilon n)) \le$$

$$\le \varepsilon(1-\varepsilon)^{-1}(y(n-\varepsilon n) - y(\alpha))(n-\varepsilon n-\alpha) < \varepsilon(1-\varepsilon)^{-1}\, y(n)$$

it follows that

$$y(n+\varepsilon n) - y(n) = o(y(n)) \quad (n \to \infty) \text{ and } y(n) - y(\varepsilon - \varepsilon n) = o(y(n)) \quad n \to \infty).$$

Combining the foregoing inequations for a_n and the asymptotic expressions for p_n and q_n we arrive at the conclusion (20) of Theorem (17). □

(29) *REMARK.* In the same spirit as in Remark (27) we can replace the $o(1)$ term in (20) by a sharper result. From the proof of Lemma (23) we have $0 \le c_n - z(n) \le \psi_1(k,c_k)$ $(n \ge k)$, and $|z(n) - y(n)| \le |z(k) - y(k)| = |c_k - y(k)|$ $(n \ge k)$. It follows that

$$(30) \qquad -|c_k - y(k)| \le c_n - y(n) \le |c_k - y(k)| + \psi_1(k,c_k) \quad (n \ge k).$$

We define $\varepsilon = \varepsilon(k) := \max\{|\psi_2(n,a_n)|(\psi_1(n,a_n))^{-1} | \ n \ge k\}$. Let k_1 be such that $0 < \varepsilon < \frac{1}{2}$ $(k \ge k_1)$. We prescribe the initial values $p_k = q_k = a_k$. Then $\beta_1 = (1+\varepsilon)\psi_1(k,a_k)$ and $\beta_2 = (1-\varepsilon)\psi_1(k,a_k)$. We apply (30) twice. First c_n, $y(n)$ and ψ_1 are replaced by p_n, $y(n+\varepsilon n)$ and $(1+\varepsilon)\psi_1$, respectively, secondly by q_n, $y(n-\varepsilon n)$ and $(1-\varepsilon)\psi_1$. Taking all together we get

$$|a_n - y(n)| \le |a_k - y(k)| + \varepsilon(1-\varepsilon)^{-1}(y(n) + y(k)) + (2+\varepsilon)\psi_1(k,a_k)$$

$$(n \ge k \ge k_1).$$

Using the fact that $\psi_1(k,a_k) = O(k^{-1}y(k))$ $(k \in \mathbb{N})$ and choosing $k = k(n) < n$ with $k(n) \to \infty$ and $y(k(n))/y(n) \to 0$ if $n \to \infty$, we can find something better than a mere $o(1)$. For instance, if $\varepsilon(k) = k^{-\frac{1}{2}}$ and $y(n) = n^{\frac{3}{4}}$ then it is not difficult to prove that $a_n = y(n)(1 + O(n^{-\frac{1}{2}}))$ $(n \to \infty)$.

(31) *EXAMPLE*. Consider the nonautonomous iteration

$$a_1 \geq 0, \quad a_{n+1} = a_n + (n^{\frac{1}{2}} + a_n)^{-1} \qquad (n \in \mathbb{N}) .$$

A particular solution of $y' = (x^{\frac{1}{2}} + y)^{-1}$ is $y(x) = x^{\frac{1}{2}}$. All other positive solutions are of the kind $x^{\frac{1}{2}} + O(x^{-\frac{1}{4}})$, $(x \to \infty)$.

Hence, according to Theorem (17)

$$a_n = n^{\frac{1}{2}} + o(1) \qquad (n \to \infty) .$$

Obviously, there is a k_1 such that $|a_k - k^{\frac{1}{2}}| \leq 1 \quad (k \geq k_1)$. Let z be a solution of $z' = (x^{\frac{1}{2}} + z)^{-1}$ with $z(k) = a_k$. Omitting details we mention that $u(n,k) := |z(n) - n^{\frac{1}{2}}| < 2k^{\frac{1}{4}}n^{-\frac{1}{4}}$, $(n \geq k \geq k_1)$. By Remark (27) we know that $|a_n - n^{\frac{1}{2}}| \leq (k^{\frac{1}{2}} + a_k)^{-1} + u(n,k)$. Choosing $k(n) = n^{1/3}$ we find that

$$a_n = n^{\frac{1}{2}} + O(n^{-1/6}) \qquad (n \to \infty) .$$

For stronger results we refer to Example (41).

(32) *EXAMPLE*. We consider

$$a_1 \geq 0, \quad a_{n+1} = a_n + (n + e^{a_n})^{-1} \qquad (n \in \mathbb{N})$$

The differential equation to be considered is $y' = (x + e^y)^{-1}$. The inverse function h of y satisfies $h' = h + x$ with solutions $h_c(x) = xe^x + Ce^x$. The inverse functions y_c of h_c are easily seen to satisfy $y_c(x) = \log x - \log\log x + o(1) \quad (x \to \infty)$. Hence,

$$a_n = \log n - \log\log n + o(1) \qquad (n \to \infty) .$$

(33) *EXAMPLE*. We consider

$$a_1 = 1,\ a_{n+1} = a_n + ([na_n])^{-1} \qquad (n \in \mathbb{N}) .$$

We can write the right-hand side as $\psi_1(n,a_n) + \psi_2(n,a_n)$ with $\psi_1(n,a_n) = (na_n)^{-1}$. Then clearly $\psi_2(n,a_n) = o(\psi_1(n,a_n))$ $(\min\{n,a_n\} \to \infty)$. A positive solution of $y' = \psi_1(x,y)$ is $y(x) = (2 \log x)^{\frac{1}{2}}$. Hence $a_n = (1 + o(1))(2 \log n)^{\frac{1}{2}}$ $(n \to \infty)$. But we can do better by solving $y' = ([xy])^{-1}$. It is not difficult to prove that $y(x) = (2 \log x)^{\frac{1}{2}} + \mathcal{O}((\log x)^{-\frac{1}{2}})$ $(x \to \infty)$.

Hence

$$a_n = (2 \log n)^{\frac{1}{2}} + o(1) \qquad (n \to \infty) .$$

(34) *EXAMPLE*. We will study

$$0 < x_1 < 1\ ,\ \ x_{n+1} = x_n - n^{-1}x_n^2 \qquad (n \in \mathbb{N}) .$$

The substitution $a_n = x_n^{-1}$ gives

$$a_1 > 1,\ \ a_{n+1} - a_n = (n - a_n^{-1})^{-1} \qquad (n \in \mathbb{N}) .$$

The analogous differential equation is $y' = (x - y^{-1})^{-1}$, $y(1) > 1$, and the differential equation for the inverse function h of y is $h' = h - x^{-1}$ with solutions $h(x) = C_1 e^x + e^x \int^{\infty} s^{-1} e^{-s} ds$ with $C_1 > 0$.
From $h(x) \approx C_1 e^x + x^{-1} - x^{-2}x + 2!x^{-3} + \ldots$ $(x \to \infty)$ it follows that

$$y(x) \approx \log x - \log C_1 + (x \log x)^{-1} + \ldots \qquad (x \to \infty) .$$

Hence

$$a_n = \log n + C + o(1) \qquad (n \to \infty)$$

and subsequently

$$x_n = \log^{-1} n - C \log^{-2} n + o(\log^{-2} n) \qquad (n \to \infty).$$

We are able to obtain sharper results as follows. Substituting the result for a_n in the right-hand side of the equation for (a_n) we get

$$a_{n+1} - a_n = n^{-1} + n^{-2}(\log n)^{-1} - Cn^{-2}(\log n)^{-2} + o(n^{-2}(\log n)^{-2})$$

$(n \to \infty)$,

whence

$$a_n = \log n + C - n^{-1}(\log n)^{-1} - (1 - C)n^{-1}(\log n)^{-2} + o(n^{-1}(\log n)^{-2})$$

$(n \to \infty)$.

By repeating this procedure one can compute arbitrarily many terms for the asymptotic behaviour of a_n.

(35) *REMARK*. This idea (substituting the asymptotic result for a_n in the right-hand side of the difference equation for (a_n)) fails to work in the examples (31) and (32).

(36) *EXAMPLE*. The asymptotic behaviour of (a_n) satisfying

$$a_2 \geq 0, \quad a_{n+1} - a_n = (a_n + \log n)^{-1} \qquad (n \geq 2),$$

can be found by solving $y' = (y + \log x)^{-1}$. The inverse function h of y satisfies $h' = \log h + x$. We easily get $h(x) = \frac{1}{2}x^2 + 2x \log x - x \log 2 + O(\log^2 x)$ $(x \to \infty)$, and subsequently $y(x) = (2x)^{\frac{1}{2}} - \log x + 2 + o(1)$ $(x \to \infty)$. Hence

$$a_n = (2n)^{\frac{1}{2}} - \log n + 2 + o(1) \qquad (n \to \infty) .$$

We remark that the insertion of this result in the right-hand side of the difference equation fails to give a sharper result. The method of Section 4 however, does improve the asymptotic formula.

(37) *EXAMPLE.* The function ψ in equation (1) is required to satisfy $\psi(x) = o(x)$ $(x \to \infty)$. For instance $\psi(x) = x^{\frac{1}{2}}$ satisfies that requirement. But Theorem (17) does not cover this situation. We have the following result. Let $\psi(u,v) := v^{-1}\varphi(u,v)$ satisfy the requirements of Theorem (17). Then the iteration $a_{n+1} = a_n + \varphi(n,a_n)$ is transformed by the substitution $\alpha_n = \log a_n$ into the iteration $\alpha_{n+1} = \alpha_n + \log(1 + e^{-\alpha_n}(n,e^{\alpha_n}))$ which is covered by Theorem (17).

(38) *EXAMPLE.* We consider

$$a_{n+1} = a_n + n^{-1}a_n^{\frac{1}{2}} .$$

The substitution $\alpha_n = \log a_n$ gives

$$\alpha_{n+1} = \alpha_n + \log(1 + n^{-1}e^{-\frac{1}{2}\alpha_n}) .$$

The differential equation $y' = \log(1 + x^{-1}e^{-\frac{1}{2}y})$ has positive solutions all of which satisfy $y(x) = 2 \log\log x - 2 \log 2 + o((\log x)^{-1})$ $(x \to \infty)$. Hence

$$\alpha_n = 2 \log\log n - 2 \log 2 + o(1) \qquad (n \to \infty),$$

and subsequently

$$a_n = \frac{1}{4}(\log n)^2(1 + o(1)) \qquad (n \to \infty) .$$

4. THE MIXED METHOD

In quite a few cases of nonautonomous iteration, one can find as many terms of an asymptotic expansion as one wishes if the asymptotic result of Theorem (17) is used in the method described in Section 2.

(39) *EXAMPLE*. We consider the interation of Example (36). Inspired by the knowledge that $h(x) \sim \frac{1}{2}x^2$ $(x \to \infty)$, we try the substitution $b_n := a_n^2$. We get

$$b_{n+1} - b_n = 2 - (a_n + \log n)^{-1} \log n + (a_n + \log n)^{-2} \quad . \tag{40}$$

Now, substituting the asymptotic result for a_n of Example (36) in the right-hand side of (40) we get

$$b_{n+1} - b_n = 2 - 2^{\frac{1}{2}} n^{-\frac{1}{2}} \log n + 2n^{-1} \log n (1 + o(1)) \qquad (n \to \infty) \quad .$$

It follows that

$$b_n = 2n - 2^{3/2} n^{\frac{1}{2}} \log n + 2^{5/2} n^{\frac{1}{2}} + (\log n)^2 (1 + o(1)) \quad (n \to \infty)$$

and subsequently

$$a_n = 2^{\frac{1}{2}} n^{\frac{1}{2}} - \log n + 2 + 3.2^{-7/2} n^{-\frac{1}{2}} (\log n)^2 (1 + o(1)) \quad (n \to \infty) \quad .$$

Using this result in the right-hand side of (40) we get a better result. Repeating this procedure we can get as many terms of an asymptotic expansion as we want.

(41) *EXAMPLE*. Again we want to study the equation of Example (31). When we try to imitate blindly the method of Example (39) we will fail to get

anything better than what we already have by Theorem (17), i.e., $a_n = n^{\frac{1}{2}} + o(1)$ $(n \to \infty)$. The substitution $b_n := a_n^2$ gives

$$b_{n+1} - b_n = 1 + (4n)^{-1}(b_n - n) + r_n \,, \tag{42}$$

where

$$r_n = (4n)^{-1}(a_n + n^{\frac{1}{2}})^{-1}(a_n + 3n^{\frac{1}{2}})(a_n - n^{\frac{1}{2}})^2 \,.$$

If we use $a_n = n^{\frac{1}{2}} + o(1)$ $(n \to \infty)$ in the right-hand side of (42) we gain nothing. But the substitution $b_n = u_n p_n$, where $p_n := \prod_{k=1}^{n} (1 + (4k)^{-1}) \approx c_1 n^{\frac{1}{4}} + c_2 n^{-\frac{3}{4}} + \ldots$ $(n \to \infty)$, and the use of $a_n = n^{\frac{1}{2}} + o(1)$ $(n \to \infty)$ in the term r_n of (42) gives better results. We get

$$u_{n+1} - u_n = \tfrac{3}{4} c_1^{-1} n^{-\frac{1}{4}} + O(n^{-5/4}) \qquad (n \to \infty) \,.$$

It follows that $u_n = c_1^{-1} n^{\frac{3}{4}} + d_1 + O(n^{-\frac{1}{4}})$ $(n \to \infty)$, whence $b_n = n + d_1 n^{\frac{1}{4}} + O(1)$ $(n \to \infty)$ and subsequently

$$a_n = n^{\frac{1}{2}} + d_2 n^{-\frac{1}{4}} + O(n^{-\frac{1}{2}}) \qquad (n \to \infty) \,.$$

By repetition of this procedure we can get a complete asymptotic series.

(43) *REMARK*. The equation of example (32) is difficult. The method of this section fails to work.

REFERENCES

[1] Bruijn,.N.G. de, Asymptotic Methods in Analysis, 1981, Dover Publications, Inc., New York.

[2] Hautus, M.L.J., Problem 191, Nieuw Archief voor Wiskunde, 16 (1968), 126.

[3] Brands, J.J.A.M., Problem 355, Nieuw Archief voor Wiskunde, 21 (1973), 282.

[4] Hautus, M.L.J., Problem 407, Nieuw Archief voor Wiskunde, 23 (1975), 175.

[5] Seidel, J.J., Lecture Notes nr. 2272, p. 13 - 21, Eindhoven University of Technology, Department of Mathematics and Computing Science, The Netherlands.

FORMALIZATION OF CONSTRUCTIVITY IN AUTOMATH

by

N.G. de Bruijn

Dedicated to J.J. Seidel on the occasion of his 65th birthday.

1. INTRODUCTION

There are various systems in which a large part of mathematical activity is formalized. The general effect of the activity of putting mathematics into such a system is what one might call the unification of mathematics: different parts of mathematics which used to be cultivated separately get united, and methods available in one part get an influence in other parts.

Very typical for twentieth century mathematics is the unifying force of the concepts of set theory. And today one might say that the language of mathematics is the one of the theory of sets combined with predicate logic, even though one might disagree about the exact foundation one should give to these two.

Not everyone thinks of set theory and logic as being parts of a single formal system. Set theory deals with objects, and logic deals with proofs, and these two are usually considered as of a different nature. Nevertheless, there are possibilities to treat these two different things in a common

system in a way that handles analogous situations analogously indeed.

A system that goes very far in treating objects and proofs alike, is the *AUTOMATH* system (see [1]). In *AUTOMATH* there are expressions on three different levels, called degrees. Each expression of degree 3 has a "type" that is of degree 2, and each expression of degree 2 has a type of degree 1. Expressions of degree 1 do not have a type. There are two basic expressions of degree 1, viz. type and prop. The underlined word type should not be confused with the word type used more ore less colloquially when saying that each expression of degree 2 or 3 has a type.

We denote typing by a semicolon. If A has B as its type, we write A : B. So we can have

$$A : B : \underline{\text{type}} \tag{1}$$

and also

$$C : D : \underline{\text{prop}}. \tag{2}$$

The interpretation of (1) is that A is the name of an object (like the number 3), and that B is the name of the class from which that object is taken (it might be a symbol for the set of integers). The interpretation of (2) is that C is a name for a proof, and that D somehow represents the statement that is proved by C.

The main profit we have from this way of describing proofs and objects is the matter of substitutivity. If we have described an object depending on a

number of parameters, that description can be used under different circumstances by means of substitution: we replace the formal parameters by explicit expressions. The same technique is applicable to theorems: a theorem is intended for many applications, and such applications can be effectuated by substitution. The conditions of the theorem are modified by these substitutions too. If we study the matter more closely, we see that some of the parameters are object-like, and others are proof-like. The substitution machinery is the same for both. All this is effectively implemented in the *AUTOMATH* system.

2. ORIENTATION ON GEOMETRICAL CONSTRUCTIONS

On the fringe of mathematics there are mathematical activities which seem to be of a kind that does not fit into the pattern of objects and proofs. One such thing is the matter of geometrical constructions, a subject that goes back to Greek mathematics. A construction is neither an object nor a proof, but constructions are discussed along with geometrical objects, and along with proofs that show that the constructions construct indeed what is claimed to be constructed.

Since these geometrical constructions can also admit substitution for formal parameters, there is a case for creating facilities which handle a new kind of things along with objects and proofs. So we can think of a system that handles objects, proofs and geometrical constructions in more or less the same way.

If we think of geometrical constructions, there is a peculiarity that may not arise easily with other kinds of constructions: it is the matter of observability. Let us study a particular example in order to stress this point. Let there be given four points A, B, C and D in the plane. We assume that A, B and C are not on a line. Let M be the centre of the circle through A, B and C. We wish to construct the point P that is defined as follows. P is obtained from D by multiplication, with M as the multiplication center, and multiplication factor 1, 2 or 3. The factor is 1 if D lies inside the circle, 2 if D lies on the circle, and 3 if D lies outside the circle. If we want to carry out the construction of P, we have to know whether we are allowed to observe what the position of D with respect to the circle is. In particular this problem comes up for the practical question what should happen if there is insufficient precision for concluding whether D is inside or outside.

If we think of a construction with actual physical means like paper, pencil, ruler and compass, then the case of D lying exactly on the circle is, of course, undecidable.

The above construction problem may seem to be very artificial, but yet its main characteristic turns up in very many geometrical constructions: it is the fact that, at some point of the construction the result of some observation will decide the further course of the construction. An example where this will happen is the case of geometrical constructions that have to be carried out inside a given finite part of the plane.

The naive approach to observability may be formulated as the slogan "truth is observable" (see Section 4). Other possibilities will be sketched in Section 8-10.

A further thing one might like to formalize is selectability: one wants to be able to select an object from a set of objects one has constructed. For example, a construction of the intersection of two circles may produce two points, and we may wish to be able to "take one of them". In this case such a selection principle is not indispensable: one might describe the effect of the construction of the intersection as giving a labelled "first point" and a labelled "second point". But there is a stronger reason for implementing a selection principle: so often we have to "take an arbitrary point" at some stage of a construction. It should be noted that in such cases the final result of the entire construction does not depend on the particular point that was taken. In Section 5 we come back to this, in particular to the matter of the difference between "giving" and "taking" arbitrary points.

A description of all these features is possible in *AUTOMATH*. We have various options for doing it. The way we present this matter is necessarily arbitrary.

It is certainly not the intention of this note to give a particular basis for geometrical construction theory. The only thing that will be attempted is to provide a framework into which such a basis might be placed.

If we formalize a thing like constructability we of course dislike to do it in the style of classical logic. We do not want to consider constructability of a point as a proposition in the ordinary sense. We do not want to admit arguments where we get a contradiction from the assumption that the point P is not constructable, and then conclude the constructability of P. Therefore we want to put constructability (and the same thing might apply to observ-

ability and selectability) in a framework of positive logic, where we have no negation at all. In fact we can be even more restrictive, and refrain from introducing the ordinary logical connectives (like $\wedge$, $\vee$, $\rightarrow$) for this logic. The only thing we want to do is to register statements about constructability, observability and selectability (possibly provided with a number of parameters), and to keep them available for later use.

We can provide facilities for such a positive logic in *AUTOMATH* by adding a new expression of degree 1, to be called pprop (the first p stands for "positive"). For this pprop we shall not proclaim any logical axioms, and we shall not introduce the notion of negation. Moreover, we do not feel the need to have abstraction in the world of pprop. That is, if u : pprop we shall not take abstractors [x : u] like we would have in cases with prop or type. Accordingly, we shall not consider application <..> ... either in this pprop world. That means: we take pprop entirely in the style of *PAL* (see [1]).

There is a case for doing something similar in the world of type. Let us create a new expression of degree 1, to be called ctype (the 'c' stands for 'construction', since we intend to use it in the world of constructions). The difference between ctype and type is similar to the difference between pprop en prop. In ctype we intend to be free from all the assumptions that might have been made about type. In particular we shall not necessarily implement set-theoretical notions. And we shall not even introduce the notion of equality. That is, if a : C : ctype and b : C : ctype, then we will not introduce the equality of a and b as a proposition. Moreover, we shall treat ctype entirely in the style of *PAL*: no application and no abstraction.

For a description of *AUTOMATH* versions where various sets of rules apply to various expressions of degree 1, we refer to [2].

It has to be admitted that geometry is not the easiest example for the study of constructions. It is not so much the fact that the geometrical universes like planes, spaces, are uncountable. It is neither the most troublesome thing that in the geometrical plane there is no fixed origin and that there are no fixed directions. The real course of trouble is that there are so many situations where we have to except some of the cases. If we want to say that points p and q have just one connecting line we have to exclude the case p = q. Such things cause a steady flow of exceptions, which even has distorted the meaning of the word "arbitrary". In past centuries the word "arbitrary" often had the meaning: "arbitrary, but avoiding some obvious exceptions", and these exceptions where usually unspecified. If one took an arbitrary point and an arbitrary line then the point should not be so arbitrary to lie accidentally on the line!

A full description of all these exceptions has the tendency to make geometrical construction theory unattractive. Yet there is still another source of irritation: so often we have to split into cases (two circles may have 0, 1 or 2 points of intersection), and these situations might pile up to an entangled mess.

Nevertheless we may be grateful to geometry for having confronted us with the notion of constructability. What we have learned from geometry might be applied to other areas. Computer science might be one of them.

Observability, as a formal element in geometrical construction theory, was considered by D. Kijne [3]. That paper also attempts a formal treatment of selectability (with selection from finite sets only), and considers "giving arbitrary points" by means of a kind of algebraical adjunction operation.

3. THE BASIS OF FORMAL GEOMETRY

Before we discuss a formal basis for geometrical constructions we have to say what "formal geometry" or more generally, formal mathematics is. Here we are not concerned about the contents of formal geometry, but just about the spirit in which it is written. It is written in an *AUTOMATH* book, using the full power of typed lambda calculus. And it is written in a setting of logic and set theory, the details of which are still open to discussion. One might or might not take the rules of classical logic (e.g. in the form of the double negation law), and we might differ in taking or not taking a thing like the axiom of choice. Such distinctions hardly influence the spirit in which geometry is presented. They might influence the content, i.e. the set of all provable geometrical statements (but it should be remarked that there are areas of mathematics which are much more susceptible to foundational differences than classical geometry seems to be). Just to give an idea of the spirit, we give a small piece of Hilbert's axiomatization of geometry. Hilbert starts with: there are things we call points and there are things we call lines (in Hilbert's system the notion of a line is not presented as a special kind of point set). In *AUTOMATH* we say this by creating primitive types "line" and "point". These types are undefined, just introduced as primitive notions

(PN's). As a primitive we also have the notion "incidence" of a point and a line. Next we can express axioms like: if two points are different, then there is exactly one line incident to both points. Something should be said about "different". We take it that our geometry text is written in a mathematics book in which for any two objects a, b of type A there is a proposition that expresses equality of a and b, and that for any proposition we can form the negation. In this way the fact that a and b are different can be expressed in *AUTOMATH* as NOT(IS(A,a,b)). But in order to keep this paper readable we shall just write a $\neq$ b instead of this.

We now give a piece of *AUTOMATH* text that can be considered as the start of a Hilbert-style geometry book:

```
point := PN : type
line := PN : type

[ p : point
  [ m : line
    incident := PN : prop
  [ q : point
    [ ass1: p ≠ q
      conn := PN : line
      ax1 := PN : incident(p,conn)
      ax2 := PN : incident(q,conn)
      [ m : line
        [ ass2 : incident(p,m)
          [ ass3 : incident(q,m)
            ax3 := PN : m = conn
```

So if p,q are points, and m is a line, then incident(p,m) is a proposition; if pr is a proof of $p \neq q$ then conn(p,q,pr) is the connecting line of p and q. In axioms 1 and 2 we have expressed that this line is incident to p and q, in axiom 3 it is stated that if a line m is incident to both p and q then m is equal to the connecting line.

Although the above fragment is still a meagre piece of geometry it is hoped that it shows the spirit of a formalization. We shall refer to such a presentation of geometry as G.

4. A NAIVE APPROACH TO OBSERVABILITY

What we shall call the naive approach is expressed by the slogan "Truth is Observable". Let us explain what this means by mentioning two cases.

In the first case we use knowledge obtained from geometrical theory G in order to prove that some object we have to construct is already in our possession. We do not bother whether that proof is "constructive" or not: truth is just truth. One might find this a poor example, since within the scope of usual geometrical theories and usual constructions it seems that "non-constructive" proofs can always be replaced by very constructive ones, but it is easy to imagine fields where the situation is different.

In the second case we have a construction that started from a point that was chosen arbitrarily. At some stage of the construction we have a point P and a circle c, and subsequently our course of actions is depending on whether P lies inside c, outside c or on c. The naive point of view says that on the

basis of the theory in G we have exactly one of the three alternatives. We can observe wich one of the three occurs, and we act accordingly.

In Sections 6 and 7 we offer two different implementations for the naive point of view.

5. TAKING ARBITRARY OBJECTS

Before going on, we have to make it clear that there are two entirely different situations where in traditional geometry it was said that an arbitrary object (like a point) was taken. Let us call these situations D and S, (these letters abbreviate "data" and "selection"). If we think of a problem where a teacher requires a pupil to construct something, then D is the case where the data have been chosen arbitrarily by the teacher. On the other hand, S is the case where the pupil, in the course of the construction, selects some point arbitrarily. Quite often the final result does not depend on the particular point that was chosen, but there may be other cases. It may happen that the final result itself has a kind of arbitrariness. An example: given points A, B and C, not on a line, construct a point inside the triangle formed by A, B and C.

In the opinion of the pupil, the points taken in situation D are not called "arbitrary": they are called "given", or possibly "arbitrarily given". The pupil has no freedom in case D. In the S-case, however, the pupil is completely free, and the teacher has no say in the matter.

In a formal presentation like in *AUTOMATH* the difference between D and S is very pronounced. D is effectuated by means of the introduction of a new

variable, S is implemented by means of a primitive notion (PN). We shall show this in detail in Sections 6 and 7.

There is something about the PN-implementation of the S-situation that might be felt as strange. If we describe a construction by such a PN, then we select exactly the same point if we are requested to do the construction a second time. If the second time we would insist on selecting a point that is actually different from the one chosen the first time, then we have to do this on the basis of some new selection principle, of course. But if we just want to take a point again, without any restriction as to its being different from or equal to the first one, our PN provides us with the same point we had before. This means that we get more information than we intended to have. Nevertheless, such information cannot possibly do any harm.

What shall we do about this weirdness of the PN-implementation? Shall we invent impopular remedies in order to cure a completely harmless disease? Let us not prescribe a definite attitude in this, and admit that there are several ways to live with the situation. Either we leave the harmless disease for what it is, or we take one of the remedies. Let us mention two remedies. The first one is to take a notion of time t, and adhere a value of t to every construction step. The arbitrarily selected points will depend on t. If we have to repeat the construction some other day, t has a different value, so nothing is known about the selected point in comparison to the one selected the previous day.

As a second remedy we suggest to implement arbitrary selection not by an axiom but by some axiom scheme. The scheme proclaims the right to create as

many copies of the axiom as one might wish, each time with a different identifier.

We leave it at these scanty remarks. The author's opinion is that unless we invent a much simpler cure, we'd better learn to live with the harmless disease.

6. FIRST IMPLEMENTATION OF THE NAIVE POINT OF VIEW

We have to express in some way or other that some of our mathematical objects have been constructed. This can be thought of as a property of those objects, but for reasons sketched in Section 2 we prefer to take this property as a pprop rather than as a prop. We shall create, for every type X and for every x of type X, the expression have(X,x) with have(X,x): pprop. In particular we can abbreviate have(point,x) to havep(x) and have(line,x) to havel(x). (Since we use "have" for points and lines only, one might think of taking just "havep" and "havel" as primitives, without taking "have" for general types.)

We now give some *AUTOMATH* text. It is supposed to be added to a book that contains geometrical theory G (see Section 3) already. First we introduce "have", and abbreviations "havep" and "havel".

(We display our *AUTOMATH* texts in a flag-and-flagpole format: the block openers are written on flags, and the poles indicate their range of validity.)

```
X : type
    x : X
        have := PN : pprop

u : point
        havep := have(point,u) : pprop

v : line
        havel := have(line,v) : pprop
```

Next we display how we take an arbitrary object in the sense of the D-situation of Section 5 ("given objects"). In order to talk about a given point we need two block openers, expressing (i) that u is a point and (ii) that havep(u) holds; inside that context the point u can be considered as given. We shall now express: if u and v are given points and if $u \neq v$ then we can construct the line connecting u and v. According to our naive point of view the condition that u and v are different is simply expressed in the terminology of G.

```
u : point
    ass11 : havep(u)
        v : point
            ass12 : havep(v)
                ass13 : u ≠ v
                    ax11 := PN : havel(conn(u,v,ass13))
```

Next we describe a case of "taking arbitrary points" in the S-situation of Section 5. We express that if m is a given line then we are able to take a point not on m (we use the identifier "ap" to suggest "arbitrary point").

```
m : line
  ass14 : havel(m)
      ap := PN : point
      ax12 := PN : NOT(incident(ap,m))
      ax13 := PN : havep(ap)
```

These pieces of text display the form in which the basic constructions are introduced. If we want to describe a more complicated construction, we mention the relevant objects one by one, in the order of the construction, and each time we express that we "have" them. We give a (still very simple) example.

```
p : point
  ass14 : havep(p)
    q : point
      ass15 : havep(q)
        ass16 : p ≠ q
            L1  := conn(p,q,ass16)              : line
            H1  := ax11(p,ass14,q,ass15,ass16)  : havel(L1)
            P1  := ap(L1,H1)                    : point
            Ni1 := ax12(L1,H1)                  : NOT(incident(P1,L1))
            H2  := ax13(L1,H1)                  : havep(P1)
```

Here L1 is an abbreviation for the line connecting p and q; H1 can be used as a reference for the fact that we actually have that line. P1 is the result of the construction, Ni1 assures us that P1 does not lie on L1, and H2 assures us that we actually have P1. Altogether the text lines with identifier P1, Ni1, H1 represent the "derived construction" expressing that if p and q are given different points then we can take a point P1 such that p, q, P1 are not on one line. This derived construction can be applied later without referring to how it came about. It can be considered as a kind of "subroutine".

The example of a derived construction we gave here is ridiculously simple, of course. Yet the pattern is the same as in more complicated cases. It shows the old idea of subroutines, which existed in constructive geometry many centuries before it came up in computer programming.

7. SECOND IMPLEMENTATION OF THE NAIVE POINT OF VIEW

In the second implementation we take a construction plane which we conceive as being different from the geometrical plane. We might think of the original geometrical plane as abstract, and the constructive plane as concrete, consisting of a piece of paper we can draw on. But, of course, our construction plane is still abstract: it is a mathematical model of a concrete plane. The objects in the construction plane will be called cpoints and clines.

In the back of our mind we think of a one-to-one mapping between the two planes: every cpoint has a point as its companion, and every cline has a

line as its companion. Yet we shall not express all of this in our mathematical formalism. We shall just talk about a mapping (to be called semp) of cpoints to points and a mapping (to be called seml) of clines to lines. The reason for this reticence lies in the interpretation. If p1 is a point, and if we are able to name a cpoint cp1 that is mapped to p1 in our mapping, then for us this means that we "have" p1. We do not want to say that every point in the geometrical plane is a point we "have" just by being able to express that point mathematically. Therefore we do not want to be able to express the inverse mapping.

Related to this reticence is the fact that we do not want to be able to discuss equality of two cpoints. Such equality has to be discussed for the companion points in the geometrical plane. And we do not want to admit as mathematical objects things like "the set of all cpoints" with some prescribed property. We achieve these restrictions by putting "cpoint" and "cline" into ctype, which is a world without equality, without set theory, without quantification. As a consequence we do not have constructability questions in our theory. A statement: "the point P is not constructable with ruler and compass" will not be a proposition in our *AUTOMATH* book. If we would be able to quantify over the construction plane we would be able to express that "there is no cpoint that is mapped onto P" and that would express the non-existence of the construction. Constructability questions belong to the meta-theory. They express that something "cannot be obtained on the basis of the PN's displayed thus far", and we cannot say such things in *AUTOMATH* itself.

What we call our second implementation starts with the introduction of cpoint, cline and the mappings semp and semf. The latter abbreviations suggest the word "semantics": we might say that the geometrical plane forms the semantics of the construction plane. If P is a cpoint then semp(P) is its semantics. Off we go:

cpoint := PN : ctype

cline := PN : ctype

```
cp : point
    semp := PN : point
```

```
cl : cline
    seml := PN : line
```

In order to take an arbitrary point in the construction plane, a single block opener "x:cpoint" plays the role of the pair "u:point", "ass 11:havep(u)" of the first implementation. We show this with the fundamental construction that connects two points:

```
x : cpoint
  y : cpoint
    ass21 : semp(x) ≠ semp(y)
      cconn := PN : cline
      ax21 := PN : seml(cconn) = conn(semp(x),semp(y),ass21)
```

The fact that cconn is the line we are looking for, is expressed (in ax21) by means of equality in G.

If we have to take an arbitrary point in the S-situation we again get one PN less than in the corresponding case of Section 6. In order to express that we can take a point outside a line, we write

```
cm : cline
    acp  := PN : cpoint
    ax22 := PN : NOT(incident(semp(acp),seml(cm)))
```

We also show the text corresponding to the one with P1, Ni1, H2 in Section 6:

```
p : cpoint
  q : cpoint
    ass 22 : semp(p) ≠ semp(q)
      CL1 := cconn(p,q,ass22) : cline
      CP1 := acp(CL1)         : cpoint
      Ni2 := ax22(CL1)        : NOT(incident(semp(CP1),seml(CL1)))
      Ni3 :=   ...            : NOT(incident(semp(CP1),
                                    conn(semp(p),semp(q),ass22)))
```

We have not displayed the proof Ni3. It will depend on applying general axioms about equality, and will make use of Ni2 and ax21.

Passages like the one from Ni2 to Ni3 might be superfluous in many cases, since it is practical to keep the discussion in the construction plane as long as possible. To that end we might copy notions from G to the construction plane. The simplest example is

```
x : cpoint
  y : cline
    cincident := incident(semp(x),seml(y))  :  prop
```

8. RESTRICTED OBSERVABILITY

In Sections 4, 6, 7 we described the naive point of view, where every truth in the geometrical theory is considered to be "observable". Observability has its meaning in the process of taking decisions about the course of our constructions.

Let us describe two different motives for restricting observability. One is practical, the other one is fundamentalistic. We shall discuss these in Sections 9 and 10, respectively.

9. PRACTICAL RESTRICTIONS ON OBSERVABILITY

The practical point of view is connected to questions of precision. This can be compared to the matter of rounding off errors in numerical analysis. If in a construction two points turn out to be so close together that our construction precision does not guarantee that they are different, then we can

not claim to be able to connect them by a line. And even if the points are different, the line will be ill-defined.

Although these practical matters give rise to quite complicated considerations, we cannot say that they are necessarily essentially different from what we did in Sections 6 and 7. Once can still go on the basis that truth is observable: the question is just a matter of which propositions we consider the truth of. Instead of claiming the possibility to connect two points p, q if $p \neq q$ in the geometrical world G, we take a thing like $d(p,q) > 1$ (distance exceeds unity) as our criterion.

Nevertheless we can make things a little livelier than this. Let us start from what we developed in the beginning of Section 7: just the four PN's that were called cpoint, cline, semp and seml. We now introduce a primitive notion "obsdif" ("observationally different") in the construction plane:

```
p : cpoint
    q : cpoint
        obsdif := PN : prop
```

And now instead of introducing the cconn, ax21, etc. of Section 7, we go on like this:

```
x : cpoint
    y : cpoint
        ass31 : obsdif(x,y)
            cconn1 := PN : cline
            ax31   := PN : seml(cconn1) = conn(semp(x),semp(y),ass31)
```

Knowledge about obsdif can come from different sources. In the first place we can axiomatize things like: if d(semp(x), semp(y)) > 1 then x and y are observationally different. A second source arises if we axiomatize in the construction plane, in some situations, that if cpoints u and v are observationally different, then the cpoints x and y, derived from u and v in one way or other, are observationally different. A very simple case of this is an axiom stating that obsdif(x,y) implies obsdif(y,x).

It will be clear that this subject will become very complicated without being very rewarding. Therefore it seems definitely unattractive.

10. FUNDAMENTALISTIC RESTRICTIONS ON OBSERVABILITY

In Section 9 we still had the uncritical acceptance of all truth that can be obtained in the geometrical world. There is a clear reason for restriction. If we have to use geometrical propositions for taking decisions in the world of constructions, it is reasonable to require that we also have a "constructive" way for actually deciding whether such propositions hold or do not hold.

We can implement such restrictions in *AUTOMATH* by selecting some "constructive" basis for logic and mathematics, like intuitionistic mathematics, and building our geometry G according to these principles. We might even mix a constructive kind of mathematics with the ordinary kind, using pprop and ctype for the constructive kind. In particular it seems to be reasonable to take the "obsdif" we had in Section 9 as a pprop rather than as a prop.

The latter remark suggests that it might be simpler to shift life entirely to the constructive plane, and to forget G altogether. But this is not what we usually want. Let us imagine that we want to describe the theory of Mascheroni constructions (constructions with compass but without ruler). The subject matter concerns both circles and straight lines, the constructions deal with circles only. This difference can be implemented by discussing both circles and straight lines in G, but just "cpoints" and "ccircles" in the construction plane.

11. COMPARISON WITH COMPUTER PROGRAM SEMANTICS

It is very natural to compare the field of geometrical constructions with the one of computer programming. In both cases there is a number of actions that produce one or more objects, and in both cases it is very essential that it is proved that these objects satisfy the problem specification that was given beforehand.

In a computer program we usually think of a "state space"; the input is an element of that state space and the output is again such an element. In the case of geometrical constructions one would say that the input is (vaguely speaking) the given figure, and the output is the required figure. Let us admit different spaces for input space and output space, and try to describe at least the specification of a geometrical construction in terms of input and output. As an example we take the following (trivial) construction problem. Given two different points P, Q and a line m. Construct a line q that intersects m, passes through P but not through Q.

Let us talk in the style of Section 7, and let us moreover decide to introduce a name R for a cpoint of intersection of q and m (otherwise we would need existential quantification). An element of the input space is a triple (P,Q,m) where P : cpoint, Q : cpoint, m : cline, and where we have semp(P) ≠ semp(Q). An element of the output space is a pair (q,R) where q : cline and R : cpoint. The problem specification is given by the conditions that seml(q) is incident with semp(P) and semp(R) but not with semp(Q).

This kind of problem specification is entirely in the style of what is called "relational semantics" in computing science.

If we deal with geometrical constructions, the role of "subroutines" is more or less the same as in computer programming. In particular we can say that descriptive geometry consists of a large body of subroutines.

In computer programs we can have loops. Sometimes pieces of a program have to be repeated until some condition is satisfied. The geometrical constructions we discussed in the previous sections have no such loops. This shows an essential restriction on the class of constructions we can describe in the various systems that were suggested in these sections. An example of a different nature is the following one. Let A, B, C be given points on a given line, B between A and C. It is required to construct a point D on that line, such that C is between B and D, and such that the length of the line segment BD is an integral multiple of the length of the segment AB. This construction requires a loop.

Our treatment of geometrical constructions in Sections 3-10 might be called "operational" or anyway "functional". All the time uniquely determined out-

puts are obtained step by step, and in the slightly more sophisticated case of the use of subroutines the only thing we actually do is taking sequences of steps together and considering them as a single step. The reason is that the treatment is based on what we shall call the *interior* approach. In the interior approach we talk in terms of the constructed objects. The constructed objects are treated in the same style as ordinary mathematical objects and (but this is a typical *AUTOMATH* feature) proofs. In our *AUTOMATH* book we discuss the objects, but the action of construction is felt as subject matter of some metalanguage.

An entirely different way to deal with constructions is that we consider constructions as objects, seemingly more abstract than the ordinary objects, but nevertheless on the same linguistic level. Let us call this the *exterior* approach. (The name is suggested by the fact that if we work in the interior approach then the metalinguistic discussion of construction is felt as being something at the outside).

With the exterior approach we can get rid of the limitations of our "functional style" of construction description. Anyway we can remove the last differences there might be between geometrical construction and computer programming.

We might try to start the exterior treatment with the introduction of a primitive notion "construction", like:

construction := PN : ctype

but it has to be more complicated than this. The notion of construction has to depend on the input space and the output space as parameters, and this is not so easy to describe.

REFERENCES

[1] de Bruijn, N.G., A survey of the project *AUTOMATH*. To H.B. Curry: Essays in combinatory logic, lambda calculus and formalism (editors: J.P. Seldin and J.R. Hindley) pp. 579-606. Academic Press, 1980.

[2] de Bruijn, N.G., A framework for the description of a number of members of the *AUTOMATH* family. Internal report, Department of Mathematics, T.H. Eindhoven, June 1974.

[3] Kijne, D., Construction geometries and construction fields. In: Algebraical and Topological Foundations of Geometry. Proceedings of a Colloquium held in Utrecht, August 1959. Pergamon Press, 1962.

OVER VERDELINGEN VAN GETALLEN IN GROEPJES

door

F.C. Bussemaker

Opgedragen aan Prof. J.J. Seidel t.g.v. zijn afscheid van de T.H. Eindhoven.

SAMENVATTING. Er wordt een "moeilijk" probleem aangekaart en een vereenvoudigde versie hiervan nader bekeken. Een met het laatste probleem in verband staande vraag wordt opgelost.

Via J.B. Dijkstra en L.S. de Jong werd ik eens geconfronteerd met het volgende probleem [1].

PROBLEEM 1: Gegeven zijn een eindig aantal reële getallen $a_1,\ldots,a_n$ en een reëel getal g zodat:

(1) $a_i > 0$ voor $i = 1,\ldots,n$,

(2) $0 < g \leq \sum_{i=1}^{n} a_i$.

Hiermee wordt een aantal reële getallen $b_1,\ldots,b_m$ gevormd zodat:

(3) $b_j \geq g$ voor $j = 1,\ldots,m$,

(4) elke b_j een a_i is of de som van enkele a_i's,

(5) elke a_i precies één keer wordt gebruikt.

Bewijs dat geldt:

$$\sum_{j=1}^{m} b_j^2 \text{ minimaal} \Rightarrow m \text{ maximaal} .$$

Dit probleem schijnt op het eerste gezicht tamelijk eenvoudig te zijn, maar schijn bedriegt!

Het gaat om het vinden van een zodanige indeling van de getallen a_i in groepjes met partiële sommen b_j, dat deze partiële sommen zo weinig mogelijk van elkaar afwijken en zo dicht mogelijk bij g liggen. Al gauw komt de gedachte op, dat het misschien mogelijk is, het probleem te beperken tot getallen $a_i < g$. Immers, bevat de verzameling getallen a_i ook nog getallen groter dan g, dan lijkt het minimaliseren van de kwadratensom gediend te zijn met het plaatsen van zo'n getal in een apart groepje.

Jammer genoeg hoeft daarmee de kwadratensom nog niet geminimaliseerd te worden, zoals het volgende eenvoudige voorbeeld laat zien:

$$n = 4,\ a_1 = 1,\ a_2 = a_3 = a_4 = 0.9\ ,\ g = 1;$$

en

$$(1)^2 + (3 * 0.9)^2 > (1 + 0.9)^2 + (2 * 0.9)^2\ .$$

Indien we dus de restrictie $a_i < g$ aanbrengen, krijgen we wel een wezenlijk "smaller" probleem, dat echter misschien ook gemakkelijker is op te lossen. Daarentegen verandert het probleem niet essentieel als we alle getallen a_i en b_j "schalen" door te delen door g.

In plaats van het oorspronkelijke probleem stellen we daarom het volgende probleem.

PROBLEEM 1': Gegeven is een verzameling V van n ($n \geq 2$) reële getallen $a_1,\ldots,a_n$ zodat:

(1) $0 < a_i < 1$ voor $i = 1,\ldots,n$.

(2) $\sum_{i=1}^{n} a_i \geq 1.$

Voor m = 1,2,... , wordt de verzameling V nu op alle mogelijke manieren gepartitioneerd in subverzamelingen V_j, waarvoor geldt:

(3) $b_j := \sum_{a_i \, V_j} a_i \geq 1, \quad j = 1,\ldots,m.$

Bewijs dan dat geldt:

$\sum_{j=1}^{m} b_j^2$ minimaal $\Rightarrow$ m maximaal.

Mocht er een bewijs voor probleem 1' gevonden worden, dan is daarmee probleem 1 nog niet algemeen bewezen. Een tegenvoorbeeld voor de bewering van probleem 1' geldt echter tevens als tegenvoorbeeld voor probleem 1. Ook dit vereenvoudigde probleem blijft heel lastig.

Voor kleinere waarden van n is het probleem nog vrij gemakkelijk te bewijzen. Neem bijvoorbeeld het geval n = 8. De 8 getallen a_i kunnen dan over ten hoogste 4 hoopjes verdeeld worden. Voor $s := \sum_{i=1}^{\infty} a_i$ geldt: $1 \leq s < 8$; we kunnen ons echter rustig beperken tot: $2 \leq s < 8$.

We voeren nu eerst wat nieuwe notaties in.

Zij V een verzameling van n reële getallen a_i met $0 < a_i < 1$, zodat $\sum_{i=1}^{n} a_i \geq 1$. Beschouw nu alle mogelijke partities van V in subverzamelingen V_j met partiële sommen $b_j \geq 1$. Bepaal bij elke partitie van V in m van deze subverzamelingen de som van de kwadraten $b_1^2 + b_2^2 + \ldots + b_m^2$.

Als dit voor alle zinvolle waarden van m gedaan wordt, krijgt men uiteindelijk één kwadratensom voor m = 1, mogelijk meerdere kwadratensommen voor m = 2, etc. De verzameling van de uitkomsten van al die mogelijke kwadratensommen bij een vaste waarde van m, waarvoor inderdaad partities bestaan, die aan alle

eisen voldoen, noemen we hier KS_m. We kunnen dan dus spreken over KS_1, KS_2, etc.

Nog één notatiekwestie: indien V_1 en V_2 twee eindige verzamelingen van reële getallen zijn, dan wordt verder met $V_1 \leq V_2$ bedoeld, dat elk getal in V_1 niet groter is dan elk getal in V_2. En als g een reëel getal is, dan wordt bijvoorbeeld met $V_1 > g$ bedoeld, dat elk getal in V_1 groter dan g is.

We gaan nu bewijzen dat, gegeven één bepaalde verzameling V van 8 getallen, altijd $KS_1 > KS_2 \geq KS_3 \geq KS_4$ (voor zover van toepassing).
Als we alle getallen op één hoop gooien, dan is $KS_1 = s^2$. Is het mogelijk de getalllen over 2 hoopjes te verdelen, dan is $KS_2 \geq 2(\frac{s}{2})^2 = \frac{s^2}{2}$. Aan de andere kant is $KS_2 \leq (s-1)^2 + 1^2 = s^2 - 2s + 2$.
Als het mogelijk is de getallen over 3 hoopjes te verdelen, dan is

$$KS_3 \geq 3(\frac{s}{3})^2 = \frac{s^2}{3},$$

terwijl ook geldt:

$$KS_3 \leq (s-2)^2 + 1^2 + 1^2 = s^2 - 4s + 6.$$

Als tenslotte een verdeling van de getallen over 4 hoopjes mogelijk is, dan is (grof geschat):

$$KS_4 \leq (s-3)^2 + 3 \cdot 1^2 = s^2 - 6s + 12.$$

Nu is voor $s \geq 2$ altijd $KS_1 > KS_2$, voor $3 \leq s \leq 6$ is $KS_2 \geq KS_3$ en voor $4 \leq s \leq 6$ is $KS_3 \geq KS_4$, zoals gemakkelijk is na te gaan.

We moeten dan nog voor $6 < s < 8$ bewijzen dat $KS_2 \geq KS_3 \geq KS_4$. Daartoe bekijken

we de verdeling van de getallen een beetje nauwkeuriger, te beginnen met de verdeling over 3 hoopjes. Elk hoopje dient steeds tenminste 2 getallen te bevatten. Dus de volgende partities van 8 getallen over 3 hoopjes getallen zijn mogelijk: $8 = 4 + 2 + 2$ en $8 = 3 + 3 + 2$. Voor de partiële som van elk hoopje met k getallen geldt: $1 \leq s_j < k$. Het is dan eenvoudig na te gaan dat een verdeling van de getallen in hoopjes met partiële sommen $s_1 = 4 - \delta$, $s_2 = s - 5 + \delta$ em $s_3 = 1$ (met δ een zo klein mogelijk positief reëel getal) leidt tot een zo groot mogelijke kwadratensom KS_3 voor $6 < s < 8$. We vinden dan: $KS_3 < s^2 - 10s + 42 < \frac{s^2}{2} \leq KS_2$ voor $6 < s < 8$.

Een verdeling van de 8 getallen over 4 hoopjes kan alleen als elk hoopje precies 2 getallen bevat. Voor de partiële som s_j van elk hoopje geldt dan: $1 \leq s_j < 2$.

Beschouw nu eerst het geval $6 < s \leq 7$. Dan krijgen we een zo groot mogelijke KS_4 met de verdeling met partiële sommen $\{(2-\delta_1), (2-\delta_2), (s-5+\delta_1+\delta_2), 1\}$ waarin δ_1 en δ_2 zo klein mogelijke positieve reële getallen zijn. Dan is $KS_4 < s^2 - 10s + 34$. En dit is kleiner dan $\frac{s^2}{3}$ voor $6 < s \leq 7$.

Tenslotte krijgen we voor $7 < s < 8$ een zo groot mogelijke KS_4 als we de verdeling van de getallen met de volgende partiële sommen hebben:
$\{(2-\delta_1), (2-\delta_2), (2-\delta_3), (s-6+\delta_1+\delta_2+\delta_3)\}$ met δ_1, δ_2 en δ_3 zo klein mogelijke positieve reële getallen. En $(2-\delta_1)^2 + (2-\delta_2)^2 + (2-\delta_3)^2 + (s-6+\delta_1+\delta_2+\delta_3)^2 < s^2 - 12s + 48$. En dit is kleiner dan $\frac{s^2}{3}$ voor $7 < s < 8$.
Waarmee dus voor alle waarden van s met $2 \leq s < 8$ bewezen is dat inderdaad $KS_1 > KS_2 \geq KS_3 \geq KS_4$.

Verder is het niet moeilijk om beweringen als bijvoorbeeld de volgende te bewijzen: Laat voor de som s van de n getallen a_i gelden dat $s = m + \delta$ met

$0 \leq \delta < 1$, en laat een partitie van de n getallen in m subverzamelingen mogelijk zijn.Dan behoort de minimale kwadratensom tot KS_m en m is maximaal. Dat m maximaal is, is triviaal. En $KS_m < (m-1) \cdot 1^2 + (1+\delta)^2 = m + 2\delta + \delta^2$. Verder zijn KS_{m-1}, KS_{m-2}, etc., allen tenminste gelijk aan $\frac{s^2}{m-1}$. Tenslotte is voor $0 \leq \delta < 1$ inderdaad $m + 2\delta + \delta^2 < (m+\delta)^2/(m-1)$.

Een bewijs vinden voor het algemene geval is echter veel moeilijker. Het is niet mijn bedoeling om dit hier te proberen. Ik wilde nu echter wel een ander probleem stellen (en oplossen) dat met het vorige probleem in verband staat.

PROBLEEM 2: Laat 2 natuurlijke getallen m en n gegeven zijn: $m \geq 1$, $n \geq 2m$. Beschouw een willekeurige verzameling van n reële getallen a_i met $0 < a_i < 1$ voor $i = 1,\ldots,n$. Laat s de som van deze getallen zijn: $s = \sum_{i=1}^{n} a_i$.
De n getallen worden nu verdeeld over m hoopjes van 2 of meer getallen elk. Geëist wordt, dat de som s_j van de getallen in elk hoopje ≥ 1 is $(j = 1,\ldots,m)$.
Intuïtief voelt men aan, dat als s maar groot genoeg is, altijd wel aan deze eis is te voldoen.
De vraag is nu: wat is de *minimale* grootte p(m,n) van s, opdat bij de gegeven waarden van m en n *altijd* (dus onafhankelijk van van de getallen a_i, $i= 1,\ldots,n$) een dergelijke verdeling in hoopjes met som $s_j \geq 1$ $(j = 1,\ldots,m)$ mogelijk is.

De bewering is nu, dat $p(m,n) = 2m - 1$. Dus de grens p(m,n) blijkt onafhankelijk van het aantal getallen a_i te zijn (in zekere zin). Dit is toch een tamelijk

onverwacht resultaat.

Het bewijs van deze bewering is simpel en valt uiteen in 2 stappen. We nemen hierbij aan dat $m \geq 2$, want voor $m = 1$ is de bewering triviaal.

i) De eerste stap is het bewijs van het geval $n = 2m$. In dit geval moeten we zelfs een sterkere bewering bewijzen: er is altijd een verdeling van de $2m$ getallen in m hoopjes van elk twee getallen met som ≥ 1 mogelijk als $s \geq 2m - 1$.

Immers, verdeel de getallen maar willekeurig in m paren. Kies er één paar uit. De som van de $2(m-1)$ getallen in de $(m-1)$ overige paren moet kleiner dan $2m-2$ zijn. Maar omdat de som van alle getallen tenminste $2m-1$ is, moet de som van de 2 getallen van het gekozen paar wel groter dan 1 zijn. Dat $p(m,2m) = 2m-1$ inderdaad minimaal is, kunnen we als volgt aantonen met een voorbeeld. Laat δ een positief reëel getal $< \frac{1}{2}$ zijn, en kies de getallen a_i als volgt:

$$\delta, \underbrace{1-2\delta,\ 1-2\delta,\ldots,1-2\delta}_{(2m-1)\text{ keer}}.$$

Er geldt: $\sum_{i=1}^{2m} a_i = (2m-1) - (4m-3)\delta$. We kunnen deze som zo dicht bij $(2m-1)$ kiezen als we willen door δ maar klein genoeg te nemen.

Hoe men deze $2m$ getallen ook in hoopjes verdeelt, er zal altijd één hoopje getallen bij zijn met som $s_j < 1$. Want het hoopje getallen waarin het getal δ terecht komt moet ook nog tenminste 2 getallen $(1-2\delta)$ bevatten, wil de som $s_j \geq 1$ worden. Dus de overige getallen, in aantal ten hoogste gelijk aan $2m-3$, moeten verdeeld worden over $m-1$ hoopjes, en dan zal tenminste één hoopje maar één getal $(1-2\delta)$ bevatten.

ii) Nu de tweede stap van het bewijs: er zijn $n > 2m$ getallen met totale som $s \geq 2m-1$.

Orden de getallen a_i in opklimmende volgorde, zodat (na eventuele hernummering) $0 < a_1 \leq a_2 \leq \ldots \leq a_{n-1} \leq a_n < 1$. We gaan nu deze getallen als volgt in groepjes indelen. Om het eerste groepje te bepalen, nemen we a_n en voegen daarbij a_1. Als blijkt dat $a_n + a_1 \geq 1$ is, laten we alleen a_n tot dit eerste groepje behoren. Is daarentegen $a_n + a_1 < 1$, dan voegen we achtereenvolgens de getallen $a_2, a_3, \ldots, a_p$ bij dit groepje totdat geldt: $a_n + a_1 + a_2 + \ldots + a_p < 1$ maar $a_n + a_1 + a_2 + \ldots + a_p + a_{p+1} \geq 1$.

Vervolgens gaan we uit de overblijvende getallen een tweede groepje samenstellen. We voegen daartoe bij het nu overgebleven grootste getal a_{n-1} achtereenvolgens de getallen $a_{p+1}, a_{p+2}, \ldots, a_q$, totdat geldt:

$a_{n-1} + a_{p+1} + a_{p+2} + \ldots + a_q < 1$ maar $a_{n-1} + a_{p+1} + a_{p+2} + \ldots + a_q + a_{q+1} \geq 1$.

Daarna gaan we derde groepje samenstellen, etc. Zo voortgaande kunnen we tenminste $2m-1$ van dergelijke groepjes met partiële som < 1 construeren. Immers, stopte dit constructieproces reeds bij een kleiner aantal, dan zou de totale som van de getallen a_i ook kleiner dan $2m-1$ zijn.

We beschouwen nu de eerste $2m-2$ groepjes, en voegen alle overige getallen bij elkaar in een restgoep. Hoe men uit de eerste $(2m-2)$ groepjes ook paren van groepjes samenstelt, de getallen van de beide groepjes bij elkaar gevoegd hebben een som > 1. En de som van de getallen in de restgroep is ook groter dan $(2m-1) - (2m-2) = 1$.

In totaal krijgen we dus op deze manier een indeling van de n getallen in m hoopjes met som > 1.

Rest ons nu nog slechts om te bewijzen dat $p(m,n) = 2m - 1$ inderdaad minimaal is. Dit tonen we aan met behulp van het volgende voorbeeld (waarvan het hiervoor gegeven voorbeeld feitelijk een bijzonder geval is).

Kies de getallen als volgt:

$$\underbrace{\delta, \delta, \ldots, \delta}_{(2-2m+1) \text{ keer}}, \underbrace{1 - (n-2m+2)\delta,\ 1 - (n-2m+2)\delta, \ldots,\ 1 - (n-2m+2)\delta}_{(2m-1) \text{ keer}},$$

waarin δ een positief reëel getal $< 1/(n-2m+2)$ is.

Het is dan gemakkelijk in te zien dat het weer niet mogelijk is om de n getallen in m hoopjes te verdelen met som ≥ 1 per hoopje.

DANKBETUIGING.

Ik dank Henk van Tilborg voor zijn waardevolle opmerkingen.

REFERENTIE

[1] Dijkstra, J.B., opgave 49, Statistica Neerlandica 2 (1977), 92.

GAP SERIES AND ALGEBRAIC INDEPENDENCE

by

P.L. Cijsouw

Dedicated to J.J. Seidel on the occasion of his retirement.

To my opinion, one of the most important concepts in mathematics is that of approximation. Approximation methods and techniques are used in almost every field of mathematics; sometimes in a very direct sense, but in most cases hidden or translated into another form. In transcendence theory, all results are obtained by the use of approximations in one way or the other. In particular, all proofs of irrationality (or transcencence) of specific numbers can be formulated such, that in the proof the knowledge is used that a certain number or function does not admit "very close" approximations of a certain kind, unless we have irrationality (or transcendence). At the same time, a construction is given of a sequence of "too close" approximations. I would like to mention two classical examples.

1. The irrationality of $e = \sum_{i=0}^{\infty} 1/i!$ follows from the observation that, in case $e = p/q$ with $p,q \in \mathbb{N}$, the number $\sum_{i=0}^{k} 1/i!$ should approximate e up to a distance $p/q - \sum_{i=0}^{k} 1/i!$; this distance clearly is a positive integral multiple of $1/k!$ when k is large enough. On the other hand, the distance is equal to $\sum_{i=k+1}^{\infty} 1/i!$, and this number is smaller than $1/k!$

when k is large. Hence, e is not rational.

2. A well-known theorem of LIOUVILLE states, that for every algebraic number θ of degree $s \geq 2$, there exists a number $c > 0$ such that for all integers p,q $(q > 0)$ we have

$$|\theta - p/q| \geq c.q^{-s} .$$

Let us apply this to $\sigma = \sum_{i=0}^{\infty} 2^{-i!}$; we then obtain that if σ were algebraic, then

$$\sum_{i=k+1}^{\infty} 2^{-i!} \geq c.2^{-s.k!} .$$

Since the latter sum is of the order of $2^{-(k+1)!}$, this inequality does not hold when k is large enough; consequently, θ is transcendental.

Note that the second example can be considered as to concern the value of a gap series $\sum_{i=0}^{\infty} a_i z^{e_i}$ with algebraic coefficients $a_i \neq 0$ and rapidly increasing exponents e_i, taken at an algebraic point within the disc of convergence of this series.

The transcendence of this kind of gap series at such points has been studied by H. COHN [3], K. MAHLER [6] and G. BARON and E BRAUNE [1], all in different special cases. Improving some of these results, P.L. CIJSOUW and R. TIJDEMAN [4] obtained a theorem that covered all earlier results. In order to formulate this theorem, we introduce some conventions and notations that will be used throughout the sequel of this paper.

The conjugates of an algebraic number α of degree s are denoted by $\alpha^{(1)} = \alpha, \alpha^{(2)}, \ldots, \alpha^{(s)}$; further, $\overline{|\alpha|} = \max_{i=1,\ldots,s} |\alpha^{(i)}|$. The denominator of α

(i.e., the smallest positive integer k such that $k\alpha$ is an algebraic integer) is denoted by $m(\alpha)$. In connection to the sequence $a_0,a_1,a_2,\ldots$ of algebraic coefficients of our gap series, we use for $k = 0,1,2,\ldots$

$$A_k = \max_{i=0,\ldots,k} |\overline{a_i}| \text{ and } S_k = [\mathbb{Q}(a_0,a_1,\ldots,a_k): \mathbb{Q}];$$

M_k will be the least common multiple of $m(a_0),m(a_1),\ldots,m(a_k)$. The sequence $(e_i)_{i=1}^{\infty}$ will be an increasing sequence of integers with $e_0 \geq 0$. We shall assume that the radius of convergence R of the power series $\sum_{i=0}^{\infty} a_i z^{e_i}$ is positive.

The mentioned theorem of CIJSOUW and TIJDEMAN from [4] can now be stated as follows.

> Suppose $\lim_{k\to\infty} (e_k + \log M_k + \log A_k)S_k/e_{k+1} = 0$. Then $\sigma(\theta) = \sum_{i=0}^{\infty} a_i\theta^{e_i}$ is transcendental for every algebraic θ with $0 < |\theta| < R$.

The proof still can be formulated as indicated above: one can show that two different algebraic numbers of a bounded complexity cannot lie very close together; the assumption that $\sigma(\theta)$ is algebraic thus leads to a certain minimum for the distance between $\sigma(\theta)$ and $\sigma_k(\theta) = \sum_{i=0}^{k} a_i\theta^{e_i}$; this is violated by the smallness of $\sigma(\theta) - \sigma_k(\theta) = \sum_{i=k+1}^{\infty} a_i\theta^{e_i}$ when k is large enough.

We are ready now to turn our attention to algebraic independence. Remember that the transcendental numbers $\sigma_1,\ldots,\sigma_p$ are said to be algebraically independent when there exists no polynomial $P \in \mathbb{Z}[z_1,\ldots,z_p]$, $P \neq 0$, with the property that $P(\sigma_1,\ldots,\sigma_p) = 0$. For very special situations, it has been known for a long time that transcendental numbers that admit "essentially

different approximations" by algebraic numbers must be algebraically independent. A rather precise description of this feature has been given by A. DURAND [5]. I formulate here a somewhat modified version of his "Corollaire"

Let $\sigma_1,\ldots,\sigma_p$ be complex numbers and assume that sequences $(\sigma_{1,n})_{n=1}^{\infty},\ldots,(\sigma_{p,n})_{n=1}^{\infty}$ of algebraic numbers exist with the following properties:

(i) $\quad 0 < |\sigma_{j+1} - \sigma_{j+1,n}| \le \frac{1}{n}|\sigma_j - \sigma_{j,n}|$ for $j = 1,\ldots,p-1$

(ii) $\quad 0 < |\sigma_1 - \sigma_{1,n}| \le \prod_{j=1}^{p}\left\{m(\sigma_{j,n})(1 + |\overline{\sigma_{j,n}}|)\right\}^{-nD_n}$

where $D_n = [\mathbb{Q}(\sigma_{1,n},\ldots,\sigma_{p,n}):\mathbb{Q}]$.

Then the numbers $\sigma_1,\ldots,\sigma_p$ are algebraically independent.

The modification consists of the replacement of DURAND's $A(\sigma_{j,n})$ by $\{m(\sigma_{j,n})(1 + |\overline{\sigma_{j,n}}|)\}^s$ with $s = [\mathbb{Q}(\sigma_{j,n}):\mathbb{Q}]$; this is easily justified by Lemma 1 of [4] and its proof.

Inequality (ii) guarantees the transcendence of σ_1 and (together with (i)) that of $\sigma_2,\ldots,\sigma_p$; inequality (i) says that the approximations to $\sigma_1,\ldots,\sigma_p$ are of sufficiently different orders, thus ensuring the algebraic independence of $\sigma_1,\ldots,\sigma_p$.

DURAND's criterion for algebraic independence has been used within our context by P. BUNDSCHUH and F.J. WYLEGALA [2] for the proof of the following theorem, in which $\sigma(z)$ is a gap series with algebraic coefficients as before.

Suppose $\lim_{k\to\infty}(e_k + \log M_k + \log A_k)\, S_k/e_{k+1} = 0$.
Then for every tuple $\theta_1,\ldots,\theta_\ell$ of algebraic numbers with $0 < |\theta_1| < \ldots < |\theta_\ell| < R$ the values $\sigma(\theta_1),\ldots,\sigma(\theta_\ell)$ are algebraically independent.

Note that the condition is exactly the same as that of the theorem of CIJSOUW-TIJDEMAN, so that the latter can be seen as the case $\ell = 1$ of the theorem of BUNDSCHUH-WYLEGALA.

In what follows, I shall show that an extension to algebraic independence of values of such a gap series and its derivatives at points that are algebraic and of different absolute values can be proved as well.

THEOREM. Let $\sigma(z) = \sum_{i=0}^{\infty} a_i z^{e_i}$ be a gap series with algebraic coefficients and with a positive radius of convergence R. Suppose

$$\lim_{k\to\infty} (e_k + \log M_k + \log A_k) S_k/e_{k+1} = 0 .$$

Then for every tuple $\theta_1,\ldots,\theta_\ell$ of algebraic numbers with $0 < |\theta_1| < \ldots < |\theta_\ell| < R$ and for every $r \in \mathbb{Z}$, $r \geq 0$, the numbers $\sigma^{(\rho)}(\theta_\lambda)$ $(\lambda = 1,\ldots,\ell;\ \rho = 0,1,\ldots,r)$ are algebraically independent.

Proof. We shall apply DURAND's criterion to the $\ell(r+1)$ complex numbers $\sigma^{(r)}(\theta_\ell),\sigma^{(r-1)}(\theta_\ell),\ldots,\sigma(\theta_\ell),\ldots,\sigma^{(r)}(\theta_1),\sigma^{(r-1)}(\theta_1),\ldots,\sigma(\theta_1)$ in this order, using the approximating algebraic numbers $\sigma^{r}_{\ell,k},\sigma^{r-1}_{\ell,k},\ldots,\sigma^{o}_{\ell,k},\ldots,\sigma^{r}_{1,k}$, $\sigma^{r-1}_{1,k},\ldots,\sigma^{o}_{1,k}$ where

$$\sigma^{\rho}_{\lambda,k} = \sum_{i=0}^{k} a_i e_i (e_i - 1)\ldots(e_i - \rho + 1)\theta_\lambda^{e_i-\rho} .$$

In fact, the sequences of approximating algebraic numbers from DURAND's criterion will be a system of subsequences $(\sigma^{\rho}_{\lambda,k(n)})_{n=1}^{\infty}$ of $(\sigma^{\rho}_{\lambda,k})_{k=1}^{\infty}$. It is clear, that $|\sigma^{(\rho)}(\theta_\lambda) - \sigma^{\rho}_{\lambda,k}|$ is of the same order as $|a_{k+1}|^{e_{k+1}(e_{k+1}-1)\dots(e_{k+1}-\rho+1)}|\theta_\lambda|^{e_{k+1}-\rho}$, i.e., that the quotient of these expressions is between 1/2 and 2 when k is sufficiently large. Using this, inequality (i) can be checked for the full sequence (and a fortiori for every subsequence k(n)) in a straightforward way when k is large enough: for $\rho = 0,1,\dots,r-1$ and $\lambda = 1,\dots,\ell$ we have

$$\left|\sigma^{(\rho)}(\theta_\lambda) - \sigma^{\rho}_{\lambda,k}\right| \le \frac{4|\theta_\lambda|}{e_{k+1}-\rho}\left|\sigma^{(\rho+1)}(\theta_\lambda) - \sigma^{\rho+1}_{\lambda,k}\right| ,$$

and for $\lambda = 1,\dots,\ell-1$ we use

$$|\sigma^{(r)}(\theta_\lambda) - \sigma^{r}_{\lambda,k}| \le$$

$$\le \frac{4e_{k+1}(e_{k+1}-1)\dots(e_{k+1}-r+1)}{|\theta_\lambda|^{r}}\left(\frac{|\theta_\lambda|}{|\theta_{\lambda+1}|}\right)^{e_{k+1}}\left|\sigma(\theta_{\lambda+1}) - \sigma^{o}_{\lambda+1,k}\right| .$$

In order to prove inequality (ii), we remark that

$$|\theta_\ell|^{-1} > R^{-1} = \limsup_{k\to\infty}|a_{k+1}|^{1/e_{k+1}} ;$$

therefore an ε, $0 < \varepsilon < 1$, exists such that for large k the inequality

$$|a_{k+1}| < \{(1-\varepsilon)/|\theta_\ell|\}^{e_{k+1}}$$

holds. Hence,

$$|\sigma^{(r)}(\theta_\ell) - \sigma^r_{\ell,k}| \le 2|a_{k+1}|e^r_{k+1}|\theta_\ell|^{e_{k+1}-r} <$$

$$< 2(1-\varepsilon)^{e_{k+1}} e^r_{k+1}|\theta_\ell|^{-r} \le \exp(-c_1^{-1}e_{k+1}) .$$

Here c_1 is a number, greater than 1, that is allowed to depend on the gap series, on $\theta_1,\dots,\theta_\ell$ and on r, but not on k. The same will be the case for $c_2,\dots,c_5$. We proceed by estimating the right-hand side of (ii).

Using inequalities like $\overline{|\alpha+\beta|} \le \overline{|\alpha|} + \overline{|\beta|}$ and $\overline{|\alpha\beta|} \le \overline{|\alpha|}\,\overline{|\beta|}$ it is easily verified that

$$1 + \overline{|\sigma^\rho_{\lambda,k}|} \le 1 + (k+1)A_k e^\rho_k \overline{|\theta_\lambda|}^{e_k-\rho} \le \exp[c_2(\log A_k + e_k)] ,$$

$$m(\sigma^\rho_{\lambda,k}) \le M_k(m(\theta_\lambda))^{e_k-\rho} \le \exp[c_3(\log M_k + e_k)] ,$$

and, with K_k denoting the field obtained by adjoining the numbers $a_0,a_1,\dots,a_k$ and $\theta_1,\dots,\theta_\ell$ to $\mathbb{Q}$,

$$D_k \le [K_k : \mathbb{Q}] \le c_4 S_k .$$

Hence,

$$\prod_{\lambda=1}^{\ell}\prod_{\rho=0}^{r}\left\{m\left(\sigma^\rho_{\lambda,k}\right)\left(1+\overline{|\sigma^\rho_{\lambda,k}|}\right)\right\}^{D_k} \le$$

$$\le \exp[\ell(r+1)\{c_3(\log M_k + e_k) + c_2(\log A_k + e_k)\}\, c_4 S_k] \le$$

$$\le \exp[c_5(\log A_k + \log M_k + e_k)\, S_k].$$

Next, we shall determine the announced subsequence $k(n)$ $(n = 1,2,\ldots)$ by taking $k(0) = 0$ and defining $k(n)$ as the smallest integer $k > k(n-1)$ for which

$$(\log A_k + \log M_k + e_k)S_k/e_{k+1} \leq 1/n^2 .$$

This choice is possible on the strength of the condition in the theorem. From now on, we use the approximating sequence $(\sigma^{\rho}_{\lambda,k(n)})_{n=1}^{\infty}$ to $\sigma^{(\rho)}(\theta_\lambda)$ $(\lambda = 1,\ldots,\ell;\ \rho = 0,1,\ldots,r)$. The inverse of the right-hand side of (ii) becomes

$$\prod_{\lambda=1}^{\ell}\prod_{\rho=0}^{r}\left\{m\left(\sigma^{\rho}_{\lambda,k(n)}\right)\left(1 + \left|\sigma^{\rho}_{\lambda,k(n)}\right|\right)\right\}^{nD_{k(n)}} \leq$$

$$\leq \exp\left[c_5 n\left(\log A_{k(n)} + \log M_{k(n)} + e_{k(n)}\right)S_{k(n)}\right] \leq$$

$$\leq \exp\left[c_5 e_{k(n)+1}/n\right] \leq \exp\left[c_5 e_{k(n+1)}/n\right] .$$

This is smaller than $\exp(c_1^{-1}e_{k(n+1)})$ when n is large enough. Thus, the proof of (ii) and that of the theorem have been completed.

It would be of considerable value to be able to prove variants of the theorem in which some of the θ_λ's are allowed to have the same absolute value. Several mathematicians payed attention to this problem, but up to now without much success. The basic difficulty seems to be, that no longer the concept of "essentially different orders of approximation" can be sufficient as tool for a proof.

REFERENCES

[1] Baron, G. and E. Braune, Zur Transzendenz von Lũckenreihen mit ganzalgebraischen Koefficienten und algebraischem Argument, Comp. Math. 22 (1970), 1 - 6.

[2] Bundschuh, P. and F.-J. Wylegala, Über algebraische Unabhängigkeit bei gewissen nichtvorsetzbaren Potenzreihen, Arch. der Math. 34 (1980), 32 - 36.

[3] Cohn, H., Note on almost-algebraic numbers, Bull. Amer. Math. Soc. 52 (1946), 1042 - 1045.

[4] Cijsouw, P.L. and R. Tijdeman, On the transcendence of certain power series of algebraic numbers, Acta Arith. 23 (1973), 301 - 305.

[5] Durand, A., Indépendance algébrique de nombres complexes et critère de transcendance, Comp. Math. 35 (1977), 259 - 267.

[6] Mahler, K., Arithmetic properties of lacunary power series with integral coefficients, J. Austral. Math. Soc. 5 (1965), 55 - 64.

OVER ENKELE INTEGRALEN DIE SAMENHANGEN MET DE ARCTANGENS-INTEGRAAL

door

P.J. de Doelder

Opgedragen aan Prof. J.J. Seidel t.g.v. zijn afscheid van de T.H. Eindhoven.

0. INLEIDING

In 1959 verzocht Seidel mij voor het tijdschrift Simon Stevin te bespreken het in 1958 verschenen boek van L.Lewin: Dilogarithms and Associated Functions [1]. Na een vluchtig doorbladeren van dit boek, om een indruk van de inhoud te krijgen, nam ik met genoegen deze recensie (die de eerste was die ik leverde) op mij.
In 1959 verscheen de bespreking in bovengenoemd tijdschrift (Jrg. 33, afl. 2, p. 94/96). Later heb ik aan deze materie nog een artikel gewijd, eveneens verschenen in Simon Stevin [2].

In de tweede, gewijzigde en aangevulde, druk van bovengenoemd boek, verschenen in 1981 worden de resultaten uit [2] door de heer Lewin vermeld onder de aanduiding: private communication.

In onderstaand artikel handelt het over een analoog soort integralen als in [2] besproken, t.w. integralen, die samenhangen met $\mathrm{Ti}_2(x)$ en integralen, die samenhangen met $\mathrm{Ti}_3(x)$.

1. INTEGRALEN DIE SAMENHANGEN MET $Ti_2(x)$

1.1. In hoofdstuk 2 van [1] worden behandeld

$$\int_0^1 \frac{Li_2(t)}{(1-xt)^2}\,dt$$

en meer in het algemeen

$$\int_0^1 f(xt)Li_2(t)dt\ ,$$

waarbij

$$Li_2(t) = -\int_0^t \frac{\log(1-u)}{u}\,du$$

en f een functie, die rond 0 ontwikkeld kan worden in een machtreeks.

Voor sommige waarden van x leidt dit tot interessante uitkomsten. De bedoeling van dit hoofdstuk is te onderzoeken of ook voor $Ti_2(t) = \int_0^t \frac{\arctan u}{u}\,du$ iets dergelijks te bereiken is.

We onderzoeken achtereenvolgens

$$\int_0^1 \frac{Ti_2(t)}{(1+xt)^2}\,dt \quad (x \geq 0)$$

en

$$\int_0^1 f(xt)Ti_2(t)dt,$$

waarbij f een machtreeks is als hiervoor aangegeven.

1.2. De integraal $I(x) = \int_0^1 \frac{Ti_2(t)}{(1+xt)^2}\,dt$, $x \geq 0$

1.2.1. $0 \leq x \leq 1$

We passen, om I te berekenen, partiële integratie toe. Er volgt voor $x > 0$:

$$I(x) = -\frac{1}{x}\frac{Ti_2(1)}{1+x} + \frac{1}{x}\int_0^1 \frac{\arctan t}{t(1+xt)}\,dt .$$

(Voor $x = 0$ volgt $I(0) = \int_0^1 Ti_2(t)dt = G - \frac{\pi}{4} + \frac{1}{2}\log 2$, waarbij $Ti_2(1) = G$ de constante van Catalan is.)

Met breuksplitsing bij de integraal vinden we

$$I(x) = \frac{G}{1+x} - \frac{1}{x}\int_0^1 \frac{\arctan t}{t+\frac{1}{x}}\,dt .$$

Definiëren we nog

$$Ti_2(\alpha,\beta) = \int_0^\alpha \frac{\arctan t}{t+\beta}\,dt,$$

dan is

$$I(x) = \frac{G}{1+x} - \frac{1}{x}\,Ti_2(1,\tfrac{1}{x}) . \qquad (1)$$

Gebruik makend van de betrekking

$$Ti_2(1,\tfrac{1}{x}) = \frac{\pi}{2}\log\frac{1+x}{(1+x^2)^{\frac{1}{2}}} - Ti_2(x) + \arctan x\cdot\log x + G - Ti_2(1,x)$$

(zie [3]) volgt er

$$I(x) = -\frac{G}{x(1+x)} - \frac{1}{x}\left[\frac{\pi}{2}\log\frac{1+x}{(1+x^2)^{\frac{1}{2}}} - Ti_2(x) + \arctan x \cdot \log x - Ti_2(1,x)\right]. \tag{2}$$

Voor $|x| < 1$ is het mogelijk $Ti_2(1,x)$ te transformeren in Ti_2-functies, die alleen van x of samenstellingen van x afhangen. Zo is

$$Ti_2(1,x) = \frac{1}{2}Ti_2\left(\frac{1-x}{1+x}\right) - \frac{1}{2}Ti_2(x) - \frac{1}{4}Ti_2\left(\frac{2x}{1-x^2}\right) + \frac{1}{2}G +$$

$$+ \arctan x \cdot \log\frac{x\sqrt{2}}{1+x} - \frac{\pi}{8}\log\frac{1-x^4}{(1+x)^4}, \quad |x| < 1 \qquad \text{(zie [4])}.$$

We merken nog op dat

$$Ti_2(y) - Ti_2\left(\frac{1}{y}\right) = \frac{\pi}{2}\,\mathrm{sgn}\, y \cdot \log|y| \qquad \text{(zie [5])}.$$

Vervangen we met [5] $Ti_2\left(\frac{2x}{1-x^2}\right)$ door $Ti_2\left(\frac{1-x^2}{2x}\right)$ dan vinden we voor $0 < x < 1$:

$$I(x) = \frac{(x-1)}{2x(x+1)}G + \frac{1}{2x}\left\{Ti_2(x) + Ti_2\left(\frac{1-x}{1+x}\right) - \frac{1}{2}Ti_2\left(\frac{1-x^2}{2x}\right) +\right.$$

$$\left. + \log 2 \cdot \arctan x - 2\arctan x \cdot \log(1+x) + \frac{\pi}{4}\log\frac{1+x^2}{2x}\right\}. \tag{3}$$

Dit resultaat is ook geldig voor $x = 1$, zodat

$$I(1) = \frac{1}{2}G - \frac{\pi}{8}\log 2.$$

1.2.2. $x > 1$

We gaan uit van (1): $I(x) = \frac{G}{1+x} - \frac{1}{x}Ti_2(1,\frac{1}{x})$ en maken gebruik van het feit

dat $\frac{1}{x} < 1$, zodat

$$\mathrm{Ti}_2(1,\tfrac{1}{x}) = \tfrac{1}{2}\mathrm{Ti}_2\left(\frac{x-1}{x+1}\right) - \tfrac{1}{2}\mathrm{Ti}_2(\tfrac{1}{x}) - \tfrac{1}{4}\mathrm{Ti}_2\left(\frac{2x}{x^2-1}\right) + \tfrac{1}{2}G +$$

$$+ \arctan\frac{1}{x} \cdot \log\frac{\sqrt{2}}{x+1} - \frac{\pi}{8}\log\frac{x^4-1}{(x+1)^4} \quad \text{wegens [4].}$$

Toepassing van [5] levert samen met $\arctan x + \arctan\frac{1}{x} = \frac{\pi}{2}$:

$$I(x) = \frac{x-1}{2x(x+1)}G + \frac{1}{2x}\left\{-\mathrm{Ti}_2\left(\frac{x-1}{x+1}\right) + \tfrac{1}{2}\mathrm{Ti}_2\left(\frac{x^2-1}{2x}\right) + \mathrm{Ti}_2(x) +\right.$$

$$\left. + \frac{\pi}{4}\log\frac{1+x^2}{2x} + \log 2 \cdot \arctan x - 2\arctan x \cdot \log(1+x)\right\}.$$

$(x > 1)$ (4)

1.2.3. Enkele toepassingen

1.2.3.1. $x = \frac{1-x}{1+x}$

De positieve wortel van deze vergelijking is $x = \sqrt{2} - 1$. Daar $\frac{1-x^2}{2x} = 1$ volgt er

$$I(\sqrt{2}-1) = -\frac{\sqrt{2}+3}{4}G + (\sqrt{2}+1)\mathrm{Ti}_2(\sqrt{2}-1) + \frac{\sqrt{2}+1}{16}\pi\log 2.$$

Jammer is dat $\mathrm{Ti}_2(\sqrt{2}-1)$ voor zover bekend niet in eenvoudige Ti_2-functies is uit te drukken.

1.2.3.2. $\frac{1-x}{1+x} = \frac{2x}{1-x^2}$

Als oplossingen van deze vergelijking vinden we $x = 2-\sqrt{3}$ en $x = 2+\sqrt{3}$.

Gebruik makend van (3) volgt

$$I(2-\sqrt{3}) = \frac{1}{6}G + \frac{2+\sqrt{3}}{48}\left(12\mathrm{Ti}_2(\tfrac{1}{3}\sqrt{3}) + \pi \log \frac{2^6}{3^5}\right), \qquad (5)$$

waarbij we bedenken dat (volgens [6]):

$$3\mathrm{Ti}_2(2-\sqrt{3}) = 2G + \frac{\pi}{4}\log(2-\sqrt{3}) .$$

Gebruik makend van (4) blijkt

$$I(2+\sqrt{3}) = \frac{1}{6}G + \frac{2-\sqrt{3}}{48}\left(-12\mathrm{Ti}_2(\tfrac{1}{3}\sqrt{3}) + \pi \log \frac{2^6}{3^7}\right), \qquad (6)$$

waarbij opgemerkt dient te worden dat (volgens [7]):

$$3\mathrm{Ti}_2(2+\sqrt{3}) = 2G + \frac{5\pi}{4}\log(2+\sqrt{3}) .$$

Uit (5) en (6) blijkt nog

$$(2-\sqrt{3})\ I(2-\sqrt{3}) + (2+\sqrt{3})\ I(2+\sqrt{3}) = \frac{2}{3}G + \frac{\pi}{4}\log\frac{2}{3} . \qquad (7)$$

Ook hier is weer een nadeel dat $\mathrm{Ti}_2(\frac{1}{3}\sqrt{3})$ niet in elementaire functies is uit te drukken.

Bekend is

$$\mathrm{Ti}_2(\tfrac{1}{3}\sqrt{3}) = -\frac{\pi}{12}\log 3 + \frac{5}{6}\mathrm{Cl}_2(\tfrac{\pi}{3})$$

met

$$\mathrm{Cl}_2(x) = \sum_{n=1}^{\infty} \frac{\sin nx}{n^2} \qquad \text{(zie [8])}.$$

Kort geleden is echter aangetoond dat $\mathrm{Cl}_2(\frac{\pi}{3}) = \frac{\sqrt{3}}{6}\left[\psi'(\frac{1}{3}) - \frac{2\pi^2}{3}\right]$ (zie [9]), zodat $\mathrm{Ti}_2(\frac{1}{3}\sqrt{3})$ in de afgeleide van een ψ-functie is uit te drukken.

2. DE INTEGRAAL $H(\alpha) = \int_0^1 t^{\alpha} Ti_2(t)\,dt$, $\alpha \geq 0$

Door partiële integratie vinden we

$$H(\alpha) = \frac{G}{1+\alpha} - \frac{1}{1+\alpha}\int_0^1 t^{\alpha} \arctan t \, dt \, .$$

Nogmaals partieel integreren geeft

$$H(\alpha) = \frac{G}{1+\alpha} - \frac{\pi}{4(1+\alpha)^2} + \frac{1}{(1+\alpha)^2}\int_0^1 \frac{t^{\alpha+1}}{1+t^2}\,dt \, .$$

Omdat

$$\int_0^1 \frac{t^{x-1}}{1+t}dt = \beta(x) = \frac{1}{2}\left[\psi\left(\frac{x+1}{2}\right) - \psi\left(\frac{x}{2}\right)\right]$$

met

$$\psi(x) = \frac{\Gamma'(x)}{\Gamma(x)} \, ,$$

waarbij $\Gamma(x)$ de gammafunctie is, kunnen we $H(\alpha)$ uitdrukken in ψ-functies, t.w.

$$H(\alpha) = \frac{G}{1+\alpha} - \frac{\pi}{4(1+\alpha)^2} + \frac{1}{(1+\alpha)^2}\left\{\psi\left(\frac{\alpha}{4}+1\right) - \psi\left(\frac{\alpha}{4}+\frac{1}{2}\right)\right\} \, . \qquad (8)$$

3. DE INTEGRAAL $\int_0^1 f(xt)Ti_2(t)\,dt$

Zij f in een Taylorreeks ontwikkelbaar rond 0. Dan is wegens (8)

$$F(x) = \int_0^1 f(xt)Ti_2(t)\,dt = \sum_{n=0}^{\infty} a_n x^n \int_0^1 t^n Ti_2(t)\,dt =$$

$$= \sum_{n=0}^{\infty} a_n x^n \left[\frac{G}{n+1} - \frac{\pi}{4(n+1)^2} + \frac{1}{4(n+1)^2} \left\{ \psi\left(\frac{n}{4}+1\right) - \psi\left(\frac{n}{4}+\frac{1}{2}\right) \right\} \right] =$$

$$= G \sum_{n=0}^{\infty} \frac{a_n x^n}{n+1} - \frac{\pi}{4} \sum_{n=0}^{\infty} \frac{a_n x^n}{(n+1)^2} + \frac{1}{4} \sum_{n=0}^{\infty} \frac{a_n \left\{ \psi\left(\frac{n}{4}+1\right) - \psi\left(\frac{n}{4}+\frac{1}{2}\right) \right\} x^n}{(n+1)^2} . \qquad (9)$$

In de gevallen waarin F eenvoudig berekenbaar is en het mogelijk blijkt de eerste twee sommaties in het rechterlid in eenvoudige functies uit te drukken, is een gesloten uitdrukking voor de laatste som te vinden, doch dat is zelden het geval.

We passen het gevondene toe op $f(x) = \frac{1}{(1+x)^2}$.

De reeksontwikkeling rond 0 is $\sum_{n=0}^{\infty} (-1)^n (n+1) x^n$, zodat $a_n = (-1)^n (n+1)$.
We vinden hieruit

$$F(x) = \int_0^1 \frac{Ti_2(t)}{(1+xt)^2}\, dt = \frac{G}{1+x} - \frac{\pi}{4} \frac{\log(1+x)}{x} + \frac{1}{4} \sum_{n=1}^{\infty} \frac{(-1)^n x^n}{n+1} \cdot$$

$$\cdot \left\{ \psi\left(\frac{n}{4}+1\right) - \psi\left(\frac{n}{4}+\frac{1}{2}\right) \right\} . \qquad (10)$$

Voor de waarden van x met $0 < x < 1$, die beschouwd zijn in 1.2.3.1 en 1.2.3.2 is de reeks dus uit te drukken in Ti_2-functies. Ook voor $x = 1$ is (10) correct, er volgt

$$\int_0^1 \frac{Ti_2(t)}{(1+t)^2}\, dt = \frac{1}{2}G - \frac{\pi}{4} \log 2 + \frac{1}{4} \sum_{n=0}^{\infty} \frac{(-1)^n}{n+1} \left\{ \psi\left(\frac{n}{4}+1\right) - \psi\left(\frac{n}{4}+\frac{1}{2}\right) \right\} .$$

Wegens (3) is

$$\int_0^1 \frac{\mathrm{Ti}_2(t)}{(1+t)^2} = \frac{1}{2}G - \frac{\pi}{8}\log 2 ,$$

zodat tenslotte

$$\sum_{n=0}^{\infty} \frac{(-1)^n}{n+1}\left\{\psi\left(\frac{n}{4}+1\right) - \psi\left(\frac{n}{4}+\frac{1}{2}\right)\right\} = \frac{\pi}{2}\log 2. \qquad (11)$$

Opmerkingen

1. (11) kan ook op een andere manier worden gevonden.

2. Mij is gebleken dat $\int_0^1 \frac{\mathrm{Li}_2(t)}{2-t}\,dt = \frac{\pi^2}{4}\log 2 - \zeta(3)$ en $\int_0^1 \frac{\mathrm{Li}_2(t)}{1+t}dt = -\frac{\pi^2}{6}\log 2 + \frac{3}{2}\zeta(3)$.

   Hierbij is ζ de zetafunctie van Riemann.

   Ik ben er echter niet in geslaagd eenvoudige gesloten uitdrukkingen te vinden voor

   $\int_0^1 \frac{\mathrm{Ti}_2(t)}{2-t}dt$ en $\int_0^1 \frac{\mathrm{Ti}_2(t)}{1+t}dt$ en ik vermoed dat die ook niet te vinden zijn.

4. GENERALISATIE

4.1. We definiëren $\mathrm{Ti}_q(x) = \int_0^x \frac{\mathrm{Ti}_{q-1}(t)}{t}dt$, $q = 2,3,\ldots$,

met $T_1(x) = \arctan x$ en bekijken

$$I(x) = \int_0^x t^p \mathrm{Ti}_q(t)\,dt , \quad p \geq 0 .$$

Door herhaald partieel integreren vinden we

$$I(x) = \frac{x^{p+1}}{p+1}\left[Ti_q(x) - \frac{T_{q-1}(x)}{p+1} + \ldots + \frac{(-1)^{q-2}}{(p+1)^{q-2}}Ti_2(x)\right] -$$

$$- \frac{(-1)^{q-2}}{(p+1)^{q-1}}\int_0^x t^p \arctan t\, dt.$$

Omdat

$$\int_0^x t^p \arctan t\, dt = \frac{x^{p+1}}{p+1}\arctan x - \frac{1}{2(p+1)}\int_0^{x^2}\frac{t^{p/2}}{1+t}\,dt,$$

volgt tenslotte

$$I(x) = \frac{x^{p+1}}{p+1}\left[Ti_q(x) - \frac{Ti_{q-1}(x)}{p+1} + \ldots + \frac{(-1)^{q-2}}{(p+1)^{q-2}}Ti_2(x)\right] -$$

$$- \frac{(-1)^{q-2}}{(p+1)^q}\left\{x^{p+1}\arctan x - \frac{1}{2}\int_0^{x^2}\frac{t^{p/2}}{1+t}\,dt\right\}.$$

Voor $x = 1$ wordt dit

$$I(1) = \frac{1}{p+1}\left(Ti_q(1) - \frac{T_{q-1}(1)}{p+1} + \ldots + \frac{(-1)^{q-2}}{(p+1)^{q-2}}G\right) -$$

$$- \frac{(-1)^{q-2}}{4(p+1)^q}\left[\pi - \psi\left(\frac{p}{4}+1\right) + \psi\left(\frac{p}{4}+\frac{1}{2}\right)\right]. \qquad (12)$$

4.2. Het is nu mogelijk de integraal

$$J(x) = \int_0^1 f(xt)Ti_q(t)\,dt$$

uit te drukken in reeksen met Ti-functies en één met ψ-functies.

We gaan weer uit van $f(x) = \sum_{0}^{\infty} a_n x^n$ en er volgt dan

$$J(x) = Ti_q(1) \sum_{n=0}^{\infty} \frac{a_n x^n}{n+1} - Ti_{q-1}(1) \sum_{n=0}^{\infty} \frac{a_n x^n}{(n+1)^2} + \ldots + Ti_2(1) \sum_{n=0}^{\infty} \frac{(-1)^{q-2} a_n x^n}{(n+1)^{q-1}} -$$

$$- \frac{\pi}{4} \sum_{n=0}^{\infty} \frac{(-1)^{q-2} a_n x^n}{(n+1)^q} + \frac{1}{4} \sum_{n=0}^{\infty} \frac{(-1)^{q-2} x^n}{(n+1)^q} \left\{ \psi\left(\frac{n}{4}+1\right) - \psi\left(\frac{n}{4}+\frac{1}{2}\right) \right\} . \qquad (13)$$

Ook hier is weer het probleem dat het slechts zelden mogelijk is $J(x)$ te berekenen, zelfs voor eenvoudige waarden van f en daarmee de laatste reeks in (13) te vinden of omgekeerd,

4.3. Een eenvoudige toepassing is:

Neem $q = 3$; $a_n = (-1)^n (n+1)^q$, zodat $f(t) = \frac{1-t}{(1+t)^3}$.

Er blijkt dan voor $0 \leq x < 1$:

$$\int_0^1 \frac{(1-xt)}{(1+xt)^3} Ti_3(t)\,dt = \frac{\pi^3}{32} \frac{1}{(1+x)^2} - \frac{G}{1+x} + \frac{\pi}{4} \frac{\log(1+x)}{x} -$$

$$- \frac{1}{4} \sum_{n=0}^{\infty} \frac{(-1)^n x^n}{n+1} \left\{ \psi\left(\frac{n}{4}+1\right) - \psi\left(\frac{n}{4}+\frac{1}{2}\right) \right\} ,$$

daar $Ti_3(1) = \frac{\pi^3}{32}$ (zie [8]).

Uit het voorgaande volgt nog

$$\int_0^1 \frac{(1-xt)}{(1+xt)^3} Ti_3(t)\,dt = - \int_0^1 \frac{Ti_2(t)}{(1+xt)^2}\,dt + \frac{\pi^3}{32} \frac{1}{(1+x)^2} ,$$

ook geldig voor $x = 1$.

We vinden tenslotte voor x = 0

$$\int_0^1 \mathrm{Ti}_3(t)\,dt = \frac{\pi^3}{32} - G + \frac{\pi}{4} - \frac{1}{2}\log 2$$

en voor x = 1:

$$\int_0^1 \frac{1-t}{(1+t)^3}\,\mathrm{Ti}_3(t)\,dt = \frac{\pi^3}{128} - \frac{1}{2}G + \frac{\pi}{8}\log 2 \ .$$

REFERENTIES

[1] Lewin, L., Dilogarithms and Assoc. Functions, McDonald, 1958.

[2] Doelder de, P.J. On some integrals of the trilogarithm type, Simon Stevin, Jrg 36, afl.2, p. 90 - 99.

[3] Lewin, L., Polylogarithms and Assoc. Functions, North Holland Publ., 1981, p. 75 (3.27).

[4] ibid p. 71 (3.9) .

[5] " p. 39 (2.6).

[6] " p. 45 (2.29).

[7] " p. 45 (2.30).

[8] " p. 106 (4.31) .

[9] Fettis, H.E., Advanced problems no. 6448, Am. Math. Monthly, Jan. 1984.

LESLIE-MATRICES

Een proeve van toegepaste wiskunde voor het V.W.O

en een didactisch dilemma

door

J.G.M. Donkers

Opgedragen aan Prof. J.J. Seidel, bij gelegenheid van diens afscheid als hoogleraar aan de Technische Hogeschool Eindhoven.

SAMENVATTING

Leslie-matrices zijn bevolkingsvoorspellings-matrices. Ze komen als voorbeeld van toegepaste wiskunde voor in het boek Matrices, dat, overeenkomstig de ideeën van de Hewet-werkgroep, is ontwikkeld ten behoeve van het nieuwe Wiskunde A programma voor de bovenbouw van het V.W.O. De wiskundige achtergrond van het Leslie-model ligt ver boven het V.W.O.-niveau. Dit veroorzaakt een didactisch dilemma.

In de inleiding wordt de reeds in het verleden gehoorde roep om toegepaste wiskunde in het V.W.O.-programma op te nemen in herinnering gebracht. Na een korte bespreking van het advies van de Hewet-werkgroep en een overzicht van het boek Matrices volgt de wiskundige achtergrond van het Leslie-model.

In de laatste paragraaf wordt de aandacht gevestigd op enkele didactische moeilijkheden die kunnen optreden wanneer niet-triviale toepassingen van de wiskunde in het V.W.O.-programma worden opgenomen.

1. TOEGEPASTE WISKUNDE IN HET V.W.O. ?

"Voor steeds meer gebieden blijkt wiskunde niet slechts een hoeveelheid kennis, maar ook een benaderingswijze, een onontbeerlijk aspect te zijn. Gebruiken en voeden gaan samen en de wisselwerking met de andere wetenschappen dringt op een steeds breder terrein door. Dit vraagt een heroriëntering van het onderwijs. De vormende taak krijgt in haar gerichtheid duidelijker gestalte en van het hulpmiddel dient vooral de levende kant op de voorgrond te komen. Voor een verantwoorde hantering van het werktuig is een overzicht van zijn mogelijkheden noodzakelijk. Aan het wiskunde-onderwijs stelt dit eisen van breedheid, efficiëntie en actualiteit". [10] Aldus de jonge hoogleraar J.J. Seidel in zijn inaugurele rede, uitgesproken op 25 februari 1958 aan de Technische Hogeschool te Eindhoven.

Deze meer dan een kwart eeuw geleden uitgesproken woorden hebben aan actualiteit niets verloren. Voor de in deze rede uitgesproken gedachte, dat van het hulpmiddel vooral de levende kant op de voorgrond dient te komen, is pas de laatste jaren in kringen van het voortgezet onderwijs belangstelling getoond. Hoe, een kwart eeuw geleden, de stand van zaken in het wiskunde-onderwijs was, wordt door de volgende passage uit Seidel's rede heel goed weergegeven:

"De vraag, hoe de functie van de wiskunde in de hedendaagse wetenschap en maatschappij bij het onderwijs moet worden voorbereid, betreft niet alleen het hoger, maar ook het middelbaar onderwijs. Weliswaar stelt de didactiek daar zijn eigen eisen, maar die mogen niet uitsluitend bepalend zijn voor de te behandelen onderwerpen. Gelukkig zoch het middelbaar onderwijs zelf nieuwe wegen. In 1954 werd een ontwerp-leerplan opgesteld, volgens hetwelk onder-

werpen uit de infinitesimaalrekening, de analytische meetkunde en de statistiek ter vervanging van uitwassen en verouderde leerstof worden ingevoerd. Wij zien deze voorstellen als een stap vooruit en betreuren het slechts dat de officiële invoering zo lang op zich laat wachten. De kennismaking met de rol van de wiskunde in de huidige maatschappij zou erdoor worden vervroegd". [10]

De officiële invoering voor de bovenbouw van het V.H.M.O. geschiedde in '58-'59, maar niet zoals Seidel, en sommigen met hem, zich dat had voorgesteld. De toegepaste wiskunde in casu de invoering van de statistiek, vond geen genade in de ogen van de toenmalige curriculumdeskundigen. Dit ondanks Seidel's uitvoerige verdediging van het vak toegepaste wiskunde voor het voortgezet onderwijs, gedaan tijdens een voordracht gehouden voor de vacantiecursus 1955 van het M.C. We citeren: "Is het dan niet gerechtvaardigd, dat ook het middelbaar onderwijs haar aandacht geeft aan de wiskunde, die toepasbaar is en daartoe haar leerplan verandert? Laten wij het ontwerp-leerplan een ogenblik in dit licht beschouwen.

Door de invoering van de statistiek zal de aanstaande wiskundestudent reeds op school kennismaken met de beginselen van een stuk toegepaste wiskunde. Dit kan van betekenis zijn voor de toekomst van de wiskunde, in die zin, dat door deze kennismaking een grondslag kan worden gelegd voor begrip tussen de beoefenaars der zuivere en der toegepaste wiskunde. Een erkenning van de toegepaste wiskunde bij middelbaar en hoger onderwijs kan helpen om de onlustgevoelens, waarmee sommige zuiveren de toepassingen beschouwen, de ondergrond, die toch vaak uit onkunde bestaat, te ontnemen". [9] De "onlustgevoelens van

zuiveren" waren te groot: op de invoering van de statistiek in het programma zou men nog bijna twee decennia moeten wachten. Pas in het nieuwe programma van 1974 komt "waarschijnlijkheidsrekening en mathematische statistiek" voor als een onderdeel van het wiskunde I programma.

Noch met de programma's uit de zestiger jaren noch met die uit de zeventiger jaren heeft men een "afglijden in rekentechnieken" [19], waarvoor Seidel in 1955 reeds waarschuwde, kunnen voorkomen.

Mede doordat de sociale-, de natuurwetenschappelijke- en de technische faculteiten aan de aankomende studenten de eis stelden dat wiskunde I onderdeel van het eindexamen diende te hebben uitgemaakt, nam het aantal leerlingen voor wiskunde I in de zeventiger jaren belangrijk toe. Door een toenemende heterogeniteit der groepen kwam het vak onder grote druk te staan. In 1978 werd door de toenmalige staatssecretaris voor Onderwijs en Wetenschappen de zogenaamde Hewet-werkgroep (Herverkaveling Wiskunde Een en Twee) ingesteld, die diende te rapporteren omtrent een nieuw wiskunde programma voor de bovenbouw van het V.W.O. In 1980 verscheen het eindrapport van deze werkgroep. [8] In dit rapport wordt de toegepaste wiskunde, in het voorgestelde wiskunde A programma, niet beperkt tot de waarschijnlijkheidsrekening en statistiek. Nadat we in de volgende paragrafen eerst enkele karakteristieken van het nieuwe programma hebben gegeven, gevolgd door een overzicht van het leerboek Matrices, zullen we de wiskundige achtergrond schetsen van het onderwerp "Leslie-matrices", een onderwerp uit de bevolkingsdynamica, dat als voorbeeld van toegepaste wiskunde in het boek Matrices voorkomt.

Dit onderwerp lijkt een goede illustratie van de didactische problemen die zich kunnen voordoen wanneer niet-triviale onderwerpen uit de toegepaste wiskunde in een leerstofpakket voor het V.W.O. worden opgenomen.

2. HET ADVIES VAN DE HEWET-WERKGROEP [8]

De aanbevelingen van de Hewet-werkgroep kunnen als volgt worden samengevat: er moeten twee nieuwe programma's komen, te weten wiskunde A en wiskunde B. Het voorgestelde programma voor wiskunde A bestaat uit de volgende onderdelen:

1. Eenvoudige analyse en toegepaste analyse.
2. Matrixrekening met toepassingen.
3. Waarschijnlijkheidsrekening en Statistiek.
4. Automatische gegevensverwerking.

Het voorgestelde programma voor wiskunde B bestaat uit:

1. Analyse.
2. Meetkunde.

Wiskunde B is bestemd voor aanstaande studenten in de technische wetenschappen en voor die in de factulteit der wiskunde en natuurwetenschappen. Het onderdeel Analyse, dat ruim tweederde deel van het programma uitmaakt, is nagenoeg gelijk aan het huidige programma wiskunde I. We zullen hier verder geen aandacht aan besteden.

Wiskunde A is bestemd voor leerlingen die in hun academische studie weinig vervolgonderwijs zullen krijgen in de wiskunde, maar wel in beperkte mate wiskunde als instrument moeten gebruiken.

Volgens het rapport "moeten de leerlingen in hun onderwijs de waarde van een wiskundig getinte presentatie leren beoordelen". "Verder moeten ze leren werken met wiskundige modellen en de relevantie van die modellen kunnen beoordelen".

Ook acht de werkgroep het "vanuit het oogpunt van een algemene wiskundige vorming van het grootste belang dat de leerling zich op verschillende wiskundige terreinen heeft georiënteerd". Het rapport vervolgt, "Het programma lijkt vanuit wiskundig standpunt misschien wat verbrokkeld. De eenheid moet echter in eerste instantie niet gezocht worden in de wiskundige begrippen, maar in het toepassingskarakter".

Tevens beveelt de werkgroep aan dat leerlingenteksten ten behoeve van experimenten en nascholingscursussen worden ontwikkeld. Aan deze laatste aanbeveling is reeds voor een groot deel voldaan. Enkele medewerkers van de vakgroep O.W & O.C van de Rijksuniversiteit Utrecht hebben inmiddels een aantal leerstofpakketten ontwikkeld. We zullen ons beperken tot het leerlingenboek Matrices. [5]

Vergelijken we de wijzigingen in de programma's van de laatste 25 jaren met het nu door de Hewet-werkgroep voorgestelde nieuwe programma wiskunde A, dan constateren we een duidelijke breuk met de traditie. Ook vroeger zijn er heftige discussies geweest over de relevantie van de nieuwe programma's en gingen voor velen de vernieuwingen niet ver genoeg. [1] In een dergelijke discussie merkte Freudenthal op: "Moderne programma's moeten m.i. allereerst dienen om het onderwijs methodisch en didactisch te verbeteren. Kan men dit

niet verwezenlijken, dan zijn de nieuwe programma's erger dan de oude" [2]. Deze door Freudenthal gelanceerde opvatting lijkt duidelijk de overhand te hebben gehad bij het bepalen en uitwerken van het nieuwe programma, want veel meer nog dan de inhoudelijke veranderingen zijn het de methodische en didactische wijzigingen die dit nieuwe programma zo doen verschillen van de traditionele programma's.

Of dit programma zó zal kunnen worden verwezenlijkt als door de Hewet-werkgroep is beoogd, valt nu nog niet te zeggen. Inmiddels is wel bepaald dat de eerste landelijke examens voor wiskunde A en B in 1987 zullen plaatsvinden.

3. HET LEERLINGENBOEK MATRICES [5]

Het leerlingenboek Matrices is een uitwerking van het onderdeel Matrixrekening en toepassingen uit het wiskunde A programma en is bedoeld voor de 5e klas van het V.W.O. In de eerste hoofdstukken worden matrices geïntroduceerd aan de hand van transportproblemen op een eiland in de Stille Zuidzee en later tussen enkele eilanden in dat gebied. Centrale begrippen zijn hier afstandsmatrices en verbindingsmatrices, die behoren bij verschillende grafen. Vervolgens leiden problemen uit het voorraadbeheer van een winkel (in spijkerbroeken) tot de zogenaamde voorraadmatrices. In deze context wordt ook het vermenigvuldigen van matrices geïntroduceerd. Na een hoofdstuk, waarin de verwerking van matrices met behulp van de computer een rol speelt en een hoofdstuk over incidentie- en kansmatrices, volgen de migratiematrices en de bevolkingsvoorspellingsmatrices.

De laatsten worden Leslie-matrices genoemd, naar de Engelse zoöloog Leslie, die deze in 1948 als een der eersten in een bevolkingsmodel heeft gebruikt [7]. Zo'n bevolkingsmodel wordt ook wel het Leslie-model genoemd. Het Leslie-model kan als volgt worden beschreven.

Verdeel een bevolking in n leeftijdsklassen met gelijke (tijds)klasse-breedten. Zo'n breedte heet een periode. Na een periode bereikt uit klasse i een fractie p_i de klasse $i+1$, de rest sterft. Uit de laatste klasse is na één periode iedereen verdwenen. Stel de gemiddelde vruchtbaarheid per individu uit klasse i over één periode is v_i. Dit betekent, als er n_i individuen in klasse i zijn, dan brengen ze na één periode $n_i v_i$ individuen in de klasse 1 voort.

Laat L de $n \times n$ matrix

$$\begin{pmatrix} v_1 & v_2 & \cdots & v_n \\ p_1 & 0 & \cdots & 0 \\ 0 & p_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots\cdots & p_{n-1} & 0 \end{pmatrix} \quad \text{en} \quad \underline{x}_0 = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

de bevolkingsvector op tijdstip t_0 zijn, dan geldt: $L\underline{x}_0 = \underline{x}_1$, waarin $\underline{x}_1$, de bevolkingsvector is op het tijdstip t_1, één periode na t_0. L heet de Leslie-matrix behorende bij het Leslie-model. Er geldt:

$$v_i \geq 0 \quad \text{en} \quad 0 < p_i \leq 1 \quad \text{voor alle} \quad i \in \{1,2,\ldots,n\} .$$

In de leerlingentekst wordt het Leslie-model geïntroduceerd aan de hand van een keverbevolking met gegeven Leslie-matrix:

$$\begin{pmatrix} 0 & 0 & 6 \\ \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{3} & 0 \end{pmatrix}.$$

Verder beperkt de leerlingentekst zich in hoofdzaak tot het interpreteren van de getallen v_i en p_i in enkele gegeven Leslie-matrices en tot het opstellen van Leslie-matrices bij enkele getallen voorbeelden van populaties.

In het laatste hoofdstuk van het boek worden de Leslie-matrices toegepast op een bevolking van klapmutsen, een soort robben, die leven in de Noordelijke IJszee. Hier wordt gebruik gemaakt van recent, door biologisch onderzoek verkregen, demografisch materiaal. De leerlingen moeten uit de gegevens de Leslie-matrix bepalen. Daarna wordt, uitgaande van een tot in procenten herleide bevolkingsvector ($\underline{x}_0 = (x_1, x_2, \ldots, x_n)$, $\sum_{i=1}^{n} x_i = 100$, $x_i \geq 0$ alle i), met behulp van de computer voor de opeenvolgende tijdstippen $t_1, t_2, \ldots$ de procentuele bevolkingsvectoren $\underline{x}_1, \underline{x}_2, \ldots$ bepaald, totdat er een stabiele procentuele bevolkingsvector $\underline{x}_s$ is verkregen. Uitgaande van deze $\underline{x}_s$ wordt met behulp van de computer nog één geïtereerde bepaald, maar nu in absolute aantallen. De computer berekent nu tevens de zogenaamde groeifactor. Vervolgens worden er verschillende jachtstrategieën besproken. De leerlingen moeten de bij de verschillende strategieën behorende Leslie-matrices bepalen. Tenslotte wordt, uitgaande van een vaste beginpopulatie $\underline{x}_0$, voor ieder der jachtstrategieën het bij de bevolkingsvector $\underline{x}_{10}$ behorende bevolkingshistogram gegeven, waaruit de leerlingen conclusies moeten trekken met betrekking tot de verschillen tussen de strategieën.

Gemakkelijk is in te zien dat de stabiele procentuele bevolkingsvector $\underline{x}_s$, eigenvector is van de Leslie-matrix, behorende bij een positieve eigenwaarde, de zogenaamde groeifactor.

Vragen omtrent de existentie en eenduidigheid van een stabiele vector, het al dan niet optreden van convergentie en dergelijke komen in de leerlingentekst niet aan de orde. Het al dan niet ingaan op dergelijke vragen in een leerlingentekst is een didactisch probleem, omdat bij de leerlingen de benodigde wiskundige voorkennis ontbreekt. We zullen hierop terugkomen, nadat we in de volgende paragraaf de wiskundige achtergrond van het Leslie-model hebben gegeven.

Een globale bespreking van het boek Matrices is te vinden in [4] en een gedetailleerde bespreking van het laatste hoofdstuk in [6].

4. DE WISKUNDIGE ACHTERGROND VAN HET LESLIE-MODEL

Een vector in $\mathbb{R}^n$ met alleen positieve respectievelijk niet-negatieve kentallen heet een positieve respectievelijk niet-negatieve vector. Een matrix met alleen positieve respectievelijk niet-negatieve elementen heet een positieve respectievelijk niet-negatieve matrix.

Voor een Leslie-matrix

$$L = \begin{pmatrix} v_1 & v_2 & \cdots & v_n \\ p_1 & 0 & \cdots & 0 \\ 0 & p_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & p_{n-1} \end{pmatrix}$$

geldt, zoals we reeds eerder zagen:

$$v_i \geq 0 \text{ voor alle } i \in \{1,\ldots,n\} \text{ en } 0 < p_i \leq 1$$

$$\text{voor alle } i \in \{1,\ldots,n-1\} .$$

We stellen voorlopig $v_n > 0$.

Is M een lineaire afbeelding van $\mathbb{R}^n$ in zichzelf en M de matrixvoorstelling van M ten opzichte van een basis $\underline{e}_1,\underline{e}_2,\ldots,\underline{e}_n$ van $\mathbb{R}^n$, dan heet M irreducibel indien er geen k basisvectoren $\underline{e}_{i_1},\ldots,\underline{e}_{i_k}$, met $0 < k < n$ bestaan, zodat de ruimte die zij opgespannen invariant is onder M.

Is de Leslie-matrix L, matrix van een lineaire afbeelding L ten opzichte van de basis $\underline{e}_1$ t/m $\underline{e}_n$ van $\mathbb{R}^n$, dan is gemakkelijk in te zien dat een invariante deelruimte met $\underline{e}_i$ ook $\underline{e}_{i+1}$ bevat voor alle $i \in \{1,\ldots,n\}$, waarbij we $\underline{e}_{n+1} = \underline{e}_1$ stellen.

GEVOLG. Een Leslie-matrix is een niet-negatieve irreducibele matrix.

Voor niet-negatieve irreducibele matrices geldt de volgende

STELLING VAN FROBENIUS-PERRON (zie [3], [11]). Zij A een $n \times n$ niet-negatieve irreducible matrix, dan geldt:

(a) Er is een positief reëel getal r, dat enkelvoudige wortel is van de karakteristieke vergelijking van A, en voor iedere wortel λ_i van de karakteristieke vergelijking van A geldt: $|\lambda_i| \leq r$.

(b) Er bestaat een positieve eigenvector van A bij de eigenwaarde r.

(c) Heeft de karakteristieke vergelijking van A h wortels met modulus r, dus $\lambda_1 = r$ en $|\lambda_2| = |\lambda_3| = \ldots = |\lambda_h| = r$, dan zijn deze de h verschillende wortels van de vergelijking $\lambda^h - r^h = 0$.

(d) Is $\rho = e^{\frac{2\pi}{h}i}$ en zijn λ_1 t/m λ_n de n wortels van de karakteristieke vergelijking van A, dan zijn $\rho\lambda_1,\ldots,\rho\lambda_n$ op volgorde na precies λ_1 t/m λ_n.

De in de bovenstaande stelling voorkomende matrix A heet primitief indien h = 1 anders imprimitief.

Uit de Stelling van Frobenius-Perron volgt dat Leslie-matrices een (maximale) positieve reële eigenwaarde bezitten met een (uiteraard op een factor na) eenduidig bepaalde positieve eigenvector. Ook is gemakkelijk aan te tonen, zie [3], dat er géén twee lineair onafhankelijke positieve eigenvectoren bestaan.

Voor een primitieve matrix A geldt (zie [3] en [11]):

er is een positieve n ×n matrix H, waarvan de kolommen veelvouden zijn van een bij de maximale positieve reële eigenwaarde r van A behorende positieve eigenvector $\underline{w}$, d.w.z. $H = [\alpha_1\underline{w},\alpha_2\underline{w},\ldots,\alpha_n\underline{w}]$ met $\alpha_i > 0$ voor alle $i \in \{1,\ldots,n\}$, zodat:

$$\lim_{k\to\infty} \frac{A^k}{r^k} = H \ .$$

M.a.w. voor een niet-negatieve vector $\underline{x} = (x_1,x_2,\ldots,x_n)$ vinden we:

$$\lim_{k\to\infty} \frac{A^k\underline{x}}{r^k} = H\underline{x} = (\alpha_1x_1 + \ldots + \alpha_nx_n)\underline{w} = \mu\underline{w} \ ,$$

waarin

$$\alpha_1x_i + \ldots + \alpha_nx_n = \mu > 0 \ .$$

Uitgaande van een op procenten herleide bevolkingsvector, hebben we hiermee voor een primitieve Leslie-matrix de convergentie naar een (op procenten herleide) positieve stabiele bevolkingsvector verkregen.

M.b.v. de stelling: (zie [3])

> Een niet-negatieve matrix A is primitief d.e.s.d. als er een natuurlijk getal $p \geq 1$ is met A^p positief

volgt nu direct dat er voor een niet-primitieve Leslie-matrix géén convergentie naar een stabiele vector optreedt.

Zij nu $r = \lambda_1 = |\lambda_2| = \ldots = |\lambda_h|$, en laat

$$\lambda^n + a_1\lambda^{n_1} + a_2\lambda^{n_2} + \ldots + a_t\lambda^{n_t} = 0$$

de karakteristieke vergelijking zijn van de niet-negatieve irreducibele $n \times n$ matrix A, met $n > n_1 > n_2 > \ldots > n_t$ en $a_1 \neq 0$, ... , $a_t \neq 0$, dan volgt uit de delen (c) en (d) van de Stelling van Frobenius-Perron (zie [3]) dat:

$$h = ggd(n - n_1,\ n_1 - n_2\ ,\ldots,\ n_{t-1} - n_t) \quad \ldots *$$

Voor de Leslie-matrix L is de karakteristieke vergelijking:

$$\lambda^n - v_1\lambda^{n-1} - p_1v_2\lambda^{n-2} - \ldots - (p_1p_2\ldots p_{n-1})v_n = 0 \ .$$

Stel $v_{j_1} > 0,\ v_{j_2} > 0\ ,\ldots,v_{j_k} = v_n > 0$ met

$$1 \leq j_1 < j_2 \ldots < j_k = n$$

en alle andere v_i's gelijk aan nul, dan is de karakteristieke vergelijking:

$$\lambda^n - a_1\lambda^{n-j_1} - a_2\lambda^{n-j_2} \ldots - a_{k-1}\lambda^{n-j_{k-1}} - a_k = 0$$

met

$$a_i = p_1 p_2 \cdots p_{j_i-1} v_{j_i} > 0 .$$

We passen nu * toe en vinden (zie ook [12])

$$h = ggd(n - (n-j_1),\ (n-j_1) - (n-j_2),\ \ldots,\ n - j_{k-1})$$

$$= ggd(j_1, j_2 - j_1,\ \ldots,\ n - j_{k-1})$$

$$= ggd(j_1, j_2, \ldots, j_{k-1}, n) .$$

De laatste gelijkheid is als volgt in te zien:

Stel

$$d = ggd(j_1, j_2, \ldots, j_k)$$

en

$$q_1 = j_1,\ q_2 = j_2 - j_1,\ \ldots,\ q_k = n - j_{k-1}$$

dan geldt

$$j_1 = q_1,\ j_2 = q_1 + q_2,\ j_3 = q_1 + q_2 + q_3,\ \ldots \text{ enz.}$$

Nu is d een positieve gehele lineaire combinatie van j_1 t/m j_k, dus ook van de q_1 t/m q_k, maar h in de kleinste positieve gehele lineaire combinatie van de q_1 t/m q_k, dus is $h \leq d$. Analoog geldt $h \geq d$. We kunnen het bovenstaande nu samenvatten in de volgende

STELLING. Een Leslie-matrix is primitief d.e.s.d. als de ggd van de indices van de getallen v_i in de eerste rij, waarvoor $v_i \neq 0$, gelijk is aan 1.

Omdat er alleen dan convergentie optreedt als de Leslie-matrix primitief is, hebben we met de laatste stelling een eenvoudig criterium voor convergentie verkregen.

Het geval van een $n \times n$ Leslie-matrix met $v_n = 0$ kunnen we herleiden tot het hierboven besproken geval met $v_n \neq 0$, en wel als volgt:

Laat L een $n \times n$ Leslie-matrix zijn met $v_{n-k} \neq 0$ en $v_{n-k+1} = \ldots = v_n = 0$ voor zeker k met $0 < k < n$, dan is L te schrijven als:

$$L = \begin{pmatrix} A & 0 \\ B & C \end{pmatrix},$$

waarin

0 de $(n-k) \times k$ 0-matrix ,

A een $(n-k) \times (n-k)$ Leslie-matrix met $v_{n-k} \neq 0$,

B de $k \times (n-k)$ matrix

$$\begin{pmatrix} 0 & \ldots & 0 & p_{n-k} \\ & & & 0 \\ & \bigcirc & & \vdots \\ & & & 0 \end{pmatrix} \quad \text{en}$$

C de $k \times k$ matrix

$$\begin{pmatrix} 0 & \cdots & \cdots & 0 \\ p_{n-k+1} & \ddots & & \vdots \\ & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \ldots & 0 \; p_{n-1} & 0 \end{pmatrix} \quad \text{is .}$$

Omdat $C^k = 0$, vinden we

$$L^k = \begin{pmatrix} A^k & 0 \\ D & 0 \end{pmatrix}$$

waarin

$$D = \sum_{i=0}^{k-1} C^i B A^{k-1-i} ,$$

zodat

$$L^{k+j} = \begin{pmatrix} A^{k+j} & 0 \\ DA^j & 0 \end{pmatrix} , \quad j \geq 0 .$$

Het zal nu duidelijk zijn dat al dan niet convergentie naar een stabiele vector voor L volledig wordt bepaald door de Leslie-matrix A.

5. EEN DIDACTISCH DILEMMA

Het leerlingenboek Matrices lijkt een goede representant van het tot nu toe ontwikkelde materiaal voor wiskunde A. Het brengt de leerlingen voortdurend in aanraking met toepassingen van de wiskunde. Het is levendig en afwisselend en confronteert de leerling tenslotte in het laatste hoofdstuk met een tot de verbeelding sprekend stuk toegepaste wiskunde. In die zin voldoet het aan Seidels criterium dat "van het hulpmiddel vooral de levende kant op de voorgrond dient te komen".

Toch is het uiteindelijk behandelde wiskundige apparaat vrij gering. Bij de wijze waarop de Leslie-matrices worden geïntroduceerd en de manier waarop ze worden gebruikt kan men vraagtekens plaatsen. De voorkennis van de leerlingen reikt niet verder dan eenvoudige matrixvermenigvuldiging, terwijl de wiskun-

dige achtergrond van het Leslie-model het V.W.O. programma ver te boven gaat. Men zou dit didactisch dilemma voor een deel kunnen ondervangen door de leerlingen te confronteren met enkele stereotype Leslie-matrices, zeg enkele 3×3 en 4×4 matrices, die de meest voorkomende gevallen illustreren. Het kan de leerlingen dan duidelijk zijn dat de bij de verschillende jachtstrategieën op de klapmutsen behorende Leslie-matrices en de daarbij berekende bevolkingsvectoren na 10 jaren (zoals in het laatste hoofdstuk van het boek Matrices gebeurt) onvoldoende zijn om de merites van de verschillende strategieën te beoordelen.

Het geven van een niet-triviale toepassing van de wiskunde in een leerlingenboek impliceert een confrontatie met een wetenschappelijke attitude. Voor aanstaande studenten van het wetenschappelijk onderwijs is dat geen luxe, maar behoort het een onontbeerlijke doelstelling van het onderwijs te zijn.

Men zal m.b.t. het Leslie-model niet zover kunnen gaan als de Hewet-werkgroep aanbeveelt nl. dat de leerling in staat moet zijn de relevantie van het model te kunnen beoordelen. Dit hoeft echter geen reden te zijn om dit onderwerp uit het leerlingenboek weg te laten.

In hoeverre het nieuwe programma uitvoerbaar is zal voor een groot deel afhangen van de mate waarin de docenten zich door de nieuwe ideeën in het programma voelen aangesproken. Van hen wordt een grote creativiteit en flexibiliteit vereist bij de verwezenlijking ervan.

LITERATUUR

[1] de Bruijn, N.G., Modernisering leerplan wiskunde Euclides, 43, 1967/68, 260-262.

[2] Freudenthal, H., Modernisering leerplan wiskunde Euclides, 43, 1967/68, 321-322.

[3] Gantmacher, F.R., Applications of the theory of matrices, Interscience Publ. Incl. N.Y. 1959, Ch. III.

[4] de Lange Jzn, J., Matrices van begin tot eind, Nieuwe Wiskrant 2e jaargang 1983, no. 3, 12-18.

[5] de Lange Jzn, J. en Kindt, M., Matrices, 3e herz. versie, Utrecht 1983.

[6] de Lange Jzn, J. en Vonk, G.A., Kapmutsen in gevaar? Nieuwe Wiskrant 2e jaargang 1982, no. 2, 44-52.

[7] Leslie, P.H., One the use of matrices in certain population mathematics. Biometrika vol. XXXIII 1945, 183-212.

[8] Rapport van de werkgroep van advies voor de herverkaveling eindexamenprogramma's wiskunde I en wiskunde II V.W.O., Staatsuitgeverij 's-Gravenhage, febr. 1980.

[9] Seidel, J.J., De betekenis van het leerplan voor de toekomstige student. Euclides, 31, 1955/56, 245-256.

[10] Seidel, J.J., Wiskunde en Technisch Hoger Onderwijs, Simon Stevin, 32e jaargang, 1958, no. 4, 145-158.

[11] Seneta, E., Non-negative matrices, George Allen & Unwin Ltd. London, 1973, Ch. 1.

[12] Sykes, Z.H., On discrete stable population theory. Biometrics vol. 25, 1969, 285-293.

ON NUMERICAL STABILITY

by

A.J. Geurts

Dedicated to J.J. Seidel on the occasion of his 65th birthday.

1. INTRODUCTION

1.1. Shortly after the beginning of the Eindhoven University of Technology in 1957 a course in numerical computations was set up by staff members of the Mathematics Department. The director of this Course for Scientific Arithmetician (Cursus Wetenschappelijk Rekenaar), as it was named, was Prof. J.J. Seidel.

In those days computations were performed on a mechanical or electrical desk-calculator. The arithmetician doing the calculations to solve for instance a 10×10 system of linear equations could watch the result of every step of the algorithm, could change the accuracy by taking more or less decimals into account, could do a part of the calculations over again if by loss of significant digits too much accuracy had been lost. He thus built up skill and intuition to judge the reliability of the computed solution.

With the advent of the computer things have changed. Desk-calculators have been replaced by personal computers, terminals or at least programmable pocket-calculators. Even if one would like to, it is not possible to follow the execution of an algorithm by the computer step by step. One only sees the result and to judge the acceptability of this result other expedients are necessary. This paper is concerned with one of the concepts that can be used to validate the results of an algorithm.

1.2. In modern computers calculations are performed in floating-point arithmetic. This means that with the execution of an algorithm the input of data of the problem as well as the arithmetic operations are subject to errors of a particular kind, namely relative errors of at most a machine-dependent quantity, named the machine-precision. It is therefore quite natural that the user is satisfied if an algorithm produces an answer that agrees with the exact solution of the problem with data that differ within a modest multiple of the machine-precision from the given data. And with respect to this agreement also a difference within machine-precision is quite acceptable. It is even unrealistic to demand for more.

One must be well aware that in this consideration nothing is said about the accuracy of an acceptable solution. It is only in a sense the best possible solution, but its accuracy may be very poor if the problem is very sensitive to variations in the data.

The idea of regarding a computed solution as satisfactory if it is "a slightly wrong solution to a slightly wrong problem" is due to Kahan [3]. An algorithm that supplies such solutions is called numerically stable. The amount of numerical stability can be expressed by a stability number of which a definition is given in this paper. Some elementary results and simple examples are given to illustrate the concept. We mention that Larson [4,5] and Miller [6] have defined similar quantities and have developed software for the calculation of these quantities in more general, i.e. nontrivial, cases based on the computational graph of an algorithm as defined by Bauer [1].

2. A STABILITY NUMBER

2.1. A numerical problem is a computable mapping, φ say, of a finite dimensional space $\mathcal{D}$, the input or data space, into a finite dimensional space X, the output or solution space; in formula

$$x = \varphi(d) \ , \tag{2.1}$$

in which $d \in \mathcal{D} \subset \mathbb{R}^m$ represents the given data and $x \in X \subset \mathbb{R}^n$ the corresponding solution.

We suppose that the norms we use in $\mathbb{R}^m$ and $\mathbb{R}^n$ are monotonic, that $\mathcal{D}$ is an open subset of $\mathbb{R}^m$ and that φ is at least a continuous function.

In this paper we only consider problems φ for which there exists an algorithm consisting of a finite number of elementary operations through which $\varphi(d)$ can be computed exactly for any $d \in \mathcal{D}$. Thus we exclude algorithms based on an infinite process, for instance an iterative process such as the Gauss-Seidel iterative method for solving linear equations.

In our considerations algorithms will be performed in floating-point arithmetic with machine-precision η. This means that the representation of the data and the performance of the arithmetic operations $(+,-,\times,/)$ are subject to a relative error of at most η, i.e., if $\bar{x}$ is the result of the representation or computation by an elementary operation of the real value x, then

$$\bar{x} = x(1+\varepsilon)^{\pm 1} , \qquad |\varepsilon| \leq \eta . \tag{2.2}$$

2.2. Let Φ be an algorithm for φ and let N be the number of operations in Φ where a rounding error of the form (2.2) may be evoked. Then the result of the execution of Φ for a given $d \in \mathcal{D}$ may be denoted by $\Phi(d,\varepsilon)$, where $\varepsilon \in \mathbb{R}^N$ is the vector of the rounding errors actually made.

DEFINITION 2.1. An algorithm for the Problem (2.1) is a mapping $\Phi : \mathcal{D} \times \mathbb{R}^N \to \mathbb{R}^n$ for which

$$\Phi(d,0) = \varphi(d) \tag{2.3}$$

holds for all $d \in \mathcal{D}$.

Let $x := \varphi(d)$ and $\bar{x}(\varepsilon) := \Phi(d,\varepsilon)$ for given $d \in \mathcal{D}$ and $\varepsilon \in \mathbb{R}^N$. Let Δd and Δx be such that $d + \Delta d \in \mathcal{D}$ and

$$\bar{x}(\varepsilon) = \varphi(d+\Delta d) + \Delta x \ . \tag{2.4}$$

Such a pair $(\Delta d, \Delta x)$ will be called a suitable pair. (2.4) expresses that Δd and Δx are perturbations of the input and output, respectively, which account for the same effect as the rounding errors. By Δd a perturbated problem is considered and Δx expresses how near the computed $\bar{x}(\varepsilon)$ is to the exact solution of this perturbed problem. The minimal relative perturbation in both input and output in order to meet $\bar{x}(\varepsilon)$ can be given by

$$r(\Phi,d,\varepsilon) := \inf \left(\max \left(\frac{\|\Delta d\|}{\|d\|} , \frac{\|\Delta x\|}{\|x\|} \right) \right) \tag{2.5}$$

where the infimum is taken over all suitable pairs. This minimal perturbation expressed in terms of the machine-precision η, will be used to express the numerical behaviour of Φ. We therefore define the following quantity.

DEFINITION 2.2. The stability number $S(\Phi,d)$ of the algorithm Φ for the problem φ in the point d is

$$S(\Phi,d) := \lim_{\eta \to 0} \sup_{0 < \|\varepsilon\|_\infty \leq \eta} (r(\Phi,d,\varepsilon)/\|\varepsilon\|_\infty) \ . \tag{2.6}$$

It is easy to see that if both d and x are non-zero and Φ is Lipschitz continuous with respect to ε, then $S(\Phi,d)$ is finite.

LEMMA 2.1. For sufficiently small ε the infimum in (2.5) is a minimum and every suitable pair $(\Delta d_0, \Delta x_0)$ for which the minimum is attained, satisfies the relations

$$\frac{\|\Delta d_0\|}{\|d\|} \leq r(\Phi,d,\varepsilon) \quad , \quad \frac{\|\Delta x_0\|}{\|x\|} = r(\Phi,d,\varepsilon) \ . \tag{2.7}$$

If, in addition, φ has property C[*)] in a neighbourhood of d, then equality applies to Δd_0 as well.

Proof. Since the infimum in (2.5) can be found in the sphere $\|\Delta d\| \leq \|d\| \, \|\bar{x}(\varepsilon) - x\| / \|x\|$, it is obvious that, if ε is sufficiently small, it is a minimum.

Suppose that $\|\Delta x_0\| / \|x\| < \|\Delta d_0\| / \|d\| = r(\Phi,d,\varepsilon)$. Define the perturbations $\Delta d_1 := (1-\vartheta)\Delta d_0$ and $\Delta x_1 := \Delta x_0 + \varphi(d+\Delta d_0) - \varphi(d+\Delta d_1)$. If ϑ is sufficiently small and positive, then Δx_1 is defined and $(\Delta d_1, \Delta x_1)$ is a suitable pair satisfying $\max(\|\Delta d_1\| / \|d\|, \|\Delta x_1\| / \|x\|) < r(\Phi,d,\varepsilon)$, in contradiction with the definition of $r(\Phi,d,\varepsilon)$. This proves (2.7).

Now suppose that φ has property C in $d + \Delta d_0$ and that $\|\Delta d_0\| / \|d\| < \|\Delta x_0\| / \|x\| = r(\Phi,d,\varepsilon)$. Define $\Delta x_1 := (1-\vartheta)\Delta x_0$ and Δd_1 such that $\varphi(d+\Delta d_1) = \varphi(d+\Delta d_0) + \vartheta \Delta x_0$. Property C implies that such a Δd_1 exists if ϑ is small enough and, if $\vartheta \to 0$, then Δd_1 can be chosen such that $\Delta d_1 \to \Delta d_0$. The pair $(\Delta d_1, \Delta x_1)$ now yields a similar contradiction as before. □

---

*) The definition of this property is given in an appendix.

COROLLARY. If Φ is performed in floating-point arithmetic with accuracy η, then there exists a suitable pair $(\Delta d_0, \Delta x_0)$ such that, in first order approximation,

$$\frac{\|\Delta d_0\|}{\|d\|} \leq S(\Phi,d)\eta \quad , \quad \frac{\|\Delta x_0\|}{\|x\|} \leq S(\Phi,d)\eta \ . \tag{2.8}$$

If, in addition, the supremum in (2.6) is attained for some ε with $\|\varepsilon\|_\infty = \eta$, which generally will be the case, and if φ has property C in a neighbourhood of d, then (2.8) is best possible in the sense that there is a $\Phi(d,\varepsilon)$ for which, in first order approximation, equality holds in both formulae for some $(\Delta d_0, \Delta x_0)$.

The following example shows that equality is not always possible in the first inequality of (2.7) and (2.8).

EXAMPLE. Let $\varphi : (0,1) \to \mathbb{R}^2$ be defined by $\varphi_1(d) = 1 + d^2$, $\varphi_2(d) = 1 - d^2$. Let $\Phi : (0,1) \times \mathbb{R}^3 \to \mathbb{R}^2$ be defined by $\bar{x}_1(\varepsilon) := \Phi_1(d,\varepsilon) = (1 + d^2(1+\varepsilon_1))(1+\varepsilon_2)$, $\bar{x}_2(\varepsilon) := \Phi_2(d,\varepsilon) = (1 - d^2(1+\varepsilon_1))(1+\varepsilon_3)$. Let the norm in $\mathbb{R}^2$ be the norm defined by (2.12) with $x := \varphi(d)$ as the given point. Then from (2.4) we see that, in first order approximation,

$$\begin{aligned} 2d\,\Delta d + \Delta x_1 &= \varepsilon_1 d^2 + \varepsilon_2(1+d^2) \\ -2d\,\Delta d + \Delta x_2 &= -\varepsilon_1 d^2 + \varepsilon_3(1-d^2) \ . \end{aligned}$$

Now it can be verified that the supremum in (2.6) is attained for $\varepsilon_1 = \varepsilon_2 = \varepsilon_3 = \eta$, that $S(\Phi,d) = 1$ and that the optimal pair $(\Delta d_0, \Delta x_0)$ satisfies $(\Delta x_0)_i / x_i = \eta$, $i = 1,2$, and $\Delta d_0 / d = \frac{1}{2}\eta$.

A small stability number does not imply that the computed solution is near the exact solution, since a small perturbation of the data may cause a large variation of the solution. It is easy to prove the following inequality

$$\frac{\|\bar{x}(\varepsilon) - x\|}{\|x\|} \leq (c(\varphi,d) + 1 + o(1))\, S(\Phi,d)\, \|\varepsilon\|_\infty , \quad \varepsilon \to 0 , \tag{2.9}$$

where $c(\varphi,d)$ is the condition number of φ in d which, roughly speaking, is the ratio between the magnitudes of a perturbation in the data and the ensuing perturbation in the solution (cf. Geurts [2]). In most cases (2.9) gives an overestimate of the error in $\bar{x}(\varepsilon)$. However, if the solution is a scalar, then (2.9) is best possible in the following sense.

LEMMA 2.2. Let $X = \mathbb{R}$ and let φ be a continuous functional on $\mathcal{D}$ and differentiable in $d \in \mathcal{D}$. Then

$$\sup_{0<\|\varepsilon\|_\infty \leq \eta} \left(\frac{|\bar{x}(\varepsilon) - x|}{|x|} \Big/ \|\varepsilon\|_\infty \right) = (c(\varphi,d) + 1 + o(1))\, S(\Phi,d) , \quad \eta \to 0 . \tag{2.10}$$

Proof. To prove that

$$\frac{|\bar{x}(\varepsilon) - x|}{|x|} \leq (c(\varphi,d) + 1 + o(1))\, r(\Phi,d,\varepsilon) , \quad \varepsilon \to 0 , \tag{2.11}$$

is easy and left to the reader. We shall prove that the inequality also holds the other way round.

It follows directly from (2.4) that for an arbitrary suitable pair $(\Delta d, \Delta x)$

$$\bar{x}(\varepsilon) - x = \varphi'(d) \cdot \Delta d + o(\Delta d) + \Delta x , \quad \Delta d \to 0 .$$

Now let Δd_1 be such that $\|\Delta d_1\| / \|d\| = r(\Phi,d,\varepsilon)$ and $\operatorname{sign}(\varphi'(d) \cdot \Delta d_1) = \operatorname{sign}(\bar{x}(\varepsilon) - x)$. Then, if η and consequently Δd_1 is small enough, it turns out that the matching Δx_1 satisfies $\operatorname{sign}(\Delta x_1) = \operatorname{sign}(\bar{x}(\varepsilon) - x)$. Consequently,

$$|\bar{x}(\varepsilon) - x| = |\varphi'(d) \cdot \Delta d_1| + |\Delta x_1| + o(\Delta d_1) \ , \quad \Delta d_1 \to 0 \ .$$

If we next require Δd_1 to be a maximizing vector of $\varphi'(d)$, then, since $c(\varphi,d) = \|\varphi'(d)\| \, \|d\| / |x|$ and obviously $|\Delta x_1| / |x| \geq r(\Phi,d,\varepsilon)$,

$$|\bar{x}(\varepsilon) - x| = \|\varphi'(d)\| \, \|\Delta d_1\| + |\Delta x_1| + o(\Delta d_1) \geq$$

$$\geq (c(\varphi,d) + 1 + o(1)) \, r(\Phi,d,\varepsilon) \, |x| \ , \quad \varepsilon \to 0 \ .$$

From this inequality we conclude that equality holds in (2.11) from which (2.10) immediately follows.

2.3. A norm in $\mathbb{R}^m$ well suited to floating-point arithmetic is the following one (cf. Geurts [2]).

Let $a \in \mathbb{R}^m$, $a \neq 0$, be a given point, then for $u \in \mathbb{R}^m$

$$\|u\| := \min(\sigma \mid |u_k| \leq \sigma |a_k| \ , \ 1 \leq k \leq m) \ . \tag{2.12}$$

It is allowed that $a_k = 0$ and $u_k \neq 0$ for some k. In that case $\|u\| = \infty$ is taken.

If this norm is used with d as the given point the condition number of a functional φ at d turns out to be $c(\varphi,d) = |\varphi'(d)| \cdot |d| / |x|$. With the help

of (2.10) we then find for the stability number the following formula

$$S(\Phi,d) = \frac{\rho(\Phi,d)}{|\varphi'(d)| \cdot |d| + |x|} \tag{2.13}$$

where $\rho(\Phi,d) := \lim_{\eta \to 0} \sup_{0<\|\varepsilon\|_\infty \le \eta} (|\bar{x}(\varepsilon) - x| / \|\varepsilon\|_\infty)$.

EXAMPLE. Let us consider the evaluation of the polynomial

$$p(a,t) := \sum_{i=0}^{k} a_i t^i ,$$

by the well-known Horner algorithm performed in floating-point arithmetic. Then the following quantities*) are computed (cf. Peters and Wilkinson [7]).

$$\begin{aligned} \bar{b}_k &= a_k \\ \bar{b}_i &= (a_i + t\bar{b}_{i+1}(1+\varepsilon_{2i+1}))(1+\varepsilon_{2i})^{-1} , \quad i = k-1,\ldots,0 , \end{aligned} \tag{2.14}$$

and the computed approximation of $x := p(a,t)$ is given by $\bar{x}(\varepsilon) = \bar{b}_0$. By elementary calculations we find the following relation

$$\bar{x}(\varepsilon) = x + \sum_{i=0}^{k} \bar{b}_i t^i (\varepsilon_{2i-1} - \varepsilon_{2i}) ,$$

if we assume the introduced ε_{-1} and ε_{2k} to be zero.

Now we regard only the coefficients a_i, $0 \le i \le k$, as input parameters for the error analysis. Then the stability number based on the norm (2.12) reads,

*) A bar over a quantity denotes the computed value of that quantity.

according to (2.13),

$$(2.15)\qquad S(\Phi,a) = \frac{2\sum_{i=0}^{k}{}'' |b_i t^i|}{\sum_{i=0}^{k} |a_i t^i| + \left|\sum_{i=0}^{k} a_i t^i\right|} = \frac{2\sum_{i=0}^{k}{}'' \left|\sum_{j=i}^{k} a_j t^j\right|}{\sum_{i=0}^{k} |a_i t^i| + \left|\sum_{i=0}^{k} a_i t^i\right|} ,$$

in which $\sum''$ denotes in the usual way that the first and the last term in the summation are to be halved. It is easy to see that $S(\Phi,a) \le 2k$. We want to show that $S(\Phi,a) \le k$ and that this upper bound is best possible.

From (2.14), written in the form

$$a_k = \bar{b}_k$$

$$a_i = \bar{b}_i(1+\varepsilon_{2i}) - t b_{i+1}(1+\varepsilon_{2i+1}) , \quad i = k-1,\ldots,0 ,$$

it follows that

$$\tilde{a}_i := a_i \prod_{j=0}^{i-1} (1+\varepsilon_{2j+1}) \prod_{j=i+1}^{k-1} (1+\varepsilon_{2j}) \qquad 0 \le i \le k$$

$$\tilde{b}_i := \bar{b}_i \prod_{j=0}^{i-1} (1+\varepsilon_{2j+1}) \prod_{j=i}^{k-1} (1+\varepsilon_{2j})$$

satisfy the relations of the Horner scheme, viz.,

$$\tilde{b}_k = \tilde{a}_k$$

$$\tilde{b}_i = \tilde{a}_i + t\tilde{b}_{i+1} , \quad i = k-1,\ldots,0 ,$$

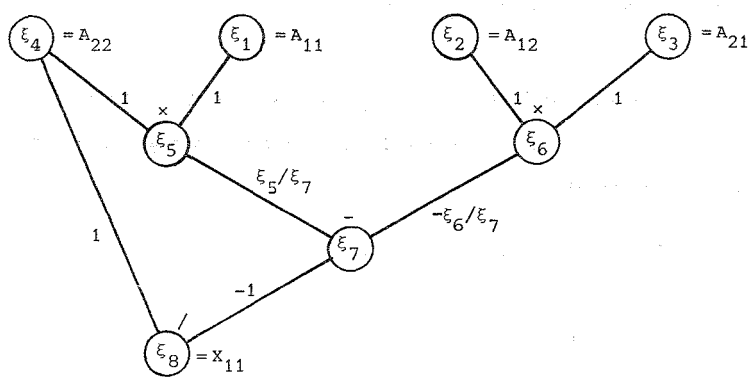
and consequently that $\tilde{b}_0 = p(\tilde{a},t)$. From this it follows that $\Delta a := \tilde{a} - a$ en $\Delta x := \bar{b}_0 - \tilde{b}_0$ is a suitable pair. This pair satisfies the inequalities $\|\Delta a\| / \|a\| \leq k \|\varepsilon\|_\infty + o(\varepsilon)$, $|\Delta x|/|x| \leq k \|\varepsilon\|_\infty + o(\varepsilon)$. So $S(\Phi,a) \leq k$ and since $S(\Phi,a) = k$ if $a_k \neq 0$ and $a_i = 0$, $0 \leq i \leq k-1$, the upper bound k is best possible.

REMARK. The inequality $S(\Phi,a) \leq k$ can also directly be derived from (2.15) by elementary calculations.

3. COMPUTATION OF THE STABILITY NUMBER

The error analysis of a non-trivial numerical algorithm, in particular the computation of a stability number is an extensive job. Among others, Larson [4,5] and Miller [6] have developed software to do the analysis by the computer. We shall give a brief sketch of the work of Larson.

3.1. The method is based on the computational graph of an algorithm as introduced by Bauer [1]. For instance, let A be a 2×2 regular matrix and let $X := A^{-1}$. Then $X_{11} = A_{22}/(A_{11}A_{22} - A_{12}A_{21})$ and the computational graph of X_{11} computed by this formula is (cf. [1, Fig. 3])



The nodes ξ_i, i = 1,2,3,4, denote the data, ξ_i, i = 5,6,7, are intermediate results in the algorithm and ξ_8 is the solution. In each node ξ_i, i ≥ 5, a rounding error ε_i may be evoked. The arcs of the graph are weighted, the weights being the propagation factors with which an error in a node is passed to the next node. The weight of the arc from ξ_ℓ to ξ_k is the partial relative derivative

$$\frac{\rho\xi_k}{\rho\xi_\ell} := \frac{\xi_\ell}{\xi_k}\,\frac{\partial\xi_k}{\partial\xi_\ell} .$$

The error-propagation factor from one node of the graph to another node obeys a computational rule similar to the chain rule for differentiation.

In order to survey the numerical stability of an algorithm, in the sense explained in Section 2, the evoked errors must be allocated among the input and the output variables. It is easy to understand that the rounding error generated at a particular node must be moved backward (to the input) if there is a large propagation factor from an input node to that node and forward if there is a small propagation factor from that node to an output node. And with moderate factors either direction may be appropriate. From the computational graph of X_{11} we can conclude immediately that the given formula has a stability number of at most 1.

3.2. For a realistic problem the computational graph is very extensive. In order to handle such a problem Larson derives from the graph an undetermined system of linear equations

$$e_x = A\,e_d + B\,\varepsilon \ , \tag{3.1}$$

where e_x is an n-vector of relative perturbations of the solution, e_d an m-vector of relative perturbations of the data and ε the N-vector of the evoked rounding errors. The elements of A and B are partial relative derivatives.

The computation of $S(\Phi,d)$ based on the norm (2.12) then implies the determination of the minimum of the quantities ϑ such that for all ε, with $\|\varepsilon\|_\infty \leq \eta$, there are e_x and e_d, both with maximum norm $\leq \vartheta$, satisfying (3.1). This problem is essentially a large-scale minimax problem. The main concern in the development of the software by Larson, cs. (cf. [5]), has been the reduction of the computational work and the search for acceptable approximation methods to solve this minimax problem.

REMARK. In the case of a scalar function (3.1) consists of just one equation. It is easy to find out that the solution of the minimax problem then equals (2.13).

APPENDIX

The mapping $\varphi : \mathcal{D} \subset \mathbb{R}^m \to \mathbb{R}^n$ is said to have property C in the point $d \in \mathcal{D}$ if there exists in $\mathbb{R}^n$ a full neighbourhood Ω_d of $\varphi(d)$, such that the following conditions hold.

1. $\Omega_d \subset \varphi(\mathcal{D})$;
2. Let $\{x_j\}$ be a sequence with $x_j \in \Omega_d$ for all j and $\lim x_j = \varphi(d)$. Then there exists a sequence $\{d_j\}$ in $\mathcal{D}$ with $x_j = \varphi(d_j)$ for all j and $\lim d_j = d$.

The mapping φ has property C in $V \subset \mathcal{D}$ if φ has property C in each point $d \in V$.

REMARK. If φ is a differentiable mapping with continuous first-order partial derivatives in $\mathcal{D}$, then φ has property C in each point in which $\operatorname{rank}(\varphi'(d)) = n$.

REFERENCES

[1] Bauer, F.L., Computational Graphs and Rounding Error. Siam J. Numer. Anal. 11, 87 - 96 (1974).

[2] Geurts, A.J., A Contribution to the Theory of Condition. Num. Math. 39, 85 - 96 (1982).

[3] Kahan, W., A Survey of Error Analysis, IFIP Congress 71, Ljubljana, 200 - 206 (1971).

[4] Larson, J.L., Methods for Automatic Error Analysis of Numerical Algorithms. Ph.D. dissertation, University of Illinois at Urbana-Champaign, 1978.

[5] Larson, J.L., M.E. Pasternak and J.A. Wisniewski, Algorithm 594, Software for Relative Error Analysis, ACM Trans. Math. Software 9, 125 - 130 (1983).

[6] Miller, W. and C. Wrathall, Software for Roundoff Analysis of Matrix Algorithms, Computer Science and Applied Mathematics, New York-London-Toronto-Sydney-San Francisco. Academic Press, 1980.

[7] Peters, G. and J.H. Wilkinson, Practical Problems Arising in the Solution of Polynomial Equations, J. Inst. Math. Appl. 8, 16 - 35 (1971).

GENERALIZED FUNCTIONS AND OPERATORS ON THE UNIT SPHERE

by

J. de Graaf *)

Dedicated to Professor J.J. Seidel on the occasion of his retirement.

Summary. Two spaces of generalized functions on the unit sphere $\Omega^{q-1} \subset \mathbb{R}^q$ are introduced. Both types of generalized functions can be identified with suitable classes of harmonic functions. Several natural classes of continuous and continuously extendible operators are discussed: Multipliers, differentiations, harmonic contractions/expansions and harmonic shifts. The latter two classes of operators are "parametrized" by the full affine semi-group on $\mathbb{R}^n$.

AMS Classifications: 46F05, 46F10, 31B05, 20G05.

*) Department of Mathematics and Computing Science, Eindhoven University of Technology.

1. INTRODUCTION AND NOTATIONS

In this paper I describe two natural theories of generalized functions on the unit sphere Ω^{q-1} in $\mathbb{R}^q$ and some natural classes of linear operators acting on those generalized functions. The test functions in both theories are restrictions to Ω^{q-1} of suitable classes of harmonic functions on open sets in $\mathbb{R}^q$. Also the generalized functions in both theories are explicitly represented by classes of harmonic functions in $\mathbb{R}^q$. The generalized functions appear to be "boundary values" of harmonic functions.

Both theories are much stronger (i.e. they contain more distributions) than Schwartz' theory.

The theory in this paper resembles hyper-function theory. There are fundamental differences however: Harmonic functions in n variables satisfy one PDE, whereas analytic functions in n variables satisfy the overdetermined system of Cauchy-Riemann equations. Further, the product of two harmonic functions is usually not harmonic, while the product of two analytic functions is always analytic, etc.

The theories we introduce here are very special concrete cases of the general functional analytic constructions in $[G_1]$, $[G_2]$, $[G_3]$, [E].

The classes of operators that we introduce are based on simple geometric considerations and on the properties of harmonic functions as derived in Section 2. For example a continuous linear operator is associated with each element of the full affine semi-group on $\mathbb{R}^q$. In the Hilbert space $L_2(\Omega^{q-1})$

these operators are (strongly) unbounded in general. The precise "representation properties" of these operators are not yet clear to me!

In the sequel I will stick to the following notations and conventions. For theory and proofs I refer to [S], [M].

$\Omega^{q-1}(\underline{a};R)$, sphere with centre $\underline{a}$ and radius R in $\mathbb{R}^q$.

$\Omega^{q-1} = \Omega^{q-1}(\underline{0};1)$, the unit sphere.

$\underline{\xi}$, $\underline{\eta}$, points on Ω^{q-1}.

$\underline{x} = r\underline{\xi}$, $\underline{y} = R\underline{\eta}$, vectors in $\mathbb{R}^q$.

$B^q(\underline{a};R)$, open ball with centre $\underline{a}$ and radius R in $\mathbb{R}^q$.

$B^q = B^q(\underline{0};1)$, the open unit ball.

$dx_1 \cdot \ldots \cdot dx_q$, the usual Lebesgue measure in $\mathbb{R}^q$.

$d\omega_q$, the usual $(q-1)$-dimensional surface measure on Ω^{q-1}.

$\omega_q = 2\pi^{q/2}(\Gamma(\frac{q}{2}))^{-1}$, the total surface measure of Ω^{q-1}.

$\int$, integrations take place over Ω^{q-1} if not indicated otherwise.

$\mathrm{Harm}(W)$, harmonic functions on $W \subset \mathbb{R}^q$, i.e. functions φ which satisfy $\frac{\partial^2\varphi}{\partial x_1^2} + \ldots + \frac{\partial^2\varphi}{\partial x_q^2} = 0$ on W.

$\mathrm{Harm}(B^q(\underline{a};R))$, the vector space of harmonic functions on the open ball $B^q(\underline{a};R)$.

$\overline{\mathrm{Harm}(B^q(\underline{a};R))} = \bigcup_{r>R} \mathrm{Harm}(B^q(\underline{a};r))$, the vector space of functions which are harmonic on an open neighbourhood of the closed ball $\overline{B^q(\underline{a};R)}$.

$\mathrm{Harm}(\mathbb{R}^q)$, the vector space of all harmonic functions on $\mathbb{R}^q$.

$\mathrm{Harm}(\underline{0}) = \bigcup_{r>0} \mathrm{Harm}(B^q(\underline{0};r))$, the vector space of functions which are defined and harmonic on an open neighbourhood of $\underline{0}$. This neighbourhood may depend on the function. ("Harmonic germs".)

$\mathrm{HHP}(q;n)$, the vector space of harmonic homogeneous polynomials of degree n in q variables.

$N(q,n) = \dim \mathrm{HHP}(q;n)$, we have, see [M], $N(q,n) \leq K_q n^{q-2}$, K_q is a constant.

$S_n(\underline{\xi})$, $S_{n,f}(\underline{\xi})$, spherical harmonics, i.e. restrictions of elements in $\mathrm{HHP}(q;n)$ to Ω^{q-1}.

$L_2(\Omega^{q-1})$, the complex Hilbert space of square integrable functions Ω^{q-1}.

$\|\cdot\|$, $(\cdot,\cdot)$, norm and inner product on $L_2(\Omega^{q-1})$

$$(f,g) = (f(\cdot),g(\cdot)) = \int f(\underline{\xi})\,\overline{g}(\underline{\xi})\,d\omega_q$$

$$\|f\|^2 = (f,f) .$$

The restriction of an arbitrary element in $\mathrm{HHP}(q;n)$ to Ω^{q-1} is orthogonal to the restriction of an arbitrary element in $\mathrm{HHP}(q;m)$ to Ω^{q-1} if $m \neq n$. The mentioned restrictions of $\mathrm{HHP}(q;n)$, $n = 0,1,2,\ldots$, establish a complete set in $L_2(\Omega^{q-1})$. We do not introduce a special orthonormal basis in $L_2(\Omega^{q-1})$. The restriction to Ω^{q-1} of any polynomial of degree m in q variables is a finite linear combination of restrictions of elements in $\mathrm{HHP}(q;n)$ with $0 \leq n \leq m$.

P_n denotes the orthogonal projection of $L_2(\Omega^{q-1})$ onto HHP(q,n). Often we write $(P_n f)(\underline{\xi}) = S_{n,f}(\underline{\xi})$.

From [M] we quote the estimate

$$|S_n(\underline{x})| \leq |\underline{x}|^n \left(\frac{N(q,n)}{\omega_q}\right)^{\frac{1}{2}} \|S_n\| , \tag{1.1}$$

with

$$\|S_n\| = \left\{\int |S_n(\underline{\xi})|^2 \, d\omega_q\right\}^{\frac{1}{2}} ,$$

for any $S_n \in HHP(q;n)$.

2. SOME LEMMAS ON HARMONIC FUNCTIONS

Let $f \in L_2(\Omega^{q-1})$. Decompose f in spherical harmonics $f(\underline{\xi}) = \sum_{n=0}^{\infty} S_{n,f}(\underline{\xi})$. In the first lemma we give conditions on f such that it can be extended to an harmonic function on $B^q(\underline{0};R)$ for some $R > 1$. The extension is again denoted by f.

LEMMA 2.1.

(i) $f \in L_2(\Omega^{q-1})$ can be extended to an element in $\mathrm{Harm}(B^q(\underline{0};R)$, $R > 1$, iff

$$\sum_{n=0}^{\infty} r^{2n} \|S_{n,f}\|^2 < \infty \qquad \text{for all } r,\ 0 \leq r < R .$$

(ii) If $f \in \mathrm{Harm}(B^q(\underline{0};R))$ then the sequence $\sum_{n=0}^{\infty} r^n S_{n,f}(\underline{\xi})$ converges uniformly to f on each ball $B^q(\underline{0};R_1)$ with $R_1 < R$.

Proof.

(i) $\Leftarrow$ With (1.1) and $r \le R_1 < R_2 < R$ we estimate

$$|r_n S_{n,f}(\underline{\xi})| = |S_{n,f}(r\underline{\xi})| \le \left(\frac{R_1}{R_2}\right)^n \left(\frac{N(q,n)}{\omega_q}\right)^{\frac{1}{2}} R_2^n \|S_{n,f}\| \le$$

$$\le \frac{K_q}{2\omega_q}\left(\frac{R_1}{R_2}\right)^{2n} n^{q-2} + \frac{1}{2}R_2^{2n}\|S_{n,f}\|^2 .$$

Both are terms of a converging sequence. Hence $\sum_{n=0}^{\infty} r_n S_{n,f}(\underline{\xi})$ converges uniformly on each $B^q(\underline{0};R_1)$ and therefore belongs to $\mathrm{Harm}(B^q(\underline{0};R))$.

(i) $\Rightarrow$ Suppose $f \in \mathrm{Harm}(B^q(\underline{0};R))$. For each $r < R$ we have $f(r\underline{\xi}) \in L_2(\Omega^{q-1})$. Hence $\sum_{n=0}^{\infty} r^{2n}\|S_{n,f}\|^2 < \infty$.

(ii) See part (i). □

If f and g belong to $\mathrm{Harm}(B^q(\underline{0};R))$ the product $f \cdot g$ is usually not harmonic. For this reason the following lemma is not a trivial result.

LEMMA 2.2. Let $f,g \in \mathrm{Harm}(B^q(\underline{0};R))$, $R > 1$. The restriction of the pointwise product $f \cdot g$ to Ω^{q-1} can be extended to a harmonic function in $\mathrm{Harm}(B^q(\underline{0};R))$. We will call this product the harmonic product of f and g.

Proof. Write

$$f(\underline{\xi}) = \sum_{n=0}^{\infty} S_{n,f}(\underline{\xi}) \quad , \quad g(\underline{\xi}) = \sum_{m=0}^{\infty} S_{m,g}(\underline{\xi}) .$$

In case of absolute convergence we can write

$$f(\underline{\xi})\, g(\underline{\xi}) = \sum_{\ell=0}^{\infty} \sum_{m+n=\ell} S_{n,f}(\underline{\xi})\, S_{m,g}(\underline{\xi}) . \tag{2.1}$$

Let $1 < R_1 < R$. Uniform convergence of (2.1) on Ω^{q-1} follows from the estimate

$$|S_{0,f}(\underline{\xi})\, S_{\ell,g}(\underline{\xi})| + |S_{1,f}(\underline{\xi})\, S_{\ell-1,g}(\underline{\xi})| + \ldots + |S_{\ell,f}(\underline{\xi})\, S_{0,g}(\underline{\xi})| \leq$$

$$\leq \frac{K_q}{2\omega_2}\, \ell^{q-2} R_1^{-\ell} \left\{ \sum_{k=0}^{\ell} R_1^{2k} \|S_{k,f}\|^2 + \sum_{k=0}^{\ell} R_1^{2k} \|S_{k,g}\|^2 \right\} \leq$$

$$\leq C_{fg}\, \ell^{q-2} R_1^{-\ell} \ .$$

Here C_{fg} is a constant which only depends on f and g. From the last inequality it also follows that

$$\| \sum_{m+n=\ell} S_{n,f}(\cdot)\, S_{m,g}(\cdot) \| \leq \omega_q^{\frac{1}{2}} C_{fg}\, \ell^{q-2} R_1^{-\ell} \ .$$

Therefore the sequence (2.1) also converges in L_2-sense. Next we estimate the norm of the projection of $f(\underline{\xi}) \cdot g(\underline{\xi})$ on the space of spherical harmonics of degree k.

$$P_k(f \cdot g) = P_k \sum_{\ell=k}^{\infty} \sum_{m+n=\ell} S_{n,f}(\cdot) \cdot S_{m,g}(\cdot) \ .$$

Note that the second sum in the above expression presents a homogeneous (not necessarily harmonic) polynomial of degree ℓ. When restriced to Ω^{q-1} this polynomial can be regarded as the restriction of a harmonic polynomial of degree $\leq \ell$ to Ω^{q-1}. So the projection P_k applied to terms with $\ell < k$

yields zero.

$$\| P_k(f \cdot g) \| \leq \sum_{\ell=k}^{\infty} \| \sum_{m+n=\ell} S_{n,f}(\cdot) \cdot S_{m,g}(\cdot) \| \leq$$

$$\leq \omega_q^{\frac{1}{2}} C_{fg} \sum_{\ell=k}^{\infty} R_1^{-\ell} \ell^{q-2} \leq R_1^{-k} \omega_q^{\frac{1}{2}} C_{fg} \sum_{\ell=k}^{\infty} R_1^{-(\ell-k)} \ell^{q-2} \leq c_1 R_1^{-k}$$

where c_1 does not depend on k.

Hence, for all R_2, $1 < R_2 < R_1 < R$

$$\sum_{k=0}^{\infty} R_2^{2k} \| P_k(f \cdot g) \|^2 < \infty .$$

Now apply Lemma 2.1. □

LEMMA 2.3. Let $f \in \mathrm{Harm}(B^q(\underline{0};R))$, $R > 1$. Let $A : \mathbb{R}^q \to \mathbb{R}^q$ be a linear mapping.

Suppose $\|A\| = R_1 < R$.

Define $g(\underline{\xi}) = f(A\underline{\xi}) \in L_2(\Omega^{q-1})$. g can be extended to a harmonic function in $\mathrm{Harm}(B^q(\underline{0};\frac{R}{R_1}))$.

Proof. Again write $f(\underline{\xi}) = \sum_{n=0}^{\infty} S_{n,f}(\underline{\xi})$. Consider $S_{n,f}(A\underline{x})$. This is a homogeneous polynomial of degree n. With (1.1) it follows

$$|S_{n,f}(A\underline{\xi})| \leq \|A\|^n \left(\frac{N(q,n)}{\omega_q}\right)^{\frac{1}{2}} \|S_{n,f}\| .$$

Hence,

$$\|S_{n,f}(A\cdot)\| \leq R_1^n (\omega_q N(q,n))^{\frac{1}{2}} \|S_{n,f}\| .$$

Since

$$(P_k g)(\underline{\xi}) = P_k \sum_{n=k}^{\infty} S_{n,f}(A\underline{\xi})$$

we have

$$\| P_k g \| \leq (\omega_q N(q,n))^{\frac{1}{2}} \sum_{n=k}^{\infty} R_1^n \| S_{n,f} \| .$$

Let $R_1 < R_2 < R$. Let $1 \leq L < \frac{R_2}{R_1}$, then

$$L^{2k} \| P_k g \|^2 \leq \omega_q \left(\sum_{n=k}^{\infty} L^k R_1^n (N(q,n))^{\frac{1}{2}} \| S_{n,f} \| \right)^2 \leq$$

$$\leq \frac{\omega_q}{2} \left\{ \sum_{n=0}^{\infty} \left(\frac{LR_1}{R_2} \right)^n N(q,n) + \sum_{n=0}^{\infty} R_2^{2n} \| S_{n,f} \|^2 \right\} < \infty .$$

Since this is true for all $L < \frac{R_2}{R_1} < \frac{R}{R_1}$ we conclude

$$\sum_{k=0}^{\infty} L^{2k} \| P_k g \|^2 < \infty \qquad \text{for all } L < \frac{R}{R_1} .$$

3. A METRIZABLE SPACE OF GENERALIZED FUNCTIONS

A theory of generalized functions on Ω^{q-1} is a Gel'fand triple

$$S(\Omega^{q-1}) \subset L_2(\Omega^{q-1}) \subset T(\Omega^{q-1}) .$$

Here $S(\Omega^{q-1})$ is the test space of smooth functions. The space $T(\Omega^{q-1})$ can be regarded as the continuous dual of $S(\Omega^{q-1})$. Moreover, $S(\Omega^{q-1})$ is embedded in $T(\Omega^{q-1})$ via $L_2(\Omega^{q-1})$. In this section we take for the elements of $S(\Omega^{q-1})$ restrictions of functions which belong to $\overline{\text{Harm}(B^q(\underline{0};1))}$. So each $f \in S(\Omega^{q-1})$ can be extended to a function $f \in \text{Harm}(B^q(0;R))$ for some $R > 1$ dependent on f. We will (somewhat loosely) identify $S(\Omega^{q-1})$ and $\overline{\text{Harm}(B^q(\underline{0};1))}$.

DEFINITION 3.1. A sequence $(f_n) \subset S(\Omega^{q-1})$ is said to converge iff $(f_n) \subset \mathrm{Harm}(B^q(\underline{0};R))$, for some $R > 1$, and (f_n) converges uniformly on $B^q(\underline{0};R)$. This is equivalent to saying that $(f_n(R\underline{\xi}))$ converges in $L_2(\Omega^{q-1})$ for some $R > 1$.

For $T(\Omega^{q-1})$ we take $\mathrm{Harm}(B^q(\underline{0};1))$. It "contains" (possibly diverging) series of spherical harmonics $\sum_{n=0}^{\infty} S_{n,f}(\underline{\xi})$ with the property that $\sum_{n=0}^{\infty} r^{2n} \| S_{n,F} \|^2 < \infty$ for all r, $0 < r < 1$.

DEFINITION 3.2. A sequence $(F_n) \subset T(\Omega^{q-1})$ is said to converge iff $(F_n(r\underline{\xi}))$ converges in $L_2(\Omega^{q-1})$ for each $0 < r < 1$.

REMARK 3.3. $S(\Omega^{q-1})$ is a space of type $S_{Y,B}$ and $T(\Omega^{q-1})$ is a space of type $T_{Y,B}$ with $Y = L_2(\Omega^{q-1})$ and $B = -\frac{1}{2}(q-1)I + \{\frac{1}{2}(q-1)^2 I - \Delta_{LB}\}^{\frac{1}{2}}$. Here Δ_{LB} denotes the Laplace-Beltrami operator on the unit sphere Ω^{q-1} and I denotes the identity operator. See [G_1], [G_2], [G_3]. All general considerations of these papers apply here. $S(\Omega^{q-1})$ and $T(\Omega^{q-1})$ are complete nuclear topological vector spaces. $T(\Omega^{q-1})$ is Fréchet (i.e. metrizable). $S(\Omega^{q-1})$ is an inductive limit of Hilbert spaces. A few general functional analytic results are presented here in an *ad hoc* manner.

DEFINITION 3.4. Let $f \in S(\Omega^{q-1})$, $F \in T(\Omega^{q-1})$. The pairing $\langle f,F \rangle$ is defined by

$$\langle f,F \rangle = (f(R\underline{\xi}), F(R^{-1}\underline{\xi})) \quad . \tag{3.1}$$

The inner product makes sense for $R > 1$ sufficiently small. The result does not depend on the choice of R. This can easily be seen by decomposing f and F in spherical harmonics.

It is a trivial exercise to prove that the mappings $f \mapsto \langle f,F\rangle$ and $F \mapsto \langle f,F\rangle$ are sequentially continuous. Moreover, all continuous linear functionals can be represented in the way of (3.1):

THEOREM 3.5. For each continuous linear functional $\ell \in S'(\Omega^{q-1})$ there exists $F_\ell \in T(\Omega^{q-1})$ such that for all $f \in S(\Omega^{q-1})$ one has $\ell(f) = \langle f,F\rangle$.

Proof. Let $\psi \in L_2(\Omega^{q-1})$. Denote the solution of the Dirichlet problem on $B^q(\underline{0};1)$ with ψ as a boundary condition again by ψ. For each r, $0 < r < 1$, $\psi(r\,\underline{\xi})$ belongs to $S(\Omega^{q-1})$. Let $\ell \in S'(\Omega^{q-1})$ be given. The functional $\psi \mapsto \ell(\psi(r\cdot))$, r fixed, is continuous on $L_2(\Omega^{q-1})$. Hence, by Riesz' theorem there exists $g_r \in L_2(\Omega^{q-1})$ such that $\ell(\psi(r\cdot)) = (\psi,g_r)$. Replacing ψ by $\psi(r_1\cdot)$ we find $\ell(\psi(r_1 r\cdot)) = (\psi(r_1\cdot),g_r) = (\psi,g_r(r_1\cdot)) = (\psi,g_{r_1 r})$. Define F_ℓ by $F_\ell(r\,\underline{\xi}) = g_r(\underline{\xi})$. It is harmonic and reproduces ℓ in the desired way. □

Now we come to some natural classes of operators which map $S(\Omega^{q-1})$ continuously into itself. Most of these operators use the harmonic extension of the test functions for their definition.

3.A. Multipliers

Let $h \in S(\Omega^{q-1})$ be fixed. Consider the mapping $f \mapsto M_h f = h \cdot f$. Following Lemma 2.2 we see that $h \cdot f \in S(\Omega^{q-1})$.

3.B. Differentiation operators

Let $\underline{a} \in \mathbb{R}^q$. The operator $f \mapsto (\underline{a} \cdot \underline{\nabla}) f$ is defined as follows. First extend f to a harmonic function, then calculate $a_1 \frac{\partial f}{\partial x_1} + \ldots + a_q \frac{\partial f}{\partial x_q}$ and restrict this to Ω^{q-1}. Instead of the constants we can also use multipliers, thus getting differential operators with variable coefficients. An interesting subclass of this type is obtained in the following way: Take a matrix $A \in \mathbb{R}^{q \times q}$. The operator $f \mapsto (\underline{x}, A\underline{\nabla}) f$ maps $S(\Omega^{q-1})$ into itself. If $A = I$ then $(\underline{x}, A\underline{\nabla}) = \frac{\partial}{\partial n}$. If A is antisymmetric, $A^T = -A$ the vector fields $(\underline{x}, A\underline{\nabla})$ are tangent to Ω^{q-1}, they are linear combinations of the moment of momentum operators in quantum mechanics. They are the infinitesimal generators of rotation operators in $L_2(\Omega^{q-1})$.

3.C. Harmonic contractions

Take a matrix $A \in \mathbb{R}^{q \times q}$ with $\|A\| \leq 1$. Define $(L_A f)(\underline{\xi}) = f(A\underline{\xi})$. In this definition the harmonic extension of f is used. From Lemma 2.3 we obtain that L_A maps $S(\Omega^{q-1})$ into itself. If A is orthogonal the harmonic extension of f is not needed because then $\|A\underline{\xi}\| = 1$. Notice that $L_{AB} \neq L_A \circ L_B$ in general!

THEOREM 3.6. The operators mentioned in 3.A, 3.B and 3.C map $S(\Omega^{q-1})$ continuously into itself.

The proof can be given by ad hoc arguments or by applying $[G_3]$.

Finally we come to the question whether the operators 3.A, 3.B and 3.C can be extended to operators from the distribution space $T(\Omega^{q-1})$ into itself. If a mapping $L : S(\Omega^{q-1}) \to S(\Omega^{q-1})$ has a $L_2(\Omega^{q-1})$-adjoint L^* which maps $S(\Omega^{q-1})$ continuously into itself, then L can be extended to $\bar{L} : T(\Omega^{q-1}) \to T(\Omega^{q-1})$ by $\langle f, \bar{L}F\rangle = \langle L^* f, F\rangle$ which is a continuous linear functional on $S(\Omega^{q-1})$. This easily proves the extendibility of the multipliers.

The extendability of differential operators with constant coefficients follows because they map $\mathrm{Harm}(B^q(\underline{0};1))$ into itself. The general differential operators are extendable because they are compositions of differential operators with constant coefficients and multipliers.

The extendability of L_A with A orthogonal follows from $L_A^* = L_{A^T}$. If $\|A\| < 1$ then

$$\| \sum_{n=0}^{\infty} R^n S_{n,f}(A\underline{\xi}) \| \leq c \| f \|$$

if $R\|A\| < 1$. This implies the extendability. Cf. [G_2]. If $\|A\| < 1$ the operator L_A is even smoothing, i.e. it maps $T(\Omega^{q-1})$ into $S(\Omega^{q-1})$.

We will not discuss the extendability of L_A for the general case $\|A\| \leq 1$ here.

4. A SPACE OF GENERALIZED FUNCTIONS WITH A METRIZABLE TESTSPACE

In this section we consider a different Gel'fand triple

$$E(\Omega^{q-1}) \subset L_2(\Omega^{q-1}) \subset U(\Omega^{q-1}) \ .$$

The test space $E(\Omega^{q-1})$ consists of restrictions to Ω^{q-1} of functions in $\mathrm{Harm}(\mathbb{R}^q)$. We will (somewhat loosely) identify $E(\Omega^{q-1})$ and $\mathrm{Harm}(\mathbb{R}^q)$.

DEFINITION 4.1. A sequence $(f_n) \subset E(\Omega^{q-1})$ is said to converge iff (f_n) converges uniformly on each ball $B^q(\underline{0};R)$ for all $R > 0$. This is equivalent to saying that $(f_n(R\,\underline{\xi}))$ converges in $L_2(\Omega^{q-1})$ for all $R > 0$.

For $U(\Omega^{q-1})$ we take $\mathrm{Harm}(\underline{0})$. It "contains" (possibly diverging) series of spherical harmonics $\sum_{n=0}^{\infty} S_{n,F}(\underline{\xi})$ with the property that $\sum_{n=0}^{\infty} r^{2n} \| S_{n,F} \|^2 < \infty$ for r sufficiently small.

DEFINITION 4.2. A sequence $(F_n) \subset U(\Omega^{q-1})$ is said to converge iff $(F_n) \subset \mathrm{Harm}(B^q(\underline{0};r))$, for some $r > 0$, and (F_n) converges uniformly on $B^q(\underline{0};r)$. This is equivalent to saying that $(F_n(r\,\underline{\xi})$ converges in $L_2(\Omega^{q-1})$ for $r > 0$ sufficiently small.

REMARK 4.3. $E(\Omega^{q-1})$ is a space of type $\tau(Y,\mathcal{B})$ and $U(\Omega^{q-1})$ is a space of type $\sigma(Y,\mathcal{B})$. See [E]. For Y and $\mathcal{B}$ see Remark 3.3. All general (topological) considerations of [E] apply here. In particular $E(\Omega^{q-1})$ and $U(\Omega^{q-1})$ are complete nuclear topological vector spaces. $E(\Omega^{q-1})$ is a Fréchet space. $U(\Omega^{q-1})$ is an inductive limit of Hilbert spaces. Some of the results in [E] are presented here in an *ad hoc* manner.

DEFINITION 4.4. Let $f \in E(\Omega^{q-1})$, $F \in U(\Omega^{q-1})$. The pairing $\langle f,F\rangle_U$ is defined by

$$\langle f,F\rangle_U = (f(R\,\underline{\xi})\,,\, F(R^{-1}\,\underline{\xi})) \tag{4.1}$$

The inner product makes sense for $R > 0$ sufficiently large and does not depend on the choice of R.

It is a simple exercise to prove that the mappings $f \mapsto \langle f,F\rangle_U$ and $F \mapsto \langle f,F\rangle_U$ are sequentially continuous. Without proof we mention, cf. [E],

THEOREM 4.5. For each continuous linear functional $\ell \in E'(\Omega^{q-1})$ there exists $F_\ell \in U(\Omega^{q-1})$ such that for all $f \in E(\Omega^{q-1})$ one has $\ell(f) = \langle f,F_\ell\rangle_U$.

Now we come to some natural classes of operators which map $E(\Omega^{q-1})$ continuously into itself. Most of these operators use the harmonic extension of the testfunctions to the whole of $\mathbb{R}^q$ for their definition.

4.A. Multipliers

Let $h \in E(\Omega^{q-1})$ be fixed. With Lemma 2.2 we see that the mapping $f \mapsto M_h f = h \cdot f$ acts from $E(\Omega^{q-1})$ into itself.

4.B. Differentiation operators

Just like in 3.B we can introduce the operators $(\underline{a} \cdot \underline{\nabla})$, $(\underline{x},A\,\underline{\nabla})$, etc. The comments in 3.B also apply here.

4.C. Harmonic contractions and expansions

Take any matrix $A \in \mathbb{R}^{q\times q}$. Define $(L_A f)(\underline{\xi}) = f(A\underline{\xi})$ with the aid of the harmonic extension of f. The comments in 3.C also apply here.

4.D. Harmonic translations

Let $\underline{w} \in \mathbb{R}^q$. Define $(T_{\underline{w}} f)(\underline{\xi}) = f(\underline{\xi} + \underline{w})$. $T_{\underline{w}}$ clearly maps $E(\Omega^{q-1})$ into itself.

THEOREM 4.6. The operators mentioned in 4.A, 4.B, 4.C and 4.D map $E(\Omega^{q-1})$ continuously into itself.

The proof can be given by ad hoc arguments or with the aid of [E].

Finally a few words on the extendibility problem. The operators 4.A and 4.B are extendible to operators from $U(\Omega^{q-1})$ into itself. The proof runs along similar lines as in the cases 3.A and 3.B. The extendibility of the operators 4.C and 4.D is an open problem.

With each element $[A;\underline{w}]$ of the affine (semi)group on $\mathbb{R}^q$ we can associate the operator $L_{[A;\underline{w}]}$ by

$$(L_{[A;\underline{w}]} f)(\underline{\xi}) = f(A\underline{\xi} + \underline{w}) \ .$$

In general we have

$$L_{[B;\underline{z}]} \circ L_{[A;\underline{w}]} \neq L_{[BA;B\underline{w}+\underline{z}]} \ .$$

As yet I do not know in which way the operators $L_{[A;\underline{w}]}$ "represent" the affine semigroup $\{[A;\underline{w}]\}$ on $\mathbb{R}^q$.

REFERENCES

[E] Eijndhoven, S.J.L. van, A theory of generalized functions based on one-parameter groups of unbounded self-adjoint operators. T.H.-Report 81-WSK-03, Eindhoven University of Technology.

[G_1] Graaf, J. de, A theory of generalized functions based on holomorphic semi-groups.
Part A: Introduction and Survey. To appear in Proc. KNAW.

[G_2] Idem. Part B: Analyticity spaces, trajectory spaces and their pairing. To appear in Proc. KNAW.

[G_3] Idem. Part C: Linear mappings, tensor products and Kernel theorems. To appear in Proc. KNAW.

[M] Müller, C., Spherical Harmonics. Springer Lecture Notes in Mathematics, Vol. 17, Springer Verlag, Berlin etc. 1966.

[S] Seidel, J.J., Spherical Harmonics and Combinatorics. Preprint, Memorandum 1981-07, June 1981, Eindhoven University of Technology.

DUAL SEIDEL SWITCHING

by

Willem Haemers

Dedicated to J.J. Seidel on the occasion of his retirement.

INTRODUCTION

From 1973 to 1980 I have studied and worked under the stimulating guidance of J.J. Seidel. Already in the beginning of this period Seidel posed the problem treated in the present note, and it has been on my mind ever since. I was never able to deal with the problem in a satisfactory way, but there are some miscellaneous results that have never been published. These results are closely related to Seidel's work. Therefore, it seemed a good idea to publish them here.

The problem treated here has to do with strongly regular graphs with $\lambda = \mu$. These graphs give rise to symmetric block designs (see for instance [2]). The question is when do non-isomorphic graphs produce isomorphic designs. The main result implies that this is impossible if the automorphism group of one of the graphs is trivial. It is also shown that the question is in a certain sense duel to the following one.
When are non-isomorphic strongly regular graphs equivalent under Seidel-switching. This led to the title of this note.

The reader is assumed to be familar with strongly regular graphs and some related topics. A good general reference is Seidel's survey [7]. The present note uses the notation from this survey.

THE ISOMORPHISM PROBLEM

Let A be the adjacency matrix of a strongly regular graph (v,k,λ,μ). It is well known and easily verified that A is the incidence matrix of a symmetric block design whenever $\lambda = \mu$. These strongly regular graphs are often called (v,k,λ) graphs (cf. [2]). Similarly, if $\lambda + 2 = \mu$, then $A + I$ represents a symmetric block design.

Here we deal with the question: when do non-isomorphic strongly regular graphs give rise to the same symmetric block design?

Let A (or $A - I$) and B (or $B - I$) be the adjacency matrices of two such graphs. Since the corresponding block designs are isomorphic, there exist permutation matrices Q and R such that

$$B = QAR .$$

Hence, $RBR^t = RQA$. Put $P = RQ$, Then $PA \sim B$ ("~" indicates that the corresponding graphs are isomorphic). This implies that the rows of A can be permuted in such a way that another strongly regular graph appears. Our first result states that this phenomenon is a privilege of (v,k,λ) graphs.

RESULT 1. Let A be the adjacency matrix of a non-trivial strongly regular graph G. Let $P \neq I$ be a permutation matrix such that PA or $PA - I$ is again the

adjacency matrix of a strongly regular graph. Then G is a (v,k,λ) graph.

Proof. Because G is strongly regular we have

$$A^2 = (\lambda - \mu)A + (k - \mu)I + \mu J.$$

A strongly regular graph with matrix PA has to have the same parameters as G: (v,k,λ,μ), and a graph with matrix PA - I has parameters $(v,k-1,\lambda-2,\mu)$. In both cases PA satisfies

$$(PA)^2 = (\lambda - \mu)PA + (k - \mu)I + \mu J.$$

By use of

$$(PA)^2 = (PA)^t(PA) = A^2$$

it now follows that PA = A, or $\lambda = \mu$. □

The two non-isomorphic (16,6,2) graphs give rise to the same block design. So they produce an example for the case $A \neq PA$. See [5], p. 10 for details.

An example where A and PA - I are both adjacency matrices of strongly regular grpahs (that are obviously non-isomorphic) can be constructed as follows. Consider the following (Hadamard) matrix:

$$H = \begin{pmatrix} 1 & -1 & 1 & 1 \\ -1 & 1 & 1 & 1 \\ 1 & 1 & 1 & -1 \\ 1 & 1 & -1 & 1 \end{pmatrix}.$$

Let H^m denote the m-th Kronecker product of H with itself, that is

$$H^m = H \otimes H \otimes \ldots \otimes H \quad (m \text{ times}) .$$

Then H^m is a regular symmetric Hadamard matrix with constant diagonal (see [4]) and $A = -\frac{1}{2}(H^m - J)$ is the adjacency matrix of a (v,k,λ) graph. Define

$$P = (J_2 - I_2) \otimes I_{2m}$$

(indices indicate the size of the matrix). Then PA is symmetric and has all-one diagonal. Therefore PA - I represents a strongly regular graph with parameters $(v,k-1,\lambda-2,\lambda)$.

In the sequel we suppose that A represents a (v,k,λ) graph, and that PA or PA - I is the adjacency matrix of a strongly regular graph. Since PA is symmetric we have

$$(*) \qquad PA = AP^t , \quad AP = P^tA, \quad PAP = P^tAP^t = A .$$

LEMMA 1. For any integer m, P^mA represents a strongly regular graph and

$$P^mA \sim \begin{cases} A , & \text{if m is even,} \\ PA , & \text{if m is odd.} \end{cases}$$

Proof. Induction on m. For m = 0, or 1 the result is obvious. Let $m > 1$, then by use of (*)

$$P^mA = PP^{m-2}(PA) = P(P^{m-2}A)P^t \sim P^{m-2}A. \qquad \square$$

LEMMA 2. Let n be the order of P. If n is odd then $PA \sim A$.
If n is even then $P^{\frac{1}{2}n}$ is an involution of the graphs represented by A and PA.

Proof. The first line follows directly from Lemma 1. To prove the second line we apply (*):

$$A = P^n A = P^{\frac{1}{2}n} A (P^{\frac{1}{2}n})^t ,$$

and similarly

$$PA = P^{n+1} A = P^{\frac{1}{2}n} A = P^{\frac{1}{2}n} (PA) (P^{\frac{1}{2}n})^t . \qquad \square$$

THEOREM 1. If the automorphism group of the graph represented by A or PA has odd order then PA ~ A.

Proof. If the order of the automorphism group is odd, the graph has no involution. So the result follows from Lemma 2. $\square$

We illustrate the use of Theorem 1 by the following example. Bussemaker, Mathon and Seidel constructed many (v,k,λ) graphs [1]. They obtained 16448 (36,15,6) graphs. They also constructed 105 strongly regular graphs with parameters (36,14,4,6) and 1853 graphs with parameters (35,16,6,8). Out of these graphs 15417, 28 and 1576 respectively, have trivial automorphism groups. Thus by Theorem 1 the first two families produce at least 15417+28+1 = 15446 non-isomorphic symmetric (36,15,6) designs. The third family produces at least 1577 symmetric (35,17,8) designs.

We finish this section with some minor results that may be of some use to those who come across the problem of the present note.

RESULT 2. Suppose PA has zero diagonal. Then P is an even permutation.

Proof. A and PA have the same eigenvalues, and hence the same determinant. So det P = 1. □

RESULT 3. Suppose PA has zero diagonal. The orbits of P correspond to cocliques of A and PA.

Proof. By Lemma 1, P^mA has zero diagonal for any integer m. However, every entry of a principal submatrix of A that corresponds to an orbit of P can be moved to the diagonal by P. □

RESULT 4. Without loss of generality we may assume that all orbit sizes of P are powers of 2.

Proof. Suppose that n, the order of P has an odd divisor. Let m be the largest odd divisor of n. Then P^m has no odd orbit sizes and $P^mA \sim PA$ by Lemma 1. □

DUAL SEIDEL SWITCHING

Permuting the rows (and not the columns) of the incidence matrix of a graph will be called dual Seidel switching. This definition will be justified by the forthcoming result. First we want to point at some similarities between Seidel switching (see [6] or [7]) and dual Seidel switching. Both concepts give a method to derive other strongly regular graphs from a given one. The parameters of the new graph are the same or slightly different (the multiplicities differ by 1 in case of dual Seidel switching). To make this derivation work, the strongly regular graphs have to be very special (regular 2-graphs (see [6]) and (v,k,λ) graphs, respectively). These observations led to the conclusion that these two types of special strongly regular graphs should be

dual to each other in the sense of Delsarte [3].

RESULT 5. If a strongly regular graph in the switching class of a regular 2-graph has a dual, then this dual is a (v,k,λ) graph, and vice versa.

Proof. If a strongly regular graph has eigenvalues k, r and s with multiplicities 1, f and g respectively, then the eigenvalues of its dual are (see [7])

$$f,\ fr/k \quad \text{and} \quad -f(r+1)/(n-k-1) \ .$$

Such a graph is a (v,k,λ) graph whenever

$$fr/k = f(r+1)/(n-k-1) \ ,$$

or, equivalently

$$k = r(n-1)/(2r+1) \ .$$

Substitution of this formula in the following identity for strongly regular graphs (see [7])

$$(n-k-1)(k+rs) = -k(r+1)(s+1)$$

leads to

$$n-1+(2r+1)(2s+1) = 0,$$

which is the condition for a strongly regular graph to be in the switching class of a regular 2-graph (see [6]). □

This duality can also be seen in terms of 3-class association schemes since

both regular 2-graphs and symmetric block designs form 3-class association schemes. It may be that such a setup produces more insight in the present duality case. We haven't worked this out.

It has turned out that Seidel switching can produce a lot of non-isomorphic strongly regular graphs, often with trivial automorphism groups (see for instance [1]). Dual Seidel switching, however, cannot produce graphs with trivial automorphism groups because of Theorem 1. It is likely that the automorphism groups of most (v,k,λ) graphs are trivial. Therefore dual Seidel switching may not be expected to be as fruitful as Seidel switching.

REFERENCES

[1] Bussemaker, F.C., R.A. Mathon and J.J. Seidel, "Tables of two-graphs", Report Techn. Univ. Eindhoven 79-WSK-05 (1979).

[2] Cameron, P.J., and J.H. van Lint, "Graphs, Codes and Designs", London Math. Soc. Lecture Note Series 43, Cambridge 1980.

[3] Delsarte, P., "An algebraic approach to the association schemes of coding theory", Philips Res. Repts. Suppl. 10 (1973).

[4] Goethals, J.-M. and J.J. Seidel, "Strongly regular graphs derived from combinatorial designs", Can. J. Math. 22 (1970) 597 - 614.

[5] Haemers, W.H., "Sterke grafen en block designs", Afstudeerverslag T.H. Eindhoven (1975).

[6] Seidel, J.J., "A survey of two-graphs", in: Proc. Intern. Colloq. Theorie Combinatorie (Roma 1973), Toma I, Accad. Naz. Lincei, 1976, pp. 481 - 511.

[7] Seidel J.J., "Strongly regular graphs", in: Surveys in Combinatorics Proc. 7th Brit. Comb. Conf., B. Bollobas (ed.), London Math. Soc. Lecture Note Series 38, Cambridge 1979, pp. 157 - 180.

NUMERICAL CALCULATION OF THE FRESNEL INTEGRAL

by

J.K.M. Jansen

Dedicated to Prof. J.J. Seidel on the occasion of his 65th birthday.

Summary

An algorithm suitable for a pocket calculator is given for the numerical computation of the Fresnel integral. The relative error in the amplitude and the absolute error in the phase is less than 0.005.

1. INTRODUCTION

There are a lot of algorithms for the computation of the Fresnel integral [1,2,3,5,6]. But these algorithms are not suitable for a pocket calculator. For practical applications [3,4] it is sufficient to calculate the amplitude of the Fresnel integral with a relative error and the phase with an absolute error of 0.005: that is within the accuracy of measurement and the drawing accuracy, respectively.

2. THE FRESNEL INTEGRAL

The Fresnel integral is for real arguments defined as

$$F_{\pm}(x) := \int_{x}^{\infty} \exp(\pm i t^2) dt \ .$$

As in [4] we start from the so-called modified Fresnel integral $K_{\pm}(x)$ defined as

$$K_{\pm}(x) := \frac{1}{\sqrt{\pi}} F_{\pm}(x) \ \exp(\mp i(x^2 + \pi/4))$$

having the following properties

$$K_{\pm}(-x) = \exp(\mp i x^2) - K_{\pm}(x) \ ,$$

$$K_{\pm}(0) = \frac{1}{2} \ ,$$

and the asymptotic behaviour is given by

$$K_{\pm}(x) \sim \frac{1}{2x\sqrt{\pi}} \exp(\pm i\pi/4) \ , \qquad (x \to \infty) .$$

It can be easily verified that

(1) $$K_{\pm}(x) = \frac{1}{2}[1 - (1 \mp i)(C(\sqrt{\frac{2}{\pi}}x) \pm iS(\sqrt{\frac{2}{\pi}}x))] \exp(\mp i x^2)$$

and

(2) $$K_{\pm}(x) = \frac{1}{2}\sqrt{2} \exp(\pm i \frac{\pi}{4}) [f(\sqrt{\frac{2}{\pi}}x) \mp i g(\sqrt{\frac{2}{\pi}}x)]$$

in which the real functions $C(x)$, $S(x)$, $f(x)$ and $g(x)$ are defined by

$$C(x) \pm iS(x) := \int_{0}^{x} \exp(\pm i \frac{\pi}{2} t^2) dt$$

and

$$f(x) \pm ig(x) := [(\frac{1}{2} - S(x) \pm i(\frac{1}{2} - C(x))] \exp(\pm i\frac{\pi}{2}x^2) \text{ , [7].}$$

Using the series expansions [7] of

$$C(x) = \sum_{n=0}^{\infty} \frac{(-1)^n (\pi/2)^{2n}}{(2n)!(4n+1)} x^{4n+1}$$

and

$$S(x) = \sum_{n=0}^{\infty} \frac{(-1)^n (\pi/2)^{2n+1}}{(2n+1)!\ (4n+3)} x^{4n+3}$$

in formula (1) we obtain

$$(3) \qquad K_{\pm}(x) = \frac{1}{2}[1 - \sqrt{\frac{2}{\pi}}x - \frac{1}{3}\sqrt{\frac{2}{\pi}}x^3 + \frac{1}{10}\sqrt{\frac{2}{\pi}}x^5$$
$$\pm i(\sqrt{\frac{2}{\pi}}x - \frac{1}{3}\sqrt{\frac{2}{\pi}}x^3 - \frac{1}{10}\sqrt{\frac{2}{\pi}}x^5) + O(x^7)] *$$
$$* \exp(\mp ix^2) \text{ , } (x \to 0)$$

Substituting the asymptotic expansions [7] of

$$f(x) = \frac{1}{\pi x}\left[1 + \sum_{m=1}^{\infty} (-1)^m \frac{1 \cdot 3 \cdots (4m-1)}{(\pi x^2)^{2m}}\right], \quad (x \to \infty)$$

and

$$g(x) = \frac{1}{\pi x} \sum_{m=0}^{\infty} (-1)^m \frac{1 \cdot 3 \cdots (4m+1)}{(\pi x^2)^{2m+1}}, \quad (x \to \infty)$$

into formula (2) we have

$$(4) \qquad K_{\pm}(x) = \frac{1}{2\sqrt{\pi x}} \exp(\pm i\frac{\pi}{4}) \left[1 - \frac{3}{4x^4} + \frac{105}{16x^8} \mp i\left(\frac{1}{2x^2} - \frac{15}{8x^6}\right) + O\left(\frac{1}{x^{10}}\right)\right], \ (x \to \infty) \text{ .}$$

3. THE COMPUTATION OF THE AMPLITUDE

From (4) it can be easily verified that

$$|K_{\pm}(x)|^2 = \frac{1}{4\pi x^2}\left[1 - \frac{5}{4x^4} + \frac{189}{16x^8} + O\left(\frac{1}{x^{12}}\right)\right], \qquad (x \to \infty) ,$$

or

$$|K_{\pm}(x)|^2 = \frac{1}{4\pi x^2}\exp\left(\frac{-5}{4x^4}\right)\left(1 + O\left(\frac{1}{x^8}\right)\right) , \qquad (x \to \infty) .$$

By means of computation it appears that for $x > 2.3$ the expressions

$$\frac{1}{4\pi x^2}\left(1 - \frac{5}{4x^4}\right) \qquad \text{and} \qquad \frac{1}{4\pi x^2}\exp\left(\frac{-5}{4x^4}\right)$$

agree with $|K_{\pm}(x)|^2$ within the relative accuracy of 0.005.

From (3) it follows that

$$|K_{\pm}(x)|^2 = \frac{1}{4}\left[1 - 2\sqrt{\frac{2}{\pi}}x + \frac{4}{\pi}x^2 - \frac{2}{3}\sqrt{\frac{2}{\pi}}x^3 + \right.$$

$$\left. + \frac{1}{5}\sqrt{\frac{2}{\pi}}x^5 - \frac{16}{45\pi}x^6\right] + O(x^7) , \qquad (x \to 0) ,$$

or

$$|K_{\pm}(x)|^2 = \frac{1}{4}\exp\left(-2\sqrt{\frac{2}{\pi}}x\right)(1 + O(x^3)) , \qquad (x \to 0) .$$

The approximation $\frac{1}{4}\exp\left(-2\sqrt{\frac{2}{\pi}}x\right)$ for $|K_{\pm}(x)|^2$ does not agree within the prescribed relative accuracy of 0.005 in the range $0 \le x \le 2.3$. By means of the least squares method we have computed a correction factor $1 + 0.105x^3$.

Summarized we have

$$|K_{\pm}(x)|^2 \sim \begin{cases} \frac{1}{4}\exp\left(-2\sqrt{\frac{2}{\pi}}x\right)\left(1 + 0.105x^3\right), & 0 \le x \le 2.3 \\ \frac{1}{4\pi x^2}\left(1 - \frac{5}{4x^4}\right) \text{ or } \frac{1}{4\pi x^2}\exp\left(\frac{-5}{4x^4}\right), & x > 2.3 \end{cases}$$

4. THE COMPUTATION OF THE PHASE

After some formula manipulations it follows from (4) that

$$\arg K_{\pm}(x) = \pm\left[-\frac{\pi}{4} + \arctan\left(2x^2\left(1 + \frac{3}{x^4} + O\left(\frac{1}{x^8}\right)\right)\right)\right], \quad (x \to \infty),$$

or

$$\arg K_{\pm}(x) = \pm\left[-\frac{\pi}{4} + \arctan\left(2x^2 \exp\left(\frac{3}{x^4}\right)\right)\left(1 + O\left(\frac{1}{x^6}\right)\right)\right], \quad (x \to \infty).$$

Again it appears by means of computation that for $x > 2.3$ the formulas

$$\pm[-\pi/4 + \arctan(2x^2 + 6/x^2)]$$

and

$$\pm[-\pi/4 + \arctan(2x^2 \exp(3/x^4))]$$

agree with $\arg K_{\pm}(x)$ within the absolute error of 0.005.

From (3) we find that

$$\arg K_{\pm}(x) = \arctan(\pm N/D) \mp x^2$$

in which

$$N := \sqrt{\frac{2}{\pi}}x - \frac{1}{3}\sqrt{\frac{2}{\pi}}x^3 - \frac{1}{10}\sqrt{\frac{2}{\pi}}x^5 + O(x^7), \quad (x \to 0)$$

and

$$D:= 1 - \sqrt{\frac{2}{\pi}}x - \frac{1}{3}\sqrt{\frac{2}{\pi}}x^3 + \frac{1}{10}\sqrt{\frac{2}{\pi}}x^5 + O(x^7) \ , \quad (x \to 0) \ .$$

Using the realtion

$$x^2 = \arctan(x^2) + O(x^6) \ , \quad (x \to 0)$$

we obtain after some elementary calculations

$$\arg K_{\pm}(x) = \pm \arctan\left(\sqrt{\frac{2}{\pi}}x + \left(\frac{2}{\pi} - 1\right)x^2 + \sqrt{\frac{2}{\pi}}\left(\frac{2}{\pi} - \frac{1}{3}\right)x^3 + O\left(x^4\right)\right), (x \to 0) \ .$$

For $0 \le x \le 2.3$ the formula

$$\pm \arctan\left(\sqrt{\frac{2}{\pi}}x + \left(\frac{2}{\pi} - 1\right)x^2 + \sqrt{\frac{2}{\pi}}\left(\frac{2}{\pi} - \frac{1}{3}\right)x^3\right)$$

does not agree with the prescribed absolute error of 0.005. But this expression led us to the idea to represent the approximation in the form

$$\pm \arctan(ax + bx^2 + cx^3)$$

in which a, b and c are computed by means of the least squares method.

Summarizing we have the following approximation of the phase

$$\arg K_{\pm}(x) \sim \begin{cases} \pm\arctan(0.773x - 0.250x^2 + 0.032x^3) \ , & 0 \le x \le 2.3 \\ \pm(-\pi/4 + \arctan(2x^2 + 6/x^2) \quad \text{or} \\ \pm(-\pi/4 + \arctan(2x^2 \exp(3/x^4)) \ , & x > 2.3 \end{cases}$$

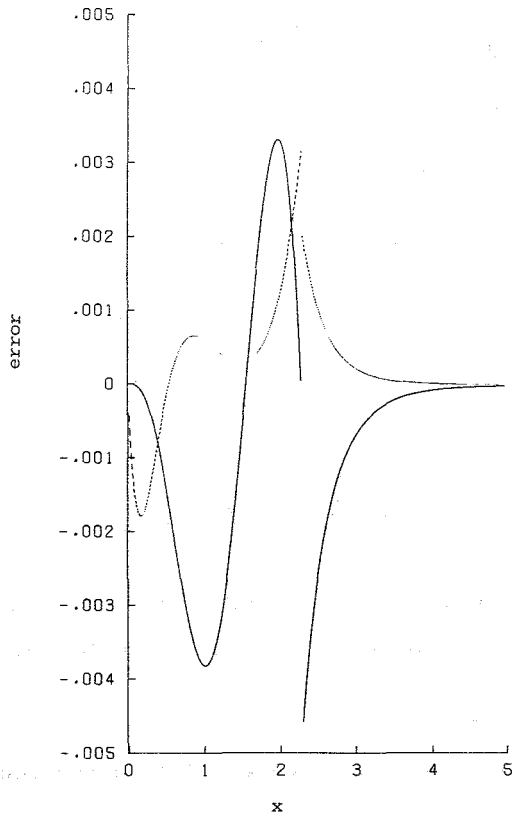


Figure 1

——— relative error of the amplitude

----- absolute error of the phase

Figure 1 shows the relative error of the amplitude and the absolute error of the phase.

All the tests were performed on the Burroughs B7700 digital computer of the Computer Center of the Eindhoven University of Technology, The Netherlands.

I am indebted to Mr. A.P.M. Baayens for the valuable assistance of this work.

5. REFERENCES

[1] Boersma, J.,Computation of Fresnel integrals. Math. Comp. 14 (1960), 380.

[2] Burlisch, R., Numerical calculation of the sine, cosine and Fresnel integrals. Num. Math. 9 (1967), 380 - 385.

[3] Henrici, P., Zur numerischen Berechnung der Fresnelschen Integrale. ZAMP. 30 (1979), 209 - 219.

[4] James, G.L., An approximation to the Fresnel integral. Proc. IEEE 67 (1979), 677 - 678.

[5] Lotsch, H., Fresnel integrals. Algorithm 244. Collected Algorithms from CACM.

[6] Németh, G., Chebyshev expansions for Fresnel integrals. Num. Math. 7 (1965), 310 - 312.

[7] ---------- Handbook of mathematical functions; with formulas, graphs and mathematical tables; ed. by M. Abramowitz and I.A. Stegun. Washington, National Bureau of Standards, 1964. (Applied Mathematical Series, nr. 55). p. 300.

FROM LINEAR ELASTICITY TO LINEAR ELASTIC RELAXATION

A first step towards a more general continuum mechanics

by

G.A. Kluitenberg

For Dr. J.J. Seidel, Professor of Mathematics in the Eindhoven University of Technology, on the occasion of his 65th birthday.

Summary. The starting point of our discussion is the strain-energy function well-known from the linear theory of elasticity. We generalize this function so that it depends not only on the temperature and the strain but also on a macroscopic hidden internal tensorial variable $Z_{\alpha\beta}$ which influences the mechanical behaviour of the medium. We introduce a flow law for $Z_{\alpha\beta}$ and it is shown that the Poynting-Thomson equation for elastic relaxation can be derived from the formalism. The latter equation includes strain relaxation at constant stress and stress relaxation at constant strain. Furthermore, we define elastic and inelastic strains and it is shown that the theory may be reformulated in terms of the temperature and these two types of strains. The thermodynamic aspects of the theory are left out of consideration. The developed theory is closely related with network theory and with the mathematical theory of systems.

§1. INTRODUCTION

The theory of elasticity is a part of mathematical physics which is well understood nowadays. A clear presentation of the foundations of this theory may be found in the lecture notes Tensorrekening by J.J. Seidel[1]. As is well-known many other mechanical phenomena may occur in continuous media such as the flow of viscous and elastic liquids, primary and secondary creep and plasticity.

Though there are ad hoc theories for many of these phenomena a more general theory for both the mechanical and thermal aspects of flow and deformation in continuous media is still lacking. The draft of such a theory is called by Truesdell and Noll the main open problem of the theory of material behaviour[2].

It is not excluded that a theory using the concept of "internal tensorial variable" may contribute to the solution of problems in continuum mechanics [3,4]. In this paper we shall discuss the use of internal variables in the theory of elastic relaxation. We shall limit ourselves to phenomena which may be described by linear relations (Poynting-Thomson equation). For simplicity thermodynamics will be left out of consideration. This implies that we must refrain from a more fundamental discussion of the law for mechanical flow phenomena and the expression for the heat dissipation due to internal friction, which we shall introduce.

Our starting point is the strain-energy function well-known from the linear theory of elasticity. We generalize this function by assuming that it depends not only on the tensor of total strain but also on the temperature and on a "hidden tensorial internal variable". Next we postulate a flow law and an expression for the heat dissipation.

We note that analogous methods may be applied to derive the Debye equation for dielectric relaxation and the Snoek equation for magnetic relaxation. We also remark that the theory discussed in the following sections is closely connected with the mathematical theory of systems and with network theory.

In §2 and §3 we give a short review of the theory of elasticity and in §4 the notions of internal variable, flow law and heat dissipation are introduced. A simple expression for the free energy (generalization of the strain-energy function) is discussed in §5. This expression leads to a completely linearized theory. We also derive the Poynting-Thomson equation in §5. Next we show in §6 how one can define the elastic and inelastic parts of the strain. Furthermore, we show that the theory may be reformulated with the aid of these two types of strain. In §7 we discuss some important inequalities and we make some final remarks.

§2. THE LINEAR THEORY OF ELASTICITY

We introduce an orthogonal Cartesian frame of axes which is fixed in space (prerelativistic point of view). The components of the metric tensor are then given by the unit matrix (Kronecker symbol) and there is no difference between contravariant and covariant components of vectors and tensors. (See Seidel[1] §1, §2 and §3 of Chapter I.)

In the theory of elasticity three quantities play a fundamental role. They are the symmetric strain tensor $\varepsilon_{\alpha\beta}$, the mechanical stress tensor $\tau_{\alpha\beta}$ which is also symmetric and the temperature T which is a scalar. Furthermore, in this theory a function f of the strain and the temperature is introduced with the property

$$\tau_{\alpha\beta} = \rho_0 \frac{\partial}{\partial \varepsilon_{\alpha\beta}} f(T, \varepsilon_{\gamma\zeta}) \ . \tag{2.1}$$

The function f is called the strain-energy function. It may be identified with the specific free energy (free energy per unit of mass) which is introduced in thermodynamics. Furthermore, ρ_0 is the mass density in the unstrained state characterized by $\varepsilon_{\alpha\beta} = 0$.

We shall assume that the deformations and rotations in the medium are small from a kinematical point of view. We shall also suppose that the deformations and temperature differences are small from a physical point of view. The latter assumption means that we suppose that f has the form

$$f(T, \varepsilon_{\alpha\beta}) = v_0 \left[\frac{1}{2} \sum_{\alpha,\beta,\gamma,\zeta=1}^{3} a_{\alpha\beta\gamma\zeta} \, \varepsilon_{\alpha\beta} \, \varepsilon_{\gamma\zeta} + (T - T_0) \sum_{\alpha,\beta=1}^{3} a_{\alpha\beta} \, \varepsilon_{\alpha\beta} \right] - \Psi(T) \ , \tag{2.2}$$

where v_0 is the volume per unit of mass in the unstrained state (specific volume), related to ρ_0 by

$$v_0 = \frac{1}{\rho_0} \ . \tag{2.3}$$

Ψ is a function of the temperature which we need not specify for our purposes. T_0 is a (constant) reference temperature (for example "room temperature").

Finally, $a_{\alpha\beta\gamma\zeta}$ and $a_{\alpha\beta}$ are tensors with the symmetry properties

$$a_{\alpha\beta\gamma\zeta} = a_{\beta\alpha\gamma\zeta} = a_{\alpha\beta\zeta\gamma} = a_{\beta\alpha\zeta\gamma} \ , \tag{2.4}$$

$$a_{\alpha\beta\gamma\zeta} = a_{\gamma\zeta\alpha\beta} \ , \tag{2.5}$$

$$a_{\alpha\beta} = a_{\beta\alpha} \ . \tag{2.6}$$

These tensors are determined by the physical properties of the material.

Using (2.3) and (2.5) one may derive from (2.1) and (2.2) that

$$\tau_{\alpha\beta} = \sum_{\gamma,\zeta=1}^{3} a_{\alpha\beta\gamma\zeta} \, \varepsilon_{\gamma\zeta} + a_{\alpha\beta}(T - T_0) \ . \tag{2.7}$$

This linear relation among the temperature, the components of the stress tensor and the components of the strain tensor is called the Duhamel-Neumann law. Hooke's law is obtained by neglecting in (2.7) the thermal effects. (See also Seidel[1] §1 and §2 of Chapter 2.)

Finally, we note that one also has

$$f = \tfrac{1}{2} v_0 \left[\sum_{\alpha,\beta=1}^{3} \tau_{\alpha\beta} \, \varepsilon_{\alpha\beta} + (T - T_0) \sum_{\alpha,\beta=1}^{3} a_{\alpha\beta} \, \varepsilon_{\alpha\beta} \right] - \Psi(T) \ . \tag{2.8}$$

This expression for the free energy (strain-energy function) follows from (2.2) and (2.7).

§3. ISOTROPIC MEDIA

The deviator $\tilde{A}_{\alpha\beta}$ and the scalar part A of an arbitrary tensor field $A_{\alpha\beta}$ of second order are defined by

$$\tilde{A}_{\alpha\beta} = A_{\alpha\beta} - \frac{1}{3}\,\delta_{\alpha\beta} \sum_{\gamma=1}^{3} A_{\gamma\gamma} \,, \tag{3.1}$$

$$A = \frac{1}{3} \sum_{\gamma=1}^{3} A_{\gamma\gamma} \,. \tag{3.2}$$

In (3.1) $\delta_{\alpha\beta}$ is the Kronecker symbol defined by

$$\delta_{\alpha\beta} = \begin{cases} 1 & \text{if } \alpha = \beta \,, \\ 0 & \text{if } \alpha \neq \beta \,. \end{cases} \tag{3.3}$$

We note that

$$A_{\alpha\beta} = \tilde{A}_{\alpha\beta} + A\,\delta_{\alpha\beta} \tag{3.4}$$

and that

$$\sum_{\gamma=1}^{3} \tilde{A}_{\gamma\gamma} = 0 \,. \tag{3.5}$$

Furthermore, it is seen that A equals $\frac{1}{3}$ of the trace of $A_{\alpha\beta}$ and that $\tilde{A}_{\alpha\beta}$ is symmetric if and only if $A_{\alpha\beta}$ is symmetric.

As noted before the tensors $a_{\alpha\beta\gamma\zeta}$ and $a_{\alpha\beta}$ which occur in the Duhamel-Neumann law (and in Hooke's law) (see (2.7)) are determined by the material properties of the medium. If the medium is isotropic $a_{\alpha\beta\gamma\zeta}$ and $a_{\alpha\beta}$ must be isotropic tensors. Isotropic tensors of order 4 which satisfy (2.4) and (2.5) can be

written in the form

$$a_{\alpha\beta\gamma\zeta} = \tfrac{1}{2}a(\delta_{\alpha\gamma}\,\delta_{\beta\zeta} + \delta_{\beta\gamma}\,\delta_{\alpha\zeta}) + \frac{1}{3}(b-a)\,\delta_{\alpha\beta}\,\delta_{\gamma\zeta}\,, \tag{3.6}$$

where a and b are scalars. Isotropic tensors of order 2 have the form

$$a_{\alpha\beta} = c\,\delta_{\alpha\beta}\,, \tag{3.7}$$

where also c is a scalar. It is trivial that this form for $a_{\alpha\beta}$ satisfies the symmetry property (2.6).

Using the two above expressions for $a_{\alpha\beta\gamma\zeta}$ and $a_{\alpha\beta}$ and the definitions (3.1) and (3.2) for deviator and scalar part, the Duhamel-Neumann law (2.7) for an isotropic thermoelastic medium reads

$$\tau_{\alpha\beta} = a\,\tilde{\varepsilon}_{\alpha\beta} + \{b\,\varepsilon + c(T-T_0)\}\,\delta_{\alpha\beta}\,, \tag{3.8}$$

where we also used the symmetry of the strain tensor $\varepsilon_{\alpha\beta}$. From this expression for $\tau_{\alpha\beta}$ we obtain for the deviator $\tilde{\tau}_{\alpha\beta}$ and the scalar part τ of $\tau_{\alpha\beta}$ the relations

$$\tilde{\tau}_{\alpha\beta} = a\,\tilde{\varepsilon}_{\alpha\beta}\,, \tag{3.9}$$

$$\tau = b\,\varepsilon + c(T-T_0)\,. \tag{3.10}$$

If we neglect in (3.8) and (3.10) the terms with the temperature, (3.8), (3.9) and (3.10) express Hooke's law for isotropic linear elastic media. (See Seidel[1] p. 20 and p. 21.)

Using the Definitions (3.1) and (3.2) for the deviator and the scalar part of a tensor of second order and the Expression (3.7) for $a_{\alpha\beta}$, the Formula (2.8)

for the free energy may be written in the form

$$f = \tfrac{1}{2}v_0 \left[\sum_{\alpha,\beta=1}^{3} \tilde{\tau}_{\alpha\beta} \tilde{\varepsilon}_{\alpha\beta} + 3\{\tau + c(T - T_0)\}\varepsilon \right] - \Psi(T) \ , \tag{3.11}$$

or

$$f = v_0 \left[\tfrac{1}{2}a \sum_{\alpha,\beta=1}^{3} (\tilde{\varepsilon}_{\alpha\beta})^2 + \frac{3}{2}b\varepsilon^2 + 3c(T - T_0)\varepsilon \right] - \Psi(T) \ , \tag{3.12}$$

where we used (3.9) and (3.10).

§4. HIDDEN TENSORIAL INTERNAL VARIABLES AND ELASTIC RELAXATION

Let us consider a crystal lattice. By mechanical stresses such a lattice may be deformed elastically. However, it is also possible that imperfections (dislocations) occur as a consequence of stress fields and such imperfections are closely connected with the occurrence of inelastic deformations. In general both types of phenomena (elastic and inelastic effects) may occur simultaneously.

In the two preceding sections we have seen that if the mechanical phenomena are of a purely thermoelastic nature we only need the macroscopic parameters strain and temperature to give a complete description of the (local) state of the medium. If, moreover, inelastic phenomena occur of the nature discussed in the preceding paragraph, we shall assume that we still need an extra parameter to give such a complete description of the macroscopic state. We shall assume that this extra variable is a symmetric second order tensor. Let us

denote this tensor by $Z_{\alpha\beta}$. Hence,

$$Z_{\alpha\beta} = Z_{\beta\alpha} \ . \tag{4.1}$$

We shall not specify the physical nature of $Z_{\alpha\beta}$. Therefore, we shall call this quantity a hidden tensorial internal variable. We shall only assume that $Z_{\alpha\beta}$ influences the mechanical properties of the medium. This terminology will be explained in the next section.

Hence, as a complete set of variables we now take T, $\varepsilon_{\alpha\beta}$ and $Z_{\alpha\beta}$ and we assume that the free energy (the generalization of the strain energy) is a function of these parameters. Furthermore, we shall suppose that

$$\tau_{\alpha\beta} = \rho_0 \frac{\partial}{\partial \varepsilon_{\alpha\beta}} f(T, \varepsilon_{\alpha\beta}, Z_{\alpha\beta}) \ . \tag{4.2}$$

This relation is analogous to (2.1). Next, we define the tensor $G_{\alpha\beta}$ by

$$G_{\alpha\beta} = -\rho_0 \frac{\partial}{\partial Z_{\alpha\beta}} f(T, \varepsilon_{\mu\nu}, Z_{\gamma\zeta}) \tag{4.3}$$

and we shall call this quantity the affinity conjugate to $Z_{\alpha\beta}$. One has

$$G_{\alpha\beta} = G_{\beta\alpha} \ . \tag{4.4}$$

We shall limit ourselves to isotropic media. For such media we shall assume that $Z_{\alpha\beta}$ satisfies the equation

$$\frac{dZ_{\alpha\beta}}{dt} = \eta' \tilde{G}_{\alpha\beta} + \eta'_v G \delta_{\alpha\beta} \ , \tag{4.5}$$

where η' and η'_v are scalars determined by the material properties of the medium.

From (4.5) it follows that

$$\frac{d\tilde{Z}_{\alpha\beta}}{dt} = \eta' \tilde{G}_{\alpha\beta} \tag{4.6}$$

and that

$$\frac{dZ}{dt} = \eta_v' G . \tag{4.7}$$

In the three preceding equations t is the time. In the References 3 and 4 a more thorough discussion of relations of the type (4.5) is presented. Equations of the type of (4.5) - (4.7) may be called flow laws. This denomination is justified by the fact that from (4.5) an equation (mechanical flow law) for the inelastic strain may be derived. This will be shown in §6.

Furthermore, we introduce the physical assumption that the heat dissipation per unit of volume and per unit of time, $\sigma^{(h)}$, is given by

$$\sigma^{(h)} = \sum_{\alpha,\beta=1}^{3} G_{\alpha\beta} \frac{dZ_{\alpha\beta}}{dt} . \tag{4.8}$$

A derivation of (4.8) based on thermodynamic arguments is given in Reference 4. Using (3.3), (3.4) and (3.5) one may also write for (4.8)

$$\sigma^{(h)} = \sum_{\alpha,\beta=1}^{3} \tilde{G}_{\alpha\beta} \frac{d\tilde{Z}_{\alpha\beta}}{dt} + 3G \frac{dZ}{dt} , \tag{4.9}$$

or, with the aid of (4.6) and (4.7),

$$\sigma^{(h)} = \eta' \sum_{\alpha,\beta=1}^{3} (\tilde{G}_{\alpha\beta})^2 + 3\eta_v' G^2 . \tag{4.10}$$

We also have

$$\sigma^{(h)} \geq 0 \ , \tag{4.11}$$

i.e. the heat dissipation is a nonnegative quantity. Thus,

$$\eta' \geq 0 \ , \quad \eta'_v \geq 0 \tag{4.12}$$

by virtue of (4.10) and (4.11).

§5. LINEAR EQUATIONS OF STATE AND THE POYNTING-THOMSON EQUATION

As we have seen in §3 the linear laws of Hooke and of Duhamel-Neumann for isotropic elastic substances are closely related with the expression (3.12) for the free energy (strain energy). We shall now generalize the expression (3.12) to the case that hidden internal tensorial variables occur as discussed in the preceding section. Starting from this generalization of the free energy and from the flow law introduced in §4 we shall derive the Poynting-Thomson equation.

The generalized expression for the free energy which we have in mind reads

$$\begin{aligned} f = v_0 \Big[\frac{1}{2} a \sum_{\gamma,\zeta=1}^{3} (\tilde{\varepsilon}_{\gamma\zeta})^2 + a' \sum_{\gamma,\zeta=1}^{3} \tilde{\varepsilon}_{\gamma\zeta} \tilde{Z}_{\gamma\zeta} + \frac{1}{2} a'' \sum_{\gamma,\zeta=1}^{3} (\tilde{Z}_{\gamma\zeta})^2 + \\ + \frac{3}{2} b \varepsilon^2 + 3 b' \varepsilon Z + \frac{3}{2} b'' Z^2 + 3(T - T_0)(c\varepsilon - c' Z) \Big] - \Psi(T) \ , \end{aligned} \tag{5.1}$$

where a, a', a", b, b', b", c and c' are constants determined by the physical properties of the material and Ψ is a function of the temperature which will not be specified since its explicit form is not of importance for the discussion of mechanical phenomena.

With the aid of (4.2) and (2.3) we may derive from (5.1)

$$\tau_{\alpha\beta} = a\,\tilde{\varepsilon}_{\alpha\beta} + a'\,\tilde{Z}_{\alpha\beta} + \{b\,\varepsilon + b'\,Z + c(T - T_0)\}\,\delta_{\alpha\beta} \quad , \tag{5.2}$$

where we also used the Definitions (3.1) and (3.2) of deviator and scalar part of a tensor field. In an analogous way we obtain from (5.1) with the help of (4.3)

$$G_{\alpha\beta} = -a'\,\tilde{\varepsilon}_{\alpha\beta} - a''\,\tilde{Z}_{\alpha\beta} - \{b'\,\varepsilon + b''\,Z - c'(T - T_0)\}\,\delta_{\alpha\beta} \quad . \tag{5.3}$$

From (5.2) and (5.3) we obtain

$$\tilde{\tau}_{\alpha\beta} = a\,\tilde{\varepsilon}_{\alpha\beta} + a'\,\tilde{Z}_{\alpha\beta} \quad , \tag{5.4}$$

$$\tilde{G}_{\alpha\beta} = -a'\,\tilde{\varepsilon}_{\alpha\beta} - a''\,\tilde{Z}_{\alpha\beta} \quad , \tag{5.5}$$

$$\tau = b\,\varepsilon + b'\,Z + c(T - T_0) \quad , \tag{5.6}$$

$$G = -b'\,\varepsilon - b''\,Z + c'(T - T_0) \quad . \tag{5.7}$$

Equations of the type of (5.2) - (5.7) will be called equations of state. From (5.2), (5.4) and (5.6) it is seen that the mechanical stress depends not only on the strain $\varepsilon_{\alpha\beta}$ but also on the internal variable $Z_{\alpha\beta}$ (unless a' and b' vanish but this case will be excluded). Thus, the internal variable influences the mechanical behaviour of the material. For instance, (5.4) is a generalization of Hooke's law (3.9).

However, we are primarily interested in the relation between the strain and the mechanical stress. To this end we eliminate the internal variable $\tilde{Z}_{\alpha\beta}$ and the conjugate affinity $\tilde{G}_{\alpha\beta}$ from the flow law (4.6) and the equation of

state (5.4). Let us write (5.4) in the form

$$a' \tilde{Z}_{\alpha\beta} = \tilde{\tau}_{\alpha\beta} - a\tilde{\varepsilon}_{\alpha\beta} \ . \tag{5.8}$$

Since we assume that a and a' are constants it follows from the above relation that

$$a' \frac{d\tilde{Z}_{\alpha\beta}}{dt} = \frac{d\tilde{\tau}_{\alpha\beta}}{dt} - a \frac{d\tilde{\varepsilon}_{\alpha\beta}}{dt} \ . \tag{5.9}$$

From (4.6) we obtain with the aid of (5.5) and (5.8)

$$\begin{aligned} a' \frac{d\tilde{Z}_{\alpha\beta}}{dt} &= a' \eta' \tilde{G}_{\alpha\beta} = \\ &= -(a')^2 \eta' \tilde{\varepsilon}_{\alpha\beta} - a' a'' \eta' \tilde{Z}_{\alpha\beta} = \\ &= -(a')^2 \eta' \tilde{\varepsilon}_{\alpha\beta} - a'' \eta' (\tilde{\tau}_{\alpha\beta} - a\tilde{\varepsilon}_{\alpha\beta}) = \\ &= \{a a'' - (a')^2\} \eta' \tilde{\varepsilon}_{\alpha\beta} - a'' \eta' \tilde{\tau}_{\alpha\beta} \ . \end{aligned} \tag{5.10}$$

We now get from the two preceding equations

$$a'' \eta' \tilde{\tau}_{\alpha\beta} + \frac{d\tilde{\tau}_{\alpha\beta}}{dt} = \{a a'' - (a')^2\} \eta' \tilde{\varepsilon}_{\alpha\beta} + a \frac{d\tilde{\varepsilon}_{\alpha\beta}}{dt} \ . \tag{5.11}$$

This equation describes elastic relaxation phenomena. It is known as the Poynting-Thomson equation. The equation is also denoted as the equation for the standard linear solid.

Equations which connect strain and mechanical stress (such as (5.11)) are called stress-strain relations, mechanical constitutive equations or rheological equations.

A state of the medium is called a stationary state if the time derivatives of mechanical stress and strain vanish. For such a state (5.11) reduces to

$$\tilde{\tau}_{\alpha\beta} = \left(a - \frac{(a')^2}{a''}\right) \tilde{\varepsilon}_{\alpha\beta} \qquad \text{(stationary state)} \ . \tag{5.12}$$

It is seen that (5.12) has the structure of Hooke's law. We also note that it follows from (5.10) and (5.12) that

$$\tilde{G}_{\alpha\beta} = 0 \qquad \text{(stationary state)} \tag{5.13}$$

and that

$$d\tilde{Z}_{\alpha\beta}/dt = 0 \qquad \text{(stationary state)} \ . \tag{5.14}$$

If also $dZ/dt = 0$ it is seen from (4.9) that

$$\sigma^{(h)} = 0 \qquad \text{(stationary state)} \ . \tag{5.15}$$

Hence, in a stationary state there is no heat dissipation and the affinity conjugate to $\tilde{Z}_{\alpha\beta}$ vanishes. It may be shown that this is one of the conditions for thermodynamic equilibrium.

In an analogous way a linear relation among τ, ε, T and their first derivatives with respect to time may be derived. This relation describes volumetric phenomena. The equation (5.11) describes shear phenomena.

§6. ELASTIC AND INELASTIC STRAINS

Let us define the deviator $\tilde{\varepsilon}^{(1)}_{\alpha\beta}$ and scalar part $\varepsilon^{(1)}$ of a tensor $\varepsilon^{(1)}_{\alpha\beta}$ by

$$\tilde{\varepsilon}^{(1)}_{\alpha\beta} = -\frac{a'}{a} \tilde{Z}_{\alpha\beta} \ , \tag{6.1}$$

$$\varepsilon^{(1)} = -\frac{b'}{b} Z \ . \tag{6.2}$$

Furthermore, we define the deviator $\tilde{\varepsilon}^{(0)}_{\alpha\beta}$ and the scalar part $\varepsilon^{(0)}$ of a tensor $\varepsilon^{(0)}_{\alpha\beta}$ by

$$\tilde{\varepsilon}^{(0)}_{\alpha\beta} = \tilde{\varepsilon}_{\alpha\beta} + \frac{a'}{a}\tilde{Z}_{\alpha\beta} , \tag{6.3}$$

$$\varepsilon^{(0)} = \varepsilon + \frac{b'}{b} Z . \tag{6.4}$$

From (6.3) and (6.4) we have

$$\varepsilon^{(0)}_{\alpha\beta} = \varepsilon_{\alpha\beta} + \frac{a'}{a}\tilde{Z}_{\alpha\beta} + \frac{b'}{b} Z \delta_{\alpha\beta} \tag{6.5}$$

and from (6.1) and (6.2) we obtain

$$\varepsilon^{(1)}_{\alpha\beta} = - \frac{a'}{a}\tilde{Z}_{\alpha\beta} - \frac{b'}{b} Z \delta_{\alpha\beta} . \tag{6.6}$$

It is seen from the two preceding equations that

$$\varepsilon_{\alpha\beta} = \varepsilon^{(0)}_{\alpha\beta} + \varepsilon^{(1)}_{\alpha\beta} . \tag{6.7}$$

Thus, with the aid of the new variable $Z_{\alpha\beta}$ the strain may be split up into two parts: $\varepsilon^{(0)}_{\alpha\beta}$ and $\varepsilon^{(1)}_{\alpha\beta}$. The physical meaning of $\varepsilon^{(0)}_{\alpha\beta}$ may be demonstrated in a very simple way. Using (6.3) one may write for (5.4)

$$\tilde{\tau}_{\alpha\beta} = a\,\tilde{\varepsilon}^{(0)}_{\alpha\beta} \tag{6.8}$$

and (5.6) becomes with the aid of (6.4)

$$\tau = b\,\varepsilon^{(0)} + c(T - T_0) . \tag{6.9}$$

Hence,

$$\tau_{\alpha\beta} = a\,\tilde{\varepsilon}^{(0)}_{\alpha\beta} + \{b\,\varepsilon^{(0)} + c(T - T_0)\}\,\delta_{\alpha\beta} . \tag{6.10}$$

From (6.8) it is seen that $\tilde{\varepsilon}^{(0)}_{\alpha\beta}$ is the part of the deviatoric strain which is proportional to the deviatoric mechanical stress. Hence, $\tilde{\varepsilon}^{(0)}_{\alpha\beta}$ may be looked upon as the deviator of the elastic strain, while (6.8) is Hooke's law for distortional phenomena. In an analogous way it follows from (6.9) that $\varepsilon^{(0)}$ is the volumetric part of the elastic (or thermoelastic) strain. In other words: $\varepsilon^{(0)}_{\alpha\beta}$ is the thermoelastic strain.

Next, we shall show that $\varepsilon^{(1)}_{\alpha\beta}$, the remaining part of the total strain, may be interpreted as the inelastic strain. To this end we consider the free energy in some more detail. From (6.1) and (6.2) we have

$$\tilde{Z}_{\alpha\beta} = -\frac{a}{a'}\tilde{\varepsilon}^{(1)}_{\alpha\beta} , \qquad Z = -\frac{b}{b'}\varepsilon^{(1)} . \tag{6.11}$$

Using these relations it follows from the Expression (5.1) for the free energy that

$$f = v_0\Big[\frac{1}{2}a\sum_{\gamma,\zeta=1}^{3}\tilde{\varepsilon}_{\gamma\zeta}(\tilde{\varepsilon}_{\gamma\zeta} - 2\tilde{\varepsilon}^{(1)}_{\gamma\zeta}) + \frac{1}{2}a^*\sum_{\gamma,\zeta=1}^{3}(\tilde{\varepsilon}^{(1)}_{\gamma\zeta})^2 + \tag{6.12}$$

$$+ \frac{3}{2}b\varepsilon(\varepsilon - 2\varepsilon^{(1)}) + \frac{3}{2}b^*(\varepsilon^{(1)})^2 + 3(T-T_0)(c\varepsilon - c^*\varepsilon^{(1)})\Big] - \Psi(T) ,$$

where

$$a^* = a''(\frac{a}{a'})^2 , \qquad b^* = b''(\frac{b}{b'})^2 , \qquad c^* = -c'\frac{b}{b'} . \tag{6.13}$$

From (6.1) and (6.3) and from (6.2) and (6.4) we obtain

$$\tilde{\varepsilon}_{\alpha\beta} = \tilde{\varepsilon}^{(0)}_{\alpha\beta} + \tilde{\varepsilon}^{(1)}_{\alpha\beta} , \qquad \varepsilon = \varepsilon^{(0)} + \varepsilon^{(1)} . \tag{6.14}$$

Using (6.14) one may write for (6.12)

$$f = v_0\Big[\frac{1}{2}a\sum_{\gamma,\zeta=1}^{3}(\tilde{\varepsilon}^{(0)}_{\gamma\zeta})^2 + \frac{1}{2}(a^*-a)\sum_{\gamma,\zeta=1}^{3}(\tilde{\varepsilon}^{(1)}_{\gamma\zeta})^2 + \tag{6.15}$$

$$+ \frac{3}{2}b(\varepsilon^{(0)})^2 + \frac{3}{2}(b^*-b)(\varepsilon^{(1)})^2 + 3(T-T_0)\{c\varepsilon^{(0)} - (c^*-c)\varepsilon^{(1)}\}\Big] - \Psi(T) .$$

It is seen that in this expression for f no cross terms occur between elastic and inelastic strains.

With the aid of (6.10) it follows from the expressions (6.12) and (6.15) for f that

$$\tau_{\alpha\beta} = \rho_0 \frac{\partial}{\partial \varepsilon_{\alpha\beta}} f(T, \varepsilon_{\alpha\beta}, \varepsilon_{\alpha\beta}^{(1)}) \tag{6.16}$$

and that

$$\tau_{\alpha\beta} = \rho_0 \frac{\partial}{\partial \varepsilon_{\alpha\beta}^{(0)}} f(T, \varepsilon_{\alpha\beta}^{(0)}, \varepsilon_{\alpha\beta}^{(1)}) \ , \tag{6.17}$$

where we also used the Definitions (3.1) and (3.2) of deviator and scalar part. Furthermore, we define the affinity stress $\tau_{\alpha\beta}^{(1)}$ by

$$\tau_{\alpha\beta}^{(1)} = -\rho_0 \frac{\partial}{\partial \varepsilon_{\alpha\beta}^{(1)}} f(T, \varepsilon_{\alpha\beta}, \varepsilon_{\alpha\beta}^{(1)}) \ . \tag{6.18}$$

With the help of (6.12) it follows from this definition that

$$\tau_{\alpha\beta}^{(1)} = a \tilde{\varepsilon}_{\alpha\beta} - a^* \tilde{\varepsilon}_{\alpha\beta}^{(1)} + \{b\varepsilon - b^* \varepsilon^{(1)} + c^*(T - T_0)\} \delta_{\alpha\beta} \ . \tag{6.19}$$

Hence,

$$\tilde{\tau}_{\alpha\beta}^{(1)} = a \tilde{\varepsilon}_{\alpha\beta} - a^* \tilde{\varepsilon}_{\alpha\beta}^{(1)} \tag{6.20}$$

and

$$\tau^{(1)} = b\varepsilon - b^* \varepsilon^{(1)} + c^*(T - T_0) \ . \tag{6.21}$$

From (6.1) and (6.13) we get

$$\frac{a'}{a} a^* \tilde{\varepsilon}_{\alpha\beta}^{(1)} = -\left(\frac{a'}{a}\right)^2 a^* \tilde{Z}_{\alpha\beta} = -a'' \tilde{Z}_{\alpha\beta} \ . \tag{6.22}$$

Next, we multiply both sides of (6.20) by $-\frac{a'}{a}$. We then obtain with the help of (6.22) and (5.5)

$$-\frac{a'}{a} \tilde{\tau}_{\alpha\beta}^{(1)} = -a' \tilde{\varepsilon}_{\alpha\beta} - a'' \tilde{Z}_{\alpha\beta} = \tilde{G}_{\alpha\beta} \ . \tag{6.23}$$

Let us define the quantity η by

$$\eta = \left(\frac{a'}{a}\right)^2 \eta' \quad . \tag{6.24}$$

With the aid of (6.1), (6.23) and (6.24) we may now write for the flow law (4.6)

$$\frac{d\tilde{\varepsilon}^{(1)}_{\alpha\beta}}{dt} = \eta \tilde{\tau}^{(1)}_{\alpha\beta} \quad . \tag{6.25}$$

The tensor $\tau^{(1)}_{\alpha\beta}$ has the physical dimension of stress (see for example (6.16) and (6.18)) and $1/\eta$ has the dimension of a viscosity i.e. η has the dimension of a fluidity.

Using (6.2) and (6.13) we derive

$$\frac{b'}{b} b^* \varepsilon^{(1)} = -\left(\frac{b'}{b}\right)^2 b^* Z = -b'' Z \quad . \tag{6.26}$$

Next, we multiply both sides of (6.21) by $-\frac{b'}{b}$. We then find with the aid of (6.26), (6.13) and (5.7)

$$-\frac{b'}{b} \tau^{(1)} = -b' \varepsilon - b'' Z + c'(T - T_0) = G \quad . \tag{6.27}$$

Furthermore, we introduce the quantity η_v, which is defined by

$$\eta_v = \left(\frac{b'}{b}\right)^2 \eta'_v \quad . \tag{6.28}$$

Using (6.2), (6.27) and (6.28) we now get from (4.7)

$$\frac{d\varepsilon^{(1)}}{dt} = \eta_v \tau^{(1)} \quad . \tag{6.29}$$

In this relation $1/\eta_v$ has the dimension of a viscosity (volume viscosity), η_v is a fluidity and $\tau^{(1)}$ has the dimension of a stress (pressure). The re-

lation (6.25) is a flow law for distortional phenomena and (6.29) is a flow law for inelastic volume changes. These equations may be considered as generalizations of Lévy's law[4]. Hence, indeed $\varepsilon^{(1)}_{\alpha\beta}$ is the inelastic strain.

Using (6.23), (6.24), (6.27) and (6.28) one may write the Formula (4.10) for the heat dissipation in the form

$$\sigma^{(h)} = \eta \sum_{\alpha,\beta=1}^{3} (\tilde{\tau}^{(1)}_{\alpha\beta})^2 + 3\eta_v (\tau^{(1)})^2 , \tag{6.30}$$

or, with (6.25) and (6.29),

$$\sigma^{(h)} = \sum_{\alpha,\beta=1}^{3} \tilde{\tau}^{(1)}_{\alpha\beta} \frac{d\tilde{\varepsilon}^{(1)}_{\alpha\beta}}{dt} + 3\tau^{(1)} \frac{d\varepsilon^{(1)}}{dt} . \tag{6.31}$$

Hence,

$$\sigma^{(h)} = \sum_{\alpha,\beta=1}^{3} \tau^{(1)}_{\alpha\beta} \frac{d\varepsilon^{(1)}_{\alpha\beta}}{dt} , \tag{6.32}$$

where we used (3.3) - (3.5).

Furthermore, one has from (6.13) and (6.24)

$$a'' \eta' = a^* \eta , \qquad (a')^2 \eta' = a^2 \eta . \tag{6.33}$$

With the aid of these relations one may write for the Poynting-Thomson equation (5.11)

$$a^* \eta \tilde{\tau}_{\alpha\beta} + \frac{d\tilde{\tau}_{\alpha\beta}}{dt} = a(a^* - a)\eta \tilde{\varepsilon}_{\alpha\beta} + a \frac{d\tilde{\varepsilon}_{\alpha\beta}}{dt} . \tag{6.34}$$

If we use the description of elastic relaxation phenomena with the aid of the hidden variable $Z_{\alpha\beta}$, the number of material constants in the equations for shear phenomena is 4: a, a', a" and η' (see (5.11)). If we use the description with elastic and inelastic strains, we have three material constants: a, a^* and η (see (6.34)). The coefficients of $\tilde{\tau}_{\alpha\beta}$, $\tilde{\varepsilon}_{\alpha\beta}$ and $d\tilde{\varepsilon}_{\alpha\beta}/dt$ in (6.34) may be determined experimentally and then a, a^* and η may be calculated. The coefficients a', a" and η' cannot be found from (5.11). This is connected with the fact that the physical meaning of $Z_{\alpha\beta}$ is not specified.

§7. INEQUALITIES AND SOME FINAL REMARKS

At the end of §4 we noted that $\eta' \geq 0$ and $\eta'_v \geq 0$. These inequalities are consequences of the fact that the heat dissipation $\sigma^{(h)}$ is a nonnegative quantity. From the expression (6.30) for $\sigma^{(h)}$ one finds in an analogous way $\eta \geq 0$ and $\eta_v \geq 0$. These inequalities also follow from (4.12) with the help of (6.24) and (6.28). Though in principle $\eta = 0$ and $\eta_v = 0$ are possible as limiting cases, this possibility can be excluded for the phenomena which we have in mind. Hence, we have

$$\eta > 0 , \qquad \eta_v > 0 . \tag{7.1}$$

Furthermore, we introduce the physical assumption that for each fixed temperature T, the free energy has a minimum for certain values of the elastic and inelastic strain tensors. It is seen from (6.15) that there exists such a

minimum if

$$a^* > a > 0 \quad \text{and} \quad b^* > b > 0 \ . \tag{7.2}$$

We shall now investigate in some more detail this minimum property. To this end we write (6.15) in the form

$$f = v_0 \Big[\frac{1}{2} a \sum_{\gamma,\zeta=1}^{3} (\tilde{\varepsilon}^{(0)}_{\gamma\zeta})^2 + \frac{1}{2}(a^* - a) \sum_{\gamma,\zeta=1}^{3} (\tilde{\varepsilon}^{(1)}_{\gamma\zeta})^2 +$$

$$+ \frac{3}{2} b \{\varepsilon^{(0)} + \frac{c}{b}(T - T_0)\}^2 + \frac{3}{2}(b^* - b) \Big\{\varepsilon^{(1)} - \frac{c^* - c}{b^* - b}(T - T_0)\Big\}^2 -$$

$$- \frac{3}{2} b \{\frac{c}{b}(T - T_0)\}^2 - \frac{3}{2}(b^* - b) \Big\{\frac{c^* - c}{b^* - b}(T - T_0)\Big\}^2 \Big] - \Psi(T) \ . \tag{7.3}$$

Hence, for given T the function f has a minimum if

$$\tilde{\varepsilon}^{(0)}_{\alpha\beta} = 0 \ , \qquad \varepsilon^{(0)} = -\frac{c}{b}(T - T_0) \ ,$$

$$\tilde{\varepsilon}^{(1)}_{\alpha\beta} = 0 \ , \qquad \varepsilon^{(1)} = \frac{c^* - c}{b^* - b}(T - T_0) \ .$$

The physical meaning of this result can be found as follows. From (6.8) and (6.9) we have

$$\tilde{\varepsilon}^{(0)}_{\alpha\beta} = \frac{1}{a} \tilde{\tau}_{\alpha\beta} \ , \qquad \varepsilon^{(0)} = \frac{1}{b}\tau - \frac{c}{b}(T - T_0) \ . \tag{7.4}$$

From (6.8) and (6.20) we have with the aid of (6.14)

$$\tilde{\varepsilon}^{(1)}_{\alpha\beta} = -\frac{1}{a^* - a}\left(\tilde{\tau}^{(1)}_{\alpha\beta} - \tilde{\tau}_{\alpha\beta}\right) . \tag{7.5}$$

Finally, using (6.14) we obtain from (6.9) and (6.21)

$$\varepsilon^{(1)} = -\frac{1}{b^* - b}\left(\tau^{(1)} - \tau\right) + \frac{c^* - c}{b^* - b}\left(T - T_0\right) . \tag{7.6}$$

With the aid of (7.4), (7.5) and (7.6) one may write for (7.3)

$$f = v_0\left[\frac{1}{2a}\sum_{\gamma,\zeta=1}^{3}\left(\tilde{\tau}_{\gamma\zeta}\right)^2 + \frac{1}{2(a^* - a)}\sum_{\gamma,\zeta=1}^{3}\left(\tilde{\tau}^{(1)}_{\gamma\zeta} - \tilde{\tau}_{\gamma\zeta}\right)^2 + \right. \tag{7.7}$$

$$\left. + \frac{3}{2b}\tau^2 + \frac{3}{2(b^* - b)}\left(\tau^{(1)} - \tau\right)^2 - \frac{3}{2}\left\{\frac{c^2}{b} + \frac{(c^* - c)^2}{b^* - b}\right\}\left(T - T_0\right)^2\right] - \Psi(T) .$$

From this expression it is seen that at a fixed temperature f has a minimum if both the mechanical stress $\tau_{\alpha\beta}$ and the affinity stress $\tau^{(1)}_{\alpha\beta}$ vanish. This result can also be obtained from (6.16) and (6.18) with the aid of a well-known mathematical theorem concerning extrema of functions of several variables. Roughly speaking the result means, that energy is stored in a body as a consequence of stresses inside the body.

If $\tilde{\tau}_{\alpha\beta}$ vanishes the Poynting-Thomson equation (6.34) reads

$$\frac{d\tilde{\varepsilon}_{\alpha\beta}}{dt} = -(a^* - a)\,\eta\,\tilde{\varepsilon}_{\alpha\beta} \qquad \text{(if } \tilde{\tau}_{\alpha\beta} \text{ vanishes)} \tag{7.8}$$

and since $(a^* - a)\eta > 0$ (see (7.1) and (7.2)) the equation (7.8) indeed de-

scribes strain relaxation. It should be noted that if $\tilde{\tau}_{\alpha\beta}$ vanishes, it follows from (6.8) that also $\tilde{\varepsilon}^{(0)}_{\alpha\beta}$ vanishes and, hence, we have from (6.14) $\tilde{\varepsilon}_{\alpha\beta} = \tilde{\varepsilon}^{(1)}_{\alpha\beta}$. Therefore, in this case, (7.8) is equivalent with the flow law (6.25) (see also (7.5)).

On the other hand if $\tilde{\varepsilon}_{\alpha\beta}$ vanishes (6.34) reduces to

$$\frac{d\tilde{\tau}_{\alpha\beta}}{dt} = -a^* \eta \tilde{\tau}_{\alpha\beta} \qquad \text{(if } \tilde{\varepsilon}_{\alpha\beta} \text{ vanishes)} \tag{7.9}$$

and since $a^* \eta > 0$ (see (7.1) and (7.2)) this equation describes stress relaxation.

The inequalities (7.1) and (7.2) also play an important role in the theory of the propagation and attenuation of acoustic waves in media with elastic relaxation phenomena (dissipative media)[5,6].

It is possible to generalize the theory so that Newtonian fluids and elastic liquids are also included in the formalism[3]. Plasticity phenomena can also be described with this type of theory[4], though in this case there is a characteristic nonlinear behaviour.

Finally, we note that the specific entropy s is related to the free energy f by

$$s = -\frac{\partial}{\partial T} f(T, \varepsilon_{\alpha\beta}, \varepsilon^{(1)}_{\alpha\beta}) \tag{7.10}$$

and that the specific internal energy u is related to f by

$$u = f + Ts = f - T \frac{\partial f}{\partial T}, \tag{7.11}$$

provided T is the absolute temperature.

Eindhoven, February 20th, 1984

REFERENCES

1) J.J. Seidel, Tensorrekening, Lecture Notes no. 2.237, Department of Mathematics and Computing Science, Eindhoven University of Technology, Eindhoven (1980).

2) C. Truesdell and W. Noll, The Non-Linear Field Theories of Mechanics, Handbuch der Physik, Band III/3, Springer-Verlag, Berlin - Heidelberg - New York (1965) p. 47.

3) G.A. Kluitenberg, Application of the Thermodynamics of Irreversible Processes to Continuum Mechanics, §6 of Non-Equilibrium Thermodynamics, Variational Techniques, and Stability, R.J. Donnelly, R. Herman and I. Prigogine Eds., University of Chicago Press, Chicago Ill. (1966) 91.

4) G.A. Kluitenberg, Plasticity and Nonequilibrium Thermodynamics, CISM Lecture Notes, Springer-Verlag, Wien - New York (to be published).

5) G.A. Kluitenberg, E. Turrisi and V. Ciancio, On the Propagation of Linear Transverse Acoustic Waves in Isotropic Media with Mechanical Relaxation Phenomena due to Viscosity and a Tensorial Internal Variable, I. General Formalism, Physica 110A (1982) 361.

6) E. Turrisi, V. Ciancio and G.A. Kluitenberg, On the Propagation of Linear Transverse Acoustic Waves in Isotropic Media with Mechanical Relaxation Phenomena due to Viscosity and a Tensorial Internal Variable, II. Some Cases of Special Interest (Poynting-Thomson, Jeffreys, Maxwell, Kelvin-Voigt, Hooke and Newton Media). Physica 116A (1982) 594.

OVER DE INVERSE VAN EEN ANALYTISCHE FUNCTIE

door

Jan Koekoek

Opgedragen aan J.J. Seidel ter gelegenheid van zijn 65e verjaardag.

1. In [1, pp. 2-5] wordt het volgende bewezen. Zij $f(z)$ analytisch op de cirkelschijf $|z| < R$, zij $f(0) = 0$, $f'(0) \neq 0$, en zij $f(z) \neq 0$ voor $0 < |z| < R$. Definieer voor $0 < r < R$ de functie $m(r)$ door

$$m(r) = \min_{|z|=r} |f(z)| .$$

Dan geldt voor elke r, $0 < r < R$: $w = f(z)$ heeft op de cirkelschijf $|w| < m(r)$ een analytische inverse $z = g(w)$ met $|g(w)| < r$, $g(w) = \sum_{n=1}^{\infty} b_n w^n$, waarin de coëfficiënten b_n worden gegeven door de formule van Bürmann-Lagrange:

$$b_n = \frac{1}{n!} \left[\frac{d^{n-1}}{dz^{n-1}} \left(\frac{z}{f(z)} \right)^n \right]_{z=0} , \quad n = 1,2,3,\ldots .$$

Als S de convergentiestraal is van de reeks $\sum_{n=1}^{\infty} b_n w^n$, dan geldt dus

$$S \geq \sup_{0<r<R} m(r) .$$

Zonder bewijs volgt dan ([1, p. 5]) een intrigerende bewering:

Indien het supremum van $m(r)$ wordt aangenomen voor $r = r_0$ met $0 < r_0 < R$, dan is er een punt z_0 op de cirkel $|z| = r_0$ met $f'(z_0) = 0$ en $S = |f(z_0)|$.

Vervolgens worden ([1, pp. 6-7]) twee voorbeelden gegeven die deze bewering lijken te bevestigen.

In dit stukje zullen we eerst (§ 2) die twee voorbeelden analyseren. Daarna geven we in § 3 een voorbeeld waaruit blijkt dat de bewering niet waar is.

2. *VOORBEELD 1.* $f(z) = z(1+z)$. Er geldt $f(0) = 0$, $f'(0) = 1$, $f(z) \neq 0$ voor $0 < |z| < 1$. Voor $0 < r < 1$ is $m(r) = r(1-r) = -f(-r)$. Uit $w = f(z) = z(1+z)$ met $z = 0$ voor $w = 0$ volgt $z = g(w) = \frac{1}{2}(\sqrt{1+4w} - 1)$, waarin $\sqrt{1+4w}$ de hoofdwaarde van de wortel voorstelt. De convergentiestraal S van de Taylor-ontwikkeling rond 0 van $g(w)$ is dus gelijk aan $\frac{1}{4}$.

Het supremum van $m(r)$ op het interval $0 < r < 1$ wordt aangenomen voor $r = \frac{1}{2}$ en op de cirkel $|z| = \frac{1}{2}$ ligt het punt $-\frac{1}{2}$ met $f'(-\frac{1}{2}) = 0$, $|f(-\frac{1}{2})| = \frac{1}{4} = S$.

VOORBEELD 2. $f(z) = z\,e^{-\alpha z}$, $\alpha \in \mathbb{C}$, $\alpha \neq 0$. Er geldt $f(0) = 0$, $f'(0) = 1$, $f(z) \neq 0$ voor $z \neq 0$. Voor $r > 0$ is $m(r) = r\,e^{-|\alpha| r} = \frac{\alpha}{|\alpha|} f\left(\frac{|\alpha|}{\alpha} r\right)$. Uit $w = f(z) = z\,e^{-\alpha z}$ volgt $z = g(w) = \sum_{n=1}^{\infty} b_n w^n$ met

$$b_n = \frac{1}{n!}\left[\frac{d^{n-1}}{dz^{n-1}}\left(e^{n\alpha z}\right)\right]_{z=0} = \frac{(n\alpha)^{n-1}}{n!}, \quad n = 1,2,3,\ldots .$$

De convergentiestraal van de reeks $\sum_{n=1}^{\infty} \frac{(n\alpha)^{n-1}}{n!} w^n$ is $S = \frac{1}{|\alpha| e}$.

Het supremum van $m(r)$ op het interval $0 < r < \infty$ wordt aangenomen voor $r = \frac{1}{|\alpha|}$ en op de cirkel $|z| = \frac{1}{|\alpha|}$ ligt het punt $\frac{1}{\alpha}$ met $f'(\frac{1}{\alpha}) = 0$, $|f(\frac{1}{\alpha})| = \frac{1}{|\alpha|e} = S$.

We merken bij deze voorbeelden het volgende op. Het supremum (maximum) van $m(r)$ wordt aangenomen voor $r = r_0$, waarin r_0 het enige nulpunt is van $m'(r)$, en er geldt $m(r) = a\,f(br)$ met $a \in \mathbb{C}$, $b \in \mathbb{C}$, $|a| = |b| = 1$, a en b *onafhankelijk van* r.

Wegens $m'(r) = a\,b\,f'(br)$ is dan br_0 een punt op de cirkel $|z| = r_0$ met $f'(br_0) = 0$. Bovendien geldt $m(r_0) = |m(r_0)| = |a\,f(br_0)| = |f(br_0)|$, dus $S \geq |f(br_0)|$. Omdat $f'(br_0) = 0$ is, is de inverse functie $g(w)$ niet analytisch in het punt $f(br_0)$. Hieruit volgt $S = |f(br_0)|$.

De bewering van § 1 is waar in de voorbeelden 1 en 2. In het algemeen echter geldt $m(r) = a\,f(br)$, $a \in \mathbb{C}$, $b \in \mathbb{C}$, $|a| = |b| = 1$, waarin a en b van r afhangen.

3. *VOORBEELD 3.* $f(z) = z(z-1)(z-\varepsilon)^{-3}$, $\varepsilon = e^{\pi i/3}$. De functie $f(z)$ is analytisch op de cirkelschijf $|z| < 1$, er geldt $f(0) = 0$, $f'(0) = -1$, $f(z) \neq 0$ voor $0 < |z| < 1$. In dit voorbeeld is er geen eenvoudige uitdrukking voor $m(r)$, $0 < r < 1$.

Wegens

$$\frac{1}{m(r)} = \max_{|z|=r} \left|\frac{1}{f(z)}\right| , \quad 0 < r < 1 ,$$

is (zie [2]) $-\log m(r)$ een convexe functie van $\log r$, dus (zie [3]) een continue functie van $\log r$ op het interval $0 < r < 1$.

Hieruit volgt dat $m(r)$ continu is op het interval $0 < r < 1$.

Er geldt $m(r) \leq |f(r)|$, $0 < r < 1$, dus $\lim_{r \downarrow 0} m(r) = 0$ en $\lim_{r \uparrow 1} m(r) = 0$.

De functie $m(r)$ heeft dus op het interval $0 < r < 1$ een globaal maximum $m(r_0)$ in een punt r_0 met $0 < r_0 < 1$.

Uit § 1 volgt dan dat $w = f(z)$ op de cirkelschijf $|w| < m(r_0)$ een analytische inverse $z = g(w)$ heeft met $|g(w)| < r_0$ voor $|w| < m(r_0)$.
Voor de convergentiestraal S van de Taylor-ontwikkeling rond 0 van $g(w)$ geldt $S \geq m(r_0)$.

Er is geen punt z_0 op de cirkel $|z| = r_0$ met $f'(z_0) = 0$, want $f'(z) = -(z - \bar{\varepsilon})^2 (z - \varepsilon)^{-4}$ heeft alleen een nulpunt op de eenheidscirkel.

Uit $w = f(z) = z(z-1)(z-\varepsilon)^{-3}$, dus $w(z-\varepsilon)^3 = z^2 - z$, met $z = 0$ voor $w = 0$ volgt met behulp van de formules van Cardano (en met enige volharding)

$$z = g(w) = \varepsilon + \frac{1}{3w}\left[1 - \varepsilon\left(1 + 3iw\sqrt{3}\right)^{1/3} + \varepsilon^2\left(1 + 3iw\sqrt{3}\right)^{2/3}\right],$$

waarin $\left(1 + 3iw\sqrt{3}\right)^{1/3}$ en $\left(1 + 3iw\sqrt{3}\right)^{2/3}$ hoofdwaarden voorstellen.
Hieruit volgt $S = \frac{1}{3\sqrt{3}} = \frac{1}{9}\sqrt{3}$.

We tonen tenslotte nog aan dat $S > m(r_0)$ is. Stel daartoe $S = m(r_0)$.
Voor $|w| < \frac{1}{9}\sqrt{3}$ geldt dan $|g(w)| < r_0$ met $0 < r_0 < 1$. Met $w \to \frac{1}{9} i\sqrt{3}$, $|w| < \frac{1}{9}\sqrt{3}$ volgt hieruit $|\bar{\varepsilon}| \leq r_0$, dus $r_0 \geq 1$: tegenspraak.

Dus geldt $S > m(r_0) = \sup_{0<r<1} m(r)$.

LITERATUUR

[1] Aantekeningen bij het college Voortgezette Functietheorie, Technische Hogeschool Eindhoven, voorjaarssemester 1977.

[2] G. Pólya - G. Szegö, Aufgaben und Lehrsätze aus der Analysis, Erster Band, Springer-Verlag, Berlin, 1970 (III. Abschn., Kap. 6, Aufgaben 304, 305, Lösung 304).

[3] A.W. Roberts and D.E. Varberg, Convex Functions, Academic Press, New York, 1973 (pp. 3-6).

ON THE RANK THEOREM FOR MATRICES

by

Dono Kijne

Dedicated to J.J. Seidel on the occasion of his retirement.

On 1 September 1950, at Delft, Jaap Seidel and I met for the first time, in the so-called Buildings for Mathematics of the University of Technology. From that date, we both had been appointed to be 'instructors' in the mathematics department, together with two other young mathematicians. By destiny, the same two-person study was assigned to Jaap and me; here we shared a lot of mathematical joys and sorrows during six years.

The instructors constituted a post-war category of the university staff with the special task to assist the freshmen from the various engineering departments in their mathematical education and to preserve them from invoking private tutors supports.

In those years we discussed several didactic problems. One of these was the problem of offering an understandable proof for the rank theorem for matrices ('row rank = column rank') in an early stage of the course in matrix theory, so without the use of properties of linear transformations, e.g., the dimension theorem, and without certain suggestive but unsatisfory arguments

concerning the invariance of the column rank (= the dimension of the column space) of a matrix under the usual elementary row operations.

In this paper we present an inductive proof for the rank theorem, using the elementary row and column operations only, in particular in the composed form of pivot operations. Pivot operations are well known, e.g., in the *GAUSS-JORDAN* reduction method for the solution of a system of linear equations, and in the simplex method for linear programming.
The proof presented here turned out to be reasonably comprehensible for a considerable part of the students.

Let $A = [a_{ij}]$ be a non-zero (m,n)-matrix with $m > 1$ and $n > 1$.

DEFINITION 1. A row pivot-operation applied to A consists of the following steps:

1° select a *pivot* $a_{rs} \neq 0$ in A ;

2° divide row r by a_{rs} ;

3° subtract a_{is} times the new row r from row i for $i = 1,\ldots,m,\ i \neq r$.

COROLLARY 1. Assume that by such a pivot-operation A changes into $A' = [a_{ij}']$. Then we have:

1° $a_{rj}' = \frac{a_{rj}}{a_{rs}}$ for $j = 1,\ldots,n$; in particular, $a_{rs}' = 1$;

2° $a_{ij}' = a_{ij} - a_{is}\frac{a_{rj}}{a_{rs}}$ for $i = 1,\ldots,m,\ i \neq r$, and $j = 1,\ldots,n$; in particular, $a_{is}' = 0$ for $i = 1,\ldots,m,\ i \neq r$.

Remark 1

It is well known that the row rank of A' equals the one of A. Now it is evident that row r of A' is not a linear combination of the other rows of A'. So if row r of A' is dropped, the row rank of the matrix obtained is 1 less than the row rank of A.

DEFINITION 2. A column pivot-operation applied to A consists of the following steps:

1° select a *pivot* $a_{rs} \neq 0$ in A ;

2° divide column s by a_{rs} ;

3° subtract a_{rj} times the new column s from column j for $j = 1,\ldots,n,\ j \neq s$.

COROLLARY 2. Assume that by such a pivot-operation A changes into $A'' = [a_{ij}'']$. Then we have:

1° $a_{is}'' = \frac{a_{is}}{a_{rs}}$ for i = 1,...,m; in particular, $a_{rs}'' = 1$;

2° $a_{ij}'' = a_{ij} - a_{rj}\frac{a_{is}}{a_{rs}}$ for j = 1,...,n, j ≠ s, and i = 1,...,m ; in particular, $a_{rj}'' = 0$ for j = 1,...,n, j ≠ s.

Remark 2

The column rank of A" equals the one of A. Now column s of A" is not a linear combination of the other columns of A". So if column s of A" is dropped, the column rank of the matrix obtained is 1 less than the column rank of A.

LEMMA. Let $A = [a_{ij}]$ be a non-zero (m,n)-matrix with m > 1 and n > 1. Assume that A changes into $A' = [a_{ij}']$ by the row pivot-operation with pivot $a_{rs} \neq 0$, and that A changes into $A'' = [a_{ij}'']$ by the column pivot-operation with the *same* pivot a_{rs}. If both in A' and in A" row r and column s are dropped, the resulting matrices are identical.

Proof. For any i ≠ r and j ≠ s we have:

$$a_{ij}' = a_{ij} - a_{is}\frac{a_{rj}}{a_{rs}} = a_{ij} - a_{rj}\frac{a_{is}}{a_{rs}} = a_{ij}'' \ . \qquad \square$$

THEOREM. The row rank of an (m,n)-matrix $A = [a_{ij}]$ equals its column rank.

Proof.

1° If A = 0 then both row rank and column rank of A are equal to 0; if A ≠ 0 and m = 1 or n = 1 then both row rank and column rank of A are equal to 1. So in both cases the theorem holds.

2° Now, let $A \neq 0$, $m > 1$, and $n > 1$; perform a pivot-operation to A with some pivot $a_{rs} \neq 0$, either by rows or by columns, at choice, and then drop row r and column s.

According to Remarks 1 and 2 and the lemma, both row rank and column rank of the matrix thus obtained are 1 less than those of A, inasmuch as dropping a zero column (row) from a matrix does not influence its column (row) rank.

This procedure may be repeated a number of times, each time by applying the operations mentioned to the matrix obtained previously, at least on condition that they are still possible. In k steps ($1 \leq k \leq \min(m-1,n-1)$) either a zero matrix or a non-zero matrix with $m = 1$ or $n = 1$ remains. According to the results of 1° the theorem holds for matrix A, and we have: row rank = column rank = k if the final matrix is a zero matrix, and k+1 in the other case. □

ON OVALS IN PG(2,4) AND THE MCLAUGHLIN GRAPH

by

J.H. van Lint

Dedicated to J.J. Seidel on the occasion of his retirement.

1. INTRODUCTION

Essentially this paper is a report on an unsuccessful attempt to either construct or prove the nonexistence of the partial geometry pg(5,28,2) connected to the unique strongly regular graph srg(275,112,30,56) known as the McLaughlin graph. This attempt led to a number of results which may be worth recording.

A combinatorial argument showing that PG(2,4) can be extended to the famous 5 - (24,8,1) design was given by Lüneburg [10]. Although it is essentially combinatorial Lüneburg's proof uses the automorphism group of the plane. In Section 2 we first show by a purely combinatorial argument that meeting in an even number of points is an equivalence relation on the ovals of PG(2,4); there are three equivalence classes of size 56. In Section 3 we then show that the Baer subplanes of PG(2,4) can be split into three classes, defined by the cardinalities of the intersections of these subplanes with ovals. It is then easy to construct the extension of PG(2,4) which is the Steiner system S(5,8,24).

In Section 4 we use the standard representation of PG(2,4) using $\mathbb{F}_4$. We first

study a special configuration of four ovals in PG(2,4) from which we find a description of GQ(3,9). This special configuration is used to reconstruct Sims' description of the Gewirtz graph and a corresponding description of points and lines in PG(2,4).

The results of Sections 2 to 4 allow us to describe the vertices of the Mc Laughlin graph as certain geometrical objects in PG(2,4), adjacency depending only on the cardinality of intersections. We classify the five types of maximal cliques. Many of the known descriptions of the Mc Laughlin graph have a high degree of symmetry. Since the graph has ten times as many cliques as we need for the lines of a partial geometry, it is difficult to select suitable candidates for the lines. In similar situations, where it turned out to be possible to find the geometry, the success was often based on an "asymmetric" description of the graph which resulted in different kinds of maximal cliques. It was this idea which led to the (sometimes cumbersome) description of Section 5.

Finally, in Section 6 we indicate how some of the construction attempts started. Maybe some of the readers will see how to continue. Much of the work described in Section 6 was done jointly with David Wales. We have not given up yet, so there may be a sequel to this paper.

Several of the results and ideas in this paper are not new. Some of these can be found in the literature; for others this is not the case even though the ideas are "known". Many ideas of Section 4 also occur (without proof) in a paper by Cameron, Delsarte and Goethals [3], dedicated to J.J. Seidel on the occasion of his sixtieth birthday! Collecting these results, their proofs, and

some new ideas in one paper is the main justification for my contribution to this volume.

One of the elegant descriptions of the Mc Laughlin graph depends on a set of *equiangular lines* in an appropriate space. The series of often important contributions to the theory of strongly regular graphs made by J.J. Seidel during the past twenty years started with our joint paper [9] on equiangular lines. Furthermore, the problem of constructing pg(5,28,2) was suggested by Cameron, Goethals and Seidel in [4]. These facts provided me with the gratifying knowledge that this paper will have at least one interested reader!

2. OVALS IN PG(2,4)

A combinatorial argument showing that PG(2,4) can be extended to a 5 - (24,8,1) design was given by Lüneburg [10]. The main idea in his construction is the use of counting arguments to analyse the intersection numbers of the ovals in PG(2,4). The 168 ovals split into three subsets of size 56, which are orbits under PSL(3,4). The arguments below are of a similar nature but we make them completely combinatorial by avoiding the use of the automorphism group altogether. In fact, even the fact that PG(2,4) can be coordinatized is not used but only the *definition* of this plane, (existence therefore assumed).

Consider four points A, B, C, D in PG(2,4), no three on a line. These points determine six lines which together contain 19 points of PG(2,4), (see fig. 1).

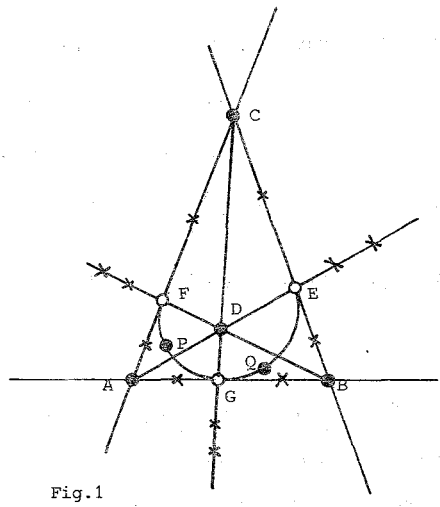


Fig.1

Since the line EF in the plane already meets four of the six lines in the figure, it contains at least one of the remaining two points P,Q. Since the same argument holds for FG and GE it follows that EFG is a line in PG(2,4), which in fact contains P and Q. It follows that {A,B,C,D,P,Q} is an *oval* and {A,B,C,D,E,F,G} form a *Baer-subplane*; (note that we use the word oval where some authors use hyperoval). These observations are of course trivial if we use coordinates. We can take the four points as (1,0,0), (0,1,0), (0,0,1) and (1,1,1). Then there are only two possibilities for P and Q, namely $(\omega,\bar{\omega},1)$ and $(\bar{\omega},\omega,1)$, where $\mathbb{F}_4 = \{0,1,\omega,\bar{\omega}\}$, etc.

We have established the following facts:

(2.1) Four points, no three collinear, determine a unique oval.

(2.2) Three " " " " three ovals.

(2.3) Through every pair of points there are twelve ovals.

(2.4) If $\mathcal{O}$ is an oval, then there are 40 ovals meeting $\mathcal{O}$ in three points.

(2.5) " $\mathcal{O}$ " " " , " " " 45 " " $\mathcal{O}$ " two "

(2.6) " $\mathcal{O}$ " " " " " " " 72 " " $\mathcal{O}$ " one point

(2.7) " $\mathcal{O}$ " " " " " " " 10 " " $\mathcal{O}$ " no "

The main result which we wish to establish in this section is the following lemma.

(2.8) *LEMMA. If we define* $\mathcal{O}_1 \sim \mathcal{O}_2$ *for ovals* $\mathcal{O}_1, \mathcal{O}_2$, *by* $\mathcal{O}_1 \sim \mathcal{O}_2$ *if* $|\mathcal{O}_1 \cap \mathcal{O}_2|$ *is even, then* $\sim$ *is an equivalence relation.*

Proof: (i) We first analyse the configuration of three ovals through a non-collinear triple of points A, B, C. Let $\mathcal{O}_1$ = {A,B,C,D,E,F}, $\mathcal{O}_2$ = {A,B,C,P,Q,R}, $\mathcal{O}_3$ = {A,B,C,U,V,W}. The remaining nine points of PG(2,4) are on the lines AB, BC, CA. (See fig.2)

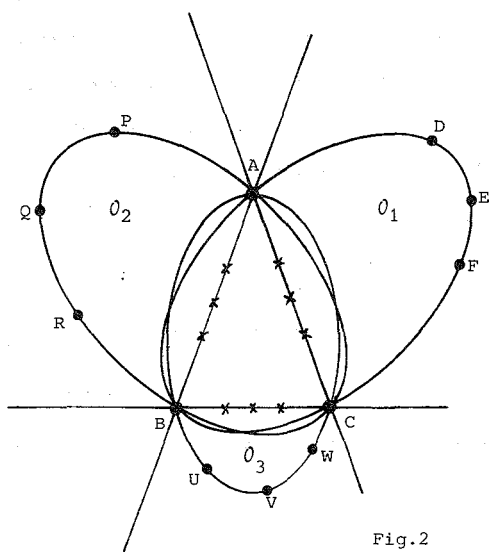


Fig.2

The secants of $\mathcal{O}_2$ through D are clearly DA, DB and DC. Hence DE does not meet $\mathcal{O}_2$. It follows that

(2.9) $\quad \mathcal{O}_1 \Delta \mathcal{O}_2$ is an oval.

(cf. Remark 2.13).

W.l.o.g. we may assume that the points have been numbered in such a way that the following sets are collinear:

(2.10) $\quad \{P,A,U,F\}\ ,\ \{R,A,V,D\}\ ,\ \{P,D,W\}\ .$

(ii) Suppose that an oval $\mathcal{O}$ meets both $\mathcal{O}_1$ and $\mathcal{O}_2$ in two points. We must distinguish between three different possibilities:

(α) $\mathcal{O} \cap \mathcal{O}_1 = \{D,E\}$, $\mathcal{O} \cap \mathcal{O}_2 = \{P,Q\}$. This is excluded by (2.9).

(β) $\mathcal{O} \cap \mathcal{O}_1 = \mathcal{O} \cap \mathcal{O}_2 = \{A,B\}$. Since $\mathcal{O}$ has at most two points on CB resp. CA, there must be two points of $\mathcal{O}$ in the set $\{U,V,W\}$. But then $\mathcal{O} = \mathcal{O}_3$ by (2.1). This contradicts the assumption.

(γ) $\mathcal{O} \cap \mathcal{O}_1 = \{A,D\}$, $\mathcal{O} \cap \mathcal{O}_2 = \{A,P\}$. By (2.10) $\mathcal{O}$ does not contain U, V, or W. Furthermore $\mathcal{O}$ has one point $\neq A,B$ on the line AB and one point $\neq A,C$ on the line AC. Therefore $\mathcal{O}$ has exactly one point on BC, a contradiction.

It follows that three ovals for which the intersections have cardinality resp. 3,2,2 cannot exist.

(iii) Suppose an oval $\mathcal{O}$ meets $\mathcal{O}_1$ in two points and $\mathcal{O}_2$ in no points. W.l.o.g. $\mathcal{O} \cap \mathcal{O}_1 = \{D,E\}$. By (2.9) $\mathcal{O}$ contains at most one of the points U, V, W. Therefore $\mathcal{O}$ contains at least three points $\notin \{A,B,C\}$ on the sides of ABC, and hence four such points. So, two sides of ABC are secants of $\mathcal{O}$ and the third secant through their point of intersection contains D, E. This contradicts the fact

that DE does not meet $\mathcal{O}_2$. Hence intersection numbers 3,2,0 cannot occur.

(iv) The same argument shows that intersection numbers 3,0,0 do not occur.

(v) Let $\mathcal{O}$ be any oval. By (2.5) and (2.7) there are 56 ovals which meet $\mathcal{O}$ in an even number of points. Let these ovals be the blocks of a "design" $\mathcal{D}$ on the 21 points of PG(2,4). By (ii), (iii), (iv) no two blocks meet in three points. Therefore there are at most four blocks through any pair of points. Since the average number of blocks through a pair of points is four, this number is 4 for every pair, i.e., $\mathcal{D}$ is indeed a design, namely a 2-(21,6,4). If α_i denotes the number of pairs of blocks which meet in i points, then we now know $\alpha_2 = \binom{21}{2} \cdot \binom{4}{2} = 1260$. Clearly $\sum i\alpha_i = 21 \cdot \binom{16}{2} = 2520$ and hence $\alpha_1 = 0$. This proves the lemma. □

(2.11) *COROLLARY.* If we take the ovals in a equivalence class of Lemma (2.8) as vertices of a graph and if we join two vertices by an edge if the corresponding ovals are disjoint.then we obtain a strongly regular graph srg(56,10,0,2). (This is one of the well known representations of the Gewirtz graph, cf. [5] page 20).

Proof. The result follows by [5] Theorem 3.2 from the fact that $\mathcal{D}$ is quasi-symmetric. □

(2.12) Remark. By Lemma 2.8 the ovals $\mathcal{O}_1,\mathcal{O}_2,\mathcal{O}_3$ of figure 2 are in three different equivalence classes. We shall call these classes the three *orbits* (or *types*) of ovals. Sometimes we denote an oval of type i by $\mathcal{O}^{(i)}$.

(2.13) Remark. The following elegant and elementary approach to (2.9) was indicated in [3]. Consider a triangle in PG(2,4), say ABC in figure 2. Let S be the set

of nine points of PG(2,4)\{A,B,C} on the sides of this triangle. A little reflection shows that these points determine twelve lines of PG(2,4), which together with S form a subplane isomorphic to AG(2,3), i.e., a unital. Each parallel class of lines in this unital determines a triangle in PG(2,4). In figure 2 these triangles are {A,B,C}, {D,E,F}, {P,Q,R}, {U,V,W}. It is now obvious that the union of two of these triangles is an oval.

3. BAER SUBPLANES IN PG(2,4) AND THE 5-(24,8,1) DESIGN

In Section 2 we saw that four non-collinear points determine a Baer subplane and an oval which meet in these four points. By standard counting we find 360 Baer subplanes in PG(2,4). These form a 2-(21,7,36) design.

Let $\mathcal{B}$ be a Baer subplane. From figure 1 we see that each line ℓ of $\mathcal{B}$ yields an oval consisting of the four points not on ℓ and the two points on $\ell\setminus\mathcal{B}$. These seven ovals pairwise meet in two points, i.e., they are all of the same type i. We shall call $\mathcal{B}$ a Baer subplane of type i in this case. Let $P_1,\ldots,P_7$ be the points of $\mathcal{B}$, P_8 to P_{21} the remaining points of PG(2,4) and let $\mathcal{O}_1,\ldots,\mathcal{O}_7$ be the seven ovals. Then figure 3 gives the adjacency matrix of the ovals and points.

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8												P_{21}
$\mathcal{O}_1$	0	0	1	0	1	1	1	1	1											
$\mathcal{O}_2$	1	0	0	1	0	1	1			1	1									
$\mathcal{O}_3$	1	1	0	0	1	0	1					1	1							
$\mathcal{O}_4$	1	1	1	0	0	1	0							1	1					
$\mathcal{O}_5$	0	1	1	1	0	0	1									1	1			
$\mathcal{O}_6$	1	0	1	1	1	0	0											1	1	
$\mathcal{O}_7$	0	1	0	1	1	1	0													1 1

Fig. 3

(3.1) *LEMMA. If O is an oval of type i and B is a Baer subplane of type j then*

$$|O \cap B| \text{ is } \begin{cases} \text{even} & \text{if } i = j, \\ \text{odd} & \text{if } i \neq j. \end{cases}$$

Proof: (i) Suppose $|O \cap B| = 1$. Then $|O \cap B \cap O_i| = 1$ for four values of i and 0 for the others. By Lemma (2.8) $|O \cap O_i|$ must have the same parity for $i = 1,\ldots,7$. From figure 3 we see that this is only possible if this parity is odd. In fact, $|O \cap O_i| = 1$ for six values of i and 3 for the remaining one.

(ii) Suppose $|O \cap B| = 2$. Now $|O \cap B \cap O_i|$ is again odd (namely 1) for four values of i. It follows that the only possibility is that $|O \cap O_i| = 2$ for six values of i, 0 for the remaining value.

(iii) If $|O \cap B| = 3$ then the intersection is not a line of B and as above we see that $|O \cap O_i| = 3$ for four values of i, 1 for the other three.

(iv) If $O \cap B = \emptyset$ then O meets three of the O_i in two points, $O \cap O_i = \emptyset$ for the other values of i. □

Lemma 3.1 and the results of Section 2 allow us to give the (well-known) description of the extension of PG(2,4) to the Steiner system by adjoining three points ∞_1, ∞_2, ∞_3. The blocks are:

(i) 21 lines of PG(2,4) with $\{\infty_1,\infty_2,\infty_3\}$ adjoined;

(ii) the ovals $O^{(i)}$ with $\{\infty_1,\infty_2,\infty_3\}\backslash\{\infty_i\}$ adjoined (i = 1,2,3);

(iii) the Baer subplanes $B^{(i)}$ with $\{\infty_i\}$ adjoined;

(iv) the $\binom{21}{2}$ sets $\ell_i \Delta \ell_j$, ℓ_i and ℓ_j lines of PG(2,4).

We have $21 + 3.56 + 3.120 + 210 = 759$ blocks. These contain $21 + 168.6 + 360 \cdot \binom{7}{2} + 210 \cdot \binom{8}{3} = \binom{21}{5}$ distinct 5-subsets of PG(2,4). To show

that we have the required Steiner system we have to consider only the 5-subsets of PG(2,4) $\cup \{\infty_1,\infty_2,\infty_3\}$ containing one or more of the adjoined points. We only treat the one non-trivial situation, the union of ∞_1 and a 4-subset of PG(2,4). If the four points are on a line, then a block of type (i) covers the 5-subset. If there are three points on a line and one not on it, then these four points derermine three Baer subplanes and by Lemma 3.1 exactly one of these has type 1 and then the 5-subset is covered by a block of type (ii). Finally, suppose we have four points, no three on a line. Then there is an oval containing these four points and a Baer subplane of the same type also containing these four points. One of these has ∞_1 adjoined to it in the set of blocks.

The results on intersections of ovals, subplanes, and lines are combined in the following table (figure 4). In the table the entry 2 (40) in the line-row and the oval i-column means that a given line meets 40 ovals of orbit i in exactly 2 points.

	line	oval i	oval j	Baer i	Baer j
line	1 (20) 5 (1)	0 (16) 2 (40)		1 (80) 3 (40)	
oval i	0 (6) 2 (15)	0 (10) 2 (45) 6 (1)	1 (36) 3 (20)	0 (15) 2 (90) 4 (15)	1 (60) 3 (60)
Baer i	1 (14) 3 (7)	0 (7) 2 (42) 4 (7)	1 (28) 3 (28)	1 (42) 3 (77) 7 (1)	0 (8) 2 (84) 4 (28)

Figure 4

4. THE FOUR-OVAL CONFIGURATION AND THE GEWIRTZ GRAPH

The following configuration is a consequence of (2.8) and (2.9), (see figure 5)

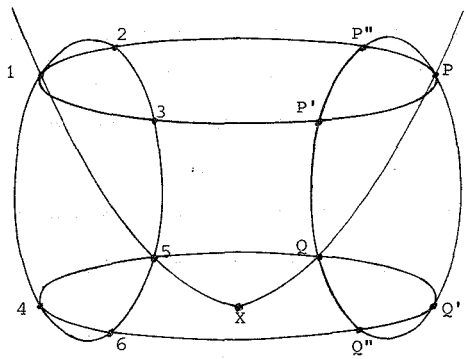


figure 5

Let $O_1^{(1)} = \{1,2,3,4,5,6\}$ be a type 1 oval. The points 1,2,3 determine a type 2 oval. Let this be $O_1^{(2)} = \{1,2,3,P,P',P"\}$. Similarly 1,2,3 determine a type 3 oval $\{1,2,3,Q,Q',Q"\}$ (cf. fig.2). Now by (2.9) the set $O_2^{(1)} = \{P,P',P",Q,Q',Q"\}$ is a type 1 oval, $O_2^{(2)} = \{4,5,6,Q,Q',Q"\}$ is a type 2 oval. In the following we shall thoroughly study these *four-oval configurations* (which by Remark 2.13 correspond to the unitals in PG(2,4)).

First, observe that each of the ten type 1 ovals which do not meet $O_1^{(1)}$ are in exactly one four-oval configuration together with $O_1^{(1)}$. The same holds for each of the twenty type 2 ovals which meet $O_1^{(1)}$ in three points. Now, consider the 112 ovals of type 1 and type 2. Let O be one of these, say of type 1, which

is not in the configuration of figure 5. Let $\alpha := |O \cap \{1,2,3\}|$, $\beta := |O \cap \{4,5,6\}|$, $\gamma := |O \cap \{P,P',P''\}|$, $\delta := |O \cap \{Q,Q',Q''\}|$. By Lemma 2.8 and Lemma 3.1 we have $\alpha \equiv \beta \pmod 2$, $\gamma \equiv \delta \pmod 2$, $\alpha \equiv \gamma + 1 \pmod 2$, $\beta \equiv \delta + 1 \pmod 2$. So, w.l.o.g. γ and δ are odd and hence $\gamma = \delta = 1$, and α and β are even. This implies that either $\alpha = \beta = 0$ or $\{\alpha,\beta\} = \{0,2\}$. In the first case O and $O_i^{(1)}$ are in some four-oval configuration; in the second case O and one of the $O_i^{(2)}$ are in a four-oval configuration.

If we take the 112 ovals as *points* in a geometry and call the four-oval configurations *lines*, then we have proved the following result.

(4.1) *LEMMA. The 112 ovals of type 1 and 2 and the four-oval configurations form a partial geometry* pg(4,10,1), *i.e.,* GQ(3,9). (cf. [2] Section 10C(ii)).

(4.2) Remark. If we also consider the type 3 ovals, then we find a system of three linked generalized quadrangles. In other words: the 168 ovals form a partial linear space with lines of size 6; leaving out one class yields GQ(3,9).

The following simple description of the Gewirtz graph is due to Sims (cf. [3] p.20). Let the vertices be ∞, the ten partitions of type (123)(456) of 1,2,3,4,5,6 into two 3-sets and the 45 pairs (ab)(cd) where a,b,c,d are four different integers $\in \{1,2,3,4,5,6\}$. Typical edges are:
∞ to (123)(456) , (123)(456) to (12)(45) and (12)(34) to (23)(56).
It is our intention to reconstruct this representation from the geometry of PG(2,4). In order to do this we introduce coordinates in PG(2,4) using $\mathbb{F}_4 = \{0,1,\omega,\bar{\omega}\}$, where $\bar{\omega} = \omega^2 = \omega + 1$. In figure 5 we take 1 = (100), 2 = (010)

3 = (001), 4 = (111), 5 = $(\omega\bar{\omega}1)$, 6 = $(\bar{\omega}\omega 1)$. The ovals $\mathcal{O}_1^{(2)}$ and $\mathcal{O}_2^{(2)}$ are now easily determined. We number the points as follows: P = $(1\bar{\omega}1)$, P' = $(\bar{\omega}11)$ P" = $(11\bar{\omega})$, Q = (11ω), Q' = $(\omega 11)$, Q" = $(1\omega 1)$.

Pick a pair, such as PQ (we have nine choices). This determines a line which in our case is {1,5,P,Q,X}, where X = $(0\bar{\omega}1)$. In this way we find the nine points of PG(2,4) not in the four-oval configuration. With the pair P,Q we now associate the partition (146)(235) of $\mathcal{O}_1^{(1)}$. The triple {1,4,6} determines the type 2 oval {1,4,6,P,(110),$(\bar{\omega}01)$} and the type 3 oval {1,4,6,Q,$(1\bar{\omega}0)$,(101)}. The symmetric difference of these is the type 1 oval $\mathcal{O}_{PQ}$:= {P,Q,(110),$(\bar{\omega}01)$,$(1\bar{\omega}0)$,(101)}, which is the type 1 oval through PQ which does not meet $\mathcal{O}_1^{(1)}$. We now give $\mathcal{O}_1^{(1)}$ the new name "∞" and call $\mathcal{O}_{PQ}$ the oval (146)(235). Of course $\mathcal{O}_2^{(1)}$ is called (123)(456). There is a unique type oval which does not meet $\mathcal{O}_2^{(1)}$ or $\mathcal{O}_{PQ}$. This is the oval containing the points 1,5, and the four points of PG(2,4) which were not used above, i.e., {1,5,$(\bar{\omega}01)$, $(\omega 01)$, (011),$(0\omega 1)$}. We must give it the name (23)(46) to be in accordance with Sims' description of the Gewirtz graph. In order to understand geometrically what is going on we must find suitable names for the points of PG(2,4) . This is easy: we already have the points 1,2,3,4,5,6 of $\mathcal{O}_1^{(1)}$. The remaining 15 are numbered according to the corresponding secants of $\mathcal{O}_1^{(1)}$, e.g., (110) is on the lines through resp. 1 and 2, 3 and 4, 5 and 6. So we say (110) = (12)(34)(56). We can now list the points of the two different kinds of type 1 ovals and see if some rule emerges:

(i) The pints of (146)(235) turn out to be the six points (1*)(4*)(6*), where the *'s are 2,3,5.

(ii) The points of (23)(46) are 1, 5 and (12)(46)(35), (13)(46)(25), (14)(23)(56), (16)(23)(45). The rule is obvious: take two missing digits and for the other four points keep one of the pairs and split the other, adding the missing digits.

It is easy to check that if we start with the 21 points numbered as above, then define as vertices: ∞ and the 6-subsets corresponding to (i), (ii) and finally call the 6-subsets adjacent if they have empty intersections, then we recover the Gewirtz graph.

We also have a description of the lines of PG(2,4) in this notation. Secants of $O_1^{(1)}$ are obvious. They look like $\{1,2,(12)(34)(56),(12)(35)(46),(12)(36)(45)\}$. The exterior lines of $O_1^{(1)}$ are more difficult. They turn out to be 1-factorisations of K_6 (numbered 1 to 6) as shown in [5] Chapter 8.

Without goning into details we remark that the other objects of PG(2,4) studied in Section 2 and 3 can also be given natural names in the terminlogy of this section. E.g., the type 2 oval $O_1^{(2)}$ would get the name (123). It consists of the points 1,2,3, (14)(26)(35), (15)(24)(36), (16)(25)(34); (here the second positions are 4,5,6 in a cyclic permutation).

5. THE MC LAUGHLIN GRAPH AND ITS CLIQUES

There are several ways to define the unque strongly regular graph srg(275,112,30,56). In [11] Mc Laughlin uses orbits of the unimodular group $U_4(3)$. There is some similarity with the definition given below. In [8] Goethals and Seidel first construct the 276-two-graph using the ternary Golay

code and find the 275-point graph from it using matrix methods. Taylor [12] uses a set of equiangular lines in the Leech lattice. Other references are [6], [7], [13]). We now define the graph using the geometrical objects studied in the previous sections.

Take as vertices: a vertex ∞, the 21 points of PG(2,4), the 21 lines of PG(2,4) the 56 ovals of type 1 and the 56 ovals of type 2, the 120 Baer subplanes of type 3. The following figure explains adjacency:

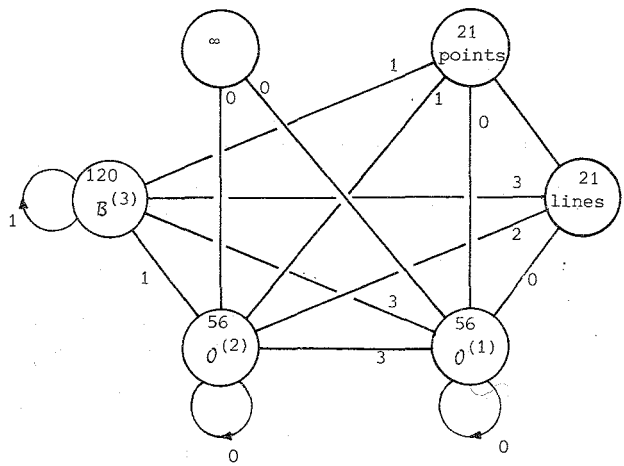


Fig. 6

Here, e.g., the 3 on the line between ovals of type 1 and Baer subplanes of type 3 means that some $\mathcal{O}^{(1)}$ is adjacent to a $\mathcal{B}^{(3)}$ if $|\mathcal{O}^{(1)} \cap \mathcal{B}^{(3)}| = 3$. Similarly a point is adjacent to a line if it is not incident with the line.

Figure 6 and figure 4 establish the fact that the graph is regular with valency 112. We also see that $\Gamma(\infty)$ is the graph of GQ(3,9) as described in Lemma 4.1. It contains two copies of the Gewirtz graph. The graph $\Delta(\infty)$ is a strongly regular graph on 162 vertices in the description given in [2] Section 10C(ii). This description has the obvious disadvantage that it is very asymetric and hence many cases have to be distinguished to prove that it is strongly regular. The reason why this may be a useful description was explained in the introduction. We shall not try to prove directly that we have indeed obtained the Mc Laughlin graph. This can be done much easier by showing that our present representation can be obtained from one of those mentioned above. Our main interest lies in a description of the different types of 5-cliques of this graph.

(5.1) Type $(\infty, O_1^{(1)}, O_2^{(1)}, O_3^{(2)}, O_4^{(2)})$

From Section 4 (fig. 5) it follows that given ∞ and an $O_1^{(1)}$ there are ten different ways to form a clique of type (5.1), namely via the ten lines of GQ(3,9) through $O_1^{(1)}$. We find 280 cliques of this type in the graph (corresponding to the 280 unitals in PG(2,4)).

(5.2) Type $(O^{(1)}, O^{(2)}, P, \ell, B^{(3)})$

These 5 cliques look like the situation of figure 7:

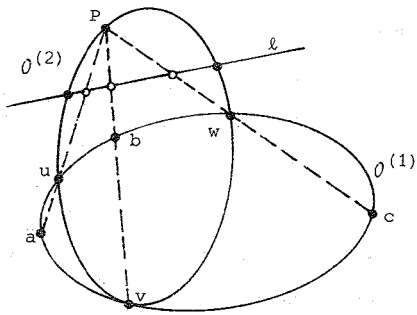


Figure 7

The Baer subplane of type 3 consists of P, the three points of ℓ not on $\mathcal{O}^{(2)}$ and the three points a, b, c of $\mathcal{O}^{(1)}$ not on $\mathcal{O}^{(2)}$. It is not difficult (using coordinates) to show that given $\mathcal{O}^{(1)}$ and $\mathcal{O}^{(2)}$ and ℓ we indeed obtain a Baer subplane in this way. Given $\mathcal{O}^{(1)}$ and $\mathcal{O}^{(2)}$ there are three choices for ℓ. It follows that an $\mathcal{O}^{(1)} - \mathcal{O}^{(2)}$ edge is in three such cliques and one of type (5.1). There are 3360 5-cliques of type (5.2).

(5.3) Type $(\mathcal{O}^{(1)}, \mathcal{O}^{(2)}, B_1^{(3)}, B_2^{(3)}, B_3^{(3)})$

The only other possibility for a 5-clique on a given $\mathcal{O}^{(1)} - \mathcal{O}^{(2)}$ edge (by figure 6) is one with three Baer subplanes. In (5.2) we used a Baer subplane meeting $\mathcal{O}^{(1)}$ in {a,b,c}. So now, w.l.o.g. the $B_1^{(3)}$ meets $\mathcal{O}^{(1)}$ in {a,b,u}, (this is one of nine possible choices). Through {a,b,u} there are two Baer subplanes of type 3, so we have 18 to choose from. The (elementary) calculations show that there are six 5-cliques of type (5.3). The total number of these is therefore 6720.

The remaining 5-cliqes must all have an $O_1^{(1)} - O_2^{(1)}$ edge or an $O_1^{(2)} - O_2^{(2)}$ edge. They are of the following types:

(5.4) Type $(O_1^{(1)}, O_2^{(1)}, P, B_1^{(3)}, B_2^{(3)})$. Total number is 2520 ,

(5.5) Type $(O_1^{(2)}, O_2^{(2)}, , B_1^{(3)}, B_2^{(3)})$. " " " 2520.

It is not difficult to show that if, for example, we start with a pair $O_1^{(1)}$, $O_2^{(2)}$ corresponding to an edge and take P not on either of the ovals, then this can be completed uniquely to a 5-clique. E.g., in figure 5 take x as the point. The $\{x,P,Q',Q'',2,3,5\}$ is easily seen to be a Baer subplane (using the coordinates given in Section 4).

(5.6) Remark. By the Hoffman bound (cf. [2]) § 4, Theorem 1) the cliques in the Mc Laughlin graph have size ≤ 5 and each point outside a clique of size 5 is joined to two points of the clique. We have pointed out that each triangle is in a 5-clique, which is therefore unique. It also follows that any 4-clique can be completed to a 5-clique. In other words, the 5-cliques are the only maximal cliques. This means that the Mc Laughlin graph is a *Zara-graph* in the sense of [1] Ch. 8. This was announced without proof in [14] and [1] p. 54.

(5.7) Remark. We could also consider co-cliques. E.g., the characterization of the regular two-graph on 276 vertices given by Taylor [12] was based on the existence of an independent set of size 23, corresponding to a co-clique of

size 22 in our graph. Two such co-cliques appear naturally in figure 6 as points ∪ {∞} resp. lines ∪ {∞}.

6. ON THE PARTIAL GEOMETRY pg(5,28,2)

As we already remarked in the introduction, the main motivation for the description of 5-cliques given in Section 5 was an attempt to construct the partial geometry of the title. In order to do this it is necessary to pick a set of 1540 maximal cliques as lines in such a way that we indeed have a partial geometry. Let us first consider the 28 lines through ∞. These are all of type (5.1). They correspond to a *matching* in the Gewirtz graph on the ovals of type 1. In this (rather disappointing) section we describe how one seems to be led in a "natural" way to such a matching, which then turns out not to be a matching at all.

In [4] it was pointed out that instead of speaking of a matching in the Gewirtz graph we could consider a *spread* of the elleptic quadric in PG(5,3). These seem to exist in great abundance, thus making our problem even more difficult.

Once again, consider figure 5. Let the four-oval configuration correspond to a line through ∞ in the partial geometry. There is a unique oval $O_3^{(1)}$ containing P and Q and joined to $O_1^{(1)}$. The line through PQ meets $O_1^{(1)}$ in the points 1 and 5. There is a unique oval $O_4^{(1)}$ through 1 and 5 and joined to $O_2^{(1)}$. Since $O_3^{(1)} - O_4^{(1)}$ is an edge in the graph, we could take this to be a "natural" candidate for a line through ∞. We now rephrase this in terms of the Gewirtz graph and then show that the idea does *not* work. Let {a, b}

be a given edge. For each x joined to a there is a unique vertex y s.t. {x,y} and {b,y} are edges (i.e., the graph has $\mu = 2$). We include the edge {x,y} in the matching.

The failure of this idea can be demonstrated using Sims' description of the Gewirtz graph. Let the edge {a,b} be {∞, (123)(456)}. Apply the above rule to (124)(356) =: x_1. We find y_1 = (12)(56). Now start with $\{x_1,y_1\}$. Take x_2:= (14)(35). Then $x_2 \sim x_1$. The rule gives us y_2 = (26)(34). Next, apply the rule to (126)(345) =: x_3. We find y_3 = (12)(45). As before, start with $\{x_3,y_3\}$ and take x_4:= (26)(34). It follows that y_4 = (15)(36) contradicting the $\{x_2,y_2\}$ matching found earlier. We did not succeed in showing that this choice for the lines through ∞ is necessary (this would prove that the partial geometry does not exist).

The following matching for the Gewirtz graph was suggested by David Wales. The fact that it is fixed under a subgroup of S_6 makes it another "natural" candidate. As before, start with a pair {∞, (123)(456)} and the second edge which we chose above, namely (124)(356) to (12)(56). The subgroup of S_6 of order 4 generated by (15 26)(34) fixes these edges. We require that all other edges of the matching are also fixed by this group. This idea produces the following matching.

(135)(246) - (13)(24)	(125)(346) - (15)(34)
(126)(245) - (54)(63)	(126)(345) - (26)(43)
(145)(236) - (23)(14)	(134)(256) - (52)(34)
(146)(235) - (64)(53)	(156)(234) - (61)(43)

(14)(26) - (12)(35)	(16)(24) - (14)(35)	
(53)(61) - (56)(42)	(51)(63) - (53)(42)	
(24)(15) - (21)(36)	(25)(14) - (24)(36)	(16)(25) - (12)(34)
(63)(52) - (65)(41)	(62)(53) - (63)(41)	(15)(26) - (56)(34)
(13)(25) - (12)(46)	(15)(23) - (13)(46)	
(54)(62) - (56)(31)	(52)(64) - (54)(31)	
(23)(16) - (21)(45)	(26)(13) - (23)(45)	
(64)(51) - (65)(32)	(61)(54) - (64)(32)	

It is not clear how to proceed once this matching has been chosen.

On the basis of conversations with others who have tried to construct pg(5,28,2) it seems safe to conjecture that the geometry does not exist.

REFERENCES

[1] Blokhuis, A., Few-distance sets, Thesis Eindhoven University of Technology, 1983.

[2] Brouwer, A.E. and J.H. van Lint, Strongly regular graphs and partial geometries, Proc. Waterloo Silver Jubilee Conference, 1982

[3] Cameron, P.J., P.Delsarte and J.M. Goethals, Hemisystems, orthogonal configurations and dissapative conference matrices, Philips Journal of Research 34 (1979), 147 - 162.

[4] Cameron, P.J., J.M. Goethals and J.J. Seidel, Strongly regular graphs having strongly regular subconstituents, J. of Algebra 55 (1978) 257 - 280.

[5] Cameron, P.J. and J.H. van Lint, Graphs, codes and designs, London Math. Soc. Lecture Note Series 43, Cambridge U.P. 1980.

[6] Conway, J.H., A group of order 8,315,553,613,086,720,000 , Bull. London Math. Soc. 1 (1969) 79 - 88.

[7] Conway, J.H., Three lectures on exceptional groups, in M.B. Powell and G. Higman, eds., Finite simple groups, Academic Press, London 1971.

[8] Goethals, J.M. and J.J. Seidel, The regular two-graph on 276 vertices, Discrete Math. 12 (1975) 143 - 158.

[9] Lint van, J.H. and J.J. Seidel, Equilateral point sets in elliptic geometry, Proc. K.N.A.W. Ser. A 69 (1966) 335 - 348.

[10] Lüneburg, H., Transitive Erweiterungen endlicher Permutationsgruppen, Lecture Notes in Math. 84, Springer Verlag, Berlin 1969.

[11] Mc Laughlin, J., A simple group of order 898,128,000, in R. Brauer and C.H. Sah, eds., Theory of Finite Groups, Benjamin New York 1969, 109 - 111.

[12] Taylor, D.E., Regular two-graphs, Proc. London Math. Soc. 35 (1977) 257 - 274.

[13] Tits, J., Groupes finis simples sporadiques, Séminaire Bourbaki No.375.

[14] Zara, F., Graphes liés aux espaces polaires (preprint).

PSEUDO-TOEVALSGETALLEN

door

W.H.J.H. van Meeuwen

Opgedragen aan J.J. Seidel ter gelegenheid van zijn afscheid van de T.H.-Eindhoven.

1. INLEIDING

In het boekje Automatische Gegevensverwerking, Verhage/Vonk [4], geschreven ten behoeve van het in ontwikkelijk zijnde programma wiskunde-A voor de bovenbouw van het V.W.O., wordt aangegeven hoe met een niet-programmeerbare zakrekenmachine pseudo-toevalsgetallen gegenereerd kunnen worden. Het besproken algoritme komt op het volgende neer:

$$x_{n+1} = 23 \cdot x_n \pmod{10^6} \qquad (n = 0,1,2,\ldots) \ . \tag{1}$$

Voorgeschreven wordt van een getal uit deze rij alleen de eerste twee cijfers te gebruiken. Daarbij dient x_n beschouwd te worden als een getal van zes cijfers, wat wij voortaan steeds ook zullen doen. Aangestipt wordt hoe x_0 geschikt gekozen kan worden, en men deelt mee dat de rij periodiek is met periode 25000 bij geschikte keuze van x_0.

In een artikel van A. Goddijn [1], bedoeld als achtergrondinformatie voor docenten bij Automatische Gegevensverwerking wordt nagegaan dat x_0 relatief priem met 10^6 gekozen dient te worden. We vinden ook uiteengezet hoe de periode berekend kan worden, en meegedeeld wordt dat van de rij met $x_0 = 7$ door

berekening van x_1 t/m x_{24999} is vastgesteld met welke frequenties combinaties van twee cijfers als begincijfers van de getallen in een cykel (d.w.z. een serie getallen ter lengte van de periode) voorkomen. Elke combinatie van twee cijfers, waarvan het tweede cijfer even is, komt voor met een frequentie 252; elke combinatie van twee cijfers met het tweede cijfer oneven, komt voor met een frequentie 248. (Met een even cijfer (oneven cijfer) zullen we steeds een cijfer bedoelen dat een even (oneven) getal voorstelt.)

2. Laat $k \in \mathbb{N}$ zijn. ($\mathbb{N} := \{1,2,3,\ldots\}$.)
Definieer $\Phi(10^k)$ door $\Phi(10^k) := \{n \in \mathbb{N} \mid n < 10^k \text{ en } ggd(n,10^k) = 1\}$. Met de vermenigvuldiging modulo 10^k als productoperatie is $\Phi(10^k)$ een groep (met $4 \cdot 10^{k-1}$ elementen). Laat $a \in \Phi(10^6)$ zijn en definieer G_a door $G_a = \{a^n \mid n \in \mathbb{N}\}$. Modulo 10^6 rekenend kunnen we G_a als een cyclische ondergroep van $\Phi(10^6)$ beschouwen. Het aantal elementen is dan een deler van $4 \cdot 10^5$, de orde van $\Phi(10^6)$. Zoals ook in Goddijn [1] beschreven is, behoeft men hoogstens voor $0 \leq k \leq 7$ en $0 \leq j \leq 5$ de getallen $23^{2^k 5^j}$ modulo 10^6 uit te rekenen om te kunnen vaststellen dat de orde van G_{23} 25000 is: In Tabel 1 blijkt dat 23^{25000} de laagste positieve macht van 23 is die modulo 10^6 gelijk is aan 1.

j\k	0	1	2	3	4	5	6	7
0	23	529	279841	985281	648961	379521	189441	892481
1	436343	213649	895201	830401	820801	281601	123201	486401
2	689943	343249	876001	752001	504001	8001	16001	32001
3	757943	591249	380001	760001	520001	40001	80001	160001
4	97943	831249	900001	800001	600001	200001	400001	800001
5	797943	31249	500001	1	1	1	1	1

Tabel 1. $23^{2^k 5^j} \pmod{10^6}$.

Er zijn dus 16 nevenklassen van G_{23} in $\Phi(10^6)$. Afhankelijk van de keuze van x_0 doorloopt de in de inleiding gegeven rij één der nevenklassen.

We stellen ons tot taak vast te stellen met welke frequenties de cijfercombinaties 00 t/m 99 als begincijfers van de getallen in een cykel voorkomen in alle zestien mogelijke gevallen. We zullen eerst aantonen dat daarvoor slechts een serie van 500 getallen onderzocht behoeft te worden.

Uit (1) volgt direct een expliciete uitdrukking voor x_n. $x_n = 23^n x_0 \pmod{10^6}$ $(n = 0,1,2,\ldots)$. Door alleen op de laatste vier cijfers van de getallen in Tabel 1 te letten lezen we af dat deze rij modulo 10^4 gezien een periode 500 heeft. Omdat $23^{500} = 380001 \pmod{10^6}$ geldt $x_{n+500} - x_n = 23^{500}x_n - x_n = 380000x_n \pmod{10^6}$. $(x_{n+500} - x_n)/10^4$ is dus een even getal. Het tweede cijfer van x_n en het tweede cijfer van x_{n+500} zijn óf beide even óf beide oneven. De deelrij $x_n, x_{n+500}, x_{n+2.500}, \cdots, x_{n+49.500}$ bestaat uit 50 verschillende getallen die in de laatste vier cijfers overeenkomen. Als het tweede cijfer van x_n even (oneven) is, komt dus elk van de 50 combinaties van twee cijfers met het tweede cijfer even (resp. oneven) precies één maal in deze deelrij als begincijfers van een getal voor. Laat f_i het aantal malen zijn dat de combinatie van twee cijfers die het getal i voorstelt, voorkomt als begincijfers van de getallen in een cykel $(0 \leq i \leq 99)$. Dan is $f_0 = f_{2i}$ en $f_1 = f_{2i+1}$ voor alle $i \in \{0,1,2,\ldots,49\}$, en om f_0 en f_1 te bepalen, behoeven we slechts een serie van 500 getallen te onderzoeken: Het aantal getallen daarin met het tweede cijfer even is f_0 en $f_1 = 500 - f_0$.

We zoeken nu representanten van de zestien nevenklassen. Uit Tabel 1 blijkt dat G_{23}, modulo 10^2 gerekend, uit 20 elementen bestaat. Gemakkelijk valt na te rekenen dat het de 20 getallen uit $\Phi(10^2)$ met even begincijfer zijn.

Voorts merken we het volgende op: Omdat $23^{20} = 895201 \pmod{10^6}$ geldt $x_{n+20} - x_n = 23^{20}x_n - x_n = 895200x_n \pmod{10^6}$. Nu is $800 \mid 895200x_n$ en $800 \mid 10^6$ dus $800 \mid (x_{n+20} - x_n)$.

Anders gezegd, de eerste vier cijfers van x_n en van x_{n+20} vormen getallen die modulo 8 gelijk zijn. De deelrij $x_n, x_{n+20}, x_{n+2.20}, \cdots, x_{n+1249.20}$ bestaat uit 1250 verschillende getallen. Er zijn precies 1250 getallen van zes cijfers waarvoor het volgende geldt: De laatste twee cijfers zijn gelijk aan die van x_n en de eerste vier vormen een getal dat modulo 8 met het getal gevormd door de eerste vier cijfers van x_n overeenkomt. Deze 1250 getallen komen dus alle in de deelrij voor. Bovenstaande overwegingen garanderen dat de zestien in Tabel 2 gekozen beginwaarden rijen opleveren die de zestien nevenklassen doorlopen. De daarbij gevonden f_0 is vermeld.

beginwaarde	f_0	beginwaarde	f_0
1	252	11	251
101	249	111	248
201	249	211	251
301	251	311	249
401	248	411	249
501	251	511	252
601	251	611	250
701	249	711	251

Tabel 2.

3. De keuze a = 23 teneinde pseudo-toevalsgetallen met behulp van het algoritme $x_{n+1} = a \cdot x_n \pmod{10^6}$ te genereren, wordt noch in Automatische Gegevensverwerking noch in het artikel van Goddijn ter discussie gesteld. Wel wordt het belang van een grote periode benadrukt en dan rijst de vraag of de keuze a = 23 in die zin wel optimaal is. Het blijkt niet zo te zijn. In $\Phi(10^k)$ bestaan, voor $k \geq 4$, cyclische ondergroepen met een orde maximaal $5 \cdot 10^{k-2}$. (Zie bijv. Ore [3, Ch. 12].) Voor elke $k \geq 5$ is de collectie getallen die een cyclische ondergroep van $\Phi(10^k)$ met maximale orde voortbrengt modulo 200 dezelfde. (Zie Knuth [2, blz. 20].) Deze collectie bestaat modulo 200 uit de getallen 3,11,13,19,21,27,29,37,53,59,61,67,69,77,83,91 en hun tegengestelden. Voor k = 4 blijkt de keuze nog groter. Dan voldoet bijv. ook nog 23, zoals we reeds eerder uit Tabel 1 opmerkten. Vanuit praktisch oogpunt zal bij keuze van een rij om pseudo-toevalsgetallen te genereren een factor twee in de lengte van de periode er wellicht niet veel toe doen. Aan de andere kant is het voor de hand liggend a zo te kiezen dat de periode maximaal is, tenzij aangetoond wordt dat een andere keuze in bepaalde opzichten voordelen biedt.

Zo een optimale keuze van a vereenvoudigt bovendien het bepalen van de frequenties f_i: Laat a zo gekozen zijn en beschouw de deelrij $x_n, x_{n+500}, x_{n+2.500}, \cdots, x_{n+99.500}$. De 100 getallen in deze deelrij zijn verschillend maar hebben alle dezelfde laatste vier cijfers. De 100 mogelijke combinaties van twee cijfers komen dus alle precies één maal als begincijfers van een getal uit deze deelrij voor. Dus $f_i = 500$ voor $i = 0,1,2,\ldots,99$.

4. De gevonden "mooie" waarden van f_i storen het toevalskarakter van een cykel van de door ons bekeken rijen sterk. We komen daar nog op terug. Wel is het mogelijk dat een serie getallen die relatief klein is ten opzichte van de periode, een "voldoende toevalskarakter" heeft. Knuth meldt in [2] dat men in de praktijk een rij op een half dozijn manieren test. Hij benadrukt dat een bevredigend toevalskarakter bij een aantal testen geen garantie biedt op het doorstaan van nog een test. De door ons onderzochte rijen hebben modulo 10^4 de betrekkelijk kleine periode 500. Om deze reden alleen al lijkt het verstandig van de getallen uit die rijen alleen de eerste twee cijfers te bekijken. Toepassing van de twee eenvoudigste door Knuth in [2] beschreven statistische toetsen blijft ook onder die beperking mogelijk. Het zijn beide χ^2-toetsen.

Neem aan dat bij een zeker experiment de uitkomstenverzameling U gepartitioneerd is in k + 1 verzamelingen $U_0, U_1, \ldots, U_k$ en dat $P(U_i) = p_i$ $(0 \leq i \leq k)$. Beschouw n (onafhankelijke) herhalingen van dit experiment en laat T_i het aantal keren zijn dat daarbij gebeurtenis U_i optreedt. De stochast T_i heeft dan een verwachtingswaarde $n\,p_i$. Definieer de stochast V_T door

$$V_T := \sum_{i=0}^{k} \frac{(T_i - n\,p_i)^2}{n\,p_i} \qquad (2)$$

dan geldt ook $V_T = \frac{1}{n} \sum_{i=0}^{k} \left(\frac{T_i^2}{p_i}\right) - n$ en V_T heeft (bij benadering) een χ^2-verdeling met k vrijheidsgraden. Als vuistregel geldt dat de benadering bruikbaar is wanneer $n\,p_i \geq 5$ voor alle $i \in \{0,1,\ldots,k\}$.

DE FREQUENTIETOETS. Laat $x_1,x_2,\ldots,x_n$ een serie van n pseudo-toevalsgetallen zijn. Laat Y_i de stochast zijn die aangeeft hoe vaak de combinatie van twee cijfers die het getal i voorstelt, als begincijfers van de getallen in deze serie voorkomt $(0 \leq i \leq 99)$. Onder de veronderstelling dat de getallen uit de serie een toevalskarakter hebben, heeft V_Y, gedefinieerd als in (2) met $k = 99$ en $p_i = \frac{1}{100}$, een χ^2-verdeling met 99 vrijheidsgraden.

DE VOLGORDETOETS. Laat $x_1,x_2,\ldots,x_{2n}$ een serie van 2n pseudo-toevalsgetallen zijn. Laat Z_i de stochast zijn die aangeeft hoe vaak in de n paren getallen (x_{2j-1}, x_{2j}) $(1 \leq j \leq n)$ het begincijfer van x_{2j-1} gevolgd door het begincijfer van x_{2j} het getal i voorstelt $(0 \leq i \leq 99)$. Onder de veronderstelling dat de getallen in deze serie een toevalskarakter hebben, heeft V_Z, gedefinieerd als in (2) met $k = 99$ en $p_i = \frac{1}{100}$, een χ^2-verdeling met 99 vrijheidsgraden.

Over blijft de keuze van n. De eerder genoemde vuistregel ($n\,p_i \geq 5$ voor alle $i \in \{0,1,\ldots,k\}$) geeft bij beide toetsen $n \geq 500$. Bij problemen aangepakt met een niet-programmeerbare zakrekenmachine zullen niet gauw series van meer dan 500 pseudo-toevalsgetallen gebruikt worden. Wanneer deze toetsen op een bepaalde serie getallen worden toegepast, is het mogelijk dat het niet-toevalsgedrag van een deelserie niet wordt gesignaleerd bijvoorbeeld doordat de "te kleine" waarden van de summanden uit die deelserie door "te grote" waarden van andere summanden worden gecompenseerd. We deden daarom de volgordetoets voor n = 500, en de serie van 1000 pseudo-toevalsgetallen die daarvoor gegenereerd dient te worden, gebruikten we tevens voor een frequentietoets met n = 1000. In Tabel 3 staan de waarden van V_Y en V_Z voor a = 21,23,27 bij vijf

vrij willekeurig gekozen beginwaarden. Toetsen we tweezijdig met een onbetrouwbaarheid van 5%, dan zijn 73,3 en 128,5 de kritieke waarden. (Zie daarvoor een tabel van de χ^2-verdeling.) De onderstreepte waarden geven dan aanleiding het toevalskarakter in twijfel te trekken. De resultaten geven geen aanleiding a = 23 te verkiezen boven a = 21 of a = 27.

Passen we de toetsen toe op een cykel dan vinden we zowel voor V_Y als voor V_Z waarden die voor een χ^2-verdeling met 99 vrijheidsgraden uiterst onwaarschijnlijk zijn. En dit bevestigt het niet-toevalskarakter van een cykel. (Zie Tabel 4.) Dat dan $V_Y = 0$ bij a = 21 en a = 27, kan met behulp van de eerder voor deze rijen berekende f_i direct worden ingezien.

Frequentietoets (n = 1000) en volgordetoets (n = 500).

x_0\a	21	21	23	23	27	27
	V_Y	V_Z	V_Y	V_Z	V_Y	V_Z
1	94,2	100	90,6	112,4	97,4	86,8
314159	86,6	102	<u>56,6</u>	95,2	107	<u>130,4</u>
271827	102,6	104,8	<u>133</u>	<u>132,4</u>	110,2	107,6
485573	118,8	105,2	112,6	118,4	101,6	118,8
513211	92,8	124	82,4	96,8	110,2	124,8

<u>Tabel 3.</u>

Frequentietoets en volgordetoets voor cykel.

a	V_Y		V_Z (bij $x_0 = 1$)	
21	0	(n = 50.000)	518,4	(n = 25.000)
23	1,6	(n = 25.000)	495,5	(n = 12.500)
27	0	(n = 50.000)	733,4	(n = 25.000)

<u>Tabel 4.</u>

LITERATUUR

1. Goddijn, A., Pseudo-toevalsgetallen. Docentenboek voor de Hewet-nascholingscursus 1983/84.

2. Knuth, D.E., The art of computer programming, Vol. II, Seminumerical algorithms (2nd edition), Addison-Wesley, 1981.

3. Ore, O., Number theory and its history, McGraw-Hill, 1948.

4. Verhage, H. en G. Vonk, Automatische Gegevensverwerking (2e herziene versie), Vakgroep OW en OC, R.U. Utrecht, 1983.

UNIFORM CONTINUITY AND THE CONTINUITY OF COMPOSITION

by

J.W. Nienhuys

Dedicated to J.J. Seidel on the occasion of his retirement.

1. INTRODUCTION

Continuity has a nice permanence property. If we compose continuous maps, the composition continues to be continuous. Hence, operations like addition and multiplication yield continuous results if the operands are continuous, and spaces of continuous maps therefore often are linear spaces. If we introduce limit operations that preserve continuity too, these linear spaces may become topological vector spaces. In this way composition of maps

$$(g,f) \mapsto g \circ f$$

can be interpreted as a map between topological vector spaces. It then makes sense to ask whether this map is continuous. Similar statements apply to differentiability. However, in that case we must restrict ourselves to Banach spaces, otherwise it is difficult to speak about derivatives. It will come as no surprise that continuity of the composition map depends on continuity of the operands f and g. For differentiability similar claims can be substantiated.

However, g should not only be continuous, but also uniformly continuous. In case of differentiability, the highest derivative of g should be uniformly continuous.

These statements will be made more precise later on, but first we discuss an example.

2. EXAMPLE

For any set A contained in $\mathbb{R}$, we denote the set of bounded continuous real functions on A by C_A. We provide C_A with the usual norm

$$\|f\|_A := \sup_{x \in A} |f(x)| \qquad \text{for all } f \in C_A .$$

In this way C_A becomes a Banach space.

For any $f \in C_A$, let us denote the ball

$$\{g \in C_A \mid \|f - g\|_A < r\}$$

by $B_A(f,r)$. If c is a constant, then $B_A(c,r)$ will denote the same as $B_A(f_c,r)$, where $f_c : x \mapsto c$ is a constant function. So, for instance, $B_A(0,r)$ is an open ball around the zero element of C_A. For all $\varepsilon > 0$, let $\mu_\varepsilon : (0,2) \to (0,2)$ be defined as follows:

$$\mu_\varepsilon(x) := \min(x/\varepsilon, 1) \qquad \text{for } x \in (0,2) ,$$

hence, $\mu_\varepsilon \in B_{(0,2)}(1,2)$. Let furthermore $\varphi \in C_{\mathbb{R}^+}$ satisfy $\varphi(x) = \sin(1/x)$

for all $x > 0$. Then the map COMP

$$\text{COMP} : C_{\mathbb{R}^+} \times B_{(0,2)}(1,2) \to C_{(0,2)}, \quad (g,f) \mapsto g \circ f$$

is not continuous at (φ,μ_1). Indeed, if we assume $\varepsilon = (1 + 2/k)$ and $x = 2/(n\pi)$, where n and k are odd positive integers, and where n is a multiple of k, then

$$|\varphi \circ \mu_\varepsilon(x) - \varphi \circ \mu_1(x)| = 2$$

hence,

$$\| \text{COMP}(\varphi,\mu_\varepsilon) - \text{COMP}(\varphi,\mu_1) \|_{(0,2)} = 2 ,$$

whereas $\| \mu_\varepsilon - \mu_1 \|_{(0,2)} \leq |1 - \varepsilon|$, for $\varepsilon > 0$.

The trick is that φ is not uniformly continuous.

Indeed, for $g_0 \in C_{\mathbb{R}^+}$ uniformly continuous and $f_0 \in B_{(0,2)}(1,2)$ arbitrary, COMP is continuous at (g_0,f_0) in both variables simultaneously. Speaking very freely, if we wiggle f in $g \circ f$ a little we would like $g \circ f$ to change little as well. But a small wiggle in f may affect any f(x) in the domain of g. Speaking more formally, it is our choice of the topology of uniform convergence on sets of functions that makes uniform continuity of g necessary for continuity of the composition operation. The choice of this topology of uniform convergence is just about the only choice we have if we want limit operations to preserve continuity.

It took mathematicians more than half a century to realize the importance of uniformity for limit processes.

3. HISTORICAL DIGRESSION

Uniform convergence came into mathematics when mathematicians started to direct their attention to limit processes involving 'arbitrary' functions. Such limit processes occur in Fourier's prize article (Mémoire sur la propagation de la Chaleur, 1812). Fourier shows how arbitrary functions ("discontinue et entièrement arbitraire") are limits of trigonometric functions.

Cauchy, to the contrary, shows in 1821 in his Cours d'Analyse de l'Ecole Royale Polytechnique that a limit of continuous functions must be continuous. Imre Lakatos [3] has suggested that Cauchy's proof was meant to imply that series like $\sin x - \frac{1}{2}\sin 2x + \frac{1}{3}\sin 3x - \ldots$ did not converge at all. Abel proves in his memoir on the binomial series (1826) that the limit of a power series $\sum a_n x^n$ is continuous on $[0,1]$ if the powerseries converges in 1. He also points out Cauchy's error. In 1837 P.G. Lejeune Dirichlet ("Ueber die Darstellung ganz willkürlicher Funktionen durch Sinus- und Cosinusreihen") gives precise convergence proofs for Fourier series of functions that satisfy what are now known as Dirichlet conditions.

In 1847 Philip Ludwig Seidel analyses Cauchy's error and proves that if a series of continuous functions has a discontinuous sum then there is a ρ such that the number of terms necessary to approximate the sum up to ρ increases beyond all bounds when we let the independent variable approach the point of discontinuity. He expresses this [4] by saying that "[man muss] in der unmittelbare Umgebung der Stelle, wo die Funktion springt, Werthe von x angeben können, für welche die Reihe *beliebig langsam* convergirt". He

remarks that this phenomenon of increasingly slow convergence also can occur when the limit is continuous, e.g. in the powerseries for e^x. If one takes for example $x = 1000000$, then even if one takes a million terms, one will make an enormous error.

About the same time G.G. Stokes in England makes similar remarks.

Whereas in 1833 Cauchy had not deemed it necessary to change his theorem, he writes in 1853 that it was erroneous but "il est facile de voir comment on doit modifier l'énoncé du théorème". Meanwhile, in 1838, Christoph Gudermann remarks that certain series expansions of elliptic functions converge "im ganzen gleichen Grad", but one of his students, Karl Weierstrass, gives in 1841 a rigorous definition of uniform convergence. In 1861 Weierstrass defines "Konvergenz in gleichem Grade", echoing Gudermann's terminology. He also gives the familiar proof by cutting ε in three and he proves the theorem about differentiability of the limit if the derivatives converge uniformly too.

In 1869 a friend of Weierstrass, E. Heine, publishes the theorem about integrals of uniformly convergent series (Weierstrass had been teaching that theorem for years). It was also Heine who introduced the notion of uniform continuity in 1872. More details about this history can be found in [1], Ch. II.

4. OUTLINE OF SEQUEL

In the sequel we discuss in what respect the composition operation is differentiable.

For this purpose we introduce Banach spaces of functions that are p times differentiable, with bounded continuous p-th derivatives. These Banach spaces will be denoted by $\mathcal{D}^{(p)}$. The composition operation can then be interpreted as a map

$$\mathrm{COMP}_k^p : \mathcal{D}^{(p)} \times \mathcal{D}^{(p)} \to \mathcal{D}^{(p-k)}, \quad \text{for } 0 \leq k \leq p .$$

One of the results will be that COMP_0^p is continuous at any (g,f) where $D^p g$ (the p-th derivative of g) is uniformly continuous.

Hence COMP_1^p is continuous, and for $2 \leq k \leq p$, COMP_k^p is k times differentiable.

These facts must be well-known. As a matter of fact, it is an exercise in Dieudonné's Foundations of Modern Analysis ([2], Ch. 8.9). However, a neat proof is difficult to locate. I meant to present such a proof here. Basically it consists of carefully keeping track of inequalities for higher derivatives, which are, after all, just covariant symmetric tensors.

5. PRELIMINARY DEFINITIONS

In Banach spaces we indicate norms generally by $\| \ \|$; in the notation we often make no distinction between norms of different spaces.

For instance, if E_1 and E_2 are Banach spaces, the norm on the space $E_1 \times E_2$ will be supposed to satisfy:

$$\| (a,b) \| = \max(\| a \| , \| b \|) \quad , \quad \text{for } (a,b) \in E_1 \times E_2 \ .$$

Obviously, the three occurrences of the symbol $\| \ \|$ here refer to different norms. The space of bounded linear mappings $E_1 \to E_2$ is denoted by $L(E_1,E_2)$ and it is normed by

$$\| \ell \| = \sup_{x \neq 0} \| \ell(x) \| / \| x \| \ .$$

Let U be an open subset of a Banach space E_1 and let $f : U \to E_2$, E_2 a Banach space. We say that f is differentiable at $x \in U$ if and only if there exists an $\ell \in L(E_1,E_2)$ such that

$$\lim_{\| h \| \to 0} \| f(x+h) - f(x) - \ell(h) \| / \| h \| = 0 \ .$$

If f is differentiable at x then we write $(Df)(x)$ instead of ℓ. If f is differentiable at every point of U, then Df is a map $U \to L(E_1,E_2)$, which may or may not be differentiable at every point of U. In the first case we write D^2f and $D^2f(x)(h_1,h_2)$ instead of $D(Df)$ and $(((DDf)(x))(h_2))(h_1)$ respectively. It is well-known that D^2f is a symmetric bilinear function. Similarly for higher derivatives.

Let E_1 and E_2 be Banach spaces and let U be an open subset of E_1. The set of maps $U \to E_2$ that are bounded and that are p times continuously differentiable and that have bounded p-th derivative is denoted by $\mathcal{D}_U^{(p)}(E_2)$. This is a linear space. We provide it by a norm as follows:

$$\|D^k f\|_U \doteq \sup_{x\in U} \|(D^k f)(x)\| \quad , \quad \text{for } 0 \le k \le p \tag{1}$$

and

$$\|f\|_p = \max(\|f\|_U, \|Df\|_U, \ldots, \|D^p f\|_U) \quad , \tag{2}$$

where $D^0 f$ denotes f, and where $f \in \mathcal{D}_U^{(p)}(E_2)$.

In this way $\mathcal{D}_U^{(p)}(E_2)$ becomes a Banach space, with $\| \ \|_p$ as norm. Usually, however, we shall omit the indexes U and p when there is no danger of confusion.

We remark that on the right hand side of (1), there occur norms of k-linear continuous map from E_1 to E_2.

DEFINITION 1. Let f be a continuous map defined on an open set U of a Banach space E_1 and with image in a Banach space E_2. A function $C : \mathbb{R}^+ \to \mathbb{R}$ is called a modulus of continuity for f if and only if

(i) $\quad \forall_{\delta>0} \forall_{x\in U} \forall_{y\in U} \ (\|x - y\| < \delta \to \|f(x) - f(y)\| < C(\delta))$

and

(ii) $\quad \lim_{\delta \downarrow 0} C(\delta) = 0 \ .$

EXAMPLE. For a differentiable map g with bounded derivative, the map $\delta \to \|Dg\|\delta$ is a modulus of continuity. A subadditive continuous function on $\mathbb{R}^+ \cup \{0\}$ is its own modulus of continuity if it is zero for zero argument.

FUNCTIONAL NOTATION. In the sequel we shall use the symbol $\Big\uparrow$ as follows: $\Big\uparrow_a b$ is to be read as: the mapping that assigns b to a. The notation $\Big\uparrow_{a\in A} b$ means that moreover the domain of this mapping is defined to be A.

6. AUXILIARY LEMMAS

LEMMA 2. Let E, F and G be Banach spaces, let U be open in F, let B(0,R) denote the ball around 0 in L(E,F). Let g be in $\mathcal{D}_U^{(k+1)}(G)$ and let $\gamma : U \times B(0,R) \to L(E,G)$ be defined as

$$\bigcup_{(x,\ell)} \bigcup_{y\in E} ((Dg)(x))(\ell(y)) \ .$$

Then

(i) $$((D^k\gamma(x,\ell))((\xi_1,h_1),\ldots,(\xi_k,h_k)))(y) =$$

$$= (D^{k+1}g)(x)(\xi_1,\ldots,\xi_k,\ell(y)) + \sum_{i=1}^{k} (D^k g(x))(\xi_1,\ldots,h_i(y),\ldots,\xi_k) \ ,$$

(ii) $$\|D^k\gamma\| \le R\|D^{k+1}g\| + k\|D^k g\| \ .$$

(iii) For x, x', ℓ, ℓ' such that $\|x - x'\| \le \delta$ and $\|\ell - \ell'\| \le \delta$

$$\|D^k\gamma(x,\ell) - D^k\gamma(x',\ell')\| \le$$

$$\le R\|(D^{k+1}g)(x) - (D^{k+1}g)(x')\| + \delta(k+1)\|D^{k+1}g\| \ .$$

iv) $$\|\gamma\|_k \le (R+k)\|g\|_{k+1} \ .$$

Proof. (i) follows by induction on k, (ii) follows by straightforward computation from (i). Likewise (iv) follows from (iii). For (iii) we remark that

$$\|D^k\gamma(x,\ell) - D^k\gamma(x',\ell')\| \le$$

$$\le \|D^k\gamma(x,\ell) - D^k\gamma(x',\ell)\| + \|D^k\gamma(x',\ell) - D^k\gamma(x',\ell')\| \le$$

$$\leq \|\ell\| \|D^{k+1}g(x) - D^{k+1}g(x')\| + \sum_{i=1}^{k} \|D^k g(x) - D^k g(x')\| +$$

$$+ \|D^{k+1}g(x')\| \|\ell - \ell'\| \leq$$

$$\leq R\|D^{k+1}g(x) - D^{k+1}g(x')\| + k\delta\|D^{k+1}g\| + \delta\|D^{k+1}g\| \ . \qquad \square$$

DEFINITION 3. For $p = -1,0,1,2,\ldots$ the polynomials M_p are defined recursively as follows:

$$M_{-1}(x) = 0$$

and

$$M_p(x) = 1 + px\,M_{p-1}(x) \qquad \text{for } p \geq 0 \ .$$

Hence

$$M_p(x) = 1 + px + p(p-1)x^2 + \ldots + p!x^p \ .$$

LEMMA 4. Let E, F and G be Banach spaces, let A be an open set in E, let R_1, R_2, δ_1, δ_2 be real numbers and suppose $R_1 \geq 1$. Denote the open ball with radius R_1 and center 0 in F by U. Let f_0 and f_1 be in $\mathcal{D}_A^{(p)}(F)$ and suppose $\|f_i\|_p < R_1$ for $i = 0,1$. Let g_0 and g_1 be in $\mathcal{D}_U^{(p)}(G)$ and suppose $\|g_i\|_p < R_2$ for $i = 0,1$. Let C be a modulus of continuity for $D^p g$. Then, if

$$\|f_0 - f_1\|_p < \delta_1 \text{ and } \|g_0 - g_1\|_p < \delta_2,$$

$$(*) \qquad \|g_0 \circ f_0 - g_1 \circ f_1\|_p \leq \delta_2 M_p(R_1) + \delta_1 R_2\, p\, R_1^{p-1} M_p(1) + R_1^p C(\delta_1) \ .$$

(See illustration.)

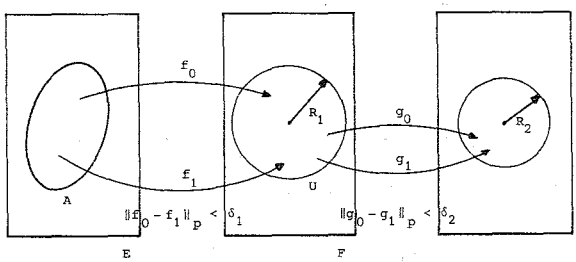


Fig. 1. The relation between the variables of Lemma 4 and Corollary 5.

Proof. Induction with respect to p. Suppose p = 0. From

$$\|g_0 \circ f_0(x) - g_1 \circ f_1(x)\| \leq \|g_0(f_0(x)) - g_0(f_1(x))\| +$$

$$+ \|g_0(f_1(x)) - g_1(f_1(x))\| \leq C(\delta_1) + \delta_2$$

it is clear that the statement holds. Now suppose the lemma is proved for p = q - 1. Then we have

$$\|g_0 \circ f_0 - g_1 \circ f_1\|_q \leq \|g_0 \circ f_0 - g_1 \circ f_1\|_{q-1} + \|\gamma_0 \circ \varphi_0 - \gamma_1 \circ \varphi_1\|_{q-1} ,$$

where γ_0 and γ_1 are defined as in Lemma 2 from g_0 and g_1 and where $\varphi_i = \overset{U}{|}_x(f_i(x), Df_i(x))$ for i = 0,1.

Now it is easily seen that $\|\varphi_1 - \varphi_0\|_{q-1} \leq \delta_1$ and $\|\varphi_i\|_{q-1} \leq R_1$ for i = 0,1 and also $\|\gamma_0 - \gamma_1\|_{q-1} \leq (q - 1 + R_1)\|g_0 - g_1\|_{q-1} \leq q R_1 \delta_2$, because of Lemma 2 and the assumption $R_1 \geq 1$.

Moreover, also because of Lemma 2, $\|\gamma_i\|_{q-1} < q R_1 R_2$.

Now Lemma 2, (iii) can be interpreted by saying that

$$\omega_\delta(R_1 C(\delta) + q\delta R_2)$$

is a modulus of continuity for $D^{q-1}\gamma_0$. Applying the induction hypothesis and straightforward computation gives the result. □

From this result we deduce a number of corollaries.

COROLLARY 5. Assumptions as in Lemma 4.

$$\|g_0 \circ f_0 - g_1 \circ f_1\|_p \leq e\, p!\, R_1^{p-1}(\delta_2 R_1 + \delta_1 R_2) + R_1^p C(\delta_1) .$$

COROLLARY 6.

$$\|g \circ f\|_p \leq e\, p!\, (\max(\|f\|_p, 1))^p \|g\|_p .$$

Proof. Take $f_0 = f_1 = f$ and $g_1 = 0$ and $g_0 = g$ in Corollary 5, furthermore $\delta_2 = R_2 = \|g\|$, $\delta_1 = 0$ and $R_1 = \max(\|f\|_p, 1)$. □

Corollary 6 tells us that the composition of two mappings with bounded p-th derivatives again yields a mapping with bounded p-th derivative. Corollary 5 tells us that a composition is continuous in any pair (f_0, g_0) where $D^p g_0$ is uniformly continuous.

We now proceed to investigate the differentiability of the composition. In the sequel we shall omit precise conditions on mappings that guarantee that they can be composed. These conditions, as a rule, are not difficult to determine, but their precise formulations interrupt easy reading.

A further Corollary to Lemma 4 illustrates our policy:

COROLLARY 7. Let g, f and u be p times continuously differentiable with bounded derivatives. Let $R_1 \geq \max(1, \|f\|_p, \|f+u\|_p)$. Then

$$\|g \circ (f+u) - g \circ f\|_{p-1} \leq R_1^{p-2}(R_1 + (p-1)!e)\|g\|_p \|u\|_p .$$

Proof. Apply Corollary 5 with $g_0 = g_1 = g$, $f_0 = f + u$, $f_1 = f$, $\delta_1 = \|u\|_p$, $\delta_2 = 0$, $R_2 = \|g\|_p$ and observe that $\delta \mapsto \|g\|_p \delta$ is a modulus of continuity for $D^{p-1}g$. □

LEMMA 8. Let g, f, u, R_1 be as in Corollary 7. There is a constant C_p such that for every modulus of continuity C for $D^p g$

$$\|g \circ (f+u) - g \circ f - Dg(f) \cdot u\|_{p-1} \leq R_1^{p-2}(R_1 \|u\|_p C(\|u\|_p) + C_p \|g\|_p \|u\|_p^2).$$

Proof. Let $\cdot$ denote any bilinear map and let K and L be k times differentiable. It is not difficult to see (use Leibniz's rule), that

$$\|K \cdot L\|_k \leq (2^k - 1)\|K\|_{k-1}\|L\|_k + \|K\|_k \|L\| \leq$$

$$\leq 2^k \|K\|_k \|L\|_k .$$

We prove the lemma by induction on p. For arbitrary k + 1 times continuously differentiable h we define

$$\psi_k(h) := \|h \circ (f+u) - h \circ f - Dh(f) \cdot u\|_k .$$

For example, if $k = 1$ and if C is a modulus of continuity for Dh, then

$$\psi_0(h) \leq \|u\| C(\|u\|) ,$$

and if h is twice continuously differentiable with bounded derivatives, then

$$\psi_0(h) \leq \|h\|_2 \|u\|^2 .$$

We now proceed with the proof of Lemma 8. We have dealt with the case $p = 1$ just a moment ago ($C_0 = 0$). Assume the lemma proved for $p < q$, and assume $q > 2$. We are going to estimate $\psi_{q-1}(g)$.

$$\psi_{q-1}(g) \leq \|g \circ (f+u) - g \circ f - Dg(f) \cdot u\| + \|D(g \circ (f+u) - g \circ f - Dg(f) \cdot u)\|_{q-2} =$$

$$= \psi_0(g) + \|Dg(f+u) \cdot (Df + Du) - Dg(f) \cdot Df - D^2g(f) \cdot (u, Df) - Dg(f) \cdot Du\|_{q-2} \leq$$

$$\leq \psi_0(g) + \|Dg(f+u) \cdot Df - Dg(f) \cdot Df - (D^2g(f) \cdot u) \cdot Df\|_{q-2} +$$

$$+ \|Dg(f+u) \cdot Du - Dg(f) \cdot Du\|_{q-2} \leq$$

$$\leq \psi_0(g) + \|Dg(f+u) - Dg(f) - D^2g(f) \cdot u\|_{q-2} \|Df\| +$$

$$+ (2^{q-2} - 1) \|Dg(f+u) - Dg(f) - D^2g(f) \cdot u\|_{q-3} \|Df\|_{q-2} +$$

$$+ 2^{q-2} \|Dg(f+u) - Dg(f)\|_{q-2} \|Du\|_{q-2} \leq$$

$$\leq \psi_0(g) + \psi_{q-2}(Dg) R_1 + 2^{q-2} \psi_{q-3}(Dg) R_1 + 2^{q-2} R_1^{q-2} (1 + (q-2)!e) \|g\|_q \|u\|_q^2 ,$$

where R_1 is defined as in Corollary 7.

Now we can estimate the last expression by using C_{q-1} and C_{q-2}. So we get

$$\psi_{q-1}(g) \le \|g\|_q \|u\|_q^2 + R_1^{q-1} \|u\|_q C(\|u\|_q) + C_{q-1} R_1^{q-2} \|g\|_q \|u\|_q^2 +$$

$$\le 2^{p-2} R_1^{q-2} \|u\|_q \{\|g\|_q \|u\|_q\} +$$

$$+ C_{q-2} R_1^{q-3} \|g\|_q \|u\|_q^2 + 2^{q-2} R_1^{q-2} (1 + (q-2)!e) \|g\|_q \|u\|_q^2 ,$$

where we have used that a modulus of continuity for $D^{q-1}(Dg)$ is a modulus of continuity for $D^q G$ and where the term in braces represents a modulus of continuity for $D^{q-1} g$.

We finally obtain as upper bound for C_q the number

$$C_{q-1} + 2^{q-2} + C_{q-2} + 2^{q-2}(1 + (q-2)!e) .$$

The proof for the case $q = 2$ is similar, but simpler: there are no terms with index $q - 3$.

The proof of the lemma is complete. We can start to reap the results. □

7. FINAL RESULTS

DEFINITION 9. COMP_k^p, for $k < p$ is the map $(g,f) \mapsto g \circ f$, interpreted as map $\mathcal{D}^{(p)} \times \mathcal{D}^{(p)} \to \mathcal{D}^{(p-k)}$.

THEOREM 10. COMP_1^p is differentiable at any point (g,f) where $D^p g$ is uniformly continuous for $p \ge 1$, and its derivative at such a point is

$$(u,v) \mapsto Dg(f) \cdot u + v \circ f .$$

Proof. Let C be a modulus of continuity for $D^p g$. Then

$$\| \mathrm{COMP}_1^p(f+u, g+v) - \mathrm{COMP}_1^p(f,g) - Dg(f) \cdot u - v \circ f \|_{p-1} =$$

$$= \| g \circ (f+u) + v \circ (f+u) - g \circ f - Dg(f) \cdot u - v \circ f \|_{p-1} \leq$$

$$\leq \psi_{p-1}(g) + \| v \circ (f+u) - v \circ f \|_{p-1} .$$

By Corollary 7 and Lemma 8 it is easy to show that the last expression is $o(\|v\|_p + \|u\|_p)$ for $\|v\|_p \to 0$ and $\|u\|_p \to 0$. □

COROLLARY 11. COMP_2^p is differentiable for $p \geq 2$.

Proof. Let J be the map $(g,f) \mapsto (g,f)$, $\mathcal{D}^{(p)} \times \mathcal{D}^{(p)} \to \mathcal{D}^{(p-1} \times \mathcal{D}^{(p-1)}$. Then $\mathrm{COMP}_2^p = \mathrm{COMP}_1^{p-1} \circ J$, and we can apply Theorem 9, as differentiability of J is no problem. Observe that for $g \in \mathcal{D}^{(p)}$, $D^{p-1}g$ is uniformly continuous. □

THEOREM 12. COMP_k^p is k - 1 times differentiable for $2 \leq k \leq p$.

Proof. The case k = 2 is just Corollary 11 above. We proceed by induction. Let the theorem be proved for all k, $2 \leq k < \ell$. The derivative of COMP_ℓ^p is the map

$$(g,f) \mapsto Dg(f) + U_f ,$$

where U_f is the linear map $v \mapsto v \circ f$. Because U_f is continuous and linear, U_f is arbitrary differentiable. The map $(g,f) \mapsto Dg(f)$ is a composition of $(g,f) \to (Dg,f)$, $\mathcal{D}^p \times \mathcal{D}^p \to \mathcal{D}^{p-1} \times \mathcal{D}^{p-1}$ and $\mathrm{COMP}_{\ell-1}^{p-1}$, and as $g \mapsto Dg$ is a norm decreasing linear map $\mathcal{D}^p \to \mathcal{D}^{p-1}$ it is also differentiable; $\mathrm{COMP}_{\ell-1}^{p-1}$ is differentiable by induction hypothesis.

Our theorem is proved. □

REFERENCES

[1] Dieudonné, Jean, (sous la direction de), Abrégé d'histoire des mathématiques (2 vols), Hermann, Paris, 1978.

[2] Dieudonné, J., Eléments d'analyse (t1), Fondements de l'analyse moderne, 3e édition, Gauthier-Villars, Paris, 1979.

[3] Lakatos, Imre, Proofs and Refutations, The logic of Mathematical Discovery, Cambridge University Press, Cambridge, 1976.

[4] Seidel, P.L., Note über eine Eigenschaft der Reihen, welche discontinuirliche Funktionen darstellen, Abh. der Math. Phys. Klasse der Kgl. Bayerische Akademie der Wissenschaften V (1847), p. 381 - 394 (= Ostwald's Klassiker der Exakten Wissenschaften 116).

SOME INVESTIGATIONS ON THE CONWAY-GOLAY GAME "LIFE"

by

C.P. van Nieuwkasteele and K.A. Post

Dedicated to J.J. Seidel on the occasion of his retirement from the Eindhoven University of Technology.

1. EXPLANATION OF THE CONWAY-GOLAY GAME "LIFE"

Conway-Golay bees live in the Euclidean plane. To be more specific, they live as solitary animals in the cells of an infinite regular honeycomb, and have specific laws of reproduction (cf. WAINWRIGHT [5]):

(i) Reproduction takes place for all animals at the same instant;

(ii) New life is born in a cell if and only if exactly two of its six neighbour cells are occupied, and these cells are not diametrically opposite with respect to the central cell;

(iii) The old generation immediately dies after its reproductive activities.

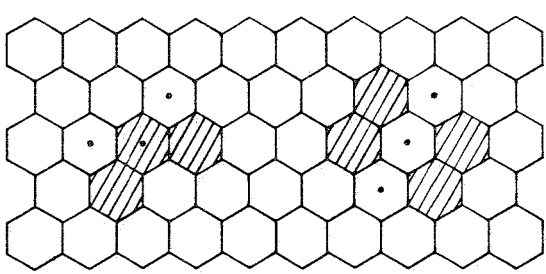


Figure 1.

Figure 1 shows two examples of the reproduction laws. The old generation is indicated by dots, the new generation by shading. Observe that the left example represents a population that is periodic with period 2 and that the right example exhibits a population that dies out.

2. SOME ALTERNATIVE DESCRIPTIONS

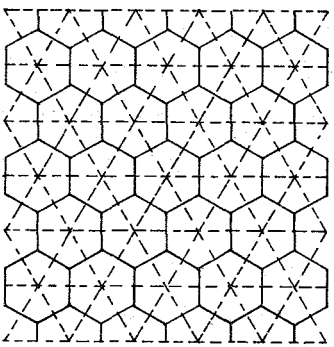


Figure 2.

It is useful to replace the infinite honeycomb by the triangular lattice that is formed by the centres of its cells. (See Figure 2; the triangles are dashed.) Examples of the reproduction laws in a cell are depicted in Figure 3.

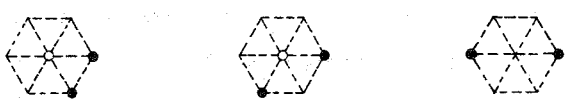


Figure 3.

The old generation is given by "solid" dots, the new generation by "open" dots.

Figure 4 illustrates a periodic population (period 4).

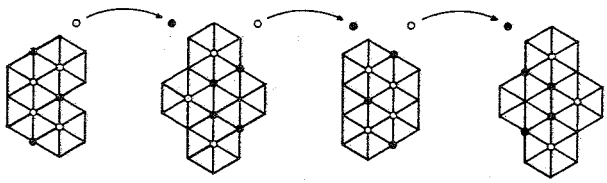


Figure 4.

The reproduction law for new life in a lattice point P can also be formulated in arithmetic terms over GF(2): Let the symbol 1 denote "life" and the symbol 0 mean "no life". If P is cyclically surrounded by six points whose life situation is given by a, b, c, d, e, f then the life in point P for the next generation is given by

$$\varphi(P) = \begin{cases} (a+1)(d+1)(b+e)(c+f) + \\ + (b+1)(e+1)(c+f)(a+d) + \\ + (c+1)(f+1)(a+d)(b+e) \ . \end{cases}$$

(General reproduction formula)

This formula can easily be checked by inspection (cf. Figure 5).

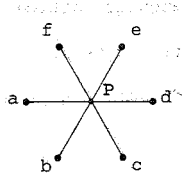


Figure 5.

The general situation is too complicated to be discussed in this paper. Therefore we shall make some restrictions: With every lattice point we can associate its vertical level. This level can be even or odd. Correspondingly, the lattice itself splits into its even and its odd sublattice (intersection points of solid lines resp.

dashed lines in Figure 6.

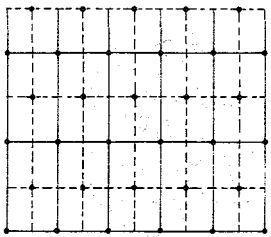


Figure 6.

It is immediately verified that a generation which lives on the even sublattice can only reproduce in odd lattice points, and conversely.

Besides, the reproduction law for a population that starts in one of the sublattices is easier to analyze. We get

$$\varphi(P) = (b+e)(c+f)$$ (Special reproduction formula)

(cf. the general reproduction formula).

From now on we assume that the starting generation (= generation 0) lives on the even sublattice.

If such a generation is finite, then we can form a rhombus with horizontal and vertical symmetry axes, edges along the grid lines of the triangular lattice, so that the whole generation lives inside or on the boundary of the rhombus.

CLAIM. The population never reproduces in points outside the rhombus.

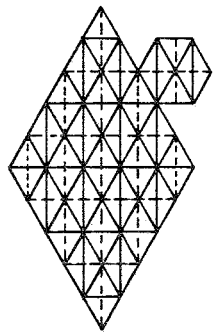


Figure 7.

Proof. A hexagon centered outside the rhombus (cf. Figure 7) has at most two vertices that are at the same time lattice points of the rhombus. However, these lattice points belong to different sublattices, so at most one of them carries life, which proves the claim. □

(1) Therefore, let us assume that we are given an $n \times n$ rhombus and a starting generation on the even sublattice of the rhombus.

(2) Our final assumption will be that in this generation for each lattice point P of the rhombus the "neighbour life sum" b + e + c + f vanishes (see Figure 8). As a consequence we can rewrite the special reproduction formula in the form (arithmetic in GF(2)!)

$$\varphi(P) = b + e = c + f \ .$$

It is also possible to visualize a generation as a binary $n \times n$ matrix $A = (a_{ij})$. This is indicated in Figure 9.

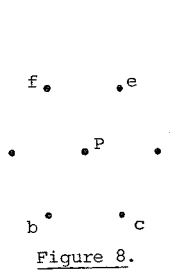


Figure 8.

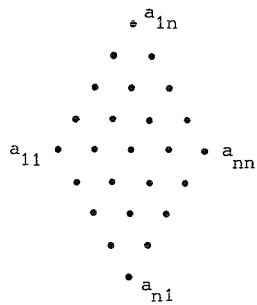


Figure 9.

CLAIM. Our assumptions imply that the starting matrix B commutes with the codiagonal matrix H given by

$$h_{ij} = \begin{cases} 1 & \text{if } |i-j| = 1 \ , \\ 0 & \text{otherwise} \end{cases} .$$

Besides, the next generation is given by the matrix BH (= HB).

Proof. By inspection, using the special reproduction formula. □

In fact, we can state even more: The history of the population can be written in its matrix form as $(BH^k)_{k\in\mathbb{N}}$.

The proof of this statement is due to the fact that a generation satisfying (1) and (2) reproduces a generation that also has the properties (1) and (2). To see this we only have to check the neighbour life sum for the new generation (cf. Figure 10). For the point P this sum equals

$$\varphi(Q) + \varphi(R) + \varphi(S) + \varphi(T) =$$

$$= (d+b) + (b+f) + (f+h) + (h+d) = 0 \ .$$

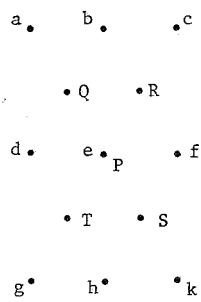


Figure 10.

Special cases of starting generations satisfying (1) and (2) are

(i) n animals in a row (B = I);

(ii) p * q blocks;

(iii) "step configurations" (trap-configuraties).

These notions were introduced by SEIDEL [4]. In Section 3 we shall restrict ourselves to the first class of examples: n animals in a row. Figure 11 illustrates the case n = 5 (period 4).

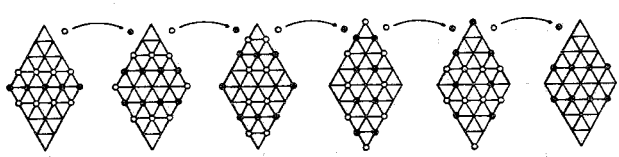


Figure 11.

3. "n ANIMALS IN A ROW"

In this section we shall investigate the development of a population whose starting generation consists of n animals in a row. The arguments in the preceding section show that the k-th generation of this population is visualized by the k-th power of the n-by-n codiagonal matrix H_n, where

$$H_n := \begin{bmatrix} 0 & 1 & & & & & \\ 1 & 0 & 1 & & & & \\ & 1 & 0 & \cdot & & & \\ & & \cdot & \cdot & \cdot & & \\ & & & \cdot & \cdot & \cdot & \\ & & & & \cdot & \cdot & 1 \\ & & & & & 1 & 0 \end{bmatrix} \tag{1}$$

and all calculations are performed in GF(2). The generating function of these k-th powers is given formally by the geometric series

$$I_n + t H_n + t^2 H_n^2 + \ldots = (I_n + t H_n)^{-1} . \tag{2}$$

The elements of the matrix $(I_n + t H_n)^{-1}$ are rational functions of t, whose denominator is given by

$$f_n(t) := \det(I_n + t H_n) . \tag{3}$$

Their numerators are polynomials of degree $\leq n - 1$ in t, one of which (viz. the top rightmost element) is exactly t^{n-1}.

We shall now focus our attention on $f_n(t)$. It is useful to define $f_0(t) := 1$. Then $f_n(t)$ clearly can be computed recursively by

$$\begin{cases} f_0(t) = 1 \\ f_1(t) = 1 \\ f_n(t) = f_{n-1}(t) + t^2 f_{n-2}(t) \qquad (n \geq 2) \end{cases} . \tag{4}$$

Now let us introduce the generating function

$$F(t,y) := \sum_{n=0}^{\infty} f_n(t)\, y^n . \tag{5}$$

From the formal identity

$$\sum_{n=2}^{\infty} f_n(t)\, y^n = \sum_{n=2}^{\infty} f_{n-1}(t)\, y^n + \sum_{n=2}^{\infty} t^2 f_{n-2}(t)\, y^n ,$$

arising from (4), using $f_0(t) = f_1(t) = 1$, we deduce that

$$F(t,y) + 1 + y = y(F(t,y) + 1) + t^2 y^2 F(t,y) ,$$

in other words

$$F(t,y) = \frac{1}{1 + y + t^2 y^2} . \tag{6}$$

From (4) we can also derive that

$$\begin{cases} f_0(t) = 1 \\ f_2(t) = 1 + t^2 \\ f_{2n}(t) = f_{2(n-1)}(t) + t^4 f_{2(n-2)}(t) \qquad (n \geq 2) \end{cases} . \tag{7}$$

and proceed in the same way to obtain the generating function

$$F_{even}(t,y) := \sum_{n=0}^{\infty} f_{2n}(t)\, y^{2n} . \tag{8}$$

It turns out that

$$F_{even}(t,y) = \frac{1 + t^2 y^2}{1 + y^2 + t^4 y^4} . \tag{9}$$

Hence, for the odd part of $F(t,y)$ we get

$$F_{odd}(t,y) := F(t,y) - F_{even}(t,y) = \frac{y}{1 + y^2 + t^4 y^4} .$$

Therefore we may state that

$$F_{odd}(t,y) = y\, F(t^2,y^2) . \tag{10}$$

We now proceed recursively as follows

$$\begin{aligned} F(t,y) &= F_{even}(t,y) + F_{odd}(t,y) = \\ &= F_{even}(t,y) + y\, F(t^2,y^2) = \\ &= F_{even}(t,y) + y\, F_{even}(t^2,y^2) + y\, F_{odd}(t^2,y^2) = \\ &= F_{even}(t,y) + y\, F_{even}(t^2,y^2) + y^3 F(t^4,y^4) = \\ &= \ldots\ldots \end{aligned}$$

so that finally

$$F(t,y) = \sum_{\ell=0}^{\infty} y^{2^\ell - 1} F_{even}(t^{2^\ell}, y^{2^\ell}) . \tag{11}$$

This formalism is fully justified because all summands in each of the expressions of the foregoing derivation comprise different powers of y, and the polynomial coefficients of these powers of y all require finitely many computations in GF(2).

We are especially interested in the question whether a population, given by its starting generation, dies out, or behaves periodically in time after some generation. This question can now be answered:

THEOREM 1. n animals in a row yield a population that dies out if and only if $n = 2^{\ell} - 1$ for some integer ℓ. If the starting generation is called generation 0 then in this case generation $(n-1)$ is the last generation that carries life.

Proof. From (7) it follows that $f_{2n}(t)$ has degree 2n and that $f_{2n}(0) = 1$ for all n. Therefore, by (11) and (8), exactly those $f_n(t)$ have degree 0 for which $n = 2^{\ell} + 1$ for some integer ℓ. In this case the inverse of $(I_n + t H_n)$ is a polynomial matrix of degree $(n-1)$. □

In the remaining cases ($n \notin \{2^{\ell} - 1 \mid \ell \in \mathbb{N}\}$) the population is periodic from its $(n-1)^{st}$ generation on (but may be periodic from a lower generation on, cf. n = 5, Fig. 11), and its period can be determined using F_{even} as follows: A typical term in Formula (11) is

$$y^{2^{\ell}-1}\left(1 + f_2(t^{2^{\ell}})\, y^{2\cdot 2^{\ell}} + f_4(t^{2^{\ell}})\, y^{4\cdot 2^{\ell}} + f_6(t^{2^{\ell}})\, y^{6\cdot 2^{\ell}} + \ldots\right).$$

So we must write $n + 1 = 2^{\ell}(2s + 1)$ and then find the period of the population to be the period of the polynomial $f_{2s}(t^{2^{\ell}})$ (cf. BERLEKAMP [1]), or, equivalently, $2^{\ell+1}$ times the period of $g_s(u)$, where the polynomial $g_s(u)$ is recursively defined by

$$\begin{cases} g_0(u) := 1 \\ g_1(u) := 1 + u \\ g_s(u) := g_{s-1}(u) + u^2 g_{s-2}(u) \qquad (s \geq 2) \end{cases} . \tag{12}$$

Therefore, we have proved the following theorem.

THEOREM 2. If $n + 1 = 2^{\ell}(2s + 1)$ then the period of a population, that starts with n animals in a row, is equal to $2^{\ell+1}$ times the period of $g_s(u)$, where $g_s(u)$ is defined by (12).

SEIDEL ([4]) proved that for $n = 2s$ the period of a population, that starts with n animals in a row, is equal to $2(2^q - 1)$, where q is the smallest integer satisfying $2^q \equiv \pm 1 \pmod{2s + 1}$.

This enables us to state the following theorem.

THEOREM 3. $g_s(u)$ is a primitive polynomial if and only if $2s + 1$ is a prime that has 2 as a primitive root.

In Figure 11 we saw that $n = 5$ yields a population that is periodic but does never return to its starting generation. It is not difficult to formulate a theorem that discusses this phenomenon:

THEOREM 4. n animals in a row yield a population that returns to its starting generation iff n is even.

Proof. If n is *odd* then there is a central point, and this point carries life in the starting generation. However, by symmetry arguments, in all generations the life in the hexagon around this point is concentrated in diametrically opposite points, so that new life never can be born in the central point. Now let n be *even*, n = 2s. The linear mapping with matrix H_{2s}^2,

$$H_{2s}^2 = \begin{bmatrix} 1 & 0 & 1 & & & & & & \\ 0 & 0 & 0 & 1 & & & & & \\ 1 & 0 & 0 & \cdot & \cdot & & & & \\ & 1 & \cdot & \cdot & \cdot & \cdot & & & \\ & & \cdot & \cdot & \cdot & \cdot & 1 & & \\ & & & \cdot & \cdot & 0 & 0 & 1 & \\ & & & & 1 & 0 & 0 & 0 & \\ & & & & & 1 & 0 & 1 & \end{bmatrix}$$

has two invariant subspaces (the spaces, spanned by the odd- and even-numbered basis vectors, respectively), on which it acts in the same way by the linear mapping with matrix K_s,

$$K_s := \begin{bmatrix} 1 & 1 & & & & & \\ 1 & 0 & 1 & & & & \\ & 1 & 0 & \cdot & & & \\ & & \cdot & \cdot & \cdot & & \\ & & & \cdot & \cdot & \cdot & \\ & & & & \cdot & \cdot & 1 \\ & & & & & 1 & 0 \end{bmatrix} .$$

The theorem is proved when we can show that the characteristic equation of K_s has s distinct roots $\neq 0$, or, equivalently, that $\det(I_s + u K_s)$ vanishes for s distinct values of u, $u \neq 0$. It is again a matter of verification that $\det(I_s + u K_s) = g_s(u)$, by checking (12). From the recurrence for $g_s(u)$ we can deduce by formal differentiation that

$$\begin{cases} g_0'(u) = 0 \\ g_1'(u) = 1 \\ g_s'(u) = g_{s-1}'(u) + u^2 g_{s-2}'(u) \qquad (s \geq 2) \end{cases} . \tag{13}$$

From (12) and (13) we easily deduce by induction that for all s, $s \geq 2$

$$g_s'(u) = g_s(u) + u\, g_{s-1}(u) . \tag{14}$$

Hence, we obtain

$$GCD(g_s'(u), g_s(u)) = GCD(g_s(u), u\, g_{s-1}(u)) = 1 ,$$

again by induction, using (12). Therefore, $g_s(u)$ has no multiple roots, which proves the theorem (cf. JACOBSON [3]). □

SEIDEL ([4]) pointed out that there exists a relation between CONWAY-GOLAY's game of life and a solitary game called "Heckenspiel", that is discussed in a paper by BUSSEMAKER and GANTER ([2]). In fact, the description that SEIDEL gives corresponds to the images of a fixed basisvector under H_n^k ($k = 1,2,3,\ldots$).

REFERENCES

[1] BERLEKAMP, E.R., Algebraic Coding Theory, New York, 1968.

[2] BUSSEMAKER, F.C. and B. GANTER, Wann hört eine Hecke zu wachsen auf Math. Phys. Semesterberichte XXIII (1976), p. 125 - 135.

[3] JACOBSON, N., Lectures in Abstract Algebra, III, Princeton, 1964.

[4] SEIDEL, J.J., Private Communication.

[5] WAINWRIGHT, R.T., Lifeline (2), june 1971, p. 14.

A FEW CONSTRUCTIONS AND A SHORT TABLE OF δ-DECODABLE CODEPAIRS FOR THE BINARY, TWO-ACCESS ADDER CHANNEL

by

H.C.A. van Tilborg

Dedicated to Jaap Seidel on the occasion of his retirement from the Eindhoven University of Technology.

Abstract. In the literature one finds only a few δ-decodable codepairs for the binary, two-access adder channel.
Here a few general constructions for such codepairs are presented. Also a short table is included, which was composed by means of a graph theoretic approach, due to Kasami et al..

1. INTRODUCTION

Consider the two-access system shown in Figure 1. Here two independent sources wish to send information to the two receivers. During the message interval, the two messages emanating from the two sources are encoded independently with two binary block codes C and D of the same length. We assume that we have bit and block synchronization.

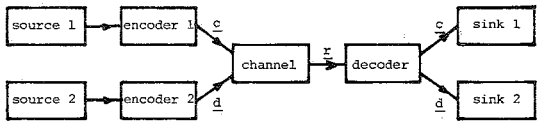


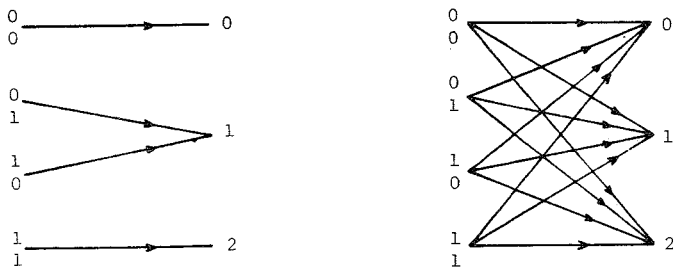
Figure 1. A 2-access communication system.

In this paper we shall deal with a particular two-access channel, known as the binary adder channel.

A two-access, binary adder channel is a channel with two binary inputs c and d and a ternary output r.

In the noiseless case r is the sum c + d, where + denotes addition over the reals. One sees from Figure 2(a) that ambiguity arises when r = 1. When r = 1 then either (c,d) = (0,1) or (1,0).

In the case of noise we say that 1 error has occurred if $|r - (c + d)| = 1$ and that 2 errors have occurred if $|r - (c + d)| = 2$.



(a) Noiseless two-user adder channel (b) Noisy two-user adder channel

Figure 2. Two-user adder channel models.

Let V_n and W_n denote n-dimensional vectorspaces over GF(2) resp. GF(3). In V_n we shall use the regular Hamming distance, denoted by d_H. In W_1 we define the distance function δ by $\delta(a,b) = |b - a|$. This distance function can be extended to W_n by

$$\delta(\underline{a},\underline{b}) = \sum_{i=1}^{n} \delta(a_i,b_i) \quad . \tag{1.1}$$

Let C and D be the codes used by the two users of the communication system. Suppose that they send the codewords $\underline{c}$ and $\underline{d}$. In the noiseless case the decoder will receive the vector

$$\underline{r} = \underline{c} + \underline{d} = (c_1 + d_1 , c_2 + d_2 , \ldots , c_n + d_n) \quad .$$

In the noisy case it may be any vector in W_n.

DEFINITION 1.1. The codepair (C,D) is called δ-decodable if for any two distinct pairs $(\underline{c},\underline{d})$ and $(\underline{c}',\underline{d}')$ in $C \times D$ one has that

$$\delta(\underline{c} + \underline{d} , \underline{c}' + \underline{d}') \geq \delta \quad .$$

It follows that if less than or equal to $\left\lfloor \frac{\delta - 1}{2} \right\rfloor$ errors have been made during a transmission, then the decoder can uniquely determine the transmitted words $\underline{c}$ and $\underline{d}$ from the received word $\underline{r}$.

EXAMPLE 1.2. If $C = \{\underline{0},\underline{1}\}$ and $D = V_n \setminus \{\underline{0}\}$ or $V_n \setminus \{\underline{1}\}$ then the pair (C,D) is uniquely decodable with $|C| = 2$ and $|D| = 2^n - 1$, as one can easily verify.

DEFINITION 1.3. The ratepair (R_1,R_2) of a codepair (C,D) of length n is defined by

$$R_1 = \frac{1}{n}\,{}^2\log|C| \ ,$$

$$R_2 = \frac{1}{n}\,{}^2\log|D| \ .$$

The ratepairs of the sequence of codepairs in Example 1.2 is given by

$$(R_1,R_2) = \left(\frac{1}{n}\ ,\ 1 + {}^2\log\,(1 - 2^{-n})\right) \ .$$

A great deal of study has been given to uniquely decodable codepairs. See [1] until [13]. For $\delta > 1$ one can find some explicit constructions in [4] and [9].

Since a systematic study of δ-decodable codes hardly can be found in the literature, we shall describe in Paragraph 2 two standard ways of making longer δ-decodable codepairs from shorter ones. Also a few general families of δ-decodable codepairs are given there.

In Paragraph 3 we describe the graph-theoretic approach of Kasami et al. and use it to compose two tables of δ-decodable codepairs.

Later on we shall need the following two lemmas. The first of these can already be found in Kasami et al. [4].

LEMMA 1.4. Let (C,D) be a δ-decodable codepair of length n. Then C and D are codes with minimum Hamming distance d at least δ.

Proof. Let $\underline{c}$ and $\underline{c}'$ be different codewords in C. Let $\underline{d} \in D$. Then

$$d_H(\underline{c},\underline{c}') = \sum_{i=1}^{n} |c_i - c_i'| = \sum_{i=1}^{n} |(c_i + d_i) - (c_i' + d_i)| =$$

$$= \delta(\underline{c} + \underline{d}, \underline{c}' + \underline{d}) \geq \delta .$$

The same argument holds for D. □

LEMMA 1.5. Let (C,D) be a δ-decodable codepair of length n. Let $\underline{u} \in V_n$ and define the codepair $(C_{\underline{u}},D_{\underline{u}})$ by

$$(C_{\underline{u}},D_{\underline{u}}) = (\{\underline{c} \oplus \underline{u} \mid \underline{c} \in C\}, \{\underline{d} + \underline{u} \mid \underline{d} \in D\}) ,$$

where $\oplus$ denotes addition modulo 2. Then $(C_{\underline{u}},D_{\underline{u}})$ is also a δ-decodable codepair.

Proof. Let $I = \{1 \leq i \leq n \mid u_i = 0\}$ and $I^* = \{1,2,\ldots,n\} \backslash I$. Then

$$\delta((\underline{c} \oplus \underline{u}) + (\underline{d} \oplus \underline{u}), (c' \oplus \underline{u}) + (\underline{d}' \oplus \underline{u})) =$$

$$= \sum_{i=1}^{n} |(c_i \oplus u_i) + (d_i \oplus u_i) - (c_i' \oplus u_i) - (d_i' \oplus u_i)| =$$

$$= \sum_{i \in I} |c_i + d_i - c_i' - d_i'| + \sum_{i \in I^*} |(1 - c_i) + (1 - d_i) - (1 - c_i') - (1 - d_i')| =$$

$$= \sum_{i=1}^{n} |c_i + d_i - c_i' - d_i'| = \delta(\underline{c} + \underline{d}, \underline{c}' + \underline{d}') .$$ □

The following theorem shows that one can also translate one of the codes C and D over the all-one vector.

THEOREM 1.6. Let (C,D) be a δ-decodable codepair of length n. Let $\underline{1}$ denote the all-one vector and $D \oplus \underline{1} = \{\underline{d} \oplus \underline{1} \mid \underline{d} \in D\}$. Then also $(C, D \oplus 1)$ is a δ-decodable codepair of length n.

Proof.

$$\delta(\underline{c} + (\underline{d} \oplus \underline{1}), \underline{c}' + (\underline{d}' \oplus \underline{1})) =$$

$$= \sum_{i=1}^{n} |\{c_i + (1 - d_i)\} - \{c_i' + (1 - d_i')\}| =$$

$$= \sum_{i=1}^{n} |(c_i + d_i') - (c_i' + d_i)| = \delta(\underline{c} + \underline{d}', \underline{c}' + \underline{d}) . \qquad \square$$

2. CONSTRUCTIONS

The following two ways of making longer δ-decodable codes from shorter ones are very trivial to check. We omit the proofs.

With $(\underline{x}|\underline{y})$ we shall denote the concatenation of the two vectors $\underline{x}$ and $\underline{y}$.

THEOREM 2.1. Let (C_i, D_i) be δ_i-decodable codepairs of length n_i, i = 1,2. Then the codepair $(\mathcal{C},\mathcal{D})$ defined by

$$\mathcal{C} = \{(\underline{c}|\underline{c}') \mid \underline{c} \in C_1, \underline{c}' \in C_2\}$$

$$\mathcal{D} = \{(\underline{d}|\underline{d}') \mid \underline{d} \in D_1, \underline{d}' \in D_2\}$$

has length $n_1 + n_2$, is δ-decodable with $\delta = \min\{\delta_1, \delta_2\}$ and satisfies $|\mathcal{C}| = |C_1| \cdot |C_2|$ and $|\mathcal{D}| = |D_1| \cdot |D_2|$.

Note that if C_i, i = 1,2, is linear with generator matrix G_i, then $\mathcal{C}$ has generator matrix G given by

$$G = \left(\begin{array}{c|c} G_1 & O \\ \hline O & G_2 \end{array}\right) \qquad (2.1)$$

THEOREM 2.2. Let (C_i,D_i) be δ_i-decodable codepairs of length n_i, $i = 1,2$. Suppose that $|C_1| = |C_2| = u$ and $|D_1| = |D_2| = v$. Let $\underline{c}_r^{(i)}$, $1 \le r \le u$, and $\underline{d}_s^{(i)}$, $1 \le s \le v$, be arbitrary numberings of the vectors in C_i resp. D_i, $i = 1,2$. Then the codepair $(\mathcal{C},\mathcal{D})$ defined by

$$\mathcal{C} = \{(\underline{c}_r^{(1)} \mid \underline{c}_r^{(2)}) \mid 1 \le r \le u\}$$

$$\mathcal{D} = \{(\underline{d}_s^{(1)} \mid \underline{d}_s^{(2)} \mid 1 \le s \le v\}$$

has length $n_1 + n_2$, is δ-decodable with $\delta = \delta_1 + \delta_2$ and satisfies $|\mathcal{C}| = u$ and $|\mathcal{D}| = v$.

If the C_i's, $i = 1,2$, are linear codes generated by $\underline{c}_r^{(i)}$, $1 \le r \le k$, then $\mathcal{C}$ is a linear code generated by the vectors $(\underline{c}_r^{(1)} \mid \underline{c}_r^{(2)})$, $1 \le r \le k$.
In the above theorems one can sometimes fruitfully make use of the degenerate codepair $(C,D) = (\{\underline{0}\},V_n)$, which is δ-decodable with $\delta = n$. This code has the ratepair $(0,1)$.
We shall now discuss several special families of δ-decodable codepairs (C,D). In view of Lemma 1.4 we have that $\delta \le n$ and in view of Lemma 1.5 we may assume that $\underline{0} \in C$.

THEOREM 2.3. Let (C,D) be a δ-decodable codepair of length $n \ge 4$ with $|C| \ge 2$ and $|D| \ge 2$. W.l.o.g. we assume that $\underline{0} \in C$. Then

i) $\delta = n$ iff n is even, $C = \{\underline{0},\underline{1}\}$ and $D = \{\underline{a},\underline{1} \oplus \underline{a}\}$ with $w_H(\underline{a}) = n/2$.

ii) $\delta = n - 1$ iff after a suitable coordinate permutation C and D are given by one of the following possibilities:

α) $C = \{\underline{0},\underline{1}\}$, $D = \{\underline{a},\underline{1} \oplus \underline{a}\}$, $w_H(\underline{a}) \in \left\{\frac{n-1}{2}, \frac{n+1}{2}\right\}$,

β) $C = \{\underline{0},(1,1,\ldots,1,0)\}$, $D = \{\underline{a},\underline{1} \oplus \underline{a}\}$,

$w_H((a_1,a_2,\ldots,a_{n-1})) \in \left\{\frac{n-2}{2}, \frac{n-1}{2}, \frac{n}{2}\right\}$,

γ) $C = \{\underline{0},\underline{1}\}$, $D = \{(a_1,a_2,\ldots,a_n),(1 \oplus a_1,1 \oplus a_2,\ldots,1 \oplus a_{n-1},a_n)\}$,

$w_H((a_1,a_2,\ldots,a_{n-1})) \in \left\{\frac{n-2}{2}, \frac{n-1}{2}, \frac{n}{2}\right\}$,

δ) $C = \{\underline{0},(1,1,\ldots,1,0)\}$, $D = \{(a_1,a_2,\ldots,a_n)$,

$(1 \oplus a_1,1 \oplus a_2,\ldots,1 \oplus a_{n-1},a_n)\}$, $w_H((a_1,a_2,\ldots,a_{n-1})) = \frac{n-1}{2}$,

ε) $C = \{\underline{0},(1,1,\ldots,1,0)\}$, $D = \{(a_1,a_2,\ldots,a_n)$,

$(1 \oplus a_1,1 \oplus a_2,\ldots,1 \oplus a_{n-2},a_{n-1},1 \oplus a_n)\}$, $w_H((a_1,a_2,\ldots,a_{n-2})) \in$

$\in \left\{\frac{n-3}{2}, \frac{n-2}{2}, \frac{n-1}{2}\right\}$.

Proof. It follows from $n \geq 4$, $\delta \geq n - 1$ and Lemma 1.4 that $|C| = |D| = 2$. One easily verifies that all the codepairs (C,D) mentioned in the theorem are indeed δ-decodable with $\delta = n$ resp. $\delta = n - 1$.

If $\delta = n$, it follows from Lemma 1.4 and the assumption that $\underline{0} \in C$, that $C = \{\underline{0},\underline{1}\}$ and $D = \{\underline{a},\underline{1} \oplus \underline{a}\}$.

The table of possible received words, if no errors have been made during the transmission, looks like

C \ D	$\underline{a}$	$\underline{1} \oplus \underline{a}$
$\underline{0}$	$(a_1,a_2,\ldots,a_n)$	$(1-a_1, 1-a_2, \ldots, 1-a_n)$
$\underline{1}$	$(1+a_1, 1+a_2, \ldots, 1+a_n)$	$(2-a_1, 2-a_2, \ldots, 2-a_n)$

From this table we see that $\underline{a}$ has to satisfy the additional requirements

$$2(a_1 + a_2 + \ldots + a_n) = \delta(\underline{0} + (\underline{1} \oplus \underline{a}), \underline{1} + \underline{a}) \geq n ,$$

$$2n - 2(a_1 + a_2 + \ldots + a_n) = \delta(\underline{0} + \underline{a}, \underline{1} + (\underline{1} \oplus \underline{a})) \geq n ,$$

i.e.

$$\frac{n}{2} \leq w_H(\underline{a}) \leq \frac{n}{2} .$$

This proves i). The proof of ii) goes analogously but involves a little case analysis. □

Of course there is no point in constructing a codepair (C,D) with $\delta = n - 1$ if n is even, since one also can construct a codepair (C',D') with $\delta = n$, $|C'| = |C|$ and $|D'| = |D|$.
We shall now study another extreme case: $|C| = 2$.

THEOREM 2.4. Let (C,D) be a uniquely decodable codepair of length n with $|C| = 2$. W.l.o.g. we assume that $\underline{0} \in C$.
Then $|D| \leq 2^n - 1$. Moreover, $|D| = 2^n - 1$ iff $C = \{\underline{0},\underline{1}\}$ and $D = V_n \backslash \{\underline{0}\}$ or $D = V_n \backslash \{\underline{1}\}$.

Proof. W.l.o.g. $C = \{\underline{0}, (0,0,\ldots,0,\overset{\longleftarrow k \longrightarrow}{1,1,\ldots,1})\}$. Now for any choice of $(a_1,a_2,\ldots,a_{n-k}) \in V_{n-k}$ one has that D cannot contain $(a_1,a_2,\ldots,a_{n-k},0,0,\ldots,0)$ and $(a_1,a_2,\ldots,a_{n-k},1,1,\ldots,1)$ together.
It follows that $|D| \leq 2^{n-k}(2^k - 1) = 2^n - 2^{n-k} \leq 2^n - 1$. Moreover, $|D| = 2^n - 1$ implies that $k = n$ and $D = V_n \backslash \{\underline{0}\}$ or $D = V_n \backslash \{\underline{1}\}$. □

Let $V_{n,i}$ be defined by

$$V_{n,i} = \{\underline{x} \in V_n \mid w_H(\underline{x}) = i\} . \tag{2.2}$$

THEOREM 2.5. Let (C,D) be a δ-decodable codepair of length n with $|C| = 2$ and $\delta = 2$. W.l.o.g. we assume that $\underline{0} \in C$.

i) If n is even, then the maximal value of $|D|$ is 2^{n-1}. Moreover, $|D| = 2^{n-1}$ iff $C = \{\underline{0},\underline{1}\}$ and $D = \bigcup_{i \text{ odd}} V_i(n,2)$.

ii) If n is odd, then the maximal value of $|D|$ is $2^{n-1} - 1$. Moreover, $|D| = 2^{n-1} - 1$ iff either

$$C = \{\underline{0},\underline{1}\} \quad \text{and} \quad D = \bigcup_{\substack{i \\ i \neq 0}} V_{n,i}$$

or

$$C = \{\underline{0},\underline{1}\} \quad \text{and} \quad D = \bigcup_{\substack{i \text{ odd} \\ i \neq n}} V_{n,i}$$

or

$$C = \{\underline{0},\underline{a}\}, \; w_H(\underline{a}) = n - 1 \quad \text{and} \quad D = \bigcup_{\substack{i \text{ odd} \\ i \neq n}} V_{n,i} .$$

Proof. It follows from Theorem 1.4 that D is a binary code with minimum Hamming distance at least 2. So $|D| \leq 2^{n-1}$ and if $|D| = 2^{n-1}$ then either $D = \bigcup_{i \text{ even}} V_{n,i}$ or $D = \bigcup_{i \text{ odd}} V_{n,i}$. Suppose that $D = \bigcup_{i \text{ even}} V_{n,i}$. Then $\underline{0} \in D$. Since $\underline{0} \in C$ it follows that the second word, say $\underline{a}$, in C is of odd weight, say ℓ. Let $\underline{b} = \underline{a} \oplus (1,0,0,\ldots,0)$. Then $\underline{b} \in D$ and we have that $\delta(\underline{0} + \underline{b}, \underline{a} + \underline{0}) = 1$ produces a contradiction with $\delta = 2$.

This leaves us with the case that $D = \bigcup_{i \text{ odd}} V_{n,i}$. Let $C = \{\underline{0},\underline{a}\}$, $w_H(\underline{a}) = \ell$. If $\ell < n$ then there exists a word $\underline{e}$ of weight 1 with $d_H(\underline{a},\underline{e}) = \ell + 1$. Now $\delta(\underline{0} + (\underline{a} \oplus \underline{e}), \underline{a} + \underline{e}) = 0$ gives a contradiction with $\delta = 2$ for even values of ℓ, while $\delta(\underline{0} + \underline{a}, \underline{a} + \underline{e}) = 1$ gives a contradiction for odd values of ℓ.

We conclude that $C = \{\underline{0},\underline{1}\}$. However, if n is odd we again contradict $\delta = 2$ with the pairs $(\underline{0},\underline{1})$ and $(\underline{1},\underline{e})$, $w_H(\underline{e}) = 1$, in (C,D). So $C = \{\underline{0},\underline{1}\}$, $D = \bigcup_{i \text{ odd}} V_{n,i}$, n even, is the only possibility left. The reader can easily check that this choice of C and D indeed gives a δ-decodable codepair with $\delta = 2$.

We can also conclude that $|D| \leq 2^{n-1} - 1$ for n odd. The reader can easily verify that the three families of codepairs mentioned in part ii) of the theorem are δ-decodable codepairs with $\delta = 2$. That there are no other choices for (C,D) can be proved in a way similar to the above. □

3. TABLES

In [7] Kasami et al. describe the following way of translating the problem of finding a code D, given a code C, such that the codepair (C,D) is δ-decodable, to the problem of finding a coclique in a graph associated with C.

DEFINITION 3.1. Let n and δ be given integers and let C be a code in V_n with minimum Hamming distance at least δ. Then the graph $\Gamma_{C,\delta}$ on the points of V_n is defined by

$$\underline{d} \sim \underline{d}' \quad \text{if there exist words } \underline{c} \text{ and } \underline{c}' \text{ in } C \text{ s.t.}$$
$$\delta(\underline{c}+\underline{d}, \underline{c}'+\underline{d}') < \delta .$$

It is obvious from this definition that $D \subset V_n$ forms a δ-decodable codepair with C iff the points in D form a coclique in $\Gamma_{C,\delta}$. For $n \leq 5$ we have made a complete list of all linear codes C of length n. For each of these and

for each $\delta \leq d_H(C)$ a computer program has determined a maximum size coclique in $\Gamma_{C,\delta}$. The results are listed in Table I. Of course codepairs (C,D) with ratepairs inferior to other ratepairs are omitted.

For n = 6 we have only investigated the "most promising" candidates for the code C. Also the algorithm used in this case does not necessarily produce the largest coclique. The results ate listed in Table II.

In the tables V_n^* will denote $V_n \setminus \{\underline{0}\}$ or $V_n \setminus \{\underline{1}\}$. If the rate of C exceeds the rate of D then C and D are interchanged.

The linear code of these two is given by a generator matrix. The other one is presented explicitly or in terms of V_n^* or $V_{n,i}$. In Table II we use the octal notation for the vectors that are stated explicitly. For example 16 denotes the vector 100011.

n	δ	C	D	|C|	|D|	R_1	R_2	Remark
2	1	(1 1)	V_2^*	2	3	0.50000	0.79248	Theorem 2.4
2	2	(1 1)	$V_{2,1}$	2	2	0.50000	0.50000	Theorem 2.5
3	1	(1 1 1)	V_3^*	2	7	0.33333	0.93578	Theorem 2.4
3	1	$\{(\underline{x}\|0) \mid \underline{x} \in V_2^*\}$	$\begin{Bmatrix}1&1&0\\0&0&1\end{Bmatrix}$	3	4	0.52832	0.66667	Theorem 2.1
3	2	(1 1 1)	$V_{3,1}$	2	3	0.33333	0.52832	Theorem 2.5
4	1	(1 1 1 1)	V_4^*	2	15	0.25000	0.97672	Theorem 2.4
4	1	$\begin{Bmatrix}1&1&0&0\\0&0&1&1\end{Bmatrix}$	$\{(\underline{x}\|\underline{y}) \mid \underline{x},\underline{y} \in V_2^*\}$	4	9	0.50000	0.79248	Theorem 2.1
4	2	(1 1 1 1)	$V_{4,1} \cup V_{4,3}$	2	8	0.25000	0.75000	Theorem 2.5
4	2	$\begin{Bmatrix}1&1&0&0\\0&0&1&1\end{Bmatrix}$	0 0 0 0, 1 0 1 0, 1 0 0 1, 0 1 1 0, 0 1 0 1.	4	5	0.50000	0.58048	
4	4	(1 1 1 1)	1 1 0 0, 0 0 1 1.	2	2	0.25000	0.25000	Theorem 2.3
5	1	(1 1 1 1 1)	V_5^*	2	31	0.20000	0.99084	Theorem 2.4
5	1	$\begin{Bmatrix}1&1&0&0&0\\0&0&1&1&1\end{Bmatrix}$	$\{(\underline{x}\|\underline{y}) \mid \underline{x} \in V_2^*, \underline{y} \in V_3^*\}$	4	21	0.40000	0.87846	Theorem 2.1
5	1	$\begin{Bmatrix}1&1&0&0&0\\0&0&1&1&0\\0&0&0&0&1\end{Bmatrix}$	$\{(\underline{x}\|\underline{y}\|0) \mid \underline{x} \in V_2^*, \underline{y} \in V_2^*\}$	8	9	0.60000	0.63399	Theorem 2.1 (2×)
5	2	0 0 0 0 0 1 1 1 1 0 1 1 1 0 1	$\begin{Bmatrix}1&0&1&0&0\\0&1&1&0&0\\0&0&0&1&1\end{Bmatrix}$	3	8	0.31699	0.60000	
5	2	$\begin{Bmatrix}1&1&0&0&0\\0&0&1&1&1\end{Bmatrix}$	0 0 1 0 0, 0 0 0 1 0, 0 0 0 0 1, 1 0 1 1 0, 0 1 1 1 0, 1 0 1 0 1, 0 1 1 0 1, 1 0 0 1 1, 0 1 0 1 1.	4	9	0.40000	0.63399	
5	3	(1 1 1 1 1)	1 1 0 0 0, 0 0 1 1 0, 1 0 1 0 1, 0 1 0 1 1.	2	4	0.20000	0.40000	
5	4	(1 1 1 1 1)	1 1 0 0 0, 0 0 1 1 1,	2	2	0.20000	0.20000	Theorem 2.3

Table I. The best δ-decodable codepairs (C,D),

with C or D linear and $n \leq 5$.

n	δ	C	D	|C|	|D|	R_1	R_2	Remark
6	1	(1 1 1 1 1 1)	V_6^*	2	63	0.16667	0.99621	Theorem 2
6	1	$\begin{pmatrix}1&1&1&0&0&0\\0&0&0&1&1&1\end{pmatrix}$	$\{(\underline{x}\|\underline{y}) \mid \underline{x},\underline{y} \in V_3^*\}$	4	49	0.33333	0.93578	Theorem 2
6	1	$\begin{pmatrix}1&1&0&0&0&0\\0&0&1&1&0&0\\0&0&0&0&1&1\end{pmatrix}$	$\{(\underline{x}\|\underline{y}\|\underline{z}) \mid \underline{x},\underline{y},\underline{z} \in V_2^*\}$	8	27	0.50000	0.79248	Theorem 2.1
6	1	$\{(0,0\|\underline{x}\|\underline{y}) \mid \underline{x},\underline{y} \in V_2^*\}$	$\begin{pmatrix}1&0&0&0&0&0\\0&1&0&0&0&0\\0&0&1&1&0&0\\0&0&0&0&1&1\end{pmatrix}$	9	16	0.52832	0.66667	Theorem 2.1
6	2	(1 1 1 1 1 1)	$V_{6,1} \cup V_{6,3} \cup V_{6,5}$	2	32	0.16667	0.83333	Theorem 2
6	2	$\begin{pmatrix}1&1&0&0&0&0\\0&0&1&1&1&1\end{pmatrix}$	10,20,40,01,02,04, 51,61,52,62,54,64, 13,23,15,25,16,26, 43,45,46,07.	4	2	0.33333	0.74324	
6	2	$\begin{pmatrix}1&1&0&0&0&0\\0&0&1&1&0&0\\0&0&0&0&1&1\end{pmatrix}$	$V_{6,1} \cup \{(\underline{x}\|\underline{y}\|\underline{z}) \mid \underline{x},\underline{y},\underline{z} \in V_{2,1}\}$	8	14	0.50000	0.63456	
6	3	(1 1 1 1 1 1)	70,13,25,46, 14,22,41.	2	7	0.16667	0.46789	
6	3	$\begin{pmatrix}1&1&0&0&1&1\\0&0&1&1&1&1\end{pmatrix}$	00,52,23,64,15.	4	5	0.33333	0.38699	
6	4	$\begin{pmatrix}1&1&0&0&1&1\\0&0&1&1&1&1\end{pmatrix}$	101010, 010110, 011001, 100101.	4	4	0.33333	0.33333	
6	6	(1 1 1 1 1 1)	111000, 000111.	2	2	0.16667	0.16667	Theorem 2

Table II. Some good δ-decodable codepairs of length 6.

ACKNOWLEDGEMENT

The author would like to thank F.C. Bussemaker for his assistance in the computerwork that was needed for composing the two tables.

REFERENCES

[1] R. Ahlswede, "Multi-way communication channels", in Proc. 2nd Int. Symp. Inform. Theory, Tsahkadsor, Armenian S.S.R. (1971), pp. 23 - 52, Hungarian Academy of Science, 1973.

[2] P.A.B.M. Coebergh van den Braak, "Constructions and an existence result of uniquely decodable codepairs for the two-access binary adder channel". Eindhoven University of Technology, Report 83-WSK-01, Eindhoven, 1983.

[3] P.A.B.M. Coebergh van den Braak and H.C.A. van Tilborg, "A family of good uniquely decodable codepairs for the 2-access, binary adder channel", to appear in IEEE Trans. Inform. Theory.

[4] T. Kasami and S. Lin, "Coding for a multiple access-channel", IEEE Trans. Inform. Theory, vol. IT-22, pp. 129 - 137, 1976.

[5] T. Kasami and S. Lin, "Bounds on the achievable rate of block coding for a memoryless multiple-access channel", IEEE Trans. Inform. Theory, vol. IT-24, pp. 187 - 197, 1978.

[6] T. Kasami and S. Lin, "Decoding of linear δ-decodable codes for a multiple-access channel", IEEE Trans. Inform. Theory, vol. IT-24, pp. 633 - 635, 1978.

[7] T. Kasami, S. Lin, V.K. Wei and S. Yamamura, "Graph theoretic approaches to the code construction for the two-user multiple-access binary adder channel", IEEE Trans. Inform. Theory, vol. IT-29, pp. 114 - 130, 1983.

[8] G.H. Khachatrian, "On the construction of codes for noiseless synchronized 2-user channels", Problems of Control and Information Theory, vol. 11, pp. 319 - 324, 1982.

[9] G.H. Khachatrian, "A class of δ-decodable pairs of codes for the binary adder 2-user channel", Proc. of the Int. Workshop "Convolutional codes; Multi-user communication", Sochi, 232 - 234, 1983.

[10] H.J. Liao, "Multiple-access channels". Ph.D. dissertation, Dept. Elec. Eng., Univ. Hawaii, Honolulu, 1972.

[11] E.C. van der Meulen, "The discrete memoryless channel with two senders and one receiver", in Proc. 2nd Int. Symp. Inform. Theory, Tsahkadsor, Armenian S.S.R. (1971), pp. 103 - 135, Hungarian Academy of Science, 1973.

[12] H.C.A. van Tilborg, "An upperbound for codes in the two-access binary erasure channel", IEEE Trans. Inform. Theory, vol. IT-24, pp. 112 - 116, 1978.

[13] H.C.A. van Tilborg, "Upperbounds on $|C_2|$ for a uniquely decodable code pair (C_1,C_2) for a two-access binary adder channel", IEEE Trans. Inform. Theory, vol. IT-29, pp. 386 - 389, 1983.

SOME EIGENVALUE INEQUALITIES

by

G.W. Veltkamp

Dedicated to J.J. Seidel on the occasion of his retirement.

1. In the propositions subjoined to the thesis [2] of Seidel's Ph.D. student W. Haemers, the following theorem is stated (without proof).

Let

$$A = \left(\begin{array}{c|c} 0 & B \\ \hline B^H & C \end{array} \right) \tag{1}$$

be hermitean with extreme eigenvalues λ^- and λ^+. Then, if B has dimensions $k \times \ell$,

$$-\lambda^+\lambda^- \geq (k\ell)^{-1} \left| \sum_{ij} B_{ij} \right|^2 \tag{2}$$

In this note we present an inequality connecting the extreme eigenvalues of a general hermitean matrix A with the Rayleigh quotient

$$\mu := x^H Ax / x^H x$$

corresponding to some non-zero vector x and the relative length σ of the residue vector $Ax - \mu x$,

$$\sigma := \| Ax - \mu x \| / \| x \| .$$

Namely,

$$(\lambda^+ - \mu)(\mu - \lambda^-) \geq \sigma^2 \ . \tag{3}$$

For the matrix (1), an appropriate choice of x yields

$$-\lambda^+\lambda^- \geq \|B\|^2 \ , \tag{4}$$

from which (2) easily follows.

It turns out that our inequality (3) enables a simple proof of a well-known inequality of Kantorovich (cf. Householder [3]):

if A is positive definite with extreme eigenvalues λ^- and λ^+ then for $y \neq 0$

$$\frac{y^H A y \cdot y^H A^{-1} y}{(y^H y)^2} \leq \frac{(\lambda^+ + \lambda^-)^2}{4\lambda^+\lambda^-} \ , \tag{5}$$

together with a characterization of the case of equality.

Finally, we show that with (3) we can sharpen some well-known inclusion theorems of Temple which give bounds for the difference between μ and the eigenvalue of A nearest to μ in terms of σ^2 and the distance of μ to neighbouring eigenvalues of A (cf. Thms 2, 3, 4 below).

Remark. Throughout this note we use the euclidean norm for vectors and the corresponding spectral norm $\|B\| := \sup\{\|Bz\| \mid \|z\| = 1\}$ for matrices.

2. *THEOREM* 1.

Let $A \in \mathbb{C}^{n\times n}$ be hermitean with distinct eigenvalues $\lambda_m < \ldots < \lambda_1$. Let $x \in \mathbb{C}^n$ and

$$x^H x = 1, \ x^H A x = \mu, \ x^H A^2 x = \mu^2 + \sigma^2 \tag{6}$$

(involving $\|Ax - \mu x\| = \sigma$).

Then

$$(\lambda_1 - \mu)(\mu - \lambda_m) \geq \sigma^2 , \tag{7}$$

with equality iff x is a linear combination of eigenvectors belonging to λ_1 and λ_m.

Proof. The matrix $(\lambda_1 I - A)(A - \lambda_m I)$ is hermitean and positive semi-definite since its eigenvalues $(\lambda_1 - \lambda_i)(\lambda_i - \lambda_m)$, $i = 1,\ldots,m$ are non-negative. Consequently

$$\begin{aligned} 0 &\leq x^H(\lambda_1 I - A)(A - \lambda_m I)x \\ &= -\lambda_1\lambda_m + (\lambda_1 + \lambda_m)\mu - \mu^2 - \sigma^2 \\ &= (\lambda_1 - \mu)(\mu - \lambda_m) - \sigma^2 , \end{aligned}$$

which is (7). The semi-definiteness of $(\lambda_1 I - A)(A - \lambda_m I)$ implies that equality holds iff $(\lambda_1 I - A)(A - \lambda_m I)\, x = 0$, i.e., if x has zero components in all eigenspaces belonging to eigenvalues other than λ_1 and λ_m. □

Remark. Theorem and proof are valid for all $m \geq 1$. If $m = 1$ then $A = \lambda_1 I$ and $\mu = \lambda_1$, $\sigma = 0$ for any x. If $m = 2$ then $(\lambda_1 I - A)(A - \lambda_2 I) = 0$ (since all its eigenvalues are zero) and we have equality in (7).

COROLLARY. Let A be given by (1). The (4) holds.

Proof. Take $x := (x_1^T | 0)^T$ with $x_1^H x_1 = 1$. Then

$$x^H x = 1, \quad \mu = 0, \quad \sigma = \|B^H x_1\| .$$

Hence (7) asserts

$$-\lambda_1\lambda_m \geq \|B^H x_1\|^2 .$$

Since x_1 is an arbitrary unit-vector, (4) follows (using that $\|B^H\| = \|B\|$). □

Remark. Lower bounds for $\|B\|$ can be obtained by taking linear combinations of the elements of B. If $B \in \mathbb{C}^{k\times\ell}$ and $j_k \in \mathbb{R}^k$, $j_\ell \in \mathbb{R}^\ell$ are vectors, all elements of which are 1, then

$$s := j_k^T B j_\ell = \sum_{i,j} B_{ij} .$$

Since $|s| \leq \|j_k\|\|B\|\|j_\ell\| = (k\ell)^{\frac{1}{2}}\|B\|$, Haemers' bound (2) follows from (4).

3. Let, in the situation of Theorem 1, A be positive definite. From (6) and (7) we have

$$\frac{x^H A^2 x \cdot x^H x}{(x^H A x)^2} = 1 + \sigma^2/\mu^2$$

$$\leq 1 + \mu^{-2}(\lambda_1 - \mu)(\mu - \lambda_m)$$

$$= (\lambda_1 + \lambda_m)\mu^{-1} - \lambda_1\lambda_m\mu^{-2} \qquad (*)$$

$$\leq (\lambda_1 + \lambda_m)^2/(4\lambda_1\lambda_m) . \qquad (8)$$

We now replace x by $A^{-\frac{1}{2}}y$ and obtain the Kantorovich inequality (5).

Equality in (5) and (8) holds iff

i) equality holds in (7), i.e., iff x is a linear combination of eigenvectors belonging to λ_1 and λ_m,

ii) the expression (*) equals its maximum as a function of μ, i.e., if

$$\mu = 2\lambda_1\lambda_m/(\lambda_1 + \lambda_m), \text{ or } \mu^{-1} = \frac{1}{2}(\lambda_1^{-1} + \lambda_m^{-1}).$$

These conditions are equivalent with

i) y is a linear combination of eigenvectors belonging to λ_1 and λ_m,

ii) $$\frac{y^H A^{-1} y}{y^H y} = \frac{1}{2}(\lambda_1^{-1} + \lambda_m^{-1}) .$$

And this, in turn, is equivalent with

$$y = y_1 + y_m ,$$

where

i) y_1 and y_m are eigenvectors belonging to λ_1 and λ_m, respectively,

ii) $$\|y_1\| = \|y_m\| .$$

In this way we have derived from our Theorem 1 a simple proof for the Kantorovich inequality as well as sufficient and necessary conditions for equality.

4. We return to the general situation of Theorem 1 but exclude the trivial case that A has only one eigenvalue. Hence, $m \geq 2$ henceforth.

The inequality (7) implies the elementary fact that

$$\lambda_m \leq \mu \leq \lambda_1 , \tag{9}$$

i.e., for any x the Rayleigh quotient μ belongs to $[\lambda_m, \lambda_1]$. Moreover, if $\sigma > 0$, μ cannot be arbitrarily close to either λ_m or λ_1. If $\Delta := (\lambda_1 - \lambda_m)/2$,

(7) implies $0 \leq \sigma \leq \Delta$ and

$$\lambda_m + \frac{\sigma^2}{\Delta + (\Delta^2 - \sigma^2)^{\frac{1}{2}}} \leq \mu \leq \lambda_1 - \frac{\sigma^2}{\Delta + (\Delta^2 - \sigma^2)^{\frac{1}{2}}} \quad .$$

5. Our basic inequality (7) is reminiscent of an inequality associated with well-known eigenvalue bounds due to Temple [4], [5]. Namely:

for $1 \leq k \leq m-1$ we have (with the notation (6))

$$(\lambda_k - \mu)(\mu - \lambda_{k+1}) \leq \sigma^2 \quad . \tag{10}$$

This inequality derives in a way similar to the proof of Theorem 1 from the fact that $(A - \lambda_k I)(A - \lambda_{k+1} I)$ is positive semi-definite.

It should be remarked that (10), although valid for any k, is non-trivial only if $\mu \in (\lambda_{k+1}, \lambda_k)$. If μ is not equal to an eigenvalue, exactly one such k exists and (10) states that if $\sigma < \frac{1}{2}(\lambda_k - \lambda_{k+1})$, μ is either near λ_k or near λ_{k+1}.

6. We now combine (10) and (7) in order to find upper and lower bounds for an eigenvalue λ_k, expressed in terms of μ, σ and the distance of μ to neighbouring eigenvalues λ_{k+1} and λ_{k-1}. In doing so it is useful to define $\lambda_0 := \lambda_m$ and $\lambda_{m+1} := \lambda_1$ (or, more generally, to interprete all indices modulo m). From (10) and (7) we obtain the following two statements:

1) The inequality

$$\lambda_k \leq \mu + \frac{\sigma^2}{\mu - \lambda_{k+1}} \tag{11}$$

holds if either ($1 \le k \le m-1$ and $\lambda_{k+1} < \mu$) or ($k = m$ and $\mu < \lambda_1$);

2) the inequality

$$\mu - \frac{\sigma^2}{\lambda_{k-1} - \mu} \le \lambda_k \tag{12}$$

holds if either ($k = 1$ and $\lambda_m < \mu$) or ($2 \le k \le m$ and $\mu < \lambda_{k-1}$).

Remark. If $m = 2$ (i.e., if A has exactly two distinct eigenvalues) then (7) and (10) (with $k = 1$) involve

$$(\lambda_1 - \mu)(\mu - \lambda_2) = \sigma^2 .$$

Hence, in this case we have either $\sigma = 0$ and ($\mu = \lambda_1$ or $\mu = \lambda_2$) or $\sigma > 0$ and $\mu \in (\lambda_1, \lambda_2)$. In the latter case both (11) and (12) hold for $k = 1,2$ with equality signs.

7. In order to formulate elegant conditions under which two-sided bounds on λ_k exist, we introduce the notion of an open internal on the projective real line $\mathbb{P}$:

if α and $\beta \in \mathbb{P}$, $\alpha \neq \beta$, then

$$(\alpha,\beta)_{\mathbb{P}} := \{\gamma \in \mathbb{P} \mid \alpha < \gamma < \beta \text{ or } \gamma < \beta < \alpha \text{ or } \beta < \alpha < \gamma\} .$$

Hence if $\alpha < \beta$ we have $(\alpha,\beta)_{\mathbb{P}} = (\alpha,\beta)$ but if $\beta < \alpha$ then $(\alpha,\beta)_{\mathbb{P}}$ is the complement in $\mathbb{P}$ of $[\beta,\alpha]$ (here and in the sequel it is useful to observe that $\mathbb{P}$ is topologically equivalent to a circle; if this circle is oriented corresponding to the orientation on $\mathbb{P}$, the relation $\gamma \in (\alpha,\beta)_{\mathbb{P}}$ can be expressed as: α,γ,β are in ascending order on $\mathbb{P}$).

In a similar way we introduce semi-closed intervals $[\alpha,\beta)_{\mathbb{P}}$ and $(\alpha,\beta]_{\mathbb{P}}$. We remark that the three statements $\gamma \in (\alpha,\beta)_{\mathbb{P}}$, $\beta \in (\gamma,\alpha)_{\mathbb{P}}$ and $\alpha \in (\beta,\gamma)_{\mathbb{P}}$ are equivalent as are the statements $\beta \in (\gamma,\alpha]_{\mathbb{P}}$ and $\alpha \in [\beta,\gamma)_{\mathbb{P}}$.

We now are prepared to state

THEOREM 2. If $1 \le k \le m$ and $\mu \in (\lambda_{k+1},\lambda_{k-1})_{\mathbb{P}}$ then

$$\mu - \frac{\sigma^2}{\lambda_{k-1} - \mu} \le \lambda_k \le \mu + \frac{\sigma^2}{\mu - \lambda_{k+1}} . \tag{13}$$

Proof. The condition $\mu \in (\lambda_{k+1},\lambda_{k-1})_{\mathbb{P}}$ implies:

a) if $k = 1$ then $\mu \in (\lambda_2,\lambda_m)_{\mathbb{P}}$ or (since $\lambda_m \le \mu$) $\lambda_m < \lambda_2 < \mu$,

b) if $2 \le k \le m-1$ then $\lambda_{k+1} < \mu < \lambda_{k-1}$,

c) if $k = m$ then $\mu \in (\lambda_1,\lambda_{m-1})_{\mathbb{P}}$ or (since $\mu \le \lambda_1$) $\mu < \lambda_{m-1} < \lambda_1$.

It follows that in all cases the conditions for the upper and lower bounds (11) and (12) on λ_k are satisfied. □

Remark. If $m = 2$ then for no value of k $\mu \in (\lambda_{k+1},\lambda_{k-1})_{\mathbb{P}}$, so in this case the theorem as stated is vacuous. However, it follows from the remark in Section 6 that in this case either $\sigma = 0$ and ($\mu = \lambda_1$ or $\mu = \lambda_2$) or $\sigma > 0$ and $\mu \in (\lambda_2,\lambda_1)$. In the latter case (13) holds for $k = 1,2$ with equality signs.

8. For many applications it is impractical to have bounds for $\lambda_k - \mu$ involving λ_{k+1} and λ_{k-1}. Rather we would use an "upper" bound α for λ_{k+1} and a "lower" bound β for λ_{k-1} and obtain an inequality like

$$\mu - \frac{\sigma^2}{\beta - \mu} \leq \lambda_k \leq \mu + \frac{\sigma^2}{\mu - \alpha} \quad . \tag{14}$$

Using again the notion of intervals on $\mathbb{P}$, the conditions for (14) to hold can be elegantly formulated.

THEOREM 3. The inequalities (14) hold if $1 \leq k \leq m$ and

i) $\mu \in (\lambda_{k+1}, \lambda_{k-1})_{\mathbb{P}}$,

ii) $\alpha \in [\lambda_{\lambda k+1}, \mu)_{\mathbb{P}}$, $\beta \in (\mu, \lambda_{k-1}]_{\mathbb{P}}$.

Proof. Since the function $\gamma \to (\mu - \gamma)^{-1}$ is ascending from $-\infty$ to 0 if $\mu < \gamma < \infty$ and from 0 to ∞ if $-\infty < \gamma < \mu$, we have $(\mu - \gamma_1)^{-1} \leq (\mu - \gamma_2)^{-1}$ iff $\gamma_1 \in (\mu, \gamma_2]$ or, equivalently, $\gamma_2 \in [\gamma_1, \mu)_{\mathbb{P}}$. Consequently, $\alpha \in [\lambda_{k+1}, \mu)_{\mathbb{P}}$ implies $(\mu - \lambda_{k+1})^{-1} \leq (\mu - \alpha)^{-1}$ and $\beta \in (\mu, \lambda_{k-1}]_{\mathbb{P}}$ implies $-(\beta - \mu)^{-1} \leq -(\lambda_{k-1} - \mu)^{-1}$. Hence, (14) is a consequence of (13). □

Remarks.

1) Using the notion of points in ascending order on $\mathbb{P}$, the conditions i) and ii) are equivalent to

$\lambda_{k+1}, \alpha, \mu, \beta, \lambda_{k-1}$ are in ascending order on $\mathbb{P}$,

coincidence of λ_{k+1} and α, and of β and λ_{k-1} being allowed.

2) If $m = 2$ the theorem is, like Theorem 2, vacuous. However, it is easy to verify that in this case (14) holds for $k = 1$ or 2 whenever condition

ii) is satisfied.

9. A variant of Theorem 3 with slightly stronger conditions in which, however, λ_{k+1} and λ_{k-1} do not occur explicitly is

THEOREM 4. The inequalities (14) hold if

i) $\mu \in (\alpha,\beta)_{\mathbb{P}}$,

ii) $\lambda_k \in (\alpha,\beta)_{\mathbb{P}}$, $\lambda_j \notin (\alpha,\beta)_{\mathbb{P}}$ for $j \neq k \pmod m$;

(in words: if $(\alpha,\beta)_{\mathbb{P}}$ contains μ and exactly one eigenvlaue λ_k).

Proof. We will show that the conditions of this theorem imply those of Theorem 3.

a) Suppose $\alpha < \mu < \beta$. Since $\lambda_m \leq \mu \leq \lambda_1$, we have $\alpha < \lambda_1$, $\lambda_m < \beta$. Hence condition ii) implies either

$$k = 1,\ \alpha < \lambda_1 < \beta,\ \lambda_m \leq \lambda_2 \leq \alpha < \mu < \beta\ ,$$

or

$$2 \leq k \leq m-1,\ \alpha < \lambda_k < \beta,\ \lambda_{k+1} \leq \alpha < \mu < \beta \leq \lambda_{k-1}\ ,$$

or

$$k = m,\ \alpha < \lambda_m < \beta,\ \alpha < \mu < \beta \leq \lambda_{m-1} \leq \lambda_1\ .$$

b) Suppose $\beta < \alpha < \mu$. Then $\alpha < \mu \leq \lambda_1$. So condition ii) implies

$$k = 1,\ \beta < \alpha < \lambda_1,\ \beta \leq \lambda_m \leq \lambda_2 \leq \alpha < \mu.$$

c) Suppose $\mu < \beta < \alpha$. Then $\lambda_m \leq \mu < \beta$. So condition ii) implies

$$k = m,\ \lambda_m < \beta < \alpha,\ \mu < \beta \leq \lambda_{m-1} \leq \lambda_1 \leq \alpha.$$

We conclude that in all cases condition ii) of Theorem 3 is satisfied. If $m = 2$ this is sufficient for the validity of (14). If $m > 2$ then $\lambda_m < \lambda_2$ and $\lambda_{m-1} < \lambda_1$ and then in all cases condition i) of Theorem 3 is satisfied as well. □

Remark. If conditions i) and ii) are strengthened by writing (α,β) instead of $(\alpha,\beta)_{\mathbb{P}}$, Theorem 4 becomes identical with Collatz' [1] formulation of Temple's inclusion theorem. Our generalization is due to our use of our inequality (7), whereas Collatz uses the more trivial inequality (9). As a consequence we may in the case $k = 1$ allow that $\beta \in (\max(\mu,\lambda_1),\lambda_m]_{\mathbb{P}}$ whereas Collatz has $\beta \in (\max(\mu,\lambda_1),\infty]$. Consequently our formulation implies the best lower bound $\mu + \sigma^2/(\mu - \lambda_n)$ for λ_1, instead of Collatz' lower bound μ. For the upper bound for λ_m we have a similar difference.

REFERENCES

[1] Collatz, L., Funktional Analyse und numerische Mathematik, Berlin, etc., 1964 (Springer).

[2] Haemers, W., Eigenvalue techniques in design and graph theory, Thesis Eindhoven, 1979.

[3] Householder, A.S., The theory of matrices in numerical analysis, New York, 1964 (Blaisdell).

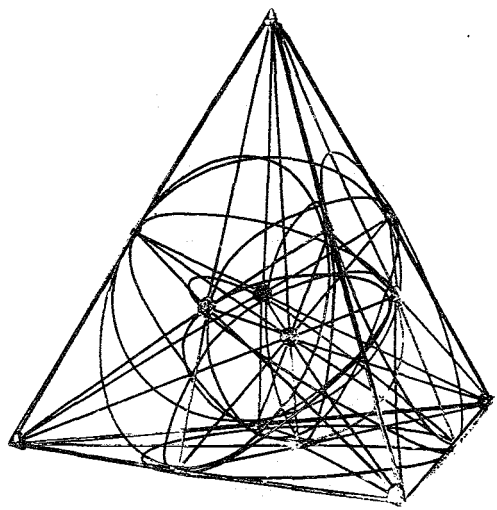
[4] Temple, G., The computation of characteristic numbers and characteristic functions, Proc. London Math. Soc. (2), 29 , 1929, 257 - 280.

[5] Temple, G., The accuracy of Rayleigh's method of calculating the natural frequencies of vibrating systems, Proc. Roy. Soc. London Ser. A, 211, 1952, 204 - 244.

LORENTZ TRANSFORMATIONS IN $V(d,\mathbb{F}_2)$ FOR $d \geq 3$ AND SOME RELATED TOPICS

by

P.G. Vroegindewey



1. General introduction
2. The exceptional case $\mathbb{F}_2$
3. Some related topics
4. The conformal closure of $V(d,\mathbb{F}_2)$

1. GENERAL INTRODUCTION

Let $V(d,\mathbb{F})$ - shortly V if no confusion is likely - denote a d-dimensional (left) vector space over a commutative field $\mathbb{F}$.

A μ-semilinear map $f : V \to V$ with respect to an automorphism $\mu : \mathbb{F} \to \mathbb{F}$ is a map such that

i) $\quad f(x+y) = f(x) + f(y)$

ii) $\quad f(\lambda x) = \mu(\lambda)\, f(x)$, $\qquad (x,y \in V\ ;\ \lambda \in \mathbb{F})$.

A quadratic form Q on V is a map $Q : V \to \mathbb{F}$ with properties:

i) $\quad Q(\lambda x) = \lambda^2 Q(x)$

ii) $\quad Q(x+y) - Q(x) - Q(y) = (x,y)$ is a symmetric bilinear form, $\qquad (x,y \in V\ ;\ \lambda \in \mathbb{F})$.

We suppose that Q is non-degenerate, i.e., $(x,a) = 0$ for all $x \in V$ and $Q(a) = 0$ imply $a = 0$.

An isotropic linear subspace D of V with respect to Q is a linear subspace with the property that $Q(x) = 0$ for all $x \in D$. As is known all maximal isotropic linear subspaces have the same dimension. This maximal dimension is called the (Witt)-index of Q. For $a \in V$ we define the cone $C(a)$ with vertex a (with respect to Q) as the set $\{x \in V \mid Q(x-a) = 0\}$.

Finally we define an *isotropic map* $f : V \to V$ as a bijection that transforms cones onto cones.

In [7] the following theorem was proved:

Assume that $d \geq 3$ and that Q has index 1. Then every isotropic map of V is a product of a translation and a semilinear map f, the latter satisfying $Q(f(x)) = c \cdot \mu(Q(x))$ for some $c \neq 0$.

The proof of this theorem uses the following lemmas:

1. Let $\ell_{a,b}$ be the line through a and b. If $b \in C(a)$ then $C(a) \cap C(b) = \ell_{a,b}$. (A line $\ell_{a,b}$ as in Lemma 1 is called an *isotropic line*.)
2. Isotropic maps transform isotropic lines onto isotropic lines.
3. Let ℓ and m be two isotropic lines without common points. Suppose that for every point $P \in \ell$ there is a point $Q \in m$ such that the line through P and Q is isotropic. Then ℓ and m are parallel.
4. Isotropic maps transform parallel isotropic lines onto parallel isotropic lines.

By a hyperbolic plane we mean a plane that contains two different isotropic directions.

5. Assume that $\mathbb{F} \neq \mathbb{F}_2$. Then isotropic maps transform hyperbolic planes onto hyperbolic planes.
6. Assume again that $\mathbb{F} \neq \mathbb{F}_2$. Then isotropic maps transform lines onto lines.

The proof of the theorem is completed after using the fundamental theorem of projective geometry. For more details and proofs of the lemmas, the reader is again referred to [7].

Finally, note that in case $d = 4$ and $\mathbb{F} = \mathbb{R}$ we recognize the classical inhomogeneous Lorentz group extended with scalar multiplications.

2. THE EXCEPTIONAL CASE $\mathbb{F}_2$

As mentioned in Section 1 the arguments used in Lemmas 5 and 6 break down for $\mathbb{F} = \mathbb{F}_2$. Therefore, in this particular case, we prove the theorem by arguing after Lemma 4 with straightforward arguments.

Note that for $\mathbb{F} = \mathbb{F}_2$ the identity is the only scalar multiplication and that semilinear maps are necessarily linear.

Furthermore, it is known (see e.g. [1] and [3]) that for $\mathbb{F} = \mathbb{F}_2$ the only possible indices are $\frac{1}{2}d - \frac{1}{2}$ (d odd) and $\frac{1}{2}d - 1$, resp. $\frac{1}{2}d$ (d even). Clearly index 1 only occurs for $d = 2,3,4$. The theorem excludes $d = 2$, hence only the cases $d = 3$ and $d = 4$ remain.

Also we deal with:

i) $d = 3$. Let $o = (0,0,0)$ $\quad p = (0,1,1)$

$a = (1,0,0)$ $\quad q = (1,0,1)$

$b = (0,1,0)$ $\quad r = (1,1,0)$

$c = (0,0,1)$ $\quad z = (1,1,1)$.

Again following [3] we represent Q by

$$Q(x) = x_1x_2 + x_3^2$$

where $x = (x_1,x_2,x_3)$.

The isotropic lines in $V(3,\mathbb{F}_2)$ are

$$oa \parallel cq \parallel pz \parallel rb ,$$

$$ob \parallel cp \parallel ar \parallel zq ,$$

$$oz \parallel cr \parallel ap \parallel bq .$$

First we consider the isotropic maps f of V that leave o invariant. These six permutations of a, b, z together with Lemma 4 give us

	o	a	b	z	c	q	p	r
f_1	o	a	b	z	c	q	p	r
f_2	o	a	z	b	c	q	r	p
f_3	o	z	b	a	c	r	p	q
f_4	o	b	a	z	c	p	q	r
f_5	o	b	z	a	c	p	r	q
f_6	o	z	a	b	c	r	q	p

Clearly, these transformations are linear and together with the 8 translations they generate the 48 elements of the inhomogeneous Lorentz group.
Hence, the theorem is proved for d = 3.

Note that c = a + b + z and hence, $f_i(c) = c$, $1 \leq i \leq 6$ and that q transforms as a $(oa \parallel cq)$, p as b $(ob \parallel cp)$ and r as z $(oz \parallel cr)$.

ii) d = 4. The elements of $V(4,\mathbb{F}_2)$ are

o = (0,0,0,0)	p = (1,1,0,0)	k = (1,1,1,0)
a = (1,0,0,0)	q = (1,0,1,0)	ℓ = (1,1,0,1)
b = (0,1,0,0)	r = (1,0,0,1)	m = (1,0,1,1)
c = (0,0,1,0)	s = (0,1,1,0)	n = (0,1,1,1)
d = (0,0,0,1)	t = (0,1,0,1)	z = (1,1,1,1)
	u = (0,0,1,1)	

We represent Q by

$$Q(x) = x_1^2 + x_2^2 + x_1x_2 + x_3x_4$$

where $x = (x_1,x_2,x_3,x_4)$.

(Compare again [3].)

The isotropic lines of C(0) are oc,od,om,on,oz and similarly as in the case d = 3 we find that the isotropic lines of $V(4,\mathbb{F}_2)$ are:

oc ‖ aq ‖ bs ‖ du ‖ pk ‖ rm ‖ tn ‖ ℓz

od ‖ ar ‖ bt ‖ cu ‖ pℓ ‖ qm ‖ sn ‖ kz

om ‖ au ‖ bz ‖ cr ‖ dq ‖ pu ‖ sℓ ‖ tk

on ‖ az ‖ bu ‖ ct ‖ ds ‖ pm ‖ qℓ ‖ rk

oz ‖ an ‖ bm ‖ cℓ ‖ dk ‖ pu ‖ qt ‖ rs .

Again referring to Lemma 4 we find that for all isotropic maps, that leave o invariant, the images of c, d, m, n, z determine the images of resp. u, r, t, ℓ, q, s, k, b, p, a and hence, we recognize the 120*) Lorentz transformations.

*) $2p^2(p^4-1)$ for p = 2.

Together with the 16 translations they generate the group of 1920 inhomogeneous Lorentz transformations representing all isotropic maps.
Hence, the theorem is proved for d = 4.

Note that, contrary to case d = 3, there is no eigenvector, for
c + d + m + n + z = 0.

3. SOME RELATED TOPICS

We return to $V(3,\mathbb{F}_2)$ representing it by $PG(2,\mathbb{F}_2)$ as in figure 1.

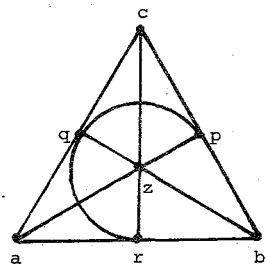


Figure 1

We meet 7 points, 7 lines, 28 conics and 7 ovals. We extend the conic abz to the oval abcz and hence, the question arises whether the theorem remains true if we replace abz (Cf. Section 1) by abcz.
This is not the case as becomes clear from the next counterexample.
Consider the map f : V → V given by

x	o	a	b	c	p	q	r	z
f(x)	o	b	c	z	q	p	r	a

f maps all cones onto cones (vertices onto vertices) but fails to be linear ($f(r) + f(z) \neq f(r+z)$).

If we check the list of lemmas we conclude that Lemmas 1 and 2 remain valid but that Lemma 3 is false, for e.g. oc, oa, cr and ar are isotropic but not oc ∥ ar.

At first sight it seems obvious to ask the similar question if we extend the quadric cdmnz in $PG(3,\mathbb{F}_2)$ to an ovoid but it soon becomes evident that in this case even the question does not make sense.
All ten candidates for extension of cdmnz are collinear with two of five elements c, d, m, n, z and hence, there is no ovoid really containing cdmnz.
Indeed, it is known that apart from quadrics, there are no other complete ovoids in $PG(3,\mathbb{F}_2)$ than the 15 maximal ovoids* consisting of $8(\neq 2^2+1)$ points outside the 4 + 6 + 4 + 1 = 15 planes. (This situation is rather exceptional in comparison with the situation in $PG(3,\mathbb{F}_{2^h})$ for $h > 1$.)

We list the ovoids and planes in $PG(3,\mathbb{F}_2)$ in the following scheme (indicate the elements of the ovoids by 0 and the elements of the planes by 1).

---

* In [4] they are called "maximal sets".

number	a	b	p	q	r	s	t	u	k	ℓ	c	d	m	n	z
1	1	1	1	1	0	1	0	0	1	0	1	0	0	0	0
2	1	1	1	0	1	0	1	0	0	1	0	1	0	0	0
3	1	0	0	1	1	0	0	1	0	0	1	1	1	0	0
4	0	1	0	0	0	1	1	1	0	0	1	1	0	1	0
5	0	0	1	0	0	0	0	1	1	1	1	1	0	0	1
6	0	1	0	1	0	0	1	0	1	0	0	1	1	0	1
7	0	1	0	0	1	1	0	0	0	1	1	0	1	0	1
8	1	0	0	0	1	1	0	0	1	0	0	1	0	1	1
9	1	0	0	1	0	0	1	0	0	1	1	0	0	1	1
10	1	1	1	0	0	0	0	1	1	1	0	0	0	1	0
11	1	0	0	0	0	1	1	1	1	1	0	0	1	0	0
12	0	1	0	1	1	0	0	1	1	1	0	0	0	1	0
13	0	0	1	0	1	0	1	0	1	0	1	0	1	1	0
14	0	0	1	1	0	1	0	0	0	1	0	1	1	1	0
15	0	0	1	1	1	1	1	1	0	0	0	0	0	0	1

It is clear from this list as well that there is no extension - as wanted - of cdmnz.

Furthermore in $V(4,\mathbb{F}_2)$ we consider the cones of index 2. Particularly the cone with vertex 0 is associated with the quadratic form Q, represented by

$$Q(x) = x_1x_2 + x_3x_4 \qquad x = (x_1,x_2,x_3,x_4)$$

As follows from [3] the 16 cones of index 2 are the complementary sets of the 16 cones of index 1. Therefore, it is immediate that our theorem in $V(4,\mathbb{F}_2)$ is also true for the cones with index 2.

Let us first give a survey of all these cones with their vertices. We indicate the elements of cones of index 1 by the number 1 (the vertex by $\underline{1}$) and similarly for the number 2.

no.	o	c	d	m	n	z	a	b	p	q	r	s	t	u	k	ℓ	no'.
1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	1'
2	1	1	2	2	2	2	2	2	2	2	1	2	1	1	2	1	2'
3	1	2	1	2	2	2	2	2	2	1	2	1	2	1	1	2	3'
4	1	2	2	1	2	2	2	1	1	1	1	2	2	2	2	2	4'
5	1	2	2	2	1	2	1	2	1	2	2	1	1	2	2	2	5'
6	1	2	2	2	2	1	1	1	2	2	2	2	2	2	1	1	6'
7	2	2	2	2	1	1	1	2	2	1	1	2	2	1	2	2	7'
8	2	2	2	1	2	1	2	1	2	2	2	1	1	1	2	2	8'
9	2	2	2	1	1	2	2	2	1	2	2	2	2	1	1	1	9'
10	2	2	1	1	2	2	1	2	2	1	2	2	1	2	2	1	10'
11	2	1	2	1	2	2	1	2	2	2	1	1	2	2	1	2	11'
12	2	2	1	2	1	2	2	1	2	2	1	1	2	2	2	1	12'
13	2	1	2	2	1	2	2	1	2	1	2	2	1	2	1	2	13'
14	2	1	1	2	2	2	1	1	1	2	2	2	2	1	2	2	14'
15	2	2	1	2	2	1	2	2	1	2	1	2	1	2	1	2	15'
16	2	1	2	2	2	1	2	2	1	1	2	1	2	2	2	1	16'

Consider in particular number 1' (it has index 2 and is complementary to number 1 with index 1).

The equation is $(x_1 - 1)(x_2 - 1) + x_3x_4 = 0$. We represent it in $PG(3, \mathbb{F}_2)$ and assign to the elements the following combinations of numbers:

a = 12; b = 13; p = 23; q = 35; r = 34; s = 25; t = 24; k = 15; ℓ = 14; u = 45.

Compare Figure 2 where the lines 234, 235, 245, 345 and 145 are omitted to avoid confusion.

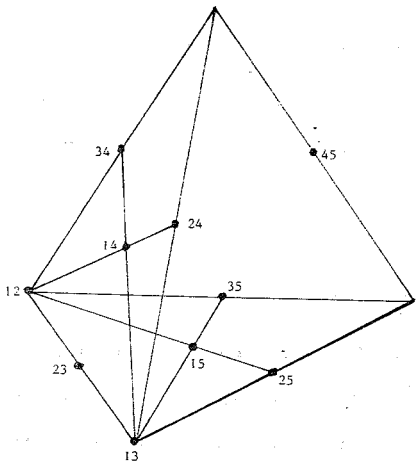


Figure 2

We shall say that two points are joined if and only if there is a third point of the configuration such that the three points are collinear. Comparison of Figure 2 and the figures 3 and 4 (Petersen graph) makes clear

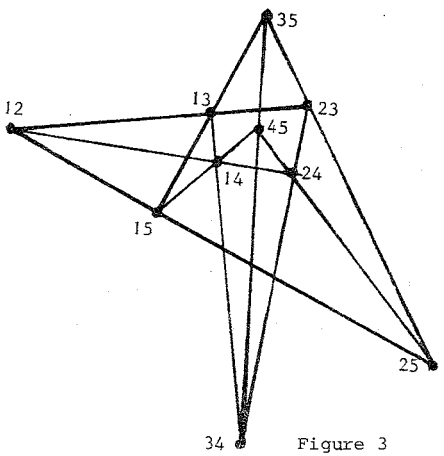


Figure 3

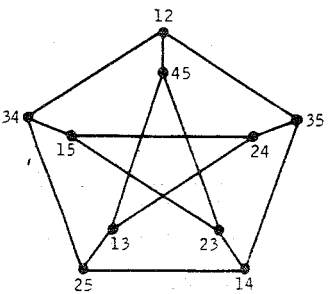


Figure 4

that there is a 1-1 correspondence between the 3 configurations (the 1-1 correspondence between the figures 3 and 4 is well known).

Remark that the 120 elements of the homogeneous Lorentz group in $V(4,\mathbb{F}_2)$ correspond to the 120 automorphisms of the Petersen graph and the Desargues configuration.

4. THE CONFORMAL CLOSURE OF $V(d,\mathbb{F}_2)$

The conformal closure $\tilde{V}$ of a space V is mostly defined by means of stereographic projection*. For details we refer the reader to e.g. [2] (only $\mathbb{F} \neq \mathbb{F}_2$) and [6] (only index 0). We start this section with some introductory remarks.

Consider $V = \mathbb{F}^d$ $(d \geq 3)$ with non-degenerate quadratic form Q with index r. Consider furthermore $\mathbb{F}^{d+2} = \mathbb{F}^d \oplus \mathbb{F}^2$; we write the elements of $\mathbb{F}^{d+2}$ as $x = (u; u_1, u_2)$, $(u \in \mathbb{F}^d)$, $u_1, u_2 \in \mathbb{F})$.

In $\mathbb{F}^{d+2}$ we introduce a quadratic form q by

$$q(x) = -Q(u) + u_1 u_2 .$$

The associated bilinear form is given by

$$\langle x,y \rangle = -(u,v) + u_1 v_2 + u_2 v_1 .$$

Note that q is also non-degenerate and has index r + 1.

* Note that the stereographic projection plays a rôle in many other branches of mathematics such as the theory of complex functions, distance geometry, Fiedler matrices (Cf. e.g. [5]).

Put $H = \{x \mid q(x) = 0\}$ and next consider projective space $P\mathbb{F}^{d+2}$. Denote the image of H in $P\mathbb{F}^{d+2}$ by PH. Consider $x = (u; u_1, u_2) \in H$ with $u_1 \neq 0$. By $\tilde{x} = (u; u_1, u_2)^{\sim}$ we denote the corresponding element of PH. The set consisting of all these $\tilde{x}$ is isomorphic to $V = \mathbb{F}^d$.

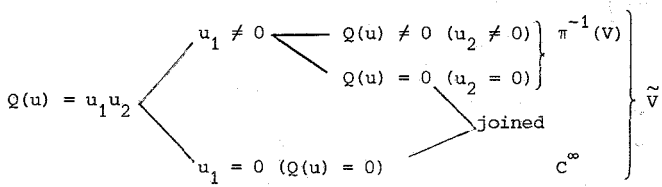
The isomorphism is given by

$$\pi(\tilde{x}) = u_1^{-1} u \ .$$

To every (isotropic) element $\tilde{x} = (u; u_1, 0)^{\sim}$ $(u_1 \neq 0,\ Q(u) = 0)$ we add the class of elements $(u; 0, u_2)$ of H as the joined element of C^{∞} to $\tilde{x}$.

By $\tilde{V}$ we mean the union of $\pi^{-1}(V)$ and C^{∞}.

It is possible to illustrate it in the following diagram



Note that the stereographic projection π^{-1} is given by

$$\pi^{-1}(t) = (t; 1, Q(t))^{\sim} \qquad t \in V \ .$$

Compare the stereographic projection in case r = 0 (e.g. the Argand plane). In that case there is a point at infinity instead of a cone at infinity. It is left to the reader to check that cones with vertex a in $V = \mathbb{F}^d$ are mapped into hyperplanes in $\mathbb{F}^{d+2}$ and that isotropic lines in $\mathbb{F}^d$ are mapped into planes in $\mathbb{F}^{d+2}$.

Lester [2] has proved that our theorem of Section 1 extends to $\tilde{V}$ in case $\mathbb{F} \neq \mathbb{F}_2$. It seems interesting to investigate what happens for $\mathbb{F} = \mathbb{F}_2$. As before we confine ourselves to index 1 and hence, to d = 3,4.

i) d = 3

$V(3,\mathbb{F}_2)$ is mapped into the quadric $x_1x_2 + x_3^2 = x_4x_5$ in $PG(4,\mathbb{F}_2)$. We find the images of o, a, b, z by adding $x_4 = 1$, $x_5 = 0$ and the images of c, p, q, r by adding $x_4 = 1$, $x_5 = 1$. The conformal closure is obtained by adding $x_4 = 0$ to o, a, b, z .

Also we obtain:

Image of $V(3,\mathbb{F}_2)$	C^∞
$(x_4 \neq 0)$	$(x_4 = 0)$
o' = (0; 1,0)	o" = (0; 0,1)
a' = (a; 1,0)	a" = (a; 0,0)
b' = (b; 1,0)	(a; 0,1)
z' = (z; 1,0)	b" = (b; 0,0)
c' = (c; 1,1)	(b; 0,1)
p' = (p; 1,1)	z" = (z; 0,0)
q' = (q; 1,1)	(z; 0,1)
r' = (r; 1,1)	

Cones with equation $Q(x - v) = 0$ transform into hyperplanes with equation $(x,v) = Q(v)x_4 + x_5$.

Hence we find:

Cone with vertex in:	Intersection of $x_1x_2 + x_3^2 = x_4x_5$ and:
o	$x_5 = 0$
a	$x_2 = x_5$
b	$x_1 = x_5$
z	$x_1 + x_2 = x_5$
c	$x_4 = x_5$
p	$x_1 + x_4 = x_5$
q	$x_2 + x_4 = x_5$
r	$x_1 + x_2 + x_4 = x_5$

Therefore the 12 isotropic lines transform in the following way:

oa	$(\alpha,0,0,\beta,0)$	o',a',a"
cq	$(\alpha,0,\beta,\beta,\beta)$	c',q',a"
pz	$(\alpha,\beta,\beta,\beta,\alpha+\beta)$	p',z',a"
rb	$(\alpha,\beta,0,\beta,\alpha)$	r',b',a"
ob	$(0,\alpha,0,\beta,0)$	o',b',b"
cp	$(0,\alpha,\beta,\beta,\beta)$	c',p',b"
ar	$(\alpha,\beta,0,\alpha,\beta)$	a',r',b"
zq	$(\alpha,\beta,\alpha,\alpha,\alpha+\beta)$	z',q',b"
oz	$(\alpha,\alpha,\alpha,\beta,0)$	o',z',z"
cr	$(\alpha,\alpha,\alpha+\beta,\beta,\beta)$	c',r',z"
ap	$(\alpha+\beta,\alpha,\alpha,\beta,\alpha)$	a',p',z"
bq	$(\alpha,\alpha+\beta,\alpha,\beta,\alpha)$	b',q',z"

Note that coplanar isotropic lines transform into planes in $\mathbb{F}^{d+2}$, intersect by a line and non-coplanar isotropic lines into planes in $\mathbb{F}^{d+2}$ with intersection (0; 0,0).

ii) d = 4

$V(4,\mathbb{F}_2)$ is mapped into the quadric $x_1^2 + x_2^2 + x_1x_2 + x_3x_4 = x_5x_6$ in $PG(5,\mathbb{F}_2)$. Similarly as in case i) we find:

Image of $V(4,\mathbb{F}_2)$ ($x_5 \neq 0$)	C^∞ ($x_5 = 0$)
o' = (0; 1,0)	o" = (0; 0,1)
c' = (c; 1,0)	c" = {(c; 0,0), (c; 0,1)}
d' = (d; 1,0)	
m' = (m; 1,0)	d" = {(d; 0,0), (d; 0,1)}
n' = (n; 1,0)	
z' = (z; 1,0)	m" = {(m; 0,0), (m; 0,1)}
a' = (a; 1,1)	
b' = (b; 1,1)	n" = {(n; 0,0), (n; 0,1)}
p' = (p; 1,1)	
q' = (q; 1,1)	z" = {(z; 0,0), (z; 0,1)}
r' = (r; 1,1)	
s' = (s; 1,1)	
t' = (t; 1,1)	
u' = (u; 1,1)	
k' = (k; 1,1)	
ℓ = (ℓ; 1,1)	

With all this data it is still a problem to prove or disprove Lester's theorem for $\tilde{V}(3,\mathbb{F}_2)$ and $\tilde{V}(4,\mathbb{F}_2)$.

Up to now I was not successful in accomplishing this.

REFERENCES

[1] Cameron, P.J. and J.J. Seidel, Quadratic forms over G(F(2). Proc. Kon. Ned. Akad. Wet. Ser. A., 76 (= Indag. Math. 35) (1973), 1-8.

[2] Lester, J.A. Conformal spaces, Journal of Geometry 14/2, (1980), 108-117.

[3] Seidel, J.J. On Two-graphs and Schult's Characterization of symplectic and orthogonal geometries over GF(2). THE-Report 73-WSK-02, Eindhoven University of Technology, Eindhoven.

[4] Seidel, J.J. Eindige Meetkunde 1979, Technische Hogeschool, Eindhoven.

[5] Seidel, J.J. Unpublished.

[6] Tits, J. Buildings of Spherical Type and finite BN-pairs, Springer-Verlag, Berlin, 1974.

[7] Vroegindewey, P.G. Some algebraic and topological investigations on space-time. Thesis, Utrecht, 1973.

ON THE (99,14,1,2) STRONGLY REGULAR GRAPH

by

H.A. Wilbrink

Dedicated to J.J. Seidel on the occasion of his retirement.

1. INTRODUCTION

One of the main open problems in the theory of strongly regular graphs is the existence question for an arbitrary set of parameters. There are strong necessary conditions known (e.g., integrality condition, Krein condition, absolute bound, see [6]) but these conditions are certainly not sufficient (cf. [7]). One of the sets of parameters for which the existence question is still open, and in which Seidel has always shown great interest, is $n = 99$, $k = 14$, $\lambda = 1$ and $\mu = 2$ (notation as in [5]). An exhaustive research for such a graph on a computer seems to be impossible. Assuming the existence of some kind of automorphism group may bring the amount of work involved back to reasonable proportions. It is the purpose of this note to study the possible automorphisms of this graph. The main result is that such a graph cannot have an automorphism of order 11, which implies that the automorphism group cannot be transitive.

The material is organized as follows. Section 2 contains some general results on automorphisms of strongly regular graphs. In Section 3 we apply these results to the hypothetical strongly regular graph on 99 points and of valency 14.

2. AUTOMORPHISMS OF STRONGLY REGULAR GRAPHS

This section contains some results of a general nature on automorphisms of strongly regular graphs. The first result gives a bound on the number of points that can be fixed by an automorphism.

THEOREM 1. Let $G = (\Omega,\Gamma)$ be a strongly regular graph (with point set Ω and Γ as the set of edges). Let G have parametrs n, k, λ and μ and let r denote the positive eigenvalue $\neq k$. If $g \neq 1$ is an automorphism of G and Fix(g) is the set of points fixed by g, then

$$|\mathrm{Fix}(g)| \leq \max\{\lambda,\mu\} \frac{n}{k-r} .$$

In addition, if g has prime order p and $p>\max\{\lambda,\mu\}$, then

$$|\mathrm{Fix}(g)| \leq \frac{n}{k-r} .$$

Proof. Partition the adjacency matrix A of G according to the sets of non-fixed and fixed points.

$$A = \left(\begin{array}{c|c} A_{11} & A_{12} \\ \hline A_{12}^{t} & A_{22} \end{array} \right).$$

Thus, A_{11} is the adjacency matrix of the subgraph induced on the nonfixed points, A_{22} is the adjacency matrix of the subgraph induced on the fixed points. By a result of Haemers (see [4, Th. 2.1.4]) ,

$$r \geq \frac{n k_1 - n_1 k}{n - n_1} ,$$

where $n_1 = n - |\mathrm{Fix}(g)|$ is the size of A_{11}, and k_1 is the average valency of the graph induced on $\Omega \backslash \mathrm{Fix}(g)$.

It follows that

$$|\mathrm{Fix}(g)| \leq (k - k_1) \frac{n}{k - r} .$$

Suppose $\alpha \in \Omega$ is nonfixed. If $\beta \in \Omega$ is fixed and adjacent to α, then β is also adjacent to α^g. Hence, if $\alpha \sim \alpha^g$ then at most λ fixed points are adjacent to α, and if $\alpha \not\sim \alpha^g$ then at most μ fixed points are adjacent to α. Therefore at least $\min\{k-\lambda, k-\mu\}$ nonfixed vertices are adjacent to α. Thus $k_1 \geq \min\{k-\lambda, k-\mu\}$ from which the first inequality follows.

Suppose the order of g is a prime $p > \max\{\lambda,\mu\}$. Then clearly, a nonfixed point cannot be adjacent to two fixed points, i.e., $k_1 \geq k-1$ in this case, from which the second inequality follows. □

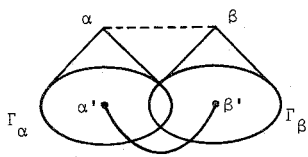
The case of an automorphism g of order $p > \max\{\lambda,\mu\}$ is worth a closer look. For any two distinct fixed points α and β, the common neighbours of α and β must also be fixed, i.e., the graph induced on $\mathrm{Fix}(g)$ is also strongly regular (with the same λ and μ) provided this graph is regular. The following (unpublished) result of A.E. Brouwer shows that this is usually the case.

THEOREM 2. Let $G = (\Omega,\Gamma)$ be a graph with the property that for some integers λ and $\mu > 0$ and all distinct $\alpha,\beta \in \Omega$

$$|\Gamma_\alpha \cap \Gamma_\beta| = \begin{cases} \lambda & \text{if } \alpha \sim \beta , \\ \mu & \text{if } \alpha \not\sim \beta . \end{cases}$$

Then either G is regular, or $\mu = 1$ and G is the union of a number of cliques of size $\lambda + 2$ all intersecting in one common point.

Proof. Take $\alpha,\beta \in \Omega$, $\alpha \neq \beta$, and count in two ways the number of pairs (α',β'), $\alpha' \neq \beta'$, $\alpha' \in \Gamma_\alpha$, $\beta' \in \Gamma_\beta$, $\alpha' \sim \beta'$.



i) If $\alpha \not\sim \beta$ this gives the equation

$$(|\Gamma_\alpha| - \mu)\mu + \mu\lambda = (|\Gamma_\beta| - \mu)\mu + \mu\lambda.$$

(For: there are $|\Gamma_\alpha| - \mu$ choices for $\alpha' \not\sim \beta$ and each such α' is adjacent to μ vertices in Γ_β, and there are μ choices for $\alpha' \sim \beta$ and each such α' is adjacent to λ vertices in Γ_β.)

Hence, $|\Gamma_\alpha| = |\Gamma_\beta|$ if $\alpha \not\sim \beta$.

ii) If $\alpha \sim \beta$ the same count yields

$$(|\Gamma_\alpha| - \lambda)(\mu - 1) + \lambda(\lambda - 1) = (|\Gamma_\beta| - \lambda)(\mu - 1) + \lambda(\lambda - 1) .$$

Hence, if $\mu > 1$, then $|\Gamma_\alpha| = |\Gamma_\beta|$ for all $\beta \in \Omega$. Therefore, if G is not regular, then $\mu = 1$ and then i) implies that $\bar{G}$, the complement of G,

is not connected, with at least one component containing more than one point. Since $\mu = 1$, there can only be one other $\bar{G}$-component necessarily containing a single point. □

COROLLARY 3. Let G be a strongly regular graph with parameters n, k, λ, and $\mu > 1$, and let g be an automorphism of G of prime order $p > \max\{\lambda,\mu\}$. Then one of the following holds.

(i) $\mathrm{Fix}(g) = \phi$,

(ii) $|\mathrm{Fix}(g)| = 1$,

(iii) The graph induced on Fix(g) is a clique of size $\lambda + 2$,

(iv) The graph induced on Fix(g) is strongly regular (with the same λ and μ).

Consider the general situation of a strongly regular graph $G = (\Omega,\Gamma)$ with an automorphism group G having t orbits $\Omega_1,\Omega_2,\ldots,\Omega_t$. Then every point in Ω_j is adjacent to c_{ij} points in Ω_i where c_{ij} is independent of the choice of $\alpha \in \Omega_j$. In matrix terminology , if the adjacency matrix A of G is partitioned according to $\Omega_1,\ldots,\Omega_t$:

$$A = \begin{pmatrix} A_{11} & \cdots & A_{1t} \\ & & \\ A_{t1} & \cdots & A_{tt} \end{pmatrix} ,$$

then the submatrices A_{ij} have constant column sums c_{ij} (and constant row sums $r_{ij} = c_{ij}$). The following result shows that the *orbit matrix* $C = (c_{ij})$ satisfies a matrix equation similar to the matrix equation $A^2+(\mu-\lambda)A+(\mu-k)I= = \mu J$ for A. Since t is usually much smaller than n, the equation for C is

easier to solve. If no solution exists, then one has shown that G cannot admit G as an automorphism group. If a solution is found one can then try to replace the numbers c_{ij} by 0/1-matrices A_{ij} to form the matrix A, a problem of much smaller complexity (see e.g. [1] where these ideas were used to construct a symmetric 2-design).

THEOREM 4. Let $G = (\Omega,\Gamma)$ be a stongly regular graph with adjacency matrix A, parameters n, k, λ and μ and eigenvalues r (>0) and s (<0). Let $G \leq \mathrm{Aut}(G)$ have t orbits Ω_i of length n_i, $1 \leq i \leq t$ on Ω. Let c_{ij} (resp. r_{ij}) be the column sums (resp. row sums) of the submatrices A_{ij} of A induced by $\Omega_1,\ldots,\Omega_t$. Define the $t \times t$-matrices C, R and N by

$$C = (c_{ij}) \ , \ R = (r_{ij}) \ , \ N = \mathrm{diag}(n_1,\ldots,n_t).$$

Then,

(i) $c_{ij} = r_{ji}$, $\sum_{i=1}^{t} c_{ij} = k$, $\sum_{j=1}^{t} r_{ij} = k$, $c_{ij}n_j = n_i r_{ij}$,

$$\sum_{k=1}^{t} c_{ik}c_{kj} + (\mu-\lambda)c_{ij} + (\mu-k)\delta_{ij} = n_i\mu;$$

(ii) $C = R^T$, $J_t C = kJ_t$, $RJ_t = kJ_t$, $CN = NR$

$$C^2 + (\mu-\lambda)C + (\mu-k)I = \mu NJ;$$

(iii) Every eigenvalue of C is also an eigenvalue of A with at least the same multiplicity;

(iv) $\mathrm{tr}(C) \equiv k + s(t-1) \mod (r-s)$.

Proof. Clearly, (ii) follows from (i). As for (i), only the last equation is not completely trivial. Express the matrix equation $A^2 + (\mu-\lambda)A + (\mu-k)I = \mu J$ in terms of the A_{ij}:

$$\sum_{k=1}^{t} A_{ik}A_{kj} + (\mu-\lambda)A_{ij} = \mu J_{n_i \times n_j} \quad , \; i \neq j \; ,$$

$$\sum_{k=1}^{t} A_{ik}A_{ki} + (\mu-\lambda)A_{ii} + (\mu-k)I_{n_i} = \mu J_{n_i} \quad .$$

Multiply these equations on the left with the all one row vector of size n_i. To prove (iii), note that if $xC = \lambda x$, then

$$\tilde{x} = (\underbrace{x_1,\ldots x_1}_{n_1}, \underbrace{x_2\ldots x_2}_{n_2}, \ldots, \underbrace{x_t\ldots x_t}_{n_t})$$

satisfies $\tilde{x}A = \lambda\tilde{x}$.

Finally, (iv) follows from the equations

$$\begin{cases} 1 + f' + g' = t \; , \\ k + rf' + sg' = \mathrm{tr}(C) \; , \end{cases}$$

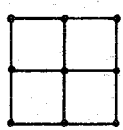
where f' and g' are the multiplicities of the eigenvalues r and s of C respectively. □

3. APPLICATION TO THE (99,14,1,2) GRAPH

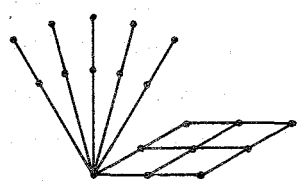
In this section, let $G = (\Omega,\Gamma)$ be a strongly regular graph on n = 99 points of valency k = 14, $\lambda = 1$ and $\mu = 2$. The eigenvalues of G are r = 3 (multi-

plicity 54) and $s = -4$ (multiplicity 44). The fact that $\lambda = 1$ makes it possible to view G as a kind of geometry with lines of size 3. In fact, G is a semi-partial geometry in the sense of [3] and also a partial quadrangle in the sense of [2], with 3 points on each line and 7 lines through each point.

Suppose $g \in \mathrm{Aut}(G)$ has prime order $p > 7$. If g fixes some $\alpha \in \Omega$, then g also fixes each of the 7 lines through α and therefore all 14 points adjacent to α. Since $\mu = 2$, this implies that g fixes all points, a contradiction. Therefore no point can be fixed by g, so $p \mid 99$ which means that $p = 11$. The only possible primes dividing the order of $\mathrm{Aut}(G)$ are therefore 2,3,5,7 and 11. By Theorem 1, $|\mathrm{Fix}(g)| \leq 18$ if g has order 2, and $|\mathrm{Fix}(g)| \leq 9$ if g has order 3,5,7 or 11. From Corollary 3 it now follows that there are only the following possibilities for an automorphism g of prime order.

order of g	Fix(g)
2	at most 17 points
3	$\emptyset$ or 
5	 $= L_2(3)$
7	• (a single point)
11	$\emptyset$

The most interesting cases seem to be p = 5, 7 and 11. Also, automorphisms of order 9 and 10 could be interesting. For an automorphism of order 10 with an $L_2(3)$ as a set of fixed points and 9 cycles of length 10 one can read off the orbit matrix from the geometric picture.



$$C = \left(\begin{array}{c|c} A_1 & I_9 \\ \hline 10I_9 & 2J_9 - I_9 - A_1 \end{array}\right) ,$$

where A_1 is the adjacency matrix of the 3×3 grid.

It would be interesting to try to extend this orbit matrix to an adjacency matrix for G. The following result shows that p = 11 is impossible.

THEOREM 5. The graph G has no automorphism of order 11.

Proof. The corresponding orbit matrix cannot exist. The first step is to determine all possible rows that can occur in the matrix C. Note that C is symmetric since $n_i = 11$ for all $i = 1,\ldots,9$. From Theorem 4 is follows that the i^{th} row of C satisfies

$$\sum_{j=1}^{9} c_{ij} = 14 \quad \text{and} \quad \sum_{j=1}^{9} c_{ij}^2 = -c_{ii} + 34 .$$

Clearly c_{ii} must be even (consider the equations mod 2 or note that a regular graph on 11 points has even valency), and $c_{ii} < 6$. There are 7 essentially different solutions of these equations (the underlined member in each solution is the diagonal element c_{ii}):

a) $\underline{4}\ 2\ 2\ 1\ 1\ 1\ 1\ 1\ 1$

b) $\underline{2}\ 4\ 2\ 2\ 1\ 1\ 1\ 1\ 0$

c) $\underline{2}\ 3\ 3\ 2\ 2\ 1\ 1\ 0\ 0$

d) $\underline{0}\ 4\ 3\ 2\ 1\ 1\ 1\ 1\ 1$

e) $\underline{0}\ 4\ 2\ 2\ 2\ 2\ 1\ 1\ 0$

f) $\underline{0}\ 3\ 3\ 3\ 2\ 1\ 1\ 1\ 0$

g) $\underline{0}\ 3\ 3\ 2\ 2\ 2\ 2\ 0\ 0$

The rest of the proof is in essence an exhaustive search, partly carried out by hand and partly done on a programmable pocket calculator. By hand it is possible to exclude rows of type a, b, d and e. A sketch of this follws.

i) *C does not contain rows of type a.*

The inner product of any two rows of C is

$$\sum_{k=1}^{9} c_{ik}c_{jk} = -c_{ij} + 22.$$

Thus a second row looks like

$$\begin{array}{cc|ccccccc} \underline{4} & 2 & 2 & 1 & 1 & 1 & 1 & 1 & 1 \\ 2 & \underline{x} & . & . & . & . & . & . & . \end{array} \Big) 12 - 2x$$

where $12 - 2x$ is the inner product the two rows must have to the right of the vertical line. The sum of the elements of the second row to the right of this line is $14 - (2 + x) = 12 - x$. Hence, $x = 0$ and $c_{23} = 0$. There remain 3 possibilities.

$$\text{ae}: \quad \begin{matrix} \underline{4}\,2\,2\,1\,1\,1\,1\,1\,1 \\ 2\,\underline{0}\,0\,4\,2\,2\,2\,2\,1 \end{matrix},$$

$$\text{af}: \quad \begin{matrix} \underline{4}\,2\,2\,1\,1\,1\,1\,1\,1 \\ 2\,\underline{0}\,0\,3\,3\,3\,1\,1\,1 \end{matrix},$$

$$\text{ag}: \quad \begin{matrix} \underline{4}\,2\,2\,1\,1\,1\,1\,1\,1 \\ 2\,\underline{0}\,0\,3\,3\,2\,2\,2\,0 \end{matrix}.$$

Continuing case ae gives

$$\left.\begin{array}{l|l|l} \underline{4}\,2\,2 & 1\,1\,1\,1\,1\,1 & \\ 2\,\underline{0}\,0 & \left.\begin{matrix} 4\,2\,2\,2\,2\,1 \\ \cdot\,\cdot\,\cdot\,\cdot\,\cdot\,\cdot \end{matrix}\right\} 18 & \Big\} 12 \\ 2\,0\,\underline{0} & & \end{array}\right. .$$

Clearly, a third row must also be of type e, f or g but none of these works. Case ag does not allow a third row either and for af the only possibility is

$$\left.\begin{array}{l|l|l} \underline{4}\,2\,2\,1 & 1\,1\,1\,1\,1 & \\ 2\,\underline{0}\,0\,3 & 3\,3\,1\,1\,1 & \Big\} 17-3x \\ \left.\begin{matrix} 2\,0\,0\,1 \\ 1\,3\,1\,\underline{x} \end{matrix}\right| & \left.\begin{matrix} 1\,1\,3\,3\,3 \\ \cdot\,\cdot\,\cdot\,\cdot\,\cdot \end{matrix}\right\} 19-x & \Big\} 9-x \end{array}\right.$$

The fourth row (and all further rows) must be of type c, d or f. Since $\mathrm{tr}(C) \equiv 3 \bmod 7$, type c must occur. The only possibility is

$\underline{4}\ 2\ 2\ 1\ 1\ 1\ 1\ 1\ 1$

$2\ \underline{0}\ 0\ 3\ 3\ 3\ 1\ 1\ 1$

$2\ 0\ \underline{0}\ 1\ 1\ 1\ 3\ 3\ 3$

$1\ 3\ 1\ \underline{2}\ 2\ 0\ 3\ 2\ 0$

and it is easy to check that no fifth row is possible.

ii) *C does not contain rows of type b.*

This is the hardest case.

$$\begin{array}{rl} & \underline{2}\ 4 \\ \text{b,d or e} \to 4 & \underline{x} \end{array} \left| \begin{array}{c} 2\ 2\ 1\ 1\ 1\ 1\ 0 \\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot \end{array} \right) 10 - 4x \ .$$

Clearly a second row of type b is impossible.

For a second row of type d there are essentially two possibilities.

$$bd_1: \begin{array}{l} \underline{2}\ 4\ 2\ 2\ 1\ 1\ 1\ 1\ 0 \\ 4\ \underline{0}\ 2\ 1\ 1\ 1\ 1\ 1\ 3 \end{array} \qquad bd_2: \begin{array}{l} \underline{2}\ 4\ 2\ 2\ 1\ 1\ 1\ 1\ 0 \\ 4\ \underline{0}\ 1\ 1\ 3\ 1\ 1\ 1\ 2 \end{array} \ .$$

For a second row of type e there are 3 possibilities.

$$be_1: \begin{array}{l} \underline{2}\ 4\ 2\ 2\ 1\ 1\ 1\ 1\ 0 \\ 4\ \underline{0}\ 2\ 0\ 2\ 2\ 1\ 1\ 2 \end{array} \qquad be_2: \begin{array}{l} \underline{2}\ 4\ 2\ 2\ 1\ 1\ 1\ 1\ 0 \\ 4\ \underline{0}\ 1\ 1\ 2\ 2\ 2\ 0\ 2 \end{array} \qquad be_3: \begin{array}{l} \underline{2}\ 4\ 2\ 2\ 1\ 1\ 1\ 1\ 0 \\ 4\ \underline{0}\ 1\ 0\ 2\ 2\ 2\ 2\ 1 \end{array}$$

All five cases die sooner or later. In each case it is advisable to take the last column as the third row.

For example bd_1:

$$\left.\begin{array}{rl} & \underline{2}\ 4\ 0 \\ & 4\ \underline{0}\ 3 \\ \text{c,f or g} \to 0\ 3 & \underline{x} \end{array} \left| \begin{array}{c} 2\ 2\ 1\ 1\ 1\ 1 \\ 2\ 1\ 1\ 1\ 1\ 1 \\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot \end{array} \right) 19 - 3x \right) 10$$

None of c, f and g can be used as a third row.

iii) *C does not contain rows of type d.*

The only possibility for a second row is

$$\underline{0}\ 4\ 3\ 2\ 1\ 1\ 1\ 1\ 1$$
$$4\ \underline{0}\ 3\ 2\ 1\ 1\ 1\ 1\ 1 \quad .$$

No further row can be added.

iv) *C does not contain rows of type e.*

The first two rows must be

$$\underline{0}\ 4\ 2\ 2\ 2\ 2\ 1\ 1\ 0$$
$$4\ \underline{0}\ 2\ 2\ 2\ 2\ 1\ 1\ 0 \quad .$$

Take the last column as the third row. Then two extensions are possible.

$$\underline{0}\ 4\ 0\ 2\ 2\ 2\ 2\ 1\ 1 \qquad \underline{0}\ 4\ 0\ 2\ 2\ 2\ 2\ 1\ 1$$
$$4\ \underline{0}\ 0\ 2\ 2\ 2\ 2\ 1\ 1 \quad \text{and} \quad 4\ \underline{0}\ 0\ 2\ 2\ 2\ 2\ 1\ 1 \quad .$$
$$0\ 0\ \underline{2}\ 3\ 3\ 2\ 2\ 1\ 1 \qquad 0\ 0\ \underline{0}\ 2\ 2\ 2\ 2\ 3\ 3$$

The second alternative is impossible: since rows 4, 5, 6 and 7 cannot be of type c, $tr(C) \not\equiv 3 \bmod 7$, a contradiction. The first alternative quickly dies if one takes the last column as the fourth row.

For the remaining cases with rows of type c, f and g only, an exhaustive search of 6 hours on a pocket calculator revealed that no solution exists containing a row of type c. Since $tr(C) \equiv 3 \bmod 7$, this shows that there can be no solution at all. □

REFERENCES

[1] Brouwer, A.E. & H.A. Wilbrink, A symmetric design with parameters 2-(49,16,5), to appear in J.C.T.

[2] Cameron, P.J., Partial quadrangles, Quart. J. Math. 26 (1975), 61 - 73.

[3] Debroey, I. & J.A. Thas, On semipartial geometries, J.C.T. (A) 25 (1978), 242 - 250.

[4] Haemers, W., Eigenvalue techniques in design and graph theory, Thesis, Eindhoven (1979).

[5] Seidel, J.J., Strongly regular graphs, pp. 185 - 197 in "Progress in combinatorics", ed. W.Tutte, Acad. Press (1969).

[6] Seidel, J.J., Strongly regular graphs, an introduction, Proc. 7th British Combin. Confer., Cambridge 1979.

[7] Wilbrink, H.A. & A.E. Brouwer, A (57,14,1) strongly regular graph does not exist. Proc. KNAW, Series A, 86 (1) (1983).

DRIEHOEKEN MET GEGEVEN SPIEGELPUNTSDRIEHOEK

door

J. van IJzeren

Opgedragen aan J.J. Seidel ter gelegenheid van zijn afscheid als hoogleraar aan de Technische Hogeschool Eindhoven.

1. Bij een willekeurige driehoek $A_1A_2A_3$ (zijden a_i, $i = 1,2,3$) kan men de spiegelpunten S_i van A_i in a_i construeren. Maar lukt het omgekeerde: gegeven een "spiegelpuntsdriehoek" $S_1S_2S_3$, is er een passer-en-lineaal constructie voor een bijbehorende "basisdriehoek" $A_1A_2A_3$?

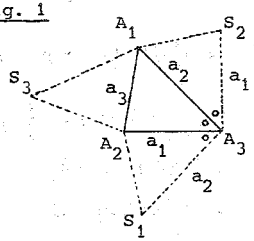
   Dit probleem, al bijna 50 jaar geleden door *Bottema* gesteld en voor een speciaal geval ten dele opgelost [1], heeft verrassende kanten. Zo kan men bij gelijkbenige $S_1S_2S_3$ makkelijk basisdriehoeken construeren, maar alléén niet-gelijkbenige [2]. Wel bestaan er bij gelijkbenige $S_1S_2S_3$ óók gelijkbenige $A_1A_2A_3$, maar die zijn nu juist in het algemeen niet construeerbaar.

   Deze eigenaardige stand van zaken wordt pas goed duidelijk bij een algemene aanpak: "gegeven een willekeurige spiegelpuntsdriehoek, bepaal *alle* basisdriehoeken en onderzoek hun construeerbaarheid".

   Deze opgave blijkt algebraïsch goed hanteerbaar, in samenhang met meetkundige overwegingen. We beginnen met wat trigonometrie.

2.

Fig. 1



In Fig. 1 is te zien dat $S_1S_2^2 = a_1^2 + a_2^2 - 2a_1a_2 \cos 3A_3$. Combinatie met $a_1^2 + a_2^2 - 2a_1a_2 \cos A_3 = a_3^2$ geeft

$$S_1S_2^2 = a_3^2 + 2a_1a_2(\cos A_3 - \cos 3A_3) =$$

$$= a_3^2 + 4a_1a_2 \sin A_3 \sin 2A_3 =$$

$$= a_3^2(1 + 8 \sin A_1 \sin A_2 \cos A_3) \ .$$

Nog eens de sinus- en cosinusregel toepassend komt men tot

$$S_1S_2^2 = a_3^2(1 + 4 \sin^2 A_1 + 4 \sin^2 A_2 - 4 \sin^2 A_3) \ .$$

Er bestaat blijkbaar een evenredigheidsfactor λ waarmee voor $i = 1,2,3$:

$$\lambda \sin^2 S_i = \sin^2 A_i (1 + 4 \sin^2 A_1 + 4 \sin^2 A_2 + 4 \sin^2 A_3) - 8 \sin^4 A_i \ . \qquad 2.1$$

Hierin fungeert $\sin^2 S_i$ als *gegeven*. We voeren in

$$\sin^2 S_1 + \sin^2 S_2 + \sin^2 S_3 =: S \ , \quad \sin^2 S_1 \sin^2 S_2 \sin^2 S_3 =: P \ .$$

Daar komt bij (kwadrateer $\sin^2 S_1 + \sin^2 S_2 - \sin^2 S_3 = 2 \sin S_1 \sin S_2 \cos S_3$)

$$\sin^2 S_1 \sin^2 S_2 + \sin^2 S_2 \sin^2 S_3 + \sin^2 S_3 \sin^2 S_1 = \tfrac{1}{4}S^2 + P \ . \qquad 2.2$$

Blijkbaar zijn $\sin^2 S_1$, $\sin^2 S_2$ en $\sin^2 S_3$ de wortels van

$$f(u) := u^3 - Su^2 + (\tfrac{1}{4}S^2 + P)u - P = 0 \ . \qquad 2.3$$

Omgekeerd: zij $S > 0$, $P > 0$ en $f(u) = u(u - \frac{1}{2}S)^2 + P(u-1)$. Voor $u \leq 0$ of $u > 1$ vindt men $f(u) \neq 0$. Complexe wortels worden uitgesloten door te eisen dat de discriminant ≥ 0 is. Daarbij een factor $P/16$ weglatend stellen we

$$D(S,P) := (9 - 4S)^3 - (8P + 2S^2 - 18S + 27)^2 \geq 0 \ . \qquad 2.4$$

Is $D(S,P) = 0$, dan heeft $f(u) = 0$ gelijke wortels (meetkundig: $S_1S_2S_3$ is gelijkbenig). Duidelijk is dat $S \leq \frac{9}{4}$ moet zijn. Is bijv. $S = 2$, dan is er voor P enige speling: $0 \leq P \leq \frac{1}{4}$, maar $S = \frac{9}{4}$ leidt tot $P = \frac{27}{64}$.

Getallen S, P die aan 2.4 voldoen bepalen $\sin^2 S_i$ en daarmee de *vorm* van $S_1S_2S_3$. Zo is $S = \frac{9}{4}$, $P = \frac{27}{64}$ kenmerkend voor gelijkzijdigheid. Bij $S = 2$ vindt men $u_1 = 1 = u_2 + u_3$, dus hoeken 90°, σ, $90^\circ - \sigma$, kortom rechthoekige $S_1S_2S_3$. Daarbij volgt σ uit $P = 1^2 \sin^2\sigma \sin^2(90^\circ - \sigma) = \frac{1}{4}\sin^2 2\sigma$, $\sigma \leq 45^\circ$.

Ook $S = 0$ blijkt zinvol. Hierbij horen "lineaire driehoeken" $S_1S_2S_3$ met hoeken van 180° en 0°. De "vorm" (verhouding der zijden) wordt wederom bepaald door S en P (namelijk door P/S^3, een functie van afstanden).

Hoe nu de onbekenden $\sin^2 A_i =: x_i$ te bepalen? We voeren in

$$x_1 + x_2 + x_3 =: s \ , \quad x_1x_2x_3 =: p \ .$$

Hiermee wordt 2.1

$$\lambda \sin^2 S_i = x_i(1 + 4s - 8x_i) \qquad 2.5$$

en overeenkomstig 2.2 geldt

$$x_1x_2 + x_2x_3 + x_3x_1 = \tfrac{1}{4}s^2 + p\,, \quad x_1^2 + x_2^2 + x_3^2 = \tfrac{1}{2}s^2 - 2p\,. \qquad 2.6$$

Doel is: de vorm van $A_1A_2A_3$ af te leiden uit die van $S_1S_2S_3$, dus s en p uit S en P. Door hun symmetrie in $\sin^2 S_i$, i = 1,2,3 zijn S en P symmetrische functies van x_1, x_2, x_3 en dus te herleiden tot uitdrukkingen in s en p. Daaruit zullen we, als S en P *gegeven* zijn, s en p moeten oplossen, om tenslotte $\sin^2 A_i$ te vinden.

Eerst dus, met het oog op S en P, aandacht voor symmetrische vormen. Daarbij verkorten we 2.5 tot $\lambda \sin^2 S_i = x_i \ell_i$ met $\ell_i := 1 + 4s - 8x_i$. Voor de hand ligt

$$\lambda S = x_1\ell_1 + x_2\ell_2 + x_3\ell_3 = s(1+4s) - 8(x_1^2 + x_2^2 + x_3^2) = s + 16p\,.$$

Vervolgens geeft 2.2

$$\lambda^2 P = x_1x_2\ell_1\ell_2 + x_2x_3\ell_2\ell_3 + x_3x_1\ell_3\ell_1 - \tfrac{1}{4}(\lambda S)^2\,.$$

Herleiding (bijv. $x_1^2x_2^2 + x_2^2x_3^2 + x_3^2x_1^2 = (\frac{1}{4}s^2 + p)^2 - 2sp$) levert

$$\lambda^2 P = p(4s-5)^2\,. \qquad 2.7$$

Gemakkelijk volgt tenslotte

$$\lambda^3 P = x_1x_2x_3\ell_1\ell_2\ell_3 = p[4s + 1 + 64p(4s-7)]\,.$$

Gebruik nu $(\lambda S)(\lambda^2 P) = S \cdot (\lambda^3 P)$ en $(\lambda^2 P)^3 = P \cdot (\lambda^3 P)^2$, dan komt er

$$(s+16p)(4s-5)^2 = S \cdot [4s + 1 + 64p(4s-7)]\,, \qquad 2.8$$

$$p(4s-5)^6 = P \cdot [4s + 1 + 64p(4s-7)]^2\,. \qquad 2.9$$

Hiermee zijn S en P uitgedrukt in s en p.

We kunnen 2.8, waarin p lineair voorkomt, herleiden tot

$$s(4s-5)^2 - S(4s+1) + 16p[(4s-5)^2 - 4S(4s-7)] = 0 \ , \qquad 2.10$$

Hiermee kan p uit 2.9 worden geëlimineerd. Dit leidt tot

$$[s(4s-5)^2 - S(4s+1)][(4s-5)^2 - 4S(4s-7)](4s-5)^2 + \\ + 16P(16s^2 - 32s - 1)^2 = 0 \ . \qquad 2.11$$

Aldus hebben we voor s een 7^e-graadsvergelijking met parameters S en P.

Kies nu een bepaalde reële wortel s. Dan geeft 2.10 de bijbehorende p. Deze voldoet niet alleen aan 2.8 maar ook aan 2.9 en is dus ≥ 0. Voor het paar s,p geldt echter nog veel meer. Zo kunnen we D(S,P) er in uitdrukken, gebruikmakend van 2.8 en 2.9:

$$D(S,P)[4s+1+64p(4s-7)]^4 = D(s,p)(4s-5)^6[4s-1+64p(4s-9)]^2 \ . \qquad 2.12$$

We zien hieruit dat $D(s,p) \geq 0$. Gevolg: bij het paar s,p hoort een derdegraadsvergelijking als 2.3, met reële wortels op (0,1). Deze geven $\sin^2 A_i$, $i = 1,2,3$, van de basisdriehoek $A_1A_2A_3$ behorend bij de wortel s.

Kennelijk leiden 7 *verschillende* reële wortels s tot 7 basisdriehoeken die onderling niet-gelijkvormig zijn (ook niet gespiegeld).

3. We beschouwen drie speciale gevallen.

Allereerst zij driehoek $S_1S_2S_3$ *gelijkzijdig*: $S = \frac{9}{4}$, $P = \frac{27}{64}$.
Dan is ook een basisdriehoek gelijkzijdig, d.w.z. 2.11 heeft een wortel $s = 2\frac{1}{4}$. Verder is er, meetkundig gezien, driezijdige symmetrie. Deze leidt tot drievoudige wortels. Inderdaad kan 2.11 (met $S = \frac{9}{4}$, $P = \frac{27}{64}$) worden teruggebracht tot

$$(16s^2 - 40s + 13)^3(4s - 9) = 0 \ . \qquad 3.1$$

Een der drievoudige wortels is $s = \frac{1}{4}(5 + 2\sqrt{3})$. Met 2.10 vindt men daarbij $p = (7 + 4\sqrt{3})/64$. Dit paar s,p bepaalt een derdegraadsvergelijking (2.3) voor $x_i = \sin^2 A_i$, $i = 1,2,3$:

$$64x^3 - (80 + 32\sqrt{3})x^2 + (44 + 24\sqrt{3})x - (7 + 4\sqrt{3}) = 0 \ .$$

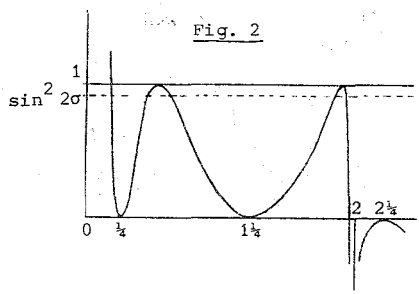
Kennelijk voldoet $x = \frac{1}{4}$. Verder is er een dubbele wortel $x = \frac{1}{4}(2 + \sqrt{3})$. De bijbehorende hoeken zijn 30°, 75°, 75°. Er zijn natuurlijk *drie* basisdriehoeken van deze vorm.
Op dezelfde manier leidt $s = \frac{1}{4}(5 - 2\sqrt{3})$ tot de hoeken 150°, 30°, 30° van weer drie basisdriehoeken. Het is niet moeilijk de figuur van een gelijkzijdige driehoek $S_1S_2S_3$ met basis-zevental te construeren.

Nu het geval van *rechthoekige* $S_1S_2S_3$: $S = 2$, $P = \frac{1}{4}\sin^2 2\sigma$. Hierbij kunnen we de 7^e-graadsvergelijking 2.11 een handiger vorm geven:

$$\frac{1}{4}(4s - 1)^2(4s - 5)^2(4s - 9)^2(2 - s)/(16s^2 - 32s - 1)^2 = \sin^2 2\sigma \ . \qquad 3.2$$

De grafiek van het linkerlid (Fig. 2) blijkt dubbel te raken aan y = 1 (bij

Fig. 2



$s = 1\frac{1}{4} \pm \sqrt{\frac{1}{2}}$). Daardoor is direct te zien dat er voor $0^\circ < \sigma < 45^\circ$ altijd *vijf* verschillende reële s voldoen. In het algemeen, d.w.z. bij willekeurige σ, zijn zij niet met vierkantswortels te bepalen. Dat betekent: geen passer-en-lineaal constructie voor de basisdriehoeken. Wèl kan men rekenenderwijs tot een schets komen.

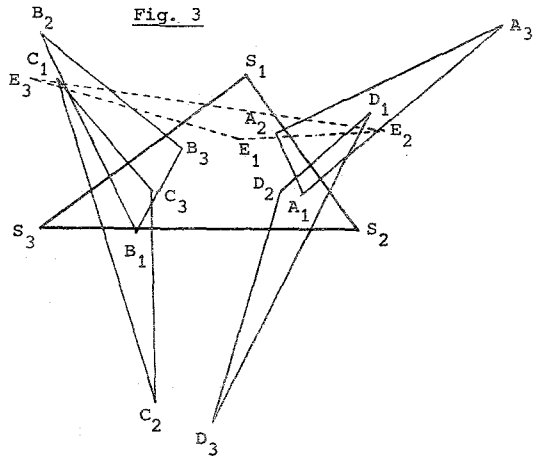
Voorbeeld: $\sigma = \arcsin 0{,}6$ (Fig.3, de bekende $5^2 = 4^2 + 3^2$ situatie). De vijf wortels s zijn numeriek benaderd. Elk bepaalt (met zijn p en met $x^3 - sx^2 + (\frac{1}{4}s^2 + p)x - p = 0$) $\sin^2 A_i$, $\sin^2 B_i$, enz. De uitkomsten voor A_i, B_i, enz. (in graden, decimale voorstelling) volgen hieronder.

s		1	2	3
1,97754	A	73,32	92,31	14,37
1,91494	B	59,74	24,91	95,34
0,63697	C	23,18	16,26	140,56
0,46348	D	21,97	147,43	10,60
0,18535	E	159,64	11,94	8,42

3.3

N.b. de groepering: A en B, C en D, E. Verklaring hiervan volgt in 4.

Fig. 3



Tenslotte zij $S_1S_2S_3$ een "*lineaire driehoek*" en wel met S_1 tussen S_2 en S_3 (zijde $s_1 = s_2 + s_3$, $\sigma_1 = 180^\circ$, $\sigma_2 = \sigma_3 = 0^\circ$). $S = 0$, $P = 0$. Vanzelfsprekend heeft een lineaire $S_1S_2S_3$ zichzelf als basisdriehoek. Deze trivialiteit komt ook te voorschijn als we in de 7^e-graadsvergelijking 2.11 invullen $S = 0$ (en dus $P = 0$). Wortels zijn dan: $s = 0$ (het triviale geval) en zesvoudig $s = 1\frac{1}{4}$. Bij de laatste waarde dienen we nog p te berekenen, maar 2.10 ontaardt in $0 + 16p \cdot 0 = 0$.

De substitutie $s = 1\frac{1}{4} + t\sqrt{S}$ geeft aan 2.10 en 2.11 een vorm, waarbij onessentiële effecten van $S \to 0$ verdwijnen (in de plaats van P komt P/S^3):

$$8t^3\sqrt{S} + 10t^2 - 2t\sqrt{S} - 3 + 64p(2t^2 - 2t\sqrt{S} + 1) = 0 , \qquad 3.4$$

$$(8t^3\sqrt{S} + 10t^2 - 2t\sqrt{S} - 3)(2t^2 - 2t\sqrt{S} + 1)t^2 + 4Q(2t^2S + t\sqrt{S} - 2)^2 = 0 , \qquad 3.5$$

waarin $Q := s_1^2 s_2^2 s_3^2/(s_1^2 + s_2^2 + s_3^2)^3$ $\quad (= P/S^3$ als $S \neq 0)$.

Samen bepalen 3.4 en 3.5 welke p bij gegeven S en Q horen. Aan de orde is S = 0. In dat geval kan t makkelijk worden geëlimineerd. Dat geeft voor p:

$$2\,\frac{64p(3-64p)}{(5+64p)^3} = Q\ . \qquad 3.6$$

Het linkerlid heeft kennelijk (bij 64p = 1) een maximum: $\frac{1}{54}$. Dit is de Q-waarde behorend bij $s_2/s_1 = \frac{1}{2}$. Bij elke andere verhouding is Q kleiner. Men vindt dan met 3.6 twee positieve waarden voor p.

Voorbeeld: $s_2/s_1 = \frac{1}{3}$. Dit geeft $Q = \frac{1}{6}(\frac{3}{7})^3$ en vervolgens de onderstaande uitkomsten. Als merkwaardigheid zijn ook de complexe vermeld (zoals steeds in graden). De bijbehorende p is < 0 en $x(x-\frac{5}{8})^2 + p(x-1) = 0$ heeft wortels $x_i = \sin^2 C_i$ buiten (0,1), namelijk $x_1 < 0$ en $x_2, x_3 > 1$.

s	64p		1	2	3
1¼	1,86165	A	40,84	101,37	17,78
1½	0,39566	B	48,13	7,36	124,51
1¾	-169,70176	C	58,85i	90 - 16,75i	90 - 42,10i

3.7

In de volledige figuur van $S_1S_2S_3$ met alle basisdriehoeken moet men beide reële gevallen dubbel tellen, omdat spiegeling in S_2S_3 mogelijk is. Werkten we met gerichte hoeken (draaizin), dan zou dit een kwestie van tekenwisseling zijn. Ook het imaginaire geval vertegenwoordigt twee tegengestelde basisdriehoeken. In totaal zijn er dus zeven ($A_i = S_i$ meegerekend).

Werken met gerichte hoeken blijkt straks verhelderend (zie 5.4 en verder). Nu al kunnen we in Fig. 3 zien dat al naar s > 1¼ dan wel < 1¼ de "omloopszin" van $S_1S_2S_3$ gelijk dan wel tegengesteld is aan die van $A_1A_2A_3$ enz. Dit tekeneffect gaat verloren in 2.7 (een kwadratische betrekking).

Het voorbeeld $s_2/s_1 = \frac{1}{3}$ is van zodanige eenvoud dat men basisdriehoeken verwacht die construeerbaar zijn. Toch zijn ze dat niet. Dit blijkt door 3.6 (waarin $Q = \frac{1}{6}(\frac{3}{7})^3$) met $5 + 64p = \frac{y}{9}$ te herleiden tot

$$4 \cdot 7^3 \cdot (y - 45)(72 - y) = y^3 \ . \qquad 3.8$$

Geen der oplossingen hiervan is geheel (zie 64p in 3.7). Gezien de coëfficiënten (geheel, 1 bij y^3) is er dus geen oplossing met vierkantswortels.

4. Alvorens over te gaan tot *gelijkbenige* driehoeken $S_1S_2S_3$, met hun in 1 aangekondigde merkwaardigheden, eerst nog de vraag: wat is een gelijkbenige driehoek? We zullen het D = 0 kenmerk hanteren. Het lijkt weliswaar evident dat gelijkbenige driehoeken $A_1A_2A_3$ dito $S_1S_2S_3$ geven. Maar het kenmerk is scherper: als $D(s,p) = 0$, dan kan $D(S,P) \neq 0$ zijn. In 2.12 heeft $D(S,P)$ namelijk een factor naast zich. Stel eens dat deze 0 is:

$$4s + 1 + 64p(4s - 7) = 0 \ . \qquad 4.1$$

Hieraan en aan $D(s,p) = 0$ blijkt uitsluitend het paar $s = 1\frac{1}{4}$, $p = \frac{3}{64}$ te voldoen. Het geeft voor A_i: 120°, 30°, 30°. Bij een dergelijke basisdriehoek vind men twee samenvallende punten S_i. Wat $S_1S_2S_3$ betreft, gaat het dus om randgevallen met één hoek 0° en een supplementair paar σ, $180^\circ - \sigma$, zodat $P = 0$, $S = 2\sin^2\sigma$. Zo'n randgeval is "ongelijkbenig" als $\sigma \neq 180^\circ - \sigma$; dan is namelijk $D \neq 0$. Alleen bij $\sigma = 90^\circ$ vindt men $D(2,0) = 0$.

Is $\sigma = 0^\circ$ (dus twee hoeken 0° en $S = 0$), dan is $S_1S_2S_3$ lineair. Hierbij spreken we van gelijkbenigheid als $s_2/s_1 = \frac{1}{2}$, dus als S_1 het midden is van

S_2S_3. In 3.6 komt dan $Q = \frac{1}{54}$ en 64p = 1 (dubbel).

s	64p		1	2	3	
1¼	1	A	45	122,85	12,15	4.2
1¼	-125	C	55,14i	90 - 27,57i	90 - 27,57i	

De eerste regel (n.b. $A_1 = 45^\circ$) vertegenwoordigt *vier* niet-gelijkbenige basisdriehoeken (Fig. 4 geeft er één). Zij zijn gespiegeld gelegen (t.o.v. s_1 en S_1) en construeerbaar ($\cot A_2 = 2 - \sqrt{7}$, $\cot A_3 = 2 + \sqrt{7}$; zie 5.8). Wèl gelijkbenig zijn de (twee) imaginaire basisdriehoeken en $S_1S_2S_3$ zelf.

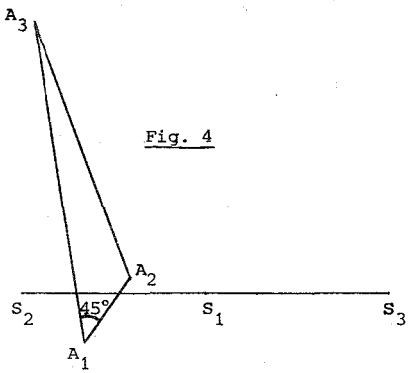


Fig. 4

Gelijkbenigheid van $S_1S_2S_3$ heeft, algebraïsch gezien, een ingrijpend effect. Dit wordt duidelijk als men driehoek $S_1S_2S_3$ van Fig. 3 in gedachten vervormt tot een rechthoekige *gelijkbenige* ($\sigma = 45^\circ$). Wat al min of meer symmetrisch is, wordt dan exact spiegelbeeldig: basisdriehoek A met B, C met D en E met zichzelf. Voor de 7^e-graadsvergelijking in s (2.11 met S = 2, P = ¼) betekent dit, dat er *twee dubbele* wortels moeten zijn. Inderdaad blijkt deze vergelijking te herleiden tot

$$(16s^2 - 40s + 17)^2(16s^3 - 72s^2 + 89s - 14) = 0 \quad . \qquad 4.3$$

Reële wortels: $s = 1\frac{1}{4} + \sqrt{\frac{1}{2}}$ (beide dubbel), $s = 1\frac{1}{4} - \frac{1}{4}(23 - \sqrt{23^2 - 19^3/27})^{1/3} - \frac{1}{4}(23 + \sqrt{23^2 - 19^3/27})^{1/3} = 0{,}18341$. Samengevat (vgl. 3.3)

s		1	2	3
$1\frac{1}{4} + \sqrt{\frac{1}{2}}$	A,B	$67\frac{1}{2}$	93,40	19,10
$1\frac{1}{4} - \sqrt{\frac{1}{2}}$	C,D	$22\frac{1}{2}$	144,14	13,36
0,18341	E	159,65	10,18	10,18

4.4

Een gelijkbenige $S_1S_2S_3$ met rechte tophoek heeft blijkbaar twee paar ongelijkbenige maar wel construeerbare basisdriehoeken (n.b. de hoeken $67\frac{1}{2}^\circ$ en $22\frac{1}{2}^\circ$) en één niet-construeerbare maar wel gelijkbenige.

5. Het voorgaande toont de algemene situatie bij gelijkbenige $S_1S_2S_3$. We nemen weer S_1 als tophoek en $S_2 = S_3 = \sigma$. Gemakkelijk volgt

$$S = (6 - 4\sin^2\sigma)\sin^2\sigma \ , \quad P = (4 - 4\sin^2\sigma)\sin^6\sigma \ . \tag{5.1}$$

Deze uitdrukkingen maken D(S,P) identiek nul. Anders gezegd: 5.1 geeft een parametrisering van de kromme D(S,P) = 0. Deze loopt van (0,0) via het keerpunt $(\frac{9}{4}, \frac{27}{64})$ naar (2,0) (Fig. 6, n.b. de S-schaal volgt S^3).

Nu geldt D(S,P) = 0 volgens 2.12 als D(s,p) = 0 ($A_1A_2A_3$ gelijkbenig) of als $s = 1\frac{1}{4}$ ($S_1S_2S_3$ lineair of randgeval), maar óók als

$$4s - 1 + 64p(4s - 9) = 0 \ . \tag{5.2}$$

Basisdriehoeken die hieraan voldoen leveren dus gelijkbenige $S_1S_2S_3$ maar zijn zelf in het algemeen niet gelijkbenig. Wèl zijn ze - bij gegeven gelijkbenige $S_1S_2S_3$ - altijd construeerbaar, zoals nu zal blijken.

Door p uit 5.2 op te lossen en het gevonden quotiënt te substitueren in 2.8 en 2.9 vindt men S en P als functie van s alléén:

$$S = (6 - 4(s - 1\tfrac{1}{4})^2)(s - 1\tfrac{1}{4})^2 , \qquad P = (4 - 4(s - 1\tfrac{1}{4})^2)(s - 1\tfrac{1}{4})^6 . \qquad 5.3$$

Samen met 5.1 geeft dit $(s - 1\frac{1}{4})^2 = \sin^2\sigma$, dus $s = 1\frac{1}{4} \pm \sin\sigma$, $0^\circ \leq \sigma < 90^\circ$. Deze uitkomst wordt beter hanteerbaar door basishoeken $\sigma < 0^\circ$ toe te laten:

$$s = 1\tfrac{1}{4} + \sin\sigma , \quad -90^\circ < \sigma < 90^\circ . \qquad 5.4$$

De bijbehorende p is volgens 2.10 (met S als in 5.1)

$$p = \frac{1}{64}\,\frac{1+\sin\sigma}{1-\sin\sigma} . \qquad 5.5$$

Het paar s,p geeft als vergelijking voor $x_i = \sin^2 A_i$ (vgl. 2.3)

$$x(x - \frac{5}{8} - \frac{1}{2}\sin\sigma)^2 + \frac{1}{64}\,\frac{1+\sin\sigma}{1-\sin\sigma}(x - 1) = 0 . \qquad 5.6$$

Hieraan voldoet $x_1 = \frac{1}{2} + \frac{1}{2}\sin\sigma$. Beide andere oplossingen - mits reëel - moeten op (0,1) liggen. Zij zijn met vierkantswortels te bepalen. Daarmee is bewezen, dat basisdriehoeken, die aan 5.2 voldoen, construeerbaar zijn.

Nader blijkt dat A_1 en S_1 direct samenhangen: $\cos 2A_1 = 1 - 2\sin^2 A_1 = 1 - 2x_1 = -\sin\sigma = \cos(90^\circ + \sigma) = \cos(180^\circ - \frac{1}{2}S_1)$, dus

$$A_1 = 90^\circ - \tfrac{1}{4}S_1 . \qquad 5.7$$

Voorbeelden hiervan (met $S_1 = 180^\circ, 90^\circ, 270^\circ$) zagen we al in 4.2 en 4.4. Voor A_2 en A_3 volgt na enige herleiding

$$\cot A_i = 2\cos\sigma \pm 2\sqrt{2 - (\sin\sigma + \tfrac{1}{2})^2} , \quad i = 2,3 . \qquad 5.8$$

Bij $\sigma > \arcsin(\sqrt{2} - \frac{1}{2}) = 66{,}09^\circ$ zijn A_2 en A_3 imaginair.

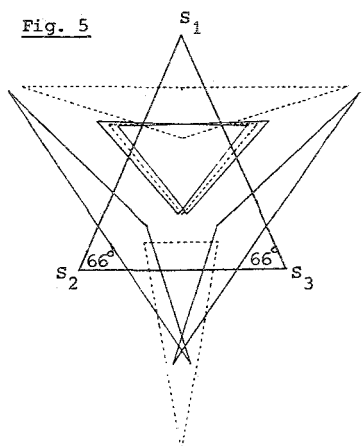
Het volgende numerieke overzicht toont de algemene samenhang.

5.9

Gelijkbenige $S_1S_2S_3$		Construeerbare $A_1A_2A_3$		
$\sigma = S_2 = S_3$	$S_1 = 180^\circ - 2\sigma$	$A_1 = 90^\circ - \frac{1}{4}S_1$	A_2	A_3
66,09	47,81	78,05	50,98	50,98
66	48	78	54	48
60	60	75	75	30
45	90	67½	93,40	19,10
0	180	45	122,85	12,15
-45	270	22½	144,14	13,36
-60	300	15	150	15
-78	336	6	156	18

Bij $\sigma < 0^\circ$ wordt de omloopszin van $S_1S_2S_3$ tegengesteld aan die van $A_1A_2A_3$. Merkwaardig zijn de "3n-gradige" gevallen $\sigma = 66^\circ$ (Fig. 5) en $\sigma = -78^\circ$. Gemakkelijk kan men nagaan dat ook bij σ resp. 54°, 42°, 30°, 6° en -18° 3n-gradige $A_1A_2A_3$ worden gevonden (zie [2]).

Fig. 5



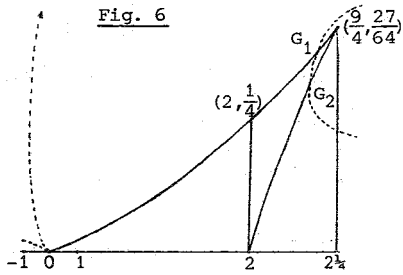
Intussen zijn er bij gelijkbenige $S_1S_2S_3$ behalve construeerbare ook nog gelijkbenige basisdriehoeken ($D(s,p) = 0$). Hun s-waarden voldoen aan een derdegraadsvergelijking, die gevonden wordt door uit de zevendegraads 2.11 twee keer $(s - 1\frac{1}{4})^2 - \sin^2\sigma$ weg te delen:

$$(s - 1\tfrac{1}{4})^2(s - S) - (s - 2\tfrac{1}{4})\sin^2 2\sigma = 0 , \quad S = (6 - 4\sin^2\sigma)\sin^2\sigma . \qquad 5.10$$

Is $S_1S_2S_3$ gelijkzijdig ($S = 2\frac{1}{4}$), dan zijn er drie oplossingen (zie 3.1). Maar het eenvoudig lijkende geval $\sigma = 45^\circ$, $S = 2$ (zie 4.3) gaf slechts één (niet-construeerbare) oplossing. Het tussenliggende grensgeval (met dubbele wortel σ) is te vinden door de discriminant van 5.10 nul te stellen. Dat geeft $\sin^2\sigma = (29 - 6\sqrt{6})/20$, $\sigma = 57{,}74^\circ$. De dubbele wortel, $s = \sqrt{6} - \frac{1}{4}$, bepaalt een basisdriehoek ($\sin^2 A_2 = \sin^2 A_3 = \frac{1}{4} + \frac{1}{4}\sqrt{6}$), die ook dubbel mag heten. Dit kan als volgt worden toegelicht. Beschouw gelijkbenige driehoeken $S_1S_2S_3$ met vaste basis S_2S_3 waarbij $\sigma \downarrow 57{,}74^\circ$. Dan naderen twee gelijkbenige basisdriehoeken tot elkaar, om bij $\sigma = 57{,}74^\circ$ samen te vallen en voor kleinere σ imaginair te worden. Iets dergelijks gebeurt als $\sigma \downarrow 66{,}09^\circ$ (Fig. 5 toont de situatie bij $\sigma = 66^\circ$). Twee ongelijkbenige (gespiegelde) basisdriehoeken naderen dan tot elkaar en dus (n.b.) ook tot een gelijkbenige. Bij $\sigma = 66{,}09^\circ$ vallen zij daarmee samen, om voor $\sigma > 66{,}09^\circ$ te verdwijnen (de gelijkbenige blijft). Een tweede grensgeval dus, nu met $\sigma > 60^\circ$.

6. We keren van gelijkbenige $S_1S_2S_3$ terug naar willekeurige.
Duidelijk is dat er bij geringe afwijking van gelijkzijdigheid 7 basisdriehoeken zijn. Maar bij welke afwijkingen loopt de grens? Bekijken we daartoe Fig. 6.

Fig. 6



Rechthoekige, scherp- en stomphoekige $S_1S_2S_3$ horen bij punten met $S = 2$, $S > 2$, $S < 2$, omgrensd door $D(S,P) = 0$. Dichtbij het keerpunt $(\frac{9}{4},\frac{27}{64})$ liggen, *op* de kromme, de punten G_1 en G_2 behorend bij de *gelijkbenige* grensgevallen ($\sigma = 57{,}74^\circ$ resp. $66{,}09^\circ$). Er moet een kromme G_1G_2 zijn, die de top der $S_1S_2S_3$ met zeven basisdriehoeken afgrenst. Bij de overige $S_1S_2S_3$ vermoeden we er vijf, soms zes. Hoe dit na te gaan en te preciseren?

Een zekere weg is de punten (S,P) te bepalen die dubbele wortels s geven in de 7^e-graadsvergelijking 2.11. Daartoe hebben we de discriminant nodig. Deze blijkt herleidbaar tot een 5×5-determinant (van Bezout), waarin elk element een veelterm in S en P is. Met de hand uitwerken is ondoenlijk. Met de computer*) verschijnt er een veelterm van de 21^e graad in S en P met coëfficiënten $> 10^{30}$! Door - weer met computerhulp - bekende factoren (o.a. $D(S,P)$ 2 keer) weg te delen vindt men als uiteindelijke, niet verder terug te brengen conditie een derdegraadsvergelijking in P:

---

*) De auteur dankt Dr.ir. J.K.M. Jansen voor desbetreffende hulp.

$$499392P^3 + (-73984S^3 + 328080S^2 - 3720000S - 5000000)P^2 + \\ + (-21888S^5 + 562500S^4 - 450000S^3)P + 432S^7 - 10125S^6 = 0 . \qquad 6.1$$

De bijbehorende kromme laat zich in Fig. 6 links en rechts even zien (stippellijnen). Voor $0 < S < 2{,}17$ is er maar één reële wortel P (> 10, vèr boven de figuur). De lus rechts, die de top van $D(S,P) \geq 0$ afsnijdt, is de grenskromme! Links is (0,0) een keerpunt. Dit representeert alle lineaire driehoeken $S_1S_2S_3$. Bij hen vonden we steeds vijf reële basisdriehoeken.

Afgezien van de "top" met zeven basisdriehoeken blijken vrijwel alle driehoeken $S_1S_2S_3$ er vijf te hebben. Ook bij het gelijkbenige grensgeval met $\sigma = 66{,}09^\circ$ vonden we er vijf. De overige punten van de grenskromme representeren $S_1S_2S_3$ met zes basisdriehoeken. Gemakkelijk vindt men numerieke voorbeelden: neem S tussen $S(G_1)$ en $S(G_2)$ ($1{,}68\sqrt{6} - 1{,}87$ resp. $12\sqrt{2} - 14\frac{3}{4}$) en bepaal met 6.1 de bijbehorende P (middelste wortel). Aldus leidt $S = \frac{1}{2}S(G_1) + \frac{1}{2}S(G_2)$ tot een driehoek $S_1S_2S_3$ met hoeken $50{,}92^\circ$, $63{,}92^\circ$, $65{,}16^\circ$. Dat deze grensdriehoeken zich door een meer directe, eenvoudige eigenschap zouden onderscheiden is, gezien de gedaante van 6.1, niet te verwachten.

Rest nog een enkel woord over de construeerbaarheid van de basisdriehoeken. Duidelijk is dat er bij willekeurige $S_1S_2S_3$ geen passer-en-lineaal constructie mogelijk is. Ook niet als men $S_1S_2S_3$ vindt door uit te gaan van een willekeurige basisdriehoek $A_1A_2A_3$. Neemt men echter $A_1A_2A_3$ gelijkbenig, dan zijn *alle* overige (reële) basisdriehoeken van $S_1S_2S_3$ construeerbaar. Voor de ongelijkbenige spreekt dat vanzelf. Voor de gelijkbenige bedenke men, dat van de betreffende derdegraadsvergelijking al een wortel s bekend is, namelijk die van de gegeven driehoek zelf.

LITERATUUR

[1] Bottema, O., De constructie van een driehoek als de spiegelpunten van de hoekpunten in de overstaande zijden gegeven zijn, Nieuw Tijdschrift voor Wiskunde 24 (1936/37), 248 - 251.

[2] IJzeren, J. v., Spiegelpuntsdriehoeken, Nieuw Tijdschrift voor Wiskunde 71 (1983/84), 95 - 106.