Printed at the Mathematical Centre, 49, 2e Boerhaavestraat, Amsterdam.

The Mathematical Centre, founded the 11-th of February 1946, is a non-profit institution aiming at the promotion of pure mathematics and its applications. It is sponsored by the Netherlands Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O).

MATHEMATICAL CENTRE TRACTS 101

PROCEEDINGS BICENTENNIAL CONGRESS WISKUNDIG GENOOTSCHAP

PART II

Edited by

P.C. BAAYEN

D. VAN DULST

J. OOSTERHOFF

MATHEMATISCH CENTRUM AMSTERDAM 1979

ISBN 90 6196 169 6

C O N T E N T S

HORDIJK, A.	*From linear to dynamic programming via shortest paths*	213
KAASHOEK, M.A.	*Recent developments in the spectral analysis of matrix and operator polynomials*	233
LENSTRA, Jr., H.W.	*Vanishing sums of roots of unity*	249
LOOIJENGA, E.J.N.	*On quartic surfaces in projective 3-space*	269
NEUENSCHWANDER, E.	*Augustin Cauchy: ein Wendepunkt in der Geschichte des Analysis*	275
PELETIER, L.A.	*The mathematical theory of clines*	295
RUNNENBURG, J.Th.	*Van Dantzig's collective remarks revisited*	309
SLUIS, A. van der	*Computation and stability of solutions of linear least squares problems*	331
SPIJKER, M.N.	*Error bounds in the numerical solution of initial value problems*	345
STIGT, W.P. van	*L.E.J. Brouwer: Intuitionism and topology*	359
TAKENS, F.	*Symmetries, conservation laws and symplectic structures; elementary systems*	375
THOMAS, E.	*Integral representations of invariant reproducing kernels*	391
TIJDEMAN, R.	*Distance sets of sequences of integers*	405
WIJNGAARDEN, A. van	*Thinking on two levels*	417
	List of addresses of authors	429

FROM LINEAR TO DYNAMIC PROGRAMMING VIA SHORTEST PATHS

A. Hordijk

1. INTRODUCTION AND SUMMARY

Linear programming and dynamic programming are two important branches of mathematical programming. In this paper we study some of the relations between the both. Especially, we will approach dynamic programming problems by converting them to linear programming problems. The intermediate is a shortest-paths problem. Shortest-paths problems are basic problems in combinatorial optimization which is another main branch of mathematical programming. Hence we try in this paper to say something about three main streams of investigation in the mathematics of operations research.

The summary of the contents we will give now is for readers already familiar with the topics of this paper. Readers who want to use this paper as an introduction might better skip the rest of this section in first reading.

In section 2 a short introduction is given to linear programming. Our approach is a geometrical one with emphasis on the theory of duality. The theory of linear programming with a finite number of variables and a finite number of constraints is well established nowadays. Section 2 is almost entirely taken from the literature, our main sources can be found in the list of references.

In our description of the simplex method we used that the active dual variables in vertex x are essentially the projections of the objective function on the extreme rays of the dual cone in x. This way of explaining the simplex method seems to be new.

In section 3 the shortest-paths problem is studied. It is explained how the problem can be solved by using the simplex method for a linear programming problem of network-flow type. An algorithmic procedure, derived from the simplex method, for solving shortest-paths problems is given, together with a tight upper bound on the number of "pivot steps" for a spe-

cialization of this procedure. The relation with the well-known relaxation procedure is made. A specialization of this procedure is essentially the nonlinear extension of Gausz-Seidel iteration. It seems that with the exception of the conversion of the shortest-paths problem to a linear programming problem the material of this section is new.

In section 4 the finite horizon dynamic programming problem is introduced. It is shown that the deterministic problem is a shortest-paths problem in an acyclic graph. Moreover, the stochastic dynamic programming problem can be seen as finding a shortest *stochastic* spanning tree. The well-known optimality principle of Bellman together with the validity of the backward recursion of dynamic programming are shown to be consequences of the duality theory of linear programming.

The relation between "existence of *pure* optimal policies" and "*integrality* of basic solutions of the corresponding linear programming problem" is made. Also these two properties can be seen as consequences of the duality theory.

It is well-known that discrete dynamic programming problems can be formulated as linear programming problems. Also, dynamic programming is often introduced via a shortest-paths problem. However, a systematic investigation does not seem to have been published before. The fifteenpage limit of this paper did not allow us to analyse here dynamic programming problems with an infinite horizon. Also for these models, often called Markov decision chains, well-known properties appear to be consequences of duality theory of linear programming. All these results seem to suggest that the right title of this paper is: "dynamic programming is linear programming". However, this title would disregard the wealth of results in dynamic programming which go far beyond the boundaries of linear programming.

2. GEOMETRY OF LINEAR PROGRAMMING

The general problem of linear programming is to find a maximal value of a linear function, the object function, in a convex region defined by linear inequalities, i.e.

$$\max\{p^T x \mid Ax \leq b,\ x \geq 0\}, \tag{2.1}$$

where $p \in \mathbb{R}^n$, $x \in \mathbb{R}^n$, $A \in \mathbb{R}^m \times \mathbb{R}^n$, $b \in \mathbb{R}^m$ and $x \geq 0$ requires that all components of the vector x are nonnegative.

Any element of $R = \{x \mid Ax \leq b, x \geq 0\}$ is called a feasible solution, point or vector. The feasible solution x^* is optimal if $p^T x^* \geq p^T x$ for any $x \in R$. The set R is a convex polyhedron in $\mathbb{R}^n$.

For any $A \in \mathbb{R}^m \times \mathbb{R}^n$ we have a convex polyhedral cone $C = \{x \mid Ax \leq 0\}$. With C we can associate the so-called dual cone: $C' = \{y \mid y = A^T u, u \geq 0\}$. C' is also a convex polyhedral cone. The outward pointing normals of the faces $a_{i\cdot}^T x = 0$ of C are the extreme rays of C' and conversily. The dual cone of the dual cone is the original cone.

A fundamental theorem in the theory of linear programming is Farkas' theorem (1902): *The vector* $p \in \mathbb{R}^n$ *makes a non-acute angle with any vector of a convex polyhedral cone if and only if* p *belongs to the dual cone,* i.e.,

$$(2.2) \qquad \forall x(Ax \leq 0 \Rightarrow p^T x \leq 0) \Leftrightarrow \exists u(p = A^T u, u \geq 0).$$

A direction s is called feasible in the feasible point x if $x + \lambda s$ is feasible for some $\lambda > 0$. The direction s is usable if in addition $p^T s > 0$. It is clear that *feasible point* x *is optimal if and only if there exists no usable direction in* x. From Farkas' theorem we conclude: *vertex* x *of* R *is optimal if and only if* p *belongs to the dual cone of the cone of feasible directions in* x.

For any feasible solution we define the vector of slack variables y by $y = b - Ax$. It is clear that $y \in \mathbb{R}^m$ and $y \geq 0$. We denote the $m \times (n+m)$-matrix (A,I) by $\bar{A}$ and the $(n+m)$-vector $(x,y)^T$, where $y = b - Ax$, by $\bar{x}$. The vector x is feasible if and only if $\bar{A}\bar{x} = b$ and $\bar{x} \geq 0$. Write $M(x) = \{i \mid a_{i\cdot}^T x = b_i\} = \{i \mid y_i = 0\}$ and $N(x) = \{j \mid x_j = 0\}$.

The feasible solution x is an extreme point or vertex of R if and only if the positive components of $\bar{x}$ correspond to columns of $\bar{A}$ which are a set of linear independent vectors in $\mathbb{R}^m$, i.e., $\{a_{\cdot j}, j \notin N(x), e_i, i \notin M(x)\}$ are linear independent. The vertex x is called nondegenerate if $\bar{x}$ has exactly m positive components. In this case, x is determined as the intersection of the n linearly independent hyperplanes $\{x_j = 0, j \in N(x), y_i = 0, i \in M(x)\}$. The variables x_j, $j \notin N(x)$, y_i, $i \notin M(x)$ are called the basic variables, hence the hyperplanes are found by equating the non basic variables to zero. The cone of feasible directions in x is: $\{s \mid s_j \geq 0, j \in N(x), a_{i\cdot}^T s \leq b_i, i \in M(x)\}$.

To check whether vertex x is optimal we have to find out whether the object vector belongs to the dual cone of the cone of feasible directions. In practice, this is done by computing the projections of p on the outward

pointing normals of the faces of the primal cone. Let v_j for $j \in N(x)$ be the projection of p on $-e_j$ and let $u_i \cdot \|a_{i\cdot}\|^{-1}$ for $i \in M(x)$ be the projection of p on $a_{i\cdot}$. If we take $v_j = 0$ for $j \notin N(x)$ and $u_i = 0$ for $i \notin M(x)$, then,

$$p = A^T u - v \tag{2.3}$$

and

$$u^T y = 0, \; v^T x = 0. \tag{2.4}$$

The components of the vectors $u \in \mathbb{R}^m$ and $v \in \mathbb{R}^n$ are called the dual variables corresponding to x. The variables v_j, $j \in N(x)$ and u_i, $i \in M(x)$ are the active ones. They are essentially the projections of p on the extreme rays of the dual cone in x. From Farkas' theorem we conclude, vertex x is optimal if and only if the corresponding dual variables are nonnegative, i.e.,

$$u \geq 0, \; v \geq 0. \tag{2.5}$$

The relations (2.3), (2.4) and (2.5) together are called the *optimality conditions*.

If vertex x is not optimal then at least one of the dual variables is negative. If $u_i < 0$ then p makes an acute angle with the extreme ray of the cone of feasible directions, found by relaxing $y_i = 0$. This extreme ray is a usable direction. Similarly, if $v_j < 0$ then making x_j positive will increase the object function

The simplex method, due to G.B. Dantzig, is a class of algorithms. All of which have as main subroutine: *given vertex* x, *check the dual variables whether* x *is optimal, if not, go to an adjacent vertex via an edge of the polyhedral set* R *by making positive a primal variable corresponding to a negative dual variable, check the dual variables of the new vertex, etc.*

The simplex method is made to an algorithm if it is specified how the choice is made on which variable will be made positive. In practice, one usually takes the primal variable corresponding to the most negative dual variable.

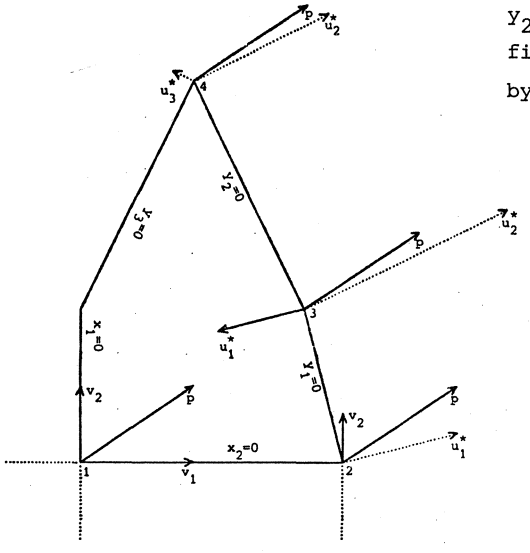
The actual computation of the new primal and dual variables will not be given here. It is enough to state that the computation is called a piv-

ot operation and that the pivoting is very similar to that of the familiar Gausz-elimination procedure to solve a set of linear equations. In this paper we will deal with linear programming problems of a special structure. For these problems the primal and dual variables are computed differently.

EXAMPLE. $\max\{3x_1+2x_2 \mid 4x_1+x_2 \le 28,\ 2x_1+x_2 \le 16,\ -2x_1+x_2 \le 4,\ x_1 \ge 0,\ x_2 \ge 0\}$.

In this example we have the variables x_1, x_2 with corresponding dual variables v_1, v_2, the variables y_1, y_2, y_3 are the slack variables with corresponding dual variables u_1, u_2, u_3. As initial vertex point we take $x_1 = 0$, $x_2 = 0$, in this point both active dual variables are negative. The choice of variable to become positive, mostly called choice of pivot column, is x_1, the adjacent vertex is then $x_2 = 0$, $y_1 = 0$ etc.

vertex	basic variables	cone of feasible directions	active dual variables
1	y_1, y_2, y_3	$x_1 \ge 0,\ x_2 \ge 0$	$v_1 < 0,\ v_2 < 0$
2	x_1, y_2, y_3	$y_1 \ge 0,\ x_2 \ge 0$	$u_1 \ge 0,\ v_2 < 0$
3	x_1, x_2, y_3	$y_1 \ge 0,\ y_2 \ge 0$	$u_1 < 0,\ u_2 \ge 0$
4	x_1, x_2, y_1	$y_3 \ge 0,\ y_2 \ge 0$	$u_3 \ge 0,\ u_2 \ge 0$.



The vertex 4, i.e. $y_3 = 0$, $y_2 = 0$ is optimal. In the figure we denote $u_i \cdot \|a_{i\cdot}\|^{-1}$ by u_i^*, $i \in M(x)$.

Any linear programming problem has an associated linear programming problem, the so-called dual problem. The dual of (2.1) is

$$\min\{b^T u \mid A^T u \geq p,\ u \geq 0\}, \tag{2.6}$$

with as vector of slack variables $v = A^T u - p$ and dual feasible region $R' = \{u \mid A^T u \geq p,\ u \geq 0\} \subset \mathbb{R}^m$. The dual problem of the dual problem is the primal problem. There is a close relationship between the primal and dual linear programming problem, as the following theorems show. The *weak duality theorem* says, *if* $(x,y)^T$ *is feasible for the primal problem and* $(u,v)^T$ *is feasible for the dual problem, then*

$$p^T x = b^T u - u^T y - v^T x \leq b^T u. \tag{2.7}$$

With $(x,y)^T$ feasible we mean x is a feasible point with y as vector of slack variables; similar for $(u,v)^T$.

The proof runs as follows,

$$p^T x = (u^T A - v^T)x = u^T A x - v^T x = u^T b - u^T y - v^T x \leq u^T b = b^T u,$$

since

$$x \geq 0,\ y \geq 0,\ u \geq 0 \text{ and } v \geq 0.$$

The optimality conditions show that if x is an optimal vertex of the primal problem then its corresponding dual variables say u, v are such that u is feasible for the dual problem and moreover v are its slack variables. From (2.4) and the above equality we conclude the *strong duality theorem, if the primal problem has an optimal solution then its corresponding dual variables are an optimal solution of the dual problem.* Suppose $(x,y)^T$ is a primal feasible and $(u,v)^T$ is dual feasible.

If $(x,y)^T$ and $(u,v)^T$ are orthogonal, i.e. (2.4) is satisfied then $(u,v)^T$ are nonnegative dual variables corresponding to $(x,y)^T$. Hence $(x,y)^T$ is primal optimal and $(u,v)^T$ is dual optimal. Conversily, if $(x,y)^T$ and $(u,v)^T$ are optimal then from (2.7) they are orthogonal. This result is known as the *orthogonality theorem.*

The primal problems of the next section will have equalities instead

of inequalities in the constraints, i.e. $a_i^T.x = b_i$, i = 1,2,...,m. By the standard trick of replacing any equality by two inequalities it can be shown that the dual variables u_i, i = 1,2,...,m become unrestricted in sign. Since y = 0 the relation (2.4) reduces to

(2.8) $\quad v^T x = 0.$

The relation (2.5) becomes

(2.9) $\quad v \geq 0.$

In order to keep this introduction short and to avoid complications we implicitly assumed that vertices were nondegenerate i.e. the number of positive components in $(x,y)^T$ is equal to the number of constraints. In section 3 all vertices are nondegenerate. In section 4 there are degenerate vertices. However, the specialization of the simplex method to the problem of section 4 is the backward recursion of dynamic programming. There are no complications, as cycling, possible with this algorithm.

3. SHORTEST PATHS

In this section we suppose to have a directed graph with $V = \{v_1,v_2,\ldots,v_n\}$ as set of nodes and A as set of arcs. Any arc $(v,w) \in A$ has assigned to it a real number l(v,w), its length. We make the graph complete by adding those arcs which are not in A giving them a length ∞. A path P from node v to node w is a sequence of arcs of the form (v,v_1), $(v_1,v_2),\ldots,(v_k,w)$, its length is $l(v,v_1) + l(v_1,v_2) +\ldots+ l(v_k,w)$. A very important problem, having many applications, is to find shortest paths from a designated node say v_1 to all other nodes.

In this section we make the assumption, *there are no cycles having negative length*. A cycle is a path from a node to itself. Suppose P_k is a path from v_1 to v_k, k = 2,3,...,n. Let x_{ij}^k denote the number of arcs in P_k which are equal to (v_i,v_j), then the following equalities hold

$$(3.1) \qquad \sum_j x_{1j}^k - \sum_j x_{j1}^k = 1,\ \sum_j x_{kj}^k - \sum_j x_{jk}^k = -1,\ \sum_j x_{ij}^k - \sum_j x_{ji}^k = 0,$$

$$i \neq 1 \text{ and } i \neq k.$$

The sum $x_{ij} = \sum_k x_{ij}^k$ satisfies

$$(3.2) \qquad \sum_j x_{1j} - \sum_j x_{j1} = n - 1, \quad \sum_j x_{ij} - \sum_j x_{ji} = -1, \quad i = 2,3,\ldots,n.$$

The total length of the paths P_k, $k = 2,3,\ldots,n$ is $\sum_{i,j} l_{ij} x_{ij}$, where l_{ij} is a short notation for $l(v_i,v_j)$. Nonnegative integers x_{ij}, $i,j = 1,2,\ldots,n$ are called a flow in our graph. If the x_{ij} satisfy the constraints (3.2) then the flow is called feasible. Any feasible flow is the sum of paths P_k from v_1 to v_k, $k = 2,3,\ldots,n$ and a number of cycles. The way to prove this is similar to the proof that an Euler graph has an Euler path, we leave it to the reader.

Since cycles with negative length do not exist, the total length of shortest paths from v_1 to all $v \neq v_1$ equals

$$\min \sum_{i,j} l_{ij} x_{ij}$$

over all feasible flows. This is almost a linear programming problem, if the condition that the x_{ij} have to be integer valued is relaxed then it would have been one.

The matrix of coefficients in the constraints is as follows,

	x_{11}	x_{12}	x_{13}	$\cdots$	$x_{1(n-1)}$	x_{1n}	x_{21}	x_{22}	x_{23}	$\cdots$	$x_{2(n-1)}$	x_{2n}	$\cdots$	x_{n1}	x_{n2}	x_{n3}	$\cdots$	$x_{n(n-1)}$	x_{nn}
y_1		1	1	$\cdots$	1	1	-1							-1					
y_2		-1					1		1	$\cdots$	1	1			-1				
y_3			-1						-1							-1			
$\vdots$				$\ddots$						$\ddots$							$\ddots$		
y_{n-1}					-1						-1							-1	
y_n						-1						-1	1	1		1	$\cdots$	1	

It is precisely the incidence matrix of the complete directed graph on n nodes. Any incidence matrix of a directed graph is totally unimodular which implies that any square submatrix has a determinant equal to 0 or ±1.

The equations (3.2) are dependent, any one can be written as a linear combination of the others. In order to make a set of independent equations we omit the first one. The extreme points or vertices of the feasible region are those feasible solutions for which the positive components correspond to a maximal set of linear independent column vectors of the coeffi-

cient matrix. The column of variable x_{ij} corresponds to arc(v_i,v_j). If the arcs of a set of columns contain a circuit then they are dependent. Hence, a maximal set of independent columns has (n-1) elements and corresponds to a subgraph having no circuits, which is a tree. However, not any tree allows a feasible flow. Therefore, it must contain paths from v_1 to any of the other nodes. Such a tree is called, directed spanning tree rooted from node v_1. There is a one-one-correspondence between the set of these trees and the vertices of the feasible region. A theorem of Cayley says, the number of undirected spanning trees in a complete graph with n nodes is n^{n-2}. Assigning to the edges the direction from v_1 we find that the set of all directed spanning trees rooted from node v_1 has n^{n-2} elements. Hence, there are that many vertices in our convex polyhedron of feasible points.

The n-1 positive components of a vertex point, also called a basic solution, are integer valued. This can be shown using Cramer's rule together with the totally unimodularity, or alternatively more directly by using that a basic solution corresponds to a directed spanning tree rooted from v_1. Indeed, the feasible flow associated with this tree has components which are integer valued.

A conclusion is that our problem of finding shortest paths can be solved by using a simplex algorithm (recall that these algorithms search for an optimal *basic* solution) for the linear programming problem

$$(3.3) \qquad \min\{\sum_{i,j} l_{ij}\, x_{ij} \mid \sum_{j} x_{ij} - \sum_{j} x_{ji} = -1,\ i = 2,3,\ldots,n,\ x \geq 0\}.$$

Another conclusion is that *there exists a spanning tree such that for all* $v \neq v_1$ *a shortest path from* v_1 *to* v *is the path of this tree*. Hence, if a shorthest path from v_1 to v leads through w then the subpath from v_1 to w is a shortest path from v_1 to w. This conclusion is a deterministic analogue of Bellamn's optimality principle, we come back to this point in the next section.

The dual of the problem (3.3) can be written as

$$(3.4) \qquad \max\{\sum_{j} u_j \mid u_1 = 0,\ u_j - u_i \leq l_{ij},\ i=1,2,\ldots,n,\ j=2,3,\ldots,n,\ i\neq j\}.$$

The slack variables of the dual problem are $v_{ij} = l_{ij} + u_i - u_j$. If x is a basic solution of the primal problem then its corresponding dual variables have to satisfy (2.8), i.e. the ortogonality relation $v^T x = 0$.

Hence,

$$x_{ij} > 0 \Rightarrow v_{ij} = 0$$

or

$$u_j = u_i + l_{ij}.$$

Since $x_{ij} > 0$ if and only if the arc (v_i,v_j) is an arc of the spanning tree corresponding to the basic solution, we conclude that u_j is exactly the length of the path from v_1 to v_j in this tree, $j = 2,3,\ldots,n$.

If basic solution x is optimal then in addition it holds that $v_{ij} \geq 0$, $\forall i \neq j$. Hence the distances in the corresponding tree satisfy,

$$u_j = \min_{j \neq i} (u_i + l_{ij}). \tag{3.5}$$

These equations from the deterministic analogue of Bellman's optimality equations of dynamic programming. We have derived these optimality equations here as a consequence of duality theory of linear programming.

What kind of algorithms does the simplex method give for the linear programming problem (3.3)? Suppose we start with the initial basic solution corresponding to the spanning tree with arcs (v_1,v_j), $j \neq 1$. The corresponding dual variables are $u_j = l_{1j}$, $j \neq 1$ and $v_{ij} = l_{ij} + u_i - u_j$, $i \neq j$. The routine of section 2 says, choose a negative dual slack variable and make the corresponding primal variable positive. Hence if we choose v_{ij} we add the arc (v_i,v_j) to the spanning tree. The new subgraph then has a circuit. We have to make one of the x_{ij}'s zero or in graphlanguage we have to delete one arc. The result has to be a new basic solution or directed spanning tree rooted from v_1. Hence we must delete the arc leading to v_j in the old tree. Having found the new tree, the new dual variables must be computed. Note that only the u's of successors of v_j in the new tree change their value.

In pseudo-algorithmic formulation the simplex method for shortest paths reads,

step 0 *(start, distances)*

$u_1 := 0$, $u_j := l_{1j}$, $j = 2,3,\ldots,n$

step 1 *(start, successors)*

$L_i := \{i\}$, $i = 2,3,\ldots,n$

step 2 *(test inequalities)*

if $u_j - u_i \leq l_{ij}$, $i = 1,2,\ldots,n$, $j = 2,3,\ldots,n$, $i \neq j$, *stop; the solution is optimal. Otherwise find* i, j *such that* $u_j - u_i > l_{ij}$.

step 3 *(change of successors)*

if $i \in L_k$ *then* $L_k := L_k \cup L_j$, *if* $i \notin L_k$ *and* $j \in L_k$ *then* $L_k := L_k \backslash L_j$, $k=2,3,\ldots,n, k\neq j$.

step 4 *(new dual variables)*

$$u_l := u_l - (u_j - u_i - l_{ij}) \text{ if } l \in L_j.$$

Return to step 2.

If the choice of i, j i.e. the dual variable v_{ij} in step 2 is specified then we obtain an algorithm.

If step 3 is omitted then the method becomes what is known as the relaxation procedure.

Let us specify the choice of $v_{ij} = l_{ij} + u_i - u_j$ in step 2;

most negative number of $(v_{1j}, v_{2j}, \ldots, v_{(j-1)j}, v_{(j+1)j}, \ldots, v_{nj})$

cyclically for $j = 2,3,\ldots,n,2,3,\ldots$

breaking ties arbitrarily if necessary.

If we pace the simplex algorithm and relaxation algorithm then the simplex u's will always be smaller than or equal to the relaxation u's.

The computation involved in the relaxation algorithm is essentially, Gausz-Seidel iteration.

Initially, set

$$u_1^1 = 0$$
$$u_j^1 = l_{1j}, \quad j \neq 1,$$

the (m+1)*st order approximations are computed from the* m*th order as follows*

$$u_j^{m+1} = \min\{\min_{i<j}(u_i^{m+1} + l_{ij}),\ u_j^m,\ \min_{i>j}(u_i^m + l_{ij})\},\ j = 2,3,\ldots,n.$$

We call this algorithm of Gausz-Seidel type since it is a straightforward nonlinear extension of the well-known iteration procedure to solve a set of linear equations.

Inductively, it can be shown that u_j^m is smaller than or equal to the shortest path from v_1 to v_j with at most m arcs.

In a spanning tree of shortest paths there are at least k nodes with paths from v_1 having at most k arcs. Consequently there are at most (n-1-k) indices j for which $u_j^{k+1} < u_j^k$. Hence, *the maximal number of improvements is* $\sum_k (n-1-k) = \frac{1}{2}(n-1)(n-2)$. Since the simplex u's are even smaller we conclude that *an upper bound for the number of pivot operations (returns to step 2) for this simplex algorithm is also* $\frac{1}{2}(n-1)(n-2)$.

This upper bound is tight as shown by the graph with $l_{1n} = 1$, $l_{j(j-1)} = 1$, $j = 3,4,\ldots,n$, $l_{jk} = 2(j-k)$ for $2 \leq k < j-1$ and the other l_{ij}'s equal to ∞.

For this algorithm we passed cyclically through the indices $j = 2,3,\ldots,n$. If we allow any order then the number of pivot steps can grow exponentially with n, an example will be published elsewhere.

4. DYNAMIC PROGRAMMING

In discrete dynamic programming we study a mathematical model specified by four objects $(I, A(i), p_{ij}(a), c(i,a))$. We are concerned with a dynamic system which at the decision epochs $t = 1,2,\ldots$ is observed to be in one of the states of state space I. After observing the state of the system, an action or decision must be chosen. For any state $i \in I$, the set A(i) denotes the set of possible actions in state i. If the system is in state i at any decision epoch and action $a \in A(i)$ is chosen, then regardless of the history of the system, the following happens:

(i) an immediate cost c(i,a) is incurred;

(ii) at the next decision epoch the system will be in state j with probability $p_{ij}(a)$ where $\sum_{j \in I} p_{ij}(a) = 1$ for all $i \in I$ and $a \in A(i)$.

A policy π for controlling the system is any (possibly randomized) rule for choosing actions. The objective is to find a policy having minimal expected cost. In this paper we will only deal with the case that there are only finitely many decision epochs, say $t = 1,2,\ldots,m$; m is often called the horizon of the problem, hence we restrict ourselves to finite horizon problems.

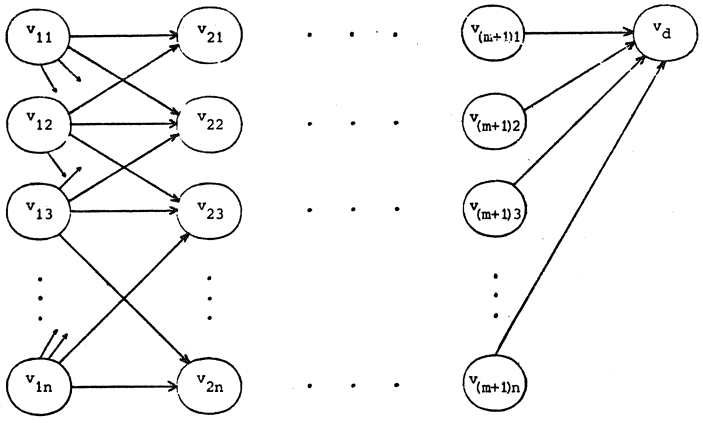
We denote by X_t resp. Y_t the state resp. the chosen action at the decision epoch t. The conditional expectation given the state at time 1 is i is denoted by $\mathbb{E}_{i,\pi}$, where π is the policy used. The problem can now be stated as, find a policy π such that

$$(4.1) \qquad \sum_{t=1}^{m} \sum_{i \in I} \mathbb{E}_{i,\pi}\, c(X_t, Y_t) \text{ is minimal.}$$

In this paper we want to focuss on the relation between dynamic programming and linear programming. Therefore we will skip in first instance the stochastics in the dynamic programming model and hence we assume that for any pair (i,a) there is exactly one j such that $p_{ij}(a) = 1$. Hence, if we take decision a in state i then at the next decision epoch the system will be in state j. Let us extend the sets of actions such that to any pair of states (i,j) there is an action a with $p_{ij}(a) = 1$, if a is not an element of A(i) then define the associated cost c(i,a) as infinite. Let us define l_{ij} as c(i,a) if $p_{ij}(a) = 1$, if there are several actions which lead from i to j then we define $l_{ij} = \min\{c(i,a) \mid p_{ij}(a) = 1\}$.

The assertion we are going to show is that the above problem of finding a best policy π is a shortest-paths problem in an acyclic graph, if we restrict the class of policies to the pure policies.

To this end let us define the directed graph G = (V,A) with set of nodes $V = \{v_d, v_{kj}, k=1,2,\ldots,m+1, j=1,2,\ldots,n\}$ and as set of arcs $A = \{(v_{ki}, v_{(k+1)j}), k=1,2,\ldots,m, i,j=1,2,\ldots,n, (v_{(m+1)i}, v_d), i=1,2,\ldots,n\}$. Moreover, the length of arc $(v_{ki}, v_{(k+1)j})$ is l_{ij} and of arc $(v_{(m+1)i}, v_d)$ is zero. The nodes v_{tj}, j=1,2,...,n correspond to all possible states at time t; we take I = {1,2,...,n}.



Suppose action a is such that $p_{ij}(a) = 1$. If we choose action a in state i at decision epoch t then the state at decision epoch (t+1) will be j. In our graph we then go from node v_{ti} to node $v_{(t+1)j}$. The associated cost is equal to l_{ij}.

A pure (also called nonrandomized) policy corresponds with assigning to any node, with the exception of v_d, an outward pointed arc. Jointly these arcs are a directed spanning tree with v_d as top i.e. from any node there is a path to v_d.

The expected cost when starting in state i and using policy π is then the length of the path from v_{1i} to v_d. Hence finding a best pure policy is equivalent to finding shortest paths from v_{1j}, $j = 1,2,\ldots,n$ to v_d.

Note that while in section 3 we had shortest paths from one designated node, we have here the problem of finding shortest paths towards one designated node v_d. It is clear that by reversing directions the problems can be converted into each other, in fact we have here another kind of duality. We could have posed the dynamic programming problem in the dual way, however we preferred the more common backward formulation. What backward does mean in this context will be clear below.

Similar as in section 3 it can be shown that finding shortest paths is equivalent to finding a minimal cost flow satisfying

$$(4.2) \qquad \sum_j x_{ij}^1 = 1, \ \sum_j x_{ij}^{t+1} - \sum_j x_{ji}^t = 0, \ t = 1,2,\ldots,m-1, \ i = 1,2,\ldots,n,$$

where the variable x_{ij}^t denotes the flow in arc $(v_{ti}, v_{(t+1)j})$.

The matrix of coefficients is again the incidence matrix of a directed graph. Hence it is totally unimodular and the basic solutions are automatically integer valued. Consequently, in order to solve our dynamic programming problem we can use the simplex method to solve the linear programming problem

$$(4.3) \qquad \min\{\sum_{t=1}^{m} \sum_{i,j} l_{ij}^t x_{ij}^t \mid \sum_j x_{ij}^1 = 1, \ \sum_j x_{ij}^{t+1} - \sum_j x_{ji}^t = 0, \ x_{ji}^t \geq 0,$$

$$t = 1,2,\ldots,m-1, \ i = 1,2,\ldots,n\},$$

where in our stationary dynamic programming problem it was assumed that $l_{ij}^t = l_{ij}$ for all t. The dual linear programming problem of (4.3) can be written as

$$\max\{\sum_j u_j^1 \mid u_j^{m+1} = 0, \ u_i^t - u_j^{t+1} \leq l_{ij}^t, \ t = 1,2,\ldots,m,$$

$$i,j = 1,2,\ldots,n\}.$$

Let us analyze how the simplex method works for the primal problem (4.3). A basic solution or vertex of the feasible region corresponds to a maximal subgraph having no circuits, such that it contains paths from nodes v_{1j}, $j = 1,2,\ldots,n$ to v_d. Note the difference with the problem in section 3, where any basic solution there had exactly (n-1) positive variables, while here the number of positive variables of basic solutions can vary from n+(m-1) to nm. These bounds are found by constructing paths from the v_{1j}, $j = 1,2,\ldots,n$ to v_d having a minimal and a maximal number of arcs, don't count the arcs from $v_{(m+1)j}$, $j = 1,2,\ldots,n$ to v_d.

Let us look at a basic solution which corresponds to a spanning tree with paths from all nodes to v_d. In the language of dynamic programming it is a pure policy.

The corresponding dual variable u_i^t is the length of the path from v_{ti} to v_d in this tree, $t = 1,2,\ldots,m$, $i = 1,2,\ldots,n$. The corresponding dual slack variables are $v_{ij}^t = l_{ij}^t + u_i^t - u_j^{t+1}$. The duality theory says, our basic solution is optimal if and only if $v_{ij}^t \geq 0$, $\forall i,j,t$. Or equivalently, a *spanning tree is optimal if and only if the distances to* v_d *satisfy,*

$$(4.4) \qquad u_i^t = \min_j (l_{ij}^t + u_j^{t+1}), \quad i = 1,2,\ldots,n, \quad t = 1,2,\ldots,m,$$

where $u_j^{m+1} = 0$, $j = 1,2,\ldots,n$.

This shows that we can find an optimal solution by solving the equations (4.4) backwards. First solve for t = m, then for t = m - 1, etc., until we find the u_i^1, $i = 1,2,\ldots,n$.

An optimal spanning tree is then found by choosing an arc in node v_{ti} *which minimizes* $(l_{ij}^t + u_j^{t+1})$ *over* j.

The total length of all paths from all v_{1i}, $i = 1,2,\ldots,n$ is by the duality theory equal to $(u_1^1 + u_2^1 + \ldots + u_n^1)$.

The backward recursion is exactly what the simplex method does if we choose the dual variables, which point to the primal variables to become positive, in the following order,

most negative number of $(v_{1j}^t, v_{2j}^t, \ldots, v_{nj}^t)$ *for all* j's *in some order, successively for* t = m,m-1,...,1.

The simplex method will also compute with any improvement new u's for all predecessors. But, if the simplex algorithm is started with the tree of all arcs from any node to v_d, this corresponds to the "big M" method with all initial basic variables artificial, then there are never predeces-

sors and hence no extra calculation is involved.

The backward recursion of (4.4) is the standard method to solve a finite horizon dynamic programming problem. We conclude that *dynamic programming can be seen as a specially structured linear programming problem. Moreover, the simplex method specified with the proper pivot rule is essentially the standard backward recursion of dynamic programming.*

Of course, this does not mean that solving a dynamic programming model with a general linear programming code is doing backward recursion. The general code will probably not use the proper pivot rule and also at any step it will probably compute all dual slack variables.

In dynamic programming the proof that an optimal policy can be computed recursively is given by induction to the horizon m. *This property of dynamic programming is called the optimality principle of Bellman. Here we derived the optimality principle from the duality theory of linear programming.*

Dynamic programmers generally, allow also randomized policies. Randomizing means that instead of choosing one outgoing arc in any node, we make a lottery over the outgoing arcs. In node v_{ti} there is for any j an arc $(v_{ti},v_{(t+1)j})$, say p_{ij}^t is the probability of drawing this arc. Then $\sum_j p_{ij}^t = 1$ for all i,t. Now suppose that for any node a lottery is fixed or in dynamic programming language the randomized policy π is given. Let us denote the expected number that arc $(v_{ti},v_{(t+1)j})$ is traversed by x_{ij}^t then,

(i) generally, the x_{ij}^t will not be integer valued

(ii) the x_{ij}^t satisfy the constraints of the linear programming problem (4.3); hence it is a feasible solution of (4.3)

(iii) $p_{ij}^t = x_{ij}^t \cdot (\sum_j x_{ij}^t)^{-1}$ if $\sum_j x_{ij}^t > 0$; define $p_{ii}^t = 1$ if $\sum_j x_{ij}^t = 0$

(iv) all feasible solutions of (4.3) can be generated in this way; *the feasible solutions are the expected flows for all pure and randomized policies*

(v) *the basic solutions correspond to pure policies.*

The backward recursion provides a constructive proof of the existence of an optimal policy which is pure. The existence of a pure optimal policy for any cost coefficients c(i,a) *implies that all basic solutions of the linear programming problem* (4.3) *are integer valued.* Indeed, if a vertex, say x, of the convex feasible region was not integer valued then by a separation theorem, there are l_{ij}'s such that the unique optimal solution is x. However, any optimal pure policy gives an optimal feasible solution

which is also integer valued.

Hence, the integrality of basic solutions of (4.3) *together with the existence of pure optimal policies are consequences of the validity of the backward recursion. Consequently, also these properties can be seen as corollaries of the duality theory of linear programming.*

Let us study the stochastic dynamic programming problem. A pure policy now means that in any node an action is given. Say in node v_{ti} we have action a then with probability $p_{ij}(a)$ the arc $(v_{ti}, v_{(t+1)j})$ will be transversed. Hence, a policy corresponds to a collection of lotteries, in any node one, with as outcomes the outgoing arcs in that node.

For a Markov policy the lotteries are independent experiments, for a history-remembering policy the action a and hence the lottery at time t may depend on the outcomes and actions at times 1,2,...,t-1.

The elementary outcome of the composition of the experiments is a spanning tree with top v_d.

Any policy induces probabilities for the elementary outcomes of the sample space of spanning trees.

Any spanning tree has assigned to it a number say L which is equal to the total length of the paths from v_{1j}, j = 1,2,...,n to v_d. The objective in stochastic dynamic programming is to find a policy for which the expectation of L is minimal.

In order to formulate the stochastic dynamic programming problem as a linear programming problem let us denote the probability of the outcome with arcs $(v_{1i_1}, v_{2i_2}), (v_{2i_2}, v_{3i_3}), \ldots, (v_{(t-1)i_{t-1}}, v_{ti_t})$ and $a_1, a_2, \ldots, a_t$ as actions by

$$x^t_{i_1,a_1,i_2,a_2,\ldots,i_t,a_t}.$$

Then the probabilistic version of the balance equation or constraint $\sum_j x^{t+1}_{ij} - \sum_j x^t_{ji} = 0$ of (4.3) becomes

$$\sum_{i_1,a_1,\ldots,i_t,a_t} \sum_a x^{t+1}_{i_1,a_1,\ldots,i_t,a_t,i,a} -$$

$$- \sum_{i_1,a_1,\ldots,i_t,a_t} x^t_{i_1,a_1,\ldots,i_t,a_t} p_{i_t i}(a_t) = 0.$$

The objective function is

$$\sum_{t=1}^{m} \sum_{i_1,a_1,\ldots,i_t,a_t} c(i_t,a_t)x^t_{i_1,a_1,\ldots,i_t,a_t}.$$

If we write x^t_{ia} for $\sum_{i_1,a_1,\ldots,i_{t-1},a_{t-1}} x^t_{i_1,a_1,\ldots,i_{t-1},a_{t-1},i,a}$

then this linear programming problem can be simplified to

$$(4.5) \qquad \min\{\sum_{t=1}^{m} \sum_{i,a} c(i,a)x^t_{ia} \mid \sum_a x^1_{ia} = 1, \sum_a x^{t+1}_{ia} - \sum_{j,a} x^t_{ja}p_{ji}(a) = 0,\ x^t_{ia} \geq 0,$$

$$t = 1,2,\ldots,m-1,\ i = 1,2,\ldots,n\}.$$

This simplification is in fact the proof that without loos of value the class of policies can be restricted to the Markov policies. Indeed, $\sum_a x^t_{ia}$ is the expected flow (flows are not integer valued here) through node v_{ti} and this flow also corresponds to the Markov policy with probability $x^t_{ia}/\sum_a x^t_{ia}$ of choosing action a in node v_{ti}.

The dual problem of (4.5) can be written as

$$(4.6) \qquad \max\{\sum_j u^1_j \mid u^{m+1}_i = 0,\ u^t_i - \sum_j p_{ij}(a)u^{t+1}_j \leq c(i,a),\ t = 1,2,\ldots,m,$$

$$a \in A(i),\ i = 1,2,\ldots,n\}.$$

In the stochastic problem as in the deterministic problem the basic solutions correspond to pure policies, backward recursion is a specialization of the simplex method etc. etc.

The backward recursion for the stochastic problem is

$$u^{m+1}_i = 0,\ i = 1,2,\ldots,n$$

and

$$u^t_i = \min_{a \in A(i)} \{c(i,a) + \sum_j p_{ij}(a)\ u^{t+1}_j\},\ i = 1,2,\ldots,n,$$

$$t = m,m-1,\ldots,1.$$

A pure optimal policy is found by choosing for any node a minimizing action.

5. COMMENTS AND REFERENCES

A standard book on linear programming is

DANTZIG, G.B. (1963), *Linear Programming and Extensions,* Princeton University Press.

The material of section 2 is partly borrowed from the book.

ZOUTENDIJK, G. (1976), *Mathematical Programming Methods,* North-Holland Publishing Company.

The writing of section 3 on shortest paths was stimulated by the book.

LAWLER, E.L. (1976), *Combinatorial Optimization: Networks and Matroids,* Holt, Rinehart and Winston.

A standard book on dynamic programming is,

BELLMAN, R. (1957), *Dynamic Programming,* Princeton University Press.

A book on dynamic programming and Markov decision chains which pays attention to the linear programming formulation of dynamic programming problems is

DERMAN, C. (1970), *Finite State Markovian Decision Processes,* Academic Press.

RECENT DEVELOPMENTS IN THE SPECTRAL ANALYSIS OF MATRIX AND OPERATOR POLYNOMIALS

M.A. Kaashoek

Operator roots play an important role in the study of operator polynomials (see [15]). For example, consider the quadratic operator differential equation

$$u^{(2)} + A_1 u^{(1)} + A_0 u = 0, \tag{0.1}$$

and let Z_1 and Z_2 be right operator roots of $\lambda^2 I + \lambda A_1 + A_0 = 0$, that is,

$$Z_i^2 + A_1 Z_i + A_0 = 0 \quad (i = 1,2).$$

Then the function

$$u(t) = e^{tZ_1}x_1 + e^{tZ_2}x_2, \tag{0.2}$$

where x_1 and x_2 are given vectors, is a solution of equation (0,1). If, in addition, the operator $Z_1 - Z_2$ is invertible, then each solution u of equation (0.1) is of the form (0.2) and this representation is unique. Moreover in that case the polynomial $\lambda^2 I + \lambda A_1 + A_0$ is uniquely determined by the roots Z_1 and Z_2, in fact

$$\lambda^2 I + \lambda A_1 + A_2 = (Z_1 - Z_2)(\lambda I - Z_2)(Z_1 - Z_2)^{-1}(\lambda I - Z_1).$$

Speaking in somewhat more general terms, we are dealing here with a monic left multiple L of a given family of monic operator polynomials $L_1,\ldots,L_r$, and we are interested to know to what extent the properties of L are determined by those of its right divisors $L_1,\ldots,L_r$. In other words, given the divisors $L_1,\ldots,L_r$ can one reconstruct the multiple L or a "part"

of L? In this form the problem is related to the following question: Given a finite family $L_1,\ldots,L_r$ of monic operator polynomials, can one construct a monic common left multiple of $L_1,\ldots,L_r$ and describe the properties of such multiples in terms of the original polynomials?

In the past few years I. GOHBERG, P. LANCASTER and L. RODMAN ([7,8,9]) have developed a theory of monic matrix and operator polynomials which is most useful in the context considered here (cf. [2,3]). The GOHBERG - LANCASTER - RODMAN theory is based on a careful analysis of the spectral properties of the polynomials concerned, in particular, it takes into account the full Jordan structure of the polynomials. Recently, I GOHBERG and L. RODMAN have carried out a similar analysis for non-monic regular matrix polynomials (see [10,11,12]). In this paper we shall give a survey of some of the main elements of both theories and show their relevance for the problems concerning multiples mentioned above.

1. PRELIMINARIES

Consider a matrix polynomial of the form

$$L(\lambda) = A_0 + \lambda A_1 + \ldots + \lambda^{\ell} A_{\ell}, \tag{1.1}$$

where the coefficients $A_0,A_1,\ldots,A_{\ell}$ are $n\times n$ complex matrices. If $A_{\ell} \neq 0$, then ℓ is called the *degree* of L. We call L *monic* if $A_{\ell} = I$, and L is said to be *comonic* if $A_0 = I$. If $\det L(\lambda)$ is not identically equal to zero, then L is called *regular*. The study of regular matrix polynomials can be reduced to that of comonic polynomials by using the transformation $L(\alpha)^{-1}L(\lambda+\alpha)$, where α has been chosen such that $\det L(\alpha) \neq 0$.

The word *operator* is used for bounded linear operators acting between complex Banach spaces. If in (1.1) the coefficients are operators acting on the same Banach space $\mathcal{B}$, then L is called an *operator polynomial*. By definition the *spectrum* of an operator polynomial is the set of all λ such that $L(\lambda)$ is not two-sided invertible. In the matrix case the spectrum of L is the set of all λ such that $\det L(\lambda) = 0$. It follows that a matrix polynomial L is regular if and only if L has discrete spectrum.

If $\mathcal{B}$ is a Banach space, then $\mathcal{B}^{\ell}$ denotes the direct sum of ℓ copies of $\mathcal{B}$ endowed with the usual normable topology. Operators from $\mathcal{B}^{\ell}$ into $\mathcal{B}^{m}$ will often be denoted by $m\times\ell$ operator matrices whose entries are operators on $\mathcal{B}$.

The symbol $\mathrm{col}(T_j)_{j=1}^r$ denotes the one column operator matrix whose entry in the j-th row is equal to T_j. Similarly, $\mathrm{row}(T_j)_{j=1}^r$ denotes the one row operator matrix with T_j in the j-th column. Further, $\mathrm{diag}(T_j)_{j=1}^r$ will denote the $r\times r$ operator matrix $(T_j\delta_{ij})_{i,j=1}^r$. Sometimes this operator will also be denoted by $T_1 \oplus T_2 \oplus \ldots \oplus T_r$. Notations of this type will also be used to describe partitions of a matrix into sub-matrices.

2. JORDAN CHAINS, EIGEN PAIRS AND SPECTRAL PAIRS

Let $L(\lambda) = \sum_{j=0}^{\ell} \lambda^j A_j$ be a regular $n\times n$ matrix polynomial. A point $\lambda_0 \in \mathbb{C}$ is called an *eigenvalue* of L if $\det L(\lambda_0) = 0$. In that case there exists a non-zero vector $x_0 \in \mathbb{C}_n$ such that

$$L(\lambda_0)x_0 = 0 \tag{2.1}$$

We call x_0 an *eigenvector* of L. A system $(x_0,x_1,\ldots,x_{k-1})$ of vectors in $\mathbb{C}_n$ is said to be a *Jordan chain* (or Keldysh chain) for L corresponding to λ_0 and x_0 if

$$L(\lambda_0)x_j + \frac{1}{1!}L^{(1)}(\lambda_0)x_{j-1} + \ldots + \frac{1}{j!}L^{(j)}(\lambda_0)x_0 = 0 \quad (0\le j\le k-1). \tag{2.2}$$

The number k is called the *length* of the chain, and if $x_0 \neq 0$ (i.e. the vector x_0 is an eigenvector), then $x_1,\ldots,x_{k-1}$ are called *generalized eigenvectors* of L. In case $L(\lambda) = A_0 - \lambda I$, the vectors $x_0,\ldots,x_{k-1}$ form a Jordan chain for L corresponding to λ_0 if and only if

$$(A_0-\lambda_0 I)x_0 = 0, \quad (A_0-\lambda_0 I)x_j = x_{j-1} \quad (1 \le j \le k-1).$$

Jordan chains for L are intimately connected with solutions of the matrix differential equation

$$L(\frac{d}{dt})u = 0. \tag{2.3}$$

If one looks for a solution of this equation of the form $u(t) = e^{\lambda_0 t}x_0$, where x_0 is a fixed vector in $\mathbb{C}_n$, then automatically one is led to consider the eigenvalue equation (2.1). In general not every solution of (2.3)

is of the form $u(t) = e^{\lambda_0 t}x_0$, because one has to allow for superpositions too. Put

$$u(t) = e^{\lambda_0 t}\left(x_{k-1} + \frac{1}{1!}x_{k-2}t + \ldots + \frac{1}{(k-1)!}x_0 t^{k-1}\right),$$

where $x_0,\ldots,x_{k-1}$ are in $\mathbb{C}_n$. Then it is not difficult to check that such a function u is a solution of the equation (2.3) if and only if $(x_0,\ldots,x_{k-1})$ is a Jordan chain for L at λ_0.

As each x_j may be considered as a column vector, a system of vectors $x_0,\ldots,x_{k-1}$ defines in a natural way an $n\times k$ matrix X_0: the entry of X_0 in the i-th row and j-th column is equal to the i-th coordinate of the vector x_j. Let J_0 be the $k\times k$ Jordan block with λ_0 on the main diagonal. Then it follows from the remark made in the previous paragraph that $(x_0,\ldots,x_{k-1})$ is a Jordan chain for L at λ_0 if and only if the $n\times k$ matrix function

$$U(t) = X_0 e^{J_0 t}$$

is a solution of the homogeneous differential equation (2.3). But this implies (see [16], Section 2) that $(x_0,\ldots,x_{k-1})$ is a Jordan chain for L corresponding to λ_0 if and only if

$$A_0X_0 + A_0X_0J_0 + \ldots + A_\ell X_0 J_0^\ell = 0. \tag{2.4}$$

This formula plays a fundamental role throughout this paper.

To organize the spectral data of the eigenvalue λ_0, we shall follow the procedure described in [13]. Let x_0 be an eigenvector of L corresponding to λ_0. By the *rank* of x_0 we shall mean the maximal length of a Jordan chain for L corresponding to λ_0 and x_0. Choose an eigenvector $x_0^{(1)}$ for L with eigenvalue λ_0 such that the rank r_1 of $x_0^{(1)}$ is maximal, and let $(x_0^{(1)},\ldots,x_{r_1-1}^{(1)})$ be a corresponding Jordan chain. Let M_1 be a direct complement in $\mathrm{Ker}L(\lambda_0)$ of the linear space spanned by $x_0^{(1)}$. Take in M_1 an eigenvector $x_0^{(2)}$ of maximal rank, r_2 say, and let $(x_0^{(2)},\ldots,x_{r_2-1}^{(2)})$ be a corresponding Jordan chain. Next we let M_2 be a direct complement in M_1 of the linear space spanned by $x_0^{(2)}$, and we repeat the procedure described above for M_2 instead of M_1. In this way we obtain a maximal linearly independent set $x_0^{(1)},\ldots,x_0^{(p)}$ of eigenvectors of L with eigenvalue λ_0 and corresponding Jordan chains

$$(x_0^{(1)},\ldots,x_{r_1-1}^{(1)}),\ldots,(x_0^{(p)},\ldots,x_{r_p-1}^{(p)}).$$

The numbers $r_1,\ldots,r_p$ (which do not depend on the particular choices made above) are the so-called partial multiplicities of λ_0. Their sum $r_1+\ldots+r_p$ is equal to the order of λ_0 as a zero of $\det L(\lambda)$.

As we have explained above, with each Jordan chain $(x_0^{(j)},\ldots,x_{r_j-1}^{(j)})$ one can associate in a canonical way a pair of matrices $(X_0^{(j)},J_0^{(j)})$, namely $J_0^{(j)}$ is the single Jordan block of order r_j with λ_0 on the main diagonal and $X_0^{(j)}$ is the $n\times r_j$ matrix whose i-th column is equal to the column vector $x_{i-1}^{(j)}$. Put

$$X_{\lambda_0} = \operatorname{row}(X_0^{(j)})_{j=1}^{p}, \quad J_{\lambda_0} = \operatorname{diag}(J_0^{(j)})_{j=1}^{p}.$$

The pair $(X_{\lambda_0},J_{\lambda_0})$ is called an *eigenpair* of L corresponding to λ_0. (The name eigenpair will also be used for any pair of matrices $(\tilde{X}_{\lambda_0},\tilde{J}_{\lambda_0})$, which is obtained from $(X_{\lambda_0},J_{\lambda_0})$ by some permutation of the blocks $J_0^{(j)}$ in J_{λ_0} and the same permutation of the corresponding blocks in X_{λ_0}.) Note that J_{λ_0} is a Jordan matrix with a single eigenvalue λ_0, the order of J_{λ_0} is equal to the order of λ_0 as a zero of $\det L(\lambda)$ and

$$A_0X_{\lambda_0} + A_1X_{\lambda_0}J_{\lambda_0} + \ldots + A_\ell X_{\lambda_0}J_{\lambda_0}^\ell = 0.$$

Further one can show that the matrix $\operatorname{col}(X_{\lambda_0}J_{\lambda_0}^{i-1})_{i=1}^{\ell}$ has full rank. These properties completely characterize the eigenpairs of L corresponding to λ_0 ([12], Theorem 1.1).

Let $\lambda_0,\lambda_1,\ldots,\lambda_m$ be the different eigenvalues of L, and let (X_j,J_j) be an eigenpair of L corresponding to λ_j $(0 \le j \le m)$. The pair of matrices (X,J), where

$$X = \operatorname{row}(X_j)_{j=1}^{m}, \quad J = \operatorname{diag}(J_j)_{j=1}^{m},$$

is called a *spectral pair* of L. It epitomizes all the information about the (finite part of the) spectrum of L.

We shall see that the notion of a spectral pair plays an important role in the study of matrix polynomials. As a first application, let us mention that the general solution of the homogeneous matrix differential equation $L(\frac{d}{dt})u = 0$ is of the form

$$u(t) = Xe^{tJ}x,$$

where (X,J) is a spectral pair of L and x is some vector in $\mathbb{C}_\nu$ with ν equal to the degree of $\det L(\lambda)$ ([16], Theorem 1.1). For other applications to (homogeneous as well as non-homogeneous) matrix differential and finite difference equations we refer to [16,17,2,4]. For some explicit examples of spectral pairs see [10].

3. INVERSE PROBLEMS AND STANDARD PAIRS

A regular matrix polynomial L is not uniquely determined by a spectral pair.

THEOREM 3.1. ([10], Theorem 5.1). *If the regular matrix polynomials* L_1 *and* L_2 *have a common spectral pair, then* $L_2(\lambda) = E(\lambda)L_1(\lambda)$, *where* $E(\lambda)$ *is an everywhere invertible matrix polynomial.*

Multiplying L on the left by an everywhere invertible matrix polynomial does not change the (finite part of the) spectrum of L, but it may change the spectrum at infinity considerably. As monic polynomials have no spectrum at infinity, one might expect that in the monic case a spectral pair completely determines the polynomial. This indeed is the case.

THEOREM 3.2. ([7], Theorem 1). *Let* (X,J) *be a spectral pair of the monic matrix polynomial* L, *and let* ℓ *be the degree of* L. *Then* $\operatorname{col}(XJ^{i-1})_{i=1}^{\ell}$ *is invertible and*

$$L(\lambda) = \lambda^{\ell}I - XJ^{\ell}(V_1+\lambda V_2 + \ldots + \lambda^{\ell-1}V_{\ell}), \tag{3.1}$$

where $\operatorname{row}(V_j)_{j=1}^{\ell} = [\operatorname{col}(XJ^{i-1})_{i=1}^{\ell}]^{-1}$.

In the special case where $L(\lambda) = T-\lambda I$, the previous theorem is nothing else than the reduction of T to Jordan normal form by a similarly transformation.

For the analysis of monic polynomials it has been important to note that representations of L, similar to the one of formula (3.1), may also be given in terms of pairs more general than spectral pairs. A pair of matrices (Q,T) is called a *standard pair* of *degree* k if Q is an $m\times nk$ matrix,

T is an nk×nk matrix and the matrix $\mathrm{col}(QT^{i-1})_{i=1}^{k}$ is invertible. Further (Q,T) is said to be a standard pair of the monic n×n matrix polynomial $L(\lambda) = \lambda^{\ell}I + \sum_{j=0}^{\ell-1} \lambda^{j} A_j$ if, in addition, its degree is equal to ℓ and

$$QT^{\ell} + A_{\ell-1} QT^{\ell-1} + \ldots + A_0 Q = 0. \tag{3.2}$$

From the definition of a spectral pair (cf. formula (2.4)) and the first part of Theorem 3.2 it is clear that a spectral pair for the monic polynomial L is a standard pair of L, but there are many other standard pairs too. For example, the pair (Y,C_L), where

$$Y = [I\ 0\ \ldots\ 0], \quad C_L = \begin{bmatrix} 0 & I & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & I \\ -A_0 & -A_1 & \ldots & -A_{\ell-1} \end{bmatrix}, \tag{3.3}$$

is a standard pair of $L(\lambda) = \lambda^{\ell}I + \sum_{j=0}^{\ell-1} \lambda^{j}A_j$, which in general is not a spectral pair. Nevertheless Theorem 3.2 remains true if in this theorem the spectral pair (X,J) is replaced by any standard pair of L. The reason for this is that any two standard pairs (Q_1,T_1) and (Q_2,T_2) of L are similar in the sense that $Q_1 = Q_2S$ and $T_2S = ST_1$ for some invertible matrix S (see [7]).

To get a representation theorem as Theorem 3.2 for arbitrary regular matrix polynomials, one has to take into account the spectrum at infinity too. For comonic polynomials this may be done as follows.

THEOREM 3.3. ([11]). *Let* L *be a comonic* n×n *matrix polynomial of degree* ℓ. *Put*

$$Q = [X\ X_{\infty}], \quad T = J^{-1} \oplus J_{\infty},$$

where (X,J) *is a spectral pair of* L *and* (X_{∞},J_{∞}) *is an eigenpair of* $\lambda^{\ell}L(\lambda^{-1})$ *corresponding to the eigenvalue* $\lambda_0 = 0$. *Then* (Q,T) *is a standard pair of degree* ℓ *and*

$$L(\lambda) = I - QT^{\ell}(Z_1\lambda^{\ell} + Z_2\lambda^{\ell-1} + \ldots + Z_{\ell}\lambda),$$

where $\mathrm{row}\,(Z_j)_{j=1}^{\ell} = [\mathrm{col}(QT^{i-1})_{i=1}^{\ell}]^{-1}$.

The notion of a standard pair carries over to the infinite dimensional case easily and so does the representation theorem for monic polynomials. A pair of operators $Q: \mathcal{B}^k \to \mathcal{B}$ and $T: \mathcal{B}^k \to \mathcal{B}^k$ is said to form a *standard pair* (Q,T) of *degree* k whenever the operator $\operatorname{col}(QT^{i-1})_{i=1}^{k}$ is two-sided invertible, and, as in the matrix case, the standard pair (Q,T) is called a standard pair of the monic operator polynomial $L(\lambda) = \lambda^{\ell}I + \sum_{j=0}^{\ell-1} \lambda^j A_j$ if its degree is equal to ℓ and formula (3.2) holds true. Also in the infinite dimensional case the pair of operators given by formula (3.3) provides an example of a standard pair of $L(\lambda) = \lambda^{\ell}I + \sum_{j=0}^{\ell-1} \lambda^j A_j$.

THEOREM 3.4. ([9], Theorem 1). *Let* (Q,T) *be a standard pair of the monic operator polynomial* L, *and let* ℓ *be the degree of* L. *Then*

$$L(\lambda) = \lambda^{\ell}I - QT^{\ell}(U_1+\lambda U_2 + \ldots + \lambda^{\ell-1}U_{\ell}), \tag{3.4}$$

where $\operatorname{row}(U_j)_{j-1}^{\ell} = [\operatorname{col}(QT^{i-1})_{i=1}^{\ell}]^{-1}$.

The representation (3.4) is the so-called *right normal form* of L. Also a left normal form may be given (see [7,9]). Further, standard pairs may be employed to give a "neutral" (neither left nor right) representation for the inverse $L(\lambda)^{-1}$ of a monic operator polynomial L (see [8,9]). In fact, using the notations of the previous theorem, we have

$$L(\lambda)^{-1} = Q(\lambda I - T)^{-1}R,$$

where $R = U_{\ell}$. This so-called resolvent form for L allows us to see $L(\lambda)^{-1}$ as a characteristic operator function associated with a new kind of operator node (see [1]).

4. DIVISIBILITY AND JORDAN CHAINS

The notion of standard pairs is most suitable to describe quotient and remainder after division on the right.

THEOREM 4.1. ([9], Theorem 12). *Let* $L(\lambda) = \sum_{j=0}^{\ell} \lambda^j A_j$ *be an operator polynomial, and let* $L_1(\lambda) = \lambda^k I - Q_1 T_1^k (V_1+\lambda V_2 + \ldots + \lambda^{k-1}V_k)$ *be a monic operator polynomial in right normal form. Suppose* $k \geq \ell$. *Then*

$$L(\lambda) = S(\lambda)L_1(\lambda) + R(\lambda),$$

where

$$\text{(i)} \qquad S(\lambda) = \sum_{j=0}^{\ell-k} \lambda^j \Big(\sum_{i=j+1}^{\ell} A_i Q_1 T_1^{i-j-1} V_k \Big),$$

$$\text{(ii)} \qquad R(\lambda) = \sum_{j=1}^{k} \lambda^{j-1} \Big(\sum_{i=0}^{\ell} A_i Q_1 T_1^{i} V_j \Big).$$

Let L and L_1 be operator (or matrix) polynomials. We call L_1 a *right divisor* of L whenever there exists an operator (or matrix) polynomial $S(\lambda)$ such that

$$L(\lambda) = S(\lambda)L_1(\lambda).$$

Now let L and L_1 be as in the previous theorem. Then L_1 is a right divisor of L if and only if

$$\text{(4.1)} \qquad \sum_{i=0}^{\ell} A_i Q_1 T_1^{i} V_j = 0 \quad (1 \le j \le k).$$

As $\operatorname{row}(V_j)_{j=1}^{k} = [\operatorname{col}(Q_1 T_1^{i-1})_{i=1}^{k}]^{-1}$ is invertible, we see that (4.1) is equivalent to the requirement that

$$\text{(4.2)} \qquad A_0Q_1 + A_1Q_1T_1 + \ldots + A_\ell Q_1 T_1^{\ell} = 0,$$

and hence L_1 is a right divisor of L if and only if (4.2) holds for an arbitrary standard pair (Q_1,T_1) of L_1.

To understand the analytical aspects of this result, let us suppose that L and L_1 are monic matrix polynomials. In that case we may assume that (Q_1,T_1) is a spectral pair of L_1. But then it is easily seen that (4.2) is equivalent to the statement that each Jordan chain of L_1 is a Jordan chain of L. It follows that L_1 is a right divisor of L if and only if each Jordan chain of L_1 is a Jordan chain of L. In this form the result carries over to the non-monic case.

THEOREM 4.2. ([11], Theorem 4.2). *Let* L *and* L_1 *be regular* n×n *matrix polynomials. Then* L_1 *is a right divisor of* L *if and only if each Jordan chain of* L_1 *is a Jordan chain of* L *corresponding to the same eigenvalue as for* L_1.

To illustrate the previous result, let us consider the following theorem, which for the monic case has been proved by H. LANGER in [19].

THEOREM 4.3. ([18], Satz 2). *Let* $L(\lambda) = \sum_{j=0}^{\ell} \lambda^j A_j$ *be a regular* n×n *matrix polynomial. Then there exists an* n×n *matrix* Z *such that*

$$A_0 + A_1 Z + \ldots + A_\ell Z^\ell = 0 \tag{4.3}$$

if and only if L *has Jordan chains* $(x_0^{(j)},\ldots,x_{k_j-1}^{(j)})$, $1 \le j \le p$, *such that* $\sum_{j=1}^{p} k_j = n$ *and the vectors* $x_i^{(j)}$ $(i = 0,\ldots,k_j-1;\ j = 1,\ldots,p)$ *are linearly independent.*

To prove Theorem 4.3, first note that formula (4.3) is equivalent to the statement that $\lambda I - Z$ is a right divisor of L. Let X_j be the $n\times k_j$ matrix whose i-th column is equal to the column vector $x_{i-1}^{(j)}$, and let J_j be the single Jordan block of order k_j with λ_j on the main diagonal. Here λ_j is the eigenvalue corresponding to the Jordan chain $(x_0^{(j)},\ldots,x_{k_j-1}^{(j)})$. Put

$$X = \operatorname{row}(X_j)_{j=1}^{p},\quad J = \operatorname{diag}(J_j)_{j=1}^{p}.$$

From our hypothesis on the given Jordan chains it follows that X is invertible. Let $Z = XJX^{-1}$. Then (X,J) is a spectral pair for Z, and it follows that each Jordan chain of $\lambda I - Z$ is a Jordan chain for L corresponding to the same eigenvalue. But then we may apply Theorem 4.2 to show that $\lambda I - Z$ is a right divisor of L. The converse statement is trivial.

Let (Q,T) be a standard pair of the monic operator polynomial L. There are many interesting relations between monic right divisors of L and certain T-invariant subspaces. First of all there is a one-one correspondence between these so-called supporting subspaces and the monic right divisors of L. Further an explicit description of the quotient may be given in terms of the corresponding supporting subspaces and the original standard pair (Q,T). For details concerning these matters we refer to [7] and [9], see also [20]. For the non-monic case similar but less complete results have been proved (see [10], [11]). For example, if (X,J) is a spectral pair of the regular matrix polynomial L, then all right divisors of L may be described in terms of the invariant subspaces of J (cf. [11], Theorem 4.1). Here the notion of extension of admissible pairs plays an important role (cf. [12]).

5. COMMON MULTIPLES AND THE VANDERMONDE OPERATOR

In this section we return to the questions about multiples referred to in the introductory paragraphs. Let $L_1,\dots,L_r$ be monic operator polynomials in right normal form, i.e.,

$$L_p(\lambda) = \lambda^{k_p} I - Q_p T_p^{k_p}(V_{p1}+\lambda V_{p2} + \dots + \lambda^{k_p-1} V_{pk_p}),$$

where $\operatorname{row}(V_{pi})_{i=1}^{k_p} = [\operatorname{col}(Q_p T_p^{i-1})_{i=1}^{k_p}]^{-1}$. From Theorem 4.1 it is clear that the operator polynomial $L(\lambda) = \sum_{j=0}^{\ell} \lambda^j A_j$ is a common left multiple of $L_1,\dots,L_r$ if and only if

$$\sum_{i=0}^{\ell} A_i Q_p T_p^i V_{pj} = 0 \quad (1 \le j \le k_p,\ 1 \le p \le r).$$

Now suppose that L is monic. Then the previous formula may be written as an operator matrix identity, as follows:

$$[\operatorname{row}(A_{j-1})_{j=1}^{\ell}][(Q_p T_p^{i-1} V_p)_{i=1,p=1}^{\ell\ \ r}] = -\operatorname{row}(Q_p T_p^{\ell} V_p)_{p=1}^{r}, \tag{5.1}$$

where $V_p = \operatorname{row}(V_{pi})_{i=1}^{k_p}$. The operator matrix, which appears as the second factor in the left hand side of (5.1), is called the *Vandermonde operator* of *order* ℓ of the polynomials $L_1,\dots,L_r$ (see [2,3], cf. [14]). It will be denoted by $V_\ell(L_1,\dots,L_r)$ or simply by V_ℓ.

The definition of the Vandermonde operator does not depend on the special choice of the standard pairs. In fact the entries $Q_p T_p^{\alpha} V_{p\beta}$ can be expressed uniquely in terms of the coefficients of $L_1,\dots,L_r$ ([2], Theorem 2.2). For example in the special case where $L_1,\dots,L_r$ have degree one, i.e., $L_p(\lambda) = \lambda I - Z_p$, we have

$$V_\ell(L_1,\dots,L_r) = (Z_p^{i-1})_{i=1,p=1}^{\ell\ \ r}. \tag{5.2}$$

In the scalar case this is just the usual Vandermonde matrix. (The right hand side of (5.2) is the Vandermonde operator as used by A.S. MARKUS and I.V. MEREUTSA in their study of operator roots [21].)

If

$$[A_0 \dots A_{\ell-1}]V_\ell = -\operatorname{row}(Q_p T_p^{\ell} V_p)_{p=1}^{r}, \tag{5.3}$$

that is, if $L(\lambda) = \lambda^{\ell} I + \sum_{j=0}^{\ell} \lambda^{j} A_j$ is a common left multiple of $L_1,\ldots,L_r$, then we must have

$$\text{Ker } V_{\ell} \subset \text{Ker}[\text{row}(Q_p T_p^{\ell} V_p)_{p=1}^{r}]. \tag{5.4}$$

Now $\text{row}(Q_p T_p^{\ell} V_p)_{p=1}^{r}$ is the last operator row in $V_{\ell+1}$, and hence it follows that condition (5.4) is equivalent to the statement $\text{Ker } V_{\ell} = \text{Ker } V_{\ell+1}$. Put

$$\text{ind}(L_1,\ldots,L_r) = \inf\{m \geq 1 \mid \text{Ker } V_m = \text{Ker } V_{m+1}\}.$$

This number is called the *index of stabilization* of $L_1,\ldots,L_r$. By definition it is infinite whenever $\text{Ker } V_m \neq \text{Ker } V_{m+1}$ for all m.

From the previous discussion it is clear that finiteness of the index of stabilization is a necessary condition for the existence of a common left multiple. In the finite dimensional case $\text{ind}(L_1,\ldots,L_r)$ is always finite, but in the infinite dimensional case it may be infinite, and hence in the infinite dimensional case common monic left multiples do not always exist (see [3], Section 2 for explicit examples). In general, finiteness of $\text{ind}(L_1,\ldots,L_r)$ is not sufficient for the existence of a monic left multiple ([3], Section 6), but in the finite dimensional case it is. This follows from the next theorem.

THEOREM 5.1. ([3], Section 5). *Suppose that* $\text{ind}(L_1,\ldots,L_r) \leq \ell < \infty$ *and* $V_{\ell}(L_1,\ldots,L_r)$ *has a generalized inverse. Then* $L_1,\ldots,L_r$ *have a common monic left multiple of degree* ℓ.

Recall that an operator S is said to have a *generalized inverse* if there exists an operator S^{+} such that $S = SS^{+}S$ and $S^{+} = S^{+}SS^{+}$. In the finite dimensional case each operator has a generalized inverse. So from Theorem 5.1 and the previous discussion we see that in the finite dimensional case $L_1,\ldots,L_r$ always have a common monic left multiple and the least degree of such a multiple is equal to $\text{ind}(L_1,\ldots,L_r)$ ([2], Theorem 13.1).

In general some sort of generalized invertibility seems indispensable. For example in the Hilbert space case we have the following theorem.

THEOREM 5.2. *Let the coefficients of* $L_1,\ldots,L_2$ *be operators on a Hilbert space, and suppose that the spectra of* $L_1,\ldots,L_r$ *are mutually disjoint. Then the left invertibility of* $V_{\ell}(L_1,\ldots,L_r)$ *is necessary and sufficient*

for the existence of a common monic left multiple of $L_1,\ldots,L_r$ *of degree* ℓ.

If the Vandermonde operator V_ℓ is surjective (or somewhat weaker has dense range), then there is at most one common left multiple of $L_1,\ldots,L_r$ of degree ℓ. This is clear from formula (5.3). Hence in that case the multiple of degree ℓ, if it exists, is uniquely determined by $L_1,\ldots,L_r$.

As the Vandermonde operator V_ℓ may be written as an operator matrix whose entries can be expressed in terms of sums and products of the coefficients of $L_1,\ldots,L_r$ ([2], Theorem 2.2), it follows that V_ℓ will depend analytically on ε whenever the coefficients of $L_1,\ldots,L_r$ are analytic in ε. This fact is heavily used in [6], where the problem of existence of "analytic" multiples is studied for polynomials whose coefficients are analytic in a second variable ε. In that case the index of stabilization $\mathrm{ind}(L_1,\ldots,L_r)$ has to be replaced by an analytic index of stabilization, which is defined in terms of Jordan chains of the Vandermonde operator V_ℓ as function of ε.

So far in this section we have restricted ourselves to the monic case, but with minor modifications most results hold for non-monic regular matrix polynomials too. For example, using the normal form of Theorem 3.3 a Vandermonde matrix may be introduced for comonic matrix polynomials, and its role here is similar to the one described for the monic case (see [5] for details).

REFERENCES

1. BART, H., I. GOHBERG and M.A. KAASHOEK, *Operator polynomials as inverses of characteristic functions,* Integral equations and Operator Theory 1 (1978), 1-18.

2. GOHBERG, I., M.A. KAASHOEK and L. RODMAN, *Spectral analysis of families of operator polynomials and a generalized Vandermonde matrix,* I. *The finite dimensional case,* Advances of Mathematics.

3. GOHBERG, I., M.A. KAASHOEK and L. RODMAN, *Spectral analysis of families of operator polynomials and a generalized Vandermonde matrix,* II. *The infinite dimensional case,* J. Funct. Anal.

4. GOHBERG, I., M.A. KAASHOEK, L. LERER and L. RODMAN, *Common multiples and common divisors,* I. *Spectral method,* to appear.

5. GOHBERG, I., M.A. KAASHOEK, L. LERER and L. RODMAN, *Common multiples and common divisors,* II. *Vandermonde and resultante,* in preparation.

6. GOHBERG, I., M.A. KAASHOEK and F. VAN SCHAGEN, *Common multiples of operator polynomials with analytic coefficients,* Rapport 74 van het Wiskundig Seminarium, Vrije Universiteit, Amsterdam, 1978.

7. GOHBERG, I. P. LANCASTER and L. RODMAN, *Spectral analysis of matrix polynomials,* I. *Canonical forms and divisors,* Linear Algebra and Appl., to appear.

8. GOHBERG, I., P. LANCASTER and L. RODMAN, *Spectral analysis of matrix polynomials,* II. *The resolvent form and spectral divisors,* Linear Algebra and Appl., to appear.

9. GOHBERG, I., P. LANCASTER and L. RODMAN, *Representations and divisibility of operator polynomials,* Dept. of Mathematics and Statistics, University of Calgary, Res. Paper 341, 1977.

10. GOHBERG, I. and L. RODMAN, *On spectral analysis of non-monic matrix and operator polynomials,* I. *Reduction to monic polynomials,* Israel J. Math., to appear.

11. GOHBERG, I. and L. RODMAN, *On spectral analysis of non-monic matrix and operator polynomials,* II. *Dependence on the finite spectral data,* Israel J. Math., to appear.

12. GOHBERG, I. and L. RODMAN, *On the spectral structure of monic matrix polynomials and the extension problem,* preprint, Tel-Aviv University, 1977.

13. GOHBERG, I.C. and E.I. SIGAL, *Operator applications of the theorem on logarithmic subtraction and the theorem of Rouché,* Mat. Sbornik, n. Ser. 84 (126) (1971), 607-629 [Russian] = Math. USSR-Sbornik 13 (1971), 603-625.

14. KABAK, V.I., A.S. MARCUS and I.V. MEREUTSA, *On spectral properties of divisors of a polynomial operator pencil,* Izv. Akad. Nauk Moldavian SSR, Kishinev (2) (1976), 24-26 [Russian].

15. KREIN, M.G. and H. LANGER, *Certain mathematical principles of the linear theory of damped vibrations of continua, Proc. Int. Symp. on Applications of the Theory of Functions in Continuum Mechanics,* Tbilisi, 1963, vol II: *Fluid and gas Mechanics,* Math. Methods, Nauka, Moscow, 1965, pp. 283-322 [Russian].

16. LANCASTER, P., *A fundamental theorem on lambda-matrices,* I, *Ordinary differential equations with constant coefficients,* Linear Algebra and Appl. 18 (1977), 189-211.

17. LANCASTER, P., *A fundamental theorem on lambda-matrices,* II, *Difference equations with constant coefficients,* Linear algebra and Appl. 18 (1977), 213-222.

18. LANCASTER, P. and H.K. WIMMER, *Zur Theorie der λ-Matrizen,* Math. Nachr. 70 (1975), 325-330.

19 LANGER, H., *Über Lancaster's Zerlegung von Matrizen-Scharen,* Arch. Rational Mech. Anal. 29 (1968), 75-80.

20. LANGER, H., *Factorization of operator pencils,* Acta Sci. Math. (Szeged) 38 (1976), 83-96.

21. MARCUS, A.S. and I.V. MEREUTSA, *On the complete n-tuple of roots of the operator equation corresponding to a polynomial operator bundle,* Izv. Akad. Nauk SSSR, Ser. Mat. 37 (1973), 1108-1131 [Russian] = Math. USSR-Izvestija 7 (1973), 1105-1128.

VANISHING SUMS OF ROOTS OF UNITY

H.W. Lenstra, Jr.

INTRODUCTION

This paper is a survey of what is known about the magnitude of coefficients appearing in linear relations between roots of unity. The special case of the cyclotomic polynomial is considered in section 1; section 2 is devoted to more general relations. Various open problems will be indicated.

By n and m we shall always mean positive integers, and by p a prime number; n is called *squarefree* if n is a product of distinct primes. By $m|n$ we mean that m divides n. An *n-th root of unity*, or simply an *n-th root*, is a complex number α for which $\alpha^n = 1$. It is called *primitive* if there exists no $m < n$ with $\alpha^m = 1$. The ring of integers is denoted by $\mathbb{Z}$, and $\mathbb{Q}$ denotes the field of rational numbers.

Research for this paper was supported by the Netherlands Organization for the Advancement of Pure Research (Z.W.O). Acknowledgements are due to the I.H.E.S. for its hospitality and to C.L. Stewart for providing ref. [10].

1. Coefficients of the cyclotomic polynomial

The n-th cyclotomic polynomial Φ_n is defined by

$$\Phi_n = \prod_{\zeta} (X - \zeta), \tag{1.1}$$

where ζ ranges over the primitive n-th roots of unity. We have

$$\prod_{d|n} \Phi_d = X^n - 1 \tag{1.2}$$

since both sides are equal to $\prod_{\zeta, \zeta^n=1} (X - \zeta)$. From (1.2) one deduces, by induction on n, that Φ_n has coefficients in $\mathbb{Z}$. Its degree is $\phi(n)$, where ϕ is Euler's function:

$$\phi(n) = \left|\{j:\ 0 \le j < n,\ (j,n) = 1\}\right|.$$

The cyclotomic polynomials are known to be irreducible in the polynomial ring $\mathbb{Q}[X]$.

By Moebius inversion it follows from (1.2) that

$$\Phi_n = \prod_{d|n} (X^d - 1)^{\mu(n/d)}. \tag{1.3}$$

Here μ denotes the Moebius-function:

$$\mu(m) = (-1)^r \quad \text{if } m \text{ is the product of } r \text{ distinct primes, } r \ge 0,$$
$$\mu(m) = 0 \quad \text{otherwise.}$$

The polynomials Φ_n can be determined inductively, using the formulae

$$\Phi_1 = X - 1$$

$$\Phi_{np} = \Phi_n(X^p) \quad \text{if } p \text{ divides } n, \tag{1.4}$$

$$\Phi_{np} = \Phi_n(X^p)/\Phi_n \quad \text{if } p \text{ does not divide } n. \tag{1.5}$$

To prove these relations, use (1.3), or check that both sides have the same zeros. In a similar way one proves that

$$\Phi_{2n} = (-1)^{\phi(n)} \cdot \Phi_n(-X) \quad \text{if } n \text{ is odd.} \tag{1.6}$$

For small n, no coefficient of Φ_n exceeds 1 in absolute value. In fact, this is true for $n = p$:

$$\Phi_p = (X^p - 1)/(X - 1) = 1 + X + X^2 + \ldots + X^{p-1},$$

and also for $n = pq$, where p and q are different primes:

$$\Phi_{pq} = \frac{(1-X)(1-X^{pq})}{(1-X^p)(1-X^q)} = \qquad \text{(by (1.3))}$$

$$= (1-X) \cdot \sum_{j=0}^{\infty} X^{jp} \cdot \sum_{k=0}^{p-1} X^{kq} = (1-X) \cdot \sum X^{\alpha}$$

where α ranges over the numbers of the form $jp+kq$, with $j \geq 0$, $0 \leq k < p$; it is easily proved that no integer has more than one such representation. Multiplying ΣX^{α} by $(1-X)$ we see that the non-zero coefficients of Φ_{pq} are alternately +1 and -1. For a different formula for Φ_{pq}, see (2.16).

From what we just proved and the formulae (1.4) and (1.6) it follows immediately that no coefficient of Φ_n exceeds 1 in absolute value if n has at most two distinct odd prime factors. The smallest number n not satisfying this condition is $3.5.7 = 105$, and in fact in Φ_{105} a coefficient -2 appears:

$$\begin{aligned}\Phi_{105} = 1 &+ X + X^2 - X^5 - X^6 - 2X^7 - X^8 - X^9 \\ &+ X^{12} + X^{13} + X^{14} + X^{15} + X^{16} + X^{17} \\ &- X^{20} - X^{22} - X^{24} - X^{26} - X^{28} \\ &+ X^{31} + X^{32} + X^{33} + X^{34} + X^{35} + X^{36} \\ &- X^{39} - X^{40} - 2X^{41} - X^{42} - X^{43} + X^{46} + X^{47} + X^{48}.\end{aligned}$$

It was first proved by Schur (see [13]) that the coefficients of the cyclotomic polynomials are arbitrarily large in absolute value. In order to present his argument it is convenient to rearrange formula (1.3) as follows:

$$\Phi_n = \Pi^{I}(1-X^d).\Pi^{II}(1+X^d+X^{2d}+\ldots) \tag{1.7}$$

$(n > 1)$, where in Π^{I} the product is over the divisors d of n with $\mu(n/d) = 1$, and in Π^{II} over those for which $\mu(n/d) = -1$.

Now let t be an odd integer ≥ 3, and let $p_1, p_2, \ldots, p_t$ be prime numbers with

$$2 < p_1 < p_2 < \ldots < p_t < p_1 + p_2;$$

such primes can be found for every t. We put $n = p_1 p_2 \ldots p_t$ and we calculate Φ_n modulo terms of degree $\geq p_t + 1$ using formula (1.7). The only divisors of n which are $< p_t + 1$ are $1, p_1, p_2, \ldots, p_t$, and since t is odd we obtain

$$\Phi_n \equiv (1-X^{p_1})(1-X^{p_2})\ldots(1-X^{p_t})(1+X+X^2+\ldots+X^{p_t})$$
$$\equiv (1-X^{p_1}-X^{p_2}-\ldots-X^{p_t})\ (1+X+X^2+\ldots X^{p_t})$$

modulo terms of degree $\geq p_t + 1$. Multiplying out we find that the coefficient at X^{p_t} equals $1-t$, thus finishing the proof.

From (1.7) and the fact that Φ_n has degree less than n, for $n > 1$, it is clear that any coefficient of Φ_n is in absolute value less than or equal to the corresponding coefficient of

$$\prod_{d|n} (1 + X^d + X^{2d} + \ldots + X^{n-d}).$$

Since the coefficients of this polynomial are positive, they are bounded from above by the value of the polynomial in 1, which equals

$$\prod_{d|n} \frac{n}{d} = n^{\tau(n)/2}.$$

Here $\tau(n)$ denotes the number of divisors of n. Using the fact that

$$\tau(n) < 2^{(1+\varepsilon)\log n/\log\log n}$$

for all $\varepsilon > 0$ and all $n > n_0(\varepsilon)$ (see [9, theorem 317]) we find, after an easy manipulation:

THEOREM (1.8) *For every real number* $\varepsilon > 0$ *there exists an integer* $n_0(\varepsilon)$ *such that for all* $n > n_0(\varepsilon)$ *the absolute value of any coefficient of* Φ_n *is less than*

$$\exp(n^{(1+\varepsilon)\log 2/\log\log n}).$$

Notice that this estimate, which is due to BATEMAN [2], is much better than the trivial upper bound

$$2^{\phi(n)}$$

for the sum of the absolute values of the coefficients of Φ_n, which one obtains from (1.1) by using $|\zeta| = 1$.

Bateman's estimate is in a sense best possible, since VAUGHAN [19] has

shown that there are infinitely many n for which Φ_n has a coefficient exceeding

$$\exp(n^{\log 2 / \log\log n})$$

in absolute value.

Using Bateman's argument and [9, theorem 432] one finds that for every $\varepsilon > 0$ the sum of the absolute values of the coefficients of Φ_n is less than

$$\exp((\log n)^{1+\log 2+\varepsilon})$$

for almost all n. Non-trivial lower bounds valid for almost all n are not known.

Bounds of a different nature have been obtained for numbers n having only a few odd prime factors. Using (1.4) and (1.6) we again restrict to the case n is odd and squarefree.

For $n = p$ and $n = pq$ we have already seen that Φ_n has no coefficient exceeding 1 in absolute value. For $n = pqr$, with p, q and r primes, $2 < p < q < r$, it was proved by BANG (see [4]) that all coefficients of Φ_n are at most

$$p - 1$$

in absolute value. This bound was improved to

$$p - k \quad \text{if } p = 4k \pm 1, \quad k \in \mathbb{Z},$$

by BEITER [3], and she conjectured that it may further be lowered to

$$\frac{p+1}{2}.$$

This result, if true, would be best possible, since MÖLLER [17] proved that for every odd prime p there exist infinitely many prime pairs q, r, with $p < q < r$, for which Φ_{pqr} has a coefficient $\frac{1}{2}(p+1)$.

For $n = pqrs$, with p,q,r,s primes, $2 < p < q < r < s$, the coefficients of Φ_n are bounded by

$$p(p - 1)(pq - 1)$$

in absolute value. This was proved by BLOOM [4]. He conjectured that, generally, for

$$n = p_1 p_2 \cdots p_t$$

with $p_1, p_2, \ldots, p_t$ primes, $2 < p_1 < p_2 < \ldots < p_t$, $t \geq 2$, the coefficients of Φ_n are bounded in absolute value by a number depending only on $p_1, p_2, \ldots, p_{t-2}$. This conjecture was proved by FELSCH and SCHMIDT [8] and JUSTIN [11]:

THEOREM (1.9) *There is a function* f *on the positive integers, such that for all* m, *and all primes* p,q *with*

$$p \neq q, \quad (pq,m) = 1,$$

the coefficients of Φ_{mpq} *are less than* f(m) *in absolute value.*

We present Justin's elegant proof of this theorem.
Define the polynomials Ψ_n by

$$\Psi_n \cdot \Phi_n = 1 - X^n.$$

Let m,p,q be as in the theorem. Applying (1.5) twice we get

$$\Phi_{mpq} = \frac{\Phi_m(X^{pq}) \cdot \Phi_m}{\Phi_m(X^p) \cdot \Phi_m(X^q)}$$

$$= \Phi_m(X^{pq}) \cdot \Phi_m \cdot \Psi_m(X^p) \cdot \Psi_m(X^q) \cdot (1 - X^{mp})^{-1} \cdot (1 - X^{mq})^{-1} \tag{1.10}$$

$$= A \cdot B$$

where A is the product of the first four factors in (1.10), and B is the power series

$$\sum_{j,k \geq 0} X^{jmp + kmq}.$$

If $\Phi_m = \sum_i a_i X^i$, $\Psi_m = \sum_i b_i X^i$, then the sum of the absolute values of the coefficients of A is clearly bounded by

$$(\sum_i |a_i|)^2 \cdot (\sum_i |b_i|)^2. \tag{1.11}$$

Further, if $B = \sum_i c_i X^i$ then $c_i \in \{0,1\}$ for all $i < mpq$, since no number less than mpq has more than one representation $jmp + kmq$, $j \geq 0$, $k \geq 0$. Multiplying A and B, and observing that the product Φ_{mpq} has degree $< mpq$, we conclude that all coefficients of Φ_{mpq} are bounded in absolute value by (1.11). Since this number depends only on m the theorem follows.

An explicit function f for which the conclusion of the theorem holds has been given by MÖLLER [17].

In the next section we shall see that there exists a positive constant C_1 such that for all squarefree $n > 1$ the number of non-zero coefficients of Φ_n exceeds

$$C_1 (\log n)^2 / \log \log n,$$

see (2.8). Schinzel has posed the problem to improve this estimate. It is known that for every $\varepsilon > 0$ there exist infinitely many squarefree n for which Φ_n has less than

$$n^{\frac{8}{13}+\varepsilon}$$

non-zero coefficients (see (2.18)). This could be improved to $(8n)^{1/2}$ if it were known that for infinitely many primes p, one of $2p + 1$ and $2p - 1$ is prime. It is an interesting problem to construct squarefree integers n for which Φ_n has substantially fewer non-zero coefficients. A question which may be related is the following: do there exist numbers n, divisible by arbitrarily many distinct primes, for which Φ_n has only coefficients -1, 0,1?

Finally we mention some results on the behaviour of the i-th coefficient - i.e., the coefficient at X^i - of the cyclotomic polynomials, for fixed i. For squarefree n, it is clear from (1.7) that the i-th coefficient of Φ_n only depends on those primes $p \leq i$ which divide n, and on the parity of the total number of primes dividing n. In particular, the i-th coefficient can assume only finitely many values, and it is easily seen that this assertion remains valid if we drop the restriction that n should be squarefree.

LEHMER [12] has given a table of the i-th coefficient of Φ_n for $i \leq 10$ and n odd and squarefree, distinguishing 16 cases according to the value of $\mu(n)$ and the greatest common divisor of n and 105. His table implies that

for $i \leq 10$ the i-th coefficient is one of 1, 0, -1, except if $n = 105\, p_1 p_2 \cdots p_{2h}$ (p_i distinct primes > 7), in which case the 7-th coefficient equals -2. Compare also MÖLLER [16].

ERDÖS and VAUGHAN [7] proved that for all i the i-th coefficient of Φ_n is bounded in absolute value by

$$\exp(C_0 \cdot i^{1/2} + C_2 i^{3/8});$$

here C_2 is some constant, and $C_0 = 2 \cdot \prod_p (1 - \frac{2}{p(p+1)})^{\frac{1}{2}} \approx 1.373580$. On the other hand, they proved that for some constant $C_3 > 0$ and all sufficiently large i there exists n for which the i-th coefficient of Φ_n exceeds

$$\exp(C_3 (i/\log i)^{\frac{1}{2}})$$

in absolute value. VAUGHAN [19] proved that for infinitely many i this can be improved to

$$\exp(C_4 \cdot i^{1/2}/(\log i)^{1/4}).$$

Here C_4 denotes a positive constant.

2. Primitive relations between roots of unity.

Let $\{\zeta_1, \zeta_2, \ldots, \zeta_k\}$ be a set of k distinct roots of unity, $k > 0$, which is linearly dependent over $\mathbb{Q}$, while no proper subset is; *proper* means: not empty, and not the whole set. Then there is a relation

$$\sum_{i=1}^{k} \lambda_i \zeta_i = 0$$

(λ_i rational, not all zero), and this relation is uniquely determined up to a rational multiple. Multiplying by a common denominator we can make the λ_i into integers, and dividing by their greatest common divisor we arrive at a relation

$$\sum_{i=1}^{k} a_i \zeta_i = 0$$

in which the coefficients a_i are non-zero integers with greatest common

divisor 1. A linear relation which arises in this way is called a *primitive* relation. It is clear that if $\Sigma_{i=1}^{k} a_i'\zeta_i = 0$ is another primitive relation between the same ζ_i, then we have either $a_i' = a_i$ for all i, or $a_i' = -a_i$ for all i.

If we have $\Sigma_{i=1}^{k} a_i\zeta_i = 0$, and ρ is a root of unity, then we have also $\Sigma_{i=1}^{k} a_i(\rho\zeta_i) = 0$; two such relations are said to be *similar*. Clearly, any relation is similar to one with $\zeta_1 = 1$.

The *exponent* of a relation $\Sigma_{i=1}^{k} a_i\zeta_i = 0$ is the smallest integer $n > 0$ for which $\zeta_i^n = 1$ for all i, and the *reduced exponent* is the smallest n for which $(\zeta_i\zeta_j^{-1})^n = 1$ for all i,j. Notice that two similar relations have the same reduced exponent, and that in the case where $\zeta_1 = 1$ the reduced exponent coincides with the exponent.

If $\Phi_n = \Sigma\, c_i X^i$ is the n-th cyclotomic polynomial, and ζ is a primitive n-th root, then we have

$$(2.1) \qquad \sum_{i,c_i\neq 0} c_i\zeta^i = 0.$$

This is a primitive relation, since Φ_n has leading coefficient 1 and is irreducible over $\mathbb{Q}$. The reduced exponent of (2.1) is the product of the distinct primes dividing n; this follows from (1.4) and the fact that $c_0 \neq 0 \neq \neq c_1$ if n is squarefree (use (1.7)).

In this section we are interested in the number of terms k and the magnitudes of the coefficients a_i in a primitive relation of reduced exponent n. The results are much less complete than those known in the special case of the cyclotomic polynomial.

In (2.2) and (2.3) we describe the general technique for dealing with vanishing sums of roots of unity, cf. [15, 6].

THEOREM (2.2) *Let* m *be the product of the different primes dividing* n, *and let* ε,ζ *denote primitive* m*-th and* n*-th roots, respectively. Then* $\{\varepsilon^i\zeta^j: 0 \le i < m,\ 0 \le j < n/m\}$ *is the set of* n*-th roots, and*

$$\sum_{i=0}^{m-1} \sum_{j=0}^{(n/m)-1} a_{ij}\varepsilon^i\zeta^j = 0 \qquad (a_{ij} \in \mathbb{Z})$$

if and only if

$$\sum_{i=0}^{m-1} a_{ij}\varepsilon^i = 0$$

for every j, $0 \le j < n/m$.

This theorem readily follows from the irreducibility of $X^{n/m} - \zeta^{n/m}$ over the field $\mathbb{Q}(\varepsilon)$; to prove this irreducibility, just notice that $[\mathbb{Q}(\zeta):\mathbb{Q}(\varepsilon)] = \phi(n)/\phi(m) = n/m$. For details we refer to [15, 6].

Theorem (2.2) reduces the analysis of vanishing sums of n-th roots to the case that n is squarefree. It follows in particular, that the reduced exponent of a primitive relation is necessarily squarefree.

Relations of squarefree exponent n can be treated by induction on the number of primes dividing n, using the following theorem.

THEOREM (2.3) *Let* $n = pm$, *where* p *is prime and* p *does not divide* m, *and let* ε,ζ *denote primitive* m*-th and* p*-th roots, respectively. Then* $\{\varepsilon^i\zeta^j: 0 \le i < m, 0 \le j < p\}$ *is the set of* n*-th roots, and*

$$\sum_{i=0}^{m-1}\sum_{j=0}^{p-1} a_{ij}\varepsilon^i\zeta^j = 0 \qquad (a_{ij} \in \mathbb{Z}) \tag{2.4}$$

if and only if

$$\sum_{i=0}^{m-1} a_{ij}\varepsilon^i - \sum_{i=0}^{m-1} a_{i0}\varepsilon^i = 0 \tag{2.5}$$

for all j, $1 \le j < p$.

The proof of this theorem depends on the irreducibility of $X^{p-1} + \ldots + X^2 + X + 1$ over $\mathbb{Q}(\varepsilon)$, which is a consequence of $[\mathbb{Q}(\varepsilon,\zeta):\mathbb{Q}(\varepsilon)] = \phi(n)/\phi(m) = p - 1$. Compare with [15, 6].

If, in (2.4), there exists j' with $a_{ij'} = 0$ for all i, then (2.5) clearly yields

$$\sum_{i=0}^{m-1} a_{ij}\varepsilon^i = 0$$

for all j, $0 \le j < p$, which means that the vanishing sum (2.4) of n-th roots decomposes in vanishing sums of m-th roots. On the other hand, if for every j there exists i with $a_{ij} \neq 0$, then (2.4) has at least p non-zero terms. In particular, it follows that if $\Sigma_{i=1}^{k} a_i\zeta_i = 0$ is a primitive relation of reduced exponent n, then $k \ge p$, where p is the largest prime divid-

ing n. A more precise result is given by the following theorem, due to CONWAY and JONES [6]. In this theorem, we call a relation $\Sigma_{i=1}^{k} a_i\zeta_i = 0$ *minimal* if there is no proper subset $I \subset \{1,2,\ldots,k\}$ with $\sum_{i\in I} a_i\zeta_i = 0$; clearly, any primitive relation is minimal.

THEOREM (2.6) *If* $\Sigma_{i=1}^{k} a_i\zeta_i = 0$ *is a minimal relation of reduced exponent* n, *then* n *is squarefree, and*

$$k \geq \sum_{p|n} (p-2) + 2, \tag{2.7}$$

the sum ranging over the primes p *dividing* n. *Conversely, for every squarefree integer* n *there exists a minimal relation of reduced exponent* n *for which equality holds in* (2.7).

For the proof of this theorem we refer to [6]. Conway and Jones used (2.6) to classify all linear relations between roots of unity of less than 10 terms.

As is remarked in [6], one can deduce from (2.6) that for every $C > 1$ there exists C' such that

$$n \leq C'.\exp(C(k \log k)^{\frac{1}{2}})$$

for all n, k as in (2.6). It follows that

$$k \geq C_1 \cdot (\log n)^2 / \log\log n \qquad (n > 1) \tag{2.8}$$

for some positive constant C_1.

Various interesting theorems in elementary geometry have been proved by the use of the technique described in (2.2) and (2.3). An appropriate one to mention at this occasion is a result appearing in G. Bol's "Beantwoording van prijsvraag no. 17" [5]:

if n is odd, $n \geq 3$, then no three diagonals of a regular n-gon pass through one point, unless they have the same endpoint.

Let the n-gon have as its vertices the n-th roots of unity in the complex plane, and suppose that the diagonals $\alpha\beta$, $\gamma\delta$, $\varepsilon\zeta$ intersect in one point. For a complex number x to be on the line through α and β it is necessary and sufficient that

$$\frac{x-\alpha}{\beta-\alpha} = \frac{\bar{x}-\bar{\alpha}}{\bar{\beta}-\bar{\alpha}}$$

which by $\bar{\alpha} = \alpha^{-1}$, $\bar{\beta} = \beta^{-1}$ simplifies to

$$x + \alpha\beta\bar{x} = \alpha + \beta.$$

Hence, if x is on all three diagonals $\alpha\beta$, $\gamma\delta$, $\varepsilon\zeta$ we must have

$$(2.9) \qquad \begin{vmatrix} 1 & \alpha\beta & \alpha+\beta \\ 1 & \gamma\delta & \gamma+\delta \\ 1 & \varepsilon\zeta & \varepsilon+\zeta \end{vmatrix} = 0.$$

Working out the determinant we see that (2.9) is a vanishing sum of twelve roots of unity. This observation makes (2.2), (2.3) applicable, and after some work we arrive at Bol's result. For more applications of (2.2), (2.3) we refer to [6].

The following theorem gives a bound for the coefficients appearing in a primitive relation.

THEOREM (2.10) *Let* $\Sigma_{i=1}^{k} a_i\zeta_i = 0$ *be a primitive relation between* k *roots of unity. Then*

$$|a_i| \leq 2^{1-k} \cdot k^{k/2}$$

for i = 1,2,...,k.

In the proof of this theorem we denote by n the reduced exponent of the relation. We know that n is squarefree, and we may assume that the ζ_i are n-th roots.

LEMMA (2.11) [cf. 18]. *Let* n *be squarefree. Then for every* n*-th root* ζ *either* ζ *or* $-\zeta$ *is a sum of distinct primitive* n*-th roots. Further, the primitive* n*-th roots are linearly independent over* $\mathbb{Q}$.

PROOF OF (2.11) We first prove by induction on the number of primes dividing n that every n-th root ζ is plus or minus a sum of primitive ones. For n = 1 this is obvious. For n = p, the case $\zeta = 1$ is dealt with by

$$1 = -\sum \alpha$$

(α ranging over the primitive p-th roots), and in the case $\zeta \neq 1$ the representation

$$\zeta = \zeta$$

works. If $n \neq 1, p$ then we can write $n = \ell \cdot m$, with $\ell, m < n$, $(\ell, m) = 1$. Every n-th root ζ has a unique representation $\zeta = \eta\theta$, where η, θ are ℓ-th and m-th roots, respectively. By the induction hypothesis, we can write

$$\eta = \pm \sum \beta, \qquad \theta = \pm \sum \gamma,$$

where β ranges over a certain set of primitive ℓ-th roots and γ over a certain set of primitive m-th roots. Multiplying we find

$$\zeta = \pm \sum \beta\gamma.$$

Each term $\beta\gamma$ is a primitive n-th root, and no primitive n-th root occurs twice. This proves our assertion that every n-th root is $\pm$ a sum of primitive ones.

It follows that the $\phi(n)$ primitive n-th roots span the $\mathbb{Q}$-vector space generated by all n-th roots. But by the irreducibility of Φ_n this vector space has dimension $\phi(n)$. We conclude that the primitive n-th roots are linearly independent over $\mathbb{Q}$. In particular, for no n-th root ζ can both ζ and $-\zeta$ be written as a sum of distinct primitive n-th roots. This proves lemma (2.11).

Continuing the proof of the theorem, we write, using the lemma

$$\pm\zeta_i = \sum_{\alpha} e_{i\alpha}\alpha, \qquad 1 \leq i \leq k,$$

with α ranging over the primitive n-th roots and $e_{i\alpha} = 0$ or 1 for all i, α. By the primitivity of the relation $\Sigma a_i \zeta_i = 0$, the $k \times \phi(n)$-matrix $(e_{i\alpha})_{i,\alpha}$ has rank $k - 1$. Choose a $k \times (k-1)$-submatrix of rank $k - 1$. If $b_1, b_2, \ldots, b_k$ denote the $(k - 1) \times (k - 1)$ determinants of this submatrix in a suitable order, and provided with suitable signs, then

$$\sum_{i=1}^{k} b_i e_{i\alpha} = 0$$

for all α, so

$$\sum_{i=1}^{k} (\pm b_i)\zeta_i = 0.$$

Here the coefficients $\pm b_i$ are in $\mathbb{Z}$, and they do not all vanish. Since the relation $\sum_{i=1}^{k} a_i\zeta_i = 0$ is primitive it follows that $\pm b_i = ca_i$ for some non-zero integer c and all i, so

$$|a_i| \le |b_i|.$$

Thus, to finish the proof of the theorem it suffices to prove the following lemma.

LEMMA (2.12) *Let* $B = (\beta_{ij})$ *be a* $(k-1) \times (k-1)$*-matrix with* $\beta_{ij} = 0$ *or* 1 *for all* i,j, $1 \le i,j \le k-1$. *Then* $|\det B| \le 2^{1-k} \cdot k^{k/2}$.

PROOF. Define the $k \times k$-matrix $C = (\gamma_{ij})$ by

$$\gamma_{ij} = 2\beta_{ij} - 1 \qquad 1 \le i,j \le k-1,$$

$$\gamma_{kj} = -1 \qquad 1 \le j \le k-1,$$

$$\gamma_{ik} = 1 \qquad 1 \le i \le k.$$

By elementary column operations, $\det C = 2^{k-1} \cdot \det B$. Further $\gamma_{ij} = \pm 1$ for all i,j, so from Hadamard's inequality

$$|\det(\gamma_{ij})| \le \prod_{i=1}^{k} \left(\sum_{j=1}^{k} \gamma_{ij}^2\right)^{\frac{1}{2}}$$

we get

$$|\det C| \le k^{k/2},$$

$$|\det B| = |2^{1-k} \cdot \det C| \le 2^{1-k} \cdot k^{k/2}.$$

This proves (2.12) and (2.10).

It is not known whether theorem (2.10) is best possible.

If $n = p$ is prime, then the only primitive relation of exponent n is

$$\pm\sum\zeta = 0,$$

ζ ranging over all n-th roots. In the case $n = pq$, p and q distinct primes, all primitive relations have been determined by MANN [15]:

THEOREM (2.13) *Let* p *and* q *be primes,* $p \neq q$, *and let* A,A',B,B' *be non-empty sets of roots of unity such that*

$$A \cup A' = \{\textit{all } p\textit{-th roots}\}, \qquad A \cap A' = \emptyset,$$

$$B \cup B' = \{\textit{all } q\textit{-th roots}\}, \qquad B \cap B' = \emptyset.$$

Then

$$\sum_{\alpha\in A}\sum_{\beta\in B}\alpha\beta - \sum_{\alpha\in A'}\sum_{\beta\in B'}\alpha\beta = 0.$$

This is a primitive relation of reduced exponent pq, *and every primitive relation of reduced exponent* pq *is similar to one of this form.*

For the proof, which is a direct application of (2.3), we refer to [15].

Theorem (2.13) suggests a representation for Φ_{pq} which is different from the one we have seen in section 1. Let $\Phi_{pq} = \sum^{pq} c_i x^i$, and let ζ be a primitive pq-th root. Then $\sum_{i,c_i\neq 0} c_i\zeta^i = 0$ is a primitive relation of reduced exponent pq, and one may wonder which sets A,A',B,B' correspond to this relation. A few trials suggest that one should take

$$A = \{\zeta^{jq}: 0 \leq j < \mu\}, \qquad A' = \{\zeta^{jq}: \mu \leq j < p\}, \tag{2.14}$$

$$B = \{\zeta^{ip}: 0 \leq i < \lambda\}, \qquad B' = \{\zeta^{ip}: \lambda \leq i < q\}, \tag{2.15}$$

where the integers λ,μ are determined by

$$\lambda p \equiv 1 \bmod q, \qquad 0 < \lambda < q,$$

$$\mu q \equiv 1 \bmod p, \qquad 0 < \mu < p.$$

Notice that $\lambda p + \mu q = 1 + pq$, since $\lambda p + \mu q \equiv 1 \bmod pq$, $1 < \lambda p + \mu q < 2pq$. Thus, the choice (2.14), (2.15) for A,A',B,B' is correct if and only if

$$\Phi_{pq} = \sum_{i=0}^{\lambda-1} \sum_{j=0}^{\mu-1} x^{ip+jq} - \sum_{i=\lambda}^{q-1} \sum_{j=\mu}^{p-1} x^{ip+jq-pq}. \tag{2.16}$$

Once discovered, this formula is easily verified: the right hand side equals

$$\frac{(1-x^{\lambda p})(1-x^{\mu q})}{(1-x^{p})(1-x^{q})} - \frac{(x^{\lambda p}-x^{pq})(x^{\mu q}-x^{pq})x^{-pq}}{(1-x^{p})(1-x^{q})}$$

and this simplifies to

$$\frac{(1-x)(1-x^{pq})}{(1-x^{p})(1-x^{q})},$$

which is Φ_{pq}, by (1.3).

From (2.16) one sees that the number of non-zero coefficients of Φ_{pq} equals $2\lambda\mu - 1$. In the case $q \equiv 1 \bmod p$ we have $\mu = 1$, $\lambda = ((p-1)q+1)/p$, so

$$2\lambda\mu - 1 = \frac{2(p-1)(q-1)}{p} + 1. \tag{2.17}$$

HOOLEY [10] has shown that for every $\varepsilon > 0$ there exist infinitely many primes q for which $q-1$ has a prime divisor p with $p > q^{(5/8)-\varepsilon}$. Putting $n = pq$ and using (2.17) we find that for every $\varepsilon > 0$ there are infinitely many squarefree n for which Φ_n has less than

$$n^{(8/13)+\varepsilon} \tag{2.18}$$

non-zero coefficients. This confirms a remark made in section 1. If $q = 2p \pm 1$, then one obtains in the same way less than $(8n)^{\frac{1}{2}}$ non-zero coefficients, with $n = pq$. It is unknown, however, whether for infinitely many primes p one of $2p+1$ and $2p-1$ is prime.

Theorem (2.13) implies that primitive relations of reduced exponent pq have no coefficients other than ± 1. Combining this observation with theorem (1.9) one is led to the following question:

> does there exist a function f on the positive integers, such that for all m, and all primes p,q with $p \neq q$, $(pq,m) = 1$, and all primitive relations $\sum_{i=1}^{k} a_i \zeta_i = 0$ of reduced exponent mpq, the coefficients a_i are bounded in absolute value by $f(m)$?

I do not know the answer to this question. Theorem (2.19) gives a partial

result.

THEOREM (2.19) *There exists a function* f *on the positive integers such that for all* m, *all primes* p *not dividing* m, *and all primitive relations* $\Sigma_{i=1}^{k} a_i \zeta_i = 0$ *of reduced exponent* mp, *the coefficients* a_i *are less than* f(m) *in absolute value.*

PROOF. Let p be an odd prime not dividing m, let R be the set of p-th roots, and let B be the set of m-th roots. Any mp-th root α has a unique expression as $\alpha = \beta\rho$, with $\beta \in B$, $\rho \in R$, so the given primitive relation is similar to one of the form

$$\sum_{\rho \in R} \sum_{\beta \in B(\rho)} a_{\beta\rho}\beta\rho = 0 \tag{2.20}$$

where $B(\rho) \subset B$ for each $\rho \in R$, and all $a_{\beta\rho} \neq 0$. Using (2.3) we find that

$$\sum_{\beta \in B(\rho)} a_{\beta\rho}\beta = \sum_{\beta \in B(\rho')} a_{\beta\rho'}\beta \tag{2.21}$$

for any two $\rho,\rho' \in R$. Thus, if some $B(\rho')$ were empty, then all these sums would vanish, contradicting that (2.20) is a primitive relation of reduced exponent mp. We conclude that the $B(\rho)$ are non-empty. Next we claim that

$$B(\sigma) = B(\sigma') \Rightarrow a_{\beta\sigma} = a_{\beta\sigma'} \quad \text{for all } \beta \in B(\sigma) \tag{2.22}$$

$(\sigma,\sigma' \in R)$. In fact, if this would not be true, then by putting

$$c_{\beta\rho} = a_{\beta\rho} \quad \text{if } \rho \in R,\ \rho \neq \sigma,\sigma',\ \beta \in B(\rho),$$

$$c_{\beta\sigma} = a_{\beta\sigma'} \quad \text{if } \beta \in B(\sigma)$$

$$c_{\beta\sigma'} = a_{\beta\sigma} \quad \text{if } \beta \in B(\sigma)$$

we would get a relation

$$\sum_{\rho \in R} \sum_{\beta \in B(\rho)} c_{\beta\rho}\beta\rho = 0,$$

which is not plus or minus the original relation (2.20) (here we use $p \geq 3$), contradicting the primitivity.

Now let q be the smallest prime larger than 2^m, and let T be the set of q-th roots. The number of different sets $B(\rho)$, $\rho \in R$, is clearly less than q, so we are able to choose, for every $\tau \in T$, a subset $C(\tau) \subset B$ such that

$$\{C(\tau): \tau \in T\} = \{B(\rho): \rho \in R\}.$$

Define $b_{\beta\tau}$ for $\tau \in T$, $\beta \in C(\tau)$ by

$$b_{\beta\tau} = a_{\beta\sigma}$$

where $\sigma \in R$ is chosen such that $B(\sigma) = C(\tau)$; by (2.22) this definition does not depend on the choice of σ. By (2.21) we have

$$\sum_{\beta \in C(\tau)} b_{\beta\tau}\beta = \sum_{\beta \in C(\tau')} b_{\beta\tau'}\beta$$

for all $\tau,\tau' \in T$, so

$$\sum_{\tau \in T} \sum_{\beta \in C(\tau)} b_{\beta\tau}\beta\tau = 0. \tag{2.23}$$

We claim that this is a primitive relation between mq-th roots of unity. Obviously the coefficients $b_{\beta\tau}$ have greatest common divisor 1, so if (2.23) is not primitive than there exist subsets $D(\tau) \subset C(\tau)$, not all empty, and not all $D(\tau) = C(\tau)$, such that $\{\beta\tau: \tau \in T, \beta \in D(\tau)\}$ is linearly dependent over $\mathbb{Q}$. Reversal of the above procedure would then, as the reader readily checks, give rise to subsets $E(\rho) \subset B(\rho)$, not all empty, and not all $E(\rho) = B(\rho)$, such that also $\{\beta\rho: \rho \in R, \beta \in E(\rho)\}$ is linearly dependent over $\mathbb{Q}$, and this would contradict that (2.20) is primitive.

Thus we have proved that any coefficient appearing in a primitive relation of reduced exponent mp, with p an odd prime not dividing m, appears in a primitive relation between mq-th roots. But q depends only on m, and there are only finitely many primitive relations of given exponent. Hence there are only finitely many coefficients, and this conclusion remains unaffected if we also allow p = 2. This proves theorem (2.19).

REFERENCES

[1] APOSTOL, T.M., *The resultant of the cyclotomic polynomials* $F_m(ax)$ *and* $F_n(bx)$, Math. Comp. 29 (1975), p. 1-6.

[2] BATEMAN, P.T., *Note on the coefficients of the cyclotomic polynomial,* Bull. Amer. Math. Soc. 55 (1949), p. 1180-1181.

[3] BEITER, M., *Magnitude of the coefficients of the cyclotomic polynomial* F_{pqr}, II, Duke Math. J. 38 (1971), p. 591-594.

[4] BLOOM, D.M., *On the coefficients of the cyclotomic polynomials,* Amer. Math. Monthly 75 (1968), p. 372-377.

[5] BOL, G., *Beantwoording van prijsvraag no.* 17, Nieuw Archief voor Wiskunde (2), 18 (1936), p. 14-66; cf. Zentralblatt 237 #50008, 244 #50009.

[6] CONWAY, J.H. & A.J. JONES, *Trigonometric diophantine equations (On vanishing sums of roots of unity),* Acta. Arith. 30 (1976), p. 229-240.

[7] ERDÖS, P. & R.C. VAUGHAN, *Bounds for the r-th coefficients of cyclotomic polynomials,* J. London Math. Soc. (2), 8 (1974), p. 393-401.

[8] FELSCH, V. & E. SCHMIDT, *Über Perioden in den Koeffizienten der Kreisteilungspolynome* $F_{np}(x)$, Math. Z. 106 (1968), p. 267-272.

[9] HARDY, G.H. & E.M. WRIGHT, *An introduction to the theory of numbers,* fourth edition, Oxford University Press 1960.

[10] HOOLEY, C., *On the largest prime factor of* $p+a$, Mathematika 20 (1973), p. 135-143.

[11] JUSTIN, J., *Bornes des coefficients du polynôme cyclotomique et de certains autres polynômes,* C.R. Acad. Sci. Paris 268 (1969), Sér. A, p. 995-997.

[12] LEHMER, D.H., *Some properties of the cyclotomic polynomial,* J. Math. Anal. Appl. 15 (1966), p. 105-117.

[13] LEHMER, E., *On the magnitude of the coefficients of the cyclotomic polynomial,* Bull. Amer. Math. Soc. 42 (1936), 389-392.

[14] LeVEQUE, W.J., *Reviews in number theory,* vol. I, Amer. Math. Soc., 1974 .

[15] MANN, H.B., *On linear relations between roots of unity,* Mathematika 12 (1965), p. 107-117.

[16] MÖLLER, H., *Über die* i*-ten Koeffizienten der Kreisteilungspolynome,* Math. Ann. 188 (1970), p. 26-38.

[17] MÖLLER, H., *Über die Koeffizienten des* n*-ten Kreisteilungspolynoms,* Math. Z. 119 (1971), p. 33-40.

[18] RÉDEI, L., *Natürliche Basen des Kreisteilungskörpers,* I, Abh. Math. Sem. Univ. Hamburg 23 (1959), p. 180-200; id., II, ibid., 24 (1960), p. 12-40.

[19] VAUGHAN, R.C., *Bounds for the coefficients of cyclotomic polynomials,* Michigan Math. J. 21 (1975), p. 289-295.

For more references to the literature about cyclotomic polynomials, one should consult [1] and [14, pp. 404-411].

ON QUARTIC SURFACES IN PROJECTIVE 3-SPACE

E.J.N. Looijenga

A nonzero homogeneous polynomial of degree four in four complex variables defines a surface in $\mathbb{CP}^3$. Two such polynomials determine the same surface (counting multiplicities) if and only if one is a scalar multiple of the other. Thus the quartic surfaces in $\mathbb{CP}^3$ are parametrised by a projective space $\mathbb{P}$ of dimension $\binom{7}{4} - 1 = 34$. A Zariski open subset $\mathbb{P}_{ns} \subset \mathbb{P}$ describes the nonsingular quartic surfaces. It is the base of an algebraic fibre bundle $\pi\colon \xi \to \mathbb{P}_{ns}$ whose fibre over $s \in \mathbb{P}_{ns}$ is just the quartic surface X_s it defines.

As $\mathbb{P}_{ns}$ is connected, the C^∞-type of the general fibre is well defined. It is not hard to show that this fibre is simply connected and that is second integral cohomology group is a free $\mathbb{Z}$-module of rank 22. Endowed with the intersection form $\langle , \rangle$ this $\mathbb{Z}$-module becomes a unimodular lattice of signature (3,19). If a polynomial $f \in \mathbb{C}[x,y,z]$ defines (an affine part of) a nonsingular quartic X, then it is easy to verify that the meromorphic 2-form $\left(\frac{\partial f}{\partial x}\right)^{-1} dydz$ restricts to X as a 2-form which has neither poles nor zeroes. If we denote this form by ω_X, then its de Rham cohomology class $[\omega_X]$ satisfies the familiar relations $\langle [\omega_X],[\omega_X]\rangle = 0$ and $\langle [\omega_X],[\bar{\omega}_X]\rangle > 0$. Any other holomorphic 2-form on X is a holomorphic function times ω_X and (as the only holomorphic functions on X are the constants) hence a scalar multiple of ω_X. So $[\omega_X]$ spans $H^{2,0}(X,\mathbb{C})$. Besides this one-dimensional subspace we distinguish the cohomology class $h_X \in H^2(X,\mathbb{Z})$ which is supported by the intersection of X with a "general" hyperplane in $\mathbb{CP}^3$. We refer to h_X as the *polarisation* of X. For elementary geometric reasons we have $\langle h_X,h_X\rangle = 4$ and $\langle \omega_X,h_X\rangle = 0$.

Our first aim is to parametrise the quartic surfaces up to isomorphism. For this purpose, the bundle π is not very appropriate. For the group $PGL_4(\mathbb{C})$ acts in a compatible way on ξ and $\mathbb{P}_{ns}$ and an orbit in $\mathbb{P}_{ns}$ describes surfaces which are projectively equivalent. Conversely, an algebraic

isomorphism between two nonsingular quartic surfaces which preserves the polarisations is the restriction of some projective automorphism. Hence the orbit space $\mathbb{P}_{ns}/PGL_4(\mathbb{C})$ parametrises the polarised nonsingular surfaces in an effective manner. However, not all the fibres of the induced map $\xi/PGL_4(\mathbb{C}) \longrightarrow \mathbb{P}_{ns}/PGL_4(\mathbb{C})$ are quartic surfaces as each quartic surface is mapped to the orbit space of its automorphism group. To remedy this, one considers quartic surfaces with some additional structure such that each nontrivial automorphism affects this structure. In this case we proceed as follows. Fix (once and for all) a lattice L isometric to the second cohomology lattice of a nonsingular quartic and an indivisible element $h \in L$ with $\langle h,h\rangle = 4$. Then a *marking* of a nonsingular quartic X will be an isometry $\phi: H_2(X,\mathbb{Z}) \longrightarrow L$ which sends h_X to h. Each nonsingular quartic can be marked and the possible markings of a given quartic are permuted in a simple transitive manner by the group

$$G := \{g \in \mathrm{Aut}(L): g \text{ preserves } \langle,\rangle \text{ and } h\}.$$

Now it can be shown that the automorphism group of a marked quartic X is trivial; in other words, the automorphism group of X acts faithfully on $H^2(X,\mathbb{Z})$. The marked quartics are described by an analytic fibre bundle $\tilde{\pi}: \tilde{\xi} \to \tilde{\mathbb{P}}_{ns}$ covering π, which is canonically endowed with a $G \times PGL_4(\mathbb{C})$-action. The actions of G and $PGL_4(\mathbb{C})$ are free, but the product action of $G \times PGL_4(\mathbb{C})$ is not. Now let $\tilde{p}: \tilde{X} \to \tilde{S}$ denote the induced map between the $PGL_4(\mathbb{C})$-orbit spaces. Then $\tilde{X}$ and $\tilde{S}$ are both nonsingular with $\dim \tilde{S} = \dim \mathbb{P}_{ns} - \dim PGL_4(\mathbb{C}) = 19$ and each fibre of $\tilde{p}$ is a marked quartic. Note that $\tilde{S}/G$ and $\mathbb{P}_{ns}/PGL_4(\mathbb{C})$ are canonical isomorphic.

Next we wish to have a better understanding of the G-manifold $\tilde{S}$. For this, the most powerful tool we have at our disposal is the so-called *period mapping* which I shall now describe.

Consider the image Ω of the set

$$\{\omega \in L_{\mathbb{C}}: \langle\omega,\omega\rangle = 0,\ \langle\omega,\bar{\omega}\rangle > 0,\ \langle\omega,h\rangle = 0\}$$

in the projective space of $L_{\mathbb{C}}$. This is clearly an open subset of a nonsingular quatric in a projective space of dimension 20. Moreover, the real Lie group

$$G_{\mathbb{R}} := \{g \in \mathrm{Aut}(L_{\mathbb{R}}): g \text{ preserves } \langle,\rangle \text{ and } h\}$$

acts in a transitive manner on Ω and realizes each of the (two) components of Ω as a bounded symmetric domain. We have a canonical G-equivariant map (the period mapping) $\tilde{P}: \tilde{S} \to \Omega$ which assigns to the marked quartic (X,ϕ) the

point in Ω defined by $\phi(H^{2,0}(X,\mathbb{C}))$. In the fifties Kodaira proved that $\tilde{P}$ satisfies the local Torelli theorem, i.e. that $\tilde{P}$ is a local isomorphism (note that dim $\tilde{S}$ = 19 = dim Ω, indeed). Around 1970 it was shown by Shafarevic and Piatetskii-Shapiro that the global Torelli theorem also holds, i.e. that $\tilde{P}$ is actually an injection. So what remained was the determination of the image of $\tilde{P}$. This was implicitly done by Shah in his thesis (1973), who succeeded in finding for each $\zeta \in \Omega$ a surface "over it" which belongs to the same family as the quartics. In order to describe his result more precisely, we need the (at first sight perhaps somewhat complicated) notion of a pseudo-polarised K3 surface of degree four: a *K3 surface* is a nonsingular connected compact surface which is simply connected and possesses a nowhere vanishing holomorphic 2-form. An element $h_X \in H^2(X,\mathbb{Z})$, where X is a K3 surface say, is called a *pseudo-polarisation* if

(i) it is the class of a divisor,

(ii) for any positive divisor D we have $\langle [D], h_X \rangle \geq 0$, and

(iii) $\langle h_X, h_X \rangle > 0$.

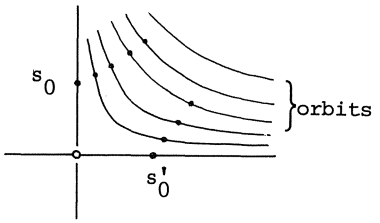
The number $\langle h_X, h_X \rangle$ is called the *degree* of the pseudo-polarisation. So any nonsingular quartic is in a canonical way a pseudo-polarised K3 surface of degree four.

Now observe that the global Torelli theorem implies that the composite map

$$P := \tilde{P}/G : \mathbb{P}_{ns}/PGL_4(\mathbb{C}) \cong \tilde{S}/G \longrightarrow \Omega/G$$

is injective. A rather general theory, due to Baily and Borel, asserts that the orbit space is in a natural way a quasi-projective variety. It can be shown that this makes of P an algebraic map. Since $\tilde{P}$ is a local isomorphism, it then follows that the image of P is Zariski-open in Ω/G. Suppose we are given a $\zeta_0 \in \Omega/G$. Then there is a sequence $\{\bar{s}_i \in \mathbb{P}_{ns}/PGL_4(\mathbb{C})\}_{i=1}^{\infty}$ such that $P(\bar{s}_i)$ converges to ζ_0. If we choose a representative sequence $\{s_i \in \mathbb{P}_{ns}\}_{i=1}^{\infty}$, then, after passing to a subsequence if necessary, it will converge to a $s_0 \in \mathbb{P}$. Clearly, there is a lot of ambiguity here for if we choose a sequence $\sigma_i \in PGL_4(\mathbb{C})$, then a limit post of $\{\sigma_i(s_i)\}$ is in general not in the same orbit as s_0 (see the example pictured below)

However, Mumford developed a theory which, roughly stated, tells you what sequences to choose. The quartics which you get if you follow Mumford's prescription have been classified by Shah. They fall into three classes:



1. The nonsingular ones.
   These require no further discussion.

2. Quartics with only rational double points as singularities.
   These are normal singular points, characterised by the property that after a minimal resolution the exceptional locus is a union of nonsingular rational curves, all having selfintersection -2. This forces the intersection matrix of these curves to be the Cartan matrix of a root system of type A, D or E. The minimal resolution of a quartic with only such singularities is in a natural way a pseudo polarised K3 surface of degree four. The set H of points in Ω corresponding to such surface is not hard to describe. For any $\ell \in L$, put $H(\ell) := \{\omega \in \Omega, \langle\omega,\ell\rangle = 0\}$ and let Σ denote the set of $\ell \in L$ with $\langle\ell,h\rangle = 0$ and $\langle\ell,\ell\rangle = -2$. Then H is open and dense in $U\{H(\ell): \ell \in \Sigma\}$. At each of its points this set is like the union of reflection hyperplanes of a complexified root system (with each irreducible component of type A, D or E and determining a rational double point of the same type on the corresponding limiting quartic).

3. The nonsingular quadrics with multiplicity two.
   This looks bad, since all such quadrics are projectively equivalent. The reason is that have completely ignored the way this quadric arises as a limit: we must not only give $s_0 \in \mathbb{P}$ but also a direction in the tangent space of $\mathbb{P}$ at s_0 (as being the limiting position of the lines s_0s_i). Geometrically, this amounts to giving a quadric Q and a curve C on Q which is the intersection of Q with a quartic surface. "In general" C will be nonsingular and then there is a two-fold cover $X(Q,C)$ of Q branched along C which is nonsingular. It can be shown that this is in a natural way a pseudo-polarised K3 surface of degree four. (Actually, the curve C is allowed to have "sinple" singularities, which yield rational double points on the branched cover.) If we let Σ' denote the set of $\ell \in L$ with $\langle\ell,\ell\rangle = 0$ and $\langle\ell,h\rangle = 2$, then the set of points in Ω corresponding to the double quadric are open-dense in $U\{H(\ell): \ell \in \Sigma'\}$.

4. The remaining cases.
   These are all a mixture of the cases 2 and 3. They correspond to points in Ω lying on an intersection $H(\ell) \cap H(\ell')$ with $\ell \in \Sigma$ and $\ell' \in \Sigma'$.

As a result we get a very nice description of both $\tilde{S}$ and $\tilde{S}/G$. For it follows that the period mapping determines an isomorphism

$$S \cong \Omega - U\{H(\ell): \ell \in \Sigma \cup \Sigma'\}$$

(G equivalently) and hence also an isomorphism

$$\left\{\begin{array}{l}\text{(projective) isomorphism}\\ \text{classes of nonsingular}\\ \text{quartics in } \mathbb{C}P^3\end{array}\right\} \cong \tilde{S}/G \xrightarrow{\cong} (\Omega - U\{H(\ell): \ell \in \Sigma \cup \Sigma'\})/G.$$

It goes without saying that (to a geometer, say) the right-hand side looks much more manageable than the left-hand side. By the way, there is some reason to believe that the space $\Omega - U\{H(\ell): \ell \in \Sigma \cup \Sigma'\}$ (or equivalently, its G-orbit space) is a $\kappa(\pi,1)$, in other words, has a contractible universal covering.

Actually, Shah classifies *all* the Mumfordian limits of quartics (not only those corresponding to points of Ω/G). The ones which have only isolated singular points are easy to describe: these are the quartics with only simple-elliptic and hyperbolic singularities. A normal surface singularity is called *simple-elliptic* if the exceptional locus of its minimal resolution is a smooth elliptic curve. The hyperbolic singularites are most conveniently characterised by giving a normal form for their equations: $x^p + y^q + z^r + xyz$ with $\frac{1}{p} + \frac{1}{q} + \frac{1}{r} < 1$. Quartics with such singularities must be included if one wants to construct an algebraic compactification of Ω/G which is geometrically meaningful. To illustrate this point, let $s \in \mathbb{P}$ represent a quartic X_s with only simple-elliptic and hyperbolic singularities. The group of projective automorphisms of X_s is finite; for the sake of exposition let us assume that it is trivial. Now, choose a transversal slice $i: \Delta(r) \longrightarrow \mathbb{P}$ with $i(0) = s$, so the $PGL_4(\mathbb{C})$-orbit of s, where $\Delta(r) = \{z \in \mathbb{C}^{19}: |z| < r\}$. We further put $\Delta(r)_{ns} := i^{-1}\mathbb{P}_{ns}$. Then, if r is small enough, the composed mapping

$$\Delta(r)_{ns} \xhookrightarrow{i} \mathbb{P}_{ns} \longrightarrow \mathbb{P}_{ns}/PGL_4(\mathbb{C}) \cong \tilde{S}/G \xrightarrow{P/G} \Omega/G$$

will be an analytic isomorphism onto an open subset of Ω/G. Therefore, with a bit of imagination, the diagram $\Delta(r) \supset \Delta(r)_{ns} \hookrightarrow \Omega/G$ may be viewed as a partial compactification of Ω/G. In case X_s has only simple-elliptic singularities this partial compactification admits quite an explicit description in terms of Ω and G only, which, as Mumford observed, is of Baily-Borel type. In the presence of hyperbolic singularities this partial compactification can be made explicit too (see my Helsinki talk); this one is neither of Baily-Borel type nor of Mumford's toroidal type (although probably related to it). It is noteworthy that the partial compactification of Ω/G thus

obtained is smooth *and* canonical, two virtues which are seldom seen together in this field. Let me finish with the (perhaps somewhat cryptic) remark that the deformation theory of the simple-elliptic and hyperbolic singularities can be very well understood by knowing how the period mapping degenerates near a quartic with these singularities. This was actually the whole point I wanted to make.

AUGUSTIN CAUCHY: EIN WENDEPUNKT IN DER GESCHICHTE DER ANALYSIS

E. Neuenschwander

EINLEITUNG

Die gesammelten Werke von Cauchy umfassen 27 dicke Bände. Eine kritische Würdigung seines Gesamtwerkes überschreitet somit den Rahmen dieses Vortrages. Im I. Teil wird versucht, Cauchy als Mensch und Mathematiker aufgrund zeitgenössischer Dokumente etwas näher zu charakterisieren; im II. Teil soll seine Bedeutung für die Geschichte der Analysis herausgearbeitet werden.

I. CAUCHYS LEBEN UND WERK

Augustin-Louis Cauchy wurde am 21. August 1789 in Paris geboren[1)]. Er stammte aus einer Familie des oberen Mittelstandes. Sein Vater hatte die alten Sprachen an der Universität Paris studiert und darauf eine Beamtenlaufbahn gewählt, wobei er im Jahre 1800 Generalsekretär des Senates wurde und dadurch in Kontakt mit Laplace und Lagrange kam. Die Mutter von Cauchy war sehr religiös. Augustin, das älteste von sechs Geschwistern, erhielt seine erste Ausbildung zusammen mit seinem um drei Jahre jüngeren Bruder Alexander im Kreise der Familie. Der Vater brachte ihnen damals die Grundlagen der Grammatik und der alten Sprachen bei, wartete aber mit der Mathematik noch zu. Lagrange hatte nämlich die mathematische Begabung von Augustin frühzeitig erkannt und den Vater gewarnt, dem Sohn vor der Vollendung des 17. Altersjahres ein Mathematikbuch in die Hand zu geben, da er sonst zwar Mathematik lernen, jedoch nicht einmal seine eigene Sprache schreiben können würde[2)]. Als die Erziehung zu Hause abgeschlossen war, trat Cauchy in die

---

*) *Der vorliegende Vortrag wurde während eines längeren Studienaufenthaltes in den Niederlanden im Frühjahr 1978 gehalten. Wir möchten bei dieser Gelegenheit dem Mathematischen Institut der Rijksuniversiteit Utrecht nochmals recht herzlich für die erwiesene Gastfreundschaft danken.*

École centrale du Panthéon ein, wo er als einer der besten Schüler sich verschiedene Auszeichnungen in Griechisch und Latein holte. Anschliessend wurde er von Professor Dinet für den Eintritt in die École Polytechnique vorbereitet, welcher ihm im Jahre 1805 nach einer Vorbereitungszeit von nur acht bis zehn Monaten glückte. Vom Jahre 1807 an besuchte er noch die École des Ponts et Chaussées.

Der junge Ingenieur Cauchy arbeitete zunächst am Canal de l'Ourcq und an der Brücke von Saint-Cloud. Später wurde er nach Cherbourg versetzt, da Napoleon hier einen neuen Hafen als Basis für die geplante Invasion von England errichten wollte. Cauchys Biograph Valson berichtet, dass Cauchy im März 1810 nach Cherbourg reiste. Im Koffer hatte er die *Mécanique céleste* von Laplace, den *Traité des Fonctions analytiques* von Lagrange, den "Vergil" sowie die *De imitatione Christi* von Thomas a Kempis[3]. Die Religion spielte eine wichtige Rolle im Leben von Cauchy, wie wir später sehen werden. In Cherbourg wartete auf den jungen Cauchy ziemlich viel Arbeit. Daneben fand er dennoch Zeit, seine mathematischen Kenntnisse zu vertiefen. Das Ziel seiner Bemühungen formulierte er in einem Brief an die Eltern folgendermassen[4]:

> *... de repasser par une étude suivie toutes les branches des Mathématiques, en commençant par l'Arithmétique et finissant par l'Astronomie; éclaircissant de son mieux les endroits obscurs, s'appliquant à simplifier les démonstrations et à découvrir des propositions nouvelles.*

Auf die Cherbourger Zeit gehen die ersten mathematischen Arbeiten von Cauchy zurück. Dazu gehören zwei Abhandlungen über Polyeder aus den Jahren 1812 und 1813, in denen Cauchy unter anderem beweist, dass ein konvexes Polyeder durch die Angabe seiner Seitenflächen bis auf Kongruenz bestimmt ist[5]. Es folgen Arbeiten über die Zahlentheorie und die Gleichungstheorie[6].

Im Jahre 1813 kehrte Cauchy aus gesundheitlichen Gründen nach Paris zurück, wo er 1814 sein berühmtes *Mémoire sur les intégrales définies*[7] verfasste. Im nächsten Jahr gewann er den grossen Preis für Mathematik der Académie des Sciences mit seiner Schrift *Théorie de la propagation des ondes à la surface d'un fluide pesant d'une profondeur indéfinie*[8], die Grundlage für seine spätere Elastizitätstheorie wurde. Schon in den ersten Arbeiten von Cauchy offenbarte sich somit die ganze Tiefe und Breite seines Geistes.

Bereits im Jahre 1813, im Alter von nur 24 Jahren, stand Cauchy erstmals auf der Kandidatenliste für die Académie des Sciences. Er hatte jedoch kein Glück, da Poinsot ihm damals vorgezogen wurde. Im nächsten Jahr fand sich sein Name erneut auf der Liste, wobei seine Kandidatur diesmal unterstützt wurde von Cuvier und Laplace, wie man aus Briefen ersehen kann[9].

Cauchy war jedoch abermals erfolglos. Sein Biograph Valson bemerkt, bei der Wahl hätten eben andere Dinge als die Verdienste eine ausschlaggebende Rolle gespielt. Cauchy kam dann zwei Jahre später dennoch in die Akademie, wenn auch nicht gerade unter rühmlichen Umständen. Nach dem Zusammenbruch des Kaisertums wurde die Akademie im Jahre 1816 reorganisiert; dabei wurden Carnot und Monge aus politischen Gründen ausgeschlossen und an ihrer Stelle Cauchy und Bréguet ernannt. Dass Cauchy die Ernennung angenommen hatte, wurde ihm von vielen zeitgenössischen Gelehrten verübelt. Cauchy hatte Monge jedoch vermutlich nicht persönlich gekannt und war zweifellos der Ansicht, dass ihm selbst ein Platz in der Akademie seit langem zustehe[10].

Im Jahre 1816 wurde Cauchy Professor an der École polytechnique und einige Jahre später auch an der Sorbonne sowie am Collège de France. 1818 heiratete er Aloïse de Bure. Bis zum Jahre 1830 verlief nun Cauchys Leben ziemlich ungestört. Als Folge seiner Vorlesungstätigkeit entstanden seine berühmten Textbücher: *Cours d'analyse de l'École Royale Polytechnique* (1821)[11], *Résumé des leçons données à l'École Royale Polytechnique, sur le calcul infinitésimal* (1823)[12] usw., die von einer bis dahin unbekannten Strenge sind. In denselben Zeitabschnitt fällt sein *Mémoire sur les intégrales définies, prises entre des limites imaginaires* (1825)[13], das heute von vielen als seine beste Arbeit betrachtet wird.

Im Juli 1830 wird der Bourbonenkönig Karl X. gestürzt und Louis-Philippe von Orléans kommt auf den Thron. Bei dieser Gelegenheit sollen sämtliche Staatsbeamten einen neuen Treueeid ablegen. Cauchy verweigert den Eid, da er einerseits als strenggläubiger Katholik seinen alten Eid nicht brechen will und andererseits sowieso nur die katholischen Bourbonen als rechtmässige Herrscher betrachtet[14]. In der Folge verliert er seine drei Lehrstühle[15] und geht freiwillig ins Exil, zunächst nach Fribourg zu den Jesuiten. Hier versucht er, was vielleicht weniger bekannt sein dürfte, in Zusammenarbeit mit anderen aus Frankreich geflohenen Wissenschaftlern eine Helvetische Akademie der Wissenschaften zu gründen[16], an der auf katholischer und königstreuer Basis sämtliche Fächer gelehrt werden sollen. Um die notwendigen finanziellen Mittel zu bekommen, schreibt Cauchy an den Papst, den Zaren, den österreichischen Kaiser, sucht in Genua den König von Sardinien auf und in Modena den dortigen Herzog. Weiter druckt er einen Werbeprospekt, um nach Gründungsmitgliedern zu suchen, die pro Jahr 1000 Fr. bezahlen müssten und dafür in den Annalen der Akademie verewigt würden. Offenbar meldeten sich zu wenige, denn die Akademie kam aus Geldmangel nie zustande.

Durch die Vermittlung der Jesuiten erhält Cauchy schliesslich einen Lehrstuhl in Turin, wo er bis ins Jahr 1833 bleibt[17]. In die Turiner Zeit fällt das berühmte *Mémoire sur la mécanique céleste et sur un nouveau calcul appelé calcul des limites* (1831-33)[18], in dem man die Cauchysche Integralformel und die Potenzreihenentwicklung einer analytischen Funktion findet. Von Turin wird Cauchy von dem aus Frankreich geflohenen Bourbonenkönig nach Prag gerufen, um bei der Erziehung des Kronprinzen mitzuhelfen[19]. Dort bleibt ihm bedeutend weniger Zeit für wissenschaftliche Arbeit[20]; dafür erhält er den Titel Baron.

Nachdem die Ausbildung des Kronprinzen abgeschlossen ist, kehrt Cauchy 1838 nach Paris zurück. Er nimmt seinen alten Platz an der Académie des Sciences wieder ein; eine Professur kann er jedoch nicht bekommen, da er sich weigert, den Treueeid zu leisten. Im Jahre 1839 wird ein Platz frei im Bureau des Longitudes. Cauchy wird einstimmig gewählt und es fehlt nur noch die Sanktion des Bürgerkönigs Louis-Philippe; doch diese wird wiederum mit dem Treueeid verbunden[21]. Zwei hochstehende Persönlichkeiten versuchen vergeblich, einen Kompromiss zu erreichen[22]; doch Cauchy ist zu keinerlei Konzessionen bereit und so wird abermals nichts aus seiner Wahl. Im Jahre 1848 wird Louis-Philippe gestürzt. Die neue republikanische Regierung schafft den Treueeid ab und Cauchy erhält wenig später seinen alten Lehrstuhl an der Sorbonne zurück, den einzigen, der noch freigeblieben ist. Vier Jahre später kommt Napoleon III. an die Macht, der Treueeid wird erneut eingeführt und Cauchy tritt auch sofort wieder von seinem Lehrstuhl zurück; er wird jedoch nach einigen Monaten vom Erziehungsminister zurückberufen und mit Arago zusammen vom Eid befreit[23]. Von da an lehrte Cauchy bis zu seinem Tode weiter an der Sorbonne. Sein Gehalt verwendete er jedoch nicht für sich, sondern stellte es meist für wohltätige Zwecke seiner Wohngemeinde Sceaux zur Verfügung. Als der Bürgermeister von Sceaux einmal Bedenken äusserte gegenüber dieser aussergewöhnlichen Grosszügigkeit, entgegnete ihm Cauchy[24]: " Ne vous effrayez pas; ce n'est que mon traitement, ce n'est pas moi, c'est l'Empereur qui paye". In Sceaux starb Cauchy am 22. Mai 1857 und wurde dort auch begraben.

Cauchy benahm sich in seinem Leben manchmal schon recht seltsam. Abel bezeichnete ihn in seinen Briefen als verdreht und bigott[25]. Hat er recht, oder ist eher Biot und Valson zuzustimmen, die Cauchy bloss als extrem naiv ansahen[26]? Die Beantwortung dieser Frage ist nicht gerade leicht. Die meisten Briefe von Cauchy sind inzwischen verlorengegangen und die ausführliche

Biographie von Valson gleicht zu sehr einem "Heldenroman", um verlässlich zu sein. Zum Abschluss des biographischen Teiles soll deshalb versucht werden, einige der wesentlichen Charakterzüge Cauchys aufgrund zeitgenössischer Dokumente herauszuarbeiten.

Ein wichtiger Punkt war sicher die religiöse Ueberzeugung von Cauchy. Als Christ und strenggläubiger Katholik organisierte Cauchy zeit seines Lebens wohltätige Werke: Er sammelte Unterschriften gegen illegitime Ehen, schickte eine Petition an den Papst zu Gunsten der Hungerleidenden in Irland, gründete eine Gesellschaft, die die Heiligung des Sonntags zum Ziel hatte, half bei der Gründung des "Institut catholique" usw.[27]. Cauchy pflegte stets gute Beziehungen zu den Jesuiten: Als diese in den letzten Regierungsjahren von Louis-Philippe stark angegriffen wurden, verfasste er zwei Verteidigungsschriften für sie[28]. Mit seiner Religiosität stiess Cauchy viele Zeitgenossen bereits in jungen Jahren vor den Kopf. Valson zum Beispiel zitiert einen längeren Brief aus dem Jahre 1810, in dem Cauchy seine Mutter von Cherbourg aus zu beruhigen versuchte. Cauchy schrieb[29]:

> *Ma chère maman, je vous remercie beaucoup de me faire part de tout ce que vous entendez dire de moi à Paris, soit en bien soit en mal...*
>
> *On dit que la dévotion me fera tourner la tête. Quelles sont les personnes qui disent cela? Ce ne sont pas celles qui ont beaucoup de religion; celles-ci ne m'en ont parlé que pour m'encourager à persister dans ma ligne de conduite, et tout ce qui m'a été rapporté à ce sujet ne me prouve pas qu'elles me blâment. Seulement, il y a quelques jours, une personne de la société de ... me dit amicalement que la religion faisait quelquefois tourner la tête aux jeunes gens. Je causai avec elle à ce sujet et je lui prouvai que je n'avais point la tête tournée. Quant aux personnes qui n'ont point de religion, j'ai résolu de ne leur en parler jamais le premier, je me contente de leur répondre quand elles veulent m'attaquer sur ce point.*

Im Jahre 1826 berichtete Stendhal im New Monthly Magazine von einem ähnlichen Vorfall in der Académie des Sciences. Nach dem Vortrag eines Naturalisten soll Cauchy mit den nachfolgenden Worten gegen den Beifall der anderen Gelehrten protestiert haben[30]:

> *Même en admettant que les choses qu'on vient de nous dire soient aussi vraies que je les crois fausses. Il n'est pas convenable de communiquer de telles vérités au public, étant donné l'état funeste où notre malheureuse Révolution a jetée l'opinion publique. De tels propos pourraient porter préjudice à notre sainte religion. Ils montrent trop nettement l'influence des "causes physiques" et ils tendent à affirmer les méchantes doctrines de Cabanis.*

Worauf natürlich alles lachte. Stendhal glaubte, dass Cauchy die Rolle eines Märtyrers spielen wollte. Wollte Cauchy dies wirklich oder war er nicht

einfach äusserst naiv? Eine Notiz aus der Turiner Zeit von Cauchy aus dem Tagebuch des Königs von Sardinien scheint eher das letztere zu bestätigen. Am 16. Januar 1832 schrieb der König[31]:

> *Je reçus aujourd'hui la visite de remercîment du célèbre professeur Cochy [sic]. Lui ayant fait quelques questions sur les sciences, sur les Universités, il me répondit cinq fois "J'avais pensé que V.M. m'aurait interrogé à ce propos, et je me suis préparé par une note à lui répondre". Et chaque fois il sortit alors un mémoire de sa poche, dont il me faisait lecture. Il manifeste des vues qui me paraissent fort sages et que je compte d'approfondir.*

Von der Mathematik war Cauchy besessen. Neu an ihn herangetragene Probleme griff er sofort auf. Wenn er eine von auswärts der Akademie eingereichte Arbeit beurteilen sollte, so liess er sich die Gelegenheit selten entgehen, selbst einige Sätze auf diesem Gebiet zu beweisen und erwähnte natürlich auch alles, was er früher schon gefunden hatte. Ueberhaupt informierte Cauchy seine Fachkollegen stets über geglückte aber auch über sämtliche misslungenen Versuche, ein Problem zu lösen. Er publizierte alles und zwar so schnell wie möglich. Als die Comptes Rendus noch nicht existierten, gab Cauchy zu diesem Zweck eine eigene Zeitschrift heraus, die *Exercices Mathématiques*. Falls es nicht anders möglich war, publizierte er seine Resultate manchmal sogar in Zeitungen. Hören wir auch hierzu wieder die Meinung eines Zeitgenossen. Bertrand schreibt in seiner Kritik des Buches von Valson[32]:

> *Le génie de Cauchy est digne de tous nos respects, mais pourquoi s'abstenir de rappeler que la trop grande abondance de ses travaux, en diminuant souvent leur précision, en a plus d'une fois caché la force? La dangereuse facilité d'une publicité immédiate a été pour Cauchy une tentation irrésistible et souvent un écueil. Son esprit, toujours en mouvement, apportait chaque semaine à l'Académie ses travaux à peine ébauchés, des projets de mémoire et des tentatives parfois infructueuses, et, lors même qu'une brillante découverte devait couronner ses efforts, il forçait le lecteur à le suivre dans les voies souvent stériles essayées et abandonnées tour à tour sans que rien vînt l'en avertir. Prenons pour exemple la théorie des substitutions et du nombre de valeurs d'une fonction... Cauchy a composé plus de vingt mémoires. Deux d'entre eux sont des chefs-d'oeuvre; que dire des dix-huit autres? Rien, sinon que l'auteur y cherche une voie nouvelle, la suit quelque temps, entrevoit la lumière, s'efforce inutilement de l'atteindre et quitte enfin, sans marquer aucun embarras, les avenues de l'édifice qu'il renonce à construire.*

Cauchy war in dieser Hinsicht das pure Gegenteil von Gauss. Er war vielleicht der flüchtigste von allen grossen Mathematikern. Er hat über 800 Arbeiten veröffentlicht; wovon etwa 400 in der Analysis, 100 in der Algebra, 40 in der Geometrie, 200 in der Mechanik und Optik und etwa 70 in der Astronomie.

Allein in der Elastizitätslehre werden 17 Theoreme oder Begriffe nach Cauchy benannt[33].

II. CAUCHYS BEITRAEGE ZUR GRUNDLEGUNG DER ANALYSIS

Es ist völlig unmöglich, die 800 Arbeiten von Cauchy hier einigermassen sachgerecht zu behandeln, weshalb sich der zweite Teil auf die Analysis beschränkt und dabei nur einige Hauptpunkte aufzeigt[34]. Zunächst soll die strenge Grundlegung der Infinitesimalrechnung betrachtet werden, wie sie Cauchy in seinem *Cours d'analyse de l'École Royale Polytechnique*[11] und in seinem *Résumé des leçons données à l'École Royale Polytechnique, sur le calcul infinitésimal*[12] geschaffen hat. Im Vorwort zum "Calcul infinitésimal" distanziert sich Cauchy deutlich von seinen Vorgängern. Er sagt[35]:

> *Mon but principal a été de concilier la rigueur, dont je m'étais fait une loi dans mon Cours d'analyse, avec la simplicité qui résulte de la considération directe des quantités infiniment petites. Pour cette raison, j'ai cru devoir rejeter les développements des fonctions en séries infinies, toutes les fois que les séries obtenues ne sont pas convergentes; et je me suis vu forcé de renvoyer au calcul intégral la formule de Taylor, cette formule ne pouvant plus être admise comme générale qu'autant que la série qu'elle renferme se trouve réduite à un nombre fini de termes, et complétée par une intégrale définie.*

Diese Stelle wendet sich hauptsächlich gegen Lagrange[36]. Cauchy kritisiert hier den Gebrauch von nicht konvergenten unendlichen Reihen. Im vorangegangenen 18. Jahrhundert hat man sich vielfach recht wenig um Konvergenzbetrachtungen gekümmert. Es ist das Verdienst von Abel, Bolzano, Cauchy, Dirichlet und Gauss auf die daraus entstehenden Fehler hingewiesen zu haben. Valson erzählt in diesem Zusammenhang[37], dass, nachdem Cauchy seine diesbezüglichen Forschungen um das Jahr 1820 zum ersten Mal vor der Akademie vorgetragen hatte, Laplace bestürzt nach Hause geeilt sei. Er sei erst wieder auf die Strasse gegangen, nachdem er die Konvergenz seiner Formeln in der "Mécanique céleste" überprüft hatte.

Cauchy baut seine Theorie der unendlichen Reihen auf dem Begriff des Limes auf, den er auf den ersten Seiten seines "Cours d'Analyse" allgemein wie folgt festlegt[38]:

> *Lorsque les valeurs successivement attribuées à une même variable s'approchent indéfiniment d'une valeur fixe, de manière à finir par en différer aussi peu que l'on voudra, cette dernière est appelée la limite de toutes les autres.*

Unter Benützung des Limesbegriffes definiert er die Konvergenz einer Reihe[39].

In seiner Theorie der unendlichen Reihen im Kapitel 6 des "Cours d'Analyse" behandelt Cauchy das berühmte Cauchysche Kriterium, dass eine unendliche Reihe genau dann konvergent ist, wenn die Differenzen ihrer Teilsummen $|s_{n+m} - s_n|$ für genügend grosse n beliebig klein werden[40]. An derselben Stelle erörtert Cauchy noch weitere Konvergenzkriterien z.B. das Wurzelkriterium und das Quotientenkriterium; sodann betrachtet er Summe und Differenz von unendlichen Reihen, studiert alternierende Reihen und entwickelt darauf eine Theorie der Potenzreihen[41]. Cauchy hat damit wesentlich zur Grundlegung der heutigen Reihentheorie beigetragen.

Auch seine Definition der Stetigkeit ist neuartig. Cauchy stützt sich hier ebenfalls auf seinen oben zitierten Limesbegriff. Er definiert[42]:

> *En d'autres termes, la fonction f(x) restera continue par rapport à x entre les limites données, si, entre ces limites, un accroissement infiniment petit de la variable produit toujours un accroissement infiniment petit de la fonction elle-même.*

Unter "infiniment petit" versteht er dabei folgendes[43]:

> *Lorsque les valeurs numériques successives d'une même variable décroissent indéfiniment, de manière à s'abaisser au-dessous de tout nombre donné, cette variable devient ce qu'on nomme un infiniment petit ou une quantité infiniment petite. Une variable de cette espèce a zéro pour limite.*

Sicher mutet diese Definition den heutigen Mathematiker noch recht archaisch an; es fehlt das vertraute ε und δ. Die sogenannte Epsilontik geht in ihren Anfängen zwar auf Cauchy zurück, systematisch ausgebaut findet sie sich jedoch erst bei Weierstrass[44]. Weiter fällt auf, dass Cauchy die Stetigkeit nicht in einem Punkt sondern in einem Intervall definiert. Dazu kommt, dass Cauchy seine Definition der Stetigkeit nicht immer konsequent benützt und zudem die Wichtigkeit der gleichmässigen Stetigkeit nie erkannte. All dies sind zweifellos Mängel. Stellt man jedoch der Definition von Cauchy den Stetigkeitsbegriff des 18. Jahrhunderts gegenüber, so ist der Fortschritt klar erkennbar. Im 18. Jahrhundert verstand man unter einer Funktion einer Variablen meist einen Rechenausdruck, in dem die Variable und Konstanten vorkommen dürfen. Dieser Rechenausdruck bestimmt eine sogenannte "stetige" Kurve nach Euler[45]. Nun kann man sich aber auch Kurven denken, die nicht mithilfe eines einzigen solchen Rechenausdruckes beschrieben werden können. Solche Kurven nannte Euler im Gegensatz zu den ersteren "unstetig" oder "gemischt". Der Begriff "stetig" hatte somit bei Euler die Bedeutung von "demselben analytischen Ausdruck genügend". Wie die weitere Entwicklung zeigt, erwies sich diese Einteilung als problematisch[46]. Zu Beginn des 19.

Jahrhunderts wurde deshalb der Stetigkeitsbegriff von Bolzano und Cauchy neu definiert. Cauchy war sich der Wichtigkeit der Aenderung durchaus bewusst. In einer späteren Arbeit aus dem Jahre 1844 schreibt er hierzu[47]:

> *Dans les Ouvrages d'Euler et de Lagrange, une fonction est appelée continue ou discontinue, suivant que les diverses valeurs de cette fonction, correspondantes à diverses valeurs de la variable, sont ou ne sont pas assujetties à une même loi, sont ou ne sont pas fournies par une seule et même équation. C'est en ces termes que la continuité des fonctions se trouvait définie par ces illustres géomètres, ... Toutefois, la définition que nous venons de rappeler est loin d'offrir une précision mathématique; car, si les diverses valeurs d'une fonction, correspondantes aux diverses valeurs d'une variable, dépendent de deux ou de plusieurs équations distinctes, rien n'empêchera de diminuer le nombre de ces équations et même de les remplacer par une équation unique, dont la décomposition fournirait toutes les autres. Il y a plus: les lois analytiques auxquelles les fonctions peuvent être assujetties se trouvent généralement exprimées par des formules algébriques ou transcendantes, et il peut arriver que diverses formules représentent, pour certaines valeurs d'une variable x, la même fonction; puis, pour d'autres valeurs de x, des fonctions différentes. Par suite, si l'on considère la définition d'Euler et de Lagrange comme applicable à toutes espèces de fonctions, soit algébriques, soit transcendantes, un simple changement de notation suffira souvent pour transformer une fonction continue en fonction discontinue, et réciproquement.*

Cauchy illustriert den Sachverhalt anhand folgender Funktion:

(i) $y = \frac{2}{\pi} \int_0^\infty \frac{x^2 dt}{t^2+x^2}$, $= \sqrt{x^2}$ "stetig" nach Euler

(ii) $y = +x$ für $x \geq 0$ "unstetig" nach Euler.
$y = -x$ für $x < 0$

Damit soll die reelle Analysis abgeschlossen und zur komplexen Funktionentheorie übergegangen werden. In der Theorie der komplexen Funktionen schreitet Cauchy zunächst nur zaghaft voran. Lange Zeit ist er - wie die meisten damaligen Mathematiker[48] - der Ansicht, dass eine Gleichung zwischen komplexen Grössen nur eine symbolische Zusammenfassung zweier reeller Gleichungen sei. Diesen Standpunkt vertritt er zunächst auch in seinem *Mémoire sur les intégrales définies* (1814-27)[7]. Sein Hauptanliegen in dieser Arbeit ist es, Gleichungen zu erhalten, die den Uebergang von einem bestimmten Integral zu einem anderen ermöglichen. Hierzu geht Cauchy von einer differenzierbaren komplexen Funktion $F(x+iy) = S + iV$ aus[49]. Aufgrund

der Cauchy-Riemannschen Differentialgleichungen, die übrigens bereits Euler und D'Alembert bekannt waren[48], erhält er:

$$\int\int \frac{\partial V}{\partial y}\, dy\, dx = \int\int \frac{\partial S}{\partial x}\, dx\, dy$$

$$\int\int \frac{\partial S}{\partial y}\, dy\, dx = -\int\int \frac{\partial V}{\partial x}\, dx\, dy.$$

Cauchy integriert diese Gleichungen über ein Rechteck $x_o \leq x \leq x_1$, $y_o \leq y \leq y_1$. Dies ergibt:

$$\int_{x_o}^{x_1} [V(x,y_1) - V(x,y_o)]dx = \int_{y_o}^{y_1} [S(x_1,y) - S(x_o,y)]dy$$

$$\int_{x_o}^{x_1} [S(x,y_1) - S(x,y_o)]dx = -\int_{y_o}^{y_1} [V(x_1,y) - V(x_o,y)]dy.$$

Hier bleibt Cauchy im Jahre 1814 zunächst stehen; er benützt diese zwei Gleichungen zur Berechnung von bestimmten Integralen. Zudem studiert er den Fall[50], wo die Funktion im Rechteck eine singuläre Stelle besitzt und versucht, das in den Gleichungen zusätzlich auftretende Glied zu bestimmen. Noch im Jahre 1823 steht Cauchy einer komplexen Integration ablehnend gegenüber. Er kritisiert Poisson, welcher die Integration durch komplexes Gebiet zur Bestimmung von gewissen reellen Integralen benutzt hatte[51]. Zwei Jahre später erkennt er jedoch den Vorteil einer solchen Betrachtungsweise und baut sie im Gegensatz zu Poisson systematisch aus. Vor dem Druck seines 1814 geschriebenen Manuskriptes im Jahre 1827 fügt Cauchy seiner Arbeit eine Randanmerkung bei, in der er darauf hinweist, dass man die beiden oben erwähnten reellen Gleichungen auch zu einer einzigen komplexen Gleichung zusammenfassen kann[52]. Multipliziert man nämlich die erste Gleichung mit i und addiert sie zur zweiten, so folgt:

$$\int_{x_o}^{x_1} [S(x,y_1) + iV(x,y_1)]dx - \int_{x_o}^{x_1} [S(x,y_o) + iV(x,y_o)]dx =$$

$$= \int_{y_o}^{y_1} [S(x_1,y) + iV(x_1,y)]idy - \int_{y_o}^{y_1} [S(x_o,y) + iV(x_o,y)]idy.$$

Es ergibt sich somit der Cauchysche Integralsatz für ein Rechteck (vgl. Abb. 1):

$$\int_{y_o}^{y_1} F(x_o+iy)idy + \int_{x_o}^{x_1} F(x+iy_1)dx = \int_{x_o}^{x_1} F(x+iy_o)dx + \int_{y_o}^{y_1} F(x_1+iy)idy.$$

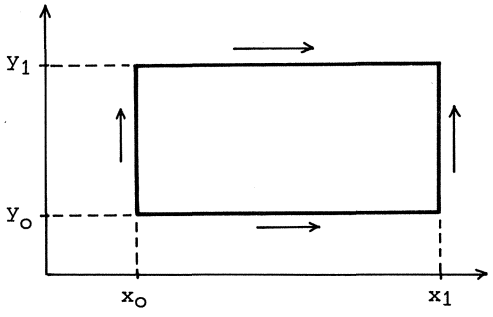


Abb. 1

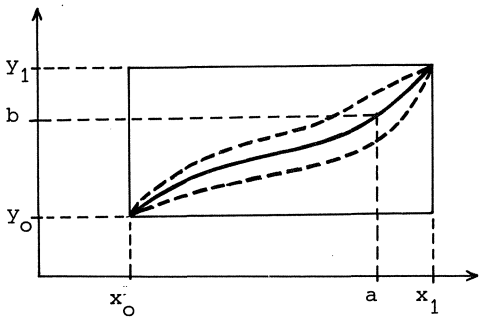


Abb. 2

Allgemeiner und deutlicher formuliert findet sich der Cauchysche Integralsatz im *Mémoire sur les intégrales définies, prises entre des limites imaginaires* (1825)[13]. Dieses Memoire gehört mit zu den schönsten Arbeiten, die Cauchy überhaupt je geschrieben hat. Cauchy definiert im § 2 zunächst, was er unter einem Integral längs eines komplexen Weges verstehen will. Den Weg gibt er dabei in Parameterdarstellung[53]:

$$A + iB = \int_{x_o+iy_o}^{x_1+iy_1} f(z)dz$$

wobei $x = \phi(t)$, $y = \chi(t)$

$\phi(t)$, $\chi(t)$ stetig[54],

monoton wachsend

$\phi(t_o) = x_o$, $\chi(t_o) = y_o$

$\phi(t_1) = x_1$, $\chi(t_1) = y_1$.

Er erhält sodann die bekannte Formel:

$$A + iB = \int_{t_o}^{t_1} [\phi'(t) + i\chi'(t)]f[\phi(t) + i\chi(t)]dt.$$

Im §3 studiert Cauchy die Frage, wann dieses Integral vom Wege unabhängig ist. Zunächst setzt er voraus, dass die gegebene Funktion in einem Rechteck "endlich[55] und stetig[54]" bleibt. In diesem Rechteck betrachtet er zwei Integrationswege (vgl. Abb. 2). Der eine Weg ist gegeben durch $x = \phi(t)$, $y = \chi(t)$ und der andere durch eine leicht variierte Kurve $x = \phi(t) + \varepsilon u(t)$, $y = \chi(t) + \varepsilon v(t)$, wobei jedoch die Anfangs- und Endpunkte der beiden Wege zusammenfallen sollen. Cauchy berechnet nun die Differenz der Integrale längs dieser beiden Wege und zeigt, dass sie unter den gegebenen Bedingungen verschwinden muss. Im Beweis benützt Cauchy, dass $f(z)$ im Rechteck eine stetige Ableitung besitzt. Dies ergibt sich jedoch seiner Ansicht nach aus den obigen Voraussetzungen. Eine solche Auffassung war damals weit verbreitet[56]; man beschränkte sich ja auch meist auf das Studium von analytischen Funktionen. Mit der Konstruktion von stetigen, "nichtdifferenzierbaren" Funktionen befasste man sich (wenn man vom damals kaum bekannten Bolzano absieht) erst zur Zeit von Weierstrass und Riemann.

Im §4 untersucht Cauchy den Fall, wo die gegebene Funktion $f(z)$ in einem Punkt $a + ib$ des nicht variierten Weges unendlich wird und dort nach der heutigen Terminologie einen Pol erster Ordnung besitzt. Cauchy bemerkt zunächst, dass in diesem Fall das Integral längs des nicht variierten Weges "unbestimmt" wird. Betrachtet man nun zwei variierte Wege, welche die Variationen $\varepsilon u(t)$, $\varepsilon v(t)$ und $-\varepsilon u(t)$, $-\varepsilon v(t)$ besitzen und die Singularität somit umschliessen (vgl. Abb. 2), so ist die Differenz der Integrale längs dieser beiden Wege nicht mehr gleich null. Sie lässt sich jedoch im vorliegenden Fall durch die untenstehende Formel ausdrücken:

$$A'' + iB'' - (A' + iB') = \pm 2\pi f i, \quad \text{wobei } f = \lim_{\substack{x\to a \\ y\to b}} [x-a+i(y-b)]f(x+iy).$$

Den Wert f nennt Cauchy später Residuum. In den nachfolgenden Paragraphen studiert er auch die Fälle, in denen die Funktion im Rechteck einen Pol m-ter Ordnung oder mehrere Pole zugleich besitzt. Die 1825er Arbeit liefert somit einen wesentlichen Beitrag zur Residuentheorie; Cauchys Formeln sind jedoch häufig komplizierter, als wir dies heute gewohnt sind. Cauchy kannte

damals die Laurentsche Reihenentwicklung einer analytischen Funktion noch nicht, diese wurde erst im Jahre 1843 durch Laurent entdeckt.

Zum Schluss soll noch kurz auf die berühmte Turiner Abhandlung *Mémoire sur la mécanique céleste et sur un nouveau calcul appellé calcul des limites* (1831-33)[18] eingegangen werden, in der Cauchy die Potenzreihenentwicklung einer analytischen Funktion behandelt. In Teil I, §2 beweist Cauchy zunächst seine Integralformel[53]:

$$f(z) = \frac{1}{2\pi} \int_{-\pi}^{+\pi} \frac{\zeta f(\zeta)}{\zeta - z}\, dp \qquad (d\zeta = \zeta i dp).$$

Er beschränkt sich dabei auf einen kreisförmigen Integrationsweg um den Nullpunkt und setzt wie üblich voraus, dass die Funktion auf dieser Kreisscheibe "endlich und stetig" bleibt. — Beim Wiederabdruck der Arbeit im Jahre 1841 hat Cauchy übrigens an dieser Stelle nach einer Kontroverse mit Sturm und Liouville zusätzlich die Existenz einer stetigen Ableitung verlangt[57]. — Anschliessend entwickelt Cauchy den Ausdruck $\frac{\zeta}{\zeta - z}$ in eine geometrische Reihe und erhält folgenden Satz: Eine analytische Funktion lässt sich innerhalb eines gewissen grössten Kreises [dem Konvergenzkreis] in eine Potenzreihe entwickeln, der dadurch bestimmt ist, dass die Funktion und ihre Ableitung in ihm gerade noch überall stetig sind. Cauchy diskutiert den Satz anhand verschiedener Beispiele und weist darauf hin, dass sich auf dem Konvergenzkreis mindestens eine singuläre Stelle befinden muss. In derselben Arbeit erwähnt Cauchy übrigens auch noch seine Abschätzungsformel für die Koeffizienten der Reihe.

Das Studium von Cauchys Werken ist nicht immer leicht. Seine Forschungen in der Funktionentheorie zum Beispiel hat Cauchy nie in Buchform zusammengefasst, sie befinden sich in einer Vielzahl von sich teilweise wiederholenden Arbeiten. Cauchy ist stets auf der Suche und mit ihm natürlich auch der Leser. Manchmal scheint Cauchy seine früheren Resultate wieder vergessen zu haben. Da definiert er zum Beispiel im "Cours d'Analyse" einen neuen Stetigkeitsbegriff, wendet diesen jedoch in der Folge keineswegs konsequent an. Oder um ein anderes Beispiel zu nennen: Cauchy hätte mit der 1825er Arbeit die Möglichkeit gehabt, seinen Integralsatz allgemein zu formulieren; er stützt sich jedoch während der nächsten 20 Jahre weiterhin auf sein 1814er Manuskript und beschränkt sich meist auf rechteckige oder kreisförmige Integrationswege. Die von Bertrand[32] kurz nach dem Tode von Cauchy geäusserte Meinung trifft somit zweifelsohne zu: "Das Genie von Cauchy verdient unseren Respekt, aber warum soll man verschweigen, dass die

Ueberfülle von Arbeiten ihre Genauigkeit häufig verringert und ihren Wert manchmal sogar verdeckt hat".

-o-o-o-o-o-o-o-o-o-o-o-o-o-o-

1) Dieser erste biographische Teil wurde vor allem auf die Biographie von VALSON gestützt. Grosse Teile von Cauchys Nachlass sind inzwischen verlorengegangen. Eine Zusammenstellung von erhaltengebliebenen Dokumenten findet man in Oeuvres, Serie II, Bd. 15, S. 611ff.

2) Vgl. VALSON, S. 18 und BIOT, S. 144.

3) Vgl. VALSON, S. 27.

4) Nach VALSON, S. 29. Brief vom 10. Dez. 1810.

5) A.-L. Cauchy, Sur les polygones et les polyèdres, J.E.P. 9, 16^{e} cah. (1813) = Oeuvres, Serie II, Bd. 1, S. 35f.

6) Vgl. die chronologische Zusammenstellung der Arbeiten von CAUCHY in Oeuvres, Serie II, Bd. 15, S. 583-607.

7) A.-L. Cauchy, Mémoire sur les intégrales définies (gelesen 1814, zum Druck eingereicht 1825), M. Sav. Étr. 1 (1827) = Oeuvres, Serie I, Bd. 1, S. 329-506.

8) A.-L. Cauchy, Théorie de la propagation des ondes à la surface d'un fluide pesant d'une profondeur indéfinie (eingereicht 1815), M. Sav. Étr. 1 (1827) = Oeuvres, Serie I, Bd. 1, S. 5-318.

9) Vgl. die Briefauszüge in VALSON, S. 55ff.

10) Vgl. die Ausführungen von BIOT, S. 147f.

11) A.-L. Cauchy, Cours d'analyse de l'École Royale Polytechnique, 1re partie, Analyse algébrique, Paris 1821 = Oeuvres, Serie II, Bd. 3.

12) A.-L. Cauchy, Résumé des leçons données à l'École Royale Polytechnique, sur le calcul infinitésimal, Bd. 1, Paris 1823 = Oeuvres, Serie II, Bd. 4.

13) A.-L. Cauchy, Mémoire sur les intégrales définies, prises entre des limites imaginaires, Paris 1825 = Oeuvres, Serie II, Bd. 15, S. 41-89.

14) Vgl. BIOT, S. 150.

15) Die erhaltengebliebenen Akten von der Amtsenthebung von Cauchy sind von Taton analysiert und publiziert worden. Vgl. TATON, S. 142f.

16) Vgl. CASTELLA und TERRACINI.

17) Für den Turiner Aufenthalt vgl. vor allem TERRACINI sowie VALSON, S. 75ff.

18) Die vollständige Fassung dieser Schrift ist in den Jahren 1832/33 als Lithographie erschienen (= Oeuvres, Serie II, Bd. 15, S. 262-411). Den Abschnitt über die Potenzreihenentwicklung hat Cauchy im Jahre 1841 mit einigen Ergänzungen in den Exercices Mathématiques abdrucken lassen (= Oeuvres, Serie II, Bd. 12, S. 48-112). Ausführliche Angaben über die weiteren Fassungen und Abdrucke gibt GRATTAN-GUINNESS 1975, S. 182f. Hinzuzufügen ist eine von Cauchy im Jahre 1840 in den Exercices Mathématiques veröffentlichte Abhandlung (= Oeuvres, Serie II, Bd. 11, S. 43-50), in der Cauchy selbst (ebenda S. 45) Angaben macht über die verschiedenen Fassungen.

19) Vgl. VALSON, S. 83. Angaben über den Aufenthalt von Cauchy in Prag machen auch D'HAUTPOUL, RYCHLÍK und SINACEUR.

20) Vgl. VALSON, S. 90.

21) Cauchy hat im Jahre 1843 schriftlich gegen diese Verknüpfung protestiert. Vgl. den diesbezüglichen Brief von Cauchy in VALSON, S. 101-104.

22) Vgl. BIOT, S. 152.

23) Vgl. BIOT, S. 153f.

24) BIOT, S. 159. Vgl. auch BIOT, S. 154 und VALSON, S. 274, wo man eine leicht variierende Formulierung findet.

25) Vgl. TATON, S. 125.

26) Vgl. z.B. BIOT, S. 152.

27) Vgl. VALSON, S. 188-242.

28) Vgl. VALSON, S. 108-121.

29) Nach VALSON, S. 36-39. Wir verweisen in diesem Zusammenhang auch auf das Memoire *Sur les limites des connaissances humaines* (= Oeuvres, Serie II, Bd. 15, S. 5-7), welches Cauchy am 14. Nov. 1811 vor der Société Académique de Cherbourg gelesen hat.

30) Nach TERRACINI, S. 192.

31) Nach TERRACINI, S. 160.

32) BERTRAND, S. 210.

33) Vgl. TRUESDELL - TOUPIN.

34) Für eine Würdigung des Gesamtwerkes von Cauchy verweisen wir auf den ausgezeichneten Artikel von FREUDENTHAL.

35) Oeuvres, Serie II, Bd. 4, S. 9.

36) Vgl. Oeuvres, Serie II, Bd. 4, S. 10.

37) VALSON, S. 127.

38) Oeuvres, Serie II, Bd. 3, S. 19.

39) Oeuvres, Serie II, Bd. 3, S. 114.

40) Oeuvres, Serie II, Bd. 3, S. 115f. Vgl. auch Oeuvres, Serie II, Bd. 7, S. 267ff.

41) Oeuvres, Serie II, Bd. 3, S. 135ff.

42) Oeuvres, Serie II, Bd. 3, S. 43.

43) Oeuvres, Serie II, Bd. 3, S. 19.

44) Vgl. SINACEUR, S. 108ff. und DUGAC, S. 63f.

45) Für Stellenangaben und detaillierte Ausführungen vgl. YOUSCHKEVITCH, S. 64ff.

46) Vgl. hierzu vor allem JOURDAIN 1914. Jourdains Gedanken sind während der letzten Jahre von verschiedenen Autoren erneut aufgegriffen worden.

47) A.-L. Cauchy, Mémoire sur les fonctions continues, C.R. 18 (1844) = Oeuvres, Serie I, Bd. 8, S. 145f.

48) Einen Ueberblick über die Anfänge der komplexen Funktionentheorie gibt z.B. MARKUSCHEWITSCH.

49) Vgl. Oeuvres, Serie I, Bd. 1, S. 336ff. Wir geben Cauchys Ausführungen hier vereinfacht und mit vereinheitlichter, leicht modernisierter Notierung wieder. Für eine ausführliche Diskussion der 1814er Arbeit verweisen wir auf ETTLINGER.

50) Vgl. Oeuvres, Serie I, Bd. 1, S. 378ff.

51) Vgl. Cauchy, Oeuvres, Serie II, Bd. 1, S. 354. Hinsichtlich der Rivalität zwischen Poisson und Cauchy vgl. u.a. GRATTAN-GUINNESS 1970, S. 28ff.

52) Vgl. Oeuvres, Serie I, Bd. 1, S. 338.

53) Cauchys Notierung wurde wiederum leicht modernisiert.

54) Vgl. unten.

55) Die Frage, ob es notwendig sei "endlich und stetig" oder bloss "stetig" zu sagen, wird von Casorati noch im Jahre 1864 in Gesprächen mit Kronecker und Weierstrass aufgeworfen. Vgl. NEUENSCHWANDER 1977, S. 7f und 16.

56) Vgl. NEUENSCHWANDER 1978.

57) Cauchy begründet diesen Schritt in Oeuvres, Serie II, Bd. 11, S. 50 und in Oeuvres, Serie II, Bd. 12, S. 58f, Randanmerkung 1.

LITERATUR

BERTRAND, J., *La vie et les travaux du baron Cauchy par C.A. Valson,* Journal des savants (1869), S. 205-215; Bulletin des sciences mathématiques 1 (1870), S. 105-116.

BERTRAND, J., *Éloge d'Augustin-Louis Cauchy,* Éloges académiques, Bd. 2, Paris 1902, S. 101-120.

BIOT, J.B., *M. le Baron Cauchy,* Correspondant, Paris 1857; Mélanges scientifiques et littéraires, Bd. 3, Paris 1858, S. 143-160.

BONCOMPAGNI, B., *La vie et les travaux du baron Cauchy,* Bollettino di bibliografia e di storia delle scienze matematiche e fisiche 2 (1869), S. 1-102.

BRILL, A. und M. NOETHER, *Die Entwicklung der Theorie der algebraischen Functionen in älterer und neuerer Zeit,* Jahresbericht der Deutschen Mathematiker-Vereinigung 3 (1894), S. 107-566.

CARRUCCIO, E., *I fondamenti dell'analisi matematica nel pensiero di Agostino Cauchy,* Bollettino Unione matematici italiani, 1957, S. 290-307; Rendiconti del Seminario Matematico (Turin) 16 (1956-1957), S. 205-216.

CASORATI, F., *Teorica delle funzioni di variabili complesse,* Pavia 1868.

CASTELLA, G., *Documents inédits sur un projet de fonder une Académie Helvétique à Fribourg en 1830,* Revue d'histoire ecclésiastique suisse 21 (1927), S. 308-313.

CAUCHY, A., *Oeuvres complètes,* 27 Bände, Paris 1882-1974.

D'HAUTPOUL, A., *Quatre mois à la cour de Prague,* Paris 1902.

DOBROVOLSKI, W.A., *Contribution à l'histoire du théorème fondamental des équations différentielles,* Archives internationales d'histoire des sciences 22 (1969), S. 223-234.

DUBBEY, J.M., *Cauchy's contribution to the establishment of the calculus,* Annals of Science 22 (1966), S. 61-67.

DUGAC, P., *Éléments d'analyse de Karl Weierstrass,* Archive for History of Exact Sciences 10 (1973), S. 41-176.

DUROSELLE, J.-B., *Les débuts du catholicisme social en France (1822-1870)*, Paris 1951.

ETTLINGER, H.J., *Cauchy's paper of 1814 on definite integrals*, Annals of Mathematics 23 (1921/22), S. 255-270.

FREUDENTHAL, H., *Cauchy, Augustin-Louis*, in: Dictionary of Scientific Biography, Bd. 3, New York 1971, S. 131-148.

FREUDENTHAL, H., *Did Cauchy plagiarize Bolzano?*, Archive for History of Exact Sciences 7 (1971), S. 375-392.

GRATTAN-GUINNESS, I., *Bolzano, Cauchy and the "New analysis" of the early nineteenth century*, Archive for History of Exact Sciences 6 (1970), S. 372-400.

GRATTAN-GUINNESS, I., *The development of the foundations of mathematical analysis from Euler to Riemann*, Cambridge Mass. 1970.

GRATTAN-GUINNESS, I., *On the publication of the last volume of the works of Augustin Cauchy*, Janus 62 (1975), S. 179-191.

YOUSCHKEVITCH, A.P., *The concept of function up to the middle of the 19th century*, Archive for History of Exact Sciences 16 (1976-77), S. 37-85.

HADAMARD, J., *Le problème de Cauchy et les équations aux dérivées partielles linéaires hyperboliques*, Paris 1932.

HAWKINS, T.W., *Cauchy and the spectral theory of matrices*, Historia Mathematica 2 (1975), S. 1-29.

IACOBACCI, R.F., *Augustin-Louis Cauchy and the development of mathematical analysis*, Dissertation: New York University 1965.

JOURDAIN, P., *The theory of functions with Cauchy and Gauss*, Bibliotheca mathematica, 3. Folge, Bd. 6 (1905), S. 190-207.

JOURDAIN, P., *The origin of Cauchy's conceptions of a definite integral and of the continuity of a function*, Isis 1 (1914), S. 661-703.

KLEIN, F., *Vorlesungen über die Entwicklung der Mathematik im 19. Jahrhundert*, Bd. 1, Berlin 1926.

LAURENT, H., *Cauchy*, École Polytechnique. Livre du Centenaire, Bd. 1, Paris 1895, S. 104-110.

LORIA, G., *A.L. Cauchy in the history of analytic geometry,* Scripta Mathematica 1 (1932), S. 123-128.

MARKUSCHEWITSCH, A.I., *Skizzen zur Geschichte der analytischen Funktionen,* Berlin 1955.

MEYER, W.F., (Hrsg.), *Encyklopädie der mathematischen Wissenschaften mit Einschluss ihrer Anwendungen,* Leipzig 1898-1935.

MOIGNO, F.N.M., *Sept leçons de physique générale par A. Cauchy,* Paris 1855.

NEUENSCHWANDER, E., *Casoratis Gespräche mit Kronecker und Weierstrass in Berlin im Jahre 1864,* Preprint: History of Science Department, University of Aarhus, Aarhus 1977.

NEUENSCHWANDER, E., *Riemann's example of a continous, "nondifferentiable" function,* Mathematical Intelligencer 1 (1978), S. 40-44.

RYCHLÍK, K., *Un manuscrit de Cauchy aux Archives de l'Académie tchécoslovaque des Sciences,* Revue d'Histoire des Sciences 10 (1957), S. 259-261.

RYCHLÍK, K., *Sur les contacts personnels de Cauchy et de Bolzano,* Revue d'Histoire des Sciences 15 (1962), S. 163-164.

SINACEUR, H., *Cauchy et Bolzano,* Revue d'Histoire des Sciences 26 (1973), S. 97-112.

STAECKEL, P., *Integration durch imaginäres Gebiet,* Bibliotheca mathematica, 3. Folge, Bd. 1 (1900), S. 109-128.

STAECKEL, P., *Beiträge zur Geschichte der Funktionentheorie im 18. Jahrhundert,* Bibliotheca mathematica, 3. Folge, Bd. 2 (1901), S. 111-121.

STUDNIČKA, F.-J., *Cauchy als formaler Begründer der Determinantentheorie. Eine literarisch-historische Studie,* Abhandlungen der königlich-böhmischen Gesellschaft der Wissenschaft, 6. Folge, Bd. 8 (1876).

TATON, R., *Sur les relations scientifiques d'Augustin Cauchy et d'Evariste Galois,* Revue d'Histoire des Sciences 24 (1971), S. 123-148.

TERRACINI, A., *Cauchy a Torino,* Rendiconti del Seminario Matematico (Turin) 16 (1956-57), S. 159-204.

TODHUNTER, I., *A history of the progress of the calculus of variations during the nineteenth century,* Cambridge 1861.

TODHUNTER, I. und K. PEARSON, *A history of the theory of elasticity and of the strength of materials*, Bd. 1, Cambridge 1886; 2. Auflage: New York 1960.

TRUESDELL, C., *The rational mechanics of flexible or elastic bodies 1638-1788*, in: Leonhardi Euleri Opera Omnia, 2. Serie, Bd. 11, 2. Hälfte, Zürich 1960.

TRUESDELL, C. und R. TOUPIN, *The classical field theories*, in: Handbuch der Physik, Bd. 3, Berlin-Göttingen-Heidelberg 1960, S. 226-793.

VALSON, C.-A., *La vie et les travaux du baron Cauchy*, Paris 1868, 2 Bände; Nachdruck: Paris 1970.

VERDET, E., *Leçons d'optique physique*, in: Oeuvres de E. Verdet, Bd. 5-6, Paris 1869-70.

THE MATHEMATICAL THEORY OF CLINES

L.A. PELETIER

1. Introduction

In recent years there has been considerable interest in the effect on the genetic composition of a population caused by an inhomogeneous environment. In some polymorphic species it is found that the frequency of one type varies monotonically in a certain direction. Huxley [11] has called such a gradient a cline, and he lists numerous examples.

A cline may be due to the fact that one particular type of the species enjoys a selective advantage in one part of the habitat and a disadvantage in another. In addition there will be the effect of migration, which will tend to keep the different types mixed. Haldane [8] mentions the example of the deer-mouse which inhabits Flordia and Alabama. On the sandy beaches of the Gulf of Mexico the lighter coloured subspecies is less visible and thus enjoys an advantage there.

The first mathematical treatment of migration and selection was given by Fisher [6]. He considered a population, distributed in a one-dimensional uniform habitat Ω in which each individual belongs to one of three possible genotypes: aa, aA and AA. Let $u(x,t)$ denote the fraction of alleles of type a amongst the total number of alleles in the population at the point $x \in \Omega$ and at time $t \geq 0$. Thus u is a measure of the genetic composition of the population which takes on values in the interval $[0,1]$. Adapting Fisher's model to a non-uniform habitat, Haldane [8] showed that under a number of simplifying assumptions u satisfies the equation

$$u_t = u_{xx} + \lambda f(x,u) \qquad x \in \Omega,\ t > 0 \tag{1}$$

in which subscripts denote partial differentiation and λ is a positive constant, which is proportional to the inverse of the rate of migration. A detailed discussion of this model was recently given by Fife [4].

The function f in (1) is related to the relative fitnesses of the three genotypes. We shall consider two particular functions f:

I. $$f(x,u) = s(x)\, u(1-u),$$

where $s: \bar{\Omega} \to \mathbb{R}$ is piecewise continuously differentiable. This choice is appropriate for a population in which the fitness of the heterozygote aA lies between the fitnesses of the homozygotes aa and AA.

II. $$f(x,u) = u(1-u)[u-a(x)],$$

where $a: \bar{\Omega} \to (0,1)$ is continuously differentiable. In this case the fitness of the heterozygote is assumed to be inferior to the fitnesses of the homozygotes.

Let us denote by Ω_a the set of points $x \in \Omega$, where the fitness of aa is superior to the fitnesses of aA and AA, and let Ω_A be the set of points $x \in \Omega$ where AA has the superior fitness. For the functions f defined above the sets Ω_a and Ω_A can be given in terms of the functions s, respectively a.

Case I. $\Omega_a = \{x \in \Omega : s(x) > 0\}$, $\Omega_A = \{x \in \Omega : s(x) < 0\}$

Case II. $\Omega_a = \{x \in \Omega : a(x) < \frac{1}{2}\}$, $\Omega_A = \{x \in \Omega : a(x) > \frac{1}{2}\}$.

We shall be interested in the situation in which neither Ω_a nor Ω_A is empty. Then the two homozygotes are competing, each of them being advantageous in part of the habitat. We shall then enquire into the possible existence of equilibrium solutions, in particular in those in which the three genotypes coexist, and into the stability of these solutions.

In the case that $\Omega = R$ the existence, uniqueness and stability of clines has been discussed by Conley [3], Nagylaki [12,13] and Fife and

Peletier [5]. In this paper we shall consider the case of a bounded habitat. We shall assume that $\Omega = (-1,1)$ and that at the boundary of Ω there is no flow of genetic material. This assumption leads to the conditions [12]

$$u_x(-1,t) = 0, \qquad u_x(+1,t) = 0 \qquad t \geq 0. \tag{2}$$

After introducing some basic notions in section 2, we shall discuss Cases I and II in sections 3 and 4 respectively. The discussion of Case I is largely based on results due to Fleming [7] although in some places we shall present somewhat different proofs. The discussion of Case II is based on [14].

2. Preliminaries

Suppose $u(x)$ is an equilibrium solution of (1) and (2). Then u is a solution of the problem

$$(A)\begin{cases} u'' + \lambda f(x,u) = 0 & -1 < x < 1 \qquad (3) \\ u'(-1) = 0,\ u'(+1) = 0. & \qquad (4) \end{cases}$$

Equation (3) is the Euler equation satisfied by the critical points of the functional

$$V(u) = \int_{-1}^{1} \{\tfrac{1}{2}(u')^2 - \lambda F(x,u)\}dx,$$

where

$$F(x,u) = \int_0^u f(x,s)ds$$

which is defined on the Sobolev space $H^1(-1,1)$. This is the space of functions $\phi \in L^2(-1,1)$ such that $\phi' \in L^2(-1,1)$, endowed with the norm

$$\|\phi\|_1 = (\|\phi\|_0^2 + \|\phi'\|_0^2)^{\frac{1}{2}} \quad , \quad \|\phi\|_0^2 = \int_{-1}^{1} \phi^2 dx.$$

If u is a critical point of V and v an arbitrary element of $H^1(-1,1)$, then

$$V(u+v) = V(u) + \tfrac{1}{2}a(v,v;u) + \rho(\|v\|_1), \tag{5}$$

where

$$a(v,v;u) = \int_{-1}^{1} \{v'^2 - \lambda f_u(x,u(x))v^2\}dx$$

and

$$\rho(s) \to 0 \quad \text{as} \quad s \to 0.$$

The variational structure of Problem A yields a simple criterion for the stability of equilibrium solutions. Suppose $\psi \in H^1(-1,1)$ and $u(x,t;\psi)$ is the solution of (1),(2) which satisfies the initial condition

$$u(x,0;\psi) = \psi(x) \qquad -1 \le x \le 1. \tag{6}$$

Suppose u^* is an equilibrium solution. Then we shall say that u^* is <u>stable</u>, if for any $\varepsilon > 0$, there exists a $\delta > 0$ such that

$$\|\psi - u^*\|_1 < \delta \Rightarrow \|u(\cdot,t;\psi) - u^*\|_1 < \varepsilon \text{ for } t \ge 0;$$

otherwise we call u^* <u>unstable</u>.

<u>THEOREM 1</u>.[7]. *Let* u^* *be an isolated equilibrium solution.*
(i) *If there exists a positive constant* ν *such that*

$$a(v,v;u^*) \ge \nu \|v\|_1^2 \qquad \textit{for all } v \in H^1(-1,1)$$

then u^* *is stable.*

(ii) *If there exists an element* $v \in H^1(-1,1)$ *such that*

$$a(v,v;u^*) < 0$$

then u^* *is unstable.*

Another criterion is based on the sign of the largest eigenvalue of the eigenvalue problem obtained by formally linearizing equation (3) about an equilibrium solution u:

$$(B_u)\begin{cases} y'' + \lambda f_u(x,u(x))y = \mu y \\ y'(-1) = 0, \quad y'(+1) = 0. \end{cases}$$

It is well known that this problem has a sequence of simple eigenvalues $\mu_1 > \mu_2 > \dots$, and that

$$-\mu_1 = \min_{y \in H^1} \frac{a(y,y;u)}{\|y\|_o^2} \qquad y \neq 0, \tag{7}$$

the minimum being attained for the eigenfunction y_1 corresponding to μ_1. Thus, if $\mu_1 > 0$,

$$a(y_1,y_1;u) = -\mu_1 \|y_1\|_o^2 < 0.$$

On the other hand, if $\mu_1 < 0$,

$$a(y,y;u) \geq \nu \|y\|_o^2 \quad \text{for all } y \in H^1(-1,1),$$

where $\nu = -\mu_1 > 0$. Since $f_u(x,u(x))$ is uniformly bounded in $[-1,1]$ it follows that there exists a constant $\nu' > 0$ such that

$$a(y,y;u) \geq \nu' \|y\|_1^2 \quad \text{for all } y \in H^1(-1,1).$$

Summarizing we have the following result.

THEOREM 2. *Let* u *be an isolated equilibrium solution, and let* μ_1 *be the largest eigenvalue of the eigenvalue problem* B_u. *Then* (i) *if* $\mu_1 > 0$, u *is unstable, and* (ii) *if* $\mu_1 < 0$, u *is stable.*

3. Case I

Since $f(x,0) = f(x,1) = 0$ for $-1 \le x \le 1$, the functions

$$u_o(x) \equiv 0 \quad , \quad u_1(x) \equiv 1$$

are equilibrium solutions. To discuss their stability we define the number

$$\lambda^* = \inf_{v \in Q} \frac{\int_{-1}^{1} v'^2 dx}{\int_{-1}^{1} q v^2 dx} , \tag{8}$$

where

$$q(x) = f_u(x,u(x)) = s(x)[1-2u(x)]$$

and

$$Q = \{v \in H^1(-1,1) : \int_{-1}^{1} q v^2 dx > 0\}.$$

In view of the assumptions we made about s in the introduction we may expect q(x) to take on both positive and negative values. It can be shown [2] that

(i) if $\int_{-1}^{1} q(x)dx \ge 0$, then $\lambda^* = 0$;

(ii) if $\int_{-1}^{1} q(x)dx < 0$, then $\lambda^* > 0$ and there exists a function $\bar{v} \in Q$ such that

$$\int_{-1}^{1} \bar{v}'^2 dx = \lambda^* \int_{-1}^{1} q\bar{v}^2 dx \ . \tag{9}$$

Let us consider the solution u_o. Then $q(x) = s(x)$ and

$$a(v,v;u_o) = \int_{-1}^{1} \{v'^2 - \lambda s(x) v^2\} dx.$$

In particular, with $\tilde{v}(x) \equiv 1$,

$$a(\tilde{v},\tilde{v};u_o) = -\lambda \int_{-1}^{1} s(x)dx.$$

Thus, if $\int_{-1}^{1} s(x)dx > 0$, $a(\tilde{v},\tilde{v};u_o) < 0$ and it follows from Theorem 1 that u_o is unstable for any $\lambda > 0$. If $\int_{-1}^{1} s(x)dx = 0$, $a(\tilde{v},\tilde{v};u_o) = 0$. However, since $\tilde{v}$ is not a solution of Problem B_{u_o}, it cannot be a minimizer of (7). Therefore $-\mu_1 < 0$ and hence, by Theorem 2, u_o is unstable for any $\lambda > 0$.

Next, suppose that $\int_{-1}^{1} s(x)dx < 0$ and $\lambda > \lambda^*$. Then, in view of (9),

$$a(\bar{v},\bar{v};u_o) = (\lambda^*-\lambda) \int_{-1}^{1} s\bar{v}^2 dx < 0$$

because $\bar{v} \in Q$. Hence u_o is unstable for $\lambda > \lambda^*$. Next, let $\lambda \in (0,\lambda^*)$. Then, assuming that $\|y_1\|_o = 1$,

$$-\mu_1 = \int_{-1}^{1} (y_1'^2 - \lambda s y_1^2) dx. \tag{10}$$

Suppose $y_1 \notin Q$. Then $\int_{-1}^{1} s y_1^2 dx \le 0$ and hence

$$-\mu_1 \ge \int_{-1}^{1} y_1'^2 dx \ge 0.$$

Thus $\mu_1 < 0$, unless $y_1(x) \equiv 2^{-\frac{1}{2}}$, when $\mu_1 \le 0$. However, in this case it

follows from Problem B_{u_o} that

$$\mu_1 = \tfrac{1}{2}\lambda \int_{-1}^{1} s(x)dx < 0.$$

Next, suppose $y_1 \in Q$. Then $\int_{-1}^{1} sy_1^2 dx > 0$ and, by (8),

$$\int_{-1}^{1} y_1'^2 dx \geq \lambda^* \int_{-1}^{1} sy_1^2 dx,$$

which yields, together with (10)

$$-\mu_1 \geq (\lambda^*-\lambda) \int_{-1}^{1} sy_1^2 dx > 0$$

implying that $\mu_1 < 0$ as well. Thus, if $\lambda \in (0,\lambda^*)$, $\mu_1 < 0$ and it follows from Theorem 2 that u_o is stable.

In an entirely analogous manner we can obtain corresponding results for u_1.

THEOREM 3. *Let* $\lambda_i^* = \lambda^*(u_i)$ $i = 0,1$ *be defined by* (8).

(i) *Let* $\int_{-1}^{1} s(x)dx < 0$. *Then* u_1 *is unstable for any* $\lambda > 0$ *and* u_o *is unstable for* $\lambda > \lambda_o^*$ *and stable for* $\lambda \in (0,\lambda_o^*)$.

(ii) *Let* $\int_{-1}^{1} s(x)dx > 0$. *Then* u_o *is unstable for any* $\lambda > 0$ *and* u_1 *is unstable for* $\lambda > \lambda_1^*$ *and stable for* $\lambda \in (0,\lambda_1^*)$.

(iii) *Let* $\int_{-1}^{1} s(x)dx = 0$. *Then* u_o *and* u_1 *are both unstable for* $\lambda > 0$.

Let us return to the case $\int_{-1}^{1} s(x)dx < 0$. Then, as we saw in the proof of Theorem 3, the largest eigenvalue $\mu_1 = \mu_1(\lambda,u_o)$ changes sign at $\lambda = \lambda_o^*$. It can be shown by means of an argument involving degree theory [15,14] that this implies that the point (λ_o^*,u_o) is a bifurcation point, i.e. in any $\mathbb{R} \times H^1$ neighbourhood of this point there exists a solution $u \neq u_o$ of Problem A. In fact, Fleming [7] showed by different means that there exists a branch of solutions

$$C_\delta = \{(\lambda,\phi(\lambda)) : \lambda_o^* < \lambda < \lambda_o^* + \delta\}$$

for some small $\delta > 0$ such that

$$\phi(\lambda) \to u_o \text{ in } H^1 \text{ as } \lambda \to \lambda_o^*$$

and

$$\mu_1(\lambda,\phi) < 0 \text{ on } C_\delta.$$

A detailed description of the behaviour of solutions of (1),(2),(6) near the point (λ_o^*,u_o) was given by Hoppensteadt[10]. Finally, it was shown by Henry [9] that the branch C_δ can be continued uniquely to the branch

$$C = \{(\lambda,\phi(\lambda)) : \lambda_o^* < \lambda < \infty\},$$

i.e. C contains no bifurcation points. In addition

$$\mu_1(\lambda,\phi) < 0 \quad \text{on } C. \tag{11}$$

and hence, by Theorem 2, solutions on C are stable.

REMARK. In the case that $s(x)$ is a nondecreasing function and $(\lambda,\phi) \in C$ it is not difficult to see that

$$\phi'(x) > 0 \qquad -1 < x < 1.$$

For in that case $f_x(x,\phi(x)) \geq 0$ and hence, writing $\phi' = w$:

$$w'' + \lambda f_u(x,\phi(x))w \leq 0,$$
$$w(-1) = 0 \ , \ w(+1) = 0.$$

Moreover, in view of (11) there exists a constant $\nu > 0$ such that

$$a(v,v;\phi) \geq \nu \|v\|_1^2 \quad \text{for all } v \in H^1(-1,1).$$

The result now follows from an application of the maximum principle [5, Appendix]. Thus if s is nondecreasing, any solution ϕ of C is a cline.

4. Case II

As in Case I, the trivial solutions u_o and u_1 are equilibrium solutions for all $\lambda > 0$. To determine whether or not they are stable we observe that

$$f_u(x,u_o) = -a(x) \quad , \quad f_u(x,u_1) = -1 + a(x).$$

Therefore, by our assumptions on a, there exists a constant $\nu > 0$ such that

$$f_u(x,u_i) < -\nu \quad \text{for} \quad -1 \leq x \leq 1 \ , \ i = 0,1 \tag{12}$$

and hence there exists a constant $\nu' > 0$ such that

$$a(v,v;u_i) \geq \nu' \|v\|_1^2 \qquad i=0,1$$

for all $v \in H^1(-1,1)$. Thus, by Theorem 1, we obtain the following result.

THEOREM 4. *The trivial solutions* u_o *and* u_1 *are stable for all* $\lambda > 0$.

It follows from (12) and a result due to Amann [1] that for each $\lambda > 0$ there exists at least one nontrivial solution ϕ of Problem A, i.e.

$$0 < \phi(x,\lambda) < 1 \quad \text{for} \quad -1 \leq x \leq 1 \ , \ \lambda > 0.$$

For small values of λ, $\phi(\cdot,\lambda)$ turns out to be the only nontrivial solution. Moreover, it depends continuously on λ and

$$\|\phi(\cdot,\lambda) - \alpha\|_1 \to 0 \quad \text{as } \lambda \to 0, \tag{13}$$

where

$$\alpha = \tfrac{1}{2}\int_{-1}^{1} a(x)dx.$$

THEOREM 5. *For small values of* λ, $\phi(\cdot,\lambda)$ *is unstable.*

Proof. Let $\tilde{v}(x) \equiv 1$. Then, by (7),

$$\mu_1(\lambda,\phi) \geq -\tfrac{1}{2}a(\tilde{v},\tilde{v};\phi) = \tfrac{1}{2}\lambda\int_{-1}^{1} f_u(x,\phi(x,\lambda))dx.$$

By (13)

$$\lim_{\lambda\to 0}\int_{-1}^{1} f_u(x,\phi(x,\lambda))dx = \int_{-1}^{1} f_u(x,\alpha)dx > 0.$$

Hence

$$\mu_1(\lambda,\phi) > 0 \qquad \text{for } \lambda \text{ small},$$

which implies - by Theorem 2 - the desired result.

REMARK. As in Case I, if $a(x)$ is a nonincreasing function, $\phi(x,\lambda)$ is strictly increasing in x, and is therefore a cline. In fact, it can be shown that any nontrivial solution $\phi(\cdot,\lambda)$, which is connected to $(0,\alpha)$ by a continuous branch in $\mathbb{R} \times H^1$ is a cline.

For large values of λ, the situation is more complicated. Assuming again that $a'(x) < 0$ it can be shown by means of the method of super and subsolutions, and a shooting method, that Problem A has at least three strictly increasing solutions $\phi_1 < \phi_2 < \phi_3$. It appears that, for large

values of λ, these are the only clines.

To conclude this section we shall show how the unique cline ϕ, which exists for small values of λ, and the three clines ϕ_i $(i = 1,2,3)$ which exists for large values of λ, fit together in a simple example.

Suppose the function $a(x)$ is given by

$$a(x) = \begin{cases} 1-a & -1 \le x < 0 \\ a & 0 < x \le 1, \end{cases}$$

where $0 < a < \frac{1}{2}$. Then $a(-x) = 1-a(x)$ for $-1 \le x \le 1$ and hence, for small values of λ, the unique cline $\phi(\cdot,\lambda)$ must have the same symmetry property:

$$\phi(-x,\lambda) = 1 - \phi(x,\lambda) \qquad -1 \le x \le 1. \tag{14}$$

In fact one can prove the following result:

THEOREM 6. *For each $\lambda > 0$, Problem A has a unique cline $\phi(\cdot,\lambda)$ which satisfies* (14), *and has the following properties*

(i) $\phi(\cdot,\lambda)\colon \mathbb{R}^+ \to H^1(-1,1)$ *is analytic;*

(ii) $\mu_1(\lambda,\phi) > 0$ *for small values of* λ;

(iii) $\mu_1(\lambda,\phi) < 0$ *for large values of* λ.

Set

$$S = \{(\lambda,\phi) : \lambda > 0,\ \phi \text{ symmetric}\},$$

where we say that ϕ is symmetric if it satisfies (14). Then it follows from Theorem 6 and the analyticity of $\mu_1(\lambda,\phi)$, that there exists a point $(\lambda_o,\phi_o) \in S$ where $\mu_1(\lambda,\phi)$ changes sign. This implies that (λ_o,ϕ_o) is a bifurcation point, from which two new clines emerge, thus accounting for the three clines found for large values of λ.

REFERENCES

1. AMANN, H., *Existence of multiple solutions for nonlinear elliptic boundary value problems*, Indiana Univ.Math.J. 21 (1972), 925-935.

2. CLÉMENT, Ph., Private communication.

3. CONLEY, C., *An application of Wazewski's method to a nonlinear boundary value problem which arises in population genetics*, J.Math.Biol. 2 (1975),241-249.

4. FIFE, P.C., *Fisher's nonlinear diffusion equation and selection-migration models*, Rocky Mtn J.of Math., to appear.

5. FIFE, P.C. & L.A. PELETIER, *Nonlinear diffusion in population genetics*, Arch.Rat.Mech.Anal. 64 (1977),93-109.

6. FISHER, R.A., *The wave of advance of advantageous genes*, Ann. of Eugenics 7 (1937),355-369.

7. FLEMING, W.H., *A selection-migration model in population genetics*, J.Math.Biol. 2 (1975),219-233.

8. HALDANE, J.B.S., *The theory of a cline*, J. Genetics 48 (1948), 277-284.

9. HENRY, D., Private communication.

10. HOPPENSTEADT, F.C., *Analysis of a stable polymorphism arising in a selection-migration model in population genetics dispersion and selection*, J.Math. Biol. 2 (1975),235-240.

11. HUXLEY, J.S., *Clines: an auxiliary taxonomic principle*, Nature 142 (1938) 219 and also *Clines: an auxiliary method in taxonomy*, Bijdr. Dierk. 27(1939) 491.

12. NAGYLAKI, T., *Conditions for the existence of clines*, Genetics 80 (1975), 595-615.

13. NAGYLAKI, T., *Clines with variable migration*, 83 (1976), 867-886.

14. PELETIER, L.A., *A nonlinear eigenvalue problem occurring in population genetics*, Proceedings "Journées d'Analyse non lineaire de Besançon", 1977. To appear in Lecture Notes in Mathematics, Springer.

15. SATTINGER, D.H., *Stability of bifurcating solutions by Leray-Schauder degree*, Arch.Rat.Mech.Anal. 43 (1971),154-166.

VAN DANTZIG'S COLLECTIVE MARKS REVISITED

J.Th. Runnenburg

0. INTRODUCTION

As this is an occasion for looking back, I decided I would have another look at the propagation of Van Dantzig's "method of collective marks" through the mathematical community. Certain problems in probability theory can be solved very elegantly by that method. Still only very few people seem to use collective marks, so too few people know about them. This may partly be due to the fact that VAN DANTZIG [1947] is in Dutch and that VAN DANTZIG [1949], [1955] and [1957] as well as VAN DANTZIG & ZOUTENDIJK [1959] are in French.

Here I discuss a new set of (not-new) examples illustrating the method (see RUNNENBURG [1965] for another set), not hesitating to show some of the shortcomings as well. I still consider it a very fruitful occupation to work through a number of these examples.

1. FLOWGRAPHS

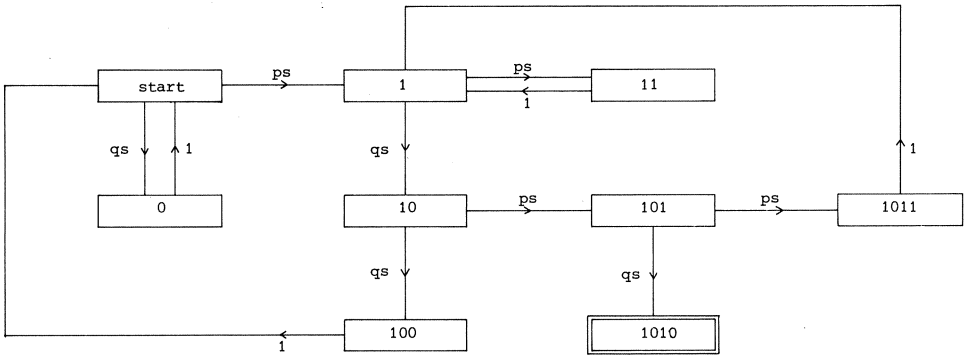
In RÅDE [1972] collective marks and flowgraphs are combined into an attractive amalgam. Just one of the simpler examples (starting on page 162 of the book) is described here, to whet your appetite.

Suppose we perform a series of independent Bernoulli trials, each with outcome 0 with probability q or outcome 1 with probability p, so $p+q = 1$. We want to find the generating function of f_n, the probability of obtaining the (complete) pattern 1010 for the first time at the n-th experiment. The collective mark approach in this case is the following. Introduce a second series of independent Bernoulli trials, now each with outcome "no catastrophe" with probability s or "catastrophe" with probability $1 - s$. Do this in such a way, that all trials involved are independent as well.

Call a trial of the first kind a "trial" and a trial of the second kind a "toss". Let us further agree to have a "toss" after each "trial" and to stop experimenting after we have performed the "toss" following the "trial" giving us the completed pattern 1010 for the first time. Then the quantity we are interested in is

$$f(s) = \sum_{n=1}^{\infty} f_n s^n, \tag{1.1}$$

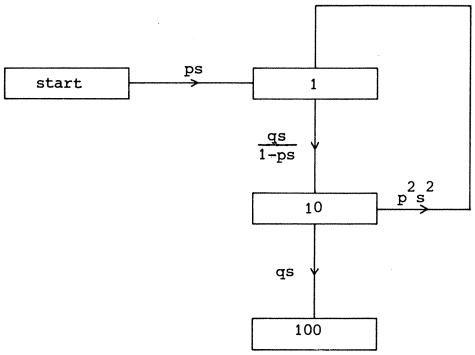
the probability of obtaining our pattern without having a catastrophe. The growth of the pattern, or the successive stages leading to the completion of the pattern, can be described in a figure called a flowgraph.



Here we have registered the last relevant outcomes from the series of "trials". If the last two trials have resulted in 10 say (and the pattern has not yet been completed), then at the next trial we reach 101 (without a catastrophe) with probability ps or 100 (without a catastrophe) with probability qs. In the first case we are nearer our (specially marked) goal, in the second case we have failed to complete the pattern and are back where we started from, in the literal sense. The arrows indicate the possible development and the added probabilities are the probabilities of making the move along that arrow without having a catastrophe on that move. The 1's near the arrows can be read "we are really back at".

Clearly in moving along the arrows from "start" we either first reach 0 or we first reach 1, in the second case we either first reach 100 or we first reach 1010. The probability of at once going to 0 and then reaching 1010 without a catastrophe is qs f(s). The probability of getting to 100 for the first time without going through 0 and without having a catastrophe

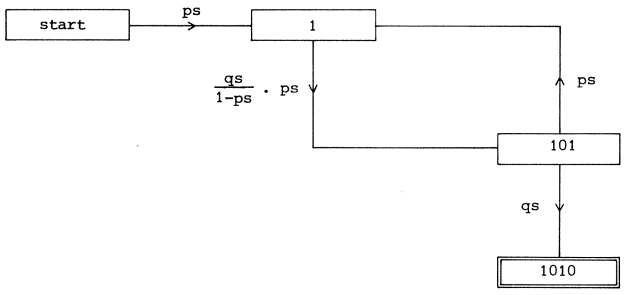
can be obtained by simplifying the relevant part of the first flowgraph to



and reading off for this probability

$$(1.2) \qquad ps \cdot \frac{qs}{1-ps}\left(1 + p^2s^2 \frac{qs}{1-ps} + \ldots\right) \cdot qs = \frac{pq^2s^3}{1-ps-p^2qs^3} .$$

The probability of reaching 1010 without a catastrophe and without going through 0 or 100 follows from the simplification



and is seen to be

$$(1.3) \qquad ps \cdot \frac{\frac{qs}{1-ps} \cdot ps}{1 - \frac{qs}{1-ps} \cdot ps \cdot ps} \cdot qs = \frac{p^2q^2s^4}{1-ps-p^2qs^3} .$$

But then we have

$$(1.4) \qquad f(s) = qs\, f(s) + \frac{pq^2s^3}{1-ps-p^2qs^3} f(s) + \frac{p^2q^2s^4}{1-ps-p^2qs^3} ,$$

from which f(s) is easily obtained.

REMARKS. Flowgraphs are used in ENGEL [1976], HOWARD [1971], LORENS [1964], MASON & ZIMMERMANN [1960], MURPHY [1957] as well. The advantage of interpreting the generating variable s as a probability clearly lies a) in the cementing together of the f_n to one quantity f(s) with a well-defined probabilistic meaning and b) in the economy of expression in the explanation of the use of the flowgraph. We have here an illustration of the statement in VAN DANTZIG [1957]: "En 1946, 1947 nous avons exposé les principes d'une méthode qui admet parfois d'obtenir une fonction génératrice (ou aussi une fonction caractéristique à argument réel) sans recourir d'abord à une équation récurrente.

L'aspect peut-être le plus caractéristique de notre méthode consiste en une interprétation probabiliste des variables auxiliaires intervenants dans la fonction génératrice, et de cette fonction elle-même."

In this example we have met with an extreme in the sense that where elsewhere most of the derivation-through-interpretation is given in words, here much is read off from a picture.

The expressions of disapproval Råde has had to listen to in discussing the above and related examples ("this is only a didactic trick") seem very short-sighted. In ENGEL [1976] flowgraphs with probabilistic interpretation are used in a textbook aimed at "senior high school and beyond". The collective marks approach is introduced on page 67 and occurs in a very natural way.

2. RUNS

In VAN DANTZIG [1949] the method of collective marks was published for the first time. Here the name arose quite naturally. Van Dantzig was very much aware of the inappropriateness of the term "collective marks" in certain contexts. Unfortunately I do not remember the new name he thought of. Råde uses something like "method of the extra event" and that is much more descriptive and indicative of what we are dealing with.

In VAN DANTZIG [1949] several "collective marks" are introduced. Among them the following one. Consider a probability space $(\Omega,\mathcal{F},P)$ and a finite number of dissections of Ω in measurable sets, i.e. for $m \in \{1,2,\ldots,r\}$ with r a positive integer,

(2.1) $\Omega = \bigcup_{n=1}^{\infty} B_{mn}$ with disjoint $B_{m1}, B_{m2}, \ldots \in \mathcal{F}$.

Take (see Note at the end of this paper)

(2.2) $p_{n_1,n_2,\ldots,n_r} = P(B_{1n_1} \cap B_{2n_2} \cap \ldots \cap B_{rn_r})$

and associate with every pair (m,n) a lottery from which the result "catastrophe E_m" is obtained with probability $1 - s_{mn}$ and hence "no catastrophe E_m" with probability s_{mn}. Catastrophes occur independently from one another and from the B_{mn}. Take a ticket from lottery (m,n) only if B_{mn} occurs. Then

(2.3) $$C \overset{\text{def}}{=} \sum_{n_1=1}^{\infty} \sum_{n_2=1}^{\infty} \ldots \sum_{n_r=1}^{\infty} p_{n_1,n_2,\ldots,n_r} \prod_{m=1}^{r} s_{mn_m}$$

is the probability of "no catastrophe of any kind" associated with taking a random element from this probability space. C is the "collective mark" of the $p_{n_1,n_2,\ldots,n_r}$ or we can also say the "mark" of the "collection" of subsets $\bigcap_{m=1}^{r} B_{mn_m}$ of Ω (provided we let the s_{mn} range independently over the real interval $[0,1]$).

On these collective marks one can operate in several ways: one can demark, substitute or draw. We shall not go into these possibilities (only dealt with in detail in the 1949 paper), but just consider the main example from that paper. Unfortunately this application is incorrect as it stands, in the sense that the final formula ((65) on page 40) is correct, but has been obtained by an incorrect interpretation (R_λ as defined above (59) seems to have no meaning whatsoever, relation (59) with R_λ as defined in (60) is the correct quantity we need analytically, the first relation after (61) is trivially not equivalent to (60) though it should be and the sentence containing that relation has other incorrect and incomplete information). In VAN DANTZIG & ZOUTENDIJK [1959] the same derivation is needed again in a more general context and now a different interpretation is used. In a footnote Van Dantzig points out that (59) and R_λ in the 1949 paper have not been interpreted correctly. One can hardly hope for a more convincing warning concerning the possible pitfalls in using the method of collective marks!

Let us now first describe and then solve the problem in question in the traditional way, using mainly a choice of the notations from VAN DANTZIG

[1949] and VAN DANTZIG & ZOUTENDIJK [1959].

We are concerned with a series of n independent experiments, where each experiment leads to just one of the marks (or outcomes) A_λ (with $\lambda \in \{1,2,\ldots,k\}$ for a finite integer k) with probability p_λ (hence $\Sigma_{\lambda=1}^{k} p_\lambda = 1$). Say on carrying out these experiments we obtain first a_1 times A_{λ_1}, then a_2 times A_{λ_2},... and finally a_r times A_{λ_r} with $\lambda_1 \neq \lambda_2 \neq \ldots \neq \lambda_r$ (although we can have $\lambda_i = \lambda_3$) where $1 \leq r \leq n$ and $\Sigma_{\rho=1}^{r} a_\rho = n$. This description is in terms of the complete iterations or runs we have obtained in our series of n experiments, where the first run consists of a_1 marks A_{λ_1}, the second of a_2 marks A_{λ_2}, etc. Say in all we obtain $m_{\lambda\ell}$ runs of A_λ of length ℓ, where

$$(2.4) \qquad m_{\lambda\ell} \geq 0 \quad \text{and} \quad \sum_{\lambda=1}^{k} \sum_{\ell=1}^{n} \ell m_{\lambda\ell} = n.$$

These same runs can occur in a different order. We are interested in the probabilities

$$(2.5) \qquad p_n(\ldots,m_{\lambda\ell},\ldots) \overset{abb}{=} p_n(m_{11},m_{21},\ldots,m_{k1},m_{12},m_{22},\ldots,m_{1n},m_{2n},\ldots,m_{kn})$$

of obtaining prescribed runs in arbitrary order. It is convenient to introduce a generating function

$$(2.6) \qquad C_n^* \overset{abb}{=} C_n^*(\ldots,s_{\lambda\ell},\ldots) \overset{def}{=} \sum_{m_{\lambda\ell}} p_n(\ldots,m_{\lambda\ell},\ldots) \prod_{\lambda=1}^{k} \prod_{\ell=1}^{n} s_{\lambda\ell}^{m_{\lambda\ell}},$$

where $s_{\lambda\ell}^{m_{\lambda\ell}}$ is the $m_{\lambda\ell}$-th power of a complex variable $s_{\lambda\ell}$ with $|s_{\lambda\ell}| \leq 1$ and $\Sigma_{m_{\lambda\ell}}$ stands for summation over all (sets of indices) $m_{\lambda\ell}$ satisfying (2.4). It is easier to take a generating function with respect to n as well, so we further introduce

$$(2.7) \qquad C^* \overset{abb}{=} C^*(t;\ldots,s_{\lambda\ell},\ldots) \overset{def}{=} 1 + \sum_{n=1}^{\infty} t^n C_n^*(\ldots,s_{\lambda\ell},\ldots),$$

where t is a complex variable with $|t| < 1$ (so we are sure of absolute convergence in (2.7) as the absolute value of C_n^* is at most 1). We are in fact trying to find (reordering the terms in (2.7) and returning to the description with successive runs)

$$(2.8)\qquad C^* = 1 + \sum_{n=1}^{\infty} \sum_{r=1}^{\infty} \overline{\sum_{\lambda_1 \neq \lambda_2 \neq \ldots \neq \lambda_r}} \overline{\sum_{\substack{a_1 \geq 1\, a_2 \geq 1, \ldots, a_r \geq 1 \\ a_1 + a_2 + \ldots + a_r = n}}} \left(p_{\lambda_1}^{a_1} t^{a_1} s_{\lambda_1 a_1}\right)\left(p_{\lambda_2}^{a_2} t^{a_2} s_{\lambda_2 a_2}\right) \ldots \left(p_{\lambda_r}^{a_r} t^{a_r} s_{\lambda_r a_r}\right).$$

Introducing

$$(2.9)\qquad \Phi_\lambda \overset{\text{def}}{=} \sum_{n=1}^{\infty} (p_\lambda t)^n s_{\lambda n},$$

we can rewrite (2.8) (after changing the order of summation)

$$(2.10)\qquad C^* = 1 + \sum_{r=1}^{\infty} \overline{\sum_{\lambda_1 \neq \lambda_2 \neq \ldots \neq \lambda_r}} \Phi_{\lambda_1} \Phi_{\lambda_2} \ldots \Phi_{\lambda_r}.$$

Clearly (collecting terms starting with Φ_λ)

$$(2.11)\qquad C^* = 1 + \sum_{\lambda=1}^{k} \Phi_\lambda D_\lambda,$$

where

$$(2.12)\qquad D_\lambda \overset{\text{def}}{=} 1 + \sum_{r=2}^{\infty} \overline{\sum_{\lambda \neq \lambda_2 \neq \lambda_3 \neq \ldots \neq \lambda_r}} \Phi_{\lambda_2} \Phi_{\lambda_3} \ldots \Phi_{\lambda_r}.$$

But then (collecting terms starting with Φ_μ)

$$(2.13)\qquad D_\lambda = 1 + \sum_{\mu \neq \lambda} \Phi_\mu D_\mu.$$

Now (using (2.13) in (2.11)) for each λ

$$(2.14)\qquad C^* = (1+\Phi_\lambda) D_\lambda$$

and so (using (2.14) in (2.11) to eliminate D_λ)

$$(2.15)\qquad C^* = \left(1 - \sum_{\lambda=1}^{k} \frac{\Phi_\lambda}{1+\Phi_\lambda}\right)^{-1}.$$

If we want to derive (2.15) with the method of collective marks, all we have to do is find a suitable interpretation for (2.11) and (2.13). As (2.8) is hard to interpret as it stands, we multiply both sides with $1 - t$ and take

(2.16) $\quad C = (1-t)C^*.$

We imagine that each time before we do an experiment (producing A_λ with probability p_λ) we toss a t-coin to decide whether we do that experiment (with probability t) or stop altogether (with probability 1 - t). We take t and each $s_{\lambda\ell}$ between 0 and 1 (with only t strictly less than 1), so each $s_{\lambda\ell}$ can be interpreted as the probability of the non-occurrence of a catastrophe. We now have a set-up, where C stands for the probability that we don't have a catastrophe in a stochastic process consisting of alternatively tossing a t-coin to decide whether we do one more experiment and doing that experiment in case of a positive decision, continuing in this way till we stop at a negative decision, with the further complication that every time we have completed a run of one kind of outcome of the experiment, say a run of marks A_λ of length ℓ, we toss a $s_{\lambda\ell}$-coin leading to "catastrophe" (with probability $1 - s_{\lambda\ell}$) or "no catastrophe" (with probability $s_{\lambda\ell}$). Of course we assume the independence of all outcomes of experiments and coin tosses. We toss a catastrophe coin every time a run of outcomes has been completed, either by the start of a new run or by the process stopping. If no experiment is done at all, we have no catastrophe with probability 1. This is the 1949 description of the stochastic process we use for the interpretation. The correct 1959 interpretation is now easy to give. Write C_λ for the probability that there is a first experiment producing A_λ and no catastrophe occurs during the process. Then clearly

(2.17) $\quad C = 1 - t + \sum_{\lambda=1}^{k} C_\lambda,$

as we either have no first experiment or a first experiment leading to some A_λ. If there is a first experiment leading to A_λ, then there is either only a first run of A_λ's or there is a first run of A_λ's followed by a second run of A_μ's for some $\mu \neq \lambda$. The probability of having only a single run of A_λ's without a catastrophe is $(1-t)\Phi_\lambda$, while the probability of having a first run of A_λ's, a second run of A_μ's for a fixed $\mu \neq \lambda$ and no catastrophe in the process is $\Phi_\lambda C_\mu$. Hence

(2.18) $\quad C_\lambda = \Phi_\lambda(1-t + \sum_{\mu\neq\lambda} C_\mu).$

The derivation through interpretation has now been obtained (without (2.8), (2.10) and (2.12)). If we write $(1-t)\Phi_\lambda D_\lambda$ for C_λ, then the difference between (2.11) and (2.17) as well as that between (2.13) and (2.18) disappears.

The main step in the interpretation is the assertion that a certain probability is given by $\Phi_\lambda C_\mu$. I find it hard to be absolutely certain of this fact without in some way, however superficial, passing through the first derivation. Having done that once, I have no difficulty in accepting the second derivation as the full story from then on and presenting it as such to an audience. Perhaps you should not try to learn to write shorthand until you have learned to spell. Van Dantzig could write shorthand.

After I had decided to include the main 1949 example in this paper, I discovered the inconsistency in that example. Before trying to rescue it I obtained (2.15) by the analytic argument I described. It showed the correctness of the main formula for this example in the 1949 paper. I then remembered that the interpretation might be given again in the Van Dantzig and Zoutendijk paper. I checked with a version of the manuscript and found the error repeated. I had a look at a reprint to see where the article was published and then noticed the very elegant new interpretation in the final version. After first changing the present analytic derivation so it differed only by a factor $1 - t$ in some places from the second derivation, I changed it back to the original one as the natural one (for me). The analytic derivation can be interpreted too, but leads to such complicated formulations that I intended to use it to show the superiority of the analytic derivation over the probabilistic one in this case (until I saw the 1959 reprint).

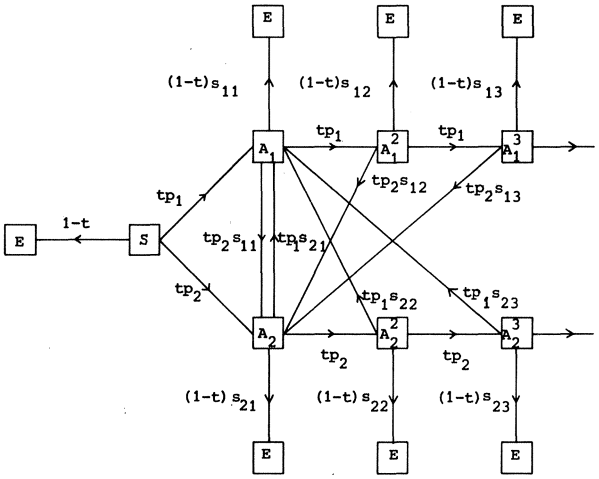
One can also use a flowgraph approach to obtain C. This leads to

$$C = 1 - t + \sum_{\lambda=1}^{k} tp_\lambda B_\lambda, \tag{2.19}$$

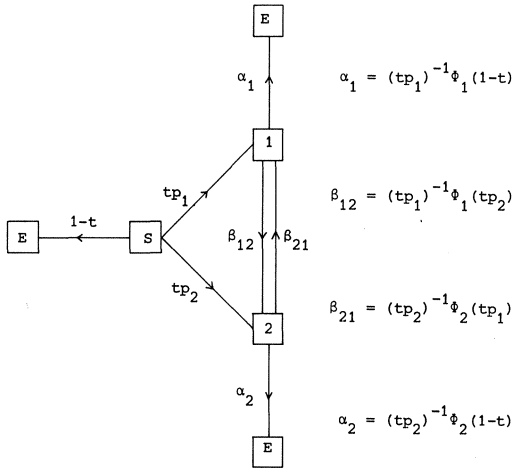
where the probability B_λ satisfies

$$B_\lambda = (tp_\lambda)^{-1}\Phi_\lambda(1-t) + \sum_{\mu\neq\lambda} (tp_\lambda)^{-1}\Phi_\lambda(tp_\mu)B_\mu, \tag{2.20}$$

or with $C_\lambda = tp_\lambda B_\lambda$ we have rederived (2.17) and (2.18). To give some indication of how these relations come about, consider the next flowgraph, where for simplicity we have taken $k = 2$. Here we use S for start and E for end. Again we have to reach (some) E from S (without a catastrophe). Notice that once the flowgraph is drawn with the correct probabilities added, we can forget about catastrophes: we only have to go from S to E.



This flowgraph can be simplified to the next one (we are lumping states in a Markov chain, see KEMENY & SNELL [1960]).



REMARKS. The problem in this section was studied in FRÉCHET [1940, 1943] to show that it is not at all unusual to find a run of one sex of length 17 in the registration of 200,000 births and that this fact certainly does not prove the inapplicability of probability theory to this kind of data (as maintained by Marbe). Whereas Fréchet obtained expectation and variance of the number of runs of length ℓ of one kind in n experiments, where each

experiment is an alternative, and conjectured that this number has a Poisson limiting distribution if ℓ and n tend to infinity in a suitable way, Van Dantzig proved this conjecture and much more on the basis of (2.15).

According to GOOD [1973] the relation (2.15) was obtained by Gontcharov in 1943 for alternatives (k=2). According to FELLER [1968], page 305 "The classical theory of runs was rather messy,..." and he therefore advocates his way of counting runs, which is quite different from what we did. One should reexamine Fréchet's conclusions on the basis of Feller's theory.

3. QUEUING

Let $\underline{x}$ be a non-negative random variable with distribution function F and take

$$(3.1) \qquad \phi(\xi) = \int_{[0,\infty)} e^{-\xi x} dF(x) \quad \text{for} \quad \xi \geq 0.$$

Assume that a stationary Poisson process (independent of $\underline{x}$) with intensity $\lambda > 0$ produces events at times $\underline{y}_1, \underline{y}_1+\underline{y}_2, \ldots$. Then $\underline{y}_1, \underline{y}_2, \ldots$ are independent and exponential with parameter λ and so $\phi(\lambda)$ stands for the probability that no event occurs in $[0,\underline{x}]$. To find the generating function

$$(3.2) \qquad f(s) = \sum_{n=0}^{\infty} f_n s^n \quad \text{for} \quad 0 \leq s \leq 1,$$

where f_n is the probability that exactly n events occur in $[0,\underline{x}]$, we can observe that

$$(3.3) \qquad f_n = \int_{[0,\infty)} e^{-\lambda x} \frac{(\lambda x)^n}{n!} dF(x),$$

because conditional on $\underline{x} = x$ we have with probability

$$(3.4) \qquad e^{-\lambda x} \frac{(\lambda x)^n}{n!}$$

exactly n events in $[0,\underline{x}]$ and hence

$$(3.5) \qquad f(s) = \sum_{n=0}^{\infty} \int_{[0,\infty)} e^{-\lambda x} \frac{(\lambda x)^n}{n!} s^n dF(x) = \phi(\lambda(1-s)).$$

Alternatively we can toss a coin every time an event occurs to decide whether that event is a marked one (with probability 1 - s) or an unmarked one (with probability s). Marking is done independently of $\underline{x}$ and the Poisson process and independently for different events. Now f(s) is the probability, that during $[0,\underline{x}]$ no marked event occurs. Marked events occur according to a Poisson process with intensity $\lambda(1-s)$. This must be proved and quickly, as otherwise we won't find a short(er) proof for (3.5) by probabilistic interpretation. Let us assume it is evident. Now we are back at the problem of finding the probability that no event occurs in $[0,\underline{x}]$, this time from a Poisson process with intensity $\lambda(1-s)$. Hence

$$f(s) = \phi(\lambda(1-s)). \tag{3.6}$$

This kind of derivation is nice, but also rather superficial. For this and similar reasons YADIN [1970] is not a serious contribution to the method of collective marks. This is not a criticism of the paper as such, which is indeed a fine piece of work. It would qualify, if we were ready to call any paper in which difference equations and differential difference equations are avoided in the determination of generating functions a paper in which the method of collective marks is used.

Next consider a passage from KINGMAN [1966] (relation (63) and beyond). The author states very definitely that he is here using the method of collective marks. And he is, in the sense that he introduces an extra event in his original problem in order to get more information about the old problem. Consider the GI/G/1 queue: customers 1,2,... arrive at a single counter to be served at times $0,\underline{y}_1,\underline{y}_1+\underline{y}_2,\ldots$, where $\underline{y}_1,\underline{y}_2,\ldots$ are non-negative random variables with the same distribution function A. The service times needed for the successive customers are $\underline{s}_1,\underline{s}_2,\ldots$, also non-negative random variables with a common distribution function B. Assume that $\underline{y}_1,\underline{s}_1,\underline{y}_2,\underline{s}_2,\ldots$ are independent. Write $\underline{u}_n$ for $\underline{s}_n-\underline{y}_n$, $\underline{t}_n$ for $\underline{s}_1+\underline{s}_2+\ldots+\underline{s}_n$ and $\underline{v}_n$ for $\underline{u}_1+\underline{u}_2+\ldots+\underline{u}_n$. Then

$$b_n \overset{abb}{=} P\{\underline{v}_1 > 0, \underline{v}_2 > 0,\ldots,\underline{v}_{n-1} > 0,\underline{v}_n \leq 0\} \tag{3.7}$$

is the probability that in the first busy period exactly n customers are served. Write $\underline{z}$ for the length of the first busy period and $\underline{n}$ for the number of customers served in that period, then $\underline{z} = \underline{t}_n$ if $\underline{n} = n$. Kingman proves

$$(3.8)\qquad \sum_{n=1}^{\infty} b_n x^n = 1 - \exp\Big(- \sum_{n=1}^{\infty} \frac{x^n}{n} P\{\underline{v}_n \le 0\}\Big)$$

and remarks that the same relation holds (with $\underline{s}_n$ replaced by $\underline{s}_n^o, \underline{u}_n$ replaced by $\underline{u}_n^o$, b_n replaced by b_n^o, etc.), if we start from non-negative random variables $\underline{s}_1^o, \underline{s}_2^o, \ldots$ that assume the value ∞ with positive probability (and have a common distribution, etc.).

Now introduce the possibility that the server may die while working, into the problem with $\underline{s}_1, \underline{s}_2, \ldots$. Let his death occur at the first event from a stationary Poisson process (independent of the queuing process) with intensity $\theta > 0$, that falls in a service period. The duration of that period is then ∞. This changes the original service times $\underline{s}_1, \underline{s}_2, \ldots$ to $\underline{s}_1^o, \underline{s}_2^o, \ldots$ with, for finite non-negative $s_1, s_2, \ldots, s_n$,

$$(3.9)\qquad P\{\underline{s}_1^o \le s_1, \underline{s}_2^o \le s_2, \ldots, \underline{s}_n^o \le s_n\} = \prod_{k=1}^{n} \int_{[0,s_k]} e^{-\theta s} dB(s).$$

We use in the ^{o} version of (3.8) only $\underline{s}_1^o, \underline{s}_2^o, \ldots$ restricted to finite values, so there is no harm in assuming $\underline{s}_1^o, \underline{s}_2^o, \ldots$ independent with

$$(3.10)\qquad P\{\underline{s}_n = \infty\} = 1 - \int_{[0,\infty)} e^{-\theta s} dB(s).$$

Now with $\underline{a}_\theta$ exponential with parameter θ and $\underline{a}_\theta, \underline{y}_1, \underline{s}_1, \underline{y}_2, \underline{s}_2, \ldots$ independent, we have (with $\chi_A(w) = 1$ for $w \in A$ and $\chi_A(w) = 0$ otherwise)

$$(3.11)\qquad P\{\underline{v}_n^o \le 0\} = P\{\underline{v}_n \le 0, \underline{t}_n \le \underline{a}_\theta\} = E e^{-\theta \underline{t}_n} \chi_{\{\underline{v}_n \le 0\}}$$

and

$$(3.12)\qquad b_n^o = P\{\underline{v}_1^o > 0, \underline{v}_2^o > 0, \ldots, \underline{v}_{n-1}^o > 0, \underline{v}_n^o \le 0\} =$$

$$= P\{\underline{v}_1 > 0, \underline{v}_2 > 0, \ldots, \underline{v}_{n-1} > 0, \underline{v}_n \le 0, \underline{t}_n \le \underline{a}_\theta\} =$$

$$= E e^{-\theta \underline{t}_n} \chi_{\{\underline{n} = n\}} = E e^{-\theta \underline{z}} \chi_{\{\underline{n} = n\}}.$$

The ^{o} version of (3.8) can now be written

$$(3.13)\qquad E x^{\underline{n}} e^{-\theta \underline{z}} = 1 - \exp\Big(- \sum_{n=1}^{\infty} \frac{x^n}{n} E e^{-\theta \underline{t}_n} \chi_{\{\underline{v}_n \le 0\}}\Big),$$

so in principle we now know the joint distribution of $\underline{n}$ and $\underline{z}$, whereas with (3.8) we only knew

$$(3.14) \qquad Ex^{\underline{n}} = 1 - \exp\Big(- \sum_{n=1}^{\infty} \frac{x^n}{n} P\{\underline{v}_n \leq 0\}\Big)$$

or the distribution of $\underline{n}$ alone.

REMARKS. Yadin told me that his paper was originally written without any attempt to use collective marks. At Neuts' suggestion he changed to the published version.

Kingman's construction is a beautiful example of operating on a given collective mark to get a more general one.

4. MORE QUEUING

D. Sikkel, one of my students, was able to translate the first chapter of Service systems with priorities by Gnedenko, Danieljan, Dimitrov, Klimov and Matvejev (in Russian). I obtained a copy of the book through J.C. Smit, Nijmegen. I am very grateful to the latter for drawing my attention to this book and to the fact that it contains applications of the method of collective marks.

So far I only know the contents of the first chapter. This part is promising, but not entirely satisfying. The explanations of interpretations are given in such a way, that much is left to the reader. In the first chapter a study is made of what happens in a busy period in an M/G/1 queue. In RUNNENBURG [1965] a result of Wishart for this busy period was derived by interpretation and it was intriguing to see that in the book slightly different aspects of the same basic situation had been focused on.

A combination of all earlier results proved possible and is contained in SIKKEL [1975], a final exam paper. I include most of the proof of the general result here, both to have it published and to demonstrate that interpretation and classical use of non-interpreted (maybe non-interpretable) summations and integrations can be combined to give seemingly complicated results in a fairly simple way.

Assume that customers arrive according to a stationary Poisson process with intensity $\lambda > 0$ at one counter to be served, where the service times $\underline{s}_1, \underline{s}_2, \ldots$ of customers $1,2,\ldots$ are independent non-negative random variables with the common distribution function B and Laplace-Stieltjes

transform β with

$$(4.1) \qquad \beta(\xi) = \underline{E}e^{-\xi \underline{s}_n} \quad \text{for} \quad \xi \geq 0.$$

Take $\underline{t}_m = \underline{s}_1 + \underline{s}_2 + \ldots + \underline{s}_m$. Arrival intervals and service times are independent. We start at time 0 with customer 1 present and his service period just starting. The server will remain at work from time 0 onwards, taking care (in some order) of all newly arriving customers one at a time till time $\underline{z}$, the end of the first busy period, at which moment he becomes idle for the first time (because there are no more waiting customers). During the first busy period $\underline{n}$ customers are served.

One would like to know at a time a with $0 < a \leq \underline{z}$

(4.2a) $\underline{m}(a)$ = number of customers fully served in $[0,a]$,

(4.2b) $\underline{z}'(a)$ = time needed for serving these $\underline{m}(a)$ customers,

(4.2c) $\underline{s}'(a) = a - \underline{z}'(a)$,

(4.2d) $\underline{s}''(a)$ = remaining service time of the customer being served at a,

(4.2e) $\underline{z}''(a) = \underline{z} - \underline{z}'(a) - \underline{s}'(a) - \underline{s}''(a)$,

(4.2f) $\underline{k}_1(a)$ = number of customers arriving in $(0,\underline{z}'(a)]$ less $(\underline{m}(a)-1)$,

(4.2g) $\underline{k}_2(a)$ = number of customers arriving in $(a-\underline{s}'(a),a]$,

(4.2h) $\underline{k}_3(a)$ = number of customers arriving in $(a,a+\underline{s}''(a)]$,

(4.2i) $\underline{k}_4(a)$ = number of customers arriving in $(\underline{z}-\underline{z}''(a),\underline{z}]$.

It seems rather hopeless to get information about all these random variables simultaneously. It is possible if we replace a by $\underline{a}_\xi$, where $\underline{a}_\xi$ has an exponential distribution (independent of the arrival intervals and service times) with parameter $\xi > 0$. We need five different markings for the customers: with probability X a customer doesn't have mark 0, with probability Y_1 (Y_2,Y_3,Y_4) he doesn't have mark 1 (2,3,4). These will lead to geometrically distributed random variables $\underline{r}_X$,$\underline{r}_{Y_1}$,$\underline{r}_{Y_2}$,$\underline{r}_{Y_3}$ and $\underline{r}_{Y_4}$ with

$$(4.3) \qquad P\{\underline{r}_X = r\} = X^r(1-X)$$

and likewise for the Y's. We also need four different kinds of catastrophe. Each kind is produced by a stationary Poisson process, the respective intensities are $\zeta_1,\eta_1,\eta_2,\zeta_2$. The first kind we call ζ_1-catastrophe, etc. We shall deal with these catastrophes through exponentially distributed random variables $\underline{a}_{\zeta_1},\underline{a}_{\eta_1},\underline{a}_{\eta_2},\underline{a}_{\zeta_2}$ with parameters $\zeta_1,\eta_1,\eta_2,\zeta_2$ respectively. All markings and Poisson processes are independent of one another and of the queuing process and $\underline{a}_\xi$. We shall write $\underline{m},\underline{z}',\ldots,\underline{k}_4$ for $\underline{m}(\underline{a}_\xi)$, $\underline{z}'(\underline{a}_\xi),\ldots,\underline{k}_4(\underline{a}_\xi)$ and call the $\underline{m}$ customers fully served before $\underline{a}_\xi$ (with $\underline{a}_\xi \le \underline{z}$) the $\underline{m}$-customers, etc. The $\underline{k}_1$-customers consist of the $(\underline{m}+1)^{st}$ customer (= the customer being served at time $\underline{a}_\xi$) plus the customers waiting just after the servicing of that customer has started. Hence $\underline{k}_1 \ge 1$.

Let now $\pi(\xi,\zeta_1,\eta_1,\eta_2,\zeta_2,X,Y_1,Y_2,Y_3,Y_4)$ denote the probability that $\underline{a}_\xi \le \underline{z}$, none of the $\underline{m}$-customers has mark 0, none of the $\underline{k}_i$-customers has mark i (for $i \in \{1,2,3,4\}$) and no ζ_1-catastrophe occurs during $(0,\underline{z}']$, no η_1-catastrophe during $(\underline{a}_\xi-\underline{s}',\underline{a}_\xi]$, no η_2-catastrophe during $(\underline{a}_\xi,\underline{a}_\xi+\underline{s}"]$ and no ζ_2-catastrophe during $(\underline{z}-\underline{z}",\underline{z}]$. This means that with independent $\underline{a}_{\zeta_1},\underline{a}_{\eta_1},\underline{a}_{\eta_2},\underline{a}_{\zeta_2},\underline{r}_X,\underline{r}_{Y_1},\underline{r}_{Y_2},\underline{r}_{Y_3},\underline{r}_{Y_4}$ and $(\underline{a}_\xi,\underline{z},\underline{z}',\underline{s}',\underline{s}",\underline{z}",\underline{m},\underline{k}_1,\underline{k}_2,\underline{k}_3,\underline{k}_4)$ we can write

$$\pi(\xi,\zeta_1,\eta_1,\eta_2,\zeta_2,X,Y_1,Y_2,Y_3,Y_4) = \tag{4.4}$$

$$= P\{\underline{z} \ge \underline{a}_\xi,\underline{z}' < \underline{a}_{\zeta_1},\underline{s}' < \underline{a}_{\eta_1},\underline{s}" < \underline{a}_{\eta_2},\underline{z}" < \underline{a}_{\zeta_2},\underline{m} \le \underline{r}_X,\underline{k}_i \le \underline{r}_{Y_i} \text{ for } 1 \le i \le 4\}.$$

We use the abbreviation

$$A \overset{abb}{=} \{\underline{t}_{\underline{m}} < \underline{a}_\xi \le \underline{t}_{\underline{m}+1},\underline{n} > m,\underline{z}' < \underline{a}_{\zeta_1},\underline{s}' < \underline{a}_{\eta_1},\underline{s}" < \underline{a}_{\eta_2},\underline{m} = m,\underline{k}_j = k_j \text{ for } 1 \le j \le 3\}. \tag{4.5}$$

Write $\delta(\zeta_2,Y_4)$ for the probability, that no Y_4-marked customer arrives and no ζ_2-catastrophe occurs during a busy period starting at time 0 with one customer present, just about to start being served at that time. Then

$$P(\{\underline{z}" < \underline{a}_{\zeta_2},\ \underline{k}_4 \le \underline{r}_{Y_4}\}|A) = \delta(\zeta_2,Y_4)^{k_1+k_2+k_3-1}, \tag{4.6}$$

because the conditional probability can be seen to be the probability, that no Y_4-marked customer arrives and no ζ_2-catastrophe occurs during a busy period starting at time 0 with $k_1+k_2+k_3-1 (\ge 0)$ customers present, the first of

which is to begin service at that time. It is well-known that this probability is the $(k_1+k_2+k_3-1)^{th}$ power of $\delta(\zeta_2,Y_4)$. To obtain $\delta(\zeta_2,Y_4)$ (assuming $B(s) < 1$ for an $s > 0$) one can determine the unique $\delta(\xi,X)$ satisfying

$$\delta(\xi,X) = \beta(\xi+\lambda-\lambda X\delta(\xi,X)) \tag{4.7}$$

with $0 < \delta(\xi,X) < 1$ for $\xi > 0$, $0 \le X \le 1$ and $\delta(0,1) = 1$. RUNNENBURG [1965] contains a proof of (4.7) by interpretation, the uniqueness follows easily by comparing the graphs of δ and $\beta(\xi+\lambda-\lambda X\delta)$ for a variable δ with $0 \le \delta \le 1$.

We now write down an expression for the right-hand side of (4.4), in which we first introduce the conditional probability obtained in (4.6) and then carry out the summations, using (4.4). Hence the left-hand side of (4.4) is equal to

$$\begin{aligned}(4.8)\qquad &\sum_{m=0}^{\infty}\sum_{k_1=1}^{\infty}\sum_{k_2=0}^{\infty}\sum_{k_3=0}^{\infty} P(A \cap \{\underline{z}'' < \underline{a}_{\zeta_2},\ \underline{k}_4 \le \underline{r}_{Y_4}\})X^m Y_1^{k_1}Y_2^{k_2}Y_3^{k_3} = \\ &= \sum_{m=0}^{\infty}\sum_{k_1=1}^{\infty}\sum_{k_2=0}^{\infty}\sum_{k_3=0}^{\infty} P(A)X^m Y_1^{k_1}Y_2^{k_2}Y_3^{k_3}\,\delta(\zeta_2,Y_4)^{k_1+k_2+k_3-1} = \\ &= \pi(\xi,\zeta_1,\eta_1,\eta_2,0,X,Y_1\delta(\zeta_2,Y_4),Y_2\delta(\zeta_2,Y_4),Y_3\delta(\zeta_2,Y_4),1)\ \delta(\zeta_2,Y_4)^{-1}.\end{aligned}$$

This means that we need not determine π as a function of all its ten arguments at once, but can first restrict to $\zeta_2 = 0$ and $Y_4 = 1$ in (4.4) and then later use (4.8) to get the general π.

From (4.4) we find, specifying the possible values of $\underline{m}$ and using $\underline{b}_{\xi,m}$ as an abbreviation for $\underline{a}_{\xi}-\underline{t}_m$,

$$\begin{aligned}(4.9)\qquad &\pi(\xi,\zeta_1,\eta_1,\eta_2,0,X,Y_1,Y_2,Y_3,1) = \\ &= \sum_{m=0}^{\infty} P\{\underline{t}_m < \underline{a}_{\xi},\ \underline{n} > m,\ \underline{t}_m < \underline{a}_{\zeta_1},\ m \le \underline{r}_X,\ \underline{k}_{1,m} \le \underline{r}_{Y_1}\}\cdot \\ &\cdot P\{\underline{s}_{m+1} \ge \underline{b}_{\xi,m}, \underline{b}_{\xi,m} < \underline{a}_{\eta_1}, \underline{s}_{m+1}-\underline{b}_{\xi,m} < \underline{a}_{\eta_2}, \underline{k}_2 \le \underline{r}_{Y_2}, \underline{k}_3 \le \underline{r}_{Y_3} \mid B\},\end{aligned}$$

where

$$B = \{\underline{b}_{\xi,m} > 0,\ \underline{n} > m,\ \underline{t}_m < \underline{a}_{\zeta_1},\ m \le \underline{r}_X,\ \underline{k}_{1,m} \le \underline{r}_{Y_1}\}. \tag{4.10}$$

Note that $\underline{k}_1$ has been replaced by $\underline{k}_{1,m}$, where $\underline{k}_{1,m} + m - 1$ is the number of customers arriving in $(0,\underline{t}_m]$ and that $\underline{b}_{\xi,m}$ has an exponential distribution with parameter ξ under the condition $\underline{b}_{\xi,m} > 0$. The conditional probability in (4.9) can be seen to equal the unconditional probability (just ignore B) in which $\underline{b}_{\xi,m}$ is an exponential random variable with parameter ξ and the relations between the random variables $\underline{s}_{m+1}, \underline{b}_{\xi,m}, \underline{a}_{\eta_1}, \underline{a}_{\eta_2}, \underline{k}_2, \underline{r}_{Y_2}, \underline{k}_3, \underline{r}_{Y_3}$ remain as they were. We can now compute this conditional probability (condition on $\underline{b}_{\xi,m} = b$ and $\underline{s}_{m+1} = s$) and find

$$(4.11)\qquad \int_0^\infty \left(\int_0^s e^{-\eta_1 b - \eta_2 (s-b)} e^{-\lambda(1-Y_2)b} e^{-\lambda(1-Y_3)(s-b)} \xi e^{-\xi b} db \right) dB(s) =$$

$$= \frac{\xi}{\xi + \eta_1 - \eta_2 - \lambda Y_2 + \lambda Y_3} \{\beta(\eta_2 + \lambda - \lambda Y_3) - \beta(\xi + \eta_1 + \lambda - \lambda Y_2)\}.$$

Note that (4.11) does not depend on m. The next step in our calculation is then to find

$$(4.12)\qquad \sum_{m=0}^{\infty} P\{\underline{t}_m < \underline{a}_\xi, \underline{n} > m, \underline{t}_m < \underline{a}_{\zeta_1}, m \le \underline{r}_X, \underline{k}_{1,m} \le \underline{r}_{Y_1}\} =$$

$$= \sum_{m=0}^{\infty} X^m P\{\underline{t}_m < \min(\underline{a}_\xi, \underline{a}_{\zeta_1}), \underline{n} > m, \underline{k}_{1,m} \le \underline{r}_{Y_1}\}.$$

In (4.12) we no longer have $\underline{z} \ge \underline{a}_\xi$ in our conditions, but that is no problem as all the random variables $\underline{z}', \underline{s}', \underline{s}'', \underline{z}'', \underline{m}, \underline{k}_1, \underline{k}_2, \underline{k}_3, \underline{k}_4$ depending on $\underline{a}_\xi$ and defined for $\underline{z} \ge \underline{a}_\xi$ only have disappeared as well. From now on we can forget the restriction $\underline{z} \ge \underline{a}_\xi$ and carry out our calculations with the help of $\underline{n}$ = number of customers served in the first busy period, $\underline{t}_m$ = total service time of the first m customers, $\underline{k}_{1,m} + m - 1$ = number of customers arriving in $(0,\underline{t}_m]$, while $\underline{a}_\xi$ and $\underline{a}_{\zeta_1}$ are exponentially distributed with parameters ξ and ζ_1 respectively and $\underline{r}_X$ and $\underline{r}_{Y_1}$ are geometrically distributed with parameters X and Y_1 respectively. Also $\underline{a}_\xi$, $\underline{a}_{\zeta_1}$, $\underline{r}_X$, $\underline{r}_{Y_1}$ and $(\underline{t}_m, \underline{n}, \underline{k}_{1,m})$ are independent.

It is now not hard to realize that for $m = 0,1,2,\ldots$ (with $P\{\underline{t}_0 < \underline{a}_\xi, \underline{n} > 0, \underline{k}_{1,0} \le \underline{r}_{Y_1}\} = Y_1$, as $\underline{k}_{1,0} = 1$)

$$(4.13)\qquad Y_1 P\{\underline{t}_{m+1} < \underline{a}_\xi, \underline{n} \ge m+1, \underline{k}_{1,m+1} \le \underline{r}_{Y_1}\} =$$

$$= P\{\underline{t}_m < \underline{a}_\xi, \underline{n} > m, \underline{k}_{1,m} \le \underline{r}_{Y_1}\} \beta(\xi + \lambda - \lambda Y_1)$$

and also

$$(4.14)\qquad P\{\underline{t}_{m+1} < \underline{a}_\xi,\ \underline{n} \geq m+1,\ \underline{k}_{1,m+1} \leq \underline{r}_{Y_1}\} =$$

$$= P\{\underline{t}_{m+1} < \underline{a}_\xi, \underline{n} > m+1, \underline{k}_{1,m+1} \leq \underline{r}_{Y_1}\} + \delta_{m+1}(\xi),$$

where (independent of Y_1)

$$(4.15)\qquad \delta_{m+1}(\xi) = P\{\underline{t}_{m+1} < \underline{a}_\xi,\ \underline{n} = m+1,\ \underline{k}_{1,m+1} \leq \underline{r}_{Y_1}\}.$$

Because

$$(4.16)\qquad \delta(\xi,X) = \sum_{m=0}^{\infty} X^m\, \delta_{m+1}(\xi)$$

is again the unique solution in $[0,1]$ of (4.7) for $\xi > 0$ and $0 \leq X \leq 1$, we can combine (4.13), (4.14) and (4.16) to

$$(4.17)\qquad \sum_{m=0}^{\infty} X^m\, P\{\underline{t}_m < \underline{a}_\xi,\ \underline{n} > m,\ \underline{k}_{1,m} \leq \underline{r}_{Y_1}\} = \frac{Y_1\{Y_1 - X\delta(\xi,X)\}}{Y_1 - X\beta(\xi+\lambda-\lambda Y_1)}$$

and so express the left-hand side in known functions. In (4.17) we need only replace ξ by $\xi + \zeta_1$ to obtain the right-hand side of (4.12). But then the left-hand side of (4.9) is known and because of (4.8) we can combine all our partial results to

$$(4.18)\qquad \pi(\xi,\zeta_1,\eta_1,\eta_2,\zeta_2,X,Y_1,Y_2,Y_3,Y_4) = \frac{Y_1\{Y_1\delta(\zeta_2,Y_4) - X\delta(\xi+\zeta_1,X)\}}{Y_1\delta(\zeta_2,Y_4) - X\beta(\xi+\zeta_1+\lambda-\lambda Y_1\delta(\zeta_2,Y_4))} \cdot$$

$$\cdot \frac{\xi}{\zeta+\eta_1-\eta_2-\lambda(Y_2-Y_3)\delta(\zeta_2,Y_4)} \{\beta(\eta_2+\lambda-\lambda Y_3\delta(\zeta_2,Y_4)) - \beta(\xi+\eta_1+\lambda-\lambda Y_2\delta(\zeta_2,Y_4))\}.$$

REMARKS. The present derivation deals with a set of random variables slightly different from the one in SIKKEL [1975]. Also $\delta(\xi,X)$ differs by a factor X from the $\delta(\xi,X)$ used in RUNNENBURG [1965].

I hesitated about including the present section in this paper, because it is hard to see how (4.17) can be used as it stands. However, one can slightly generalize the foregoing derivation to a calculation of the probability of having the exponentially distributed $\underline{a}_\xi$ land in some busy period, rather than in the first one only (and not having any of the marks and catastrophes). We then obtain

(4.19) $$\frac{\xi+\lambda}{\xi+\lambda-\lambda\delta(\xi,1)}\,\pi(\xi,\zeta_1,\eta_1,\eta_2,\zeta_2,X,Y_1,Y_2,Y_3,Y_4)$$

and can take the limit for $\xi \downarrow 0$. This produces a useful result concerning nine random variables $(\underline{z}',\underline{s}',\underline{s}'',\underline{z}'',\underline{m},\underline{k}_1,\underline{k}_2,\underline{k}_3,\underline{k}_4)$ in the stationary limiting process (provided we add $\lambda E\underline{s}_1 < 1$ to our assumptions).

Note concerning (2.2) and (2.3).

In the section on runs I intended to give an example of a collective mark from VAN DANTZIG [1949] to indicate his manner of introducing such marks. As pointed out to me by A.J. Lenstra and A.H. Hoekstra, the mark (2.3) hardly makes sense for infinitely many dissections of Ω. Hence I have added the restriction to finitely many (say r) dissections. With all s_{mn} equal to 1 in (2.3) the mark C is now equal to 1 as well, as it should be. The mark in (2.3) becomes an ordinary generating function, if we replace $\prod_{m=1}^{r} s_{mn_m}$ by $\prod_{m=1}^{r} s_m^{n_m}$, where $s_m^{n_m}$ is the n_m^{th} power of the complex variable s_m (with absolute value at most one).

REFERENCES

VAN DANTZIG, D. [1947], *Syllabus Mathematische Statistiek*, Mathematisch Centrum, Amsterdam (in Dutch, lecture notes).

———————— [1949], *Sur la méthode des fonctions génératrices*, Colloques internationaux du CNRS 13, 29-45.

———————— [1955], *Chaînes de Markof dans les ensembles abstraits et applications aux processus avec régions absorbantes et au problème des boucles*, Ann. Inst. H. Poincaré 14^3, 145-199.

———————— [1957], *Les fonctions génératrices liées à quelques tests non-paramétriques*, Report S 224, Mathematical Centre, Amsterdam.

VAN DANTZIG, D. & G. ZOUTENDIJK [1959], *Itérations markoviennes dans les ensembles abstraits*, J. Math. Pures Appl. 38^2, 183-200.

ENGEL, A. [1976], *Wahrscheinlichkeitsrechnung und Statistik, Band 2*, Ernst Klett, Stuttgart.

FELLER, W. [1968], *An introduction to probability theory and its applications, Volume 1*, third edition, Wiley, New York.

FRÉCHET, M. [1940,1943], *Les probabilités associées à un système d'événements compatibles et dépendants,* première partie (1940), deuxième partie (1943), Hermann, Paris.

GNEDENKO, B.V., E.A. DANIELJAN, B.N. DIMITROV, G.P. KLIMOV & W.F. MATVEJEV [1973], *Service systems with priorities,* Moskow University Publication, Moskow (in Russian).

GOOD, I.J. [1973], *The joint probability generating function for run-lenghts in regenerative binary Markov chains, with applications,* Ann. Statist. 1, 933-939.

HOWARD, R.A. [1971], *Dynamic probabilistic systems,* 2 volumes, Wiley, New York.

KEMENY, J.G. & J.L. SNELL, [1960], *Finite Markov chains,* Van Nostrand, Princeton.

KINGMAN, J.F.C. [1966], *On the algebra of queues,* J. Appl. Probability 3, 285-326.

LORENS, C. [1964], *Flowgraphs for the modeling and analysis of linear systems,* McGraw-Hill, New York.

MASON, S.J. & H.J. ZIMMERMANN, [1960], *Electronic circuits, signals and systems,* Wiley, New York.

MURPHY, G.J. [1957], *Basic automatic control theory,* Van Nostrand, Princeton.

RÅDE, L. [1972], *Thinning of renewal point processes,* Teknologtryck, Göteborg.

RUNNENBURG, J.Th. [1965], *On the use of the method of collective marks in queueing theory, Chapter 13 in Congestion Theory,* edited by W.L. Smith and W.E. Wilkinson, University of North Carolina Press, Chapel Hill.

SIKKEL, D. [1975], *Simultane verdelingen in wachtrijen met één bediener, afgeleid met de methode van de collectieve kenmerken,* final exam paper, University of Amsterdam (doctoraalscriptie).

YADIN, M. [1970], *Queueing with alternating priorities, treated as random walk on the lattice in the plane,* J. Appl. Probability 7, 196-218.

COMPUTATION AND STABILITY OF SOLUTIONS OF LINEAR LEAST SQUARES PROBLEMS

A. van der SLUIS

Notation

A will denote a real or complex m×n-matrix;
b will denote a real or complex m-vector;
x will denote a real or complex n-vector;
$\|\cdot\|$ will denote the euclidean vector norm as well as its associated matrix norm;
* used as a superscript of a matrix will denote transposition followed by complex conjugation.

1. INTRODUCTION

In this paper we discuss a number of properties of least squares problems and their solution methods in a historic setting, with emphasis on their backgrounds and implications rather than on their formal derivation. For reasons of time and space we restrict ourselves to the *full rank* least squares problem, i.e. given an overdetermined system of linear equations

$$Ax = b \tag{1.1}$$

where A has rank n, determine the vector x which minimizes

$$\|Ax - b\|. \tag{1.2}$$

The vector x is then called the *least squares solution* of (1.1) and the vector

$$r = b - Ax \tag{1.3}$$

is called the *residual vector*.

A typical situation in which this problem may arise is the following. A physical quantity β is known (or supposed) to depend linearly on quantities $\alpha_1,\ldots,\alpha_n$:

$$\beta = \alpha_1 x_1 + \ldots + \alpha_n x_n, \tag{1.4}$$

but the coefficients $x_1,\ldots,x_n$ are unknown and are to be determined. Now suppose that for any given n-tuple $\alpha_1,\ldots,\alpha_n$ the corresponding value of β may be measured. Then each of these measurements gives a linear equation (1.4) for $x_1,\ldots,x_n$. This leads to a system (1.1) which, due to measuring errors, will usually be contradictory if $m > n$. By minimizing the functional (1.2) one then hopes to get a better solution than by just solving n of these equations.

The idea of minimizing (1.2) goes back to Gauss, who also noted that the value of x so obtained is the *most probable* value, what is called nowadays the maximum likelihood value (cf. GAUSS [5], § 179). He also introduced the normal equations method for solving the minimization problem (l.c.§180).

Actually, Gauss states he hit on these ideas as early as 1795. However, he did not publish them until 1809, and then ran into a heated priority debate, with Legendre, who had found the same method independently and published it in 1806 (cf. GAUSS [6], p. 196).

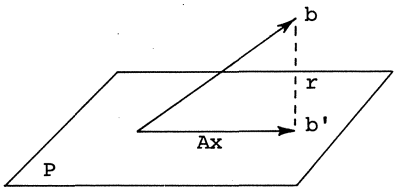
2. THE NORMAL EQUATIONS

A rather simple geometrical explanation of Gauss' way of solving the least squares problem, to which we shall refer in the sequel, is the following.

Let P denote the hyperplane spanned by the columns of A, and let b' be the orthogonal projection of b on P (cf. the figure on the next page). Then the vector x which minimizes $\|Ax - b\|$ should obviously be such that $Ax = b'$. Hence $b - Ax \perp P$, i.e. the inner product $(b - Ax, Au)$ should equal 0 for all u, or $(A^*b - A^*Ax, u) = 0$ for all u. Hence

$$A^*Ax = A^*b, \tag{2.1}$$

which is a system of n equations in n unknowns, the well-known *normal equations*, which we know how to solve.



This method has worked well for a century and a half. However, when high speed electronic computers came into use after WW II and much larger systems could be handled than before, difficulties arose: the normal equations used to be *ill-conditioned*.

3. THE CONDITION OF EQUATIONS

Already in 1950 the notion of *condition* of a matrix was the subject of lively discussion, as appears from a remark in TAUSSKY [22]. It amounts to the following. Let

$$By = c \tag{3.1}$$

denote a non-singular system of n equations in n unknowns and let

$$\tilde{\tilde{B}}\tilde{y} = c \tag{3.2}$$

denote a non-singular system whose matrix is somewhat perturbed with respect to B. Then, defining $\Delta y = \tilde{y} - y$, $\Delta B = \tilde{B} - B$, it is easily verified that

$$\frac{\|\Delta y\|}{\|y\|} \leq \frac{\|B^{-1}\|\|B\|}{1 - \|B^{-1}\Delta B\|} \frac{\|\Delta B\|}{\|B\|} ; \tag{3.3}$$

moreover it may be shown (cf. VAN DER SLUIS [19]) that for any B and c and any $\delta > 0$, $\varepsilon > 0$ there exists a $\Delta B \neq 0$, $\|\Delta B\| \leq \delta$, such that the quotient of the right- and left-hand sides of (3.3) is less than $1 + \varepsilon$. Hence, a relative error (to be understood in the sense of norms) β in B may lead to a relative error η in y which is about a factor

$$\kappa(B) \underset{D}{=} \|B\|\|B^{-1}\| \tag{3.4}$$

times as large as β. The quantity $\kappa(B)$ (which is never less than 1) is called the *spectral condition number* of β, and if it is large then obviously the solution of (3.1) is very sensitive to errors in the matrix B. The matrix and the system of equations are then called *ill-conditioned*.

For the effect of perturbing c by Δc we have

$$\frac{\|\Delta y\|}{\|y\|} \leq \kappa(B,y)\ \frac{\|\Delta c\|}{\|c\|} \tag{3.5}$$

where $\kappa(B,y) = \|B^{-1}\|\|By\|/\|y\|$, whence $1 \leq \kappa(B,y) \leq \kappa(B)$, and the equality sign in (3.5) may occur for any B and c.

The condition number of a matrix is closely related to a geometric property of its column vectors: if B_i denotes the i-th column of B and δ_i denotes its euclidean distance from the hyperplane spanned by the other columns then we have

$$\frac{\max\|B_i\|}{\min\ \delta_i} \leq \kappa(B) \leq n\cdot\frac{\max\|B_i\|}{\min\ \delta_i}\ . \tag{3.6}$$

Thus, for moderate values of n we have that $\kappa(B)$ is large if and only if at least one of the δ_i is small with respect to at least one of the $\|B_j\|$.

This implies that large condition numbers have two possible sources: (a) the row- or column-vectors of the matrix have widely differing norms; (b) the directions of the row- or column-vectors are not very well separated, i.e. the span of one of the rows or columns makes a small angle with the span of the others.

As an illustration consider a 2×2 matrix B. If $v \leq 1$ denotes the quotient of the norms of the columns (rows) and ϕ is the angle between the columns (rows), then (3.6) implies

$$\frac{1}{v\cdot\sin(\phi)} \leq \kappa(B) \leq \frac{2}{v\cdot\sin(\phi)}\ . \tag{3.7}$$

A condition number can be defined equally well for a full rank m×n-matrix B, $m > n$. For, considered as a mapping from $\mathbb{R}^n$ or $\mathbb{C}^n$ to the range of B it has a unique inverse whose norm is given by

$$\max_{x\neq 0}\frac{\|x\|}{\|Bx\|}\ , \tag{3.8}$$

and substituting this into (3.4) for $\|B^{-1}\|$ then defines $\kappa(B)$. The quantity in (3.8) may also be written as $\|B^{+}\|$, where B^{+} denotes the generalized inverse of B (for a definition cf. STEWART [21] p. 221 and p. 325).

Then (3.6) still holds. Moreover we have

$$\kappa(B^{*}B) = \kappa^{2}(B). \tag{3.9}$$

What becomes out of (3.3) and (3.5) in this case will be the subject of a latter section.

4. THE ILL-CONDITIONING OF THE NORMAL EQUATIONS

From (3.9) we see that as soon as A is moderately ill-conditioned, $A^{*}A$ is terribly ill-conditioned, which makes it understandable, to some extent, why normal equations are so often ill-conditioned.

A very bad example of ill-conditioning is given by the normal equations encountered in determining a least squares polynomial fit to a function which is sampled more or less uniformly on the interval [0,1]. The matrix $A^{*}A$ then resembles

$$\begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \cdots & \frac{1}{n} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \cdots & \frac{1}{n+1} \\ \vdots & & & & \\ \frac{1}{n} & \cdots & \cdots & & \frac{1}{2n-1} \end{pmatrix} \tag{4.1}$$

(a finite segment of the infinite Hilbert matrix), whose columns obviously become more nearly dependent the more right-most they are. And indeed, $\kappa(A^{*}A) \approx 10^{1.5n}$ (cf. GREGORY & KARNEY [9]).

Now suppose that $A^{*}A$ is computed with relative rounding errors of 10^{-15}, say, after each arithmetic operation. Then one should at least expect a relative error (in the sense of norms, cf. § 3) in $A^{*}A$ of the order of 10^{-15}, and hence the value n = 10 is no longer tolerable (cf. (3.3) and subsequent text).

It should be noted, however, that not always, when A has a large condition number, the normal equations perform as badly as has just been described. Indeed, if the columns of A have widely differing norms, and if D is a diagonal matrix such that the columns of AD have about equal norms,

then it may be shown that the relative error in x is controlled by $\kappa(AD)\kappa(A)$ (private communication by Prof. Å. Björck) and if the directions of the columns of A are well separated then $\kappa(AD)$ will be very much smaller than $\kappa(A)$ and hence $\kappa(AD)\kappa(A) << \kappa^2(A)$. However, this does obviously not apply to the example earlier in this section.

5. BETTER SOLUTION METHODS

It is most disturbing, of course, that in solving the least squares problem via the normal equations, the vector x is determined from a system whose sensitivity is governed by $\kappa^2(A)$ whereas, when ordinary systems of equations are solved, the vector x is determined right away from a system whose sensitivity is governed by $\kappa(A)$.

Thus a search went underway for methods which would not suffer the squaring of the condition number (without bothering, what would have been more natural, about the question whether the true vector x minimizing (1.2) would not already have a sensitivity with respect to perturbations of A and b for which $\kappa^2(A)$ is relevant). For discussions and early results cf. e.g. LÄUCHLI [13], VON HOLDT [11], OSBORNE [16].

The idea was to avoid the use of normal equations altogether or to transform the least squares problem in such a way that the normal equations could not be arbitrarily ill-conditioned. Recalling our geometric treatment in § 2, a rather obvious way of doing this is to construct an orthonormal basis in P, then the coordinates of b' with respect to this basis can be easily calculated, and by expressing these basis vectors in terms of the columns of A we get x. In formulae $Q = AU$, Q an $m \times n$-matrix with orthonormal columns and U an $n \times n$-matrix, then b' has Q^*b as its vector of coordinates with respect to the columns of Q, i.e. $b' = QQ^*b = AUQ^*b$, hence $x = UQ^*B$.

The most obvious way to find Q is by applying Gram-Schmidt orthogonalization to A, but if one actually does so then one is in trouble since this orthogonalization method is not numerically stable (cf. WILKINSON [23], p. 243), and this is dramatically confirmed by the results of a least squares solution method based on this principle (cf. JORDAN [12]).

In 1965, however, GOLUB published a method (cf. [7], also BUSINGER & GOLUB [4]) where in fact the orthogonalization process was carried out using Householder transformations, and this is known to be a stable process. His method is still the one in widest use.

Then, in 1967, BJÖRCK (cf. [1]) proved that a modified version of

Gram-Schmidt also is numerically stable. A least squares solution method based on it may, according to the experiments of JORDAN [12], even be slightly more accurate than Golub's method.

The amount of work for Golub's and Björck's methods are about the same, and about double what is required for the normal equations method.

6. THE SENSITIVITY OF LEAST SQUARES SOLUTIONS

In 1966 Golub and Wilkinson raised the question about (what they called) the *inherent sensitivity* of the least squares solution (cf. GOLUB & WILKINSON [8]).

They showed that if $\|A\| = \|b\| = 1$ and A and b are perturbed by ΔA and Δb with $\|\Delta A\| \leq \varepsilon$, $\|\Delta b\| \leq \varepsilon$, then the vector $\hat{x}$ which minimizes the perturbed expression (1.2) satisfies

$$\|\hat{x} - x\| \leq \varepsilon[\kappa^2(A)\|r\| + \kappa(A)\|x\| + \kappa(A)] + O(\varepsilon^2) \tag{6.1}$$

(for r cf. (1.3)), and to our dismay we see that $\kappa^2(A)$ still plays a role. Now (6.1) gives only an upper bound, but Golub and Wilkinson added that it is easy to verify by means of a 3×2 matrix that this bound is realistic.

A heuristic explanation, that $\varepsilon\kappa^2(A)r$ should indeed be expected to be of importance for the sensitivity of the least squares solution, can be derived from geometrical arguments. Indeed, for given $\varepsilon > 0$ and any ΔA with $\|\Delta A\| \leq \varepsilon$ the angle between span(A+ΔA) and span(A) is at most $\gamma = \arcsin(\varepsilon\kappa(A))$, and this value is attained for a suitable ΔA. Tilting span(A) over this angle will cause b' (see the figure in §2) to move by a distance of about $\gamma\|r\|$ (if γ is small) across the tilting plane. Since x satisfies Ax = b' this means changing the right hand side by $\varepsilon\kappa(A)\|r\|$, and this may lead to an error in x which is $\max_{y\neq 0} \frac{\|y\|}{\|Ay\|}$ times this quantity (cf. (3.8)), and this accounts for the second factor $\kappa(A)$ in $\varepsilon\kappa^2(A)r$ (cf. (6.1)) because of $\|A\| = 1$.

However, the normalization A = b = 1 obscures something. Indeed, condition numbers are expected to link *relative* perturbations, whereas in (6.1) something is said about absolute perturbations. Returning to our geometrical explanation, and looking at (3.5), we would rather expect something like $\varepsilon\kappa(A,x)\kappa(A)\|r\|/\|b'\|$ in an estimate for the relative error in x. This geometrical argument can be made exact, and leads to the following statement (cf. VAN DER SLUIS [20]):

Assumptions:

- let $\|\Delta A\| \leq \alpha\|A\|$ and $\|\Delta b\| \leq \beta\|b\|$;
- let $\mu \underset{D}{=} \alpha\kappa(A) < 1$ (this is a natural requirement since otherwise there would be a ΔA such that $\operatorname{rank}(A+\Delta A) < n$);
- let ϕ denote the angle between b and span(A);
- let $\kappa(A,x)$ denote $\frac{\|Ax\|}{\|x\|}\|A^+\|$ (where $\|A^+\| = \max_{y\neq 0}\frac{\|y\|}{\|Ay\|}$, cf. § 3; obviously $1 \leq \kappa(A,x) \leq \kappa(A)$).

Then

$$\frac{\|\hat{x}-x\|}{\|x\|} \leq \alpha\frac{\kappa(A)}{1-\mu^2}\kappa(A,x)\tan(\phi) + \frac{\kappa(A)}{1-\mu}\left(\alpha + \frac{\beta}{\cos(\phi)}\right) \tag{6.2}$$

and there exist ΔA, Δb with $\|\Delta A\| = \alpha\|A\|$, $\Delta b = 0$, such that

$$\frac{\|\hat{x}-x\|}{\|x\|} \geq \alpha\frac{\kappa(A)}{1-\mu^2}\kappa(A,x)\tan(\phi). \tag{6.3}$$

We note that in (6.2) and (6.3) the "dangerous" terms containing $\kappa(A)\kappa(A,x)$ coincide. For earlier results in the direction of (6.2) and (6.3) we refer to LAWSON & HANSON [14] and to HANSON & LAWSON [10].

Formulae (6.2) and (6.3) make a few things clear:

(a) In order that the sensitivity (in the sense of relative perturbations of A, b and x) is controlled by $\kappa^2(A)$ it is necessary that $\|Ax\|/\|x\|$ is of the order of $\|A\|$ and $\tan(\phi)$ is of the order 1 (unless one is willing to consider systems with $\phi \approx \pi/2$). Now $\|Ay\|/\|y\| \approx \|A\|$ is true for a large fraction of all vectors y, but if $\kappa(A)$ is large than $\|Ax\|/\|x\| \approx \|A\|$ is true for a small fraction of all vectors b only.

(b) Nevertheless a sensitivity which is quite a bit greater than corresponding to $\kappa(A)$ may already happen as soon as $\kappa(A)\tan(\phi)$ is noticeably larger than 1, and this, in turn, may already be the case for very small ϕ, i.e. for systems (1.1) which are only slightly incompatible.

7. THE ACCURACY OF LEAST SQUARES SOLUTIONS

Now what does the sensitivity analysis of § 6 mean for the accuracy with which least squares solutions may be obtained?

Let us assume we use Golub's method (cf. § 5), carrying out the arithmetic operations with relative rounding errors of the order of $\bar{\xi}$ (typical

values for $\bar{\xi}$ lie between 10^{-7} and 10^{-15}). Then it can be proved that the computed solution x is the true least squares solution of a system

$$\tilde{\tilde{A}}x = \tilde{\tilde{b}}, \tag{7.1}$$

and $\Delta A = \tilde{A} - A$, $\Delta b = \tilde{b} - b$ satisfy

$$\|\Delta A_i\| \leq \varepsilon \|A_i\|, \quad \|\Delta b\| \leq \varepsilon \|b\|, \quad \varepsilon = c(m,n)\bar{\xi}, \tag{7.2}$$

where A_i denotes the i-th column of A and c(m,n) is independent of A and b (cf. GOLUB & WILKINSON [8]).

Hence

$$\|\Delta A\| \leq \varepsilon \|A\| \sqrt{n} \tag{7.3}$$

which is an overestimate, but there are ΔA satisfying (7.2) such that

$$\|\Delta A\| \geq \varepsilon \|A\|. \tag{7.4}$$

Thus (6.2) may be applied with $\alpha = \sqrt{n}\cdot\varepsilon$, $\beta = \varepsilon$, but may (6.3) be applied with $\alpha = \varepsilon$? That depends entirely on whether *all* ΔA satisfying (7.2) are possible, and for the kind of relative errors we are considering this will not, in general, be the case. It will certainly not be the case if the columns of A have widely differing norms, and thus (6.3) will not be applicable.

Indeed, suppose that the columns of A have well separated directions (cf. § 3), but strongly differing norms. Then $\kappa(A)$ will be large. Reflection on our heuristic argument in §6 makes it clear, however, that small *relative* perturbations of the columns of A, i.e. small perturbations of the *directions* of the columns of A, can now only result in a small tilting angle of the plane P (cf. the figure in § 2), irrespective of $\kappa(A)$. That is to say that if D is a diagonal matrix such that the columns of AD have about equal norms, then the sensitivity of the solution is expected to depend on $\kappa(AD)\kappa(A,x)\tan(\phi)$ rather than on $\kappa(A)\kappa(A,x)\tan(\phi)$. This is actually proved in (more detail) in VAN DER SLUIS [20].

Comparing this result with the one mentioned at the end of §4 one notes that in this case, too, Golub's method is superior to the normal equations method.

The result also reminds us that condition numbers say something about sensitivity only for certain specific perturbation classes.

Obviously, speaking about the accuracy of least squares solutions one should also take into account possible errors in the data. If these errors satisfy properties like $\|\Delta A\| \leq K\|A\|$ or $\|\Delta A_i\| \leq K\|A_i\|$ then the perturbation results mentioned before are clearly applicable. One should not be mistaken, however, in thinking that the effects of these errors also are smaller when Golub's method is used than when normal equations are used. Since the normal equations are mathematically equivalent to the least squares problem they must give the same result when no rounding errors are made.

8. STABILITY

In § 7 we noted that the computed solution satisfies a slightly perturbed least squares problem. This may be used for estimating the error in the computed solution (as we did in § 7), but now we consider this from a different point of view.

Actually, the elements of A and b will usually be subject to some error or uncertainty, if only because of their being rounded to machine precision. This in turn will lead to an uncertainty in the solution, which we will call the *intrinsic uncertainty*. Then, if the elements of ΔA and Δb in (7.2) exceed the uncertainties in the corresponding elements of A and b by at most a factor q, it follows that the error in the solution caused by rounding errors during the computation, large as it may be, will not exceed the intrinsic uncertainty by more than the same factor q. This is a very desirable *stability* property provided q has a reasonable value, and in fact no numerical algorithm may then be expected to perform essentially better.

From (7.2) we see that we will have this kind of stability if for all j all elements in A_j have uncertainties which are not small with respect to $\varepsilon\|A_j\|$, and similarly for b.

One situation in which the latter condition will certainly not be satisfied is when A is obtained from a matrix which does satisfy this condition by multiplying its rows by widely differing constants, and likewise b (so as to put widely differing weights on the various equations). And indeed, as has been observed by POWELL & REID [18], in this case Golub's method may lead to awkward results. They suggested to add a *pivoting procedure* to Golub's method in order to circumvent these difficulties (column pivoting combined with row pivoting).

Roughly speaking, this pivoting procedure has the effect that for all i the elements in the i-th row of ΔA are of the order of ε times the largest element in the i-th row of A, and the i-th element of Δb is about $\|x\|$ times the latter amount (although sometimes the situation may be worse). In order to appreciate this, suppose that $\|A_j\| \approx \|b\|$ for all j, which is just a matter of scaling, and that then $\|x\|$ is not large with respect to 1, which is not unreasonable (see however §6, under (a)). Then, if all elements in the i-th row of $(A\vdots b)$ have an uncertainty which is not small with respect to ε times the largest element in the i-th row of A - a rather reasonable assumption - we have again the desired kind of stability.

Column pivoting alone, as suggested in GOLUB [7], does not have this beneficial effect. In fact, it is hard to see when column pivoting may be expected to have a positive effect, and this is confirmed by experiments of JORDAN [12]. Column pivoting may be useful as a tool for detecting near rank-deficiency, however, but that is quite another story.

9. TOPICS NOT TREATED

Again for reasons of time and space we have not dealt with alternatives for Golub's method and with iterative methods. However, a few references to this effect have been added (cf. [1], [2], [3], [8], [15], [17]).

REFERENCES

[1] BJÖRCK, Å., *Solving linear least squares problems by Gram-Schmidt orthogonalization,* BIT 7 (1967) p. 1-21.

[2] BJÖRCK, Å., *Iterative refinement of linear least squares solutions I,* BIT 7 (1967) p. 257-278.

[3] BJÖRCK, Å. & G.H. GOLUB, *Iterative refinement of linear least squares solutions by Householder Transformation,* BIT 7 (1967) p. 322-337.

[4] BUSINGER, P. & G.H. GOLUB, *Linear least squares solutions by Householder transformations,* in: *Handbook for automatic computation,* vol. II, *linear algebra,* J.H. Wilkinson and C. Reinsch (ed.), p. 111-118, Springer, Berlin etc. 1971.

[5] GAUSS, C.F., *Theoria motus corporum coelestium in sectionibus conicis solem ambientium,* in: *Werke,* vol. VII, Perthes, Gotha, 1871, or in: *Abhandlungen zur Methode der kleinsten Quadrate,* ch. II, Physica Verlag, Würzburg, 1964.

[6] GAUSS, C.F., *Gedenkband anlasslich des 100. Todestages am 23. Februar 1955,* H. Reichardt (ed.), Teubner, Leipzig, 1957.

[7] GOLUB, G.H., *Numerical methods for solving linear least squares problems,* Numer. Math. 7 (1965) p. 206-216.

[8] GOLUB, G.H. & J.H. WILKINSON, *Note on the iterative refinement of least squares solution,* Numer. Math. 9 (1966) p. 139-148.

[9] GREGORY, R.T. & D.L. KARNEY, *A collection of matrices for testing computational algorithms,* Wiley-Interscience, New York etc. 1969.

[10] HANSON, R.J. & C.L. LAWSON, *Extensions and applications of the Householder algorithm for solving linear least squares problems,* Math. Comp. 23 (1969) p. 787-812.

[11] HOLDT, R. VON, *More accurate linear least squares,* in: *Proc. Western Joint Comput. Conf.,* Mar. 3-5, 1959.

[12] JORDAN, T.L., *Experiments on error growth associated with some linear least squares procedures,* Math. Comp. 22 (1968) p. 579-588.

[13] LÄUCHLI, P., *Jordan-Elimination und Ausgleichung nach kleinsten Kwadraten,* Numer. Math. 3 (1961) p. 226-240.

[14] LAWSON, C.L. & R.J. HANSON, *Solving least squares problems,* Prentice Hall, Englewood Cliffs, 1974.

[15] MARTIN, R.S., G. PETERS & J.H. WILKINSON, *Iterative refinement of the solution of a positive definite system of equations,* Numer. Math. 8 (1966) p. 203-216.

[16] OSBORNE, E.E., *On least squares solutions of linear equations,* JACM 8 (1961) p. 628-636.

[17] PETERS, G. & J.H. WILKINSON, *The least squares problem and pseudo-inverses,* Comput. J. 13 (1970) p. 309-316.

[18] POWELL, M.J.D. & J.K. REID, *On applying Householder's method to linear least squares problems,* in: Proc. IFIP Congress, 1968, p. 122-126, North-Holland, Amsterdam, 1969.

[19] SLUIS, A. VAN DER, *Stability of solutions of linear algebraic systems,* Numer. Math. 14 (1970) p. 246-251.

[20] SLUIS, A. VAN DER, *Stability of the solutions of linear least squares problems,* Numer. Math. 23 (1975) p. 241-254.

[21] STEWART, G.W., *Introduction to matrix computations,* Academic Press, New York, 1973.

[22] TAUSSKY, O., *Note on the condition of matrices,* Math. Tables Aids Comput. 4 (1950) p. 111-112.

[23] WILKINSON, J.H., *The algebraic eigenvalue problem,* Clarendon, Oxford, 1965.

ERROR BOUNDS IN THE NUMERICAL SOLUTION OF INITIAL VALUE PROBLEMS

M.N. Spijker

1. INTRODUCTION

In this paper we consider initial value problems for systems of ordinary differential equations and integro-differential equations. We shall deal with numerical methods for approximating the solutions of such initial value problems. Many authors have been concerned with the problem of finding *bounds for the errors* which are usually present in the approximations produced by these numerical methods. Roughly speaking the bounds to be found in the literature fall into two groups.

The *first group* consists of computable a-posteriori bounds, the purpose of which is to give the users of a numerical method precise quantitative information about the meaning and accuracy of the numbers actually produced by the computer. Obtaining bounds of this type often requires some new numerical process to be executed on a computer. So-called interval-arithmetic is one of the techniques that are used in arriving at these bounds - see e.g. [7], [8].

The *second group* consists of a-priori bounds originating from a mathematical analysis of the numerical method under consideration. Usually these bounds are of little value for arriving at a useful quantitative statement about the accuracy of the approximations actually obtained. On the other hand the bounds within our second group may lead to an understanding of the mechanism by which the errors in the approximations are built up. Further these bounds can be used to obtain useful qualitative information about the accuracy of the approximations. Moreover, these bounds have often been a guiding principle in constructing new efficient numerical methods.

In this paper we shall deal with error bounds belonging to the second group. In particular we shall be concerned with so-called *two-sided error bounds*.

In the, still introductory, chapter 2 it will be shown how a shortcoming of some of the classical (one-sided) error bounds leads in a

natural way to the concept of a two-sided error bound. In chapter 3 we consider the general problem of finding easily verifiable conditions that are necessary and sufficient for the existence of two-sided error bounds. In the rest of the paper we review some results obtained with regard to this problem during the last few years.

In the chapters 2, 3 and 4 we deal with numerical methods for solving ordinary differential equations. In chapter 5 we consider the numerical solution of integro-differential equations of Volterra.

2. THE NEED OF TWO-SIDED ERROR BOUNDS

2.1. We consider an initial value problem for a system of s ordinary differential equations which, by using vector notation, can be written in the form

$$\frac{d}{dt}U(t) = f(t,U(t)) \quad (0 \le t \le T), \quad U(0) = c. \tag{2.1}$$

Here c denotes a given vector in the s-dimensional real vectorspace $\mathbb{R}^s$, and f is a given continuous function from $[0,T] \times \mathbb{R}^s$ to $\mathbb{R}^s$. We assume that f satisfies a Lipschitz condition

$$|f(t,\tilde{\xi}) - f(t,\xi)| \le L\cdot|\tilde{\xi} - \xi| \quad (\xi,\tilde{\xi} \in \mathbb{R}^s,\ 0 \le t \le T) \tag{2.2}$$

where $|\cdot|$ denotes any norm on vectors in $\mathbb{R}^s$, and where L is independent of ξ, $\tilde{\xi}$ and t. Under these conditions there exists a unique solution $U(t) \in \mathbb{R}^s$ (for $0 \le t \le T$) to the initial value problem (2.1) - see e.g. [4, p.113].

In order to introduce the concept of a two-sided error bound we are going to consider the numerical solution of the problem (2.1) by the following simple method, due to Euler.

$$h^{-1}\cdot(u_n - u_{n-1}) = f(t_{n-1},u_{n-1}) \quad (n = 1,2,\ldots,N), \quad u_0 = c. \tag{2.3}$$

Here h denotes the so-called stepsize, which is chosen such that $0 < h \le T$. The vectors u_n are computed recursively from (2.3) and stand for approximations of U(t) at $t = t_n = nh$. With N we denote the greatest integer satisfying $Nh \le T$.

2.2. Along with (2.3) we consider a perturbed version of Euler's method

$$(2.4)\qquad h^{-1}\cdot(\tilde{u}_n-\tilde{u}_{n-1}) = f(t_{n-1},\tilde{u}_{n-1}) + w_n \quad (n=1,2,\ldots,N),\ \tilde{u}_0 = c + w_0,$$

where $\tilde{u}_n$ denotes the approximation of $U(t_n)$ obtained in the presence of some local perturbations $w_0,w_1,\ldots,w_N$. For instance, in case f is a complicated function, w_n may stand for the error introduced in the n^{th} stage of the computations if the value $f(t_{n-1},\tilde{u}_{n-1})$ is approximated by some, more easily computable, quantity $\tilde{f}(t_{n-1},\tilde{u}_{n-1}) = f(t_{n-1},\tilde{u}_{n-1}) + w_n$. Likewise, the perturbations w_n may be understood to be caused by rounding-off only. Finally, w_n may also stand for the so-called local discretization error, which means that w_n is defined by the relations (2.4) with $\tilde{u}_n = U(t_n)$ $(n=0,1,\ldots,N)$.

From (2.1) and Taylor's theorem one obtains the following expressions for the local discretization errors

$$(2.5)\qquad w_0 = 0, \quad w_n = \frac{h}{2}\cdot\frac{d^2}{dt^2}U(t_n) + O(h^2) \quad (n = 1,2,\ldots,N).$$

Clearly, in each of the three cases just mentioned it is desirable to have an a-priori bound by means of which the effect of the perturbations w_n on the differences $\tilde{u}_n - u_n$ can be estimated.

By subtracting the relations (2.3) from (2.4) and by using the Lipschitz condition (2.2) one arrives, after a short calculation, at the (classical) error bound

$$(2.6)\qquad \max_{0\le n\le N}|\tilde{u}_n-u_n| \le e^{LT}\cdot\left\{|w_0| + h\sum_{j=1}^{N}|w_j|\right\}$$

(see e.g. [6], [16]).

An important application of this bound is obtained by choosing $\tilde{u}_n = U(t_n)$ and w_n as in (2.5). The right-hand member of (2.6) then reduces to

$$e^{LT}\cdot h\sum_{j=1}^{N} O(h)$$

and since $Nh \le T$, it follows from (2.6) that

$$\max_{0\le n\le N}|U(t_n)-u_n| = O(h).$$

The bound (2.6) can thus be used to prove that Euler's method produces approximations u_n which, if $h \downarrow 0$, converge to the corresponding values of the true solution $U(t)$.

2.3. Although the bound (2.6) thus has a useful application, it suffers from an imperfection, which manifests itself if we consider the behaviour of (2.6), when T is fixed and $h \downarrow 0$, for perturbations different from (2.5). We choose $f(t,\xi) \equiv 0$, $L = 0$ and $w_0 = 0$, $w_j = (-1)^{j-1} w_1$ $(2 \leq j \leq N)$, $|w_1| = \varepsilon > 0$.

The right-hand member of (2.6) now reduces to $N \cdot h\varepsilon$. An easy calculation shows that the actual quantity $\max_{0\leq n\leq N}|\tilde{u}_n - u_n|$ equals $h \cdot \varepsilon$. Hence the factor by which the right-hand member of (2.6) overestimates its left-hand member, equals $N \approx \frac{T}{h}$, which tends to infinity if $h \downarrow 0$.

In the next section it will be shown how this shortcoming of (2.6), which is also present for other, less trivial, choices for $f(t,\xi)$, can be overcome by using the concept of a two-sided error bound.

For the sake of completeness we conclude this section by noting that the bound (2.6) suffers from a second imperfection, which manifests itself if we keep $h > 0$ fixed and let $T \to \infty$. Then, it can be seen from (2.6) that its right-hand member may increase exponentially in cases where $\max_{0\leq n\leq N}|\tilde{u}_n - u_n|$ remains bounded. This imperfection can, to some extent, be overcome by replacing the factor $\exp(LT)$ in (2.6) by a factor in which the so-called logarithmic norm of the Jacobian matrix $\frac{\partial}{\partial \xi} f(t,\xi)$ enters (see [3], [5]). We shall not deal any further with this imperfection of (2.6).

2.4. Suppose we have an arbitrary error bound, for Euler's method, which can be written in the form

$$\max_{0\leq n\leq N} |\tilde{u}_n - u_n| \leq E_1 \tag{2.7}$$

where $E_1 = E_1[w_0, w_1, \ldots, w_N; h]$ depends on the perturbations $w_n \in \mathbb{R}^s$ and on $h > 0$. The factor by which E_1 overestimates the actual quantity $\max_{0\leq n\leq N}|\tilde{u}_n - u_n|$ equals $[\max_{0\leq n\leq N}|\tilde{u}_n - u_n|]^{-1} \cdot E_1$, and is thus bounded by some constant $\beta > 0$ if $\frac{1}{\beta} \cdot E_1 \leq \max_{0\leq n\leq N}|\tilde{u}_n - u_n|$. Hence, the factor by which (2.7) overestimates the actual error, is bounded uniformly for all $h > 0$ and all $w_n \in \mathbb{R}^s$, if and only if (2.7) can be completed to a *two-sided error bound*

$$E_0 \leq \max_{0\leq n\leq N} |\tilde{u}_n - u_n| \leq E_1 \tag{2.8}$$

with a left-hand member $E_0 = E_0[w_0,w_1,\ldots,w_N;h]$ that can be written in the form

$$E_0[w_0,w_1,\ldots,w_N;h] \equiv \frac{1}{\beta} \cdot E_1[w_0,w_1,\ldots,w_N;h].$$

Here the constant $\beta > 0$ is independent of $w_0,w_1,\ldots,w_N$ and h.

It has been proved (cf. [12], [16], [17]) that the error $\tilde{u}_n - u_n$, caused by the perturbations w_n in (2.4), admits the following bound of type (2.8)

$$(2.9) \qquad \gamma_0 \cdot \max_{0\le n\le N} |w_0+h\sum_{j=1}^{n} w_j| \le \max_{0\le n\le N} |\tilde{u}_n-u_n| \le \gamma_1 \cdot \max_{0\le n\le N} |w_0+h\sum_{j=1}^{n} w_j|.$$

Here and in the following we use the convention that $\sum_{j=1}^{n} \ldots = 0$ if $n < 1$. For γ_0,γ_1 one can obtain the expressions $\gamma_0 = (1+LT)^{-1}$, $\gamma_1 = \exp(LT)$.

From the two-sided error bound (2.9) it thus follows that the error bound (2.7) holds with

$$E_1 = e^{LT} \cdot \max_{0\le n\le N} |w_0 + h\sum_{j=1}^{n} w_j|,$$

and that, with this choice of E_1, the bound (2.7) overestimates $\max_{0\le n\le N}|\tilde{u}_n-u_n|$ by a factor which remains bounded if $h \downarrow 0$. Consequently, the imperfection of (2.6) mentioned above has been removed by replacing the expression

$$|w_0| + h\sum_{j=1}^{N} |w_j|$$

appearing in (2.6) by

$$\max_{0\le n\le N} |w_0 + h\sum_{j=1}^{n} w_j|.$$

3. GENERAL TWO-SIDED ERROR BOUNDS

3.1. In this chapter we consider the numerical solution of the initial value problem (2.1) by the general step-by-step method

$$(3.1) \qquad h^{-1} \cdot (u_n-u_{n-1}) = F(t_{n-1},u_{n-1};h) \quad (n = 1,2,\ldots,N), \quad u_0 = c,$$

and along with (3.1) we consider the perturbed method

$$(3.2)\qquad h^{-1}\cdot(\tilde{u}_n-\tilde{u}_{n-1}) = F(t_{n-1},\tilde{u}_{n-1};h) + w_n \quad (n = 1,2,\ldots,N),$$
$$\tilde{u}_0 = c + w_0.$$

It is assumed that the function F satisfies the following condition (3.3), in which h_0 and λ denote arbitrary, but otherwise fixed, numbers with $0 < h_0 \le T$, $\lambda > 0$.

(3.3) For all $h \in (0,h_0]$, $\xi \in \mathbb{R}^s$ the elements $F(t_{n-1},\xi;h)$ belong to $\mathbb{R}^s$ $(n = 1,2,\ldots,N)$. Further F satisfies the Lipschitz condition

$$|F(t_{n-1},\tilde{\xi};h) - F(t_{n-1},\xi;h)| \le \lambda\cdot|\tilde{\xi}-\xi|$$

$(\xi,\tilde{\xi} \in \mathbb{R}^s;\ h \in (0,h_0];\ n = 1,2,\ldots,N)$.

Clearly, if $F(t,\xi;h) \equiv f(t,\xi)$ and the Lipschitz constant L in (2.2) satisfies $L \le \lambda$, then condition (3.3) is fulfilled, and the methods (3.1), (3.2) reduce to Euler's method (2.3) and to its perturbed version (2.4), respectively. Further examples of the general method (3.1) include Runge-Kutta methods and Taylor expansion methods (see [4]).

3.2. In order to investigate general bounds for the errors $\tilde{u}_n - u_n$, resulting from the perturbations w_n occurring in (3.2), it is convenient to introduce the vectors

$$u = (u_0,u_1,\ldots,u_N), \quad \tilde{u} = (\tilde{u}_0,\tilde{u}_1,\ldots\tilde{u}_N), \quad w = (w_0,w_1,\ldots,w_N).$$

These vectors belong to the vectorspace X_h given by

$$X_h = \{x \mid x = (x_0,x_1,\ldots,x_N) \text{ where } N = N(h) \text{ and all } x_n \in \mathbb{R}^s\}.$$

Here N(h) denotes the greatest integer N with $N\cdot h \le T$. In X_h addition and multiplication with real numbers are defined coordinate-wise.

The difference between $\tilde{u}$ and u will be measured by means of a given *seminorm* $\|x\|_h$ for vectors $x \in X_h$ (see [9], p.24). We thus assume that $\|x\|_h$ is a real number, and that $\|x+y\|_h \le \|x\|_h + \|y\|_h$, $\|\alpha\cdot x\|_h = |\alpha|\cdot\|x\|_h$, for all vectors x, $y \in X_h$ and all real α.

Special attention will be paid to *absolute seminorms*, i.e. to seminorms such that $\|x\|_h = \|y\|_h$ for any pair of vectors $x = (x_0,x_1,\ldots,x_N)$,

$y = (y_0, y_1, \ldots, y_N)$ the coordinates of which satisfy $|x_n| = |y_n|$ $(n = 0,1,\ldots,N)$.

We list four examples of seminorms in X_h.

EXAMPLE 1. $\|x\|_h = \max_{0\le n\le N} |x_n|$,

EXAMPLE 2. $\|x\|_h = |x_N|$,

EXAMPLE 3. $\|x\|_h = \max[\max_{0\le n\le N} |x_n|, \max_{1\le n\le N} |h^{-1} \cdot (x_n - x_{n-1})|]$,

EXAMPLE 4. $\|x\|_h = \max[|x_N|, |h^{-1} \cdot (x_N - x_{N-1})|]$.

Clearly, the seminorms in the examples 1, 2 are absolute, and those in the examples 3, 4 are not.

The following definition formalizes the concept of a two-sided error bound discussed in chapter 2.

DEFINITION. Let γ_0, γ_1 be positive constants, and assume ϕ_h is a mapping from X_h to the set of real numbers $\mathbb{R}$. If for all $h \in (0,h_0]$ and all $w_n \in \mathbb{R}^s$ the relations (3.1), (3.2) imply that

$$\gamma_0 \cdot \phi_h[w] \le \|\tilde{u} - u\|_h \le \gamma_1 \cdot \phi_h[w], \tag{3.4}$$

then (3.4) is called a *two-sided error bound for method* (3.1).

The error bound (2.9) provides an example of (3.4). In this case the functional ϕ_h has the remarkable property that it is independent of the function f appearing in (2.3).

In the rest of this paper we consider the problem under what conditions the result (2.9) allows a generalization for the case of numerical methods that are more general than Euler's method (2.3), and of arbitrary seminorms $\|x\|_h$.

3.3. In order to formulate concisely a criterion for the existence of two-sided error bounds, we introduce the *summation operator* S. For any $h \in (0,h_0]$ and any $x = (x_0, x_1, \ldots x_N) \in X_h$ we define

$$y = Sx$$

by $y = (y_0, y_1, \ldots, y_N) \in X_h$ with

$$y_0 = 0, \quad y_n = h \cdot (x_0 + x_1 + \ldots + x_{n-1}) \quad (n = 1,2,\ldots,N).$$

The following theorem gives a simple condition on the seminorm $\|x\|_h$ (viz. statement 3 in theorem 1) which is necessary and sufficient in order that suitable two-sided error bounds exist for all methods of type (3.1).

THEOREM 1. (Two-sided error bounds with absolute seminorms). *Assume* $\|x\|_h$ *is a given absolute seminorm in* X_h *(for all* $h \in (0,h_0]$*). Then the following three propositions are equivalent.*

1. *For any method of type* (3.1) *there exists a two-sided error bound* (3.4) *with a functional* ϕ_h *which is independent of the function* F *appearing in* (3.1).
2. *For any method of type* (3.1) *there exists a two-sided error bound* (3.4) *of the special form*

$$\gamma_0 \cdot \|v\|_h \le \|\tilde{u} - u\|_h \le \gamma_1 \cdot \|v\|_h \tag{3.5}$$

   where $v = (v_0,v_1,\ldots,v_N)$ *is defined by* $v_n = w_0 + h\sum_{j=1}^{n} w_j$ $(n = 0,1,\ldots,N)$.
3. *There is a constant* δ *such that for all* $h \in (0,h_0]$, $x \in X_h$ *we have the inequality* $\|Sx\|_h \le \delta \cdot \|x\|_h$.

This theorem can be proved, for instance, by a straightforward application of lemma 6 in [14].

It is easily verified that e.g. the seminorm of example 1 satisfies condition 3. Hence by virtue of theorem 1 there exists an error bound (3.5) where $\|x\|_h$ stands for $\|x\|_h = \max_{0\le n\le N}|x_n|$. Note that the bound thus obtained is of the same form as (2.9).

On the other hand, the seminorm of example 2 is easily seen to violate condition 3. Theorem 1 thus proves the surprising fact that the error bound (2.9) allows no generalization in case $\max_{0\le n\le N}|\tilde{u}_n-u_n|$ would be replaced by $|\tilde{u}_N-u_N|$.

4. GENERALIZATIONS OF THEOREM 1

4.1. In this chapter we review some extensions of theorem 1. First of all it should be noted that the class of numerical methods covered by formula (3.1) is still rather restricted. Therefore we replace (3.1) by the more general formula

$$(4.1)\qquad h^{-1}\cdot(u_n - Ku_{n-1}) = F(t_{n-1},u_{n-1};h) \qquad (n = 1,2,\ldots,N),$$
$$u_0 = c_0(h),$$

where K is a fixed square matrix and $c_0(h)$ is a starting vector depending on $h \in (0,h_0]$. We also allow that the vectors U(t), f(t,U(t)), c, of the initial value problem (2.1), belong to some vectorspace whose dimension is smaller than the dimension of the space to which u_n, $F(t_{n-1},u_{n-1};h)$, $c_0(h)$, occurring in (4.1), belong. The class of methods (4.1) now includes for instance (linear) multistep methods, provided the latter are rewritten as one-step methods (see e.g. SKEEL [10]).

We consider the following perturbed version of (4.1).

$$(4.2)\qquad h^{-1}\cdot(\tilde{u}_n - K\tilde{u}_{n-1}) = F(t_{n-1},\tilde{u}_{n-1};h) + w_n \qquad (n = 1,2,\ldots,N),$$
$$\tilde{u}_0 = c_0(h) + w_0.$$

Under mild conditions on the matrix K theorem 1 allows a generalization which deals with the methods (4.1), (4.2), instead of (3.1), (3.2) (see [13]). According to this generalization there still exists a suitable two-sided error bound for (4.1) if the maximum-norm (see example 1) is used, and there exists no such error bound in case the seminorm from example 2 is used.

SKEEL [10] recently derived a simple functional ϕ_h which can be used (in a two-sided error bound of type (3.4) for method (4.1)) in case $\|\tilde{u}-u\|_h$ stands for the maximum norm $\max_{0\le n\le N}|\tilde{u}_n - u_n|$. This functional reads

$$\phi_h[w] = \max_{0\le n\le N} |v_n|,$$

with

$$v_0 = w_0,\ v_n = E(w_0 + h\sum_{j=1}^{n-1} w_j) + hw_n \qquad (n = 1,2,\ldots,N),$$

where E denotes the so-called component of K corresponding to its eigenvalue 1 (see [10]). For a related result we refer to ALBRECHT [1, section 5.1].

4.2. Returning to the numerical method (3.1) we consider another generalization of theorem 1.

An obvious restriction of theorem 1 consists in the fact that it only applies to seminorms $\|x\|_h$ that are absolute. For instance, from theorem 1 it cannot be concluded whether two-sided error bounds for the methods (3.1) exist with the seminorm $\|x\|_h$ from example 3.

In [18] STUMMEL proved that for some special non-absolute seminorms, including the one from example 3, there do exist two-sided error bounds for (3.1).

The following theorem is a generalization of theorem 1 that applies to arbitrary seminorms, which may be absolute or not. For the proof and applications of the theorem we refer to [15], [2].

We note that the proof in [15] of the theorem differs in many respects from the proof of theorem 1 (essentially contained in [14]). The proof in [15] allows no straightforward generalization so as to apply also to the more general methods of type (4.1).

THEOREM 2. (Two-sided error bounds with arbitrary seminorms). *Assume the maximal stepsize* h_0, *appearing in condition* (3.3), *is so small that* $\lambda h_0 < 1$. *Let* $\|x\|_h$ *denote an arbitrary seminorm in* X_h (*for all* $h \in (0,h_0]$). *Then the following three propositions are equivalent.*

(i) *For any method of type* (3.1) *there exists a two-sided error bound* (3.4) *with a functional* ϕ_h *which is independent of the function* F *appearing in* (3.1).

(ii) *For any method of type* (3.1) *there exists a two-sided error bound* (3.4) *of the special form*

$$\gamma_0 \cdot \|v\|_h \le \|\tilde{u} - u\|_h \le \gamma_1 \cdot \|v\|_h \tag{4.3}$$

where $v = (v_0,v_1,\ldots,v_N)$ *with* $v_n = w_0 + h\sum_{j=1}^{n} w_j$ $(n = 0,1,\ldots,N)$.

(iii) *There is a constant* δ *such that for all* $h \in (0,h_0]$ *and all* $x = (x_0,x_1,\ldots,x_N)$, $\tilde{x} = (\tilde{x}_0,\tilde{x}_1,\ldots,\tilde{x}_N) \in X_h$ *with* $|\tilde{x}_n| \le |x_n|$ $(n = 0,1,\ldots,N)$, *we have the inequality* $\|S\tilde{x}\|_h \le \delta \cdot \|x\|_h$.

It is easily verified that the seminorm of example 3 satisfies condition (iii), while the one of example 4 does not. Consequently, with the seminorm of example 3 we have the error bound (4.3), which is equivalent to the bound given by STUMMEL [18]. Further, we may conclude that with the seminorm of example 4, statement (i) is not valid.

5. INTEGRO-DIFFERENTIAL EQUATIONS

5.1. In this chapter we consider the initial value problem for a system of s integro-differential equations of Volterra

$$(5.1) \qquad \frac{d}{dt}U(t) = f\Big(t,U(t),\int_0^t g(t,\tau,U(\tau))d\tau\Big) \quad (0 \le t \le T), \quad U(0) = c.$$

Here c has the same meaning as in the initial value problem (2.1), and f, g are given continuous functions from $[0,T] \times \mathbb{R}^s \times \mathbb{R}^s$ to $\mathbb{R}^s$, and from the set

$$\{(t,\tau,\xi) \mid 0 \le t \le T, \quad 0 \le \tau \le t, \quad \xi \in \mathbb{R}^s\}$$

to $\mathbb{R}^s$, respectively. We assume that f and g satisfy the Lipschitz conditions

$$|f(t,\tilde{\xi},\tilde{\eta}) - f(t,\xi,\eta)| \le L_1 \cdot |\tilde{\xi}-\xi| + L_2 \cdot |\tilde{\eta}-\eta|,$$
$$|g(t,\tau,\tilde{\xi}) - g(t,\tau,\xi)| \le L_3 \cdot |\tilde{\xi}-\xi|.$$

Here L_1, L_2, L_3 denote positive constants independent of $t \in [0,T]$, $\tau \in [0,t]$ and ξ, $\tilde{\xi}$, η, $\tilde{\eta} \in \mathbb{R}^s$.

In [11] SMIT considered the problem under what conditions there exist two-sided error bounds in the numerical solution of the initial value problem (5.1).

We shall discuss two simple, but typical, applications of the theory contained in [11].

For the ease of presentation we confine ourselves to the following numerical method, which is a straightforward generalization of Euler's method (2.3).

$$(5.2) \qquad h^{-1} \cdot (u_n - u_{n-1}) = f\Big(t_{n-1},u_{n-1},h\sum_{j=1}^{n-1} g(t_{n-1},t_j,u_j)\Big),$$
$$(n = 1,2,\ldots,N), \quad u_0 = c.$$

We note that the use of a finite sum in (5.2) to approximate the integral $\int_0^{t_{n-1}} g(t_{n-1},\tau,U(\tau))d\tau$, generates an error which is not present in Euler's method (2.3). Therefore, it is realistic to consider a

perturbed version of (5.2) which, in addition to the w_n occurring in (2.4), also contains perturbations w_n^* in the third argument of the function f. We thus arrive at the following generalization of (2.4)

$$(5.3)\qquad h^{-1}\cdot(\tilde{u}_n-\tilde{u}_{n-1}) = f\Big(t_{n-1},\tilde{u}_{n-1},h\sum_{j=1}^{n-1} g(t_{n-1},t_j,\tilde{u}_j)+ w_{n-1}^*\Big) + w_n$$

$$(n = 1,2,\ldots,N),\quad \tilde{u}_0 = c + w_0 .$$

5.2. Suppose $u_n,\tilde{u}_n$ satisfy (5.2), (5.3), respectively. We consider the problem whether there exists a two-sided bound for $\max_{0\le n\le N}|\tilde{u}_n-u_n|$ similar to the bound (2.9). To be more exact, we pose the question whether the $u_n,\tilde{u}_n$ computed from (5.2), (5.3) satisfy a relation of type

$$(5.4)\qquad \gamma_0\cdot\phi_h[w,w^*] \le \max_{0\le n\le N}|\tilde{u}_n-u_n| \le \gamma_1\cdot\phi_h[w,w^*].$$

Here the positive constants γ_0,γ_1 should be independent of $h\in(0,T]$, $w = (w_0,w_1,\ldots,w_N)$, $w^* = (w_0^*,w_1^*,\ldots,w_{N-1}^*)$, $w_n\in\mathbb{R}^s$, $w_n^*\in\mathbb{R}^s$. The constants γ_0,γ_1 may depend on f and g, but the functional ϕ_h should be independent of f, g.

It follows from [11, Theorem 6.2] that such a two-sided error bound of type (5.4) does not exist.

5.3. The above application of the theory in [11] has thus led to a negative result. We consider a second application yielding a positive one.

We define for $n = 0,1,\ldots,N-1$

$$y_n = h\sum_{j=1}^{n} g(t_n,t_j,u_j),\qquad \tilde{y}_n = h\sum_{j=1}^{n} g(t_n,t_j,\tilde{u}_j) + w_n^*.$$

It follows from [11, Theorem 6.20] that the $u_n,\tilde{u}_n$, computed from (5.2), (5.3), satisfy the two-sided error bound

$$\gamma_0\cdot\phi_h[w,w^*] \le \max\Big[\max_{0\le n\le N}|\tilde{u}_n-u_n|,\ \max_{0\le n\le N-1}|\tilde{y}_n-y_n|\Big] \le \gamma_1\cdot\phi_h[w,w^*].$$

Here $\gamma_0,\gamma_1 > 0$ are constants independent of $h\in(0,T]$, $w_n\in\mathbb{R}^s$, $w_n^*\in\mathbb{R}^s$. The functional ϕ_h is defined by

$$\phi_h[w,w^*] = \max\Big[\max_{0\le n\le N}|w_0+h\sum_{j=1}^{n} w_j|,\ \max_{0\le n\le N-1}|w_n^*|\Big].$$

We note that ϕ_h is independent of f and g.

REFERENCES

[1] ALBRECHT, P., *Numerische Behandlung gewöhnlicher Differentialgleichungen*, Report Jül-1274, Kernforschungsanlage Jülich, 1976.

[2] CALVO, M., *Two-sided error bounds for one-step methods in the numerical solution of differential equations*, Part II. Report No. 77-15, Institute of Applied Mathematics and Computer Science, University of Leiden, Leiden, 1977.

[3] DAHLQUIST, G., *Stability and error bounds in the numerical integration of ordinary differential equations*, Trans. Roy. Inst. Technol., Stockholm, Nr. 130 (1959).

[4] HENRICI, P., *Discrete variable methods in ordinary differential equations*, New York, J. Wiley & Sons, 1962.

[5] HENRICI, P., *Problems of stability and error propagation in the numerical integration of ordinary differential equations*, in: Proceedings of the International Congress of Mathematicians 1962, Djursholm: Inst. Mittag-Leffler 1963.

[6] HULL, T.E. & W.A.J. Luxemburg, *Numerical methods and existence theorems for ordinary differential equations*, Numer. Math. 2 (1960), 30-41.

[7] HUNGER, S., *Intervallanalytische Defektabschätzung zur Lösung mit exakter Fehlererfassung bei Anfangswertaufgaben für Systeme gewöhnlicher Differentialgleichungen*, Z. Angew. Math. und Mech. 52 (1973), T 208-T 209.

[8] MARCOWITZ, U., *Fehlerabschätzung bei Anfangswertaufgaben für Systeme von gewöhnlichen Differentialgleichungen mit Anwendung auf das Reentry-Problem*, Numer. Math. 24 (1975), 249-275.

[9] RUDIN, W., *Functional analysis*, New York: Mc Graw-Hill Book Comp. 1973.

[10] SKEEL, R., *Analysis of fixed-stepsize methods*, SIAM J. Numer. Anal. 13 (1976), 664-685.

[11] SMIT, J.H., *Two-sided error bounds in the numerical solution of*

generalized Volterra integro-differential equations. Ph.D. Thesis, Leiden University, 1976.

[12] SPIJKER, M.N., *On the structure of error estimates for finite-difference methods,* Numer. Math. 18 (1971), 73-100.

[13] SPIJKER, M.N., *Two-sided error estimates in the numerical solution of initial value problems,* In: Numerische Behandlung nicht-linearer Integro-differential- und Differentialgleichungen, Lecture Notes in Mathematics 395, pp.109-122, Berlin: Springer 1974.

[14] SPIJKER, M.N., *On the possibility of two-sided error bounds in the numerical solution of initial value problems,* Numer. Math. 26 (1976), 271-300.

[15] SPIJKER, M.N., *Two-sided error bounds for one-step methods in the numerical solution of differential equations,* Part I. Report No. 77-14, Institute of Applied Mathematics and Computer Science, University of Leiden, Leiden 1977.

[16] STETTER, H.J., *Analysis of discretization methods for ordinary differential equations,* Berlin: Springer-Verlag 1973.

[17] STUMMEL, F., *Approximation methods in analysis,* Lecture Notes Series of Aarhus Universitet (1973).

[18] STUMMEL, F., *Biconvergence, bistability and consistency of one-step methods for the numerical solution of initial value problems in ordinary differential equations,* In: Topics in numerical analysis II, pp.197-211, London: Academic Press 1975.

L.E.J. BROUWER: INTUITIONISM AND TOPOLOGY

W.P. van Stigt

For the various sections of the mathematical community Brouwer's fame seems confined to their own specialisms. For the topologist he is the man of the fixed-point theorem, who introduced revolutionary methods and solved the problem of the invariance of dimension. Alexandroff claims "that the power of these methods reached far beyond the proof of invariance of dimension; they were the kind of creation which have made Brouwer with Cantor and Poincaré the founder of modern topology" [1]. To the logician Brouwer represents the extreme challenge of the traditional principles of logic in the practice of mathematics, in particular of the Principle of the Excluded Middle. This challenge and his more stringent definition of the negative formed the basis of a completely new practice of mathematical logic, the modern intuitionist school.

For the serious mathematician-philosopher Brouwer's intuitionist challenge extends beyond the subtleties of mathematical logic; to him Brouwer is perhaps the last of that rare breed of mathematicians and philosophers who concerned themselves with the fundamental, metaphysical questions of the nature of mathematics, the nature of mathematical truth and reality. An uncompromising reformer, who passionately tried to halt the increasingly formalistic trend in the practice of mathematics and made gallant attempts to systematically re-construct the whole of mathematics in accordance with his constructive philosophy. At the height of Brouwer's international fame Hermann Weyl renounced his own contribution to the solution of the continuum-problem and proclaimed 'a crisis in the foundations of mathematics resulting from Brouwer's findings: "Und Brouwer, das ist die Revolution!"

---

*) The author gratefully acknowledges the generous help of ZWO and the hospitality of Utrecht University during the academic year 1976/77 which enabled him to gather biographical information and set up the Brouwer Archief.

.. the one mathematician who at last had solved the problem of the continuum which since ancient times had defeated even the greatest minds [2].

There is no doubt as to the prominent place Brouwer holds both in the field of Foundation studies and in Topology. Perhaps due to some professional jealousy topologists and foundationalists each claim Brouwer as their champion and play down his role and contribution in the other field, or it may be that in our age of specialization we cannot but doubt the seriousness and value of contributions in what we regard as two completely separate specialisms. It was the declared policy of the editors of Brouwer's *Collected Works* to separate his foundational and topological work. Heyting speaks in the Introduction of "two almost disjoint parts".

No attempt has yet been made to relate Brouwer's activity in these two fields. Foundationalists hardly mention his contribution to Topology or use it only as evidence and proof of Brouwer's genius as a mathematician and express their doubts about Brouwer's own claim that in his topological work he "was always careful to derive only such results as could be expected to find their place in the new system after the systematic construction of an intuitionist set theory" [3]. Topologists express surprise and disappointment at Brouwer's "loss of interest in Topology" after a brilliant but brief blaze of activity in the period 1909 - 1912. Newman and Kreisel diagnose "a shift in Brouwer's interest to philosophy after the first world-war" [4] and Freudenthal and Heyting blame "national isolation during World-war I for Brouwer's loss of interest in Topology" [5].

As part of our own attempt at a Brouwer Biography*) I venture to answer some of these questions' the reasons for the "shift" in Brouwer's interest, the timing of this shift and the relation between these apparently disjoint interests. Considering all the evidence now available, I think that the answer to these questions is to be found mainly in the person and character of Brouwer and the particular circumstances he found himself in at the time. We have been fortunate in having found and gathered together in the past year most of Brouwer's papers including his correspondence with D.J. Korteweg, his 'promotor' in more sense than one, i.e. the professor supervising a doctoral thesis.

I have no doubt that Brouwer's chief mathematical interest throughout his life has been the Foundations of Mathematics as philosophical reflection

*) Dirk van Dalen and the author in a joint effort have started to gather biographical information in preparation of a Brouwer biography.

on the nature of mathematics, its origin, its fundamental principles and concepts and its relation to other human activities, and that his passionately held views on these fundamental aspects were the inspiration and driving force behind most of his mathematical work.

It is no coincidence that on all important occasions in his academic career he chose to speak out on foundational matters rather than topology even at a time when his topological fame reached its climax; both his doctoral thesis (1907) and his Inaugural Address (1912) were major contributions in his intuitionist campaign. This preference goes back to his student days. On the occasion of the conferment of the Honorary Doctorate on Gerrit Mannoury, Brouwer recalled his disillusionment with the undergraduate mathematics course: "I could see the figure of the mathematician only as a servant of natural science or as a collector of truths, fascinating by their immovability but horrifying by their lifelessness, like stones from barren mountains of disconsolate infinity" [6]. He gratefully acknowledged his indebtedness to the man who had opened his eyes to the possibility of extending his mathematical activities to the domain of foundational studies. Korteweg had hoped that his talented student would use the opportunity of his doctoral thesis to produce a brilliant piece of mathematics. Brouwer, however, saw it as the occasion to launch the manifesto of his deeply felt views on the nature and practice of mathematics which would shock the world. When after two years of reading and studying Brouwer produced his magnum opus Korteweg was stunned. He crossed out and rejected the greater part of Chapter II, the most important part of the thesis, and angrily wrote to Brouwer: "I have again considered whether I can accept it as it stands, but honestly Brouwer, I cannot ... it has nothing to do with mathematics ... it is bizarre [7]. In the following heated exchange Brouwer pleaded passionately: "You know that when two years ago I chose my subject it was not for lack of ability on my part to tackle a more 'normal' topic, but only because I felt strongly drawn towards this subject ... how mathematics is rooted in life and what should be the starting point of all our theories" [8]. Korteweg, however, remained adamant, Brouwer had to agree to the removal of practically the whole of his treatise on the nature of mathematics and the physical sciences; all that was left was a sober statement on the nature of mathematics together with a brief account of the Primordial Intuition of Time. The main body of the all-important Chapter II is now taken up by a survey of the physical sciences and a criticism of Russell. When some months later Korteweg made some critical comment on mathematical

detail of what was left of Chapter II, Brouwer, still resentful, blamed the general incoherence of the chapter on the removal of the main theme which these mathematical references only served as props: ".. in my mind they were originally only incidental offshoots of a fundamental idea which held them together - and which is not any more to be found in the dissertation; they only had secondary importance. After their sudden appearance in the full limelight, substituting for their former leader, it was not possible to doll them up quickly in such a way that they together by themselves could save the entire performance. At least that is my impression when I look at the chapter" [9].

Brouwer's second attempt to set the world of Foundations of Mathematics ablaze met with a similar fate. Towards the end of 1907 he submitted his now famous article on the Unreliability of the Principles of Logic to Bierens de Haan, the chief editor of the newly founded *Tijdschrift voor Wijsbegeerte*. It is his first public rejection of the Principle of the Excluded Middle. Far from receiving an enthusiastic response for his revolutionary ideas Brouwer was told that his contribution was unintelligible. Only after much persuasion could he convince the secretary to plead on his behalf with the editorial board; they only reluctantly and conditionally agreed to its publication. Kohnstamm wrote: "In today's meeting the editorial board decided to publish your paper in the February issue. But, as I expected, only after considerable resistance; most members stated they had not understood a word of it.. I only succeeded in overcoming their resistance by repeating your promise in your letter of 7 December to Bierens de Haan to elaborate on your point of view in a series of articles a little more comprehensible to non-mathematicians" [10].

The promised articles never materialized, Brouwer became more and more disillusioned: even those who recognized his mathematical talents were not prepared to listen to his philosophical views. At the International Congress of Mathematicians in Rome in the spring of 1908 he contributed two papers, using some of the ideas of his thesis. In a letter from Italy he complained to Korteweg about the poor reception these papers received. But he also speaks of the inspiration he experienced from the company of great men, "the admiration and awe I felt in the presence of these heroes of abstraction", in particular of the inspiring presence of Poincaré: "To be able to raise oneself to a view from where one can produce a lecture such as Poincaré's l'Avenir des Mathématiques, whose truthfulness everyone experiences and accepts as a guide in his work, this to me seems to be the

highest ideal for any mathematician" [11].

Brouwer began to realize that his own revolutionary views on mathematics would only be listened to if pronounced from such a height and with an authority such as Poincaré had earned through his mathematical work; he first had to gain the respect of his fellow-mathematicians on account of his own purely mathematical work.

The need to prove his mathematical worth became the more urgent as the down-to-earth necessities of life began to make themselves felt: he had to earn himself a living. For more than a year he had been completely dependent on his wife's earnings (and it would be another four years before he earned his first salary).

Brouwer's letters to Korteweg leave no doubt as to his ambition: he wanted no less than a professorship or at least a lectureship. Yet he was not prepared to follow the traditional Dutch slow route to academic prominence via a successful teaching career in a gymnasium, gradually and carefully preparing one's candidature for some future academic vacancy through frequent contacts with academics in the meetings of the Wiskundig Genootschap, publications in the national academic press and by giving one's free service to the University as a 'privaat-docent'. Brouwer knew his limitations and his unsuitability for teaching in schools. He hated teaching, "submitting to the 'bon plaisir' of one's audience" [12]. Korteweg's advice was "to let no opportunity pass to make yourself known and show what you would be worth as a lecturer or professor" [13]. Brouwer reluctantly and resentfully accepted Korteweg's advice to apply for the humble and unpaid licence to teach at the University of Amsterdam as a 'privaat-docent'. He was determined and used all his energies to force a short cut to academic high office. In a letter to Korteweg already in November 1908 he advanced the revolutionary idea that such high office should be awarded solely on academic merit i.e. international reputation as a scholar*). He purposely set out in quest of international recognition, bypassing the parochial national routes and wooing the mathematical giants of the time. He was shrewd enough to realize that he had to bury the controversy on the Foundations of Mathematics for the time being and he was prepared to remain

---

*) About ten years later Brouwer returned to this theme in a one-man campaign on the appointment of professors and the establishment of mathematics departments as research institutes in the national press, the Wiskundig Genootschap and the Royal Academy.

silent on this issue, closest to his heart, until he had achieved his ambition. He would not speak out on the matter of Foundations until the very day that he acceeded to the chair of mathematics at Amsterdam University in 1912.

In the isolation of his 'Hut' on the Laren heath Brouwer now began his pure-mathematical research. The unfinished problems of his thesis were an obvious and natural subject for further investigations. In the heat of the argument with Korteweg he had claimed "to have solved Hilbert's fifth Paris problem (The Lie notion of continuous Transformation groups without the assumption of differentiability) for the simple linear case" [14]. In his Rome lecture he posed the more general problem of "determining all finite continuous groups of the n-dimensional manifold" [15]. His thorough study of Lie Groups, approaching them through manifolds and one-to-one mappings [16] was his first work which he considered good enough to offer to Hilbert for publication in the 'Mathematische Annalen'. It set him firmly on the road to Topology. In October 1908 Brouwer gave his first lecture to the Wiskundig Genootschap [17], a general survey of research to date, *On Plane Curves and Plane Domains*, still firmly based on Schoenfliess's results but posing many problems and a programme of his work for the years to come. It was, however, his discovery of flaws in Schoenfliess's work during the winter of 1908/09 that gave him the break he had been waiting for, an opportunity to conduct a searching mathematical investigation in the international limelight and under the eyes of 'the great'. Early in 1909 he writes to Hilbert, "When during the last winter I had the second part of my *Finite Continuous Groups* ready to send in for publication in the Mathematische Annalen I suddenly noticed that Schoenfliess's investigations into the Analysis Situs of the plane, on which I so entirely based my work, cannot in all its parts be sustained; this also calls into question my own group-theoretical results. To clear this matter it was necessary to work thoroughly through the relevant parts of Schoenfliess's theory and determine precisely on which results we can rely in full confidence. That's how the enclosed work originated" [18]. Brouwer enclosed the original, rather aggressive draft of his *Zur Analysis Situs* with a request to publish it in the Mathematische Annalen. Another copy he sent to Schoenfliess. Hilbert's reply was a simple acceptance and concerned only the size of the illustrations. But it was Schoenfliess's reaction that pleased Brouwer most: "At last some fish has taken the bait! .. I am so glad at last to receive something more than just a polite postcard about my work .. Schoenfliess has

gone into my paper in considerable detail, but I had to put the thumbscrews on rather hard" [19].

Zur Analysis Situs marks the beginning of a series of many papers of Brouwer's hand during the years 1909-1912 which completely changed the course and role of Topology. He had found his strength and in that time achieved his ambition to rub shoulders with the 'great', drawing their attention by his sharp and critical analysis of the work of others*) and the simplicity and beauty of his own methods. His contacts with Hadamard of the 1908 Rome Congress had developed into a personal friendship. In the summer of 1909 he met Hilbert for the first time in The Hague, and the relation Hilbert-Brouwer at that time can only be described as mutual admiration and friendship; he had the great satisfaction of receiving a personal tribute from Poincaré and was admired and encouraged by Klein. Success, ambition and pressing financial need spurred him on into frantic activity over a wide front of topology, culminating in his work on dimension. Casually but proudly he writes to Hilbert at the end of a letter, "Apart from a new paper on group theory I am preparing for submission to the editorial board of the Annalen a paper in which I solve the problem of the invariance of dimension, showing that there cannot be a one-to-one mapping between spaces of even and odd dimensions" [20]. In October that same year he divulged his remarkable proof of the invariance of dimension for the first time in public at a meeting of the Wiskundig Genootschap [21].

The fulfilment of his other ambition, a professorship with its financial security, was greatly due to the relentless efforts of his promotor, D.J. Korteweg, who in two long campaigns lobbied the university authorities, the city fathers and his colleagues of sister universities. It is amusing to unravel the intrigue and enthusiastic plotting of Korteweg and his fellow-mathematics-professor Hendrik de Vries to secure Brouwer's membership of the Royal Academy and so to advance his candidature for a professorship; and to read their long and carefully worded petitions supported by glowing references from such giants as Hilbert, Poincaré and

---

*) A letter of Carathéodory to Hilbert is evidence of Brouwer's reputation for rigour: "The length of the manuscript is due to my attempt to be absolutely rigorous (you know how in this part of mathematics one sins against this - with the exception of course of Brouwer)" .. (Niedersachs. Staatsbibl. Göttingen).

Borel*). As a true mathematician Brouwer must have enjoyed the satisfaction and excitement of his creative topological work, but his mood in those years is overwhelmingly one of resentment and impatience, hurt pride and resentment at not being offered an academic position any earlier. His letters to Korteweg are full of angry outbursts .., "a second slap in the face .. after being passed over in Delft for someone younger and inferior" [22]. In another letter he sneeringly suggests that Korteweg perhaps ought to tell the University Authorities who Poincaré - his referee - is [23]. There was resentment too and the beginning of bitter enmity towards those who did not immediately and openly acknowledge his mathematical superiority, especially towards the topological school of Lebesgue.

On the other hand the absence of lecturing duties and the administrative commitments of a university post gave him the freedom and the opportunity to devote all his time and energy to research - and he needed the strong pressure to apply himself and force his energies in the direction of pure-mathematical research. In private he often confessed to his friends that pure mathematics "bored" him. But he maintained his calculated silence and managed to keep his revolutionary ideas on the nature of mathematics from those whose support he still needed. In a long letter to Hilbert [24] - seven pages foolscap - he volunteered his criticism on *die Grundlagen der Geometrie*, but his suggested improvements and criticisms only concern the technical details of Hilbert's addendum as published in 1902, not a word of criticism on the most fundamental issue, Hilbert's formal axiomatic method, which was launched in the first chapter of *die Grundlagen* as published in 1899 and which Brouwer had so severely criticized in his thesis. (That Brouwer did not expect Hilbert to know about his contributions in Dutch is clear from a letter to Korteweg in which he expresses surprise and his "mixed feelings" at hearing from Hilbert that he had seen some of his Dutch contributions [25].)

---

*) Hilbert's testimony speaks of Brouwer as "a scholar of unusual talent, of the most rich and wide-ranging knowledge and rare sharpness of mind.. It is characteristic of Brouwer never to be satisfied with the easy results which the usual research offers; he always becomes involved in especially difficult and deep problems and only leaves them when he has succeeded in a solution which completely satisfies him. I think in particular of his solution of the problem of finite continuous groups and his marvellous proof of the Jordan Curve theorem" (Hilbert to Korteweg, 6-2-1911, DJK 80).

Brouwer finally broke his silence on the Foundations of Mathematics on the day that he officially acceeded to the chair of mathematics. He did not use the occasion of his inaugural address to reveal to the world yet another of his brilliant innovations in the field of topology for which he had received early fame and was awarded his chair nor did he choose to speak about the two areas of his new responsibility, the theory of functions and analytic geometry. The newly found security gave him the freedom at last to speak out publicly on deeper, fundamental issues and to resume his campaign against an increasingly mechanistic and formalistic trend in the whole field of mathematics. He now could speak also confidently and authoritatively, knowing that his voice would be listened to. There is an element of defiance in his opposition to the established policy of world famous mathematicians like Peano and Hilbert whom he now knew personally; perhaps even an element of mischievous revenge in Brouwer's publicly repeating views - in some cases literally -which had been rejected by his promotor and removed from the thesis. (It is a tribute to Korteweg's magnanimity that within a year he would voluntarily step down to give Brouwer the chance to become 'Ordinarius' and so be retained for Amsterdam.)

Brouwer's accession to his professorial chair marks the beginning of his full-scale intuitionist campaign and the end of his involvement in topology. He did not purposely abandon topology. However, his university duties and other interests began to make increasing demands on his time and energy: In his first year he had to prepare courses in coordinate geometry and function theory, in 1913 he became fully responsible for all courses in mechanics; at the death of Schouten he also started a long term of office on the committee of the Wiskundig Genootschap. But above all, having secured his academic position he was now free to pursue his interests in the foundations of mathematics, in 'Significs' and in national and university politics.

The biographical evidence leaves no doubt that Brouwer's first love and the chief concern throughout his life has been the fundamental problems, the philosophy of mathematics. This is also clear from the style of his writing which betrays his passionate involvement in the struggle for the purity of mathematics. At a time when mathematical practice underwent perhaps its most fundamental change he recognized in the increasing linguistic and formalistic emphasis a threat to what he believed to be the true nature of mathematics, a threat also to a practice of mathematics which

came most natural to him, through profound thought, "inner vision", rather than mechanical manipulation of symbols on paper.

I do not wish to maintain that the reasons for Brouwer's involvement in topology are solely mercenary and accidental and that there is no connection between his topology and intuitionism other than a means to an end. If there is to be a more fundamental reason for his dual interests, the key perhaps may be found in Brouwer's own character, his misanthropic and solipsistic preference for 'inner vision'. In his first public lecture as a 'privaat-docent', *The Nature of Geometry*, Brouwer proclaimed the 'problems of analysis situs' as the most topical and urgent; however, his emphasis on the visual perceptive simplicity of geometric topology is the clearest indication of its appeal to Brouwer's intuitive conception of mathematics. In the final lines he equates 'topological' to 'geometric' and 'formula-less': "Therefore also in other theories (- i.e. besides projective geometry -) even if one succeeds in founding them on analysis situs, coordinates and formulae need not entirely be banned; but the 'formula-less', the 'geometric' treatment will be the starting point while the analytic treatment becomes a dispensable expedient. It is to the possibility and desirability of the priority of the geometric treatment that I have wanted to draw your attention" [26]. The geometric medium allowed the 'intuitive' approach and promoted a simplicity which is the hall-mark of Brouwer's topological methods and solutions. This approach and his exceptional genius produced the master strokes, the brilliant and grand ideas which made Brouwer one of the great topological pioneers. He was ill-suited nor prepared to undertake the more laborious and mundane task of constructing a detailed and systematic follow up. Writing to Korteweg about his own more technical-mathematical contributions he remarks apologetically, "I consider these publications to be of a much lower standard than my other work; anyone could have obtained these results" [27].

As to Brouwer's move away from topology in 1912 and his return to the field of foundations, there is some continuity and a direct link between Brouwer's 'topological' interests in 1912 and his intuitionist preoccupation during the following years. It is quite remarkable that Brouwer's first breakthrough in 1909 and his dramatic entry into the field of topology was sparked off by his critical analysis of Schoenflies's work and that again in 1912 the change back to foundations was partly due to his involvement in Schoenflies's work in set theory.

In both cases Brouwer's work in the following years was determined by it. Early in 1912 Brouwer's attention began to centre on set theory. He describes the reason for this interest in a letter to Hilbert :

" You know probably that for some time I have been busy with the re-edition of Schoenflies's Bericht über die Mengenlehre. What happened in this: I was repeatedly pressed from different quarters to write a book on the theory of sets, since the present textbooks and articles in Encyclopediae are inadequate and superficial. When I was in Göttingen in the summer of 1911 I was asked again and about the same time I heard that Schoenflies was preparing a new edition of his book. I thought it would solve the problem and save my time if I could control the edition during the printing stage and where necessary improve it and supplement it. The problem of persuading Schoenflies to allow me the control was soon solved when Fricke, who knows him personally, offered to mediate. Schoenflies was very pleased with my proposal made by Fricke. But I am now in some difficulty with hum on the nature and the degree of my cooperation; we differ on the fundamental issue. Schoenflies likes to restrict my role to correcting false theorems and proofs, whereas I, of course, also aim improving and complementing them" 28).

Brouwer's "cooperation" had in his own imagination grown into a "right of control". A different version emerges from his letter to Schoenflies in December 1911, in which Brouwer offers his services: "When I discussed with Fricke the new edition of your Bericht I mentioned that my collaboration might be helpful. I now hear that Fricke has discussed the matter with you and that you like the idea. I am wholly at your service " 29).

During 1912 Brouwer becomes more and more involved in editing Schoenflies's Bericht 30) and more and more frustrated that Schoenflies did not do as he was told. Mrs. Brouwer writes to her sister-in-law, " At the moment all his time is taken up with correcting the second print of the German standard work in his special field, the work by Schoenflies. He was requested by the German authorities to offer his generous services... . that same Schoenflies with whom he quarrelled before, just as obstinate and headstrong as then .. " 31).

Brouwer's letters to Hilbert throughout 1913 are dominated by his obsession with the edition of Schoenflies's work, he repeatedly urges Hilbert to put pressure on Schoenflies:

On 16th March 1913 : " Schoenflies presses me to hurry up. I feel it all as a Sisyphus task, after my corrections Schoenflies tries to improve it

again and brings in new errors. The solution perhaps might be pressure from a third party. I know that Schoenflies will be visiting Göttingen next week and will probably discuss the Bericht with you. Perhaps you could suggest that he should allow me more freedom. His great respect for you might help. Please, drop a hint in that direction. It will be good for set theory and mathematics" 32).

On the 23rd April he again asks Hilbert to intervene: " I am not his assistent, nor do I do it out of frienship; my concern is a good book on the theory of sets ... and at the moment I am the most competent person .. " 33).

On the 16th June " Your suggestions to Schoenflies have not worked very long ... He makes errors any student would be ashamed of. But I remain at my post trying to salvage what can be saved " 34).

On the 4th July " Schoenflies gets worse. If there is not a complete change I shall have to give up the work on which I spent more than eight months ... I have not done any work of my own " 35) and on the 5th September: " Once more I ask for your help against Schoenflies in the interest of science. On one of the most important points after endless efforts I succeeded in making him delete a proof and replace it by one of mine. But now at the last moment, blind to his mistake (and honestly you don't need much intelligence to see them) he wants to re-introduce it .. I beg you to send him a telegram immediately and tell him that he must give in to me " 36) etc.

Brouwer's complaint that he had not been able to do any work of his own is echoed by Mrs Brouwer in a letter to her sister-in-law in which she specifically blames Schoenflies for Brouwer losing the contest in proving Poincaré's last theorem : " Bertus thinks he cannot stand it any longer. It also made him lose the priority in solving the famous last problem of Poincaré " 37). From the Brouwer - Korteweg correspondence it appears that during the summer of 1912 Brouwer had been working on the solution of Poincaré's problem: Just before his Inaugural Address he writes to Korteweg: " As to my solution of Poincaré's problem, it will take me some more weeks. Please don't speak to anyone about it. When I have a fully worked out version I shall present it to the Academy " 38).

Instead of using the occasion of his Inaugural Address to reveal his solution of Poincaré's last problem Brouwer returns to the field of Foundations; after years of silence he speaks out openly again on Intuitionism. But his emphasis has shifted away from logic, the

Principle of the Excluded Middle is not even mentioned. Neither do we hear another diatribe against science or the application of mathematics. The offensive is mainly directed at Zermelo's axiomatic treatment of Set Theory.

In his review of Schoenflies's book Brouwer elaborates on the deficiencies in Schoenflies's work and in Set Theory at the time. He describes Schoenflies's Bericht 39) as a useful survey of Set Theory to date. His criticism are that " that Schoenflies keeps away from the fundamental philosophical problems" and that because of this lack of a proper foundation the work fails to satisfy the intuitionist as well as the formalist: " for the formalist there is too little, for the intuitionist.. who only recognizes well-constructed sets .. there is too much ". Brouwer diagnosed a general lack of a theory of sets which questions assumptions and goes back to fundamental philosohical issues. The construction on such a theory of sets became his task in the years ahead and would result in his intuitionist set-theoretical contributions in 1917 and 1918.

As to Brouwer's own opinion about his contribution to mathematics, he saw his own rôle primarily as that of a reformer of general mathematical practice, the father of intuitionism. History may well bear him out. After the rapid advances in topology his name and his contributions already seem to fade into oblivion. In the field of foundations, however, his extreme and consistent challenge of a formal and mechanistic conception of mathematics has remained as the only viable alternative philosophy of mathematics which still intrigues the seekers of mathematical truth all over the world.

NOTES AND REFERENCES

[1] ALEXANDROFF-HOPF, *Topologie*, Springer, 1930.

[2] *Ueber die neue Grundlagenkrise der Mathematik*, Math. Zeitschr. Band 10, Heft 1/2, p.47.
(Cf. also Weyl's tribute in his *Philosophy of Mathematics and Natural Science*, Princeton 1949, p.54: "Mathematics with Brouwer gains its highest untuitive clarity. He succeeds in developing the beginnings of analysis in a natural manner, all the time preserving the contact with intuition much more closely than had been done before".)

[3] *Intuitionistische Verzamelingsleer*, KNAW Versl. 29 (1921), p.798.

[4] Biographical Memoirs of the Royal Society (London).

[5] *Levensbericht L.E.J. Brouwer*, Jaarboek KNAW, 1966-1967.

[6] *Jaarboek van de Universiteit van Amsterdam 1946/1947*.

[7] Korteweg to Brouwer, 5-11-1906, Brouwer Archief DJK 22.
(Among Korteweg's papers we found a first draft of this letter, even stronger: "It must be removed, with root and branch!" ("Het moet eruit, met wortel en tak!), DJK 21.)

[8] Brouwer to Korteweg, 5-11-1906, DJK 20.

[9] Brouwer to Korteweg, 11-1-1907, DJK 32.
(For the complete text of the rejected parts see W.P. VAN STIGT, *The Rejected Parts of Brouwer's Dissertation*, publication expected in Historia Mathematica.)

[10] Kohnstamm to Brouwer, 3-1-1908 (Brouwer Archief).

[11] Brouwer to Korteweg, April 1908, DJK 39.

[12] Brouwer to Korteweg, 8-6-1909, DJK 52.

[13] Korteweg to Brouwer, 16-2-1909, DJK 43.

[14] Brouwer to Korteweg, 5-11-1906, DJK 20.

[15] *Die Theorie der endlichen kontinuierlichen Gruppen, unabhängig von den Axiomen von Lie*, Atti IV Congr. Intern. Mat. Roma II, p.109.

[16] *Die Theorie der endlichen kontinuierlichen Gruppen, unabhängig von den Axiomen von Lie*, I. M A 67, pp.246-267.

[17] Brouwer Archief BMS 4.

[18] Brouwer to Hilbert, 14-5-1909, DHI 1

[19] Brouwer to Korteweg, 18-6-1909, DJK 58.

[20] Brouwer to Hilbert, 18-3-1910, DHI 7.

[21] *De invariantie van het aantal dimensies eener ruimte,* Brouwer Archief, BMS 6A.

[22] Brouwer to Korteweg, 28-11-1910, DJK 75.

[23] Brouwer to Korteweg, 16-4-1912, DJK 112.

[24] Brouwer to Hilbert, 28-10-1909, DHI 6.

[25] Brouwer to Korteweg, 22-5-1909, DJK 50.

[26] *Het Wezen der Meetkunde* (The Nature of Geometry), Amsterdam, 1909.

[27] Brouwer to Korteweg, 22-5-1909, DJK 50.

[28] Brouwer to Hilbert, 16-1-1913, DHI 16.

[29] Brouwer to Schoenflies, 12-12-1912, ASC 10.

[30] A. Schoenflies, *Die Entwickelung der Mengenlehre und ihrer Anwendungen,* Leipzig-Berlin 1913 (2nd edition of *Die Entwickelung der Lehre von Punktmannigfaltigkeiten,* Jahresber. der D.M.V., vol 8 (1900).

[31] Mrs. E. Brouwer-de Holl to Mrs. L. Brouwer-van der Spil, 12-4-1913. That Mrs. Brouwer was under the impression that Hilbert had more or less told Brouwer to undertake the control of the work is evident from an added note of her hand in a letter by Brouwer to Hilbert "If he goes mad because of Schoenflies it will be thanks to you" (11.9.1913 DHI 22).

[32] Brouwer to Hilbert, 16-3-1913, DHI 16.

[33] Brouwer to Hilbert, 23-4-1913, DHI 17.

[34] Brouwer to Hilbert, 16-6-1913, DHI 18.

[35] Brouwer to Hilbert, 4-7-1913, DHI 20.

[36] Brouwer to Hilbert, 5-9-1913, DHI 21.

[37] Mrs. E. Brouwer-de Holl to Mrs. L. Brouwer-van der Spil, 12-4-1913.

[38] Brouwer to Kórteweg, 30-9-1912, DJK 116.

39 Review of: A. Schoenflies und H. Hahn, *Die Entwickelung der Mengenlehre und ihrer Anwendungen*, Jahresber. der D.M.V., vol. 23 (1914).

SYMMETRIES, CONSERVATION LAWS AND SYMPLECTIC STRUCTURES; ELEMENTARY SYSTEMS

F. Takens

1. INTRODUCTION

The foundations of rational mechanics can be based on various "basic assumptions", e.g. the variational structure leading to the Euler-Lagrange equations or the canonical structure leading to the Hamilton equations, but whatever choice one makes, in one way or the other the "phase space", or rather "orbit space" [6] gets in a natural way a symplectic structure which plays a dominant role in the whole theory. It is a well known consequence of the existence of this symplectic structure that, roughly speaking, for each symmetry there is a conservation law; see for example [2]. In [7] I considered the problem whether the existence of a certain amount of symmetry together with the corresponding conservation laws imply the existence of an underlying variational principle. The results, obtained in that paper, concern field equations; the present paper deals with the corresponding problem for mass points or, as Souriau calls them *elementary systems* [6]. The fact that this time we are interested in symplectic structures rather than variational principles has a technical reason: for "particles" with spin there is no obvious variational principle but there is still a good symplectic structure of its orbit space.

Although our formal definitions are somewhat different from those of Souriau's we also define an elementary system as a mechanical system which is "minimal" in some sense. The symmetry groups which are relevant, are the Huygens group (usually called the Galilei group, but see [3]), and the Poincaré group depending on whether we consider Newtonian or relativistic mechanics. The notion of "conservation law corresponding to a given symmetry" leads to two possible definitions namely "momentum" (2.5) or "momentum with agreeing force form" (2.4), (2.6). The main result of this paper states that an elementary system with symmetry group G and corresponding conservation laws admits an associated symplectic structure (on its

orbit space see (2.7)); if G is the Poincaré group but does not necessarily admit such a structure if G is the Huygens group.

We conclude this introduction with some miscellaneous comments: Although the main problem is this paper is closely related with [7], the methods are completely different and much more related with [6] Chapitre II, III; the treatment of elementary systems in this last source motivated much of the present work; the reader is assumed to be familiar with that treatment.

The classification of elementary systems, as given by SOURIAU [6] becomes incomplete with our definitions; this has various reasons most of which are trivial but one of which is interesting. This will be illustrated in example (6.3).

2. FORMALIZATION OF THE PROBLEM

We shall first define what we consider as a (particle) system. The definition is formal but we indicate the physical meaning in brackets.

DEFINITION 2.1. A *particle system* is a structure consisting of:
an affine space V (V plays the role of space-time and will be 4-dimensional in all examples);
a Lie group G of affine transformations of V containing the vector space $\sigma(V)$ of all affine translations of V (it is this group which, in the applications will be the Huygens or the Poincaré group);
a subset T of $T(V)$, the tangent bundle of V, such that for $t \in T$, $\lambda \in \mathbb{R}$, $g \in G$, $dg(t) \in T$ and $\lambda \cdot t \in T$ if and only if $\lambda > 0$ (the vectors $t \in T$ correspond to motion in *positive time direction*); a translation $\Sigma \in \sigma(V)$ is said to be a translation in positive time direction if the tangent vector of $s \to (s \cdot \Sigma)(v)$ is in T; $T \cap T_V(V)$ is denoted by $T(v)$;
a differentiable bundle $\pi: E \to V$ over V with a G-action on E (notation: $g \in G$ and $e \in E$ then $(g,e) \to g_E(e)$) such that for all $e \in E$, $g \in G$, $g(\pi(e)) = \pi(g_E(e))$ (each point e of E is supposed to represent a possible state of our particle when its position / time is $\pi(e)$);
an *evolution field* E which assigns to each $e \in E$ an open halfline $E(e)$ in $T_e(E)$ such that $d\pi(E(e)) \subset T(\pi(e))$ and $dg_E(E(e)) = E(g_E(e))$ (E should be interpreted as the direction of evolution with time, which occures without external forces);
a *restriction field* R which assigns to each $e \in E$ a linear subspace

$R(e) \in T_e(E)$, transverse to $E(e)$, such that for each $g \in G$, $e \in E$, $X_1 \in E(e)$ and $X_2 \in R(e)$, $dg_E(R(e)) = R(g_E(e))$ and $d\pi(X_1+X_2) \in T(\pi(e))$ (these vectors $X_1 + X_2$, with $X_1 \in E(e)$ and $X_2 \in R(e)$, represent possible evolution directions in the situation where external forces are present).

DEFINITION 2.2. A curve $\gamma: \mathbb{R} \to E$ is called a *free orbit* if for each t, $\dot{\gamma}(t) \in E(\gamma(t))$; it is called a *possible orbit* if each $\dot{\gamma}(t)$ is the sum of a vector in $E(\gamma(t))$ and $R(\gamma(t))$.

DEFINITION 2.3. A particle system (V,G,T,E,π,E,R) is an *elementary system* if G acts transitively on E and if there is no proper submanifold $E' \subset E$ such that for each $e' \in E'$, $E(e')$ and $R(e')$ are contained in $T_{e'}(E')$.

DEFINITION 2.4. A *force form* F for an elementary system (V,G,T,E,π,E,R) is a map which assigns to each $e \in E$ an injection $F(e): R(e) \to T_e^*(E)$ such that for $g \in G$, $e \in E$, $X_1 \in R(e)$ and $X_2 \in T_e(E)$, $(F(e)X_1)X_2 = (F(g_E(e))dg_E(X_1))dg_E(X_2)$ (if $\gamma: \mathbb{R} \to E$ is a possible orbit, and $\dot{\gamma}(t) = X_1 + X_2$, $X_1 \in R(\gamma(t))$ and $X_2 \in E(\gamma(t))$, then $F(\gamma(t))\ X_1$ "is the force needed at time t to get this orbit").

DEFINITION 2.5. A *momentum* M for an elementary system (V,G,T,E,π,E,R) is smooth map $M: E \to J^*$ (J^* is the dual of the Lie algebra J of G) such that

(i) M is constant on each free orbit;

(ii) the rank of dM is everywhere equal to dim(E)-1;

(iii) there is a map $\theta: G \to J^*$ such that for each $e \in E$ and $g \in G$, $M(g_E(e)) = g_{J^*}(M(e)) + \theta(g)$, where g_{J^*} refers to the canonical linear representation of G on J^*, determined by $(g_{J^*}\alpha)(X) = \alpha(g_J^{-1}X) = \alpha((\mathrm{Ad}\ g^{-1})X)$ ($g \in G$, $\alpha \in J^*$, $X \in J$ and Ad the adjoint representation [1]).

COMMENT 2.5a. The notion of momentum is introduced as a concept describing a system of conservation laws corresponding to a group (namely G) of symmetries. Namely for each $X \in J$, i.e., for each infinitesimal symmetry X, $M_X: E \to \mathbb{R}$, defined by $M_X(e) = (M(e))X$ is a function which is constant on free orbits and hence the value of M_X may be considered as a conserved quantity.

The fact that the rank of dM equals dim(E)-1 means that (locally) the free orbits can be distinguished by the values of the conserved quantities along them.

The third condition would, for $\theta \equiv 0$, mean that M is equivariant with respect to the given G-actions on E and J^*. The reason for adding the term $\theta(g)$ is the following: in general, conserved quantities are only defined up to some additive constant. Hence we should expect that not M, but "M modulo an additive constant" has prime significance. Requiring "M modulo an additive constant" to be compatible with the G-actions is for our purpose equivalent with requiring dM to be compatible with these actions, which leads to (iii).

DEFINITION 2.6. Let (V,G,T,E,π,E,R) be an elementary system with force form F and momentum M. F and M are said to *agree* if for each $e \in E$, $X \in R(e)$ and $Z \in J$, $(F(e)X)Z_E(e) = ((dM)_e X)Z$, where $Z_E(e)$ is the tangent vector of the curve $s \to (\mathrm{Exp}(s.Z))_E e$, and where we identify $T_{M(e)}(J^*)$ with J^* so that $(dM)_e$ becomes a map from $T_e(E)$ to J^*. A force form F, resp. a momentum M, is said to be *agreeble* if there is a momentum $\tilde{M}$, resp. a force form $\tilde{F}$, such that F and $\tilde{M}$, resp. $\tilde{F}$ and M agree; in this case $\tilde{M}$ is unique up to a constant, resp. $\tilde{F}$ is unique.

DEFINITION 2.7. Let (V,G,T,E,π,E,R) be an elementary system; a *symplectic structure* for this system consists of a closed 2-form Ω on E such that

(i) if $X \in E(e)$, $Y \in T_e(E)$, then $\Omega(X,Y) = 0$;

(ii) $\dim(E) = 2n+1$, for some integer n, and $\Omega \wedge \ldots \wedge \Omega$ (n-times) is nowhere zero;

(iii) for each $g \in G$, $g_E^* \Omega = \Omega$.

Note that such Ω does not define a symplectic structure (in the usual sense) on the manifold E, but only on the (maybe only locally defined) "manifold of free orbits".

DEFINITION 2.8. Let Ω define a symplectic structure for an elementary system (V,G,T,E,π,E,R). The *force form* F_Ω, *induced by* Ω, is defined by $(F_\Omega(e)X)Y = \Omega(X,Y)$ for all $X \in R(e)$, $Y \in T_e(E)$. A *momentum* M is said to be *related with a symplectic structure* Ω if M agrees with the force form F_Ω.

REMARK 2.9. If E is connected and $H^1(E;\mathbb{R}) = 0$ then there is for each symplectic structure Ω a related momentum M which is unique up to an additive constant; this momentum M is of course agreeable.

3. STATEMENT OF THE RESULTS

H, resp. P, denotes the Huggens, resp. Poincaré group, defined in §5. They are both groups of affine transformations of $\mathbb{R}^4$. We define T_H, resp. T_P, as follows:

$$T_H = \{\Sigma\ \alpha_i \frac{\partial}{\partial x_i} \Big|_x \in T_x(\mathbb{R}^4) \mid \alpha_4 > 0\}$$

$$T_P = \{\Sigma\ \alpha_i \frac{\partial}{\partial x_i} \Big|_x \in T_x(\mathbb{R}^4) \mid \alpha_4 > 0 \text{ and } \alpha_4^2 \geq \alpha_1^2 + \alpha_2^2 + \alpha_3^2\}.$$

Note that, up to the arbitrary choice $\alpha_4 > 0$, T_H and T_P are the maximal subsets of $T(\mathbb{R}^4)$ satisfying the conditions in definition (2.1) G = H resp. G = P.

THEOREM 3.1. *Let* $(\mathbb{R}^4, P, T_P, E, \pi, E, R)$ *be an elementary system. If* M: E $\to P^*$ *is a momentum then* M *is agreeable and* M *is related with a symplectic structure.*

THEOREM 3.2. *Let* $(\mathbb{R}^4, H, T_H, E, \pi, E, R)$ *be an elementary system. If* M: E $\to H^*$ *is a momentum then is does not follow that* M *is agreeable; if* M *is agreeable, it does not follow that* M *is related with some symplectic structure.*

4. THE PROOF OF THEOREM 3.1.

In this § we let (V, G, T, E, π, E, R) be an elementary system and M: E $\to J^*$ some momentum. We investigate the obstructions to M being agreeable and to M admitting a related symplectic structure.

OBSERVATION 4.1. Let M be a momentum and $\alpha \in J^*$ a constant. M is agreeable and/or related with a symplectic structure if and only if M+α is; $(M+\alpha)(e) := M(e) + \alpha$. For this reason we say that M and $\tilde{M}$ are *equivalent* if and only if $M - \tilde{M}$ is constant.

LEMMA 4.2. *Let* M *be a momentum with* θ: G $\to J^*$ *as in* (2.5). *Then* θ *defines an affine representation* g $\to g_{J^*,\theta}$ *of* G *on* J^*:

$$g_{J^*,\theta}(\alpha) = g_{J^*}(\alpha) + \theta(g).$$

PROOF. From the above definition it follows that $g_{J^*,\theta}$ is an invertible affine map which is linearly related with g_{J^*} (i.e. $g_{J^*,\theta}$ is g_{J^*} plus a translation) and that $(id)_{J^*}$ = identity on J^*. So we only have to show that $g_{1_{J^*,\theta}} \circ g_{2_{J^*,\theta}} = (g_1 . g_2)_{J^*,\theta}$. From (2.5) we have $(g_1 . g_2)_{J^*}(M(e) + \theta(g_1 . g_2)) = M((g_1 . g_2)_E e) = M((g_{1_E}(g_{2_E} e))) = g_{1_{J^*}}(M(g_{2_E}(e))) + \theta(g_1) = g_{1_{J^*}} g_{2_{J^*}}(M(e)) + g_{1_{J^*}}(\theta(g_2)) + \theta(g_1)$, or $\theta(g_1 . g_2) = g_{1_{J^*}} \theta(g_2) + \theta(g_1)$. From this last formula the lemma follows directly.

OBSERVATION 4.3. If M and $\tilde{M}$ are equivalent momenta and $\theta, \tilde{\theta}: G \to J^*$ the corresponding mappings then there is some $\alpha \in J^*$ such that for each $g \in G$, $\theta(g) - \tilde{\theta}(g) = \alpha - g_{J^*}(\alpha)$.

DEFINITION 4.4. We define $Z^1(G;J^*)$ to be the vectorspace of those smooth maps $\theta: G \to J^*$ such that for $g_1, g_2 \in G$, $\theta(g_1 . g_2) = g_{1_{J^*}} \theta(g_2) + \theta(g_1)$. $B^1(G;J^*) \subset Z^1(G;J^*)$ is the vectorsubspace of those θ for which there is a fixed $\alpha \in J^*$ such that for all $g \in G, \theta(g) = \alpha - g_{J^*}(\alpha)$. $H^1(G;J) = Z^1(G;J^*) / B^1(G;J^*)$.

Note that a momentum M determines a unique $\theta \in Z^1(G;J^*)$; if $\tilde{\theta} \in Z^1(G;J^*)$ with $\theta - \tilde{\theta} \in B^1(G;J^*)$ then there is a unique momentum $\tilde{M}$, equivalent with M, such that $\tilde{\theta}$ corresponds with $\tilde{M}$. An equivalence class of momenta determines an equivalence class $[\theta] \in H^1(G;J^*)$.

DEFINITION 4.5. We define $Z^1(J;J^*)$ to be the vectorspace of bilinear maps $f: J \times J \to \mathbb{R}$ such that, for all $X_1, X_2, X_3 \in J$,

$$f([X_1,X_2],X_3) = f(X_1,[X_2,X_3]) - f(X_2,[X_1,X_3]);$$

$B^1(J;J^*) \subset Z^1(J;J^*)$ is the subspace of those bilinear maps such that for some $\alpha \in J^*$ and all $X_1, X_2 \in J$, $f(X_1,X_2) = \alpha([X_1,X_2])$; $H^1(J;J^*) = Z^1(J;J^*)/B^1(J;J^*)$.

OBSERVATION 4.6. It should be noted that the bilinear maps $f: J \times J \to \mathbb{R}$ are in 1-1 correspondence with linear maps $\tilde{f}: J \to J^*$ $(f(X_1,X_2) = (\tilde{f}(X_1))X_2)$. For each bilinear $f: J \times J \to \mathbb{R}$ there is a corresponding affine vectorfield X_f on J^* defined by: $(X_f(\alpha))Y = -\alpha([X,Y]) + f(X,Y)$ (using the identification $T_\alpha(J^*) \tilde{=} J^*$); $\alpha \in J^*, X, Y \in J$. The linear part of X_f ($-\alpha([X,Y])$ in the above formula) corresponds to the infinitesimal representation of $g \to g_{J^*}$ in the sense that $((\mathrm{Exp}\ t.X)_{J^*}\alpha - \alpha)Y = \alpha((\mathrm{Exp}\ t.-X)_J Y - Y) = -t\alpha([X,Y]) + O(t^2)$.

From a simple computation one finds that for bilinear maps $f: J \times J \to \mathbb{R}$, $f \in Z^1(J;J^*)$ if and only if $X \to X_f$ is an infinitesimal affine representation of G on J^*. In case G is connected and $H^1(G;\mathbb{R}) = 0$, there is a 1-1 correspondence between affine representations of G on J^*, which are linearly equivalent with $g \to g_{J^*}$, and infinitesimal affine representations of G on J^*, which are linearly equivalent with the infinitesimal co-adjoint representation.

Since both H and P (see §5) have the homotopy type of SO(3) they are both connected and both have $H^1(\ ;\mathbb{R})$ equal to zero, we may and do *assume* from now on *that* G *is connected and that* $H^1(G;\mathbb{R}) = 0$.

From the above remarks we have that $Z^1(G;J^*) \cong Z^1(J,J^*)$. From a simple calculation one sees that this isomorphism induces an isomorphism $B^1(G;J^*) \cong (J;J^*)$ and hence $H^1(G;J^*) \cong H^1(J;J^*)$.

PROPOSITION 4.7. *Let* M *be a momentum,* $\theta \in Z^1(G;J^*)$ *the corresponding map from* $G \to J^*$ *and* $f \in Z^1(J;J^*)$ *the related bilinear map on* J,i.e., $f(X,Y) = ((d\theta)_{id}X)Y$ *see* (4.6). *We take some* $e_0 \in E$ *and define* $J_{e_0} = \{X \in J \mid X_E(e_0) = 0\}$, $J_{e_0,R+E} = \{X \in J \mid X_E(e_0)$ *is contained in the linear of* $T_{e_0}(e)$ *spanned by* $R(e_0)$ *and* $E(e_0)\}$.

Then M *is agreeable if and only if for each pair* $X_1 \in J_{e_0}, X_2 \in J_{e_0,R+E}$, $f(X_1,X_2) = -f(X_2,X_1)$.

PROOF. Let $X \in E(e_0)$ and $Y \in R(e_0)$. Then M is agreeable (at least in e_0, but since G is transitive on E this implies agreeable everywhere) if and only if for each such X,Y and $Z \in J_{e_0}$, $((dM)_{e_0}(X+Y))Z = 0$.

Since G is transitive, there is a $Z' \in J_{e_0,R+E}$ such that $Z'_E(e_0) = X+Y$. Hence we have to show that for all $Z^1 \in J_{e_0,R+E}$, $Z \in J_{e_0}$, $((dM)_{e_0}(Z'_E(e_0)))Z = 0$, if and only if $f(Z',Z) = -f(Z,Z')$.

$$((dM)_{e_0}(Z'_E(e_0))Z = \frac{d}{dt}(M((\mathrm{Exp}\ t\ Z')_E e_0))Z =$$

$$\frac{d}{dt}(((\mathrm{Exp}\ t\ Z')_{J^*}M(e_0))Z) + \frac{d}{dt}(\theta(\mathrm{Exp}\ t\ Z')Z) =$$

$$-M(e_0)([Z',Z]) + f(Z',Z).$$

In the same way one obtains that $((dM)_{e_0}(Z_E(e_0)))Z' = -M(e_0)([Z,Z']) + f(Z,Z')$. Since $Z_E(e_0) = 0$, this last expression is zero, so $M([Z,Z']) = f(Z,Z')$ and hence $((dM)_{e_0}(Z'_E(e_0)))Z = f(Z,Z') + f(Z',Z)$.

This proves the position.

CONSTRUCTION 4.8. Let F be an agreeable force form; we shall show how to reconstruct the momentum. From (2.6) we know that F determines the derivative of the momentum restricted to R; in the direction of E the derivative of the momentum has to be zero. From (2.3) it now follows that two momenta can only differ by a constant: if M_1, M_2 do not, differ by a constant, let $\alpha \in J^*$ be the image of $M_1 - M_2$; $(M_1 - M_2)^{-1}(\alpha)$ is a manifold (because due to equivariance $\mathrm{rk}(d(M_1 - M_2))$ is constant) which is not allowed to exist by (2.3). Hence F determines dM.

If F is moreover induced by some symplectic structure Ω, then by (2.8) and the above observation Ω is uniquely determined by F: $\Omega(X, Z_E(e)) = ((dM_e(X))Z$.

PROPOSITION 4.9. *Let* M *be an agreeable momentum* with corresponding $\theta: G \to J^*$ *and* $f: J \times J \to \mathbb{R}$. *Then the agreeing force form* F *is induced by a symplectic structure if and only if* f *is anti-symmetric.*

PROOF. If Ω is a symplectic structure inducing the force form F of M, $\Omega(Z_E(e), Z'_E(e)) = -\Omega(Z'_E(e), Z_E(e))$. Using the arguments in the proof of (4.7) we see that the following expression has to be anti-symmetric in Z, Z' : $\Omega(Z_E(e), Z'_E(e)) = ((dM)_e Z_E(e))Z' = -M(e)([Z, Z']) + f(Z, Z')$; hence f has to be anti-symmetric.

If f is anti-symmetric then $((dM)_e(Z_E(e)))Z'$ is anti-symmetric in Z, Z'; since this expression is zero for $Z_E(e) = 0$, it is also zero if $Z'_E(e) = 0$. Hence there is a 2-form Ω on E such that for all $e \in E, Z, Z' \in J$, $\Omega(Z_E(e), Z'_E(e)) = ((dM)_e Z_E(e))Z'$. If there is a symplectic structure inducing F, it must be Ω; Ω defines indeed a symplectic structure:

- Ω is G-invariant because dM is G-equivariant;
- for $X \in E(e)$, $\Omega(X, -) = 0$ because M is constant along free orbits;
- for X transversal to $E(e)$, $\Omega(X, -) \neq 0$ because otherwise the rank of dM would be $\leq \dim(E) - 2$;
- $d\Omega = 0$; this results from the following computation: because of the invariance of Ω, $L_{Z_E}\Omega = 0$ for all $Z \in J$ (L stands for the Lie derivative) hence $\imath_{Z_E} d\Omega + d\imath_{Z_E}\Omega = 0$, but $\imath_{Z_E}\Omega = -((dM)(-))Z = -d(M(Z))$ and hence $d\imath_{Z_E}\Omega = 0$, so $\imath_{Z_E} d\Omega = 0$ for all $Z \in J$; from this and the transitivity of the G-action on E, it follows that $d\Omega$ 0.

REMARK (4.10). Theorem (3.1) now follows, using (4.7) and (4.9) from the fact that $H^1(P;P^*) = 0$ which we shall prove in §5.

5. THE HUYGENS AND THE POINCARÉ GROUP.

(a) The Huygens group.

The Huygens group H consists of affine transformations of $\mathbb{R}^4$ which, as 5x5 matrices, have the form

$$g = \begin{pmatrix} \tilde{A} & \beta & \gamma \\ 0 & 1 & \delta \\ 0 & 0 & 1 \end{pmatrix}$$

with $\tilde{A} \in SO(3); \beta,\gamma \in \mathbb{R}^3$ and $\delta \in \mathbb{R}$. The action of g on $\mathbb{R}^4$ is determined by:

$$g\begin{pmatrix} x_1 \\ \vdots \\ x_4 \end{pmatrix} = \begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_4 \end{pmatrix} \text{ if } \begin{pmatrix} \tilde{A} & \beta & \gamma \\ 0 & 1 & \delta \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} x_1 \\ \vdots \\ x_4 \\ 1 \end{pmatrix} = \begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_4 \\ 1 \end{pmatrix}$$

The Lie algebra $\mathcal{H}$ of H consists of those 5x5 matrices

$$X = \begin{pmatrix} A & b & c \\ 0 & 0 & d \\ 0 & 0 & 0 \end{pmatrix}$$

with A skew symmetric; $b,c \in \mathbb{R}^3$ and $d \in \mathbb{R}$.

The elements of the dual Lie algebra $\mathcal{H}^*$ can be represented by 5x5 matrices of the form

$$\sum = \begin{pmatrix} A^* & 0 & 0 \\ b^* & 0 & 0 \\ c^* & d^* & 0 \end{pmatrix}$$

with A skew symmetric; $b^*,c^* \in (\mathbb{R}^3)^*$ and $d^* \in \mathbb{R}$.

$$\sum(X) = (\text{trace } A^*A) + b^*b + c^*c + d^*d.$$

THEOREM 5.1. $H^1(H;H^*) \tilde{=} H^1(H;H^*) \tilde{=} \mathbb{R}^3$. *Representatives of* $f \in Z^1(H;H^*)$ *can be given as*

$$f\begin{pmatrix} j(a) & b & c \\ 0 & 0 & d \\ 0 & 0 & 0 \end{pmatrix}\begin{pmatrix} j(a') & b' & c' \\ 0 & 0 & d' \\ 0 & 0 & 0 \end{pmatrix} =$$

$$\lambda_0 \cdot \langle b,a' \rangle + \lambda_1 \{\langle b,c' \rangle - \langle b',c \rangle\} + \lambda_2 d.d', \quad \textit{where}$$

$$j(a) = j\begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} = \begin{pmatrix} 0 & -a_3 & a_2 \\ a_3 & 0 & -a_1 \\ -a_2 & a_1 & 0 \end{pmatrix}, \textit{ or } (j(a))b = a \times b.$$

The corresponding $\theta_f : H \to H^*$ *is given by*

$$\theta_f\begin{pmatrix} A & \beta & \gamma \\ 0 & 1 & \delta \\ 0 & 0 & 1 \end{pmatrix} =$$

$$\begin{pmatrix} \lambda_0 \cdot j(\beta) + \tfrac{1}{2}\lambda_1(\gamma\beta^T - \beta\gamma^T) & 0 & 0 \\ \lambda_1 \cdot (-\gamma^T + \delta \cdot \beta^T) & 0 & 0 \\ \lambda_1 \cdot \beta^T & & \\ & \tfrac{1}{2}\lambda_1 \beta^T\beta + \lambda_2\delta & 0 \end{pmatrix}$$

For the sake of completeness we add:

$$\begin{pmatrix} A & \beta & \gamma \\ 0 & 1 & \delta \\ 0 & 0 & 1 \end{pmatrix}_{H^*}\begin{pmatrix} A^* & 0 & 0 \\ b^* & 0 & 0 \\ c^* & d^* & 0 \end{pmatrix} :=$$

$$\begin{pmatrix} \tilde{A}.A^*.\tilde{A}^T + \tfrac{1}{2}[\beta\, b^*.\tilde{A}^T - \tilde{A}.b^*.{}^T\beta^T + \gamma.c^*.\tilde{A}^T - \tilde{A}.c^{*T}.\gamma^T] & 0 & 0 \\ \delta.c^*.\tilde{A}^T + b^*\tilde{A}^T & 0 & 0 \\ c^*.\tilde{A}^T & d^* - c^*.\tilde{A}^T.\beta & 0 \end{pmatrix}$$

PROOF. The fact that $\theta_f \in Z^1(H;H^*)$, $f \in Z^1(H;H^*)$ and that for different $\lambda_0,\lambda_1,\lambda_2$ the corresponding f's represent different elements in $H^1(H;H^*)$ follows from tedious but straightforward calculations.

The proof that dim $H^1(H;H^*) \leq 3$ is harder; we shall not use and also not prove this last fact in this paper.

(b) The Poincaré group.

The Poincaré group P consists of affine transformations of $\mathbb{R}^4$ which, as 5x5 matrices, have the form

$$g = \begin{pmatrix} A & M \\ 0 & 1 \end{pmatrix}$$

with $M \in \mathbb{R}^4$ and A a 4x4 matrix such that $AA^T = ad$, however, in this section on the Poincaré group, A^T is defined by

$$\begin{pmatrix} a_{11} & \cdots & a_{14} \\ \vdots & & \vdots \\ a_{41} & \cdots & a_{44} \end{pmatrix}^T = \begin{pmatrix} a_{11} & a_{21} & a_{31} & -a_{41} \\ a_{12} & a_{22} & a_{32} & -a_{42} \\ a_{13} & a_{23} & a_{33} & -a_{43} \\ -a_{14} & -a_{24} & -a_{34} & a_{44} \end{pmatrix},$$

and such that $a_{44} > 0$, $\det(A) > 0$. The corresponding action on $\mathbb{R}^4$ is determined by

$$g\begin{pmatrix} x_1 \\ \vdots \\ x_4 \end{pmatrix} = \begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_4 \end{pmatrix} \quad \text{if} \quad \begin{pmatrix} A & M \\ 0 & 1 \end{pmatrix}\begin{pmatrix} x_1 \\ \vdots \\ x_4 \\ 1 \end{pmatrix} = \begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_4 \\ 1 \end{pmatrix}.$$

The Lie algebra $\mathcal{P}$ of P consists of 5x5 matrices

$$X = \begin{pmatrix} B & N \\ 0 & 0 \end{pmatrix},$$

with $B + B^T = 0$ and $N \in \mathbb{R}^4$. We shall prove:

THEOREM 5.2. $H^1(P;P^*) = 0$ *and hence also* $H^1(P;P^*) = 0$.

PROOF. We introduce two Lie subalgebras of P: L consists of those elements of P which have the form:

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

with $B + B^T = 0$; σ consists of the infinitesimal translations, in matrix notation:

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

with $N \in \mathbb{R}^4$.

Let $f: P \times P \to \mathbb{R}$ be an element of $Z^1(P;P^*)$, i.e., f satisfies for all $X_1, X_2, X_3, \in P$:

$$(1) \ldots . f([X_1,X_2],X_3) = f(X_1,[X_2,X_3]) - f(X_2,[X_1,X_3]).$$

We have to show that $f \in B^1(P;P^*)$. First we observe that $f \mid L \times L : L \times L \to \mathbb{R}$ is in $Z^1(L;L^*)$. Due to Whitehead's lemma [5] there is a $\mu \in L^*$ such that, for $X_1, X_2 \in L$, $f(X_1,X_2) = \mu([X_1,X_2])$. We extend μ to P (as linear map to $\mathbb{R}$) and replace $f(\ ,\ )$ by $f(\ ,\ ) - \mu([\ ,\ ])$; this makes no difference modulo $B^1(P;P^*)$ and enables us to assume that for $X_1, X_2 \in L$, $f(X_1,X_2) = 0$.

We apply (1) tot the case where $X_1 \in \sigma$ and $X_2 = X_3 \in L$ and find $f([X_1,X_2],X_2) = -f(X_2,[X_1,X_2])$. Since for generic $X_2 \in L$, $[\sigma,X_2] = \sigma$, we conclude that $f(X_1,X_2) = -f(X_2,X_1)$ whenever $X_1 \in \sigma$ and $X_2 \in L$.

Next we apply (1) to the case where $X_1, X_2 \in L$ and $X_3 \in \sigma$:

$$f([X_1,X_2],X_3) + f([X_2,X_3],X_1) + f([X_3,X_1],X_2) = 0.$$

From this we want to conclude that if $X_1' \in L$, $X_3' \in \sigma$ and $[X_1',X_3'] = 0$, we have $f(X_1',X_3') = 0$. We represent X_1', X_3' as

$$\begin{pmatrix} B_1 & 0 \\ 0 & 0 \end{pmatrix}, \text{ resp. } \begin{pmatrix} 0 & N_3 \\ 0 & 0 \end{pmatrix};$$

$[X_1',X_3'] = 0$ is then equivalent with $B_1N_3 = 0$ or $N_3 \in \mathrm{Ker}\,(B_1)$. We assume that $N_3 \neq 0$ (otherwise $X_3' = 0$ and hence $f(X_1',X_3') = 0$) and conclude that $\mathrm{Ker}\,(B_1)$ has at least dimension 2, see [4], Chapter XV, §6,15.26. Hence, arbitrarily close to X_3' there is some $X_3'' \in \sigma$, $X_3'' = \begin{pmatrix} 0 & N_3' \\ 0 & 0 \end{pmatrix}$, with $[X_1',X_3'] = 0$, or $B_1 N_3' = 0$, and $N_3' = \begin{pmatrix} y_1 \\ \vdots \\ y_4 \end{pmatrix}$ with $y_1^2 + y_2^2 + y_3^2 - y_4^2 \neq 0$.

It clearly suffices to show that $f(X_1',X_3'') = 0$. For this we define $L^{X_3''} = \{X \in L \mid [X,X_3''] = 0\}$; because $y_1^2 + y_2^2 + y_3^2 - y_4^2 \neq 0$, $L^{X_3''}$ is isomorphic with the Lie algebra of $SO(3)$. $X_1' \in L^{X_3''}$ and hence there are $X_1'',X_2'' \in L^{X_3''}$ such that $[X_1'',X_2''] = X_1'$. Now the above formula implies that $f([X_1'',X_2''],X_3'') = 0$. From this we conclude that for some linear $\mu:\sigma \to \mathbb{R}$, $f(X_1,X_3) = \mu([X_1,X_3])$ whenever $X_1 \in L$ and $X_3 \in \sigma$. We extend μ to a linear function on P so that $\mu \mid L = 0$ and replace (again) $f(\ ,\ )$ by $f(\ ,\ ) - \mu([\ ,\ ])$. This means that we may assume that $f(X_1,X_2) = 0$ whenever $X_1,X_2 \in L$, or $X_1 \in L$ and $X_2 \in \sigma$, or $X_1 \in \sigma$ and $X_2 \in L$.

Finally we apply (1) with $X_1,X_2 \in \sigma$ and $X_3 \in L$: $f(X_1,[X_2,X_3]) = f(X_2,[X_1,X_3])$ and (interchanging X_2 and X_3 in (1)):

$$f([X_1,X_3],X_2) = f(X_1,[X_3,X_2]).$$

From this we conclude that $f \mid \sigma \times \sigma$ is invariant under the adjoint action of $L = \mathrm{Exp}(L)$ on σ and that $f \mid \sigma \times \sigma$ is anti-symmetric. From this we conclude that $f \mid \sigma \times \sigma = 0$ and hence that $f = 0$.

6. EXAMPLES AND THE PROOF OF THEOREM (3.2)

We give some examples of elementary systems with momenta. In all these examples $V = \mathbb{R}^4$, $G = H$, the Huygens group, and $T = T_H$ see §3. To describe such an example we give:

(a) the affine respresentation Θ of H on H^*; this is done by specifying the $\lambda_1,\lambda_2,\lambda_3$ in theorem (5.1);

(b) the image of some $e_0 \in \pi^{-1}(0)$ under M in H^*; these first two data determine the Lie algebra $H_{e_0,E} = \{X \in H \mid X_E(e_0)$ is in the linear subspace spanned by $E(e_0)\}$ = Lie algebra of $\{h \in H \mid (\Theta(h))\, M(e_0) = M(e_0)\}$

(c) the linear subspace $H_{e_0,R+E} \supset H_{e_0,E}$, see proposition (4.7).

If the representation Θ, the element $M(e_0) \in H^*$ and the subspace $H_{e_0,R+E}$ are given then there is a (unique) elementary particle model if:

1. $\dim(H_{e_0,E}) = \dim(H_{e_0}) + 1$, $H_{e_0,E}$ as above (determined by Θ and $M(e_0)$) and $H_{e_0} = \{X \in H_{e_0,E} \mid X(0) = 0\}$, the elements of H are here identified with affine vectorfield on $\mathbb{R}^4$;
2. for some $Y \in H_{e_0,E} \setminus H_{e_0}$, $Y(0)$ is in positive time direction;
3. $H_{e_0,R+E}$ is not contained in a proper subalgebra of H.

If these conditions are satisfied, the momentum M is agreeable if and only if for each $X_1 \in H_{e_0}$, $X_2 \in H_{e_0,R+E}$, $f(X_1,X_2) = -f(X_2,X_1)$, see (4.7). The momentum is related with a symplectic structure if and only if $f(X_1,X_2) = -f(X_2,X_1)$ for all $X_1,X_2 \in H$, see (4.9); f is here related with the representation Θ as in §4. The verification, along the above lines, that the examples below have the announced properties, is left to the reader.

EXAMPLE 6.1. Representation: $\lambda_0 \neq 0$, $\lambda_1 \neq 0$, $\lambda_2 = 0$; $M(e_0) = 0$;

$$H_{e_0,R+E} = \{ \begin{pmatrix} A & b & c \\ 0 & 0 & d \\ 0 & 0 & 0 \end{pmatrix} \Big| \; c = 0\}.$$

In this case the momentum is not agreeable; if we replace λ_0 by zero we get the classical mass point with the usual momentum.

EAMPLE 6.2. Representation: $\lambda_0 \neq 0$, $\lambda_1 \neq 0$, $\lambda_2 = 0$;

$M(e_0) = \begin{pmatrix} j\begin{pmatrix}1\\0\\0\end{pmatrix} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$, for the meaning of $j\begin{pmatrix}1\\0\\0\end{pmatrix}$ see theorem (5.1);

$$H_{e_0,R+E} = \{ \begin{pmatrix} A & b & c \\ 0 & 0 & d \\ 0 & 0 & 0 \end{pmatrix} \Big| \; b = \begin{pmatrix} 0 \\ b_2 \\ b_3 \end{pmatrix}, \; d = 0\}.$$

In this case, the momentum is agreeable but not related with any symplectic structure. Theorem (3.2) follows from the examples (6.1) and (6.2).

EXAMPLE 6.3. Representation: $\lambda_0 = 0$, $\lambda_1 \neq 0$, $\lambda_2 = 0$;

$$M(e_0) = \begin{pmatrix} j\begin{pmatrix}1\\0\\0\end{pmatrix} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

$$H_{e_0, R+E} = \left\{ \begin{pmatrix} A & b & c \\ 0 & 0 & d \\ 0 & 0 & 0 \end{pmatrix} \;\middle|\; c = 0 \right\}.$$

Here the momentum is related with a symplectic structure. However, we have another stange phenomenon. If $\pi: E \to \mathbb{R}^4$ is the fibration in the above example, one can get a new example by modifying π: take e.g. $\tilde{\pi}(e_0) = \begin{pmatrix}1\\0\\0\\0\end{pmatrix}$ and extend $\tilde{\pi}$ equivariantly (which is possible in a unique way). If one adjusts the definition of R we have again an elementary system with momentum and related symplectic structure. In Souriau's classification [6] these two elementary systems (with π and $\tilde{\pi}$) are considered equal.

REFERENCES

1. ADAMS, J.F., *Lectures on Lie groups*, Benjamin New York, 1969.
2. CARTAN, E., *Lecons sur les invariants intégraux*, Hermann, Paris, 1922.
3. DIJKSTERHUIS,E.J., *De mechanisering van het wereldbeeld*, Meulenhoff, Amsterdam, 1950.
4. GREUB, W.H., *Linear Algebra*, Springer, Berlin, 1963.
5. JACOBSON, N., *Lie algebras*, Wiley, New York, 1962.
6. SOURIAU, J.M., *Structure des systèmes dynamiques*, Dunod, Paris, 1970.
7. TAKENS, F., *Symmetries, conservation laws and variational principles*, proceedings of III. ELAM, Lecture notes in mathematics 597, Springer, Berlin.

INTEGRAL REPRESENTATIONS OF INVARIANT REPRODUCING KERNELS

E. Thomas

INTRODUCTION

Let V be a C^∞-manifold*), and let $\mathcal{D}(V)$ and $\mathcal{D}'(V)$ be respectively the spaces of Schwartz test functions and distributions.

Let G be a group of diffeomorphisms $\sigma: V \to V$. G operates in a natural way on $\mathcal{D}'(V)$ and on $\mathcal{D}'(V\times V)$.

We shall consider distributions $K \in \mathcal{D}'(V\times V)$ which satisfy the condition

$$K(\phi\otimes\bar{\phi}) \geq 0 \qquad \forall\phi \in \mathcal{D}(V),$$

the so-called kernels of positive type, and which are invariant under G:

$$K(\phi\circ\sigma \otimes \psi\circ\sigma) = K(\phi\otimes\psi) \qquad \forall\phi,\psi \in \mathcal{D}(V).$$

These kernels form a closed convex cone Γ_G in $\mathcal{D}'(V\times V)$.

The object of this paper is to prove the following theorems, the precise content of which will be made clear below:

A) Every element in Γ_G possesses an integral representation by means of extreme generators of Γ_G (sometimes called zonal kernels). In particular Γ_G is the closed convex hull of its extreme generators.

B) Every $K \in \Gamma_G$ possesses a *unique* integral representation by means of extreme generators if and only if Γ_G is a lattice in its own order.

C) If $G_1 \subset G_2$ and Γ_{G_1} is a lattice in its own order, so is Γ_{G_2}.

Thus for instance the uniqueness of the integral representations of translation invariant kernels of positive type $K \in \mathcal{D}'(\mathbb{R}^n \times \mathbb{R}^n)$, (theorem

*) V will be assumed throughout to possess a countable basis of open sets.

of Bochner-Schwartz) actually implies the uniqueness in the case of any group of diffeomorphisms of $\mathbb{R}^n$ containing the translations on $\mathbb{R}^n$ (e.g. the Euclidean motion group, the inhomogeneous Lorentz group, etc.).

D) If V is a unimodular Lie group (or more generally a locally compact group (second countable), $\mathcal{D}'(V)$ being the space of Bruhat distributions) the cone of bi-invariant distributions of positive type (traces) is a lattice in its own order. Thus every trace possesses a unique integral representation by means of characters.

The existence of such integral representations has been previously proved by K. MAURIN [6].

The question of the uniqueness has not been previously taken up. The precise meaning of the uniqueness, for cones not necessarily possessing a base, will be defined in section 3.

In certain contexts (e.g. quantum mechanics) it is desirable to consider instead of $\mathcal{D}'(V)$ a space of vector valued distributions $\mathcal{D}'(V;F)$ where for instance F is a finite dimensional vector space over $\mathbb{C}$. Positive kernels are then themselves vector valued distributions $K \in \mathcal{D}'(V\times V;F\otimes\bar{F})$, $\bar{F}$ being a space anti-isomorphic to F with a given anti-isomorphism $x \to \bar{x}$ (cf. [8]).

The consideration only of transformations of $\mathcal{D}'(V)$, defined by diffeomorphisms of V, is unnecessarily restrictive. For instance many examples involve transformations defined, by transposition to $\mathcal{D}(V)$, through maps $\phi \to \alpha\phi\circ\sigma$ where σ is a diffeomorphism and $\alpha \in C^\infty(V)$, $\alpha(p) \neq 0$ for all p, $\alpha(p)$ being an invertible linear operator in the vector valued case.

We shall therefore consider more generally groups of automorphisms of $\mathcal{D}'(V)$ not necessarily implemented by diffeomorphisms.

Finally it will be useful in most of what follows, to replace $\mathcal{D}'(V)$ by a general locally convex space.

CONTENTS

In section 1 we recall facts from Schwartz's theory of kernels and associated Hilbert spaces [7].

In section 2 we study cones of invariant kernels.

In section 3 we present the necessary facts from integral representation theory.

Finally, in section 4, we deduce the theorems mentioned above and we show by an elementary counter example the need to use distributions rather than for instance measures.

1. KERNELS AND HILBERT SUBSPACES

Let E be a quasi-complete locally convex space over $\mathbb{C}$, e.g. $\mathcal{D}'(V)$.

Let E^* be the space of continuous linear forms on E, together with a nondegenerate sesquilinear form $\langle x,\phi\rangle$, linear with respect to $x \in E$, antilinear with respect to $\phi \in E^*$. (In the case $E = \mathcal{D}'(V)$ $E^* = \mathcal{D}(V)$ and we put $\langle T,\phi\rangle = T(\bar{\phi})$.)

If F,F^* is a second similar pair and $u: F \to E$ is a continuous linear map we define the adjoint linear operator u^* by

$$\langle ux,\phi\rangle = \langle x,u^*\phi\rangle. \tag{1}$$

Let H be a Hilbert subspace of E, i.e. a linear subspace with a given Hilbert space inner product such that the inclusion

$$j: H \hookrightarrow E$$

is continuous.

We identify H^* with H, the sesquilinear form being replaced by the inner product (x,y). (The adjoint of a linear operator $u: H \to H$ thus becomes the usual adjoint operator.)

This implies that for $\phi \in E^*$, $j^*\phi$ is the element of H satisfying the equation

$$(x,j^*\phi) = \langle jx,\phi\rangle. \tag{2}$$

Note that the image of $j^*: E^* \to H$ in H is a dense subspace, o being the only vector orthogonal to it by (2).

Now let $K = jj^*$. Then K is a linear operator:

$$K: E^* \to E$$

called the *reproducing kernel* of H (see [7] for the background to this definition). Replacing x by $j^*\psi$ in (2) one obtains

$$(j^*\psi,j^*\phi) = \langle K\psi,\phi\rangle. \tag{3}$$

which shows that K is hermitian symmetric:

$$\langle K\psi,\phi\rangle = \overline{\langle K\phi,\psi\rangle}. \tag{4}$$

Putting $\psi = \phi$ in (3) we see that

(5) $\quad \langle K\phi,\phi\rangle \geq 0 \qquad \forall \phi \in E^*,$

which is expressed by saying that K is a kernel of positive type.

Let F be the set of all hermitian kernels (i.e. linear operators $K: E^* \to E$ satisfying relation (4)) and let Γ be the set of all positive hermitian operators (i.e. those satisfying (5)). Then F is a linear space over $\mathbb{R}$, the addition of kernels being defined by $(K_1+K_2)\phi = K_1\phi + K_2\phi$, and Γ is a convex cone in F for which $\Gamma \cap -\Gamma = (0)$ (this follows from the inequality $|\langle K\phi,\psi\rangle|^2 \leq \langle K\phi,\phi\rangle\langle K\psi,\psi\rangle$ valid for $K \in \Gamma$).

We put $K_1 \leq K_2$ if $K_2-K_1 \in \Gamma$.

Also H_1 and H_2 being Hilbert subspaces we put $H_1 \leq H_2$ if $H_1 \subset H_2$ and the inclusion map is norm decreasing (i.e. the unit ball of H_1 is included in the unit ball of H_2).

Let us now recall some results of L. SCHWARTZ [7] in the form of some lemmas:

LEMMA 1. *To each* $K \in \Gamma$ *there corresponds a unique Hilbert subspace* $H_K \hookrightarrow E$ *of which* K *is the reproducing kernel.*

LEMMA 2. $K_1 \leq K_2$ *if and only if* $H_1 \leq H_2$, K_i *being the reproducing kernel of* H_i.

LEMMA 3. *Let* $H \hookrightarrow E$ *be a Hilbert subspace with reproducing kernel* $K = jj^*$. *Let* $T \in L(H)$ *be a bounded hermitian operator in* H *such that* $0 \leq T \leq I$, *and let* $K_1 = jTj^*$. *Then* K_1 *is a hermitian kernel and* $0 \leq K_1 \leq K$. *Conversely every* K_1 *with* $0 \leq K_1 \leq K$ *is obtained in this manner from a unique operator* T.

LEMMA 4. *The space* H_1 *associated with* K_1 *in lemma* 3 *is a closed linear subspace of* H, *with the norm inherited from* H, *if and only if* T *is an orthogonal projection. In that case* T *is the orthogonal projection on* H_1.

Let $u: E \to E$ be continuous, linear, and one-to-one (in the sequel bijective). Then H is said to be *invariant* under u if the following two conditions are satisfied:

a) $uH = H$;

b) $\|ux\|_H = \|x\|_H \qquad \forall x \in H.$

In this case $u|_H$ is a unitary map.

A kernel $K \in \Gamma$ is said to be invariant under u if

(6) $\quad \langle Ku^*\phi,u^*\psi\rangle = \langle K\phi,\psi\rangle \qquad \forall \phi,\psi \in E^*.$

This is obviously equivalent to

(7) $\qquad uKu^* = K.$

LEMMA 5. *K is invariant under* u *if and only if* H_K *is invariant under* u.

For the proofs of these facts we refer to L. SCHWARTZ [7]. An elementary introduction to Hilbert subspaces and their kernels can also be found in [8].

In the sequel we shall equip the space F with a locally convex topology. Several topologies are in fact reasonable:

1. the topology of pointwise convergence on E^* (the kernel topology);
2. the topology of pointwise weak convergence on E^* (the weak kernel topology);
3. the topology of uniform convergence on $\sigma(E^*,E)$ bounded subsets of E^* (the strong kernel topology).

Finally let us briefly consider the case where $E = \mathcal{D}'(V)$, $E^* = \mathcal{D}(V)$ and $\langle T,\phi\rangle = T(\bar{\phi})$. In this case E^* is itself endowed with a topology and $\langle K\phi,\psi\rangle = \overline{\langle K\psi,\phi\rangle}$ is separately (hence jointly) continuous.

Let $K \in \mathcal{D}'(V\times V)$ be a distribution of positive type, i.e. such that $K(\phi\otimes\bar{\phi}) \geq 0$ for all $\phi \in \mathcal{D}(V)$. Then the equation

(8) $\qquad \langle K\phi,\psi\rangle = K(\phi\otimes\bar{\psi})$

defines a kernel $K \in \Gamma$, $K\phi$ being the distribution $\psi \to K(\phi\otimes\psi)$.

Conversely Schwartz's kernel theorem implies that every kernel $K \in \Gamma$ is obtained in this manner for a unique distribution K. We may thus identify F and Γ with subsets of $\mathcal{D}'(V\times V)$, and drop the notational distinction. The topology induced by $\mathcal{D}'_b(V\times V)$ on F is the strong kernel topology. The weak kernel topology is the one induced by $\sigma(\mathcal{D}'(V\times V),\mathcal{D}(V)\otimes\mathcal{D}(V))$.

Consider now an automorphism $u: \mathcal{D}'(V) \to \mathcal{D}'(V)$ defined by putting

(9) $\qquad \langle u(T),\phi\rangle = \langle T,\alpha\phi\circ\sigma\rangle,$

σ being a diffeomorphism and $\alpha \in C^\infty(V)$ with $\alpha(p) \neq 0$ $\forall p \in V$. Then $u^*(\phi) = \alpha\phi\circ\sigma$ and K is invariant if and only if

(10) $\qquad K(\alpha\phi\circ\sigma \otimes \overline{\alpha\psi\circ\sigma}) = K(\phi\otimes\bar{\psi})$

or equivalently:

(11) $\quad K_{\sigma\times\sigma} = \alpha\otimes\bar{\alpha}K$

$K_{\sigma\times\sigma}$ being the distribution defined (by virtue of the kernel theorem) by

(12) $\quad K_{\sigma\times\sigma}(\phi\otimes\psi) = K(\phi\circ\sigma^{-1} \otimes \psi\circ\sigma^{-1})$.

Similar results hold in Bruhat's theory of distributions for locally compact groups (cf. BRUHAT [1]).

2. INVARIANT KERNELS

Let G be a group (or set) of automorphisms of E (i.e. linear homeomorphisms).

We shall say that H (resp. K) is invariant under G if H (resp. K) is invariant under each $u \in G$. We denote by $G|H$ the set of restrictions $u|H$, for $u \in G$.

Let Γ_G denote the set of $K \in \Gamma$ invariant under G. Then Γ_G is clearly a closed convex cone in Γ (for any of the three topologies considered above).

PROPOSITION 1. *Let H be invariant and let $K = jj^*$ be its reproducing kernel, j being the inclusion of H in E. Let $H_1 \leq H$. Recall (lemma 3) that the reproducing kernel of H_1 has the form $K_1 = jTj^*$, where $0 \leq T \leq I$ in $L(H)$. Then H_1 is invariant if and only if T belongs to the commutant $(G|H)'$.*

PROOF. Let $u \in G$ and let $U = u|H$. Then $uj = jU$, hence $j^*u^* = U^*j^*$, U^* being the adjoint of the unitary operator U. Thus $uK_1u^* = ujTj^*u^* = jUTU^*j^*$. Now j being injective and j^* dense, this kernel equals $K_1 = jTj^*$ if and only if $UTU^* = T$, i.e. $UT = TU$.

Before proceeding further let us observe that the proper order of Γ_G is just the order induced by Γ in Γ_G, i.e. if K_1 and K_2 belong to Γ_G we have $K_2-K_1 \in \Gamma_G$ if and only if $K_2 - K_1$ belongs to Γ. This follows immediately from the fact that the difference of two G-invariant kernels is G-invariant.

Now let $\mathrm{ext}(\Gamma_G)$ denote the set of extreme generators of Γ_G, i.e. the set of kernels $K \in \Gamma_G$ such that $0 \leq K_1 \leq K$ implies $K_1 = \lambda K$ for some $\lambda \geq 0$.

PROPOSITION 2. $K \in \mathrm{ext}(\Gamma_G)$ *if and only if* $G|H_K$ *is irreducible.*

PROOF. By $G|H_K$ irreducible we mean that H_K contains no closed subspaces also invariant under G except for (0) and H. By Schur's lemma this is equivalent to $(G|H_K)' = \mathbb{C}I$, which again is equivalent to the fact that there are no operators T between 0 and I commuting with $G|H$ other than the λI, with $0 \leq \lambda \leq 1$. By proposition 1 this means that the only invariant kernels K_1 with $0 \leq K_1 \leq K$ are the multiples of K, i.e. K is extremal.

REMARK. In the case of function kernels, i.e. $E = \mathbb{C}^Q$, Q being some set, and the elements of G being defined by a group of permutations of Q, these propositions have been proved by KREIN [5]. In principle the above more general propositions can be deduced from these particular cases by noting that E is a subspace of $\mathbb{C}^{E^*}$, and that the maps u^* permute E^*. The direct proof seems simpler however.

COROLLARY. *The extremal elements of* $\Gamma = \Gamma_{\{Id\}}$ *are the degenerate kernels of rank one, the reproducing kernels of one dimensional Hilbert subspaces:* $\langle K\phi,\psi\rangle = \langle \overline{e,\phi}\rangle\langle e,\psi\rangle$, *denoted by* $\bar{e} \otimes e$.

Anticipating somewhat the next section observe that every $K \in \Gamma$ possesses an "integral representation" by means of extreme kernels: for any orthogonal basis $(e_i)_{i\in I}$ of H_K one has $K = \sum_{i\in I} \bar{e}_i \otimes e_i$. The obvious non-uniqueness of these decompositions is related to the fact, observed by SCHWARTZ [7], that Γ is not a lattice. Cones Γ_G do not necessarily have any extreme generators at all however.

PROPOSITION 3. Γ_G *is a lattice if and only if* $(G|H)'$ *is commutative for every* G*-invariant* H.

Let us introduce the following notation: For any convex cone Γ let $\Gamma(a)$ denote the set of elements of Γ dominated by a, i.e.

$$\Gamma(a) = \{x \in \Gamma: \exists\lambda \geq 0,\ x \leq \lambda a\} = \bigcup_{\lambda\geq 0} \Gamma \cap (a-\Gamma).$$

LEMMA 6.

a) *Let* Γ *be a proper convex cone. The proper order of* $\Gamma(a)$ *equals the order induced by* Γ;

b) Γ *is a lattice if and only if* $\Gamma(a)$ *is a lattice for all* a.

This is clear without proof.

LEMMA 7. *Let* $K \in \Gamma_G$ *and let* $H = H_K$. *Let* $A = (G|H)'$ *and let*

$$A^+ = \{T \in A: (Tx,x) \geq 0 \ \forall x \in H\}.$$

Then $\Gamma_G(K)$ *is linearly isomorphic to* A^+.

This is an immediate consequence of proposition 1. The proof of proposition 3 therefore results from:

LEMMA 8. (SHERMAN [10]). *Let* H *be a Hilbert space and let* $A \subset L(H)$ *be a Von Neumann algebra. Let* A^+ *be the set of positive hermitian operators belonging to* A. *Then* A^+ *is a lattice in its own order if and only if* A *is commutative.*

PROOF. If A is commutative A is isomorphic to a space $C(K)$, A^+ corresponding to $C(K)^+$ (Guelfand). Thus A^+ is a lattice.

Conversely assume A^+ is a lattice. It suffices to prove that two orthogonal projections P and Q in A commute. Let $A = \inf(P,Q)$ (relative to A). Let R be the orthogonal projection on $\overline{\text{Im } A}$. Then R belongs to A^+ and it is easy to see that $A \leq R \leq P,Q$; hence $R = A = \inf(P,Q)$. Let $P_1 = P-R$, $Q_1 = Q-R$. Then $\inf(P_1,Q_1) = 0$, and $P_1+Q_1 = \sup(P_1,Q_1) \leq I$, i.e. $P_1 \leq I-Q_1$ which means that P_1 and Q_1 are projections on mutually orthogonal subspaces, in particular $P_1Q_1 = Q_1P_1$, i.e. $(P-R)(Q-R) = (Q-R)(P-R)$. Expanding this and using the fact that $PR = RP (= R)$ and $QR = RQ$, we see that $PQ = QP$.

Now consider two groups (or sets) of automorphisms $G_1 \subset G_2 \subset GL(E)$. Then clearly $\Gamma_{G_2} \subset \Gamma_{G_1}$.

PROPOSITION 4. *Assume* $G_1 \subset G_2$. *Then, if* Γ_{G_1} *is a lattice,* Γ_{G_2} *is a lattice. Moreover, for* K' *and* K'' *in* Γ_{G_2}, $\sup(K',K'')$ *is the same whether calculated in* Γ_{G_1} *or in* Γ_{G_2}.

PROOF. The first assertion is an immediate consequence of proposition 3. For the proof of the second assertion, let us label with subscripts the a priori different suprema. Since $\sup_2(K',K'')$ belongs to Γ_{G_1} we have $\sup_1(K',K'') \leq \sup_2(K',K'')$. Therefore it is sufficient to show that $\sup_1(K',K'')$ belongs to Γ_{G_2}. Let K be an element in Γ_{G_2} majorising K' and K'' (e.g. $K'+K''$). Then $\sup_i(K',K'')$ is also the supremum of K' and K'' calculated in $\Gamma_{G_i}(K)$. Now $\Gamma_{G_2}(K) \subset \Gamma_{G_1}(K)$ and, by proposition 1, these cones are isomorphic respectively to A_2^+ and A_1^+, where $A_i = (G_i|H)'$, $H = H_K$. Thus we have $A_2^+ \subset A_1^+ \subset L(H)$ and it is sufficient to prove that for T',T'' in A_2^+,

$\sup_1(T',T'')$ belongs to A_2^+. Equivalently, it suffices to prove that for any hermitian $T \in A_2$, $|T| = \sup_1(T,-T)$ belongs to A_2. But this is obvious since $|T|$ is the limit in norm of polynomials in T, and A_2 is a closed algebra.

3. INTEGRAL REPRESENTATION THEORY

Let F be a quasi-complete locally convex space over $\mathbb{R}$. We denote by ℓ the elements of the dual space F'. We assume that F' contains a countable set of linear forms separating the points of F. Let us recall some definitions from the theory of conical measures of G. CHOQUET [2,3]. Let h(F) be the smallest subspace of $\mathbb{R}^F$ containing F' and containing, together with f and g, sup(f,g). The elements $f \in h(F)$ can be written in the form

$$f = \sup_i \ell_i - \sup_j \ell'_j,$$

i.e. as difference of two finite suprema of continuous linear forms. According to Choquet a conical measure is a linear form $\mu: h(F) \to \mathbb{R}$ such that $\mu(f) \geq 0$ for all $f \geq 0$. The resultant of a conical measure μ is the point a in the weak completion of F, such that $\ell(a) = \mu(\ell)$ for all $\ell \in F'$.

EXAMPLE. Let m be a positive Radon measure (cf. [9]) on $F\setminus\{0\}$ such that $\int |\ell(x)|dm(x) < +\infty$ for all $\ell \in F'$, and consequently $h(F) \subset L^1(m)$. Let us put

$$\mu(f) = \int f dm. \tag{13}$$

Then μ is a conical measure. In general not all conical measures can be defined in this way. A conical measure defined by relation (13) will be said to be *localizable*, and m will be said to be a localization of μ.

A localizable conical measure $\mu \neq 0$ possesses infinitely many localizations since from (13) follows that for any $\lambda > 0$

$$\mu(f) = \lambda^{-1} \int f(\lambda x)dm(x) = \int f d\tilde{m},$$

$\tilde{m}$ being λ^{-1} times the image of m under the map $x \to \lambda x$.

We have the following result however:

PROPOSITION 5. *Let* μ *be a localizable conical measure.*

1. *Let* Γ *be any cone in* F *(with vertex* 0*). Let* m_1 *and* m_2 *be two localizations of* μ. *Then* m_1 *is concentrated on* Γ *if and only if* m_2 *is concentrated on* Γ.
   In this case we say that μ *is concentrated on* Γ.
2. *Let* μ *be concentrated on* Γ *and let* $p: \Gamma \to [0,+\infty)$ *be any Borel function, positively homogeneous of degree* 1, *and such that* $p(x) > 0$ *for* $x \neq 0$ *(such functions exist though not necessarily linear and continuous). Let* $A = \{x \in \Gamma: p(x) = 1\}$. *Then there exists one and only one localization of* μ *concentrated on* A.
3. *Let* m *be any localization of* μ *and let* a *be the resultant of* μ. *Then*

$$a = \int x\,dm(x) \quad (i.e.\ \ell(a) = \int \ell(x)\,dm(x) \quad \forall \ell \in F').$$

Only the last statement is trivial. We shall write symbolically $a = \int x\,d\mu(x)$ *(cf.* [12]*).*

DEFINITION. Let Γ be a closed convex cone and let $\mathrm{ext}(\Gamma)$ be the cone of its extreme generators. We shall say that a point $a \in \Gamma$ possesses a (unique) integral representation by means of extreme generators if there exists a (unique) localizable conical measure μ concentrated on the cone $\mathrm{ext}(\Gamma)$ such that $a = \int x\,d\mu(x)$.

REMARKS. From proposition 5, 2. follows that if Γ possesses a basis $A = \{x \in \Gamma: \ell(x) = 1\}$, ℓ being a continuous linear form such that $\ell(x) > 0$ for all $x \in \Gamma\setminus\{0\}$, the point a possesses a (unique) integral representation by means of extremal generators if and only if a is the resultant of a (unique) Radon measure m concentrated on $A \cap \mathrm{ext}(\Gamma)$, i.e. on the set of extreme points of A.

On the other hand, whether Γ possesses a basis or not, there always exists a topological Hausdorff space T and a continuous function $t \to e(t) \in \mathrm{ext}(\Gamma)$, taking but one value on each ray, such that $a \in \Gamma$ possesses a (unique) integral representation by means of extremals if and only if there exists a (unique) Radon measure on T such that $\int |\ell(e(t))|\,dm(t) < +\infty$ for all $\ell \in F'$, and $a = \int e(t)\,dm(t)$. It suffices to choose $T = \{x \in \mathrm{ext}(\Gamma): p(x) = 1\}$, p being a Borel function of the kind described in 2. proposition 5, and to take e to be the identity map.

DEFINITION (CHOQUET). Let Γ be a convex cone. A *cap* of Γ is a convex compact subset $C \subset \Gamma$ containing 0 and such that $\Gamma\setminus C$ is convex. A cone Γ is

said to be *well capped* if Γ is the union of its caps (e.g. a cone with compact base is well capped).

A closed convex cone of a well capped cone is well capped, in fact, if Γ_1 is a closed subcone of Γ and C is a cap of Γ, $C \cap \Gamma_1$ is a cap of Γ_1.

THEOREM ([11] or [12]). *Let* F *be a quasi-complete space such that* F' *contains a countable system of linear forms separating the points of* F. *Let* $\Gamma \subset F$ *be a closed convex well capped cone. Then we have:*

A) *every point in* Γ *possesses an integral representation by means of extreme generators;*

B) *every point in* Γ *possesses a unique integral representation by means of extreme generators if and only if* Γ *is a lattice in its own order.*

PROOF. See [11] for an outline, [12] for details.

The application of this theorem to the situation considered here depends on the following proposition:

PROPOSITION 6. *Let* E *be a quasi-complete barralled conuclear space*[*]. *Let* Γ *be the cone of hermitian positive kernels* $K: E^* \to E$ *equipped with the topology of uniform convergence on bounded sets (strong kernel topology). Then* Γ *is well capped.*

COROLLARY. Γ_G *is well capped for any* $G \subset GL(E)$.

PROOF. Let $K \in \Gamma$ and let C_K be the set of kernels $K' \in \Gamma$ satisfying the following conditions:

1) $H_{K'} \subset H_K$,
2) the inclusion map is of Hilbert Schmidt type, with Hilbert Schmidt norm at most equal to 1.

Then C_K is a cap in Γ. This can be proved by making use of the fact that the set of positive operators in $L(H)$, of trace at most 1, forms a cap in $L^+(H)$ equipped with the weak operator topology. On the other hand, E being conuclear, for every Hilbert suspace $H' \subset E$ there exists $H \supset H'$ such that the inclusion operator $H' \subset H$ is of Hilbert Schmidt type (cf. [9], p.230).

---

*) These spaces are precisely the strong duals of barralled nuclear spaces (cf. SCHWARTZ [9]). Thus $E = \mathcal{D}'(V)$ is such a space; moreover $\mathcal{D}(V) = E'$ is separable.

This implies that $\Gamma = \bigcup_{K\in\Gamma} C_K$.

REMARK. Assume moreover that E^* contains a countable set (ϕ_n) separating the points of E. Then the linear forms $K \to \langle K\phi_n, \phi_m\rangle$ separate F and so the previous theorem may be applied to Γ and to Γ_G for any set $G \subset GL(E)$.

4. APPLICATIONS

Let us summarize the results in a somewhat more symmetric fashion with a slight change of notation.

THEOREM. *Let $\mathcal{D}$ be a separable, barralled, nuclear space. Let $G \subset GL(\mathcal{D})$ be a group of automorphisms of $\mathcal{D}$. Let Γ_G be the set of hermitian positive kernels on $\mathcal{D} \times \mathcal{D}$ such that $K(u\phi,u\psi) = K(\phi,\psi)$ for all $u \in G$. Then we have:*

1. *Every element of Γ_G possesses an integral representation by means of extreme elements.*
2. *The following conditions are equivalent:*
   - i) *Γ_G is a lattice;*
   - ii) *the integral representations are unique;*
   - iii) *the representations of G associated with the kernels K are multiplicity free.*
3. *If $G_1 \subset G_2 \subset GL(\mathcal{D})$ and Γ_{G_1} is a lattice, so is Γ_{G_2}, with the lattice operations inherited from Γ_{G_1}.*

This justifies the assertions A,B,C in the Introduction. As for D) we still have to prove the following:

THEOREM. *Let G be a unimodular locally compact group (second countable). Let $\mathcal{D}(G)$ be the space of Bruhat test functions. Let Γ_{bi} be the cone of bi-invariant distributions $K \in \mathcal{D}'(G\times G)$ of positive type. Then Γ_{bi} is a lattice in its own order. Consequently each element in Γ_{bi} (trace) has a unique integral representation by means of extreme elements (characters).*

PROOF. It suffices to prove that if $H \subset \mathcal{D}'(G)$ is any bi-invariant Hilbert space the set U of operators, commuting with both right and left translation operators, is commutative. Now let $L(R)$ be the von Neumann algebra generated by the left (right) translation operators in H. Then $U = L' \cap R'$. But by the Godement-Segal commutativity theorem ([4], p.71) $L' = R$ and so

$U = R \cap R'$ is commutative.

We terminate with an elementary example proving the need for the introduction of distributions.

Let $V = \mathbb{R}$ and let $G \subset \mathcal{D}'(\mathbb{R})$ be the group generated by the translations and by the operator defined by

$$u^*(\phi)(t) = \alpha^{-\frac{1}{2}}\, \phi(\alpha^{-1}t), \qquad \alpha > 1$$

so that $u(fdx) = \alpha^{\frac{1}{2}} f(\alpha x)dx$ and L^2 is invariant). Then the irreducible Hilbert subspaces of $\mathcal{D}'(\mathbb{R})$ can be shown to be the spaces H_λ of distributions of the form

$$f = F \sum_{n\in\mathbb{Z}} f_n \alpha^n \delta_{(\alpha^n\lambda)} \qquad \text{with} \qquad \sum_n \alpha^n |f_n|^2 < +\infty,$$

i.e.

$$f(t) = \sum_n e^{i\alpha^n\lambda t}\, \alpha^n\, f_n \qquad \text{(convergence in } S')$$

$$= \frac{1}{i\lambda}\frac{d}{dt}\sum_n e^{i\alpha^n\lambda t}\, t_n = \frac{1}{i\lambda}\frac{d}{dt} F.$$

Now if $\alpha > 1$ is a sufficiently large odd integer and $f_n = 0$ for $n < 0$, $f_n = \beta^n$ for $n \geq 0$ with $\alpha\beta^2 < 1$ and $\alpha\beta > 1 + \frac{3}{2}\pi$, F is a continuous function which has been shown by Weierstrass to be nowhere differentiable (Hardy also proved this for $\alpha\beta > 1$). It follows that F is of bounded variation in no interval, so that f is a distribution which reduces to a measure in no open interval. A fortiori the reproducing kernels of the spaces H_λ cannot be measures.

REFERENCES

[1] BRUHAT, F., *Distributions sur un groupe localement compact etc.*, Bull. Soc. Math. France 89 (1961) 43-75.

[2] CHOQUET, G. & P.A. MEYER, *Existence et unicité de representations integrales dans les convexes compacts quelconques*, Ann. Inst. Fourier 13 (1963) 139-154.

[3] CHOQUET, G., *Lectures on Analysis* (Benjamin 1969).

[4] GODEMENT, R., *Théorie des caracteres I & II,* Ann. of Math. 59 (1954) 47-85.

[5] KREIN, M.G., *Hermitian positive kernels on homogeneous spaces I & II,* Translations Amer. Math. Soc. Series 2 Vol 34.

[6] MAURIN, K., *General eigenfunction expansions and unitary representations of topological groups,* Monografie Matematyczne Tom 48.

[7] SCHWARTZ, L., *Sous espaces hilbertiens d'espaces vectoriels topologiques et noyaux associés,* J. Analyse Math. 13 (1964) 115-256.

[8] SCHWARTZ, L., *Application of distributions to the theory of elementary particles in quantum mechanics* (Gordon & Breach).

[9] SCHWARTZ, L., *Radon measures on arbitrary topological spaces and cylindrical measures* (Oxford Univ. Press) 1973.

[10] SHERMAN, S., *Order in Operator Algebras,* Amer. Jour. Math. Vol. 73, 1951, 227-232.

[11] THOMAS, E., *Integral representations in conuclear spaces,* Proc. Conf. on vector space measures and applications, Dublin 1977 (Springer Lecture Notes no. 645).

[12] THOMAS, E., *Integral representations in convex cones,* Report ZW-7703.

DISTANCE SETS OF SEQUENCES OF INTEGERS

R. Tijdeman

1. Let $A = \{a_1, a_2, \ldots\}$ be an infinite, strictly increasing sequence of non-negative integers. The *distance set* $\mathcal{D}(A)$ of A is defined to be the set of all non-negative integers which occur as the difference between two elements of A. The *distance sequence* of A, also denoted by $\mathcal{D}(A)$, is obtained by arranging the elements of the distance set of A in increasing order. By $N(A,n)$ we denote the number of elements of A not exceeding n. We call $\bar{d}(A) = \limsup_{n\to\infty} N(A,n)/n$ the *upper density* of A and $\underline{d}(A) = \liminf_{n\to\infty} N(A,n)/n$ the *lower density* of A. If $\bar{d}(A) = \underline{d}(A)$ this value is said to be the *density* $d(A)$ of A.

It is not difficult to show that $\underline{d}(\mathcal{D}(A)) \geq \bar{d}(A)$ for every sequence A. If $d(A) = 0$, this inequality is trivial. It was conjectured by Erdös and proved by RUZSA [11] that in this case even

$$\lim_{n\to\infty} \frac{N(\mathcal{D}(A),n)}{N(A,n)} = \infty.$$

On the other hand, it can happen that $\mathcal{D}(A) = A$. Take for example for A the sequence of all non-negative multiples of a fixed integer.

Our aim is to find necessary and sufficient conditions for subsets of the set of non-negative integers $\mathbb{N}_0$ in order to be the distance set of some sequence A. Since it seems to be difficult to formulate a simple criterion, we shall consider questions like these:

a) What can be said about the distances of consecutive elements of a distance sequence?

b) Does every distance sequence of a dense sequence contain an arithmetical progression?

c) Is the intersection of two distance sets also a distance set?

d) How dense can a sequence A be, if the distance set $\mathcal{D}(A)$ does not contain any elements from a given set K? An interesting choice for K is the set of all positive squares.

These questions arose as natural questions during the investigations or originated from problems in other fields.

Because of possible applications it was interesting also to investigate the above mentioned questions for certain subsets of $\mathcal{D}(A)$. We define the *infinite distance set* $\mathcal{D}_\infty(A)$ of A as the set of all non-negative integers which occur infinitely often as the difference of two terms of A. Further we define the *density distance set* $\mathcal{D}_0(A)$ of A as the set of all non-negative integers d such that the upper density of the integers a with both $a \in A$ and $a + d \in A$ is positive. It is obvious that $\mathcal{D}_0(A)$ may be empty. For example $\mathcal{D}_\infty(A) = \{0\}$ if and only if $a_{j+1} - a_j \to \infty$ as $j \to \infty$ and in this case $\mathcal{D}_0(A) = \emptyset$. On the other hand, it is not difficult to show that $\mathcal{D}_\infty(A)$ and $\mathcal{D}_0(A)$ are non-empty if $\bar{d}(A) > 0$, and that, moreover, $\underline{d}(\mathcal{D}_\infty(A)) \geq \underline{d}(\mathcal{D}_0(A)) \geq \bar{d}(A)$ for every sequence A.

In many aspects $\mathcal{D}(A)$, $\mathcal{D}_\infty(A)$ and $\mathcal{D}_0(A)$ behave in a similar way. In this respect the following result is very useful.

LEMMA 1. *For every sequence* A *with* $\bar{d}(A) > 0$ *there exists a sequence* B *with* $\underline{d}(B) \geq \bar{d}(A)$ *such that* $\mathcal{D}(B) \subset \mathcal{D}_0(A)$.

In view of this lemma it suffices for establishing that $\underline{d}(\mathcal{D}_0(A)) \geq \bar{d}(A)$ for every sequence A to prove that $\underline{d}(\mathcal{D}(A)) \geq \bar{d}(A)$ for every sequence A. The latter result is a consequence of another useful lemma.

LEMMA 2. *Let* A *be a sequence with* $\bar{d}(A) = \alpha > 0$. *For any positive integer* b *there are at least* $[\alpha r]$ *of the integers* b,2b,...,rb *in* $\mathcal{D}(A)$.

2. According to RUZSA [11] it was proved by Erdös and Sárközy that if $\bar{d}(A) > 0$, then the differences between consecutive terms of $\mathcal{D}(A)$ are bounded by a number M. For a proof see [15]. By Lemma 1 the corresponding results hold for $\mathcal{D}_\infty(A)$ and $\mathcal{D}_0(A)$. It is not possible to find an upper bound M which depends only on $\bar{d}(A)$. For let A_t denote the sequence of integers of the form $3nt + i$ for $i = 1,\ldots,t$ and $n = 0,1,2,\ldots$. Then $\mathcal{D}(A_t)$ consists of the non-negative integers of the form $3nt \pm i$ for $i = 1,\ldots,t$ and $n = 0,1,2,\ldots$ and so contains infinitely many gaps of length t. On the other hand, $d(A_t) = 1/3$ for every t. This fact was observed for the first time

by SZEMERÉDI, [11].

3. Erdős asked whether every distance sequence of a sequence A with $\bar{d}(A) > 0$ contains an arithmetical progression. More generally, one might ask whether there exists a countable set $\mathcal{E}$ of infinite subsets of $\mathbb{N}_0$ such that every distance set $\mathcal{D}(A)$ contains at least one element of $\mathcal{E}$. It is well known (see e.g. [8] Ch.1, Theorem 4.1) that for any sequence $E = \{e_1, e_2, \ldots\}$ the sequence $\{\eta e_k\}_{k=1}^{\infty}$ is uniformly distributed modulo 1 for almost all real numbers η. Hence, given countably many sequences $E^{(i)} = \{e_k^{(i)}\}$ we can find an irrational number θ for which $\{\theta e_k^{(i)}\}$ is uniformly distributed modulo 1 for all i. By applying the following result to this number θ we see that Erdős' question is answered in the negative.

THEOREM 1. *Let θ be an irrational number and let α be a number between 0 and 1. There exist uncountably many sequences A with density α for which*

$$\bar{d}(\mathcal{D}(A) \cap E) \leq 2\alpha\, \bar{d}(E)$$

for every sequence $E = \{e_1, e_2, \ldots\}$ such that $\{\theta e_k\}_{k=1}^{\infty}$ is uniformly distributed modulo 1.

This result is non-trivial if $\alpha < \frac{1}{2}$. However, if $\alpha > \frac{1}{2}$, then $\mathcal{D}(A) = \mathbb{N}_0$. It would be interesting to have a simple, not very thick, set $\mathcal{E}$ of infinite subsets of $\mathbb{N}_0$ such that every distance set $\mathcal{D}(A)$ contains at least one element of $\mathcal{E}$.

4. It was proved by RUZSA [11] that the union of two distance sets need not be a distance set. Let A_1 and A_2 be sequences of positive upper density. STEWART and TIJDEMAN [15] showed that $\mathcal{D}(A_1) \cap \mathcal{D}(A_2)$ need not be a distance set and one can show by a similar example that $\mathcal{D}(A_1) \cup \mathcal{D}(A_2)$ need not be a distance set. Suprisingly it is true that both $\mathcal{D}_0(A_1) \cup \mathcal{D}_0(A_2)$ and $\mathcal{D}_0(A_1) \cap \mathcal{D}_0(A_2)$ are density distance sets. This can be proved by constructing sequences B_1 and B_2 with $\mathcal{D}_0(B_1) = \mathcal{D}_0(A_1) \cup \mathcal{D}_0(A_2)$ and $\mathcal{D}_0(B_2) = \mathcal{D}_0(A_1) \cap \mathcal{D}_0(A_2)$. For infinite distance sets the situation is even more striking. We have

THEOREM 2. *The collection of infinite distance sets associated with sequences of positive upper density is a filter on the set of all subsets of $\mathbb{N}_0$.*

It is not true that it is also an ultrafilter. For there exist

disjoint sets D_1 and D_2 satisfying $D_1 \cup D_2 = \mathbb{N}_0$ and $\underline{d}(D_1) = \underline{d}(D_2) = 0$. Neither D_1 nor D_2 is an infinite distance set of a sequence of positive upper density, since $\underline{d}(\mathcal{D}_\infty(A)) \geq \bar{d}(A)$ for every sequence A. The proof of Theorem 2 essentially falls into two parts, the proof of the extension property and the proof of the intersection property.

The *extension property* says that if A is a sequence with $\bar{d}(A) > 0$ and D is a set with $\mathcal{D}_\infty(A) \subseteq D$, then there exists a sequence B with $\mathcal{D}_\infty(B) = D$. In fact we have

THEOREM 3. *Let* A *be a sequence and let* D *be a set with* $\mathcal{D}_\infty(A) \subset D$. *Then there exists a sequence* B *with* $\bar{d}(B) = \bar{d}(A)$ *and* $\underline{d}(B) = \underline{d}(A)$ *whose infinite distance set is* D.

An immediate consequence of this result is that there exist sequences A with $\bar{d}(A) = \underline{d}(A) > 0$ for which $\bar{d}(\mathcal{D}_\infty(A)) > \underline{d}(\mathcal{D}_\infty(A))$. The extension property does not hold for the other types of distance sets. Take for A the non-negative even integers and put $D = A \cup \{1\}$. It is obvious that there does not exist a sequence B with $\mathcal{D}(B) = D$. A more complicated argument shows that there is neither a sequence B with $\mathcal{D}_0(B) = D$. Thus neither the distance sets nor the density distance sets form a filter in the sense of Theorem 2.

5. The *intersection property* says that if $A_1,\ldots,A_h$ are sequences with positive upper density, then $\mathcal{D}(A_1) \cap \ldots \cap \mathcal{D}(A_h)$ contains an infinite distance set $\mathcal{D}(B)$ of a sequence B with $\bar{d}(B) > 0$. Erdös posed the problem to prove that $\mathcal{D}(A_1) \cap \ldots \cap \mathcal{D}(A_h)$ is non-empty. This was independently of each other solved by Prikry and by Stewart and Tijdeman. PRIKRY [9] deduced this result by means of a theorem of HINDMAN [7], [1], [3], which says that if $\mathbb{N}_0$ is divided into two sets then there is a sequence drawn from one of these sets such that all finite sums of distinct numbers of this sequence remain in the same set. He even proved the much stronger assertion that $\mathcal{D}_0(A_1) \cap \ldots \cap \mathcal{D}_0(A_h)$ does not contain gaps of arbitrary length. He further applied his result to a problem in chromatic graph theory. The proof of Stewart and Tijdeman is elementary and uses cyclic shifts. In this way they obtained quantitative results which imply the results of Prikry. Let $X + k$ denote the set $\{x + k \mid x \in X\}$.

THEOREM 4. *Let* $A_1,\ldots,A_h$ *be sequences with positive upper densities* $\alpha_1,\ldots,\alpha_h$ *respectively. Put* $C_1 = \alpha_1$ *and* $C_h = \prod_{i=1}^{h} (\alpha_i/5 \log(h+1))$ *for*

$h \geq 2$. *Then there exists a sequence* A *with* $\underline{d}(A) \geq C_h$ *such that* $\mathcal{D}(A) \subseteq \cap_{i=1}^{h} \mathcal{D}_0(A_j)$. *Furthermore, there exist* r *integers* $k_1,\ldots,k_r$ *such that*

$$\bigcup_{j=1}^{r} \left\{\left(\bigcap_{i=1}^{h} \mathcal{D}_0(A_i)\right) + k_j\right\} \supseteq \mathbb{N}_0$$

with $r \leq C_h^{-\log 3/\log 2}$.

Apart from the factor $5\log(h+1)$ in the definition of C_h the first assertion is best possible. For let $n_1,\ldots,n_h$ be positive integers and $A_1 = \{a \mid a \geq 0 \text{ and } a \equiv 0 \pmod{n_1}\}$ and $A_i = \{a \mid a \geq 0 \text{ and } a \equiv 0,1,\ldots,n_1\ldots n_{i-1} \pmod{n_1\ldots n_i}\}$ for $i = 2,\ldots,h$. Then

$$d\left(\bigcap_{i=1}^{h} \mathcal{D}_0(A_i)\right) = \prod_{i=1}^{h} d(A_i) = \prod_{i=1}^{h} \alpha_i.$$

It further follows from the second assertion that $\cap_{i=1}^{h} \mathcal{D}_0(A_i)$ cannot contain gaps of size larger than twice the maximum in absolute value of the k_j's. For if there was a larger gap, the integer(s) closest to the middle of the gap would not be in the union of the sets $(\cap_{i=1}^{h} \mathcal{D}_0(A_i)) + k_j$. In particular the theorem implies the results mentioned in section 2. Observe that because of the intersection property and the extension property both $\mathcal{D}_\infty(A_1) \cup \mathcal{D}_\infty(A_2)$ and $\mathcal{D}_\infty(A_1) \cap \mathcal{D}_\infty(A_2)$ are infinite distance sets.

It is very likely that the first assertion of Theorem 4 remains valid if C_h is replaced by $\Gamma_h = \prod_{i=1}^{h} \alpha_i$. The example given above shows that the inequality in the second assertion cannot be better than $r \leq \Gamma_h^{-1}$. It is an open problem whether the second assertion with this inequality holds.

6. Let K be any given set of positive integers. It is clear that there exists a sequence A with $\mathcal{D}(A) \cap K = \emptyset$ and $d(A) = t^{-1}$, if no multiple of t is in K. Therefore A can be chosen to have positive density if $|K| < \infty$. Here $|K|$ denotes the cardinality of K. Let $\mu(K) = \sup \bar{d}(A)$, where the supremum is taken over all sequences A with $\mathcal{D}(A) \cap K = \emptyset$. T.S. Motzkin posed the problem to compute $\mu(K)$. CANTOR and GORDON [2] determined the exact value of $\mu(K)$ if $|K| \leq 2$. Of course, $\mu(K) = \frac{1}{2}$ if $|K| = 1$. They found for $K = \{k_1,k_2\}$ with $(k_1,k_2) = d$,

$$\mu(K) = \left(d\left[\frac{k_1+k_2}{2d}\right]\right) / (k_1+k_2).$$

For the general case they obtained the following result.

THEOREM 5. (CANTOR and GORDON [2]).
Let $|x|_\ell$ *denote the absolute value of the absolutely least remainder of* x (mod.ℓ). *Then*

$$\mu(K) \geq \sup_{(c,\ell)=1} \frac{1}{\ell} \min_j |ck_j|_\ell,$$

where the supremum is taken over all positive integers c *and* ℓ *with* $(c,\ell) = 1$.

They wondered whether equality holds if the elements of K are relatively prime. Recently HARALAMBIS [6] established the value of $\mu(K)$ for several classes of sets K with $|K| \leq 4$.

7. Natural sets K for which it is not obvious that there exist sequences A with $\bar{d}(A) > 0$ and $\mathcal{D}(A) \cap K = \emptyset$ are for example the factorials and the squares. It was proved by STEWART and TIJDEMAN [15] that $\mu(K) \geq 1/9$ if $K = \{k! \mid k \in \mathbb{N}\}$. M. Voorhoeve observed that this implies that the above question of Cantor and Gordon has a negative answer, at least for infinite sets K. Taking $k_j = j!$ for all j gives a bound 0 in Theorem 5. The lower bound 1/9 for $\mu(K)$ follows from a quantitative version of the following result.

THEOREM 6. *If* $k_1, k_2, \ldots$ *is a sequence of positive integers satisfying* $\liminf_{j\to\infty} k_{j+h}/k_j > 1$ *for some fixed* h, *then there exists a sequence* A *with* $\underline{d}(A) > 0$ *for which* $k_j \notin \mathcal{D}(A)$ *for* $j = 1,2,\ldots$.

This result is in a sense best possible. It is not difficult to prove that if $\liminf_{j\to\infty} k_{j+h}/k_j = 1$ for every h, then there exists a sequence $L = \{\ell_1, \ell_2, \ldots\}$ with $\ell_{j+1}/\ell_j \geq k_{j+1}/k_j$ for $j = 1,2,\ldots$ such that $\mathcal{D}(A) \cap L \neq \emptyset$ for every sequence A with $\bar{d}(A) > 0$.

8. In case K is the set of all positive squares, $\mu(K) = 0$. This was shown by Furstenberg and Sárközy independently. FURSTENBERG [5] deduced his result as a by-product of his ergodic proof of the theorem of Szemerédi that every sequence A with $\underline{d}(A) > 0$ contains arbitrarily long finite arithmetical progressions. SÁRKÖZY [12] used an adaptation of the Hardy-Littlewood method elaborated by Roth. He also obtained quantitative results.

THEOREM 7. (SÁRKÖZY [12], [13])
If $\mathcal{D}(A)$ does not contain any positive squares, then

$$N(A,x) = O(x(\log\log x)^{2/3} (\log x)^{-1/3}).$$

On the other hand, there exists a sequence A with

$$N(A,x) > x^{1/2} \exp\{\tfrac{1}{3} \log x \log\log\log x (\log\log x)^{-1}\}$$

such that $\mathcal{D}(A)$ does not contain any positive squares.

Sárközy conjectured that $N(A,x) = O(x^{\frac{1}{2}+\varepsilon})$ for any $\varepsilon > 0$ for such sequences A. In a third paper he considered other sets K for which his method works. The case $K_t = \{k^t \mid k \in \mathbb{N}\}$ for any $t \in \mathbb{N}$, $t \geq 3$ is quite similar to the case of the squares K_2. He further treated the sets $K = \{p-1 \mid p \text{ prime}\}$, in which case he obtained $N(A,x) = O(x (\log\log\log x)^4 (\log\log x)^{-2})$ for every sequence A with $\mathcal{D}(A) \cap K = \emptyset$ and he mentioned that the case $K = \{k^2-1 \mid k \in \mathbb{N}, k > 1\}$ can be dealt with similarly. On the other hand, it is almost trivial that in the cases $K = \{p \mid p \text{ prime}\}$ and $K = \{k^2+1 \mid k \in \mathbb{N}\}$ one has $\mu(K) > 0$.

9. If K is still denser, it is even uncertain whether there exists an infinite set A with $\mathcal{D}(A) \cap K = \emptyset$. It is simple to show that $\limsup_{j\to\infty} (k_{j+1}-k_j) = \infty$ is a sufficient condition.

This result is best possible in a similar sense as Theorem 6 is. If $k_{j+1} - k_j$ is bounded for all j, then there exists a sequence $L = \{\ell_1,\ell_2,\ldots\}$ with $\ell_{j+1} - \ell_j \geq k_{j+1} - k_j$ for $j = 1,2,\ldots$ such that $\mathcal{D}(A) \cap L \neq \emptyset$ for every infinite sequence A.

A closely related problem was studied by ERDÖS and HARTMAN [4]. Let A be an infinite sequence of positive integers. The set $B \subset \mathcal{D}(A)$ is said to be *avoidable* if there is an infinite subsequence A_1 of A such that $\mathcal{D}(A_1)$ and B are disjoint. They proved for example that for every ε with $0 < \varepsilon < 1$ there is a B which is not avoidable but of relative density at most ε in A. On the other hand, if $\underline{d}(A) > 0$ and $\underline{d}(B) = 0$, then B is avoidable. ROTENBERG [10] proved that $B = \{b_j\}_{j=1}^{\infty}$ is avoidable if $b_{j+1} - b_j \to \infty$ as $j \to \infty$, thereby solving a problem of Erdös and Hartman.

The results mentioned in the sections 6-9, found during independent investigations, fit together to a global answer to question d). It would be very interesting to have a general theorem complementing Theorem 6 by describing the structural properties of a sequence $K = \{k_1, k_2, \ldots\}$ such that $\mathcal{D}(A) \cap K \neq \emptyset$ for every sequence A with $\bar{d}(A) > 0$. The results of Sárközy for particular sequences are important steps in this direction.

10. Proof of all numbered theorems on distance sets and on infinite distance sets can be found in the paper by STEWART and TIJDEMAN [15], unless stated otherwise. For density distance sets most results follow directly from an application of Lemma 1 to these results. The proof of this lemma is unpublished yet.

There is another type of distance sets which might be studied. Let, for any ε with $0 < \varepsilon < 1$, $\mathcal{D}_\varepsilon(A)$ be the set of all non-negative integers d such that the upper density of the integers a with both $a \in A$ and $a+d \in A$ is at least ε. Then $\mathcal{D}_\varepsilon \subset \mathcal{D}_0$, but the behaviour of $\mathcal{D}_\varepsilon$ is quite different from $\mathcal{D}_0$ and only a few trivial results about $\mathcal{D}_\varepsilon$ are known.

Even if you did not like the topic it might be worthwhile to remember that number theory, combinatorics, graph theory, logic and ergodic theory have common facets and that the interaction between these field led to new research and results.

REFERENCES

[1] BAUMGARTNER, J., *A short proof of Hindman's theorem*, J. Comb. Th. Ser. A 17 (1974), 384-386.

[2] CANTOR, D.G. & B. GORDON, *Sequences of integers with missing differences*, J. Comb. Th. Ser. A 14 (1973), 281-287.

[3] COMFORT, W.W., *Ultrafilters: some old and new results*, Bull. Amer. Math. Soc. 83 (1977), 417-455.

[4] ERDÖS, P. & S. HARTMAN, *On sequences of distances of a sequence*, Colloq. Math. 17 (1967), 191-193.

[5] FURSTENBERG, H., *Ergodic behavior of diagonal measures and a theorem of Szemerédi on arithmetic progressions*, J. Analyse Math. 31 (1977), 204-256.

[6] HARALAMBIS, N.M., *Sets of integers with missing differences*, J. Comb. Th. Ser. A 23 (1977), 22-33.

[7] HINDMAN, N., *Finite sums from sequences within cells of a partition of* N, J. Comb. Th. Ser. A 17 (1974), 1-11.

[8] KUIPERS, L. & H. NIEDERREITER, *Uniform distribution of sequences*, Wiley, New York etc., 1974.

[9] PRIKRY, K., *Private communication*.

[10] ROTENBERG, D., *Sur une classe de parties évitables*, Colloq. Math. 20 (1969), 67-68.

[11] RUZSA, I.Z., *On difference-sequences*, Acta Arith. 25 (1974), 151-157.

[12] SÁRKÖZY, A., *On difference sets of sequences of integers* I, Acta Math. Acad. Sci. Hungar, to appear.

[13] _____, *On difference sets of sequences of integers* II, Annales Univ. Sci. Budapest. Eötvös, to appear.

[14] _____, *On difference sets of sequences of integers* III, to appear.

[15] STEWART, C.L. & R. TIJDEMAN, *On infinite difference sets*, Preprint ZW 100/77, Math. Centr. Amsterdam, 1977.

Mathematical Institute
Wassenaarseweg 80
Leiden, The Netherlands.

ADDENDUM

The problems mentioned at the end of section 5 have been solved. Both Katznelson and Ruzsa have given elementary proofs of Theorem 4 with $C_h = \pi_{i=1}^{h} \alpha_i$ and $r \leq C_h^{-1}$. It has also been noted that these results can be deduced by ergodic theory.

ERDÖS & SÁRKOZY [16] proved a result which is closely related to Theorem 6. Their proof is essentially the same as the one obtained independently by Stewart and the author, but the quantitative forms of the results differ.

Important progress is made concerning the problem mentioned at the end of section 9. Let $P \in \mathbb{Z}[x]$ with $P(0) = 0$ and $P(n) \to \infty$ as $n \to \infty$. Put $K = \{P(n) \mid n \in \mathbb{N},\ P(n) > 0\}$. H.L. Montgomery proved that if A is a sequence with $\mathcal{D}(A) \cap K = \emptyset$, then $d(A) = 0$. This result can be extended to the following theorem which is the best possible.

THEOREM 8. (KAMAE & MENDES FRANCE, [18]). *Let* $P \in \mathbb{Z}[x]$. Put $K = \{P(n) \mid n \in \mathbb{N},\ P(n) > 0\}$. *If for every* $r \in \mathbb{N}$ *there exists an* $m \in K$ *such that* $r \mid m$, *then every sequence* A *with* $\mathcal{D}(A) \cap K = \emptyset$ *has density* 0.

The counterpart of this theorem is obvious. If there exists an $r \in \mathbb{N}$ such that no element of K is divisible by r, then there exists a sequence A with $d(A) > 0$ such that $\mathcal{D}(A) \cap K = \emptyset$, namely $A = \{0,r,2r,\ldots\}$. By a theorem of FROBENIUS [17] every irreducible polynomial $P \in \mathbb{Z}[x]$ of degree at least two has a prime p with $p \nmid P(n)$ for all $n \in \mathbb{Z}$ and does therefore not fulfil the condition of the theorem. Examples of polynomials satisfying the condition of Theorem 8 and not having an integral zero are $(x^3-19)(x^2+3)$ and $(x^2-13)(x^2-17)(x^2-221)$.

KAMAE and MENDES FRANCE [18] showed an interesting relation with the theory of uniform distribution. The set $H \subset \mathbb{N}$ is said to be a Van der Corput set if the uniform distribution (mod. 1) of the sequences $(u_{n+h} - u_n)_{n=1}^{\infty}$ for all $h \in H$ implies the uniform distribution (mod. 1) of $(u_n)_{n=1}^{\infty}$. They proved that sets K which satisfy the condition of Theorem 8 are Van der Corput sets. It further turns out that H cannot be a Van der Corput set, if there exists a sequence A with $\bar{d}(A) > 0$ such that $\mathcal{D}(A) \cap H = \emptyset$. The converse question is still open. Is it true that H is a Van der Corput set, if $\mathcal{D}(A) \cap H \neq \emptyset$ for every sequence A with $\bar{d}(A) > 0$?

[16] ERDÖS, P. & A. SÁRKÖZY, *On differences and sums of integers*, II. To appear.

[17] FROBENIUS, F.G., *Über Beziehungen zwischen den Primidealen eines algebraischen Körpers und den Substitutionen seiner Gruppe*, S.-B. Preuss. Akad. Wiss. 1896, 689-703 = Gesamm. Abh. II, Springer-Verlag, Berlin etc., 1968, pp. 719-733.

[18] KAMAE, T. & M. MENDES FRANCE, *Van der Corput's difference theorem*. To appear.

THINKING ON TWO LEVELS

A. van Wijngaarden

1. INTRODUCTION

Two-level grammars, introduced in [1], have been mainly used to define programming languages [2]. However, they can be used to express algorithms directly, without the intervention of a programming language. Since the grammars imply only a play with letters without any meaning, the programmer is forced to go into minute details, on the one hand, but, on the other hand, does not need to conform his way of thinking to the idiosyncrasies of a particular language.

Some experience shows that the two-level mechanism suits the human way of thinking well. The lower level of the grammar enables us to depict a specific situation, the higher level to express an abstraction, viz., a class of situations. These are precisely the tools of human thought.

In order to illustrate this point, this paper gives a self-contained exposition. Section 2 defines the two-level concept and gives some practical conventions. If the reader has grasped the idea, then he may forget that some elementary mathematical concepts and notations have been used in this section in order to define the tool, since they are of no relevance later on; it is only the tool that matters.

In the following sections some examples of grammars are given in quickly ascending order of complexity. In Section 3 the natural number and decimal notation are introduced; in Section 4 the prime numbers are constructed and in Section 5 sorting is described.

One should notice that not only the definition of the concepts and algorithms is very short but also easy to grasp, and that the correctness is easily verifiable since there is no language that obscures their expression.

2. TWO-LEVEL GRAMMARS

A "vocabulary" is a set; its elements are termed "letters". A "word" over a vocabulary V is a mapping $[1:n] \to V$, for any $n \in \mathbb{N}_0$, and is thus a set of n ordered pairs $(1,v_i)$, for $i = 1,\ldots,n$, $v_i \in V$. Therefore, v_i is termed the "i-th letter" and n the "length" of the word. If n = 0, then the word is the empty set, also termed the "empty word". For any vocabulary V, V^* denotes the set of words over V, and V^+ the set of nonempty words over V. A "sentence" over a vocabulary V is a word over the vocabulary whose letters are the words over V; hence, V^{**} is the set of sentences over V.

A "rule" is an ordered pair (v,w) where v and w are owrds over certain vocabularies.

A "two-level grammar" VWG is an ordered sextuple $(V_m,V_o,V_t,R_m,R_h,w_s)$, where V_m, V_o, V_t are finite vocabularies, whose letters are termed "metaletters", "ortholetters" and "terminal letters" respectively, R_m and R_h are finite sets of rules, termed "metarules" and "hyperrules" respectively, and w_s is some word over V_o, termed the "start word". Let $V_h := V_m \cup V_o$. It is required that $V_m \cap V_o = \{\ \}$, $V_t \subset V_o^+$, $R_m \subset V_m \times V_h^*$, $R_h \subset V_h^+ \times V_h^{**}$, $w_s \in V_o^+$.

The grammar VWG "generates" a "language" L defined as follows:

Let $R_{mo} := V_m \times V_o^*$, $R_{oo} := V_o^+ \times V_0^{**}$, $R_{so} := \{w_s\} \times V_o^{**}$ and $R_{st} := \{w_s\} \times V_t^{**}$. A set R_m' identical with R_m and a set R_h' identical with R_h are introduced and then extended by arbitrarily often, if possible, applying the following extension, where at each application, of the three alternatives enclosed by "( )" and separated by "/", consistently either the first, or the second or the third must be chosen:

Extension: To $(R_m'/R_h'/R_h')$ a rule is added, obtained by replacing in a copy of some rule $(v,w) \in (R_m'\backslash R_{mo} \;/\; R_h'\backslash R_{oo} \;/\; R_h'\backslash R_{oo})$ and for some rule $(v',w') \in (R_m' \cap R_{mo} \;/\; R_m' \cap R_{mo} \;/\; R_h' \cap R_{oo})$, (some / each / some) occurrence of v' in (w/v and w/w) by w'.

Then, $L := \{w \mid (w_s,w) \in R_h' \cap R_{st}\}$.

In order to create a useful tool from two-level grammars, one must lay down some conventions about the choice of the different vocabularies and the way of writing the rules.

It has become a tradition to use as the letters of V_m specific words over the vocabulary of "metamarks", i.e., conventional capital letters and some other convenient marks. In [2], e.g., SOME and MODE 2 are letters of

V_m. Similarly, traditionally, the letters of V_o are specific words over the vocabulary of "orthomarks", i.e., conventional lower-case letters. In [2], e.g., long and real are letters of V_o.

Four other marks play a role, viz., colon, point, comma and semicolon.

Since V_m is a finite vocabulary its letters can, or rather should be, chosen, in such a way that a word over V_m cannot be misread as another word over V_m. In the simple grammars of this paper, each metaletter consists of a capital letter followed by zero or more apostrophs, so that this condition is fulfilled. It is more difficult to ensure that a sentence over V_o cannot be misread.

Marks are "written" one after the other in such a way that the order in which they have been written is clear, in this paper conventionally to the right of the mark lastly written before, or on the next line or on the next page, whatever this may mean.

A word is written when its writing starts by writing its fisrt letter, if it exists, and, when its i-th letter has been written by writing its (i+1)-th letter, if it exists.

A metarule (v,w) is written by writing v, then writing twice a colon, then writing w and then writing a point (cf. the rules M1 up to M4).

A hyperrule (v,w) is written by writing v, then writing a colon, then writing w and then writing a point. Writing w, however, poses the problem that a sentence over V_o might later on be misread. In natural languages this problem is overcome by separating the words of the sentence by blanks. Here, traditionally, one separates the words by writing a comma after a word has been written and before the next word of the sentence is going to be written (cf. the rules H1', H3, H9, H12, H20) what leaves open the use of blanks for display purposes inside the words.

The grammar mechanism as defined so far is, however, not yet complete. The terminal letters are words over V_o but this internal structure is of no relevance to the user. Therefore, V_t is mapped onto another vocabulary W_t, the set of "representations" of the terminal letters, which may be marks chosen by the user at his convenience as long as they differ from all marks mentioned above.

A terminal production of a VWG, i.e., an element of the language that it generates, is obtained by first constructing any such element by the process described above, and then replacing any terminal letter in that element by its representation and taken out the comma that follows it, if any.

A useful shorthand notation for rules is the following one: If two rules have the same left-hand side up to and including the colon or double colon, then they may be combined into one rule consisting of the first rule in which the point has been replaced by a semicolon, followed by the right-hand side of the second rule.

Thus L::a;b;c. stands for L::a. L::b. L::c.

Another useful convention stems from the fact that one frequently needs metarules differing only in the left-hand side, because one wants to circumvent the effect of the, utterly necessary, word "each" in the Extension. Therefore, by convention, it holds:

Let *M* stand for any element of V_m. Then any occurrence of *M'* in a VWG tacitly implies that the metarule *M'::M.* is an element of R_m.

Thus, the occurence of N" implies the rule N"::N'., which implies again the rule N'::N., so that N, N' and N" have the same terminal productions over V_o.

In the following sections some examples of two-level grammars are given. In all those grammars.

$V_m \subset$ {D,N,N',N",P,P',P",S,S',S"},

$V_o \subset$ {c,d,i,j,n,p,q,s},

$V_t \subset$ {cs,ds,dis,diis,diiis,diiiis,diiiiis,diiiiiis,diiiiiiis,
diiiiiiiis,diiiiiiiiis,es,ps}.

The representations of the terminal letters are highly suggestive for their intended function, viz.,

cs	,
ds	*0*
dis	*1*
diis	*2*
diiis	*3*
diiiis	*4*
diiiiis	*5*
diiiiiis	*6*
diiiiiiis	*7*
diiiiiiiis	*8*
diiiiiiiiis	*9*
es	=
ps	+

3. NATURAL NUMBERS, DECIMAL REPRESENTATION AND ADDITION.

In this section a two-level grammar concerning natural numbers, their decimal representation and their addition is discussed. Let the first rule of R_m be

N::;Ni. (M1)

which stands therefore, more precisely for the two metarules

N::. (M1.1)

N::Ni. (M1.2)

This set can be extended by virtue of the first alternative of the Extension by replacing the second occurrence of N in (M1.2) by the right-hand side of (M1.1), yielding

N::i. (M1.3)

and then again by replacing that occurrence of N by the right-hand side of (M1.3), yielding

N::ii. , (M1.4)

and so on. The terminal productions of N are therefore words of any length, possibly zero, all of whose letters are i. If one interprets this i as "the successor of", then iii is interpreted as "the successor of the successor of the successor of nothing". Therefore, M1 defines the concept "natural number"; it is actually equivalent to the first two Peano axioms whereas the other Peano axioms are automatically included by the operational character of the grammar.

If iii suggests a specific natural number, viz., three, then on a higher level N suggests any natural number. The two levels may be mixed; e.g., Ni suggests any positive natural number. This is immediately abstracted again by the second rule of R_m:

P::Ni. (M2)

which is logically not necessary but provides a convenient abbreviation.

Let the start word w_s of the grammar be n and let

n:jN. (H1)

be the first rule of R_h.

The second alternative of the Extension with the now extended R'_m permits the extension of R_h with the rules

n:j. (H1.1)

n:ji. (H1.2)

n:jii. (H1.3)

and so on. The letter j in H1 serves as a handle to the sequence of i's preventing the right-hand side of H1.1 to be empty which would cause difficulties later on. In fact, we are not through yet. Unlike in ordinary mathematics, it is not sufficient to define, by means of axioms, natural numbers, what we did in some sense with the rules given above, but they also should be represented by representations of terminal letters. In order to make our task somewhat realistic and not too easy, we choose to represent the natural numbers by their decimal representation. Let the second rule of R_m be

D::;i;ii;iii;iiii;iiiii;iiiiii;iiiiiii;iiiiiiii;iiiiiiiii. (M3)

which stands, more precisely for ten metarules. Just like N introduced the concept of the natural number, so does D introduce the concept of the decimal digit.

Moreover, we complete R_h with the three hyperrules

NjN'iiiiiiiiii : NijN'. (H2)

PjD : jP,jD. (H3)

jD: dDs (H4)

In order to see how this works let us use i^k as a shorthand notation for a word of length k all of whose letters are i. Consider the rule found by the extension of R_h:

$n:ji^{691}$.

Using the rules

$N::i^{0}$.

$N::i^{681}$.

and the second alternative of the Extension, H2 yields the rule

$ji^{691}:ij^{681}$.

Using

$N::i$.

$N::i^{671}$.

H2 yields the rule

$iji^{681}:i^{2}ji^{671}$.

and so on. The third alternative of the Extension of Section 1 now extends R_h' with

$n:ij^{681}$.

and then with

$n:i^{2}ji^{671}$.

and finally with

$n:i^{69}ji$.

Then the rules M2 and

$N::i^{68}$.

extend R_m' with

$P::i^{69}$.

which then with

$D::i$.

and H3 extend R_h' with

n:ji^{69},ji.

Proceeding similarly, the rule

n:ji^{6},ji^{9},ji.

is obtained which together with H4 then yields

n:diiiiiis,diiiiiiiiis,dis.

which after replacing the terminal letters by their representations and removing the separating comma's yields the result *691*.

Of course, even a context-free grammar, i.e., a two-level grammar with empty V_m and, hence, empty R_m and representations equal to the terminal symbols themselves could produce the same result.

E.g., the context-free grammar

n:*0*;p.

p:d;p*0*;pd.

d:*1*;*2*;*3*;*4*;*5*;*6*;*7*;*8*;*9*.

does the job. However, our two-level grammar does not only produce *691* but does it via the detour of the word constituted by 691 times the letter i generated by the metalevel of the grammar. This opens the door to arithmetic as is shown by changing the grammar into one which produces all elementary-school sums like

3+4 = 7 or *597+94 = 691*.

The only thing one has to do is to replace H1 by

n:jN,ps,jN',es,jNN'. (H1')

The addition is now actually performed on the metalevel and no context-free grammar can achieve this.

4. PRIME NUMBERS

We shall now give, as a more complicated and interesting example, a grammar producing any prime number in decimal notation.

R_m consists again of the metarules M1, M2 and M3.

R_h consists of the hyperrules H2, H3 and H4, together with

NjNPjP':NP'NPjP'. (H5)

PP"jPjP':jPjP'i. (H6)

PjPjP':jPijii. (H7)

jPjP:jP. (H8)

ws is jiijii.

Consider hypernotions of the form NjPjP'. The start word is of this form with N empty and P = P' = ii, and H5, H6 and H7 all transform a hypernotion of this form into another one of this form. Let us now interpret such a hypernotion as follows: The number P is tested for primality, P' is a candidate for a divisor, at least 2, or course, and N is some multiple of P'. To start with, N is 0 and P and P' are 2.

If both the candidate and its multiple are less than the number, then only H5 applies and the multiple is increased by the candidate. If, after doing so, the multiple is larger than the number, i.e., the candidate is not a divisor, then only H6 applies, the candidate is increased by 1 and the multiple is reset to 0. If, however, after doing so, the multiple equals the number, i.e., the candidate is a divisor, then only H7 is applicable, the number, which is considered as composite, is increased by 1, the candidate is reset to 2 and the multiple to 0. If, at last, the candidate equals the number and its multiple is 0, i.e. the number is prime, then both H5 and H8 apply. Application of H5 leads to application of H7, i.e., the fact that the number is prime is ignored and the next number is tested. Application of H8, however, accepts the number as a prime number and H2, H3 and H4 take care of its representation in decimal notation.

One should notice that the grammar, although extremely short, produces no blind alleys since, on the one hand, each hypernotion can be produced further and, on the other hand, in view of Euclid's theorem of the nonexistence of a largest prime number, there is no danger of ignoring H8 too often.

A typical terminal production of jiijii is *691*.

5. SORTING

As last example we treat the sorting of a sequence of natural numbers in ascending order according to the so-called Quick-Sort method of C.A.R. HOARE as modified by M.H. VAN EMDEN.

R_m consists of the metarules M1, M2 and M3 together with

S:::;SjN. (M4)

R_m consists of the hyperrules H2, H3 and H4 together with

q:pS,es,S. (H9)

p:. (H10)

pjN:jN. (H11)

pjNSjN':jN,cs,pSjN'. (H12)

jNPSjN:jNSjNP. (H13)

jNSjNN':jNpNSpNN'jNN'. (H14)

SpNN'jNS'pNN'N"S":SjNpNN'S'pNN'N"S". (H15)

SpNS'jNN'N"pNN'S":SpNS'pNN'jNN'N"S". (H16)

SpNjNPS'pNN'PS": SjNPpNPS'pNN'PS". (H17)

SpNS'jNN'pNN'PS":SpNS'pNN'jNN'S". (H18)

SpNPjNN'PP'S'jNpNN'PS":SjNpNPS'pNN'PjNN'PP'S". (H19)

SpNpNN'S':S,cs,S'. (H20)

w_s is q.

The metanotion S stands according to M4 for the abstract concept of the, possibly empty, sequence of natural numbers. Hyperrule H9 produces the start word q into a sequence produced by p, a separating symbol cs and the same sequence without the preceding p. The hyperrules H10, H11 and H12 produce the abstract sequence preceded by p into a concrete sequence of its terms separated by cs. The sequence without the preceding p is sorted and split by the hyperrules H12 up to H20.

First of all, if the sequence is empty or if it has only one element, then there is nothing to sort and split. If it has at least two elements then only H13 or H14 may apply, since these are the only rules whose left-hand side does not contain p. If the first element is larger than the last

element, then H13 interchanges these two elements. Now, anyhow, the first element is at most equal to the last element, H14 applies and introduces a lower bound pN and an upper bound pNN' which shall also serve as pointers. Now, the following property holds: All elements (actually the only element) to the left of the lower-bound pointer are at most (is actually) equal to the lower bound, all elements (actually the only element) to the right of the upper-bound pointer are at least (is actually) equal to the upper bound and the upper bound is at least equal to the lower bound. The hyperrules H15 up to H19 all preserve this property as is seen by inspection. Moreover, at least one of them, or otherwise H20 applies.

If the element to the (right of the lower / left of the upper)-bound pointer is not (larger/smaller) than that bound, then (H15/H16) applies and shifts that pointer one element to the (right/left).

If the element to the (right of the lower / left of the upper)-bound pointer is (larger/smaller) than that bound but not (larger/smaller) than the (upper/lower)-bound, then (H17/H18) applies, (increases/decreases) the (lower/upper)-bound to that element and shifts the pointer one element to the (right/left).

If, at last, the element to the right of the lower-bound pointer is larger than the upper bound and the element to the left of the upper-bound pointer is smaller then the lower-bound, then there are obviously at least two elements between the pointers and H19 applies, swaps the two elements and shifts the lower-bound pointer one element to the right and the upper-bound pointer one element to the left.

Therefore, each application of some hyperrule preserves the property and moreover, decreases the number of elements between the two pointers by one or two. After a finite number of steps, therefore, the two pointers come together and H20 applies and splits the sequence into a left subsequence, all of whose elements are at most equal to the lower-bound, and a right subsequence of all whose elements are at least equal to the upper bound and, hence, at least equal to the lower bound. These subsequences can, therefore, be sorted independently. Since each subsequence contains fewer elements than the sequence, after a finite number of steps each subsequence contains at most one element whereupon sorting and splitting are no longer necessary.

Finally, each element will be represented in decimal notation by means of H2, H3 and H4, as before.

A typical terminal production of q is

13,1,4,691,4 = 1,4,4,13,691.

REFERENCES

[1] WIJNGAARDEN, A. VAN, *Orthogonal design and description of a formal language,* Mathematical Centre, Amsterdam, MR76 (1966).

[2] WIJNGAARDEN, A. VAN, et al., eds., *Revised Report on the algorithmic language ALGOL 68*, Acta Informatica 5 (1975) 1-234.

LIST OF ADDRESSES OF AUTHORS

BOREL, A.
Mathematical Department
Institute of Advanced Studies
Princeton, N.J. 08450
U.S.A.

BRAAKSMA, B.L.J.
Mathematical Institute
University of Groningen
P.O. Box 800
9700 AV Groningen
The Netherlands

CRAMER, J.S.
Department of Econometrics
University of Amsterdam
Jodenbreestraat 23
1011 NH Amsterdam
The Netherlands

DALEN, D. van
Mathematical Institute
University of Utrecht
Budapestlaan 6
3584 CD Utrecht
The Netherlands

DILLER, J.
Institut für mathematische Logik und
Grundlagenforschung
Westfälische Wilhelms-Universität
Roxeler Strasze 64
44 Münster (Westf.)
West Germany

DUISTERMAAT, J.J.
Mathematical Institute
University of Utrecht
Budapestlaan 6
3584 CD Utrecht
The Netherlands

DIJKSTRA, E.W.
Department of Mathematics
Eindhoven University of Technology
P.O. Box 513
5600 MB Eindhoven
The Netherlands

EST, W.T. van
Mathematical Institute
University of Amsterdam
Roetersstraat 15
1018 WB Amsterdam
The Netherlands

FREUDENTHAL, H.
Frans Schubertstraat 44
3533 GW Utrecht
The Netherlands

HEMELRIJK, J.
Institute for Applied Mathematics
University of Amsterdam
Roetersstraat 15
1018 WB Amsterdam

Mathematical Centre
2e Boerhaavestraat 49
1091 AL Amsterdam
The Netherlands

HIGMAN, D.G.
Department of Mathematics
University of Michigan
Ann Arbor, Michigan 48104
U.S.A.

HORDIJK, A.
Institute for Applied Mathematics and Informatics
University of Leiden
Wassenaarseweg 80
2333 AL Leiden
The Netherlands

KAASHOEK, M.A.
Mathematical Department
Free University
de Boelelaan 1081
1081 HV Amsterdam
The Netherlands

KUIPER, N.H.
Institut des Hautes Etudes Scientifiques
35, Route de Chartres
91440 Bures-sur-Yvette
France

LENSTRA jr., H.W.
Mathematical Institute
University of Amsterdam
Roetersstraat 15
1018 WB Amsterdam
The Netherlands

LOOIJENGA, E.J.N.
Mathematical Institute
University of Nijmegen
Toernooiveld
Nijmegen
The Netherlands

NEUENSCHWANDER, E.
Mathematisches Institut
Universität Zürich
Freiestrasse 36
8032 Zürich
Switzerland

PELETIER, L.A.
Mathematical Institute
University of Leiden
Wassenaarseweg 80
2333 AL Leiden
The Netherlands

RUNNENBURG, J.Th.
Institute for Applied Mathematics
University of Amsterdam
Roetersstraat 15
1018 WB Amsterdam
The Netherlands

SEIDEL, J.J.
Department of Mathematics
Eindhoven University of Technology
P.O. Box 513
5600 MB Eindhoven
The Netherlands

SLUIS, A. van der
Mathematical Institute
University of Utrecht
Budapestlaan 6
3584 CD Utrecht
The Netherlands

SPIJKER, M.N.
Institute for Applied Mathematics and Informatics
University of Leiden
Wassenaarseweg 80
2333 AL Leiden
The Netherlands

STIGT, W.P. van
Roehampton Institute of Higher Education
London University
West Hill - London SW 15 3SN
United Kingdom

TAKENS, F.
Mathematical Institute
University of Groningen
P.O. Box 800
9700 AV Groningen
The Netherlands

THOMAS, G.E.F.
Mathematical Institute
University of Groningen
P.O. Box 800
9700 AV Groningen
The Netherlands

TIJDEMAN, R.
Mathematical Institute
University of Leiden
Wassenaarseweg 80
2333 AL Leiden
The Netherlands

WIJNGAARDEN, A. van
Mathematical Centre
2e Boerhaavestraat 49
1091 AL Amsterdam
The Netherlands

OTHER TITLES IN THE SERIES MATHEMATICAL CENTRE TRACTS

A leaflet containing an order-form and abstracts of all publications mentioned below is available at the Mathematisch Centrum, Tweede Boerhaavestraat 49, Amsterdam-1005, The Netherlands. Orders should be sent to the same address.

MCT 1 T. VAN DER WALT, *Fixed and almost fixed points*, 1963. ISBN 90 6196 002 9.

MCT 2 A.R. BLOEMENA, *Sampling from a graph*, 1964. ISBN 90 6196 003 7.

MCT 3 G. DE LEVE, *Generalized Markovian decision processes, part I: Model and method*, 1964. ISBN 90 6196 004 5.

MCT 4 G. DE LEVE, *Generalized Markovian decision processes, part II: Probabilistic background*, 1964. ISBN 90 6196 005 3.

MCT 5 G. DE LEVE, H.C. TIJMS & P.J. WEEDA, *Generalized Markovian decision processes, Applications*, 1970. ISBN 90 6196 051 7.

MCT 6 M.A. MAURICE, *Compact ordered spaces*, 1964. ISBN 90 6196 006 1.

MCT 7 W.R. VAN ZWET, *Convex transformations of random variables*, 1964. ISBN 90 6196 007 X.

MCT 8 J.A. ZONNEVELD, *Automatic numerical integration*, 1964. ISBN 90 6196 008 8.

MCT 9 P.C. BAAYEN, *Universal morphisms*, 1964. ISBN 90 6196 009 6.

MCT 10 E.M. DE JAGER, *Applications of distributions in mathematical physics*, 1964. ISBN 90 6196 010 X.

MCT 11 A.B. PAALMAN-DE MIRANDA, *Topological semigroups*, 1964. ISBN 90 6196 011 8.

MCT 12 J.A.TH.M. VAN BERCKEL, H. BRANDT CORSTIUS, R.J. MOKKEN & A. VAN WIJNGAARDEN, *Formal properties of newspaper Dutch*, 1965. ISBN 90 6196 013 4.

MCT 13 H.A. LAUWERIER, *Asymptotic expansions*, 1966, out of print; replaced by MCT 54 and 67.

MCT 14 H.A. LAUWERIER, *Calculus of variations in mathematical physics*, 1966. ISBN 90 6196 020 7.

MCT 15 R. DOORNBOS, *Slippage tests*, 1966. ISBN 90 6196 021 5.

MCT 16 J.W. DE BAKKER, *Formal definition of programming languages with an application to the definition of ALGOL 60*, 1967. ISBN 90 6196 022 3.

MCT 17 R.P. VAN DE RIET, *Formula manipulation in ALGOL 60, part 1*, 1968. ISBN 90 6196 025 8.

MCT 18 R.P. VAN DE RIET, *Formula manipulation in ALGOL 60, part 2*, 1968. ISBN 90 6196 038 X.

MCT 19 J. VAN DER SLOT, *Some properties related to compactness*, 1968. ISBN 90 6196 026 6.

MCT 20 P.J. VAN DER HOUWEN, *Finite difference methods for solving partial differential equations*, 1968. ISBN 90 6196 027 4.

MCT 21 E. WATTEL, *The compactness operator in set theory and topology*, 1968. ISBN 90 6196 028 2.

MCT 22 T.J. DEKKER, *ALGOL 60 procedures in numerical algebra, part 1*, 1968. ISBN 90 6196 029 0.

MCT 23 T.J. DEKKER & W. HOFFMANN, *ALGOL 60 procedures in numerical algebra, part 2*, 1968. ISBN 90 6196 030 4.

MCT 24 J.W. DE BAKKER, *Recursive procedures*, 1971. ISBN 90 6196 060 6.

MCT 25 E.R. PAERL, *Representations of the Lorentz group and projective geometry*, 1969. ISBN 90 6196 039 8.

MCT 26 EUROPEAN MEETING 1968, *Selected statistical papers, part I*, 1968. ISBN 90 6196 031 2.

MCT 27 EUROPEAN MEETING 1968, *Selected statistical papers, part II*, 1969. ISBN 90 6196 040 1.

MCT 28 J. OOSTERHOFF, *Combination of one-sided statistical tests*, 1969. ISBN 90 6196 041 x.

MCT 29 J. VERHOEFF, *Error detecting decimal codes*, 1969. ISBN 90 6196 042 8.

MCT 30 H. BRANDT CORSTIUS, *Excercises in computational linguistics*, 1970. ISBN 90 6196 052 5.

MCT 31 W. MOLENAAR, *Approximations to the Poisson, binomial and hypergeometric distribution functions*, 1970. ISBN 90 6196 053 3.

MCT 32 L. DE HAAN, *On regular variation and its application to the weak convergence of sample extremes*, 1970. ISBN 90 6196 054 1.

MCT 33 F.W. STEUTEL, *Preservation of infinite divisibility under mixing and related topics*, 1970. ISBN 90 6196 061 4.

MCT 34 I. JUHÁSZ, A. VERBEEK & N.S. KROONENBERG, *Cardinal functions in topology*, 1971. ISBN 90 6196 062 2.

MCT 35 M.H. VAN EMDEN, *An analysis of complexity*, 1971. ISBN 90 6196 063 0.

MCT 36 J. GRASMAN, *On the birth of boundary layers*, 1971. ISBN 90 6196 064 9.

MCT 37 J.W. DE BAKKER, G.A. BLAAUW, A.J.W. DUIJVESTIJN, E.W. DIJKSTRA, P.J. VAN DER HOUWEN, G.A.M. KAMSTEEG-KEMPER, F.E.J. KRUSEMAN ARETZ, W.L. VAN DER POEL, J.P. SCHAAP-KRUSEMAN, M.V. WILKES & G. ZOUTENDIJK, *MC-25 Informatica Symposium*, 1971. ISBN 90 6196 065 7.

MCT 38 W.A. VERLOREN VAN THEMAAT, *Automatic analysis of Dutch compound words*, 1971. ISBN 90 6196 073 8.

MCT 39 H. BAVINCK, *Jacobi series and approximation*, 1972. ISBN 90 6196 074 6.

MCT 40 H.C. TIJMS, *Analysis of (s,S) inventory models*, 1972. ISBN 90 6196 075 4.

MCT 41 A. VERBEEK, *Superextensions of topological spaces*, 1972. ISBN 90 6196 076 2.

MCT 42 W. VERVAAT, *Success epochs in Bernoulli trials (with applications in number theory)*, 1972. ISBN 90 6196 077 0.

MCT 43 F.H. RUYMGAART, *Asymptotic theory of rank tests for independence*, 1973. ISBN 90 6196 081 9.

MCT 44 H. BART, *Meromorphic operator valued functions*, 1973. ISBN 90 6196 082 7.

MCT 45 A.A. BALKEMA, *Monotone transformations and limit laws*, 1973. ISBN 90 6196 083 5.

MCT 46 R.P. VAN DE RIET, *ABC ALGOL, A portable language for formula manipulation systems, part 1: The language*, 1973. ISBN 90 6196 084 3.

MCT 47 R.P. VAN DE RIET, *ABC ALGOL, A portable language for formula manipulation systems, part 2: The compiler*, 1973. ISBN 90 6196 085 1.

MCT 48 F.E.J. KRUSEMAN ARETZ, P.J.W. TEN HAGEN & H.L. OUDSHOORN, *An ALGOL 60 compiler in ALGOL 60, Text of the MC-compiler for the EL-X8*, 1973. ISBN 90 6196 086 x.

MCT 49 H. KOK, *Connected orderable spaces*, 1974. ISBN 90 6196 088 6.

MCT 50 A. VAN WIJNGAARDEN, B.J. MAILLOUX, J.E.L. PECK, C.H.A. KOSTER, M. SINTZOFF, C.H. LINDSEY, L.G.L.T. MEERTENS & R.G. FISKER (Eds), *Revised report on the algorithmic language ALGOL 68*, 1976. ISBN 90 6196 089 4.

MCT 51 A. HORDIJK, *Dynamic programming and Markov potential theory*, 1974. ISBN 90 6196 095 9.

MCT 52 P.C. BAAYEN (ed.), *Topological structures*, 1974. ISBN 90 6196 096 7.

MCT 53 M.J. FABER, *Metrizability in generalized ordered spaces*, 1974. ISBN 90 6196 097 5.

MCT 54 H.A. LAUWERIER, *Asymptotic analysis, part 1*, 1974. ISBN 90 6196 098 3.

MCT 55 M. HALL JR. & J.H. VAN LINT (Eds), *Combinatorics, part 1: Theory of designs, finite geometry and coding theory*, 1974. ISBN 90 6196 099 1.

MCT 56 M. HALL JR. & J.H. VAN LINT (Eds), *Combinatorics, part 2: graph theory, foundations, partitions and combinatorial geometry*, 1974. ISBN 90 6196 100 9.

MCT 57 M. HALL JR. & J.H. VAN LINT (Eds), *Combinatorics, part 3: Combinatorial group theory*, 1974. ISBN 90 6196 101 7.

MCT 58 W. ALBERS, *Asymptotic expansions and the deficiency concept in statistics*, 1975. ISBN 90 6196 102 5.

MCT 59 J.L. MIJNHEER, *Sample path properties of stable processes*, 1975. ISBN 90 6196 107 6.

MCT 60 F. GÖBEL, *Queueing models involving buffers*, 1975. ISBN 90 6196 108 4.

* MCT 61 P. VAN EMDE BOAS, *Abstract resource-bound classes, part 1*. ISBN 90 6196 109 2.

* MCT 62 P. VAN EMDE BOAS, *Abstract resource-bound classes, part 2*. ISBN 90 6196 110 6.

MCT 63 J.W. DE BAKKER (ed.), *Foundations of computer science*, 1975. ISBN 90 6196 111 4.

MCT 64 W.J. DE SCHIPPER, *Symmetric closed categories*, 1975. ISBN 90 6196 112 2.

MCT 65 J. DE VRIES, *Topological transformation groups 1 A categorical approach*, 1975. ISBN 90 6196 113 0.

MCT 66 H.G.J. PIJLS, *Locally convex algebras in spectral theory and eigenfunction expansions*, 1976. ISBN 90 6196 114 9.

* MCT 67 H.A. LAUWERIER, *Asymptotic analysis, part 2.*
ISBN 90 6196 119 X.

MCT 68 P.P.N. DE GROEN, *Singularly perturbed differential operators of second order*, 1976. ISBN 90 6196 120 3.

MCT 69 J.K. LENSTRA, *Sequencing by enumerative methods*, 1977.
ISBN 90 6196 125 4.

MCT 70 W.P. DE ROEVER JR., *Recursive program schemes: semantics and proof theory*, 1976. ISBN 90 6196 127 0.

MCT 71 J.A.E.E. VAN NUNEN, *Contracting Markov decision processes*, 1976.
ISBN 90 6196 129 7.

MCT 72 J.K.M. JANSEN, *Simple periodic and nonperiodic Lamé functions and their applications in the theory of conical waveguides*, 1977.
ISBN 90 6196 130 0.

MCT 73 D.M.R. LEIVANT, *Absoluteness of intuitionistic logic*, 1979.
ISBN 90 6196 122 x.

MCT 74 H.J.J. TE RIELE, *A theoretical and computational study of generalized aliquot sequences*, 1976. ISBN 90 6196 131 9.

MCT 75 A.E. BROUWER, *Treelike spaces and related connected topological spaces*, 1977. ISBN 90 6196 132 7.

MCT 76 M. REM, *Associons and the closure statement*, 1976. ISBN 90 6196 135 1.

MCT 77 W.C.M. KALLENBERG, *Asymptotic optimality of likelihood ratio tests in exponential families*, 1977 ISBN 90 6196 134 3.

MCT 78 E. DE JONGE, A.C.M. VAN ROOIJ, *Introduction to Riesz spaces*, 1977.
ISBN 90 6196 133 5.

MCT 79 M.C.A. VAN ZUIJLEN, *Empirical distributions and rankstatistics*, 1977.
ISBN 90 6196 145 9.

MCT 80 P.W. HEMKER, *A numerical study of stiff two-point boundary problems*, 1977. ISBN 90 6196 146 7.

MCT 81 K.R. APT & J.W. DE BAKKER (eds), *Foundations of computer science II*, part I, 1976. ISBN 90 6196 140 8.

MCT 82 K.R. APT & J.W. DE BAKKER (eds), *Foundations of computer science II*, part II, 1976. ISBN 90 6196 141 6.

* MCT 83 L.S. VAN BENTEM JUTTING, *Checking Landau's "Grundlagen" in the AUTOMATH system*, ISBN 90 6196 147 5.

MCT 84 H.L.L. BUSARD, *The translation of the elements of Euclid from the Arabic into Latin by Hermann of Carinthia (?) books vii-xii*, 1977.
ISBN 90 6196 148 3.

MCT 85 J. VAN MILL, *Supercompactness and Wallman spaces*, 1977.
ISBN 90 6196 151 3.

MCT 86 S.G. VAN DER MEULEN & M. VELDHORST, *Torrix I*, 1978.
ISBN 90 6196 152 1.

* MCT 87 S.G. VAN DER MEULEN & M. VELDHORST, *Torrix II*,
ISBN 90 6196 153 x.

MCT 88 A. SCHRIJVER, *Matroids and linking systems*, 1977.
ISBN 90 6196 154 8.

MCT 89 J.W. DE ROEVER, *Complex Fourier transformation and analytic functionals with unbounded carriers*, 1978. ISBN 90 6196 155 6.

* MCT 90 L.P.J. GROENEWEGEN, *Characterization of optimal strategies in dynamic games*, . ISBN 90 6196 156 4.

* MCT 91 J.M. GEYSEL, *Transcendence in fields of positive characteristic*, . ISBN 90 6196 157 2.

* MCT 92 P.J. WEEDA, *Finite generalized Markov programming*, . ISBN 90 6196 158 0.

MCT 93 H.C. TIJMS (ed.) & J. WESSELS (ed.), *Markov decision theory*, 1977. ISBN 90 6196 160 2.

MCT 94 A. BIJLSMA, *Simultaneous approximations in transcendental number theory*, 1978 . ISBN 90 6196 162 9.

MCT 95 K.M. VAN HEE, *Bayesian control of Markov chains*, 1978. ISBN 90 6196 163 7.

* MCT 96 P.M.B. VITÁNYI, *Lindenmayer systems: structure, languages, and growth functions*, 1978 . ISBN 90 6196 164 5.

* MCT 97 A. FEDERGRUEN, *Markovian control problems; functional equations and algorithms*, 1978 . ISBN 90 6196 165 3.

MCT 98 R. GEEL, *Singular perturbations of hyperbolic type*, 1978. ISBN 90 6196 166 1

MCT 99 J.K. LENSTRA, A.H.G. RINNOOY KAN & P. VAN EMDE BOAS, *Interfaces between computer science and operations research*, 1978. ISBN 90 6196 170 X.

MCT 100 P.C. BAAYEN, D. VAN DULST & J. OOSTERHOFF (Eds), *Proceedings bicentennial congress of the Wiskundig Genootschap, part 1*, 1979. ISBN 90 6196 168 8.

MCT 101 P.C. BAAYEN, D. VAN DULST & J. OOSTERHOFF (Eds), *Proceedings bicentennial congress of the Wiskundig Genootschap, part 2*, 1979. ISBN 90 9196 169 6.

MCT 102 D. VAN DULST, *Reflexive and superreflexive Banach spaces*, 1978. ISBN 90 6196 171 8.

MCT 103 K. VAN HARN, *Classifying infinitely divisible distributions by functional equations*, 1978 . ISBN 90 6196 172 6.

* MCT 104 J.M. VAN WOUWE, *Go-spaces and generalizations of metrizability*, . ISBN 90 6196 173 4.

* MCT 105 R. HELMERS, *Edgeworth expansions for linear combinations of order statistics*, . ISBN 90 6196 174 2.

AN ASTERISK BEFORE THE NUMBER MEANS "TO APPEAR"