PROCEEDINGS

OF THE

INTERNATIONAL CONGRESS

OF

MATHEMATICIANS

Cambridge, Massachusetts, U. S. A.

August 30—September 6, 1950

PUBLISHED BY THE

AMERICAN MATHEMATICAL SOCIETY

80 Waterman Street, Providence, R. I.

1952

PROCEEDINGS OF THE
INTERNATIONAL CONGRESS OF MATHEMATICIANS

Cambridge, Massachusetts, U. S. A.

1950

Under the Auspices of the

AMERICAN MATHEMATICAL SOCIETY

VOLUME II

CONFERENCE IN ALGEBRA

CONFERENCE IN ANALYSIS

CONFERENCE IN APPLIED MATHEMATICS

CONFERENCE IN TOPOLOGY

INTERNATIONAL CONGRESS

OF

MATHEMATICIANS

Cambridge, Massachusetts, U. S. A.

1950

CONFERENCE IN ALGEBRA

Committee

A. A. Albert (*Chairman*)

Richard Brauer

Nathan Jacobson

Saunders MacLane

Oscar Zariski

The central theme for the Conference in Algebra was the Theory of Rings. Because of the complexities of programming, the session on Algebraic Geometry took place as the first session of the Conference although that session, consisting of two stated addresses, was logically the fourth and final session of the Conference. In the first session, on Groups and Universal Algebra, modern aspects of lattice theory and group theory were presented. In the second session, on the Structure Theory of Rings and Algebras, recent results on the structure and representation of various types of rings and algebras were given, and in the third session, on Arithmetic Algebra, the arithmetic aspects of ring and field theory were discussed. The fourth session on Algebraic Geometry, may then be regarded as a continuation of the session on Arithmetic Algebra. The Committee on the Conference in Algebra regrets that it was not possible for H. Zassenhaus and M. Deuring to attend and to present their papers which were listed in the program.

A. A. ALBERT

GROUPS AND UNIVERSAL ALGEBRA

SOME PROBLEMS OF LATTICE THEORY

GARRETT BIRKHOFF

I feel no need to stress the importance of lattice theory here. Instead, I should like to discuss some of its unsolved problems, with special reference to the list of 111 compiled in [1].

Such a large list is naturally uneven; three problems on it (nos. 24, 82, 93) even have trival solutions. I am glad to report that interesting solutions have been given to nine others in the last two years: no. 11 by Nakayama, no. 20 by Arnold, nos. 32 and 36 by M. Hall, no. 34 by F. B. Thompson, no. 49 by Dilworth, nos. 64–65 by Croisot, Sholander, and Vassiliou, and no. 101 by Iwasawa. *Added in proof.* Solutions of nos. 1, 31, 33, 35, 39, 44, and 46–48 have now been reported.

At this rate, about half the others should be solved in the next decade. But there are two problems which have already resisted so much effort, that they may well defy solution for much longer.

First, is every finite lattice isomorphic with a sublattice of the lattice of all partitions of a suitable *finite* set (no. 48)? Whitman has shown that every lattice is isomorphic with a sublattice of *some* lattice of partitions. But, just as all finite subfields are commutative, there may be some identity true of partitions of finite sets, and not true of partitions in general. Although I doubt this, a definitive answer should stimulate new developments in combinatory analysis.

Second, a solution of the decision problem for the free modular lattice generated by four elements—or even by $\mathbf{2} + \mathbf{1} + \mathbf{1}$,—would be most important (nos. 28, 29). The nature of the lattice of subspaces, even in three-dimensional vector space, generated by four "general" subspaces a, b, c, d is not now known —not even assuming $a \cap b = c \cap d = 0$, $a \cup b = c \cup d = I$. The general problem may have no solution in finite terms, but progress on it should throw light on representation theory, as well as being interesting for its own sake.

Various other lattice-theoretic problems seem to require radically new ideas, although their difficulty is harder to appraise as they have not yet been seriously attacked. Can every proper sublattice of any lattice be extended to a *maximal* proper sublattice (no. 18)? Does every infinite Boolean algebra admit a proper automorphism (no. 74)? How can one tell whether the congruence relations on a given lattice form a Boolean algebra (no. 72)? Can one enumerate all lattices L which are highly *symmetric*, in the sense that their groups of automorphisms admit (say) at most six sets of transitivity? Many such lattices should be isomorphic with classical configurations of points, lines, and planes (or of curves and surfaces), and a study of these isomorphisms should give a fresh approach to configurational geometry.

A similar fresh stimulus to group theory has been provided by studies of the lattice $L(G)$ of all subgroups of a finite group G. Thus Dedekind's enumeration of all groups whose subgroups were permutable has been extended by Iwasawa, A. W. Jones, and M. Hall, who have enumerated all groups for which $L(G)$ is modular or semi-modular. Again, it is known that if G is Abelian, then $L(G)$ is self-dual, but it is not known what non-Abelian finite groups have this property (no. 37). Neither is it known when $L(G)$ is complemented (no. 38). In the hypercentral case, $L(G)$ is complemented if and only if G is the direct product of cyclic groups of prime order—for if G is not Abelian, its center can have no complement, and if Abelian, the subgroup of all elements of square-free order can have no complement. However, the general case of solvable G is less easy, while the case of finite simple G promises to be really difficult. Straightforward enumeration of subgroups reveals that if G is the alternating group of degree five, $L(G)$ is complemented but not relatively complemented. This suggests trying to prove that if $L(G)$ is relatively complemented, G cannot be simple. (All Sylow subgroups of G must then be elementary.)

I should like to raise the analogous questions, for the lattice $L(A)$ of all subalgebras of a Lie algebra A. Every element of $L(A)$ is a join of "points". Any $(n+1)$-dimensional complex Lie algebra A, such that $L(A)$ is semi-modular, can be shown to belong to one of the two types

$$[a, b_i] = \lambda b_i \quad \text{and} \quad [b_i, b_j] = 0 \qquad (i, j = 1, \cdots, n),$$

where $\lambda = 0$ or 1, for suitable basis-elements $a, b_1, \cdots, b_n$. In these cases, $L(A)$ is a projective geometry. Moreover, any other A contains a three-dimensional subalgebra S such that $L(S)$, which is an interval sublattice of $L(A)$, is not semi-modular. On the other hand, if A is a solvable complex Lie algebra, it seems that $L(A)$ must be of length n and satisfy the Jordan-Dedekind chain condition, though I have not checked this carefully. But what about other complex Lie algebras, and what about real and rational Lie algebras?

It would also be interesting to know when $L(A)$ was complemented or relatively complemented. A theorem of Knebelman shows that if A is semi-simple, the I of $L(A)$ is the join of two points. I suspect that, by an extension of this reasoning, one can show that $L(A)$ is complemented, and even that every element not 0 or I has a point-complement. On the other hand, I suspect that $L(A)$ is rarely relatively complemented.

Although lattice theory suggests the preceding problems, it does not provide the technique for solving them. I shall now turn to possible *applications* of lattice-theoretic techniques.

The theory of modular lattices has already been successfully applied to give the structure theory of groups (and loops) with operators. One might hope that the decomposition theory of multiplicative and additive (Dedekind) ideals could be derived by using similarly the theory of lattice-ordered semigroups. But in spite of encouraging recent progress by Lorenzen and Dubreil (based

on Artin's earlier ideas), much remains to be done before the derivation can be considered fully satisfactory.

The algebra of relations suggests other interesting problems involving lattice-ordered semigroups. Thus, no set of postulates for the algebra of binary relations is known, fully consistent with the spirit of modern algebra (no. 94).

Other interesting unsolved problems concern lattice-ordered groups. Although all finite-dimensional vector lattices are known ([1, p. 240]), the theory of Lie l-groups is still fragmentary. Thus it is not even known which abstract Lie groups are group-isomorphic with l-groups (see nos. 98, 100, 106). In attacking this problem, it may be useful to remark that the centralizer of any subgroup of an l-group is an l-subgroup. More generally, nothing is known about "analytic", locally Euclidean lattices. Neither is it known whether every vector lattice is a homomorphic image of a subdirect union of replicas of the ordered additive group of rational numbers (no. 107).

Logically related is the following question. Tarski has shown that any complete, completely distributive Boolean algebra is isomorphic with the field of all subsets of a suitable set. One can also show that any closed sublattice of a direct union of complete chains is a complete, completely distributive lattice. The question is (no. 69), are there any other complete, completely distributive lattices? This is only one of many interesting problems (nos. 25, 57, 58, 76, 79, 104) involving *infinitary* operations.

But I do not wish to emphasize only the *difficult* problems of lattice theory. The hardest problems may well be sterile. Thus, in lattice theory, the problems least likely to be solved in the next decade are those inherited from set theory (nos. 3, 4, 12, 15–17, 19, 86). But, without radical new ideas, their study is apt to be sterile. The most important contributions to mathematics often consist in the discovery and clear formulation of simple results. Carathéodory's abstract measure theory is an excellent example of such a discovery.

Perhaps the most fruitful problem of lattice theory is to develop a consequential general theory of *lattice-ordered rings*. At present, such a theory exists only for the ring of continuous real-valued functions on a compact space. Using fragments of a general theory, I have already [2] obtained a decomposition theory for averaging operators. I have also satisfied myself that the theory is full of interesting problems.

In particular, it is not known whether there exists any "natural" simple ordering of the ring $\mathfrak{R}$ of all real, single-valued functions of one real variable (nos. 2, 14). This problem has many variants. For example, does there exist a simple ordering which makes (a) $\mathfrak{R}$ into an ordered group, under the addition of functions, (b) any order-automorphism α of the real field R induce one on $\mathfrak{R}$, through $f(x) \to \alpha f(\alpha x)$, (c) $f(x) \geqq 0$ for all real x imply $f \geqq 0$ in $\mathfrak{R}$, (d) $f \geqq 0$ and $g \geqq 0$ imply $fg \geqq 0$? If so, what additional properties of the substitution operation $f(g(x))$ relative to order can also be postulated compatibly? If not, what combinations of (a)–(d) are compatible? What if R is replaced by the rational field, or by the domain of integers?

As an example of the type of simple result which seems to me worth stating, I shall outline a new formulation of Duhamel's principle. This formulation is applicable to both Riemann and Lebesgue integrals, if one uses the appropriate Boolean algebra A of sets, and the appropriate directed set of partitions of the unit I of A.

To formulate Duhamel's principle, define a *measure estimate* as a bounded real-valued function $m[x]$ on A, such that

$$\text{(1)} \qquad \text{Lim}_{\pi \downarrow} \sum m[a \cap \Delta x_i] = \int_a dm$$

exists for all $a \in A$. Here the Δx_i are the parts into which π divides I. The limiting "integral" so defined is necessarily additive. Two measure estimates which define the same limiting integral may be called *equivalent*; this corresponds to the usual notion of "equivalent infinitesimals".

Next, define a multiple-valued function $f(a)$ to be *consistent* if and only if $a \leqq b$ in A implies that the set of $f(a)$ is contained in the set of $f(b)$; this will clearly be the case if A represents a field of sets and $f(a)$ is the set of values assumed by a given function on the set a. Define $f(a)$ to be *measurable* if and only if, given $\epsilon > 0$, there exists a partition of I into elements Δx_i, on each of which the set of $f(\Delta x_i)$ has diameter at most ϵ.

The following results are then easily proved. If m is any measure estimate, and if $f(a)$ is consistent, bounded, and measurable, then $m^*[a] = f(a)m[a]$ defines a measure estimate, and all choices of $f(a)$ give equivalent measure estimates. Further, the product of any two consistent, bounded, measurable functions f and g has the same properties, and we have:

$$\text{(2)} \qquad \text{If } dm^* = \int f\, dm, \quad \text{then} \quad \int g\, dm^* = \int (gf)\, dm.$$

This is Duhamel's principle.

1. G. Birkhoff, *Lattice theory*, Rev. ed., Amer. Math. Soc. Colloquium Publications vol. 25, New York, 1948.
2. G. Birkhoff, *Moyennes des fonctions bornées*, Procédé d'un Colloque d'Algèbre, Paris, 1950.

Harvard University,
Cambridge, Mass., U. S. A.

COHOMOLOGY THEORY OF ABELIAN GROUPS

SAUNDERS MACLANE[1]

This note will present certain algebraic results obtained by Samuel Eilenberg and the author in a study of the relations between homotopy and homology groups of a topological space.[2] These results yield a homology theory for any abelian group Π, in which the low dimensional homology and cohomology groups of Π correspond to familiar constructions on Π. They depend upon the application of the methods of algebraic topology to algebraic systems. Some of the topological aspects of these constructions are presented by Eilenberg in another note in these Proceedings.

An abstract cell complex K is a sequence

$$(1) \qquad C_1 \xleftarrow{\partial_2} C_2 \xleftarrow{\partial_3} C_3 \cdots C_{n-1} \xleftarrow{\partial_n} C_n \cdots$$

of abelian groups C_n and homomorphisms ∂_n such that the composition $\partial_n\partial_{n+1}: C_{n+1} \to C_{n-1}$ of two successive homomorphisms is the zero homomorphism; furthermore each group C_n is a free abelian group with a specified set of free generators σ, called the n-dimensional cells of K. The group Z_n of n-dimensional cycles of K is the kernel of ∂_n in C_n, for $n > 1$, and is C_1 for $n = 1$. The group B_n of n-dimensional boundaries is the image of ∂_{n+1} in C_n. Since $\partial_n\partial_{n+1} = 0$, $B_n \subset Z_n \subset C_n$. The n-dimensional integral homology group $H_n = H_n(K)$ is defined as Z_n/B_n.

The cohomology groups of the complex K may be defined for any abelian "coefficient group" G. The group $C^n(K; G)$ of n-dimensional cochains of K is the group of all homomorphisms $f: C_n \to G$, or equivalently the group of all functions f on the n-cells of K to G. The coboundary $\delta_n f$ is an $(n + 1)$-cochain defined as the composite homomorphism $f\partial_{n+1}: C_{n+1} \to G$. These definitions yield a sequence of groups and homomorphisms

$$C^1(K; G) \xrightarrow{\delta_1} C^2(K; G) \xrightarrow{\delta_2} C^3(K; G) \longrightarrow \cdots$$

with $\delta_{n+1}\delta_n = 0$. As before, one defines the cohomology group $H^n(K, G)$ as Z^n/B^n, where Z^n is the kernel of δ_n, B^n the image of δ_{n-1} for $n > 1$, and $B^1 = 0$.

Any group Q (not necessarily abelian) has a standard homology theory which is the homology theory of the cell complex $A^0(Q)$ constructed as follows. The n-cells of $A^0(Q)$ are all n-tuples $[x_1, \cdots, x_n]$ of elements x_i of Q, and the boundary homomorphisms ∂_n (we omit the subscript n) are obtained by setting

$$(2) \qquad \partial[x, y] = [y] - [xy] + [x],$$

$$(3) \qquad \partial[x, y, z] = [y, z] - [xy, z] + [x, yz] - [x, y],$$

[1] These investigations were started while the author held a John Simon Guggenheim Memorial Fellowship.

[2] S. Eilenberg and S. MacLane, *Cohomology theory of abelian groups and homotopy theory* I *and* II, Proc. Nat. Acad. Sci. U. S. A. vol. 36 (1950) pp. 443–447, 657–663.

and, more generally,

$$\partial[x_1, \cdots, x_n] = [x_2, \cdots, x_n] + \sum_{i=1}^{n-1} (-1)^i [x_1, \cdots, x_{i-1}, x_i x_{i+1}, x_{i+2}, \cdots, x_n] + (-1)^n [x_1, \cdots, x_{n-1}], \tag{4}$$

in agreement with (2) and (3) for $n = 2$ and $n = 3$. The verification that $\partial\partial = 0$ depends only on the associative law for Q.

The groups of the complex $A^0(Q)$ are known as the homology and cohomology groups of Q. For example, the group Z_1 of one-dimensional cycles is the free abelian group with generators $[x]$, for $x \in Q$. If $[Q, Q]$ is the commutator group of Q, the homomorphism of Z_1 into $Q/[Q, Q]$ given by mapping each generator $[x]$ into the coset $x[Q, Q]$ has as kernel the group B_1 of one-dimensional boundaries $[y] - [xy] + [x]$. Hence the isomorphism

$$H_1(A^0(Q)) \cong Q/[Q, Q], \qquad \text{under } [x] \to x[Q, Q]. \tag{5}$$

A one-dimensional cochain is a function $f(x)$ defined on the 1-cells $[x]$ with values in G. It is a cocycle if $\delta f = 0$; by (2) this means that $f(y) - f(xy) + f(x) = 0$; i.e., that f is a homomorphism. Hence

$$H^1(A^0(Q), G) \cong \text{Hom } (Q, G), \tag{6}$$

the group of all homomorphisms of Q into G. Similarly, a two-dimensional cocycle f is a function $f(x, y)$ on 2-cells $[x, y]$ with values in G which satisfies, according to (3), the functional equation

$$f(y, z) + f(x, yz) = f(xy, z) + f(x, y). \tag{7}$$

Any such function is a "factor set" of Q in G. Each factor set determines a central group extension E of G by Q; specifically, E is the group of all pairs (x, g), for $x \in X$, $g \in G$, with the composition

$$(x, g)(y, h) = (xy, ghf(x, y)) \tag{8}$$

and the homomorphism $(x, g) \to x$ onto Q. The equation (7) insures that the product is associative. This extension E is equivalent to the direct product extension $Q \times G$ if and only if the cocycle f is a coboundary. In this fashion one may prove that

$$H^2(A^0(Q), G) \cong \text{Extcent } (Q, G),$$

where "Extcent" denotes the group of all central extensions of G by Q.

Instead of using explicit boundary formulas such as (4) we may characterize the cohomology groups of the complex $A^0(Q)$ by certain formal properties, using the special case when Q is the free group F with a fixed denumerable set of generators $g_1, g_2, \cdots$. In the complex $A^0(F)$ a cell $[x_1, \cdots, x_n]$ is called *generic* if each x_i is a product of zero or more generators, such that any one generator

g_j appears in at most one of these products x_i. Inspection of the boundary formula (4) shows that the boundary of a generic cell is a linear combination of generic cells. Hence the generic cells alone span a subcomplex $A^0(F^*)$ of $A^0(F)$. This "generic subcomplex" has the homology groups

$$H_n(A^0(F^*)) = 0, \qquad n > 1. \tag{9}$$

This property, together with (5), can be used to characterize the homology and cohomology groups of any group Q without reference to the specific complex $A^0(Q)$. It gives implicitly a rule for the construction of suitable complexes like A^0: given the cells through dimension n, enough cells in dimension $n + 1$ must be added to make every n-dimensional generic cycle a boundary.

There are several indications that the homology theory appropriate to an abelian group Π will not be given by the complex $A^0(\Pi)$. In the appropriate complex, the proof that $\partial\partial = 0$ should use both the associative and commutative laws valid in Π. In dimension 2, the cohomology groups of Π should correspond to extensions of G by Π which are abelian. For Q abelian, the extension E described in (8) will be abelian if the factor set f satisfies the additional condition

$$f(x, y) = f(y, x); \tag{10}$$

this indicates that two-dimensional "abelian" cocycles f should satisfy both (7) and (10).

We thus propose to construct complexes $A(\Pi)$ for additive abelian groups Π in such fashion that the generic subcomplex $A^0(F_a^*)$ for a free abelian group F_a will have vanishing higher homology groups, as in (9). The complex $A^0(\Pi)$ itself does not enjoy this property; indeed, if g and h are distinct generators of F_a, the 2-chain $[g, h] - [h, g]$ has boundary zero, hence is a generic cycle but not a generic boundary. We therefore adjoin to $A^0(\Pi)$ a new 3-cell $[x \mid y]$ with boundary

$$\partial[x \mid y] = [x, y] - [y, x]. \tag{11}$$

After this adjunction, the generic cycle $[g, h] - [h, g]$ becomes a boundary and the 2-dimensional cocycles $f(x, y)$ must satisfy not only the "associativity" condition (7), but also the "commutativity" condition (10). In dimension 4, we adjoin two more types of four cells $[x, y \mid z]$ and $[x \mid y, z]$, with boundaries

$$\partial[x, y \mid z] = [y \mid z] - [x + y \mid z] + [x \mid z] - [x, y, z] + [x, z, y] - [z, x, y], \tag{12}$$

$$\partial[x \mid y, z] = [x \mid z] - [x \mid y + z] + [x \mid z] + [x, y, z] - [y, x, z] + [y, z, x]. \tag{13}$$

Indeed, the expressions on the right in (12) and (13) would otherwise be non-bounding generic 3-cycles when x, y, z are distinct generators of F_a.

These formulas may be written more conveniently if we define a "shuffle" of m letters $x_1, \cdots, x_m$ through n letters $y_1, \cdots, y_n$ to be any list of these $m + n$ letters in an order which preserves both the order of the x's alone and that of

the y's alone. The sign of the shuffle is the sign of the permutation required to bring the shuffled letters back to the standard order $x_1, \cdots, x_m, y_1, \cdots, y_n$. Finally, the "star" product $[x_1, \cdots, x_m]*[y_1, \cdots, y_n]$ is defined as the signed sum of all shuffles of the letters x through the letters y. The boundary formula (12) becomes, in this notation,

$$\partial[x, y \mid z] = [\partial(x, y) \mid z] - (x, y)* z,$$

with similar expressions for (13).

Generalizing the formulas (12) and (13), we construct the complex $A^1(\Pi)$ in which the cells are symbols $\sigma = [\alpha_1 \mid \cdots \mid \alpha_p]$, with each α_i a cell of $A^0(\Pi)$. The dimension of σ is $p - 1$ plus the sum of the dimensions of the α_i, and the boundary of σ is

$$\begin{aligned}\partial\sigma = \sum_{i=1}^{p} (-1)^{\epsilon_i - 1} [\alpha_1 \mid \cdots \mid \partial\alpha_i \mid \cdots \alpha_p] \\ + \sum_{i=1}^{p-1} (-1)^{\epsilon_i} [\alpha_1 \mid \cdots \mid \alpha_i * \alpha_{i+1} \mid \cdots \mid \alpha_p],\end{aligned} \tag{14}$$

where $\epsilon_i = 1 + \dim [\alpha_1 \mid \cdots \mid \alpha_i]$. The proof that $\partial\partial = 0$ uses both the associative and the commutative laws in Π.

In the so constructed complexes $A^1(\Pi)$ all the generic cycles of $A^0(\Pi)$ become boundaries, but there are new generic cycles, such as the cycle $[g \mid h] + [h \mid g]$ of dimension 3. To kill this cycle, we add a new 4-cell $[x \mid\mid y]$ with boundary

$$[x \mid\mid y] = -[x \mid y] - [y \mid x]. \tag{15}$$

This is the first step in the construction of a new complex $A^2(\Pi)$ which has cells $[\sigma_1 \mid\mid \cdots \mid\mid \sigma_q]$, with the σ_i cells of $A^1(\Pi)$, and a boundary formula resembling (14) (but with a new shuffling operation which shuffles only the "blocks" α_j of each σ_i). The same construction[2] is then repeated inductively to give complexes $A^k(\Pi)$. The final complex $A(\Pi)$ is the union of all the complexes $A^k(\Pi)$; for a fixed dimension q the cells of $A(\Pi)$ are simply those of A^k with $k = q - 2$.

In the complex $A(\Pi)$ we may again define for the free abelian group F_a the generic subcomplex $A(F_a^*)$.

THEOREM 1. *For the construction A the generic subcomplex has the homology groups*

$$H_1(A(F_a^*)) \cong F_a, \qquad H_n(A(F_a^*)) = 0, \qquad n > 0. \tag{16}$$

There are many alternative constructions K of complexes $K(\Pi)$, one for each abelian group Π, which have as cells various types of n-tuples of elements of Π, with a boundary "formula" valid for all groups, and such that the generic cells of $K(F_a)$ form a subcomplex $K(F_a^*)$. We require also that the 1-cells of $K(\Pi)$ be the 1-cells $[x]$ of $A(\Pi)$.

THEOREM 2. *Let K be any such construction which has generic homology groups as in* (15). *Then for any abelian group* Π, *the homology and cohomology groups of* $K(\Pi)$ *in dimension* n *are isomorphic to those of* $A(\Pi)$.

We are thus justified in referring to the cohomology groups of $A(\Pi)$ as *the* cohomology groups of the *abelian* group Π.

These groups may be explicitly computed for any given dimension if Π is a finitely generated group. For certain low dimensions they have been determined for any Π. Thus, as in (6),

$$H^1(A(\Pi), G) = H^1(A^0(\Pi), G) \cong \text{Hom } (\Pi, G),$$

while, as already indicated in (10) and (11)

$$H^2(A(\Pi), G) = H^2(A^1(\Pi), G) \cong \text{Ext } (\Pi, G),$$

the group of all abelian group extensions of G by Π.

A 3-dimensional cochain of $A^1(\Pi)$ is a pair of functions $f(x, y, z) \in G$, $d(x \mid y) \in G$; they form a cocycle if they vanish on the boundaries (10), (11), and (4); that is, if they satisfy the identities

$$(17)\quad d(x + y \mid z) - d(y \mid z) - d(x \mid z) + f(x, y, z) - f(x, z, y) + f(z, x, y) = 0$$

$$(18)\quad d(x \mid y + z) - d(x \mid y) - d(x \mid z) - f(x, y, z) + f(y, x, z) - f(y, z, x) = 0$$

$$(19)\quad f(y, z, t) - f(x + y, z, t) + f(x, y + z, t) + f(x, y, z + t) - f(x, y, z) = 0.$$

A two-dimensional cochain is a function $h(x, y) \in G$; its coboundary is the pair (f, d) with

$$(20)\qquad f(x, y, z) = h(y, z) - h(x + y, z) + h(x, y + z) - h(x, y),$$

$$(21)\qquad d(x \mid y) = h(x, y) - h(y, x).$$

To any cocycle (f, d) we assign the function $t(x) = d(x \mid x) \in G$ as its trace. By (21), the trace of a coboundary is zero, and one may show that any trace satisfies the identities

$$(22)\quad t(x + y + z) - t(x + y) - t(y + z) - t(z + x) + t(x) + t(y) + t(z) = 0,$$

$$(23)\qquad t(x) = t(-x).$$

These are incidentally the formal identities satisfied by a "square" $t(x) = x^2$; hence we call any such function t a quadratic function.

THEOREM 3. *The function assigning to each cocycle its trace induces an isomorphism of* $H^3(A^1(\Pi), G)$ *to the group of all quadratic functions on* Π *to* G.

In the complex $A(\Pi)$ the three-cochains are the same pairs (f, d), but a cocycle must satisfy one additional identity $d(x \mid y) + d(y \mid x) = 0$, derived from (15). Thus $2t(x) = 0$, and we have

$$H^3(A(\Pi), G) = H^3(A^2(\Pi), G) \cong \text{Hom } (\Pi, {}_2G),$$

where ${}_2G$ is the subgroup of all elements of order 2 in G.

There are parallel results for the homology groups; one has the isomorphisms

$$H_1(A(\Pi)) \cong \Pi, \qquad H_2(A(\Pi)) = 0, \qquad H_3(A(\Pi)) \cong \Pi/2\Pi,$$

while $H_3(A^1(\Pi))$ is the group[3] $\Gamma(\Pi)$ which has the generators $[x]$ for $x \in \Pi$ and the relations

$$[x + y + z] - [x + y] - [y + z] + [z + x] + [x] + [y] + [z] = 0,$$

$$[x] = [-x].$$

In dimension 4, $H_4(A^1(\Pi))$ is isomorphic to the abelian group with the following generators: a generator $[s]$ for each element s of ${}_2\Pi$ and a generator $[x, y; h]$ for each pair of elements $x, y \in \Pi$ and each integer h such that $hx = hy = 0$. These generators satisfy the following relations for x, y, z with $hx = hy = hz = 0$

$$[x, y + z; h] = [x, y; h] + [x, z; h], \tag{24}$$

$$[x + y, z; h] = [x, z; h] + [y, z; h], \tag{25}$$

$$[x, x; h] = 0, \tag{26}$$

and the relations

$$[x, y; kh] = [kx, y; h], \quad \text{if} \quad (kh)x = 0, \qquad hy = 0, \tag{27}$$

$$[x, y; 2] = [x + y] - [x] - [y], \quad \text{if} \quad 2x = 2y = 0. \tag{28}$$

Conditions (24), (25), and (28) imply that $[s]$, for $s \in {}_2\Pi$, satisfies the relation used to define $\Gamma({}_2\Pi)$, while conditions (24), (25), and (26) imply that $[x, 0; h] = [0, y; h] = 0$ and that $[x, y; h] = -[y, x; h]$.

For the complex $A^2(\Pi)$, we have an isomorphism

$$H_4(A^2(\Pi)) \cong {}_2\Pi + \Lambda(\Pi),$$

where $\Lambda(\Pi)$ is the abelian group with generators $\langle x, y \rangle$ for all $x, y \in \Pi$, and relations

$$\langle x, y + z \rangle = \langle x, y \rangle + \langle x, z \rangle; \qquad \langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle; \qquad \langle x, x \rangle = 0.$$

Finally, $H_4(A(\Pi)) = H_4(A^3(\Pi)) \cong {}_2\Pi$.

Closely connected with Theorem 3 is the fact that the symmetric three-dimensional cohomology group is zero, in the following sense.

THEOREM 4. *If the function $f(x, y, z) \in G$ on Π satisfies* (19) *and the symmetry condition*

$$f(x, y, z) - f(x, z, y) + f(z, x, y) = 0,$$

then there is a function $h(x, y) \in G$ with $h(x, y) - h(y, x) = 0$, and

$$f(x, y, z) = h(y, z) - h(x + y, z) + h(x, y + z) - h(x, y).$$

[3] This is the group $\Gamma(\Pi)$ used by J. H. C. Whitehead; see his article in these Proceedings.

In other words $\delta f = 0$ implies $f = \delta h$, when f and h both satisfy symmetry conditions, and δ is the coboundary operator of $A^0(\Pi)$.

For the case when Π is the additive group of integers the homology groups of $A(\Pi)$ can be expressed by means of direct sums of cyclic groups (m) of order m, as follows:

$$H_2 = H_4 = H_6 = H_8 = 0, \qquad H_{10} = (2)$$

$$H_1 = (\infty), \qquad H_3 = (2), \qquad H_5 = (2) + (3)$$

$$H_7 = (2) + (2), \qquad H_9 = (2) + (2) + (3) + (5), \qquad H_{11} = (2) + (2).$$

UNIVERSITY OF CHICAGO,
CHICAGO, ILL., U. S. A.

THE COHOMOLOGY THEORY OF A PAIR OF GROUPS

REINHOLD BAER

1. The appropriate framework for our discussion is provided by the general concept of group representation. Such a representation has three component parts: an arbitrary group Q, the group to be represented; an abelian group G, the representation modulus (which could be an operator group without changing an iota of our discussions); the way Q operates on G, i.e., a multiplication $qg = g'$ such that for q in Q and g in G the product qg is a uniquely determined element g' in G with the usual properties of distributivity $[q(g + h) = qg + qh]$ and associativity $[p(qg) = (pq)g, 1g = g]$. The principal problem is to find invariants of such representations (Q, G, qg). The cohomology groups $H^n(Q, G)$ constitute a family of invariants which is of great interest not only because of their applications in topology, but much more so because of the great, and apparently rather disparate, significance of the first members of this chain: H^0 is the group of fixed elements of Q in G [satisfying $qf = f$ for every q in Q]; Z^1 is the group of crossed homomorphisms and B^1 the principal genus of Q in G; and H^2 is the group of group extensions of G by Q (and we note that Z^1 is closely connected with the automorphisms of the individual group extensions). For a comprehensive introduction into this theory see, for instance, Eilenberg [2] whose notations we shall use.

It is only natural to ask whether these invariants suffice to distinguish between nonisomorphic representations. This is not the case, as becomes apparent from the following (well-known) theorem which is easily verified by using I. Schur's "summation

If Q has finite order m, and if the mapping $g \to mg$ is an automorphism of G, then $H^n(Q, G) = 0$ for $0 < n$.

As a matter of fact it is not astonishing that these invariants fail to distinguish between different representations, since, for instance, the groups $C^n(Q, G)/Z^n(Q, G) \cong B^{n+1}(Q, G)$ form a further, and as yet apparently neglected, set of invariants. But once we have given up hope of using these invariants to distinguish different representations, we ask immediately the complementary question: which representations have the same cohomology groups?

2. Clearly we cannot expect that the cohomology groups of two random representations have any similarity with each other. Thus it seems justified to compare the cohomology groups of two representations only in case these representations are, in some fashion, related to each other. With this in mind we are going to consider representations (R, G, rg) and (S, G, sg) with the same representation modulus G which operate on each other as follows.

If r and s are elements in R and S respectively, then r^s and s^r are uniquely determined elements in R and S respectively, subject to the following rules:

(1) $r^{ss'} = (r^s)^{s'}$, $r^1 = r$; $s^{rr'} = (s^r)^{r'}$, $s^1 = s$;

(2′) $(rr')^s = r^s r'^s$; $(ss')^r = s^r s'^r$;

(2″) $r^{s^{r'}} = r^s$; $s^{r^{s'}} = s^r$;

(3) $rs^r g = sr^s g$ for every g in G.

If these conditions are satisfied, then it is possible to construct *the product of this pair of representations*. This is a representation $(Q, G, q \circ g)$ with the following properties.

(a) $Q = RS = SR$, $1 = R \cap S$.

(b) $sr^s = rs^r$.

(c) $rg = r \circ g$, $sg = s \circ g$.

Condition (a) has to be understood as requiring that R and S are subgroups of G and that every element in G may be written in one and only one way in the form rs and in one and only one way in the form $s'r'$. The existence of a group Q meeting requirements (a) and (b) is a special case of a theorem of Zappa [6]; and using this theorem it is possible to prove the equivalence of conditions (2′) and (2″).

The importance of products Q [with properties (a) and (b)] has been emphasized by various authors [Miller, Neumann, Ore, Szép, Zappa]; and this would justify an independent interest in the cohomology groups of $(Q, G, q \circ g)$; see, for instance, Lyndon [3] for an investigation of the cohomology groups of a direct product. For us, however, the cohomology theory of the product representation will only be a means for obtaining relations between the cohomology groups of the "factors." As a matter of fact, it will often prove useful to substitute for $(Q, G, q \circ g)$ suitably selected representations $(Q/N, G, q \circ g)$ where the elements in N operate trivially on G.

3. If (R, G, rg) and (S, G, sg) is a pair of representations and (Q, G, qg) their product [in the sense of §2], then it is not difficult to prove the equivalence of the following properties:

(I. R) There exists an idempotent endomorphism ρ of Q such that $Q^\rho = R$ and such that $q^\rho = 1$ implies $qg = g$ for every g in G.

(II. R) There exists a normal subgroup N of Q such that $Q = NR$, $1 = N \cap R$ and such that $ng = g$ for n in N and g in G.

(III. R) There exists a homomorphism η of S into R such that $sg = s^\eta g$ for s in S and g in G.

[By an obvious interchange of R and S we obtain a similar set of equivalent properties (I. S) to (III. S).]

Property (II. R) shows the importance of the representation $(Q/N, G, qg)$ mentioned before; and Property (III. R) may be used to see that the following

important class of pairs of representations (studied by Lyndon [3]) is a subclass of the class under consideration:

S is a normal subgroup of R and the elements in S operate on G as they do as elements in R; $r^s = r$ and $s^r = r^{-1}sr$.

The class of pairs meeting requirement (I. R) has the following important property.

(R) If the pair (R, G, rg), (S, G, sg), and their product (Q, G, qg) meet requirement (I. R), then $H^n(Q, G)$ possesses [for $0 < n$] a direct summand isomorphic to $H^n(R, G)$; in symbols:

(R*) $H^n(Q, G) = H^n(R, G) \oplus D$ *for* $0 < n$.

We sketch the proof. The cochain $f(\cdots r_i \cdots)$ in $C^n(R, G)$ may be mapped upon the cochain $f^\rho(\cdots q_i \cdots) = f(\cdots q_i^\rho \cdots)$ in $C^n(Q, G)$; and this mapping ρ is a homomorphism of $C^n(R, G)$ into $C^n(Q, G)$ which commutes with the coboundary operator, since $qg = q^\rho g$. A homomorphism ρ of $H^n(R, G)$ into $H^n(Q, G)$ is consequently induced by the homomorphism ρ of $C^n(R, G)$ into $C^n(Q, G)$. Next we map the cochain h in $C^n(Q, G)$ "by restriction" upon the cochain $h^{\rho'}$ in $C^n(R, G)$ which coincides on R with h. Clearly ρ' is a homomorphism of $C^n(Q, G)$ into $C^n(R, G)$ which commutes with the coboundary operator; and consequently ρ' induces a homomorphism of $H^n(Q, G)$ into $H^n(R, G)$ which we also denote by ρ'. Since the endomorphism ρ of Q is idempotent and maps Q upon R, it leaves invariant every element in R. The endomorphisms $\rho\rho'$ of $C^n(R, G)$ and of $H^n(R, G)$ are therefore equal to the identity automorphism. The homomorphism ρ of $H^n(R, G)$ into $H^n(Q, G)$ is consequently an isomorphism; and the endomorphism $\rho'\rho$ of $H^n(Q, G)$ is idempotent and maps $H^n(Q, G)$ upon the direct summand $H^n(R, G)^\rho$ which is isomorphic to $H^n(R, G)$. This proves (R).

The principal weakness of this result (R*) is the almost complete absence of information concerning the complementary direct summand D. To remedy this situation we have to introduce an important new concept.

4. We consider, as in §2, a pair of representations (R, G, rg) and (S, G, sg) linked by the operations r^s and s^r. Functions of the form

$$f_{i,j} = f(r_1, \cdots, r_i, s_1, \cdots, s_j) \quad \text{for} \quad r_h \text{ in } R \quad \text{and} \quad s_k \text{ in } S$$

may be considered both as i-dimensional cochains of R in $C^j(S, G)$ and as j-dimensional cochains of S in $C^i(R, G)$. This fact we try to exploit.

The elements in R and S operate on these functions $f_{i,j}$ according to the following rule:

$$r \circ f_{i,j} = rf(\cdots r_h \cdots s_k^r \cdots),\ s \circ f_{i,j} = sf(\cdots r_h^s \cdots s_k \cdots);$$

and these operations are related by the formula:

$$r \circ (s^r \circ f_{i,j}) = s \circ (r^s \circ f_{i,j}).$$

Next we define coboundary operators δ_R and δ_S in the obvious way; for instance,

$$\begin{aligned}\delta_S f_{i,j} &= (\delta_S f_{i,j})(r_1, \cdots, r_i, s_1, \cdots, s_{j+1}) \\ &= s_1 \circ f_{i,j}(r_1, \cdots, r_i, s_2, \cdots, s_{j+1}) \\ &+ \sum_{k=1}^{j} (-1)^k f_{i,j}(r_1, \cdots, r_i, s_1, \cdots, s_k s_{k+1}, \cdots, s_{j+1}) \\ &+ (-1)^{j+1} f_{i,j}(r_1, \cdots, r_i, s_1, \cdots, s_j);\end{aligned}$$

and we note the relations: $\delta_R^2 = 0$, $\delta_R \delta_S = \delta_S \delta_R$, $\delta_S^2 = 0$.

Now we define the n-dimensional cochains of our pair of representations as sequences

$$h = [h_{0,n}, \cdots, h_{i,n-i}, \cdots, h_{n,0}]$$

Their totality forms the group $C^n(R, S; G)$, if we define addition of cochains "coordinatewise". For these cochains we define an operator δ by the rule

$$\delta h = [\delta_S h_{0,n}, \cdots, \delta_R h_{i-1,n-i+1} + (-1)^i \delta_S h_{i,n-i}, \cdots, \delta_R h_{n,0}].$$

This operator is a homomorphism of $C^n(R, S; G)$ into $C^{n+1}(R, S; G)$ which satisfies $\delta^2 = 0$. Thus it is a true coboundary operator; and we may define as usual n-dimensional cocycles and coboundaries and the n-dimensional cohomology group $H^n(R, S; G)$.

This construction made its appearance in various places. The Parisian school of topologists [Cartan, Koszul, Leray] has used it [according to a communication of S. MacLane]; a complete exploitation of I. Schur's summation may be effected within this framework; and a very similar construction has recently been given by J. H. C. Whitehead [5].

The importance of this construction for our present needs stems from the validity of the following theorem.

If (Q, G, qg) is the product of the pair of representations (R, G, rg), (S, G, sg), then there exists a homomorphism φ of $C^n(Q, G)$ into $C^n(R, S; G)$ which commutes with the coboundary operator and satisfies

$$f^\varphi = [f(s_1, \cdots, s_n), \cdots, f(r_1, \cdots, r_n)] \quad \textit{for} \quad f \quad \textit{in} \quad C^n(Q, G).$$

Clearly φ induces a homomorphism of $H^n(Q, G)$ into $H^n(R, S; G)$.

Our construction of this homomorphism φ is rather intricate and does not seem to give any useful information concerning the "inner" coordinates of f^φ. This lack of information seems to be the principal obstacle encountered when using the constructions of this §4. [The reader is again referred to Lyndon's investigations of the cohomology groups of group extensions. There use is made of a certain fundamental homomorphism which may or may not be related to the homomorphism φ.]

5. Inserting the homomorphism φ [of §4] into the argument used when proving the proposition (R) [in §3] we obtain now the following result.

If the pair of representations (R, G, rg), (S, G, sg), and their product (Q, G, qg) satisfy condition (I. R), then

$$H^n(R, S; G) = H^n(R, G)^* \oplus H_0^n(R, S; G) \quad \textit{for} \quad 0 < n$$

where $H^n(R, G)^*$ is isomorphic to $H^n(R, G)$ and where the group H_0^n consists exactly of those elements in $H^n(R, S; G)$ which may be represented by cocycles of the form $[h_{0,n}, \cdots, h_{n-1,1}, 0]$.

To gain further insight into the structure of the "complement" H_0^n, denote by $H_{i-1}^n(R, S; G)$ the totality of elements in $H^n(R, S; G)$ which may be represented by cocycles of the form $[h_{0,n}, \cdots, h_{n-i,i}, 0, \cdots, 0]$. We obtain a chain of subgroups

$$0 = H_{n-1}^n \leqq \cdots \leqq H_i^n \leqq H_{i-1}^n \leqq \cdots \leqq H_0^n \leqq H^n(R, S; G);$$

and one verifies without any trouble that mapping the cocycle

$$[h_{0,n}, \cdots, h_{n-i,i}, 0, \cdots, 0]$$

upon the cocycle $h_{n-i,i}$ in $Z^{n-i}[R, C^i(S, G)]$ induces an isomorphism of H_{i-1}^n/H_i^n into $H^{n-i}[R, C^i(S, G)]$. (The reader should compare this with Lyndon [3, Theorem 4′, p. 291] where similar, though apparently different, subgroup chains make their appearance.)

6. We state now without proof a number of theorems that may be obtained with the help of the methods developed in the preceding sections. These theorems are not the most general results that may be derived in this fashion; but they are quite typical.

A. *If the pair of representations* (R, G, rg), (S, G, sg), *and their product* (Q, G, qg) *satisfy condition* (I. R), *and if furthermore* S *operates trivially on* R [*so that* $r^s = r$], *then*

$$H^i(S, G) = 0 \textit{ for } 0 \leqq i \leqq n \textit{ implies } H^i(R, G) = 0 \textit{ for } 0 \leqq i \leqq n.$$

The general hypotheses of this Theorem A are satisfied, for instance, whenever S is a normal subgroup of R. Consequently the first $n + 1$ cohomology groups of R vanish whenever the first $n + 1$ cohomology groups of some normal subgroup S of R do vanish; and this rather striking result may also be obtained as a special case of a theorem of Lyndon [3, Theorem 4′, p. 291].

B. *If the pair of representations* (R, G, rg), (S, G, sg), *and their product* (Q, G, qg) *satisfy conditions* (I. R) *and* (I. S), *and if furthermore* R *and* S *operate trivially upon each other* [*so that* $r^s = r$ *and* $s^r = s$], *then*

$$0 = H^i(R, G) = H^i(S, G) \textit{ for } 0 \leqq i < n \textit{ implies } H^n(R, G) \cong H^n(S, G).$$

It is a weakness of these two theorems that their hypotheses include the vanishing of $H^0(S, G)$, since this group is just the totality of elements g in G such that $sg = g$ for every s in S [the fixed element group of S in G]. The following result which does not make use of this hypothesis may therefore be of interest.

C. *If the pair of representations* (R, G, rg), (S, G, sg), *and their product* (Q, G, qg) *satisfy condition* (I. R), *and if* S *operates trivially on* R, *then*

$$0 = H^1(S, G) = \cdots = H^n(S, G)$$

implies that every cocycle in $Z^n(R, G)$ *is cohomologous to a cocycle with values in* $H^0(S, G)$.

If in particular S is a normal subgroup of R, then we obtain again a result that is a special case of Lyndon [3, Theorem 4′, p. 291]. The special cases $n = 2$ and $n = 3$ of Theorem C have been derived in the framework of Galois theory by Bergström [1] and Nakayama [4].

Bibliography

1. Harald Bergström, *Über abelsche Erweiterungen mit zerfallenden Faktorensystemen*, Mathematische Nachrichten vol. 1 (1948) pp. 350–356.

2. Samuel Eilenberg, *Topological methods in abstract algebra. Cohomology theory of groups*, Bull. Amer. Math. Soc. vol. 55 (1949) pp. 3–37.

3. Roger C. Lyndon, *Cohomology theory of group extensions*, Duke Math. J. vol. 15 (1948) pp. 271–292.

4. Tadasi Nakayama, *Note on 3-factor sets*, Tokyo Kodai Math. Rep. vol. 1 (1949) pp. 11–14.

5. J. H. C. Whitehead, *On group extensions with operators*, Quart. J. Math. Oxford Ser. vol. 1 (1950) pp. 219–228.

6. Guido Zappa, *Sulla costruzione dei gruppi prodotto di due dati sottogruppi permutabili tra loro*, Atti secondo Congresso Unione Mat. Ital., Bologna 1940 [Roma 1942] pp. 119–125.

University of Illinois,
Urbana, Ill., U. S. A.

THE BETTI NUMBERS OF THE EXCEPTIONAL SIMPLE LIE GROUPS

C. CHEVALLEY

The first explicit determination of the Poincaré polynomials of the exceptional simple Lie groups has been accomplished by Yen Chih Ta (C. R. Acad. Sci. Paris vol. 228 pp. 628–630). We propose to describe briefly the main steps in another procedure to achieve the same result.

I. REDUCTION TO AN ALGEBRAIC PROBLEM

It has been known for a long time that the Betti numbers of an arbitrary compact connected Lie group G can be determined by solving a purely algebraic problem. Denote by $\mathfrak{E}^p$ the space of closed differential forms of class C^∞ on G (i.e., of those differential forms ω for which $d\omega = 0$) and by $\mathfrak{F}^p$ the space of those elements in $\mathfrak{E}^p$ which may be written in the form $d\zeta$, ζ being a differential form of degree $p-1$ and class C^∞. Then de Rham's theorem asserts that the pth Betti number B_p of G is equal to the dimension of the factor space $\mathfrak{E}^p/\mathfrak{F}^p$.

If s is in G, let L_s and R_s be respectively the left and right translation by s. If ω is in $\mathfrak{E}^p$, and s and t in G, then $L_sR_t\omega$ is still in $\mathfrak{E}^p$, and congruent to ω (mod $\mathfrak{F}^p$), because G is connected. The form $\omega_1 = \int (L_sR_t\omega)\, ds \cdot dt$ (the integration being made relatively to the product by itself of the Haar measure on G) is again in $\mathfrak{E}^p$, is congruent to ω modulo $\mathfrak{F}^p$ and is two-sided invariant (i.e., $L_s\omega_1 = R_s\omega_1 = \omega_1$ for every s in G). On the other hand, it can be proved that a two-sided invariant differential form of degree p on G is in $\mathfrak{E}^p$ and cannot be in $\mathfrak{F}^p$ unless it is 0. The pth Betti number of G is therefore equal to the dimension of the space of two-sided invariant differential forms on G.

Let $\mathfrak{g}$ be the Lie algebra of G. Then the left-invariant differential forms of degree p of G are in a one-to-one correspondence with the elements of degree p of the exterior algebra $A(\mathfrak{g})$ over the dual vector space $\mathfrak{g}^*$ of $\mathfrak{g}$. Any right translation of G permutes among themselves the left-invariant differential forms of G; this gives rise to a linear representation ρ of G by automorphisms of the algebra $A(\mathfrak{g})$, and the two-sided invariant forms are the invariants of this representation of G.

There corresponds to the representation ρ of G by automorphisms of $A(\mathfrak{g})$ a representation θ of $\mathfrak{g}$ by derivations of the algebra $A(\mathfrak{g})$. This derivation may be obtained directly from $\mathfrak{g}$ as follows; if $X \in \mathfrak{g}$, the adjoint operation $\mathrm{ad}\, X$ of X is an endomorphism $Y \to [X, Y]$ of the vector space $\mathfrak{g}$; the restriction of $\theta(X)$ to $\mathfrak{g}^*$ is then the operation on $\mathfrak{g}^*$ contragredient to the operation $\mathrm{ad}X$ on $\mathfrak{g}$, and $\theta(X)$ itself is the unique extension to a derivation of $A(\mathfrak{g})$ of its restriction to $\mathfrak{g}^*$. Thus, the operations $\theta(X)$ can be determined as soon as the constants of structure of $\mathfrak{g}$ are given, and the determination of the Betti numbers of G is reduced to an algebraic problem (evaluating the ranks of certain systems of linear equations).

II. The theorem of Hopf

The two-sided invariant differential forms on G form a subalgebra $H(\mathfrak{g})$ of $A(\mathfrak{g})$. The structure of this algebra has been determined by Hopf (H. Hopf, *Über die Topologie der Gruppen-Mannigfaltigkeiten und ihre Verallgemeinerungen*, Ann. of Math. vol. 42 (1941)): it is isomorphic to the exterior algebra constructed over a certain vector subspace $P(\mathfrak{g})$ of $H(\mathfrak{g})$, and $P(\mathfrak{g})$ has a base composed of a certain number of forms of odd degrees $p_1, \cdots, p_l$. It follows that the Poincaré polynomial $\sum_p B_p t^p$ of G is $\prod_{k=1}^{l} (1 + t^{p_k})$, and the determination of the Betti numbers is reduced to that of the exponents p_k $(1 \leqq k \leqq l)$, which are called the *primitive exponents* of G.

J. L. Koszul has proved the theorem of Hopf in a purely algebraic manner (*Homologie et cohomologie des algèbres de Lie*, Bull. Soc. Math. France (1950)) and has defined explicitly a certain space $P(\mathfrak{g})$ which has the property stated above; the elements of this space are called the *primitive elements* of $A(\mathfrak{g})$.

III. Reduction to a problem of symmetric invariants

Besides $A(\mathfrak{g})$, let us now introduce the symmetric algebra $S(\mathfrak{g})$ over $\mathfrak{g}^*$; $S(\mathfrak{g})$ may be thought of as being the algebra of all functions of an argument X in $\mathfrak{g}$ which are polynomial functions in the sense that they can be expressed as polynomials in the coefficients of the expression of X with respect to some base of $\mathfrak{g}$. We have constructed above a linear representation of G whose space was $A(\mathfrak{g})$; the operations of this representation indicate how the right translations of G permute among themselves the left-invariant differential forms on G. Similarly, the elements of $S(\mathfrak{g})$ may be identified with the left-invariant symmetric tensors on G, and it follows immediately that we have a linear representation σ of G by automorphisms of $S(\mathfrak{g})$; the operations of this representation indicate how the right translations of G permute among themselves the left-invariant symmetric tensors on G. We shall denote by $I(\mathfrak{g})$ the algebra formed by those elements of $S(\mathfrak{g})$ which are invariant with respect to the operations of the representation we have just described; $I(\mathfrak{g})$ is therefore the algebra of two-sided invariant symmetric tensors on G. This algebra contains in particular the fundamental quadratic form $\mathrm{Tr}(\mathrm{ad}X)^2$ of Cartan; more generally, the coefficients of the Killing equation of an $X \in \mathfrak{g}$ (i.e., of the characteristic equation of ad X), considered as functions of X, are in $I(\mathfrak{g})$.

A. Weil has discovered the fact that the structure of the space $P(\mathfrak{g})$ of primitive elements is intimately related to that of the algebra $I(\mathfrak{g})$. Let $(\omega_1, \cdots, \omega_n)$ be a base of $\mathfrak{g}^*$, and let P be an element of $S(\mathfrak{g})$; we may express $P(X)$ in the form of a polynomial $P(x_1, \cdots, x_n)$ in the quantities $x_i = \omega_i(X)$ $(1 \leqq i \leqq n)$. On the other hand, the ω_i's are differential forms of degree 1 whose differentials $d\omega_i$ are of degree 2 and therefore belong to the center of $A(\mathfrak{g})$; it follows that we may substitute the $d\omega_i$ for the variables x_i in the polynomial P. It is clear that, if $P \in I(\mathfrak{g})$, then $P(d\omega_1, \cdots, d\omega_n)$ is in $H(\mathfrak{g})$, and that it is the coboundary (or the differential) of the element

$$\eta = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial P}{\partial x_i}\right)(d\omega_1, \cdots, d\omega_n)\omega_i.$$

But no element $\neq 0$ of $H(\mathfrak{g})$ can be a coboundary; it follows that $P(d\omega_1, \cdots, d\omega_n) = 0$ and that η is a cocycle. Moreover, it is easily seen that η is itself in $H(\mathfrak{g})$. Thus, we obtain a linear mapping $P \to \eta$ of $I(\mathfrak{g})$ into $H(\mathfrak{g})$. The study of this mapping shows that its range is precisely the space $P(\mathfrak{g})$ of primitive elements, and that its kernel is the subspace of $I(\mathfrak{g})$ spanned by 1 and by the products of homogeneous elements of degrees >0.

Let $p_k = 2q_k - 1$ $(1 \leqq k \leqq l)$ be the primitive exponents of G. Observe that, if P is of degree q, η is of degree $2q-1$; it then follows immediately from the result quoted above that the algebra $I(\mathfrak{g})$ is generated by l algebraically independent elements $I_1, \cdots, I_l$ of degrees $q_1, \cdots, q_l$; in order to find the primitive exponents, it will therefore be sufficient to determine the structure of the algebra $I(\mathfrak{g})$.

IV. Reduction to an invariant problem for a finite group

Let H be a maximal toroïdal subgroup of G, and let $\mathfrak{h}$ be the Lie algebra of H. It is well known that every element of G is conjugate in G to some element of H, and therefore that every element of $\mathfrak{g}$ may be transformed by the adjoint group into some element of $\mathfrak{h}$. If P is an element of $I(\mathfrak{g})$, then we have $P(sX) = P(X)$ for every element s of the adjoint group; it follows that P cannot vanish on $\mathfrak{h}$ without being identically zero. Let $I'(\mathfrak{g})$ be the ring of polynomial functions on $\mathfrak{h}$ which are obtained by taking the restrictions to $\mathfrak{h}$ of the functions in $I(\mathfrak{g})$; $I'(\mathfrak{g})$ is therefore isomorphic to $I(\mathfrak{g})$.

If N is the normalizator of H in G, then it is known that N/H is a finite group W, which I have proposed to call the *Weyl group* of G. This group may be represented as a finite group of linear transformations on $\mathfrak{h}$: it is the group of operations induced on $\mathfrak{h}$ by those operations of the adjoint group which transform $\mathfrak{h}$ into itself. It is then clear that, for any $P \in I(\mathfrak{g})$, the restriction P' of P to $\mathfrak{h}$ is an invariant of the finite group W. It turns out that the converse is also true: if P' is a polynomial function on $\mathfrak{h}$ which is invariant with respect to the operations of the group W, then P' is the restriction to $\mathfrak{h}$ of some function P belonging to $I(\mathfrak{g})$.

Thus, the determination of the primitive exponents of G is reduced to the solution of the "Formenproblem" for the finite group W. We know in advance that the ring of invariants will be generated by l algebraically independent elements; this fact can also be established directly by making use of the fact that the group W can be generated by a certain number of operations of order 2 which are symmetries with respect to certain hyperplanes in $\mathfrak{h}$, relatively to a suitable euclidean metric in $\mathfrak{h}$.

The groups W relative to the various exceptional Lie groups are known. Moreover, it can be proved *a priori* that the sum of the primitive exponents of G is equal to the dimension of G, and that the product $q_1 \cdots q_l$ is equal to the

order of the group W. Making use of these facts, it is possible to compute explicitly the primitive exponents in every individual case. The results are the following:

for G_2: 3, 11;

for F_4: 3, 11, 15, 23;

for E_6: 3, 9, 11, 15, 17, 23;

for E_7: 3, 11, 15, 19, 23, 27, 35;

for E_8: 3, 15, 23, 27, 35, 39, 47, 59.

It will be observed that, in every case, it turns out that we have $p_{k+1} - p_k = p_{l-k+1} - p_{l-k}$ $(1 \leqq k < l)$. This is also true for the primitive exponents of the classical groups; but we lack any rational explanation of this general fact.

COLUMBIA UNIVERSITY,
NEW YORK, N. Y., U. S. A.

STRUCTURE THEORY OF RINGS AND ALGEBRAS

POWER-ASSOCIATIVE ALGEBRAS

A. A. ALBERT

An *algebra* is a mathematical system $\mathfrak{A}$ consisting of an n-dimensional vector space over a field $\mathfrak{F}$ and a product xy which is a bilinear function of its arguments x and y. When every element x of $\mathfrak{A}$ generates an associative subalgebra $\mathfrak{F}[x]$ of $\mathfrak{A}$ the algebra $\mathfrak{A}$ is said to be *power-associative*. All of the major classes of algebras which have been studied so far are power-associative. The classes are, of course, the associative, alternative, Lie, and Jordan classes of algebras.

Lie algebras of characteristic not two are defined by the identities

$$xy = -yx, \qquad x(yz) + y(zx) + z(xy) = 0. \tag{1}$$

The structure theory for this class of algebras is due to Elie Cartan,[1] W. Landherr, and N. Jacobson.[2] It exists only for the case where $\mathfrak{F}$ has characteristic zero. Lie algebras are trivially power-associative since $x^2 = 0$, $x^n = 0$ for $n \geqq 2$. Indeed such algebras are *nilalgebras*, that is, all elements are nilpotent. As a consequence there is a sharp divergence between the methods used in the structure theory for Lie algebras and those used in the other major theories.

Alternative algebras may be defined by the identity $(yx)x = y(xx)$ and the *flexible* law $x(yx) = (xy)x$. The structure theory is due to M. Zorn[3] and, like the associative theory, has been developed for the case where $\mathfrak{F}$ can have characteristic $p \neq 0$. The alternative laws imply the *theorem of Artin*[4] which states that $\mathfrak{A}$ is alternative if and only if the subalgebra $\mathfrak{F}[x, y]$ generated by any two elements of $\mathfrak{A}$ is associative. The end result of the theory is that *all simple algebras* of the class *are either associative or are eight-dimensional Cayley algebras.*

A much richer theory exists in the case of the algebras of P. Jordan.[5] These algebras are defined by the identities

$$xy = yx, \qquad x(yx^2) = (xy)x^2. \tag{2}$$

[1] The structure theory is due to Killing and Cartan and was published in Cartan's thesis, Paris 1894. Real simple Lie algebras were determined by Cartan in his paper *Les groupes réels simples et continues*, Ann. École Norm. vol. 31 (1914) pp. 263–265.

[2] Some of the principal classes of simple Lie algebras over an arbitrary field of characteristic zero were determined by W. Landherr in his papers *Über einfache Liesche Ringe*, Abh. Math. Sem. Hamburgischen Univ. vol. 11 (1937) pp. 41–64 and *Liesche Ringe von Typsus* A, Abh. Math. Sem. Hansischen Univ. vol. 12 (1938) pp. 200–241. The remaining classes were determined by N. Jacobson, *A class of normal simple Lie algebras of characteristic zero*, Ann. of Math. vol. 38 (1937) pp. 508–517.

[3] *Theorie der alternativen Ringe*, Abh. Math. Sem. Hamburgischen Univ. vol. 8 (1930) pp. 123–147.

[4] A simple proof of a generalization of this theorem was given by M. Smiley, *The radical of an alternative ring*, Ann. of Math. vol. 49 (1948) pp. 702–709.

[5] These algebras were first defined by P. Jordan in his paper entitled *Über eine Klasse nichtassoziativer hyperkomplexer Algebren*, Göttingen Nachrichten, 1932, pp. 569–575.

The first structure theory of such algebras was given in 1934 by Jordan, E. Wigner, and J. von Neumann.[6] They made the highly restrictive hypotheses that $\mathfrak{F}$ is the field of all real numbers and that $\mathfrak{A}$ is formally real, that is, a sum $x_1^2 + \cdots + x_r^2 = 0$ only if $x_1 = \cdots = x_r = 0$. These hypotheses imply that $\mathfrak{A}$ contains no nilpotent elements and can have no radical. The present structure theory, which is due to myself,[7] G. Kalisch and N. Jacobson,[8] assumes only the basic identities (2) and that $\mathfrak{F}$ has characteristic zero. However the properties of a Jordan nilalgebra[7] were derived for the case where $\mathfrak{F}$ has any characteristic $p \neq 2$.

The reason why the present theories of Jordan and Lie algebras are restricted to the characteristic zero case is that the basic tool in both theories is a trace argument. In the characteristic p case of the associative theory the structure theorems are proved by using the characterization of the radical as the set of all properly nilpotent elements[9] of the algebra. The resulting arguments are strictly associative, and all attempts at extending the characterization to obtain a tool useful for the study of power-associative systems have been fruitless. It is for this reason that a trace argument has remained the only available tool for so long a time.

I have recently analyzed the trace arguments which have been used in the studies of algebras similar to associative algebras, and have carried out[10] an abstraction of the theory as follows. A power-associative algebra $\mathfrak{A}$ over a field $\mathfrak{F}$ is said to be *trace-admissible* if there exists a bilinear function $\tau(x, y)$, with arguments x and y in $\mathfrak{A}$ and values in $\mathfrak{F}$ (an *admissible trace function* for $\mathfrak{A}$), such that

I. τ is *symmetric*, that is, $\tau(x, y) = \tau(y, x)$;

II. τ is *associative*, that is, $\tau(x, yz) = \tau(xy, z)$;

III. $\tau(e, e) \neq 0$ if $e^2 = e \neq 0$;

IV. $\tau(x, y) = 0$ if xy is zero or is nilpotent.

Define the *radical* $\mathfrak{N}$ of $\mathfrak{A}$ to be the maximal nilideal of $\mathfrak{A}$. Then the principal structure theorems will hold for all power-associative trace-admissible algebras and so a great deal of structure theory may be deleted. Let us now see what structure theorems are like and what the results are in the trace-admissible case.

[6] *On an algebraic generalization of the quantum mechanical formalism*, Ann. of Math. vol. 35 (1934) pp. 29–64.

[7] *A structure theory for Jordan algebras*, Ann. of Math. vol. 48 (1947) pp. 546–567.

[8] The principal classes of special simple Jordan algebras over an *arbitrary* field of characteristic zero were determined by G. Kalisch, *On special Jordan algebras*, Trans. Amer. Math. Soc. vol. 61 (1947) pp. 482–494, together with F. D. Jacobson and N. Jacobson, *Classification and representation of semi-simple Jordan algebras*, Trans. Amer. Math. Soc. vol. 65 (1949) pp. 141–169.

[9] Cf. Chapter 2 of my *Structure of algebras*, Amer. Math. Soc. Colloquium Publications vol. 24, New York, 1939.

[10] *Trace-admissible algebras*, Proc. Nat. Acad. Sci. U. S. A. vol. 35 (1949) pp. 317–322.

When the radical $\mathfrak{N}$ of a power-associative algebra $\mathfrak{A}$ is defined to be the maximal nilideal of $\mathfrak{A}$, all Lie algebras become radical algebras and their study is then excluded. Every algebra $\mathfrak{A}$ not a radical algebra contains an idempotent element $e = e^2 \neq 0$ and we may define $\mathfrak{A}_e(\lambda)$ to be the subspace of $\mathfrak{A}$ consisting of all elements x in $\mathfrak{A}$ such that $ex + xe = 2\lambda x$ for λ in F. It is known[11] that $\lambda = 0$, 1 or $\frac{1}{2}$ and that $\mathfrak{A}$ is the supplementary sum

$$\mathfrak{A} = \mathfrak{A}_e(1) + \mathfrak{A}_e(\tfrac{1}{2}) + \mathfrak{A}_e(0) \tag{3}$$

of the three subspaces $\mathfrak{A}_e(\lambda)$. Moreover $ex = xe = \lambda x$ for $\lambda = 0$, 1 and x in $\mathfrak{A}_e(\lambda)$, the subspaces $\mathfrak{A}_e(1)$ and $\mathfrak{A}_e(0)$ are *orthogonal* and *are subalgebras of* $\mathfrak{A}$ *in the commutative case.* The desired basic structure theorems for all power-associative algebra theories are then:

THEOREM 1. *If e is principal (that is, $\mathfrak{A}_e(0)$ contains only nilpotent elements), then $\mathfrak{A}_e(\frac{1}{2}) + \mathfrak{A}_e(0) \subseteq \mathfrak{N}$.*

THEOREM 2. *If $\mathfrak{A}$ is semisimple, that is, $\mathfrak{N} = 0$, then $\mathfrak{A}$ has a unity quantity.*

THEOREM 3. *Every ideal of a semisimple algebra is semisimple.*

Theorems 1 and 2 are actually equivalent. Of course the final goal of any structure theory is a theorem stating the nature of the simple algebras.

The decomposition properties stated above are easy to prove for all power-associative rings $\mathfrak{R}$ whose characteristic is prime to 30 and which are such that $\frac{1}{2}x$ is a unique element of $\mathfrak{R}$ for every x of $\mathfrak{R}$. The mysterious integer 30 enters because of our use of the property that *a commutative ring $\mathfrak{R}$ is power-associative if and only if $x^2x^2 = (x^2x)x$*, and this property holds[12] if and only if the characteristic of $\mathfrak{R}$ is prime to 30. However the decomposition and its properties are now known[13] also for rings whose characteristic is merely prime to two if we assume associativity of fifth and sixth powers as well as the associativity of fourth powers. The decomposition is extended from the commutative to the non-commutative case by the following observation. Let $\mathfrak{A}$ be any power-associative algebra over a field $\mathfrak{F}$ whose characteristic is not two. Then there is an attached commutative algebra $\mathfrak{A}^{(+)}$ which is the same vector space as $\mathfrak{A}$ and is defined relative to a product $x \cdot y$ expressible in terms of the product xy of $\mathfrak{A}$ by $2x \cdot y = xy + yx$. The algebra $\mathfrak{A}^{(+)}$ is power-associative when $\mathfrak{A}$ is, and indeed powers in $\mathfrak{A}$ coincide with powers in $\mathfrak{A}^{(+)}$. The provable properties of $\mathfrak{A}$ are then derivable from those of $\mathfrak{A}^{(+)}$ by using the linearized form

[11] See Theorem 2 of my *Power-associative rings*, Trans. Amer. Math. Soc. vol. 64 (1948) pp. 552–593.

[12] *On the power-associativity of rings*, Summa Brasiliensis Mathematicae vol. 2 (1948) pp. 21–33.

[13] Mr. Louis Kokoris has proved these results as a part of an investigation in which he is trying to extend all of my theorems on commutative power-associative algebras to algebras of characteristic 3 and 5.

$$x(yz + zy) + y(zx + xz) + z(xy + yx) = (yz + zy)x + (xz + zx)y + (xy + yx)z, \tag{4}$$

of the identity $xx^2 = x^2x$.

Let us now return to the results in the trace-admissible case. When $\mathfrak{A}$ is trace-admissible, it has been shown that $\mathfrak{A}$ and $\mathfrak{A}^{(+)}$ have the same radical and so $\mathfrak{A}$ is semisimple if and only if $\mathfrak{A}^{(+)}$ is semisimple. Also the difference algebra $\mathfrak{A}-\mathfrak{N}$ is trace-admissible. When $\mathfrak{A}$ is semisimple, it has a unity quantity e and $\tau(x, y)$ is an admissible trace function for $\mathfrak{A}^{(+)}$ as well as for $\mathfrak{A}$. Then it is easy to show that $\mathfrak{A}^{(+)}$ is a Jordan algebra, that $\mathfrak{A}$ is *flexible*, and that $\mathfrak{A}$ is simple if and only if $\mathfrak{A}^{(+)}$ is simple. The simple trace-admissible algebras are then known if one knows the nature of all algebras $\mathfrak{A}$ such that $\mathfrak{A}$ is flexible and $\mathfrak{A}^{(+)}$ is a known simple Jordan algebra. Such algebras have actually been determined[14] and I shall describe the result later. Let us now list all known simple power-associative algebras which are not nilalgebras.

I have not stated yet what I mean when I say that an algebra is *simple* and I always mean *more* than the obvious assumption that it has no nontrivial ideals. In the associative case it is customary to add the hypothesis that $\mathfrak{A}$ is not a one-dimensional zero algebra, that is, if $\mathfrak{A} = u\mathfrak{F}$, then $u^2 \neq 0$. This hypothesis of the associative case is equivalent in that case to the assumption that $\mathfrak{A}$ is not a simple nilalgebra, and we shall adjoin the assumption that $\mathfrak{A}$ is not a nilalgebra to our definition of a simple power-associative algebra. When $\mathfrak{A}$ is a Jordan algebra, this assumption reduces to the assumption about one-dimensional algebras as in the associative case, and so we are led to our first important unsolved question. Do there exist simple commutative power-associative nilalgebras of dimension $n > 1$? The question may be rephrased as follows. If $\mathfrak{A}$ is any algebra we define $\mathfrak{A}^{(1)}$ to be the vector subspace of $\mathfrak{A}$ spanned by all products xy for x and y in $\mathfrak{A}$. Then is it true that if $\mathfrak{A}$ is a commutative power-associative nilalgebra, then $\mathfrak{A}$ contains $\mathfrak{A}^{(1)}$ properly? This result would imply that a nilalgebra is solvable, that is, that $\mathfrak{A} \supset \mathfrak{A}^{(1)} \supset \mathfrak{A}^{(2)} \cdots \supset \mathfrak{A}^{(r)} = 0$. A beginning in the study of this question has been made by M. Gerstenhaber who has shown that if x is a nilpotent element of a commutative power-associative algebra of characteristic zero, the linear transformation $a \to ax$ is nilpotent.

The study of simple nonassociative algebras $\mathfrak{A}$ is reducible[15] to the *central simple* case, that is, to the case where every scalar algebraic extension $\mathfrak{K}$ of the ground field $\mathfrak{F}$ yields a simple algebra $\mathfrak{A}_{\mathfrak{K}}$. If $\mathfrak{A}$ is central simple, we define the degree t of $\mathfrak{A}$ to be the maximal number of pairwise orthogonal idempotents in any $\mathfrak{A}_{\mathfrak{K}}$. As yet even central simple power-associative algebras of degrees one and two have not been completely classified. A class of algebras with a unity quantity e and $t = 1$ is the sum $\mathfrak{A} = e\mathfrak{F} + \mathfrak{B}_1 + \cdots + \mathfrak{B}_m$ where $\mathfrak{B}_i = u_i\mathfrak{F}$

[14] See Theorem 5.13 of the paper of footnote 11.

[15] See N. Jacobson, *A note on non-associative algebras*, Duke Math. J. vol. 3 (1937) pp. 544–548.

$+ v_i\mathfrak{F}$, $u_i^2 = v_i^2 = u_iu_j = u_iv_j = 0$ for $i \neq j$, $u_iv_i = -v_iu_i = \alpha_i e$ for $\alpha_i \neq 0$ in $\mathfrak{F}$. These algebras are trivially power-associative since every element $x = \alpha e + y$ where $y^2 = 0$. They are easily shown to be central simple. A class of algebras with $t = 2$ and a unity quantity e are the algebras $\mathfrak{A} = e\mathfrak{F} + u_2\mathfrak{F} + \cdots + u_s\mathfrak{F}$

$$u_i^2 = \alpha_i e, \qquad u_ju_i = -u_iu_j \qquad (i \neq j; \quad i, j = 2, \cdots, s), \tag{5}$$

where the $\alpha_i \neq 0$ are in $\mathfrak{F}$. These algebras are power-associative for all definitions of the products u_iu_j as long as $u_ju_i = -u_iu_j$. They are also easily seen to be central simple if $s > 2$. They become the central simple Jordan algebras of degree $t = 2$ if $s > 2$ and $u_iu_j = 0$ for $i \neq j$.

The algebras of (5) can be alternative only for the values $s = 1, 2, 4$, and 8 and we shall use the notation $\mathfrak{C}_s$ for these alternative algebras. Every $\mathfrak{C}_s$ has an involution (involutorial anti-automorphism) $x \to \bar{x}$ determined by $\bar{e} = e$, $\bar{u}_i = -u_i$ for $i = 2, \cdots, s$. The algebra $\mathfrak{C}_2$ may be only semisimple. Every $\mathfrak{C}_s$, except the Cayley algebra $\mathfrak{C}_8$, is actually associative.

Let us now turn to a description of what we shall call the classical central simple Jordan algebras of degree $t > 2$. We shall *describe* these algebras *only* in the case where $\mathfrak{F}$ *is algebraically closed* since the description in this case will be needed later. If $\mathfrak{S}$ is a central simple Jordan algebra of degree $t > 2$, there is an attached algebra $\mathfrak{G}$ consisting of all t-rowed square matrices X with elements x_{ij} in one of the alternative algebras $\mathfrak{C}_s$ described above. When $s = 8$, we must use only the value $t = 3$. The algebra $\mathfrak{G}$ has an involution J defined by $X = (x_{ij}) \to X^J = (\bar{x}_{ji})$, and $\mathfrak{S}$ is the subspace of $\mathfrak{G}$ consisting of all $X = X^J$. Indeed $\mathfrak{S}$ is then a subalgebra of $\mathfrak{G}^{(+)}$. The attached algebra $\mathfrak{G}$ is associative except when $s = 8$ and the other central simple Jordan algebras, defined for $s = 1$, 2, and 4, are *special* Jordan algebras, that is, they may be imbedded in a Jordan algebra $\mathfrak{G}^{(+)}$ where $\mathfrak{G}$ is associative. We shall call an algebra $\mathfrak{A}$ a *classical* Jordan central simple algebra if there exists a scalar extension $\mathfrak{K}$ of the ground field $\mathfrak{F}$ such that $\mathfrak{A}_{\mathfrak{K}}$ is either one of the algebras $\mathfrak{S}$ given above or has degree two. All classical Jordan algebras have been determined by Kalisch and Jacobson, and all central simple Jordan algebras of characteristic zero are classical Jordan algebras.

Our list of simple power-associative non-nil algebras consists, at this point, of the examples of algebras of degrees one and two, and the Jordan algebras. The arbitrary associative simple algebra is the set of all t-rowed square matrices with elements in an associative division algebra. In the case where the ground field is algebraically closed this reduces to a total matric algebra. We now extend the list by including the algebras which arise as a solution of the problem of finding all flexible algebras $\mathfrak{A}$ such that $\mathfrak{A}^{(+)}$ is a classical simple Jordan algebra.

Let $\mathfrak{A}$ be an algebra over a field $\mathfrak{F}$ and let $\lambda \neq \frac{1}{2}$ be in $\mathfrak{F}$. We may then define an algebra $\mathfrak{A}(\lambda)$ which is the same vector space as $\mathfrak{A}$ but is defined relative to the product $(x, y) = \lambda xy + (1 - \lambda)yx$ where xy is the product in $\mathfrak{A}$. An algebra $\mathfrak{A}$ over $\mathfrak{F}$ is now said to be *quasi-associative* if there exists a scalar extension $\mathfrak{K}$ of $\mathfrak{F}$ and an element λ in $\mathfrak{K}$ such that $\mathfrak{B} = \mathfrak{A}_{\mathfrak{K}}(\lambda)$ is associative. The algebra $\mathfrak{A}$ is

central simple if and only if $\mathfrak{B}$ is central simple and these are the algebras $\mathfrak{A} \neq \mathfrak{A}^{(+)}$ such that $\mathfrak{A}$ is flexible and $\mathfrak{A}^{(+)}$ is a simple Jordan algebra. The principal result on trace-admissible algebras $\mathfrak{A}$ of characteristic zero then states that a trace-admissible algebra $\mathfrak{A}$ is simple if and only if $\mathfrak{A}$ is either a simple Jordan algebra or a simple quasi-associative algebra.

The results outlined so far are quite well known and I want to continue with some new results. I am glad to be able to announce here that I have now been able to obtain a structure theory[16] for arbitrary commutative power-associative algebras over a field of characteristic $p \neq 2$, 3, or 5. As might be expected, I define the radical to be the maximal nilideal, and note that this is really no restriction in the case of Jordan algebras of characteristic p, a structure theory which is included in my theory. The theory succeeds because it begins with the study of simple algebras whereas these algebras are usually studied at the end of a structure theory. Assume first that $\mathfrak{A}$ is a commutative power-associative algebra which is not a nilalgebra and thus that $\mathfrak{A}$ contains an idempotent u. Then we have the decomposition $\mathfrak{A} = \mathfrak{A}_u(1) + \mathfrak{A}_u(\frac{1}{2}) + \mathfrak{A}_u(0)$. Let x_λ represent the arbitrary element of $\mathfrak{A}_u(\lambda)$ and write $z = xy$ for any two elements x and y of $\mathfrak{A}$. Then the subspaces $\mathfrak{A}_u(\lambda)$ have multiplicative properties which may be expressed by the formulas

$$\begin{aligned} x_\lambda y_\lambda &= z_\lambda, \qquad x_\lambda y_{1-\lambda} = 0, \\ x_{\frac{1}{2}} y_{\frac{1}{2}} &= z_0 + z_1, \qquad x_\lambda y_{\frac{1}{2}} = z_{\frac{1}{2}} + z_{1-\lambda} \end{aligned} \tag{6}$$

for $\lambda = 0, 1$. It follows that we may write

$$\begin{aligned} x_1 y_{\frac{1}{2}} &= y_{\frac{1}{2}}[S_{\frac{1}{2}}(x_1) + S_0(x_1)], \\ x_0 y_{\frac{1}{2}} &= y_{\frac{1}{2}}[T_{\frac{1}{2}}(x_0) + T_1(x_0)], \end{aligned} \tag{7}$$

where $S_{\frac{1}{2}}(x_1)$ and $T_{\frac{1}{2}}(x_0)$ are linear transformations[17] of $\mathfrak{A}_u(\frac{1}{2})$ as well as linear functions of x_1 and x_0 respectively, $S_0(x_1)$ is a linear mapping of $\mathfrak{A}_u(\frac{1}{2})$ into $\mathfrak{A}_u(0)$, $T_1(x_0)$ is a linear mapping of $\mathfrak{A}_u(\frac{1}{2})$ into $\mathfrak{A}_u(1)$. The linearized form of the relation $x^2x^2 = (x^2x)x$ may then be used to show that

$$S_{\frac{1}{2}}(x_1y_1) = S_{\frac{1}{2}}(x_1)S_{\frac{1}{2}}(y_1) + S_{\frac{1}{2}}(y_1)S_{\frac{1}{2}}(x_1), \tag{8}$$

$$\tfrac{1}{2}S_0(x_1y_1) = S_{\frac{1}{2}}(x_1)S_0(y_1) + S_{\frac{1}{2}}(y_1)S_0(x_1), \tag{9}$$

$$S_{\frac{1}{2}}(y_1)T_{\frac{1}{2}}(x_0) = T_{\frac{1}{2}}(x_0)S_{\frac{1}{2}}(y_1), \tag{10}$$

$$2w_{\frac{1}{2}}S_{\frac{1}{2}}(y_1)T_1(x_0) = [w_{\frac{1}{2}}T_1(x_0)]y_1. \tag{11}$$

The relation (8) implies that the mapping

$$x_1 \to 2S_{\frac{1}{2}}(x_1)$$

[16] The results appear in a paper entitled *A theory of power-associative commutative algebras*, Trans. Amer. Math. Soc. vol. 69 (1950) pp. 503–527.

[17] In the case where $\mathfrak{A}$ is a ring, $S_{\frac{1}{2}}$ and $T_{\frac{1}{2}}$ are endomorphisms and S_0 and T_1 are additive mappings. The results are actually derived in the ring case but their extension to the algebra case is immediate.

is a homomorphism of $\mathfrak{A}_u(1)$ onto the special Jordan algebra consisting of the linear transformations $S_i(x_1)$. The kernel of this homomorphism is an ideal $\mathfrak{B}_u$ of $\mathfrak{A}_u(1)$, and $\mathfrak{B}_u^2 = \mathfrak{C}_u$, where $\mathfrak{C}_u$ is the ideal of $\mathfrak{A}$ of all elements x_1 such that $y_i x_1 = 0$ for every y_i of $\mathfrak{A}_u(\frac{1}{2})$. When $\mathfrak{A}$ is simple, the ideal $\mathfrak{C}_u = 0$. It is these inner structural properties that have enabled me to prove the following rather remarkable ring theorem.

THEOREM. *Let $\mathfrak{A}$ be a simple commutative power-associative ring whose characteristic is prime to* 30, *and let $\mathfrak{A}$ contain a pair of orthogonal idempotents whose sum is not the unity quantity of $\mathfrak{A}$. Then $\mathfrak{A}$ is a Jordan ring.*

The proof of this result begins with a proof of the property that if u and v are orthogonal idempotents of a power-associative ring $\mathfrak{R}$, then $(au)v = (av)u$ for every a of $\mathfrak{R}$. Let $\mathfrak{A}$ be simple and e_1 and e_2 be orthogonal idempotents of $\mathfrak{A}$. Then $\mathfrak{A} \subseteq \mathfrak{R}$ where $\mathfrak{R}$ is a commutative power-associative ring with a unity quantity e and the same characteristic as $\mathfrak{A}$. Moreover $\mathfrak{R} = \mathfrak{A}$ or every nonzero ideal of $\mathfrak{R}$ contains $\mathfrak{A}$. In either case $e = e_1 + e_2 + e_3$ for pairwise orthogonal idempotents e_i and we may show that $\mathfrak{R} = \mathfrak{R}_{11} + \mathfrak{R}_{22} + \mathfrak{R}_{33} + \mathfrak{R}_{12} + \mathfrak{R}_{13} + \mathfrak{R}_{23}$ where $\mathfrak{R}_{ii} = \mathfrak{R}_{e_i}(1)$, and $\mathfrak{R}_{ij}$ is the intersection of $\mathfrak{R}_{e_i}(\frac{1}{2})$ and $\mathfrak{R}_{e_j}(\frac{1}{2})$ for $i \neq j$. Also $\mathfrak{R}_{ij}\mathfrak{R}_{jk} \subseteq \mathfrak{R}_{ik}$. We then show that if $g = e_i + e_j$, the intersection of the kernel $\mathfrak{B}_g$ and $\mathfrak{R}_{ij}$ is zero. It follows readily that

$$\mathfrak{B} = \mathfrak{B}_{e_1+e_2} + \mathfrak{B}_{e_1+e_3} + \mathfrak{B}_{e_2+e_3}$$

is an ideal of $\mathfrak{R}$, and that $\mathfrak{B} \supseteq \mathfrak{A}$ if $\mathfrak{B} \neq 0$. This is easily seen to be impossible and so $\mathfrak{B} = 0$, the subrings $\mathfrak{R}_{e_i+e_j}(1)$ are Jordan rings. A computation using the property $x^2x^2 = (x^2x)x$ will then yield the theorem.

The result of this theorem and further arguments about ideals may be used to show that *every simple commutative power-associative algebra has a unity quantity*. Moreover, every such algebra of degree $t > 2$ is a classical Jordan algebra and also every Jordan algebra of degree $t = 2$ is a classical Jordan algebra. The major structure theorems for commutative power-associative algebras then follow readily. However two important unsolved problems remain in the study of algebras of low-degree and we shall present them now.

The first of these problems is that of the nature of a commutative power-associative algebra $\mathfrak{A}$ with a unity quantity e over an algebraically closed field $\mathfrak{F}$. Assume that e is primitive so that $x = \alpha e + y$, where α is in $\mathfrak{F}$ and y is nilpotent for every x of $\mathfrak{A}$. I have shown that if $\mathfrak{A}$ is simple and $\mathfrak{F}$ has characteristic zero, then $\mathfrak{A} = e\mathfrak{F}$. I have also proved this result for *special* Jordan algebras of characteristic p. There remains the general case of algebras of characteristic p.

The second problem is concerned with the nature of simple commutative power-associative algebras of degree $t = 2$ *even in the case of algebras of characteristic zero*. It may be true that all such algebras are classical Jordan algebras, but there is no strong indication that this is actually true.

The results we have described are, in a sense, negative results. For the simple algebras are really the end results of any structure theory of a class of algebras, and the quasi-associative algebras, which are the only nonclassical algebras we have obtained, are only minor distortions of associative algebras. It therefore seems reasonable to propose the question as to whether there are any simple algebras behaving like associative algebras in respect to the existence of idempotents and which are also power-associative. The first real attack on this question might then be an attempt to find all power-associative algebras $\mathfrak{A}$ such that $\mathfrak{A}^{(+)}$ is a central simple Jordan algebra. I have solved this problem and obtained[18] the construction given in the following theorem.

THEOREM. *Let $\mathfrak{A}$ be a power-associative algebra over a field $\mathfrak{F}$ of characteristic prime to 30, $\mathfrak{S} = \mathfrak{A}^{(+)}$ be a central simple special Jordan algebra so that there exists a scalar extension $\mathfrak{K}$ of $\mathfrak{F}$ and an associative algebra $\mathfrak{G}$ over $\mathfrak{K}$ such that $\mathfrak{S}_{\mathfrak{K}}$ is the set of all J-symmetric elements of $\mathfrak{G}$. Then $\mathfrak{K}$ may be selected so that there exists a linear mapping* T *of the set of all J-skew elements of $\mathfrak{G}$ into $\mathfrak{S}_{\mathfrak{K}}$ such that the product $x \cdot y$ of $\mathfrak{A}$ is expressible in terms of the product xy of $\mathfrak{G}$ by the formula*

$$x \cdot y = \tfrac{1}{2}(xy + yx) + (xy - yx)\mathrm{T}. \tag{12}$$

Conversely if $\mathfrak{G}$ is an associative algebra attached to a central simple Jordan algebra $\mathfrak{S}$ and $\mathfrak{A}$ is the vector space $\mathfrak{S}$ of all J-symmetric elements of $\mathfrak{G}$, the algebra $\mathfrak{A}$ defined by (12) *is a central simple power-associative algebra.*

It should be evident that powers in $\mathfrak{A}$ coincide with powers in $\mathfrak{G}$ and in $\mathfrak{S}$. In the case where $\mathfrak{S} = \mathfrak{A}^{(+)}$ is not a *special* Jordan algebra the algebra $\mathfrak{G}$ is not associative, $xx^2 \neq x^2x$, in $\mathfrak{G}$ even for the elements x of $\mathfrak{S}$. It is then necessary to adjoin the hypothesis that T shall annihilate diagonal skew elements of the algebra $\mathfrak{G}$ of all three-rowed matrices of Cayley elements.

In the case where $\mathfrak{F}$ is the real number field and $\mathfrak{S}$ is a formally real algebra the algebras defined by (12) are also formally real. Such algebras might have physical applications. In any event the new classes of algebras defined by (12) should have interesting properties and provide a starting point for new problems of structure and representation for power-associative algebras.

UNIVERSITY OF CHICAGO,
CHICAGO, ILL., U. S. A.

[18] These results will appear in a paper entitled *New simple power-associative algebras* which will be published in Summa Brasiliensis Mathematicae.

ON THE REPRESENTATIONS OF GROUPS OF FINITE ORDER

RICHARD BRAUER

1. By a representation of a group G of finite order n, we shall always mean a representation of G by linear transformations of a finite-dimensional vector space over a given field. We are interested in the following question: To what extent are the properties of representations of G determined by properties of representations of suitable subgroups?

As a first result in this direction, we state a theorem which gives the necessary and sufficient conditions that a function $\chi(g)$ defined over G be an irreducible character. As in the classical theory, the underlying field is assumed to be the field of complex numbers or, more generally, an algebraically closed field of characteristic 0. We shall call a group an elementary group, if it is the direct product of a cyclic group and a p-group (i.e., a group whose order is a power of a prime p). Then the conditions are as follows:

I. If H is any elementary subgroup of G and if the argument g is restricted to H, then $\chi(g)$ is a (reducible or irreducible) character of H.

II. For $g \in G$, the function $\chi(g)$ is a class function, that is, the value of $\chi(g)$ depends only on the class of conjugate elements of G to which g belongs.

III. $$(1/n) \sum_{g} | \chi(g) |^2 = 1.$$

The necessity of these conditions is clear. The sufficiency can be deduced from results concerning induced representations.[1] The condition I can be replaced by the weaker condition that $\chi(g)$ for $g \in H$ be a linear combination of the characters of H with integral rational coefficients, if the condition $\chi(1) > 0$ is added. The result can also be formulated as a theorem on representations rather than on characters. If for each elementary subgroup H of G a representation of H is given, we have the necessary and sufficient conditions that, after similarity transformations, these representations can be pieced together to a representation of G.

In the special case where $\chi(1) = 1$, the condition III is a consequence of I. Thus, the linear characters of G can be characterized as the class functions which yield linear characters for every elementary subgroup of G. Since the linear characters are closely related to the commutator subgroup G' of G, this leads to necessary and sufficient conditions that G' be different from G. By applying these conditions to all subgroups of G, we also obtain necessary and sufficient conditions for the solubility of G. However, it should be mentioned that these results can also be obtained by direct methods developed by Burnside, Frobenius, and Schur.

[1] R. Brauer, Ann. of Math. vol. 48 (1947) p. 502.

2. It is a disadvantage of our result that a knowledge of the characters of the elementary groups, that is, essentially of the p-groups, is required if one wants to apply it for the construction of characters. We give therefore a second theorem where this difficulty is avoided. We keep conditions II and III and replace I by another condition. Consider all pairs of subgroups L, M of G such that M is normal in L and L/M is cyclic. Let $M\tau$ be a generating element of L/M and denote the orders of L and M by l and m respectively. The new condition I then is that equations hold

$$\sum_{\mu \in M} \chi(\mu\tau^j) = m \sum_{i=0}^{l/m-1} x_i \rho^{ij}, \qquad (j = 0, 1, \cdots, l/m - 1),$$

where ρ is a primitive (l/m)th root of unity and where the x_i are rational integers independent of j.

Again, we obtain a set of necessary and sufficient conditions for irreducible characters. For arbitrary groups, the new condition is difficult to handle. However, because of our first theorem, it is only necessary to apply the new theorem in the case of an elementary group. In this case, the result can actually be used for the investigation of the characters.

3. The theory of modular representations of groups furnishes further theorems which connect properties of representations of G with properties of representations of suitable subgroups. Since at least an outline of these results has already been published,[2] we shall not go into any details but indicate only the type of problems in which these connections appear. If p is a fixed prime number, the modular as well as the ordinary absolutely irreducible characters appear distributed into a certain number of "blocks". This distribution is related closely to the decomposition of the modular group ring into a direct sum of indecomposable rings. It turns out that the structure of the blocks of G is determined largely by the structure of the blocks of subgroups. These subgroups are the normalizers of the p-subgroups of G and related groups.

4. We return again to fields of characteristic 0. The theory of group representations in a field K which is not algebraically closed has been developed by I. Schur.[3] Let $\bar{K}$ be an algebraically closed extension field of K. Then in $\bar{K}$ the classical theory applies. Each irreducible representation T of G in K breaks up completely in $\bar{K}$ into a certain number of distinct irreducible representations $F_1, F_2, \cdots, F_r$ and each F_i appears in T with the same multiplicity m. This m is the *Schur index* of the representations F_i. The characters $\chi_1, \chi_2, \cdots, \chi_r$ of $F_1, F_2, \cdots, F_r$ form a full family of absolutely irreducible characters of G which are algebraically conjugate with respect to K. Conversely, each such

[2] R. Brauer, Proc. Nat. Acad. Sci. U. S. A. vol. 30 (1944) p. 109; vol. 32 (1946) pp. 182, 215.

[3] I. Schur, Preuss. Akad. Wiss. Sitzungsber. (1906) p. 164 and Trans. Amer. Math. Soc. vol. 15 (1909) p. 159.

family of conjugate characters gives rise to one and only one irreducible representation of G in K. Finally, for representations in K, the theorem of complete reducibility holds. Thus, if the characters of G in the classical sense are known, it remains only to determine the Schur index m for each absolutely irreducible character of G in order to have a complete theory of group representations in K.

Schur also gave a second characterization of m. The representation F_i can be written in suitable fields of degree m over the field $K(\chi_i)$ obtained from K by adjunction of the character χ_i of F_i, and m is the minimal degree for which this is possible. In fact, if F_i can be written in a field L, then L must of course contain χ_i. If the degree $[L:K(\chi_i)]$ is finite, it is even divisible by m.

Later, the theory of algebras provided still another interpretation of m. Every representation T of G can be extended to a representation of the corresponding group ring. If T is irreducible in K, this defines a homomorphism of the group algebra Λ over K on a simple algebra Γ. Every simple homomorphic image Γ of Λ corresponds to one and only one representation T. The center Z of Γ is isomorphic to $K(\chi_i)$ with respect to K. If we now write Γ as a complete matric algebra over a division algebra D, then D as a central algebra over Z has rank m^2. Thus, the Schur indices of the representations are of fundamental importance for the study of the group algebra.

These different characterizations of the Schur index m do not provide a method of determining m, and as a matter of fact, this question remained open in Schur's theory. We shall deal with it in the following sections.

5. Let χ be a fixed absolutely irreducible character of G. Without restriction, we may assume that K contains χ since adjunction of χ does not change the Schur index m. Let p be an arbitrary prime. We need the following lemma.

LEMMA. *There exist elements g of G such that the group H^* generated by g and a p-Sylow subgroup of the normalizer of the cyclic group $\{g\}$ possesses absolutely irreducible characters ω with the following two properties*: 1. *The degree $[K(\omega):K]$ is prime to p.* 2. *If χ is considered as a character of H^* by restricting the argument to H^*, then ω appears in χ with a multiplicity prime to p.*

The proof of the lemma can be obtained by a refinement of the method of the paper quoted in footnote 1. If the absolutely irreducible characters of G and the relations to the characters of subgroups (at least of the type of H^*) are known, then it is actually possible to select ω. All we have to know about the field K (as given originally) is the manner in which the characters are distributed into classes of algebraically conjugate characters.

It follows from the lemma that the highest power of p dividing the Schur index m of χ is equal to the index m^* of the character ω of H^* with respect to K. Since p was an arbitrary prime, it will be sufficient to obtain the Schur indices of the representations of H^* in order to find m itself.

The group H^* contains a normal cyclic subgroup $\{g\}$ whose factor group is a

p-group. We shall call a group of this type a *semi-elementary group*. In particular, every semi-elementary group is soluble. We now have succeeded in reducing the problem of the Schur index to the case where the group in question is semi-elementary. In this case, the degrees of the absolutely irreducible representations are all powers of the same prime.

We have again a result of the type in which we are interested: A property of the representations of G is determined by the corresponding property of representations of suitable subgroups.

6. In the case of the Schur index, it is possible to obtain a further reduction. The basic fact here is that if ω is an absolutely irreducible character of a semi-elementary group H^* and if the degree of ω is greater than 1, there exist normal subgroups of H^* of prime index for which the character ω becomes reducible. The final result is that it is possible to reduce the whole problem to the case of a group R which has a cyclic normal subgroup S such that R/S is an *abelian* p-group. Even further conditions can be imposed. In the sense indicated above, this reduction is independent of the field K.

The representations of the group R can be constructed explicitly without difficulty. The only irrationalities needed are the qth roots of unity where q is the least common multiple of the orders of the elements of R. Further, we can determine the factor sets of the corresponding central division algebras.

It follows from the preceeding statements that if n^* is the least common multiple of the orders of the elements of the original group G and if the field K contains the n^*th roots of unity, then every representation of G can be written in the field K.[4] Indeed, our reduction leads to groups R which can be written in the field K. Hence their Schur indices are all 1, and then the Schur indices of G have the same value 1. As is well known, there exist in general other fields K which do not contain the n^*th roots of unity but which are such that every irreducible representation of the given group G can be written in K. The particular role of the n^*th cyclotomic field can be explained by the fact that it is the field obtained by adjunction of all characters of G and of all subgroups of G.

As a special result, we mention that the Schur index is always equal to 1 for a p-group with odd p. For p-groups with $p = 2$, we have $m = 1$ or $m = 2$.

7. The problem of the Schur index was reduced in §6 to the case of certain metabelian groups and as was mentioned, the corresponding factor sets can be obtained explicitly in this case. If we now assume that K is an algebraic number field, the theory of algebras provides methods of determining the indices. Actually, we obtain more information concerning the division algebras in question. Indeed, since the method works for an arbitrary field K, we can also determine the splitting fields of the division algebras.

UNIVERSITY OF MICHIGAN,
ANN ARBOR, MICH., U. S. A.

[4] This has already been proved in R. Brauer, Amer. J. Math. 69 (1947) p. 709.

REPRESENTATION THEORY FOR JORDAN RINGS[1]

N. JACOBSON

It is well known that the theory of Lie algebras is equivalent to the study of subspaces of associative algebras which are closed relative to the composition $[ab] \equiv ab - ba$. Similarly, the theory of Jordan algebras has arisen in the attempt to study subspaces of an associative algebra which are closed relative to the composition $\{ab\} = ab + ba$. The characterization of the Lie composition by identities is well known ([5] and [21]). (Besides the bilinearity, the axioms which characterize $[ab]$ are $[aa] = 0$ and the Jacobi identity.) On the other hand, we do not possess a set of identities for $\{ab\}$ which characterize this composition. It is easy to see that $\{ab\}$ satisfies

$$\{ab\} = \{ba\}, \qquad \{\{aa\}\{ba\}\} = \{\{\{aa\}b\}a\}, \tag{1}$$

and this observation has led to the definition of an (*abstract*) *Jordan algebra* as a (nonassociative) algebra whose composition ab satisfies the properties (1) of $\{ab\}$, that is,

$$ab = ba, \qquad a^2(ba) = (a^2b)a. \tag{1'}$$

The algebras obtained from subspaces of associative algebras closed relative to $\{ab\}$ are called *special Jordan algebras*. It is known that there exist Jordan algebras which are not special [1] so that the properties (1) do not give an exact characterization of the special systems. However, this may be the best that one can do in the way of identities; for it is conceivable that every Jordan algebra is a homomorphic image of a special one. At any rate, an extensive theory can be built on the axioms (1') (see particularly [16] and [3]). In this note we shall be concerned primarily with an extension of this theory of abstract Jordan algebras.

1. Definition and elementary properties of representations of Jordan algebras. The second identity $(a^2b)a = a^2(ba)$ for Jordan algebras is cubic in a. If the characteristic is not two or three (and we assume this unless otherwise stated), then this identity is equivalent to the multilinear identity:

$$xybz + yzbx + zxby = (xy)(bz) + (yz)(bx) + (zx)(by) \tag{2}$$

in which we have abbreviated $((xy)b)z$ to $xybz$, etc. If we denote the right multiplication mapping $x \to xa$ by R_a, then we obtain from (2) the following relations:

[1] The main results stated in the first five sections of this paper can be found in [14]. The results in the last section are contained in the joint paper [15] by Rickart and the present author.

$$
\begin{gathered}
[R_aR_{bc}] + [R_bR_{ac}] + [R_cR_{ab}] = 0 \\
R_aR_bR_c + R_cR_bR_a + R_{(ac)b} = R_aR_{bc} + R_bR_{ac} + R_cR_{ab} .
\end{gathered} \tag{3}
$$

We define a (general) representation of a Jordan algebra in the following way.

DEFINITION 1. A mapping $a \to S_a$ of a Jordan algebra into a set of linear transformations of a vector space $\mathfrak{M}$ is called a *representation* if (1) S_a is linear in a and (2) S_a satisfies

$$
\begin{gathered}
[S_aS_{bc}] + [S_bS_{ac}] + [S_cS_{ab}] = 0 \\
S_aS_bS_c + S_cS_bS_a + S_{(ac)b} = S_aS_{bc} + S_bS_{ac} + S_cS_{ab} .
\end{gathered} \tag{4}
$$

This concept is equivalent to that of a Jordan module which has been given by Eilenberg [7]. We shall not require the alternative formulation here.

Evidently the mapping $a \to R_a$ is a representation. As usual we call this mapping *the regular representation*. This representation plays a fundamental role in the structure theory. On the other hand, if one recalls the origin of the Jordan theory, one is led to consider also mappings $a \to U_a$ of a Jordan algebra into linear transformations such that U_a is linear in a and

$$U_{ab} = U_aU_b + U_bU_a . \tag{5}$$

Such mappings, which are just the homomorphisms of Jordan rings into the special Jordan rings of linear transformations, have been considered before ([6] and [11]). It is noteworthy that these mappings are representations in the sense defined here. We call these representations *special*. Thus the general concept of representations can be used to unify to some extent the structure theory and the theory of special representations.

There is another interesting connection between special and general representations. Let $a \to U_a$, $a \to V_a$ be two special representations acting in the same vector space. Assume that these *commute* in the sense that $[U_aV_b] = 0$ for all a, b. Then it can be verified that $a \to S_a \equiv U_a + V_a$ is a representation. In particular, if $a \to U_a$ and $a \to V_a$ are arbitrary special representations, then $a \to S_a = U_a \times 1 + 1 \times V_a$, where the $\times$ denotes Krönecker multiplication, is a representation.

2. Universal associative algebras. Let $\mathfrak{A}$ be a Jordan algebra and let $\mathfrak{F} = \mathfrak{A} \oplus \mathfrak{A} \times \mathfrak{A} \oplus \mathfrak{A} \times \mathfrak{A} \times \mathfrak{A} \oplus \cdots$ be the free associative algebra determined by the vector space $\mathfrak{A}$. Multiplication in $\mathfrak{F}$ is defined by the distributive laws and the rule

$$(x_1 \times x_2 \times \cdots \times x_r) \times (x_{r+1} \times \cdots \times x_s) = x_1 \times x_2 \times \cdots \times x_s .$$

Let $\mathfrak{R}$ be the ideal in $\mathfrak{F}$ generated by the elements

$$
\begin{gathered}
a \times bc - bc \times a + b \times ac - ac \times b + c \times ab - ab \times c, \\
a \times b \times c + c \times b \times a + (ac)b - a \times bc - b \times ca - c \times ab,
\end{gathered} \tag{6}
$$

and let $\mathfrak{U} = \mathfrak{F}/\mathfrak{R}$. We shall call $\mathfrak{U}$ the *universal associative algebra (of the representations)* of $\mathfrak{A}$. The algebra $\mathfrak{U}$ has the following properties: (1) Any representation of $\mathfrak{A}$ can be extended to one of $\mathfrak{U}$ (in the associative sense). (2) Any (associative) representation of $\mathfrak{U}$ defines a representation of $\mathfrak{A}$. A basic result in this connection is that $\mathfrak{U}$ *is finite-dimensional if* $\mathfrak{A}$ *is finite-dimensional.* This has the consequence that a finite-dimensional Jordan algebra has only a finite number of inequivalent irreducible representations. A fundamental problem which is as yet unsolved is the determination of these representations.

We can define also a *special universal associative algebra* $\mathfrak{U}_s$ for the special representations of $\mathfrak{A}$. This is obtained by replacing the ideal $\mathfrak{R}$ by the ideal $\mathfrak{R}_s$ generated by the elements $a \times b + b \times a - ab$. The algebra $\mathfrak{U}_s$ bears the same relation to special representations that $\mathfrak{U}$ does to all the representations. It is of interest to introduce also another universal associative algebra which can be used to study Kronecker sums $(U_a \times 1 + 1 \times V_a)$ of special representations. We obtain this algebra by adjoining a (new) identity 1 to $\mathfrak{U}_s$ to obtain $\mathfrak{U}_s^*$. Let $\mathfrak{U}_s^{(2)}$ be the subalgebra of $\mathfrak{U}_s^* \times \mathfrak{U}_s^*$ generated by the elements $a_s \times 1 + 1 \times a_s$ where a_s is the coset of $a + \mathfrak{R}_s$ in $\mathfrak{U}_s$. Any Kronecker sum of special representations defines a representation of $\mathfrak{U}_s^{(2)}$.

The algebras $\mathfrak{U}_s$ and $\mathfrak{U}_s^{(2)}$ are homomorphic images of $\mathfrak{U}$. Also $\mathfrak{U}_s$ is a homomorphic image of $\mathfrak{U}_s^{(2)}$. It would be interesting to have more precise results on the relation between $\mathfrak{U}$ and $\mathfrak{U}_s^{(2)}$ for special Jordan algebras $\mathfrak{A}$. At the present time we know of no example of a special Jordan algebra for which $\mathfrak{U} \ncong \mathfrak{U}_s^{(2)}$. Such examples could be used to construct extensions of $\mathfrak{A}$ which are not special.

3. Associator structure Lie triple systems. If $\mathfrak{A}$ is a special Jordan algebra with composition $\{ab\}$, a direct verification shows that the Jordan associator

$$A(b, c, a) \equiv \{\{bc\}a\} - \{b\{ca\}\} = [[ab]c]. \tag{7}$$

Hence $\mathfrak{A}$ is closed also relative to the Lie ternary composition $[abc] \equiv [[ab]c]$. A subspace of an associative algebra having this property is called a *(special) Lie triple system.*

It is possible to give a characterization by identities of Lie triple systems. Thus let $\mathfrak{T}$ be a vector space in which a ternary composition $[abc]$ is defined. Assume that this composition is trilinear and that it satisfies the following relations:

$$\begin{gathered}
[aab] = 0 \\
[abc] + [bca] + [cab] = 0 \\
[[abc]de] + [[bad]ce] + [ba[cde]] + [cd[abe]] = 0 \\
[[abc]de] + [[bad]ce] + [[dcb]ae] + [[cda]be] = 0 \\
[[[abc]de]fg] + [[[bac]df]eg] + [[[bad]ce]fg] \\
+ [[[abd]cf]eg] + Q + R = 0
\end{gathered} \tag{8}$$

where Q and R are obtained from the displayed terms by cyclic permutation of the pairs (a, b), (c, d), (e, f). Then it can be shown that $\mathfrak{T}$ can be imbedded in a Lie algebra $\mathfrak{L}$ in such a way that $\mathfrak{T}$ becomes a subspace of $\mathfrak{L}$ closed relative to $[[ab]c]$ and that $[[ab]c] = [abc]$. Since every Lie algebra can be identified with a subspace of an associative algebra closed under $[ab] \equiv ab - ba$, $\mathfrak{T}$ can be identified with a subspace of an associative algebra closed relative to $[[ab]c]$.

One can also associate a Lie triple system with every abstract Jordan algebra. If $\mathfrak{A}$ is such an algebra, we define $[abc] = A(b, c, a) = (bc)a - b(ca)$. Then it can be proved that this composition satisfies (8). This can be established easily by noting the following consequence of the second relation in (3):

$$[[R_a R_b]R_c] = R_{[abc]} . \tag{9}$$

The system consisting of the vector space $\mathfrak{A}$ and the composition $[abc]$ is called the *associator* (*Lie triple*) *system* of the Jordan algebra. As a generalization of (9) we have

$$[[S_a S_b]S_c] = S_{[abc]} , \tag{10}$$

and this shows that if $a \to S_a$ is a representation of $\mathfrak{A}$, then it is also a representation of the associator system, that is, it is a Lie triple system homomorphism of $\mathfrak{A}$ into the Lie triple system of linear transformations.

It is possible to define a universal Lie algebra $\mathfrak{L}_u$ for any Lie triple system $\mathfrak{T}$ in a manner similar to that indicated in the preceding section. If $\mathfrak{T}$ is a Lie triple system contained in a Lie algebra, the Lie algebra generated by $\mathfrak{T}$ is the set $\mathfrak{T} + [\mathfrak{T}\mathfrak{T}]$ of sums $a + \sum[b_i c_i]$. It is clear from this that if $\mathfrak{T}$ is finite-dimensional, then so is $\mathfrak{T} + [\mathfrak{T}\mathfrak{T}]$. In particular the universal Lie algebra of a finite-dimensional Lie triple system is finite-dimensional. Thus we can associate with every finite-dimensional Jordan algebra a finite-dimensional Lie algebra, the universal Lie algebra of its associator system. This association is very useful in the representation theory.

4. Representation theory for finite-dimensional algebras. At the present time we are in possession of some of the basic facts on representations of finite-dimensional Jordan algebras. These have been obtained by making extensive use of Lie algebra theory. Consequently we have had to assume that the base fields of our algebras are of characteristic 0.

We state now two of the main results of the representation theory of finite-dimensional algebras.

THEOREM A. *Every representation of a finite-dimensional semi-simple Jordan algebra $\mathfrak{A}$ of characteristic 0 is completely reducible.*

The analogous result for Lie algebras is required for the proof of this theorem. We recall that the Lie algebra result is best proved by establishing first a certain cohomology lemma due to Whitehead [9]. In the Jordan case we have found it

more convenient to reverse this procedure and prove Theorem A first. Using this result we can establish the following analogue of Whitehead's (first) lemma.

THEOREM B. *Let $\mathfrak{A}$ be as in the preceding theorem and let $a \to S_a$ be a representation of $\mathfrak{A}$ acting in the vector space $\mathfrak{M}$. Let $a \to f(a)$ be a linear mapping of $\mathfrak{A}$ into $\mathfrak{M}$ such that*

$$f(ab) = f(a)S_b + f(b)S_a .$$

Then there exist elements $w_i \in \mathfrak{M}$ and $b_i \in \mathfrak{A}$ such that

$$f(a) = \sum w_i(S_aS_{b_i} - S_{ab_i}).$$

The analogue for Jordan algebras of Levi's theorem on Lie algebras has been proved recently by Penico [20]: If $\mathfrak{A}$ is a finite-dimensional Jordan algebra of characteristic 0, then $\mathfrak{A} = \mathfrak{S} + \mathfrak{N}$ when $\mathfrak{N}$ is the radical (maximal solvable ideal) and $\mathfrak{S}$ is semi-simple. The analogues of the supplementary results on the Levi decomposition which are due to Malcev and Harish-Chandra ([8] and [18]) can be obtained for Jordan algebras by using Theorem B.

5. Jordan homomorphisms of rings. At the present time little is known on the structure of Jordan algebras of infinite dimensions or of Jordan rings (see, however, [19]). It is therefore somewhat premature to consider the representation theory for such systems. However, it does seem to be of interest to study certain special Jordan rings obtained from associative rings. For example, the problem of semi-automorphisms of rings which was introduced by Ancochea [4] can be regarded as a problem on automorphisms of Jordan rings. If $\mathfrak{A}$ is an associative ring, a *semi-automorphism* S of $\mathfrak{A}$ is a 1–1 mapping of $\mathfrak{A}$ onto itself satisfying

$$(a + b)^S = a^S + b^S, \qquad (ab + ba)^S = a^Sb^S + b^Sa^S. \tag{11}$$

Evidently this is just an automorphism of the special Jordan ring $\mathfrak{A}_J$ obtained from $\mathfrak{A}$ by replacing the associative composition ab by $\{ab\}$. Because of this it seems to be more appropriate to call S a *Jordan automorphism* of $\mathfrak{A}$. Also we can generalize this notion and consider Jordan homomorphisms of one associative ring into a second one. The second condition in (11) loses a good deal of its force if the rings have elements of additive order 2 and in order to be able to treat this case, too, it is necessary to replace this condition by

$$(a^2)^S = (a^S)^2, \qquad (aba)^S = a^Sb^Sa^S. \tag{12}$$

Mappings of this type have been studied by a number of writers ([4], [10], [12], and [17]). Recently C. E. Rickart and the present author have undertaken a systematic study of Jordan homomorphisms in the sense of (11) [15]. Some rather surprising results have been obtained. Thus, for example, we have shown that if $\mathfrak{A} = \mathfrak{B}_n$ is a ring of n by n, $n \geqq 2$, matrices over *any* ring $\mathfrak{B}$ with an identity, then any Jordan homomorphism of $\mathfrak{A}$ is obtained by combining in an obvious way an associative homomorphism and an associative anti-homomor-

phism. A similar result holds also for rings which are *locally matrix* in the sense that any finite subset of elements can be embedded in a subring $\mathfrak{B}_n$, $n \geqq 2$, $\mathfrak{B}$ with an identity. A noteworthy class of locally matrix rings is the class of simple rings possessing minimal one-sided ideals. Hence the Jordan homomorphisms of these rings have been completely determined. We have also obtained the Jordan automorphisms of primitive rings having minimal one-sided ideals. We recall that a ring is *primitive* if it has a 1–1 irreducible module. The Jordan automorphisms (and more generally the Jordan homomorphisms of one such ring onto another) are either automorphisms (homomorphisms) or anti-automorphisms (anti-homomorphisms).

The methods which we have used are based in part on matrix calculations, in part, on the identity

$$[(ab)^s - a^s b^s][(ab)^s - b^s a^s] = 0, \tag{13}$$

and in part on Lie ring techniques.

In addition to the Jordan rings obtained from associative rings by using $\{ab\}$, another "classical" type of Jordan ring is the following: Let $\mathfrak{A}$ be an associative ring which has an anti-automorphism I of period two. Then the subset $\mathfrak{H}(\mathfrak{A}, I)$ of self-adjoint elements of $\mathfrak{A}$ is a special Jordan ring. The set of ordinary symmetric matrices and the set of hermitian matrices are obvious examples of this type. Rickart and I have recently considered the problem of the homomorphisms of Jordan rings of this type and we have found that a substantial portion of the results for the rings $\mathfrak{A}_I$ have analogues for the rings $\mathfrak{H}(\mathfrak{A}, I)$. We hope to publish these results shortly.

Bibliography

1. A. A. Albert, *On a certain algebra of quantum mechanics*, Ann. of Math. vol. 35 (1934) pp. 65–73.

2. ———, *On Jordan algebras of linear transformations*, Trans. Amer. Math. Soc. vol. 59 (1946) pp. 524–555.

3. ———, *A structure theory for Jordan algebras*, Ann. of Math. vol. 48 (1947) pp. 446–467.

4. G. Ancochea, *On semi-automorphisms of division algebras*, Ann. of Math. vol. 48 (1947) pp. 147–154.

5. G. Birkhoff, *Representability of Lie algebras and Lie groups by matrices*, Ann. of Math. vol. 38 (1937) pp. 526–532.

6. G. Birkhoff and P. Whitman, *Representation of Jordan and Lie algebras*, Trans. Amer. Math. Soc. vol. 61 (1942) pp. 116–136.

7. S. Eilenberg, *Extensions of general algebras*, Annales de la Société Polonaise de Mathématique vol. 21 (1948) pp. 125–134.

8. Harish-Chandra, *On the radical of a Lie algebra*, Proc. Amer. Math. Soc. vol. 1 (1950) pp. 14–17.

9. G. Hochschild, *Semi-simple algebras and generalized derivations*, Amer. J. Math. vol. 64 (1941) pp. 677–694.

10. L. K. Hua, *On the automorphisms of a field*, Proc. Nat. Acad. Sci. U. S. A. vol. 35 (1949) pp. 386–389.

11. F. D. Jacobson and N. Jacobson, *Classification and representation of semi-simple Jordan algebras*, Trans. Amer. Math. Soc. vol. 65 (1949) pp. 141–169.

12. N. JACOBSON, *Isomorphisms of Jordan rings*, Amer. J. Math. vol. 70 (1948) pp. 317–328.

13. ———, *Lie and Jordan triple systems*, Amer. J. Math. vol. 71 (1949) pp. 149–170.

14. ———, *General representation theory of Jordan algebras*, Trans. Amer. Math. Soc. vol. 70 (1951) pp. 510–530.

15. N. JACOBSON and C. E. RICKART, *Jordan homomorphisms of rings*, Trans. Amer. Math. Soc. vol. 69 (1950) pp. 479–502.

16. P. JORDAN, E. WIGNER, and J. VON NEUMANN, *On an algebraic generalization of the quantum mechanical formalism*, Ann. of Math. vol. 35 (1934) pp. 29–64.

17. I. KAPLANSKY, *Semi-automorphisms of rings*, Duke Math. J. vol. 14 (1947) pp. 521–527.

18. A. MALCEV, *On the representation of an algebra as direct sum of the radical and a semi-simple algebra*, C. R. (Doklady) Acad. Sci. URSS. N.S. vol. 36 (1942) pp. 42–45.

19. J. VON NEUMANN, *On an algebraic generalization of the quantum mechanical formalism (Part I)*, Rec. Math (Mat. Sbornik) N.S. 4 (43) (1936) pp. 415–484.

20. A. J. PENICO, *The Wedderburn principal theorem for Jordan algebras*, to appear.

21. E. WITT, *Treue Darstellung Liescher Rings*, J. Reine Angew. Math. vol. 177 (1937) pp. 152–160.

YALE UNIVERSITY,
NEW HAVEN, CONN., U. S. A.

LES IDÉAUX MINIMAUX DANS LES ANNEAUX ASSOCIATIFS[1]

JEAN DIEUDONNÉ

1. On sait que les idéaux minimaux jouent un rôle fondamental dans l'étude classique des algèbres (associatives) de rang fini, et plus généralement dans celle des anneaux d'Artin (anneaux satisfaisant à la condition minimale pour les idéaux). J'ai montré en 1942 [2] comment la considération de ces idéaux dans un anneau (associatif) *quelconque* conduit aussi à des résultats intéressants, et permet entre autres de définir une classe nouvelle d'anneaux *simples*, qui généralisent directement les anneaux de matrices classiques, mais où la condition minimale n'est plus vérifiée. Au cours de ses importantes recherches sur la notion de "radical" et de "semi-simplicité", N. Jacobson est parvenu indépendamment, en 1945, à une partie de mes résultats, qu'il a étendus en 1947 à la catégorie des anneaux "primitifs" introduite par lui ([3], [4], et [5]). Je me propose dans ce qui suit de rappeler brièvement les points essentiels de la théorie des idéaux minimaux, et de montrer comment on peut déduire directement de mes résultats de 1942 ceux obtenus par Jacobson en 1947, ainsi que des théorèmes s'appliquant à des catégories d'anneaux plus générales.

2. Un idéal minimal à droite (resp. gauche) d'un anneau A est un sous-module *simple* de A, considéré comme A-module à droite (resp. gauche). Si $\mathfrak{r}$ est un tel idéal, on a $\mathfrak{r}^2 = \mathfrak{r}$ ou $\mathfrak{r}^2 = 0$; si $\mathfrak{r}$, $\mathfrak{r}'$ sont deux idéaux à droite minimaux *isomorphes* (en tant que A-modules), tout isomorphisme de $\mathfrak{r}$ sur $\mathfrak{r}'$ est de la forme $x \to ax$ (avec $a \in \mathfrak{r}'$); en outre, on a $\mathfrak{r}\mathfrak{r}' = \mathfrak{r}$ si $\mathfrak{r}'$ est idempotent, $\mathfrak{r}\mathfrak{r}' = 0$ si $\mathfrak{r}'$ est nilpotent.

La somme de *tous* les idéaux minimaux à droite (resp. gauche) d'un anneau A est ce que j'ai appelé le *socle droit* (resp. *gauche*) de A. Si $\mathfrak{r}_0$ est un idéal minimal à droite de A, la somme des idéaux à droite minimaux de A isomorphes à $\mathfrak{r}_0$ est appelée un *pied* du socle droit S. On démontre les propriétés suivantes [2]:

(a) Tout pied du socle droit S est un idéal bilatère de A, somme directe d'idéaux à droite minimaux de A tous isomorphes; S est un idéal bilatère de A, somme directe de ses pieds.

(b) Si $\mathfrak{a}$ est un pied du socle droit S, la somme $\mathfrak{b}$ des idéaux à droite minimaux nilpotents contenus dans $\mathfrak{a}$ est un idéal bilatère de A, qui est l'intersection de $\mathfrak{a}$ et de l'annulateur *à droite* de $\mathfrak{a}$ dans A.

(c) Si $\mathfrak{a}^2 \neq 0$, l'intersection de $\mathfrak{a}$ et de son annulateur *à gauche* dans A est réduite à 0.

(d) Si $\mathfrak{a}^2 \neq 0$, tout idéal à droite dans l'anneau $\mathfrak{a}$ est aussi un idéal à droite dans A.

3. Un anneau A est dit *quasi-simple à droite* (resp. *à gauche*) s'il est identique à un des pieds de son socle droit (resp. gauche) et si $A^2 \neq 0$. Soit K l'opposé du

[1] Cette communication était mentionnée sur le programme imprimé sous le titre *Minimal ideals*.

corps des endomorphismes d'un idéal à droite minimal de A. On démontre qu'il existe un espace vectoriel à gauche E sur K et un sous-espace vectoriel E' de l'espace dual E^* de E (espace vectoriel à droite sur K), non réduit à 0, tels que A soit *isomorphe* à l'anneau $\mathfrak{F}(E, E')$ des endomorphismes u de E, tels que $u^{-1}(0)$ soit l'intersection d'un nombre fini d'hyperplans de la forme $x'^{-1}(0)$, où $x' \in E'$ (ce qui implique que u est de rang *fini*); inversement, tout anneau $\mathfrak{F}(E, E')$ défini de cette façon est quasi-simple à droite. Tout *idéal à droite* de $\mathfrak{F}(E, E')$ peut être défini de la façon suivante: c'est l'ensemble des endomorphismes $u \in \mathfrak{F}(E, E')$ tels que $u(E) \subset H$, où H est un sous-espace vectoriel quelconque de E.

Si on tient compte des propriétés (a) et (d) du §2, on voit que la structure du socle droit d'un anneau quelconque peut être considérée comme complètement déterminée.

4. Pour qu'un anneau $\mathfrak{F}(E, E')$ soit *simple*, il faut et il suffit que la relation $\langle x, x' \rangle = 0$ pour tout $x' \in E'$ entraîne $x = 0$ dans E. Alors E peut être considéré comme sous-espace du dual E'^* de E', et on montre que l'application qui, à tout endomorphisme $u \in \mathfrak{F}(E, E')$, fait correspondre son *transposé* ${}^t u$, est un isomorphisme de $\mathfrak{F}(E, E')$ sur l'anneau *opposé* de $\mathfrak{F}(E', E)$. Cela montre en particulier que $\mathfrak{F}(E, E')$ a alors des idéaux minimaux *à gauche*, tous isomorphes, et est somme de ses idéaux minimaux à gauche (autrement dit, est identique à un pied de son socle gauche).

Si on considère sur E la topologie $\sigma(E, E')$ (topologie de la convergence simple dans E'), $\mathfrak{F}(E, E')$ peut encore être défini comme l'anneau des *endomorphismes continus* de E qui sont de *rang fini*. La topologie de la *convergence uniforme* dans E est compatible avec la structure d'anneau de $\mathfrak{F}(E, E')$, et un système fondamental de voisinages de 0 pour cette topologie est formé des *annulateurs à droite des parties finies* de $\mathfrak{F}(E, E')$ (qui sont des idéaux à droite de $\mathfrak{F}(E, E')$). Soit $\bar{E}$ le *complété* de E (pour la topologie $\sigma(E, E')$), identique au dual algébrique E'^* de E'; le *complété* de l'anneau topologique $\mathfrak{F}(E, E')$ peut être identifié à l'anneau $\mathfrak{C}(\bar{E}, E')$ de *tous les endomorphismes de $\bar{E}$ continus pour la topologie* $\sigma(\bar{E}, E')$. Il convient de noter que le socle droit de $\mathfrak{C}(\bar{E}, E')$ est $\mathfrak{F}(\bar{E}, E')$, qui contient $\mathfrak{F}(E, E')$ et en est en général distinct; on vérifie en outre aisément que tout idéal à droite minimal dans $\mathfrak{C}(\bar{E}, E')$ est *fermé* dans cet anneau. Enfin, les sous-anneaux de $\mathfrak{C}(\bar{E}, E')$, dont le socle est $\mathfrak{F}(E, E')$, peuvent être caractérisés comme les sous-anneaux (contenant $\mathfrak{F}(E, E')$) de l'anneau $\mathfrak{C}(E, E')$ des endomorphismes continus *de l'espace* E: en effet, si u est un endomorphisme continu de $\bar{E}$ tel que, pour un $x \in E$, on ait $u(x) = a \notin E$, pour l'idéal à droite minimal $\mathfrak{r} \subset \mathfrak{F}(E, E')$ formé des endomorphismes v tels que $v(E) \subset Kx$, $u\mathfrak{r}$ est un idéal minimal de $\mathfrak{C}(\bar{E}, E')$ qui n'est pas contenu dans $\mathfrak{F}(E, E')$.

5. Après cette étude des anneaux quasi-simples et des anneaux simples ayant des idéaux minimaux, revenons à l'étude des socles d'un anneau quelconque A. Si A possède des idéaux minimaux nilpotents, les relations entre socle droit et socle gauche de A sont assez complexes (voir [1] pour l'étude du cas où A est

un anneau d'Artin). Toutefois, si $\mathfrak{a}$ est un pied du socle droit S de A *ne contenant pas d'idéal nilpotent*, $\mathfrak{a}$ est aussi un pied du socle gauche S' de A: en effet, $\mathfrak{a}$ est alors un anneau simple, donc contient des idéaux à gauche minimaux (§4); en outre, si $\mathfrak{l}$ est un idéal à gauche minimal *dans l'anneau* $\mathfrak{a}$, $\mathfrak{a}\mathfrak{l}$ est un idéal à gauche dans A, contenu dans $\mathfrak{l}$, et qui ne peut être réduit à 0, puisque l'annulateur à droite de $\mathfrak{a}$ a une intersection avec $\mathfrak{a}$ réduite à 0 (§2 b); on a donc $\mathfrak{a}\mathfrak{l} = \mathfrak{l}$, ce qui prouve que tout idéal à gauche dans $\mathfrak{a}$ est aussi un idéal à gauche dans A, et par suite que $\mathfrak{a}$ est un pied de S'.

6. Nous allons nous borner désormais à l'étude des anneaux A *ne contenant pas d'idéaux nilpotents*. Il résulte du §5 que, dans un tel anneau, les socles droit et gauche sont identiques à un même idéal bilatère S, et que S est *somme directe d'anneaux simples de la forme* $\mathfrak{F}(E_\alpha, E'_\alpha)$. Désignons par T *l'annulateur à droite* de S dans A; c'est un idéal bilatère de A, et on a $S \cap T = 0$ puisqu'il n'y a pas d'idéaux nilpotents dans A; donc $TS \subset S \cap T = 0$, ce qui montre que T est aussi l'annulateur à gauche de S.

Cela étant, considérons sur A la topologie $\mathfrak{T}$ dont un système fondamental de voisinages de 0 est formé des *annulateurs à droite des parties finies de* $S + T$. Cette topologie est *compatible* avec la structure d'anneau de A: en effet, comme tout voisinage de 0 du système précédent est un idéal à droite, pour tout $x_0 \in A$, $y \to yx_0$ est continue au point $y = 0$, et $(x, y) \to xy$ continue au point $(0, 0)$; enfin $y \to x_0 y$ est continue au point $y = 0$, car si $a \in S + T$, la relation $(ax_0)y = 0$ équivaut à $a(x_0 y) = 0$, et $ax_0 \in S + T$.

La topologie $\mathfrak{T}$ est *séparée*, car l'intersection de tous les voisinages de 0 est un idéal annulant à droite $S + T$, donc contenu dans T; mais comme il n'y a pas d'idéal nilpotent dans A, cette intersection est réduite à 0.

Soit alors $\bar{A}$ le *complété* de l'anneau topologique A, $\bar{S}$ et $\bar{T}$ les adhérences de S et T dans $\bar{A}$ (qui sont isomorphes aux complétés des sous-anneaux S et T de A). On a les propriétés suivantes:

(a) Un système fondamental de voisinages de 0 dans $\bar{A}$ est encore constitué par les annulateurs à droite *dans* $\bar{A}$ des parties finies de $S + T$; en effet, l'annulateur à droite $\mathfrak{r}_1$ dans $\bar{A}$ d'une partie finie F de $S + T$ contient l'adhérence de l'annulateur à droite $\mathfrak{r}$ de F dans A; comme $\mathfrak{r}$ est un idéal ouvert dans A, son adhérence dans $\bar{A}$ est un idéal ouvert dans $\bar{A}$, et a fortiori $\mathfrak{r}_1$ est ouvert dans $\bar{A}$; mais comme A est partout dense dans $\bar{A}$, cela entraîne que $\mathfrak{r}_1$ est l'adhérence de $\mathfrak{r}$.

(b) Comme la topologie de $\bar{A}$ est séparée, l'annulateur à droite de $S + T$ dans $\bar{A}$ est réduit à 0. Soit alors $\mathfrak{l} \neq 0$ un idéal à gauche dans $\bar{A}$, et $x \neq 0$ un élément de $\mathfrak{l}$; par hypothèse, il existe $a \in S + T$ tel que $ax \neq 0$; comme A est partout dense dans $\bar{A}$, il existe $y \in A$ tel que $a(x - y) = 0$, d'où $ay = ax \neq 0$; mais $ax \in \mathfrak{l}$, donc $ay \in \mathfrak{l} \cap A^2$; en d'autres termes, $\mathfrak{l} \cap A^2$ *n'est pas réduit à* 0. Cela montre aussitôt en particulier que $\bar{A}$ *ne contient pas d'idéaux nilpotents*.

(c) La topologie induite par $\mathfrak{T}$ sur S est celle pour laquelle un système fondamental de voisinages de 0 est formé des annulateurs à droite (dans S) des parties

finies de S. D'après le §4, $\bar{S}$ est isomorphe à l'anneau *produit* des anneaux topologiques $\mathfrak{C}(\bar{E}_\alpha, E'_\alpha)$. En particulier, $\bar{S}$ admet un *élément unité* e, et la décomposition de Peirce $x = ex + (x - ex)$ montre que $\bar{A}$ est somme directe de $\bar{S}$ et de son annulateur (à gauche et à droite) $T_1 \supset \overline{T}$.

(d) *L'anneau T_1 ne contient pas d'idéaux minimaux* (en d'autres termes, tous les idéaux minimaux de $\bar{A}$ sont contenus dans $\bar{S}$). Supposons le contraire, et soit $\mathfrak{r}$ un idéal à droite de longueur finie dans T_1. Les voisinages de 0 dans $\mathfrak{r}$ pour la topologie induite par celle de $\bar{A}$ sont des idéaux à droite dans T_1, donc leur longueur est bornée par celle de $\mathfrak{r}$, et une suite strictement décroissante de tels idéaux est donc finie, ce qui prouve que la topologie induite sur $\mathfrak{r}$ est *discrète.* Nous allons en déduire que $\mathfrak{r} \cap T = 0$. En effet, dans le cas contraire, $\mathfrak{r} \cap T$ serait un idéal à droite non nul de T; comme $\mathfrak{r}$ est discret, tout idéal à droite $\mathfrak{r}'$ de T contenu dans $\mathfrak{r} \cap T$ est identique à son adhérence *dans* $\mathfrak{r}$; mais cette adhérence est l'intersection de $\mathfrak{r}$ et de l'adhérence de $\mathfrak{r}'$ *dans* $\bar{A}$, et par suite c'est un idéal à droite dans T_1 ; $\mathfrak{r} \cap T$ serait par suite un idéal de longueur finie dans T. Or, il n'existe aucun idéal $\neq 0$ de cette nature dans T: en effet, si $\mathfrak{r}_1$ est un idéal à droite minimal *dans* T, $\mathfrak{r}_1 T$ n'est pas nul, puisqu'aucun élément $\neq 0$ de T n'annule T; $\mathfrak{r}_1 T$ est alors un idéal à droite *dans* A, contenu dans T et par suite égal à $\mathfrak{r}_1$, et $\mathfrak{r}_1$ est alors un idéal à droite minimal *dans* A, ce qui est contraire à la définition de T.

Considérons alors le *socle* U de T_1 ; s'il n'était pas nul, chacun de ses éléments $\neq 0$ appartiendrait à un idéal à droite de T_1, de longueur finie, donc ne serait pas dans T; autrement dit, on aurait $U \cap T = 0$; mais cela est contradictoire avec (b), puisque U est un idéal à gauche dans $\bar{A}$.

(e) *Si l'anneau A est sans radical* (au sens de Chevalley-Jacobson) *il en est de même de $\bar{A}$.* Remarquons d'abord que le radical de A est toujours contenu dans T, et celui de $\bar{A}$ dans T_1 ; tout revient donc à prouver que si T est sans radical, il en est de même de T_1. Soit R_1 le radical de T_1 ; s'il est $\neq 0$, comme c'est un idéal à gauche de $\bar{A}$, sa trace sur T n'est pas nulle d'après (b), donc $R = R_1 \cap \overline{T}$, qui est le radical de $\overline{T}$, n'est pas nul; nous allons en déduire que le radical de T n'est pas nul. D'après (b), $R \cap T^2$ n'est pas nul; soit $z \neq 0$ un de ses éléments. Par hypothèse, lorsque y parcourt $\overline{T}$, l'ensemble des $y + yz$ est identique à $\overline{T}$ [4, pp. 302–303]; lorsque x parcourt T, l'ensemble des $x + xz$ est donc *partout dense* dans $\overline{T}$. Soient alors a et b deux éléments quelconques de T; par définition de la topologie de $\overline{T}$, il existe $x \in T$ tel que $a(b - x - xz) = 0$, c'est-à-dire $ab = (ax) + (ax)z$. L'ensemble des $u + uz$ où $u \in T$, contient donc T^2, et en particulier, il existe $u \in T$ tel que $-z = u + uz$, autrement dit, z a un *quasi-inverse à gauche* dans T. Pour tout $c \in T$, on a encore $cz \in R \cap T^2$, donc cz a aussi un quasi-inverse à gauche dans T, ce qui prouve que z appartient au radical de T [4, pp. 302–303].

7. Si $T = 0$, on a $T_1 = 0$, car T_1, idéal à gauche de $\bar{A}$, doit avoir d'après le §6 (b), une intersection avec A non réduite à 0 s'il est lui-même $\neq 0$; dans ce cas, on a $\bar{A} = \bar{S}$, d'où $S \subset A \subset \bar{S}$, on retrouve un résultat récent de P.

Jaffard [6]; en particulier, si A est primitif et admet des idéaux minimaux, on obtient la caractérisation de ces anneaux donnée par N. Jacobson [5]. Mais en général, lorsque $T \neq 0$, on peut avoir $\overline{T} \neq T_1$.[2]

Remarquons que, dans $\bar{A}$, l'annulateur de $\overline{T}$ (à droite et à gauche) est $\overline{S}$, car il contient $\bar{S}$ et ne peut contenir aucun élément $\neq 0$ de T_1, sans quoi son intersection avec T_1 serait un idéal à gauche $\neq 0$ dans $\bar{A}$, et aurait donc une intersection $\neq 0$ avec T (§6 (b)) ce qui est contraire à l'hypothèse. L'intersection $\overline{S} \cap A = S_0$ est donc *l'annulateur de T dans A* (à droite et à gauche); on peut aussi le caractériser comme *le plus grand idéal de A (à droite ou à gauche) contenant S et ne contenant aucun élément annulant S* (à droite ou à gauche): car si un idéal à gauche $\mathfrak{a}$ contient S et un élément $x_0 \notin S_0$, il contient Tx_0, qui par hypothèse n'est pas nul et est contenu dans T, donc annule S. Remarquons encore que S est le *socle de l'anneau* S_0, car si $\mathfrak{r}$ est un idéal à droite minimal *dans* S_0, il contient $\mathfrak{r}S_0$, qui n'est pas nul puisque $\mathfrak{r}S \neq 0$, donc $\mathfrak{r} = \mathfrak{r}S_0$, et $\mathfrak{r}$ est un idéal minimal *dans* A, donc contenu dans S.

Bibliographie

1. J. Dieudonné, *Sur les systèmes hypercomplexes*, J. Reine Angew. Math. t. 184 (1942) pp. 178–192.

2. ———, *Sur le socle d'un anneau et les anneaux simples infinis*, Bull. Soc. Math. France t. 70 (1942) pp. 46–75.

3. N. Jacobson, *Structure theory of simple rings without finiteness assumptions*, Trans. Amer. Math. Soc. t. 57 (1945) pp. 228–245.

4. ———, *The radical and semi-simplicity for arbitrary rings*, Amer. J. Math. t. 67 (1945) pp. 300–320.

5. ———, *On the theory of primitive rings*, Ann. of Math. t. 48 (1947) pp. 8–21.

6. P. Jaffard, *Détermination de certains anneaux*, C. R. Acad. Sci. Paris t. 229 (1949) pp. 805–806.

University of Nancy,
Nancy, France.

[2] Soit E un espace vectoriel admettant une base dénombrable (e_n), et soit E' le sous-espace du dual E^* engendré par les formes coordonnées e'_n. Dans le produit $\mathfrak{L}(E, E') \times \mathbf{Z}$, où $\mathbf{Z}$ est l'anneau des entiers rationnels, le socle S est $\mathfrak{F}(E, E')$; soit A le sous-anneau de ce produit engendré par les éléments $(s, 2n)$, où $s \in \mathfrak{F}(E, E')$ et $n \in \mathbf{Z}$, et par l'élément $(u, 1)$, où u est l'application linéaire de E dans lui-même définie par $u(e_n) = e_{n+1}$ pour tout n. Comme u^n tend vers 0 dans $\mathfrak{L}(E, E')$ lorsque n croît indéfiniment, on voit aisément que $\overline{T}$ se compose des éléments $(0, 2n)$, tandis que T_1 est formé des éléments $(0, n)$.

ON TWO TOPICS IN THE STRUCTURAL THEORY OF RINGS (GALOIS THEORY OF RINGS AND FROBENIUS ALGEBRAS)

TADASI NAKAYAMA

The present address is concerned with two topics in the theory of rings. The topics are rather different from each other, but they have in common the fact that the writer's study of them is mainly concerned with non-semisimple rings.

The first is the *Galois theory of rings*. Galois theory was first extended to skew-fields by Jacobson [12] some ten years ago; the theory may be called an outer one and generalizes the classical Galois theory of fields. A further advance of noncommutative Galois theory was made by Cartan ([8], cf. [7]) and Jacobson [13], independently, so as to include the theory of commuters (commutators). The study was pushed further by Dieudonné [9] in a work which happened to have much in common with one by Azumaya and the writer [22]. Hochschild [11] has extended to the noncommutative case the non-normal Galois theory, so to speak, of Kaloujnine and Jacobson [14]. These are however all concerned with skew-fields, or matric rings over skew-fields. On the other hand, Azumaya [4] succeeded in extending the (outer) Galois theory (first to simple rings and then) to uni-serial rings, in the sense of Köthe (see e.g. [2]). The writer then established it for general rings with minimum condition [17]. As for the inner Galois theory, i.e., the theory of commuters, the writer studied it also for non-semisimple rings [18] and obtained a result for them from which the main 1–1 duality in the Brauer-Noether-Shoda-Artin-Whaples theory (see [1]) of commuters in simple rings may be drawn. He has then combined these to obtain the mixed Galois theory for primary rings [19], which generalizes the Cartan-Jacobson one for skew-fields. The writer has also studied the non-normal Galois theory for general rings [20]. Let him present some features of these studies.

Firstly, the outer Galois theory is rather satisfactory. Let, namely, R be a ring with unit element and satisfying the minimum condition. For an R-two-sided module $\mathfrak{m}$ and an automorphism σ of R, we can introduce a new R-two-sided module $(\sigma, \mathfrak{m})$ which is isomorphic, or rather identical, with $\mathfrak{m}$ as R-right-module while the left operation of R on it is defined by $a \cdot u = a^{\sigma}u$. Then we call a finite group $G = \{1, \sigma, \cdots, \tau\}$ of automorphisms of R a *Galois group* of R when no module (σ, R) with $\sigma \in G$, $\neq 1$ has a composition residue-module isomorphic to a composition residue-module of $R = (1, R)$. Our theory centers on the Galois correspondence:[1] *If R is directly indecomposable and G is a Galois group of R, then the subgroups of G are in 1–1 correspondence with those subrings of R which contain the invariant system S of G in R and over which R has independent right or left bases. G exhausts the automorphisms of R leaving S elementwise fixed.* Our notion of Galois group thus seems appropriate, though rather

[1] Here we neglect another feature in Galois theory which is concerned with the extendability of isomorphisms of subrings. This the writer wants however to discuss elsewhere.

special, if we intend to extend the classical Galois theory to general rings. We have also the *theorem of the normal basis* that the group ring (G, S) and R are (G, S)- (right-, or left-) isomorphic; we have even the (G, R)-isomorphism of R^g and (G, R), where R^g denotes the direct sum of g copies of R, g being the order of G, and (G, R) is the semilinear group-ring (or, crossed-product with unit factor set) of G over R. This holds in fact without the assumption of direct indecomposability of R and under the assumption that G induces a Galois group in the residue-ring of R modulo the radical N. One of the arguments used in our proof may be formulated in the following generalized crossed-product argument [17, Lemma 5]: Let R be a ring with unit element and with minimum condition, and N be its radical. Let a finite group of automorphism-classes of R induce a Galois *class*-group in the residue-ring R/N, where the Galois class-group is defined similarly as a Galois group. Consider a crossed product (G, R), with perhaps a nonunit factor set, and two (G, R)-right-moduli $\mathfrak{r}$, $\mathfrak{s}$, which are direct sums of R-submoduli R-isomorphic to directly indecomposable right-ideal components of R. Now, *if* $\mathfrak{r}/\mathfrak{r}N$ *and* $\mathfrak{s}/\mathfrak{s}N$ *are* R*-isomorphic, then* $\mathfrak{r}$ *and* $\mathfrak{s}$ *are* (G, R)*-isomorphic*; observe that both the moduli and the operator domain are enlarged in the conclusion. The notion of regular moduli is also effectively used; we call a right-, say, module $\mathfrak{m}$ of a ring R with unit element *regular* when $\mathfrak{m}^v$ and R^u, with suitable integers u, v, are isomorphic; the number u/v, uniquely determined, is called its rank. In the above Galois correspondence we may replace the requirement of the existence of bases with respect to subrings, which are associated with subgroups of the Galois group, by mere regularity. Rather general considerations of regular moduli and their endomorphism rings together with the above crossed-product theorem are sufficient to give half of our Galois duality, while further investigations making use of the full property of the Galois group are needed in order to obtain the other half [17, §2].

As for the theory of commuters, we have the following theorem: Let K be a subring of R, a ring with unit element and with minimum condition, containing the center of R and such that R_lK_r has an independent finite basis over R_l consisting of elements of K_r, where R_l and K_r denote, respectively, the ring of left multiplications of elements of R, and the ring of right multiplications of elements of K on R and where the product is formed in the absolute endomorphism ring of the module R. Let further R be regular with respect to R_lK_r. Let $S = V_R(K)$ be the commuter system of K in R. (It amounts to considering a subring S of R such that the S_r-endomorphism ring of R has an independent finite R_l-basis consisting of elements of R_r and such that R is S_r-regular (or, more specifically, R has an independent S_r-basis).) Then, *subrings* L, T *of* K, R *such that, respectively,* R *is* R_lL_r*-regular,* $R_l L_r \cap K_r = L_r$, *and* R *is* T*-right-regular,* $T \supseteq S$, *are mutually in* 1–1 *correspondence according to* $V_K(T) = V_R(T) = L$, $V_R(L) = T$.

Let now R be a *primary* ring, with unit element and minimum condition, and Φ be a group of automorphisms of R which satisfies the following conditions: (i) the subring K of R defined by $K_r = R_l\Phi \cap R_r$ has an independent finite basis over the center Z of R, (ii) K mod $K \cap N$ is a skew-field, where N denotes

the radical of R, (iii) R is R_lK_r-regular, (iv) every (ring-) automorphism of R contained in R_lK_r is contained in Φ, (v) the (normal) subgroup $\Phi_0 = \Phi \cap R_lK_r$ has a finite index in Φ, and (vi) for a representative system $\{\rho_i\}$ $(\rho_1 = 1)$ of Φ mod Φ_0, no R-two-sided module (ρ_i, R) with $i \neq 1$ has a composition residue-module isomorphic to a composition residue-module of R. *Then Φ exhausts the automorphisms of R which leave its invariant system S elementwise fixed, and R has an independent S-right-basis of $(\Phi:\Phi_0)(K:Z)$ terms. Further, subrings T of R, which contain S and over which R has independent right-bases, are in 1–1 correspondence with subgroups Ψ of Φ such that L defined by $L_r = R_l\Psi \cap R_r$ have independent bases over Z and all automorphisms of R contained in R_lL_r are contained in Ψ.* Though the setting is rather complicated, this specializes to the Cartan-Jacobson theory in case R is a skew-field.[2] It is desirable to replace here the primarity assumption on R by mere direct indecomposability (and the skew-field property of $K/K \cap N$ by simplicity).[3]

In general, subrings $(\ni 1)$ of R, a ring with unit element and minimum condition, over which R has independent right-bases are in 1–1 correspondence with subrings of the absolute endomorphism ring of R, as module, which contain R_l and for which R is regular with the inverse of an integer as rank. This may already be considered as the non-normal Galois theory of R, such a subring of the absolute endomorphism ring may be called a *Galois ring* of module-endomorphisms of R. It has an independent right-basis over R_l, and in fact one which can be obtained by constructing the direct self-product of R over the corresponding subring S. This construction is a special instance of what we can formulate by means of R-double-moduli: Let namely $\mathfrak{m}$ be an R-double-module having an R-right-basis $(u_1, u_2, \cdots, u_m)$, and u_0 be an element of $\mathfrak{m}$. Let m endomorphisms μ_i of R be defined by $au_0 = u_1a^{\mu_1} + u_2a^{\mu_2} + \cdots + u_ma^{\mu_m}$ $(a \in R)$. We call the R-right- (in fact, two-sided) module generated by $\mu_1, \mu_2, \cdots, \mu_m$ the relation module of u_0 in $\mathfrak{m}$; it is independent of the special choice of our R-basis of $\mathfrak{m}$. If $\mathfrak{n}$ is a second R-double-module having R-right-basis and $v_0 \in \mathfrak{n}$, then the relation module of $u_0 + v_0$ (resp. u_0v_0) in $\mathfrak{m} \oplus \mathfrak{n}$ (resp. $\mathfrak{m} \times_R \mathfrak{n}$) is the sum (resp. product) of those of u_0, v_0 in $\mathfrak{m}, \mathfrak{n}$. Now, whenever $u_1, u_2, \cdots, u_m \in Ru_0$, $au_0 = 0$ $(a \in R)$ implies $a = 0$, and our relation module forms a ring; it is a Galois ring, of (module-)endomorphisms of R (and conversely). We can further study the structure of such a Galois ring in terms of the decomposition of the underlying module $\mathfrak{m}$.

However, the difficulties in our above Galois theories of specified types lie in showing that certain given rings of endomorphisms of R are really Galois rings and moreover that certain special subrings of them have a certain specified particular structure, as, for instance, having an R_l-basis consisting of ring-automorphisms of R rather than general module-endomorphisms of R. In many parts of our above theories, not to speak of the non-normal one, the assumption of the minimum condition may be replaced by some weaker conditions. Leaving

[2] Cf. footnote 1.

[3] This latter will be treated in a forthcoming work of the writer.

however the axiomatic analysis aside, we only mention the case of closed primitive rings ([8], [22]).[4]

Now we turn to our second topic, *Frobenius algebras*. A Frobenius algebra is an algebra, over a certain field, which has a unit element and whose left and right regular representations are equivalent. Its importance was observed by Brauer and Nesbitt in connection with their work on modular representations of finite groups. In collaborating with them, the writer obtained the structural theory of Frobenius and related algebras [16]. It was found convenient to introduce a quasi-Frobenius algebra as an algebra having a unit element and such that the totality of distinct directly indecomposable components of its left regular representation coincides with that of its right regular representation; in other words, it is an algebra whose core-algebra (cf. e.g. [16, I, §3] or [18, §4][5]) is Frobeniusean. Our study starts with a certain structural characterization of Frobenius and quasi-Frobenius algebras; they are characterized by means of certain simple properties of minimal left and right ideals, contained in directly indecomposable left and right components. The characterization enables us also to extend the notions to the case of rings, satisfying the minimum condition, rather than algebras. It turns out, now, that a ring satisfying the minimum condition is quasi-Frobeniusean if and only if its left ideal lattice is *dual-isomorphic* to its right ideal lattice. In fact, the *annihilator correspondence*: $\mathfrak{l} \to r(\mathfrak{l})$, $\mathfrak{r} \to l(\mathfrak{r})$ gives such a dual-isomorphism, where $\mathfrak{l}$, $\mathfrak{r}$ represent, respectively, left and right ideals, and r, l are the right and the left annihilator operators in our ring. The ring is, furthermore, Frobeniusean if and only if there exists such a dual-isomorphism satisfying a certain dimension relation, which amounts to the duality of the ranks of corresponding left and right ideals in case the ring is an algebra. It seems of interest that the representation-theoretical properties defining Frobenius and quasi-Frobenius algebras are equivalent to these lattice and annihilation properties of ideals.[6] A residue ring of a Frobenius ring, modulo a certain two-sided ideal $\mathfrak{z}$, is Frobeniusean if and only if $l(\mathfrak{z})$, $r(\mathfrak{z})$ are left-, right-principal. This indicates the significance of Frobenius rings for the theory of principal ideal rings and provides a viewpoint of looking at the latter.[7]

Another interesting feature of Frobenius algebras is the *orthogonality relation*. There exists a certain particular class of automorphisms in a Frobenius algebra which describe, in a sense, the representation-theoretical significance of the above annihilation duality in it. By making use of those automorphisms, Nesbitt and the writer obtained orthogonality relations for the coefficients of the (normalized) regular representations of Frobenius algebras (see [6], [24]), which generalize the well-known orthogonality relations for the coefficients of (ordinary) irreducible representations of (finite) groups. Brauer [6] discussed further

[4] Primitive rings were called irreducible rings in [22].

[5] It is called also the basic algebra.

[6] The dual-isomorphism of left and right ideal lattices has been considered from somewhat different standpoints in [5] and [15]. For the annihilation duality see [10] too.

[7] For principal ideal rings with minimum condition see [2].

the arithmetical significance of those orthogonality relations and obtained a generalization of Speiser's theorem on modular behaviour of an irreducible representation.

These are, however, somewhat older results. Somewhat newer is the study of almost symmetric rings by Azumaya [3]; the notion forms a ring approximation of *symmetric algebras* which are Frobenius algebras whose left and right regular representations with respect to a common basis can be transformed into each other by a symmetric matrix. A ring with unit element, and with minimum condition, is Frobeniusean if and only if the left and the right annihilators of the radical are right- and left-principal, respectively; in fact, they coincide when the condition holds, i.e., when the ring is Frobeniusean. Now, an almost symmetric ring is, by definition, a Frobenius ring in which our (two-sided) annihilator of the radical is a principal ideal generated by an element of the center. In the case of algebras the notion turns out, as Azumaya and the writer showed, to be equivalent to that of absolutely weakly symmetric algebras, where weakly symmetric algebras are those Frobenius algebras such that the (directly indecomposable) representations obtained by the left and the right ideals generated by the same primitive idempotent element are equivalent (see [23]).

Some of the properties of Frobenius and quasi-Frobenius algebras may be assumed individually and separately as postulates to introduce their weakened modifications. A recent work of Thrall [25] is along this line. The properties considered by him are mainly representation-theoretical. It seems of some interest to carry out a similar study with respect to the ideal-lattice and annihilator properties.

In closing, and in our context, the writer may cite a recent result, of Ikeda and himself [21], that in the case of algebras our above necessary and sufficient condition for Frobenius or quasi-Frobenius algebras may be cut in half. Namely, if $l(r(\mathfrak{l})) = \mathfrak{l}$ for every left ideal $\mathfrak{l}$ in an algebra, then $r(l(\mathfrak{r})) = \mathfrak{r}$ for every right ideal $\mathfrak{r}$ in it (whence the algebra is quasi-Frobeniusean. If moreover the sum of the ranks of $\mathfrak{l}$ and $r(\mathfrak{l})$, over the ground-field, is equal to the rank of the algebra for every $\mathfrak{l}$, then the algebra is Frobeniusean).

It seems to the writer that our topics possess a somewhat deeper connection with each other than was said in the beginning.

References

1. E. Artin and G. Whaples, *The theory of simple rings*, Amer. J. Math. vol. 65 (1943).

2. K. Asano, *Über Hauptidealringe mit Kettensatz*, Osaka Mathematical Journal vol. 1 (1949).

3. G. Azumaya, *On almost symmetric algebras*, Japanese Journal of Mathematics vol. 19 (1948).

4. ———, *Galois theory for uni-serial rings*, Journal of the Mathematical Society of Japan vol. 1 (1949).

5. R. Baer, *Rings with duals*, Amer. J. Math. vol. 65 (1943).

6. R. Brauer, *On hypercomplex arithmetic and a theorem of Speiser*, A. Speiser-Festschrift, Zürich, 1945.

7. ———, *On a theorem of H. Cartan*, Bull. Amer. Math. Soc. vol. 55 (1949).

8. H. CARTAN, *Théorie de Galois pour les corps non commutatifs*, Ann. École Norm. vol. 65 (1948).

9. J. DIEUDONNÉ, *La théorie de Galois des anneaux simples et semi-simples*, Comment. Math. Helv. vol. 21 (1948).

10. M. HALL, *A type of algebraic closure*, Ann. of Math. vol. 40 (1939).

11. G. HOCHSCHILD, *Double vector spaces over division rings*, Amer. J. Math. vol. 71 (1949).

12. N. JACOBSON, *The fundamental theorem of Galois theory for quasi-fields*, Ann. of Math. vol. 41 (1940).

13. ———, *An extension of Galois theory to non-normal and non-separable fields*, Amer. J. Math. vol. 66 (1944).

14. ———. *A note on division rings*, Amer. J. Math. vol. 69 (1947).

15. I. KAPLANSKY, *Dual rings*, Ann. Math. vol. 49 (1948).

16. T. NAKAYAMA, *On Frobeniusean algebras*, I, Ann. of Math. vol. 40 (1939); II, Ann. of Math. vol. 42 (1941); III, Japanese Journal of Mathematics vol. 18 (1942).

17. ———, *Galois theory for general rings with minimum condition*, Journal of the Mathematical Society of Japan vol. 1 (1949).

18. ———, *Commuter systems in a ring with radical*, Duke Math. J. vol. 16 (1949).

19. ———, *Generalized Galois theory for rings with minimum condition*, forthcoming in Amer. J. Math.

20. ———, *Non-normal Galois theory for non-commutative and non-semisimple rings*, forthcoming in Canadian Journal of Mathematics.

21. ———, *Supplementary remarks on Frobeniusean algebras*, I, Proceedings of Japan Academy vol. 25 (1949).

T. NAKAYAMA and M. IKEDA, II, Osaka Mathematical Journal vol. 2 (1950).

22. T. NAKAYAMA and G. AZUMAYA, *On irreducible rings*, Ann. of Math. vol. 48 (1947).

23. T. NAKAYAMA and C. NESBITT, *Note on symmetric algebras*, Ann. of Math. vol. 39 (1938).

24. C. NESBITT and R. THRALL, *Some ring theorems with applications to modular representations*, Ann. of Math. vol. 47 (1946).

25. R. THRALL, *Some generalizations of quasi-Frobenius algebras*, Trans. Amer. Math. Soc. vol. 64 (1948).

NAGOYA UNIVERSITY,
NAGOYA, JAPAN.

ARITHMETIC ALGEBRA

MODERN DEVELOPMENT OF ALGEBRAIC NUMBER THEORY AND CLASS FIELD THEORY

EMIL ARTIN

This address was given as part of the Conference in Algebra, but no manuscript has been received by the editors.

JACOBSONSCHES RADIKAL UND HILBERTSCHER NULLSTELLENSATZ[1]

WOLFGANG KRULL

Bekanntlich hat man für das *Wedderburnsche Theorem*, daß jedes einfache hyperkomplexe System ein voller Matrizenring über einem Schiefkörper ist, eine denkbar weitgehende und völlig befriedigende Verallgemeinerung gefunden, bei der an Stelle des einfachen hyperkomplexen Systems ein einfacher Ring im Sinne von *Chevalley* und *Jacobson* tritt. Die Einführung des *Jacobsonschen Radikals* lieferte dann auch zu dem zweiten Wedderburnschen Satz, nach dem jedes hyperkomplexe System mit verschwindendem Radikal die direkte Summe von endlich vielen einfachen Systemen ist, ein ganz allgemeines und äußerlich sehr ähnliches Gegenstück: Ein beliebiger (i.A. nichtkommutativer, aber assoziativer) Ring ist dann und nur dann eine *subdirekte* Summe von einfachen Ringen, wenn sein Jacobsonsches Radikal verschwindet.[2] Leider ist aber in diesem zweiten Fall die Analogie zwischen dem speziellen und dem allgemeinen Theorem nur formal. Der Wedderburnsche Satz enthält eine völlig befriedigende Strukturaussage; denn nach ihm übersieht man, angesichts der Definition der direkten Summe, alle halbeinfachen Systeme vollständig, wenn man nur die einfachen Systeme beherrscht. Dagegen ist der Begriff der subdirekten Summe so allgemein und in gewissem Sinne nichtssagend, daß man von vornherein nicht hoffen darf, bei einer subdirekten Zerlegung aus der bekannten Struktur der Komponentenringe tiefergehende Schlüsse über den Bau des Gesamtringes ziehen zu können. In der Tat kann man schon im einfachen Spezialfall der kommutativen Integritätsbereiche leicht Ringe von ganz verschiedenem arithmetischem Typus angeben, die sich als subdirekte Summen ein und derselben Körpermenge darstellen lassen.[3]

Unter diesen Umständen ist es besonders bemerkenswert, daß gerade bei den Integritätsbereichen bzw. allgemeiner bei beliebigen *kommutativen Ringen mit Einheitselementen*, auf deren Betrachtung wir uns weiterhin beschränken wollen, eine gewisse Verschärfung des Jacobsonschen Begriffes eines Ringes mit verschwindendem Radikal, wenn auch nicht zu Strukturs̈atzen im Sinne des zweiten Wedderburnschen Theorems, so doch zu anderen, in ihrer Art bemerkenswerten Ergebnissen führt.[4] Wir gehen aus von der Bemerkung, daß in einem kommutativen Ring $\mathfrak{R}$ mit Einheitselement ϵ das Jacobsonsche Radikal gleich

[1] Dieser Bericht wurde aufgeführt im gedruckten Programm unter dem Titel *Jacobsonches Radikal, Hilbertscher Nullstellensatz, Dimensionstheorie.*

[2] Zum Fundamentaltheorem über einfache Ringe vergl. die besonders einfache und elegante Ableitung bei Artin [1]. Zur Definition und den Grundeigenschaften des Jacobsonschen Radikals vergl. Jacobson [1]. Über subdirekte Summen vergl. auch McCoy [1].

[3] Vergl. hierzu die ausführliche Darstellung in Krull [4].

[4] Die Terminologie (Ober- und Unterideal, minimales Primoberideal bzw. Radikal eines Ideals usw.) ist die von Krull [1].

dem Durchschnitt aller maximalen Ringprimideale ist, wobei wie üblich unter einem maximalen Primideal ein Ideal $\mathfrak{m}$ verstanden wird, das (abgesehen von dem Gesamtring $\mathfrak{R}$) kein echtes Oberideal besitzt, oder, was auf dasselbe herauskommt, ein Ideal $\mathfrak{m}$, für das der Restklassenring $\mathfrak{R}/\mathfrak{m}$ ein Körper ist.[5] Diese Bemerkung legt es nahe, jedem Ideal $\mathfrak{a}$ aus $\mathfrak{R}$ den Durchschnitt $\mathfrak{r}_j(\mathfrak{a})$ aller maximalen Primoberideale als zugehöriges *Jacobsonsches Radikal* zuzuordnen; denn es stellt ja $\mathfrak{r}_j(\mathfrak{a})/\mathfrak{a}$ das Jacobsonsche Radikal des Restklassenringes $\mathfrak{R}/\mathfrak{a}$ dar. Man hat dann zu jedem $\mathfrak{a}$ *zwei* Radikale, einerseits das schon definierte $\mathfrak{r}_j(\mathfrak{a})$, andererseits das Radikal $\mathfrak{r}(\mathfrak{a})$ im üblichen Sinne des Wortes, das der Durchschnitt aller minimalen Primoberideale von $\mathfrak{a}$ ist und somit stets der Beziehung $\mathfrak{r}(\mathfrak{a}) \leqq \mathfrak{r}_j(\mathfrak{a})$ genügt. Eine ausgezeichnete Rolle spielen nun offenbar die Ringe, in denen kein Unterschied zwischen gewöhnlichem und Jacobsonschem Radikal existiert, also durchweg $\mathfrak{r}(\mathfrak{a}) = \mathfrak{r}_j(\mathfrak{a})$ wird. Wir wollen derartige Bereiche kurz als *Jacobsonsche Ringe* bezeichnen. Wie mühelos aus der Definition von $\mathfrak{r}(\mathfrak{a})$ und $\mathfrak{r}_j(\mathfrak{a})$ zu ersehen, gilt der Satz: $\mathfrak{R}$ ist dann und nur dann ein Jacobsonscher Ring, wenn jedes Primideal $\mathfrak{p}$ aus $\mathfrak{R}$ gleich dem Durchschnitt seiner maximalen Primoberideale wird. Daß auch für Jacobsonsche Ringe kein Strukturtheorem im Sinne des Wedderburnschen Satzes über halbeinfache hyperkomplexe Systeme zu erwarten ist, überlegt man sich leicht an Hand der Bemerkung, daß jeder Polynomring in endlich vielen Variablen mit Körperkoeffizienten einen Jacobsonschen Ring darstellt. Auf der anderen Seite weist aber gerade die Betrachtung eines Polynomringes $\mathfrak{R} = K[x_1, \cdots, x_n]$ über einem Körper K den Weg zur fruchtbaren Weiterarbeit. Es ist bekannt, daß der für $\mathfrak{R}$ gültige *Hilbertsche Nullstellensatz* gleichwertig ist mit den folgenden beiden Aussagen: 1. $\mathfrak{R}$ ist ein Jacobsonscher Ring. 2. Jeder Körperhomomorphismus von $\mathfrak{R}$ bildet $\mathfrak{R}$ auf einen algebraischen Oberkörper von K ab.[6] In dieser Fassung kann nun der Hilbertsche Nullstellensatz zu einem Theorem über beliebige Jacobsonsche Ringe erweitert werden, das an Allgemeinheit und Einfacheit nichts zu wünschen übrig läßt.

PERMANENZSATZ. *Ist $\mathfrak{R}$ ein Jacobsonscher Ring, so ist es auch jede endliche Ringerweiterung $\mathfrak{S} = \mathfrak{R}[\alpha_1, \cdots, \alpha_n]$. Darüber hinaus ist jeder Körperhomomorphismus $\mathfrak{K}$ von $\mathfrak{S}$ eine algebraische Fortsetzung eines Körperhomomorphismus, d.h. ist das Bild von $\mathfrak{S}$ bei* K *ein Körper* Λ, *so ist das Bild von $\mathfrak{R}$ bei* K *gleichfalls ein Körper K und es wird Λ algebraisch über K.*

Um einzusehen, daß der Permanenzsatz tatsächlich eine Verallgemeinerung des Hilbertschen Nullstellensatzes bildet, braucht man nur zu beachten, daß der im Hilbertschen Nullstellensatz auftretende Koeffizientenkörper K ein

[5] Wegen der Existenz des Einheitselements muß jedes in unserem Sinne maximale Ideal von selbst Primideal sein.

[6] Unter einem Körperhomomorphismus von $\mathfrak{R}$ verstehen wir eine homomorphe Abbildung von $\mathfrak{R}$ *auf* (nicht nur *in*) einen Körper. Zu der Reduktion des Hilbertschen Nullstellensatzes auf die beiden angegebenen Bedingungen vergl. Zariski [1].

Jacobsonscher Ring einfachster Art ist, und daß jeder Körperhomomorphismus von K notwendig K auf sich selbst abbildet. Der Beweis des Permanenzsatzes ist ganz elementar. Zunächst sind drei Reduktionen möglich, die auch weiterhin eine große Rolle spielen werden: 1. Da jede endliche Ringerweiterung $\mathfrak{R}[\alpha_1, \cdots, \alpha_n]$ Restklassenring eines Polynomringes $\mathfrak{R}[x_1, \cdots, x_n]$ ist, genügt es, Polynomringe über $\mathfrak{R}$ zu betrachten. 2. Es muß sogar nur der Fall $\mathfrak{R}[x]$ betrachtet werden, da dann Induktion möglich ist. 3. Ist $\mathfrak{p}$ ein Primideal aus $\mathfrak{S} = \mathfrak{R}[x]$, $\mathfrak{p}^{(r)} = \mathfrak{p} \cap \mathfrak{R}$, so kann man im Falle $\mathfrak{p}^{(r)} \neq (0)$ die Betrachtung von $\mathfrak{R}$ und $\mathfrak{S}$ nach dem Integritätsbereich $\mathfrak{F} = \mathfrak{R}/\mathfrak{p}^{(r)}$ und dem zu $\mathfrak{S}/\mathfrak{p}^{(r)} \cdot \mathfrak{S}$ isomorphen Polynomring $\mathfrak{F}[x]$ verlegen. (Man beachte, daß $\mathfrak{F}$ gleichzeitig mit $\mathfrak{R}$ ein Jacobsonscher Ring ist.) Das $\mathfrak{p}$ in $\mathfrak{F}[x]$ entsprechende Primideal $\mathfrak{p}_0$ genügt der Gleichung $\mathfrak{p}_0 \cap \mathfrak{F} = (0)$. Derartige $\mathfrak{F}[x]$-Primideale $\mathfrak{p}_0$ können aber leicht genauer charakterisiert werden: Ist $\mathfrak{k}$ der Quotientenkörper von $\mathfrak{F}$, $a(x)$ ein Polynom aus $\mathfrak{F}[x]$, so möge unter dem *Quasihauptideal* $\{a(x)\}$ das Ideal aller der $c(x) \in \mathfrak{F}[x]$ verstanden werden, die in $\mathfrak{k}[x]$ durch $a(x)$ teilbar sind. Man sieht dann sofort, daß die Primideale $\mathfrak{p}_0$ nichts anderes sind als die *Primquasihauptideale* $\{p(x)\}$, bei denen das erzeugende Polynom $p(x)$ in $\mathfrak{k}[x]$ irreduzibel ist. Aus 1–3 folgt, daß es genügt, Primquasihauptideale im Polynomring $\mathfrak{F}[x]$ über dem Integritätsbereich $\mathfrak{F}$ zu betrachten. Zu zeigen ist offenbar zweierlei: (a) Ist $\{p(x)\}$ in $\mathfrak{F}[x]$ maximal, so muß $\{p(x)\} \cap \mathfrak{F} \neq (0)$ in $\mathfrak{F}$ maximal, d.h. es muß $\mathfrak{F}$ ein Körper sein. (b) Ist $\{p(x)\}$ nicht maximal und $a(x) \notin \{p(x)\}$, so gibt es stets ein zwar $\{p(x)\}$, aber nicht $a(x)$ enthaltendes, maximales Primideal in $\mathfrak{F}[x]$. (a) Es sei $p(x) = p_0x^n + p_1x^{n-1} + \cdots + p_n$. Ist dann der Jacobsonsche Ring $\mathfrak{F}$ kein Körper, so gibt es in $\mathfrak{F}$ ein p_0 nicht enthaltendes, maximales Primideal $\mathfrak{m}^{(i)} \neq (0)$ und man rechnet unschwer nach, daß $\{p(x)\}$ das von $\mathfrak{F}[x]$ verschiedene, echte Oberideal $(\mathfrak{m}^{(i)}, \{p(x)\}) \cdot \mathfrak{F}[x]$ besitzt, also sicher nicht maximal ist. (b) Es sei $p(x) = p_0x^n + \cdots + p_n$, $a(x) = a_0x^m + \cdots + a_m$, r sei die wegen $a(x) \in \{p(x)\}$ von 0 verschiedene Sylvestersche Resultante von $p(x)$ und $a(x)$, $\mathfrak{m}^{(i)}$ sei ein maximales, $p_0 \cdot r$ nicht enthaltendes Primideal aus $\mathfrak{F}$. Dann gibt es nach (a) ein maximales Primoberideal $\mathfrak{m}$ von $\{p(x)\}$ in $\mathfrak{F}[x]$ mit $\mathfrak{m} \cap \mathfrak{F} = \mathfrak{m}^{(i)}$, und da beim Übergang von $\mathfrak{F}$ zum Körper $K = \mathfrak{F}/\mathfrak{m}$ aus $a(x)$ und $p(x)$ zwei teilerfremde Restklassenpolynome von $K[x]$ entstehen, kann $\mathfrak{m}$ das Polynom $a(x)$ nicht enthalten.

Daß der Permanenzsatz als eine *vollbefriedigende, denkbar weitgehende* Verallgemeinerung des Hilbertschen Nullstellensatzes angesehen werden darf, zeigen die folgenden Bemerkungen: 1. Die Voraussetzung, daß $\mathfrak{R}$ selbst ein Jacobsonscher Ring ist, ist schon im Falle einer einfachen transcendenten Ringerweiterung $\mathfrak{S} = \mathfrak{R}[u]$ nicht nur für die Gültigkeit des ersten, sondern auch für die Gültigkeit des zweiten Teiles des Permanenzsatzes notwendig. Die Überlegung, die diese Notwendigkeit zeigt, ist nichts anderes als eine geschickte Verallgemeinerung des bekannten, für die meisten einfachen Beweise des gewöhnlichen Hilbertschen Nullstellensatzes grundlegenden Schlusses von *Rabinowitsch*. 2. Die Beschränkung auf *endliche* Ringerweiterungen $\mathfrak{S} = \mathfrak{R}[\alpha_1, \cdots, \alpha_n]$ ist keine unnötige Spezialisierung. Es ist leicht einzusehen, daß für unendliche Ringerweiterungen

eines Jacobsonschen Ringes weder der erste noch der zweite Teil des Permanenzsatzes zu gelten braucht. (Zum ersten Teil braucht man nur zu beachten, daß es Oberringe von Körpern gibt, die ihrerseits keine Jacobsonschen Ringe sind. Zum zweiten Teil wähle man etwa für $\mathfrak{R}$ den Ring der ganzen, für $\mathfrak{S}$ den Körper der rationalen Zahlen und betrachte den identischen Automorphismus von $\mathfrak{S}$.) 3. Nur in *einem* Fall scheint es von vornherein aussichtsvoll, den Permanenzsatz auf unendliche Ringerweiterungen $\mathfrak{S}$ auszudehnen, nämlich dann, wenn $\mathfrak{S}$, d.h. jedes Element von $\mathfrak{S}$ vom Ausgangsring $\mathfrak{R}$ *ganz abhängt*. Wirklich gelingt diese Verallgemeinerung des Permanenzsatzes unschwer mit Hilfe der wohlbekannten Fundamentalsätze über den Zusammenhang zwischen den Primidealen eines beliebigen Ringes $\mathfrak{R}$ und eines ganz abhängigen Oberringes $\mathfrak{S}$.[7]

Der Permanenzsatz bedeutet seiner Natur nach einen gewissen Abschluß. Über ihn hinaus führt die Überlegung, daß in der Theorie der kommutativen Ringe, insbesondere der Integritätsbereiche, die *Stellenringe* (local rings) eine fundamentale Rolle spielen, die nur ein einziges maximales Primideal enthalten und infolgedessen bestimmt *nicht* Jacobsonsche Ringe sind. Hier kommt man mit dem *"Jacobsonschen Prinzip"*, das in der Betrachtung der Durchschnitte geeigneter Mengen von *maximalen* Ringprimidealen besteht, nicht mehr weiter. Es liegt nahe, das Prinzip dadurch zu verallgemeinern, daß man passende Mengen von beliebigen Primidealen zuläßt und insbesondere überall dort, wo, wie im Spezialfall der Stellenringe, ein bestimmtes Primideal $\mathfrak{p}$ ausgezeichnet ist, die aus *unmittelbaren Primunteridealen*[8] *von* $\mathfrak{p}$ gebildeten Mengen untersucht. Man stößt dann sofort auf die folgende Frage: Es seien $\mathfrak{p}$ und $\mathfrak{q} \subset \mathfrak{p}$ zwei Primideale aus dem Ringe $\mathfrak{R}$. Welche Aussagen können, allgemein oder unter besonderen Voraussetzungen für $\mathfrak{R}$, über die Möglichkeit der Durchschnittsdarstellung von $\mathfrak{q}$ durch unmittelbare Primunterideale von $\mathfrak{p}$ gemacht werden? Hier zeigt sich nun, zum mindesten bei oberflächlicher Überprüfung, kein Weg, der bei beliebigen Ringen $\mathfrak{R}$ zu einem irgendwie befriedigenden Theorem, etwa im Sinne unseres Permanenzsatzes führte. Beschränkt man sich aber auf *Noethersche Ringe* (Ringe mit Maximalbedingung), so erhält man ohne Schwierigkeit allgemein:

LOKALER DURCHSCHNITTSSATZ. *In einem Noetherschen Ring* $\mathfrak{R}$ *ist jedes Primunterideal* $\mathfrak{q}$ *des Primideals* $\mathfrak{p}$ *Durchschnitt von unmittelbaren Primunteridealen von* $\mathfrak{p}$.

Der Beweis, bei dem man sich, wie unmittelbar zu sehen, auf die Betrachtung eines nullteilerfreien Stellenringes $\mathfrak{R}$ mit dem maximalen Primideal $\mathfrak{p} = \mathfrak{m}$ beschränken darf, hat zur Grundlage einerseits den *Hauptidealsatz*, nachdem in

[7] Gemeint sind die "drei ersten Primidealsätze" von Krull [1, §1].

[8] Das echte Primoberideal $\mathfrak{p}_1$ von $\mathfrak{p}$ heißt unmittelbar, wenn zwischen $\mathfrak{p}$ und $\mathfrak{p}_1$ kein echtes Zwischenprimideal eingeschaltet werden kann. Man beachte, daß ein unmittelbares Primoberideal von $\mathfrak{p}$ kein minimales Primoberideal ist, das einzige minimale Primoberideal von $\mathfrak{p}$ ist $\mathfrak{p}$ selbst!

einem nullteilerfreien Noetherschen Stellenring jedes minimale Primoberideal eines Hauptideals (a) ringminimal ist, also kein von (0) verschiedenes echtes Primunterideal besitzt. Andererseits stützt er sich auf den *Schrankensatz*, nach dem die Gliederzahl m einer in $\mathfrak{R}$ gebildeten Primunteridealkette $\mathfrak{p} = \mathfrak{p}_0 \supset \mathfrak{p}_1 \supset \mathfrak{p}_2 \supset \cdots \supset \mathfrak{p}_m$ eine feste Schranke M nicht überschreiten kann.[9] Der Hauptidealsatz, zusammen mit der Bemerkung, daß in einem Noetherschen Ring jedes Ideal nur endlich viele minimale Primoberideale besitzt, gestattet es, zu jedem nicht in $\mathfrak{q}$ liegenden Element a ein a nicht enthaltendes unmittelbares Primoberideal von $\mathfrak{q}$ zu konstruieren. Zieht man dann noch den Schrankensatz heran, so liefert ein einfacher Induktionsschluß rasch den gewünschten lokalen Durchschnittssatz.

Gleichzeitig mit $\mathfrak{R}$ ist auch jede endliche Ringerweiterung $\mathfrak{S} = \mathfrak{R}[\alpha_1, \cdots, \alpha_n]$ ein Noetherscher Ring. Der lokale Durchschnittssatz gilt also nicht nur für $\mathfrak{R}$ sondern auch für jede endliche Ringerweiterung von $\mathfrak{R}$, der "Permanenzsatz" ist hier kein Problem. Dafür erhebt sich bei den Noetherschen Ringen die weitergehende Frage, ob nicht etwa die Dimensionstheorie der Primideale, wie sie für einen Polynomring $K[x_1, \cdots, x_n]$ in endlich vielen Variabeln mit Körperkoeffizienten gilt, in passender Form auf den entsprechenden Polynomring $\mathfrak{S} = \mathfrak{R}[x_1, \cdots, x_n]$ über einem beliebigen Noetherschen Ring $\mathfrak{R}$ ausgedehnt werden kann.[10] Es zeigt sich, daß eine derartige Ausdehnung tatsächlich möglich ist, wenn wir die schärfere Voraussetzung machen, daß $\mathfrak{R}$ *nicht nur ein Noetherscher, sondern gleichzeitig auch ein Jacobsonscher Ring* ist. Unter dieser doppelten Annahme erhält man den:

DIMENSIONSSATZ. *Es sei $\mathfrak{p}$ ein beliebiges Primideal aus $\mathfrak{S} = \mathfrak{R}[x_1, \cdots, x_n]$, $\mathfrak{p}^{(r)} = \mathfrak{p} \cap \mathfrak{R}$ sei das darunter liegende Primideal aus $\mathfrak{R}$. Hat dann $\mathfrak{p}^{(r)}$ die Dimension ∞, so hat auch $\mathfrak{p}$ die Dimension ∞. Hat aber $\mathfrak{p}^{(r)}$ eine endliche Dimension $d^{(r)}$, so gilt für die stets endliche Dimension von $\mathfrak{p}$ die Ungleichung $d^{(r)} + n \geqq d \geqq n$ wobei, sowohl der Fall $d^{(r)} + n = d$ als auch der Fall $d^{(r)} = d$ wirklich vorkommt. Die Dimensionsdifferenz $d - d^{(r)}$ kann (für $d^{(r)} \neq \infty$) genau so körpertheoretisch gedeutet werden wie die Dimension $d = d - 0$ im Spezialfall eines Polynomringes $\mathfrak{S} = K[x_1, \cdots, x_n]$.*

Natürlich ist beim Dimensionssatz in erster Linie der Fall einer endlichen Dimension $d^{(r)}$ wichtig. $d^{(r)} = d = \infty$ kann als eine praktisch unwesentliche Ausnahme angesehen werden. Was die präzise Fassung des letzten Teiles des Dimensionssatzes angeht, so sei $K(\mathfrak{p}^{(r)})$ bzw. $K(\mathfrak{p})$ der Quotientenkörper des Restklassenringes $\mathfrak{R}/\mathfrak{p}^{(r)}$ bzw. $\mathfrak{S}/\mathfrak{p}$. Dann kann man $K(\mathfrak{p})$ als Oberkörper von

[9] Zum Hauptideal- und Beschränktheitssatz vergl. Krull [1]. Nr. 8.

[10] Zum Dimensionsbegriff der Idealtheorie und speziell den für Ringe aus ganzen algebraischen Funktionen gültigen Dimensionssätzen vergl. Krull [2]. Dem Primideal $\mathfrak{p}$ wird bekanntlich die Dimension m zugeschrieben, wenn in $\mathfrak{R}$ zwar eine m-gliedrige Primoberidealkette $\mathfrak{p} \subset \mathfrak{p}_1 \subset \cdots \subset \mathfrak{p}_m$ aber keine $(m+1)$-gliedrige Kette $\mathfrak{p} \supset \mathfrak{p}_1 \supset \cdots \supset \mathfrak{p}_{m+1}$ existiert. Gibt es in $\mathfrak{R}$ Primoberidealketten $\mathfrak{p} \supset \mathfrak{p}_1 \supset \cdots \supset \mathfrak{p}_n$ von beliebig großer Gliederzahl, so setzt man die Dimension von $\mathfrak{p}$ gleich ∞.

$\mathfrak{R}(\mathfrak{p}^{(r)})$ auffassen, und es wird $d - d^{(r)}$ gerade gleich dem Transzendenzgrade von $K(\mathfrak{p})$ über $K(\mathfrak{p}^{(r)})$.

Beim Beweise des Dimensionssatzes ist eine ähnliche Reduktion möglich wie früher beim Permanenzsatz. Man kommt zu dem Ergebnis, daß nur gezeigt zu werden braucht: *Ein Primquasihauptideal* $\{p(x)\}$ *aus dem Polynomring* $\mathfrak{F}[x]$ *über dem gleichzeitig Noetherschen und Jacobsonschen Integritätsbereich* $\mathfrak{F}$ *besitzt stets dieselbe Dimension wie das Nullideal von* $\mathfrak{F}$. Ist nun d bzw. $d^{(i)}$ die Dimension von $\{p(x)\}$ in $\mathfrak{F}[x]$ bzw. von (0) in $\mathfrak{F}$, so gewinnt man die Ungleichungen $d \leqq d^{(i)}$ und $d \geqq d^{(i)}$ auf zwei völlig verschiedenen Wegen: (a) Bei $d \leqq d^{(i)}$ liegt die Hauptschwierigkeit in dem Nachweis, daß in $\mathfrak{F}[x]$ kein Primoberideal $\mathfrak{p}_1$ von $\mathfrak{p}_0 = \{p(x)\}$ existieren kann, das Erweiterungsideal eines ringminimalen Primideals $\mathfrak{p}_0^{(i)}$ aus $\mathfrak{F}$ ist, $\mathfrak{p}_1 = \mathfrak{p}_0^{(i)} \cdot \mathfrak{F}[x]$. Gäbe es nun ein solches $\mathfrak{p}_1$, so wäre $\mathfrak{p}_2 = \mathfrak{p}_1 + (x) \cdot \mathfrak{S}$ ein echtes Primoberideal von $\mathfrak{p}_1$, und es müßte, wie leicht zu sehen, $\mathfrak{p}_1$ das *einzige* zwischen $\mathfrak{p}_0$ und $\mathfrak{p}_2$ liegende Primideal sein. Das aber widerspricht dem für $\mathfrak{F}[x]$ geltenden lokalen Durchschnittssatz. (b) Bei $d \geqq d^{(i)}$ braucht man, anders als im Fall (a), die Tatsache, daß $\mathfrak{F}$ ein Jacobsonscher Ring ist. Daraus und aus der Maximalbedingung folgt nämlich sofort, daß in $\mathfrak{F}$ für $1 \leqq m \leqq d^{(i)}$ jedes m-dimensionale Primideal unendlich viele $(m-1)$-dimensionale Primoberideale besitzt, und auf Grund dieser Bemerkung kann in $\mathfrak{F}[x]$ mit ähnlichen Überlegungen, wie sie beim Beweise des lokalen Durchschnittssatzes benutzt wurden, die Existenz mindestens einer $d^{(i)}$-gliedrigen Primoberidealkette $\{p(x)\} = \mathfrak{p}_0 \subset \mathfrak{p}_1 \subset \cdots \subset \mathfrak{p}_{d^{(i)}}$ bewiesen werden.

Daß die Ungleichung $d \geqq d^{(i)}$ nicht mehr zu gelten braucht, wenn $\mathfrak{F}$ kein Jacobsonscher Ring ist, kann leicht durch Beispiele belegt werden. Was die Maximalbedingung angeht, so dürfte es einerseits nicht einfach sein, einen Jacobsonschen Ring ohne Maximalbedingung zu konstruieren, für den der Dimensionssatz *nicht* gilt. Andererseits könnten nur *wesentlich neue* Überlegungen die Ausdehnung dieses Theorems auf Jacobsonsche, aber nicht-Noethersche Ringe ermöglichen. Weit wichtiger als die Frage nach der Entbehrlichkeit der Maximalbedingung scheint im übrigen die andere, ob es nicht möglich ist, unter schärferen Voraussetzungen über $\mathfrak{R}$ zu schärferen Dimensionssätzen zu kommen. In einem Polynomring $K[x_1, \cdots, x_n]$ mit Körperkoeffizienten gilt für alle Primoberidealketten der:

LÄNGENSATZ. *Zwei Primoberidealketten* $\mathfrak{p}_0 \subset \mathfrak{p}_1 \subset \cdots \subset \mathfrak{p}_{m-1} \subset \mathfrak{p}_m$ *und* $\mathfrak{p}_0 = \mathfrak{p}_0' \subset \mathfrak{p}_1' \subset \cdots \subset \mathfrak{p}_{m'-1}' \subset \mathfrak{p}_{m'}' = \mathfrak{p}_m$ *mit gleichem Anfangs- und Endglied, die beide nicht durch Einschaltung von Zwischengliedern verlängert werden können, besitzen stets dieselbe Gliederzahl* $m = m'$.[10]

Es liegt nun auf der Hand zu fragen: *Es sei* $\mathfrak{R}$ *ein gleichzeitig Noetherscher und Jacobsonscher Ring, für den der Längensatz gilt. Gilt dann der Längensatz stets auch für* $\mathfrak{R}[x_1, \cdots, x_n]$? (Problem der "Permanenz des Längensatzes".)

Hier scheint es nicht mehr möglich zu sein, mit den bekannten Hilfsmitteln zu einer Entscheidung zu kommen. Doch läßt sich wenigstens die aufgeworfene

Frage auf eine einfachere zurückführen, die vor allem deshalb bemerkenswert ist, weil bei ihr die eigentliche Schwierigkeit der Aufgabe besonders deutlich hervortritt. Es sei $\mathfrak{F}$ ein Noetherscher Integritätsbereich, für den der Längensatz gilt. (Der Jacobsonsche Charakter von $\mathfrak{F}$ spielt keine Rolle, man dürfte sogar $\mathfrak{F}$ als Stellenring annehmen.) Ist in $\mathfrak{F}[x]$ das Primideal $\mathfrak{p}_2$ ein echtes Oberideal des Primquasihauptideals $\{p(x)\}$, und setzt man $\mathfrak{p}_2 \cap \mathfrak{F} = \mathfrak{p}_2^{(i)}$, so zeigt man leicht mit Hilfe der beim Beweis des Dimensionssatzes unter (a) skizzierten Überlegungen: Soll zwischen $\mathfrak{p}_2$ und $\{p(x)\}$ ein echtes Zwischenprimideal $\mathfrak{p}_1$ existieren, so muß es in $\mathfrak{F}$ im Falle $\mathfrak{p}_2^{(i)} \cdot \mathfrak{F}[x] \subset \mathfrak{p}_2$ mindestens eine Primidealkette $\mathfrak{p}_2^{(i)} \subset \mathfrak{p}_1^{(i)} \subset (0)$ und im Falle $\mathfrak{p}_2^{(i)} \cdot \mathfrak{F}[x] = \mathfrak{p}_2$ sogar mindestens eine Kette $\mathfrak{p}_2^{(i)} \subset \mathfrak{p}_1^{(i)} \subset \mathfrak{p}_0^{(i)} \subset (0)$ geben. Die Betrachtung spezieller Beispiele legt nun die Vermutung nahe, daß diese für die Existenz eines Zwischenprimideals $\mathfrak{p}_1$ notwendigen Bedingungen immer auch hinreichend sind, und die Anwendung der schon beim Permanenzsatz benutzten Reduktionen, sowie eine zwar umständliche aber grundsätzlich nicht schwierige Diskussion zeigt: *Der Längensatz ist dann und nur dann permanent, wenn die Vermutung über die nicht unmittelbaren Primoberideale der Quasihauptideale im Polynomring $\mathfrak{F}[x]$ richtig ist.*

Mit dieser Feststellung dürfte der eigentliche Kern des Problems erfaßt sein. Will man die Primidealvermutung beweisen, so hat man als Hilfsmittel zur Konstruktion eines Zwischenprimideals $\mathfrak{p}_1$ im wesentlichen nur den bei der Beweisskizze des lokalen Durchschnittssatzes erwähnten *Hauptidealsatz* zur Verfügung. Aus diesem folgt sofort, daß in einem Noetherschen Ring $\mathfrak{R}$ stets jedes minimale Primoberideal von $\mathfrak{p} + (a)$ ein unmittelbares Primoberideal von $\mathfrak{p}$ ist, falls $\mathfrak{p}$ ein beliebiges Primideal bedeutet, für das weder $\mathfrak{p} + (a) = \mathfrak{p}$ noch $\mathfrak{p} + (a) = \mathfrak{R}$ wird, und mit Hilfe dieser Bemerkung kann tatsächlich die Zwischengliedidealkonstruktion wenigstens in Spezialfällen durchgeführt werden, z.B. da, wenn $\mathfrak{F}$ ganz abgeschlossen und $\mathfrak{p}_2^{(i)} \cdot \mathfrak{F}[x] \subset \mathfrak{p}_2$ ist. Versucht man aber allgemein durchzukommen, so steht man der Notwendigkeit gegenüber, die Folgerung aus dem Hauptidealsatz zu verschärfen und man erkennt leicht, daß es mehr als ausreichen würde, wenn man in ihr unter Beschränkung auf Stellenringe das Hauptideal (a) durch ein ringminimales Primideal $\mathfrak{q}^*$ ersetzen dürfte, wenn man also den Satz beweisen könnte:

SATZ. *Ist $\mathfrak{q}^*$ ein ringminimales, $\mathfrak{p}$ ein beliebiges Primideal aus dem Noetherschen Stellenring $\mathfrak{R}$, so ist für $\mathfrak{q}^* + \mathfrak{p} \neq \mathfrak{p}$ stets jedes minimale Primoberideal von $\mathfrak{q}^* + \mathfrak{p}$ ein unmittelbares Primoberideal von $\mathfrak{p}$.*

[11] Bei den algebraischen Funktionenringen würde unser Theorem besagen, daß auf einer beliebigen irreduzibeln $(n+1)$-dimensionalen Mannigfaltigkeit M_{n+1} der Schnitt einer irreduzibeln m-dimensionalen Untermannigfaltigkeit mit einer irreduzibeln n-dimensionalen Untermannigfaltigkeit stets in endlich viele nicht eingebettete, irreduzible Teilmannigfaltigkeiten der genauen Dimension $m-1$ zerfällt. Für den Beweis der Permanenz des Längensatzes brauchte man übrigens von dem fraglichen Theorem nur den Spezialfall, daß auch $\mathfrak{p} = \mathfrak{p}^*$ in $\mathfrak{F}$ minimal ist, der bei den algebraischen Funktionenringen dem Schnitt zweier genau n-dimensionalen irreduzibeln Untermannigfaltigkeiten von M_{n+1} entspricht.

Damit sind wir aber auf ein bisher unbewiesenes Theorem über Noethersche Stellenringe gestoßen, das, abgesehen von seiner im Spezialfall der algebraischen Funktionenringe möglichen geometrischen Deutung,[11] schon durch die grundsätzliche Betrachtung nahegelegt wird, daß der auf die multiplikative Idealtheorie zugeschnittene Hauptidealbegriff in rein "additiven" Untersuchungen, wie denen über die halbeordnete Menge der Primideale, wenn irgend möglich, eliminiert werden sollte. Die Analyse der beim Permanenzproblem des Längensatzes auftretenden Schwierigkeiten hat also jedenfalls den einen Vorteil, daß sie neues Licht auf die fundamentale Bedeutung des Hauptidealsatzes wirft, und insbesondere klar zeigt, daß eine Verschärfung dieses Theorems und seiner Folgerungen nicht nur aus Gründen der Methodenreinheit dringend zu wünschen ist.

Fassen wir die Grundgedanken unserer Ausführungen kurz zusammen, so können folgende Hauptpunkte hervorgehoben werden. Die Definition des Jacobsonschen Radikals führt (bei kommutativen Ringen mit Einheitselement) zu dem speziellen bzw. verallgemeinerten Jacobsonschen Prinzip, d.h. zum systematischen Studium geeigneter Mengen von maximalen bzw. beliebigen Primidealen. Daneben tritt in unseren Sätzen und ihren Beweisen ein weiterer Leitgedanke hervor, nämlich der, eine Theorie der Polynomringe in endlich vielen Variabeln über einem völlig allgemeinen oder wenigstens beliebig Noetherschen Ringe zu entwickeln unter Ausnützung der Tatsache, daß wegen der möglichen Induktion in der Regel nur Polynomringe in *einer* Variabeln betrachtet werden müssen, wobei überdies der Koeffizientenring meistens als Integritätsbereich angenommen werden darf. Das spezielle Jacobsonsche Prinzip liefert den Begriff des Jacobsonschen Ringes und den Permanenzsatz, ein denkbar allgemeines und in gewisser Hinsicht abschließendes Theorem. Beim verallgemeinerten Jacobsonschen Prinzip treten die Noetherschen Ringe in den Vordergrund. Man gewinnt für sie den lokalen Durchschnittssatz, der sich zwar weitgehend auf bekannte Tatsachen stützt, sie aber unter einem neuen Gesichtswinkel zeigt. Der die allgemeinen Polynomringe betreffende Leitgedanke gibt zunächst Anlaß zur Entwicklung der Grundlagen einer Dimensionstheorie der Primideale über beliebigen, gleichzeitig Noetherschen und Jacobsonchen Ringen. Der Versuch, darüber hinaus zu tiefer liegenden Ergebnissen zu kommen ("Permanenzproblem des Längensatzes"), führt auf ein bemerkenswertes, aber bisher unlösbares Problem hinsichtlich eines Integritätsbereiches $\mathfrak{J}[x]$ in einer Variabeln mit nullteilerfreiem Noetherschen Koeffizientenring. Dabei werden gleichzeitig die Punkte sichtbar, an denen die allgemeine Theorie der Noetherschen Stellenringe noch wesentliche Lücken aufweist. Unsere an das Jacobsonsche Radikal anknüpfenden Betrachtungen liefern also nicht nur abschließende Resultate, sie geben auch Anlaß zu neuen bisher unbeantworteten Fragen. Man kann wohl diese Tatsache verschieden beurteilen; dem Verfasser scheint sie entschieden *zugunsten* der behandelten Ansätze zu sprechen.

Was die Frage angeht, wie weit unsere Überlegungen Anregungen für die Theorie der *nichtkommutativen* Ringe bieten können, so sei auf drei Punkte

hingewiesen. 1. Klar dürfte sein, daß auch im Nichtkommutativen überall dort, wo nichtmaximale (zweiseitige) Primideale auftreten, die Ringe mit verschwindendem Jacobsonschen Radikal an Bedeutung hinter den Jacobsonschen Ringen in unserem Sinne wesentlich zurücktreten werden. 2. Es scheint nicht ausgeschlossen, den Permanenzsatz des Textes auf einen Polynomring $\mathfrak{R}[x_1, \cdots, x_n]$ mit nichtkommutativem Koeffizientenbereich $\mathfrak{R}$ auszudehnen, wobei allerdings die x_i untereinander und mit den Elementen von $\mathfrak{R}$ kommutieren müssen. 3. Nachdem über nichtkommutative Ringe mit Minimalbedingung in vieler Hinsicht abschließende Resultate vorliegen, ist es grundsätzlich an der Zeit, die Untersuchung beliebiger nichtkommutativer Ringe mit Maximalbedingung systematisch in Angriff zu nehmen. Die vorliegenden Betrachtungen scheinen nun darauf hinzuweisen, daß dabei von vornherein nicht nur die Übertragung von Zerlegungssätzen im Sinne der ursprünglichen Noetherschen Theorie ins Auge gefaßt werden sollte, sondern auch die Frage, wie weit sich Struktursätze, die im Kommutativen für die halbgeordnete Menge der Primideale eines Noetherschen Ringes gelten, auf den nichtkommutativen Fall ausdehnen lassen.

Literatur

E. Artin

1. *The influence of J.H.M. Wedderburn on the development of modern algebra*, Bull. Amer. Math. Soc. Bd. 56 (1950) S. 65–72.

N. Jacobson

1. *The radical and semisimplicity for arbitrary rings*, Amer. J. Math. Bd. 67 (1945).

W. Krull

1. *Idealtheorie*, Ergebnisse der Mathematik und ihrer Grenzgebiete, Berlin, Springer, 1935.
2. *Beiträge zur Arithmetik kommutativer Integritätsbereiche.* III *Zum Dimensionsbegriff der Idealtheorie*, Math. Zeit. Bd. 42 (1937) S. 425–466.
3. *Dimensionstheorie in Stellenringen*, Journal für Mathematik Bd. 179 (1938) S. 204–226.
4. *Subdirekte Summen von Integritätsbereichen*, Math. Zeit. Bd. 52 (1950) S. 810–826.

Neal H. McCoy

1. *Subdirect sums of rings*, Bull. Amer. Math. Soc. Bd. 53 (1947) S. 856–877.

O. Zariski

1. *A new proof of Hilbert's Nullstellensatz*, Bull. Amer. Math. Soc. Bd. 53 (1947) S. 362–68.

University of Bonn,
Bonn, Germany.

ARITHMETICS OF ORTHOGONAL GROUPS[1]

MARTIN EICHLER

A general survey of algebraic arithmetics, especially of those branches which have had the most outstanding success during the last decades, shows that with few exceptions research work has centred around the classical groups. In the theory of class fields the multiplicative group of the numbers of an algebraic number field is distinguished in a conspicuous manner. The multiplicative groups of nonsingular elements of central simple algebras are subgroups of the linear group. There is some justification in considering the algebraic and arithmetic results of the theory of simple algebras as results on the linear group.

In a similar sense there exists the closest connection between the theories of quadratic forms and of the orthogonal group and, analogously, between the theories of Hermitian and skew-symmetric forms and the unitary and the symplectic groups. This analogy is of essential nature; a number of theorems on simple algebras and quadratic forms can be traced back to the same source. For example, there are the theorems that the number of classes of ideals in a central simple algebra and the number of classes of indefinite quadratic forms over the field of rational numbers are in general unity. The proof depends in the first case on the simplicity of the linear group and in the second case on the simplicity of the orthogonal group in a Galois field.

In this address I shall give a brief sketch of some features of the theory of quadratic forms as seen from this angle. We shall notice at first a close parallelism with the elementary arithmetic of algebras. Later we shall be led to new ideas which are put forward by Hecke's work on modular forms and ϑ-functions.

Once and for all an algebraic number field k is to be fixed from which all coefficients of quadratic forms and linear substitutions will be taken.

The orthogonal group in k is given naturally not as an abstract group but as a so-called *group-pair*. In fact, two groups are given. The first is the additive group of vectors $\alpha, \beta, \cdots$ of an n-dimensional vector space R. There exists a basis (ι_ν) of R with respect to k, so that the general vector is

$$\xi = \sum_{\nu=1}^{n} \iota_\nu x_\nu$$

with n variables x_ν. Between the vectors a scalar multiplication is defined:

$$\alpha\beta = \beta\alpha = \text{number of } k,$$

with the property

$$(a_1\alpha_1 + a_2\alpha_2)\beta = a_1\alpha_1\beta + a_2\alpha_2\beta.$$

The square of ξ is a quadratic form

[1] This address replaced the address by M. Deuring listed in the printed program under the title *Singularities of commutative rings*.

$$\xi^2 = \sum_{\mu,\nu=1}^{n} \iota_\mu \iota_\nu x_\mu x_\nu = \sum_{\mu,\nu=1}^{n} f_{\mu\nu} x_\mu x_\nu .$$

Conversely each quadratic form gives rise to a space furnished with a scalar multiplication. Such spaces are called *metrical spaces*. It is trivial to remark that we have to make the assumption $| \iota_\mu \iota_\nu | \neq 0$.

The second group S_R which we have to consider consists of all similarity transformations of R:

$$\alpha \longrightarrow \Sigma\alpha,$$

$$\Sigma(a_1\alpha_1 + a_2\alpha_2) = a_1\Sigma\alpha_1 + a_2\Sigma\alpha_2 , \qquad \Sigma\alpha \cdot \Sigma\beta = n(\Sigma)\alpha\beta.$$

$n(\Sigma)$ will be called the *norm* of Σ. We have

$$n(\Sigma\mathrm{T}) = n(\Sigma)n(\mathrm{T}).$$

A pair of groups with the connecting equation $\Sigma(\alpha + \beta) = \Sigma\alpha + \Sigma\beta$ has been called a group-pair by A. Kurosh.

The elements P of S_R with $n(\mathrm{P}) = 1$ form the *orthogonal group*. If, in this address, we shall exclusively deal with the group-pair $\{R, S_R\}$, we hope not to be accused of having made false promises.

Arithmetic enters in connection with the following notion: a modul $\mathfrak{J}$ of vectors with respect to the order $\mathfrak{o}$ of all integers of k is called a *lattice* if $\mathfrak{J}$ contains n linearly independent vectors and if $\mathfrak{J}$ can be generated by a finite number of its elements. A basis with respect to $\mathfrak{o}$ does not always exist; this is in general the case only if the class number of ideals in k is one.

The *norm* of $\mathfrak{J}$ is defined as follows. Let ι range over $\mathfrak{J}$, and $\mathfrak{b} = \text{l.c.d. } (\iota^2/2)$. Then

$$n(\mathfrak{J}) = \frac{\text{g.c.d. } (\mathfrak{b}\iota^2/2)}{\text{l.c.d. } (\iota^2/2)} = \frac{\mathfrak{a}}{\mathfrak{b}}, \qquad \iota \in \mathfrak{J}.$$

$n(\mathfrak{J})$ is an ideal in k. It is the first task to investigate the local properties of lattices. Let $\mathfrak{p}$ be a prime ideal in k and $k_\mathfrak{p}$, $\mathfrak{o}_\mathfrak{p}$, $\mathfrak{J}_\mathfrak{p}$ the $\mathfrak{p}$-adic extensions of k, $\mathfrak{o}$, $\mathfrak{J}$ respectively. $\mathfrak{J}_\mathfrak{p}$ possesses a basis $[\iota_\nu]$ with respect to $\mathfrak{o}_\mathfrak{p}$. The ideals

$$\mathfrak{d}_\mathfrak{p}(\mathfrak{J}) = \left| \frac{\iota_\mu \iota_\nu}{n(\mathfrak{J}_\mathfrak{p})} \right|, \qquad \mathfrak{d}(\mathfrak{J}) = \prod_\mathfrak{p} \mathfrak{d}_\mathfrak{p}(\mathfrak{J})$$

are called *discriminant ideals* of $\mathfrak{J}_\mathfrak{p}$ and $\mathfrak{J}$ respectively. They are integral ideals. The discriminant ideals and the elementary divisors of the discriminant matrices $(n(\mathfrak{J}_\mathfrak{p})^{-1}\iota_\mu\iota_\nu)$ are invariants of a lattice. Lattices for which these invariants coincide form an *order* of lattices. These brief remarks on the purely additive arithmetic of $\{R, S_R\}$ may suffice.

In the next step the multiplicative arithmetic is developed. Let two lattices $\mathfrak{J}$ and $\mathfrak{R}$ of the same order be given. The set of all $\Sigma \in S_R$ with

$$\Sigma\mathfrak{R} \subset \mathfrak{J}$$

is called an *ideal* and denoted by $\mathfrak{J}/\mathfrak{R}$. Ideals have the following properties:

(1) $\mathfrak{J}/\mathfrak{J}$ is a semigroup.

(2) $\mathfrak{J}/\mathfrak{R}\cdot\mathfrak{R}/\mathfrak{L} = \mathfrak{J}/\mathfrak{L}$; in consequence the ideals form a groupoid with the units $\mathfrak{J}/\mathfrak{J}$.

(3) If Σ is an arbitrary element of S_R, then

$$(\Sigma\mathfrak{J})/\mathfrak{J} = \Sigma(\mathfrak{J}/\mathfrak{J}); \qquad \mathfrak{J}/\Sigma\mathfrak{J} = (\mathfrak{J}/\mathfrak{J})\Sigma^{-1};$$

such ideals are called *principal ideals*. If k is replaced by its $\mathfrak{p}$-adic extension for any prime ideal $\mathfrak{p}$, all ideals become principal ideals.

A norm of ideals can be defined on the basis of the norm of similarity transformations. It is always

$$n(\mathfrak{J}/\mathfrak{R}) = n(\mathfrak{J})/n(\mathfrak{R}).$$

The theory of factorization of ideals may be passed by.

The lattices of an order belong to a finite number of classes, represented by, say, $\mathfrak{J}_1, \cdots, \mathfrak{J}_h$, where each lattice of this order is of the form $\mathfrak{J} = \mathrm{P}\mathfrak{J}_i$, P in S_R. The classes of lattices lead to a definition of classes of ideals in an obvious manner.

From this point we make the restricting assumption that k is the field of rational numbers and the quadratic form defining the metric of R is positive definite. This assumption is not always necessary, but it allows us to simplify definitions and results even where the restriction is avoidable. In consequence of this assumption there exists for every lattice $\mathfrak{J}$ only a finite number of *units* which are defined as the elements Σ of S_R with the property $\Sigma\mathfrak{J} = \mathfrak{J}$.

Let now a be an integer in k and $\pi_{ik}(a)$ the number of ideals of the form $\mathrm{P}\mathfrak{J}_i/\mathfrak{J}_k$, $\mathrm{P}\mathfrak{J}_i \subset \mathfrak{J}_k$, with norm a; this number is finite. The h-rowed matrices

$$P(a) = (\pi_{ik}(a))$$

are of great importance for the further theory. An elementary consequence of the factorization theory is

$$P(a)P(b) = P(ab) \qquad (\text{for } (a, b) = 1).$$

The matrices $P(a)$ can be generalized under the restricting assumption just made. Let $R^{(r)}$ be the space of all tensors of rank r and $M^{(r)}(\Sigma)$ the representation of S_R in $R^{(r)}$. Instead of the numbers $\pi_{ik}(a)$ we now form the sums

$$U_{ik}^{(r)}(a) = \sum_{\mathrm{P}} M^{(r)}(\mathrm{P}) \qquad \text{for} \qquad \mathrm{P}\mathfrak{J}_i \subset \mathfrak{J}_k, \quad n(\mathrm{P}\mathfrak{J}_i/\mathfrak{J}_k) = a,$$

which are matrices of a certain number $t^{(r)}$ of rows. Now the $h\cdot t^{(r)}$-rowed matrices

$$P^{(r)}(a) = (U_{ik}^{(r)}(a))$$

are introduced. Again

$$P^{(r)}(a)P^{(r)}(b) = P^{(r)}(ab) \qquad (\text{for } (a, b) = 1)$$

holds. The $P^{(r)}(a)$ with an odd r are identically 0 and possibly some of them for an even r but not all as we shall see soon.

The meaning of the $P(a) = P^{(0)}(a)$, $P^{(r)}(a)$ is that they lead to a connection between the additive and the multiplicative properties of the group-pair $\{R, S_R\}$. Up to this point the arithmetics of vectors and of similarity transformations have been dealt with separately. No use has yet been made of the multiplication $\Sigma\alpha$ between elements of S_R and R. If α, β are vectors in a lattice $\mathfrak{J}$ and $\Sigma \in \mathfrak{J}/\mathfrak{J}$ such that

$$\beta = \Sigma\alpha,$$

we shall say: β is *divisible* by Σ. A theory of divisibility of vectors by similarity transformations and, more generally, by ideals is easily established. This is the point where we meet characteristic differences between the arithmetics of algebras and group-pairs. In a quadratic number field, for example, there exists not more than one ideal of given norm which divides a given "primitive" number (i. e., a number which is not divisible by a rational integer); but there exists in general a large number of ideals in S_R which divide a given "primitive" vector.

Considerations on divisibility of vectors by ideals lead eventually to the following result: let $\delta_i(a)$ be the number of vectors α in $\mathfrak{J}_i$ with the property

$$\frac{1}{2n(\mathfrak{J}_i)}\,\alpha^2 = a \tag{1}$$

and introduce a vector $\mathfrak{d}(a)$ (matrix with 1 column); let furthermore $p(a)$ be the number of all ideals of the form $\mathfrak{J}/\mathfrak{R}$, $\mathfrak{J} \subset \mathfrak{R}$, with a fixed $\mathfrak{R}$ and arbitrary $\mathfrak{J}$ and of norm a ($p(a)$ is finite and depends only on the order of $\mathfrak{J}$ and $\mathfrak{R}$); then

$$p(a)^{-1}P(a)\mathfrak{d}(b) = p(b)^{-1}P(b)\mathfrak{d}(a) = \mathfrak{d}(ab) \qquad (\text{for } (a, b) = 1). \tag{2}$$

Similar though less simple equations hold if $(a, b) > 1$. Analogously, vectors $\mathfrak{d}^{(r)}(a)$ of $h \cdot t^{(r)}$ components can be defined such that

$$p(a)^{-1}P^{(r)}(a)\mathfrak{d}^{(r)}(b) = p(b)^{-1}P^{(r)}(b)\mathfrak{d}^{(r)}(a) = \mathfrak{d}^{(r)}(ab) \qquad (\text{for } (a, b) = 1). \tag{3}$$

These equations have wide consequences. The vectors $\mathfrak{d}^{(r)}(a)$ (including $\mathfrak{d}(a) = \mathfrak{d}^{(0)}(a)$) span certain vector spaces $\mathfrak{Z}^{(r)}$, the dimensions $m^{(r)}$ of which need not coincide with the numbers $h \cdot t^{(r)}$ of components of these vectors. To the matrices $p(a)^{-1}P^{(r)}(a)$, therefore, correspond $m^{(r)}$-rowed matrices $Z^{(r)}(a)$ which operate in $\mathfrak{Z}^{(r)}$. The elements of the $Z^{(r)}(a)$ are linear functions of the components of the $\mathfrak{d}^{(r)}(a)$. Furthermore, instead of (2) and (3) the formula

$$Z^{(r)}(a)Z^{(r)}(b) = Z^{(r)}(ab) \qquad (\text{for } (a, b) = 1) \tag{4}$$

holds, and generalizations in the case $(a, b) > 1$. Conversely, the $\mathfrak{d}^{(r)}(a)$ are uniquely determined by the $Z^{(r)}(a)$, and all vectors with property (1) are uniquely determined by the $\mathfrak{d}^{(r)}(a)$ for all r. Equation (4) allows us to find all vectors α of composite a if all vectors with $a = p =$ prime are known. The

procedure consists of five steps: (1) calculation of $\mathfrak{d}^{(r)}(p)$; (2) calculation of $Z^{(r)}(p)$; (3) application of (4); (4) calculation of $\mathfrak{d}^{(r)}(a)$; (5) calculation of all α with property (1). It is interesting to compare this result with a similar fact for an algebraic number field K with class number one, for example. In order to find all integers α in K with composite norm a, one has to find all integers with prime norm and to multiply them. If $\{\alpha(a)\}$ is the set of all integers in K with norm a, then

$$\{\alpha(a)\}\cdot\{\alpha(b)\} = \{\alpha(ab)\} \qquad \text{(for } (a, b) = 1\text{)}. \tag{4'}$$

A restricting remark is necessary: according to our procedure, (4) has a meaning only for such vectors α for which numbers a in (1) are norms of ideals; not all a have this property.

The matrices $Z(a)$ are closely connected with those of Hecke's theory. The ζ-functions of the classes $\mathfrak{S}_i$ are

$$\zeta_i(s) = \sum_{\alpha\in\mathfrak{S}_i}\left(\frac{1}{2n(\mathfrak{S}_i)}\,\alpha^2\right)^{-s} = \sum_{m=1}^{\infty}\delta_i(m)m^{-s}.$$

The $Z(m)$ are linear functions of the $\delta_i(m)$:

$$Z(m) = \sum_{i=1}^{h} Z_i\cdot\delta_i(m).$$

We now introduce the infinite series of $m^{(0)}$-rowed matrices:

$$\zeta(s) = \sum_{i=1}^{h} Z_i\zeta_i(s) = \sum_{m=1}^{\infty} Z(m)m^{-s}.$$

This function can be split up into an Euler product:

$$\zeta(s) = \prod_{p}\zeta_p(s),$$

the factors of which are of a certain quality; Hecke has called this product a *canonical* Euler product. In the case of an order of lattices of discriminant 1 each number in k is the norm of an ideal, and our Euler product is exactly the same as in Hecke's theory. If the discrimininant is greater than 1, Hecke takes into account general lattices of more than one order. His Euler factors $\zeta_p^H(s)$ split up into matrices of smaller degrees:

$$(\zeta_p^H(s)) = \begin{pmatrix} \zeta_p^1(s) & & \\ & \zeta_p^2(s) & \\ & & \ddots \end{pmatrix}$$

corresponding to the different orders which enter into his theory and, possibly, to such modular forms which cannot be represented by ϑ-functions. However, this is true only in the case when p does not divide the discriminant and when ideals of norm p exist. For primes dividing the discriminant the theory becomes complicated: we have to distinguish between different kinds of representations

by a quadratic form of a number which is not prime to the discriminant. Moreover, we have to remember that statements have been made only for those numbers which are norms of ideals. In this regard Hecke's theory is more complete. On the other hand, our theory has the advantage of yielding results on orders of quadratic forms immediately without the necessity of translating theorems on modular forms into the language of ϑ-functions.

Let us close with mentioning an open problem: When we wish to apply Hecke's theory to quadratic forms, we are confronted with the task of representing all ϑ-functions of a certain kind by a set of linearly independent modular forms, which includes the task of finding all linear relations among the ϑ-functions. The same problem arises when we try to eliminate analytical methods from the theory of representations of numbers by quadratic forms. It seems desirable to have a purely arithmetical theory of formal ϑ-series. This would establish a bridge between two fields of specific algebraic interest: the arithmetic of quadratic forms and the arithmetical theory of algebraic functions which are given by the elliptic modular functions. Although the problem of finding all linear equations among ϑ-functions is far from its solution, I am glad to report that at least some of them have been found by a more detailed study of the matrices $Z(a)$, and there remains the hope that the whole problem may be solved within our time.

UNIVERSITY OF MÜNSTER,
MÜNSTER, GERMANY.

GÉNÉRALISATIONS NON-ABÉLIENNES DE LA THÉORIE LOCALE DES CORPS DES CLASSES[1]

MARC KRASNER

On sait que la théorie des corps des classes fournit une déscription de l'ensemble des extensions abéliennes K d'un corps k de nombres $\mathfrak{p}$-adiques (théorie locale) ou algébriques (théorie globale), en leur attachant, dans le cas local, le groupe des normes $N_{K/k}(\alpha)$ de K à k des éléments non-nuls α de K, et, dans le cas global, un groupe multiplicatif assez analogue d'idéaux de k, le groupe de Takagi de K/k. La position de ces groupes dans le groupe multiplicatif k^* de k (cas local) ou dans celui des idéaux de k (cas global) permet de déterminer à priori, pour l'extension correspondante K/k, ses propriétés algébriques (son groupe de Galois $G_{K/k}$) et arithmétiques (nombres de ramification v_q et ordres v_q des groupes de décomposition, d'inertie et des groupes de ramification successifs des diviseurs premiers $\mathfrak{P}$ dans K de tout idéal premier $\mathfrak{p}$ de k; et, en particulier, le degré f et l'ordre e de ces $\mathfrak{P}$).

Cette caractérisation des extensions, telle quelle, ne peut guère être étendue au delà du cas abélien, car le groupe analogue pour une extension algébrique finie arbitraire K/k coincide avec celui de sa plus grande sous-extension abélienne $K^{(a)}/k$. Mais, depuis le dernier Congrès, plusieurs tentatives (dont certaines par le conférencier) ont été faites pour généraliser certaines parties de la théorie des corps de classes, aussi bien locale que globale, aux extensions non-abéliennes (la théorie des fonctions L d'Artin, qui est une généralisation partielle de la théorie globale, est la seule tentative plus ancienne dans cette diréction). Jusqu'a présent, elles ont réussi surtout dans le cas local. Le but de cette conférence est de donner l'idée des ces généralisations non-abéliennes de la théorie locale des corps de classes.

A. THÉORIE DE KRASNER

La première ébauche de cette théorie, où je me borne aux extensions complètement ramifiées des corps $\mathfrak{p}$-adiques, se trouve dans mon travail *Sur la primitivité des corps $\mathfrak{P}$-adiques*, Mathematica (Cluj) (1937). En partant de l'idée que la norme $N_{K/k}(\alpha)$ d'un $\alpha \in K$ primitif n'est que le dernier coefficient du polynôme minimal $f_{\alpha/k}(x)$ de α par rapport à k, j'y considère l'ensemble $E_{K/k}$ des polynômes d'Eisenstein définissant une telle extension K/k, avec l'espoir de voir apparaître dans sa structure ce qui est éffacé dans celle du groupe des normes de K/k, dès qu'on sort du cas abélien. Pour celà, j'introduis, dans l'ensemble $E_{k,n}$ des polynômes d'Eisenstein de k de degré n, une notion convenable de congruence (mod $\mathfrak{p}^u$) (où $\mathfrak{p}$ est l'idéal premier de k, et où u est rationnel), ce qui revient à y définir une distance $d(f, g)$ telle que $d(f, h) \leqq \text{Max}\,[d(f, g), d(g, h)]$ (si, dans

[1] Cette communication était mentionnée sur le programme imprimé sous le titre *Essai d'une théorie nonabélienne des corps de classes.*

un espace métrique, la distance satisfait à cette condition, il sera dit *ultramétrique*). Alors, on voit apparaître, dans la structure de $E_{K/k}$, en tant que sous-espace de $E_{k,n}$, un conducteur $\mathfrak{f}_{K/k}$, analogue à celui du cas abélien et se calculant par les mêmes formules à partir des v_q et ν_q. L'expression de $|\mathfrak{f}_{K/k}|$ fournissait la réponse à la question: de combien près doivent être $f, g \in E_{k,n}$ pour définir sûrement une même extension? D'autres résultats ont été également obtenus sur la structure de $E_{K/k}$, qui donnaient un critère de normalité de K/k et permettaient de calculer le nombre des surcorps K de k (contenus dans sa clôture algébrique $\mathfrak{R}$) de degré $n = (K:k)$ donné.

Ce n'est que plus tard que j'ai vu la véritable raison de ces résultats. Il se trouve que la question, "de combien près doivent être $f(x)$ et $g(x)$ pour définir une même extension?" se décompose en deux questions: de combien près doivent être les zéros le plus proches α, β des f, g dans $\mathfrak{R}$ pour définir un même surcorps de k? Comment la distance des f, g depend de celle des α, β? La réponse à la première question est fournie par le principe suivant.

PRINCIPE FONDAMENTAL. *k étant un corps valué[2] complet, et $\mathfrak{R}$ étant sa clôture algébrique valuée, si $\alpha \in \mathfrak{R}$ est séparable par rapport à k, et si C_α est le plus grand cercle de centre α dans $\mathfrak{R}$ ne contenant aucun conjugué $\alpha' \neq \alpha$ de α par rapport à k, $\beta \in C_\alpha$ entraîne $k(\beta) \supseteq k(\alpha)$. En particulier, $(k(\beta):k) \geqq (k(\alpha):k)$ et $(k(\beta):k) = (k(\alpha):k)$ entraîne $k(\beta) = k(\alpha)$.*

DÉMONSTRATION. Tout conjugué de $\alpha - \beta$ par rapport à $k(\beta)$ est de la forme $\alpha' - \beta$, où α' est un conjugué de α par rapport à k. Mais, si $\alpha' \neq \alpha$, on a $|\beta - \alpha| < |\alpha - \alpha'|$, d'où, puisque $|\alpha - \alpha'| \leqq \mathrm{Max}(|\alpha - \beta|, |\alpha' - \beta|)$, on a $|\alpha - \beta| < |\alpha' - \beta|$. Or, k étant complet, $k(\beta)$ l'est aussi, et les conjugués par rapport à $k(\beta)$ ont une même valuation (Ostrowski). Donc, $\alpha - \beta$ et, par suite, α n'ont pas d'autres conjugués par rapport à $k(\beta)$ qu'eux-mêmes. α étant séparable, on a $\alpha \in k(\beta)$. C.q.f.d.

Sauf si le corps de restes r de k est, en un certain sens, "trop proche" de sa clôture algébrique $\mathfrak{R}$ (et il ne l'est sûrement pas si k est localement compact, et, à fortiori, s'il est $\mathfrak{p}$-adique), le réciproque de ce principe, *C_α est le plus grand cercle C de centre α tel que $\beta \in C$ entraîne $k(\beta) \supseteq k(\alpha)$*, est aussi exact. Sinon, il suffit de modifier légèrement le principe fondamental pour qu'il admette un réciproque. Je me borne au cas, où le réciproque précédent a lieu.

Soit $F(x) = \sum_1^{+\infty} a_i x^i$ une série de Taylor (pouvant être un polynôme) à coefficients dans un surcorps $K \subseteq \mathfrak{R}$ de k et soit Π_F son polygone de Newton dans un plan $O\xi\eta$. Si v est tel qu'il existe une droite L_v de pente $-v$, qui touche Π_F, l'ordonnée de l'intersection de L_v avec $O\eta$ sera notée $\varphi_F(v)$. On montre que les pentes et les longueurs de projections des cotés de Π_F sur $O\xi$ donnent, exactement comme dans le cas de polynômes, les valuations des zéros de $F(x)$ dans $\mathfrak{R}$ et leur nombre; d'où, si $O = a_0 = a_1 = \cdots = a_{i-1} \neq a_i$, on a $-\varphi_F(-\mathrm{Log}\, r) =$

[2] Les mots "corps valué" sont employés ici avec le sens "corps valué ultramétrique," autrement dit, on suppose que $|a + b| \leqq \mathrm{Max}(|a|, |b|)$

$i \operatorname{Log} r + \sum_i \operatorname{Log}(r/|z_i|)$, où z_i parcourt les zéros de $F(x)$ dans $\mathfrak{R}$ tels que $|z_i| < r$. Si α est un élément primitif de K/k, posons $F(x) = f_{\alpha/k}(x+\alpha)$ (donc $a_1 = (df_{\alpha/k}(x+\alpha)/dx)_{x=0} = (df_{\alpha/k}(x)/dx)_{x=\alpha}$, et l'idéal (a_1) est la différente $\mathfrak{d}_{\alpha/k}$ de α/k). Désignons, dans ce cas, Π_F et $\varphi_F(v)$ aussi par $\Pi_{\alpha/k}$ et $\varphi_{\alpha/k}(v)$. Si $\Pi_{\alpha/k}$ est la ligne brisée $P_0^{(\alpha)}P_1^{(\alpha)} \cdots P_{m_\alpha}^{(\alpha)}(+\infty)$, soient $-v_q^{(\alpha)}$ la pente de $P_q^{(\alpha)}P_{q+1}^{(\alpha)}$ et $\nu_q^{(\alpha)}$ l'abcisse de $P_q^{(\alpha)}$ (donc $\nu_0^{(\alpha)} = n = (K:k)$ et $\nu_{m_\alpha}^{(\alpha)} = 1$). $\mathfrak{d}_{K/k}$ étant le p.g.c.d. des $\mathfrak{d}_{\alpha/k}$, où α parcourt les entiers de K (il est dit la *différente arithmétique* de K/k), quand $|\mathfrak{d}_{\alpha/k}|$, où α est entier, tend vers $|\mathfrak{d}_{K/k}|$, $\Pi_{\alpha/k}$ et $\varphi_{\alpha/k}(v)$ tendent vers un polygone $\Pi_{K/k}$ (dit *polygone de ramification de* K/k) et une fonction $\varphi_{K/k}(v)$. Si $\Pi_{K/k} = P_0P_1 \cdots P_m(+\infty)$, et si $-v_q$ et ν_q sont la pente de P_qP_{q+1} et l'abcisse de P_q, les v_q sont les valeurs limites (et les seules) des $v_q^{(\alpha)}$, pour les α précédents. Elles sont toutes atteintes si et seulement s'il existe des *éléments discriminantiels* de K/k, c'est-à-dire, des $\alpha \in K$ entiers tels que $\mathfrak{d}_{\alpha/k} = \mathfrak{d}_{K/k}$.

Une théorie de fonctions analytiques dans les corps valués complets que j'ai construite montre que *si $F(x)$ converge pour un $\beta \in \mathfrak{R}$ qui n'est à la distance $< |\beta|$ d'aucun zéro de $F(x)$, on a* $-\operatorname{Log}|F(\beta)| = \omega(F(\beta)) = \varphi_F(\omega(\beta)) = \varphi_F(-\operatorname{Log}|\beta|)$ (ce résultat est, visiblement, un raffinement de l'analogue du théorème de croissance de fonctions entières de Hadamard). En particuliers, *si d est la valuation du plus petit zéro non nul de $F(x)$ et si, en plus, $|\beta| \leqq d$, on a* $|F(\beta)| = |F'(0)|\,|\beta| = |a_1|\,|\beta|$. a étant un nombre positif, et $R(f, g)$ désignant le résultant des $f(x)$, $g(x)$, organisons l'ensemble S_k des polynômes normés irréductibles en x à coefficients dans k en un espace métrique (qui se trouve être ultramétrique) $S_k^{(a)}$ à l'aide de la distance

$$d_a(f, g) = |R(f, g)|^{a:nu} = |f(\beta)|^{a:n} = |g(\alpha)|^{a:u},$$

où n, u sont les degrés des f, g, et où α, β sont des zéros quelconques des f, g. On note $S_{k,n}$ le sous-espace de $S_k^{(n)}$, formé de ses polynômes de degré n. Si $f \in S_{k,n}$ et si $g \in S_k$, on a $d_n(f, g) = |f(\beta)|$ (si, en plus, $g \in S_{k,n}$, on a aussi $d_n(f, g) = |g(\alpha)|$). Par suite, l'application

$$(T) \qquad \beta \longrightarrow f_{\beta/k}(x)$$

de $\mathfrak{R}$ sur S_k applique le cercle $\bar{C}: |\beta - \alpha| \leqq$ (ou $<$) $\operatorname{Exp}(-v)$, de centre α dans $\mathfrak{R}$ sur le cercle $C: d_n(g(x), f_{\alpha/k}(x)) \leqq$ (ou $<$) $\operatorname{Exp}(-\varphi_{\alpha/k}(v))$ de centre $f_{\alpha/k}(x)$ dans $S_k^{(n)}$. Mais l'image inverse de ce dernier cercle est constitué par la réunion des cercles $|\beta - \alpha'| \leqq$ (ou $<$) $\operatorname{Exp}(-v)$ de centres α', où α' parcourt tous les conjugués de α par rapport à k. En particulier, si $\bar{\rho} = \operatorname{Exp}(-v) < \operatorname{Exp}(-v_{m_\alpha-1}^{(\alpha)})$, $C = T \cdot \bar{C}$ est un cercle de même espèce (c'est-à-dire, circonférencié ou non) que $\bar{C}$ et de rayon $\rho = |\mathfrak{d}_{\alpha/k}|\,\bar{\rho}$, et l'image inverse $T^{-1} \cdot C$ de cercle C est constitué (puisque la distance de conjugués de α est $\geqq \operatorname{Exp}(-v_{m_\alpha-1}^{(\alpha)})$) par la réunion de n cercles disjoints de même espèce et de rayon $\rho\,|\mathfrak{d}_{\alpha/k}|^{-1}$.

Soient $\mathfrak{S}_{K/k}^*$ l'ensemble des $f \in S_k$ qui définissent des surextensions de K/k, $\mathfrak{S}_{K/k}$ celui des $f \in \mathfrak{S}_{K/k}^*$ à coefficients entiers, $S_{K/k} = \mathfrak{S}_{K/k} \cap S_{k,n}$, [où $n = (K:k)$] celui des $f \in S_k$ à coefficients entiers, qui définissent K/k, $S_{K/k}^{(\epsilon)}$ celui des

$f \in S_{K/k}$ tels que $| \mathfrak{d}_f | = | (df(x)/dx)_{x=\alpha} | \geqq | \mathfrak{d}_{K/k} | (1 - \epsilon)$, où $\epsilon > 0$ et $f(\alpha) = 0$. Chacun de ces ensembles, ainsi que, quand existent des $\alpha \in K$ discriminantiels, celui $S^{(0)}_{K/k}$ de leurs $f_{\alpha/k}(x)$, caractérisent K/k, et on a $K/k \subseteq K'/k$ si et seulement si $\mathfrak{S}_{K/k} \supseteq \mathfrak{S}_{K'/k}$.

Une subdivision P d'un espace ultramétrique E en cercles d'un même rayon (noté $| P |$) et d'une même espèce est dit un *diviseur de* E de valuation $| P |$ et de l'espèce considérée. J'appelle *conducteur de* K/k le diviseur le plus fin $\mathfrak{f}_{K/k}$ de $S^{(n)}_k$ tel que, pour tout $f \in S_{K/k}$, il existe un $g \in S_k$ tel que $g \equiv f(\mathfrak{f}_{K/k})$ et que $g \notin \mathfrak{S}^*_{K/k}$, et j'appelle *conducteur de* K/k *en* α le diviseur le plus fin $\mathfrak{f}^{(\alpha)}_{K/k}$ de $S^{(n)}_k$, satisfaisant à la même condition pour $f = f_{\alpha/k}$ seulement. Comme, en vertu du principe fondamental et de son réciproque, le cercle circonférencié de rayon $\mathrm{Exp}\,(-v^{(\alpha)}_{m_\alpha - 1})$ est le plus petit cercle de centre α contenant des $\beta \in \mathfrak{K}$ tels que $k(\beta) \not\equiv k(\alpha)$, son image par T, qui est le cercle $d_n(g(x), f_{\alpha/k}(x)) \leqq \mathrm{Exp}\,(-\varphi_{\alpha/k}(v^{(\alpha)}_{m_\alpha - 1}))$ est une classe (mod $\mathfrak{f}^{(\alpha)}_{K/k}$); donc, $\mathfrak{f}^{(\alpha)}_{K/k}$ est aussi circonférencié et $| \mathfrak{f}^{(\alpha)}_{K/k} | = \mathrm{Exp}\,(-\varphi_{\alpha/k}(v^{(\alpha)}_{m_\alpha - 1}))$. Comme $\mathfrak{f}_{K/k}$ est le p.g.c.d. des $\mathfrak{f}^{(\alpha)}_{K/k}$, pour les α entiers, on a $| \mathfrak{f}_{K/k} | = \mathrm{Exp}\,(-\varphi_{K/k}(v_{m-1}))$ et $\mathfrak{f}_{K/k}$ est circonférencié si et seulement s'il existe des $\alpha \in K$ discriminantiels. Si $\eta > 0$, il existe un $\epsilon > 0$ tel que $S^{(\epsilon)}_{K/k}$ soit une réunion de cercles de $S_{k,n}$ de rayon $| \mathfrak{f}_{K/k} | (1 - \eta)$.

D'autres valeurs critiques apparaissent également dans la structure de $\mathfrak{S}_{K/k}$ autour d'un $f_{\alpha/k} \in S_{K/k}$ (ce sont les $\varphi_{\alpha/k}(v^{(\alpha)}_q)$) et autour de $S_{K/k}$ (ce sont les $\varphi_{K/k}(v_q)$). Si K/k est galoisienne, elles apparaissent déjà dans celle de $S_{K/k}$.

Si k est localement compact, $S^{(\epsilon)}_{K/k}$ est une réunion *finie* de cercles de rayon $| \mathfrak{f}_{K/k} | (1 - \eta)$ et un calcul *fini* permet de décider si une réunion finie de tels cercles est un $S^{(\epsilon)}_{K/k}$ ou non. K/k étant, en plus, complètement ramifiée, et ρ étant $| \mathfrak{f}_{K/k} |$, l'ensemble $E_{K/k}$ des polynômes d'Eisenstein définissant K/k (qui est $\subseteq S^{(0)}_{K/k}$) est une réunion d'un nombre fini $c = c(\rho; K/k)$ de cercles de rayon ρ dans $S_{k,n}$. $T^{-1} \cdot E_{K/k}$ est, comme on a vu, la réunion de nc cercles disjoints de rayon $\rho \,|\, \mathfrak{d}_{K/k} \,|^{-1}$ dans le sous-espace $\mathfrak{K}^{(n)}$ de $\mathfrak{K}$, formé des $\alpha \in \mathfrak{K}$ de degré n par rapport à k. K' etant un corps conjugué de K/k et $\langle \mathfrak{P}' \rangle$ étant l'ensemble de ses éléments d'ordre 1 par rapport à l'idéal premier $\mathfrak{P}'$ de K' (donc, $\langle \mathfrak{P}' \rangle$ est le complémentaire de $\mathfrak{P}'^2$ dans $\mathfrak{P}'$), le même ensemble est, d'autre part, la réunion de ces $\langle \mathfrak{P}' \rangle$ pour tous les corps conjugués $K' \subset \mathfrak{K}$ de K/k. Pour tous ces $\langle \mathfrak{P}' \rangle$, le nombre des cercles de rayon $\rho \,|\, \mathfrak{d}_{K/k} \,|^{-1}$, qui les composent, est, évidemment, le même; il ne dépend que des k, $\rho \,|\, \mathfrak{d}_{K/k} \,|^{-1}$, et n, et peut s'écrire $\nu(k, \rho \,|\, \mathfrak{d}_{K/k} \,|^{-1}, n)$; et on a $nc = \nu(k, \rho \,|\, \mathfrak{d}_{K/k} \,|^{-1}, n)\, l_{K/k}$, où $l_{K/k}$ est le nombre des corps conjugués distincts de K/k dans $\mathfrak{K}$. Donc, K/k est galoisienne si et seulement si $c = n^{-1} \nu(k, \rho \,|\, \mathfrak{d}_{K/k} \,|^{-1}, n)$; et, si E est une réunion des $E_{K/k}$, formés de polynômes d'un même degré n et d'une même différente $\mathfrak{d}$, le nombre des surcorps $K \subset \mathfrak{K}$ de k tels que $E_{K/k} \subseteq E$ est le quotient du nombre de cercles de rayon ρ composant E par $n^{-1} \nu(k, \rho \,|\, \mathfrak{d} \,|^{-1}, n)$. Les nombres $\mu(k, \rho, n, \mathfrak{d})$ et $\mu(\rho, \Pi)$ de tous les cercles de rayon ρ, composés de polynômes d'Eisenstein f de degré n et de différente $\mathfrak{d}$ respectivement de polygone de ramification $\Pi_{\alpha/k} = \Pi$ (où $f(\alpha) = 0$) donnés, est facilement calculable, ce qui m'a permis de calculer explicitement les nombres $N_{k,n,\mathfrak{d}}$ et $N_{k,\Pi}$ des

surcorps complètement ramifiés $K \subset \mathfrak{R}$ de k de degré n et de différente $\mathfrak{d}$ respectivement de polygone de ramification Π donnés. Dans le cas $\mathfrak{p}$-adique, par sommations convenables, j'ai calculé le nombre $\mathfrak{N}_{k,n}$ de tels corps qui sont de degré n par rapport à k. Les extensions primitives et les extensions metagaloisiennes (c'est-à-dire, obténus par une succéssion d'extensions galoisiennes) des corps valués complets pouvant (comme je l'ai montré) se caractériser par leurs propriétés arithmétiques, j'ai pu, également, calculer le nombre de tels surcorps $K \subset \mathfrak{R}$ de k de degré et de différente donnés (et, dans le cas $\mathfrak{p}$-adique, de degré donné).

k et k' étant deux corps valués, et $\mathfrak{O}, \mathfrak{O}'$ étant leurs idéaux d'une même valuation $a = |\mathfrak{O}| = |\mathfrak{O}'|$ et d'une même espèce, une application biunivoque de l'ensemble des classes de congruence multiplicative $(\mathrm{mod}^{\times} \mathfrak{O}')$ dans k' sur l'ensemble des celles de congruence $(\mathrm{mod}^{\times} \mathfrak{O})$ dans k est dite un *isomorphisme résiduel de norme a de k' sur k* si elle conserve l'addition (non-univoque) et la multiplication de ces classes, ainsi que la valuation de leurs éléments. Un corps valué complet k sera dit la *limite* d'une suite de corps valués $k_1, k_2, \cdots, k_m, \cdots$, si, pour tout m, sont donnés des isomorphismes résiduels λ_m de k_m sur k de normes $a_m > 0$ tels que $\lim a_m = 0$. On définit, d'une manière évidente, la convergence d'une suite d'éléments $\alpha_m \in k_m$ ou de polynômes $f_m(x)$ à coefficients dans k_m vers un $\alpha \in k$ ou vers un polynôme $f(x)$ à coeffiicients dans k. Si les k_m sont également complets, soient K/k une extension séparable, $f(x)$ un polynôme qui la définit et $f_m(x) \to f(x)$ $(f_m(x) \in k_m[x])$. Alors, le principe fondamental permet de montrer qu'à partir d'un certain rang, $f_m(x)$ est irréductible, l'extension K_m/k_m qu'il définit ne dépend pas des choix des f, f_m et est galoisienne en même temps que K/k et, dans ce dernier cas, on a $G_{K_m/k_m} = G_{K/k}$, cet isomorphisme s'établissant d'une manière canonique et avec conservation d'objets de la théorie de la ramification de Hilbert. On a, ainsi, une théorie d'approximation des corps valués complets par des suites de tels corps, et il est possible de prouver que tout corps valué complet de caractéristique $p \neq 0$ peut s'approximer, au sens précédent, par des suites convenables de tels corps de caractéristique 0 (et, en particulier, les corps des séries de puissances sur un champ de Galois par des suites convenables de corps $\mathfrak{p}$-adiques).

B. Théorie de Shafarevitch

Cette théorie (*Sur les p-extensions*, Rec. Math. (Mat. Sbornik) N. S. (1947)) détermine, pour un corps $\mathfrak{p}$-adique k, ne contenant aucune racine p-ième primitive de l'unité (toutefois, cette détermination devrait pouvoir se faire par des méthodes voisines sans cette restriction), un groupe-facteur essentiel du groupe de Galois de $\mathfrak{R}/k$; à savoir, si $\mathfrak{R}_p/k$ est le composé de toutes les extensions galoisiennes K/k de degré puissance de p, la théorie détermine le groupe de Galois (avec sa topologie de Krull) $G_{\mathfrak{R}_p/k}$ de $\mathfrak{R}_p/k$.

G étant un groupe, soit G' son sous-groupe engendré par les puissances p-ièmes et par les commutateurs de ses éléments, et soit $G^{(i)} = (G^{(i-1)})'$. Si les indices $(G^{(i-1)} : G^{(i)})$ sont tous finis, la topologie uniforme de G, définie par les

$G^{(i)}$, est la même à gauche et à droite, et sera dite sa *p-topologie*. Soit n_0 le degré de k par rapport au corps p-adique rationnel. Alors, $G_{\mathfrak{R}_p/k}$ *est isomorphe*, en tant que groupe topologique, *au complété, par rapport à sa p-topologie, du groupe libre* L_{n_0+1} *à* $n_0 + 1$ *générateurs*.

Idée de démonstration: Soient k' le composé des extensions de degré p de k dans $\mathfrak{R}$, et $k^{(i)} = (k^{(i-1)})'$. Si $G = G_{\mathfrak{R}_p/k}$, on a $G^{(i)} = G_{\mathfrak{R}_p/k^{(i)}}$ et $\bigcap_i G^{(i)} = 1$. $G_{k^{(i)}/k^{(i-1)}} = G^{(i-1)}/G^{(i)}$ est, en vertu de la théorie locale des corps de classes et de la structure du groupe multiplicatif $k^{(i-1)*}$ de $k^{(i-1)}$ (Hensel), un groupe de type $(p, p, \cdots, p)$ de rang $n_0(k^{(i-1)}: k) + 1 = n_0(G: G^{(i-1)}) + 1$. Ainsi, G/G' est de rang $n_0 + 1$ et il existe un système de $n_0 + 1$ générateurs de G (mod G'). Comme $(G/G^{(i)})' = G'/G^{(i)}$ et $G/G^{(i)}$ est un p-groupe, le groupe $\overline{L}$ que ce système engendre est dense dans G, et sa topologie induite est définie par les $\overline{L} \cap G^{(i)}$. Or, $\overline{L}$ peut s'identifier avec le quotient de $L = L_{n_0+1}$ par un sous-groupe invariant convenable, et alors, avec les identifications habituelles, on a $\overline{L} \cap G^{(i)} \supseteq \overline{L}^{(i)} \supseteq L^{(i)}$. Puisque on a $(\overline{L}: \overline{L} \cap G^{(i)}) = (G: G^{(i)})$, si l'on prouve, pour tout i, l'égalité $(G: G^{(i)}) = (L: L^{(i)})$, on aura $\overline{L} = L$ et $\overline{L}^{(i)} = \overline{L} \cap G^{(i)}$, ce qui prouverait le théorème. Or, en vertu d'un théorème de Schreier, $L^{(i-1)}$ est un groupe libre de $[(n_0 + 1) - 1]\,(L: L^{(i-1)}) + 1 = n_0(L: L^{(i-1)}) + 1$ générateurs, et $L^{(i-1)}/L^{(i)}$ est un groupe de type $(p, \cdots, p)$ de ce rang. Si l'on suppose $(G: G^{(i-1)}) = (L: L^{(i-1)})$, on a $G^{(i-1)}/G^{(i)} \simeq L^{(i-1)}/L^{(i)}$, d'où $(G^{(i-1)}: G^{(i)}) = (L^{(i-1)}: L^{(i)})$ et $(G: G^{(i)}) = (L: L^{(i)})$.

g étant un p-groupe, ce résultat permet de calculer le nombre des extensions galoisiennes $K \subset \mathfrak{R}$ de k de groupe donné g (qui est >0 si et seulement si le rang de g est $\leqq n_0 + 1$), ainsi que celui des extensions $K \subset \mathfrak{R}$ de k de degré n donné, dont le degré $n = (K: k)$ et l'ordre du groupe de Galois sont des puissances de p. Or, ce nombre est celui des extensions métagaloisiennes $K \subset \mathfrak{R}$ de k d'un tel degré. Il se calcule aussi par ma théorie.

National Center of Scientific Research,
Paris, France.

ALGEBRAIC GEOMETRY

THE FUNDAMENTAL IDEAS OF ABSTRACT ALGEBRAIC GEOMETRY

OSCAR ZARISKI

1. Introductory remarks. The past 25 years have witnessed a remarkable change in the field of algebraic geometry, a change due to the impact of the ideas and methods of modern algebra. What has happened is that this old and venerable sector of pure geometry underwent (and is still undergoing) a process of arithmetization. This new trend has caused consternation in some quarters. It was criticized either as a desertion of geometry or as a subordination of discovery to rigor. I submit that this criticism is unjustified and arises from some misunderstanding of the object of modern algebraic geometry. This object is not to banish geometry or geometric intuition, but to equip the geometer with the sharpest possible tools and effective controls. It is true that the lack of rigor in algebraic geometry has created a state of affairs that could not be tolerated indefinitely. Effective controls over the free flight of geometric imagination were badly needed, and a complete overhauling and arithmetization of the foundations of algebraic geometry was the only possible solution. This preliminary foundational task of modern algebraic geometry can now be regarded as accomplished in all its essentials.

But there was, and still is, something else and more important to be accomplished. It is a fact that the synthetic geometric methods of classical algebraic geometry, operating from a narrow and meager algebraic basis and faced by the extreme complexity of the problems of the theory of higher varieties, were gradually losing their power and in the end became victims to the law of diminishing returns, as witnessed by the relative standstill to which algebraic geometry came in the beginning of this century. I am speaking now not of the foundations but of the superstructure which rests on these foundations. It is here that there was a distinct need of sharper and more powerful tools. Modern algebra, with its precise formalism and abstract concepts, provided these tools.

An arithmetic approach to the geometric theories which we were fortunate to inherit from the Italian school could not be undertaken without a simultaneous process of generalization; for an arithmetic theory of algebraic varieties cannot but be a theory over arbitrary ground fields, and not merely over the field of complex numbers. For this reason, the modern developments in algebraic geometry are characterized by great generality. They mark the transition from classical algebraic geometry, rooted in the complex domain, to what we may now properly designate as *abstract algebraic geometry*, where the emphasis is on abstract ground fields.

2. Revision of the concept of a variety. My object is to present some of the fundamental ideas of abstract algebraic geometry. I must begin with the very

concept of a variety, since the arithmetic point of view led to a subtle revision of this concept and revealed some of its aspects that were not visible in the classical case. What I want to discuss in this connection concerns the following two topics: (1) the set-theoretic modifications in our conception of a variety as a set of points, modifications which were made methodologically necessary by the introduction of the well-known notion of a general point of an irreducible variety, due to Emmy Noether and van der Waerden; (2) the distinction between *absolute* and *relative* properties of a variety, a distinction which was made possible only by the admission of arbitrary ground fields.

If we wish to arrange matters so that the general point of an irreducible variety be an actual point of the variety, we must allow point coördinates which are elements of some transcendental extension of the ground field. Furthermore, in the theory of algebraic correspondences it is essential to operate simultaneously with any finite number of independent general points of one and the same variety. It follows that we must have a reservoir of infinitely many transcendentals for the point coördinates in our geometry. For these reasons, it was found convenient, following Busemann and André Weil, to fix once and for always a *universal coördinate domain*; this is to be an algebraically closed field having infinite transcendence degree over the particular ground field k in which we happen to be interested. Once this universal domain has been fixed, only such ground fields will be allowable which are subfields of the universal domain and over which the universal domain has infinite transcendence degree. We deal then with projective spaces over the universal domain, and all our varieties will be immersed in these spaces.

This being so, if k is any allowable ground field and if a variety V admits a system of defining equations with coefficients in k, then k is said to be a *field of definition of* V. Naturally, any variety V has infinitely many fields of definition. A property of V is *relative* or *absolute* according as it does or does not depend on the choice of the field of definition of V. For example, irreducibility of a variety is a relative property. But we also have the so-called *absolutely irreducible* varieties which are irreducible over each one of their fields of definition. The concept of the general point (x) of an irreducible variety V/k is a relative concept. On the other hand, *the dimension* of that irreducible variety V/k, i.e., the transcendence degree of the function field $k(x)$ of V/k, is an absolute concept.

A necessary and sufficient condition that a variety V be absolutely irreducible is that it be irreducible over some algebraically closed field of definition. An equivalent condition is the following: an irreducible variety V/k is absolutely irreducible if the ground field k is quasi-maximally algebraic in the function field $k(x)$ of V/k, i.e., if every element of $k(x)$ which is separably algebraic over k belongs to k. In theory, it would be sufficient to restrict the study of varieties to absolutely irreducible varieties, since any variety has a unique representation as a sum of absolutely irreducible varieties. However, in practice, and especially in the foundations, such a restriction introduces unnecessary complications.

It may be advisable to give a special name to those varieties which admit every (allowable) ground field as field of definition. Obviously, these are the varieties which are defined over the prime field of the given characteristic p. I propose to call them *universal varieties*. The projective space and the Grassmannian varieties are examples of universal varieties. Another important class of universal varieties is obtained by considering the set of all algebraic varieties, of a given order and dimension, in the n-dimensional universal projective space and introducing in that set an algebro-geometric structure based on the Chow coördinates of a cycle. The study of these varieties (of which the Grassmannian varieties are special cases) is closely connected with the outstanding problem of developing a theory of algebraic equivalence of cycles on a given variety, and will no doubt be a fundamental object of future research.

The definition of a variety as a set of points having coördinates in the universal domain has some startling, and perhaps unpleasant, set-theoretic implications. We have populated our varieties with points having coördinates which are transcendental over k. Thus, if x and y are independent variables the pair (x, y) is a legitimate point of the plane; and—what is worse—if x' and y' are other independent variables, then (x', y') is another point of the plane, quite distinct from the point (x, y). This *is* shocking, especially if we recall that our universal domain has infinite transcendence degree and that consequently we have created infinitely many replica of that ghostlike point (x, y). However, we are dealing here with a methodological fiction which is extremely useful in proving very real theorems. For instance, the entire theory of specializations is based on this set-theoretic conception of a variety, and the entire elementary theory of algebraic correspondences can be developed on that basis in the most effortless and simple fashion. Furthermore, most results concerning irreducible subvarieties of a given variety can be best expressed and derived as results concerning the general points of these subvarieties.

Nevertheless fiction remains fiction even if it is useful, and I feel that perhaps our varieties have altogether too many points to be good geometric objects. As the theory progresses beyond its foundational stage, some cuts and reductions may become necessary. Thus, one may begin first of all by eliminating isomorphic replica of points, by identifying points which are isomorphic over the given ground field k. Or one may restrict the coördinate domain to the algebraic closure of k. Or one may do both of these things at the same time. I have no strong convictions on these issues, and I am quite content in leaving their settlement to the future development of the theory of algebraic varieties. But to round up this discussion, let me indicate briefly some topological aspects of these issues.

Given a variety V and given any field k (not necessarily a field of definition of V), there is a natural topology on V, *relative to* k: it is the topology in which the closed sets are intersections of V with varieties which are defined over k. In particular, if V itself is defined over k, then the closed sets on V are the subvarieties of V which are also defined over k, and in terms of this topology

general points and specialization of points are easily defined. Thus, a general point of an irreducible variety V/k is a point whose closure is the entire variety V; a point Q is the specialization of a point P, over k, if Q belongs to the closure of P. It is clear that in this topology even the weakest separation axioms are not valid. The only points of V which constitute closed sets are the points having coördinates which are pure inseparable over k. Hence V is not a T_1-space. It is not even a T_0-space, for if P and Q are k-isomorphic points, then each belongs to the closure of the other. However, if we identify k-isomorphic points of V, we restore the separation axiom T_0. If, moreover, we restrict the coördinate domain to the algebraic closure of the ground field k, then V becomes a T_1-space.

An even more radical revision of the concept of a variety has been offered by André Weil. His so-called *abstract varieties* are not defined as subsets of the projective space, but are built out of pieces of ordinary varieties, pieces that must hang together in some well-defined fashion. It is still an open question whether the varieties of Weil can be embedded in the projective space.

In all that precedes I have used deliberately the term "general point" rather than that of "generic point". When the Italian geometers speak of a property enjoyed by the *generic* point of an irreducible variety, they mean a property that is enjoyed by all points of V, except perhaps those which belong to some proper subvariety of V. It is clear that this is not equivalent to saying that the *general* point of V has that given property. There *is* equivalence if and only if we are dealing with a property of points that can be expressed by equations and inequalities (with coefficients in k) connecting the point coördinates. But not every algebro-geometric property is of this category. For instance, it is possible to define algebraically the notion of analytical irreducibility of a variety V at a point. Now if W is, say, an algebraic curve on V, V may be analytically reducible at the generic algebraic point of the curve and analytically irreducible at all the general points of the curve. From our point of view, according to which W consists of both algebraic and transcendental points, either one of the following statements is false: (1) V is analytically irreducible at the generic point of W; (2) V is analytically reducible at the generic point of W. In the complex domain this corresponds to the following state of affairs: at the generic (complex) point of the curve W the variety V decomposes into several analytical branches, but these branches are permuted transitively along closed paths traced on the Riemann surface of the curve W. This is a good example of the difference between the meanings of general and generic.

I shall pass now, without delay, to more concrete topics dealing with the major developments in abstract algebraic geometry. Roughly speaking, these major developments come under the following headings: (1) theory of specializations; (2) normal varieties; (3) analytical methods; (4) theory of valuations; (5) Abelian varieties. Time will not allow me to discuss the very general and elegant theory of Abelian varieties which we now possess and for which André

Weil is entirely responsible. Let me, then, first make a few remarks about the theory of specializations.

3. The theory of specializations. Specialization arguments in abstract algebraic geometry are the arithmetic substitute, or analogue, of continuity arguments of classical algebraic geometry, and have been largely developed by van der Waerden. The theory of specializations centers around one basic fact, concerning *extensions of specializations:* if an algebraic function f is defined at the general point of an irreducible variety V/k, then it is possible to extend the domain of the function f to the entire variety V, including therefore also those points of V at which the explicit expression of f is indeterminate. In this statement, the function f need not be a numerical function; the values of f may be points of another variety V', and when that is so, we are dealing with an algebraic correspondence between V and V'. It is well-known that the theorem on extensions of specializations is equivalent to the existence of resultant systems in elimination theory. Without advocating the elimination of elimination theory, it may be pointed out that also the Hilbert Nullstellensatz, *in its homogeneous form*, can be used as a foundation for the theory of specializations (and hence also of algebraic correspondences). In fact, the Nullstellensatz provides a key to the whole of the elementary theory of algebraic varieties, including such topics as the dimension theory, the principal ideal theorem (in its geometric formulation), the decomposition of a variety under ground field extensions, etc. In my forthcoming Colloquium book, this part of the theory of varieties is built entirely around the Hilbert Nullstellensatz.

One of the most important applications of the theory of specializations was the development of the general intersection theory. This application is due to van der Waerden and André Weil, with the work of Severi serving as a geometric background for the general plan of this undertaking. At present, then, we have a complete intersection theory which is valid for any nonsingular variety over an arbitrary algebraically closed ground field. A parallel development is the intersection theory for algebraic manifolds due to Chevalley. Chevalley's theory is an outstanding example of the arithmetization of some of the concepts and methods of the theory of analytic functions which are used in algebraic geometry. As far as the local analytical treatment is concerned, modern algebra has provided us with the necessary tools. I refer to the theory of local rings and their completions, due to Krull and Chevalley and further enriched by important contributions by I. S. Cohen and P. Samuel.

The present intersection theories all have an absolute character, since they refer to an algebraically closed ground field. It is still an open question whether there exists a consistent *relative* intersection theory, i.e., an intersection theory relative to a given ground field. The fact that there exists such a thing as the *relative order* of a variety seems to indicate the possible existence of a relative intersection theory. Another unsolved question is whether there exists a rea-

sonable intersection theory on varieties which have singularities, for instance—and above all—on normal varieties. The example of algebraic cones shows that in the case of singular varieties one may have to use fractional intersection numbers and—more generally—fractional cycles.

4. Normal varieties. I shall now discuss briefly the concept of a normal variety, especially from the standpoint of the theory of specializations. This concept, which is purely arithmetic in character, turned out to be a useful contribution even to classical algebraic geometry.

An irreducible variety V/k is said to be *normal at a point* Q if the local ring of V at that point is integrally closed. The variety V is *normal* if it is locally normal at every point. Normality is a relative property. The *normalization* of a variety V consists in passing from V to a birationally equivalent variety $\bar{V}$ such that: (1) $\bar{V}$ is normal; (2) the birational transformation between V and $\bar{V}$ has no fundamental points on V, i.e., to every point of V there corresponds on $\bar{V}$ at most a finite number of points. By these two conditions the normal variety $\bar{V}$ is uniquely determined by V, to within a regular birational transformation (a birational transformation is regular if it is (1,1) without exceptions and if corresponding points have the same local ring). Normal varieties were originally introduced in connection with the problem of the resolution of singularities, for a normal variety of dimension r has the property that its singular locus is of dimension at most $r - 2$. This property of normal varieties is connected with the well-known fact that as far as the minimal prime ideals are concerned, the ideal theory of integrally closed Noetherian domains does not differ essentially from the classical ideal theory of Dedekind domains. At any rate, the process of normalization does have the effect of resolving all the singular loci of V, of dimension $r - 1$.

But there are other properties of normal varieties which are of particular interest for the theory of specializations, and hence also for the general theory of algebraic correspondences. Suppose that V is a normal variety and that T is a birational transformation of V into some other variety V'. Then the following theorem holds: *if to a given point Q of V there corresponds on V' more than one point, then* (1) *the point Q is fundamental for the birational transformation T, i.e., to Q there corresponds on V' an infinite set of points, and* (2) *the set of points of V' which correspond to Q is a variety, all irreducible components of which are of positive dimension.* The really significant and nontrivial part of this theorem is the second part. It is this that I have given in a Transactions paper as the "main theorem" on birational transformations and for which I gave a short and simple proof, based on valuation theory and the theory of local rings, in a recent note in Proc. Nat. Acad. Sci. U. S. A. In terms of specializations, the above "main theorem" signifies, roughly speaking, that, in the case of normal varieties, the presence of an isolated specialization implies the uniqueness of that specialization. As was pointed out by André Weil, this property of normal varieties leads at once to the proof of the uniqueness of the intersection mul-

tiplicity of two varieties at a common isolated intersection. The uniqueness of intersection multiplicity is, on the other hand, the crucial point of the whole of intersection theory, and the fact that this point can be disposed of in such a casual manner by the use of a general theorem on normal varieties illustrates the usefulness of the concept of normality. The treatment of the intersection theory would be further simplified, in fact the whole theory would become almost trivial, if one could prove the normality of any complete algebraic system of cycles in the projective space, i.e., the normality of the Chow-van der Waerden representative variety of such a system. It would even be sufficient to prove that this variety is analytically irreducible everywhere, for in that case the normalization process would lead to another representative variety, whose points are still in (1,1) correspondence with the cycles of the system (without exceptions) but which is *normal.*

Another important aspect of normal varieties has to do with the theory of complete linear systems. Any normal variety V has the following *characteristic* property: *the hypersurfaces of a sufficiently high order n cut out on V a complete linear system* $|nC|$, where C is any hyperplane section of V. This result, in conjunction with Hilbert's postulation formula and the existence of derived normal models, leads at once to an expression of the dimension of the complete system $|nC|$ (n large), whether V is normal or not. In a joint paper of Muhly and myself, now in course of publication in Trans. Amer. Math. Soc., we define the *virtual arithmetic genus* $p(V)$ as $\pm$ the constant term in Hilbert's postulation formula of V. In the case of algebraic surfaces, we prove that, in any birational class $\{V\}$ of normal surfaces V, the numerical character $p(V)$ is a monotone nonincreasing function of V, with respect to the following partial ordering of the class: $V < V'$ if the birational transformation from V' to V is single-valued without exceptions. An essential ingredient of the proof is the remark that, since a normal surface V has only a finite number of singular points, the generic *hyperplane section of V is normal* (since it is a curve free from singularities). Now the normality of a generic hyperplane section of *any* normal variety has recently been established by Seidenberg. In virtue of this interesting result of Seidenberg, the monotone character of the virtual arithmetic genus $p(V)$ can now be regarded as established for normal varieties of any dimension. In particular, $p(V)$ is invariant under regular birational transformations.

Added in proof: In view of the technical difficulties of the proof of Seidenberg's theorem, we point out that the results of our joint paper with Muhly do not actually require Seidenberg's theorem. All that is needed is the following statement: the *general* hyperplane section of a normal variety V/k (k algebraically closed) is absolutely normal. This statement is an immediate consequence of the theorem of Bertini, of the classical Jacobian criterion for simple points, and of Weil's characterization of absolutely normal varieties.

In the joint paper of Muhly and myself it is proved that (a) the function $p(V)$, defined in a given birational class of varieties of dimension $\leqq 3$, has a

minimum; (b) this minimum is reached for the nonsingular varieties of the class; (c) this minimum is equal to the effective arithmetic genus of the field of algebraic functions determined by the given birational class. Each of these statements represents an unsolved problem for varieties of dimension greater than three.

5. Holomorphic functions and the principle of degeneration. The "main theorem" on birational transformations is a special case of a much more general "connectedness theorem" on algebraic correspondences, a theorem which in its turn contains as a special case a principle of degeneration for varieties over arbitrary ground fields. The proof of this theorem is based on a theory of abstract holomorphic functions which I have developed in a paper now in course of publication in the Memoirs of the American Mathematical Society, and which represents an extension of the analytical methods of abstract algebraic geometry from a local theory to a theory in the large. The use of normal varieties is essential in this theory. I shall now briefly outline the geometric background and the underlying ideas of this work.

In the classical case, the principle of degeneration (first formulated by Enriques) asserts that if an irreducible variety V varies continuously and degenerates in the limit into a reducible variety V_0, then this limit variety is connected. This principle is almost self-evident, since V_0 is a continuous image of the irreducible—and therefore connected—variety V. I say "almost evident", because in order to assert that V_0 is a continuous map of V it would be necessary to show that the continuous variation of V can be accompanied by a continuous *deformation* of V into V_0. The existence of such a deformation has always been taken more or less for granted.

Now, the principle of degeneration can easily be transformed into an equivalent statement in which no reference is made to continuity or limits and which therefore makes sense also in the abstract case. First of all, if V is a variety and k is any ground field (not necessarily a field of definition of V), then the expression "V is connected over k" has a meaning, since V has a natural topology over k. In particular, if k is an algebraically closed field of definition of V and if V is connected over k, then it is easy to see that V is connected over every one of its fields of definition. We say then that V is *absolutely connected.* This definition can then be extended in an obvious fashion to effective algebraic *cycles*, i.e., to formal linear combinations, with positive integral coefficients, of absolutely irreducible varieties of the same dimension r. Now let M be an irreducible algebraic system of r-dimensional cycles, and suppose that M is defined and irreducible over a given ground field k. Then I prove the principle of degeneration under the following form:

If the general cycle of M/k is absolutely irreducible, then every cycle in M is absolutely connected.

It is easy to transform this principle into a statement concerning the incidence correspondence associated with the system M, i.e., the correspondence

in which to every cycle of M correspond all the points which belong to that cycle. The transformed statement can be itself incorporated into the following more general *connectedness theorem on algebraic correspondences*:

THEOREM. *Let T/k be an irreducible algebraic correspondence between two varieties V and V' and let (P, P') be a general point pair of T/k. We make the following assumptions*:

(1) *T^{-1} is rational [i.e., $k(P) \subset k(P')$] and semi-regular.*

(2) *The field $k(P)$ is maximally algebraic in $k(P')$. Let W/k be any connected subvariety of V. Then if V is analytically irreducible (in particular, locally normal) at each point of W, the total transform of W under T is a k-connected subvariety of V'.*

This last theorem covers a good deal more ground than does the principle of degeneration. The incidence correspondence of an algebraic system has the special property (not shared by arbitrary correspondences) that it has no fundamental points: to every point of the representative variety V of the system M there corresponds, on the carrier variety V', a variety of the same fixed dimension r, r = the dimension of the cycles of the system M. If we deal, however, with an arbitrary correspondence T and if a point Q of V happens to be a fundamental point of T, then the variety $T\{Q\}$, whose connectedness is being claimed, may very well have dimension higher than that of the total transform of the general point of V/k. The connectedness of that variety $T\{Q\}$ is, in that case, not at all trivial even in the classical case (and—to our knowledge—has never been proved in the classical case, even for *birational* transformations T).

It is in the proof of this theorem that the holomorphic functions come directly into play. With the given subvariety W of V we associate "functions" on V which "are defined and holomorphic" along W. These functions are, by definition, certain specified elements of the direct product of the completions of the local rings of V at the various points of W. They are those elements of this direct product which can be represented by a finite number of sequences of elements of the function field of V, in such a manner that (1) each sequence converges uniformly on some open subset Γ_i of W and (2) the sets Γ_i cover W. These functions form a ring which we denote by o_W^*. As a first tangible evidence of the nonartificiality of this new concept, we have the following:

CONNECTEDNESS CRITERION. *If V is analytically irreducible at each point of W, then W/k is connected if and only if the ring o_W^* of holomorphic functions along W is an integral domain.*

With this theorem, we are still very far from the proof of the connectedness theorem for algebraic correspondences. The high point, and also the most difficult part, of the whole theory is still to come. It is represented by a theorem of

invariance of rings of holomorphic functions under rational transformations. It is noteworthy that while the proof of the principle of degeneration in the classical case is essentially a simple exercise in topology, our proof of this principle in the abstract case can be given only after a long and difficult journey. The end of this journey is as follows:

Let T be an algebraic correspondence between two irreducible varieties V/k and V'/k such that T^{-1} is a rational transformation, semi-regular at each point of V'. Let W' be the total transform $T\{W\}$ of W, where W is any subvariety of V. We have the ring $o^*_{W'}$ of functions on V', defined and holomorphic along W'. We first show that there *always exists a natural isomorphism* $H_{W,W'}$ *of* o^*_W *into* $o^*_{W'}$. The fundamental theorem of invariance asserts the following:

THEOREM. *Let* (P, P') *be a general point pair of* T/k. *If* (a) $k(P)$ *is maximally algebraic in* $k(P')$ *and* (b) *if* V *is locally normal at each point of* W, *then* $H_{W,W'}$ *is an isomorphism of* o^*_W *onto* $o^*_{W'}$. [Note that condition (a) is automatically satisfied if T is a birational transformation.]

This theorem, together with the above connectedness criterion, gives immediately the connectedness theorem for algebraic correspondences.

There is a number of very difficult problems suggested by the theory of holomorphic functions and which are still open even in the classical case. One of them is to prove that:

I. *The ring* o^*_W *is Noetherian.*

The elements of o^*_W which belong to the field of rational functions on V form themselves a ring, denoted by o_W. This ring is the intersection of all the local rings of the points of W. The second problem is to prove the following:

II. *The ring* o_W *is Noetherian.*

The non units in o_W form an ideal, say m. Whether or not o_W is Noetherian, one may consider the completion of o_W with respect to the powers of this ideal. The next problem is to prove the following:

III. *If* W *is connected, then* o^*_W *is the completion of* o_W.

A special case of III is the following conjecture: *if* o_W *consists only of constants, then* $o_W = o^*_W$. The existence of nonconstant holomorphic functions on V, defined along W, is closely connected with the existence of a rational transformation T of V into some other variety V', such that W is the total transform of a point of V'. It is obvious that the latter implies the former, but I have no proof of the converse. At any rate, if W can be transformed into a point Q' of some variety V' by a rational transformation, then it is easily seen that $o_W = o_{Q'}$, and from the theorem of invariance of rings of holomorphic functions it follows that $o^*_W = o^*_{Q'}$. Hence in this case the conjectures I, II, III are true.

6. Transcendental theory of specializations. My report is very incomplete as it stands, but it would be glaringly incomplete if I had not said anything about the role of valuation theory in abstract algebraic geometry. The ordinary theory

of specializations applies to finite sets of quantities; it is a finite theory of specializations. This theory does not do everything that continuity does in classical geometry, for it contains nothing that corresponds to the notion of a branch, whether algebraic or transcendental. It does not tell us anything about the different modes of approach to a point on a given variety. For this reason, the finer differential aspects of the local geometry of a variety, in particular the analysis of the neighborhood of a singular point, are outside the province of the finite theory of specializations. What was needed here was a theory which deals with the simultaneous specialization of all the rational functions on a given variety, therefore a transcendental theory of specializations. Valuation theory meets precisely this requirement. It is to be observed that it was precisely the *general* valuation theory, as developed by Krull, i.e., the theory of valuations having arbitrary value groups (nondiscrete as well as discrete, non-Archimedian value groups as well as Archimedian), that turned out to be the necessary tool for the solution of such a concrete algebro-geometric problem as the local uniformization of algebraic varieties and for partial progress in the problem of the resolution of singularities. Nothing less than the general valuation theory would have served that purpose, and that is so for the following two reasons:

(1) Without the *general* concept of a valuation it is not possible even to formulate the problem of local uniformization in pure algebraic terms. The nearest algebraic substitute for the neighborhood of a point Q of a variety V is the set of all modes of approach to the point Q, therefore essentially, the set of all valuations of the function field of V which have *center* at Q. Here, by the center of a valuation v we mean the point whose coördinates are the v-residues of the coördinates of the general point of the variety V. And the only statement which can reproduce in algebraic terms, and without loss of power, that what has been a classical conjecture, namely that the complete neighborhood of the point Q can be represented by a finite number of power series expansions, is the statement that *any valuation v of center Q can be uniformized with respect to V*. By this I mean that there exists a birational transform V' of V such that (a) the center of v on V' is a simple point Q' of V' and (b) the local ring of V at Q is contained in the local ring of V' at Q' (this second condition implies that every element of the function field of V which is holomorphic at Q is also holomorphic at Q').

(2) The set of all valuations of the function field of V is a compact space in a suitable natural topology of that set. This space is called the *Riemann surface* of the function field of V. Now it is the compactness of the Riemann surface that makes it possible to apply the theorem of local uniformization of abstract algebraic geometry (i.e., the uniformization of a single valuation) to the classical problem of local uniformization of the complete neighborhood of a point of a variety. No reasonable proper subset of the set of all valuations has that compactness property. Thus, the set of all discrete valuations is not compact (except in the case of *curves*, in which case all valuations are, of course, discrete),

and neither is the set of all algebraic valuations (represented by algebraic branches).

The problem of local uniformization has been settled, so far, only in the case of characteristic zero. The extension of the present proof to the case of nonzero characteristic will call for considerable algebraic skill and ingenuity. It is not a problem for geometers; it is a problem for algebraists with a feeling and intuition for all the unpleasant things that can happen in the case of nonzero characteristic. The essential difficulties of this problem are already apparent in the case of algebraic surfaces over an algebraically closed ground field. In this case I can prove the theorem of local uniformization for every valuation except those which are nondiscrete and have rational rank 1 (i.e., those whose value group is a dense set of rational numbers). The case which is particularly difficult is the one in which the value group of the valuation contains rational numbers having in the denominator arbitrarily high powers of the characteristic p. In this case, already such a simple surface as $z^p = f(x, y)$ becomes untractable.

Also the problem of the resolution of singularities is still unsolved; or—to put it into more cautious terms—no solution of this problem has ever come to my direct attention. For characteristic zero, this problem has been solved, so far, only for varieties of dimension $\leqq 3$.

7. Concluding remarks. The Italian geometers have erected, on somewhat shaky foundations, a stupendous edifice: the theory of algebraic surfaces. It is the main object of modern algebraic geometry to strengthen, preserve, and further embellish this edifice, while at the same time building up also the theory of algebraic varieties of higher dimension. The bitter complaint that Poincaré has directed, in his time, against the modern theory of functions of a real variable cannot be deservedly directed against modern algebraic geometry. We are not intent on proving that our fathers were wrong. On the contrary, our whole purpose is to prove that our fathers were right.

The arithmetic trend in algebraic geometry is not in itself a radical departure from the past. This trend goes back to Dedekind and Weber who have developed, in their classical memoir, an arithmetic theory of fields of algebraic functions of one variable. Abstract algebraic geometry is a direct continuation of the work of Dedekind and Weber, except that our chief object is the study of fields of algebraic functions of more than one variable. The work of Dedekind and Weber has been greatly facilitated by the previous development of classical ideal theory. Similarly, modern algebraic geometry has become a reality partly because of the previous development of the general theory of ideals. But here the similarity ends. Classical ideal theory strikes at the very core of the theory of algebraic functions of one variable, and there is in fact a striking parallelism between this theory and the theory of algebraic numbers. On the other hand, the general theory of ideals strikes at most of the foundations of algebraic geometry and falls short of the deeper problems which we face in the post-foundational stage. Furthermore, there is nothing in modern commutative

algebra that can be regarded even remotely as a development parallel to the theory of algebraic function fields of more than one variable. This theory is after all itself a chapter of algebra, but it is a chapter about which modern algebraists knew very little. All our knowledge here comes from geometry. For all these reasons, it is undeniably true that the arithmetization of algebraic geometry represents a substantial advance of algebra itself. In helping geometry, modern algebra is helping itself above all. We maintain that abstract algebraic geometry is one of the best things that happened to commutative algebra in a long time.

HARVARD UNIVERSITY,
CAMBRIDGE, MASS., U. S. A.

NUMBER-THEORY AND ALGEBRAIC GEOMETRY

André Weil

Mr. Chairman, Ladies and Gentlemen,

The previous speaker concluded his address with a reference to Dedekind and Weber. It is therefore fitting that I should begin with a homage to Kronecker. There appears to have been a certain feeling of rivalry, both scientific and personal, between Dedekind and Kronecker during their life-time; this developed into a feud between their followers, which was carried on until the partisans of Dedekind, fighting under the banner of the "purity of algebra", seemed to have won the field, and to have exterminated or converted their foes. Thus many of Kronecker's far-reaching ideas and fruitful results now lie buried in the impressive but seldom opened volumes of his Complete Works. While each line of Dedekind's XIth Supplement, in its three successive and increasingly "pure" versions, has been scanned and analyzed, axiomatized and generalized, Kronecker's once famous *Grundzüge* are either forgotten, or are thought of merely as presenting an inferior (and less pure) method for achieving part of the same results, viz., the foundation of ideal-theory and of the theory of algebraic number-fields. In more recent years, it is true, the fashion has veered to a more multiplicative and less additive approach than Dedekind's, to an emphasis on valuations rather than ideals; but, while this trend has taken us back to Kronecker's most faithful disciple, Hensel, it has stopped short of the master himself.

Now it is time for us to realize that, in his *Grundzüge*, Kronecker did not merely intend to give his own treatment of the basic problems of ideal-theory which form the main subject of Dedekind's life-work. His aim was a higher one. He was, in fact, attempting to describe and to initiate a new branch of mathematics, which would contain both number-theory and algebraic geometry as special cases. This grandiose conception has been allowed to fade out of our sight, partly because of the intrinsic difficulties of carrying it out, partly owing to historical accidents and to the temporary successes of the partisans of purity and of Dedekind. It will be the main purpose of this lecture to try to rescue it from oblivion, to revive it, and to describe the few modern results which may be considered as belonging to the Kroneckerian program.

Let us start from the concept of a point on a variety, or, what amounts to much the same thing, of a specialization. Take for instance a plane curve C, defined by an irreducible equation $F(X, Y) = 0$, with coefficients in a field k. A point of C is a solution (x, y) of $F(X, Y) = 0$, consisting of elements x, y of some field k' containing k. In order to define the function-field on the curve, we identify two polynomials in X, Y if they differ only by a multiple of F, i.e., we build the ring $k[X, Y]/(F)$, and we take the field of fractions $\mathfrak{K}$ of that ring: in particular, X and Y themselves determine the elements $\bar{X} = X \bmod F$,

$\bar{Y} = Y \bmod F$, of $\mathfrak{R}$, and $(\bar{X}, \bar{Y})$ is a point of C, called generic since it does not satisfy any relation over k except $F(\bar{X}, \bar{Y}) = 0$ and its consequences. Then any point (x, y) of C, with coordinates in an extension k' of k, determines a homomorphism σ of the ring $k[X, Y]/(F)$ into k', defined by putting $\sigma(\bar{X}) = x$, $\sigma(\bar{Y}) = y$, and $\sigma(\alpha) = \alpha$ for every $\alpha \in k$; this homomorphism is also called a *specialization* of that ring, and a *generic* one if it is an isomorphism of it into k'; consequently, (x, y) will be called a specialization of $(\bar{X}, \bar{Y})$, and will be called generic if σ is generic.

Our homomorphism σ has been so defined as to preserve the elements of the "ground-field" k; but this restriction, usual as it is in algebraic geometry, may well prove too narrow for some purposes. If, for example, we consider a curve $F(X, Y, t) = 0$, depending upon a parameter t, where F is a polynomial in X, Y, t with coefficients in a field k, then the coefficients of the equation of the curve are in the field $k(t)$. However, with our curve, we naturally associate the surface $F(X, Y, T) = 0$; the curve then appears as a plane section of that surface by the plane $T = t$. Because of this changed point of view, the parameter t, previously frozen by its inclusion in the field of "constants", is now liberated and available for specialization; and so we are free now to consider as a specialization of our ring $k[X, Y, t]/(F(X, Y, t))$ any homomorphism of that ring into an extension k' of k, still preserving the elements of k, but mapping $\bar{X} = X \bmod F, \bar{Y} = Y \bmod F, \bar{t} = t \bmod F$ onto any three elements x', y', t' of k' satisfying $F(x', y', t') = 0$. Thus no longer restricted to the exclusive consideration of the "generic" curve belonging to the family $F(X, Y, t) = 0$, we are enabled to consider any specialization $F(X, Y, t') = 0$ of that curve, and the whole surface $F(X, Y, t) = 0$ spanned by that family.

This shifting of our point of view necessitates a re-examination of the concept of ground-field and of the field of definition of a variety. The previous speaker has mentioned, as one of the main achievements of modern algebraic geometry, the possibility of operating over quite arbitrary ground-fields. One should not be blind, however, to the somewhat illusory nature of this achievement. As our knowledge of algebraic curves is fairly extensive, there is, it is true, a great deal that we can say on the curve $F(X, Y, t) = 0$ depending upon the parameter t, in the example discussed above; and we should not possess that knowledge if our methods of proof were not valid over the ground-field $k(t)$. But as we have pointed out, all we can say on the curve $F(X, Y, t) = 0$ is but part of the theory of the surface $F(X, Y, T) = 0$. This may be at the present moment, and it is in fact, one of the best ways of acquiring some knowledge of the geometry on that surface; but the fact remains that, in the final analysis, any statement on a variety with a larger ground-field boils down to a statement on a variety (of higher dimension, and therefore intrinsically more difficult to study) over a smaller ground-field.

Now consider, with Kronecker, that, in most problems of algebraic geometry, only a finite number of points and varieties occur at a time; these will necessarily have a common field of definition which is finitely generated over the prime

field, i.e., which is generated over the prime field (the field Q of rational numbers if the characteristic is 0, and otherwise a finite field) by a finite number of quantities $(t_1, \cdots, t_N)$; if these are considered as parameters, and are made available for specialization, then, in the final analysis, *every statement we can make can be thought of as a theorem in algebraic geometry over an absolutely algebraic ground-field*, i.e., either *over a finite field* or *over an algebraic number-field* of finite degree. While this realization, of course, cannot in any way detract from the methodological importance of arbitrary ground-fields as one of the chief tools of modern algebraic geometers, it gives us some insight into the deep meaning of Kronecker's view, according to which the absolutely algebraic fields are the natural ground-fields of algebraic geometry, at any rate as long as purely algebraic methods (as distinct from analytical or topological methods) are being used. Now these are fields with strongly marked individual features, which will undoubtedly have to be taken more and more into account as algebraic geometry develops along more Kroneckerian lines. For instance, the field with q elements can be characterized by the fact that its elements are invariant under the automorphism $x \to x^q$ of any field containing it; this must have a profound influence on the geometry over that field; and recent work connected with the Riemann hypothesis ([11e]) fully confirms that expectation. Another fact, so far an isolated one, in the same direction, is the existence of matrices, associated with curves over a finite field, which bear a curious resemblance with the period-matrices of abelian integrals in the classical theory (cf. [11d]).

We are now in a position to discuss specializations again from our broadened point of view. If e.g. $F(X, Y) = 0$ is the equation of a curve, with coefficients in a subring R of a field k, then any homomorphism σ of the ring $R[X, Y]/(F)$ into a field k' will be called a specialization of that ring; if in particular it preserves (or at least if it maps isomorphically) the elements of R, then it can be extended to a homomorphism of $k[X, Y]/(F)$ which preserves the elements of k.

As Kronecker realized, this affects our concept of dimension. Take for instance, instead of our curve, a hypersurface $F(X_1, \cdots, X_n) = 0$ in n-dimensional space, with coefficients in a subring R of a field k; let $\mathfrak{R}$ denote the ring $R[X_1, \cdots, X_n]/(F)$. Then the dimension $n - 1$ of that variety can be defined as the degree of transcendency, over the ground-field k, of the function-field on the variety, i.e., of the field of fractions of $\mathfrak{R}$, or, equivalently, as the maximum number of successive specializations σ, σ', σ'', $\cdots$, of $\mathfrak{R}$ onto a ring $\mathfrak{R}'$, of $\mathfrak{R}'$ onto a ring $\mathfrak{R}''$, etc., each one of which preserves the elements of R, and none of which is an isomorphism; the rings $\mathfrak{R}'$, $\mathfrak{R}''$, $\cdots$ are understood to be "integral domains" (i.e., subrings of fields). If we remove the condition that the specializations σ must preserve the elements of R, but merely require that they should preserve the elements of the "prime ring" (the ring Z of integers if the characteristic is 0, the ring of integers mod p if it is $p > 1$), this gives us the dimension over the prime field, or absolute dimension. So far, we have not crossed the boundaries of ordinary algebraic geometry, even though we may have pushed down the ground-field to an absolutely algebraic field. In particular, if the char-

acteristic is $p > 1$, every homomorphism must preserve the elements of the prime field, and so there is no temptation, nor even any possibility, for us to cross those boundaries. However, if the characteristic is 0, there are homomorphisms which do not preserve the characteristic; as soon as we allow these to enter the picture, we are within a wider area, where algebraic geometry and number-theory commingle and cannot be kept apart; and, as a consequence, the proper concept of dimension is the Kroneckerian concept. Since our sequences of specializations $\sigma, \sigma', \cdots$ can now be increased by one which changes the characteristic from 0 to some $p > 1$, it follows that the Kroneckerian dimension is higher by 1 than that of algebraic geometry proper. For instance, a curve over an algebraic number-field has the Kroneckerian dimension 2.

In this sense, the only two cases of dimension 1 are those of a curve over a finite field, and of an algebraic number-field. In fact, it has been well known, ever since Kronecker and Dedekind, that there are far-reaching analogies between these two cases, and these have been among the chief sources of progress in both directions; indeed, we have reached a stage where we can deal simultaneously with large segments of both theories, not merely the more elementary ones, but also class-field theory and part of the theory of the zeta-function. It is true that these analogies are still incomplete at some crucial points; new concepts are clearly needed before we can transport to number-fields, even conjecturally, the facts about the Jacobian variety of a curve which have recently led to the proof of the Riemann hypothesis ([11e], [11f]). Nevertheless, our knowledge of these topics is fairly extensive, whereas the same can hardly be said of the problems in higher dimensions.

It is true that the theory of local rings has been extensively developed, largely by its initiator Krull (cf. e.g. [6]), and more recently by Chevalley ([2]), I. Cohen ([4]), and others. Such rings arise as follows: σ being, as above, a specialization, say, of the ring $\mathfrak{R} = R[X, Y]/(F)$ defined by a curve $F(X, Y) = 0$, it can be extended to a homomorphic mapping of the ring $\mathfrak{R}'$ of those elements u/v of the field of fractions $\mathfrak{K}$ of $\mathfrak{R}$, for which u, v are in $\mathfrak{R}$ and $\sigma v \neq 0$, by putting $\sigma(u/v) = \sigma u/\sigma v$; $\mathfrak{R}'$ is the specialization-ring, and the ideal of non-units in $\mathfrak{R}'$, which is the kernel of σ, is the specialization-ideal; $\mathfrak{R}'$ is called a local ring, and its completion, with respect to the topology defined on it by the powers of the specialization-ideal, is a complete local ring; experience shows that it is desirable to confine oneself to integrally closed specialization-rings, and this leads to Zariski's fundamental concept of normality. Up to now geometers have used only characteristic-preserving specializations; therefore all their local rings contain a field, and have the same characteristic as their residue-class ring. Fortunately algebraists have not confined themselves to that case, so that their work is immediately available for the more general geometry that we are envisaging here.

We are thus led to modify the Kroneckerian view that the "true" or "natural" ground-fields in algebraic geometry are the absolutely algebraic fields; this is so as long as ground-fields are considered from the purely algebraic point of view,

without any additional structure. However, it is now clear that the study of a family of varieties at, or rather in the neighborhood of, a given specialization of the parameter leads at once to the consideration of algebraic varieties over complete local rings and their fields of fractions; some recent work by Chow ([3]) may be considered as pertaining to this subject, of which the "geometry on a variety in the neighborhood of a subvariety" (as exemplified chiefly by Zariski's theory of holomorphic functions ([12]) forms a natural extension. That this does not contradict the Kroneckerian outlook, but has its root in it, is clearly shown by the fact that the theory of local rings was originated by Hensel; his p-adic rings, in fact, are the complete local rings attached to the specializations of the rings of integers in algebraic number-fields. Hence the local study, say, of an algebraic curve $F(X, Y) = 0$ with coefficients in the ring Z of rational integers, "at" the specialization of Z determined by a prime p, amounts to enlarging the ground-field to the p-adic field. Thus the p-adic fields appear as another kind of "natural" ground-field, and one may expect that the geometry over such fields will acquire more and more importance as it learns to develop its own methods. One may quote here E. Lutz's results on elliptic curves ([7]), showing that the group of points on such a curve has a subgroup of finite index, isomorphic to the additive group of integers in the ground-field; similar results undoubtedly hold for Abelian varieties of any dimension. In his beautiful thesis, Chabauty ([1]), following ideas of Skolem ([10]), has shown how the method of p-adic completion, with respect to a more or less arbitrary prime p, can yield deep results about varieties over an algebraic number-field; there, as already in Skolem's work, the problem concerns the intersection of an algebraic variety and of a multiplicative group; by p-adic completion, the latter becomes an algebroid variety defined by linear differential equations. Of course geometry over finite fields may in a certain sense be obtained from the geometry over p-adic fields by reduction modulo p, so that the latter may be said to contain all that the former contains, and a good deal more; but little use has been made so far of the relations between these two kinds of geometries, and little is known about them.

But the geometry over p-adic fields, and more generally over complete local rings, can provide us only with local data; and the main tasks of algebraic geometry have always been understood to be of a global nature. It is well known that there can be no global theory of algebraic varieties unless one makes them "complete", by adding to them suitable "points at infinity," embedding them, for example, in projective spaces. In the theory of curves, for instance, one would not otherwise obtain such basic facts as that the numbers of poles and of zeros of a function are equal, or that the sum of residues of a differential is 0. One way of doing this (which, however, is effective only in the case of dimension 1) consists in considering the valuations of the field of functions on the curve; on a given affine model $F(X, Y) = 0$, each simple point defines a valuation, viz., that one which assigns, to each function on the curve, the order of the pole or zero it may have at that point; and all valuations, with a finite num-

ber of exceptions, can be so obtained; the exceptions correspond to the "multiple points" and to the "points at infinity", and give an invariant definition for these. Correspondingly, if we apply this idea to an algebraic number-field (also a one-dimensional problem), we obtain satisfactory formulations for global theorems, entirely analogous to the theorems on algebraic curves, provided we allow for "archimedean" valuations with somewhat weaker properties than those of algebraic geometry and than the p-adic valuations on number fields, viz., those for which the completed field is the field of real or that of complex numbers. Thus it appears that algebraic geometry over the complex number-field is, after all, a legitimate object of study, no less necessary or useful than geometry over p-adic fields; and so the door is opened to topology, function-theory, differential geometry, and partial differential equations. This, at any rate, is the logical way in which algebraic geometry over complex numbers ought to have been born, had mathematics consisted solely of number-theory and algebra. That it came into being quite differently, and that it developed so far ahead of other branches of geometry, is a historical accident; it is indeed a fortunate one, having allowed free play to a tool which is invaluable as long as one is aware of its limitations; I need hardly tell you that I am referring to our spatial intuition.

We are now ready to consider in more specific terms the few known results in the "geometry over integers", which, following Kronecker, I have been trying to define; and for this we must turn first of all, naturally, to Kronecker himself. His great work on elliptic functions ([5b]), or rather its algebraic part (as distinct from the equally profound analytical theory), gives us a first example of an investigation of that kind; this consists in *the study of the equation* $Y^2 = 1 - \rho X^2 + X^4$ *over the ring* $Z[\rho]$, where ρ is an indeterminate, and is chiefly concerned with the transformation of elliptic functions. Jacobi's results on this subject are interpreted as defining, for every *odd* prime p, a correspondence between two generic points (x, y, ρ), (x', y', ρ') of the surface $Y^2 = 1 - \rho X^2 + X^4$, where $x' = x^p F(1/x)/F(x)$, $y' = G(x)/F(x)^2$, ρ' is algebraic over $Q(\rho)$, and F, G are polynomials with algebraic coefficients over $Q(\rho)$. Let σ be a root of $1 - \rho X^2 + X^4 = 0$, and σ' a root of $1 - \rho' X^2 + X^4 = 0$. Then Kronecker proves the following facts. The coefficients of F, G are in the field $Q(\sigma, \sigma')$; if divisibility relations are understood in the sense of integral algebraic elements over $Z[\rho]$, then σ' and all the roots of $G(X)$ are *units*; and $F(X)$ is of the form

$$F(X) = \pi X^{p-1} + \pi \sum_{i=1}^{p-2} \gamma_i X^{p-i-1} + 1,$$

where π *and the* γ_i *are integral over* $Z[\rho]$; furthermore, π is of degree $p + 1$ over $Q(\rho)$, *and has the norm* p over that field. The main results on complex multiplication, and its application to the class-field theory of imaginary quadratic fields, can be derived from these facts by specialization of the parameter ρ. It is very probable that a reconsideration of this splendid work from a modern point of view would not merely enrich our knowledge of elliptic function-fields, but

would also reveal principles of great importance for any further development of algebraic geometry over integers.

Now, coming back to the *Grundzüge*, take Kronecker's well-known and supposedly outmoded device for the introduction of ideals. This consists in associating with the elements $a_0, a_1, \cdots, a_m$ of a ring the linear expression $a_0 + \sum_{i=1}^{m} u_i a_i$, or, when the homogeneous notation happens to be more suitable, the linear form $\sum_{i=0}^{m} u_i a_i$, in the indeterminates u_i; thus the u_i are new variables adjoined to the ring, a feature which, in the eyes of orthodox Dedekindians, is a fatal blemish of this procedure. If, for instance, the a_i are in the ring $k[X, Y]/(F)$ determined by a plane curve $F(X, Y) = 0$ with coefficients in a field k, the ideal generated by them means substantially the same as the set of common zeros of the a_i, counted with their multiplicities; and this is again nothing else than the "fixed part" of the linear series cut by the variable linear variety $\sum_{i=0}^{m} u_i X_i = 0$ through the point 0, in the affine space of dimension $m + 1$, on the model of the given curve which is the locus of the point $(a_0, \cdots, a_m)$. If we translate this into the projective language, we find ourselves at the heart of the theory of linear series; and a slight extension of Kronecker's idea could lead us very naturally to such thoroughly "modern" topics as, for example, the associated form of a variety in projective space (the "Chow coordinates"). There is thus every reason to believe that the same idea will reacquire its full meaning in number-theory as soon as the interpenetration of number-theory and algebraic geometry, which Kronecker sought to realize, has been accomplished. Let us for instance try to define for number-fields a concept corresponding to the degree of a projective curve. If $f_0, f_1, \cdots, f_m$ are the coordinates of a generic point of a curve, and the u_i are indeterminates, the degree is the number of "variable" zeros of $\sum_i u_i f_i(x)$; this must be equal to the number of fixed poles minus the number of fixed zeros; in other words, if at every point P of the curve we put $n(P) = \inf_i \omega_P(f_i)$, where $\omega_P(f)$ indicates the order of f at P, then the degree of the curve is $d = -\sum_P n(P)$. If we replace the f_i by numbers ξ_i belonging to an extension k of degree n of the rational number-field Q and if ξ is the point $\xi = (\xi_0, \cdots, \xi_m)$ in the projective m-space, we are thus led to consider the number $H(\xi) = \prod_v \sup_i v(\xi_i)$, where the product is taken over all absolute values (p-adic or archimedean) of k; $H(\xi)$ does not change if the ξ_i are replaced by $\rho\xi_i$, with $\rho \in k$. This concept is essentially due to Siegel [9][1]; as D. G. Northcott indicates [8a], it is more convenient, for arithmetical purposes, to introduce the number $h(\xi) = H(\xi)^{1/n}$, which depends only upon the point ξ and not upon the field k. We shall call $h(\xi)$ the *height* of the point ξ. Following Kronecker, we may associate with the point ξ, with coordinates in k, the form $F(u) = r \cdot N_{k/Q}(\sum \xi_i u_i)$, where the u_i are indeterminates, and the rational number r is so chosen that the coefficients of $F(u)$ are rational integers without common divisor. Then we have $F(u) \ll (h(\xi) \cdot \sum u_i)^n$ (which means

[1] Cf. also H. Hasse, Monatshefte für Mathematik vol. 48 (1939) p. 205. Actually there is a slight discrepancy between Northcott's definition of $H(\xi)$ and that of Siegel and Hasse; we follow the latter.

that every coefficient of F is at most equal in absolute value to the corresponding one in the right-hand side): hence, if n_0 and h_0 are given, there is at most a finite number of points ξ for which $n \leqslant n_0$, $h(\xi) \leqslant h_0$. This is Northcott's lemma ([8a]; cf. [11h]), which is at the bottom of the application of the "infinite descent" to elliptic curves, and, more generally, to Abelian varieties over algebraic number-fields ([11a]; cf. [8b] and [11c]).

The height of a variable point on a curve or on a variety can best be studied by means of the theory of distributions; this is the only chapter of Kroneckerian geometry which has been developed beyond the rudiments. Let us first consider a curve C, defined over an algebraic number-field k; if it is rational, i.e., if its function-field is the field $k(t)$ generated over k by a single variable t, every function $f(t)$ on it can be written as $f(t) = \gamma\prod_i(t - a_i)^{m_i}$, where γ is a constant, the a_i are the poles and zeros of the function, and the integers m_i are their multiplicities (counted positively for a zero, negatively for a pole). If the curve is not rational, such a representation is not possible, except in a merely symbolical manner, or else by means of transcendental multi-valued functions which cannot be used for arithmetical purposes. Let us, however, consider for a moment a definite embedding of k in the field of complex numbers, so that C is defined over that field; and consider merely absolute values. Then one can attach to each point A of C a continuous real-valued function $d_A(M)$ on C, with $0 \leqslant d_A(M) \leqslant 1$, which is 0 when $M = A$ and only then, in such a way that if a function $f(M)$ belonging to the function-field of C (over k or even over the field of complex numbers) has the zeros and poles A_i with the multiplicities m_i, then

$$f(M) = \gamma(M) \prod_i d_{A_i}(M)^{m_i},$$

where $\gamma(M)$ is an *inessential* factor in the sense that there are constants γ_1, γ_2, both > 0, such that $\gamma_1 \leqslant \gamma(M) \leqslant \gamma_2$ for all M. This can easily be verified by elementary topological methods. It can also be proved by an algebraic argument, which remains valid if the field of complex numbers is replaced by the algebraic closure of the p-adic field, and also if the curve C is replaced by a variety. Reduced to its essential features, this argument can be described as follows. If V is a variety in an affine space, defined over the complex number-field or over a p-adic field, and if it does not contain the origin, then there is a polynomial $P(X_1, \cdots, X_n)$, vanishing on V and not at 0, with coefficients in the ground-field; this means that all points of V must satisfy an equation

$$1 = \sum a_{\nu_1 \dots \nu_n} X_1^{\nu_1} \cdots X_n^{\nu_n},$$

where all terms in the right-hand side are of degree $\geqslant 1$; therefore, if $(x_1, \cdots, x_n)$ is such a point, $\sup_i |x_i|$ cannot be arbitrarily small, and precisely it must be $\geqslant 1$ or $\geqslant (\sum |a_{\nu_1 \dots \nu_n}|)^{-1}$; here $|\ |$ denotes of course the ordinary or the p-adic absolute value, as the case may be.

So far we have considered only one absolute value, ordinary or p-adic, at a time, and so we have obtained, in this sense, merely "local" results; global re-

sults come from the consideration of all absolute values simultaneously; or else, what amounts to the same thing, one can treat the archimedean absolute values separately, in the manner indicated above, and then deal simultaneously with all the others. This is done by remarking that, if a variety V is defined over an algebraic number-field k, and does not go through 0, its points must, as above, satisfy an equation

$$a_0 = \sum a_{\nu_1 \cdots \nu_n} X_1^{\nu_1} \cdots X_n^{\nu_n}$$

whose coefficients are algebraic integers in k; and then if $(x_1, \cdots, x_n)$ is a point on V, with algebraic coordinates, the G.C.D. of the numerators of the fractional principal ideals $(x_1), \cdots, (x_n)$ must divide the principal ideal (a_0) of k. Out of this very simple fact one derives all the known results of the theory of distributions, one of whose main results is the following "theorem of decomposition":

Let C be a curve, defined over an algebraic number-field k. One can attach to each algebraic point A on C a function $\mathfrak{a}_A(M)$, defined at all algebraic points M of C, whose value at M is an ideal of the algebraic number-field $k(A, M)$, so that the following properties hold: $\mathfrak{a}_A(M)$ is 0 when $M = A$ and only then; and whenever f is a function on C, having the zeros and poles A_i with the multiplicities m_i, then the principal fractional ideal $(f(M))$ has the expression

$$(f(M)) = \mathfrak{c}(M) \prod \mathfrak{a}_{A_i}(M)^{m_i},$$

where $\mathfrak{c}(M)$ is an inessential fractional ideal in the sense that both $\mathfrak{c}(M)$ and $\mathfrak{c}(M)^{-1}$ divide a fixed natural integer. Furthermore, exactly the same result holds for every nonsingular projective variety V of any dimension r, except that, of course, the ideal-valued functions $\mathfrak{a}(M)$ are then attached, not to the points of V, but to the subvarieties of V of dimension $r - 1$.

As we have said above, this becomes a truly global result if we combine it with the corresponding result over complex numbers. When this is done, one finds inequalities for the height of a variable point on a projective variety, which is found to depend essentially only upon the class of the divisors in the linear series determined on the variety by its hyperplane sections. In particular, let C be a curve of degree d in a projective space; let C' be a curve, birationally equivalent to C, of degree d', in the same or in another projective space; let M, M' be corresponding points on C, C', with algebraic coordinates; then, to every ϵ, there are constants γ_1, γ_2, both > 0, such that

$$\gamma_1 h(M)^{1/d-\epsilon} \leqslant h(M')^{1/d'} \leqslant \gamma_2 h(M)^{1/d+\epsilon}$$

for all pairs of corresponding points M, M' on C; in this sense, the "order of magnitude" of $h(M)^{1/d}$ is independent of the projective model chosen for C. This is the decisive inequality for Siegel's proof of the fact that a nonrational curve can have at most a finite number of points with integral coordinates in a given algebraic number-field ([9]; cf. [11h]). The same approach also leads very simply to Northcott's inequalities ([8]; cf. [11h]); these contain as special cases

the inequalities by which it was first proved that the points on an Abelian variety, with coordinates in a given algebraic number-field, form a finitely generated group ([11a]; cf. [11c]), so that a thoroughly "modernized" version of that proof could now be given.

I should like to conclude with a brief discussion of a very interesting conjecture, due, I believe, to Hasse. As we have said, from the Kroneckerian point of view the fields of dimension 1 are the number-fields and the function-fields of curves over finite fields; to each one of these there belongs a zeta-function, the properties of which may be said to epitomize in analytic garb some of the more important properties of the field. It is therefore reasonable to guess that similar functions can be attached to fields of higher dimension, and in the first place to the fields of dimension 2, i.e., to the curves over an algebraic number-field, and to the surfaces over a finite field. Consider the latter problem first: let S be a surface over the finite field k of q elements; and define N_ν, for each ν, as the number of points on the surface with coordinates in the extension k_ν of degree ν of the ground-field; the analogy with curves, as well as the consideration of some special cases, makes it very natural ([11g]) to introduce the function $Z(q^{-s})$, where $Z(U)$ is defined by $Z(0) = 1$, $d \log Z(U)/dU = \sum_1^\infty N_\nu U^{\nu-1}$, and to expect that this will have the essential properties of a zeta-function over a finite field; i.e., that it is a rational function of U, that it satisfies a functional equation, and that it satisfies a suitably modified Riemann hypothesis; even the first property seems exceedingly difficult to prove at present, except in special cases. Now, suppose that we have on S a family of curves $C(t)$ depending upon a parameter t; for simplicity assume that $C(t)$ depends rationally upon t, and that no two curves $C(t)$ have a point in common. If we give to t a value which is algebraic over k, $C(t)$ will be defined over $k(t)$, and a zeta-function will be attached to it, defined in a manner similar to that employed for S. As the number of points on S with coordinates in k_ν is obviously the sum of the same numbers for all the curves $C(t)$, it follows at once that $Z(U)$ is the product of the zeta-functions attached to the curves $C(t)$, provided that we take only one representative for each set of curves conjugate to each other over k. Now this definition may at once be transported to number-fields: if C is a curve over the algebraic number-field K, given by an equation $F(X, Y) = 0$, then, for almost all prime ideals $\mathfrak{p}$ of K, the equation $F = 0$, reduced modulo $\mathfrak{p}$, will define a curve of the same genus as C over the finite field of $q = N(\mathfrak{p})$ elements; this has a zeta-function; and we are thus led to consider *the product of these zeta-functions for all* $\mathfrak{p}$, which is precisely the function previously defined by Hasse, of which he conjectured that it can be continued analytically over the whole plane, that it is meromorphic, and that it satisfies a functional equation. In a few simple cases, this function can actually be computed; e.g., for the curve $Y^2 = X^3 - 1$ it can be expressed in terms of Hecke's L-functions for the field $k(\sqrt[3]{1})$; this example also shows that such functions have infinitely many poles, which is a clear indication of the very considerable difficulties that one may expect in their study.

REFERENCES

1. C. CHABAUTY, *Sur les équations diophantiennes liées aux unités d'un corps de nombres algébriques fini*, Annali di Matematica (IV) vol. 17 (1938) p. 127.

2. C. CHEVALLEY, *On the theory of local rings*, Ann. of Math. vol. 44 (1943) p. 690.

3. W. L. CHOW, *Algebraic systems of positive cycles in an algebraic variety*, Amer. J. Math. vol. 72 (1950) p. 247.

4. I. S. COHEN, *On the structure and ideal theory of complete local rings*, Trans. Amer. Math. Soc. vol. 59 (1946) p. 54.

5. L. KRONECKER, a) *Grundzüge einer arithmetischen Theorie der algebraischen Grössen*, Werke, vol. II, Leipzig, Teubner, 1897, pp. 237–387.

b) *Zur Theorie der elliptischen Funktionen*, Werke, vol. IV, Leipzig-Berlin, Teubner, 1929, pp. 345–495.

6. W. KRULL, *Dimensionstheorie in Stellenringen*, J. Reine Angew. Math. vol. 179 (1938) p. 204.

7. E. LUTZ, *Sur l'équation* $y^2 = x^3 - Ax - B$ *dans les corps* $\mathfrak{p}$*-adiques*, J. Reine Angew. Math. vol. 177 (1937) p. 238.

8. D. G. NORTHCOTT, a) *An inequality in the theory of arithmetic on algebraic varieties*, Proc. Cambridge Philos. Soc. vol. 45 (1949) p. 502.

b) *A further inequality in the theory of arithmetic on algebraic varieties*, ibid. p. 510.

9. C. L. SIEGEL, *Über einige Anwendungen diophantischer Approximationen*, Abhandlungen der Preussischen Akademie der Wissenschaften (Jahrgang 1929, no. 1), Berlin, 1930.

10. TH. SKOLEM, *Einige Sätze über* $\mathfrak{p}$*-adische Potenzreihen mit Anwendung auf gewisse exponentielle Gleichungen*, Math. Ann. vol. 111 (1935) p. 399.

11. A. WEIL, a) *L'arithmétique sur les courbes algébriques*, Acta Math. vol. 52 (1928) p. 281 (Thèse, Paris 1928).

b) *Arithmétique et géométrie sur les variétés algébriques*, Actualités Scientifiques et Industrielles, no. 206 (Exp. math. à la mém. de J. Herbrand XI), Paris, Hermann et Cie, 1935.

c) *Arithmetic of algebraic varieties* (Russian), Uspehi Matematičeskih Nauk, vol. 3 (1937) p. 101.

d) *Sur les fonctions algébriques à corps de constantes fini*, C. R. Acad. Sci. Paris vol. 210 (1940) p. 592.

e) *Sur les courbes algébriques et les variétés qui s'en déduisent*, Actualités Scientifiques et Industrielles, no. 1041 (Publ. Inst. Math. Strasb. VII), Paris, Hermann et Cie, 1948.

f) *Variétés abéliennes et courbes algébriques*, Actualités Scientifiques et Industrielles, no. 1064 (Publ. Inst. Math. Strasb. VIII), Paris, Hermann et Cie, 1948.

g) *Numbers of solutions of equations in finite fields*, Bull. Amer. Math. Soc. vol. 55 (1949) p. 497.

h) *Arithmetic on algebraic varieties*, Ann. of Math. vol. 53 (1951) pp. 412–444.

12. O. ZARISKI, *Theory and applications of holomorphic functions on algebraic varieties*, Memoirs of Amer. Math. Soc., no. 5 (1951).

UNIVERSITY OF CHICAGO,
CHICAGO, ILL., U. S. A.

INTERNATIONAL CONGRESS

OF

MATHEMATICIANS

Cambridge, Massachusetts, U. S. A.
1950

CONFERENCE IN ANALYSIS

Committee

Marston Morse (*Chairman*)

L. V. Ahlfors
Salomon Bochner
G. C. Evans
Einar Hille

The Committee on the Conference in Analysis consisted of Marston Morse (Chairman), L. V. Ahlfors, Salomon Bochner, G. C. Evans, and Einar Hille, with Marshall Stone originally a member. The directive from the Program Committee requested emphasis on the relations of algebra and topology to analysis, and specifically recommended three subdivisions of the Conference eventually formulated as: I, Algebraic Tendencies in Analysis; II, Analysis and Geometry in the Large; and III, Extremal Methods and Geometric Theory of Functions of a Complex Variable. Marshall Stone, Marston Morse, and Lars Ahlfors were assigned the task of organizing the corresponding subconferences. Of the twelve hours made available for the Conference, six were assigned to Subconference I, three each to II and III. The three hours assigned to II was subsequently enlarged to four hours.

Marshall Stone originated the idea of conducting Subconference I in the form of panels. On his resignation, Einar Hille continued with this interesting but somewhat difficult task, aided by von Neumann and Hildebrandt. Subconference II was broken down into two two-hour subconferences on Analysis in the Large, and Analysis and Geometry in the Large. Bochner was associated with Morse in selecting the speakers for these subconferences, with advice from H. Whitney as to the second.

The first conceptions of the Conference on Analysis included the use of the eminent Russian analysists Gelfand, Kolmogoroff, Vinogradow, Lusternik, S. Bernstein and others, either in the Conference or as invited to give stated addresses. Unfortunately these mathematicians could not come to the Congress. The stated addresses of Morse and Bochner separately invited by the Organizing Committee were regarded as associated with the subconferences derived from II.

Further comments on the nature of the subconferences will precede the program of the subconferences.

Attention is called to the publication of the panel reports in the Mathematical Surveys of the American Mathematical Society, New York.

MARSTON MORSE

ALGEBRAIC TENDENCIES IN ANALYSIS

The task of organizing Subconference I on Analysis was delegated to Marshall H. Stone who proposed a symposium on the algebraic tendencies in recent research in analysis, to consist of a series of reports, each to be prepared by a committee or panel of experts and to be presented by a spokesman for the panel. This proposal was accepted by the Organizing Committee and six panels were formed. Four of these were ready in October 1949, the others got under way by the end of the year. Stone resigned in November from the chairmanship of the subconference and from the panel on spectral theory. The direction of the subconference was then taken over by the present writer, ably assisted by T. H. Hildebrandt and J. von Neumann.

The composition of the panels was made as international as conveniently possible, due account being taken of the nature of the subject matter and the need of effective collaboration. Invitations sent to four Russian mathematicians through proper channels to join the panels led to no response, however. The membership of the panels at the time of the Congress was as follows, spokesmen being indicated by asterisks.

Panel on Group Representations: H. CARTAN, A. M. GLEASON, R. GODEMENT*, G. W. MACKEY, F. I. MAUTNER, L. SCHWARTZ.

Panel on Topological Algebra: J. DIEUDONNÉ, I. KAPLANSKY*, I. E. SEGAL.

Panel on Measure Theory: J. DIEUDONNÉ, P. R. HALMOS*, J. C. OXTOBY, D. MAHARAM STONE, S. M. ULAM.

Panel on Spectral Theory: W. AMBROSE, J. DIXMIER, N. DUNFORD*, F. J. MURRAY, J. VON NEUMANN, F. RELLICH, B. DE SZ. NAGY, K. YOSIDA.

Panel on Applied Functional Analysis: N. ARONSZAJN*, S. BERGMAN, J. W. CALKIN, K. FRIEDRICHS, K. KODAIRA, A. WEINSTEIN.

Panel on Ergodic Theory: N. DUNFORD, W. EBERLEIN, G. A. HEDLUND, E. HILLE, S. KAKUTANI*, J. C. OXTOBY.

The reports to be prepared by the panels were intended to give a thorough survey of the state of knowledge in the field under consideration with adequate bibliography, emphasis being placed on the interplay of algebra and analysis typical for the field. A preliminary survey showed the task to be a staggering one. From the beginning it was expected that each spokesman might select from the report of the panel those items which he wished to emphasize in a necessarily brief oral presentation before the Congress. It was indeed obvious that nothing like a complete exposition of the fields covered by the panel could be expected then. By the end of 1949 it also became clear that publishing detailed reports with extensive bibliographies would call for much more space than was available in the Proceedings of the Congress. Thus it became necessary to find another outlet for the reports and, on the initiative of the Organizing Committee, the Board of Editors of the Mathematical Surveys, published by the American Mathematical Society, agreed to issue in their series such reports as

would be finished within reasonable time. It is hoped that all reports will appear in the Surveys in the near future.

The spokesmen were invited to address the Congress on the subject matter of their reports so that their lectures naturally find a place in these Proceedings. The following six papers are addresses or abstracts of such presented by the spokesmen to the Congress. They are not the final reports of the panels which are still to come. As a matter of fact, in several cases no reports were available and for this reason, or in order to avoid duplication, the spokesmen chose to speak as individuals, discussing some special feature of the field, regardless of how it would ultimately fit into the final report. There were also cases where reports had taken definite form and the spokesman could really speak for his panel. It will be obvious from the context to which category the individual addresses belong.

The Committee hopes that these addresses will stimulate interest in the forthcoming reports. Though the latter are not intended to cover functional analysis as a whole, they may well prove to be indispensible to workers in this field during the next decade.

EINAR HILLE

SOME UNSOLVED PROBLEMS IN THE THEORY OF GROUP REPRESENTATIONS

ROGER GODEMENT

There are still many unsolved problems in the theory of locally compact groups, and we intend to state some of them, without any claim for completeness, of course. Some of the problems are related to already solved questions, and have only a purely methodological interest; the explanation for this lies in the experimental fact that it is always useful to find more and more general methods.

(I) It has been proved by K. Iwasawa [18], with the help of the theory of semi-simple Lie groups, that the component of the identity in the group of automorphisms of a compact group contains only inner automorphisms. Can a direct proof for this theorem be constructed?

(II) Is it possible to determine all irreducible representations of the "classical" compact groups [32] by purely integral methods? In particular, is it possible to compute explicitly the characters of these groups by using only their functional equations, and to deduce the representations of these groups from the knowledge of corresponding characters?

(III) Let $s \to U_s$ be a unitary representation of a locally compact group $G = \{s\}$ in a Hilbert space $\mathfrak{H}$; such a representation is called a factor representation if the weakly closed ring of operators generated by $\{U_s \mid s \in G\}$ is a factor in the sense of F. J. Murray and J. von Neumann [25] (i.e., if the center of the ring is one-dimensional). (Factors have been classified by F. J. Murray and J. von Neumann [25] into the classes (I_n), $n = 1, 2, \cdots$, (I_∞), (II_1), (II_∞), and (III_∞). Factor representations of all types exist if we consider discrete groups. Factor representations of connected groups and Lie groups were recently discussed by F. J. Mautner [24] and I. E. Segal and J. von Neumann [30].) The question: Do there exist connected groups which admit factor representation of type (III_∞)?

(IV) Let $f(s)$ be a positive definite continuous function defined on a topological group $G = \{s\}$. Let us denote by $\{\mathfrak{H}, U_s\}$ the unitary representation of G defined by means of $f(s)$ (cf. R. Godement [11]), and by R the weakly closed ring of operators in $\mathfrak{H}$ generated by the operators $\{U_s \mid s \in G\}$. If one performs in a suitable way the decomposition of R with respect to its center, then one gets a corresponding decomposition of the representation $\{\mathfrak{H}, U_s\}$ into a continuous sum of factor representations $\{\mathfrak{H}(t), U_s(t)\}$:

$$\{\mathfrak{H}, U_s\} = \int \{\mathfrak{H}(t), U_s(t)\}\, d\mu(t),$$

where t runs through a certain locally compact space with a positive measure μ (cf. F. I. Mautner [23], [24], J. von Neumann [26], R. Godement [15]). Is it

possible to find in almost every space $\mathfrak{H}(t)$ a (possibly bounded) Hermitian operator $F(t)$ in such a way that the following is true:

$$f(s) = \int \mathrm{Tr}_t(U_s(t)F(t))\, d\mu(t)$$

where Tr_t denotes some relative trace in the factor generated by the operators $\{U_s(t) \mid s \in G\}$? In case the answer to this question is yes, is this decomposition of $f(s)$ essentially unique? (Of course, the problem (IV) depends on the problem (III).)

(V) Is it possible to find a simple characterization for the locally compact groups all the factor representations of which are of type (I_n), $n = 1, 2, \cdots$, or (I_∞)? Is there any relation between this problem and the decomposition of the Haar measure into ergodic parts with respect to the group of inner automorphisms?

(VI) Can the classical theory of characters be extended to general groups? The first task is to find a definition of characters which applies to every case (the definition proposed by myself [12], [13], [14], is not general enough), and then to solve the following problems: Are the characters defined by some extremal properties? Do the characters possess some simple functional properties? Are they in correspondence with irreducible two-sided representations (in the sense of R. Godement [12] and I. E. Segal [28])? Can every positive definite measure which is invariant under inner automorphisms be decomposed into some kind of continuous sum of characters? The last problem includes a generalization of the Plancherel theorem to noncommutative and noncompact groups (cf. R. Godement [9], [10] and I. E. Segal [29]).

(VII) In many special cases it is observed that the Plancherel theorem can be proved as follows: One decomposes the left regular representation into a continuous sum of irreducible representations; then one brings together the components which belong to a given class of irreducible representations. Can this process be extended to arbitrary groups?

(VIII) Let E be a locally compact space and G a locally compact group which acts on E. Let μ be a positive measure on E such that the family of all μ-null sets of E is carried onto itself by the transformations of G. Then one can define in an obvious way a unitary representation of G in the corresponding L^2-space on E. The problem is to find a condition under which such a representation is irreducible and to find conditions under which two representations of that kind are similar. Some progress in this direction was recently made by G. W. Mackey [20], [21], [22]. Similar problems occur for representations which are defined by the method of positive definite kernels (as for instance in the case of supplementary series in the Lorenz group). (Cf. I. Gelfand and M. Neumark [5], [6].)

(IX) To extend to general groups the duality theory already known for locally compact abelian groups (L. Pontrjagin [27]) and for compact groups (T. Tannaka [31]).

(X) A theory of "spherical harmonics in symmetric Riemann spaces" was recently published by I. Gelfand [3]; but Gelfand's results are related only to functions which are invariant under the group of rotations around a fixed point of the space. Can these results be extended to systems of spherical functions, associated with the irreducible representations of that group? It is quite important to solve these problems in a general way. But it is also equally important to investigate the situation in detail in every special case and to get precise results about Bessel functions, hypergeometric functions, and so on. One may believe that this part of the problem is not the easiest one.

(XI) To study unitary representations of real semi-simple Lie groups. I. Gelfand and M. Neumark [4], [5], [6], [7], [8] obtained many results in the complex case. But the case of real unimodular groups was discussed by V. Bargmann [1] only in the case of two variables so that there is still much to be done.

(XII) Let us denote by G a connected semi-simple Lie group, by $\mathfrak{G}$ its Lie algebra, by A the universal envelopping associative algebra of $\mathfrak{G}$. From a well-known result of L. Gårding [2], it follows that to every unitary representation of G in a Hilbert space $\mathfrak{H}$, there corresponds a linear representation of A in an everywhere dense subspace of $\mathfrak{H}$. The problem is to find relations between irreducible representations (or more generally factor representations) of G and the corresponding linear representations of A. Important results were obtained recently in this direction by Harish-Chandra [16], [17]. But there are still unsolved problems. For instance: when two unitary representations of G give rise to two algebraically equivalent representations of A, are these two representations unitary equivalent? (This problem was solved by Harish-Chandra in the meantime.)

(XIII) We use the notations of problem (XII). One can consider in an obvious way the elements of the algebra A as being differential operators on the manifold G. The center Z of A then becomes a commutative algebra of operators which are both left and right invariant. Let us say that a function—or more precisely, a distribution in the sense of L. Schwartz—is a character of G if (i) it is invariant under inner automorphisms of G, (ii) it is an eigenfunction for each element of Z. What are the properties of characters thus defined? Are they associated with (not necessarily) irreducible representations of G? Could the characters be used in order to construct some kind of Laplace transforms? (Cf. G. W. Mackey [19].)

(XIV) In many cases the determination of all irreducible representations of a semi-simple Lie group can be performed by two different methods: the infinitesimal method and the integral method. So far, it seems that the integral method is more powerful since it leads to explicit formulas for decomposing reducible representations into irreducible ones. Could the same results be obtained by the infinitesimal method?

(XV) Let $G = \{s\}$ be a Fuchsian group in the unit circle of the complex z-plane. Can the theory of zeta-Fuchsian functions of Poincaré be extended to

infinite-dimensional representations of G? In particular, let $s \to U_s$ be an irreducible representation of G in a Hilbert space $\mathfrak{H}$. Is it possible to realize $\mathfrak{H}$ by means of some vector space of meromorphic functions in the unit circle in such a way that (i) the operators U_s are given by a formula like

$$U_s f(z) = \left[\frac{d(s^{-1}z)}{dz}\right]^{\gamma} f(s^{-1}z)$$

(ii) the scalar product of any two elements of $\mathfrak{H}$ can be computed in a simple way (this is obviously the more important part of the problem)?

(XVI) Let $G = \{s\}$ be a Fuchsian group in the upper half z-plane. Let $\mathfrak{H}$ be the Hilbert space of all analytic functions $f(z)$ defined on the upper half z-plane such that

$$\iint_{\Im(z)>0} |f(z)|^2 \, dx\, dy < \infty, \qquad z = x + iy.$$

One defines a unitary representation $s \to U_s$ of G in $\mathfrak{H}$ by

$$U_s f(z) = \left[\frac{d(s^{-1}z)}{dz}\right] f(s^{-1}z).$$

The problem is to find a decomposition of this representation into irreducible parts. It is easy to suggest many other problems analogous to (XV) and (XVI).

(XVII) Let $G = \{s\}$ be a locally compact group, n a positive integer, $\hat{G}_n =$ the set of equivalence classes of irreducible n-dimensional unitary representations of G. Since there exists a one-to-one correspondence between members ϑ of G_n and their characters $\chi_\vartheta(s)$, one may introduce a topology in $\hat{G}_n$ by saying that a sequence of classes $\{\vartheta_p \mid p = 1, 2, \cdots\} \subset \hat{G}_n$ converges to a class $\vartheta_0 \in \hat{G}_n$ if the sequence of characters $\{\chi_{\vartheta_p}(s) \mid p = 1, 2, \cdots\}$ converges to the character $\chi_{\vartheta_0}(s)$ uniformly on every compact subset of G. It turns out that $\hat{G}_n$ becomes a locally compact space with respect to this topology. Have these topological spaces interesting properties from the point of view of algebraic geometry? In this argument, it is of course necessary to assume that G possesses "sufficiently many" finite-dimensional unitary representations. (This is true, for example, if G is a finitely generated discrete group.)

(XVIII) Let $G = \{s\}$ be a compact Lie group. Let us denote by $s \to U_s$ the adjoint representation of G. Then it is known that

$$P(z) = \int_G \det(U_s + z \cdot 1) ds$$

is the Poincaré polynomial of G. Is it possible to compute this integral explicitly by means of analytical methods? This problem is stated by H. Weyl in [32].

(XIX) To find new applications of the theory of group representations to other parts of mathematics.

Bibliography

1. V. Bargmann, *Irreducible unitary representations of the Lorentz group*, Ann. of Math. vol. 48 (1947) pp. 568–640.

2. L. Gårding, *Note on continuous representations of Lie groups*, Proc. Nat. Acad. Sci. U. S. A. vol. 33 (1947) pp. 331–332.

3. I. Gelfand, *Spherical functions in symmetric Riemann spaces*, Doklady Akad. Nauk SSSR vol. 70 (1950) pp. 5–8.

4. I. Gelfand and M. Neumark, *The principal series of irreducible representations of of the complex unimodular group*, C. R. (Doklady) Acad. Sci. URSS vol. 56 (1947) pp 3–4.

5. ———, *Supplementary and degenerate series of representations of the complex unimodular group*, C. R. (Doklady) Acad. Sci. URSS vol. 58 (1947) pp. 1577–1580.

6. ———, *Unitary representations of the Lorentz group*, Bull. Acad. Sci. URSS. Sér. Math. vol. 11 (1947) pp. 411–504.

7. ———, *Unitary representations of semi-simple Lie groups*, I, Rec. Math. (Mat. Sbornik) N. S. vol. 21 (1947) pp. 405–434.

8. ———, *On the connection between the representations of a complex simple group and those of its maximal compact subgroups*, Doklady Akad. Nauk. SSSR vol. 63 (1948) pp. 225–228.

9. R. Godement, *Analyse harmonique dans les groupes centraux* I. *Fonctions centrales et caractères*, C. R. Acad. Sci. Paris vol. 225 (1947) pp. 19–21.

10. ———, *Analyse harmonique dans les groupes centraux* II. *Formule d'inversion de Fourier*, C. R. Acad. Sci. Paris vol. 225 (1947) pp. 221–223.

11. ———, *Les fonctions de type positif et la théorie des groupes*, Trans. Amer. Math. Soc. vo. 63 (1948) pp. 1–84.

12. ———, *Sur la théorie des caractères* I. *Définition et classification des caractères*, C.R. Acad. Sci. Paris vol. 229 (1949) pp. 967–969.

13. ———, *Sur la théorie des caractères* II. *Mesures et groupes de classe finie*, C.R. Acad. Sci. Paris vol. 229 (1949) pp. 1050–1051.

14. ———, *Sur la théorie des caractères* III. *Un exemple de mesure-caractère de classe* (I), C.R. Acad. Sci. Paris vol. 229 (1949) pp. 1107–1109.

15. ———, *Sur la théorie des representations unitaries*, Ann. of Math. (2) vol. 53 (1951) pp. 68–124.

16. Harish-Chandra, *Faithful representation of Lie algebras*, Ann. of Math. vol. 50 (1949) pp. 68–76.

17. ———, *On representations of Lie algebras*, Ann. of Math. vol. 50 (1949) pp. 900–915.

18. K. Iwasawa, *On some types of topological groups*, Ann. of Math. (2) vol. 50 (1949) pp. 507–558.

19. G. W. Mackey, *The Laplace transforms for locally compact abelian groups*, Proc. Nat. Acad. Sci. U. S. A. vol. 34 (1948) pp. 156–162.

20. ———, *Imprimitivity for representations of locally compact groups*, Proc. Nat. Acad. Sci. U. S. A. vol. 35 (1949) pp. 537–545.

21. ———, *Imprimitivité pour les representations des groupes localement compacts*, C.R. Acad. Sci. Paris vol. 230 (1950) pp. 808–809.

22. ———, *Imprimitivité pour representations des groupes localement compacts* III. *Produits de Kronecker et nombres d'entracement forts*, C.R. Acad. Sci. Paris vol. 230 (1950) pp. 908–909.

23. F. I. Mautner, *Unitary representations of locally compact groups* I, Ann. of Math. (2) vol. 51 (1950) pp. 1–25.

24. ———, *Unitary representations of locally compact groups* II, Ann. of Math. vol. 52 (1950) pp. 528–556.

25. F. J. Murray and J. von Neumann, *On rings of operators*, Ann. of Math. (2) vol. 37 (1936) pp. 116–229.

26. J. VON NEUMANN, *On rings of operators. Reduction theory*, Ann. of Math. (2) vol. 50 (1949) pp. 401–485.

27. L. PONTRJAGIN, *The theory of topological commutative groups*, Ann. of Math. (2) vol. 35 (1934) pp. 361–388.

28. I. E. SEGAL, *The two-sided regular representation of a unimodular locally compact group*, Ann. of Math. (2) vol. 51 (1950) pp. 293–298.

29. ———, *An extension of Plancherel's formula to separable unimodular groups*, Ann. of Math. (2) vol. 52 (1950) pp. 272–292.

30. I. E. SEGAL and J. VON NEUMANN, *A theorem on unitary representations of semi-simple Lie groups*, Ann. of Math. (2) vol. 52 (1950) pp. 509–517.

31. T. TANNAKA, *Dualität der nicht-kommutativen Gruppen*, Tohoku Math. J. vol. 53 (1938) pp. 1–12.

32. H. WEYL, *Classical groups*, Princeton, Princeton University Press, 1939.

UNIVERSITY OF NANCY,
NANCY, FRANCE.

TOPOLOGICAL ALGEBRA[1]

I. KAPLANSKY

Topological algebra made its first appearance in the paper of Kürschak [9], where the definition of an abstract field with a valuation is clearly set forth. The foundation was completed in the thesis of van Dantzig [3]; topological groups, rings, fields, and linear spaces are there defined, and their basic properties are established.

In 1933 an important advance occurred when Haar [7] demonstrated the existence of an invariant measure in any locally compact group. Since then, our understanding of the structure of locally compact groups has rapidly grown. Most notably, Hilbert's fifth problem has been solved for compact groups and for solvable groups, and the general case appears to be within reach.

With the appearance of Banach's book [1] in 1932, the work of the Polish school on functional spaces reached its climax. With the perspective of two more decades, it seems to be fair to say that the expectation that Banach spaces would be a definitive unifying concept has been partly disappointed. Banach spaces are at once too general and too special; they are too general in that our detailed knowledge of Hilbert space shows no prospect of being extended to Banach spaces. They are too special in that they do not cover important instances of topological linear spaces that are met in applications.

Somewhat crudely, we may describe the present situation in the theory of topological linear spaces as follows. Banach's results fall into two types. In the first type the arguments are of a highly algebraic nature; here a great clarification was achieved by Mackey's program [10] of studing pairs of dual linear spaces in a purely algebraic way. In the second type of theorem, an important part of the proof rests on a topological device, or above all on a category argument. Here considerable progress has been made recently by Dieudonné and Schwartz [2]; they investigate topological linear spaces admitting a complete metric and direct limits of such spaces. The results have applications in Schwartz's theory of distributions.

In the present decade, much work has been devoted to Banach algebras (normed rings) since the publication of Gelfand's paper [5]. Gelfand attracted attention by his observation that a useful theorem due to Wiener could be regarded as a statement about inverses and maximal ideals in a suitable Banach algebra. Segal [12] and independently Godement [6] have pushed this line of investigation further.

In the study of Banach algebras themselves, there are at present two directions in which work is being done. One may attempt to probe more deeply

[1] This is a brief summary of the oral report delivered by the spokesman, who bears sole responsibility for it. A full account, prepared jointly by all three members of the panel, will appear elsewhere.

into the structure of a fairly general commutative Banach algebra; Silov, in a series of papers of which [14] is typical, has made some noteworthy contributions to this program. In the noncommutative case, study is at present largely confined to self-adjoint uniformly closed algebras of operators on Hilbert space (called C*-algebras by Segal in [13]). A commutative C*-algebra (say with unit) is completely known: it admits a faithful representation as the algebra of all continuous complex functions on a compact Hausdorff space. In a mild generalization of the commutative case, Kaplansky [8] finds that a suitably weakened version of this theorem is still valid.

Concerning the structure of the most general C*-algebra, there is as yet little that can be said as long as we merely assume uniform closure. If however we strengthen this hypothesis to weak closure, then a fairly definitive structure theory becomes available; it was given by Murray and von Neumann in a series of five memoirs, beginning with [11]. The theory of weakly closed algebras has again been taken up recently by Dixmier [4] and others, and there is every indication that our knowledge of them will be considerably increased.

Bibliography

1. S. Banach, *Théorie des opérations linéaires*, Warsaw, 1932.

2. J. Dieudonné and L. Schwartz, *La dualité dans les espaces (F) et (LF)*, Annales de l'Institut Fourier vol. 1 (1949) pp. 61–101.

3. D. Van Dantzig, *Studiën über topologische Algebra*, Dissertation, Amsterdam, 1931.

4. J. Dixmier, *Les anneaux d'opérateurs de classe finie*, Ann. École Norm. vol. 66 (1949) pp. 209–261.

5. I. Gelfand, *Normierte Ringe*, Rec. Math. (Mat. Sbornik) N.S. vol. 9 (1941) pp. 3–24.

6. R. Godement, *Théorèmes taubériens et théorie spectrale*, Ann. École Norm. vol. 63 (1948) pp. 119–138.

7. A. Haar, *Der Massbegriff in der Theorie der kontinuierlichen Gruppen*, Ann. of Math. vol. 34 (1933) pp. 147–169.

8. I. Kaplansky, *The structure of certain operator algebras*, Trans. Amer. Math. Soc. vol. 70 (1951) pp. 219–255.

9. J. Kürschák, *Über Limesbildung und allgemeine Körpertheorie*, J. Reine Angew. Math. vol. 142 (1913) pp. 211–253.

10. G. W. Mackey, *On infinite-dimensional linear spaces*, Trans. Amer. Math. Soc. vol. 57 (1945) pp. 155–207.

11. F. J. Murray and J. von Neumann, *On rings of operators*, Ann. of Math. vol. 37 (1936) pp. 116–229.

12. I. E. Segal, *The group algebra of a locally compact group*, Trans. Amer. Math. Soc. vol. 61 (1947) pp. 69–105.

13. ———, *Irreducible representations of operator algebras*, Bull. Amer. Math. Soc. vol. 53 (1947) pp. 73–88.

14. G. Silov, *On regular normed rings*, Travaux de l'Institut Mathématique, Stekloff, Moscow, 1947.

University of Chicago,
Chicago, Ill., U. S. A.

MEASURE THEORY

PAUL R. HALMOS

The purpose of this report (prepared in collaboration with J. Dieudonné, D. Maharam, and J. C. Oxtoby) is to describe some of the recent developments in measure theory. The report falls into three sections concerned with problems of existence, differentiability, and decomposability, respectively.

The existence problems treated are those discussed by Banach, Kuratowski, and Ulam (nonexistence of well-behaved measures defined on the class of all subsets of a set), Kolmogoroff, Doob, and Jessen (extension of measures to product spaces), Carathéodory, Oxtoby, and Ulam (characterizations of Lebesgue measure), Markoff, Alexandroff, and Nikodým (measures in topological spaces), Haar, Cartan, and Loomis (measures in groups and uniform structures), and Szpilrajn, Kakutani, and Kodaira (extensions of Lebesgue measure).

The fundamental result on differentiability is the Radon-Nikodým theorem. The important problems here concern the exact domain of applicability of the theorem. Conditions for its validity have been obtained by Oxtoby, Godement, and Segal, and generalizations to Banach space valued integrals were given by Dunford, Pettis, and Dieudonné. In order to state precisely the results of the latter named authors, the theories of abstract integration (initiated by Bochner and Gelfand) are also discussed in the report.

The problem of decomposing a measure space was first treated (in a special case) by von Neumann; since then it has received extensive attention from Halmos, Maharam, and Dieudonné. Particularly relevant here is the work of Maharam characterizing measure algebras and their generalizations; the report gives a summary of her results.

Only this somewhat telegraphic table of contents is offered here because, unlike some of the other lectures at the Conference on Algebraic Tendencies in Analysis, the oral presentation of the spokesman was essentially a subset (selected by the spokesman) of the collaboratively prepared written report, which is soon to be published elsewhere.

UNIVERSITY OF CHICAGO,
CHICAGO, ILL., U. S. A.

THE REDUCTION PROBLEM IN SPECTRAL THEORY

NELSON DUNFORD

As spokesman for the panel on spectral theory I have been given a choice of either presenting a survey of progress in this field or of speaking on some special aspect of the subject. Since the committee has, so far, been able only to summarize a small proportion of the research in spectral theory and since the report of the panel will be published in detail elsewhere, I have chosen to talk to you about a problem in spectral theory which has been of particular interest to me. The problem is that of the complete reduction of a linear operator. What is sought is something that will parallel closely the classical theories of the canonical reduction of a finite matrix. The interpretations to be given for this problem will be more readily understood if the known results for finite matrices are first stated in a form similar to that which is to be expected for certain more extensive classes of linear operators. Consequently let us first consider a finite-dimensional linear vector space $\mathfrak{X}$ over the field of complex numbers. Let $\lambda_1, \cdots, \lambda_k$ be the characteristic numbers of the linear transformation T in $\mathfrak{X}$ and let ν_i be the multiplicity of λ_i as a root of the minimal equation for T. If the symbol $\mathfrak{M}_\lambda^\nu$ is used for the set of vectors x for which $(T - \lambda I)^\nu x = 0$, then a geometrical form of the canonical reduction of T is given by the equation

$$\mathfrak{X} = \mathfrak{M}_{\lambda_1}^{\nu_1} \oplus \cdots \oplus \mathfrak{M}_{\lambda_k}^{\nu_k}. \tag{1}$$

This equation means that every $x \in \mathfrak{X}$ has a unique decomposition of the form $x = x_1 + \cdots + x_k$ with $x_i \in \mathfrak{M}_{\lambda_i}^{\nu_i}$, $i = 1, \cdots, k$. If we define $E_{\lambda_i} x = x_i$, then E_{λ_i} is a projection of $\mathfrak{X}$ onto $\mathfrak{M}_{\lambda_i}^{\nu_i}$ and equation (1) becomes

$$I = E_{\lambda_1} + \cdots + E_{\lambda_k}. \tag{2}$$

Let us define, for every set σ of complex numbers, the operator E_σ to be zero if σ contains none of the numbers $\lambda_1, \cdots, \lambda_k$, otherwise E_σ is the sum $\sum E_{\lambda_i}$ taken over those i for which $\lambda_i \in \sigma$. It follows in an entirely elementary fashion that the operators E_σ have the following properties

$$\begin{aligned} &\text{(i)} \quad E_\sigma \cap E_\tau = E_{\sigma \cap \tau}, \qquad E_\sigma \cup E_\tau = E_{\sigma \cup \tau}, \qquad E_\emptyset = 0, \qquad E_{\mathrm{pl}} = I, \\ &\text{(ii)} \quad E_\tau T = T E_\tau, \qquad \sigma[T, E_\tau \mathfrak{X}] \subset \tau, \end{aligned} \tag{3}$$

where here we have written $A \cap B$ for AB, $A \cup B$ for $A + B - AB$, $\emptyset$ for the void set, pl for the whole complex plane and $\sigma[T, \mathfrak{M}]$ for the spectrum of T when considered as an operator in the subspace $\mathfrak{M}$. Property (3(i)) states that E_σ, $\sigma \in B$, is a homomorphic map of the Boolean algebra $\mathfrak{B}$ of all the subsets of the plane into a Boolean algebra of projections. Since, as may be readily shown, conditions (3) determine E_σ, $\sigma \in \mathfrak{B}$, uniquely and since conditions (3) do not explicitly exhibit the finite character of T as is exhibited in equations (1) and (2), we are led by conditions (3) close to the formulation of the problem. There

are two features however of the conditions (3) which should be modified if we are thinking, as we are, of interpreting the problem for a bounded linear operator T in a complex Banach space $\mathfrak{X}$. First of all it is too much to expect that E_σ be defined for every set σ of complex numbers. Secondly, since the spectrum of a bounded operator is always a closed point set, it seems desirable to write the condition (3(ii)) as $E_\tau T = TE_\tau$, $\sigma[T, E_\tau\mathfrak{X}] \subset \bar{\tau}$. Thus we are led to the following rough formulation of the problem.

PROBLEM (first part). *Find a sufficiently large Boolean algebra $\mathfrak{B}$ of sets in the complex plane and a homomorphic map E_τ, $\tau \in \mathfrak{B}$, of $\mathfrak{B}$ into a Boolean algebra of projections such that for every $\tau \in \mathfrak{B}$*

$$E_\tau T = TE_\tau, \qquad \sigma[T, E_\tau\mathfrak{X}] \subset \bar{\tau}.$$

The second part of the problem will make clear what is meant by "a sufficiently large Boolean algebra". We are going to demand that the algebra $\mathfrak{B}$ be large enough so that an operational calculus may be based upon the *resolution of the identity* E_τ, $\tau \in \mathfrak{B}$.

Before making this statement more precise let us again turn our attention to the case where $\mathfrak{X}$ is finite-dimensional. In this case it follows immediately from the properties of the projections E_{λ_i} that for any polynomial f we have

$$f(T) = \sum_{i=1}^{k} \sum_{\nu=0}^{\nu_i-1} \frac{f^{(\nu)}(\lambda_i)}{\nu!} (T - \lambda_i I)^\nu E_{\lambda_i}. \tag{4}$$

This formula, which is an analytical representation of a homomorphism between the algebra of polynomials in T and a direct sum of reduced polynomial algebras, enables one to calculate $f(T)$ in terms of the projections E_{λ_i} and the values taken on by the scalar function f and its derivatives on the spectrum $\sigma(T)$ of T. In terms of the resolution of the identity E_σ, the formula (4) may be written as

$$f(T) = \sum_{\nu=0}^{\infty} \int_{\sigma(T)} \frac{f^{(\nu)}(\lambda)}{\nu!} (T - \lambda I)^\nu \, dE_\lambda. \tag{5}$$

If the indices ν_i for T satisfy $\nu_i \leqq m$, then (5) takes the form

$$f(T) = \sum_{\nu=0}^{m-1} \int_{\sigma(T)} \frac{f^{(\nu)}(\lambda)}{\nu!} (T - \lambda I)^\nu \, dE_\lambda. \tag{6}$$

In particular, Hermitian symmetric matrices have $\nu_i = 1$, and for such T, equation (5) reduces to

$$f(T) = \int_{\sigma(T)} f(\lambda) \, dE_\lambda. \tag{7}$$

The formula (4) or (5) may be used to define its left-hand side when f is a function in the class $F(\sigma(T))$ of all complex-valued functions single-valued and analytic on the spectrum $\sigma(T)$ of T. It then sets up a homomorphic mapping of

the algebra $F(\sigma(T))$ onto the algebra of polynomials in T. From this fact it may be deduced immediately that

$$f(T) = \frac{1}{2\pi i}\int_C (\lambda I - T)^{-1} f(\lambda)\, d\lambda, \qquad f \in F(\sigma(T)), \tag{8}$$

where C is a contour surrounding $\sigma(T)$ and chosen so that f is analytic[1] on and within C.

Now in case T is a bounded linear operator on the complex Banach space $\mathfrak{X}$, the class $F(\sigma(T))$ may be defined as before and the formula (8) may be used[2] to define the operator $f(T)$ which corresponds to a scalar function f in the algebra $F(\sigma(T))$. Such a correspondence is an algebraic homomorphism. The formula (8) has proved to be of considerable value in some discussions[3] but of little or no help in certain problems which might be readily handled with the aid of a formula of the type (5), (6), or (7). This is because formula (8) gives $f(T)$ in terms of the values $f(\lambda)$ with λ on a contour C surrounding the spectrum, whereas (5), (6), or (7) gives $f(T)$ in terms of the scalar function f and its derivatives on the spectrum. Thus for an operator T satisfying an equation of the type (6) one would expect to be able to show that $f_n(T) \to 0$ providing f_n and its first $m - 1$ derivatives converge (in some sufficiently strong sense[4]) to zero on the spectrum $\sigma(T)$. This is one of the types of problems encountered in ergodic theory. The second part of the problem is then that of representing the homomorphism given by (8) of the algebra $F(\sigma(T))$ into an algebra of operators in terms of the values of the scalar function and its derivatives on the spectrum. Or, more explicitly, we may state the following.

PROBLEM (second part). *Express the integral*

$$\frac{1}{2\pi i}\int_C (\lambda I - T)^{-1} f(\lambda)\, d\lambda, \qquad f \in F(\sigma(T)),$$

in terms of the quantities

$$E_\tau,\ T^n,\ f^{(n)}(\lambda); \qquad \tau \in \mathfrak{B}, \qquad \lambda \in \sigma(T), \qquad n = 0, 1, \cdots.$$

[1] In the case where f is a power series and C is a circle, formula (8) has been given by H. Poincaré, *Sur les groupes continus*, Transactions of the Cambridge Philosophical Society vol. 18 (1900) pp. 220–255.

[2] See, for example, A. E. Taylor, *Analysis in complex Banach spaces*, Bull. Amer. Math. Soc. vol. 49 (1943) pp. 652–669, N. Dunford, *Spectral theory*, Bull. Amer. Math. Soc. vol. 49 (1943) pp. 637–651, and N. Dunford, *Spectral theory* I, Trans. Amer. Math. Soc. vol. 54 (1943) pp. 185–217.

[3] To mention but one of many uses to which formula (8) has been put, we cite the extension of the Fredholm-F. Riesz-T. H. Hildebrandt theory of integral equations to the case where the resolvent $T(\xi)$ is a meromorphic function of $1/\xi$.

[4] If for example E_σ is countably additive, one would expect the bounded pointwise convergence of $f_n^{(\nu)}(\lambda)$, $\nu = 0, \cdots, m-1$, $\lambda \in \sigma(T)$, to yield the desired conclusion, whereas if only $|E_\sigma| \leqq M$, one would expect to have to assume the uniform convergence of these functions.

It is through the contemplation of the plausible methods of attack on this second part of the problem that one is led to several interpretations of the first part. For, in the light of the various solutions of the second part in the finite-dimensional case (the solutions expressed by (5), (6), and (7)), it is clear that some form of integration will be needed in the construction of the solution of the second part of the problem. If, for example, the norms of the projections E_τ, $\tau \in \mathfrak{B}$, are unbounded, a different definition of integration would be needed than that which might suffice otherwise. In the bounded case, perhaps an integral of the Hildebrandt-Fichtenholz-Kantorovich type would suffice to throw light on the problem. If the resolution of the identity E_τ, $\tau \in \mathfrak{B}$, is countably additive (in the weak operator topology), it is of course bounded and a Lebesgue or a Riemann integral may be used to integrate the required functions provided the Boolean algebra $\mathfrak{B}$ contains enough sets. The family of Borel sets is sufficiently large for this purpose, but for some operators it is much larger than necessary. Of course it is conceivable that the problem may be given a fair degree of precision even without the assumption that E_τ, for a fixed $\tau \in \mathfrak{B}$, be a bounded linear operator. Thus there are several distinct problems determined by the metric or topological conditions imposed upon the homomorphism E_τ, $\tau \in \mathfrak{B}$. For operators T having a countably additive resolution of the identity E_τ defined for τ in the algebra of Borel sets, the properties of the Lebesgue integral yield such elegant results concerning T (especially if the resolvent $T(\xi) = (\xi I - T)^{-1}$ has a finite rate of growth for ξ near $\sigma(T)$) that it seems highly desirable to characterize this class of operators. Consequently I shall devote the remaining part of this lecture to summarizing a few of the known[5] results concerning operators which possess a countably additive resolution of the identity E_τ defined for τ in the family $\mathfrak{B}$ of Borel sets in the plane. Such operators will, for the sake of brevity, be called *spectral operators*. For spectral operators the solution of the second part of the problem is immediate and is summarized in the following theorem.

THEOREM. *If T is a spectral operator, the resolution of the identity E_τ, for τ a Borel set, is unique and for every $f \in F(\sigma(T))$ we have*

$$f(T) = \sum_{n=0}^{\infty} \int_{\sigma(T)} \frac{f^{(n)}(\lambda)}{n!} (T - \lambda I)^n \, dE_\lambda, \tag{5}$$

where the integral exists as a Riemann integral in the uniform topology of operators and the series converges in the uniform topology of operators.

The principal problem is that of characterizing spectral operators, and very little is known concerning this. There are, however, a few necessary conditions

[5] The formulation of the reduction problem as outlined here as well as the results to be mentioned were obtained in connection with the research done under contract N7onr-448 with, and reported to, the Office of Naval Research during the period between Sept. 1, 1947 and May 1, 1948.

and, for operators in a restricted class, conditions which are necessary and sufficient in order that an operator T be a spectral operator. These conditions may all be stated in terms of the analytic properties of the resolvent $T(\xi) = (\xi I - T)^{-1}$. We shall first describe three necessary conditions. For a vector $x \in \mathfrak{X}$ the vector-valued function $T(\xi)x$ defined on the resolvent set $\rho(T)$ of T is analytic and may have analytic extensions to a larger open set containing $\rho(T)$. Let $x(\xi)$ be the *maximal* analytic extension of $T(\xi)x$ and let $\rho(x)$ be the open set upon which $x(\xi)$ is defined. Then if T is a spectral operator, $x(\xi)$ is single-valued. If $\sigma(x)$ is the complement of $\rho(x)$ and if $[\sigma] = [x \mid \sigma(x) \subset \sigma]$, then a second necessary condition is that $[\sigma]$ be closed if σ is closed. The third condition is stated in terms of a sort of *residue*. Let C be a rectifiable Jordan curve in the domain of analyticity of $x^*x(\xi)$ where x^* is a point of the conjugate space $\mathfrak{X}^*$. Let σ be the set of singularities of $x^*x(\xi)$ which are contained within C, and let

$$(x^*, x)_\sigma \equiv \frac{1}{2\pi i}\int_C x^*x(\xi)\, d\xi.$$

Then an obvious necessary condition (even if T is to have only a bounded resolution of the identity) is the existence of a constant K such that $|(x^*, x)_\sigma| \leqq K|x|\cdot|x^*|$. To summarize we state the theorem:

THEOREM. *In order that T be a spectral operator the following conditions are necessary.*

N(1) $x(\xi)$ *is single-valued,*

N(2) $[\sigma] = [x \mid \sigma(x) \subset \sigma]$ *is closed if σ is closed,*

N(3) $|(x^*, x)_\sigma| \leqq K|x|\cdot|x^*|$ *(the boundedness condition).*[6]

Before stating conditions which are sufficient to insure that T be a spectral operator, we should like to discuss briefly the following question. If an operator T has its spectrum $\sigma(T)$ nowhere dense in the complex plane (in this case N(1) is automatically satisfied), how near do conditions N(2) and N(3) come to being sufficient for the existence of a countably additive resolution of the identity E_τ defined for τ in the family $\mathfrak{B}$ of Borel sets? Let us suppose then, for the moment, that T satisfies N(2) and N(3) and has its spectrum $\sigma(T)$ nowhere dense. We shall describe how a Borel field $m(T)$ of sets "measurable T" may be defined which is, in a certain sense, a maximal domain of definition of a countably additive resolution of the identity for T. We shall then state a condition on the rate of growth of the resolvent (condition S(1) to follow) which will be sufficient to insure that $m(T)$ contains all Borel sets.

[6] The geometrical interpretation of the boundedness condition is that it demands a minimum positive angle between the various manifolds upon which we are going to project. In case T is a self-adjoint operator in Hilbert space, the spectrum is real and, by choosing the contour C to be symmetrical with respect to the real axis, it may be shown that K in N(3) is at most one.

We begin by defining the class s_1 as the class of all sets σ of complex numbers with the property that vectors of the form $x + y$ with $\sigma(x) \subset \sigma$ and $\sigma(y) \subset \tilde{\sigma}$ are dense in $\mathfrak{X}$. For such σ there is clearly one and only one bounded projection E_σ in $\mathfrak{X}$ with the properties that $E_\sigma x = x$ if $\sigma(x) \subset \sigma$, $E_\sigma x = 0$ if $\sigma(x) \subset \tilde{\sigma}$. The operator E_σ also has the properties $TE_\sigma = E_\sigma T$, $\sigma[T, E_\sigma X] \subset \bar{\sigma}$.

Next we define the class s_2 of all sets σ of complex numbers which have the following property. For every $z \in \mathfrak{X}$ and every $\epsilon > 0$ there are vectors x and y with

$$| x + y - z | < \epsilon, \qquad \sigma(x) \subset \sigma(z)\sigma, \qquad \sigma(y) \subset \sigma(z)\tilde{\sigma}.$$

This family s_2 is contained in s_1, is a Boolean algebra, and E_σ, $\sigma \in s_2$, is a bounded resolution of the identity for T, i.e., E_σ, $\sigma \in s_2$, satisfies the conditions

(i) E_σ, $\sigma \in s_2$, is a homomorphism of the Boolean algebra s_2 into an algebra of projection operators.

(ii) $TE_\sigma = E_\sigma T$, $\sigma[T, E_\sigma X] \subset \bar{\sigma}$.

(iii) $| E_\sigma | \leqq K$.

The resolution of the identity E_σ is countably additive on the subalgebra $s_3 \subset s_2$ defined as follows. The symbol s_3 will be used for those sets $\sigma \subset s_1$ for which there exist closed sets μ_n, ν_n in s_2 with $\nu_n \subset \sigma$, $\mu_n \subset \tilde{\sigma}$, $n = 1, 2, \cdots$, and

$$x = \lim_n (E_{\nu_n} + E_{\mu_n})x, \qquad x \in \mathfrak{X}.$$

The family s_3 is a Boolean algebra and E_σ, $\sigma \in s_3$, is a countably additive (in the strong topology) resolution of the identity for T. The smallest Borel algebra of sets containing the Boolean algebra s_3 will be called the family of *sets measurable* T and will be denoted by $m(T)$.

THEOREM. *If $\sigma(T)$ is nowhere dense and if $\mathfrak{X}$ is weakly complete, a countably additive resolution of the identity E_σ, $\sigma \in m(T)$, exists provided the conditions* N(2) *and* N(3) *are satisfied.*

We shall now restrict our attention to an operator T whose spectrum $\sigma(T)$ is contained in a Jordan curve Γ of class C''. Conditions which are sufficient to insure that T is a spectral operator will be described. The first of these is the following.

S(1) (*The growth condition*). *For each $\zeta \in \Gamma$ there is an integer $\nu(\zeta)$ and a number $K(\zeta)$ such that*

$$| (\zeta - \xi)^{\nu(\zeta)} T(\xi) | \leqq K(\zeta)$$

for ξ near ζ and in the normal to Γ at ζ.

A more restrictive condition which will occur in some of the following remarks is the following.

S(1)$_m$. *If $d(\xi)$ is the distance from ξ to the spectrum $\sigma(T)$, then*

$$d(\xi)^m | T(\xi) | \leqq K$$

for ξ near the spectrum.

Unfortunately, as elementary examples show, neither of these conditions is sufficient to insure that T be a spectral operator. However, in reflexive spaces, we have the following two theorems.

THEOREM. *Let $\mathfrak{X}$ be reflexive and let T satisfy the growth condition* S(1). *Then T is a spectral operator if and only if the boundedness condition* N(3) *is satisfied.*

A spectral operator is said to be of *type m* in case

$$f(T) = \sum_{n=0}^{m-1} \int_{\sigma(T)} \frac{f^{(n)}(\lambda)}{n!} (T - \lambda I)^n \, dE_\lambda, \qquad f \in F(\sigma(T)).$$

In reflexive spaces such operators may be characterized according to the following theorem.

THEOREM. *The bounded linear operator T in the reflexive space $\mathfrak{X}$ is a spectral operator of type m if and only if the conditions* $S(1)_m$ *and* N(3) *are satisfied.*

In case the space $\mathfrak{X}$ is assumed only to be weakly complete, there are results analogous to the two preceding theorems but they require another condition which will now be described. Let $\mathfrak{M}_\zeta^\nu$, $\mathfrak{N}_\zeta^\nu$ be the zeros and range of the operator $(T - \zeta I)^\nu$ respectively. Then it turns out that the manifolds $\mathfrak{M}_\zeta^{\nu(\zeta)}$ and $\overline{\mathfrak{N}}_\zeta^{\nu(\zeta)}$ are independent of the function $\nu(\zeta)$ satisfying S(1). If these manifolds are designated by the symbols $\mathfrak{M}_\zeta$, $\mathfrak{N}_\zeta$ respectively, then the additional condition referred to is the following.

S(2) (*The density condition*). *For every ζ in a set dense on Γ we have $M_\zeta + N_\zeta$ dense in $\mathfrak{X}$.*

The density condition will, of course, be satisfied providing no subarc (of positive length) of Γ contains entirely points in the point spectrum of T^*, the adjoint of T. Thus we have the following theorem.

THEOREM. *In weakly complete spaces, an operator T is a spectral operator provided it satisfies the growth, density, and boundedness conditions, and it will be an operator of type m if it satisfies, in addition, condition* $S(1)_m$.

The role played by the boundedness condition N(3), as a sufficient condition, may be replaced by a type of mean rate of growth condition. In order to state this, suppose that $\xi(\lambda, 0)$, $-1 \leqq \lambda \leqq 1$, gives Γ and that for $\delta > 0$ the points $\xi = \xi(\lambda, \delta)$ and $\xi^- = \xi(\lambda, -\delta)$ are on the normal to Γ at $\xi(\lambda, 0)$, ξ being on the exterior normal and ξ^- on the interior normal and both at a distance δ from the point $\xi(\lambda, 0)$.

S(3) (*The mean rate of growth condition*). *For every $x \in \mathfrak{X}$ and $x^* \in \mathfrak{X}^*$ the resolvent $T(\xi)$ or some operator function $U(\xi)$ having the same residues satisfies the condition*

$$\limsup_{\delta \to 0} \int_{-1}^{+1} \left| x^* \left\{ U(\xi) \frac{\partial \xi}{\partial \lambda} - U(\xi^-) \frac{\partial \xi^-}{\partial \lambda} \right\} x \right| d\lambda < \infty.$$

For certain curves Γ the mean rate of growth condition may replace the boundedness condition as a necessary as well as a sufficient condition. We shall summarize the known facts for the case where the spectrum $\sigma(T)$ is contained in the interval $(-1, 1)$. In this case the growth and mean growth conditions may be more readily comprehended. They become the following.

G(1) (*The growth condition in case* $\sigma(T) \subset (-1, 1)$). *There is an integer* $\nu(\lambda)$ *such that for* δ *real*

$$\limsup_{\delta \to 0} | \delta^{\nu(\lambda)} T(\lambda + i\delta) | < \infty, \qquad -1 < \lambda < 1.$$

G(2) (*The mean rate of growth condition in case* $\sigma(T) \subset (-1, 1)$). *For some* $U(\xi)$ *having the same residues as* $T(\xi)$ *we have*

$$\limsup_{\delta \to 0} \int_{-1}^{+1} | x^*\{U(\lambda + i\delta) - U(\lambda - i\delta)\}x | \, d\lambda < \infty.$$

We may summarize as follows:

THEOREM. *If* $\mathfrak{X}$ *is reflective,* $\sigma(T) \subset (-1, 1)$, *and* T *satisfies the growth condition* G(1), *then* T *is a spectral operator if and only if the mean rate of growth condition* G(2) *is satisfied.*

THEOREM. *If* $\mathfrak{X}$ *is reflexive and* $\sigma(T) \subset (-1, 1)$, *then* T *is a spectral operator of type* m *if and only if it satisfies the mean rate of growth condition* G(2) *and the growth condition*

$$G_m \qquad | \delta^m T(\lambda + i\delta) | \leqq K, \qquad -1 < \lambda, \delta < 1.$$

Furthermore T *will be of type* 1 *if and only if it satisfies* G_1 *and* G(2) *with* $U(\xi) = T(\xi)$.

If $\mathfrak{X}$ is not necessarily reflexive but weakly complete, the two preceding theorems remain valid provided T satisfies the density condition S(2). This density condition will be satisfied if, for example, the union of the continuous spectrum and the resolvent set is dense in $(-1, 1)$.

The results we have outlined for spectral operators in weakly complete or reflexive spaces have analogues in an arbitrary complex Banach space $\mathfrak{X}$. The chief role played by the assumption of completeness was to assure that E_σ was defined for every Borel set σ and not only for those σ in a "sufficiently large" Boolean algebra $\mathfrak{B}$.

YALE UNIVERSITY,
NEW HAVEN, CONN., U. S. A.

APPLIED FUNCTIONAL ANALYSIS

N. ARONSZAJN

At the preliminary meetings of the panel it was decided to limit the subject to: "Results in differential problems connected with problems in applied mathematics and obtained by essential use of the theory of abstract spaces". In view of the great amount of material under this title, the spokesman decided to further limit the address to: *Application of Hilbert spaces to linear differential problems*, especially boundary value and eigenvalue problems.

In this address an attempt is made to give a unified presentation of the different ways in which Hilbert spaces were applied, and to state the basic difficulties and problems encountered in this application.

The report of the panel which is to appear in a volume of Mathematical Surveys will treat the results obtained in this application as well as results pertaining to subjects under the more comprehensive title decided upon by the panel.

In the presentation given here all references and bibliography have been omitted. These will be given in the report of the panel.

1. Generalities about differential problems.[1] In a domain D with boundary C in a ν-dimensional euclidean space we consider boundary value problems of the form

$$Au = 0 \quad \text{in} \quad D, \qquad \Lambda_i u = \phi_i \quad \text{on} \quad C, \tag{1.1}$$

and eigenvalue problems

$$Au = \mu u \quad \text{in} \quad D, \qquad \Lambda_i u = 0 \quad \text{on} \quad C, \tag{1.2}$$

where A is a linear differential operator of order m with coefficients defined in $\bar{D}$; Λ_i are linear differential boundary operators. The boundary C and the coefficients of the operators will be supposed sufficiently regular. The above equations have a meaning for functions u belonging to the class $\mathcal{C}^{(m)}$ in $\bar{D}$ (i.e., continuous in $\bar{D}$ with all their derivatives of order less than or equal to m). In boundary value problems the boundary values ϕ_i will be supposed as given by a function $q \in \mathcal{C}^{(m)}$ so that

$$\phi_i = \Lambda_i q. \tag{1.3}$$

We shall also treat the more general eigenvalue problem where the equation $Au = \mu u$ is replaced by $Au = \mu Bu$, with the operator B defined in $\bar{D}$.

An operator A defined in $\bar{D}$ together with a system $\{\Lambda_i\}$ of boundary operators on C forms a linear differential system $\{A; \Lambda_i\}$. To avoid complication we shall suppose $\{\Lambda_i\}$ a *normal* system, i.e., the Λ_i are normal boundary operators of

[1] Some developments of this section are not proved completely as yet in all generality for dimensions greater than or equal to 3. For the sake of brevity and a unified presentation, all developments are stated without qualification.

strictly increasing orders. A boundary operator Λ, say of order k, is called normal when represented in terms of normal and tangential derivatives it contains the pure normal derivative $\partial^k/\partial n^k$ with coefficient 1—all its other terms containing pure normal derivatives of order less than k.

The notions of elliptic, parabolic, and hyperbolic operators on a set $\mathcal{E}$ will be defined in the usual way; also the formal adjoint A^* of A. A is formally self-adjoint if $A = A^*$.

A linear differential system $\{A; \Lambda_i\}$ with a normal system $\{\Lambda_i\}$ composed of l operators of orders less than or equal to $m-1$ always possesses an adjoint system $\{A^*; \Lambda_j^*\}$ with a system $\{\Lambda_j^*\}$ (not necessarily normal) composed of at most $m-l$ operators such that

$$\int_D Au\,\bar{v}\,dz = \int_D u\overline{A^*v}\,dz + \int_C \left(\sum_i \Lambda_i u\overline{\Lambda_i' v} + \sum_j \Lambda_j'' u\overline{\Lambda_j^* v}\right) ds. \tag{1.4}$$

The Λ_i' and Λ_j'' are some boundary operators such that the bilinear differential form in $\int_C$ is of total order less than or equal to $m-1$. The system $\{\Lambda_j^*\}$ is well determined except for an equivalence.[2] If an adjoint system $\{A^*; \Lambda_j^*\}$ can be formed with a normal system $\{\Lambda_j^*\}$ (which is always the case when A is elliptic on C), then the adjoint system of $\{A^*; \Lambda_j^*\}$ is, except for an equivalence, the system $\{A; \Lambda_i\}$.

If A is formally self-adjoint and $\{\Lambda_j^*\}$ is equivalent to $\{\Lambda_i\}$, the linear differential system $\{A; \Lambda_i\}$ is called self-adjoint. If A is elliptic on C and of order $2t$, the simplest example of a self-adjoint system is obtained with $\Lambda_i = \partial^i/\partial n^i$, $i = 0, 1, 2, \cdots, t-1$. This is a generalized Dirichlet system and the corresponding boundary value problems are called generalized Dirichlet problems.

The system $\{A; \Lambda_i\}$ is positive or positive definite when the quadratic form $\int_D Au\,\bar{u}\,dz$ is positive or positive definite in the class of all functions $u \in \mathcal{C}^{(m)}$ satisfying the boundary conditions $\Lambda_i u = 0$. In such a case A must be formally self-adjoint and elliptic or parabolic; hence it must be of even order $m = 2t$. The system is called formally positive with respect to operators A_k and Ω_j if the quadratic form is representable as a sum of formally positive terms:

$$\int_D Au\,\bar{u}\,dz = \sum_k \int_D |\,A_k u\,|^2\,dz + \sum_j \int_C |\,\Omega_j u\,|^2\,ds, \tag{1.5}$$

for all $u \in \mathcal{C}^{(m)}$ satisfying $\Lambda_i u = 0$. We shall suppose always that the Ω_j's are of orders less than or equal to $t-1$.

Formula (1.5) implies that all A_k's are of orders less than or equal to t and that

$$A = \sum_k A_k^* A_k. \tag{1.6}$$

When A satisfies such a formula, it is called *formally positive*.

[2] Two systems $\{\Lambda_i\}$ and $\{\Lambda_j'\}$ are equivalent if the corresponding sets of boundary conditions $\Lambda_i u = 0$ and $\Lambda_j' u = 0$ are equivalent. We define similarly weaker and stronger systems of boundary operators.

For an elliptic (in $\bar{D}$) formally positive operator A of order $2t$, we call normal boundary operators of orders less than or equal to $t-1$ *stable boundary* operators, those of orders greater than or equal to t, *unstable*.

In the considerations of formally positive linear differential systems $\{A; \Lambda_i\}$ we can always suppose that $\{\Lambda_i\}$ is what we call a *minimal system*. A minimal system is a normal system composed of t operators of orders less than or equal to $2t-1$. Except for an equivalence, its unstable boundary operators are determined by the stable ones.[3] The name "minimal" is justified by the fact that if we consider all normal systems $\{\Lambda_i\}$ such that $\{A; \Lambda_i\}$ is formally positive, each of them is stronger than some minimal system and that two minimal systems are either equivalent or noncomparable.

2. Generalities about the application of Hilbert spaces to differential problems. Formally, this application began with the introduction of abstract Hilbert spaces in the late twenties. However, many classical methods concerning variational problems equivalent to differential problems can be interpreted as belonging to Hilbert space theory. The basic idea of the application consists in translating the differential problem in terms of a well determined operator in a chosen Hilbert space. It is especially fruitful if this operator is self-adjoint—the theory of such operators being the most completely analyzed.[4]

Two methods can be distinguished for choosing a Hilbert space for the treatment of a differential problem.

(1) *First method.* Usually the space $\mathfrak{L}^2$ in the domain D is taken as the Hilbert space. A differential system $\{A; \Lambda_i\}$ determines directly the operator L in $\mathfrak{L}^2$ with domain composed of all functions $u \in \mathfrak{C}^{(m)}$ satisfying the boundary conditions $\Lambda_i u = 0$. For such functions Lu is defined as equal to Au. The operators treated here are always nonbounded.

If we consider two adjoint systems $\{A; \Lambda_i\}$ and $\{A^*; \Lambda_j^*\}$ and the corresponding operators L and L' in $\mathfrak{L}^2$, it can be proved that the smallest closed extensions $\tilde{L}$ and $\tilde{L}'$ exist and that $\tilde{L}'$ is always a restriction of the adjoint operator $\tilde{L}^*$. Two basic problems arise: 1°. *To characterize the functions u belonging to the domain $\mathfrak{D}_{\tilde{L}}$ of $\tilde{L}$ and to determine the value $\tilde{L}u$*; in particular, to determine when and in what sense $Lu = Au$. 2°. *To determine the systems for which $\tilde{L}^* = \tilde{L}'$.* It is to be expected that this equality is usually true; this would mean that for a self-adjoint system $\{A; \Lambda_i\}$ the operator $\tilde{L}$ is self-adjoint.

The main problem treated by this method for formally self-adjoint A is the following: we can denote by A the operator L in $\mathfrak{L}^2$ corresponding to the system $\{A; 0\}$ (i.e., without any Λ_i's). The operators $\tilde{L}$ for any system $\{A; \Lambda_i\}$ are restrictions of $\tilde{A}$. The problem is then to determine all self-adjoint restrictions of $\tilde{A}$.

[3] In variational problems corresponding to the system $\{A; \Lambda_i\}$ with a minimal system $\{\Lambda_i\}$, the unstable boundary operators appear as the "natural" boundary conditions.

[4] The normal operators have also been thoroughly investigated but until now an application to differential problems does not seem to have been found for them.

The boundary value problem is then treated in the form

$$Av = Aq \quad \text{in} \quad D, \qquad \Lambda_i v = 0 \quad \text{on} \quad C, \tag{2.1}$$

where q is the function realizing the boundary values ϕ_i by (1.3) and v gives the solution u of (1.1) by $u = q - v$. This leads to the inversion of the corresponding operator $\tilde{L}$. The eigenvalue problem leads to the spectral decomposition of $\tilde{L}$. When $\tilde{L}$ is self-adjoint, both problems are completely solvable.

The advantage of the method is that it permits the treatment of the most general differential systems and in many cases allows the establishment of the most general self-adjoint systems. However, the more general eigenvalue problem concerning the equation $Au = \mu Bu$ is not easily translated by the first method into a standard spectral problem for operators in the space $\mathfrak{L}^2$.

(2) *Second method.* This method is closely connected with classical variational methods. It is applicable only to positive systems $\{A; \Lambda_i\}$ and especially to formally positive systems with minimal $\{\Lambda_i\}$. The Hilbert space introduced here depends on the system. At first we consider the space $\mathfrak{C}^{(2t)}$ ($2t$ = order of A) with a quadratic norm, $\| u \|^2$, being given by (1.5). In general this norm will not be a proper norm ($\| u \|$ may be equal to 0 for $u \neq 0$) but usually it can be made into a proper norm by adjoining to the Ω_j's in (1.5) some of the Λ_i's.[5] The space $\mathfrak{C}^{(2t)}$ is then an incomplete Hilbert space. The essential difficulty here is to find a suitable completion of this space.

The advantage of this method vis-à-vis the previous one is that when it can be applied, it leads to much simpler problems in the Hilbert space.

Consider the completion $\overline{\mathfrak{C}^{(2t)}}$ of $\mathfrak{C}^{(2t)}$ and the space $\mathfrak{K}$ composed of all $u \in \mathfrak{C}^{(2t)}$ satisfying all the boundary conditions $\Lambda_i u = 0$. Form the closure $\overline{\mathfrak{K}}$ of $\mathfrak{K}$ in $\overline{\mathfrak{C}^{(2t)}}$ and the orthogonal complement $\mathfrak{K}^\perp = \overline{\mathfrak{C}^{(2t)}} \ominus \overline{\mathfrak{K}}$. A basic problem here is to characterize the spaces $\overline{\mathfrak{K}}$ and $\mathfrak{K}^\perp$. In usual cases the functions w of $\overline{\mathfrak{K}}$ are characterized essentially by the stable boundary conditions $\Lambda_i w = 0$ and those of $\mathfrak{K}^\perp$ by the equation $Aw = 0$ in D and the unstable boundary conditions $\Lambda_i w = 0$. If then $\{\Lambda_i\}$ is a generalized Dirichlet system, the solution u of the boundary value problem (1.1) is the projection of q on $\mathfrak{K}^\perp$, the solution v of (2.1) is the projection of q on $\overline{\mathfrak{K}}$. For other systems $\{\Lambda_i\}$ another simple interpretation can be given for the solutions u and v.

The general eigenvalue problem with the equation $Au = \mu Bu$ in D is treated by considering in $\overline{\mathfrak{K}}$ the quadratic form $\int_D Bu\bar{u}\, dz$. This form determines a linear operator K in $\overline{\mathfrak{K}}$ satisfying $(Ku, \bar{v}) = \int_D Bu\bar{v}\, dz$ where (f, g) is the scalar product in $\overline{\mathfrak{K}}$. The eigenvalue problem is then solved by the spectral decomposition of K.

The connection between the second method and the classical variational methods allows us to apply Hilbert space theory to variational approximation procedures.

[5] Sometimes, as in the Neumann's problem for the Laplacian Δ, we may have to restrict the class $\mathfrak{C}^{(2t)}$ in some suitable way.

3. Concluding remarks. We can enumerate briefly the following advantages obtained by applying Hilbert space theory to linear differential problems.

(1) It allows the problems to be stated and treated under very general assumptions concerning the order of the problem, the nature of the operators, regularity conditions, etc.

(2) Even if it does not permit one to solve the boundary value and eigenvalue problems in such generality,[6] it often clarifies the problems and determines essential properties of the differential operators which permit a complete analysis of these problems.

(3) Concerning concrete problems, it has proved to be most advantageous for approximation methods—previously established variational methods have been improved and new methods devised. It does not seem likely that any of these results could have been attained without the background of the abstract theory of Hilbert spaces.

OKLAHOMA AGRICULTURAL AND MECHANICAL COLLEGE,
STILLWATER, OKLA., U. S. A.

[6] This is to be expected since these problems are not always well defined. This is generally so when A is hyperbolic, and also when the boundary C of the domain is not sufficiently regular.

ERGODIC THEORY

SHIZUO KAKUTANI

It is impossible to give a complete report on ergodic theory in a talk of one hour. It is therefore not my intention to cover all parts of ergodic theory in the present talk. I have to be satisfied with the discussion of two or three topics from ergodic theory. A more detailed and more systematic report on recent development of ergodic theory compiled jointly by N. Dunford, W. F. Eberlein, G. A. Hedlund, E. Hille, J. C. Oxtoby, and myself will be published elsewhere.[1]

Let $(S, \mathfrak{B}, m)$ be a measure space, i.e., a triple consisting of a space $S = \{s\}$, a σ-field $\mathfrak{B} = \{B\}$ of subsets B of S, and a countably additive measure $m(B)$ defined on $\mathfrak{B}$. We do not assume that $m(S) < \infty$, but it is assumed that S is the union of a countable number of subsets from $\mathfrak{B}$ with a finite m-measure. The space $L^p(S) = L^p(S, \mathfrak{B}, m)$ $(p \geqq 1)$ of all complex-valued measurable functions $x(s)$ defined on S with

$$(1) \qquad \| x \|_p = \left(\int_S | x(s) |^p m(ds) \right)^{1/p} < \infty$$

as its norm is defined as usual.

A one-to-one mapping $s' = \varphi(s)$ of S onto itself is called a measure preserving transformation if $B \in \mathfrak{B}$ implies $\varphi(B) \in \mathfrak{B}$, $\varphi^{-1}(B) \in \mathfrak{B}$ and $m(\varphi(B)) = m(\varphi^{-1}(B)) = m(B)$. The main problem of ergodic theory is to discuss the properties of measure preserving transformations. In particular, we are interested in the asymptotic behavior of the iterations $s' = \varphi^n(s)$ of $s' = \varphi(s)$ as n tends to ∞.

Let $s' = \varphi(s)$ be a measure preserving transformation defined on a measure space $(S, \mathfrak{B}, m)$. For any real- or complex-valued function $x(s)$ defined on S, let us put

$$(2) \qquad x_n(s) = \frac{1}{n} \sum_{k=0}^{n-1} x(\varphi^k(s)),$$

$n = 1, 2, \cdots$. The following two theorems are fundamental in ergodic theory:

MEAN ERGODIC THEOREM. *For any function $x(s) \in L^2(S)$, there exists a function $\bar{x}(s) \in L^2(S)$ such that*

$$(3) \qquad \lim_{n \to \infty} \| x_n - \bar{x} \|_2 = 0.$$

[1] The bibliography at the end is by no means complete. It contains only those papers which were quoted in the main text. The address given at the Congress contained more topics than this note. Owing to the limitation in printing space, it was necessary to give up the discussions of many interesting problems. These will be contained in a forthcoming report mentioned above.

INDIVIDUAL ERGODIC THEOREM. *For any function* $x(s) \in L^1(S)$, *there exists a function* $\bar{x}(s) \in L^1(S)$ *such that*

$$\lim_{n\to\infty} x_n(s) = \bar{x}(s) \tag{4}$$

almost everywhere on S.

The mean ergodic theorem is due to J. von Neumann [51] and the individual ergodic theorem is due to G. D. Birkhoff [8]. Birkhoff discussed the case when S is a manifold and when $s' = \varphi(s)$ is a measure preserving homeomorphism of S onto itself. The general case when S is a space with measure but without topology was discussed by A. Khintchine [41] (the case $m(S) < \infty$) and by W. Stepanoff [62] (the case $m(S) \leqq \infty$).

These two theorems were published in the years 1931–1932, and became the starting point of all development of ergodic theory in the following 18 years 1932–1950. Most of the results obtained in the earlier part of this period by G. D. Birkhoff, G. A. Hedlund, E. Hopf, A. Khintchine, B. O. Koopman, J. von Neumann, W. Seidel, W. Stepanoff, N. Wiener, A Wintner, and others are collected in a monograph *Ergodentheorie* by E. Hopf [33] which appeared in 1937. This book contains not only the summary of previously published results, but also many new results and new proofs which had not been published elsewhere.

J. von Neumann's proof [51] of the mean ergodic theorem is based on the observation due to B. O. Koopman [44] that

$$x(s) \to Vx(s) = x(\varphi(s)) \tag{5}$$

defines a unitary operator V of $L^2(S)$ onto itself. The mean ergodic theorem then states that the arithmetic mean $T_n = (1/n)\sum_{k=0}^{n-1} V^k$ of the iterations V^k of V converges strongly to a certain bounded linear operator P of $L^2(S)$ into itself, i.e., that

$$\lim_{n\to\infty} \left\| \frac{1}{n} \sum_{k=0}^{n-1} V^k x - Px \right\|_2 = 0 \tag{6}$$

for any function $x(s) \in L^2(S)$. (It is easy to see that P is a projection operator of $L^2(S)$ onto the linear subspace of $L^2(S)$ consisting of all functions which are invariant under φ, and satisfies $PV = VP = P^2 = P$.) It turned out that the theory of spectral resolution of unitary operators in a Hilbert space developed by J. von Neumann [50] and M. H. Stone [63] in the years 1929–1932 was a powerful tool which was not only useful for the proof of the mean ergodic theorem, but also was indispensable in carrying out the Fourier analysis of measure preserving transformations.

The mean ergodic theorem was generalized in two directions: (i) to more general Banach spaces and (ii) to more general classes of linear operators. The first result concerning the convergence of the arithmetic means $T_n = (1/n)\sum_{k=0}^{n-1} T^k$ of the iterates T^k of a general bounded linear operator T defined on a Hilbert space

was obtained by C. Visser [66] who proved weak convergence for the case when $\{ \| T_n \| \mid n = 1, 2, \cdots \}$ is bounded, and strong convergence for the case when, in addition to this, T is Hermitian, unitary, or completely continuous. It was then observed by F. Riesz [56] that the mean ergodic theorem holds in $L^p(S)$ $(p \geqq 1)$, i.e., if V is a bounded linear operator defined on $L^p(S)$ by (5), then $T_n = (1/n) \sum_{k=0}^{n-1} V^k$ converges strongly on $L^p(S)$. (We need to assume that $m(S) < \infty$ if $p = 1$, while this assumption is unnecessary if $p > 1$.) At the same time, independently from F. Riesz, the following result was obtained by K. Yosida [69] (also S. Kakutani [36], K. Yosida and S. Kakutani [75]).

MEAN ERGODIC THEOREM IN BANACH SPACES. *Let T be a bounded linear operator of a Banach space X into itself such that* (a) $\{ \| T^n \| \mid n = 1, 2, \cdots \}$ *is bounded. If x is an element of X such that* (b) $\{x_n = (1/n) \sum_{k=0}^{n-1} T^k x \mid n = 1, 2, \cdots \}$ *contains a weakly convergent subsequence (i.e., a subsequence which converges weakly to an element x of X), then the sequence $\{x_n \mid n = 1, 2, \cdots \}$ itself converges strongly to the same limit $\bar{x}$.*

It is remarkable that the existence of a weakly convergent subsequence implies the strong convergence of the sequence itself. It is easy to see that under the condition (a) which is obviously satisfied by $T = V$ in $L^p(S)$ $(p \geqq 1)$, the second condition (b) is satisfied by every element x of $L^p(S)$. In case $p > 1$ this follows from the fact that the unit sphere of $L^p(S)$ is sequentially weakly compact (i.e., from any sequence $\{x_n \mid n = 1, 2, \cdots \}$ of elements of $L^p(S)$ with $\| x_n \| \leqq 1$, $n = 1, 2, \cdots$, we can find a weakly convergent subsequence), and in case $p = 1$ this is a consequence of the equi-uniform integrability of the sequence $\{x_n(s) \mid n = 1, 2, \cdots \}$ defined by (2) (i.e., for any $\epsilon > 0$ there exists a $\delta > 0$ such that $m(B) < \delta$ implies $\int_B | x_n(s) | \, m(ds) < \epsilon$ for $n = 1, 2, \cdots$. (Here we need the assumption $m(S) < \infty$. In fact, equi-uniform integrability implies sequential weak compactness only if $m(S) < \infty$.) Thus the results of F. Riesz [56] mentioned above follows from the mean ergodic theorem in Banach spaces. G. Birkhoff [5] discussed the case of a bounded linear operator in an abstract (L)-space (AL) satisfying the condition (a) such that, for any $x \in (AL)$, the sequence $\{x_n = (1/n) \sum_{k=0}^{n-1} T^k x \mid n = 1, 2, \cdots \}$ is bounded above by an element x^* of (AL). The mean ergodic theorem holds in this case since, when the abstract (L)-space (AL) is represented as a concrete (L)-space, the corresponding sequence $\{x_n(s) \mid n = 1, 2, \cdots \}$ will be equi-uniformly integrable and hence contains a weakly convergent subsequence (S. Kakutani [37; 39], F. Riesz [58]).

Since the unit sphere of a reflexive Banach space is sequentially weakly compact, we have the following result (E. R. Lorch [49]): Let T be a bounded linear operator of a reflexive Banach space X into itself such that $\{ \| T^n \| \mid n = 1, 2, \cdots \}$ is bounded. Then there exists a projection operator P such that $PT = TP = P^2 = P$ and $T_n = (1/n) \sum_{k=0}^{n-1} T^k$ converges strongly to P. It is easy to see that P is a projection to the linear subspace of X consisting of all elements of X which are invariant under T. Further, since uniformly convex space is re-

flexive, the same theorem holds in any uniformly convex Banach space. In case T is a contraction (i.e., $\| T \| \leqq 1$ and hence $\| T^n \| \leqq 1$, $n = 1, 2, \cdots$) a very simple proof of the mean ergodic theorem in a uniformly convex Banach space was given by G. Birkhoff [6] and F. Riesz [57]. This proof is geometrical and has the advantage of avoiding the use of the notion of weak convergence. Since the spaces $L^p(S)$ $(p > 1)$ are uniformly convex, this may be considered as a generalization of the mean ergodic theorem in $L^p(S)$ $(p > 1)$. In case $p = 2$, F. Riesz [59] (cf. E. Hopf [33, p. 23]), F. Riesz and B. von Sz. Nagy [61] gave very simple proofs of the mean ergodic theorem for unitary operators or contractions which are completely free from the notion of spectral resolution, while essential use is made of some fundamental properties of the inner product.

One of the most important facts about the mean ergodic theorem in Banach spaces is that the strong convergence of $x_{m,n} = (1/(n-m)) \sum_{k=m}^{n-1} T^k x$ as $n - m \to \infty$ (to the same limit $\bar{x}$) follows from it immediately. In other words, the strong convergence of $x_{p,p+n} = (1/n) \sum_{k=p}^{p+n-1} T^k x$ as $n \to \infty$ (to the same limit $\bar{x}$) is uniform in p. It is to be noticed that a similar fact does not hold for the case of the individual ergodic theorem.

Another important fact about the mean ergodic theorem in Banach space is that the existence of the inverse T^{-1} of T is not assumed. This fact makes it possible to apply the mean ergodic theorem to the problems in probability and statistical mechanics where irreversible processes are discussed. The mean ergodic theorem in Banach spaces was applied by K. Yosida and S. Kakutani [73; 75] to the problems of Markoff processes, and it was shown that many of the results in the theory of Markoff processes previously obtained by W. Doeblin [12], J. L. Doob [14], M. Fréchet [24; 25], and N. Kryloff and N. Bogoliouboff [45; 46] can be obtained operator-theoretically by using the mean ergodic theorem in Banach spaces and the following theorem:

UNIFORM ERGODIC THEOREM. *Let T be a bounded linear operator of a Banach space X into itself such that $\{ \| T^n \| \mid n = 1, 2, \cdots \}$ is bounded, and assume that there exist an integer $m \geqq 1$ and a completely continuous linear operator V of X into itself such that $\| T^m - V \| < 1$. Then T has only a finite number of proper values of absolute value 1 (each with a finite multiplicity) and if we denote these by λ_i, $i = 1, \cdots, N$, then T can be expressed in the form:*

$$T = \sum_{i=1}^{N} \lambda_i P_i + S, \tag{7}$$

where P_i is a projection operator which maps X onto a finite-dimensional proper subspace of X corresponding to the proper value λ_i of T such that $P_i^2 = P_i$, $P_i P_j = 0$ $(i \neq j)$, $P_i T = T P_i = \lambda_i P_i$, $i = 1, \cdots, N$, and S is a bounded linear operator of X into itself such that $P_i S = S P_i = 0$, $i = 1, \cdots, N$ and $\| S^n \| < \alpha(1+\beta)^{-n}$, $n = 1, 2, \cdots$ (α and β are positive constants). From (7) it follows that

$$T^n = \sum_{i=1}^{N} \lambda_i^n P_i + S^n, \tag{8}$$

$n = 1, 2, \cdots$, *and hence there exists a constant* M *such that*

$$\left\| \frac{1}{n} \sum_{k=0}^{n-1} \left(\frac{T}{\lambda_i} \right)^k - P_i \right\| \leqq \frac{M}{n} \tag{9}$$

$i = 1, \cdots, N; n = 1, 2, \cdots$.

If we consider $n \to T^n$ as a bounded representation of the additive semi-group of all non-negative integers n by bounded linear operators T^n of a Banach space X into itself, then the mean ergodic theorem may be considered as a result concerning the strong convergence of the means of this representation. This interpretation leads to the following general formulation (L. Alaoglu and G. Birkhoff [1; 2]): Let $G = \{g\}$ be a group or a semi-group (commutative or not), and let $g \to T^g$ be a bounded representation of G by bounded linear operators T^g of a Banach space X into itself. Let $\{\mu_n(E) \mid n = 1, 2, \cdots\}$ be a sequence of measures defined on a certain σ-field $\mathfrak{E} = \{E\}$ of subsets E of G with $\mu_n(G) = 1$, $n = 1, 2, \cdots$, and let $T_n = \int_G T^g \mu_n(dg)$ be the mean of T^g with respect to μ_n. Under what conditions on X, G, $\{T^g \mid g \in G\}$, and $\{\mu_n \mid n = 1, 2, \cdots\}$ can we conclude the strong convergence of $\{T_n \mid n = 1, 2, \cdots\}$? Mean ergodic theorems of N. Dunford [19; 20] and N. Wiener [68] concerning a finite commutative system $\{T_1, \cdots, T_N\}$ of bounded linear operators correspond to the case when $G = \{g\}$ is a free abelian semi-group generated by N elements $\{g_1, \cdots, g_N\}$ (i.e., G consists of all elements of the form $g = g_1^{k_1} \cdots g_N^{k_N}$ where $k_i = 0, 1, 2, \cdots; i = 1, \cdots, N$; two elements $g = g_1^{k_1} \cdots g_N^{k_N}$ and $g = g_1^{l_1} \cdots g_N^{l_N}$ are equal if and only if $k_i = l_i$, $i = 1, \cdots, N$) and when

$$\mu_n(E) = \frac{|E \cap D_n|}{|D_n|} \tag{10}$$

where D_n is the set of all elements $g = g_1^{k_1} \cdots g_N^{k_N}$ of G with $k_1^2 + \cdots + k_N^2 \leqq n^2$ and $|E \cap D_n|$, $|D_n|$ denote the number of elements in $E \cap D_n$, D_n respectively. In fact, if $g_i \to T^{g_i} = T_i$, $i = 1, \cdots, N$, then

$$T_n = \int_G T^g \mu_n(dg) = \frac{1}{|D_n|} \sum_{k_1^2 + \cdots + k_N^2 \leqq n^2} T_1^{k_1} \cdots T_N^{k_N}. \tag{11}$$

N. Dunford [19; 20] proved the strong convergence of $\{T_n \mid n = 1, 2, \cdots\}$ when X is a reflexive Banach space. N. Wiener [68] discussed the case when $X = L^p(S)$ $(p > 1)$ and when $T_i x(s) = x(\varphi_i(s))$, $i = 1, \cdots, N$, where $\{\varphi_1, \cdots, \varphi_N\}$ is a commutative system of measure preserving transformations. Asymptotic invariance of the measures μ_n (i.e., the property that, for any fixed g, the total variation of $\mu_n(gE) - \mu_n(E)$ converges to 0 as $n \to \infty$), or the asymptotic invariance of the means T_n (i.e., the property that for any fixed g,

$$\lim_{n \to \infty} (T^g T_n - T_n) = 0 \tag{12}$$

with respect to the uniform topology of operators) plays an essential role in the argument. This fact makes it rather difficult to obtain a general mean

ergodic theorem for the case when G is not commutative. In fact, in case G is a free group or a free semi-group generated by more than one element, there is no such sequence of measures defined on G with the property of asymptotic invariance. M. M. Day [10; 11] discussed mean ergodic theorems for general commutative or noncommutative groups or semi-groups. G. Birkhoff [7] discussed the case when the sequential convergence $\lim_{n\to\infty} T_n$ is replaced by the Moore-Smith convergence $\lim_\alpha T_\alpha$, where α runs through a certain directed set. Further, W. F. Eberlein [22; 23] discussed the case when in the definition of asymptotic invariance of means T_α the uniform convergence of $\lim_\alpha(T^g T_\alpha - T_\alpha)$ is replaced by the strong convergence, and when the condition of boundedness of $\{\| T^g \| \mid g \in G\}$ is somewhat relaxed. W. F. Eberlein also discussed the uniform ergodic theorem for noncyclic, noncommutative semi-groups.

Further investigations on semi-groups of linear operators and in particular, discussion of differentiability and analyticity of semi-groups with real or complex parameters, and discussions of relations between the ergodic theorem (C_1 summability) and other summation methods (e.g., Abel, Cesàro, and Poisson summability) were carried out in detail by E. Hille [30; 31].

Generalization of the individual ergodic theorem was made in two directions: (i) to more precise results and (ii) to more general class of transformations. N. Wiener [68] proved the following:

DOMINATED ERGODIC THEOREM. *If $x(s)$ is a real-valued function from $L^p(S)$ $(p > 1)$, then*

$$x^*(s) = \sup_n x_n(s) \tag{13}$$

(where $x_n(s)$ is defined by (2)) belongs to the same $L^p(S)$; if $x(s)$ is a real-valued function from $L^1(S)$ such that $x(s) \log (1 + | x(s) |)$ also belongs to $L^1(S)$, then $x^*(s)$ belongs to $L^1(S)$. ($x^*(s)$ does not necessarily belong to $L^1(S)$ if we only assume that $x(s) \in L^1(S)$.)

It is easy to see that both the individual ergodic theorem and the dominated ergodic theorem follow from the following lemma called the maximal ergodic theorem:

MAXIMAL ERGODIC THEOREM. *If $x(s)$ is a real-valued function from $L^1(S)$, and if $x^*(s)$ is defined by (13), then*

$$\alpha m(E^*(\alpha)) \leqq \int_{E^*(\alpha)} x(s) m(ds) \tag{14}$$

where

$$E^*(\alpha) = \{s \mid x^*(s) > \alpha\}. \tag{15}$$

This inequality was obtained by K. Yosida and S. Kakutani [74] by using the method of A. Kolmogoroff [43] by which he gave a very simple proof of the

individual ergodic theorem. The same inequality was also discussed by C. Carathéodory [9] and H. R. Pitt [53]. A similar inequality in which $x^*(s) = \sup_n x_n(s)$ is replaced by $\bar{x}(s) = \limsup x_n(s)$ was obtained earlier by G. D. Birkhoff [8], and this inequality was the essential key to his original proof of the individual ergodic theorem.

As was observed by F. Riesz [59], there is a close connection between the maximal ergodic theorem and the following lemma of G. H. Hardy and J. E. Littlewood [28] (also F. Riesz [55]) called the maximal theorem:

MAXIMAL THEOREM. *Let $p > 1$. If $\{a_1, \cdots, a_n\}$ is a finite sequence of positive numbers (n is arbitrary), then*

$$\sum_{k=1}^{n} \left\{ \max\left(a_k, \frac{a_k + a_{k-1}}{2}, \cdots, \frac{a_k + \cdots + a_1}{k} \right) \right\}^p \leqq \left(\frac{p}{p-1} \right) \sum_{k=1}^{n} a_k^p, \tag{16}$$

where $(p/(p-1))^p$ is the best possible constant independent of n.

M. Fukamiya [26] gave a direct proof of the dominated ergodic theorem by using this maximal theorem. Further, it is to be noticed that there is a close analogy between the proof of the maximal ergodic theorem and the proof of the differentiation theorem (to the effect that a real-valued nondecreasing function of a real variable is almost everywhere differentiable) due to F. Riesz [54]. There is no wonder that such an analogy exists. In fact, as was observed by N. Wiener [68], if we consider the case of a flow (i.e., a one-parameter group of $\{\varphi_t \mid -\infty < t < \infty\}$ of measure preserving transformations $s' = \varphi_t(s)$ such that $\varphi_t(\varphi_u(s)) = \varphi_{t+u}(s)$ for all real numbers t, u and for all $s \in S$, satisfying the measurability condition that $y(s, t) = x(\varphi_t(s))$ is measurable as a function of two variables (s, t) if $x(s)$ is a measurable function of s), then the individual ergodic theorem is a statement concerning the existence almost everywhere of the limit

$$\frac{1}{T} \int_0^T x(\varphi_t(s))\, dt \tag{17}$$

as $T \to \infty$, while the differentiation theorem is concerned with the limit of the same expression (17) as $T \to 0$. P. Hartman [29] gave a proof of the maximal ergodic theorem for the case of the flow by using the method of F. Riesz [54].

Another important fact worth mentioning is a similarity between the maximal ergodic theorem and the inequality due to A. Kolmogoroff [42] which plays a fundamental role in the theory of sums of independent random variables. This analogy becomes more striking if we observe the following fundamental lemma in the theory of Martingale due to J. L. Doob [14], P. Lévy [47], and J. Ville [65] which we may call the Martingale theorem:

MARTINGALE THEOREM. *If $\{x_k(s) \mid k = 1, \cdots, n\}$ is a finite ordered system of real-valued random variables defined on a probability space (S, B, m) (i.e., a measure space with the normalization condition $m(S) = 1$) satisfying the condition of*

Martingale: for any $k = 1, \cdots, n-1$, *and for any system of real numbers* $\alpha_1, \cdots, \alpha_k$, *the conditional expectation of* $x_{k+1}(s)$ *under the conditions* $x_i(s) = \alpha_i$, $i = 1, \cdots, k$, *is equal to* α_k, *then*

$$\alpha m(E^*(\alpha)) \leqq \int_{E^*(\alpha)} x_n(s) m(ds) \tag{18}$$

where

$$E^*(\alpha) = \{s \mid \max_{1 \leqq k \leqq n} x_k(s) > \alpha\}. \tag{19}$$

The existence of this analogy is very interesting in view of the fact that the strong law of large numbers (concerning the arithmetic means of independent random variables which are integrable and have the same distribution) follows immediately both from the individual ergodic theorem (J. L. Doob [13] and E. Hopf [32]) (and hence from the maximal ergodic theorem) and from the Martingale theorem. This fact was also recently observed by M. Loève [48]. In spite of the fact that the proofs of these theorems are also very similar to each other, we have thus far no satisfactory formulation of a general theorem which contains both the maximal ergodic theorem and the Martingale theorem as a special case. Finally, it is to be noticed that the proofs of fundamental limit theorems in the theory of measures in infinite product spaces due to S. Andersen and B. Jessen [3] have also a very similar form. The reason for this similarity lies in the fact that these limit theorems can, in most of the cases, be proved by using the method of Martingale.

F. Riesz [59] gave a simple proof of the maximal ergodic theorem which is based on the following combinatorial lemma concerning a sequence of real numbers: if $\{a_i \mid i = 0, 1, 2, \cdots\}$ is a sequence of real numbers such that

$$\max_{1 \leqq k \leqq N} \sum_{j=0}^{k-1} a_{i+j} > 0 \tag{20}$$

for $i = 0, 1, 2, \cdots$, where N is a fixed positive integer, then

$$\max_{1 \leqq k \leqq N} \sum_{j=0}^{i+k-1} a_j > 0 \tag{21}$$

for $i = 0, 1, 2, \cdots$. Similar proofs were given to the maximal ergodic theorem by E. Hopf [34] and H. R. Pitt [53].

Birkhoff's individual ergodic theorem was generalized to the following form by E. Hopf [33, p. 49]: If $x(s)$ is any function from $L^1(S)$, and if $y(s)$ is a positive measurable function defined on S such that

$$\sum_{k=0}^{\infty} y(\varphi^k(s)) = \infty \tag{22}$$

almost everywhere on S, then

$$\lim_{n\to\infty} \frac{\sum_{k=0}^{n-1} x(\varphi^k(s))}{\sum_{k=0}^{n-1} y(\varphi^k(s))} \tag{23}$$

exists almost everywhere on S. We do not assume that $m(S) < \infty$. But it is worthwhile to observe that the condition (22) is satisfied by every positive measurable function $y(s)$ if $m(S) < \infty$. In case $m(S) = \infty$, this condition is satisfied by every positive measurable function $y(s)$ if S has no wandering set of positive measure, i.e., if for any measurable set B with $m(B) > 0$ there exists a positive integer n such that $m(\varphi^n(B) \cap B) > 0$. If $y(s) \equiv 1$, the condition is obviously satisfied, and this theorem is reduced to the individual ergodic theorem of Birkhoff. The proof of this ergodic theorem of Hopf can be carried out in exactly the same way as that of the individual ergodic theorem of Birkhoff.

The first generalization of the individual ergodic theorem to the case when $s' = \varphi(s)$ is not necessarily measure preserving was given by W. Hurewicz [35] who also discussed the case of absolutely continuous set functions instead of integrable measurable functions. P. R. Halmos [27] then gave a general formulation of the individual ergodic theorem which contains Hurewicz's ergodic theorem and Hopf's ergodic theorem at the same time: a one-to-one mapping $s' = \varphi(s)$ of S onto itself is called measurability preserving if $B \in \mathfrak{B}$ implies $\varphi(B) \in \mathfrak{B}$, $\varphi^{-1}(B) \in \mathfrak{B}$; $s' = \varphi(s)$ is called nonsingular if $m(B) = 0$ implies $m(\varphi(B)) = m(\varphi^{-1}(B)) = 0$. For any nonsingular measurability preserving transformation $s' = \varphi(s)$, there exist positive measurable functions $\omega_n(s)$ such that

$$m(\varphi^n(B)) = \int_B \omega_n(s) m(ds) \tag{24}$$

for any $B \in \mathfrak{B}$ and $n = 0, 1, 2, \cdots$. For any measurable function $x(s)$ defined on S, let us put

$$x^{(n)}(s) = \sum_{k=0}^{n-1} x(\varphi^k(s))\omega_k(s), \tag{25}$$

$n = 1, 2, \cdots$. Then the general ergodic theorem of Halmos reads as follows:

GENERAL ERGODIC THEOREM. *If $x(s)$ is any function from $L^1(S)$ and if $y(s)$ is a positive measurable function defined on S such that $y^{(n)}(s) \to \infty$ (as $n \to \infty$) almost everywhere on S, then*

$$\lim_{n\to\infty} \frac{x^{(n)}(s)}{y^{(n)}(s)} \tag{26}$$

exists almost everywhere on S.

It is clear that $\omega_n(s) \equiv 1, n = 0, 1, \cdots$, if $s' = \varphi(s)$ is measure preserving, and the general ergodic theorem is reduced to Hopf's ergodic theorem in this case. J. C. Oxtoby [52] obtained a further generalization of ergodic theorems of Hurewicz and Halmos in which the absolute continuity of the set functions in

question is not assumed and the nonexistence of wandering sets is not required.

It is interesting to observe that the proof of this general ergodic theorem does not involve any further difficulty than that of the maximal ergodic theorem. In fact as was shown by Y. N. Dowker [18], the same method used by F. Riesz [58] for the proof of the maximal ergodic theorem can be applied to obtain the following inequality:

$$\alpha \int_{E^*(\alpha)} y(s)m(ds) \leqq \int_{E^*(\alpha)} x(s)m(ds) \tag{27}$$

where

$$E^*(\alpha) = \left\{ s \,\middle|\, \sup_n \frac{x^{(n)}(s)}{y^{(n)}(s)} > \alpha \right\}. \tag{28}$$

The general ergodic theorem follows from this immediately.

Individual ergodic theorems concerning a commutative system $\{\varphi_1, \cdots, \varphi_N\}$ of measure preserving transformations were discussed by N. Wiener [68]. N. Wiener proved that for any function $x(s) \in L^p(S)$ $(p > 1)$ there exists a function $\bar{x}(s) \in L^p(S)$ (same p) such that

$$\lim_{n\to\infty} \frac{1}{|D_n|} \sum_{k_1^2+\cdots+k_N^2 \leqq n^2} x(\varphi_1^{k_1} \cdots \varphi_N^{k_N}(s)) = \bar{x}(s) \tag{29}$$

almost everywhere on S, where $|D_n|$ means the same thing as in (11). The commutativity of the system $\{\varphi_1, \cdots, \varphi_N\}$ is essential in the proof. In case $\{\varphi_1, \cdots, \varphi_N\}$ is not commutative, the problem of obtaining a natural generalization of the individual ergodic theorem seems to be difficult. In view of this fact, the following recent unpublished result of N. Dunford and A. Zygmund is remarkable: Let $\{\varphi_1, \cdots, \varphi_N\}$ be a finite system of measure preserving transformations, commutative or not. Then, for any function $x(s) \in L^p(S)$ $(p > 1)$, there exists a function $\bar{x}(s) \in L^p(S)$ (same p) such that

$$\lim \frac{1}{n_1 \cdots n_N} \sum_{k_1=0}^{n_1-1} \cdots \sum_{k_N=0}^{n_N-1} x(\varphi_1^{k_1} \cdots \varphi_N^{k_N}(s)) = \bar{x}(s) \tag{30}$$

almost everywhere on S, as $n_1 \to \infty, \cdots, n_N \to \infty$ independently of one another. It is to be noticed that the measure preserving transformations $s' = \varphi_1^{k_1} \cdots \varphi_N^{k_N}(s)$, $k_i = 0, 1, 2, \cdots$; $i = 1, \cdots, N$, do not exhaust the group generated by φ_i, $i = 1, \cdots, N$.

It is an interesting problem to discuss the relations between the mean ergodic theorem and the individual ergodic theorem. N. Wiener [67; 68] was the first to discuss this problem and was led to the notion of the dominated ergodic theorem. It is easy to see that the individual ergodic theorem and the dominated ergodic theorem together imply the mean ergodic theorem in $L^p(S)$ $(p > 1)$. In fact, for any function $x(s) \in L^p(S) \cap L^1(S)$, the sequence $\{x_n(s) \mid n = 1, 2, \cdots\}$ converges to a function $\bar{x}(s) \in L^1(S)$ almost everywhere on S, and because of the dominated ergodic theorem, there exists a function $x^*(s) \in L^p(S)$ such that

$|x_n(s)| \leqq x^*(s)$ for almost all $s \in S$ and $n = 1, 2, \cdots$. Thus $\bar{x}(s) \in L^p(S)$ and $\lim_{n\to\infty} \| x_n - x \|_p = 0$. The mean ergodic theorem in $L^p(S)$ then follows immediately from the facts that $L^p(S) \cap L^1(S)$ is dense in $L^p(S)$ and that $T_n = (1/n)\sum_{k=0}^{n-1} V^k$ satisfies $\| T_n \| \leqq 1$, $n = 1, 2, \cdots$.(The argument above is interesting only when $m(S) = \infty$. If $m(S) < \infty$, we can avoid the use of the dominated ergodic theorem if we consider the class $M(S)$ of all bounded measurable functions defined on S instead of $L^p(S) \cap L^1(S)$.)

The problem whether conversely the individual ergodic theorem can be proved from the mean ergodic theorem was discussed by K. Yosida [71; 72] who succeeded in giving an operator-theoretical proof of the individual ergodic theorem. Yosida's idea is to consider $Vx(s) = x(\varphi(s))$ as a linear operator in the (F)-space of all real-valued measurable functions $x(s)$ defined on S with the quasi-norm

$$|||x||| = \int_S \frac{|x(s)|}{1 + |x(s)|}\, m(ds). \tag{31}$$

[9] The completeness of this space and the fact that $x^*(s) = \sup_n x_n(s) < \infty$ (or that $\bar{x}(s) = \limsup x_n(s) < \infty$) almost everywhere on S for any real-valued function $x(s)$ from $L^1(S)$ are essential in Yosida's proof. It is, however, to be noticed that the proof of this last fact, which is obviously an immediate consequence of the maximal ergodic theorem, is by no means easy and requires almost the same amount of arguments as the proof of the maximal ergodic theorem itself.

In case $s' = \varphi(s)$ is not necessarily measure preserving, but is only measurability preserving and nonsingular, our problem takes a different form: the problem is no more how to prove one theorem from the other, but is whether or not one theorem holds true when we know that the other is true. N. Dunford and D. S. Miller [21] proved that the strong convergence of $T_n = (1/n)\sum_{k=0}^{n-1} V^k$ in $L^1(S)$ implies the almost everywhere convergence of $T_n x(s)$ for any $x(s) \in L^1(S)$. Again the following fact is essential in the proof: if $T_n x$ is strongly convergent in $L^1(S)$ for any $x(s) \in L^1(S)$, then for any real-valued function $x(s) \in L^1(S)$, $x^*(s) = \sup_n x_n(s) < \infty$ almost everywhere on S. F. Riesz [60] later gave another proof for this result by using again the same lemma concerning a sequence of real numbers quoted above which he had used for the proof of the maximal ergodic theorem. As was shown by Y. N. Dowker [17] the converse of this result is not true, i.e., there exists a nonsingular measurability preserving transformation for which the individual ergodic theorem holds while the mean ergodic theorem does not hold. This phenomenon is quite natural since the individual ergodic theorem is preserved by any nonsingular measurability preserving transformation while the mean ergodic theorem can be affected by such a transformation.

Finally, an interesting and important generalization of the individual ergodic theorem called the random ergodic theorem was recently obtained by S. M. Ulam and J. von Neumann [64]. The random ergodic theorem may be formulated as follows: Let $(S, \mathfrak{B}, m)$ and $(\Gamma, \mathfrak{E}, \mu)$ be two measure spaces of which the

second is a probability space (i.e., $\mu(\Gamma) = 1$). For any $\gamma \in \Gamma$, let $s' = \varphi_\gamma(s)$ be a measure preserving transformation of S onto itself. We assume that $s' = \varphi_\gamma(s)$ depends measurably on γ, i.e., that $y(s, \gamma) = x(\varphi_\gamma(s))$ is a measurable function of two variables (s, γ) if $x(s)$ is a measurable function of s. Let $\Omega = \mathrm{P}_{n=-\infty}^{\infty}\Gamma_n$ be a two-sided infinite direct product space of Γ_n, where $\Gamma_n = \Gamma$, $n = 0, \pm 1, \pm 2, \cdots$, and let $(\Omega, \mathfrak{E}^*, \mu^*) = \mathrm{P}_{n=-\infty}^{\infty}(\Gamma_n, \mathfrak{E}_n, \mu_n)$ be the direct product measure space of $(\Gamma_n, \mathfrak{E}_n, \mu_n) = (\Gamma, \mathfrak{E}, \mu)$, $n = 0, \pm 1, \pm 2, \cdots$, defined on Ω. For any $\omega = \{\gamma_n \mid n = 0, \pm 1, \pm 2, \cdots\} \in \Omega$, $\gamma_n = \gamma_n(\omega)$ is called the nth coordinate of ω. Then we have the following theorem.

RANDOM ERGODIC THEOREM. *For any function $x(s) \in L^1(S)$ and for almost all ω (the exceptional null set of ω may depend on $x(s)$), there exists a function $\bar{x}_\omega(s) \in L^1(S)$ such that*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} x(\varphi_{\gamma_{k-1}(\omega)} \cdots \varphi_{\gamma_1(\omega)} \varphi_{\gamma_0(\omega)}(s)) = \bar{x}_\omega(s) \tag{32}$$

almost everywhere on S.

In a very special case when Γ consists of two elements $+1$ and -1 each with μ-measure $1/2$, we may consider $\{\gamma_n(\omega) \mid n = 1, 2, \cdots\}$ as Rademacher's system, i.e., the sequence of functions defined on the unit interval $\Omega = \{\omega \mid 0 \leqq \omega < 1\}$ by $\gamma_n(\omega) = (-1)^k$ if $k2^{-n} \leqq \omega < (k+1)2^{-n}$, $k = 0, 1, \cdots, 2^n - 1$; $n = 1, 2, \cdots$. In this case the random ergodic theorem may be stated as follows: Let φ_{+1}, φ_{-1} be any two measure preserving transformations, commutative or not. Let $\{\psi_n \mid n = 1, 2, \cdots\}$ be an independent random sequence of measure preserving transformations such that the probability that $\psi_n = \varphi_\gamma$ is $1/2$ for $\gamma = \pm 1$. Then, for any function $x(s)$ from $L^1(S)$, the individual ergodic theorem:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} x(\psi_k \cdots \psi_2 \psi_1(s)) = \bar{x}(s) \tag{33}$$

(almost everywhere on S) holds with probability 1.

It is interesting to observe that the random ergodic theorem can be proved by considering the mapping $(s', \omega') = \bar{\varphi}(s, \omega)$ of $S \times \Omega$ onto itself defined by $\bar{\varphi}(s, \omega) = (\varphi_{\gamma_0(\omega)}(s), \sigma(\omega))$, where $\omega' = \sigma(\omega)$ is a shift transformation defined on Ω by $\gamma_n(\sigma(\omega)) = \gamma_{n+1}(\omega)$, $n = 0, \pm 1, \pm 2, \cdots$. It is easy to see that $(s', \omega') = \bar{\varphi}(s, \omega)$ is a measure preserving transformation of $S \times \Omega$ onto itself (with respect to the direct product measure on $S \times \Omega$). The individual ergodic theorem applied to the transformation $(s', \omega') = \bar{\varphi}(s, \omega)$ and to the function $x(s, \omega) = x(s)$ will immediately give (by Fubini's theorem) the required random ergodic theorem.

It was recently observed (H. Anzai [4], S. Kakutani [40]) that the random ergodic theorem has close connections with the theory of Markoff processes with a stable distribution previously discussed by J. L. Doob [14; 16], S. Kakutani [38], and K. Yosida [70].

It is easy to see that the convergence (32) is a strong convergence in $L^p(S)$

if $x(s)$ belongs to $L^p(S)$ ($p > 1$ if $m(S) = \infty$ and $p \geqq 1$ if $m(S) < \infty$). But the following generalization of the mean ergodic theorem in Banach spaces is still open: Let T_{+1}, T_{-1} be two bounded linear transformation operators of a Banach space into itself such that $\| T_i \| \leqq 1$, $i = +1, -1$. We do not assume that T_{+1} and T_{-1} commute. Let $\{\gamma_n(\omega) \mid n = 1, 2, \cdots\}$ be a Rademacher's system. Is it true then that, for almost all ω, there exists a projection operator P_ω such that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} T_{\gamma_0(\omega)} T_{\gamma_1(\omega)} \cdots T_{\gamma_{k-1}(\omega)} = P_\omega \tag{34}$$

strongly?

Bibliography

1. L. Alaoglu and G. Birkhoff, *General ergodic theorem*, Proc. Nat. Acad. Sci. U. S. A. vol. 25 (1939) pp. 628–630.

2. ———, *General ergodic theorems*, Ann. of Math. vol. 41 (1940) pp. 293–309.

3. S. Andersen and B. Jessen, *Some limit theorems on integrals in an abstract set*, Det. Kgl. Danske Videnskabernes Selskab. Matematisk-Fysiske Meddelelser vol. 22 no. 14 (1946).

4. H. Anzai, *Random ergodic theorem with finite possible states*, Osaka Mathematical Journal vol. 2 (1950) pp. 43–49.

5. G. Birkhoff, *Dependent probabilities and spaces (L)*, Proc. Nat. Acad. Sci. U. S. A. vol. 24 (1938) pp. 134–139.

6. ———, *The mean ergodic theorem*, Duke Math. J. vol. 5 (1939) pp. 19–20.

7. ———, *An ergodic theorem for general semi-groups*, Proc. Nat. Acad. Sci. U. S. A. vol. 25 (1939) pp. 625–627.

8. G. D. Birkhoff, *Proof of the ergodic theorem*, Proc. Nat. Acad. Sci. U. S. A. vol. 17 (1932) pp. 656–660.

9. C. Carathéodory, *Bemerkungen zur Riesz-Fischerschen Satz und zur Ergodentheorie*, Abh. Math. Sem. Hansischen Univ. vol. 14 (1941) pp. 351–389.

10. M. M. Day, *Ergodic theorems for abelian semi-groups*, Trans. Amer. Math. Soc. vol. 51 (1942) pp. 399–412.

11. ———, *Means for the bounded functions and ergodicity of the bounded representation of semi-groups*, Trans. Amer. Math. Soc. vol. 69 (1950) pp. 276–291.

12. W. Doeblin, *Sur les propriétés asymptotiques de mouvements régis par certains types de chaînes simples*, Bull. Soc. Math. Roumaine vol. 39-1 (1937) pp. 57–115; vol. 39-2 (1938) pp. 3–61.

13. J. L. Doob, *Note on probability*, Ann. of Math. vol. 37 (1936) pp. 363–367.

14. ———. *Stochastic processes with an integral-valued parameter*, Trans. Amer. Math. Soc. vol. 44 (1938) pp. 87–150.

15. ———, *Regularity properties of certain families of chance variables*, Trans. Amer. Math. Soc. vol. 47 (1940) pp. 455–486.

16. ———, *Asymptotic properties of Markoff transition probabilities*, Trans. Amer. Math. Soc. vol. 63 (1948) pp. 393–421.

17. Y. N. Dowker, *A note on the ergodic theorem*, Bull. Amer. Math. Soc. vol. 55 (1949) pp. 379–383.

18. ———, *A new proof of the general ergodic theorem*, Acta Szeged vol. 12 (1950) pp. 162–166.

19. N. Dunford, *An ergodic theorem for n-parameter groups*, Proc. Nat. Acad. Sci. U. S. A. vol. 25 (1939) pp. 195–196.

20. ———, *A mean ergodic theorem*, Duke Math. J. vol. 5 (1939) pp. 635–646.

21. N. Dunford and D. S. Miller, *On the ergodic theorem*, Trans. Amer. Math. Soc. vol. 60 (1946) pp. 538–549.

22. W. F. Eberlein, *Abstract ergodic theorems*, Proc. Nat. Acad. Sci. U. S. A. vol. 34 (1948) pp. 43–47.

23. ———, *Abstract ergodic theorems and weak almost periodic functions*, Trans. Amer. Math. Soc. vol. 67 (1949) pp. 217–240.

24. M. Fréchet, *Sur l'allure asymptotique de la suite des itérés d'un noyau de Fredholm*, Quart. J. Math. Oxford Ser. vol. 5 (1934) pp. 106–144.

25. ———, *Sur l'allure asymptotique des densités itérées dans le problème des probabilités en chaîne*, Bull. Soc. Math. France vol. 62 (1934) pp. 68–83.

26. M. Fukamiya, *On dominated ergodic theorems in* L_p ($p \geqq 1$), Tôhoku Math. J. vol. 46 (1939) pp. 150–153.

27. P. R. Halmos, *An ergodic theorem*, Proc. Nat. Acad. Sci. U. S. A. vol. 32 (1946) pp. 156–161.

28. G. H. Hardy and J. E. Littlewood, *A maximal theorem with function-theoretic applications*, Acta Math. vol. 54 (1930) pp. 81–116.

29. P. Hartman, *On the ergodic theorems*, Amer. J. Math. vol. 69 (1947) pp. 193–199.

30. E. Hille, *Remarks on ergodic theorems*, Trans. Amer. Math. Soc. vol. 57 (1945) pp. 246–269.

31. ———, *Functional analysis and semi-groups*, Amer. Math. Soc. Colloquium Publications, vol. 31, New York, 1948.

32. E. Hopf, *On causality, statistics and probability*, Journal of Mathematics and Physics vol. 13 (1934) pp. 51–102.

33. ———, *Ergodentheorie*, Berlin, 1937.

34. ———, *Über eine Ungleichung der Ergodentheorie*, Sitzungsberichte der Mathematisch-Naturwissenschaftlichen Abteilung der Bayerischen Akademie der Wissenschaften zu München (1944) pp. 171–176.

35. W. Hurewicz, *Ergodic theorems without invariant measure*, Ann. of Math. vol. 45 (1944) pp. 192–206.

36. S. Kakutani, *Iteration of linear operations in complex Banach spaces*, Proc. Imp. Acad. Tokyo vol. 14 (1938) pp. 295–300.

37. ———, *Mean ergodic theorem in abstract* (L)*-spaces*, Proc. Imp. Acad. Tokyo vol. 15 (1939) pp. 121–123.

38. ———, *Ergodic theorems and Markoff processes with a stable distribution*, Proc. Imp. Acad. Tokyo vol. 16 (1940) pp. 49–54.

39. ———, *Concrete representation of abstract* (L)*-spaces and the mean ergodic theorem*, Ann. of Math. vol. 42 (1941) pp. 523–537.

40. ———, *Random ergodic theorem and Markoff process with a stable distribution*, Symposium on Probability, Berkeley, 1950.

41. A. Khintchine, *Zur Birkhoffschen Lösung des Ergodenproblems*, Math. Ann. vol. 107 (1932) pp. 485–488.

42. A. Kolmogoroff, *Über die Summen durch den Zufall bestimmter unabhängiger Grössen*, Math. Ann. vol. 99 (1928) pp. 309–319; vol. 102 (1929) pp. 484–488.

43. ———, *Ein vereinfachter Beweis des Birkhoff-Khintchinschen Ergodensatzes*, Rec. Math. (Mat. Sbornik) N.S. vol. 2 (1937) pp. 366–368.

44. B. O. Koopman, *Hamiltonian systems and transformations in Hilbert space*, Proc. Nat. Acad. Sci. U. S. A. vol. 17 (1931) pp. 315–318.

45. N. Kryloff and N. Bogolioubolf, *Les propriétés ergodiques des suites des probabilités en chaîne*, C. R. Acad. Sci. Paris vol. 204 (1937) pp. 1454–1456.

46. ———, *Sur les propriétés en chaîne*, C. R. Acad. Sci. Paris vol. 204 (1937) pp. 1386–1388.

47. P. Lévy, *Théorie de l'addition des variables aléatoires*, Paris, 1937.

48. M. Loève, *Fundamental limit theorems in probability theory*, Ann. Math. Statist. vol. 21 (1950) pp. 321–338.

49. E. R. Lorch, *Means of iterated transformations in reflexive vector spaces*, Bull. Amer. Math. Soc. vol. 45 (1939) pp. 945–957.

50. J. von Neumann, *Allgemeine Eigenwerttheorie Hermitescher Funktionaloperatoren*, Math. Ann. vol. 102 (1929) pp. 49–131.

51. ———, *Proof of the quasi-ergodic hypothesis*, Proc. Nat. Acad. Sci. U. S. A. vol. 18 (1932) pp. 70–82.

52. J. C. Oxtoby, *On the ergodic theorem of Hurewicz*, Ann. of Math. vol. 49 (1948) pp. 872–884.

53. H. R. Pitt, *Some generalizations of the ergodic theorem*, Proc. Cambridge Philos. Soc. vol. 38 (1942) pp. 325–343.

54. F. Riesz, *Sur l'existence de la dérivée des fonctions monotones et sur quelques problèmes qui s'y rattachent*, Acta Szeged vol. 5 (1930–1932) pp. 208–221.

55. ———, *Sur un théorème de maximum de MM. Hardy et Littlewood*, J. London Math. Soc. vol. 7 (1932) pp. 10–13.

56. ———, *Some mean ergodic theorems*, J. London Math. Soc. vol. 13 (1938) pp. 274–278.

57. ———, *Another proof of the mean ergodic theorem*, Acta Szeged vol. 10 (1941) pp. 75–76.

58. ———, *Sur la théorie ergodique des espaces abstraits*, Acta Szeged vol. 10 (1941) pp. 1–20; vol. 10 (1941) p. 141.

59. ———, *Sur la théorie ergodique*, Comment. Math. Helv. vol. 17 (1945) pp. 221–239.

60. ———, *On a recent generalization of G. D. Birkhoff's ergodic theorem*, Acta Szeged vol. 11 (1946–1948) pp. 193–200.

61. F. Riesz and B. von Sz. Nagy, *Über Kontraktionen des Hilbertschen Raumes*, Acta Szeged vol. 10 (1943) pp. 202–205.

62. W. Stepanoff, *Sur une extension du théorème ergodique*, Compositio Math. vol. 3 (1936) pp. 239–253.

63. M. H. Stone, *Linear transformations in Hilbert space*, Amer. Math. Soc. Colloquium Publications, vol. 15, New York, 1932.

64. S. M. Ulam and J. von Neumann, *Random ergodic theorems*, Bull. Amer. Math. Soc. vol. 51 (1945) p. 660.

65. J. Ville, *Étude critique de la notion de collectif*, Paris, 1938.

66. C. Visser, *On the iteration of linear operations in a Hilbert space*, Neder. Akad. Wetensch. vol. 41 (1938) pp. 487–495.

67. N. Wiener, *The homogeneous chaos*, Amer. J. Math. vol. 60 (1938) pp. 897–936.

68. ———, *The ergodic theorem*, Duke Math. J. vol. 5 (1939) pp. 1–18.

69. K. Yosida, *Mean ergodic theorem in Banach spaces*, Proc. Imp. Acad. Tokyo vol. 14 (1938) pp. 292–294.

70. ———, *The Markoff process with a stable distribution*, Proc. Imp. Acad. Tokyo vol. 16 (1940) pp. 43–48.

71. ———, *Ergodic theorems of Birkhoff-Khintchine's type*, Jap. J. Math. vol. 17 (1940) pp. 31–36.

72. ———, *An abstract treatment of the individual ergodic theorem*, Proc. Imp. Acad. Tokyo vol. 16 (1940) pp. 280–284.

73. K. Yosida and S. Kakutani, *Applications of mean ergodic theorems to the problems of Markoff process*, Proc. Imp. Acad. Tokyo vol. 14 (1938) pp. 333–339.

74. ———, *Birkhoff's ergodic theorem and the maximal ergodic theorem*, Proc. Imp. Acad. Tokyo vol. 15 (1939) pp. 165–168.

75. ———, *Operator-theoretical treatment of Markoff process and mean ergodic theorem*, Ann. of Math. vol. 42 (1941) pp. 188–228.

Yale University,
New Haven, Conn., U. S. A.

ANALYSIS IN THE LARGE

RECENT ADVANCES IN VARIATIONAL THEORY IN THE LARGE

MARSTON MORSE

1. The theory in 1936. At the time of the Congress in Oslo a general structure had been laid for the theory. It was developed in papers from 1925 on to Morse [1] in 1934, and in those aspects which could be reached by a study of "category" by Lusternik and Schnirelmann. In the Morse theory a space Ω of admissible elements was associated with the different boundary value problems arising in variational theory. Significant subtheories developed prior to 1936 included the following:

(A) The critical points of a function f of the point on a compact differentiable manifold M;

(B) The extremals joining two fixed points of M, for a positive definite regular simple integral J;

(C) The existence of closed extremals of J in the case where M is a compact differentiable manifold without boundaries.

In (B) Morse used the Fréchet space $\Omega(P, Q)$ of sensed arcs joining two fixed points P and Q of M. When M was an n-sphere S_n, the homology groups of $\Omega(P, Q)$ were specifically determined (Morse [1, p. 247]). In (C) the space Ω was a space of special closed curves and when $M = S_n$, the homology groups of Ω were again determined (Morse [1, p. 349]). (Throughout this report we may suppose that in the definition of chains the group of coefficients is an arbitrary finite field.) The methods employed led to the first determination of the homology groups of the symmetric quare of an n-sphere, when the locus of coincident pairs of points is taken as a modulus. See Morse [1, p. 181].

In the category theory a most significant topological theory was initiated by Lusternik and Schnirelmann and in particular the category of the n-dimensional projective plane and of the product of n-circles was obtained and applied to critical point theory. The existence of at least three closed geodesics on a differentiable topological image of a 2-sphere was affirmed, and a proof outlined. Also see Fox for category theory.

The early limitations on critical values. The application of the topological theory to the theory of critical points (including extremals) was sharply limited by differentiability conditions and conditions such as the discreteness of the set of critical values. Such limitations were inherent in the use of the orthogonal trajectories of the level manifolds of f in (A) to define deformations, and in similar procedures in (B) and (C) where J was approximated by its value on families of broken extremals with a limited number of vertices. In the absence of restrictive conditions on the set of critical values, neither the orthogonal tra-

jectories nor the deformations would have been adequately defined. This is shown by an example due to Whitney.

Whitney's example. Whitney has shown the existence of a function F defined and continuous on a unit square in the (x, y)-plane, with F_x and F_y continuous, yet with F possessing a connected set X of critical points whose critical values *fill an interval.* One might say that the mission of the critical point theory was already fulfilled in the Whitney example since the existence of infinitely many critical points is granted. However, this view would be mistaken since the Morse theory also aims at the relations between the critical sets σ (each at one f-level and closed at that level) classified according to the nature of associated local relative homology groups. See Morse [2; 3]. We shall term a function f whose critical values fill some interval a function of Whitney type.

Limitations imposed by the use of singular cycles. The Morse theory aims to associate a critical set σ with each nontrivial homology class H on M (or Ω). Thus for a cycle $z \in H$, let $|z| \subset M$ be a compact carrier (always minimal) of z and set

$$c(H) = \inf_{z \in H} [\sup_{x \in |z|} f(x)] \tag{1.1}$$

$$N_\sigma = \sigma \cup x \mid (f(x) < f(\sigma)). \tag{1.2}$$

We say that H "*causes*" σ if $c(H) = f(\sigma)$, if $N_\sigma \supset z$ for some $z \in H$, and if no proper closed subset of σ has this property. If singular cycles (rather than Vietoris cycles) are used, an homology class H need cause no critical set when M and f are not analytic, even when N is a compact differentiable manifold of class $C^{(n)}$ and f of class $C^{(n)}$ on M, with n arbitrarily large. This is shown in Morse [5] and is not equivalent to the well-known fact that singular homology groups are not always isomorphic to the corresponding Vietoris homology groups.

A theorem on dim σ. Suppose that H_n and H_r are different homology classes of cycles of dimension n and r respectively. If $c(H_n) = c(H_r)$, simple examples will show that H_n and H_r may cause a common critical set σ consisting of just one point. If, however, $n > r$, there are simple conditions sufficient that the critical set σ caused by H_n be such that

$$\dim \sigma \geqq n - r. \tag{1.3}$$

In the special case in which the critical values of f are *isolated and f is of class C''*, such conditions have been given by Lusternik and Schnirelmann. If, however, f is a function of the Whitney type or if the orthogonal trajectories of the level manifolds of f fail to have the usual field properties, then the earlier proofs of (1.3) are not applicable. However, a general theorem with (1.3) as a conclusion can be stated in terms which are purely topological and proved without reference to the category theory. One may assume that the compact manifold M is merely locally euclidean (not in general differentiable) and that the function f is merely continuous. One uses a purely topological definition of a critical point of f. Such a theorem is stated for the first time in §5.

Multiple integrals. Prior to 1936 the theory had not been applied to obtain any general theorems on the existence of unstable (non-minimizing) extremals of multiple integrals J. The reasons for this are clear. There is as yet no workable generalization for J of the orthogonal trajectories of the level manifolds of f. In addition, the theorem on the existence of minimizing extremals in the small for positive definite regular simple integrals has no apparent counterpart for multiple integrals. Finally, it turned out that the concept of lower semi-continuity of a positive definite functional J, while adequate in the theory of the absolute minimum (if accompanied by conditions implying compactness of the set of admissible elements), requires the addition of the concept of *upper reducibility* of J (defined in §2) if unstable extremals are sought. The general grounds preparing for an attack on multiple integral problems were laid in Morse [2].

Minimal surfaces. Using these general concepts Morse and Tompkins, and Shiffman, independently and at essentially the same time, proved the existence of unstable minimal surfaces of disc type spanning a simple closed curve g. The conditions initially imposed on g were somewhat heavier than rectifiability. These conditions have been progressively reduced until Shiffman has established the existence of a minimal surface of minimax type for a rectifiable g. Morse has verified Shiffman's result by an independent proof not yet published. Some of the results of Morse and Tompkins cover more general topological aspects and have not yet been reduced to the hypothesis of rectifiability.

To escape the limitations of the earlier development the critical point theory has advanced at three levels in

(a) *the general causal theory* Morse [2],

(b) *the span theory* Morse [3],

(c) *the nondegenerate theory* Morse [4].

These three levels are distinguished by their objectives and hypotheses. They are all concerned with a positive lower semi-continuous function F on an abstract metric space S. We further distinguish these theories as follows.

2. The causal theory Morse [2]. This theory imposes minimum conditions on F and S and is concerned with critical sets as *caused* by homology or homotopy classes of various types. Extreme types of deformations are the isotopies (deformations in which the images at any one time are topological) and the F-deformations which require no derivatives of F for their definition. See §5. The function F is assumed positive and lower semi-continuous over the metric space S.

As distinguished from the span theory and nondegenerate theory, the causal theory does not aim at a complete set of *relations* between the classified critical sets. Its hypotheses are too general and the critical sets too numerous and complex in most problems to make a theory of relations feasible. On the other

hand Whitney functions can be treated under this theory. The causal homology theory imposes two additional conditions on the function F and metric space S:

(i) *the F-accessibility of S*;

(ii) *the upper-reducibility of F*.

We shall define these conditions.

The set of points $x \in S$ on which $F(x) \leqq c < \infty$ will be denoted by S_c. The space S is termed *F-accessible* if any nonbounding Vietoris k-cycle z, given as homologous to zero mod S_{c+e} for each $e > 0$, is homologous to a k-cycle in S_c. If the subsets S_c are compact for each c, S can be shown to be F-accessible, making use of the lower semi-continuity of F.

A continuous deformation of a subset $A \subset S$ which replaces each point $x \in A$ by a point $x^t \in S$ at the time t is called an *F-deformation* of A if for each t $(0 \leqq t \leqq 1)$ and $x \in A$

$$F(x) - F(x^t) \geqq 0. \tag{2.1}$$

This deformation is termed *proper* over A if the difference (2.1) is bounded from zero whenever the distance $d(x, x^t)$ is bounded from zero. The integrals of variational theory are ordinarily lower semi-continuous but not upper semi-continuous. Upper reducibility in some form is, however, satisfied in general and serves in place of upper semi-continuity.

Let p be a point of S with $F(p) < \infty$. We term F *upper reducible* at p if for any set S_b, with $b > F(p)$, there exists an F-deformation D of a neighborhood N_b of p relative to S_b such that

$$\limsup_{(t,x)\to(1,p)} F(x^t) \leqq F(p) \qquad (x \in N_b, 0 \leqq t \leqq 1) \tag{2.2}$$

and such that D is a *proper* F-deformation of any subset of N_b on which $F(x)$ exceeds $F(p)$ by a positive constant. *Note that the deletion of t in* (2.2) *yields the definition of upper semi-continuity of F at p*. It is easy to show that lower semi-continuity and upper reducibility are independent conditions on F.

To state the principal theorem one must define a homotopic critical point. A point p of S at which F is finite will be called *homotopically ordinary* if some neighborhood of p, relative to some S_b with $b > F(p)$, admits a proper F-deformation $(0 \leqq t \leqq 1)$ which ultimately (for some t) displaces p. The point p will be termed *homotopically critical* if not homotopically ordinary.

THEOREM 2.1. *Suppose that F is positive, lower semi-continuous and upper reducible on an F-accessible metric space S. Let H be a nontrivial homology class of Vietoris cycles on S. If there is a k-cycle $z \in H$ on some set S_a with finite a, there is a least value of a such that there is a k-cycle of H on S_a. If c is this minimum value of a, there is at least one homotopic critical point p with $F(p) = c$.*

In establishing the existence of unstable minimal surfaces it was fundamental to show that the Douglas-Dirichlet integral

$$\iint_D \left[\sum_i \left(\frac{\partial x_i}{\partial u}\right)^2 + \left(\frac{\partial x_i}{\partial v}\right)^2\right] du\, dv \qquad (i = 1, 2, 3)$$

taken over the circular disc D was upper reducible, admitting harmonic surfaces $x_i(u, v)$, $i = 1, 2, 3$, spanning the given simple closed curve g. A modification of the above theorem was then used in which the homology class H was replaced by a suitable relative homology class. See Morse-Tompkins or Shiffman.

3. The span theory Morse [3]. In seeking the *totality* of relations between the critical values clarified by means of the associated groups of "caps," one alternative is to turn to nondegenerate functions (see §4) for which the critical points are isolated. Another alternative is to classify "caps" according to their "*spans*" e. It turns out very remarkably that on employing only caps of given fixed span greater than e, a consistent topological theory of critical values results which behaves formally as if F were a real analytic function for which the number of critical values $< c < 1$ is finite. An infinitely complex problem is thus reduced to an essentially finite problem.

It is convenient to suppose that $0 \leqq F \leqq 1$. If this were not the case, the functions

$$\frac{F}{1 + F}$$

could be used in place of F. Certain new terms needed here must be defined.

Let $F(p) = c < \infty$. The space S will be said to be *locally F-connected* of order r at p if corresponding to each positive constant e there exists a positive constant δ such that each singular r-sphere on the δ-neighborhood of p and on $S_{c+\delta}$ bounds an $(r + 1)$-cell of diameter at most e on S_{c+e}. We say that S is *F-reducible* at $c = 1$ if corresponding to any compact subset A of S there exists an F-deformation D^A of A into some subset S_c of S for which $c < 1$. The principal hypotheses in the span theory are then as follows.

(i) *The function F is positive and lower semi-continuous on S.*

(ii) *The sets S_a are compact for each $c < 1$, S is F-reducible at $c = 1$ and locally F-connected of all orders at each point x at which $F(x) < 1$.*

It remains to define *cap-spans*. Given a with $0 \leqq a < 1$, we say that a set A (for example, the compact carrier of a Vietoris cycle) lies definitely in S_a (written d-on S_a) if $A \subset S_{a-e}$ for some $e > 0$. The phrase d-mod S_a shall mean mod S_{a-e} for some $e > 0$.

k-cap-spans. Let u be a relative Vietoris k-cycle d-mod S_a, with a carrier $|u|$ in S_a. If $u \nsim 0$ on S_a, d-mod S_a, u is called a *k-cap* with *cap-height $a(u)$*. The boundary βu of such a k-cap is d-on $S_{a(u)}$. The cycle u is termed *linkable* or *nonlinkable* according as $\beta u \sim 0$ or $\beta u \nsim 0$, d-on $S_{a(u)}$. If u is linkable, set $\sigma(u) = \sup b$ for all b such that

$$u \nsim 0 \qquad [\text{on } S_b,\ d\text{-mod } S_{a(u)}] \tag{3.1}$$

and set

$$\text{span } u = \sigma(u) - a(u) \geqq 0. \tag{3.2}$$

If u is nonlinkable, set $\tau(u) = \inf b$ for all b such that

$$\beta u \sim 0 \qquad [d\text{-on } S_{a(u)},\ d\text{-mod } S_b] \tag{3.3}$$

and set

$$\text{span } u = a(u) - r(u) > 0. \tag{3.4}$$

Corresponding to each nontrivial homology class H there is a k-cycle $z \in H$ which is also a k-cap and for which the cap-height $a(z)$ is a minimum for all k-cycle-caps $\in H$. This k-cycle z will be termed *canonical.* Morse [3].

Recall that the group of coefficients which we are using is a field G. We shall be concerned with classes A of k-caps or k-cycles such that the conditions $u \in A$, $\delta \in G$, and $\delta \neq 0$, imply $\delta u \in A$. A "*maximal group*" of elements in A is a group B every element of which, except the null element, is in A, while B is a proper subgroup of no other such group of elements in A. With this understood, let $e > 0$ be given and fixed. We introduce maximal groups

M_k^e of k-caps with span greater than e,

N_k^e of nonlinkable k-caps with span greater than e,

P_k of canonical k-cycles.

If A stands for any of these three defining properties, it is a theorem in Morse [3] that any two maximal groups with property A are isomorphic, with corresponding elements u and u' such that $u - u'$ does not have property A. It is also shown in Morse [3] that the group N_k^e is isomorphic with βN_k^e where $u \in N_k^e$ corresponds to βu. It is remarkable that dim N_k^e is finite, and that P_k is also a maximal group of nonbounding k-cycles. We have the following fundamental theorem. Morse [3, Corollary 12.2].

THEOREM 3.1. *The maximal groups N_k^e can be so chosen that the direct sum*

$$N_k^e + \beta N_{k+1}^e + P_k \qquad (k = 0, 1, \cdots) \tag{3.3}$$

is a maximal group of k-caps with span greater than e.

It is easy to show that the cap-heights $a(u)$ of k-caps with span greater than e have at most the cluster point $a = 1$. Moreover a maximal group of k-caps with cap-height a and span greater than e always has a *finite* dimension. A maximal group of k-caps with span greater than e is seen to be the direct sum of maximal groups of such k-caps with the respective cap-heights. On setting

$$\dim M_k^e = m_k^e, \qquad \dim N_k^e = n_k^e, \qquad \dim P_k = p_k$$

we have the following corollary.

COROLLARY. $m_k^e - p_k = n_k^e + n_{k+1}^e \quad (k = 0, 1, \cdots)$.

The numbers p_k are of course the connectivities of S. One has the relations $m_k^e \geqq p_k$, and if one sets $E_k = m_k^e - p_k$ whenever $p_k < \infty$, and defines E_k as the right member in the corollary when $p_k = \infty$, the members E_k satisfy the infinite set of inequalities

$$E_n - E_{n+1} + \cdots + (-1)^n E_0 \geqq 0, \qquad (n = 0, 1, \cdots). \tag{3.4}$$

The preceding results hold for each $e > 0$. It is clear that m_k^e increases monotonically as e decreases. If F is an analytic function of the point on an analytic compact manifold, the numbers m_k^e are finite and independent of $e > 0$ for e sufficiently small. The span theory is a theory of *critical values*. It easily yields a theory of critical points and sets on adding the hypothesis of upper reducibility of F.

4. The nondegenerate theory Morse [4]. As distinguished from the causal theory, the objective here is to obtain the *totality of relations* between the critical points classified according to their indices. As distinguished from the span theory, the topological hypotheses that the critical points be nondegenerate makes it possible to treat *all the critical points together* rather than the generic subset of critical points associated with caps of span greater than e. The conditions on F are here necessarily more restricting, but there is a sense in which the nondegenerate F may be everywhere dense among all F admitted in the preceding sections. This has been established in important cases of considerable generality.

The *nondegenerate* theory owes much of its importance to the fact that it is through this theory that analysis extends topology as contrasted with the aid topology usually gives to analysis. The homology characters of Ω in (B) and (C) of §1 and of the symmetric square of the n-sphere were first obtained in this theory by a principle which we shall describe. Striking relations of this theory with homotopy theory have long been apparent and are now beginning to be explored. (See Morse [1, pp. 231–243] and Morse [6].)

The index. If there is just one nondegenerate critical point p at an F-level c, then as one passes from S_{c-e} to S_{c+e} for $e > 0$ sufficiently small only one homology group changes, and that by the addition of a k-cycle or subtraction of a $(k-1)$-cycle as a generator. We say that p then has the *index* k, and, if a k-cycle is added, that p is of *increasing* type. We state a fundamental theorem. Morse [1, p. 230] and Morse [4].

THEOREM 4. *If there exists on the abstract metric space S a nondegenerate function F all of whose critical points are of increasing type, then the k-homology group has a minimum base which includes just one k-cycle associated with each critical point of index k, and no other k-cycles.*

If f is a function of class C'' defined in an n-dimensional local coordinate system, an ordinary *differential critical point* p of f was termed nondegenerate if the Hessian H of f at p was nonvanishing; otherwise put, if no characteristic root of the determinant of H vanished. This generalizes for variational problems as follows. Given a critical extremal g in a variational boundary value problem, the Jacobi equations and the given boundary conditions give rise to a classical *characteristic value problem* associated with g. The critical extremal g is termed *nondegenerate* if and only if there is no vanishing characteristic value. The writer

has shown, Morse [1, p. 230], that the integral of length $J(P, Q)$, along curves joining two fixed points P and Q on a compact differential n-manifold M_n of class C''' without boundary, admits no degenerate extremals for any fixed P and almost all Q on M_n. It is in this sense that nondegenerate curve-functions $J(P, Q)$ are dense among all admissible functions $J(P, Q)$.

A general problem. It turns out that a closed extremal g of J on M_n is degenerate in the above sense if and only if the Jacobi equations based on g have no non-null periodic solution. The question arises, is it possible to give meaning and validity to the statement: "Among admissible manifolds of class C''', near M_n in a suitably restricted sense, those manifolds on which every closed extremal is nondegenerate are everywhere dense"? The writer has established such a theorem when $n = 2$, but the case $n > 2$ is open. More generally it should be possible in the case of variational problems in the large of general type to show that in some sense the nondegenerate function is everywhere dense.

The most useful principle of this sort is the following. On the above manifold M_n point functions f of class C'' which are nondegenerate are everywhere dense among functions f of class C''. This follows from work of the writer (cf. Morse [6]) and will be more explicitly elaborated and used in a later paper. The theory of nondegenerate functions parallels the theory of analytic functions in many remarkable ways.

Nondegeneracy topologically defined. Morse [4]. We shall start with a homotopic critical point p of F when F is a positive lower semi-continuous function on the metric space S. Suppose that $F(p) = c < \infty$. We shall be concerned with an *F-bounded* neighborhood U of p, that is, a neighborhood of p relative to some S_b for which $b > F(p)$. If D^t is a deformation of U on S with time parameter t, $0 \leqq t \leqq 1$, the *terminal mapping* of U into S at the time 1 is D^1. We shall refer to a topological image K_r in S of a euclidean r-disc. We take K_0 as a point.

DEFINITION D. *A homotopic critical point p of F will be termed nondegenerate if there exists a proper F-deformation D^t of some F-bounded neighborhood U of p such that*

(i). *D^t leaves p invariant and deforms U into a topological r-disc K_r which contains p as an interior point when $r > 0$, and on which $F(x) < F(p)$ when $x \neq p$.*

(ii). *The terminal mapping D^1, as applied to $K_r \cap U$, is F-deformable in K_r into the identity holding p fast.* Morse [4, p. 50].

It has been shown that an ordinary nondegenerate critical point of a point function f in a local n-dimensional coordinate system is nondegenerate in the above topological sense. See Morse [2, pp. 43–46]. The condition (ii) can be satisfied in the case of this f by choosing D^t so that the mapping in (ii) is the identity. It has also been shown that a critical extremal (an arc c) which is nondegenerate in the earlier sense of this section is also nondegenerate in our topological sense (Morse [4, p. 72]).

In Morse [4] the function F is termed nondegenerate if its homotopic critical points are topologically nondegenerate and finite in number below any finite F-level, and if certain F-deformations exist. All these conditions are topological.

The subscript r of the r-disc K_r appearing in the preceding definition is shown to be the index of the critical point. The homology theory used is the singular theory of Eilenberg. Let p_k be the kth connectivity of S and m_k the number of critical points of index k. Then $m_k \geqq p_k$ and there exist integers b_k, with $0 \leqq b_k \leqq \infty$, and $b_0 = 0$, such that

$$m_k - p_k = b_k + b_{k+1} \qquad (k = 0, 1, \cdots). \tag{4.1}$$

The numbers $E_k = m_k - p_k$, if finite, satisfy the relations (3.4).

Lacunary index sequences. One can obviously derive many properties of the connectivities p_k from the index sequence $I(F)$

$$m_0, m_1, m_2, \cdots. \tag{4.2}$$

If each integer $m_k \neq 0$ in (4.2) has vanishing adjacent integers, $I(F)$ will be termed *lacunary*. From (4.1) one obtains the new theorem:

THEOREM 4.1. *If* (4.2) *is a lacunary sequence, then* $m_k = p_k$ $(k = 0, 1, \cdots)$ *where* p_k *is the* k*th connectivity of* S.

An important use of this theory is the determination of the homology groups of the space $S = \Omega_M(P, Q)$ of sensed arcs joining two fixed points on a differentiable manifold M of class C'''. The homology groups are independent (up to isomorphisms) of the choice of P and Q so that one can take P and Q so that the length integral $F = J(P, Q)$ is nondegenerate. The index m_k is then the number of geodesics joining P to Q on which there are k conjugate points of P preceding Q. Cf. Morse [1, p. 229]. In the case of an n-sphere the index sequence is known to consist of zeros except that $m_k = 1$ when $k \equiv 0 \bmod (n-1)$. Morse [1, p. 247]. Hence in the case of the n-sphere $(n > 2)$ the connectivities p_k of $\Omega_M(P, Q)$ are zero except that $p_k = 1$ when $k \equiv 0 \bmod (n-1)$.

The question arises, what geometric manifolds admit a Riemann metric such that the nondegenerate length integral $J(P, Q)$ possess a lacunary sequence? There are infinitely many geometric manifolds with this property. In particular, the writer has shown in an unpublished paper that the cartesian product

$$S_{n_1} \times \cdots \times S_{n_r} = M \tag{4.3}$$

of any finite number of m-spheres with $n_i > 2$ admit such lacunary sequences. Thus a knowledge of the conjugate points of the geodesics joining P to Q on such M suffices to determine the homology groups of $\Omega_M(P, Q)$. This is consistent with, but not equivalent to, the theorem that the homology groups of $\Omega_M(P, Q)$ are obtained from the homology groups of $\Omega_{S_{n_i}}(P, Q)$, $i = 1, \cdots, r$, by the combinatorial processes usual for products. The latter theorem, proposed by the writer to Pitcher during the preparations of this report, was confirmed by Pitcher and later verified by the writer.

It is of interest to note that if $m_{n+1} = 0$, and if $E_i = m_i - p_i$ is finite for $i = 1, \cdots, n$, then

$$E_n - E_{n-1} + \cdots + (-1)^n E_0 = 0. \tag{4.4}$$

Thus the absence of geodesics g joining P to Q with $n + 1$ points on g conjugate to P implies (4.4).

5. A theorem on the dimension of a critical set. The results of this section will be published in detail in a later memoir. The principal theorem will be stated under much weaker conditions than the theorem on continuous functions on a manifold suggested at the end of §1. To this end, let F be defined over a metric space S, with

(i) *F positive, lower semi-continuous, and upper reducible,*

(ii) *the sets S_c compact for each $c < \infty$.*

Given a nonempty closed subset $A \subset S_c$, an infinite sequence

$$(D) = D_1, D_2, D_3, \cdots \tag{5.1}$$

of F-deformations will be regarded as *applicable* to A if D_1 is an F-deformation of A yielding a terminal image A_1 of A, if then D_2 is an F-deformation of A_1 yielding a terminal image A_2 of A_1 then, D_3 an F-deformation of A_2, etc. Let

$$\Delta_n = D_n D_{n-1} \cdots D_1 \tag{5.2}$$

be the resultant deformation of A, obtained on applying D_1 to A, D_2 to A_1, $\cdots$, D_n to A_{n-1}. Let T_n be the terminal transformation of A under Δ_n.

It is always possible to choose a sequence (5.1) of F-deformations applicable to A, together with a sequence (e_n) of numbers with $e_n > 0$ and $e_n \to 0$ as $n \uparrow \infty$, such that the following holds. If one sets

$$\lim_{n \uparrow \infty} [\sup_{x \in T_n(A)} F(x)] = v(A), \tag{5.3}$$

the set σ of homotopic critical points at the level $v(A)$ is not empty; if B_n is the subset of $T_n(A)$ on which $x > v(A) - e_n$, then B_n is not empty and

$$0 = \lim_{n \uparrow \infty} [\sup_{x \in B_n} d(x, \sigma)] \qquad (d = \text{distance}). \tag{5.4}$$

We suppose the sequence (5.1) and the sequence (e_n) so chosen, and term $v(A)$ an *F-barrier* of A.

We shall define an intrinsic property of a compact set A. Let r and n be integers with $0 < r < n$. We say that A is *(r, n)-admissible* if corresponding to an arbitrary closed subset $X \subset A$,

(1) *the bounding in A of each Vietoris $(n - r)$-cycle in X implies*

(2) *every Vietoris r-cycle in A is homologous in A to an r-cycle in $A - X$.*

The theory of manifolds contains explicit conditions for the existence of (r, n)-admissible sets A. It is clear, however, that A need not be restricted to manifolds. The theory of characters is involved. In this connection it should be recalled that our group of coefficients is a finite field. The fundamental theorem follows:

THEOREM 5.2. *Let A be a closed (r, n)-admissible subset of S_c with an F-barrier $v(A)$ and such that any $(n - r)$-cycle in A which bounds in S_c bounds in A. Let*

H be a nontrivial r-homology class in A. If $v(A) = c(H)$, *then the set of homotopic critical points at the level* $v(A)$ *carries an* $(n - r)$*-Vietoris cycle which is non-bounding in* S_c, *so that*

$$\dim \sigma \geqq n - r.$$

The proof of this theorem makes no use of category. In general, it seems to be possible to obtain many results on the minimum number of critical points which have been obtained by a use of the category theory without using that theory, and to add to these results a causal relation between various homology, homotopy, and isotopy classes and the respective critical points.

It should be noted that the critical point theory suggests many other numerical topological invariants in addition to the category: for instance, the minimum number N of *nondegenerate* critical points of a continuous nondegenerate function f defined over a geometric manifold, as f ranges over all such functions. If R is the minimum number of *isolated* homotopic critical points of a continuous function f defined over a geometric manifold, as f ranges over all such functions, it is clear that $N \geqq R$. Both N and R are topological invariants over geometric manifolds for which N is defined. When is $N > R$?

6. Other advances. I shall refer first of all to the unpublished work of E. Pitcher which makes use of the mechanism (Morse [1, pp. 244–247]) whereby the homology groups of the space $\Omega(P, Q)$ of §4 were determined for the space of curves joining P to Q on an n-sphere by explicitly giving a base for the non-bounding cycles. Pitcher's work makes it clear that these models will be useful in analyzing the homology groups $\pi_r(S^n, x)$ of the n-sphere. Here x is the point in S^n into which the fixed point of the antecedent r-sphere S^r is mapped. In results announced at the Congress, Pitcher has used these new geometric methods to verify the result of Whitehead that $\pi_5(S^3)$ yields the integers mod 2. The variational methods are capable of great extension in the direction of determining models for use in homotopy theories. For example, one can replace curves joining two fixed points on an n-sphere by disc-type surfaces spanning a circle.

The papers by Morse and Ewing introduce a new approach to the restricted problem of three bodies. The Jacobi least action integral J which is studied is neither regular nor positive definite. Nevertheless Morse and Ewing have established the upper reducibility of J under suitable conditions and prepared for the advances to follow. Ewing has used the Weierstrass generalized integral to give an essential simplification of one of the proofs. In the general direction of fundamental definition of integral and length see Menger, *What Paths Have Length?*

Special attention is called to the remarkable work of the Russian school. The recently published paper by L. Lusternik and Schnirelmann [4] gives more detail concerning early results and continues their program. The paper by Seifert cited

refers to the general problem of periodic motion and runs into the same problem of one motion covering another.

Reference should be made to the forthcoming Colloquium Lectures, American Mathematical Society, New York, by Arnold Hedlund where his researches on flow, transitivity, symbolic dynamics, etc., have much to do with variational theory. A paper by Rauch to appear in the Annals of Mathematics uses variational theory in the large, and in particular uses generalized comparison theorems to obtain sufficient conditions on the variation in ratio of the Riemann curvature on a compact simply-connected manifold of positive curvature in order that the manifold be the topological image of an n-sphere.

A fundamental paper by C. B. Morrey first solves the problem of Plateau for a general Riemannian manifold in the case where the manifold is not coverable by a single coordinate system. This result should accelerate variational theory in the large for multiple integrals. In this direction is the penetrating work of Shiffman who has attacked the crucial problems of the multiple integral theory with great ingenuity and success. Courant has aided the general advance by his papers on minimal surfaces and conformal mapping. His book contains other references.

In a basic topological study Leray has initiated a theory of mappings which embraces part of the critical point theory and suggests unsuspected relations. One may expect striking developments along this line in the near future.

The extensive work of McShane and L. C. Young is in another direction but has introduced new power and completeness into the foundations of the theory of generalized curves and surfaces.

The recent work of Morse and Transue gives an abstract representation of a generalization of the second variation. The generalized Euler equations include classical Euler equations as well as integral and integro-differential equations of general type. The relation to the variational theory in the large is in connection with the unpublished characteristic value theory and index theory.

S. Bergman has made use of the theory of level manifolds and of critical points in his study of pseudo-conformal mapping. In particular the theory of equivalence of Reinhardt domains clearly calls for such analysis.

Bibliography

S. Bergman

Sur la fonction noyau d'un domaine et ses applications dans la théorie des transformations pseudo-conformes, Mémorial des Sciences Mathématiques, vol. 108, Paris, Gauthier-Villars, 1948.

R. Courant

Dirichlet's principle, conformal mapping, and minimal surfaces, New York, Interscience Publishers, Inc., 1950.

S. Eilenberg

Singular homology theory, Ann. of Math. vol. 45 (1944) pp. 407–449.

L. Elsholz

1. *Zur Theorie der Invarianten, die zur Bestimmung der unteren Grenze der Anzahl der kritischen Punkte einer stetigen Funktion, die auf einer Mannigfaltigkeit bestimmt ist, dienen können*, Rec. Math. (Mat. Sbornik) N.S. vol. 5 (1939) pp. 551–558 (in Russian).

2. *Die Länge einer Mannigfaltigkeit und ihre Eigenschaften*, Rec. Math. (Mat. Sbornik) N.S. vol. 5 (1939) pp. 565–571 (in Russian).

J. Ewing
Minimizing an integral on a class of continuous curves, Duke Math. J. vol. 10 (1943) pp. 471–477.

R. Fox
On the Lusternik-Schnirelmann category, Ann. of Math. vol. 42 (1941) pp. 333–370. A bibliography for the category theory is found here.

S. Froloff and L. Elsholz
Limite inférieure pour le nombre des valeurs critiques d'une fonction, donnée sur une variété, Rec. Math. (Mat. Sbornik) vol. 42 (1935) pp. 637–643.

T. Gordon
On the minimal number of critical points of a real function defined on a manifold, Rec. Math. (Mat. Sbornik) N.S. vol. 4 (1938) pp. 105–113.

T. L. Kelley and E. Pitcher
Exact homomorphism sequences in homology theory, Ann. of Math. (2) vol. 48 (1947) pp. 682–709.

J. Leray
L'anneau spectral et l'anneau filtré d'homologie d'un espace localement compact et d'une application continue, J. Math. Pures Appl. vol. 29 (1950) pp. 2–139.

L. Lusternik
On the number of solutions of a variational problem, C. R. (Doklady) Acad. Sci. URSS N.S. vol. 40 (1943) pp. 215–217.

L. Lusternik and L. Schnirelmann
1. *Existence de trois lignes geodésiques fermées sur toute surface de genre* 0, C. R. Acad. Sci. Paris vol. 188 (1929) pp. 534–536; vol. 189 (1929) pp. 269–271.
2. *Méthodes topologiques dans les problèmes variationnels*, Institute for Mathematics and Mechanics Moscau, 1930, (in Russian).
3. *Méthodes topologiques dans les problèmes variationnels*, Paris, Hermann, 1934.
4. *Topological methods in variational problems and their application to the differential geometry of surfaces*, Uspehi Matematičeskih Nauk N.S. vol. 2 no. 1 (17) (1947) pp. 166–217 (in Russian).

E. J. McShane
Generalized curves, Duke Math. J. vol. 6 (1940) pp. 513–536.

K. Menger
What paths have length?, Fund. Math. vol. 36 (1949) pp. 109–118.

C. B. Morrey
The problem of Plateau on a Riemannian manifold, Ann. of Math. (2) vol. 49 (1948) pp. 807–851.

M. Morse
1. *The calculus of variations in the large*, Amer. Math. Soc. Colloquium Publications vol. 18, New York, 1934.
2. *Functional topology and abstract variational theory*, Mémorial des Sciences Mathématiques vol. 92, Paris, Gauthier-Villars, 1939.
3. *Rank and span in functional topology*, Ann. of Math. vol. 41 (1940) pp. 419–454.
4. *A positive lower semi-continuous non-degenerate function on a metric space*, Fund. Math. vol. 35 (1948) pp. 47–78.
5. *Sur le calcul des variations*, Bull. Soc. Math. France (1939).
6. *The analysis and analysis situs of regular n-spreads in* $(n + r)$*-space*, Proc. Nat. Acad. Sci. U.S.A. vol. 13 (1927) pp. 813–817.

M. Morse and M. Heins
Causal isomorphisms in the theory of pseudoharmonic functions, Ann. of Math. vol. 46 (1945) pp. 600–624.

M. MORSE and G. EWING
1. *The variational theory in the large including the non-regular case*, First paper, Ann. of Math. vol. 44 (1943) pp. 329–353.
2. *The variational theory in the large including the non-regular case*, Second Paper, Ann. of Math. vol. 44 (1943) pp. 354–374.

M. MORSE and C. TOMPKINS
1. *The existence of minimal surfaces of general critical types*, Ann. of Math. vol. 40 (1939) pp. 443–472. (Corrections for this paper appear in the Ann. of Math. vol. 42 (1941) p. 331.
2. *Minimal surfaces of unstable type by a new mode of approximation*, Ann. of Math. vol. 42 (1941) pp. 62–72.
3. *Unstable minimal surfaces of higher topological structure*, Duke Math. J. vol. 8 (1941) pp. 350–375.
4. *The continuity of the area of harmonic surfaces as a function of the boundary representations*, Amer. J. Math. vol. 63 (1941) pp. 825–838.

M. MORSE and W. TRANSUE
A characterization of the bilinear sums associated with the classical second variation, Annali di Matematica Pura ed Applicata vol. 28 (1949). References to ten related papers by Morse and Transue can be found here.

H. SEIFERT
Periodische Bewegungen mechanischer Systeme, Math. Zeit. vol. 52 (1945) pp. 197–216.

H. SEIFERT and W. THRELFALL
Variationsrechnung im Grossen, Berlin, Teubner, 1938.

M. SHIFFMAN
1. *The problem of Plateau for non-relative minima*, Ann. of Math. vol. 40 (1939) pp. 834–854.
2. *Unstable minimal surfaces with several boundaries*, Ann. of Math. vol. 43 (1942) pp. 197–222.
3. *Unstable minimal surfaces with any rectifiable boundary*, Proc. Nat. Acad. Sci. U.S.A. vol. 28 (1942) pp. 103–108.
4. *Instability for double integral problems in the calculus of variations*, Ann. of Math. vol. 45 (1944) pp. 543–576.

L. C. YOUNG
Generalized surfaces in the calculus of variations, Ann. of Math. vol. 43 (1942) pp. 530–554.

INSTITUTE FOR ADVANCED STUDY,
PRINCETON, N. J., U. S. A.

SINGULARITIES OF MINIMAL SURFACES

LIPMAN BERS

1. Introduction. Minimal surfaces are of interest in various branches of mathematics. In the calculus of variations they appear as surfaces of least area, in differential geometry as surfaces of vanishing mean curvature. In gas dynamics the equation of minimal surfaces,

$$(1) \qquad (1 + \varphi_y^2)\varphi_{xx} - 2\varphi_x\varphi_y\varphi_{xy} + (1 + \varphi_x^2)\varphi_{yy} = 0,$$

is interpreted as the potential equation of a hypothetical gas, which yields flows closely approximating adiabatic flows of low Mach number. This interpretation, due to Chaplygin,[1] has been used extensively in recent aerodynamical literature. In the general theory of partial differential equations, finally, equation (1) appears as the simplest nonlinear equation of elliptic type. From the point of view of this theory it is natural to restrict ourselves to minimal surfaces in three-space admitting a non-parametric representation.

In this report we deal with solutions $\varphi(x, y)$ of (1) possessing isolated singularities. The kind of results one should look for is suggested by the beautiful theorem by S. Bernstein which asserts that *every solution of* (1) *defined and regular*[2] *for all finite values of* z ($z = x + iy$) *is a linear function.*[3] For harmonic functions the same conclusion would hold only under the additional assumption that $\varphi_x^2 + \varphi_y^2$ is uniformly bounded. This situation is typical. Contrary to what one might expect a priori, the theory of singularities for equation (1) is simpler than that for the Laplace equation or for any linear elliptic equation.

We shall consider both single-valued and multiple-valued solutions $\varphi(x, y)$, but make once and for all the assumption that φ_x and φ_y are at most finitely many-valued.

2. Classification of singularities.[4] The following result bears the same relation to Riemann's theorem on removable singularities as Bernstein's theorem does to that of Liouville.

[1] S. A. Chaplygin, Učenye Zap. Imp. Mosk. Univ., Sek. Mat. Fiz. vol. 21 (1902) pp. 1–121. An English translation of this celebrated paper appeared as the National Advisory Committee for Aeronautics, Technical Memoir vol. 1063 (1944).

[2] By a regular solution we mean one which possesses continuous partial derivatives of second order. Such solutions are necessarily analytic.

[3] Bernstein obtained this result as a corollary of a general geometrical theorem (cf. Comm. Soc. Math. Kharkov vol. 15 (1915–1917) pp. 38–45, or Math. Zeit. vol. 26 (1927) pp. 551–558). A topological gap in his proof has been corrected by E. Hopf (Proc. Amer. Math. Soc. vol. 1 (1950) pp. 80–85) and by E. J. Mickle (ibid. pp. 86–89). T. Radó (Math. Zeit. vol. 26 (1927) pp. 551–558) gave a function-theoretical proof of the theorem on minimal surfaces. Another such proof is given in the author's paper *Isolated singularities of minimal surfaces* which will appear in the Ann. of Math.

[4] All theorems of this section are proved in the author's paper referred to in footnote 3.

(a) *A solution $\varphi(x, y)$ of* (1) *which is regular and single-valued in a deleted neighborhood of $z_0 = x_0 + iy_0 \neq \infty$ is also regular at z_0.*

For harmonic functions the same conclusion would hold only under the additional assumption that φ is bounded.

From now on we consider also multiple-valued solutions. We say that $z_0 \neq \infty$ ($z_0 = \infty$) is a branch-point of order $(m - 1)$ if $\varphi(x, y)$ is regular for $0 < |z - z_0| < R$ (for $R < |z| < +\infty$), and φ_x and φ_y are m-valued functions.

(b) *If z_0 is a branch-point of order $(m - 1)$ of a solution $\varphi(x, y)$ of* (1), *then the limit*

$$w_0 = \lim_{z \to z_0} (\varphi_x - i\varphi_y)$$

exists. This limit is necessarily finite if $z_0 = \infty$.

If $\varphi(x, y)$ is not a linear function (a trivial case which shall be neglected), then $\varphi_x - i\varphi_y \neq w_0$ at all points sufficiently close to and distinct from z_0. Hence if Γ is a simple closed curve sufficiently close to z_0 and containing z_0 in its interior, and if $z = x + iy$ goes m times around Γ in the direction positive with respect to z_0, the argument of $\varphi_x - i\varphi_y$ increases by $2n\pi$, the integer n being independent of Γ. We call n the index of the branch-point.

(c) *Under the hypothesis of* (b), $n > 0$ *if* $w_0 = 0$, $n = 0$ *if* $0 < |w_0| < +\infty$, $-m \leqq n < 0$ *if* $w_0 = \infty$.

Branch-points with vanishing or positive index are called ordinary, those of negative index polar.

(d) *Let z_0 be an ordinary branch-point of a solution $\varphi(x, y)$ of* (1) *of order $(m - 1)$ and index n. If $z_0 \neq \infty$, then*

$$\varphi(x, y) = \mathbf{Re}\{A + B(z - z_0)^{1+n/m}\} + O(|z - z_0|^{1+(n+1)/m}),$$

$$\varphi_x - i\varphi_y = \left(1 + \frac{n}{m}\right) B(z - z_0)^{n/m} + O(|z - z_0|^{(n+1)/m}), \qquad z \to z_0,$$

where A and $B \neq 0$ are complex constants. If $z_0 = \infty$, then

$$\varphi(x, y) = \mathbf{Re}\{A + Bz^{1-n/m} + C \log z\} + O(|z|^{1-(n+1)/m}),$$

$$\varphi_x - i\varphi_y = \left(1 - \frac{n}{m}\right) Bz^{-n/m} + (C/z) + O(|z|^{-(n+1)/m}), \qquad z \to \infty,$$

where A, B, C are complex constants and $B \neq 0$, $C = 0$ for $n > m$, $C \neq 0$ for $n = m$.

Introduce the auxiliary variable $Z = (z - z_0)^{1/m}$ if $z_0 \neq \infty$, $Z = z^{-1/m}$ if $z_0 = \infty$. In the Z-plane φ behaves like a regular harmonic function (in a neighborhood of $Z = 0$) if $z_0 \neq \infty$, like a harmonic function which may have a pole and a logarithmic singularity if $z_0 = \infty$. This justifies the name "ordinary branch-point." In the case of a polar branch-point, however, $\varphi(x, y)$ is, in general, topologically distinct from a harmonic function.

(e) *Let z_0 be a polar branch-point of order $(m - 1)$ and index n of a solution $\varphi(x, y)$ of* (1). *Set $Z = (z - z_0)^{1/m}$ and consider the functions φ, $\varphi_x^2 + \varphi_y^2$ in the*

Z-plane. The level-curves of $\varphi_x^2 + \varphi_y^2$ *are simple closed curves around* $Z = 0$. *The level-curves of* φ *form a pattern*[5] *shown in Fig.* 1, *the number of "hyperbolic sectors" being exactly* $2(m + n)$.

If φ_x and φ_y are single-valued in the neighborhood of a polar branch-point, then $m = 1$, so that $n = -1$ and there are no "hyperbolic sectors." In this case $\varphi(x, y)$ behaves like the function $\varphi = \arctan(y/x)$. In view of the gas-dynamical interpretation of equation (1) a polar branch-point of order 0 may be called a vortex-point. We note that in general $\varphi(x, y)$ may be either m-valued or infinitely many-valued in the neighborhood of a branch-point z_0 of order $(m - 1)$ except that it must be m-valued if $z_0 \neq \infty$ and $n \geqq 0$, or if $z_0 = \infty$ and $n > m$, and it must be infinitely many-valued if $z_0 \neq \infty$ and $n = -m$.

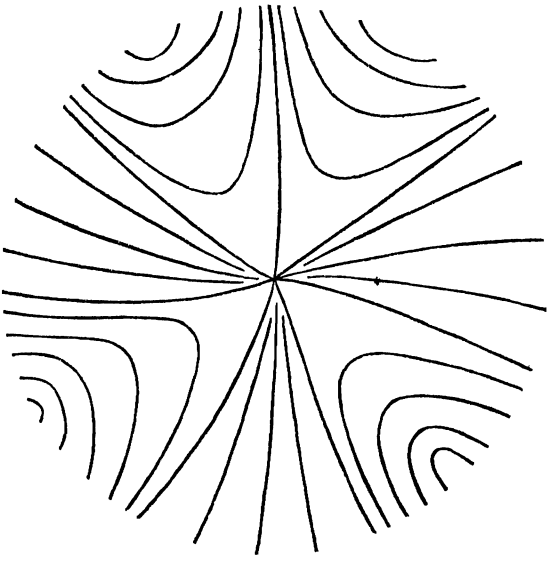

FIGURE 1

Solutions of (1) with finitely many-valued derivatives can have no isolated singularities other than those described above.[6]

3. Abelian minimal surfaces.[7] The local theory of isolated singularities given above may serve as a foundation for the theory in the large. It is natural to begin with an inquiry into the nature of non-parametric minimal surfaces with only a finite number of singularities. More precisely, we want to find all solutions $\varphi(x, y)$ of (1) such that (α) $\varphi(x, y)$ may be continued analytically along every finite path in the z-plane which avoids certain excluded points $z_0, \cdots, z_K$,

[5] A precise description of this pattern is given in the paper referred to above. The level-curves which enter $Z = 0$ possess there tangents. Each "hyperbolic segment" has an "opening" π/m. For a given m and n there exist, in general, several topologically distinct types of polar branch-points, since these parameters determine only the number but not the relative position of the "hyperbolic sectors". For $m = 3$, $n = -1$, for instance, there are 5 distinct types, one of which is shown in Fig. 1.

[6] If one asks for point-singularities of the analytic functions giving the Monge-Weierstrass representation of a minimal surface, rather than for the point-singularities of the solutions of (1), one arrives at results of a quite different nature. Such an investigation was carried out by Y. W. Chen (Ann. of Math. vol. 65 (1949) pp. 790–806).

[7] A paper with the same title, containing proofs of the statements made in this section, will appear elsewhere.

and (β) $\varphi_x - i\varphi_y$ is a finitely many-valued, say M-valued, function. From (β) it follows that (γ) along all paths leading to the excluded points or to the point at infinity $\varphi_x - i\varphi_y$ approaches finite or infinite limits.

If a harmonic function $\varphi(x, y)$ has properties (α), (β), (γ), it is the real part of an Abelian integral (integral of an algebraic function). We therefore call a minimal surface $\mathbf{S}: \varphi = \varphi(x, y)$, in the (x, y, φ)-space an Abelian surface if $\varphi(x, y)$ satisfies conditions (α), (β).

We state now a generalization of Bernstein's theorem.

(f) *An Abelian minimal surface is a plane if $\varphi_x - i\varphi_y$ omits any finite value, or if there are no polar branch-points.*

From now on we assume that $\mathbf{S}$ is not a plane. Over the points $z_1, \cdots, z_K, \infty$ the function $\varphi(x, y)$ has branch-points of the kind described in the preceding section.[8] We denote by B the sum of the orders of all branch-points, by N the sum of the indices of the ordinary branch-points, and by P the number of polar branch-points. Upon proper definition of terms Theorem (f) can be strengthened as follows.

(g) *If $\mathbf{S}: \varphi = \varphi(x, y)$ is an Abelian minimal surface, then $\varphi_x - i\varphi_y$ takes on every finite value N times, and the sum of the indices of all polar branch-points is $(-N)$.*

The totality of Abelian minimal surfaces is described in the following statement.

(h) *To every Abelian minimal surface $\mathbf{S}: \varphi = \varphi(x, y)$, there corresponds a uniquely determined real plane algebraic curve $\mathfrak{C}$ of genus*

$$p = B - 2M + P + 1$$

in the (s, t)-plane, with P real branches, and with no points with real s and imaginary t. The surface $\mathbf{S}$ admits the parametric representation

$$(2) \quad x = \mathbf{Re} \int \frac{2s}{1 + s^2}\, it\, ds, \qquad y = \mathbf{Re} \int \frac{1 - s^2}{1 + s^2}\, it\, ds, \qquad \varphi = \mathbf{Re} \int t\, ds$$

in terms of Abelian integrals attached to $\mathfrak{C}$. The regular points and ordinary branch-points of $\varphi(x, y)$ are in a one-to-one correspondence with the points of $\mathfrak{C}$ with $\mathbf{Im}\, s > 0$; the polar branch-points are in a one-to-one correspondence with the real branches of $\mathfrak{C}$.

The converse of this theorem is also true.

(i) *Let $\mathfrak{C}$ be a real plane algebraic curve in the (s, t)-plane such that* (i) *for every point (s, t) on $\mathfrak{C}$ the reality of s implies that of t,* (ii) *the functions x, y defined by* (2) *are single-valued on the part of the Riemann surface $\mathbf{F}$ of $\mathfrak{C}$ where $\mathbf{Im}\, s \geqq 0$. Then the surface $\mathbf{S}$ defined by* (2) *is an Abelian minimal surface.*

A curve $\mathfrak{C}$ satisfying the hypotheses of this theorem will be said to belong to the class $\mathfrak{A}$. If two Abelian minimal surfaces equivalent under a translation are considered as identical, then the correspondence between Abelian minimal

[8] Several branch-points of different types may be located over the same point in the z-plane.

surfaces and algebraic curves of class $\mathfrak{A}$ is one-to-one. The Riemann surface $\mathbf{F}$ of a curve $\mathfrak{C}$ of class $\mathfrak{A}$ is necessarily orthosymmetric. It can be shown that to every orthosymmetric Riemann surface belong infinitely many curves of class $\mathfrak{A}$ and hence infinitely many Abelian minimal surfaces.

4. Boundary value problems. The discussion of minimal surfaces with only isolated singularities reduces completely to a problem in algebraic function theory. The theory of solutions of (1) which possess at given points prescribed singularities and satisfy along given curves prescribed boundary conditions presents greater difficulties. Thus far, only one class of such boundary value problems has been treated successfully.

Let Γ be a given simple closed curve in the z-plane. We want to find a solution $\varphi(x, y)$ of (1) which is regular at all finite points exterior to Γ, possesses single-valued partial derivatives which are continuous on Γ, and satisfies on Γ either of the homogeneous boundary conditions

$$\text{A: } \frac{\partial \varphi}{\partial n} = 0, \qquad \text{B: } \varphi = 0,$$

as well as the condition

$$\max (\varphi_x^2 + \varphi_y^2) = \mu^2 \tag{3}$$

where μ is a given number. At $z = \infty$ the function $\varphi(x, y)$ will have a singularity, and according to (d) we shall have that

$$\varphi(x, y) = ax + by + c \log (x^2 + y^2)^{1/2} + d \arctan (y/x) + O(1),$$

$$\varphi_x - i\varphi_y \to a - ib, \qquad z \to \infty ,$$

where a, b, c, d are real constants. In order to obtain a well-defined problem we must specify the character of the singularity. This can be done, for instance, by means of either of the following conditions.[9]

1. $a > 0, b = 0, \varphi_x - i\varphi_y = 0$ at a given point z_1 of Γ.
2. $a > 0, b = c = d = 0$.
3. $a = b = 0$.
4. $c = d = 0, \varphi_x - i\varphi_y = 0$ at a given point z_1 of Γ.
5. $\varphi_x - i\varphi_y = 0$ at two given points, z_1 and z_2 , of Γ.

Problems involving boundary condition A are best interpreted in the language of gas dynamics. Problem A1, for instance, requires the determination of a Chaplygin flow past the profile Γ, possessing a given direction at infinity, attaining a given maximum speed (condition (3)), and satisfying at the "trailing edge" z_1 the Kutta-Joukowsky condition. Problem A3 involves a purely circulatory flow around Γ. Problems with boundary condition B are best interpreted geometrically. Thus Problem B3 is that of finding a one-sheeted minimal surface

[9] These conditions are suggested by analogy with Laplace's equation. Note that condition A implies that $c = 0$, and condition B that $d = 0$.

bounded by a plane curve, extending to infinity and becoming there parallel to the plane of the curve, and possessing a given maximum slope (condition (3)).

We make the following assumptions on the curve Γ. (i) Γ possesses a piecewise continuous curvature. (ii) Γ has at most finitely many intruding corners, and no protruding ones, except that the point z_1 in Problem 1 and 4 and the points z_1, z_2 in Problem 5 may be protruding corners. (iii) If a point goes once around Γ in the counter-clockwise direction, the tangent to Γ is turned in the clockwise direction by less than π. Under these assumptions we can prove:

(j) *All ten problems*, A1 — B5, *have solutions.*

Condition (i) can be somewhat weakened. Condition (ii) is actually necessary for the existence of a solution. We do not know whether or not condition (iii) is essential. The proof of (j) is rather involved. The boundary value problem is first reduced to a mapping problem and then to a nonlinear integro-differential equation which is treated by the topological method due to Leray and Schauder.[10] A uniqueness proof is still lacking.

Problems 1, 2, 4, 5 can be modified by replacing condition (3) by

$$(3^*) \qquad a = \mu_0 ,$$

μ_0 being a given positive number, i.e., by prescribing the "speed at infinity" instead of the "maximum speed." Some such problems have been discussed by several authors. Thus Frankl and Keldysh[11] showed that Problem A1* has a unique solution if μ_0 is sufficiently small. They considered an equation more general than (1). Slioskin[12] proved that Problem A2* has a unique solution for symmetrical profiles Γ and for sufficiently small μ_0. Chen[13] observed that a solution of A2* also solves a certain variational problem, and proved that this variational problem has a solution.

There are, of course, many other boundary value problems for minimal surfaces with singularities. Thus it would be of interest to know whether there exists a solution of (1) which is regular in the domain interior to a smooth closed curve, except for a single vortex-point, and which satisfies along the curve condition A. Such a solution would be the minimal surface analogue of Green's function.

5. Other partial differential equations. The fact that solutions of (1) are relatively poor in isolated singularities is due to the nonlinearity of this equation.

Indeed, let $\mathfrak{L}(\varphi) = 0$ be any linear homogeneous elliptic partial differential equation of second order for an unknown function of two independent variables.

[10] A sketch of the proof of the existence theorem for A1 will be found in the author's note, Proceedings of Symposia in Applied Mathematics vol. 1 (1949) pp. 41–46. This note also contains the proof of the necessity of condition (ii). A detailed presentation will appear elsewhere, presumably in the Trans. Amer. Math. Soc.

[11] F. Frankl and M. Keldysh, Bull. Acad. Sci. URSS vol. 12 (1934) pp. 561–601.

[12] N. Slioskin, Učenye Zapiski Moskovskogo Gosudarstvennogo Universiteta vol. 7 (1937) pp. 43–69.

[13] Y. W. Chen, Trans. Amer. Math. Soc. vol. 65 (1949) pp. 331–347.

This equation possesses solutions which behave (in a sense which we shall not make precise here) like the harmonic functions $\mathrm{Re}(z - z_0)^{n/m}$, $\mathrm{Im}(z - z_0)^{n/m}$, $m = 1, 2, \cdots ; n = 0, \pm 1, \pm 2, \cdots$. From these, poles and essential singularities can be built up by superposition. The situation is not more complicated, however, than in the case of harmonic functions, since every solution which is regular and m-valued in the neighborhood of z_0 can be uniquely expanded in a "Laurent series" in terms of the particular solutions mentioned above.

These theorems are consequences of a general theory of *"pseudo-analytic"* functions of which only a brief sketch dealing with self-adjoint equations has appeared thus far.[14] We can establish these results without restrictive regularity assumptions on the equation $\mathfrak{L}(\varphi) = 0$. If it is of the form

$$\varphi_{xx} + \varphi_{yy} + a(x, y)\varphi_x + b(x, y)\varphi_y = 0, \tag{4}$$

for instance, we require only that a and b be Hölder-continuous functions.

If $\mathfrak{L}(\varphi) = 0$ is an equation with analytic coefficients, then the existence of poles follows from the theory of the fundamental solution. For an analytic elliptic equation of any order in n-space the construction of the fundamental solution and an analysis of poles has been accomplished by John.[15] For the case of equation (4) with analytic a, b a "Laurent expansion" seems to be contained in the work of Vekua.[16]

On the other hand, the equation of minimal surfaces may well turn out to be typical for a wide class of nonlinear equations of the form

$$(\rho\varphi_x)_x + (\rho\varphi_y)_y = 0, \qquad \rho = \rho(q), \qquad q^2 = \varphi_x^2 + \varphi_y^2 . \tag{5}$$

Such an equation may be interpreted as the potential equation of a gas flow with density ρ. It will be elliptic ("subsonic") if $q\rho(q)$ is an increasing function of q. We assume this to be true for all positive values of q and set $L = \lim_{q\to+\infty}(\rho q)$. (For equation (1), $\rho = (1 + q^2)^{-1/2}$, $L = 1$.) Equation (5) possesses a solution of the form $\varphi(x, y) = \Phi(r)$, $r^2 = x^2 + y^2$ (potential of a source). If $L < +\infty$, then $\Phi(r)$ becomes singular for a positive value of r. I would conjecture that if $L < +\infty$, then Bernstein's theorem and Theorem (a) are valid for equation (5). But it is easy to find functions $\rho(q)$ with $d[q\rho(q)]/dq > 0$ and $L = +\infty$, such that $\Phi(r)$ is continuous for $r = 0$. Hence there exist elliptic equations of the form (5) for which even Riemann's theorem on removable singularities ceases to be true. Such equations possess solutions which may remain continuous at isolated singularities.

Note added in proof (July 18, 1951): 1. I am now able to prove the existence theorem (j) without assuming that the boundary curve Γ is nearly convex (that is, without condition (iii)).

2. R. S. Finn proved[17] the theorem on removable singularities (Theorem (a))

[14] L. Bers, Proc. Nat. Acad. Sci. U.S.A. vol. 36 (1950) pp. 130–136.

[15] F. John's paper will appear in Comm. Pure Appl. Math.

[16] See a revue of N. Vekua's work in Uspehi Matematičeskih Nauk 5 (1950) pp. 167–169.

[17] R. S. Finn, Bull. Amer. Math. Soc. Abstract 57–3–212.

for all elliptic equations of the form (5) with $L < +\infty$, and even for the more general case when ρ depends also on x and y.

3. The papers referred to in footnotes 2, 7, 10, 15 have appeared[18] and so has a sketch of the general theory of pseudo-analytic functions.[19]

SYRACUSE UNIVERSITY,
SYRACUSE, N. Y., U. S. A.
INSTITUTE FOR ADVANCED STUDY,
PRINCETON, N. J., U. S. A.

[18] L. Bers, Ann. of Math. vol. 53 (1951) pp. 364–386; Journal d'Analyse Mathématiques (Jerusalem) vol. 1 (1951) pp. 43–58; Trans. Amer. Math. Soc. vol. 70 (1951) pp. 465–491. F. John, Comm. Pure Appl. Math. vol. 3 (1951) pp. 273–304.

[19] L. Bers, Proc. Nat. Acad. Sci. U.S.A. vol. 37 (1951) pp. 42–47.

GEOMETRIC AND POTENTIAL-THEORETICAL METHODS IN THE THEORY OF FUNCTIONS OF SEVERAL COMPLEX VARIABLES[1]

STEFAN BERGMAN

1. The kernel function as a tool for defining an invariant metric. A generalization of the principle of hyperbolic measure. Within the past two decades or so, attempts have been made to reformulate the methods of the theory of functions of one variable[2] so that they may be applied to other fields such as the theory of functions of several variables, differential geometry, and partial differential equations, to obtain unified methods in all these fields.

In this connection, it was particularly useful to develop certain new tools which arose from the study of the class $L^2(\mathfrak{B})$ of functions in the domain[3] $\mathfrak{B}$. Using some properties of functions of this class, we determine for every domain $\mathfrak{B}$, the so-called kernel function which has various useful properties. Connecting this approach with certain variational problems, the "principle of the minimum integral" was obtained. Using these procedures, results have been derived which, in the classical approach, are usually obtained by employing the Riemann mapping theorem and the principle of hyperbolic measure. In this new form, the methods can be generalized to the theory of several variables. [Bergman 1, 3, 4].[4]

2. Complications arising in attempting to generalize potential-theoretical methods. A large chapter in the theory of functions of one variable is based on potential-theoretical methods. In the case of one variable, this direction has been connected with the theory of the kernel function. The establishment of this connection leads to a series of new results for one variable, as well as making it possible to generalize these procedures to the theory of partial differential equations of elliptic type [Bergman 4, Bergman and Schiffer {1, 2, 3, 6, 7}].

[1] Paper done under Contract N5ori 76/16 NR 043 046 with the Office of Naval Research.

[2] Unless otherwise indicated, here and in the following, "variables" will be used to mean complex variables. As a rule, we shall use z to denote (z_1, z_2), $z_k = x_{2k-1} + ix_{2k}$, when referring to a function of two variables. Z will denote (x_1, x_2, x_3, x_4). For the sake of simplicity, we shall formulate our results for two variables, but almost all considerations can easily be generalized to the case of n variables.

[3] As a rule, manifolds are designated by German or script characters. The superscript n, where $n = 3, 2, 1, 0$, indicates the dimension of the manifold. In the case of 4-dimensional manifolds, the superscript is omitted. By surface, we shall mean a two-dimensional variety; the term hypersurface will be used for three-dimensional varieties. A sequence of intersections of a domain with distinguished boundary surface is discussed in another article in these Proceedings, on p. 363ff. Note: Sometimes the same manifolds are denoted in the present paper and in the article on p. 363ff by different symbols.

[4] Within the brackets, the numbers following a name refer to the bibliography at the end of the paper unless they appear within ( ) or{ }, in which case they refer to the bibliography in [Bergman 3 or 4], respectively.

Attempting to generalize the potential theoretical methods to the case of two variables, we meet difficulties which largely arise from the following two factors:

(1) In contrast to the situation in the case of one variable, geometrical manifolds which are of interest in the theory of functions of several complex variables differ from those usually considered when investigating the Euclidean space of four real variables (see §3);

(2) The differential equations which are satisfied by the real or imaginary parts of a function of two variables are no longer of elliptic type.

3. Domains with distinguished boundary surface. While a function of one variable assumes a value at a point, a function of two variables assumes a constant value on an analytic surface, so that the analytic surfaces or segments of these surfaces (and not isolated points), represent those basic elements in the geometry with the help of which other manifolds have to be generated. As a consequence, a number of new geometrical notions arise in the theory of several variables [Bergman 1, Chapter I].

The following manifolds of the space of two variables play an essential role: Analytic hypersurfaces ($\mathfrak{h}^3$), i.e., one-parameter family of analytic surfaces, and domains whose boundary consists of finitely many segments of analytic hypersurfaces $\mathfrak{h}_k^3$, $k = 1, 2, \cdots, n$.

In considering analytic functions in a domain $\mathfrak{G}$, it is of interest to study the connection between the value distribution of a function inside the domain, and on its boundary. In particular, let us consider a simply-connected domain

$$\mathfrak{G} = \sum_{|\zeta|<1} \mathfrak{L}^2(\zeta), \tag{1}$$

where $\mathfrak{L}^2(\zeta) = [z_\nu = \varphi_\nu(\mathbf{Z}, \zeta), \nu = 1, 2]$, for every fixed[5] ζ, represents a segment of an analytic surface, and a sequence of domains $\mathfrak{G}_n$ which exhaust[6] $\mathfrak{G}$. Let $\sum_{k=1}^{M_n} \mathfrak{h}_k^3$, $\mathfrak{h}_k^3 = \sum_{\lambda_k=0}^{2\pi} \mathfrak{K}_k^2(\exp(i\lambda_k))$; $\mathfrak{K}_k^2(\exp(i\lambda_k)) = [z_\nu = \varphi_\nu^{(k)}(Z_k, \exp(i\lambda_k))$, $\nu = 1, 2]$, $0 \leqq \lambda_k \leqq 2\pi$, (see Fig. 1)[7] be a system of analytic hypersurfaces whose intersection with the boundary $\mathfrak{b}^3$ of $\mathfrak{G}$ is the surface $\mathfrak{s}_n^2 = \mathfrak{b}^3 \cap \sum_{k=1}^{M_n} \mathfrak{h}_k^3$ (which possibly consists of a number of disconnected parts, $\mathfrak{s}_{n,1}^2, \mathfrak{s}_{n,2}^2, \cdots$). If the $\mathfrak{h}_k^3$ are chosen in such a way that every surface $\mathfrak{L}^2(\zeta)$, which passes through a point of $\mathfrak{G}_n$, intersects in a closed curve $\mathfrak{l}_n^1(\zeta)$, the system $\sum_{k=1}^{M_n} \mathfrak{h}_k^3$, and $\mathfrak{G}_n \cap \mathfrak{L}^2(\zeta)$ lies

[5] $\varphi_\nu(\mathbf{Z}, \zeta)[\varphi_\nu^{(k)}(\mathbf{Z}_k, \exp(i\lambda_k))]$ are analytic functions of the variable $\mathbf{Z} = X + iY$ [$\mathbf{Z}_k = X_k + iY_k$], which are regular in a conveniently chosen domain of the $\mathbf{Z}$ [$\mathbf{Z}_k$]-plane.

[6] We obtain, for instance, a sequence of such domains $\mathfrak{G}_n$ by considering the subdomains $K(z, \bar{z}) = n, n_0 \leqq n < \infty$, where K is the kernel function of the domain $\mathfrak{G}$ and $z = (z_1, z_2)$.

[7] In Fig. 1 the surface $\mathfrak{S}_n^2$ appears as four lines $\mathfrak{S}_{n,k}^2$, $k = 1, 2, 3, 4$. The four-dimensional domains $\mathfrak{G}$ and $\mathfrak{G}_n$ appear as two three-dimensional bodies ($\mathfrak{G}_n$ lying inside $\mathfrak{G}$); the boundary hypersurfaces $\mathfrak{b}_3^3$ and $\mathfrak{b}_n^3$ appear as surfaces. The segments of analytic hypersurfaces $\mathfrak{h}_k^3$, $k = 1, 2$, appear as mantle surfaces of two truncated cones, segments of analytic surfaces appear as straight lines on the mantle surfaces. The boundary curve $\mathfrak{h}_k^1(\exp(i\lambda_k))$ of each $\mathfrak{K}_k^2(\exp(i\lambda_k))$ appear as two points. $\mathfrak{L}^2(\zeta)$ appears as a segment of a surface, $\mathfrak{l}_n^1(\zeta)$ is the intersection of $\mathfrak{h}_2^3$ with $\mathfrak{L}^2(\zeta)$. It should be noted that some of the characters in Fig. 1 differ slightly from those used in the text.

inside $\mathfrak{l}_n^1(\zeta)$, then the knowledge of values of $f(z)$, $z \in \mathfrak{S}_n^2$, permits us to determine $f(z)$ at every point of $\mathfrak{G}_n$. Indeed, the boundary curve of each $\mathfrak{K}_k^2(\exp(i\lambda_k))$ lies in $\mathfrak{S}_n^2$, and since $f(z)$ is an analytic function of the variable Z_k in $\mathfrak{K}_k^2(\exp(i\lambda_k))$, we can determine $f(z)$ in every $\mathfrak{K}_k^2(\exp(i\lambda_k))$ and thus in $\sum_{k=1}^{M_n} \mathfrak{h}_k^3$. Further, since in every $\mathfrak{L}^2(\zeta)$, $f(z)$ is an analytic function of one variable $\mathbf{Z}$, we can determine from the values of $f(z)$ in $\mathfrak{l}_n^1(\zeta)$, the value of $f(z)$ in the interior of $\mathfrak{l}_n^1(\zeta)$ which includes $\mathfrak{G}_n \cap \mathfrak{L}^2(\zeta)$.

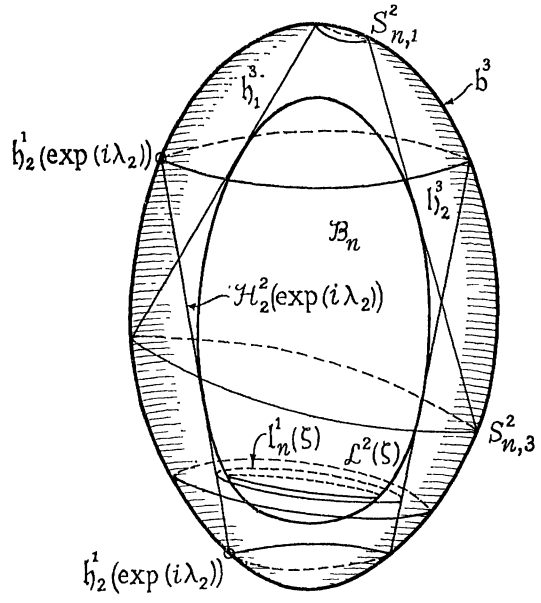

FIG. 1

If $\mathfrak{G}_n \to \mathfrak{G}$, then in general the $\mathfrak{S}_n^2$ will fill out the whole boundary $\mathfrak{b}^3$. For certain special domains, the analytic hypersurfaces $\mathfrak{h}_k^3$ can be chosen in such a way that M_n remains bounded for all n and the $\mathfrak{S}_n^2$ converge to a surface $\mathfrak{S}^2$, which is called the distinguished boundary surface. This occurs in particular in the case of a domain $\mathfrak{G}$ bounded by finitely many segments of analytic hypersurfaces. In this and some other respects, the distinguished boundary surface plays the role of the boundary curve, and in studying the relations between the value distribution inside the domain, and on its boundary, it is sometimes of interest to replace the boundary curve by the whole boundary $\mathfrak{b}^3$, and sometimes by the distinguished boundary surface $\mathfrak{S}^2$. Two particular questions have been studied in this connection:

(1) The determination of integral operators[8] transforming $f(z)$, $z \in \mathfrak{b}^3$ (or $z \in \mathfrak{S}^2$) into $f(z)$, $z \in \mathfrak{G}$ (or for values of z belonging to a certain subdomain of $\mathfrak{G}$).

(2) The question of the introduction of the "extended classes" $E(\mathfrak{G})$ of func-

[8] Most of the operators considered in the following are obtained by using theorems of Green's type or various generalizations of it, relating integrals over the boundary $\mathfrak{b}^3$ or the distinguished boundary surface $\mathfrak{S}^2$ with integrals over the domain.

tions, which class includes B-harmonic functions (real parts of analytic functions of two variables) and possesses some useful properties. In particular, $\boldsymbol{E}(\mathfrak{G})$ have to possess the property that to every (real) function $b(Z)$, $Z \in \mathfrak{b}^3$ (or $\mathfrak{s}^2$), there exists a unique function $F(Z)$, $Z \in \mathfrak{G}$, $F \in \boldsymbol{E}(\mathfrak{G})$, possessing on $\mathfrak{b}^3$ (or $\mathfrak{s}^2$) the prescribed boundary value $b(Z)$. Here and in the following, Z will denote a point of (four-dimensional) x_1, x_2, x_3, x_4-space.

4. Operators transforming boundary data into functions of two variables. Operators solving the problem (1) in the case where $f(z)$ is given on the boundary $\mathfrak{b}^3$ have been discussed in [Bergman (23, p. 389); 2, Supplementary Note III. p. 9; Bochner-Martin {2}]. The derivation of these operators is based on the fact that f is a harmonic function of four real variables and can be represented, using Green's formula, in terms of the boundary value of itself and its derivatives on $\mathfrak{b}^3$. Since the real and imaginary parts of f are connected by the generalized Cauchy-Riemann equations, we can replace the integrals over the derivatives of f by the integrals over f itself. Another operator solving the same problem can be obtained by using complex orthonormal functions which are orthogonal with respect to the boundary $\mathfrak{b}^3$.

The formulas representing the values of the function inside the domain in terms of its value on the distinguished boundary surface $\mathfrak{s}^2$ have been obtained independently by [Bergman (13), 5, 6] and [Weil (1), (2)]. In attempting to develop a unified treatment which will refer simultaneously to problems (1) and (2) of §3, the author of the present paper considers domains bounded by finitely many analytic hypersurfaces of the form

$$\phi_k(z_1, z_2; \lambda_k) = 0, \qquad \lambda_k \text{ real.} \tag{2}$$

It is assumed that the ϕ_k can be uniformized by writing

$$z_\nu = \varphi_\nu^{(k)}(\mathbf{Z}_k, \exp(i\lambda_k)), \qquad \nu = 1, 2, \qquad k = 1, 2, \cdots, n, \qquad \lambda_k \text{ real,} \tag{2a}$$

where $\varphi_\nu^{(k)}$ are, for a fixed λ_k, analytic functions of the auxiliary variable $\mathbf{Z}_k$.

5. Operators transforming boundary data into (real) "functions of the extended class." Considering question (2) for functions F given on the whole boundary $\mathfrak{b}^3$, it is most natural to introduce harmonic functions of four real variables as the "extended" class. On the other hand, this class is not invariant with respect to pseudo-conformal transformations, and the geometry of the space of harmonic functions of four real variables is that of Euclidean space, and not the space of two complex variables. This is the reason why this "extended" class can be used efficiently for comparatively few purposes.

As we stressed before in the case of the domain bounded by finitely many analytic hypersurfaces, the distinguished boundary surface $\mathfrak{s}^2$ in some respects plays a role similar to that of the boundary curve in the theory of one variable.

We shall assume in the following that $\zeta_k = \rho_k \exp(i\lambda_k)$, and that the functions, $\varphi_\nu^{(k)}$ in (2a) are defined for $1 - \epsilon \leqq |\zeta_k| \leqq 1$, $\epsilon > 0$, for $k = 1, 2, \cdots, n$, and one pair, say $\varphi_\nu^{(1)}$, is defined for $|\zeta_1| \leqq 1$, so that we may use the surfaces $\mathfrak{IC}_1^2(\zeta_1)$ as the surfaces $\mathfrak{L}^2(\zeta)$ appearing in (1). In the case of the bicylinder $|z_\nu| < 1$, $\nu = 1, 2$, it is natural to introduce the class of doubly-harmonic functions (i.e., functions F of four real variables x_1, y_1, x_2, y_2 satisfying $\partial^2 F/\partial z_\nu \partial \bar{z}_\nu = 0$, $\nu = 1, 2$) as the extended class $\boldsymbol{E}$. The surface $\mathfrak{S}^2 = [|z_1| = 1, |z_2| = 1]$ is the distinguished boundary surface.

We proceed now to define an analogous "extended class" $E(\mathfrak{B})$ for more general domains $\mathfrak{B}$ bounded by finitely many segments $\mathfrak{h}_k^3$ of analytic hypersurfaces. Since each boundary component $\mathfrak{h}_k^3$ is a sum of segments $\mathfrak{IC}_k^2(\exp(i\lambda_k))$ of analytic surfaces, and the boundary curve $\mathfrak{h}_k^1(\exp(i\lambda_k))$ of each segment $\mathfrak{IC}_k^2(\exp(i\lambda_k))$ lies in the distinguished boundary surface, we can, in every segment $\mathfrak{IC}_k^2(\exp(i\lambda_k))$ determine that harmonic function of the auxiliary variables X_k, Y_k ($\mathbf{Z}_k = X_k + iY_k$) which assumes the prescribed value on the boundary curve $\mathfrak{h}_k^1(\exp(i\lambda_k))$. By this rule, the function of the extended class is defined in the whole boundary $\mathfrak{b}^3$ of the domain. With rather general assumptions, it is possible to show that the function defined in $\mathfrak{b}^3$ in this manner depends only upon the values prescribed on $\mathfrak{S}^2$ [Bergman (27)].

Further, we assumed that the domain $\mathfrak{B}$ can be represented as a sum, $\mathfrak{B} = \sum_{|\zeta_1|<1} \mathfrak{IC}_1^2(\zeta_1)$, of analytic surfaces, see (1). The boundary curve $\mathfrak{h}_1^1(\zeta_1)$ of each segment $\mathfrak{IC}_1^2(\zeta_1)$, $|\zeta_1| \leqq 1$, lies in the three-dimensional boundary $\mathfrak{b}^3$, and therefore we can determine in each $\mathfrak{IC}_1^2(\zeta_1)$ that harmonic function of X_1, Y_1 ($\mathbf{Z}_1 = X_1 + iY_1$) which assumes on the boundary $\mathfrak{h}_1^1(\zeta_1)$ the values obtained previously. In this manner, we define for every real integrable function $f(Z)$, defined on the distinguished boundary surface $\mathfrak{S}^2$, a function $\mathrm{P}(f)$, defined in the domain $\mathfrak{B}$ and assuming on $\mathfrak{S}^2$ the boundary values[9] f. The linear operator $\mathrm{P}(f)$ is a generalization of the Poisson formula. It has, among other properties, the properties:

(1) If $f \geqq 0$, $\mathrm{P}(f) \geqq 0$.

(2) If f are the boundary values of a B-harmonic function $F(Z)$, regular in $\mathfrak{B}$ and satisfying certain conditions, then $\mathrm{P}(f) = F$ (see p. 168, l. 1).

(3) The functions of the extended class are invariant with respect to transformations which are pseudo-conformal in $\bar{\mathfrak{B}}$.

REMARK. For some purposes (for which the invariance with respect to pseudo-conformal transformations is not essential), one can define the function of the extended class by replacing the second step by forming (after the function is defined on $\mathfrak{b}^3$ in the above way) that harmonic function of four real variables which assumes on $\mathfrak{b}^3$ the previously determined values.

[9] However, we wish to stress that it could easily happen that the above representations of the domain could occur in a different manner, so that we could obtain different "extended classes." The question of the relation between the extended classes has not yet been sufficiently clarified.

6. Applications of functions of the extended class. The introduction of functions of the extended class permits us to employ potential-theoretical methods in the theory of functions of two complex variables. In particular, we can introduce Green's function which behaves as $-\log|g(z_1, z_2)|$ along the segment of the surface $g(z_1, z_2) = 0$, lying in $\mathfrak{B}$, and vanishes on the distinguished boundary surface. (Here g is an analytic function of two complex variables.) Using these results, it is possible, for some special domains, to derive generalizations of the Fatou theorem [Bergman and Marcinkiewicz {1}, Bers {1}] and of the theories of Nevanlinna and Ahlfors for value distribution [Bergman {19}].

More exactly, it was possible to obtain relations between certain geometrical notions, from the geometry of space of two variables, such as certain "measures" (i.e., quantities characteristic of some properties) of the intersection $\mathfrak{f}_b^1 = [(f(z_1, z_2) = v) \cap \mathfrak{b}^3]$ of v-surfaces of a function $f(z_1, z_2)$ with the (three-dimensional) boundary $\mathfrak{b}^3 = \sum_{k=1}^{n} \mathfrak{b}_k^3$ of a domain with a distinguished boundary surface $\mathfrak{s}^2$, on one side, and the values of f on $\mathfrak{s}^2$ on the other side [Bergman (7), (14), (20), 7; Bergman and Martin (1); Gelbart {1}]. In particular, one can consider a one-parameter family of domains $\mathfrak{B}(\rho)$, $\rho_0 \leqq \rho < \infty$, with distinguished boundary surfaces $\mathfrak{s}^2(\rho)$, and obtain relations between the growth of the above "measures" referring to $\mathfrak{f}^1(\rho)$ and the growth of f on $\sum_{\rho=\rho_0}^{\infty} \mathfrak{s}^2(\rho)$. These results represent a generalization of the first and second theorems of Nevanlinna and the theorem of Hadamard on zeros of entire functions [Bergman {19}].

Let $\mathfrak{B} = [z_1 \in \mathfrak{K}_1^2(Z_2),\ Z_2 = z_2\ ;\ |Z_2| < 1]$ be a domain bounded by $\mathfrak{h}_1^3 = [z_2 = \exp(i\lambda_2)]$ and $\mathfrak{h}_2^3 = [z_1 = h(z_2, \exp(i\lambda_1)]$, $0 \leqq \lambda_k \leqq 2\pi$, $z_k = x_{2k-1} + ix_{2k}$. Here $z_1 = h(z_2, \exp(i\lambda_1))$ represents for every fixed z_2 a simple closed curve $\mathfrak{h}_1^1(z_2)$, whose interior is denoted by $\mathfrak{K}_1^2(z_2)$.

Let f and g be three times differentiable functions of x_1, x_2, x_3, x_4, such that $f_{z_1\bar{z}_1} = 0$ ($f_{z_1\bar{z}_1} \equiv (\partial^2 f/\partial z_1 \partial \bar{z}_1)$) for every fixed x_3, x_4, $x_3^2 + x_4^2 \leqq 1$.

Let $d\omega_k = dx_{2k-1}\, dx_{2k}$, $\mathfrak{K}_2^2 = [|z_2| < 1]$, $\mathfrak{h}_2^1 = [|z_2| = 1]$. Then

$$\iint_{\mathfrak{K}_2^2} \iint_{\mathfrak{K}_1^2(z_2)} f_{z_1 z_2} g_{\bar{z}_1 \bar{z}_2}\, d\omega_1\, d\omega_2 = -\frac{1}{4} \int_{\mathfrak{h}_2^1} \int_{\mathfrak{h}_1^1(z_2)} f_{z_1 z_2} [g - \bar{h} g_{\bar{z}_1}]\, dz_1\, dz_2$$

$$+ \left(\frac{1}{2i}\right) \iint_{\mathfrak{K}_2^2} \int_{\mathfrak{h}_1^1(z_2)} f_{z_1 z_2} \bar{h} \left\{\frac{\partial(g_{\bar{z}_1})}{\partial \bar{z}_2}\right\} dz_1\, d\omega_2 - T,$$

$$T \equiv \left(\frac{1}{2i}\right) \iint_{\mathfrak{K}_2^2} \int_{\mathfrak{h}_1^1(z_2)} \left\{\frac{\partial(f_{z_1 z_2})}{\partial \bar{z}_2}\right\} [g - \bar{h} g_{\bar{z}_1}]\, dz_1\, d\omega_2 .$$

Here $(\ ) \equiv (\ )_{z_1 = h(z_2, \lambda_1),\, \bar{z}_1 = \bar{h}(\bar{z}_2, \lambda_1)}$, $\{\ \} \equiv \{\ \}_{\lambda_1 = \lambda_1(z_1, \bar{z}_1, z_2, \bar{z}_2)}$.

REMARK. If $\partial(f_{z_1 z_2})/\partial \bar{z}_2 = 0$ for every λ_1 (which will take place, e.g., if f is an analytic function of z_1, z_2), then $T = 0$. (In this case, in the formula (3) we can replace $\bar{h}$ by $\{\bar{h}(\bar{z}_2, \exp(i\lambda_1)) - \bar{h}(0, \exp(i\lambda_1))\}$.)

Green's formulas of similar kind can be obtained for various integrands, and for general distinguished boundary surfaces. Using them and substituting for f conveniently chosen functions, we obtain various generalized "residue theorems" which can be among others used to generalize considerations of [Bergman {19}].

A further application of functions of the extended class $E(\mathfrak{G})$ is the possibility of obtaining invariants with respect to pseudo-conformal transformations. In this connection, it is of considerable interest to study the topological structure of the domains with distinguished boundary surface described in §3. On the boundary $\mathfrak{b}^3$ of a domain $\mathfrak{G}$ which is bounded by finitely many segments $\mathfrak{h}_k^3$ of analytic surfaces, we can distinguish the following types of points: (1) those which belong only to one $\mathfrak{h}_k^3$; (2) those through which go at least two segments $\mathfrak{h}_k^3$ (they form the distinguished boundary surface $\mathfrak{s}^2$); (3) those through which go at least three $\mathfrak{h}_k^3$ (distinguished boundary line $\mathfrak{l}^1$); (4) those through which go at least four $\mathfrak{h}_k^3$ (the distinguished boundary set $\mathfrak{P}^0$).

Applying the Meyer-Vietoris theorem, and assuming that there are no intersections of more than four[10] $\mathfrak{h}_k^3$, one obtains relations between the Betti numbers B^p of the segments $\mathfrak{h}_k^3$ of the boundary $\mathfrak{b}^3 = \sum_{\nu=1}^n \mathfrak{h}_\nu^3$, segments $\mathfrak{s}_{\nu j}^2 = \mathfrak{h}_\nu^3 \cap \mathfrak{h}_j^3$ of the distinguished boundary surface, segments $\mathfrak{l}_{\nu jk}^1 = \mathfrak{h}_\nu^3 \cap \mathfrak{h}_j^3 \cap \mathfrak{h}_k^3$ of the distinguished boundary line, discrete point set $\mathfrak{P}_{\nu jks}^0$ belonging to at least four segments $\mathfrak{h}_k^3$. We have

$$
\begin{aligned}
B^p(\mathfrak{b}^3) = {} & \sum_{\nu=1}^{n} B^p(\mathfrak{h}_\nu^3) - \sum_{\nu=1}^{n} \sum_{j>\nu} B^p(\mathfrak{s}_{\nu j}^2) + \sum_{\nu=1}^{n} \sum_{j>\nu} \sum_{k>j} B^p(\mathfrak{l}_{\nu jk}^1) - B^p(\mathfrak{P}_{\nu jks}^0) \\
& + \sum_{k=1}^{n-1} N^p\left[\sum_{\nu=1}^{k} \mathfrak{h}_\nu^3, \mathfrak{h}_{k+1}^3\right] - \sum_{k=1}^{n-2} \sum_{j=2}^{n-k} N^p\left[\sum_{\nu=1}^{k} \mathfrak{h}_\nu^3 \cap \mathfrak{h}_{k+j}^3, \mathfrak{h}_{k+1}^3 \cap \mathfrak{h}_{k+j}\right] \\
& + \sum_{k=1}^{n-3} \sum_{j=2}^{n-k-1} \sum_{s=j+1}^{n-k} N^p\left[\sum_{\nu=1}^{k} \mathfrak{h}_\nu^3 \cap \mathfrak{h}_{k+j}^3 \cap \mathfrak{h}_{k+s}^3, \mathfrak{h}_{k+1}^3 \cap \mathfrak{h}_{k+j}^3 \cap \mathfrak{h}_{k+s}^3\right].
\end{aligned} \tag{4}
$$

Here $N^p(\mathfrak{A}, \mathfrak{B}) = n^p(\mathfrak{A}, \mathfrak{B}) + n^{p-1}(\mathfrak{A}, \mathfrak{B})$, where $n^p(\mathfrak{A}, \mathfrak{B})$ is the number of homologously independent p-cycles of $\mathfrak{A} \cap \mathfrak{B}$ which bound in $\mathfrak{A}$ and in $\mathfrak{B}$ but not in $\mathfrak{A} \cap \mathfrak{B}$.

As a consequence of the above formula, we obtain for the Euler-Poincaré number E, the relation

$$
E(\mathfrak{b}^3) = \sum_{\nu=1}^{n} E(\mathfrak{h}_\nu^3) - \sum_{\nu=1}^{n} \sum_{j>\nu} E(\mathfrak{s}_{\nu j}^2) + \sum_{\nu=1}^{n} \sum_{j>\nu} \sum_{k>j} E(\mathfrak{l}_{\nu jk}^1) - E(\mathfrak{P}_{\nu jks}^0). \tag{5}
$$

Since by pseudo-conformal transformation, defined in the closed domain, the topological structure of the distinguished boundary manifolds is preserved, the above numbers are invariant with respect to pseudo-conformal transformations. In the case where the boundary consists of more than two segments $\mathfrak{h}_k^3$, the distinguished boundary line $\mathfrak{l}^1$ divides the distinguished boundary surface into parts $\mathfrak{s}_{\nu k}^2$. In this case, we can introduce generalized harmonic measures $\omega_{\nu k}(Z)$, i.e., functions of the extended class which assume on one of the segments, say

[10] In the case where there are intersections of more than four $\mathfrak{h}_k^3$, the formulas (4), (5) have to be altered somewhat.

on $S^2_{\nu k}$, the value one, and vanish on the remaining part of S^2. Using Morse's theory of critical points, one can show that for various combinations $[\alpha_{\nu k}\omega_k(Z) + \beta_\nu G(Z)]$, $\alpha_{\nu k}$, β_ν constants, critical surfaces and critical points must exist. Since functions of the extended class are invariant with respect to the pseudo-conformal transformation, the topological structure of the above critical manifolds must be preserved in pseudo-conformal transformations, so that in this manner we obtain further invariants with respect to pseudo-conformal transformations.

Further, if one considers the projection of the above-mentioned functions (the generalized harmonic measures, Green's functions, kernel functions, etc.) in the space of B-harmonic functions, and introduces corresponding analytic functions of two variables, then one obtains certain functions which can be considered as generalizations of functions of the first, second, and third kinds, in the theory of functions of one variable.[11] (Projections of this kind can be obtained by using the kernel function for B-harmonic functions.) Using the relation between harmonic functions of four variables and the class of B-harmonic functions, we can make a number of conclusions for the kernel function of the B-harmonic function, which refer to the behavior of the kernel function near the boundary. With these results, it is possible to establish different properties of the above-mentioned analogues of functions of the first, second, and third kinds, and of some of their combinations.

REMARK. The theory of functions of several complex variables can be used successfully in a number of fields. In particular, many results in this theory lead immediately to theorems in the theory of linear and certain nonlinear partial differential equations of n variables, $n \geqq 2$.

In substituting $z^* = 0$ into the harmonic function $h(x_1, x_2) = [f(z) + f(z^*)]/2$, $z \equiv x_1 + ix_2$, $z^* = x_1 - ix_2$, i.e., considering h in the characteristic plane $x_1 = ix_2$, we obtain $f(z)/2 +$ const. (a function of one complex variable). Conversely, the operator "Re" transforms f back into h. Analogously, substituting

$$x_1 = 2(ZZ^*)^{1/2}, \qquad x_2 = -i(Z + Z^*), \qquad x_3 = Z - Z^*$$

into a harmonic function $h(x_1, x_2, x_3)$, i.e., considering h in a subdomain of the characteristic space $x_1^2 + x_2^2 + x_3^2 = 0$, we obtain the function $\chi = \chi_1 + (ZZ^*)^{1/2}\chi_2$ where $\chi_\kappa(Z, Z^*)$, $\kappa = 1, 2$, are analytic functions of two complex variables, Z, Z^*, which are regular at the origin. The inverse operator transforming χ back into the harmonic function h (i.e., the analogue of "Re") is

$$\frac{1}{\pi i}\int_{|\zeta|=1}\int_{T=0}^{1}\frac{u^{1/2}}{\zeta}\,\frac{d[u^{1/2}\chi(u\zeta^{-1}T^2, u\zeta(1-T)^2)]}{du}\,dT\,d\zeta,$$

where $u = x_1 + ((ix_2 + x_3)\zeta/2) + ((ix_2 - x_3)\zeta^{-1}/2)$.

[11] Similar to the situation in one variable, these functions are solutions of certain variational problems.

Thus (in analogy to the situation for harmonic functions of two variables) many results in the theory of functions of two complex variables can be interpreted as results in the theory of harmonic functions and harmonic vectors[12] of *three* variables. (E. g. an analogue of Abel's theorem [8, §III, 4].) Analogously, using operators which transform χ into solutions of differential equations, the results in the theory of functions of two variables can also be interpreted as theorems relating to functions of three variables which satisfy certain linear partial differential equations. For details, see [8]. Further, it is possible to derive various relations between properties of subsequences of the coefficients of the series development of a solution and its behavior in the large.

[12] I. e. vector $\mathfrak{q}$ for which curl $\mathfrak{q} = 0$, div $\mathfrak{q} = 0$.

Bibliography

1. S. Bergman, *Sur les fonctions orthogonales* . . . , New York, Interscience Publishers, 1941, and Mémorial des Sciences Mathématiques vol. 106, Paris, 1947.

2. ———, *Partial differential equations, advanced topics*, Providence, Brown University, 1941.

3. ———, *Sur les fonctions-noyau* . . . , Mémorial des Sciences Mathématiques vol. 108, Paris, 1948.

4. ———, *The kernel function and conformal mapping*, Mathematical Surveys, vol. 5, New York, Amer. Math. Soc., 1950.

5. ———, *Über eine in gewissen Bereichen gültige Integraldarstellung* . . . , Math. Zeit. vol. 39 (1934) pp. 76–94, 605–608.

6. ———, *Über eine Integraldarstellung* . . . , Rec. Math. (Mat. Sbornik) N.S. vol. 1 (43) (1936) pp. 851–862.

7. ———, *Über das Verhalten der Funktionen von zewi komplexen Veränderlichen in Gebieten mit einer ausgezeichneter Randfläche*, Mathematica vol. 14 (1938) pp. 107–123.

8. ———, *On solutions with algebraic character of linear partial differential equations*, Trans. Amer. Math. Soc. vol. 68 (1950) pp. 461–507.

Harvard University,
Cambridge, Mass., U. S. A.

APPLICATIONS OF AREA THEORY IN ANALYSIS

L. CESARI AND T. RADÓ

1. General background. During the first half of the present century, the so-called direct method in calculus of variations has been developed to a high degree of completeness for single integral problems, both in the parametric and in the non-parametric case. The foundations of the theory include general concepts in functional analysis in the sense of Volterra and the general principle of semi-continuity, introduced and systematically exploited in this field by Tonelli. Furthermore, constant and essential use is made of a set of basic theorems which state relationships between the fundamental analytic concepts of derivative, integral, bounded variation, and absolute continuity for functions of a single real variable, and the fundamental geometric concepts of curve and arc length. It is to be expected that the general program of developing a corresponding comprehensive theory of double integral problems in calculus of variations will depend upon appropriate two-dimensional extensions of these one-dimensional concepts and theorems. The purpose of this lecture is to sketch the two-dimensional theory developed for these purposes in the course of the past fifty years, with particular reference to recent applications in calculus of variations. The literature of this theory and of its applications in calculus of variations is very extensive, and we have to restrict ourselves to the following comments concerning bibliography. During the last three years, there appeared three comprehensive expository presentations. 1. L. Cesari, *Area and representation of surfaces*, Bull. Amer. Math. Soc. vol. 56 (1950). 2. T. Radó, *Length and area*, Amer. Math. Soc. Colloquium Publications, vol. 30, New York, 1948. 3. J. W. T. Youngs, *Topological methods in the theory of Lebesgue area*, Bull. Amer. Math. Soc. vol. 56 (1950). These expository presentations include comprehensive bibliographical and historical references, and in particular they give a picture of the fundamental initial contributions of Lebesgue, Geöcze, Banach, and Vitali. Let us also call attention to about fifty research papers in our field which appeared in the course of the last two years. Due to limitations of space, no further references to individual authors can be made in this lecture.

As stated above, our main objective is to review a two-dimensional extension of a one-dimensional theory. Manifestly, the ultimate aim is to develop a corresponding n-dimensional theory. Actually, intensive and promising research is already going on in this direction. While the picture is very far from being complete as yet, it is clear that the constant interplay of analysis and topology is even more evident in the n-dimensional case. In fact, it appears that a number of new and difficult topological problems are bound to arise. We restrict ourselves in the sequel to the two-dimensional case, but in setting forth the basic concepts we shall give preference to those variants of the definitions which seem to be best suited for generalization to the n-dimensional case, in the light of recent research.

2. The concept of surface area. Since the principle of semi-continuity is basic in the direct method in calculus of variations, it is natural that the concept of the Lebesgue area $L(S)$ of a surface S, proposed by Lebesgue in 1900, has been generally accepted in the theory with which we are concerned. Indeed, the Lebesgue area is a lower semi-continuous functional. The Lebesgue area $L(S)$ is defined as the limit inferior of the elementary areas of all sequences of polyhedra (not necessarily inscribed) approaching S, in the sense of the distance introduced by Fréchet. The concept of surface employed here is based on ideas of Fréchet. Let us merely note that a surface, in this sense, is not a point set, but rather a class of continuous mappings (from a 2-cell into Euclidean three-space) which are equivalent in a certain sense defined by Fréchet. It is well known that there exist many definitions of surface area. While each one of these various definitions is based upon some postulate derived from properties of surface area in the elementary range, these definitions are known to lead to different values for the area if applied to general surfaces in the sense of Fréchet (quite precisely, the term general surface S will mean Fréchet surface of the type of the 2-cell in this lecture). However, a number of results have been obtained in recent years to the effect that if the postulates underlying the various definitions of surface area are relaxed in a certain natural and appropriate manner, then the resulting modified areas agree with the Lebesgue area $L(S)$ for all surfaces S. In particular, this is true for various definitions operating with the orthogonal projections of the surface upon planes. Such definitions include those proposed by Banach, Peano, and Cauchy. The main result in this direction is that the so-called lower area $a(S)$ agrees with the Lebesgue area $L(S)$. Similarly, it was found recently that an area-definition proposed by Favard, operating with the average number of intersections of a line with the surface S (in the sense of integral geometry) may be made to agree with the Lebesgue area by means of an appropriate modification. If one modifies the Lebesgue definition of area by permitting only the use of inscribed polyhedra, one obtains an area $L^*(S)$. The celebrated problem of Geöcze consists of showing that $L^*(S)$ agrees with the Lebesgue area $L(S)$. It is now known that this is generally true, even though the so-called strong form of the problem requires further study in the parametric case. While the beautiful and difficult researches concerned with the relationships between the Lebesgue area $L(S)$ and other relevant definitions justify the hope that ultimately a unified picture of area theory will emerge, a great deal remains to be done yet. In particular, the relationships with area definitions based upon the concept of n-dimensional Hausdorff measure in N-dimensional space seem to require further fundamental study, even though several important results (both published and unpublished) are known.

3. Plane mappings of bounded variation. Let D be a bounded domain (connected open set) in the uv plane and let $T\colon x = x(u, v),\ y = y(u, v),\ (u, v) \in D$, be a bounded continuous mapping from D into the xy plane. A basic concept is that of the (essential) multiplicity function $K(x, y, T, D)$. Of the various equiva-

lent definitions the following one seems most suitable from the point of view of extension to higher dimension. Given a point (x_0, y_0) in the xy plane, and a finitely connected Jordan region R in D, let $\gamma_1, \cdots, \gamma_n$ be the boundary curves of R, oriented in the usual sense, and let $C_1, \cdots, C_n$ be their images under T. Let $O(x_0, y_0, C_i)$ be the topological index of (x_0, y_0) with respect to C_i, and let us set

$$\mu(x_0, y_0, T, R) = \sum_{i=1}^{n} O(x_0, y_0, C_i),$$

with the understanding that $\mu = 0$ if (x_0, y_0) lies on $C_1 + \cdots + C_n$. Then R is termed an indicator region for (x_0, y_0), T, D if $\mu(x_0, y_0, T, R) \neq 0$. An indicator region is termed positive (negative) if $\mu > 0$ ($\mu < 0$). Let σ be a generic notation for a finite system of pairwise disjoint indicator regions for (x_0, y_0) under T in D. One defines then

$$K(x_0, y_0, T, D) = \operatorname{lub}_{\sigma} \sum_{R \in \sigma} | \mu(x_0, y_0, T, R) | .$$

If there are no nonempty systems σ, then of course one sets $K = 0$. If each R occurring in σ is required to be a positive (negative) indicator region, then one obtains multiplicity functions to be denoted by $K^+(x_0, y_0, T, D), K^-(x_0, y_0, T, D)$ respectively. The following comments are in order. There are a number of other definitions for an essential multiplicity function, but it can be shown that any two of these differ at most at a countable set of points (x, y). Hence, all of them are equivalent as regards the following theory, since only their definite integrals appear in the fundamental definitions and formulas. However, this is not true any more in the n-dimensional case ($n > 2$). In a similar manner, further fundamental concepts admit of equivalent definitions in the case $n = 2$, while a judicious selection must be made if $n > 2$. In the sequel, only the variant best suited for generalization to the case $n > 2$ will be given in connection with basic concepts.

DEFINITION. The mapping T is BV (of bounded variation) in D if $K(x, y, T, D)$ is summable. The integral of K (taken over the whole xy plane) is then the total variation $V(T, D)$ of T in D.

4. Absolutely continuous plane mappings. A continuum $c \subset D$ is termed an essential maximal model continuum for (x_0, y_0), T, D if (i) c is a component of the set $T^{-1}(x_0, y_0)$, and (ii) every neighborhood of c contains an indicator region for (x_0, y_0), T, D which contains c. The essential set $E(T, D)$ is the set sum of all the essential maximal model continua corresponding to all the points (x, y).

DEFINITION. The mapping T is termed AC (absolutely continuous) in D if (i) it is BV, and (ii) subsets of measure zero of $E(T, D)$ are mapped by T into sets of measure zero.

5. The generalized Jacobian. Assume that T is BV in D. Then the multiplicity functions $K^+(x, y, T, D)$, $K^-(x, y, T, D)$ are finite almost everywhere, and hence we can define (almost everywhere) a signed multiplicity function $n(x, y, T, D) = K^+(x, y, T, D) - K^-(x, y, T, D)$. If r is a generic notation for a closed rectangle in D, with sides parallel to the axes, then $n(x, y, T, r)$ denotes the signed multiplicity function corresponding to the interior of r. The Lebesgue integral of $n(x, y, T, r)$, taken over the xy plane, is a rectangle function which can be shown to possess a derivative almost everywhere in D. This derivative is the generalized Jacobian $J(u, v, T)$. Similarly, we define $W(u, v, T)$ as the derivative of the rectangle function obtained by using K instead of n in the preceding definition. In the sequel, we shall not display T in the notations.

6. Relationships between the basic concepts. In analogy with the classical one-dimensional case, we have the following set of theorems.

(i) If T is BV in D, then the generalized Jacobian $J(u, v)$ exists almost everywhere in D, is summable in D, and satisfies the relations

$$| J(u, v) | = W(u, v) \text{ a.e. in } D, \qquad V(D) \geqq \iint_D | J(u, v) | \, du \, dv.$$

(ii) In the preceding inequality, the sign of equality holds if and only if T is AC in D.

(iii) If T is BV in D, and if the partial derivatives x_u , x_v , y_u , y_v exist a.e. in D, then $J(u, v) = x_u y_v - x_v y_u$ a.e. in D.

(iv) $V(D) = V(D, T)$ is a lower semi-continuous functional of T.

7. Applications to the transformation of double integrals. For simplicity, let the continuous mapping T be defined on the unit square Q: $0 \leqq u \leqq 1$, $0 \leqq v \leqq 1$. The interior Q^0 of Q will then take the role of the domain D. Let $f(x, y)$ be a finite, real, single-valued function in the xy plane, and let us set $F(u, v) = f[x(u, v), y(u, v)]$ for $(u, v) \in Q$. Assume that T is AC in Q^0. Then the following transformation formulas hold:

$$\iint_Q F(u, v) \, | J(u, v) | \, du \, dv = \iint_{T(Q)} f(x, y) K(x, y) \, dx \, dy,$$

$$\iint_Q F(u, v) J(u, v) \, du \, dv = \iint_{T(Q)} f(x, y) n(x, y) \, dx \, dy,$$

provided that at least one of the products FJ, fK is measurable and summable. These formulas contain, as very special instances, a large number of previous results. The wide scope of the above formulas is due to the fact that the class of AC mappings is closed under certain very general limit processes.

8. Applications to the Lebesgue area. Given a surface S: $x = x(u, v)$, $y = y(u, v)$, $z = z(u, v)$, $(u, v) \in Q$: $0 \leqq u \leqq 1$, $0 \leqq v \leqq 1$, let us denote by T_1 ,

T_2, T_3 the plane continuous mappings obtained by the formulas T_1: $y = y(u, v)$, $z = z(u, v)$; T_2: $z = z(u, v)$, $x = x(u, v)$; T_3: $x = x(u, v)$, $y = y(u, v)$. In analogy with classical theorems on arc length, we have the following theorems.

(i) The Lebesgue area is finite if and only if T_1, T_2, T_3 are BV.

(ii) If $L(S) < \infty$, then the generalized Jacobians J_1, J_2, J_3, corresponding to the mappings T_1, T_2, T_3, exist a.e. in Q and are summable in Q, and satisfy the inequality

$$L(S) \geqq \iint_Q (J_1^2 + J_2^2 + J_3^2)^{1/2} \, du \, dv.$$

(iii) In the preceding inequality, the sign of equality holds if and only if the mappings T_1, T_2, T_3 are AC.

(iv) Assume that $L(S) < \infty$. Introduce a new set of Cartesian coordinates $\bar{x}$, $\bar{y}$, $\bar{z}$, and denote by $\bar{J}_1$, $\bar{J}_2$, $\bar{J}_3$ the generalized Jacobians corresponding to the new representation of S. Then the old and new generalized Jacobians are related by the same linear transformation which leads from x, y, z to $\bar{x}$, $\bar{y}$, $\bar{z}$, and indeed these relations hold *uniformly almost everywhere in* Q in the sense that the exceptional set of measure zero can be chosen independently of the choice of the system $\bar{x}$, $\bar{y}$, $\bar{z}$.

(v) If $L(S) < \infty$, then S admits of a representation where the mappings T_1, T_2, T_3 are AC, and indeed one has AC representations which are *almost conformal* in a certain appropriate sense.

REMARK. These statements imply that a number of basic facts in differential geometry continue to hold, in a certain approximate sense, for surfaces of finite Lebesgue area.

9. Applications in calculus of variations. In view of the striking analogies between the preceding two-dimensional theory and the corresponding classical one-dimensional theory, corresponding analogies may be expected in applications to double integral problems in calculus of variations. The following advances have been achieved up to now in this direction. First, a general concept of a Weierstrass-type integral $I(S)$ has been introduced under the only assumption that the Lebesgue area of the surface S is finite. This integral is independent of the particular representation of S, but if the representation is such that the corresponding mappings T_1, T_2, T_3 are AC, then $I(S)$ coincides with the usual variational integral calculated in terms of generalized Jacobians. General theorems concerning the lower semi-continuity of double integrals were obtained. The following existence theorem has been proved: *Each positive definite semi-regular integral $I(S)$ has an absolute minimum in the class of all surfaces S satisfying the following conditions.* (i) *S has finite Lebesgue area.* (ii) *S is contained in a given closed and bounded convex subset A of xyz space.* (iii) *S is bounded by a given Jordan curve which bounds at least one surface of finite Lebesgue area.* Let us note also that two classical double integral problems, the Plateau problem

and the isoperimetric problem for closed surfaces, have been solved in terms of the Lebesgue area $L(S)$ and the principle of semi-continuity.

UNIVERSITY OF BOLOGNA,
BOLOGNA, ITALY.
OHIO STATE UNIVERSITY,
COLUMBUS, OHIO, U. S. A.

THE PROBLEM OF PLATEAU ON A RIEMANNIAN MANIFOLD AND RELATED TOPICS

CHARLES B. MORREY, JR.

1. Introduction. In a recent paper entitled *The problem of Plateau on a Riemannian manifold* [9],[1] the writer generalized the results of Courant [3] for the case of surfaces "of type k" bounded by k Jordan curves to the case where the surfaces were embedded in a Riemannian manifold of considerable generality. The writer knows of only one paper in which the Plateau problem has been solved in any space other than Euclidean space, namely the recent paper by Lonseth [4] where the space considered is "hyperbolic space". However, Bochner [1] has proved some interesting results concerning "harmonic" functions in a general Riemannian metric of sufficient differentiability.

In this address the writer will first explain in more detail the results obtained and indicate briefly some of the methods used and difficulties encountered in obtaining the results. The writer will also discuss briefly some of his previous work bearing on the differentiability of the resulting solutions. For the sake of brevity, a discussion of the history of this problem and extensive references to the work of other writers in this field will be omitted. The writer hopes that he will be pardoned for such omissions. The writer's paper referred to above and some of the references cited therein contain extensive bibliographies.

2. Preliminaries. By a *surface of type* k is meant a Fréchet variety which possesses a representation on a plane region of type k, i.e., a closed region bounded by k mutually exclusive Jordan curves. If $x = x(u, v)$, $(u, v) \in \bar{B}$, is a representation of such a surface S, the boundary of S is the system Γ of k closed curves which has as one of its representations $x = x(u, v)$, $(u, v) \in B^*$, B^* denoting the boundary of B.

We assume the usual definition of a manifold $\mathfrak{M}$ (without boundary) of class $C^{(n)}$ in terms of a finite or denumerable system of preferred neighborhoods $\mathfrak{M}_i$ and transformations T_i from the unit cube R: $| x^j | < 1$, $j = 1, \cdots, N$. The manifold is Riemannian if it is also a connected complete metric space in which the metric can be defined in the usual way on each $\mathfrak{M}_i$ through the transformations T_i and functions $g_{p,q}^{(i)}(x)$ of class $C^{(n-1)}$. Vector functions $x(u, v)$ of class $C^{(m)}$, $m \leqq n$, with values in a Riemannian manifold $\mathfrak{M}$ of class $C^{(n)}$, $n \geqq 1$, are defined as usual and the definitions of absolute continuity of a function of one variable and absolute continuity in the sense of Tonelli (ACT) of $x(u, v)$ can be carried over immediately. For functions $x(u, v)$ which are ACT, we define

$$E = g_{\alpha\beta}(x)x_u^\alpha x_u^\beta, \qquad F = g_{\alpha\beta}(x)x_u^\alpha x_v^\beta, \qquad G = g_{\alpha\beta}(x)x_v^\alpha x_v^\beta$$

$$| x_u | = E^{1/2}, \ | x_v | = G^{1/2}$$

[1] Numbers in brackets refer to the bibliography at the end of the paper.

wherever the partial derivatives of the components exist; it is clear that these functions are independent of which T_i is used to determine the components. If $x(u, v)$ is ACT with E and G summable over B, we say that x is of class $\mathfrak{B}_2''$ on B and define the Dirichlet and area integrals, respectively by

$$D_2(x, B) = \iint_B (E + G)\, du\, dv, \qquad \tilde{L}(x, B) = \iint_B (EG - F^2)^{1/2}\, du\, dv.$$

From work of McShane [5] and the writer [6], it follows that these integrals are lower semi-continuous with respect to uniform convergence over the class of vectors of class $\mathfrak{B}_2''$ on a given region B of type k.

Using the lower semi-continuity of $\tilde{L}(x, B)$ we define the functional $L(x, B)$ for any continuous mapping (x, B) as the largest lower semi-continuous functional which coincides with $\tilde{L}(x, B)$ whenever x is of class $\mathfrak{B}_2''$. It is easily shown that $L(x_1, B_1) = L(x_2, B_2)$ whenever (x_1, B_1) and (x_2, B_2) are Fréchet equivalent. Accordingly, if S is any surface of type k, we define its Lebesgue area $L(S) = L(x, B)$ where (x, B) is any representation of S.

3. The boundary value problem. Following the general idea of Courant, the writer replaces the problem of least area by a Dirichlet problem with variable domains and boundary values and then shows that the second problem has a solution which is also a solution of the first.

To do this, we begin by defining

$$l(\Gamma) = \text{greatest lower bound of } \liminf_{n\to\infty} L(S_n)$$

for all sequences $\{S_n\}$ of surfaces whose boundaries $\Gamma_n \to \Gamma$; and

$$d(\Gamma) = \text{greatest lower bound of } \liminf_{n\to\infty} D_2(x_n, B_n)$$

for all sequences $\{(x_n, B_n)\}$ of mappings of class $\mathfrak{B}_2''$ in which $\Gamma_n \to \Gamma$, each Γ_n being the boundary of the surface S_n defined by (x_n, B_n); if no sequences exist satisfying the conditions, we define $l(\Gamma)$ or $d(\Gamma) = +\infty$. Finally, we define

$$d^*(\Gamma) = \begin{cases} +\infty \text{ if } k = 1 \\ \min \sum_{i=1}^{p} d(\Gamma^{(i)}), \qquad k > 1, \end{cases}$$

for all subdivisions of Γ into logically mutually exclusive systems of curves $\Gamma^{(1)}, \cdots, \Gamma^{(p)}$ for $2 \leqq p \leqq k$.

From the definitions of $L(S)$ and $L(x, B)$, it follows that $l(\Gamma)$ may be defined in exactly the same way as is $d(\Gamma)$, using the integral $\tilde{L}(x, B)$ instead of $D_2(x, B)$. The usual inequality between these integrals shows that $d(\Gamma) \geqq 2l(\Gamma)$. On the other hand, it is shown that if (x, B) is a representation of class $\mathfrak{B}_2''$ of the surface S, then S possesses another representation (y, D) of class $\mathfrak{B}_2''$ such that

$D_2(y, D) < 2L(S) + \epsilon$. Hence it follows that $d(\Gamma) = 2l(\Gamma)$. It is easily seen that both functionals are lower semi-continuous in Γ. An argument very similar to that of Courant shows that $d(\Gamma) \leqq d^*(\Gamma)$.

From the definition of $d(\Gamma)$, it follows if $d(\Gamma)$ is finite, that there is a sequence (x_n , B_n) of vectors of class $\mathfrak{P}_2''$ such that the boundaries Γ_n of the corresponding surfaces S_n tend to Γ and $D_2(x_n , B_n) \to d(\Gamma)$. Let Γ consist of the mutually exclusive Jordan curves $C_1 , \cdots , C_k$. Since $\Gamma_n \to \Gamma$, we may label the curves $C_{1,n}, \cdots , C_{k,n}$ of Γ_n so that $C_{i,n} \to C_i , i = 1, \cdots , k$. By using various mappings and certain lemmas of Courant, which carry over to the present case, we may assume that each B_n is bounded by fixed mutually exclusive circles $b_1 , \cdots , b_l , l \leqq k$, and possible variable circles $b_{i,n} , i = l + 1, \cdots , k$, in which (i) b_1 is the outer boundary and is the unit circle, (ii) x_n carries b_i into $C_{i,n} , i \leqq l$, and carries $b_{i,n}$ into $C_{i,n} , i > l$, (iii) the x_n are equicontinuous along b_1 , and (iv) the circles $b_{i,n} , i > l$, tend to points b_i .

The x_n and B_n having been normalized as above, the following undesirable situations might occur:

(i) $l < k$ and one of the b_i with $i > l$ is interior to the region B bounded by $b_1 , \cdots , b_l$,

(ii) $l < k$ and one of the $b_i , i > l$, is on a b_j with $j \leqq l$ along which the x_n are equicontinuous, and

(iii) $l \leqq k$ and the x_n fail to be equicontinuous along some b_i with $i \leqq l$.

If any of these cases occurs, the methods of Courant generalize almost verbatim to show that $d(\Gamma) = d^*(\Gamma)$.

Hence, if we impose Courant's condition that $d(\Gamma) < d^*(\Gamma)$, none of the cases of degeneracy can occur, $l = k$, and each B_n is just the limiting region B. By choosing a subsequence, we may assume that the $x_n(u, v)$ tend uniformly along B^* to a continuous function $x^*(u, v)$ which is a representation of Γ.

4. A solution of the problem of Plateau. In the case where the underlying manifold $\mathfrak{M}$ is ordinary Euclidean N-space, the solution follows immediately from the results of the preceding section. Since harmonic functions solve the Dirichlet problem, we may assume each x_n to be harmonic on B. Then the x_n converge uniformly on $\bar{B}$ to a harmonic function $x(u, v)$ which coincides with x^* on B^*. From the lower semi-continuity of the Dirichlet integral, it follows that $D_2(x, B) = d(\Gamma) = 2l(\Gamma) = 2\tilde{L}(x, B)$ so that $E = G, F = 0$ on B and the corresponding surface S is a surface of least area bounded by Γ.

In the case of the general manifold $\mathfrak{M}$, the writer was unable to demonstrate the existence of any minimizing sequence $\{x_n\}$ which is equicontinuous on $\bar{B}$. Accordingly, in order to solve the problem, the writer introduced and studied vector functions "of class $\mathfrak{P}_2$" (also of class $\mathfrak{P}_\alpha , \alpha \geqq 1$) with values in the Riemannian space $\mathfrak{M}$. These functions are entirely analogous to the ordinary functions of class $\mathfrak{P}_\alpha$ studied by Calkin and the writer in the joint paper [2] and [7]. In the case of a "homogeneously regular" (see below) manifold $\mathfrak{M}$, the writer found it possible to solve the problem without using the more general functions.

However, the developments were just as long and the results obtained were less general than those obtained in the paper [9]. Moreover the writer felt that the study of the more general functions would prove useful in connection with other problems.

We now present some definitions and facts concerning functions of class $\mathfrak{P}_2$, $\mathfrak{P}_2'$, $\mathfrak{P}_2''$, etc., with values in a general Riemannian space $\mathfrak{M}$. First of all, we say that $x(u, v)$ is *measurable* on the set U provided that for any open set $O \subset \mathfrak{M}$, the subset of U where $x(u, v) \in O$ is measurable. In this case U is measurable and the distance $| x(u, v) - x_0 |$ is measurable on U for any x_0 in $\mathfrak{M}$. We say that $x(u, v)$ is in L_2 on U if and only if $x(u, v)$ is measurable on U and $| x(u, v) - x_0 |^2$ is summable on U. The space L_2 is a complete metric space if equivalent functions are identified and the obvious metric is introduced. With each function x in L_2 on the set U, we associate the function $\bar{x}$ uniquely defined almost everywhere on U by the condition that $\bar{x}(u_0, v_0) = x_0$ for each (u_0, v_0) of U for which there exists a point x_0 in $\mathfrak{M}$ such that the Lebesgue derivative at (u_0, v_0) of the set function

$$\int_e | x(u, v) - x_0 |^2 \, du \, dv$$

is zero; $\bar{x}(u, v) = x(u, v)$ almost everywhere on U.

A function $x(u, v)$ is said to be *of class* $\mathfrak{P}_2'$ on the domain B if

(i) $x(u, v)$ is in L_2 on B,

(ii) $x(u_0, v)$ is absolutely continuous (AC) in v along each segment in B for almost all u_0 and is similarly AC in u for almost all v_0,

(iii) $| x_u |^2$ and $| x_v |^2$ are summable.

Without defining functions of class $\mathfrak{P}_2$ we state the result that $x(u, v)$ is of class $\mathfrak{P}_2$ on B if and only if the associated function $\bar{x}(u, v)$ is of class $\mathfrak{P}_2'$ on B and is such that $\bar{x}[u(s, t), v(s, t)]$ is also of class $\mathfrak{P}_2'$ in (s, t) whenever $u = u(s, t)$, $v = v(s, t)$ is a regular (i.e., it and its inverse have bounded derivatives) transformation of class C'; it is further shown that if $\bar{y}(s, t) = \bar{x}[u(s, t), v(s, t)]$, then the quantities E, F, and G for $\bar{x}$ and $\bar{y}$ are connected almost everywhere by the usual rules of the Calculus. Any function of class $\mathfrak{P}_2'$ or $\mathfrak{P}_2''$ is of class $\mathfrak{P}_2$ and any function of class $\mathfrak{P}_2$ which is continuous is of class $\mathfrak{P}_2''$.

Now suppose that $x(u, v)$ is of class $\mathfrak{P}_2$ on a domain B of class D' and suppose x is extended arbitrarily to $\bar{B}$. Then $\bar{x}(u, v)$ is defined almost everywhere on any regular arc or simple closed curve of class D' in $\bar{B}$ and is of class L_2 there (with respect to arc length). These values of $\bar{x}(u, v)$ on B^* are the *boundary values* of x; $\bar{x}(u, v)$ tends to these boundary values along almost all lines parallel to any given direction (but not necessarily in the two-dimensional sense). From the preceding paragraph one sees that the boundary values transform in the usual way under transformations of independent variables. Finally, a "substitution theorem" (like [7, Theorem 7.4]) is proved.

We say that x_n *tends weakly in* $\mathfrak{P}_2$ *to* x *on* B (strong convergence is not defined in [9]) if and only if

(i) $x(u, v)$ and each $x_n(u, v)$ are of class $\mathfrak{P}_2$ on B,

(ii) $D_2(x_n , B)$ is uniformly bounded, and

(iii) x_n converges to x in L_2 on each cell interior to B. If B is of class D', then weak convergence in $\mathfrak{P}_2$ on B of x_n to x implies strong convergence in L_2 on the whole of B and strong convergence of $\bar{x}_n$ to $\bar{x}$ on any regular arc or simple closed curve of class D' in $\bar{B}$. Of course weak convergence on B implies weak convergence on subregions and is preserved by changes of variable. Also the Dirichlet integral is lower semi-continuous with respect to weak convergence in $\mathfrak{P}_2$. Of course the integrals are to be formed using $\bar{x}$, etc. Finally, suppose $\{x_n\}$ is any sequence of class $\mathfrak{P}_2$ on B with $D_2(x_n , B)$ and

$$L_2(x_n , x_0 , B) = \iint_B | x_n(u, v) - x_0 |^2 \, du \, dv$$

uniformly bounded for some x_0 in $\mathfrak{M}$; then a subsequence converges weakly in $\mathfrak{P}_2$ to some function x on B. If we merely know that the $D_2(x_n , B)$ are uniformly bounded, $L_2(x_n , x_0 , B)$ will be also provided merely that either $L_2(x_n , x_0 , R)$ for some cell $R \subset B$ or the single integral $L_2^*(\bar{x}_n , x_0 , C)$ for some arc C of class D' in $\bar{B}$ is uniformly bounded.

Finally, it is shown that a certain "Dirichlet growth" condition implies a certain type of continuity on B; this material is entirely analogous to that found in [8, pp. 13–15].

Now, let us define $\tilde{d}(\Gamma)$ and $\tilde{d}^*(\Gamma)$ as $d(\Gamma)$ and $d^*(\Gamma)$ were defined, but allowing the more general functions; however, we assume that the boundary values of the $\bar{x}_n(u, v)$ are continuous and the B_n are of class D'. Then the results of the preceding section are unchanged. Since the boundary values are uniformly bounded, we may extract a further subsequence which converges weakly in $\mathfrak{P}_2$ to some function x in $\mathfrak{P}_2$ with boundary values x^*. From lower semi-continuity we find that $\tilde{d}(\Gamma) = D_2(x, B)$. Obviously x also solves the Dirichlet problem for the fixed boundary values x^* and thus is "harmonic".

In order to show that if x is harmonic, then $\bar{x}$ is continuous, the writer found it expedient to require that $\mathfrak{M}$ be *homogeneously regular*. By this is meant that there exist numbers m and M, independent of x_0 , with $0 < m \leqq M$, such that if x_0 is any point of $\mathfrak{M}$, there is a neighborhood $\mathfrak{N}$ of x_0 which can be mapped on the unit cube $| x^i | < 1$ by means of a transformation T of class C' in such a way that x_0 corresponds to the origin and the corresponding $g_{p,q}(x)$ satisfy the condition that

$$m \sum_{i=1}^{N} (\xi^i)^2 \leqq g_{p,q}(x)\xi^p\xi^q \leqq M \sum_{i=1}^{N} (\xi^i)^2$$

for all ξ and all x in the unit cube. It is easily seen that every compact manifold of class $C^{(n)}$, $n \geqq 1$, has this property so that the restriction refers only to the behavior of $\mathfrak{M}$ at infinity. The hyperbolic space used by Lonseth also has this property.

Thus, if $\mathfrak{M}$ is homogeneously regular, we conclude that $\tilde{d}(\Gamma) = d(\Gamma)$, $\tilde{d}^*(\Gamma) =$

$d^*(\Gamma)$, and $D_2(\bar{x}, B) = 2L(\bar{x}, B)$ so that $E = G$ and $F = 0$ almost everywhere and $(\bar{x}, B)$ defines a surface S of least area bounded by Γ.

That some restriction on $\mathfrak{M}$ is necessary to ensure the continuity of harmonic functions is seen by the following example: Let $\mathfrak{M}$ be the 3-dimensional manifold

$$\mathfrak{M}: \; w = f(x, y, z) = \frac{1}{2}\log\,(x^2 + y^2), \qquad x^2 + y^2 > 0.$$

Let

$$x(r, \theta) = R(r)\cos\theta, \qquad y(r, \theta) = R(r)\sin\theta, \qquad z(r, \theta) \equiv 0,$$

$$w(r, \theta) = \log R(r), \qquad 0 < r \leqq 1,$$

where $R(r)$ is determined by the relations

$$R^{-2}(R^2 + 1)^{1/2}R' = r^{-1}, \qquad R(1) = 1.$$

Then $R(0^+) = 0$ and the vector is of class $\mathfrak{P}_2$ and harmonic but there is no continuous vector whatever with the given boundary values.

5. The differentiability theory. In the writer's paper [9], the following results are proved concerning the differentiability of harmonic functions x (i.e., $\bar{x}$): If $\mathfrak{M}$ is of class C''' and homogeneously regular, then the harmonic functions are of class C''_γ on interior regions for each γ, $0 < \gamma < 1$; if $\mathfrak{M}$ if of class $C^{(n)}_\gamma$, $n \geqq 3$, $0 < \gamma < 1$, then the harmonic functions are of class $C^{(n)}_\gamma$ on interior regions. $\mathfrak{M}$ is of class $C^{(n)}_\gamma$ if $\mathfrak{M}$ is of class $C^{(n)}$, the functions in the connecting transformations for overlapping $\mathfrak{M}_i$ are of class $C^{(n)}_\gamma$, and, for each i, the $g^i_{p,q}(x)$ are of class $C^{(n-1)}_\gamma$; a function $\varphi(x)$ is of class $C^{(n)}_\gamma$ on the domain D if it is of class $C^{(n)}$ and its nth derivatives satisfy uniform Hölder conditions with exponent γ on D. Since each minimizing vector $(\bar{x}, B)$ obtained above is harmonic, these results apply to them.

To prove these results, we choose any circle $\overline{C(p_0, a)}$ interior to B with radius a so small that $\bar{x}(u, v)$ remains in some one of the neighborhoods $\mathfrak{M}_i$ for all $(u, v) \in \overline{C(p_0, a)}$. Let $\mathfrak{M}_i$ be mapped on the unit cube R by T_i. Since $\overline{C(p_0, a)}$ is closed, the components $x^j(u, v)$ with respect to T_i remain within the slightly smaller unit cube R_δ: $|x^j| < 1 - \delta$ on which

$$m\sum_{j=1}^{N}(\xi^j)^2 \leqq g^{(i)}_{\alpha\beta}(x)\xi^\alpha\xi^\beta \leqq M\sum_{j=1}^{N}(\xi^j)^2$$

for all ξ and some numbers m and M with $0 < m \leqq M$ and the $g_{\alpha\beta}(x)$ are of class C'' with uniformly bounded second derivatives, $\mathfrak{M}$ being of class C'''. Since $\bar{x}$ is harmonic on B, it is harmonic on $\overline{C(p_0, a)}$ and hence minimizes the Dirichlet integral

$$D_2[\bar{x}, C(p_0, a)] = \iint_{C(p_0,a)} g_{\alpha\beta}(x)(x^\alpha_u x^\beta_u + x^\alpha_v x^\beta_v)\,du\,dv$$

at least among all $y = \bar{x} + \lambda\xi$ for all ξ which satisfy uniform Lipschitz conditions on $\overline{C(p_0, a)}$ and vanish on $C^*(p_0, a)$, if λ is sufficiently small. This is sufficient to show that the first variation is zero for such ξ which in turn guarantees that $\bar{x}$ satisfies Haar's equations

$$\int_{R^*} (f_{p^i}\, dv - f_{q^i}\, du) = \iint_R f_{x^i}\, du\, dv, \qquad 1 \leqq i \leqq N,$$

$$f(x, p, q) = g_{\alpha\beta}(x)(p^\alpha p^\beta + q^\alpha q^\beta)$$

on almost all cells $R \subset C(p_0, a)$. Since x satisfies a Dirichlet growth condition (see [8, p. 13]) $A_2(2\lambda, P)$ on B, this implies in turn the existence of functions y and Y with the same property such that

$$(1)\quad y_{iu} - Y_{iv} = -f_{q^i}, \quad y_{iv} + Y_{iu} = f_{p^i}, \quad \int_{R^*} Y_{iu}\, dv - Y_{iv}\, du = \iint_R f_{x^i}\, du\, dv.$$

At this point, we confine ourselves to a circle $C(p_0, b)$ with $0 < b < a$, and apply the device of Lichtenstein: For each h with $0 < |h| < a - b$, we subtract equations (1) for (u, v) from those for $(u + h, v)$ [and also $(u, v + h)$ in turn] and divide by h, obtaining a system of the form

$$(2)\quad \begin{aligned} \eta_{iu} - H_{iv} &= -(b_{\beta i}\xi^\beta_u + c_{i\beta}\xi^\beta_v + e_{i\beta}\xi^\beta + k_i) \\ \eta_{iv} + H_{iu} &= a_{i\beta}\xi^\beta_u + b_{i\beta}\xi^\beta_v + d_{i\beta}\xi^\beta + g_i \end{aligned}$$

$$\int_{R^*} (H_{iu}\, dv - H_{iv}\, du) = \iint_R (d_{\beta i}\xi^\beta_u + e_{\beta i}\xi^\beta_v + f_{i\beta}\xi^\beta + l_i)\, du\, dv$$

where, for instance,

$$a_{ij}(u, v) = \int_0^1 f_{p^ip^j}[(1 - t)x(u, v) + tx(u + h, v), \cdots]\, dt.$$

The coefficients a_{ij}, b_{ij}, $\cdots$, l_i are all measurable, the a_{ij}, b_{ij}, and c_{ij} are bounded (continuous in our case) and satisfy

$$(3)\quad m\sum_{i=1}^N [(\zeta^i)^2 + (\omega^i)^2] \leqq a_{\alpha\beta}\zeta^\alpha\zeta^\beta + 2b_{\alpha\beta}\zeta^\alpha\omega^\beta + c_{\alpha\beta}\omega^\alpha\omega^\beta \leqq M\sum_{i=1}^N [(\zeta^i)^2 + (\omega^i)^2]$$

for almost all (u, v) and all ζ and ω, and the other coefficients satisfy

$$(4)\quad \iint_{C(q,r)\cap C(p_0,b)} \sum [d_{ij}^2 + e_{ij}^2 + |f_{ij}| + g_i^2 + k_i^2 + |l_i|]\, du\, dv \leqq M_2 r^{2\lambda}$$

for all q and r (since x satisfies a condition $A(2\lambda, P)$). These inequalities hold independently of h for $0 < |h| < a - b$. In (2) we have set

$$\xi^i = [x(u + h, v) - x(u, v)]/h \qquad (\text{or } [x(u, v + h) - x(u, v)]/h), \text{ etc.}$$

Moreover, we have

$$\iint_{C(p_0,b)} \sum [(\xi^i)^2 + (\eta_i)^2 + (H_i)^2]\, du\, dv \leqq M_3, \tag{5}$$

independently of h. Actually, for our special integral, we evidently have

$$a_{ij} = c_{ij}, \qquad b_{ij} = 0.$$

In my memoir [8], it was proved that any solution of (2) subject to hypotheses (3), (4), and (5), satisfies conditions A[2μ, $M(c, d)$] and B[μ, $N(c, d)$] on $C(p_0, b)$ where the functions $M(c, d)$ and $N(c, d)$ depend only on m, M, λ, M_2, M_3, μ, and N, and μ can be any positive number less than $m/4M$ and $\lambda/2$. Thus ξ, η, and H are equicontinuous on regions interior to $C(p_0, b)$ and so we may let $h \to 0$ and conclude that the derivatives $\bar{x}^i_u$, etc., satisfy such conditions. Another theorem in this memoir shows that the limiting ξ^i, η_i, and H_i satisfy the limiting equations (almost everywhere). Now if $\mathfrak{M}$ is of class C'''_γ, the $g_{ij}(\bar{x})$ are of class C''_γ and hence all the coefficients a_{ij} to l_i satisfy Hölder conditions. Other theorems in the memoir show then that the first derivatives of the ξ^i, i.e., the second derivatives of the $\bar{x}^i$, and those of η_i and H_i satisfy Hölder conditions. Knowing this much, we make use of the special form of our integral to show that the $\bar{x}^i$ are of class C'''_γ. The process may be repeated to obtain the higher differentiability properties.

Unfortunately the proof of the results stated concerning systems of type (2) is long and roundabout. Shiffman has succeeded in a recent paper [10] in simplifying this proof for the case that the coefficients d_{ij}, e_{ij}, and f_{ij} are all absent but has not succeeded in treating the general case. I have also tried without success to simplify the proof in the general case. I have however succeeded in simplifying and generalizing part of Shiffman's results by giving a very short proof of the following theorem for functions of any number of variables: Suppose the vector function (in Euclidean space) $x(u)$ is of class L_2 on $C(p_0, a)$ and is of class $\mathfrak{B}_2$ on any sphere $C(p_0, r)$ with $r < a$ with

$$D_2[x, C(p_0, r)] \leqq KD_2[H_r, C(p_0, r)], \qquad 0 < r < a,$$

H_r denoting the harmonic function coinciding with x on $C^*(p_0, r)$; then

$$D_2[x, C(p_0, r)] \leqq 4K^2(a - r)^{-2} \int_{C(p_0,a)} |x|^2\, du, \qquad 0 < r < a.$$

If we take $K = M/m$, this shows immediately, and without introducing the "conjugate functions" η_i and H_i, that any solution of the equivalent Haar equations to (2), in the case the d_{ij}, etc. all vanish, which satisfies (5) has uniformly bounded Dirichlet integrals on interior circles from which the A and B conditions follow easily by simple methods in the case of two independent variables. It is to be hoped that some similar device will be found which will handle the general case.

Bibliography

1. S. Bochner, *Harmonic surfaces in Riemann metric*, Trans. Amer. Math. Soc. vol. 47 (1940) pp. 146–154.

2. J. W. CALKIN, *Functions of several variables and absolute continuity*, I, Duke Math. J. vol. 6 (1940) pp. 170–186.

3. R. COURANT, *Plateau's problem and Dirichlet's principle*, Ann. of Math. (2) vol. 38 (1937) pp. 679–724.

4. A. LONSETH, *The problem of Plateau in hyperbolic space*, Amer. J. Math. vol. 64 (1942) pp. 229–259.

5. E. J. MCSHANE, *Integrals over surfaces in parametric form*, Ann. of Math. (2) vol. 34 (1933) pp. 815–838.

6. C. B. MORREY, JR., *A class of representations of manifolds* I, Amer. J. Math. vol. 55 (1933) pp. 683–707.

7. ———, *Functions of several variables and absolute continuity*, II, Duke Math. J. vol. 6 (1940) pp. 187–215.

8. ———, *Multiple integral problems in the calculus of variations and related topics*, University of California Publications in Mathematics, N. S., vol. 1. no. 1, 1943, pp. 1–130.

9. ———, *The problem of Plateau on a Riemannian manifold*, Ann. of Math. (2) vol. 49 (1948) pp. 807–851.

10. M. SHIFFMAN, *Differentiability and analyticity of solutions of double integral variational problems*, Ann. of Math. (2) vol. 48 (1947) pp. 274–284.

UNIVERSITY OF CALIFORNIA,
BERKELEY, CALIF., U. S. A.

ANALYSIS AND GEOMETRY IN THE LARGE

LAPLACE OPERATOR ON MANIFOLDS

S. BOCHNER

During the last fifteen years or so the so-called Laplace-Beltrami operator has been steadily gaining in importance and prominence in many areas of analysis, and I shall try to summarize here some of its applications with which I am familiar.

We take a compact differentiable manifold M. A Laplacian on it has the form

$$\Delta f = \frac{1}{g^{1/2}} \frac{\partial}{\partial x_i} \left(g^{1/2} g^{ij} \frac{\partial f}{\partial x_j} \right) \tag{1}$$

$$\equiv g^{ij} \frac{\partial^2 f}{\partial x_i \partial x_j} - g^{ij} \Gamma_{ij}^k \frac{\partial f}{\partial x_k}$$

where g^{ij} is a positive definite tensor field (without any "singularities") and g is the determinant of its inverse g_{ij}. The quantity $g^{1/2}$ is a scalar density. If we replace it by any other scalar density, and, correspondingly, replace (1) by the more general expression

$$Lf = \frac{1}{b} \frac{\partial}{\partial x_i} \left(b g^{ij} \frac{\partial f}{\partial x_j} \right) \tag{2}$$

$$\equiv g^{ij} \frac{\partial^2 f}{\partial x_i \partial x_j} + \frac{1}{b} \frac{\partial}{\partial x_i} (b g^{ij}) \frac{\partial f}{\partial x_j},$$

then the latter operator is self-adjoint with regard to the volume element

$$dv = b dx_1 \cdots dx_n , \tag{3}$$

in the sense that for any two functions φ, ψ both differentiable twice, we have

$$\int_M (\varphi L\psi - \psi L\varphi)\, dv = 0.$$

Conversely, if we envisage the most general elliptic operator having in each coordinate system the form

$$g^{ij} \frac{\partial^2 f}{\partial x_i \partial x_j} + a^k \frac{\partial f}{\partial x_k} \tag{4}$$

and if we demand that it be self-adjoint, in the manner just stated, with respect to a prescribed volume element (3), then we must have

$$a^k \equiv \frac{1}{b} \frac{\partial}{\partial x_i} (b g^{ik})$$

and (4) has of necessity the form (2).

PART I. VECTOR FIELDS

We shall first very briefly touch upon an application to differential geometry in which the only property of the Laplacian used explicitly will be the following lemma.

LEMMA 1. *If for a scalar function $f(x)$ on our compact manifold we have*

$$\Delta f \geqq 0 \tag{5}$$

everywhere, then we must have equality everywhere, that is,

$$\Delta f \equiv 0. \tag{6}$$

Now, the expression (1) can be written as

$$\Delta f = g^{ij} f_{,i,j} \tag{7}$$

where $f_{,i,j}$ is the second covariant derivative as formed with the Christoffel symbols based on g_{ij}. If now we take an arbitrary vector field ξ_a on M, form its square length

$$f = g^{ab}\xi_a\xi_b \tag{8}$$

and apply the Laplacian to it, then we obtain

$$\tfrac{1}{2}\Delta f = g^{ij}g^{ab}\xi_{a,i}\xi_{b,j} + g^{ab}(g^{ij}\xi_{a,i,j})\xi_b\,. \tag{9}$$

The first term on the right is always $\geqq 0$, and if by chance the second is also $\geqq 0$, then both must be 0 by the lemma. We now assume that ξ_a is the solution of an equation

$$g^{ij}\xi_{a,i,j} = A_a^c\xi_c \tag{10}$$

in which A_{ac} is a given symmetric tensor, so that the second term in (9) has then the value

$$A^{bc}\xi_b\xi_c\,, \tag{11}$$

and we are led to the following conclusion.

THEOREM 1. *If our symmetric tensor A_{ac} is strictly positive definite, then there is no solution of the differential equation* (10) *other than $\xi_a = 0$.*

If a vector field is harmonic (div $\xi = 0$ and curl $\xi = 0$), then it satisfies (10) with A_{ac} being $-R_{ac}$, where R_{ac} is the Ricci-tensor based on g_{ij}; if it is a Killing tensor ($\xi_{i,j} + \xi_{j,i} = 0$), then it satisfies it with $A_{ac} = R_{ac}$. We thus obtain the following theorem.

THEOREM 2. *If $-R_{ac}$ is positive definite (sphere, etc.), then there is no harmonic vector field, that is, the first Betti number is* 0, *and if $-R_{ac}$ is negative definite (constant negative curvature, etc.), then there is no Killing vector, and there exists no continuous group of isometries.*[1]

In the two-dimensional case there is even no continuous group of conformal homeomorphisms. In particular, for a closed Riemann surface of the classical type, of genus $p > 1$, there is only a finite number of conformal homeomorphisms, since by uniformizing it into the unit circle the surface acquires a non-Euclidean metric for which $-R_{ac} < 0$.[1]

More generally, in the case of several complex variables, no matter how many, if a compact manifold can be uniformized by a bounded domain in E_{2k}, then on such a compact manifold there exist only a finite number of complex isomorphisms. Algebraically put, if the complex domain is an algebraic variety and if we introduce the algebraic field of meromorphic functions on it, then the field has only a finite number of automorphisms.[2]

There are more complicated conclusions of this type involving higher Betti numbers and tensors of arbitrary order for both real and complex manifolds, compact ones or covering spaces of compact ones, but the principal point of reasoning is always Lemma 1 or a "localized" substitute thereof.[3]

PART II. ANALYTIC IMBEDDING

If we introduce the Hilbert space of L_2-functions relative to the volume element (3), then our formal operator (1) or, more generally, the formal operator (2) can be completed (that is, closed) to an operator which is self-adjoint, that is hypermaximal, in the technical sense,[4] and for the present we use only the following operational property of it.

It has a pure point-spectrum $\{-\rho_r\}$, and we thus obtain in our Hilbert space of L_2-functions a basis $\{\psi_r\}$ whose elements are twice differentiable solutions of the partial differential equations

$$-\Delta\psi_r = \rho_r\psi_r, \tag{12}$$

and we apply this in the following set-up.

Assume that M is a real analytic space. Nothing is known in general about the existence of analytic functions, scalars, or tensors, in the large. Assume, however, that we are also given the existence of an analytic positive definite tensor field g^{ij}, or what is the same, of an analytic Riemannian line-element

$$ds^2 = g_{ij}dx^i dx^j, \tag{13}$$

[1] *Vector fields and Ricci curvature*, Bull. Amer. Math. Soc. vol. 52 (1946) pp. 776–797.

[2] *On compact complex manifolds*, J. Indian Math. Soc. vol. 51 (1947) pp. 1–21.

[3] *Curvature and Betti numbers* (I) *and* II, Ann. of Math. vol. 49 (1948) pp. 379–390; vol. 50 (1949) pp. 77–93.

[4] *Analytic mapping of compact Riemann spaces into Euclidean space*, Duke Math. J. vol. 3 (1937) pp. 339–354.

without any singularities, that is. We then envisage the equations (12), and since all its coefficients are analytic, then by a theorem originally due to Hadamard, the functions ψ_r themselves are likewise analytic. Thus the one analytic tensor field g_{ij} known to exist generates a multitude of analytic scalar functions $\{\psi_r\}$, and it can be shown that the latter are dense in every differentiability class C^s on M in the following very precise sense.

LEMMA 2. *To any f in C^s there exists a sequence of finite linear combinations of the ψ_r which converge uniformly towards f, together with all partial derivatives of order $\leqq s$.*

By a general theorem of H. Whitney, our M_k can be topologically imbedded (C^1) in the Euclidean E_{2k+1}, and by applying Lemma 2 for $s = 1$ we see that by a small deformation Whitney's imbedding can be made into an analytic one.

Also, once this analytic imbedding has been achieved the previous statement for arbitrary s (including $s = \infty$) follows trivially, and not only for scalars, but for arbitrary tensors and tensor densities

$$\xi_{a_1 \cdots a_p}^{b_1 \cdots b_q}$$

as well. That is to say, the following conclusion holds.

THEOREM 3. *If on a compact analytic M_k there exists an analytic positive definite g_{ij}, then M_k can be mapped analytically topologically into the Euclidean E_{2k+1}, and in any type of scalars or tensors of differentiability class C^s, the analytic ones are C^s-dense.*[4]

PART III. STOCHASTIC PROCESSES[5]

We are returning to a differentiable M, and we are stating the operational properties of our operators much more fully than heretofore. The operator $-Lf$ has distinct eigenvalues λ_r such that

$$0 = \lambda_0 < \lambda_1 < \lambda_2 < \cdots \to \infty. \tag{14}$$

The smallest eigenvalue has multiplicity 1, and every other λ_r has a finite multiplicity μ_r at most. There is a complete orthonormal set of eigenfunctions $\varphi_{r\mu}$ pertaining to class C^2 (with φ_{01} being a constant) such that the equations

$$-L\varphi_r = \lambda_r \varphi_{r\mu}, \qquad \mu = 1, \cdots, \mu_r, \tag{15}$$

$r = 0, 1, \cdots$, are satisfied in the ordinary sense. If we put

$$g_r(x, y) = \sum_{\mu=1}^{\mu_r} \varphi_{r\mu}(x)\varphi_{r\mu}(y) \tag{16}$$

and introduce the kernel function

[5] In connection with parts III and IV see the author's paper *Quasi-analytic functions, Laplace operator, positive kernels*, Ann. of Math. vol. 51 (1950) pp. 68–91.

$$G(t; x, y) = \sum_{r=0}^{\infty} e^{-t\lambda_r} g_r(x, y) \tag{17}$$

for

$$0 < t < \infty, \tag{18}$$

then it has the important property

$$G(t; x, y) \geqq 0 \tag{19}$$

first established by Kolmogoroff in 1933,[6] and also the properties

$$\int_M G(t; x, y)\, dv = 1 \tag{20}$$

$$\int_M G(t_1; x, y) G(t_2; y, z)\, dv_y = G(t_1 + t_2; x, z), \tag{21}$$

which three properties make it into a so-called stationary stochastic process on M. Also, if for fixed y we denote $f(t; x, y)$ by $f(t; x)$, then it satisfies the "heat equation"

$$L_x f(t; x) = \frac{\partial f}{\partial t}, \tag{22}$$

the left side of (22) being uniquely determined by it, and if we take an arbitrary solution of (22) having boundary values $f(0; x) \equiv f(x)$, then its values for $t > 0$ are given by the formula

$$f(t; x) = \int_M f(0; y) G(t; x, y)\, dv_y .^{7} \tag{23}$$

We next take in $0 \leqq \lambda < \infty$ a continuous function $\Phi(\lambda)$, $\Phi(0) = 0$, such that for every $t > 0$ we have a Stieltjes representation

$$e^{-t\Phi(\lambda)} = \sum e^{-\rho\lambda} d\gamma_t(\rho) \tag{24}$$

with $d\gamma_t(\rho) \geqq 0$ [that is, $\gamma_t(\rho)$ monotonely nondecreasing in $0 \leqq \rho \leqq +\infty$], and this amounts to assuming that we have $\Phi(\lambda) > 0$, $(-1)^{n-1}\Phi^{(n)}(\lambda) > 0$, $n = 1, 2, \cdots$, in $0 < \lambda < \infty$. If we now set up the expression

[6] A. Kolmogoroff, *Zur Theorie der stetigen zufälligen Prozesse*, Math. Ann. vol. 108 (1933) pp. 149–160.

[7] In the work of Kolmogoroff the self-adjointness of the operator is not required. To this compare K. Yosida, *Integration of Fokker-Planck's equation in a compact Riemannian space*, Arkiv för Matematik vol. 1 (1949) no. 9. However, under self-adjointness, the kernel with the properties (14)–(21) has the additional property of converging towards "identity" as the time variable tends to infinity. It would be interesting to decide to what extent self-adjointness is also necessary for this to take place. See also the author's note, *Diffusion equation and stochastic processes*, Proc. Nat. Acad. Sci. U. S. A. vol. 35 (1949) pp. 368–370.

$$G_\Phi(t; x, y) = \int_0^\infty e^{-t\Phi(\lambda_n)} g_n(x, y), \tag{25}$$

then owing to the representation

$$G_\Phi(t; x, y) = \int_0^\infty G(\rho; x, y) d\gamma_t(\rho) \tag{26}$$

it also satisfies the "stochastic" properties (19), (20), (21); and instead of (22) it satisfies the "heat equation"

$$-\Phi(-L)f = \frac{\partial f}{\partial t} \tag{27}$$

where $\Phi(-L)$ is the self-adjoint operator resulting from applying the numerical function $\Phi(\lambda)$ to the operator $-L$.

For $\Phi(\lambda) = \lambda$, G_Φ is the original kernel itself, and other very important cases arise on putting $\Phi(\lambda) = \lambda^\sigma$, $0 < \sigma < 1$. For instance, if M is the ordinary circle $-\pi \leqq x < \pi$, and (22) is

$$\frac{d^2 f}{dx^2} = \frac{\partial f}{\partial t}, \tag{28}$$

then we have $G(t; x, y) = G(t; x - y)$ where

$$G(t; x) = \left(\frac{\pi}{t}\right)^{1/2} \sum_{m=-\infty}^{\infty} e^{-(2\pi m - x)^2/t} \tag{29}$$

and

$$G_{\lambda^{1/2}}(t; x) = \frac{1 - e^{-2t}}{1 - 2e^{-t}\cos x + e^{-2t}}. \tag{30}$$

We now take on our M a fixed point P_0 and we consider on M all possible paths emanating from it, $x = P(t)$, $0 \leqq t \leqq 1$, $P(0) = P_0$, whether continuous or not. Now, the function $G(t; x, y)$, or any other one of our functions $G_\Phi(t; x, y)$, since it satisfies (19)–(21), can be used, by a fixed procedure, to introduce a probability (that is, a Lebesque measure of total measure 1) into the space of all such paths,[8] and we are going to state two properties of this probability, the first due to Paul Lévy,[9] and the second to Paul Lévy[9] and the author.[10]

(i) In the original case $\Phi(\lambda) = \lambda$, *and only in this case*, almost all paths are continuous in $0 \leqq t \leqq 1$.

(ii) However for $\Phi(\lambda) = \lambda$, and more generally for $\Phi(\lambda) = \lambda^\sigma$, $1/2 \leqq \sigma < 1$, almost all paths are *not* rectifiable, and thus in particular will have simple discontinuities only. More generally, if

[8] *Partial ordering in theory of stochastic processes*, Proc. Nat. Acad. Sci. U. S. A. vol. 36 (1950) pp. 439–443.

[9] P. Lévy, *Théorie de l'addition des variables aléatoires*, Paris, 1937, chapter VII, and in particular pp. 180, 199, 201.

[10] *Stochastic processes*, Ann. of Math. vol. 48 (1947) pp. 1014–1060, in particular pp. 1031–1037.

$$\int_1^\infty \frac{\Phi(\lambda)}{\lambda^{3/2}}\, d\lambda < \infty, \tag{31}$$

then almost all paths are rectifiable, and if

$$\int_1^\infty \frac{\Phi(\lambda)}{\lambda^{3/2}}\, d\lambda = \infty, \tag{32}$$

they are not.

Part IV. Quasi-analytic functions

We again assume that the space M and the coefficients of Lf are analytic or at least infinitely often differentiable. Any (continuous) function $f(x) \equiv f(0; x)$ has an expansion

$$f(x) \sim \sum_{r=0}^{\infty} \sum_{\mu=1}^{\mu_r} c_{r\mu}\, \varphi_{r\mu}, \tag{33}$$

and on putting

$$f_r(x) = \sum_{\mu=1}^{\mu_r} c_{r\mu}\, \varphi_{r\mu} \tag{34}$$

we write for this

$$f(x) \sim \sum_{r=0}^{\infty} f_r(x), \tag{35}$$

where each $f_r(x)$ is some eigenfunction

$$-Lf_r(x) = \lambda_r f_r(x). \tag{36}$$

The solution $f(t; x)$ of (22) as given by (23) has the (convergent) expansion

$$f(t; x) = \sum_{r=0}^{\infty} e^{-t\lambda_r} f_r(x), \tag{37}$$

and from (19) we conclude

$$\max_{t,x} | f(t; x) | \leqq \max_x | f(0; x) |. \tag{38}$$

If L^n is the nth iterate of our operator, then (22) implies

$$L_x^n f = \frac{\partial}{\partial t} L^{n-1} f = \frac{\partial^n f}{\partial t^n}, \tag{39}$$

and this gives the relation

$$\left| \frac{\partial^n f(t; x)}{\partial t^n} \right| \leqq \max_x \left| L^n f(x) \right| \tag{40}$$

which we shall employ in connection with

$$\left. \frac{\partial^n f(t; x)}{\partial t^n} \right|_{t=0} = L^n f(x) \tag{41}$$

in the following manner. For fixed $x = \xi$ we shall view $f(t; \xi)$ as a function of t in $0 \leqq t < \infty$, and to this function we shall apply Carleman's theorem on quasi-analytic functions. Also, $f(t; \xi)$ is a Dirichlet series in t, and if it vanishes in t, all its coefficients must be zero, and we therefore obtain the following lemma.

LEMMA 3. *If for an infinitely differentiable function* (35) *we have*

$$\sum_{n=0}^{\infty} \max_x \left| L^n f(x) \right|^{-1/n} = \infty \tag{42}$$

and if for a point $x = \xi$ *we have*

$$L^n f(\xi) = 0, \qquad n = 0, 1, 2, \cdots, \tag{43}$$

then at this point we have

$$f_r(\xi) = 0, \qquad r = 0, 1, 2, \cdots. \tag{44}$$

We call a pointset U on M a set of uniqueness if any eigenfunction of $-Lf$ vanishing on U must vanish identically, and in this way we obtain the following theorem.

THEOREM 4. *If in addition to* (42), (43) *holds for all points of a pointset of uniqueness, then* $f(x)$ *vanishes identically.*

For the operator $\Delta f = d^2f/dx^2$ on the torus $-\pi \leqq x < \pi$, the eigenfunctions are

$$a \cos rx + b \sin rx = c \cos r\,(x - x_0),$$

and a pointset U is therefore a set of uniqueness if it contains infinitely many points or if it contains two points whose geodesic distance is an irrational part of 2π. Hence the following conclusion.

THEOREM 5. *If for an infinitely differentiable periodic function* $f(x)$ *we have*

$$\sum_{n=0}^{\infty} \max_x \left| \frac{d^{2n} f(x)}{dx^{2n}} \right|^{-1/n} = \infty \tag{45}$$

and if we have

$$\frac{d^{2n} f(\xi)}{d\xi^{2n}} = 0, \qquad n = 0, 1, \cdots \tag{46}$$

for infinitely many points, ξ, *or for two points whose distance is irrational on the torus of length one, then*

$$f(x) = 0. \tag{47}$$

Further remark: If we take a function $\Phi(\lambda)$ as in §4, then we may replace $-L$ by $\Phi(-L)$, and we obtain the following conclusion.

THEOREM 6. *If in addition to*

$$\sum_{n=0}^{\infty} \max_x \left| \Phi(-L)^n f \right|^{-1/n} = \infty \tag{48}$$

we have

$$\Phi(-L)^n f(\xi) = 0 \tag{49}$$

for all ξ of a uniqueness set, then

$$f(x) \equiv 0. \tag{50}$$

PART V. A THEOREM OF D. V. WIDDER AND P. LELONG

D. V. Widder[11] has proved the following. If a function $f(x)$ is infinitely differentiable in $0 \leqq x \leqq 1$ and if

$$(-1)^n \frac{d^{2n} f(x)}{dx^{2n}} \geqq 0 \qquad \text{in} \qquad 0 \leqq x \leqq 1, \tag{51}$$

then $f(x)$ is analytic in $0 \leqq x \leqq 1$.

More precisely, $f(x)$ can be analytically continued to an entire function $f(z)$, $z = x + iy$, and its order of magnitude is, roughly speaking, comparable to that of the function $\sin \pi z$, which is the standard function for satisfying (51), as it were.

Now P. Lelong[12] has generalized this to a Laplacian on a noncompact M. If we have, for a function $f(x)$, the condition

$$(-1)^n L^n f \geqq 0, \qquad n = 0, 1, \cdots, \tag{52}$$

then $f(x)$ is a real analytic function.

Also it can be continued into a complex analytic environment of M, $z_j = x_j + iy_j$, and Lelong has some statements on the size of this environment, and of the magnitude of $f(z_1, \cdots, z_k)$ in it.

PART VI. MEROMORPHIC FUNCTIONS ON ELLIPTIC SPACES

In the theorem of Widder and Lelong use has to be made of a principal solution of the Laplacian with the properties

$$\Delta_x H(x; \xi) = 0, \qquad H(x; \xi) = H(\xi; x)$$

among others, such a principal solution being the means for solving the boundary value problem of the ordinary kind.

On a compact manifold M_n there is only a generalized solution with

$$\Delta_x H(x; \xi) = -1,$$

[11] D. V. Widder, *Completely convex functions and Lidstone series*, Trans. Amer. Math. Soc. vol. 51 (1942) pp. 387–398.

[12] P. Lelong, *Sur les fonctions indéfiniment dérivables de plusieurs variables dont les Laplaciens successifs ont des signes alternés*, Duke Math. J. vol. 14 (1947) pp. 143–149.

and the corresponding type of boundary value problem, as first envisaged by Riemann, is the construction of harmonic functions having prescribed saltuses on designated pieces of hypersurfaces. The principal solution available is indeed suitable for treating this type of problem and one significant application is as follows.[1]

Suppose M_n is real analytic, say, and assume that it is covered by a finite number of functional elements $\{\varphi_a\}$, $a = 1, \cdots, N$, such that whenever two functional elements φ_a, φ_b overlap, their difference

$$\varphi_{ab} = \varphi_a - \varphi_b \tag{53}$$

shall be a solution of the Laplacian, that is

$$\Delta\varphi_{ab} = 0.$$

The individual elements φ_a by themselves need not be so, and furthermore the individual φ_a may have "singularities" on "exceptional" pieces of space, provided such singularities happen to cancel out from the differences φ_{ab} by ordinary analytic continuation. "Meromorphic" character expresses itself in this type of singularities, but there are singularities admissible for us which are much more general than meromorphic ones.

Now, the generalized principal solution can be used for constructing a new family of functional elements $f_{(a)}$ having the following properties: (i) each $f_{(a)}$ is defined where φ_a was defined, and has no singularities whatsoever; (ii) each $f_{(a)}$ is harmonic individually; and (iii) for all a, b we have

$$f_{(a)} - f_{(b)} = \varphi_a - \varphi_b .$$

By (iii) the differences

$$\varphi_a - f_{(a)} = \varphi_b - f_{(b)} = \varphi_c - f_{(c)} = \cdots$$

will then merge into *one* function F on M_n, and this function F will then have the same singularities as the given elements $\{\varphi_a\}$, the latter being "principal singular parts" of F, as it were.

We now make the special assumption that M_n is complex analytic, $n = 2k$; that the metric on it is Hermitian

$$ds^2 = g_{\alpha\bar{\beta}} dz_\alpha d\bar{z}_\beta ,$$

with the Kaehler property

$$\frac{\partial g_{\alpha\bar{\beta}}}{\partial z_\gamma} = \frac{\partial g_{\gamma\bar{\beta}}}{\partial z_\alpha},$$

in which case

$$\Delta\varphi \equiv g^{\alpha\bar{\beta}} \frac{\partial^2 \varphi}{\partial z_\alpha \, \partial \bar{z}_\beta};$$

that the φ_a are complex-meromorphic, that is, $\partial\varphi/\partial\bar{z}_\alpha = 0$, and a fortiori $\Delta\varphi_a = 0$;

and that their differences φ_{ab} are complex holomorphic. This in itself does not yet sharpen any of the properties of the $f_{(a)}$ previously stated. If however we are adding the assumption: $-R_{\alpha\bar{\beta}} > 0$, (complex projective space, complex Grassmann varieties, etc.), then by the method of part I it can be proved that the vector field

$$\zeta_{\bar{a}} = \frac{\partial f_{(a)}}{\partial \bar{z}_\alpha} \left(\equiv \frac{\partial f_{(b)}}{\partial \bar{z}_\alpha} \right)$$

is identically zero, so that each $f_{(a)}$ is not only harmonic but also complex-holomorphic and the following precise conclusion can be drawn:[1]

THEOREM 7. *If $-R_{\alpha\bar{\beta}}$ is positive definite (complex projective space, complex Grassmann varieties, etc.), and if the space is covered by a finite number of meromorphic functional elements $\{\varphi_a\}_{a=1,\dots,N}$, and if the differences $\varphi_a - \varphi_b$ are holomorphic wherever φ_a, φ_b overlap; then there exists one joint meromorphic function Φ on M_{2k} having precisely the singularities of the given functional elements; meaning that $\Phi - \varphi_a$ is holomorphic for each a on the piece of space on which φ_a is defined.*

This generalizes the "trivial" theorem of the ordinary Gaussian sphere, $k = 1$, stating that there exists on it a rational function having at prescribed poles z_a prescribed principal parts

$$\varphi_a = \sum_{\nu=1}^{n_a} \frac{c_{a\nu}}{(z - z_a)^\nu},$$

$[\varphi = \sum_{\nu=1}^{n} c_\nu z^\nu \text{ for } z_a = \infty]$, the desired function Φ being then trivially the sum $\sum_a \varphi_a$. For $k > 1$ the construction of Φ as indicated above is very much more complicated than that, and its very existence is far from obvious.

PART VII. MEROMORPHIC FUNCTIONS ON MULTI-TORI

In Theorem 7, the restriction on $-R_{\alpha\bar{\beta}}$ to be positive definite is a very essential one even for $k = 1$, since already for the ordinary elliptic torus [having a flat metric with $R_{1\bar{1}} = 0$] the known theorem of Liouville for the existence of elliptic functions with given principal parts demands that the sum of the residues of the principal parts shall have value 0. Now, for $k = 1$, Liouville's condition is not only necessary but also sufficient and for arbitrary k it can be generalized in the following manner.[18]

We take a complex multi-torus M_{2k} whose complex universal covering space is the complex Euclidean E_{2k}, and on it a simplicial decomposition $\{S_a\}$, $a = 1, \cdots, N$, and for each S_a a complex-meromorphic element φ_a is a neighborhood of its closure such that whenever two simplices S_a, S_b intersect in a $(2k-1)$-dimensional "oriented" face B_{ab}, the difference (53) is complex-holomorphic in a neighborhood of its closure. In order that there exist a meromorphic function

[18] *Analytic and meromorphic continuation by means of Green's formula*, Ann. of Math. vol. 44 (1943) pp. 652–673, especially Theorem 15 on p. 672.

Φ having the given φ_a's as principal parts as in Theorem 7, it is necessary and sufficient that for each $\alpha = 1, \cdots, k$ we have

$$\sum_{a,b} \int_{B_{a,b}} \varphi_{ab}\, d\zeta_\alpha\, \omega^{k-1} = 0,$$

where $\omega \equiv dz_1 d\bar{z}_1 + \cdots + dz_k d\bar{z}_k$ is the basic external form associable with the flat Euclidean metric given naturally.

Part VIII. Analytic completion

One of the most striking differences between functions in one and several complex variables is as follows. For $k = 1$, to any domain D there exists a function $f(z)$ which is analytic in D and in no larger domain; whereas for $k \geqq 2$ there are pairs of domains D, $\tilde{D}$, with $D \subset \tilde{D}$, such that any function of D is continuable into $\tilde{D}$. Also, perhaps the most significant statement bearing on the latter phenomenon is the following theorem which was first stated by F. Hartogs (virtually without a proof by himself). For $k \geqq 2$, if a bounded domain D has a boundary B consisting of one piece, and if a function $f(z_1, \cdots, z_k)$ is given in a neighborhood S of B, then it can be continued into $D + S$.

Now, analyticity means

$$\frac{\partial f}{\partial \bar{z}_\alpha} = 0, \qquad \alpha = 1, \cdots, k,$$

and thus our function in S satisfies in particular the Laplace equation

$$\frac{\partial^2 f}{\partial x_1^2} + \frac{\partial^2 f}{\partial y_1^2} + \cdots + \frac{\partial^2 f}{\partial x_k^2} + \frac{\partial^2 f}{\partial y_k^2} = 0$$

in $n = 2k$ variables, and also the equation

$$\frac{\partial^2 f}{\partial x_1^2} + \frac{\partial^2 f}{\partial y_1^2} = 0$$

in $m = 2$ variables, and in the paper cited before[13] we have given a sweeping generalization of Hartog's theorem to the following effect.

Theorem 8. *If a real analytic function $f(x_1, \cdots, x_n)$ is given in a neighborhood S of the connected boundary of a bounded domain D, then it can be continued into all of $D + B$ provided it is in S the joint solution of the Laplace equation*

$$\frac{\partial^2 f}{\partial x_1^2} + \cdots + \frac{\partial^2 f}{\partial x_n^2} = 0 \tag{54}$$

in all n-variables, and of some other (inhomogenous) linear partial differential equation with constant coefficients

$$\sum_{0 \leqq r_1 + \cdots + r_m \leqq R} a_{r_1 \cdots r_m} \frac{\partial^{r_1 \cdots r_m} f}{\partial x_1^{r_1} \cdots \partial x_m^{r_m}} = 0 \tag{55}$$

in the first m variables, with $m < n$.

In the paper cited we have assumed that the second operator (55) shall be of elliptic type in a certain sense, but this assumption can be abandoned altogether, as we shall show in a future publication, and we shall then also show that the first operator may be one of elliptic type with constant coefficients with which a suitable Green's kernel can be associated, as is, for instance, the case with the powers of the Laplacian Δf.

We also generalized Hartog's theorem in other directions and one of the conclusions obtained was as follows.

THEOREM 9. *If $f(z_1, \cdots, z_k)$ is analytic in a domain which is of the type of a cell, and if the set of the zeros of the function in the domain contains a bounding cycle of dimension $2k - 2$, then the function vanishes identically.*

PRINCETON UNIVERSITY,
PRINCETON, N. J., U. S. A.

LA THÉORIE DES POINTS FIXES ET SES APPLICATIONS EN ANALYSE

JEAN LERAY

À la mémoire du profond mathématicien polonais JULES SCHAUDER, victime des massacres de 1940.

I. INTRODUCTION

1. Soit $\phi(x)$ une application continue d'un espace X en lui-même; on nomme points fixes de $\phi(x)$ les solutions de l'équation

$$x = \phi(x). \tag{1}$$

Nous ne parlerons pas de *l'étude locale* de l'équation (1). Cette étude fut faite d'abord par E. Picard [11], à l'aide de la méthode des approximations successives; puis par E. Schmidt [15], à l'aide de développements en séries, $\phi(x)$ étant supposée holomorphe. La notion d'espace de Banach permit à T. H. Hildebrandt et L. M. Graves [3] de systématiser la méthode de E. Picard; il est aisé [9] de systématiser de même celle de E. Schmidt.

C'est de *l'étude globale* de l'équation (1) que nous nous occuperons. Cette étude fut faite d'abord par Fredholm [4], F. Riesz [12], quand $\phi(x)$ est linéaire et transforme les parties bornées de X en parties compactes; puis, quand $\phi(x)$ n'est pas linéaire, par L. E. J. Brouwer [2], Birkhoff et Kellogg [1], Lefschetz [5], Schauder [14], Leray [6], [7], Rothe [13], Tychonoff [16], Nielsen [10], et Wecken [17]; deux types d'hypothèses[1] furent utilisés et conduisirent à des théories bien différentes: certains auteurs supposèrent que X est un espace vectoriel et que $\phi(x)$ prend ses valeurs dans un compact; d'autres supposèrent que X est compact et vérifie des hypothèses appropriées. Ces hypothèses compliquent ce second point de vue, que nous n'aurons pas le temps d'analyser en détail; c'est d'ailleurs le premier point de vue qui se présente quand on applique la théorie des points fixes à celle des équations aux dérivées partielles. Exposons-le d'abord, en résumant [9], qui synthétise [2], [1], [14], [6], [16], [10], [17].

II. LES POINTS FIXES D'UNE APPLICATION COMPLÈTEMENT CONTINUE D'UN ESPACE VECTORIEL À VOISINAGES CONVEXES

2. Définitions. Soit X *un espace vectoriel à voisinages convexes*: c'est un espace vectoriel (sur le corps des nombres réels) possédant une topologie de Hausdorff, qui puisse être définie par un système fondamental de voisinages convexes. Soit V un voisinage symétrique de 0; les points x_1 et x_2 de X sont dits *voisins d'ordre* V quand

$$x_1 - x_2 \in V.$$

[1] D'autres hypothèses furent utilisées avec succès par E. Rothe; nous ne disposons malheureusement pas de la place qu'exigerait l'exposé de ses recherches.

Soit $\phi(x)$ une application de X en lui-même *complètement continue*, c'est-à-dire qui applique continûment X dans une partie compacte de X; nous posons

$$\Phi(x) = x - \phi(x).$$

G désignera *une partie ouverte* de X, $\dot{G}$ sa *frontière* et $\bar{G} = G \cup \dot{G}$ son *adhérence*.

3. Les propriétés de $\Phi(x)$, du point de vue de la topologie générale. $\Phi(F)$ est *fermé*, quand F est *fermé* (autrement dit: l'application $\Phi(x)$ est fermée). $\Phi^{-1}(C)$ est *compact*, quand C est *compact*.

4. La définition du degré topologique de Φ (x). *Supposons X de dimension finie et $\phi(x)$ linéaire par morceaux*, c'est-à-dire linéaire au voisinage de tout point n'appartenant pas à la réunion d'un ensemble d'hyperplans P_λ, n'ayant pas d'élément d'accumulation. Ces hyperplans décomposent X en domaines, que nous noterons D_μ^+, D_ν^0, D_ρ^- suivant que le déterminant de $\Phi(x)$ y est >0, $=0$, <0. Soit y un point de X étranger aux $\Phi(\dot{G})$, $\Phi(P_\lambda)$, et $\Phi(D_\nu^0)$; soit p [et n] le nombre des $\Phi(G \cap D_\mu^+)$ [et des $\Phi(G \cap D_\rho^-)$] contenant y; $p - n$ est une fonction de y constante sur chacun des domaines d en lesquels $\Phi(\dot{G})$ décompose X; sa valeur est nommée *degré topologique* sur d de la restriction de Φ à G et est notée

$$d^0(\Phi, G, d);$$

on définit

$$d^0(\Phi, G, y) = d^0(\Phi, G, d) \qquad \text{si } y \in D$$

même si y appartient à $\Phi(P_\lambda)$ ou $\Phi(D_\nu^0)$.

Supposons X de dimension finie et $\phi(x)$ complètement continue; soit $y \notin \Phi(\dot{G})$; soit V_1 un voisinage convexe et symétrique de 0 tel que le voisinage d'ordre V_1 de y soit étranger à $\Phi(\dot{G})$; soit $\Phi_1(x)$ une application linéaire par morceaux telle que $\Phi(x)$ et $\Phi_1(x)$ soient voisins d'ordre V_1 ; $d^0(\Phi_1, G, y)$ est indépendant des choix de V_1 et Φ_1 ; c'est, par définition, $d^0(\Phi, G, y)$.

Cas général. Soit $y \notin \Phi(\dot{G})$; soit V_1 un voisinage convexe et symétrique de 0, tel que le voisinage d'ordre V_1 de y soit étranger à $\Phi(\dot{G})$; soit $\phi_1(x)$ une application complètement continue, voisine d'ordre V_1 de $\phi(x)$ et telle que $\phi_1(X)$ appartienne à un sous-espace X_1 de dimension finie, contenant y; $d^0(\Phi_1, G \cap X_1, y)$ est indépendant des choix de V_1, $\Phi_1(x) = x - \phi_1(x)$, X_1 ; c'est par définition $d^0(\Phi, G, y)$.

5. Les propriétés du degré topologique.

PROPRIÉTÉ 5.1. *$d^0(\Phi, G, y)$ est un entier positif, nul ou négatif, défini quand $\Phi(x) - x$ est complètement continue et que $y \notin \Phi(\dot{G})$; $d^0(\Phi, G, y)$ reste constant quand Φ, G, y varient continûment, en sorte que $y \notin \Phi(\dot{G})$.*

En particulier, $d^0(\Phi, G, y)$ ne dépend que de y et de la restriction de Φ à $\dot{G}$; on peut même, ce qu'utilise le §7, définir $d^0(\Phi, G, y)$ en supposant Φ défini seulement sur $\dot{G}$.

PROPRIÉTÉ 5.2. *Si $d^0(\Phi, G, y) \neq 0$, alors $y \in \Phi(\bar{G})$.*

PROPRIÉTÉ 5.3. *Si les G_α sont des parties ouvertes de G, deux à deux sans point commun et telles que $\Phi(x) \neq y$ quand $x \in G$, $x \notin G_\alpha$, alors*

$$d^0(\Phi, G, y) = \sum_\alpha d^0(\Phi, G_\alpha, y).$$

PROPRIÉTÉ 5.4. *Soit $\Psi(x) - x$ une seconde application complètement continue, définie sur $\Phi(\bar{G})$; soient d_α les domaines en lesquels $\Phi(\dot{G})$ décompose X [autrement dit: les d_α sont les composantes connexes du complémentaire de $\Phi(\dot{G})$]; si $y \notin \Psi\Phi(\dot{G})$, alors*

$$d^0(\Psi\Phi, G, y) = \sum_\alpha d^0(\Phi, G, d_\alpha) \cdot d^0(\Psi, d_\alpha, y).$$

PROPRIÉTÉ 5.5. *Supposons X somme directe de deux espaces X_1 et X_2 :*

$$X = X_1 + X_2 ;$$

on a

$$x = x_1 + x_2 ; \qquad \Phi(x) = \Phi_1(x) + \Phi_2(x), \qquad \text{où } x_\alpha \in X_\alpha, \Phi_\alpha \in X_\alpha ;$$

on a $\Phi_1(x) = \Phi_1(x_1, x_2)$; supposons $\Phi_2(x) = \Phi_2(x_2)$ fonction seulement de x_2. Soit G_1 une partie ouverte de X_1 et D_2 un domaine de X_2 ; si $y_1 \notin \Phi_1(\dot{G}_1, D_2)$ et $y_2 \notin \Phi_2(\dot{D}_2)$, alors

$$d^0(\Phi, G_1 + D_2, y_1 + y_2) = d^0(\Phi_1, G_1, y_1) \cdot d^0(\Phi_2, D_2, y_2),$$

$d^0(\Phi_1(x_1, x_2), G_1, y_1)$ devant être calculé en supposant que x_2 est un point fixe, arbitraire de D_2.

6. L'indice des points fixes d'une application complètement continue $\phi(x)$. Soit *F un ensemble isolé de point fixes* de $\phi(x)$: F a un voisinage G ne contenant d'autres points fixes que les points de F; F est compact; $d^0(\Phi, G, 0)$ est indépendant du choix de G, est nommé *indice de F* et est noté $i(F)$. Les propriétés du degré ont pour conséquences immédiates les propriétés suivantes de l'indice:

PROPRIÉTÉ 6.1. *Soit F l'ensemble des points fixes de $\phi(x)$ contenus dans une partie ouverte G de X; F est compact et $i(F)$ est défini, quand $\dot{G}$ ne contient aucun point fixe; $i(F)$ est un entier positif, négatif, ou nul, qui reste constant quand $\phi(x)$ et G varient continûment, sans que $\dot{G}$ ne contienne jamais de point fixe de $\phi(x)$.*

COROLLAIRE 6.1. *$i(F)$ ne dépend que de la restriction de $\phi(x)$ à $\dot{G}$.*

COROLLAIRE 6.2. *Si $\phi(x)$ possède au point fixe a une différentielle[2] complètement continue $\lambda(x - a)$, telle que $a + \lambda(x - a)$ ait pour seul point fixe a, alors a est un point fixe isolé de $\phi(x)$ ayant les mêmes indices comme point fixe de $\phi(x)$ et comme point fixe de $a + \lambda(x - a)$.*

PROPRIÉTÉ 6.2. *F n'est pas vide, si $i(F) \neq 0$.*

PROPRIÉTÉ 6.3. *Si F est la réunion d'un nombre fini de compacts F_α, deux à deux sans point commun, alors $i(F) = \sum_\alpha i(F_\alpha)$.*

[2] $\lambda(y)$ est linéaire homogène; il existe un voisinage V de 0 tel que, si ϵ est un nombre réel tendant vers 0, le transformé par $\epsilon^{-1}[\phi(x) - a - \lambda(x - a)]$ du voisinage de a d'ordre ϵV tende vers 0.

PROPRIÉTÉ 6.4. *Soient deux espaces X_1 et X_2, une partie ouverte G_1 de X_1, un domaine D_2 de X_2, une application complètement continue $\phi_1(x_1, x_2)$ de $X_1 + X_2$ dans X_1 et une application complètement continue $\phi(x_2)$ de X_2 en lui-même. Supposons*

$$x_1 \neq \phi_1(x_1, x_2) \text{ pour } x_1 \in \dot{G}_1, \qquad x_2 \in D_2; \qquad x_2 \neq \phi(x_2) \text{ pour } x_2 \in \dot{D}_2;$$

soit i l'indice des points fixes $x_1 + x_2 \in G_1 + D_2$ de $\phi_1(x_1, x_2) + \phi_2(x_2)$; soit i_1 l'indice des points fixes $x_1 \in G_1$ de $\phi_1(x_1, x_2)$, quand $x_2 \in D_2$; soit i_2 l'indice des points fixes $x_2 \in D_2$ de $\phi_2(x_2)$. On a

$$i = i_1 \cdot i_2 .$$

Ces propriétés de l'indice permettent de prouver des *théorèmes d'existence* (d'après la propriété 6.2, il existe au moins un point fixe quand l'indice de l'ensemble des points fixes diffère de zéro; les propriétés 6.1 et 6.3 permettent de déterminer l'indice de l'ensemble des points fixes) et des *théorèmes d'unicité* (si l'indice de l'ensemble des points fixes est $\epsilon = \pm 1$ et si le corollaire 6.2 et la propriété 9.1 permettent de prouver que chaque point fixe est isolé et a l'indice ϵ, alors il existe un point fixe unique).

7. Le théroème de Jordan-Brouwer. *Soient F et F' deux parties fermées de X, entre lesquelles existe un homéomorphisme $x \leftrightarrow x'$; F et F' décomposent X en le même nombre de domaines, s'il existe un compact contenant toutes les valeurs prises par $x - x'$.* (On sait que cette hypothèse est essentielle: la sphère de Hilbert F: $x_1^2 + x_2^2 \cdots = 1$ décompose l'espace en deux domaines; on peut l'appliquer isométriquement sur F': $x_1 = 0$, $x_2^2 + x_3^2 + \cdots = 1$, dont le complémentaire constitue un seul domaine.

PREUVE. Soient D_λ et D'_μ les domaines en lesquels F et F' décomposent X. Posons $\Phi(x) = x', \Psi(x') = x$; on a $\Psi\Phi(x) = x$ et $\Phi\Psi(x') = x'$; d'après la propriété 5.4 les matrices $d^0(\Phi, D_\lambda, D'_\mu)$ et $d^0(\Psi, D'_\mu, D_\lambda)$ sont inverses l'une de l'autre; elles sont donc carrées.

On prouve de même:

8. L'invariance du domaine. *L'image $\Phi(D)$ d'un domaine D par un homéomorphisme $\Phi(x)$ est un domaine si $\Phi(x) - x$ est complètement continue (hypothèse essentielle).*

9. Les équations linéaires. Soit $\lambda(x)$ une application linéaire et homogène de X en lui-même, qui soit complètement continue sur un voisinage de l'origine convenablement choisi; soit ρ un nombre réel; soit

$$\Lambda_\rho(x) = x - \rho\lambda(x).$$

L'invariance du domaine a pour conséquence immédiate l'alternative de Fredholm: ou bien $\Lambda_\rho(x)$ a d'autres zéros que $x = 0$; ou bien $\Lambda_\rho(x)$ applique X sur lui-même. Il est aisé [8], en simplifiant des raisonnements de F. Riesz [12], d'en déduire les autres théorèmes de Fredholm. D'où:

PROPRIÉTÉ 9.1. *Soit n_ρ la dimension de l'espace vectoriel constitué par les zéros de $\Lambda_\rho(x)$, $\Lambda_\rho(\Lambda_\rho(x))$, $\cdots$; soit $n = \sum_{0<\rho<1} n_\rho$; si $x = 0$ est le seul point fixe de $\lambda(x)$, son indice est $(-1)^n$. En particulier cet indice est le signe, pour $\rho = 1$, de la déterminante de Fredholm, si $\lambda(x)$ est une application du type de Fredholm: $x(s) \to \int K(s, t)x(t)\, dt$.*

10. Les classes de points fixes. Le procédé de Nielsen [10] et Wecken [17] permet de classer les points fixes de $\phi(x)$ contenus dans G: les points fixes x_1 et x_2 sont placés dans une même classe quand on peut les joindre par un chemin l tel que l et $\phi(l)$ appartiennent à G et soient homotopes dans G. Chaque classe constitue évidemment un ensemble isolé de points fixes; donc son indice est défini et reste constant quand $\phi(x)$ et G varient continûment, en sorte qu'aucun point fixe n'appartienne jamais à $\dot{G}$.

III. LES POINTS FIXES D'UNE APPLICATION CONTINUE D'UN COMPACT

11. *Soit $\xi(x)$ une application continue en lui-même d'un espace compact C; supposons que C soit un rétracte d'une partie ouverte G d'un espace vectoriel à voisinages convexes X*: il existe une application continue $\pi(x)$ de G sur C dont la restriction à C, supposé intérieur à G, est l'identité. Il est clair que les points fixes de $\xi(x)$ sont ceux de l'application complètement continue $\phi(x) = \pi\xi(x)$: *les point fixes de $\xi(x)$ ont un indice possédant les propriétés énoncées au* §6.

Si X est l'espace de Hilbert, C est un espace LC^*; rappelons les deux définitions équivalentes de ces espaces (Lefschetz): ce sont les compacts métrisables et localement connexes pour toutes les dimensions; ce sont les rétractes absolus de voisinages.

[7] généralise et complète les résultats précédents: *l'indice de l'ensemble des points fixes de $\xi(x)$ est le nombre de Lefschetz de $\xi(x)$*; plus généralement $i(f)$ est le nombre de Lefschetz de restrictions convenables de $\xi(x)$ quand f est l'ensemble des points fixes de $\xi(x)$ contenus dans une partie ouverte g de C telle que

$$\lim_{n\to+\infty} \phi^n(\bar{g}) \subset g.$$

On connaît *le théorème de Lefschetz* [5]: $\xi(x)$ a au moins un point fixe quand son nombre de Lefschetz diffère de zéro. Ce théorème est une conséquence de la théorie précédente; mais il s'applique à certains espaces compacts auxquels cette théorie n'a pas été étendue. Le problème est ouvert de savoir si cette théorie est un cas particulier d'une théorie plus générale, applicable à tout espace compact.

IV. LES APPLICATIONS DE LA THÉORIE DES POINTS FIXES

La théorie des points fixes a des applications variées:

Équations intégrales non linéaires: [24, Chapitre I].

Problème de Dirichlet pour les équations non linéaires, du type elliptique à deux variables indépendantes: [22].

Calcul des variations: [13], [22].

Problème de Dirichlet posé par la théorie des fluides visqueux: [24, Chapitres II, III].

Équations linéaires, du type elliptique, à conditions aux limites non linéaires: [20].

Problèmes de représentation conforme du type d'Helmholtz posés par les écoulements de fluides parfaits avec jets ou sillages: [21], [23], [25].

Problèmes posés par les écoulements des fluides parfaits et compressibles: [18], [19].

Bibliographie

La théorie des points fixes:

1. G. D. Birkhoff et O. D. Kellogg, *Invariant points in function space*, Trans. Amer. Math. Soc. t. 23 (1922) pp. 96–115.

2. L. E. J. Brouwer, *Über Abbildungen von Mannigfaltigkeiten*, Math. Ann. t. 71 (1912) pp. 97–115.

———, *Beweis der Invarianz des n-dimensionalen Gebiets*, Math. Ann. t. 71 (1912) pp. 305–314.

3. T. H. Hildebrandt et L. M. Graves, *Implicit functions and their differentials in general analysis*, Trans. Amer. Math. Soc. t. 29 (1927) pp. 127–153, 514–552.

4. I. Fredholm, *Sur une classe d'équations fonctionelles*, Acta Math. t. 27 (1903) pp. 365–390.

5. S. Lefschetz, *Algebraic topology*, Amer. Math. Soc. Colloquium Publications, vol. 27, New York, 1942, Chap. VIII, §§5 et 6, pp. 318–326.

6. J. Leray et J. Schauder, *Topologie et équations fonctionnelles*, Ann. École Norm. t. 51 (1934) pp. 45–78.

J. Leray, *Topologie des espaces de Banach*, C. R. Acad. Sci. Paris t. 200 (1935) pp. 1082–1084.

7. J. Leray, *Sur les équations et les transformations*, Journal de Mathématiques t. 24 (1945) pp. 201–248.

8. ———, *Valeurs propres et vecteurs propres d'un endomorphisme complètement continu d'un espace vectoriel à voisinages convexes*, Acta Univ. Szeged. t. 12 (1950) pp. 177–186.

9. ———, *Théorie des équations dans les espaces vectoriels à voisinages convexes* (à paraître; professé au Collège de France en 1948–1949).

10. J. Nielsen, *Untersuchungen zur Topologie der geschlossenen zweiseitigen Flächen* I, Acta Math. t. 50 (1927) pp. 189–358.

11. E. Picard, *Traité d'analyse*, t. 3, Chap. V.

12. F. Riesz, *Über lineare Funktionalgleichungen*, Acta Math. t. 41 (1918) pp. 71–98.

13. E. Rothe, *Über Abbildungsklassen von Kugeln des Hilbertschen Raumes*, Compositio Math. t. 4 (1937) pp. 294–307.

———, *Zur Theorie der topologischen Ordnung und der Vektorfelder in Banachschen Räumen*, Compositio Math. t. 5 (1937) pp. 166–176.

———, *Über den Abbildungsgrad bei Abblidungen von Kugeln des Hilbertschen Raumes*, Compositio Math. t. 5 (1937) pp. 177–197.

———, *The theory of topological order in some linear topological spaces*, Iowa State College Journal of Science t. 13 (1939) pp. 373–390.

———, *Gradient mappings and extrema in Banach spaces*, Duke Math. J. t. 15 (1948) pp. 421–431.

———, *Gradient mappings in Hilbert spaces*, Ann. of Math. t. 47 (1946) pp. 580–592.

———, *Completely continuous scalar and variational methods*, Ann. of Math. t. 49 (1948) pp. 265–280.

14. J. SCHAUDER, *Der Fixpunktsatz in Funktionalräumen*, Studia Mathematica t. 2 (1930) pp. 170–179.

———, *Über den Zusammenhang zwischen der Eindeutigkeit und Lösbarkeit partieller Differentialgleichungen zweiter Ordnung von elliptischen Typus*, Math. Ann. t. 106 (1932) pp. 667–721.

15. E. SCHMIDT, *Über die Auflösung der nichtlinearen Integralgleichungen und die Verzweigung ihrer Lösungen*, Math. Ann. t. 65 (1908) pp. 370–399.

16. A. TYCHONOFF, *Ein Fixpunktsatz*, Math. Ann. t. 111 (1935) pp. 767–776.

17. F. WECKEN, *Fixpunktklassen*, Math. Ann. t. 117 (1941) pp. 659–671; t. 118 (1943) pp. 216–234, 544–577.

Les applications de la théorie des points fixes:

18. L. BERS, *An existence theorem in two-dimensional gas dynamics*, à paraître.

19. C. JACOB, *Sur l'écoulement lent d'un fluide parfait, compressible, autour d'un obstacle-circulaire*, Mathematica t. 17 (1941) pp. 1–18.

———, *Sur les mouvements lents des fluides parfaits compressibles*, Portugaliae Mathematica t. 1 (1939) pp. 209–257, §13.

20. ———, Thèse (Paris), Mathematica t. 11 (1935) pp. 1–118, §§22 à 28.

21. J. KRAVTCHENKO, Thèse (Paris), Journal de Mathématiques t. 21 (1942).

22. J. LERAY, *Discussion d'un problème de Dirichlet*, Journal de Mathématiques t. 18 (1939) pp. 249–284.

23. ———, *Les problèmes de représentation conforme de Helmholtz*, Comment. Math. Helv. t. 8 (1935) pp. 149–180.

———, *Théorie des sillages et des proues*, Comment. Math. Helv. t. 8 (1935) pp. 250–263.

24. ———, Thèse, Journal de Mathématiques t. 12 (1933) pp. 1–82.

25. J. LERAY et A. WEINSTEIN, *Sur un problème de représentation conforme posé par la théorie de Helmholtz*, C. R. Acad. Sci. Paris t. 198 (1934) pp. 430–432.

COLLÈGE DE FRANCE,
PARIS, FRANCE.

INTÉGRALES HARMONIQUES ET THÉORIE DES INTERSECTIONS

GEORGES DE RHAM

Dans la première partie de cet exposé, je parlerai du concept de *courant*, que j'ai introduit pour expliquer les relations entre la théorie des chaînes et celle des formes différentielles extérieures. La définition des courants adoptée ici m'a été suggerée par la lecture de l'article [8] où M. Laurent Schwartz a introduit sa notion de distribution.[1] Dans la seconde partie, je montrerai comment la théorie des formes différentielles harmoniques conduit à une formule intégrale remarquable pour représenter l'indice de Kronecker de deux chaînes différentiables dans un espace de Riemann.

Soit M une variété à n dimensions, que pour simplifier je supposerai compacte et orientable, donnée avec une structure différentiable C^∞. Par définition, un *courant de dimension* p sur M est une fonctionnelle linéaire $T[\varphi]$, définie sur l'espace linéaire de toutes les formes différentielles extérieures φ de degré p et de classe C^∞ sur M, qui est *continue* dans le sens suivant: si φ tend vers zéro en restant nulle en dehors d'un compact fixe contenu dans le domaine d'un système de coordonnées locales de manière que chaque dérivée (de n'importe quel ordre $\geqq 0$) de chaque coefficient de la forme φ (représentée au moyen des coordonnées locales) tende uniformément vers zéro, alors $T[\varphi]$ tend vers zéro.

Le plus petit ensemble fermé en dehors duquel une forme différentielle φ est nulle est appelé le *support* ("carrier" en anglais) de φ. Un courant T est dit nul dans l'ensemble ouvert D, si $T[\varphi] = 0$ pour toute forme φ dont le support est contenu dans D. On démontre que, parmi les ensembles ouverts dans lesquels $T = 0$, il y en a un, le plus grand, qui contient tous les autres. L'ensemble fermé complémentaire de ce plus grand ensemble ouvert est appelé le *support* de T.

Voici trois exemples de courants.

1. Soit α une forme différentielle de degré $n - p$ sur M. En posant

$$\alpha[\varphi] = \int \alpha \wedge \varphi$$

où $\alpha \wedge \varphi$ est le produit extérieur de α et de φ (ce qui est une forme de degré n) et où l'intégrale est étendue à toute la variété M qu'on suppose orientée une fois pour toutes, on définit un courant qui sera dit égal à la forme différentielle α. C'est un courant de dimension p.

D'une manière générale, un courant de dimension p dans une variété de dimension n sera dit de *degré* $n - p$.

2. Soit c une chaîne différentiable de dimension p dans M. En posant

[1] J'ai introduit les courants dans [3] et [6], sous une forme moins précise et moins générale. La nouvelle définition adoptée ici est déjà développée dans [7]. M. Schwartz m'a d'ailleurs communiqué qu'il a aussi envisagé cette même définition, qui est une extension naturelle de sa notion de distribution.

$$c[\varphi] = \int_c \varphi$$

on définit un courant de dimension p qui sera dit égal à la chaîne c.

3. Soit V un p-vecteur contravariant donné en un point donné y de M. Si $V^{i_1 \cdots i_p}$ sont ses composantes relativement à un système de coordonnées locales et si $\varphi_{i_1 \ldots i_p}$ sont les valeurs au point y des coefficients de la forme φ représentée au moyen des mêmes coordonnées locales, en posant

$$V[\varphi] = \sum V^{i_1 \cdots i_p} \varphi_{i_1 \cdots i_p}$$

on définit un courant de dimension p, qui sera dit égal au p-vecteur V.

Les deux premiers exemples montrent que les chaînes de dimension p et les formes de degré $n - p$, sur une variété à n dimensions, sont des cas particuliers de courants de dimension p.

Pour $p = 0$, l'argument φ de la fonctionelle est une fonction et les courants de degré n (ou de dimension zéro) ne sont pas autre chose que les distributions de M. Schwartz. Le terme "courant" a été suggéré par le cas particulier $p = 1$ et $n = 3$, qui est adapté à la notion physique de courant électrique dans l'espace ordinaire. Un courant électrique dans un réseau de fils conducteurs est représenté par une chaîne à une dimension, tandis qu'un courant électrique dans un volume conducteur est représenté par une densité vectorielle, ce qui est équivalent à une forme différentielle de degré 2.

La définition de la *différentielle extérieure* d'une forme peut facilement s'étendre aux courants. Si α est une forme de degré $n - p$ et si ψ est une forme de degré $p - 1$, on obtient par intégration par parties la formule

$$\int d\alpha \wedge \psi = (-1)^{n-p+1} \int \alpha \wedge d\psi$$

qui peut aussi s'écrire

$$d\alpha[\psi] = (-1)^{n-p+1} \alpha[d\psi].$$

Cela étant, pour chaque courant T on définit dT en posant

$$dT[\psi] = (-1)^{n-p+1} T[d\psi].$$

En tenant compte de la formule générale de Stokes, on voit que si T est une chaîne, dT est au signe près le *bord* de T. Ainsi, la différentielle extérieure d'une forme et le bord d'une chaîne apparaissent comme deux cas particuliers d'un même concept général défini pour tous les courants.

Il n'est pas aussi simple d'étendre aux courant la définition du produit extérieur de deux formes. Toutefois, on peut définir le produit extérieur $T \wedge \alpha$ d'un courant quelconque T par une forme α de classe C^∞ en posant

$$(T \wedge \alpha)[\varphi] = T[\alpha \wedge \varphi].$$

Grâce à cette définition, on peut représenter les courants par des formes diffé-

rentielles généralisées. Si en effet $x^1, \cdots, x^n$ est un système de coordonnées locales dans le domaine D est si les $T_{i_1 \cdots i_p}$ sont des courants de degré zéro,

$$\sum T_{i_1 \cdots i_p} dx^{i_1} \wedge \cdots \wedge dx^{i_p}$$

est un courant de degré p dans D, et l'on prouve que tout courant de degré p dans D peut être ainsi représenté par une forme différentielle dont les coefficients sont des courants de degré zéro.

Pour obtenir une définition du produit extérieur de deux courants dans des cas plus généraux, on peut utiliser l'approximation des courants par des formes. Supposons que si, pour $m \to \infty$, les formes différentielles α_m et β_m tendent respectivement vers les courants T et S, dans un sens qui devra être précisé, leur produit $\alpha_m \wedge \beta_m$ tende vers un courant déterminé R; il sera naturel de convenir alors que le produit $T \wedge S$ est égal à R.

On est ainsi amené à étudier l'approximation des courants par des formes. Supposons que l'on ait une métrique sur M et appelons voisinage (ϵ) d'un ensemble E l'ensemble des points dont la distance à E est moindre que ϵ. On peut alors établir le théorème suivant (pour la démonstration, voir [7, Chapitre IV]).

THÉORÈME. *Il existe des opérateurs linéaires R_ϵ et A_ϵ, définis pour tout $\epsilon > 0$, ayant les propriétés suivantes.*

(1) *Si T est un courant de dimension p dans M, $R_\epsilon T$ et $A_\epsilon T$ sont des courants de dimensions p et $p + 1$ respectivement, dont les supports sont contenus dans le voisinage (ϵ) du support de T et qui satisfont à la relation*

$$R_\epsilon T - T = (-1)^{n-p}(dA_\epsilon T - A_\epsilon dT).$$

(2) *$R_\epsilon T$ est une forme différentielle de classe C^∞ et $R_\epsilon T[\varphi] \to T[\varphi]$ pour $\epsilon \to 0$, quelle que soit la forme φ de classe C^∞.*

(3) *Si T est une forme de classe C^∞, $A_\epsilon T$ est aussi une forme de classe C^∞.*

Il résulte de (1) que les opérateurs R_ϵ et d sont permutables: $R_\epsilon dT = dR_\epsilon T$. Par suite, si T est fermé ($dT = 0$), $R_\epsilon T$ est aussi fermé et la formule (1) qui se réduit alors à $R_\epsilon T - T = \pm dA_\epsilon T$ montre que T est homologue à $R_\epsilon T$. On voit ainsi que tout courant fermé est homologue à une forme différentielle.

Pour $\epsilon \to 0$, la forme $R_\epsilon T$, considérée comme courant et fonctionnelle linéaire, tend vers T, d'après (2), et d'après (1) les supports de $R_\epsilon T$ et $dR_\epsilon T$ tendent respectivement vers les supports de T et dT.

Considérons un exemple simple. Soit $f_\epsilon(t)$ une fonction de classe C^∞ de la variable réelle t, telle que

$$f_\epsilon(t) = 0 \quad \text{si} \quad |t| > \epsilon \quad \text{et} \quad \int_{-\infty}^{+\infty} f_\epsilon(t)\, dt = 1.$$

Dans l'espace ordinaire $Oxyz$, considérons les formes

$$\alpha_\epsilon = f_\epsilon(x)\,dx, \qquad \beta_\epsilon = f_\epsilon(y)\,dy,$$

désignons par a et b les chaînes à deux dimensions constituées respectivement par les plans $x = 0$ et $y = 0$, convenablement orientés, et par c la chaîne à une dimension intersection au sens topologique de a et b, chaîne constituée par l'axe Oz. Pour $\epsilon \to 0$, on a

$$\alpha_\epsilon[\varphi] \to a[\varphi] = \int_a \varphi, \qquad \beta_\epsilon[\varphi] \to b[\varphi] \quad \text{et} \quad (\alpha_\epsilon \wedge \beta_\epsilon)[\varphi] \to c[\varphi]$$

quelle que soit la forme φ à support compact. On voit que les formes α_ϵ et β_ϵ tendent respectivement vers les chaînes a et b, tandis que leur produit extérieur tend vers l'intersection de a avec b. La notion d'intersection apparait ainsi comme un cas limite de la notion de produit extérieur. Cela semble montrer en même temps l'utilité d'étendre cette notion, et l'impossibilité d'une définition raisonnable et tout à fait générale, puisque l'intersection de deux chaînes peut être indéterminée.

Particulièrement intéressant est le cas où l'on a, dans la variété M de dimension n, deux chaînes c^p et c^{n-p} de dimensions complémentaires p et $n - p$. Leur intersection, si elle est déterminée, est une chaîne de dimension zéro, $c^0 = c^p \wedge c^{n-p}$, combinaison linéaire d'un nombre fini de points; la somme des coefficients de ces points est l'*indice de Kronecker* ou *nombre algébrique d'intersections* de c^p avec c^{n-p}, $I(c^p, c^{n-p})$, ce qui est aussi la valeur de la fonctionnelle $c^0[\varphi]$ pour la fonction $\varphi = 1$. Du théorème ci-dessus, on peut déduire que l'on a

$$I(c^p, c^{n-p}) = \int R_\epsilon c^p \wedge R_{\epsilon'} c^{n-p} = \int_{c^p} R_{\epsilon'} c^{n-p} = (-1)^{pn+p} \int_{c^{n-p}} R_\epsilon c^p$$

pourvu que $\epsilon + \epsilon'$ soit inférieur à la distance de chacune des chaînes c^p et c^{n-p} au bord de l'autre.

La théorie des formes différentielles harmoniques permet d'obtenir, pour cet indice de Kronecker, une autre représentation où interviennent les intégrales étendues aux deux chaînes d'une certaine forme différentielle parfaitement déterminée dès qu'on a donné sur M une métrique riemannienne. Supposons donc donné sur M un ds^2 défini positif, de classe C^∞. Comme on sait, à chaque forme α de degré p est alors associée une forme *adjointe* $*\alpha$ de degré $n - p$, deux formes α et β de même degré ont un produit scalaire

$$(\alpha, \beta) = (\beta, \alpha) = \int \alpha \wedge *\beta = \beta[*\alpha]$$

et à côté de la différentiation d s'introduit un autre opérateur différentiel, la codifférentiation $\delta = \pm *d*$, qui satisfait à

$$(\delta\varphi\,, \psi) = (\varphi, d\psi).$$

Si T est un courant et φ une forme de classe C^∞ du même degré p que T, on définit (T, φ) et $*T$ en posant

$$(T, \varphi) = T[*\varphi] = (-1)^{pn+p} {}*T[\varphi]$$

et la codifférentielle $\delta T = \pm *d*T$ est déterminée par l'identité $(\delta T, \psi) = (T, d\psi)$. Avec ces notations, si c^p est une chaîne à p dimensions et α une forme de degré p, on a

$$(*c^p, \alpha) = c^p[\alpha] = \int_{c^p} \alpha$$

et les relations (1) peuvent s'écrire de la manière suivante

$$(2) \qquad I(c^p, c^{n-p}) = (*R_\epsilon c^p, R_{\epsilon'} c^{n-p}) = (*c^p, R_{\epsilon'} c^{n-p}) = (*R_\epsilon c^p, c^{n-p}).$$

Cela suggère que l'indice de Kronecker $I(c^p, c^{n-p})$ peut être interpreté comme le produit scalaire $(*c^p, c^{n-p})$ des deux courants $*c^p$ et c^{n-p}.

Or, on peut étendre d'une autre manière la définition du produit scalaire. Appelons *ensemble singulier* d'un courant T le plus petit ensemble fermé en dehors duquel T est égal à une forme de classe C^∞. Supposons que les ensembles singuliers de S et de T n'ont aucun point commun. On peut alors décomposer ces courants en sommes $S = S_1 + S_2$, $T = T_1 + T_2$, de manière que S_2 et T_2 soient des formes de classe C^∞ et que les supports de S_1 et T_1 n'aient aucun point commun. Les produits scalaires (S_1, T_2) et (S_2, T) sont alors déterminés et l'on peut définir (S, T) en posant $(S, T) = (S_1, T_2) + (S_2, T)$. Ainsi, *le produit scalaire de deux courants est déterminé lorsqu'ils n'ont pas de point singulier commun.*

Pour aller plus loin, utilisons la théorie des formes différentielles harmoniques. D'après un théorème de cette théorie, toute forme α peut être décomposée, d'une manière unique, en la somme de trois formes, $\alpha = H_1\alpha + H_2\alpha + H_3\alpha$, dont la première est homologue à zéro ($H_1\alpha = d\beta$), la seconde est cohomologue à zéro ($H_2\alpha = \delta\gamma$) et la troisième est harmonique ($dH_3\alpha = 0$, $\delta H_3\alpha = 0$). Les opérateurs H_1, H_2, et H_3 définis par cette décomposition sont des projecteurs deux à deux orthogonaux. Il en résulte en particulier la formule

$$(3) \qquad (\alpha, \beta) = (H_1\alpha, \beta) + (\alpha, H_2\beta + H_3\beta).$$

Le théorème de décomposition ci-dessus s'étend aux courants. En effet, en définissant H_jT ($j = 1, 2, 3$) par $(H_jT, \varphi) = (T, H_j\varphi)$, on a $T = H_1T + H_2T + H_3T$ et l'on prouve que les ensembles singuliers de H_1T et H_2T sont respectivement identiques à ceux de δT et dT, tandis que celui de H_3T est vide, H_3T étant une forme harmonique. Pour la démonstration, voir [7, Chapitre III]. Si alors, dans la formule (3) on remplace α par $*c^p$ et β par c^{n-p}, on obtient

$$(4) \qquad (*c^p, c^{n-p}) = (H_1*c^p, c^{n-p}) + (*c^p, H_2c^{n-p} + H_3c^{n-p}).$$

L'ensemble singulier du courant $H_1*c^p = *H_2c^p$ est contenu dans le support de dc^p, c'est-à-dire dans le bord de c^p, qui par hypothèse ne rencontre pas c^{n-p}; ainsi, le premier terme au second membre de (4) est bien défini. Il en est de même du second terme, car l'ensemble singulier de $H_2c^{n-p} + H_3c^{n-p}$ se réduit à

celui de H_2c^{n-p} et il est contenu dans le bord de c^{n-p} qui par hypothèse ne rencontre pas c^p. Ainsi, la formule (4) donne une valeur bien déterminée pour $(*c^p, c^{n-p})$ et l'on peut vérifier directement qu'elle est bien égale à $I(c^p, c^{n-p})$.

Considérons un point déterminé y de M et supposons que l'on attribue aux C_n^p produits $dy^{i_1} \wedge \cdots \wedge dy^{i_p}$ des valeurs numériques déterminées. Il existe alors un courant bien déterminé de degré p, Y, tel que

$$(Y, \varphi) = \sum \varphi_{i_1 \cdots i_p}(y)\, dy^{i_1} \wedge \cdots \wedge dy^{i_p}.$$

C'est le courant adjoint au courant égal au p-vecteur contravariant situé en y de composantes $dy^{i_1} \wedge \cdots \wedge dy^{i_p}$ (troisième exemple de courant considéré plus haut). Pour abréger, on dira que (Y, φ) est la valeur de la forme φ au point y et l'on écrira $(Y, \varphi) = \varphi(y)$. Désignons de même par X le courant tel que $(X, \varphi) = \varphi(x) =$ valeur de φ au point x.

Le support de Y se réduisant au point y, les ensembles singuliers de H_1Y et H_2Y se réduisent aussi au point y. Par suite, en dehors de ce point, H_1Y est égal à une forme de classe C^∞ dont la valeur $e(x, y)$ au point $x \neq y$ est $e(x, y) = (X, H_1Y)$. Si l'on considère x et y comme variables, $e(x, y)$ est une double forme de degré p par rapport à x et par rapport à y, symétrique parceque $(X, H_1Y) = (H_1X, Y)$, de classe C^∞ pour $x \neq y$. Elle est fermée parceque H_1Y est un courant homologue à zéro, donc fermé. Elle est aussi cofermée, car, en dehors de y, $Y = 0$ et par suite $H_1Y = -H_2Y - H_3Y$ et l'on sait que H_2Y et H_3Y sont cofermés. Ainsi, $e(x, y)$ est une double forme symétrique, harmonique pour $x \neq y$, parfaitement déterminée par la métrique riemannienne donnée sur M.

En dehors du bord de c^p, le courant $H_1{*}c^p$ est égal à une forme de classe C^∞ dont la valeur au point y est donnée par

$$(Y, H_1{*}c^p) = (H_1Y, *c^p) = \int_{c_x^p} e(x, y),$$

l'intégration étant effectuée sur c^p par rapport à x. De là résulte

$$(H_1 *c^p, c^{n-p}) = \int_{c_y^{n-p}} \int_{c_x^p} e(x, y^*)$$

formule où $e(x, y^*)$ est l'adjointe relativement à y de $e(x, y)$, ce qui est une forme de degré p en x et de degré $n - p$ en y, l'intégration étant effectuée d'abord par rapport à x sur c^p et ensuite par rapport à y sur c^{n-p}.

Pour $x \neq y$, on a $(H_2X + H_3X, Y) = -(H_1X, Y) = -e(x, y)$. Par suite, en un point x non situé sur le bord de c^{n-p}, le courant $H_2c^{n-p} + H_3c^{n-p}$ est égal à une forme dont la valeur est

$$(H_2c^{n-p} + H_3c^{n-p}, X) = (c^{n-p}, H_2X + H_3X) = -\int_{c_y^{n-p}} e(x, y^*).$$

Il en résulte

$$(*c^p, H_2c^{n-p} + H_3c^{n-p}) = -\int_{c_x^p} \int_{c_y^{n-p}} e(x, y^*)$$

d'où finalement la formule cherchée pour l'indice de Kronecker

$$I(c^p, c^{n-p}) = \int_{c_y^{n-p}} \int_{c_x^p} e(x, y^*) - \int_{c_x^p} \int_{c_y^{n-p}} e(x, y^*).$$

On peut montrer que, pour $x = y$, la forme $e(x, y)$ devient infinie d'ordre n, ce qui explique qu'un changement dans l'ordre des intégrations puisse changer la valeur de l'intégrale, sauf naturellement dans le cas où les deux chaînes c^p et c^{n-p} ne se coupent pas, auquel cas évidemment $I(c^p, c^{n-p}) = 0$.

Cette formule a été établie dans [5], pour le cas d'un espace de Riemann compact et orientable. Comme cela est indiqué dans [7], on peut l'étendre au cas d'un espace de Riemann non orientable, et même non compact. Dans le cas non compact, le théorème de décomposition pour les formes est encore valable, ainsi que l'a montré M. Kodaira, pourvu que l'on se restreigne aux formes α de carré sommable, c'est-à-dire telles que l'intégrale (α, α) converge. On l'étend ensuite aux courants T tels que la fonctionnelle (T, φ) de φ soit définie pour toutes les formes φ de classe C^∞ et de carré sommable. Le courant Y satisfait évidemment à cette condition, ce qui permet de définir la forme $e(x, y) = (X, H_1 Y)$.

Bibliographie

1. P. Bidal et G. de Rham, *Les formes différentielles harmoniques*, Comment. Math. Helv. t. 19 (1946) pp. 1–49.

2. K. Kodaira, *Harmonic fields in Riemannian manifolds*, Ann. of Math. t. 50 (1949) pp. 587–665.

3. G. de Rham, *Relations entre la topologie et la théorie des intégrales multiples*, L'Enseignement Mathématique, 1936, pp. 213–228.

4. ———, *Remarque au sujet de la théorie des formes différentielle harmoniques*, Annales de l'Université de Grenoble t. 23 (1947–1948) pp. 55–56.

5. ———, *Sur la théorie des formes différentielles harmoniques*, Annales de l'Université de Grenoble t. 22 (1946) pp. 135–152.

6. ———, *Ueber mehrfache Integrale*, Abh. Math. Sem. Hansischen Univ. t. 12 (1938) pp. 313–339.

7. G. de Rham et K. Kodaira, *Harmonic integrals*, Lectures delivered at the Institute for Advanced Study, Princeton, 1950.

8. L. Schwartz, *Généralisation de la notion de fonction ...*, Annales de l'Université de Grenoble t. 21 (1945) pp. 57–74.

9. ———, *Théorie des distributions*, t. 1, Paris, Hermann and Cie, 1950.

University of Geneva,
Geneva, Switzerland.
University of Lausanne,
Lausanne, Switzerland.

COURBURE, NOMBRES DE BETTI, ET ESPACES SYMÉTRIQUES[1]

A. LICHNEROWICZ

1. Introduction. Si V_n est une variété riemannienne compacte, de métrique définie positive, il est connu depuis longtemps que ses propriétés topologiques et géométriques sont étroitement reliées. Par exemple, si V_n est orientable et de courbure constante positive, on peut établir directement qu'elle admet pour espace universel de recouvrement la sphère S_n ; par suite ses nombres de Betti sont ceux de la sphère S_n.

Récemment des résultats beaucoup plus étendus ont été obtenus parallèlement par Bochner et moi-même, sur certaines relations reliant la courbure riemannienne de V_n et ses nombres de Betti. Les résultats de Bochner et les miens, dans cet ordre d'idées, seront l'objet principal de cet exposé. Bochner avait surtout en vue, semble-t-il, des applications au domaine assez riche des variétés à structure analytique complexe, en particulier des variétés kähleriennes. Je me limiterai au contraire ici à l'étude, peut-être plus ingrate, des propriétés qui ressortent strictement du domaine réel. Dans toute la suite, V_n désignera une variété riemannienne à n dimensions *compacte et de métrique définie positive.*

Notre méthode repose essentiellement sur l'étude d'opérations elliptiques portant sur un scalaire défini sur V_n. La même méthode fournit également des conditions nécessaires et suffisantes nouvelles pour qu'un espace riemannien compact soit *symétrique* au sens de Elie Cartan, avec tout ce que cela comporte de propriétés géométriques et topologiques liées. C'est grâce à la technique des formes harmoniques de Hodge-de Rham que les propriétés d'homologie de la variété V_n peuvent être atteintes. Les relations classiques d'harmonicité peuvent être exprimées en termes de dérivation covariante sur V_n de la manière suivante: un p-tenseur antisymétrique $T_{\beta_1\beta_2\cdots\beta_p}$ est *harmonique* s'il satisfait aux deux conditions

$$(1.1) \qquad \text{(a)}\ \epsilon^{\beta\beta_1\cdots\beta_p}_{\alpha_1\alpha_2\cdots\alpha_{p+1}}\nabla_\beta T_{\beta_1\beta_2\cdots\beta_p} = 0; \qquad \text{(b)}\ \nabla_\alpha T^\alpha_{\beta_1\cdots\beta_{p-1}} = 0$$

où ∇_α est l'opérateur de dérivation covariante et ϵ l'indicateur classique de permutation. Cela posé, on sait que, *si V_n est orientable, le nombre de p-tenseurs harmoniques linéairement indépendants définis sur V_n est égal au $p^{\text{ème}}$ nombre de Betti de V_n.*

2. Existence d'un tenseur symétrique à dérivée covariante nulle. De la forme (1.1) des conditions d'harmonicité, on peut déjà déduire des résultats simples de géométrie différentielle globale. Supposons qu'il existe sur V_n un tenseur symétrique $T_{\alpha\beta}$ d'ordre 2, non proportionnel à $g_{\alpha\beta}$ et de dérivée covariante nulle ($\nabla_\gamma T_{\alpha\beta} = 0$). La variété V_n est alors localement réductible en ce sens que la

[1] Cette communication était mentionnée sur le programme imprimé sous le titre *Curvature and Betti numbers.*

métrique de V_n peut être décomposée, aux différents points de V_n, en une somme de plusieurs métriques locales telle que:

$$ds^2 = g_{\alpha_1\beta_1}(x^{\lambda_1})\, dx^{\alpha_1} dx^{\beta_1} + g_{\alpha_2\beta_2}(x^{\lambda_2})\, dx^{\alpha_2} dx^{\beta_2}$$

$$(\alpha_1, \beta_1, \lambda_1 = 1, \cdots, r;\ \alpha_2, \beta_2, \lambda_2 = r+1, \cdots, n).$$

Le polynôme caractéristique de la matrice $(T_{\alpha\beta})$ par rapport à la matrice $(g_{\alpha\beta})$ est à coefficients constants sur V_n. À toute valeur caractéristique d'ordre p, correspond un champ de p-plans parallèles qui définit un p-vecteur représentatif $u_{\alpha_1\alpha_2\cdots\alpha_p}$ à dérivée covariante nulle et par suite *une forme harmonique sur* V_n. De plus, à des valeurs caractéristiques distinctes de même ordre de multiplicité, correspondent manifestement des formes harmoniques linéairement indépendantes.

On peut remarquer en outre que, *sur un espace compact, la nullité d'une dérivée covariante d'ordre quelconque d'un tenseur entraîne la nullité de la dérivée première de ce tenseur.* En effet si, par exemple,

$$\nabla_\lambda \nabla_\mu Q_{\alpha\beta} = 0$$

en multipliant par $Q^{\alpha\beta}$ et contractant, il vient:

$$\frac{1}{2}\Delta[Q^{\alpha\beta}Q_{\alpha\beta}] = g^{\lambda\mu}\nabla_\lambda Q^{\alpha\beta}\nabla_\mu Q_{\alpha\beta}$$

où Δ désigne le laplacien attaché à la métrique riemannienne et où le second membre est essentiellement positif ou nul. Comme le scalaire $(Q^{\alpha\beta}Q_{\alpha\beta})$ atteint nécessairement son maximum en un point de la variété compacte V_n, $\Delta[Q^{\alpha\beta}Q_{\alpha\beta}]$ ne peut être qu'identiquement nul et $\nabla_\mu Q_{\alpha\beta} = 0$. Le même raisonnement s'étend manifestement par récurrence à une dérivée covariante d'ordre quelconque d'un tenseur arbitraire. Nous pouvons donc énoncer:

THÉORÈME. *S'il existe sur une variété orientable V_n un tenseur symétrique $T_{\alpha\beta}$, non proportionnel à $g_{\alpha\beta}$, et dont l'une des dérivées covariantes est nulle, la variété V_n est localement réductible et ses nombres de Betti satisfont aux inégalités*

$$b_p(V_n) \geqq k_p \qquad (p = 1, 2, \cdots, n-1) \tag{2.1}$$

où k_p désigne le nombre de valeurs caractéristiques de $(T_{\alpha\beta})$ qui admettent l'ordre de multiplicité p.

En particulier, il ne peut exister de tel tenseur sur une variété V_n dont les nombres de Betti sont ceux de la sphère.

3. Une formule fondamentale relative aux tenseurs harmoniques. La démonstration précédente nous fournit un exemple simple du mode de raisonnement que nous allons utiliser. Nous établirons d'abord une formule fondamentale

relative au carré d'un tenseur harmonique. Partons de la formule de commutation des dérivations covariantes

$$\nabla_\lambda(\nabla_\mu T_{\alpha_1\cdots\alpha_p}) - \nabla_\mu(\nabla_\lambda T_{\alpha_1\cdots\alpha_p}) = -R_{\alpha_1\rho,\lambda\mu}T^\rho{}_{\alpha_2\cdots\alpha_p} - \cdots - R_{\alpha_p\rho,\lambda\mu}T_{\alpha_1\cdots\alpha_{p-1}}{}^\rho$$

où nous supposerons le tenseur T harmonique. En contractant les indices α_1 et λ il vient, compte tenu de (1.1) (b),

$$\nabla_\lambda(\nabla_\mu T^\lambda{}_{\alpha_2\cdots\alpha_p}) = -R_{\rho\mu}T^\rho{}_{\alpha_2\cdots\alpha_p} - R_{\alpha_2\rho,\lambda\mu}T^{\lambda\rho}{}_{\alpha_3\cdots\alpha_p} - \cdots - R_{\alpha_p\rho,\lambda\mu}T^\lambda{}_{\alpha_2\cdots\alpha_{p-1}}{}^\rho.$$

Si nous multiplions par $T^{\mu\alpha_2\cdots\alpha_p}$ et contractons, nous obtenons en transformant le premier membre au moyen de l'autre relation d'harmonicité, la formule fondamentale:

$$\frac{1}{2p}\Delta[T^2] = \frac{1}{p}g^{\lambda\mu}\nabla_\lambda T^{\alpha_1\cdots\alpha_p}\nabla_\mu T_{\alpha_1\cdots\alpha_p} - R_{\rho\sigma}T^\rho{}_{\alpha_2\cdots\alpha_p}T^{\sigma\alpha_2\cdots\alpha_p} + \frac{p-1}{2}R_{\lambda\mu,\rho\sigma}T^{\lambda\mu}{}_{\alpha_3\cdots\alpha_p}T^{\rho\sigma\alpha_3\cdots\alpha_p}, \tag{3.1}$$

où

$$T^2 = T^{\alpha_1\cdots\alpha_p}T_{\alpha_1\cdots\alpha_p} \tag{3.2}$$

est le carré du tenseur harmonique T envisagé.

4. Quelques applications simples. (a) *Le cas où* $p = 1$. Dans ce cas, la formule fondamentale (3.1) s'écrit:

$$\frac{1}{2}\Delta[T^2] = g^{\lambda\mu}\nabla_\lambda T^\alpha \nabla_\mu T_\alpha - R_{\rho\sigma}T^\rho T^\sigma. \tag{4.1}$$

Au second membre apparaît la courbure de Ricci de l'espace dans la direction du vecteur T^ρ.[2] Supposons que la courbure de Ricci de V_n soit strictement positive, c'est-à-dire que la forme quadratique $R_{\rho\sigma}T^\rho T^\sigma$ soit partout définie négative. $\Delta[T^2]$ étant partout positif ou nul doit être nul. Il en est donc nécessairement de même du vecteur harmonique T_ρ. Nous obtenons le théorème de Bochner (voir [1, p. 780]).

THÉORÈME. *Si une variété riemannienne compacte, orientable admet une courbure de Ricci strictement positive, son premier nombre de Betti est nul.*

Ce résultat a été aussi obtenu par Myers (voir [9]) par une méthode toute différente.

(b) *Espace à courbure de Ricci nulle.* Supposons V_n tel que $R_{\rho\sigma} = 0$. La formule

[2] La courbure de Ricci en M et dans la direction du vecteur $\vec{T}$ est $C(M, \vec{T}) = -R_{\rho\sigma}T^\rho T^\sigma / G_{\rho\sigma}T^\rho T^\sigma$.

(4.1) s'écrit

$$\frac{1}{2}\nabla[T^2] = g^{\lambda\mu}\nabla_\lambda T^\alpha \nabla_\mu T_\alpha$$

et un raisonnement identique au précédent montre que si T_α est harmonique, il est à dérivée covariante nulle. Par un raisonnement tout à fait analogue, dû à Bochner (voir [1, pp. 781, 782]), on voit qu'il en est de même si l'on suppose que T_α, au lieu d'être harmonique, définit un groupe d'isométrie à 1 paramètre de V_n. Ainsi:

THÉORÈME. *Si V_n est de courbure de Ricci nulle, son premier nombre de Betti est égal au nombre de champs indépendants de vecteurs parallèles et au nombre de générateurs de groupes d'isométrie à un paramètre de cette variété.*

L'étude approfondie de ces variétés intéressantes est encore peu avancée. Dans cet ordre d'idées, je signalerai le résultat suivant.[3]

THÉORÈME. *Si un V_4 de courbure de Ricci nulle admet une caractéristique d'Euler-Poincaré inférieure à 2, il est localement euclidien.*

En effet si $x(V_4) < 2$, le premier nombre de Betti de V_4 est au moins égal à 1 et V_4 admet un champ de vecteurs parallèles. Par suite la métrique de V_4 peut localement se mettre sous la forme

$$ds^2 = (dx^4)^2 + g_{ij}(x^h)\,dx^i\,dx^j \qquad (i,j,h = 1,2,3).$$

La nullité de la courbure de Ricci de V_4 entraîne, d'après les équations de Gauss-Codazzi, la nullité de celle de la métrique à 3 dimensions $g_{ij}\,dx^i\,dx^j$ et par suite le caractère localement euclidien de l'espace. Un résultat analogue peut être obtenu à partir de la formule de Chern donnant la caractéristique d'une V_n.

(c) *Espace à courbure constante positive.* De la formule (3.1), on déduit immédiatement le théorème classique:

THÉORÈME. *Si un V_n orientable admet une courbure constante positive, ses nombres de Betti sont ceux de la sphère.*

5. Un théorème général. Les exemples précédents montrent qu'il convient d'étudier le signe du second membre de la formule (3.1) et en particulier le signe de la quantité

$$Q_p(T) = R_{\lambda\mu,\rho\sigma}T^{\lambda\mu}{}_{\alpha_3\cdots\alpha_p}T^{\rho\sigma\alpha_3\cdots\alpha_p} - \frac{2}{p-1}R_{\rho\sigma}T^{\rho}{}_{\alpha_2\cdots\alpha_p}T^{\sigma\alpha_2\cdots\alpha_p}.$$

Il est souhaitable que le tenseur T intervienne de la même façon dans les deux termes de $Q_p(T)$. À cet effet nous introduirons le tenseur

[3] J'ai établi antérieurement la même conclusion pour un V_4 de métrique hyperbolique normale admettant un groupe d'isométrie à trajectoires orientées "dans le temps" [5], ce qui est un résultat fondamental pour la théorie relativiste de la gravitation.

$$2A_{\lambda\mu,\rho\sigma} = -(R_{\lambda\rho}g_{\mu\sigma} + R_{\mu\sigma}g_{\lambda\rho} - R_{\lambda\sigma}g_{\mu\rho} - R_{\mu\rho}g_{\lambda\sigma}) \tag{5.1}$$

et écrirons

$$-2R_{\rho\sigma}T^{\rho}{}_{\alpha_2\cdots\alpha_p}T^{\sigma\alpha_2\cdots\alpha_p} = A_{\lambda\mu,\rho\sigma}T^{\lambda\mu}{}_{\alpha_3\cdots\alpha_p}T^{\rho\sigma\alpha_3\cdots\alpha_p}. \tag{5.2}$$

Il en résulte

$$Q_p(T) = \left[R_{\lambda\mu,\rho\sigma} + \frac{1}{p-1}A_{\lambda\mu,\rho\sigma}\right] T^{\lambda\mu}{}_{\alpha_3\cdots\alpha_p}T^{\rho\sigma\alpha_3\cdots\alpha_p}.$$

Nous sommes ainsi conduits à étudier dans quels cas la forme quadratique

$$Q_p(t) = \left[R_{\lambda\mu,\rho\sigma} + \frac{1}{p-1}A_{\lambda\mu,\rho\sigma}\right] t^{\lambda\mu}t^{\rho\sigma} \quad (t^{\lambda\mu} = -t^{\mu\lambda})$$

où $(\lambda\mu)$ joue le rôle d'un indice composé, est définie positive. Supposons la courbure de Ricci strictement positive. Selon (5.2), la forme quadratique de coefficients $A_{\lambda\mu,\rho\sigma}$ est définie positive et les valeurs caractéristiques $\rho_{\lambda\mu}$ de la matrice $(R_{\lambda\mu,\rho\sigma})$ par rapport à la matrice $(A_{\lambda\mu,\rho\sigma})$ sont réelles. La forme Q_p est définie positive si

$$\max \mid \bar{\rho}_{\lambda\mu} \mid < \frac{1}{p-1}$$

où $\bar{\rho}_{\lambda\mu}$ désigne celles des valeurs caractéristiques qui sont négatives. S'il en est ainsi, $\Delta[T^2]$ est strictement positif sauf pour $T^2 = 0$. De la formule (3.1), il résulte alors que tout p-tenseur harmonique est identiquement nul et que, par suite, $b_p(V_n) = 0$. Nous dirons que V_n satisfait à l'hypothèse (H_p) pour l'entier p si

HYPOTHÈSE (H_p): (a) *la courbure de Ricci est positive*; (b) $\max \mid \bar{\rho}_{\lambda\mu} \mid < 1/(p-1)$.

Il est clair que si (H_p) est satisfaite, (H_q) est satisfaite pour $0 < q < p$. Nous énoncerons:

THÉORÈME. *Si V_n satisfait à l'hypothèse (H_p) et est orientable, $b_q(V_n) = 0$ pour $q = 1, 2, \cdots, p$.*

6. Deux corollaires. Du théorème du §5, on déduit aisément les deux corollaires suivants:

(a) Supposons la variété riemannienne V_n localement réductible. *Si, pour un entier déterminé p, chacune des métriques locales satisfait à l'hypothèse (H_p), il en est de même pour la métrique de V_n et $b_q(V_n) = 0$ pour $q = 1, 2, \cdots, p$.*

(b) *Si une V_n orientable est de courbure de Ricci positive et de courbure projective ou conforme nulle, ses nombres de Betti sont ceux de la sphère.*

7. Les hypersurfaces d'un espace euclidien. Supposons que V_n puisse être localement plongée dans un espace euclidien V_{n+1}. Nous dirons que V_n est une hypersurface *localement convexe* si sa seconde forme quadratique fondamentale est définie. Dans ce cas, les courbures principales ρ_λ ont toutes le même signe,

par exemple, le signe +, et la courbure de Ricci de V_n est positive. En rapportant les tenseurs au repère orthonormé tangent aux lignes de courbure, le coefficient de $(t^{\alpha\beta})^2$, dans la forme $Q_p(t)$, s'écrit à un facteur constant positif près

$$(\sum\rho_\nu)(\rho_\alpha + \rho_\beta) - (\rho_\alpha^2 + \rho_\beta^2) - 2(p-1)\rho_\alpha\rho_\beta .$$

Ces coefficients sont tous positifs si

$$(\sum \rho_\nu) - \rho_\alpha - (p-1)\rho_\beta > 0$$

pour tout couple d'indices α, β *distincts*. Ainsi l'hypothèse (H_p) est satisfaite quand, pour α et β distincts,

$$\frac{1}{p-1} > \frac{\rho_\beta}{\sum\limits_{\nu\neq\alpha}\rho_\nu}.$$

De ce résultat et du théorème du §5 on déduit le théorème suivant:

THÉORÈME. *Si une V_n orientable est une hypersurface localement convexe d'un espace euclidien V_{n+1}, ses deux premiers nombres de Betti sont nuls. Si la plus petite courbure principale est supérieure à la moitié de la plus grande, les nombres de Betti de V_n sont ceux de la sphère.*

On notera que ce résultat subsiste si l'on substitue à l'espace euclidien V_{n+1} un espace de courbure constante *positive*.

8. Une formule relative au carré du tenseur de courbure. Grâce aux identités de Bianchi, on peut établir pour le tenseur de courbure une formule analogue à la formule (3.1) établie pour les tenseurs harmoniques. Partons de la formule de commutation des dérivations covariantes

$$\nabla_\lambda(\nabla_\mu R_{\alpha\beta,\gamma\delta}) - \nabla_\mu(\nabla_\lambda R_{\alpha\beta,\gamma\delta}) = H_{\alpha\beta\gamma\delta,\lambda\mu} \tag{8.1}$$

où $H_{\alpha\beta\gamma\delta,\lambda\mu}$ est un tenseur introduit par Elie Cartan dans la théorie des espaces symétriques (voir [4, p. 265])

$$H_{\alpha\beta\gamma\delta,\lambda\mu} = R^\rho{}_{\alpha,\lambda\mu}R_{\rho\beta,\gamma\delta} + R^\rho{}_{\beta,\lambda\mu}R_{\alpha\rho,\gamma\delta} + R^\rho{}_{\gamma,\lambda\mu}R_{\alpha\beta,\rho\delta} + R^\rho{}_{\delta,\lambda\mu}R_{\alpha\beta,\gamma\rho} . \tag{8.2}$$

En contractant α et λ dans (8.1) et multipliant par $R^{\mu\beta,\gamma\delta}$, on obtient, grâce aux identités de Bianchi qui jouent ici le rôle précédemment joué par les conditions d'harmonicité, la formule générale

$$\frac{1}{2}\Delta[P^2] = g^{\lambda\mu}\nabla_\lambda R^{\alpha\beta,\gamma\delta}\nabla_\mu R_{\alpha\beta,\gamma\delta} + 4R^{\alpha\beta,\gamma\delta}\nabla_\alpha(\nabla_\gamma R_{\beta\delta}) + 2K \tag{8.3}$$

où l'on a posé

$$P^2 = R^{\alpha\beta,\gamma\delta}R_{\alpha\beta,\gamma\delta} ; \qquad K = H_{\lambda\beta\gamma\delta,\mu}{}^{\lambda}R^{\mu\beta,\gamma\delta}.$$

De la formule (8.3) et des propriétés classiques des espaces symétriques, on déduit le théorème suivant.

THÉORÈME. *Si un espace V_n est tel que $K \geqq 0$ et si de plus*
(1) *le tenseur dérivé du tenseur de Ricci est nul, l'espace V_n est symétrique,*
(2) *le tenseur de Ricci est nul, l'espace est localement euclidien.*

Il en résulte par exemple que les relations

$$H_{\alpha\beta\gamma\delta,\lambda\mu} = 0 \qquad \nabla_\gamma R_{\alpha\beta} = 0$$

fournissent une condition nécessaire et suffisante pour qu'un espace riemannien *compact* soit symétrique.

9. Application aux espaces récurrents compacts. À propos de la théorie des espaces riemanniens harmoniques, Ruse et A. G. Walker ont introduit récemment la notion d'espace récurrent. Un espace riemannien est dit *récurrent* s'il existe un vecteur k_λ tel que

$$\nabla_\lambda R_{\alpha\beta,\gamma\delta} = k_\lambda R_{\alpha\beta,\gamma\delta}\,. \tag{9.1}$$

Le vecteur de récurrence k_λ est le gradient de $\log P$ et l'on a manifestement $R = cP$ (c = const.). On notera que tout espace à 2 dimensions est récurrent. Pour un espace récurrent, le tenseur H est identiquement nul et la formule (8.3) peut être mise sous la forme simple

$$(1 - c^2)\left[\frac{1}{2}\Delta(P^2) + \Delta_1 P\right] = 0. \tag{9.2}$$

On en déduit le théorème suivant.

THÉORÈME 1. *Tout espace récurrent compact est ou bien symétrique, ou bien tel que $R^2 = P^2$. Si $R_{\alpha\beta} = 0$, l'espace est localement euclidien.*

Ce théorème peut être précisé en étudiant la réductibilité locale des espaces récurrents. On établit, par une étude directe, que tout espace récurrent localement réductible est soit symétrique (et à composantes symétriques), soit localement la somme d'une métrique récurrente irréductible et d'une métrique localement euclidienne. Or toute métrique récurrente irréductible est nécessairement métrique d'Einstein:

$$R_{\alpha\beta} = \frac{R}{n}\, g_{\alpha\beta}.$$

Il en résulte que pour une telle métrique trois cas sont possibles: ou bien $n = 2$, ou bien $R = \text{const.} \neq 0$ (et la métrique est symétrique) ou bien $R_{\alpha\beta} = 0$. Or, des identités de Bianchi, il résulte que dans ce cas $R_{\alpha\beta\gamma\delta}(k^\lambda k_\lambda) = 0$.

Ainsi tout espace récurrent est soit symétrique, soit de métrique réductible à la somme d'une métrique arbitraire à deux dimensions et d'une métrique euclidienne [11].

Un *espace récurrent du second ordre* est un espace tel que

$$\nabla_\lambda(\nabla_\mu R_{\alpha\beta,\gamma\delta}) = a_{\lambda\mu} R_{\alpha\beta,\gamma\delta} . \tag{9.3}$$

Si $R \neq 0$, $a_{\lambda\mu} = \nabla_\lambda k_\mu + k_\lambda k_\mu$, où k_λ est le gradient de $\log | R |$ et en transformant légèrement la formule (8.3), il vient:

THÉORÈME 2. *Tout espace récurrent du second ordre compact, de courbure scalaire non nulle, est récurrent et par suite est soit symétrique, soit de métrique localement réductible à la somme d'une métrique arbitraire à deux dimensions et d'une métrique localement euclidienne.*

BIBLIOGRAPHIE

1. S. BOCHNER, *Vector fields and Ricci curvature*, Bull. Amer. Math. Soc. vol. 52 (1946) pp. 776–797.

2. ———, *Curvature in hermitian metric*, Bull. Amer. Math. Soc. vol. 53 (1946) pp. 179–195.

3. ———, *Curvature and Betti numbers*, Ann. of Math. vol. 49 (1948) pp. 379–390.

4. E. CARTAN, *Leçons sur la géométrie des espaces de Riemann*, 2è ed., Paris, 1946.

5. A. LICHNEROWICZ, *Problèmes globaux en mécanique relativiste*, Paris, 1939.

6. ———, *Courbure et nombres de Betti d'une variété riemannienne compacte*, C. R. Acad. Sci. vol. 226 (1948) pp. 1678–1680.

7. ———, *Dérivation covariante et nombres de Betti*, C. R. Acad. Sci vol. 230 (1950) pp. 1248–1250.

8. ———, *Sur certaines classes d'espaces riemanniens compacts*, C. R. Acad. Sci. vol. 230 (1950) pp. 2146–2148.

9. S. B. MYERS, *Riemannian manifolds with positive mean curvature*, Duke Math. J. vol. 8 (1941) pp. 401–404.

10. H. S. RUSE, *Three dimensional spaces of recurrent curvature*, Proc. London Math. Soc. vol. 50 (1947) pp. 438–446.

11. A. G. WALKER, *On Ruse's spaces of recurrent curvature*, Proc. London Math. Soc. vol. 52 (1950) pp. 36–64.

THE SORBONNE,
PARIS, FRANCE.

EXTREMAL METHODS AND GEOMETRIC THEORY OF FUNCTIONS OF A COMPLEX VARIABLE

COEFFICIENT REGIONS FOR SCHLICHT FUNCTIONS

A. C. SCHAEFFER AND D. C. SPENCER

Our purpose is to give a survey of various methods which have been developed in order to attack a certain type of extremal problem in conformal mapping. Although the methods are quite general, we shall confine ourselves for the most part to a concrete problem—the so-called coefficient problem for functions which are regular and schlicht in the interior of the unit circle.

A function $f(z)$ is said to be schlicht in a domain of the z-plane if it assumes distinct values for distinct values of z. We consider functions which are regular and schlicht in $|z| < 1$ and which therefore map $|z| < 1$ onto subdomains of the plane, and we assume that the developments in powers of z have the form

$$f(z) = z + a_2 z^2 + \cdots + a_n z^n + \cdots, \qquad |z| < 1. \tag{1}$$

The class of these functions will be denoted by S.

For each n, $n \geqq 2$, let the precise region in complex space of $n-1$ dimensions occupied by points $(a_2, a_3, \cdots, a_n)$ corresponding to functions of class S be denoted by V_n. The coefficient problem is the problem of finding for each n, $n \geqq 2$, the region V_n or, in other words, of expressing the points of V_n in terms of $2n-2$ real parameters. The regions V_n are closed cells (see [14c]).

This problem had its roots in the investigations of Koebe on the uniformization principle (see [8]). The first significant method applied to the coefficient problem is the so-called "area-principle" discovered by Gronwall [4] in 1914 and rediscovered two years later by Faber [3] who, together with Bieberbach, Pick, and others, applied it to prove that the region V_2 is the circle $|a_2| \leqq 2$. The next method, which penetrated much deeper, is the representation developed by Löwner [9]. Löwner gave a representation for the coefficients a_ν of a class of schlicht functions which lie everywhere dense in S, and this representation can be extended to include all functions of S (see [14a], [17]). In fact, the region P_n of points $(c_1, c_2, \cdots, c_{n-1})$ belonging to functions

$$p(z) = 1 + 2\sum_{\nu=1}^{\infty} c_\nu z^\nu \tag{2}$$

of positive real part in $|z| < 1$ has been characterized by Carathéodory [2], and to each point $(a_2, a_3, \cdots, a_n)$ of V_n there is a curve $(c_1(\tau), \cdots, c_{n-1}(\tau))$ in P_n such that

$$a_k = \sum (-1)^\mu \Gamma_{\alpha_1 \alpha_2 \cdots \alpha_\mu} \int_0^\infty \cdots \int_0^\infty \exp\left[-\sum_1^\mu \alpha_\nu \tau_\nu\right] \prod_1^\mu c_{\alpha_\nu}(\tau_\nu)\, d\tau_1 \cdots d\tau_\mu, \tag{3}$$

$$k = 2, \cdots, n,$$

where

$$\Gamma_{\alpha_1\alpha_2\cdots\alpha_\mu} = 2^\mu(k-\alpha_1)(k-\alpha_1-\alpha_2)\cdots(k-\alpha_1-\alpha_2\cdots-\alpha_\mu),$$

the $\alpha_1, \alpha_2, \cdots, \alpha_\mu$ being positive integers with sum $k-1$. However, since the formula (3) expresses each point of V_n in terms of a curve in P_n, the representation is not one involving finitely many parameters.

The coefficient problem was considered by Peschl [10]. If $f(z)$ belongs to the point $(a_2, a_3, \cdots, a_n)$ of V_n, the function

$$g(z) = \frac{f(z)}{zf'(z)} = 1 + 2\sum_{\nu=1}^{\infty} u_\nu z^\nu \tag{4}$$

defines a point $(u_1, u_2, \cdots, u_{n-1})$ of a region U_n, and the mapping from V_n onto U_n is one–one and analytic. A star-like schlicht function (1) maps the interior of the unit circle onto a domain every point of which can be connected to the origin by a ray which lies interior to the region, and it is characterized by the property that the corresponding $g(z)$ defined by (4) is regular with positive real part in $|z| < 1$. Let $S(z)$ be star-like, S^{-1} its inverse. For $t \geqq 0$ the function $S^{-1}[e^{-t}S(z)]$ maps $|z| < 1$ onto a domain contained in the unit circle, so that

$$F(z) = e^t f\{S^{-1}[e^{-t}S(z)]\} \tag{5}$$

is of class S. Hence

$$G(z) = \frac{F(z)}{zF'(z)} = 1 + 2\sum_{\nu=1}^{\infty} u_\nu(t)z^\nu \tag{6}$$

defines a point $(u_1(t), u_2(t), \cdots, u_{n-1}(t))$ of U_n, and as t varies from 0 to ∞ this point traces a curve which extends from the given point $(u_1, \cdots, u_{n-1}) = (u_1(0), \cdots, u_{n-1}(0))$ belonging to $g(z)$ to the point $(c_1, \cdots, c_{n-1}) = (u_1(\infty), \cdots, u_{n-1}(\infty))$ of P_n belonging to the function $S(z)/(zS'(z))$. Let $(u_1, u_2, \cdots, u_{n-1})$ be a boundary point of U_n. As $(c_1, \cdots, c_{n-1})$ varies over the boundary of P_n, the tangent vectors of these curves at the initial point $(u_1, u_2, \cdots, u_{n-1})$ sweep out a cone $C_n(u_1, u_2, \cdots, u_{n-1})$. More generally, given any closed region U_n^* containing P_n, these direction cones can be formed at its boundary points. Peschl characterizes U_n in terms of these tangent cones.

Schlicht functions (1) with all coefficients real form a subclass of the typically-real functions considered by Rogosinski [11]. A function

$$h(z) = z + t_2z^2 + \cdots + t_nz^n + \cdots \tag{7}$$

is said to be typically-real if it is regular in $|z| < 1$ and if $\operatorname{Im} h(z)$ and $\operatorname{Im}(z)$ have the same sign there. If $h(z)$ is typically-real, then

$$p(z) = \frac{1-z^2}{z}h(z) \tag{8}$$

is a function (2) of positive real part with real coefficients, and conversely.

Hence the region T_n of points $(t_2, \cdots, t_n)$ is obtained from the region P_n by the simple mapping

$$(9) \quad 2c_k = t_{k+1} - t_{k-1}, \qquad t_k = \begin{cases} 2c_1 + 2c_3 + \cdots + 2c_{k-1}, & k \text{ even} \\ 1 + 2c_2 + 2c_4 + \cdots + 2c_{k-1}, & k \text{ odd.} \end{cases}$$

If R_n is the region of points $(a_2, \cdots, a_n)$ belonging to schlicht functions (1) with all coefficients real, it is readily shown that T_n is simply the convex closure of R_n.

Grunsky [5] gave an interesting set of inequalities which characterize the points $(a_2, \cdots, a_n)$ of V_n but these inequalities, infinite in number, do not define V_n in terms of a finite number of parameters. Grunsky's inequalities have recently been given a new significance (see [1]), and in their most general form for Riemann surfaces they express conditions under which one surface may be imbedded in another (see [16b]).

A representation of the points of V_n in terms of finitely many parameters is achieved by using variational methods (see [14b, c]).[1] If $F(a_2, \bar{a}_2, \cdots, a_n, \bar{a}_n)$ is a real-valued, continuously differentiable function defined in some region containing V_n in its interior and if grad $F \neq 0$ in V_n, then F attains its maximum in V_n at some boundary point $(a_2, a_3, \cdots, a_n)$. Let $f(z)$ belong to this boundary point, and consider a neighboring function $f^*(z)$ of the class S with coefficients a_k^*, $k = 2, 3, \cdots$. Writing $\delta F = F(a_2^*, \bar{a}_2^*, \cdots, a_n^*, \bar{a}_n^*) - F(a_2, \bar{a}_2, \cdots, \bar{a}_n, a_n)$, we see that $\delta F \leqq 0$, and we then find that the extremalizing function $w = f(z)$ satisfies a differential equation of the form

$$(10) \qquad \left(\frac{z}{w}\frac{dw}{dz}\right)^2 P(w) = Q(z)$$

where

$$(11) \qquad P(w) = \sum_{\nu=1}^{n-1} \frac{A_\nu}{w^\nu},$$

$$(12) \qquad Q(z) = \sum_{\nu=-(n-1)}^{n-1} \frac{B_\nu}{z^\nu}, \qquad B_0 > 0, B_{-\nu} = \bar{B}_\nu.$$

The function $Q(z)$ is non-negative on $|z| = 1$ with at least one zero there. Thus $Q(z)(dz/z)^2$ is a negative quadratic differential of the region $|z| < 1$, while $P(w)(dw/w)^2$ is a quadratic differential of the w-sphere.

The variational method gives necessary conditions in order that a function f of class S should extremalize an arbitrary function of its first n coefficients, and a theorem of Teichmüller [18] complements this result by proving that the necessary conditions are sufficient. To each point $(a_2, a_3, \cdots, a_n)$ on the boundary of V_n there belongs precisely one function $w = f(z)$ of class S and this function satisfies an equation of the form (10). Conversely, any function $w = f(z)$ of the form (1) which is regular in $|z| < 1$ and satisfies an equation of type (10) necessarily belongs to a boundary point of V_n.

[1] The application of variational methods began with the paper [15].

The study of the boundary of V_n therefore turns on a detailed investigation of the differential equations (10). We observe that, since $Q(z)$ is non-negative on $|z| = 1$ with at least one zero there, the coefficient B_0 is uniquely determined by $B_1, B_2, \cdots, B_{n-1}$; thus $Q(z)$ is determined by the vector $(B_1, B_2, \cdots, B_{n-1})$. The first question which arises concerns the relationship between the vectors $(A_1, A_2, \cdots, A_{n-1})$ and $(B_1, B_2, \cdots, B_{n-1})$ in order that the equation (10) should have a solution which is regular in $|z| < 1$. This relationship is investigated in detail in [14c].

The equation (10), written in the form

$$P(w)\left(\frac{dw}{w}\right)^2 = Q(z)\left(\frac{dz}{z}\right)^2, \tag{13}$$

expresses the equality of two quadratic differentials. Since $Q(z)(dz/z)^2 \leqq 0$ on $|z| = 1$, it is important to investigate the hyperelliptic trajectories of the w-sphere along which $P(w)(dw/w)^2 \leqq 0$. From each zero and simple pole[2] of the quadratic differential $P(w)(dw/w)^2$ prolong all loci on which $P(w)(dw/w)^2 \leqq 0$. These finitely many loci plus their limit points define a set Γ_w. It is proved in [14c] that Γ_w consists of finitely many analytic arcs, each arc being analytic up to and including its end-points. An important property of Γ_w is that any subcontinuum of it which does not contain the origin $w = 0$ is necessarily simply-connected. The corresponding set Γ_z defined by $Q(z)(dz/z)^2$ is also composed of finitely many closed analytic arcs; it is symmetric about $|z| = 1$ and, apart from $|z| = 1$, its only Jordan curves pass through $z = 0$ or $z = \infty$.

The trajectories defined by $R(w)\,dw^2 \leqq 0$, where $R(w)$ is an arbitrary rational function, are investigated in the paper [6]. In particular, it is proved that if $R(w)dw^2$ has at most three poles, then the set Γ_w is composed of finitely many closed analytic arcs. This result is not generally true when there are more than three poles.

A function $w = f(z)$ of form (1) which is regular in $|z| < 1$ and satisfies (13) maps $|z| < 1$ onto a domain D_w which is obtained from the w-sphere by removing a subcontinuum C_w of Γ_w, where C_w contains $w = \infty$ but not $w = 0$. The differential $P(w)(dw/w)^2$ is a negative quadratic differential of D_w, and equation (13) expresses its invariance under a conformal mapping $w = f(z)$.

On the other hand, let z' and z'' be a pair of points of $|z| = 1$ which map by $w = f(z)$ into points w', w'' which define the same point of the sphere but lie on opposite edges of a slit of C_w. If we identify all such pairs z', z'' on $|z| = 1$, we make $|z| \leqq 1$ into a closed Riemann surface R of genus zero and $w = f(z)$ maps R onto the w-sphere. Equation (13) expresses the invariance of the quadratic differential $P(w)(dw/w)^2$, which is regarded as belonging to the sphere.

We thus have dual interpretations of the equation (13) according as we regard the quadratic differential $P(w)(dw/w)^2$ as belonging to D_w or to the sphere. The point of view adopted in [14c] is to regard $P(w)(dw/w)^2$ as belonging to the sphere.

[2] If $A_1 \neq 0$, $w = \infty$ is a simple pole; this is seen by introducing $t = 1/w$ as uniformizer.

Given $(B_1, B_2, \cdots, B_{n-1})$, $Q(z)$ is determined and with it the metric defined by

$$ds^2 = |Q(z)| \cdot \left|\frac{dz}{z}\right|^2. \tag{14}$$

Let a Riemann surface R be formed from the region $|z| \leqq 1$ by identifying m pairs of arcs on $|z| = 1$ such that: (i) the circle $|z| \leqq 1$ with $2m$ arcs on $|z| = 1$ identified in pairs is equivalent to the sphere; (ii) in the identification of any pair of arcs the metric (14) is preserved; (iii) if two identified arcs abut, the common end-point is a zero of $Q(z)$ with at most one exception. If this exceptional point exists, we call it Σ; otherwise let Σ be any class of equivalent points on $|z| = 1$. The function $w = f(z)$, of the form (1) for $|z| < 1$, which maps R onto the sphere with Σ going into $w = \infty$ defines a boundary point of V_n and it satisfies an equation (13) where $Q(z)$ is determined by the given vector $(B_1, B_2, \cdots, B_{n-1})$. The identification of arcs on $|z| = 1$ can be expressed in terms of finitely many parameters, and these plus the parameters defining $(B_1, B_2, \cdots, B_{n-1})$ yield a parametrization of the boundary of V_n. This is the method used in [14c].

A function $v(z)$ will be said to realize the identification of a pair of arcs on $|z| = 1$ if it takes the same value at identified points of the arcs. The identification of arcs on $|z| = 1$ may be realized continuously by a function

$$v(z, t) = e^{-t}\{z + \beta_2(t)z^2 + \cdots\}, \qquad v(z, 0) = z, \tag{15}$$

which maps $|z| < 1$ onto $|v| < 1$ minus slits whose edges correspond to pairs of arcs on $|z| = 1$ the identification of which is realized by $v(z, t)$. The function $v(z, t)$ satisfies an equation

$$Q(v, t)\left(\frac{dv}{v}\right)^2 = Q(z)\left(\frac{dz}{z}\right)^2, \tag{16}$$

where

$$Q(z, t) = \sum_{\nu=-(n-1)}^{n-1} \frac{B_\nu(t)}{z^\nu}, \qquad B_0(t) > 0, \qquad B_{-\nu}(t) = \overline{B_\nu(t)}, \tag{17}$$

$$Q(z, 0) = Q(z).$$

For each t the function $Q(z, t)$ is non-negative on $|z| = 1$ with at least one zero there. In its dependence on t, $v(z, t)$ satisfies an equation of Löwner type, namely

$$\frac{\partial v}{\partial t} = -v\frac{1 + \kappa(t)v}{1 - \kappa(t)v} \tag{18}$$

where $|\kappa(t)| = 1$, $Q(1/\kappa(t), t) = 0$. Furthermore,

$$\frac{\partial Q(z, t)}{\partial t} = z\frac{\partial Q(z, t)}{\partial z}\frac{1 + \kappa(t)z}{1 - \kappa(t)z} + \frac{4\kappa(t)z}{(1 - \kappa(t)z)^2}Q(z, t). \tag{19}$$

As t tends to $+\infty$,

$$e^t v(z, t) \to f(z), \qquad Q(e^{-t}w, t) \to P(w). \tag{20}$$

Let $z(v) = z(v, t)$ be the function inverse to $v(z, t)$, and define

$$w(v, t) = f(z(v)) = e^t\{v + a_2(t)v^2 + \cdots + a_n(t)v^n + \cdots\}. \tag{21}$$

The function $w(z, t)$ maps $|z| < 1$ onto a subcontinuum $C_w(t)$ of the set Γ_w defined by $P(w)(dw/w)^2$, and in its dependence on z and t it satisfies the equations

$$P(w)\left(\frac{dw}{w}\right)^2 = Q(z, t)\left(\frac{dz}{z}\right)^2, \tag{22}$$

$$\frac{\partial w}{\partial t} = z\frac{\partial w}{\partial z}\frac{1 + \kappa(t)z}{1 - \kappa(t)z}. \tag{23}$$

The equation (23) is the Löwner equation which is the dual of (18). Thus the Löwner equations arise automatically from the continuous realization of the identification (see [14c]).

Interior points of V_n can be characterized by the property that bounded functions belong to them, and the parametrization of the interior of V_n is accomplished by considering the class of functions

$$v(z) = b_1 z + b_2 z^2 + \cdots + b_n z^n + \cdots, \qquad 0 < b_1 \leqq 1, \tag{24}$$

which are regular, schlicht, and bounded by 1 in $|z| < 1$. As $v(z)$ ranges over all such functions, the point $(b_1, b_2, \cdots, b_n)$ sweeps out a region B_n in euclidean space of $2n - 1$ real dimensions. The region B_n is not compact, but is made so by addition of the point $b_1 = 0, b_2 = 0, \cdots, b_n = 0$. To each point $(b_1, b_2, \cdots, b_n)$ of B_n other than $b_1 = 0, b_2 = 0, \cdots, b_n = 0$ there corresponds an interior point $(a_2, a_3, \cdots, a_n)$ of V_n where $a_k = b_k/b_1$, and the correspondence between the boundary of B_n and the interior of V_n is topological (see [12]). Thus the interior of V_n is described by parametrizing the boundary of B_n. The boundary functions $v(z)$ of B_n satisfy equations of the form

$$R(v)\left(\frac{dv}{v}\right)^2 = S(z)\left(\frac{dz}{z}\right)^2 \tag{25}$$

where $R(z)(dz/z)^2$ and $S(z)(dz/z)^2$ are negative quadratic differentials of $|z| < 1$ which are of the same form as $Q(z)(dz/z)^2$. If $v(z)$ belongs to the boundary point $(b_1, b_2, \cdots, b_n)$ of B_n, the function $f(z) = v(z)/b_1$, which belongs to the point $(a_2, a_3, \cdots, a_n)$, $a_k = b_k/b_1$, is characterized by the property that, among all functions of class S whose coefficients $a_2, a_3, \cdots, a_n$ have the given values, it is the function whose maximum modulus in $|z| < 1$ is a minimum. So far as the region V_n is concerned, it is therefore sufficient to consider only functions which satisfy equations of form (13) (boundary functions) or functions which satisfy an equation (25) (interior functions). Since equation (16) is of the same form as equation (25), the continuous realization of the identification on $z = 1$ yields curves which sweep out the boundary of B_n as well as the interior of V_n.

The curve $(a_2(t), a_3(t), \cdots, a_n(t))$ defined by (21) lies on the boundary of V_n and is the dual of the curve defined by (15) which lies on the boundary of B_n. These curves may be regarded as characteristic curves of the partial differential equations for the boundaries of V_n, B_n (see [13], [16a]). From this point of view the boundaries appear as integral conoids of their characteristic curves.

By way of illustration, let us find the equations for the characteristic curves which lie on the boundary of B_n. The quadratic differentials appearing in (25) have the forms (see [16a])

$$R(v)\left(\frac{dv}{v}\right)^2 = \left(\frac{dv}{v}\right)^2 \sum_{k=1}^{n} [F_k D_k(v) + \bar{F}_k \tilde{D}_k(v)], \tag{26}$$

$$S(z)\left(\frac{dz}{z}\right)^2 = \left(\frac{dz}{z}\right)^2 \sum_{k=1}^{n} [F_k d_k(z) + \bar{F}_k \tilde{d}_k(z)], \tag{27}$$

where

$$D_k(v) = b_k + 2\sum_{\rho=2}^{k} \frac{b_k^{(\rho)}}{v^{\rho-1}}, \qquad \tilde{D}_k(v) = \bar{b}_k + 2\sum_{\rho=2}^{k} \bar{b}_k^{(\rho)} v^{\rho-1}, \tag{28}$$

$$d_k(z) = kb_k + 2\sum_{\nu=1}^{k-1} \frac{\nu b_\nu}{z^{k-\nu}}, \qquad \tilde{d}_k(z) = k\bar{b}_k + 2\sum_{\nu=1}^{k-1} \nu \bar{b}_\nu z^{k-\nu}. \tag{29}$$

Here $b_k^{(\rho)}$ denotes the kth coefficient of $\{v(z)\}^\rho$ and, at a regular point of the boundary of B_n, $(F_1, F_2, \cdots, F_n)$ is the normal vector.

Now $R(v)$ has a double zero at a point $v = 1/\kappa$, $|\kappa| = 1$. That is,

$$\sum_{k=1}^{n} \left[F_k D_k\left(\frac{1}{\kappa}\right) + \bar{F}_k \tilde{D}_k\left(\frac{1}{\kappa}\right)\right] = 0, \tag{30}$$

$$\sum_{k=1}^{n} \left[F_k D_k'\left(\frac{1}{\kappa}\right) + \bar{F}_k \tilde{D}_k'\left(\frac{1}{\kappa}\right)\right] = 0. \tag{31}$$

The quantity $1/\kappa$ may be regarded as eliminated from (30) by means of the equation (31), so that (30) is an equation of the form

$$E(b_1, b_2, \bar{b}_2, \cdots, b_n, \bar{b}_n; F_1, F_2, \bar{F}_2, \cdots, F_n, \bar{F}_n) = 0. \tag{32}$$

If the boundary of B_n were given by an equation $F(b_1, b_2, \bar{b}_2, \cdots, b_n, \bar{b}_n) = 0$, F real, we would have

$$F_k = \frac{\partial F}{\partial b_k}, \qquad \bar{F}_k = \frac{\partial F}{\partial \bar{b}_k}, \tag{33}$$

and (32) would be a partial differential equation of the first order for the boundary of B_n. The differential equations for the characteristics are

$$\frac{db_k}{dt} = \frac{\partial E}{\partial F_k}, \qquad \frac{d\bar{b}_k}{dt} = \frac{\partial E}{\partial \bar{F}_k}, \qquad \frac{dF_k}{dt} = -\frac{\partial E}{\partial b_k}, \qquad \frac{\partial \bar{F}_k}{\partial t} = -\frac{\partial E}{\partial \bar{b}_k},$$
$$\sum_{k=1}^{n} \left\{F_k \frac{\partial E}{\partial F_k} + \bar{F}_k \frac{\partial E}{\partial \bar{F}_k}\right\} = 0, \tag{34}$$

where t is a real parameter. In applying these formulas, it is necessary to take into account that κ depends on the F_k and the b_k. However, since the differentiation with respect to κ gives a zero contribution by (31), we may treat κ as a constant in deriving the characteristic equations. From the first equation (34) we obtain, after replacing t by $-t$,

$$\frac{db_k}{dt} = -b_k - 2\sum_{\rho=2}^{k} b_k^{(\rho)} \kappa^{\rho-1}. \tag{35}$$

Multiplying both sides of this equation by z^k and summing on k from 1 to ∞, we obtain equation (18). The function $v(z, t)$ satisfies the equation

$$R(v, t)\left(\frac{dv}{v}\right)^2 = S(z)\left(\frac{dz}{z}\right)^2 \tag{36}$$

where

$$\frac{\partial R(v, t)}{\partial t} = v\frac{\partial R(v, t)}{\partial v}\frac{1 + \kappa(t)v}{1 - \kappa(t)v} + \frac{4\kappa(t)v}{(1 - \kappa(t)v)^2} R(v, t); \qquad \frac{\partial S}{\partial t} = 0. \tag{37}$$

Since $S(z)$ has a double zero at a point $z = 1/\eta$, $|\eta| = 1$, we obtain, in an analogous manner, as law of variation along the dual set of characteristics:

$$\frac{\partial v}{\partial t} = z\frac{\partial v}{\partial z}\frac{1 + \eta(t)z}{1 - \eta(t)z}. \tag{38}$$

The function $v(z, t)$ satisfies the equation

$$R(v)\left(\frac{dv}{v}\right)^2 = S(z, t)\left(\frac{dz}{z}\right)^2 \tag{39}$$

where

$$\frac{\partial R}{\partial t} = 0, \qquad \frac{\partial S(z, t)}{\partial t} = z\frac{\partial S(z, t)}{\partial z}\frac{1 + \eta(t)z}{1 - \eta(t)z} + \frac{4\eta(t)z}{(1 - \eta(t)z)^2} S(z, t). \tag{40}$$

In the variation defined by (18) the identification of points on $|z| = 1$ is fixed and is being continuously realized; in the variation (38) the set Γ_v defined by $R(v)(dv/v)^2$ is fixed and the function v maps $|z| < 1$ onto $|v| < 1$ minus slits of varying lengths which belong to Γ_v.

The coefficient problem may be generalized to mappings of a fixed Riemann surface M onto subdomains of another fixed surface R (see [16a, b]). In the special cases discussed above, M is the interior of the unit circle and R is either the punctured sphere (mappings (1)) or the interior of the unit circle (mappings (24)). In particular, we may take M and R to be multiply-connected domains of the plane. The theory of characteristic curves provides an apparatus for generalizing the Löwner equations (18) and (34), and for multiply-connected domains of the plane we obtain equations of the form

$$\frac{\partial v}{\partial t} = \frac{\partial v}{\partial z} h(z, t) - H(v, t) \tag{41}$$

where $h(z, t)$ is a reciprocal differential in its dependence on z. The term involving $h(z, t)$ provides a variation of the type (38) while the term $H(v, t)$ provides a variation of the type (18). From another point of view, equation (41) is a generalization not only of Löwner's variation but also of Julia's (see [7]).

References

1. S. Bergman and M. Schiffer, *Kernel functions and conformal mapping*, Compositio Math. vol. 8 (1951) pp. 205–249.

2. C. Carathéodory, *Über den Variabilitätsbereich der Fourier'schen Konstanten von positiven harmonischen Funktionen*, Rend. Circ. Mat. Palermo vol. 12 (1911) pp. 193–217.

3. G. Faber, *Neuer Beweis eines Koebe-Bieberbachschen Satzes über konforme Abbildung*, K. B. Akademie der Wissenschaften zu München, Sitzungsberichte der Mathematisch-physische Klasse (1916) pp. 39–42.

4. T. H. Gronwall, *Some remarks on conformal representation*, Ann. of Math. vol. 16 (1914–1915) pp. 72–76.

5. H. Grunsky, *Koeffizientenbedingungen für schlicht abbildende meromorphe Funktionen*, Math. Zeit. vol. 45 (1939) pp. 29–61.

6. J. A. Jenkins and D. C. Spencer, *Hyperelliptic trajectories*, Ann. of Math. vol. 53 (1951) pp. 4–35.

7. G. Julia, *Sur une équation aux dérivées fonctionnelles liée à la représentation conforme*, Ann. École Norm. (3) vol. 39 (1922) pp. 1–28.

8. P. Koebe, *Über die Uniformisierung beliebiger analytischer Kurven*, Nachr. Ges. Wiss. Göttingen (1907) pp. 191–210.

9. L. Löwner, *Untersuchungen über schlichte konforme Abbildungen des Einheitskreises*, Math. Ann. vol. 89 (1923) pp. 103–121.

10. E. Peschl, *Zur Theorie der schlichten Funktionen*, J. Reine Angew. Math. vol. 176 (1937) pp. 61–94.

11. W. Rogosinski, *Über positive harmonische Entwicklungen und typisch-reele Potenzreihen*, Math. Zeit. vol. 35 (1932) pp. 93–121.

12. H. L. Royden, *The coefficient problem for bounded schlicht functions*, Proc. Nat. Acad. Sci. U. S. A. vol. 35 (1949) pp. 657–662.

13. A. C. Schaeffer, M. Schiffer, and D. C. Spencer, *The coefficient regions of schlicht functions*, Duke Math. J. vol. 16 (1949) pp. 493–527.

14. A. C. Schaeffer and D. C. Spencer, (a) *The coefficients of schlicht functions*, II, Duke Math. J. vol. 12 (1945) pp. 107–125.

(b) *The coefficients of schlicht functions*, III, Proc. Nat. Acad. Sci. U. S. A. vol. 32 (1946) pp. 110–116.

(c) *Coefficient regions for schlicht functions*, Amer. Math. Soc. Colloquium Publications, vol. 35, New York, 1950.

15. M. Schiffer, *A method of variation within the family of simple functions*, Proc. London Math. Soc. (2) vol. 44 (1938) pp. 432–449.

16. M. Schiffer and D. C. Spencer, (a) *The coefficient problem for multiply-connected domains*, Ann. of Math. vol. 52 (1950) pp. 362–402.

(b) *Functionals of finite Riemann surfaces*, Princeton University Press (to appear)

17. D. C. Spencer, *Some problems in conformal mapping*, Bull. Amer. Math. Soc. vol. 53 (1947) pp. 417–439.

18. O. Teichmüller, *Ungleichungen zwischen den Koeffizienten schlichter Funktionen*, Preuss. Akad. Wiss. Sitzungsber. (1938) pp. 363–375.

University of Wisconsin,
Madison, Wis., U. S. A.
Princeton University,
Princeton, N. J., U. S. A.

VARIATIONAL METHODS IN THE THEORY OF CONFORMAL MAPPING

MENAHEM SCHIFFER

1. Let D be a domain in the complex z-plane which is bounded by n smooth curves C_ν ($\nu = 1, 2, \cdots, n$) which form the boundary C of D. A major problem in conformal mapping consists in determining univalent analytic functions $f(z)$ in D which map it on a canonical domain of specified type. The most important canonical maps can be easily expressed in terms of the Green's function of the domain. Green's function $g(z, \zeta)$ depends on its two argument points and on the domain D (or on its boundary curve system C). The latter dependence is of transcendental character and its investigation belongs to functional analysis as pointed out by Volterra and Hadamard. The theory of schlicht functions is also closely related to this field of research; for, every schlicht function, say, in the unit circle can be characterized by the domain upon which it maps and can easily be expressed in terms of the Green's function of the image domain. Thus, extremum problems with respect to schlicht functions can be reduced to extremum problems for Green's function and are incorporated into the theory of the latter.

A fundamental tool in the study of functionals are variation formulas which express the rate of change of the functional under given deformation of its domain. The value of such formulas for extremum problems, comparison formulas, etc. is evident. We shall discuss here various types of variation methods and compare their relative merits. Actual applications and illustrations by concrete problems will not be given here for lack of space.

2. The first and already very general variational formula for Green's function was given by Hadamard [5]. Let the domain D be deformed into D^* by subjecting every boundary point to a normal shift δn (counted positive along the interior normal vector n); the variation of the Green's function is that expressed by Hadamard's formula

$$\delta g(z, \zeta) = -\frac{1}{2\pi} \int_C \frac{\partial g(z, t)}{\partial n_t} \frac{\partial g(t, \zeta)}{\partial n_t} \delta n_t \, ds_t . \tag{1}$$

Immediate applications of (1) are obvious; let $g(z, \zeta) = \log (1 / |z - \zeta|) + h(z, \zeta)$. We observe from (1) that the functionals $h(z, z) + h(\zeta, \zeta) - 2h(z, \zeta)$ and $\sum_{i,k=1}^{N} h(z_i, z_k)\alpha_i\alpha_k$ vary monotonically with the domain and useful inequalities and comparison formulas follow directly.

Let us next define for z and ζ on C the functional of the curve system C

$$\mathfrak{G}(z, \zeta) = -\frac{1}{2\pi} \frac{\partial^2 g(z, \zeta)}{\partial n_z \partial n_\zeta}. \tag{2}$$

It is easily seen that $\mathfrak{G}$ is negative on C and that knowledge of it is sufficient to solve Dirichlet's problem with respect to D. It has the elegant variational formula

$$\delta\mathfrak{G}(z,\zeta) = \int_C \mathfrak{G}(z,t)\mathfrak{G}(t,\zeta)\,\delta n_t\,ds_t. \tag{3}$$

The monotony of $\mathfrak{G}(z,\zeta)$ as a functional of D is obvious from (3) and the definiteness of $\mathfrak{G}$, and, hence, useful comparison formulas can be established for it. Formula (3) must, however, be used with caution since $\mathfrak{G}(z,\zeta)$ has an infinity for $z=\zeta$. A permissible deformation must leave arcs of C around z and ζ unchanged if (3) is to be used. Otherwise, corrective terms of rather complicated nature arise and the integral in (3), moreover, becomes improper.

One can avoid these difficulties by introducing the two complex functions [13]

$$K(z,\bar\zeta) = -\frac{2}{\pi}\frac{\partial^2 g(z,\zeta)}{\partial z\partial\bar\zeta}, \quad L(z,\zeta) = -\frac{2}{\pi}\frac{\partial^2 g(z,\zeta)}{\partial z\partial\zeta} = \frac{1}{\pi(z-\zeta)^2} - l(z,\zeta), \tag{4}$$

where $\partial/\partial z = (\partial/\partial x - i\partial/\partial y)/2$ and $\partial/\partial\bar z = (\partial/\partial x + i\partial/\partial y)/2$. Both functions are analytic in all their arguments; K is regular throughout D while L has a double pole at $z=\zeta$ as exhibited in (4). These two new functionals of D satisfy the following variational equations which are derived from (1) by differentiation and using the boundary behavior of Green's function:

$$\delta K(z,\bar\zeta) = \int_C K(z,\bar t)K(t,\bar\zeta)\,\delta n_t\,ds_t, \quad \delta L(z,\zeta) = \int_C K(z,\bar t)L(t,\zeta)\,\delta n_t\,ds_t. \tag{5}$$

3. The functions K and L to which we are led quite naturally by the formalism of the variational formulas play an important role in conformal mapping [3]. They satisfy the following equations on the boundary C of D:

$$\overline{z'K(z,\bar\zeta)} = z'L(z,\zeta), \qquad z\in C, \qquad \zeta\in D, \qquad z' = \text{tangent vector at } z \text{ to } C; \tag{6}$$

$$z'K(z,\bar\zeta)\bar\zeta' = z'L(z,\zeta)\zeta' = \mathfrak{G}(z,\zeta), \qquad z\in C, \qquad \zeta\in C. \tag{6'}$$

The connection between the functional (2) and the two new functions K and L is thus established. $K(z,\bar\zeta)$ satisfies the identity

$$\iint_D K(z,\bar\zeta)f(\zeta)\,d\tau_\zeta = f(z), \qquad d\tau_\zeta = \text{area element in } \zeta, \tag{7}$$

for every analytic function $f(z)$ in D with $\iint_D |f|^2\,d\tau < \infty$. From this reproducing property follows easily a representation of $K(z,\bar\zeta)$ in terms of any complete orthonormal set $\{\varphi_\nu(z)\}$ in D:

$$K(z,\bar\zeta) = \sum_{\nu=1}^{\infty}\varphi_\nu(z)\overline{\varphi_\nu(\zeta)}. \tag{7'}$$

The kernel function $K(z,\bar\zeta)$ was defined by Bergman [1] just by this formula and

various applications to conformal mapping were given by him from this definition. (7′) lends itself also easily to a generalization for the case of analytic functions of several complex variables. The kernel $L(z, \zeta)$ can be easily constructed in terms of $K(z, \bar{\zeta})$ by means of the identity

$$L(z, \zeta) = \frac{1}{\pi(z - \zeta)^2} - \frac{1}{\pi} \iint_D K(z, \bar{t})(t - \zeta)^{-2} \, d\tau_t, \tag{8}$$

which shows that from $K(z, \bar{\zeta})$ ultimately Green's function can be obtained.

Let D_0 be a domain containing D and denote its corresponding kernels by K_0 and L_0. We observe that $L(z, \zeta) - L_0(z, \zeta)$ is regular analytic in D and we obtain by use of (6), after easy calculation, the identity, valid for any two points z and ζ in D [15]:

$$\begin{aligned} &\iint_D (L(z, t) - L_0(z, t))\overline{(L(t, \zeta) - L_0(t, \zeta))} \, d\tau_t \\ &\qquad = K(z, \bar{\zeta}) - K_0(z, \bar{\zeta}) - \iint_{D_0 - D} L_0(z, t) \, \overline{L_0(t, \zeta)} \, d\tau_t . \end{aligned} \tag{9}$$

Suppose now that the boundary C_0 of D_0 lies in a Fréchet neighborhood of order ϵ to the corresponding curve system C. Then, we shall have $K - K_0 = O(\epsilon)$ and $L - L_0 = O(\epsilon)$. Introducing these estimates into (9), we arrive at

$$\delta K_0(z, \bar{\zeta}) = \iint_{D_0 - D} L_0(z, t)\overline{L_0(t, \zeta)} \, d\tau_t + O(\epsilon^2) \tag{10}$$

which, in view of (6), can be easily brought into the form of the first formula (5). Thus, (9) represents a finite comparison formula which goes for small deformations into the Hadamard form. One can use (9) as the basic formula in a process of successive approximation for $K - K_0$ and express this difference as an infinite series of iterated integrals containing L_0 and K_0 only [15]. Formula (9) gives also rise to numerous inequalities; in particular, Grunsky's inequalities [4] for the coefficients of schlicht functions can be easily derived from it.

4. An essential assumption for the application of Hadamard's formula is the smoothness of the boundary C, since the normal of C plays a distinguished role in the formula. Thus, the formula is inapplicable for very many domains possessing a Green's function; this fact precludes the use of Hadamard's formula in extremum problems of conformal mapping. One is never sure, a priori, if the extremum domain in question permits the use of Hadamard's formula and one cannot compare it in this way with neighbor domains. One may, however, transform Hadamard's formula into such a form that this difficulty is removed. We introduce an infinitesimal deformation of the whole z-plane by the formula

$$z^* = z \mp \frac{e^{i\varphi}\rho^2}{z - z_0}; \qquad \rho > 0, \quad z_0 \in D. \tag{11}$$

The curve system C is deformed into a system C^* which is, for ρ sufficiently small, very near to C and of the same character. In fact, the map (11) is univalent and analytic outside of the circle $\gamma\colon |z - z_0| < \rho$. The domain D^* determined by the curve system C^* has the Green's function $g^*(z, \zeta)$. Let $D_{z_0,\rho}$ be the domain obtained from D by removing the circle γ; we remark that $g^*(z^*(z), \zeta^*(\zeta))$ is harmonic in $D_{z_0,\rho}$, except for $z = \zeta$, and vanishes on C. Hence, we may apply Green's identity to the functions $g(z, \zeta)$ and $g^*(z^*, \zeta^*)$ with respect to $D_{z_0,\rho}$; we easily obtain:

$$g^*(z^*, \zeta^*) = g(z, \zeta) - \frac{1}{2\pi} \cdot \oint_\gamma \left\{ g^*(z^*, t^*) \frac{\partial}{\partial n_t} g(t, \zeta) - g(t, \zeta) \frac{\partial}{\partial n_t} g^*(z^*, t^*) \right\} ds_t . \tag{12}$$

Thus, by means of Green's identity we expressed the difference between the two Green's functions in terms of an integral extended over a little circle γ entirely in D and removed from the dangerous boundary C. Hadamard's general formula is also an immediate consequence of Green's identity, but our special type of variation (11) allows the further transformation from (1) to (12). It is also obvious that a description of a variation in terms of normal shift on C is impossible for general curve systems and that a variation should rather be described by a rule determining the shift of all points in the neighborhood of C. (11) is a very special case of such a rule [cf. 8], but the most general variation of the described form can be approximated arbitrarily by superposition of elementary deformations (11).

Formula (12) can easily be evaluated by series developments in the neighborhood of the point z_0 and one arrives at [11; 12]:

$$g^*(z^*, \zeta^*) = g(z, \zeta) + \operatorname{Re}\left\{ 4e^{i\varphi}\rho^2 \frac{\partial g(z_0, z)}{\partial z_0} \frac{\partial g(z_0, \zeta)}{\partial z_0} \right\} + o(\rho^2) \tag{13}$$

where $o(\rho^2)$ can be estimated uniformly in each closed subdomain Δ of $D_{z_0,\rho}$ with respect to all domains in a sufficiently small Fréchet neighborhood of D. So far, we have still to make the assumption that C is a smooth curve system in order to apply Green's identity. However, using the continuity of Green's function and all its derivatives in dependence of the domain D, we may extend the validity of (13) to the most general type of domains D bounded by n continua and possessing a Green's function. Thus, (13) is generally valid and may be used in extremum problems concerning Green's function, even if nothing is known about the nature of the boundary of the sought extremal domain. (13) has been used in the coefficient problem for schlicht and p-valued functions [11] and various related extremum problems for conformal functionals [2]. In most cases, one can show by using (13) that the extremum domain possesses piecewise analytic boundary curves C which satisfy certain differential equations [9]. A finer investigation of this domain and a study of higher order variations is then more

conveniently carried out by Hadamard type variations which become permissible, once that the analytic character of the boundary has been proved. This remark explains why the type (11) of variation is general enough for most applications and why a greater generality in variational formulas of this type would often lead to unnecessary complications.

Consider the domain $D_{z_0,\rho}$ where the boundary $| z - z_0 | = \rho$ has been removed by identifying points on it which have the same image points under the map (11). The domain thus obtained is topologically equivalent to D and D^*. Its Green's function $g^*(z, \zeta)$ is equal at each point to the right-hand side of (13). D^* is the realization of this Riemann manifold in the z-plane and (11) is the map which performs this realization. In this interpretation, we can easily extend the above type of interior variation to arbitrary domains on Riemann surfaces by changing their structure slightly through boundary correspondence along artificial holes and by realizing the domain again over the complex plane. Two further types of variation are also suggested by this point of view: (a) Hole punching: remove from D the domain $g(z, z_0) \geqq \log (1/\rho)$ which transforms D approximately into $D_{z_0,\rho}$. The Green's function of the new domain can again be computed by proper use of Green's identity. We obtain [15]:

$$g^*(z, \zeta) = g(z, \zeta) + (\log \rho)^{-1} g(z_0, z)\, g(z_0, \zeta) - \operatorname{Re}\left\{4\rho^2 \frac{\partial g(z_0, z)}{\partial z_0} \frac{\partial g(z_0, \zeta)}{\partial \bar{z}_0}\right\} + o(\rho^2) \tag{14}$$

(b) Sewing on of handles: eliminate from D the two domains $G(z; z_0, z_1) > \log (1/\rho)$ and $G(z; z_0, z_1) < \log \rho$ with $G(z; z_0, z_1) = g(z, z_0) - g(z, z_1)$. Pairs w, ω of points on the newly created boundaries c and γ near z_0 and z_1 are identified if they satisfy

$$(c) \int_{w_0}^{w} \frac{\partial}{\partial n_t} G(t; z_0, z_1)\, ds_t = (\gamma) \int_{\omega_0}^{\omega} \frac{\partial}{\partial n_t} G(t; z_0, z_1)\, ds_t. \tag{15}$$

We obtain thus a new domain D^* with larger genus than D. Its corresponding Green's function is obtained again by means of Green's identity [15]:

$$g^*(z, \zeta) = g(z, \zeta) + \frac{1}{2} (\log \rho)^{-1} G(z; z_0, z_1) G(\zeta; z_0, z_1) - \operatorname{Re}\left\{4e^{i\varphi} \rho^2 \left[\frac{\partial g(z_0, z)}{\partial z_0} \frac{\partial g(z_1, \zeta)}{\partial z_1} + \frac{\partial g(z_1, z)}{\partial z_1} \frac{\partial g(z_0, \zeta)}{\partial z_0}\right]\right\} + o(\rho^2). \tag{16}$$

Relation (15) establishes a relation $(w - z_0)(\omega - z_1) = \rho^2 e^{i\varphi} + o(\rho^2)$ between the identified points and the real constant φ in (16) is just defined by this relation. Variations (a) and (b) permit us to change connectivity and genus of the domain D and, together with the previous topology preserving variations, give a great freedom in changing the initial domain D.

5. Let $f(z)$ be schlicht in the unit circle E, $f(0) = 0$, and let it map E upon the simply-connected domain Δ with boundary curve Γ. If each point $\omega \in \Gamma$ is shifted along a vector $\delta\omega$ which points into the direction of the normal at ω and varies continuously along Γ, we obtain a domain Δ^* which is mapped by means of a schlicht function $f^*(z)$. Julia derived from Hadamard's formula (1) the following variational law [6]:

$$\delta f(z) = f'(z) \cdot \frac{1}{2\pi i} \oint_\Gamma \frac{z}{\zeta^2} \frac{\zeta + z}{\zeta - z} \frac{\delta\omega \, d\omega}{f'(\zeta)^2}, \qquad \omega = f(\zeta). \tag{17}$$

This formula contains Loewner's differential equation [7] as a limit case, namely if we let $\delta\omega$ converge to zero everywhere on Γ except near one single point where it is made to grow beyond any limit. The great advantage of Loewner's form of variation is that it permits the introduction of a simple and natural parameter and thus reduces many problems of functional analysis to problems concerning ordinary differential equations.

Let us generalize Julia's formula to the following case: A domain D in the z-plane is given and the class of schlicht functions in D is to be studied. The variation δf of a map function under a Julia variation $\delta\omega$ of the image domain Δ is to be determined. We set up the formula

$$\delta f(z) = f'(z) \cdot \frac{1}{2\pi i} \oint_\Gamma n(z, \zeta) \frac{\delta\omega \, d\omega}{f'(\zeta)^2}, \qquad \omega = f(\zeta), \tag{18}$$

where the integration is now to be extended over the whole boundary curve system Γ of Δ. In order to study $n(z, \zeta)$ suppose that $\delta\omega = 0$ along a subarc $\gamma \subset \Gamma$. Let $c \subset C$ be the corresponding arc on the boundary of D and let $z \to c$. Clearly, the image of z must lie on γ even after the variation and hence $\delta f(z)$ must have tangential direction. Thus $\delta f(z) \cdot f'(z)^{-1}$ must be a vector of tangential direction at $z \in c$ and denoting this vector by z', we easily recognize that

$$\frac{n(z, \zeta)\zeta'^2}{z'} = \text{real}, \qquad z, \zeta \in C, \tag{19}$$

is the characteristic property of the Julia kernel $n(z, \zeta)$; i.e., $n(z, \zeta)$ must be a reciprocal differential in its first argument and a quadratic differential in the second. Besides, it must have a simple pole for $z = \zeta$ as is seen from (17). Let us assume for sake of simplicity that D has $n \geqq 3$ boundary curves C_ν. There will exist two linearly independent harmonic functions $\omega_1(z)$ and $\omega_2(z)$ which are constant on each C_ν. By means of these functions, let us define

$$\Lambda(z, \zeta) = \begin{vmatrix} \frac{\partial\omega_1(z)}{\partial z} & \frac{\partial\omega_2(z)}{\partial z} \\ \frac{\partial\omega_1(\zeta)}{\partial\zeta} & \frac{\partial\omega_2(\zeta)}{\partial\zeta} \end{vmatrix}, \quad \Lambda(z, \bar{\zeta}) = \begin{vmatrix} \frac{\partial\omega_1(z)}{\partial z} & \frac{\partial\omega_2(z)}{\partial z} \\ \frac{\partial\omega_1(\zeta)}{\partial\bar{\zeta}} & \frac{\partial\omega_2(\zeta)}{\partial\bar{\zeta}} \end{vmatrix}, \quad T(z) = \begin{vmatrix} \frac{\partial\omega_1}{\partial z} & \frac{\partial\omega_2}{\partial z} \\ \frac{\partial^2\omega_1}{\partial z^2} & \frac{\partial^2\omega_2}{\partial z^2} \end{vmatrix} \tag{20}$$

and construct the kernels [14]

$$n(z,\zeta) = \pi L(z,\zeta)\frac{\Lambda(z,\zeta)}{T(z)}, \qquad m(z,\bar{\zeta}) = \pi K(z,\bar{\zeta})\frac{\Lambda(z,\bar{\zeta})}{T(z)}. \tag{21}$$

One easily verifies that $n(z,\zeta)$ has all the properties required above and that one has for ζ in D and z on C the relation:

$$\frac{n(z,\zeta)}{z'} = \overline{\left(\frac{m(z,\bar{\zeta})}{z'}\right)}. \tag{22}$$

The functions n and m are not regular throughout D but have (besides the simple pole of n for $z=\zeta$) exactly $3n-6$ poles at the zeros of T. Thus, not every variation (18) is permissible but only such which make δf regular in $D+C$; this gives $3n-6$ conditions which are just the number of moduli determining the conformal type.

We are now able to attack the following important problem of conformal mapping. Two domains D and R are given; consider the family $\mathfrak{F}$ of all functions $f(z)$ in D which map D into a domain Δ which is schlicht relative to R and develop a calculus of variations for $\mathfrak{F}$. This problem includes the problem of schlicht functions within D (R = complex plane), of bounded schlicht functions (R = circle) and of p-valued functions in D (R = Riemann surface with p sheets). We can define to every given $f(z)\in\mathfrak{F}$ a neighbor function of the same family by means of the kernels n and m of D and the corresponding kernels N and M of R. We have [14]:

$$\begin{aligned} f^*(z) = f(z) &+ \epsilon f'(z)\sum_{\nu=1}^{k}\left(r_\nu n(z,\zeta_\nu) + \bar{r}_\nu m(z,\zeta_\nu)\right) \\ &- \epsilon\sum_{\nu=1}^{k}\left(\mathrm{P}_\nu N(w,\omega_\nu) + \bar{\mathrm{P}}_\nu M(w,\bar{\omega}_\nu)\right) + o(\epsilon), \qquad \epsilon \text{ real}, \end{aligned} \tag{23}$$

where $w=f(z)$, $\omega_\nu = f(\zeta_\nu)$, and $\mathrm{P}_\nu = r_\nu f'(\zeta_\nu)^2$. One sees easily that the addition of the second right-hand term represents a tangential shift of each boundary point of Δ and that the third right-hand term has on the boundary of R again tangential direction. The last term $o(\epsilon)$ which is necessary in order to make $f^*(z)$ precisely univalent and its image domain Δ^* exactly lie in R can be estimated uniformly in each closed subdomain of D for any compact subclass of $\mathfrak{F}$. Thus, the variation formula (23) can be used in order to characterize the functions of $\mathfrak{F}$ solving significant extremum problems. It should be observed that we are not quite free in the choice of the values r_ν in (23). The formula has been constructed such that the right-hand side is regular at all points ζ_ν; however, it could become infinite at the fixed zeros of the kernels n, m and N, M which are independent of the ζ_ν. The r_ν have to be chosen such that these poles just cancel; since the number k of the arbitrary pole points ζ_ν is not bounded, one has ample possibility to keep this and finitely many other side conditions which one likes to impose on the variation. Numerous applications of this method to the coefficient problem for function classes $\mathfrak{F}$ are possible.

The coefficient problem for schlicht functions in a multiply-connected domain

has previously been attacked by a different variational method [10]. The exterior of small pieces of the boundary in the image domain was mapped conformally, and, in view of the group property of schlicht conformal mappings, this led to a variation of the original map function. Differential equations for the boundary curves of the extremal domains considered were readily derived from this procedure. Comparing it with the above general method, we may say that as long as this boundary variation is applicable it is much easier to handle and more elegant. It breaks down, however, if many side conditions are to be observed and in this case one is obliged to use the heavier but more adaptable variation (23).

Bibliography

1. S. Bergman, *Ueber die Entwicklung der harmonischen Funktionen der Ebene und des Raumes nach Orthogonalfunktionen*, Math. Ann. vol. 96 (1922) pp. 237–271.

2. G. Golusin, *Method of variations in the theory of conformal mapping*, Rec. Math. (Mat. Sbornik) N. S. vol. 19 (1946) pp. 203–236, vol. 21 (1947) pp. 83–117, 119–132.

3. P. R. Garabedian and M. Schiffer, *Identities in the theory of conformal mapping*, Trans. Amer. Math. Soc. vol. 65 (1949) pp. 187–238.

4. H. Grunsky, *Koeffizientenbedingungen fuer schlicht abbildende meromorphe Funktionen*, Math. Zeit. vol. 45 (1939) pp. 29–61.

5. J. Hadamard, *Mémoire sur le problème d'analyse relatif à l'équilibre des plaques élastiques encastrées*, Mém. d. Sav. étr. vol. 33 (1908) pp. 1–128.

6. G. Julia, *Sur une équation aux dérivées fonctionelles liées à la représentation conforme*, Ann. École Norm. (3) vol. 39 (1922) pp. 1–28.

7. K. Loewner, *Untersuchungen ueber schlichte konforme Abbildungen des Einheitskreises*, Math. Ann. vol. 89 (1923) pp. 103–121.

8. A. C. Schaeffer and D. C. Spencer, *A variational method in conformal mapping*, Duke Math. J. vol. 14 (1947) pp. 949–966.

9. ———, *Coefficient regions for schlicht functions*, Amer. Math. Soc. Colloquium Publications vol. 35, New York, 1950.

10. M. Schiffer, *A method of variation within the family of simple functions*, Proc. London Math. Soc. (2) vol. 44 (1938) pp. 432–449.

11. ———, *Variation of the Green function and theory of p-valued functions*, Amer. J. Math. vol. 65 (1943) pp. 341–360.

12. ———, *Hadamard's formula and variation of domain functions*, Amer. J. Math. vol. 68 (1946) pp. 417–448.

13. ———, *The kernel function of an orthonormal system*, Duke Math. J. vol. 13 (1946) pp. 529–540.

14. M. Schiffer and D. C. Spencer, *The coefficient problem for multiply-connected domains*, Ann. of Math. vol. 52 (1950) pp. 362–402.

15. ———, *Lectures on conformal mapping and extremal methods*, Princeton Lectures, 1949–1950.

Hebrew University,
Jerusalem, Israel.

ÜBER TSCHEBYSCHEFFSCHE PROBLEME

H. GRUNSKY

Unter einem Tschebyscheffschen Problem verstehen wir einen bestimmten Typus von Extremalproblemen. Zu einem Extremalproblem gehört eine Klasse $\mathfrak{K}$ von Funktionen $F(z)$, jedem $F(z)$ ist eine reelle Zahl $A(F)$ zugeordnet; gesucht wird eine Funktion $F_0(z) \in \mathfrak{K}$ mit $A(F_0) = \min_{\mathfrak{K}}$. Wir denken hierbei an analytische Funktionen einer komplexen Veränderlichen, obwohl die Problemstellung natürlich auch in andern Fällen möglich ist. Von einem Tschebyscheffschen Problem reden wir, wenn $A(F)$ das Maximum des Betrages von F (oder auch einer mit F in gegebener Weise zusammenhängenden Funktion) auf einer gegebenen abgeschlossenen[1] Punktmenge $\mathfrak{M}$ ist: $A(F) = U(|F|, \mathfrak{M})$. Der einfachste Fall ist der der zu einem $\mathfrak{M}$ gehörigen Tschebyscheffschen Polynome (im folgenden kurz T-Polynome genannt). $\mathfrak{K}$ ist hier die Klasse aller normierten Polynome n-ten Grades: $P_n(z) = z^n + \sum_{\nu=0}^{n-1} a_\nu z^\nu$. Es soll nun im folgenden eine Methode zur Behandlung solcher Probleme entwickelt werden, die eine naturgemässe Verallgemeinerung der geläufigen Methode für gewöhnliche Maximum- und Minimumprobleme bei endlich vielen Veränderlichen darstellt. Wir stellen sie zunächst am Beispiel der T-Polynome dar und wenden sie dann auf ein in neuerer Zeit behandeltes Extremalproblem an.

I. Wir setzen Existenz und Einzigkeit des T-Polynoms $P_n^{(0)}(z)$ als bekannt voraus.[2] Mit $\mathfrak{M}_1$ bezeichnen wir die abgeschlossene Teilmenge von $\mathfrak{M}$, auf der $P_n^{(0)}(z)$ das Maximum des Betrages annimmt.

HILFSSATZ 1. *Das zu $\mathfrak{M}_1$ gehörige T-Polynom ist identisch mit dem zu $\mathfrak{M}$ gehörigen.*[3]

BEWEIS. Bei Annahme des Gegenteils folgt aus der Einzigkeit des zu $\mathfrak{M}_1$ gehörigen T-Polynoms $P_n^{(1)}(z)$: $|P_n^{(1)}(z)| < |P_n^{(0)}(z)|$ für $z \in \mathfrak{M}_1$ und daraus schliesst man leicht, dass bei hinreichend kleinem τ für

$$P_n^{(\tau)}(z) = \tau P_n^{(1)}(z) + (1 - \tau) P_n^{(0)}(z) \; (0 \leqq \tau \leqq 1) \; |P_n^{(\tau)}(z)| < P_n^{(0)}(z)|$$

überall auf $\mathfrak{M}$ gelten würde.

Jedes Polynom aus $\mathfrak{K}$ ist bestimmt durch seine Nullstellen, $\zeta_1, \cdots, \zeta_n$, oder durch irgend welche Funktionen $\omega_\nu(\zeta_1, \cdots, \zeta_n)$, $\nu = 1, \cdots, n$, durch die sich die ζ_ν wieder eindeutig ausdrücken lassen; wir setzen die ω_ν als analytisch voraus. Ihre Real- und Imaginärteile bezeichnen wir mit $\tau_1, \cdots, \tau_{2n}$; ein oberer Index 0 an einer dieser Grössen soll den speziell zu $P_n^{(0)}(z)$ gehörigen Wert bezeichnen.

[1] Die Problemstellung ist natürlich mit den erforderlichen Abänderungen auch für nicht abgeschlossenes $\mathfrak{M}$ möglich, doch legen wir hier keinen Wert auf jede mögliche Verallgemeinerung.

[2] J. C. de la Vallée Poussin, Bulletin de l'Académie Royale de Belgique (1911) pp. 199–211.

[3] J. C. de la Vallée Poussin, a. a. O.

Wir schreiben gelegentlich ausführlicher: $P_n(z;\omega)$ statt $P_n(z)$, wo ω die Menge der ω_ν repräsentiert. Ferner setzen wir

$$l_n(z;\omega) = \log|P_n(z;\omega)| .$$

Aus Hilfssatz 1 folgt nun unmittelbar:

HILFSSATZ 2. *Bei willkürlicher Wahl von* $d\tau_1, \cdots, d\tau_n$ *gibt es stets mindestens ein* $z \in \mathfrak{M}_1$, *derart, dass*

$$\sum_{\nu=1}^{2n} \frac{\partial l_n(z;\omega)}{\partial \tau_\nu} d\tau_\nu \geqq 0.$$

Indem wir für den aus den partiellen Ableitungen von l_n gebildeten Vektor die Bezeichnung einführen:

$$\eta(z) = \left(\frac{\partial l_n(z;\omega)}{\partial \tau_\nu}\right) \qquad \nu = 1, \cdots, 2n \tag{1}$$

können wir das vorige Resultat auch so aussprechen: zu beliebigem $\zeta = (d\tau_\nu)$, $\nu = 1, \cdots, 2n$, gibt es unter den Vektoren $\eta(z)$ mit $z \in \mathfrak{M}_1$ mindestens einen mit

$$\eta\zeta \geqq 0. \tag{2}$$

Daraus folgert man weiter: Ist $\mathfrak{P}$ die konvexe Hülle aller $\eta(z)$ mit $z \in \mathfrak{M}_1$, so enthält $\mathfrak{P}$ einen linearen Raum $\mathfrak{R}_m$ positiver Dimensionszahl m: $1 \leqq m \leqq 2n$. Unter den $\eta(z) \subset \mathfrak{R}_m$ gibt es dann ein System η_κ, $\kappa = 1, \cdots, k \leqq m+1 \leqq 2n+1$, derart, dass gilt

$$\sum_{\kappa=1}^{k} \lambda_\kappa \eta_\kappa = 0 \quad \text{mit} \quad \lambda_\kappa \geqq 0, \qquad \sum_{\kappa=1}^{k} \lambda_\kappa > 0.$$

Wir gehen nun auf die Bedeutung der $\eta_\kappa = \eta(z_\kappa)$ zurück, vgl. (1). Mit der Abkürzung

$$\mathfrak{h}_n = \sum_{\kappa=1}^{k} \lambda_\kappa l_n(z_\kappa;\omega)$$

erhalten wir

$$\left.\frac{\partial \mathfrak{h}_n}{\partial \tau_\nu}\right|_{\omega=\omega_0} = 0 \qquad \nu = 1, \cdots, 2n. \tag{3}$$

Wählen wir zunächst $\omega = \zeta$, so ist

$$\mathfrak{h}_n = R_n \mathfrak{B}_n \quad \text{mit} \quad \mathfrak{B}_n = \sum_{\nu=1}^{n} \sum_{\kappa=1}^{k} \lambda_\kappa \log(z_\kappa - \zeta_\nu),$$

und (3) bedeutet

$$R(\zeta) = \sum_{\kappa=1}^{k} \frac{\lambda_\kappa}{\zeta - z_\kappa} = 0 \qquad \text{für } \zeta = \zeta_1^{(0)}, \cdots, \zeta_n^{(0)}.$$

Das heisst also: Die Nullstellen des T-Polynomes $P_n^{(0)}(z)$ sind zugleich Nullstellen einer rationalen Funktion, die an $k \leqq 2n+1$ Stellen $z_\kappa \in \mathfrak{M}_1$ Pole erster Ordnung mit positiven Residuen hat. Es gilt aber auch die entsprechende Aussage über die Vielfachheit der Nullstellen: Jede q-fache Nullstelle von $P_n^{(0)}(z)$ ist mindestens q-fache Nullstelle von $R(\zeta)$. Ohne Beschränkung der Allgemeinheit nehmen wir für den Beweis[4] an, $z = 0$ sei eine q-fache Nullstelle von $P_n^{(0)}(z)$, d.h., $\zeta_1^{(0)} = \cdots = \zeta_q^{(0)} = 0$. Wir entwickeln den zu einem ζ_ν gehörigen Summanden von $\mathfrak{B}_n$ in eine Potenzreihe nach ζ_ν:

$$\mathfrak{F}(\zeta_\nu) = \sum_{\kappa=1}^{k} \lambda_\kappa \log(z_\kappa - \zeta_\nu) = \alpha_0 + \alpha_1 \zeta_\nu + \cdots + \alpha_q \zeta_\nu^q + \cdots .$$

Wählen wir nun

$$\omega_\nu = \sum_{\kappa=1}^{q} \zeta_\kappa^\nu ,$$

so ist

$$\mathfrak{B}_n = \alpha_0 + \alpha_1 \omega_1 + \cdots + \alpha_q \omega_q + \cdots$$

und (3) bedeutet

$$\frac{\partial \mathfrak{B}_n}{\partial \omega_\nu} = \alpha_\nu = 0, \qquad \nu = 1, \cdots, q,$$

also das Verschwinden der q ersten Ableitungen von $\mathfrak{F}(\zeta)$ bei $\zeta = 0$; dieser Punkt ist also q-fache Nullstelle von $\mathfrak{F}'(\zeta) = R(\zeta)$. Wir haben also den folgenden Satz.

SATZ. *Die Nullstellen des T-Polynomes $P_n^{(0)}(z)$ zu einer gegebenen Menge $\mathfrak{M}$ sind zugleich Nullstellen mindestens gleicher Vielfachheit einer rationalen Funktion $R(\zeta)$, die an $k \leqq 2n+1$ Stellen z_κ, in denen $P_n^{(0)}(z)$ das Maximum des Betrages auf $\mathfrak{M}$ annimmt, Pole erster Ordnung mit positivem Residuum hat.*

Da $R(\zeta)$ genau $k-1$ Nullstellen hat, so folgt:

$$n+1 \leqq k \leqq 2n+1.$$

Die erste dieser Ungleichungen besagt (was bekannt ist), dass $P_n^{(0)}(z)$ das Maximum seines Betrages auf $\mathfrak{M}$ in mindestens $n+1$ verschiedenen Stellen annimmt, die zweite stimmt mit der von de la Vallée Poussin bewiesenen Tatsache überein, dass $P_n^{(0)}(z)$ auch das T-Polynom einer Teilmenge von $\mathfrak{M}_1$ mit höchstens $2n+1$ Punkten ist; doch folgt das noch nicht aus Obigem. Unser Satz ergibt weiter, aufgrund einer bekannten Verallgemeinerung des Gauss-Lucasschen Satzes die ebenfalls bekannte Tatsache, dass die Nullstellen von $P_n^{(0)}(z)$ der konvexen Hülle von $\mathfrak{M}$ angehören.

II. Wir wenden diese Methode an auf ein von J. L. Walsh gestelltes und von M. Heins bearbeitetes Problem,[5] das wir zunächst allgemeiner formulieren: $\mathfrak{M}$ sei eine abgeschlossene Punktmenge, z_0 ein beliebiger Punkt aus $|z| < 1$,

[4] Vgl. P. R. Garabedian, Trans. Amer. Math. Soc. vol. 67 (1949) pp. 27-28.

[5] M. Heins, Trans. Amer. Math. Soc. vol. 55 (1944) pp. 349-372.

$z_0 \notin \mathfrak{M}$. Ohne Beschränkung der Allgemeinheit setzen wir voraus: $z_0 = \rho > 0$. Ferner sei $\mu > 0$ gegeben. Wir betrachten Funktionen mit folgenden Eigenschaften:

(1a) $F(z)$ sei regulär in $|z| < 1$, und es sei daselbst $|F(z)| \leqq 1$.

(1b) $F(z)$ sei ausserdem rational mit $|F(z)| = 1$ auf $|z| = 1$.

(1c) $F(z)$ sei ausserdem höchstens vom Grade n.

(2) $|F(z_0)| \geqq \mu$.

$\mathfrak{K}'$ sei die durch (1a) und (2), $\mathfrak{K}_\infty$ die durch (1b) und (2), $\mathfrak{K}_n$ die durch (1c) und (2) bestimmte Klasse. Es ist $\mathfrak{K}_n \subset \mathfrak{K}_\infty \subset \mathfrak{K}'$. $\mathfrak{K}$ sei irgend eine der drei Klassen. Mit $\mathfrak{PK}$ bezeichnen wir das Problem: Gesucht sind Funktionen $F(z) \in \mathfrak{K}$ mit

$$M(|F|, \mathfrak{M}) = \min_{\mathfrak{K}}.$$

Wir betrachten zunächst $\mathfrak{PK}_n$. Die Bezeichnungsweise ist der in I völlig analog. Zwei Fälle sind zu unterscheiden: (a) $|F_0(z_0)| > \mu$; (b) $|F_0(z_0)| = \mu$. Im Falle (a) verlaufen die Überlegungen ganz wie in I,[6] da jetzt die Bedingung (2) keinerlei Einschränkung für die zulässigen Variationen $\mathfrak{z}$ der τ_ν von $\tau_\nu^{(0)}$ aus bedeutet. Im Falle (b) dagegen kann nur die Zulässigkeit derjenigen Variationen behauptet werden, für die

$$\mathfrak{x} \cdot \mathfrak{z} > 0 \quad \text{mit} \quad \mathfrak{x} = \left(\frac{\partial l(z_0\,;\,\omega)}{\partial \tau_\nu}\right) \qquad \nu = 1, \cdots, 2n.$$

Andererseits bedeutet die Minimalforderung: Für alle zulässigen $\mathfrak{z}$ muss für mindestens ein $\mathfrak{y}(z)$, $z \in \mathfrak{M}_1$, (2) gelten. Für ein $\mathfrak{z}$, das alle Ungleichungen

$$\mathfrak{y}\mathfrak{z} < 0 \quad \text{mit} \quad \mathfrak{y} = \mathfrak{y}(z), \qquad z \in \mathfrak{M}_1,$$

erfüllt, muss also gelten

$$\mathfrak{x}\mathfrak{z} \leqq 0.$$

Daraus folgt, dass $\mathfrak{x}$ der konvexen Hülle $\mathfrak{Y}$ der $\mathfrak{y}(z)$, $z \in \mathfrak{M}_1$, angehören muss. Es gilt also eine Gleichung der Form:

$$\begin{gathered} \lambda_\xi - \sum_{\kappa=1}^{k} \lambda_\kappa \mathfrak{y}_\kappa = 0, \qquad \mathfrak{y}_\kappa = \mathfrak{y}(z_\kappa), \qquad z_\kappa \in \mathfrak{M} \\ \lambda = 1, \qquad \lambda_\kappa \geqq 0, \qquad \sum \lambda_\kappa > 0, \qquad k \leqq 2n + 1. \end{gathered} \tag{4}$$

Eine ebensolche Gleichung mit $\lambda = 0$ gilt in dem Fall (a). Setzen wir nun

$$\mathfrak{b} = \lambda l(z_0\,;\,\omega) - \sum_{\kappa=1}^{k} \lambda_\kappa l(z_\kappa\,;\,\omega)$$

so besagt (4):

$$\left.\frac{\partial \mathfrak{b}}{\partial \tau_\nu}\right|_{\omega=\omega^{(0)}} = 0, \qquad \nu = 1, \cdots, 2n.$$

[6] Abgesehen von dem Beweis für Hilfssatz 1.

Bezeichnen wir ferner mit $g(z;\zeta)$ die Greensche Funktion für $|z|<1$ mit Singularität ζ, mit $G(z;\zeta)$ eine analytische Funktion mit $\mathrm{Rn}G=g$, so ist

$$\log|F(z)| = l(z;\omega) = -\sum_{\nu=1}^{n} g(z;\zeta_\nu) = -\mathrm{Rn}\sum_{\nu=1}^{n} G(z;\zeta_\nu).$$

Setzen wir endlich

$$\mathfrak{G}(\zeta) = -\lambda G(\zeta;z_0) + \sum_{\kappa=1}^{k}\lambda_\kappa G(\zeta;z_\kappa),$$

so ergibt sich

$$\mathfrak{G}'(\zeta_\nu^{(0)}) = 0, \qquad \nu=1,\cdots,n. \tag{5}$$

$\mathfrak{G}'(\zeta)$ ist rational mit den einfachen Polen $z_0, z_1, \cdots, z_k$ mit bzw. den Residuen $\lambda, -\lambda_1, \cdots, -\lambda_k$. Ferner ist $\zeta\mathfrak{G}'(\zeta) = d\mathfrak{G}(\zeta)\,|\,d\log\zeta$ reell auf $|\zeta|=1$. Ebenso wie in I folgt die entsprechende Feststellung über die Vielfachheit der Nullstellen. Daraus schliesst man, da die Anzahl der Nullstellen von $\mathfrak{G}'$ in $|\zeta|<1$ höchstens k ist:

$$k \geqq n. \tag{6}$$

Betrachten wir nun $\mathfrak{PR}_\infty$, so muss die Extremalfunktion $F_0(z)$, wenn sie existiert (was im Gegensatz zu $\mathfrak{PR}_n$ und zu $\mathfrak{PR}'$ nicht von vornherein feststeht) und vom Grade n ist, denselben Bedingungen genügen wie bisher, ausserdem jedoch noch einer weiteren, die sich aus der Tatsache ergibt, dass es in $\mathfrak{R}_\infty$ zu $F_0(z)$ benachbarte Funktionen höheren Grades gibt. Es gilt nämlich $F(z)\to F_0(z)$, wenn die Nullstellen von $F_0(z)$ (mit Berücksichtigung ihrer Vielfachheiten) Häufungsstellen von Nullstellen von $F(z)$ sind und wenn die übrigen Nullstellen von $F(z)$ gegen $|z|=1$ konvergieren. Nehmen wir also an, $F(z)$ habe ausser den Nullstellen $\zeta_1,\cdots,\zeta_n$ noch eine weitere ζ_0, die radial gegen $\zeta_0^{(0)}$ mit $|\zeta_0^{(0)}|=1$ konvergiere, während die $\zeta_\nu\to\zeta_\nu^{(0)}$ $(\nu=1,\cdots,n)$ streben. Dann kommt zu den bisherigen Parametern $\tau_1,\cdots,\tau_{2n}$ noch ein weiterer τ_0 hinzu, und wir können etwa $\tau_0 = 1-|\zeta_0|$ wählen. Zu der hinreichenden Bedingung $\mathfrak{x}_\mathfrak{z}>0$ für Zulässigkeit eines $\mathfrak{z}$ tritt also mit $\mathfrak{w}=(1,0,\cdots,0)$: $\mathfrak{w}_\mathfrak{z}\geqq 0$ und entweder $\mathfrak{x}$ oder $\mathfrak{w}$ muss zu $\mathfrak{H}$ gehören, sicher also eine gewisse lineare Kombination $\lambda_\mathfrak{x}+\alpha\mathfrak{w}$ mit $\lambda\geqq 0, \alpha\geqq\nu$, $\alpha+\lambda>0$, in der, wie man leicht überlegt $\lambda=1$ gewählt werden kann. D.h. es gilt eine Gleichung der Form

$$\lambda_\mathfrak{x} - \sum_{\kappa=1}^{k}\lambda_\kappa\mathfrak{y}_\kappa = -\alpha\mathfrak{w}.$$

Die Interpretation für $\mathfrak{G}'(\zeta)$ ist: $\zeta\mathfrak{G}'(\zeta)\geqq 0$, $|\zeta|=1$.

Betreffend $\mathfrak{PR}'$ bemerken wir zunächst, dass die Einzigkeit der Extremalfunktion leicht nachzuweisen ist. Ist nun $F_0(z)\notin\mathfrak{R}_\infty$, so gibt es eine Folge von Funktionen $F_\nu(z)\in\mathfrak{R}_\nu$ wachsenden Grades, die gegen $F_0(z)$ konvergiert. Ersetzen wir jedes $F_\nu(z)$ durch die Extremalfunktion $F_{\nu 0}(z)$ in $\mathfrak{R}_\nu$ so konvergiert die Folge $F_{\nu 0}(z)$ wieder gegen $F_0(z)$. Daher muss auch der Grad der $F_{\nu 0}$ monoton wachsen und müssen fast alle Nullstellen der $F_{\nu 0}$ ausserhalb eines beliebig vorgegebenen

abgeschlossenen Teilbereiches von $|z| < 1$ liegen. Unter gewissen, ziemlich stark einschränkenden, Voraussetzungen lässt sich nun zeigen, dass das nicht möglich ist, dass vielmehr alle Nullstellen von $\mathfrak{G}'(\zeta)$ bis auf höchstens eine innerhalb eines nur von $\mathfrak{M}$ abhängenden abgeschlossenen Teilbereiches von $|\zeta| < 1$ liegen. In einem derartigen Fall muss also die Extremalfunktion von $\mathfrak{PR}'$ zu $\mathfrak{R}_\infty$ gehören.

Wir specialisieren nun das Problem auf das von Walsh und Heins: $\mathfrak{M}$ sei der Kreis $|z| = r < \rho$. In diesem Falle ist für $\mathfrak{PR}'$, $F_0(z) \in \mathfrak{R}_\infty$. Allgemein gilt, dass die Anzahl der Maxima einer Funktion aus $\mathfrak{R}_\infty$ auf $|z| = r$ höchstens gleich ihrem Grad ist, [7] und insbesonder ist also $k \leqq n$, also wegen (6) $k = n$.

Es fragt sich, wie n von den Gegebenheiten des Problems abhängt. Falls $\mu = \rho^p$ mit ganzem nicht negativem p so ist die Lösung von $\mathfrak{PR}'$ angebbar: $F_0(z) = z^p$. Setzen wir allgemein $\mu = \rho^q$ so kann sich für $p < q < p + 1$ n nicht ändern, da dies für k gilt. Für hinreichend kleines r lässt sich ferner die vorhin erwähnte Feststellung betreffend die Nullstellen von $\mathfrak{G}'(\zeta)$ machen, wodurch gezeigt ist, dass, wenn q eine ganze Zahl überschreitet, sich n höchstens um 1 ändern kann. Damit ist $n = [q] + 1$ für nicht ganzes q bewiesen. Die Allgemeingültigkeit dieser Aussage, unabhängig von r, folgt nun, indem wir r bei festem q wachsen lassen; wiederum kann sich k und damit n nicht ändern.

Eine ähnliche Behandlung lässt der ebenfalls von Heins untersuchte Fall zu: $\mathfrak{M}$ sei durch $-r \leqq z \leqq r < \rho$ erklärt. Ist dann $z_1 < \cdots < z_k$, so ist zunächst leicht zu sehen: $z_k = r$ und ferner ist nach (5) klar, dass zwischen je zwei $z_\varkappa$ ein ζ_ν liegen muss, ausserdem eines links von z_1, das auch mit -1 zusammenfallen kann. Also ist entweder $n = k$ oder $n = k - 1$. Wenn μ von 1 aus abnimmt, so muss n wachsen. Eine Änderung von n ist nur bei entsprechender Änderung von k möglich, und diese wiederum kann nur dadurch eintreten, dass in $-r$ für gewisse Werte von μ ein Maximum auftaucht (ein Randmaximum). Gerade in diesem Falle muss -1 Nullstelle von $\mathfrak{G}'(\zeta)$ sein, bei weiterer Abnahme von μ wird $z_1 > -r$, $\zeta_1 > -1$, sodass also nun n um 1 zugenommen hat.

UNIVERSITY OF TÜBINGEN,
TÜBINGEN, GERMANY.

[7] M. Heins, a. a. O.

SURFACES DE RIEMANN OUVERTES

ROLF NEVANLINNA

1. La conception fondamentale de la théorie géométrique de la théorie des fonctions monogènes consiste dans la notion d'une surface de Riemann. Dans le développement de cette théorie on peut distinguer deux idées directrices. Il s'agit d'une part d'une étude des surfaces mêmes au point de vue de leurs propriétés topologiques et conformes invariantes. D'autre part, l'on cherche de déterminer et d'étudier les classes les plus caractéristiques des fonctions analytiques ou harmoniques définies sur une surface donnée, classes simples au point de vue de leurs propriétés d'uniformité et de la nature des singularités des fonctions correspondantes. Dans la théorie des surfaces compactes les résultats classiques de Riemann concernant les classes topologiques et les classes conformes des surfaces closes constituent le fondament pour les grands problèmes de la représentation conforme et de l'uniformisation, qui occupaient les mathématiciens jusqu'au commencement de notre siècle. La solution définitive de ces questions d'existence fut donnée par Poincaré, Koebe, et Carathéodory dans une suite de travaux classiques, précedés par les recherches fondamentales de Schwarz, Neumann, et Klein.

Quant au problème concernant les fonctions qui existent sur une surface donnée, la théorie classique des intégrales abéliennes en donne une solution complète pour le cas d'une surface close. Ces fonctions jouissent des properiétés particulièrement simples: elles sont intégrales des différentielles *uniformes* et ne présentent qu'un nombre fini de pôles commes points singuliers.

2. Pour les *surfaces ouvertes* on a obtenu des résultats assez généraux dans le cas le plus simple où la surface donnée est de genre zéro et, par conséquent, en vertu du théorème de Riemann, conformément équivalente soit au plan complexe ponctué (cas parabolique), soit à l'intérieur d'un cercle fini (cas hyperbolique). Les recherches classiques de Weierstrass, Poincaré, Hadamard, et Borel constituent la base à la théorie moderne des singularités d'une fonction analytique uniforme dans une region connexe du plan complexe, théorie développée surtout par l'école française et scandinave (Wiman, Lindelöf, Fatou, Denjoy, Valiron, Julia, Milloux, Henri Cartan, Carleman, Ahlfors, Beurling, et autres).

3. Les propriétés conformes des surfaces ouvertes quelconques et surtout des surfaces de *genre infini* sont jusqu'ici très peu etudiées. Dans ce qui suit je donnerai un petit résumé de certains résultats généraux obtenus pendant les dernières dix années. Si je parle des surfaces ouvertes, les surfaces fermées ne seront pas exclues: en effet, en enlevant un point on transforme une surface close dans une surface ouverte.

4. Un problème général se pose ici. Est-il possible de déterminer des classes de surface ouvertes et des fonctions correspondantes pour lesquelles se conservent les théorèmes, fondamentaux de la théorie classique des intégrale abéliennes, de sorte que la théorie généralisée contienne la théorie classique comme un cas spécial si la surface donnée est fermée?

On se borne au premier lieu aux intégrales de première espèce. Soit R une surface de Riemann quelconque et soit $d\Phi = \phi(z)\,dz$ une différentielle analytique, holomorphe et uniforme sur R et définie d'une manière invariante par rapport à un changement du paramètre local $z = x + iy$. Une telle différentielle soit dite *abélienne de première espèce*. L'intégrale Φ est holomorphe, mais multiforme en général. Elle présente un nombre de périodes, correspondant soit aux cycles non homologues à zéro, soit aux cycles homologues à zéro mais dans un sens non compact, les chemins correspondantes divisant la surface en deux portions non compactes.

5. Les différentielles de première espèce forment une variété linéaire A. On doit à Behnke et Stein[1] quelques résultats très généraux concernant la classe A. Ils ont montré qu'il existe sur une surface R ouverte quelconque des intégrales A admettant des périodes donnée d'avance d'une manière arbitraire. En particulier, il existe toujours une intégrale *uniforme* et *holomorphe* sur R.

6. Or, si l'on cherche à développer une théorie des intégrales A pour lesquelles resteront valables les *théorèmes d'unicité* qui jouent un rôle si important dans la théorie classique, la classe A est évidemment trop étendue, et il faut la restreindre, par des conditions supplémentaires qui limitent d'une manière convenable la croissance des covariants $\phi(z)$ au voisinage de la frontière idéale de la surface R.

Pour quelques classes particulières de surfaces transcendantes (en premier lieu pour certaines classes de surfaces à un nombre fini de feuillets), Myrberg et ces élèves ont généralisé la théorie classique des intégrales des fonctions algébriques. Dans ce qui suit nous ne pouvons pas insister à ces résultats intéressantes; nous considérons de problème défini ci-dessus dans une forme plus générale, en admettant que les surfaces données soient définies comme espaces topologiques et conformes abstraites.

7. Une sous-classe (classe D) remarquable de la variété A définie ci-dessus est formée par les différentielles $d\Phi = \phi\,dz$ pour lesquelles l'intégrale

$$\text{(Classe } D\text{)} \qquad \iint_R |\phi^2|\,dxdy$$

est *finie*. Pour une surface fermée cette classe se confonds avec l'ensemble de différentielles abéliennes de première espèce.[2]

[1] On trouvera une bibliographie dans une monographie sur la théorie de l'uniformisation qui paraîtra bientôt.

[2] La classe D est un cas particulier ($\rho = 1$) de la classe D_ρ plus générale pour laquelle

Dans cette classe $D \subset A$, encore assez générale, on a considéré deux sous-classes importantes.

1°. Classe $D_1 \subset D$, formée par les différentielles $d\Phi = dU + idU'$ pour lesquelles dU est *exacte* (c'est-a-dire: U est uniforme et les périodes de l'intégrale Φ sont purement imaginaires).

2°. Classe $D_2 \subset D_1 \subset D$ pour laquelle $d\Phi$ est exacte (Φ uniforme).

8. Indiquons encore deux autres variétés, sous-classes de la variété A, qui ont été l'objet de nombreuses recherches dans le dernier temps.

3°. Classe BH, composée par les différentielles $d\Phi$ exactes de première espèce telles que le potentiel U est borné en valeur absolue

(Classe BH) $$|U| < M < \infty.$$

4°. Classe $BA \subset BH$, dont les différentielles $d\Phi$ sont exactes de sorte que

(Classe BA) $$|\Phi| < M < \infty.$$

9. Parmi les propriétés conformes invariantes d'une surface la suivante joue un rôle important.

Soit R une surface quelconque et E une classe des différentielles $d\Phi$ quelconques définies par certaines propriétés conformément invariantes et contenant l'élément nul $d\Phi \equiv 0$.

Nous disons qu'une surface donnée R appartient à *la classe* $R_0(E)$ si la classe E ne contient que l'élément $d\Phi \equiv 0$.

Autrement la surface R soit de *la classe* $R_1(E)$; sur une surface $R \subset R_1(E)$ il existe donc au moins une différentielle $d\Phi \not\equiv 0$ de la classe E.

10. Pour illustrer cette notion considérons d'abord le cas le plus simple où R est une surface fermée ponctuée dans un point P. Il est évident que ce point singulier est enlevable pour les classes D_1 et B. Chaque intégrale correspondante Φ se réduit donc à une constante, en vertu du principe de maximum, et l'on conclut que $R \subset R_0(D_1)$, $R \subset R_0(B)$. Ceci reste vrai si l'on remplace D_1 et B par les sous-classes D_2 et B_1.

Il en est de même pour la classe D si le genre p de la surface R est zéro. Mais pour $p > 0$ il n'en est plus ainsi: en effet, la classe D se confond la variété linéaire de dimension p des différentielles abéliennes de première espèce.

11. Revenons avec surfaces ouvertes quelconques. Pour mesurer la frontière idéale d'une surface on a introduit, au délà des classifications dont je viens de parler, la notion de *capacité* ou de mesure harmonique, généralisant la notion connue et classique de Wiener. Étant donnée une surface R, faisons une exhaustion de R à l'aide d'une suite

---

l'intégrale $\iint_R |\phi|^2 \rho dx\, dy$ reste finie, ρ étant une scalaire positive, uniforme sur R. Nous reviendrons dans un autre article à cette variété.

$$R_0 \subset R_1 \subset \cdots \subset R_n \subset \cdots; \qquad R_n \to R$$

des domaines compactes, R_n étant limité par un ensemble compacte Γ_n, constitué d'un nombre fini de courbes fermées Γ_n^2 :

$$\Gamma_n = \sum_i \Gamma_n^i.$$

Soit ω_n la fonction harmonique dans $R_n - R_0$, prenant la valeur 0 sur Γ_0 et la valeur 1 sur Γ_n (la mesure harmonique de Γ_n par rapport à $R_n - R_0$). Pour $n \to \infty$ cette fonction tend vers une limite ω. Si $\omega \equiv 0$, on dit que la mesure harmonique ou la capacité de la frontière idéale Γ de R est nulle:

(Classe C_0) $\qquad$ mes harm $\Gamma = 0$.

Si la limite harmonique $\omega > 0$ à l'intérieur de $R - R_0$, la mesure harmonique (et la capacité) de Γ est dite positive (classe C_1 de surfaces R).

12. On doit à Myrberg le résultat qu'une condition nécessaire et suffisante pour que mes harm Γ soit positive est que la surface R admet une fonction de Green. Voici une autre propriété qui rapproche le type d'une surface, mesuré par la capacité, aux classifications considérées plus haut.

Une surface R du type mes harm $\Gamma = 0$ (classe C_0) appartient toujours à la classe $R_0(D_1) \subset R_0(D_2)$. De même, une surface de la classe C_0 est aussi de la classe $R_0(BH) \subset R_0(BA)$, de sorte qu'on a

$$(1) \qquad C_0 \subset \begin{cases} R_0(D_1) \subset R_0(D_2) \\ R_0(BH) \subset R_0(BA). \end{cases}$$

13. Est-ce que ces conclusions restent vraies dans le sens opposé? Il est facile de voir qu'il n'en est pas ainsi pour les deux dernières relations: les classes $R_0(D_1)$ et $R_0(BH)$ sont des sous-classes *effectives* des classes $R_0(D_2)$ et $R_0(BA)$. Une question plus difficile est celle concernant l'étendue des trois premières classes dans le tableaux (1).

Pour des classes étendues des surfaces R, par exemple pour les surfaces de genre fini, mais d'un ordre de connexion quelconque (fini ou infini), les conditions C_0, D_1, et BH sont *équivalentes*. Or, Ahlfors a récemment indiqué un exemple d'une surface R de genre infini qui montre que la classe C_0 est une sous-classe effective de la classe BH.

14. Les définitions considérée ci-dessus sont d'un caractère assez implicite, car il reposent sur l'existence ou non existence de certaines fonctions sur la surface, définies à l'aide des propriétés dont la valabilité n'est pas simple à vérifier pour une surface de Riemann, donnée d'une manière particulière (comme par exemple une sous-region ou une surface de récouvrement d'une autre surface). C'est pourquoi on a essayé de mettre ces conditions en relation aux propriétés structurelles qui entrent d'une façon plus directe dans la définition

d'une surface. On connaît un grand nombre de telles critères particulières, surtout concernant le "problème de type" (les classes C_0 et C_1).

15. Je ne peux pas insister à ces recherches[3] et je me borne à rappeler une méthode générale qui a conduit aux résultats remarquables. Supposons qu'on ait défini sur la surface R une métrique conforme

$$d\sigma = \lambda(z) \mid dz \mid$$

de sorte que le domaine R_ρ formé par les points ayant une distance au plus égale à ρ d'un point donné z_0 est compact et limité par un ensemble C_ρ, composé d'un nombre fini de courbes fermées C_ρ^i :

$$C_\rho = \sum_i C_\rho^i ,$$

les longueurs correspondantes étant

$$L(\rho) = \sum_i L^i(\rho).$$

On suppose que $C_\rho \to \Gamma$ pour $\rho \to \infty$. Cela étant, un théorème de Ahlfors et Laasonen montre que la divergence de l'intégrale

$$\int^\infty \frac{d\rho}{L(\rho)}$$

est suffisante pour que la surface R soit de la classe C_0.

Remplaçons dans cette intégrale $L(\rho)$ par le nombre $1(\rho) \leqq L(\rho)$, défini comme le maximum des longueurs L_ρ^i des courbes C_ρ^i qui constituent l'ensemble C_ρ. Si cette intégrale

$$\int^\infty \frac{d\rho}{1(\rho)}$$

diverge, on peut affirmer que la surface R est de la classe $R_0(DA)$; dans ce cas donc il n'existe aucune fonction Φ non constant analytique et *uniforme* sur R et telle que l'intégrale (D) soit finie. Ce résultat, dû à Sario, s'obtient de l'inégalité

$$\iint_{R_\rho} |\phi|^2 \, dx\, dx \geqq \text{const.}\, e^{4\pi} \int_0^\infty \frac{d\rho}{1(\rho)} .$$

La valeur du facteur 4π est la meilleure possible. Sous cette forme précise l'inégalité a été donnée par Pfluger.

16. Les métriques conformes jouent aussi un rôle important pour diverses problèmes extrémales concernant les invariants conformes attachés à une surface. Quelques-uns de ces résultats, dûs à Ahlfors et Beurling, font l'objet de

[3] Dans un article de LeVan, paru dans les Comment. Math. Helv. (1948), on trouve une bibliographie assez complète concernant le problème de type.

certains autres conférences du congrès, et je n'insiste pas à ces questions. C'est pourquoi je laisse à coté aussi les extensions du problème classique de Painlevé, concernant la classe BA (résultats récents de Grunsky, Ahlfors, Beurling, Garabedian).

17. Passons enfin à la classe de différentielles (D), définies par cette propriété que l'intégrale de Dirichlet correspondante reste finie. Si l'on introduit le produit scalair et le norme

$$(\phi, \psi) = \iint_R \phi\bar{\psi}\, dx\, dy, \qquad \| \phi \|^2 = (\phi, \phi),$$

il est facile de voir, suivant les méthodes de Bergman et Bochner, que la variété D est un espace de Hilbert. On peut donc développer les différentielles ϕ dans une série de Fourier suivant un système $\phi^1, \cdots, \phi^n, \cdots$ orthonormale.

18. Dans cet espace D on peut introduire trois "axes" orthogonaux l'un à l'autre.

Le premier axe, le sous-espace S_1, est composé par les différentielles *exactes* D (les différentielles D_2). Les différentielles du second axe (sous-espace S_2) ont cette propriété que les périodes des intégrales correspondantes s'évanouissent sur les cycles homologues à zéro (dans un sens compact ou non compact). Le troisième axe S_3 est le complément orthogonal de S_1 et S_2 dans l'espace D.

Si la surface R appartient à la classe $R_0(D_2)$, le sous-espace S_1 s'évanouit. Il en est de même pour l'axe S_2, si $R \subset R_0(D_1) \subset R_0(D_2)$. Dans ce cas l'espace D se réduit à l'axe S_3 ; il en est ainsi pour toute surface de la classe C_0.

19. On doit ces résultats à Virtanen. Independemment de lui, Ahlfors a établi une théorie qui conduit à une partie du théorème ci-dessus. La méthode de Ahlfors consiste dans une approximation d'une surface ouverte R par une suite de surfaces compactes symétriques, et il obtient les différentielles sur R comme limites des différentielles de Schottky. Ce procédé généralise une méthode qui a conduit aux résultats importants dans la théorie des surfaces multiplement connexes de genre zéro (Groetsch, Schiffer, Lehto, Lokki, Garabedian, et autres).

Dans toutes ces recherches, la "kernelfunction" de Bergman joue un rôle remarquable.

UNIVERSITY OF ZÜRICH,
ZÜRICH, SWITZERLAND.

ON CERTAIN SET FUNCTIONS DEFINED BY EXTREMUM PROPERTIES IN THE THEORY OF FUNCTIONS AND IN MATHEMATICAL PHYSICS

G. SZEGÖ

This brief survey deals with certain set functions which are of interest for the theory of functions and for mathematical physics. Most of them occur in classical problems as we are going to point out later. Some of them have been investigated during the past six years or so by a group of mathematicians at Stanford University. My purpose is to name a few of the principal concepts, problems, and results of this investigation.

For the sake of general formulation, we denote the set function under consideration by $f(E)$ where the set E is arbitrarily chosen from a class K of sets. The nature of these sets may be very different. We can consider curves in the plane or surfaces in space, restricted by various conditions of smoothness, or by some kind of boundary conditions or by the condition that another given set function $g(E)$ has a preassigned value. We may consider arbitrary closed sets in plane or space. *The general problem is to find inequalities for the set function $f(E)$ valid for all sets of the given class K.* A more precise problem is to find the absolute maximum or the absolute minimum of the set function $f(E)$ when E varies on K. Finally we may deal with the question of relative extrema; in this case the concept of a distance or deviation of two sets within K must be fixed.

As perhaps the oldest example we mention the class K of all closed curves E in the plane with given area $g(E)$, the set function $f(E)$ being the length of E. Another important example is given by Hamilton's principle in analytical mechanics; in the most elementary case, K consists then of all curves E joining two given points and $f(E)$ is defined by Hamilton's integral. Other examples are the electrostatic capacity of a conductor E with given volume, the fundamental frequency of a membrane with fixed boundary and given area, and the fundamental frequency of a clamped plate of given area.

In these examples (except in Hamilton's principle) we characterize the class K by the condition that a certain set function $g(E)$ has a given value for all sets E. Such problems are called *isoperimetric problems* by extension of the usual meaning of the term which refers to the best known special case, namely to the classical isoperimetric problem concerning length and area of a plane curve.

With a plane domain E bounded by a simple closed curve γ we associate the following set functions:

the area A;

the perimeter L;

the inner (conformal) radius r_a with respect to an interior point a; this is the radius of the circle onto the interior of which the given domain can be mapped conformally, so that a is transformed into the center of the circle without change of scale at the point a;

the outer (conformal) radius $\bar{r}$ of the given curve; this is the radius of the circle onto the exterior of which the exterior of the given curve γ can be mapped conformally so that the point at infinity is transformed into itself without change of scale at this point;

the principal frequency Λ of a membrane with fixed boundary having the given shape;

the principal frequency Λ' of a clamped plate having the given shape,

the torsional rigidity P or stiffness of an infinite beam with the given plane domain as cross section; this is the resistance to a twisting torque divided by the angle of the twist and by a constant depending on the elastic nature of the beam.

With a space domain E bounded by a simple closed surface we associate the following set functions:

the volume V;

the surface area S;

the integral M of the mean curvature (Minkowski's constant); dealing with this constant we assume that the given domain is convex;

the electrostatic capacity C (with respect to the infinitely large sphere).

These are just a few examples of interesting set functions. Some of them are of geometrical nature in the sense that they can be computed by direct integration (for instance, the area, length, volume, etc.), some others depend on the solution of a boundary value problem. In the case of the electrostatic capacity this boundary value problem is a Dirichlet problem. A more complicated quantity is the virtual mass W of a solid immersed in a uniform flow. It is the energy of the disturbance of a uniform flow caused by the presence of the given solid and can be determined by solving a Neumann problem. The quantity W thus defined depends of course not only on the given solid but also on the given flow.

Our main interest is in such isoperimetric problems in which, using the notation introduced before, the set function $g(E)$ is a geometric quantity while $f(E)$ is a quantity depending on the solution of a boundary value problem.

Other interesting set functions depend on one or more points which play the role of parameters. The most familiar example is the Green function $G(p, q)$ of the domain E corresponding to the Laplace equation $\nabla^2 u = 0$ with the boundary condition $u = 0$ in plane or space, and the more complicated Green function $\Gamma(p, q)$ of the clamped plate satisfying the differential equation $\nabla^4 u = 0$ (which is of the fourth order) with the boundary conditions $u = \partial u/\partial n = 0$. The principal term of G at the source point q is $\log r$ where r is the distance of the points p and q; the principal term of Γ is $r^2 \log r$.

Now I mention a few set functions arising in the theory of functions whose definition is based on certain maximum-minimum problems. With a closed set E in the plane and with an arbitrary integer n we associate:

the polynomial $f(z)$ of degree n with the leading term z^n determined by the condition that the maximum modulus of $f(z)$ on E is a minimum. We denote this minimum by R_n^n; the set function $f(E) = f(E; n)$ under consideration is R_n;

the polynomial $f(z)$ of degree n whose roots z_i are all on E, determined by the condition that the discriminant $\prod_{i<k}(z_i - z_k)$ has a maximum modulus. We denote this maximum by $D_n^{C_{n,2}}$. A geometric definition of this quantity D_n is the following: It is the maximum of the geometric mean of the $C_{n,2}$ mutual distances of n arbitrary points z_i on E.

I leave aside a geometric formulation of the definition of R_n.

Finally let E be an arbitrary closed and rectifiable curve γ, and a an arbitrary point in the interior of γ. We consider the polynomial $f(z)$ of degree n that satisfies the normalizing condition

$$\int_\gamma |f(z)|^2 |dz| = 1 \tag{1}$$

and for which $|f(a)|^2$ is maximum. This maximum is a set function $f(E) = f(E; n, a)$ depending, besides on n, on the parameter a. The integral over the curve γ in the normalizing condition (1) can be replaced by an area integral extended over the interior of γ. In both cases appropriate weight functions can be introduced in these integrals.

These problems can be further generalized in various directions and they have analogues in space.

Let us formulate now a few particularly simple results and some open problems.

The ratio L^2/A for all closed curves in the plane, and the ratio S^3/V^2 for all closed surfaces in space, is a minimum for the circle and the sphere, respectively. These are the classical isoperimetric theorems. (Another formulation is: A given, L is a minimum for the circle.)

The ratios S/M^2 and MV/S^2 for arbitrary *convex* surfaces are maximal for the sphere. This result is due to Minkowski and represents an important addition to the isoperimetric theorems.

The ratios r_a^2/A and $A/\bar{r}^2$ are maximal for a circle. These are the area theorems in the theory of conformal mapping.

The ratio V/C^3 is maximal for the sphere. This statement was made by Poincaré [8]; the first indication of a satisfactory proof is due to Faber [4], the first complete proof to myself [16].

The ratio C/M is maximal for the sphere. This I have proved in 1931 [17].

Let C be the capacity of an arbitrary solid of revolution, $\bar{r}$ the outer radius of its meridian section. The ratio $C/\bar{r}$ is maximal for a circular disk.

The ratio P/A^2 is maximal for a circle. This was stated by de Saint-Venant [13]. The two latter theorems are proved in a joint investigation of Professor Pólya and myself which will be published in the Annals of Mathematics Studies [11]. I refer to this book concerning further literature on the subjects mentioned above.

Some of these maximum-minimum theorems can be obtained by the geometric process of symmetrization, some others by resorting to methods of the theory of functions. A very important factor is the interpretation of the quantities in question as maxima or minima of certain problems in the calculus of variations.

In some cases the relative extremum property of the circle or sphere can be shown by computing the first and second variations of the quantities involved.

Here are a few further results.

The quantity $\Lambda^2 A$ is minimal for a circle. This was conjectured by Lord Rayleigh in 1894 [12] and proved by Faber [4] and Krahn [7] in 1923.

The corresponding quantity $\Lambda'^2 A$ for the clamped plate is also minimal for a circle. This I proved recently under the hypothesis that the first eigenfunction of the clamped plate does not change its sign [19].

The last assertion would follow from another one which was formulated by Hadamard in 1908 [6]. Hadamard stated the conjecture that the Green function $\Gamma(p, q)$ of the differential equation $\nabla^4 u = 0$ with the boundary conditions $u = \partial u/\partial n = 0$ does not change its sign in the given domain. This Green function is certainly positive for $p = q$ as Hadamard himself has shown by using variational methods. A discussion of the limiting case of an infinite strip given lately by Duffin [3] strongly indicated that the answer to the question of Hadamard is negative in general. Indeed Loewner and myself constructed recently several examples of finite closed analytic curves for which the Green function becomes negative, i.e., it changes its sign. Finally Garabedian has shown by direct discussion that this is the case for a sufficiently flat ellipse; here Γ becomes negative if p and q are sufficiently near to the opposite end points of the major axis. Needless to say, the question of the first non-vanishing eigenfunction is *not* decided by these considerations. It would follow for instance from the positivity of *any* kernel arising from $\Gamma(p, q)$ by repeated iteration.

Finally, let us discuss in a few words the set function of the theory of functions introduced above.

The quantities R_n and D_n have been defined and studied by Professor Fekete for an arbitrary closed set [5]. Both sequences R_n and D_n tend for $n \to \infty$ to the so-called *transfinite diameter* of E which coincides with the outer radius $\bar{r}$ in the case the set E is a simple curve. The polynomials $f(z)$ having the minimal property which serves as the definition of R_n, generalize the classical Tchebychev polynomials (which arise in the case of a linear segment). For an analytic curve they were studied by Faber. A geometric approach to the quantities R_n and D_n, their generalizations, and also their analogue in space were studied by Professor Pólya and myself in 1931 [9].

The polynomial $f(z)$ of degree n maximizing $|f(a)|^2$ under the normalizing condition (1) was introduced by myself in 1921 [15]. It is closely related to the polynomials $\{q_n(z)\}$ orthonormal on the curve γ and can be represented in the form

$$f(z) = [K_n(a, a)]^{-1/2} K_n(a, z)$$

where

$$K_n(a, z) = \sum_{\nu=0}^{n} \overline{q_\nu(a)} q_\nu(z).$$

The maximum itself can be written in the form $K_n(a, a)$. As $n \to \infty$, and a and

z are arbitrary points in the interior of the curve γ, $K_n(a, z)$ tends to an important function $K(a, z)$ which is called the *kernel function*. It is a set function associated with a given curve and a given metric on it. The analogous function arising when the integral in the normalizing condition (1) is replaced by an area integral, was introduced by Bergman [1] and Bochner [2] about 1922. These concepts and results were the starting points of important recent investigations on kernel functions associated with other manifolds and other metrics by Bergman, Garabedian, Schiffer and Spencer.

Bibliography

1. S. Bergman, *Über die Entwicklung der harmonischen Funktionen der Ebene und des Raumes*, Math. Ann. vol. 86 (1922) pp. 238–279.

2. S. Bochner, *Über orthogonale Systeme analytischer Funktionen*, Math. Zeit. vol. 14 (1922) pp. 180–207.

3. R. J. Duffin, *On a question of Hadamard concerning super-biharmonic functions*, Journal of Mathematics and Physics vol. 27 (1949) pp. 253–258.

4. G. Faber, *Beweis, dass unter allen homogenen Membranen von gleicher Fläche und gleicher Spannung die kreisförmige den tiefsten Grundton gibt*, Sitzungsberichte der Bayrischen Akademie der Wissenschaften, 1923, pp. 169–172.

5. M. Fekete, *Über die Verteilung der Wurzeln bei gewissen algebraischen Gleichungen mit ganzzahligen Koeffizienten*, Math. Zeit. vol. 17 (1923) pp. 228–249.

6. J. Hadamard, *Mémoire sur le problème d'analyse rélatif à l'équilibre des plaques élastiques encastrées*, Mémoires Présentés par Divers Savants à l'Académie des Sciences (2) vol. 33 (1908) 128 pp.

7. E. Krahn, *Über eine von Rayleigh formulierte Minimaleigenschaft des Kreises*, Math. Ann. vol. 94 (1924) pp. 97–100.

8. H. Poincaré, *Figures d'équilibre d'une masse fluide*, Paris, 1903.

9. G. Pólya and G. Szegö, *Über den transfiniten Durchmesser (Kapazitätskonstante) von ebenen und räumlichen Punktmengen*, J. Reine Angew. Math. vol. 165 (1931) pp. 4–49.

10. ———, *Inequalities for the capacity of a condenser*, Amer. J. Math. vol. 67 (1945) pp. 1–32.

11. ———, *Isoperimetric inequalities in Mathematical Physics*, Annals of Mathematics Studies, in print.

12. Lord Rayleigh, *The theory of sound*, 2d ed., London, 1894–1896.

13. B. de Saint-Venant, *Mémoire sur la torsion des prismes*, Mémoires Présentés par Divers Savants à l'Académie des Sciences vol. 14 (1856) pp. 233–560.

14. M. Schiffer and G. Szegö, *Virtual mass and polarization*, Trans. Amer. Math. Soc. vol. 67 (1949) pp. 130–205.

15. G. Szegö, *Über orthogonale Polynome, die zu einer gegebenen Kurve der komplexen Ebene gehören*, Math. Zeit. vol. 9 (1921) pp. 218–270.

16. ———, *Über einige Extremalaufgaben der Potentialtheorie*, Math. Zeit. vol. 31 (1930) pp. 583–593.

17. ———, *Über einige neue Extremaleigenschaften der Kugel*, Math. Zeit. vol. 33 (1931) pp. 419–425.

18. ———, *On the capacity of a condenser*, Bull. Amer. Math. Soc. vol. 51 (1945) pp. 325–350.

19. ———, *On membranes and plates*, Proc. Nat. Acad. Sci. U. S. A. vol. 36 (1950) pp. 210–216.

Stanford University,
Stanford, Calif., U. S. A.

INTERNATIONAL CONGRESS

OF

MATHEMATICIANS

Cambridge, Massachusetts, U.S.A.
1950

CONFERENCE IN APPLIED MATHEMATICS

Committee

John von Neumann (*Chairman*)

Walter Bartky
R. V. Churchill
Richard Courant
G. C. Evans
William Prager
Mina Rees

The Conference in Applied Mathematics held three sessions, devoted respectively to the following general topics:

A. Random Processes in Physics and Communications;
B. Partial Differential Equations;
C. Statistical Mechanics.

The session on partial differential equations was devoted to problems in fluid dynamics, except for the lecture by W. Prager on the theory of plasticity.

RANDOM PROCESSES IN PHYSICS AND COMMUNICATIONS

SOME TOPICS IN INFORMATION THEORY

C. E. SHANNON

Previous work in communication theory [2] has shown that *amount* of information for purposes of communication has a natural measure in terms of entropy type formulas $H = -\Sigma p \log p$. This has led to theorems giving the most efficient encoding of the messages produced by a stochastic process into a standard form, say a random sequence of binary digits, and for the most efficient use of an available communication channel. However, no concept of information itself was defined. It is possible to formulate an approach to the theory in which the information sources in a communication network appear as elements of a lattice.

The leading idea is that any reversible translation of the messages produced by a stochastic process, say by a non-singular finite state transducer, should be regarded as containing the same information as the original messages. From the communications point of view, knowledge of the Morse code translation of the text originating at a telegraph office is equivalent to knowledge of the text itself. Thus we consider the information of a source to be the equivalence class of all reversible translations of the messages produced by the source. Each particular translation is a representative of the class, analogous to describing a tensor by giving its components in a particular coordinate system.

Various theories may be obtained depending on the set of translation operations allowed for equivalence. Two choices lead to interesting and applicable developments: (1) the group of all finite state transducers (allowing effectively positive or negative delays), (2) the group of *delay free* finite state transducers, in which it is required that the present output symbol be a function of the present and past history of the input, and similarly for the reverse transducer.

The first case is the simplest and relates most closely to previous work in which unlimited encoding delays at transmitter and receiver were allowed. A transitive inclusion relation between information elements, $x \geqq y$, (inducing a partial ordering) means that y can be obtained by operating on x with some finite state transducer (not necessarily reversible). The entropy of a source (which is invariant under the group of reversible transducers) appears as a norm monotone with the ordering. The least upper bound for two elements is the total information in both sources, a representation being the sequence of ordered pairs of letters from the two sources. A greatest lower bound can also be defined, thus resulting in an information lattice. There will always be a universal lower bound, and if the set of sources considered is finite, a universal upper bound. The lattices obtained in this way are, in general, non-modular. In fact, an information lattice can be constructed isomorphic to any finite partition lattice.

A metric can be defined by $\rho(x, y) = H_x(y) + H_y(x)$ satisfying the usual

requirements. This introduces a topology and the notion of Cauchy convergent sequences of information elements and of limit points. If convergent sequences are annexed to the lattice as new points, with corresponding modifications of the definition of equality, etc., there result continuous lattices, for example the set of all the abstractions of the total information in the system by finite state transducers, or limiting sequences of such transducers.

The delay free theory leads also to a lattice but the problems, while perhaps more important in the applications, are less well understood. The entropy of a source is no longer sufficient to characterize the source for purposes of encoding, and in fact an infinite number of independent invariants have been found. Certain of them are related to the problem of best prediction of the next symbol to be produced, knowing the entire past history. The delay free theory has an application to the problem of communication over a channel where there is a second channel available for sending information in the reverse direction. The second channel can, in certain cases, be used to improve forward transmission. Upper bounds have been found for the forward capacity in such a case. The delay free theory also has an application to the problem of linear least square smoothing and prediction [1]. A minimum phase filter has an inverse (without delay) and therefore belongs to the delay free group of translations for continuous time series. The least square prediction problem can be solved by translating the time series in question to a canonical form and finding the best prediction operator for this form.

References

1. H. W. Bode and C. E. Shannon, *A simplified derivation of linear least square smoothing and prediction theory*, Proceedings of the Institute of Radio Engineers vol. 38 (1950) pp. 417–425.

2. C. E. Shannon and W. Weaver, *A mathematical theory of communication*, University of Illinois Press, 1949.

Bell Telephone Laboratories, Inc.,
Murray Hill, N. J., U. S. A.

RANDOM PROCESSES AND TRANSFORMATIONS

S. ULAM

I

It is intended to present here a general point of view and specific problems connected with the relation between descriptions of physical phenomena by random processes and the theory of probabilities on the one hand, and the deterministic descriptions by methods of classical analysis in mathematical physics, on the other. We shall attempt to formulate procedures of random processes which will permit heuristic and also quantitative evaluations of the solutions of differential or integral-differential equations. Broadly speaking, such methods will amount to construction of statistical models of given physical situations and statistical *experiments* designed to evaluate the behavior of the physical quantities involved.

The role of probability theory in physics is really manifold. In classical theories the role of *initial conditions* is consciously idealized. In reality these initial conditions are known only within certain ranges of values. One could say that probability distributions are given for the actual values of initial parameters. The influence of the variation of initial conditions, subject to "small" fluctuations, on the properties of solutions has been studied in numerous cases and forms one subject of the theories of "stability."

In a more general way, not only the initial constants, but even the operators describing the behavior of a given physical situation may not be known exactly. We might assume that, for example, the forces acting on a given mechanical system are known only within certain limits. They might depend, for example, to some extent on certain "hidden" parameters and we might again study the influence of random terms in these forces on the given system.

In quantum theory, of course, the role of a stochastic point of view is even more fundamental. The variables describing a physical system are of higher mathematical type. They are sets of points or sets of numbers (real, complex, or still more general) rather than the numbers themselves. The probability distributions enter from the beginning as the primitive notions of the theory. The observable or measurable quantities are values of certain functionals or eigenvalues of operators acting on these distributions. Again, in addition to this fundamental role of the probabilities formulation, there will enter the fact that the nature of forces or conditions may not be known correctly or exactly, but the operators corresponding to them will depend on "hidden" parameters in a fashion similar to that in classical physics. In fact, at the present time considerable latitude exists in the choice of operators corresponding to "forces" in nuclear physics.

There is, in addition, another reason for the recourse to descriptions in the spirit of the theory of probabilities which permit from the beginning, a flexibility and, therefore, greater generality of formulations. It is obvious that a general mathematical formalism for dealing with "complications" in models of reality

is needed already on a heuristic level. This need is mainly due to the lack of simplicity in the presently employed models for the behavior of matter and radiation. The combinatorial complexity alone, present in such diverse problems as hydrodynamics, the theory of cosmic rays, the theory of nuclear reaction in heterogenous media, is very great. One has to remember that even in the present theories of so-called elementary particles themselves one employs rather complicated models for each of these particles and their interactions. Often the complications relate already to the qualitative topological and algebraic structure even before one attempts to pursue analysis of these models. One reason for these complications is that such problems involve a considerable number of independent variables. The infinitesimal analysis, i.e., the methods of calculus, become, for the case of many variables, unwieldy and often only purely symbolic. The class of "elementary" functions within which the operators of the calculus act in an algebraically tolerable fashion is restricted in the main to functions of one variable (real or complex). Mathematical physics deals with this increasing complexity in two opposite limiting methods. The first is the study of systems of differential or integral-differential equations describing in detail the behavior of each element of the system under consideration. The second, an opposite extreme in treatment, is found in theories like statistical mechanics dealing with only a few total or integral properties of systems which consist of enormous numbers of objects. There we resign ourselves to the study of only a few functionals or operators on such ensembles.

Systems involving, so to say, an intermediate situation have been becoming, in recent years, more and more important in both theory and practice. A mechanical problem of a system of N bodies with forces acting between them (we think here of N as having a value like, say, 10 or 20) would present an example of this kind. Similarly one can think of a continuum, say a fluid subject to given forces in which, however, we are interested in the values of N quantities describing the whole continuum of the fluid. Neither of the two extreme approaches which we mentioned is very practical in such cases. It will be impractical to try to solve exactly the deterministic equations. The purely statistical study of the system, in the spirit of thermodynamics, will not be detailed enough. The approach should be rather a combination of the two extreme points of view, the classical one of following step by step in time and space the action of differential and integral operators and the stochastic method of averaging over whole classes of initial conditions, relations, and interactions. We propose a way to combine the deterministic and probability method by some general mathematical algorithms.

In mathematics itself combinatorial analysis lacks general methods, and methodologically resembles an experimental science. Its problems are suggested by arrangements and combinations of physically existing situations and each requires for solution specific ingenuity. In analysis the subject of functional equations is in a similar position. There is a variety of special cases, each treated by special methods. According to Poincaré it is even impossible to define, in general, functional equations.

We shall now give examples of heuristic approaches all based on the same principle: of an equivalent random process through which one can examine the various problems of mathematical physics alluded to above.

One should remember that mathematical logic itself or the study of mathematics as a formal system can be considered a branch of combinatorial analysis. Metamathematics introduces a class of games—"solitaires"—to be played with symbols according to given rules. One sense of Gödel's theorem is that some properties of these games can be ascertained only by playing them.

From the practical point of view, investigation of random processes by playing the corresponding games is facilitated by the electronic computing machines. (In this connection: a simple computational device for production of a sequence of numbers with certain properties of randomness is desirable. By iterating the function $x' = 4x(1 - x)$ one obtains, for almost all x, the same ergodic distribution of iterates in $(0,1)$ [10; 12].)

II

One should remember that the distinction between a probabilistic and deterministic point of view lies often only in the interpretation and not in the mathematical treatment itself. A well-known example of this is the comparison of two problems, (1) Borel's law of large numbers for the sequence of the throws of a coin, and (2) a simple version of the ergodic theorem of Birkhoff: if one applies this ergodic theorem to a very special situation, namely, the system of real numbers in a binary expansion, the transformation T of this set on itself being a shift of the binary development by 1, one will realize that the theorems of Borel and Birkhoff assert in this case the same thing (this was noticed first, independently, by Doob, E. Hopf, and Khintchine.) In this case a formulation of the theory of probability and a deterministic one of iterating a well-defined transformation are mathematically equivalent.

In simple situations one might combine the two points of view: the one of probability theories, the other of iterating given transformations as follows. Given is a space E; given also are several measure preserving transformations $T_1, T_2, \cdots, T_n$. We start with a point p and apply to it in *turn* at *random* the given transformations. Assume for simplicity that at each time each of the N given transformations has an equal chance $= 1/N$ of being applied. It was proved by von Neumann and the author that the ergodic theorem still holds in the following version: for almost every sequence of choices of these transformations and for almost every point p the ergodic limit will exist [10; 12]. The proof consists in the use of the ergodic theorem of Birkhoff in a suitably defined space embodying, as it were, the space of all choices of the given transformations over the space E. The question of metric transitivity of a transformation, i.e., the question whether the limit in time is equal to the space average, can be similarly generalized from the iteration of a given transformation to the situation dealt with above; that is, the behavior of a sequence of points obtained by using several trans-

formations at random. One can again show, similarly to the case of one transformation [11], that metric transitivity obtains in very general cases.

III

A very simple practical illustration of a statistical approach to a mathematically well-defined problem is the evaluation of integrals by a sampling *procedure*: suppose R is a region in a k-dimensional space defined by the inequalities:

$$f_1(x_1, \cdots, x_k) < 0$$
$$f_2(x_1, \cdots, x_k) < 0$$
$$\cdots\cdots\cdots\cdots$$
$$f_l(x_1, \cdots, x_k) < 0.$$

The region is contained, say, in the unit cube. The problem is to evaluate the volume of this region. The most direct approximation is from the definition of the integral: one divides each of the k axes into a number N of, say, equidistant points. We obtain in our cube, N^k lattice points and by counting the fraction of those which do belong to the given region we obtain an approximate value of its volume. An alternative procedure would be to produce, at random, with uniform probability a number M of points in the unit cube and count again the fraction of those belonging to the given region. From Bernoulli's law of large numbers it follows that as M tends to infinity this fraction will, with probability 1, tend to the value of the volume in question. It is clear from the practical point of view that for large values of k, the second procedure will be, in general, more economical. We know the probability of an error in M tries and given the error, the necessary value of M will be for large k much smaller than N^k. Thus it can be seen in this simple problem that by playing a game of chance (producing the points at random) we may obtain quantitative estimates of numbers defined by strictly deterministic rule. Analogously, one can evaluate by such statistical procedures, integrals occurring in more general problems of "geometric probabilities."

IV

Statistical models, that is, the random processes equivalent to the deterministic transformations, are obvious in the case of physical processes described by differential diffusion equations or by integral differential equations of the Boltzmann type. These processes are, of course, the corresponding "random walks". One finds in extensive literature dealing with stochastic processes the foundations for construction and study of such models, at least for simple problems of the above type. It is known that limiting distributions resulting from such processes obey certain partial differential equations. Our aim is to invert the usual procedure. Given a partial differential equation, we construct models of suitable

games and obtain distributions or solutions of the corresponding equations by playing these games, i.e., by experiment. As an illustration consider the problem of description of large cosmic ray showers. It can be schematized as follows:

An incoming particle produces with certain probabilities new particles; each of these new particles, which are of several kinds, is, moreover, characterized by additional indices giving its momenta and energies. These particles can further multiply into new ones until the energies in the last generation fall under certain given limits. The problem is first: to predict, from the given probabilities of reactions, the statistical properties of the shower; secondly, a more difficult one, the inverse problem, where the elementary probabilities of transformation are not known but statistics of the showers are available, to estimate these probabilities from the properties of the shower. Mathematically, the problem is described by a system of ordinary differential equations or by a matrix of transitions, which has to be iterated.

A way to get the necessary statistics may be, of course, to "produce" a large number of these showers by playing a game of chance with assumed probabilities and examine the resulting distributions statistically. This may be more economical than the actual computation of the powers of the matrices describing the transition and transmutation probabilities: the multiplication of matrices corresponds to evaluation of all contingencies at each stage, whereas by playing a game of chance we select at each stage only *one* of the alternatives.

Another example: given is a medium consisting of several nuclearly different materials, one of which is uranium. One introduces one or several neutrons which will cause the generation of more neutrons through fissions in uranium. We introduce types, i.e., indices of particles corresponding to different kinds of nuclei present. In addition, the positions and velocities of particles of each type can be also characterized by additional indices of the particle so that these continuous variables are also, approximately, represented by a finite class of discrete indices. The given geometrical properties of the whole assembly and nuclear constants corresponding to the probabilities of reaction of particles (they are, in general, functions of velocities) would give us a matrix of transitions and transmutations. Assuming that time proceeds by discrete fixed intervals, we can then study the powers of the matrix. These will give us the state of the system at the nth interval of time. It is important to remember that the Markoff process involved here has infinitely many states because the numbers of particles of each type are not a priori bounded. A very schematized mathematical treatment would be given by the partial differential equation

$$\frac{\partial w}{\partial t} = a\Delta w + b(x)w.$$

This equation describes the behavior of a diffusing and multiplicative system of particles of one type, x denoting the "index" of position. For a mathematical description of this system it is preferable, instead of picturing it as an infinite-

dimensional Markoff process, to treat it as an iteration of a transformation of a space given by the generating functions [2; 3; 5; 6; 9]. (Considerable work has been done on a theory of such processes also by Russian mathematicians [8].) The transformation T, given by the generating functions which is of the form $X_i' = f_i(x_1, \cdots, x_n), i = 1, \cdots, n$, where the f_i are power series with non-negative coefficients, will define a linear transformation A whose terms a_{ij} will be the *expected values* of the numbers of particles of type j produced by starting with a particle of type i. Ordinarily, to interpret a matrix by a probabilistic game, one should have all of the terms non-negative, and the sum of each row should be equal to 1. One can generalize the interpretation of matrices, however, by playing a probability game, considering the terms not as transition probabilities but rather as the first moments or expected values of the numbers of particles of type j produced by one particle of type i. (The probabilities, of course, can be fixed in many different ways so as to yield the same given values of the moments.) One can go still further. Multiplication of matrices with arbitrary real coefficients can be studied by playing a probability game if we interpret the real numbers in each term as matrices with non-negative coefficients over two symbols:

$$1 \approx \begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix} \quad -1 \approx \begin{vmatrix} 0 & 1 \\ 1 & 0 \end{vmatrix}.$$

The negative and positive numbers require then each its own "particles" with separate indices. This correspondence preserves, of course, both addition and multiplication on matrices. Obviously, more general matrices with complex numbers as general terms admit, therefore, also of analogous probabilistic interpretation, each complex number requiring 4 types of "particles" in this correspondence [4].

The following theorem provides one mathematical relation between the properties of the iterates of the transformation given by generating functions and the iterates of the associated linear transformation (given by the expected values): With probability 1 the ratios of the numbers of particles of any two types will approach the ratios defined by the direction of the *invariant* vector given by Frobenius' theorem for the linear matrix [2; 3; 5; 6; 9].

It is possible to interpret the "particles" in a rather general and abstract fashion. Thus, for example, one may introduce an auxiliary particle whose role is that of a clock [2, part 2]. A distribution in the 4-dimensional time-space continuum can be investigated by an iteration of transition and transmutation matrices. The parameter of iteration will then be a purely mathematical variable τ, having no direct physical meaning since physical time is now one of the dependent variables.

V

In some cases one could deal with a partial differential equation as follows.

First, purely formally, we transform it into an equation of the diffusion-multiplication type. We then interpret this equation as describing the behavior of a system consisting of a large number of particles of various types which diffuse and transmute into each other. Finally we study the behavior of such a system empirically by playing a game with these particles according to prescribed chances of transitions. Suppose, for example, we have the time independent Schrödinger equation:

$$a\Delta\psi + (E - V(x, y, z))\psi = 0.$$

By introducing a new variable τ, and the function

$$u = \psi e^{-E\tau},$$

we shall obtain the equation

$$\frac{\partial u}{\partial t} = a\Delta u - Vu.$$

This latter is of the desired type. The potential $V(x, y, z)$ plays the role of expected value of the multiplication factor at the position given by the vector x [1]. Dirac's equation can also be treated in a similar fashion. (We have to introduce at least 4 types of particles since the description is not by means of real numbers but through Dirac's matrices. Again the parameter τ, as in Schrödinger's equation, is a purely auxiliary variable not interpretable as time.) Such probability models certainly have heuristic value in cases where no analytical methods are readily applicable to obtain solutions of the corresponding equations in closed form. This is, for example, the case when the potential function is not of simple enough type or in problems dealing with three or more particles. The result of a probability game will, of course, never give us the desired quantities accurately but could only allow the following possible interpretation: Given $\epsilon > 0$, $\eta > 0$, with probability $1 - \eta$, the values of quantities which we try to compute lie within ϵ of the constants obtained by our random process for sufficiently great number n of the sampling population.

One should remember that in reality the integral or partial differential equations often describe only the behavior of averages or expected values of physical quantities. Thus, for example, if one assumes as fundamental a model of the fluid as does the kinetic theory, the equations of hydrodynamics will describe the behavior of average quantities; velocities, pressures, etc., are defined by averaging these over very large numbers of atoms near a given position. The results of a probability game will reflect, to some extent, the deviation of such quantities from their average values. That is to say, the fluctuations unavoidably present as a result of the random processes performed may not be purely mathematical but may reflect, to some extent, the physical reality.

VI

One economy of a statistical formulation is this: often, in a physical problem, one is merely interested in finding the values of only a few *functionals* of an unknown distribution. Thus, for example, in a hydrodynamic problem we would like to know, say, the average velocity and the average pressure in a certain region of the fluid. In order to compute these one has to know, in an analytic formulation of the problem, the positions of all the particles of the fluid. One needs then the knowledge of the functions for all values of the independent variables. In an abstract formulation the situation is this: given is an operator equation $U(f) = 0$ where f is a function of k variables; what we want to know is the value of several given functionals $G_1(f), G_2(f), \cdots, G_l(f)$. (Sometimes, of course, even the existence of a solution of the equation $U(f) = 0$ or, which is the same, of the equation $V(f) = U(f) + f = f$, that is, the fixed point of the operator $V(f)$, is not a priori guaranteed.) The physical problem, however, consists merely in finding the values of $G_i(f)$. Mathematically it amounts to looking for functions f for which $G_i(V(f)) = G_i(f)$. We might call such f *quasi-fixed* points of the transformation V (with respect to the given functionals G_i). Obviously, the existence of quasi-fixed points is, a priori, easier to establish than the existence of a solution in the strict sense. A simple mathematical illustration follows: let T be a continuous transformation of the plane onto itself given by $x' = f(x, y)$; $y' = g(x, y)$. There need not, of course, exist a fixed point. There will always exist a point (x_0, y_0) such that $|x'_0| = |x_0|$; $|y'_0| = |y_0|$, analogously in n dimensions. Similar theorems in function spaces would permit one to assert the existence of quasi-solutions of operator equations $V(f) = f$. A quasi-solution (for given functionals) is then a function which possesses the same first n moments or the same first n coefficients in its Fourier series as its transform under V. For each n there should exist such quasi-solutions.

In a random process "equivalent" to a given equation, the values of *functionals* of the desired solution or, more generally, quasi-solutions, are obtained quite automatically as the process proceeds. The convergence in probability of the data, obtained during the process, to their true value may, in some cases, be much more rapid than the convergence of the data describing the functions themselves. This will be in general the case for functionals which have the form of integrals over the distributions.

VII

The role of "small" variations introduced in the operators which describe physical processes is discussed in elementary cases in the theories of stability. In the simplest cases one deals with the influence which variations of constants have on the behavior of solutions, say, of linear differential equations. In many purely mathematical theories one can conceive the problem of stability in a very general way. One can, for example, study instead of functional *equations*, functional *inequalities* and ask the question whether the solutions of these inequalities

are, of necessity, close to the solutions of the corresponding equations. Perhaps the simplest example would be given by the equation

$$T(x + y) = T(x) + T(y)$$

for all x, y which are elements of a vector space E, and the corresponding functional inequality:

$$|| S(x + y) - S(x) - S(y) || < \epsilon$$

for all x, y.

A result of Hyers is that there exists a T satisfying the equation such that for all x, we have then

$$|| T(x) - S(x) || < \epsilon.$$

Or, more generally, one could ask the question: given an ϵ-isomorphism F of a metric group, is there always an actual isomorphism G within, say, k times ϵ of the given F. Another example is the question of ϵ-isometric transformations T, i.e., transformations T such that for all p, q:

$$| \rho(p, q) - \rho(T(p), T(q)) | < \epsilon.$$

Here again one can show that such T differ only by $k \cdot \epsilon$ from strictly isometric transformations. To give still another example one can introduce a notion of almost convex functions and almost convex sets. Again it is possible to show that such objects differ little from strictly convex bodies which, one proves, will exist in their vicinity.

All this is mentioned here because, in order to establish rigorously the comparison between random process models of physical problems and their classical descriptions by analysis, mathematical theorems will be needed which will allow us to estimate more precisely the influence of variations not merely of constants but of the operators themselves.

In many mathematical theories it is natural to subject the definitions themselves to ϵ-variations. Thus, for example, the notion of the homeomorphic transformation can be replaced by a notion of a continuous transformation which is up to ϵ a one-to-one transformation. Again one finds that many theorems about one-to-one transformations can be generalized to hold for the almost one-to-one case.

Little is known at present about solutions of functional inequalities. One needs, of course, beyond theorems on stability, more precise information on the rapidity of the convergence in probability.

VIII

In theories which would deal with *actually* infinite assemblies of points—the probability point of view can become axiomatic and more fundamental rather than only of the approximative character evident in the previous discussion. Let us indicate as an example a purely schematic set-up of this sort. We want to treat a dynamic system of an infinite number of mass points interacting with each other. Imagine that on the infinite real axis we have put, with probability equal

to $\frac{1}{2}$, on each of the integer points a material point of mass 1. That is to say, for each integer we decided by a throw of a coin whether or not to put such a mass point on it. Having made infinitely many such decisions, we shall obtain a distribution of points on the line. It can be denoted by a real number in binary development, e.g., the indices corresponding to *ones* give us, say, for odd places, the non-negative integers where mass points are located, for the even indices of ones, we obtain the location of the mass points on the negative part of the line. Imagine that this binary number represents our system at the time $T = 0$. Assume further that the mass points attract each other with forces proportional to the inverse squares of the distances. (It is obvious that forces on each point are well-defined at all times since the sum of the inverse squares of integers converges absolutely.) Motions will now ensue. We propose to study properties of the motion common to almost all initial conditions, or theorems valid for almost all binary sequences (normal numbers in the sense of Borel). As representing initial conditions one may make the assumption that as the two points collide they will from then on stay together and form a point with a greater mass whose motion will be determined by the preservation of the momentum. It is interesting to note here that, because the total mass of the system is infinite, the various formulations of mechanics which are equivalent to each other in the case of finite systems cease to be so in this case. One can use, however, Newton's equations quite legitimately in our case. The interesting thing to notice is that the behavior of our infinite system will not be obtainable as a limiting case of the behavior of very large but finite systems approximating it. One shows, for example, that the average density of the system will remain constant equal to $\frac{1}{2}$ for all time. One can prove that collisions will lead to formations or condensations of arbitrarily high orders. For all time T there will be particles which have not yet collided with another particle. On the other hand, given a particle, the probability that it will collide at some time tends to 1. We might add that one could treat similarly systems of points distributed on integer-valued lattice points in the plane or in 3-dimensional space. The forces will not be determined any more by absolute convergence, but in 2 and 3 dimensions one can show that if we sum over squares or spheres the forces acting on a point from all the other points in the spheres whose radii tend to infinity, the limits will exist for each point with probability 1. That is, for almost every initial condition of the whole system the force is defined everywhere. In a problem of this sort it is obvious that the role of probability formulation is fundamental. Actually infinite systems of this kind may be thought of, however, as a new kind of idealization of systems already considered in present theories. This is so if we allow in advance for an infinity of hidden parameters present in the physical system, and which are not so far treated explicitly in the model. An important case in which the idealization to an actual infinity of many degrees of freedom interacting with each other seems to be useful is the recent theory of turbulence of Kolmogoroff, Onsaeger, and Heisenberg.

An interesting field of application for models consisting of an infinite number of

interacting elements may exist in the recent theories of automata.[1] A general model, considered by von Neumann and the author, would be of the following sort:

Given is an infinite lattice or graph of points, each with a finite number of connections to certain of its "neighbors." Each point is capable of a finite number of "states." The states of neighbors at time t_n induce, in a specified manner, the state of the point at time t_{n+1}. This rule of transition is fixed deterministically or, more generally, may involve partly "random" decisions.

One can define now closed finite subsystems to be called *automata* or *organisms*. They will be characterized by a periodic or almost periodic sequence of their states as function of time and by the following "spatial" character: the state of the neighbors of the "organism" has only a "weak" influence on the state of the elements of the organism; the organism can, on the contrary, influence with full generality the states of the neighboring points which are not part of other organisms.

One aim of the theory is to establish the existence of subsystems which are able to multiply, i.e., create in time other systems identical ("congruent") to themselves.

As time proceeds, by discrete intervals, one will generate, starting from a finite "activated" region, organisms of different types. One problem is again to find the equilibrium ratios of the numbers of individual species, similarly to the situation described in §IV. The generalization of Frobenius' theorem mentioned there gives one basis for the existence of limits of the ratios.

The existence of finite universal organisms forms one of the first problems of such theory. These would be closed systems able to generate arbitrarily large (or "complicated") closed systems.

One should perhaps notice that any metamathematical theory has, to some extent, formally a character of the above sort: one generates, by given rules, from given classes of symbols, new such classes.

Mathematically, the simplest versions of such schemes would consist simply of the study of iterates of infinite matrices, having nonzero elements in only a finite number of terms in each row. The problems consist of finding the properties of the finite submatrices appearing along the diagonal, as one iterates the matrix.

References

1. M. D. Donsker and M. Kac, *A sampling method for determining the lowest eigenvalue and the principal eigenfunction of Schrödinger's equation*, Journal of Research of the National Bureau of Standards vol. 44 (1950) pp. 551–557.

2. C. J. Everett and S. Ulam, Los Alamos Report L. A. D. C. 533, 534, and 2532.

3. ———, *Multiplicative systems*, I, Proc. Nat. Acad. Sci. U. S. A. vol. 34 (1948) pp. 403–405.

4. ———, *On an application of a correspondence between matrices over real algebras and matrices of positive real numbers*, Bull. Amer. Math. Soc. Abstract 56-1-96.

5. T. E. Harris, *Some mathematical models for branching processes*, RAND Corp. Report, September, 1950.

[1] J. von Neumann lectures at the University of Illinois, December, 1949.

6. D. HAWKINS and S. ULAM, Los Alamos Report L. A. D. C. 265, 1944.

7. G. W. KING, *Stochastic methods in quantum mechanics*, A. D. Little Co. Report, February, 1950.

8. A. N. KOLMOGOROFF and M. A. SEVATYANOV, *The calculation of final probabilities for branching random processes*, C. R. (Doklady) Acad. Sci. URSS (N.S.) vol. 56 (1947) pp. 783–786.

9. N. METROPOLIS and S. ULAM, *The Monte Carlo method*, Journal of the American Statistical Association vol. 44 (1949) pp. 335–341.

10. J. VON NEUMANN and S. ULAM, *Random ergodic theorems*, Bull. Amer. Math. Soc. Abstract 51-9-165.

11. J. C. OXTOBY and S. ULAM, *Measure preserving homeomorphisms and metrical transitivity*, Ann. of Math. vol. 42 (1941) pp. 874–920.

12. S. M. ULAM and J. VON NEUMANN, *On combination of stochastic and deterministic processes*, Bull. Amer. Math. Soc. Abstract 53-11-403.

LOS ALAMOS SCIENTIFIC LABORATORY,
LOS ALAMOS, N. M., U. S. A.

PARTIAL DIFFERENTIAL EQUATIONS

SHOCK INTERACTION AND ITS MATHEMATICAL ASPECTS

J. VON NEUMANN

This address was given as part of the Conference in Applied Mathematics, but no manuscript has been received by the editors.

BOUNDARY VALUE PROBLEMS IN MODERN FLUID DYNAMICS

(SUMMARY)

RICHARD COURANT

New aspects arising in boundary value problems of fluid dynamics, and largely connected with the nonlinearity of these problems were discussed in this paper.

According to Hadamard a problem of mathematical physics, to represent reality in an adequate way, should satisfy three postulates: (1) The solution should exist, (2) it should be uniquely determined by the data, (3) it should be stable, i.e., small changes in the data should produce only small changes in the solution. Classical mathematical physics or mechanics does satisfy these postulates if they are properly interpreted. Specifically, classical theories proceed by establishing general laws (differential equations) for classes of phenomena and by singling out individual phenomena by additional boundary or initial or similar conditions. The unique and stable determination of events in mechanics and physics by this scheme has long been a main tenet of natural philosophy.

Of course, modern quantum theory has shown the limitation of this classical approach. But, even more strikingly, classical dynamics of compressible fluids illuminate the fact that differential equations supplemented by boundary conditions, etc. are not always a sufficiently complete framework for an adequate description of physical reality.

Questions of this type were treated in a way similar to that used in a previous publication (Courant and Friedrichs, *Supersonic flow and shock waves*, Interscience Publishers, 1948, pp. 367 ff.). In particular the following points were discussed, though not beyond the initial stage of stating open questions: Under what conditions does an initial-boundary value problem of fluid dynamics lead asymptotically to a steady state? And most important for theoretical insight and practical applications: Assuming there is a steady state, what are the appropriate intrinsic boundary conditions that should and could be imposed in order to determine the flow?

Even the simplest problems of gas dynamics illustrate difficulties in stating such a priori conditions. For example, we may consider the problem of finding the flow in an infinite duct which originally contains gases under different pressures at rest in different parts of the duct; these gases are separated by membranes which at the time $t = 0$ are suddenly removed. The effect of the resulting interaction process will spread through the whole duct and eventually reach every part of it. Therefore we cannot expect the asymptotic steady state to be determined by the same conditions that prevailed at the far ends originally.

An illuminating example is steady flow in a duct infinitely extended in both directions. Very different types of boundary conditions, such as prescription of pressures at one or both ends, are appropriate according to the supersonic or subsonic character of the flow. However, particularly when the duct is not parallel, this character in turn has to be ascertained as a part of the solution.

The most trenchant complication in the field of nonlinear problems arises in connection with the fact that in general, even under continuous boundary and initial conditions, solutions do not exist which are continuous in the whole space-time region under consideration. As Riemann already discovered, problems of gas dynamics are mathematically solvable only if "shock discontinuities" are admitted; across the discontinuity surfaces the same conservation laws must be stipulated which in domains of continuity are expressed by differential equations, but the position of the discontinuities cannot be prescribed a priori. In this way considerable difficulties appear in the construction of solutions, and an even more serious problem arises with respect to Hadamard's second and third postulates. Apart from some of the most elementary cases, the solutions of problems involving shocks are not uniquely determined by the data which we might perhaps consider "natural," and the question of stability has hardly been taken up satisfactorily even in quite simple cases. For example, the problem of supersonic flow through a duct, such as a rocket motor, can, it is true, be tackled by the actual construction of flow patterns satisfying given conditions; yet shocks, their reflections on a wall, their Mach intersections, etc. imply a high degree of mathematical ambiguity. Hardly any serious attempt has been made so far to find out what the data and stipulations are which correspond to physical reality and ensure uniqueness and stability of the solution.

The importance of such investigations from the applied point of view is obvious; a theory which provides solutions not identified as the only possible solutions cannot be a firm basis for engineering construction. For the mathematician the challenge may seem enhanced by the suspicion that even deeper issues may be involved than the experience of a groping and imperfect mathematical description of reality. If one admits in the solution a fixed number of discontinuities, it may well turn out that we can always formulate a pertinent problem, for which at least one of the three principles of Hadamard is violated. Thus it may be that in gas dynamics or fields of related structure a paradoxical situation exists similar to that which has been observed to exist in mathematical logic. A clarification of this suspicion would indeed be a challenge to mathematical ingenuity.

In the second part of the paper the problem of numerically computing solutions of nonlinear boundary and initial value problems was discussed, in particular, a method using characteristics and based on finite differences, valid for two independent variables but for arbitrary order. A method was pointed out, including convergence and estimation of errors, which will be published shortly in collaboration with Dr. E. Isaacson and Dr. Mina Rees, and which should lend itself to use of the automatic computing machines which for such a long time have been promised to the world for the immediate future.

In some final remarks a very different type of boundary value problems of a somewhat unorthodox type was briefly mentioned, problems arising from supersonic flow around thin airfoils assuming the flow to be almost constant in

speed. The linearized theories of supersonic airfoils have led Evvard and others to remarkable new types of problems for the classical wave equation and have been recently further clarified by Ward, Gardner, and other authors. Here again, questions of uniqueness are still open, as is the generalization to more dimensions and to other types of differential equations.

NEW YORK UNIVERSITY,
NEW YORK, N. Y., U. S. A.

SELECTED PROBLEMS IN GAS DYNAMICS

S. GOLDSTEIN

1. Introduction. Originally I was asked to talk about the linearized equations of supersonic motion. I had doubts about the suitability of this topic for a talk to this Congress, and would have preferred to talk about the shortcomings of the linearized theory and also about viscosity; but on preparation I find that it is necessary to explain so much about the linearized theory that this talk will, in the main, be on that subject.

Moreover I shall restrict what I have to say to steady, supersonic motion, so that I am dealing with hyperbolic equations in three variables, x, y, z.

First, I would say that we should, in the main, be content if we could usefully discuss—I shall not say "solve" at this stage—the equations of gas dynamics for a viscous, heat-conducting gas, on the Navier-Stokes assumption that the stress tensor is a linear isotropic function of the rate-of-strain tensor. Certainly there are situations where further approximations—or a fuller use of the kinetic theory of gases—is necessary, but it would be an enormous step forward, and I think interesting and in some situations imperative to discuss the influence of viscosity and heat conduction. But that would have to be the subject of a separate talk.

Thus we neglect viscosity and heat conduction. But we must pay a price for that. Neglecting viscosity and heat conduction means neglecting molecular transfer phenomena. Even though the inviscid gas equations are what we may call the zeroth approximation of the kinetic theory, if we make such drastic simplifications, we cannot really argue from the molecular nature of the gas. But the molecular nature of the gas is the only physical argument we have for assuming the velocity components continuous, otherwise we should be tearing molecules apart. For an inviscid gas we must forego this argument and allow surfaces of discontinuity; the solution is then not unique. We may proceed in two ways. The first thing we could do would be to try to find solutions with continuous velocities and see if they are physically acceptable. In the second place we can use knowledge from experiment and from the theory of viscous gases to postulate the kind and position of the discontinuities that we might expect or accept. Thus we must be prepared to allow the appearance at any rate of vortex sheets and shock waves—and only previous experience will tell us much about vortex sheets, at any rate.

2. Fundamental equations. In the steady motion of an inviscid gas the entropy S and the total energy $I + q^2/2 + \Omega$, which is the sum of the enthalpy I, the kinetic energy $q^2/2 = (u^2 + v^2 + w^2)/2$, and the potential energy Ω, per unit mass, are constant along each streamline. The usual theory takes them constant throughout the field. With the extra assumption that the velocities are continuous, the

motion is then irrotational by the theorems of Kelvin and Helmholtz, and there is a velocity potential ϕ such that the velocity components are

$$(u, v, w) = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z}\right)\phi \tag{1}$$

and ϕ satisfies the equation

$$(c^2 - u^2)\phi_{xx} + (c^2 - v^2)\phi_{yy} + (c^2 - w^2)\phi_{zz} - 2uv\phi_{xy} - 2vw\phi_{yz} - 2wu\phi_{zx} = 0, \tag{2}$$

where c is the local sound velocity. S being constant, $c^2 = (\partial p/\partial \rho)_S$ is a known function of I and is given by the constancy of the total energy. It is also usual to specialize to a perfect gas with constant specific heats c_p, c_v with a ratio of $\gamma = c_p/c_v$ (= 1.4 for air), and also to neglect Ω, the neglect of Ω being permissible except on a meteorological scale or in other very special circumstances. Then

$$\frac{c^2}{\gamma - 1} + \tfrac{1}{2}(u^2 + v^2 + w^2) = \text{constant}. \tag{3}$$

But these last assumptions are not always needed, and some progress can be made without them.

In the linearized theory we consider only small perturbations of a uniform velocity U, which we take to start with along the axis of x, and if now $U + u$, v, w are the velocity components, terms of the second and higher degree in u, v, w and their derivatives are omitted. ϕ is now used for the perturbation potential, so that $u = \partial\phi/\partial x$, etc., and the linearized equation for ϕ is simply

$$\frac{\partial^2\phi}{\partial y^2} + \frac{\partial^2\phi}{\partial z^2} = \frac{\partial^2\phi}{\partial x^2} \cdot \cot^2\mu, \tag{4}$$

where μ is the Mach angle, $\cot^2\mu = M^2 - 1$, M is the Mach number, $M^2 = U^2/c_0^2$, and c_0 is the undisturbed velocity of sound.

Equation (4), of course, is simply the two-dimensional wave equation. I shall not pause here to discuss the possible analogies; my business is to discuss the application to supersonic aerodynamics. But it is both interesting and efficient that research on supersonic aerodynamics, accoustics, electromagnetic waves, water waves, etc., should be carried out in the same place and that everyone should learn something of everyone else's business.

3. Thin body and slender body. We usually consider the perturbation produced by a thin or slender body as an obstacle in the flow, and it is convenient to distinguish between a thin and a slender body. A thin body is one whose thickness is small in one direction only, and a slender body is one whose thickness is small in two directions; the limit of a thin body is a surface, the limit of a slender body is a line. A typical thin body is a supersonic airfoil; a typical slender body is a pointed body of revolution.

4. Boundary conditions. The first boundary condition is that of zero normal velocity on the body. For a thin body this may also be completely linearized and applied at the mean surface of the body; for a nearly plane airfoil at a small incidence lying near the plane $z = 0$, for example, it may be applied on the projection of the airfoil surface on $z = 0$. The mean surface, or the projection on $z = 0$ for a nearly plane airfoil, will be denoted by S.

For a slender body, however, this boundary condition may not be completely linearized in the same way, because of the nature of the singularity at the axis when the solution is continued into the body.

Other boundary conditions follow from the physical picture that for flow past an obstacle there is no disturbance at all upstream of the obstacle, and for a finite obstacle no disturbance beyond a finite distance at any section of the obstacle by a plane $X =$ constant; in fact the disturbance must have the character of waves proceeding from the obstacle, except in so far as these are reflected at shock waves, and this reflection is neglected in the linearized theory.

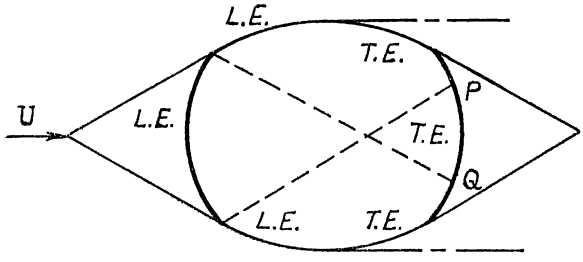

FIGURE 1

For an airfoil, or any thin body, we have to distinguish between leading and trailing edges and between supersonic and subsonic edges (see Fig. 1). Supersonic edges are inclined to the stream direction at an angle greater than the Mach angle, subsonic edges at a smaller angle.

At a trailing edge of a lifting surface, it is now usual to take the Kutta-Jonkowski condition to be satisfied, that the velocities shall be finite; grad ϕ is then continuous at a subsonic trailing edge, but not necessarily so at a supersonic trailing edge. There is a trailing vortex sheet whose mean or standard position (or projection on $z = 0$ for a nearly plane airfoil) is denoted by T. It is usual to make Prandtl's assumption that the vortex lines are exactly parallel to the undisturbed velocity U, thereby, even on a linearized theory, neglecting the perturbation velocity normal to the sheet; at the sheet the correct boundary conditions would be that the pressure must be continuous and the velocity tangential.

In the linearized theory there may be, corresponding with shock waves in a more exact theory, characteristic surfaces at which the perturbation velocity is discontinuous; at such surfaces the tangential velocity and the normal mass-flux are continuous. Except on such surfaces, on vortex sheets extending downstream from trailing edges, and on any jet boundaries, for example, that may be present

as an assigned part of the problem, the perturbation velocities are assumed to be continuous at all points actually in the fluid.

As another condition, to make the solution definite, it is found necessary in supersonic flow to impose the additional condition that the overall aerodynamic forces on a body are finite, which is a condition of integrability on the singularities.

The velocity components have infinities, on linearized theory, at a subsonic leading edge, and the forces and moments must be found by a limiting process. If ψ is the angle between the tangent to the edge and the direction of the undisturbed stream, and r is distance, in a tangent plane, normal to the edge, the velocity at almost all points on the edge asymptotes $Kr^{-1/2}$ as $r \to 0$, where K is finite and independent of r. The contribution to the force is then $\pi\rho_0 K^2/(1 - M^2 \sin^2 \psi)^{1/2}$ per unit length of the edge and normally outwards from the edge, where ρ_0 is the undisturbed density; this acts on subsonic edges only, where $M^2 \sin^2 \psi < 1$.

According to the usual linearized theory, there are also infinities of the velocity at the edge of the vortex sheet, which invalidate the solution in that neighborhood, and also make rigorous calculations from momentum considerations of the forces and moments on an airfoil difficult; this problem has, in fact, not been completely solved; in particular, it is difficult to calculate the lateral force on a thin wing which is not nearly plane. In a more exact theory the vortex sheet would roll up at the edges behind the body, and the rolling-up proceeds the more slowly the smaller the strength of the vortex sheet, so the usual assumptions of linearization represent an approximation valid in a region which increases as the strength decreases, but not uniformly valid in the whole space.

5. Some general theorems in the linearized theory. The equations may be expressed in a rather different form. If $\boldsymbol{U}$ is the vector undisturbed velocity, $\boldsymbol{v}$ the vector perturbation velocity, and $\boldsymbol{w}$ the perturbation of the mass flux (first introduced by Robinson) so that

$$\rho_0(\boldsymbol{w} + \boldsymbol{U}) = \rho(\boldsymbol{U} + \boldsymbol{v}) \tag{5}$$

where ρ is the density and ρ_0 the undisturbed density, then the field equations are simply

$$\operatorname{curl} \boldsymbol{v} = 0 \qquad \operatorname{div} \boldsymbol{w} = 0 \tag{6}$$

and, to a linear approximation,

$$\boldsymbol{w} = \boldsymbol{v} - \frac{\boldsymbol{U}\cdot\boldsymbol{v}}{c_0^2}\boldsymbol{U}. \tag{7}$$

The pressure perturbation is given by

$$\begin{aligned} \frac{\Delta p}{\rho_0} &= -\boldsymbol{U}\cdot\boldsymbol{v} \text{ to the first order} \\ &= -\boldsymbol{U}\cdot\boldsymbol{v} - \tfrac{1}{2}\,\boldsymbol{v}\cdot\boldsymbol{w} \text{ to the second order.} \end{aligned} \tag{8}$$

If now v_1, w_1; v_2, w_2 satisfy the linearized equations and are bounded and continuous on and inside a closed surface except that v_1, v_2 may be discontinuous at a characteristic surface of the linearized equation, then

$$\int_S \{(w_2 \cdot n)v_1 + (w_1 \cdot n)v_2 - (v_1 \cdot w_2)n\}\, dS = 0 \tag{9}$$

where n is unit normal vector to S. This quadratic identity, given by Ursell and Ward, is used by them to prove uniqueness with the Kutta-Jonkowski condition imposed, and a general reversed flow theorem. Reversed flow theorems had, in fact, been previously given by many authors. What Ursell and Ward show is mathematically expressed as

$$U \cdot \int_\Sigma (\Delta p_1 n_2 + \Delta p_2 n_1)\, dS = 0. \tag{10}$$

We consider two boundary surfaces for thin bodies having the same mean surface, Σ, or, for nearly plane bodies, the same projection on $z = 0$. They are in opposite flows, U, $-U$; v_1, v_2 are the perturbation velocities on linearized theory; Δp_1, Δp_2 are the linearized perturbation pressures ($\Delta p_i = \rho_0 U \cdot v_i$, $i = 1, 2$). Now the integrals are related to the forces and moments on the thin body, but since Δp_1, Δp_2 are first order linearized perturbation pressures, the forces and moments so calculated do not include contributions from the suction forces on sharp subsonic edges. Usually n_1 and n_2 are unit normals to the surface of the thin body, but they may in fact be any vector functions of position so long as, on Σ, v_1, v_2 satisfy the boundary conditions

$$n \cdot v_i = -U \cdot n_i \tag{11}$$

($i = 1, 2$), and this allows all, or any part, of either surface to be given a rotation with a small angular velocity. Since also the surfaces need not be the same, having only the same mean surface or projection, parts of either surface may be warped to obtain the effects of control surfaces, etc. Thus the theorem is very general. Simple examples are:

Drag forces (with leading edge suction neglected) are the same for the same body in the reversed flow, and so is the initial slope of the curve of lift against incidence. Again, the rolling moment due to a given rate of yaw is equal to the yawing moment in the reversed flow due to the same rate of roll.

6. Conical flows. I should say a few words next on methods of solution. A mathematically simple and elegant method of solution may be found for the so-called conical flows (originally discussed by Busemann) in which u, v, w are homogeneous functions of x, y, z of degree zero, and so are constant along all radii vectors through the origin. The downstream half of the cone of semi-vertical angle μ with vertex at the origin is the Mach cone of the origin. Outside the Mach cone, u, v, w satisfy the simple wave equation

$$\frac{\partial^2 u}{\partial \sigma^2} = \frac{\partial^2 u}{\partial \theta^2} \tag{12}$$

and inside, the harmonic equation

$$\frac{\partial^2 u}{\partial s^2} + \frac{\partial^2 u}{\partial \theta^2} = 0 \tag{13}$$

where

$$\frac{y}{x \tan \mu} = \left.\begin{matrix} \sec \sigma \cos \theta \\ \operatorname{sech} s \cos \theta \end{matrix}\right\}; \quad \frac{z}{x \tan \mu} = \left.\begin{matrix} \sec \sigma \sin \theta \\ \operatorname{sech} s \sin \theta \end{matrix}\right\}. \tag{14}$$

Thus inside the Mach cone, u, v, w are the real parts of three functions g_1, g_2, g_3 of $\zeta = e^{s+i\theta}$ and it may be shown that

$$g_1'(\zeta) : g_2'(\zeta) : g_3'(\zeta) = 2\zeta \tan \mu : -(\zeta^2 + 1) : i(\zeta^2 - 1) \tag{15}$$

where primes represent derivatives. Similarly outside the Mach cone

$$u, v, w, = f_{1,2,3}(\sigma + \theta) + F_{1,2,3}(\sigma - \theta)$$

and

$$\begin{aligned} f_1' : f_2' : f_3' &= -\tan \mu : \cos(\sigma + \theta) : \sin(\sigma + \theta) \\ F_1' : F_2' : F_3' &= -\tan \mu : \cos(\sigma - \theta) : -\sin(\sigma - \theta). \end{aligned} \tag{16}$$

This method is used to solve problems of flow past thin cones, past plane triangular airfoils, and, by superposition, past plane polygonal airfoils, and to solve many other problems of physical and engineering interest. In some cases (e.g., cone with wing attached inside the Mach cone) the boundary conditions are mixed and the problem remains of some difficulty, but most problems are easy.

The method has been generalized by Lagerstrom to include the method of the integration of conical field solutions with different vertices, and by Germain to the case when ϕ is a homogeneous function of degree n.

Now in linearized conical field solutions there are discontinuities at the Mach cone. The velocities are finite, but their derivatives are infinite like $1/(1 - r)^{1/2}$, where $r = 1$ is the Mach cone of the vertex. We therefore expect attempts to improve linearized theory by successive approximation to fail, since the ratio of succeeding approximations, or rather of succeeding terms added to form a series, will have a factor $1/(1 - r)^{1/2}$, and any such series will diverge near the Mach cone.

7. The method of sources and sinks. For nearly plane thin bodies (airfoils), the problem is solved by the use of Hadamard's solution of the general second order hyperbolic equation.

For $z > 0$

$$\phi(x, y, z) = -\frac{1}{\pi}\iint \left[\frac{\partial\phi(x', y', z')}{\partial z'}\right]_{z'=+0} \cdot \frac{dx'\,dy'}{((x'-x')^2 - \cot^2\mu[(y-y')^2 + z^2])^{1/2}} \tag{17}$$

over the part of $z = 0$ for which $x' \leqq x - \cot\mu \mid y - y' \mid$.

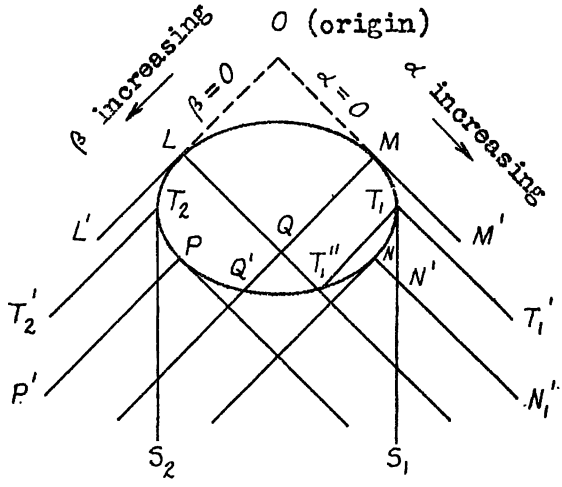

FIGURE 2

The problem is divided into two parts, for one of which ϕ is even and $\partial\phi/\partial z$ odd in z, and for the other ϕ is odd and $\partial\phi/\partial z$ even. The above formula solves the symmetrical problem immediately, since $\partial\phi/\partial z$ is known from the boundary condition on S and $\partial\phi/\partial z = 0$ elsewhere on $z = 0$. For the antisymmetrical problem $\partial\phi/\partial z$ is given on S, $\partial\phi/\partial x = 0$ on T and $\phi = 0$ on the rest R of the plane $z = 0$, ϕ is continuous in $z \geqq +0$ and grad ϕ is finite on the common boundary of S and T. The problem is solved if ϕ is found on S, since ϕ is then known everywhere on $z = 0$. The formula is transferred to characteristic coordinates $\alpha = x - y\cot\mu$, $\beta = x + y\cot\mu$ so that

$$(\phi)_{z=0} = \iint \frac{f(\alpha', \beta')\,d\alpha'\,d\beta'}{((\alpha-\alpha')(\beta-\beta'))^{1/2}} \tag{18}$$

where $f(\alpha', \beta')$ is known on S, and the integral is over $\alpha' \leqq \alpha$, $\beta' \leqq \beta$. The solution is completed, step by step, by solving an integral equation of Abel's type. This type of solution is originally due to Evvard, and has been further considered by Ward and by Germain.

A very neat solution is possible if the subsonic trailing edges are independent (P and Q are not on the subsonic edges in Fig. 1). There is no time to go into the details of the solution at any length, but the first step may be briefly sketched.

ϕ and all its derivatives are identically zero upstream of the envelope surface of all the downstream Mach cones with vertices on the supersonic leading edge LM (Fig. 2), so that $f(\alpha', \beta') = 0$ on the plane $z = 0$ forward of $L'LMM'$. For a point (α, β) on S in LQM, $f(\alpha, \beta)$ is therefore known in the whole region of integration, and the formula (18), as it stands, may be used to find $(\phi)_{z=0}$.

Now consider a point (α, β) in the region $M'MT_1T_1'$ (Fig. 2). In this region $(\phi)_{z=0} = 0$. Consider now only the case in which β is a single-valued function, $A_1(\alpha)$, of α on PLM, and a single-valued function, $A_2(\alpha)$, of α on MNP. Then in the integral for $(\phi)_{z=0}$ in (18) the limits for α' are 0 and α, and the limits for β' are $A_1(\alpha')$ and β; when β' is between $A_1(\alpha')$ and $A_2(\alpha')$, $f(\alpha', \beta')$ is known (since the relevant point is on S), and is equal, say, to $\lambda(\alpha', \beta')$; when β' is between $A_2(\alpha')$ and β, $f(\alpha', \beta')$ is unknown; let its value be $\mu(\alpha', \beta')$. Then we must have

$$(19) \qquad \int_0^{\alpha} \frac{d\alpha'}{(\alpha - \alpha')^{1/2}} \left\{ \int_{A_1(\alpha')}^{A_2(\alpha')} \frac{\lambda(\alpha', \beta')\, d\beta'}{(\beta - \beta')^{1/2}} + \int_{A_2(\alpha')}^{\beta} \frac{\mu(\alpha', \beta')\, d\beta'}{(\beta - \beta')^{1/2}} \right\} = 0 .$$

This is an integral equation for the quantity in braces; its solution is that this quantity is zero. (This gives another integral equation, this time for $\mu(\alpha', \beta')$, which may be solved for $\mu(\alpha', \beta')$.)

The value of ϕ is now found at once for a point on S in $T_1T_1''Q'M$. We suppose that on LMN, α is a single-valued function, $B_1(\beta)$, of β. Then for a point in $T_1T_1''Q'M$

$$(20) \qquad \begin{aligned} (\phi)_{z=0} &= \int_0^{\alpha} \int_{A_1(\alpha')}^{\beta} \frac{f(\alpha', \beta')\, d\alpha'\, d\beta'}{(\alpha - \alpha')^{1/2}\, (\beta - \beta')^{1/2}} \\ &= \int_{B_1(\beta)}^{\alpha} \frac{d\alpha'}{(\alpha - \alpha')^{1/2}} \int_{A_1(\alpha')}^{\beta} \frac{\lambda(\alpha', \beta')\, d\beta'}{(\beta - \beta')^{1/2}} \\ &\quad + \int_0^{B_1(\beta')} \frac{d\alpha'}{(\alpha - \alpha')^{1/2}} \left\{ \int_{A_1(\alpha')}^{A_2(\alpha')} \frac{\lambda(\alpha', \beta')\, d\beta'}{(\beta - \beta')^{1/2}} + \int_{A_2(\alpha')}^{\beta} \frac{\mu(\alpha', \beta')\, d\beta'}{(\beta - \beta')^{1/2}} \right\}. \end{aligned}$$

From the previous result, the quantity in braces is zero, so that only the first line of this last expression for ϕ survives, involving only the known function $\lambda(\alpha', \beta')$.

To find ϕ for a point on NT_1T_1'', the result for $N'T_1N$ must be used. But here we are on the trailing vortex sheet, so that ϕ is not zero; it is a function of y only, and therefore of $\alpha - \beta$. This function vanishes on T_1S_1. We may proceed as before if we assume this function to be known, and later this function is, in fact, determined by the condition that there shall be no infinity in the velocity on the boundary T_1N.

8. Heaviside operational method. The most interesting problem considered by this method is the flow outside and inside a nearly circular duct, calculated by G. N. Ward.

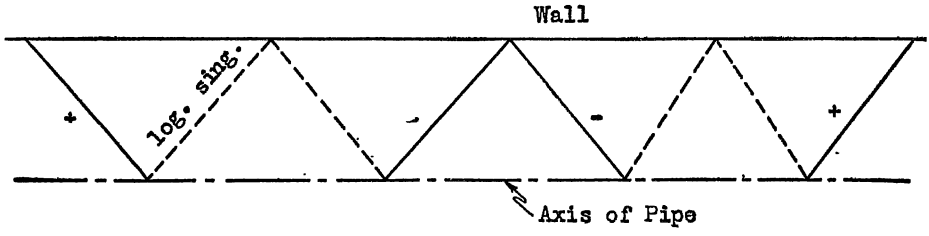

FIGURE 3

The most interesting result relates to the singularities for the flow inside the duct. The same singularities occur for flow along a circular pipe, initially of constant cross-section, when the cross-section changes in such a way that the boundary of a meridian section has a discontinuous tangent. Along the leading characteristic we have a discontinuity, either a compression or an expansion. This is reflected, according to linear theory, at the axis as a logarithmic singularity, reflected again at the wall as a logarithmic singularity, then reflected at the axis as a discontinuity of opposite sign, and so on (Fig. 3). This is quite different from the result on linearized theory for two-dimensional flow, where all discontinuities are of the same sign. Although we know that the flow inside breaks down, with the appearance of shock waves, the relation of the linearized solution to the exact inviscid solution is still an interesting unsolved problem. Is the linearized solution here completely wrong after the leading characteristic?

9. Improvement of the linearized approximation. To improve the linear approximation we might first try a method of iteration equivalent to expanding ϕ, as nearly as possible, in a power series of the thickness t. This can be attempted both for flow past a thin body and for flow past a slender body. For a pointed body of revolution, Broderick found it necessary to use a series of the form

$$\phi = t^2\phi_2 + t^3\phi_3 + t^4\{\phi_4 + \bar{\phi}_4 \log t\} + t^5\phi_5 + t^6(\phi_6 + \bar{\phi}_6 \log t + \bar{\bar{\phi}}_6(\log t)^2) + \cdots. \tag{21}$$

For a body of revolution ϕ_3 and ϕ_5 are identically zero, ϕ_2, ϕ_4, and $\bar{\phi}_4$ were found explicitly. The assumption of a potential neglects the entropy change at the bow shock wave, but it appears that this entropy change is in fact $O(t^{12})$. An approach like this fails not only when the Mach number M is too near unity, or is too large, but also near the bow shock wave for any Mach number. The solution obtained differs from uniform flow only inside the Mach cone of the vertex, and gives no better approximation to the position of the shock wave; the series diverges on the Mach cone. In fact it is valid only near the body. Nevertheless to obtain it the condition must be used that the solution represents waves outward from the body with no inward-moving waves, and this is a boundary condition at a large distance. If this is not done, the only other available method is to use boundary conditions at the bow shock wave. But for this, the solution at and near the shock wave must be improved, and this cannot be done by successive approxi-

mation. It has been done, by an entirely different technique, for conical flows. For flow past a cone, for example, for which an exact numerical solution is also known, full analytical and numerical comparisons may be carried out, and the series solution (21) appears to be correct under the conditions assumed.

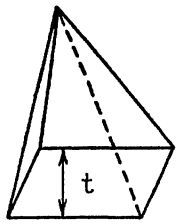

FIGURE 4

Extensions to bodies of revolution at yaw, and to bodies of revolution the slope of whose meridian section is discontinuous, have been made by Lighthill; and Ward tells me that it can be extended to the general slender body.

For a *thin* body the expansion in powers of the thickness is different: it begins

$$\phi = t\phi_1 + t^2\phi_2 + \cdots. \tag{22}$$

A second approximation has been found, by Moore at Cornell University under Professor Sears, for an arrow-head wing of finite thickness (Fig. 4), i.e., a tetrahedron of which one side of a section is the thickness parameter that is considered small. This flow is a conical flow, and Moore uses boundary condition at the shock on the improved theory to complete his solution.

10. Improvement near a singular characteristic. To improve the linearized conical flow solutions, for example, and in other cases, the mathematical problem therefore arises of improving the solution of a hyperbolic partial differential equation near a singular characteristic. A method that has been developed and used by Lighthill is an extension of that used for an ordinary differential equation near a singular point, and, in particular, of a method used in discussing nonlinear oscillations and the appearance of a limit cycle, a method that dates back to Poincaré—in the *Mécanique céleste*—except that the question there is, for an ordinary differential equation, rather harder than is needed, since the singularity is, in fact, an irregular singularity in such cases. A simple example would be

$$(x + \alpha u)\frac{du}{dx} + q(x)\cdot u = r(x) \tag{23}$$

with α small and $x = 0$ as a singular point of the linearized equation. The method is to introduce a new variable z for which

$$x = z + \alpha x_1(z) + \alpha^2 x_2(z) + \cdots \tag{24}$$

$$u = u_0(z) + \alpha u_1(z) + \cdots. \tag{25}$$

Equation (24) defines z, and we may choose x_1, etc. in lefining z so that the series for u converges as well near the singular point as elsewhere. This can be extended to singular characteristics of hyperbolic partial differential equations. An example might be the equations

$$\frac{\partial u}{\partial y} = \alpha\left(u + \frac{\partial v}{\partial y}\right)\frac{\partial u}{\partial x}, \qquad \frac{\partial v}{\partial x} = u \tag{26}$$

with α small. For the linearized equation, $\partial u/\partial y = 0$, of which the solution is $u = u_0(x)$, and this may have a singularity at $x = 0$, say $u_0 \sim A/x^q$, where $q > 0$.

We write

$$\begin{aligned} x &= z + \alpha x_1(z, y) + \alpha^2 x_2(z, y) + \cdots \\ u &= u_0(z, y) + \alpha u_1(z, y) + \cdots \\ v &= v_0(z, y) + \alpha v_1(z, y) + \cdots \end{aligned} \tag{27}$$

and $x_1(z, y)$, etc. may be found so that the expansions for u and v are valid in a region including $x = 0$.

11. Application to conical flows. Lighthill has applied this method to conical flows. If $r = 1$ is the Mach cone of the origin on linearized theory (r being equal to sec σ or sech s in equation (14)), the velocity potential ϕ is expressed as

$$\phi = Ux[1 + f(r, \theta)]. \tag{28}$$

The variable is then changed from r to R, and f expanded in a series

$$\begin{aligned} r &= R + r_1(\theta) + r_2(R, \theta) + r_3(R, \theta) + \cdots \\ f &= f_1(R, \theta) + f_2(R, \theta) + f_3(R, \theta) + \cdots \end{aligned} \tag{29}$$

where $f = f_1$ is the linearized solution, f_2 is quadratic in f_1 and its derivatives, and so on, and r_k is of the same order as f_k. Then r_1, r_2, etc. are to be found so that the expansion for f is valid in a region near $R = 1$. The solutions are found separately for $R > 1$ and for $R < 1$, and, for given r and θ, R has different meanings in the two cases; the regions of validity of the two solutions overlap, and in the region of overlap, boundary conditions may be applied, appropriate to the appearance of a shock wave, to determine the approximate position of the shock at which these conditions hold.

12. Improvement of solution at infinity. For hyperbolic partial differential equations, approximation solutions may also be nonuniform at infinity owing to the deviation of the characteristics. As we go to infinity along one set of characteristics, the slope may be everywhere nearly the same as on linear theory,

but the characteristic curve itself may be changed in position by an infinite amount. If we go to infinity along a linearized characteristic $y =$ constant, it is sufficient, in order to improve the approximation, to replace y by a more nearly exact characteristic coordinate z, which is itself found by an expansion in y. The simplest example, considered by Whitham, is

$$2\frac{\partial u}{\partial r} + 2b\frac{\partial u}{\partial x} + \frac{u}{r} + ku\frac{\partial u}{\partial x} = 0 \tag{30}$$

$$\text{i.e.,} \quad \frac{\partial}{\partial r}(ur^{1/2}) + b\frac{\partial}{\partial x}(ur^{1/2}) + \tfrac{1}{2}kur^{1/2}\frac{\partial u}{\partial x} = 0.$$

For the linearized equation the solution is

$$u = \frac{1}{r^{1/2}} f(x - br). \tag{31}$$

(Here we suppose, not that k is small, but that u is small because r is large.) In this particular case the full equation may be solved exactly and the solution is

$$u = -\frac{z}{kr^{1/2}} \tag{32}$$

with

$$x = br - zr^{1/2} - h(z), \tag{33}$$

where $h(z)$ is an arbitrary function of z. Hence, along $z =$ constant

$$\frac{dx}{dr} = b - \frac{z}{2r^{1/2}} = b + \tfrac{1}{2}ku \tag{34}$$

and $z =$ constant is an exact characteristic.

The above type of analysis has been applied to study the flow at a large distance from a body of revolution, where linearized theory fails. Mach lines intersect the bow shock wave, which is thereby weakened and curved inwards. The equation of the bow shock wave is found to be of the form

$$x = r\cot\mu_0 - br^{1/4} - c - \frac{d}{r^{1/4}} + \cdots \tag{35}$$

and that of the rear shock wave of the form

$$x = r\cot\mu_0 + br^{1/4} + \cdots. \tag{36}$$

At a constant value of r, the pressure falls approximately linearly between the shocks.

THE TECHNION,
HAIFA, ISRAEL.

ON THE STABILITY OF LAMINAR FLOW[1]

W. HEISENBERG

The stability of laminar flow has for a long time been a subject of considerable dispute, and it is only recently that clarity has been achieved in the essential points. Recently two surveys of the problem, by Lin [3] and by Tollmien [13] have been published. Since these surveys and the calculations contained in them have contributed essentially to the clarity in this problem, I need not here repeat the survey in detail. I would like however to go shortly through the history of the problem and to add a few remarks at those points where I think that a further clarification is necessary. At the same time I would like to discuss the physical interpretation of the mathematical results and to compare it with the physical interpretation of the statistical isotropic turbulence.

1. We confine ourselves in the usual way to the two-dimensional flow between parallel walls, and we know from the work of Squire that the extension to three-dimensional perturbations would not alter the problem essentially. Let us call the original velocity distribution $w(y)$, the perturbation in velocity u and v, with

$$u = \frac{\partial \psi}{\partial y}, \qquad v = -\frac{\partial \psi}{\partial x},$$

and put

$$\psi(x,y) = \varphi(y) e^{i\alpha(x-ct)},$$

then

$$(1) \qquad (w - c)(\varphi'' - \alpha^2 \varphi) - w''\varphi = -\frac{i}{\alpha R}(\varphi'''' - 2\alpha^2 \varphi'' + \alpha^4 \varphi)$$

according to Orr [6] and Sommerfeld [9].

This differential equation represents an eigenvalue problem of a similar type to those one meets in wave mechanics. It is therefore natural to use similar methods for solution in both cases, and I would like to mention that the asymptotic method [1] which was used long ago in treating equation (1) was essentially the same method which later in wave mechanics has been worked out by Wentzel, Kramers, and Brillouin.

If one wants to know the stability at very large Reynolds numbers, it is natural to put the right side in the Orr-Sommerfeld-equation equal to zero and to omit two boundary conditions. The liquid then is allowed to slip along the walls and one gets the equation used by Rayleigh:

$$(2) \qquad (w - c)(\varphi'' - \alpha^2 \varphi) - w''\varphi = 0.$$

[1] This address was listed on the printed program under the title *Die Stabilitätsfragen der Flüssigkeitsdynamik im Zusammenhang mit der statischen Turbulenztheorie.*

This equation is sufficient to decide the instability in all cases where damped or amplified solutions of (2) exist, because then (2) contains no singular point in which viscosity could come into play. It should also be emphasized—and this is a point in which I cannot agree entirely with the mathematical formulation in Lin's analysis—that whenever there is a damped solution of (2), there must also

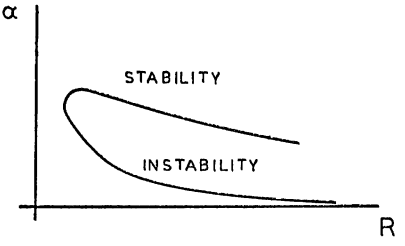

FIGURE 1

be an amplified one with the same wave length, because one may simply change the sign of α in (2). Physically you may say that one may always simply reverse the sign of time in any mechanical problem in which viscosity plays no role. It does play no role here for damped or amplified solutions since the inertia forces (left side of (1)) are at every point much stronger than the viscosity (right side of (1)). Therefore for every damped solution one can find an amplified one, and every profile of such type is unstable at very high Reynolds numbers.

The real problem, however, arises from the fact that equation (2) allows such solutions only for very special profiles, namely profiles with a point of inflection ($w'' = 0$ for $y = y_0$). As early as 1880 Lord Rayleigh has shown that one can have amplified solutions for such profiles, and starting from the complete equation (1) the limits of instability for finite Reynolds numbers are given by a curve of the type of Fig. 1. (Compare [14], [3], and [13].)

2. But some of the most important profiles have no point of inflection, and it is for those that one must from the beginning go back to the complete equation (1). This is already necessary in any case where one considers neutral disturbances; because for real values of c, equation (2) contains the singular point $w = c$ where viscosity becomes important. The simplest profile of this type is the Couette motion $w(y) = y$, where (2) has no solution, and v. Mises [4] and Hopf [2] accordingly had shown very early that (1) leads only to damped vibrations. The linear profile has been therefore known to be stable long ago. It should be emphasized in this connection, that a damped solution of equation (1) does generally *not* go over into a solution of equation (2) in the limit $R \to \infty$, since equation (2) has generally no solution. In fact, as Lin has pointed out, there is a radical difference between the damped and the amplified solutions of equation (1), contrary to the behaviour of equation (2). Generally one should expect that

in the limit $R \to \infty$ a solution of (1) will not go over into a solution of (2) which fulfills the boundary conditions; but for every solution of (2) there will be a solution of (1) which approaches it in the limit $R \to \infty$.

The next simplest profile is the Poiseuille motion between parallel walls

$$w = 1 - y^2 \qquad \text{(walls at } y = \pm 1\text{)}.$$

For this distribution equation (2) has one solution corresponding to a somewhat

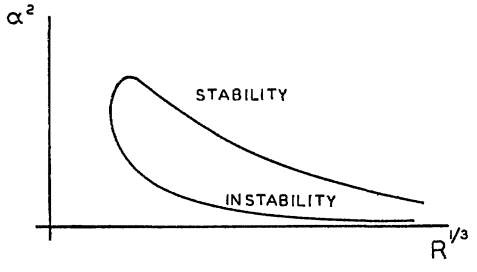

FIGURE 2

degenerate neutral disturbance: $\varphi(y) = w$, $\alpha = c = 0$. It looked natural to try whether this solution would change into an amplified solution when viscosity was taken into account. When this problem was attacked in 1924 by the method of asymptotic expansion [1], it turned out that one actually did get instability, and recently the much more accurate calculations of Lin [3] have confirmed this view. One gets an instability range of the type shown in Fig. 2. Naturally both branches of the neutral curve go to $\alpha \to 0$ for very large R, because they must tend towards the solution $\varphi(y) = w$, $\alpha = c = 0$. (In the paper of 1924 only one branch was calculated, the other one was very roughly estimated; the curve above has been given by Lin [3].)

Thus the calculations seemed to give the very natural result that whenever equation (2) has any solution at all (amplified, damped, or neutral), then the profile becomes unstable at sufficiently large Reynolds numbers. It should be mentioned, however, that not all recent calculations on the Poiseuille motion have led to this result; e.g., Pekeris [7] seems to get stability, at least for very small values of α; so that it would certainly be worthwhile to extend the calculations, possibly also on the higher harmonics.

The result described would probably have been generally accepted as plausible if shortly after 1924 a paper of Noether [5] had not appeared, in which he claimed to show that a curve of neutral disturbance can never exist for any continuous velocity distribution. The mathematical methods of Noether were better than those in the earlier papers, and the whole situation became rather obscure. Then in 1929 Tollmien [14] took up the question of stability for profiles of the boundary layer type. He got instability for such profiles and was able to calculate the limiting neutral curve. Quite recently, in 1944, Tollmien [13] has improved

the mathematical methods considerably and got similar results, and in 1947 Schubauer and Skramstad [8] could prove by careful experiments that Tollmien's calculation of the stability limits was correct.

Therefore the paper of Noether, which in his time had made the whole theory of instability suspicious, seems to contain some mistake, but this mistake has not yet been found.

Concerning the experimental side of the problem one should add that in any given experiment the limit of stability may be quite different from the theoretical one as soon as the experiments introduce new sources of instability which are not accounted for by the simple theory.

For instance the primary turbulence of the wind tunnel in which the experiment is done can according to Taylor [11] produce instability, and the same may be true for any perturbations of the stream at the entrance into the region of measurement. But this does not change the validity of the simple theory.

3. Taking now for granted that instability arises generally even in those cases in which the inviscid equation allows only a neutral solution, the question arises how viscosity can cause instability. From simple arguments one would expect damping rather than amplifying. But here one should remember that an inviscid fluid is a system of an infinite number of degrees of freedom, which normally interact so that the energy is dissipated among all modes of vibration. It is only for very special geometrical conditions that this transfer of energy does not take place. Therefore, if a neutral disturbance is possible in the inviscid fluid, the viscosity may easily change the phases of the vibration in such a manner that the transfer of energy begins, which then means amplification of the vibration.

There it is very probable, though not certain, that the amplification of the perturbation leads at once to complete turbulence, that is, to the statistical distribution of energy among all degrees of freedom. This is not quite certain since we know from the work of Taylor [10] on the cylindrical Couette case that special modes of vibration may be developed which have been described as cellular motions and which do not mean real turbulence. But this seems to be a special consequence of the centrifugal forces. In most cases the instability can safely be assumed to lead directly to turbulence. Such turbulence need not be isotropic since the walls introduce deviations from isotropy, but it will be a statistical distribution of energy among many degrees of freedom.

Finally the question remains: what happens to those profiles where the inviscid equation (2) has *no* solution at very high Reynolds numbers? These profiles certainly will be stable at extremely high Reynolds numbers if one can keep away all outer perturbations, e.g. at the entrance of the flow. But for a finite perturbation one should expect that there will be a Reynolds number from which on these perturbations will start the exchange of energy between the different degrees of freedom going, and thereby will cause turbulence. This has, however, never been followed mathematically.

REFERENCES

1. W. HEISENBERG, Annalen der Physik vol. 74 (1924) p. 577.
2. L. HOPF, Annalen der Physik vol. 44 (1914) p. 1.
3. C. C. LIN, Quarterly of Applied Mathematics vol. 3 (1946) pp. 117, 217, 277.
4. R. v. MISES, Heinrich Weber-Festschrift, 1912, p. 252.
5. F. NOETHER, Zeitschrift für Angewandte Mathematik und Mechanik vol. 6 (1926) pp. 232, 339, 428, 497.
6. W. M. F. ORR, Proceedings of the Royal Irish Academy vol. 27 (1906–1907) pp. 9–26, 69.
7. C. L. PEKERIS, The Physical Review vol. 74 (1948) p. 191.
8. G. B. SCHUBAUER and H. K. SKRAMSTAD, Journal of the Aeronautical Sciences vol. 14 (1947) p. 69.
9. A. SOMMERFELD, Atti del IV Congresso internazionale dei Matematici, Rome, 1909.
10. G. J. TAYLOR, Proc. Roy. Soc. London Ser. A vol. 102 (1923) p. 541.
11. ———, Proc. Roy. Soc. London Ser. A vol. 156 (1936) p. 307.
12. ———, Proceedings of the Fifth International Congress of Applied Mechanics, Cambridge, 1938.
13. W. TOLLMIEN, Fiat Review of German Science 1939–1946 vol. 11 Hydro- and Aerodynamics (1948) pp. 21–53.
14. ———, N. A. C. A. Technical Memoire no. 609 (1931), no. 792 (1936).

MAX PLANCK INSTITUTE FOR PHYSICS,
GÖTTINGEN, GERMANY.

ON THE BOUNDARY VALUE PROBLEMS OF THE MATHEMATICAL THEORY OF PLASTICITY

WILLIAM PRAGER

1. Introduction. The mathematical theory of elasticity is based on Hooke's law which establishes a linear relation between the tensors of stress and strain. To a given state of strain this law lets correspond a uniquely determined state of stress without regard to the process of deformation which led to the considered state of strain. The mathematical theory of plasticity takes account of the fact that real solids, in particular, structural metals, exhibit the mechanical behavior stipulated by Hooke's law only as long as the *stress intensity* remains sufficiently small. For an *isotropic* material the stress intensity must be a scalar invariant of the stress tensor, e.g., the maximum shearing stress (Tresca [10]) or the octahedral shearing stress (Nadai [7]).

When the stress intensity first reaches a certain critical value, the material leaves the *elastic range* and enters the *plastic range*. In the plastic range, the total strain is the sum of the elastic strain, which disappears upon return to the stress-free state, and the permanent plastic strain. For the *ideal* plastic material considered in the following, the stress is related to the elastic strain by Hooke's law and to the rate of change of the plastic strain by Mises' law [6]. The latter is homogeneous of the order zero in the components of the plastic strain rate. Accordingly, the state of stress reached by a certain deformation process is independent of the speed with which this process is performed. In other terms, the plastic material considered here is *inviscid*. The stress intensity can never exceed the critical value mentioned above, and changes in *plastic* strain can occur only while the stress intensity maintains this value; as soon as the stress intensity sinks below this value, any change of strain is of a purely *elastic* nature.

2. Contained plastic deformation and impending plastic flow. Since an exhaustive discussion of the boundary value problems of the mathematical theory of plasticity is not possible in the time allotted to this talk, the following remarks will be concerned with a typical boundary value problem in plane strain.

Fig. 1 shows a tensile specimen with semicircular grooves which is supposed to be tested under conditions of plane strain. As the surface traction T applied to the end sections AA' and BB' is gradually increased starting from zero, the specimen is first stressed in a purely elastic manner. Eventually, the stress intensity reaches the critical value k at the bottoms C, C' of the grooves, and plastic regions begin to form there. The value T_e of the surface traction for which this is the case marks the end of the elastic range of loading.

According to the numerical work of Allen and Southwell [1] $T_e = .66k$, where k is the critical value of the stress intensity (maximum shearing stress). For $T =$

$1.06k$ the plastic regions, as determined by these authors, are indicated by shading in Fig. 1. While finite amounts of material are stressed plastically in this

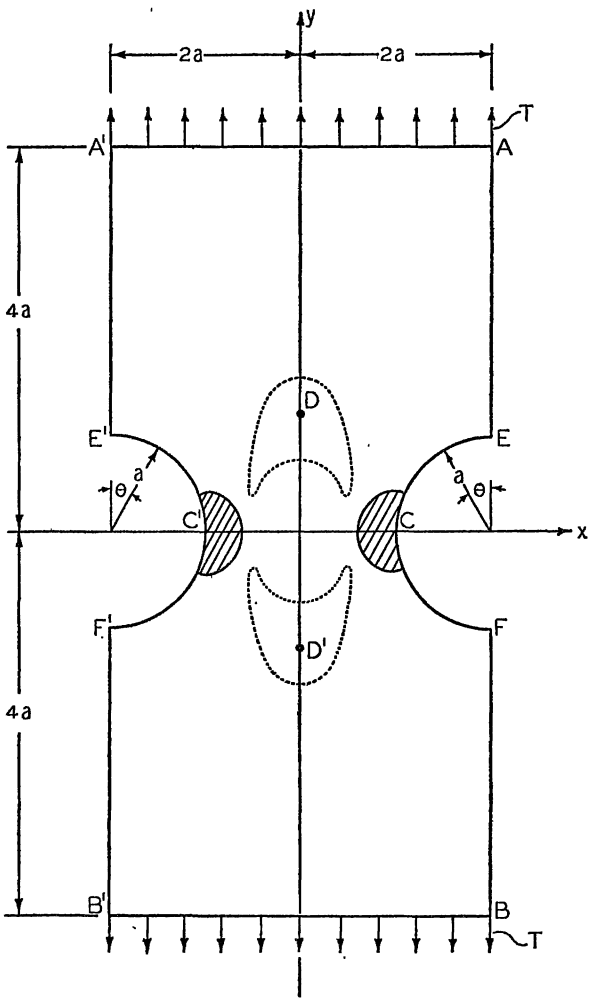

FIGURE 1

case, large plastic deformations are ruled out by the fact that there is a central elastic strip separating the peripheral plastic regions. The overall elongation of the test specimen under this surface traction is determined primarily by the elastic properties of this strip. In particular, to increase the elongation we must increase the surface traction T. This type of elastic-plastic behavior characterizes the range of *contained plastic deformation*. Throughout this range, the specimen assumes a definite overall elongation for each value of the applied surface traction just as in the elastic range. Whereas, in the elastic range, the

elongation is proportional to the surface traction T, this is no longer the case in the range of contained plastic deformation.

As we keep increasing the surface traction T, the plastic regions increase in size, and eventually a bridge of plastic material is formed which extends all the way across the specimen. At this instant of *impending plastic flow*, the specimen is first able to extend under constant surface traction. The surface traction T_f for impending plastic flow characterizes the *load carrying capacity* of the specimen, and is therefore of major importance to the structural engineer.

Before entering upon a discussion of the boundary value problem in the range of contained plastic deformation and the determination of bounds for the load carrying capacity, let us remark that the development of the plastic regions from the elastic limit to the instant of impending plastic flow may take place in a rather unexpected manner. According to the computations of Allen and Southwell [1], new plastic regions begin to form at the points marked D, D' in Fig. 1, when the surface traction T reaches the value $1.20k$. At this instant the plastic regions which developed from the points C, C' are not much larger than the shaded regions in Fig. 1 which correspond to $T = 1.06k$. The new plastic regions originating at the points D, D' spread extremely fast; for $T = 1.22k$, they attain already the size indicated by the dotted lines in Fig. 1 and are about to merge with the plastic regions which originated at the points C, C'.

3. The boundary value problem in the range of contained plastic deformation. For a given surface traction T satisfying $T_e \leqq T \leqq T_f$, the boundary value problem for the stresses is most readily formulated in terms of Airy's *stress function* $\varphi(x, y)$. This must be of class C^2 throughout the region which the specimen occupies in the x, y plane (Fig. 1). Moreover, the function φ and its normal derivative $\partial\varphi/\partial n$ (n = interior normal) must satisfy the following boundary conditions:

along AA' and BB':

$$\varphi = Tx^2/2, \qquad \partial\varphi/\partial n = 0,$$

along AE, $A'E'$, BF, and $B'F'$:

$$\varphi = 2Ta^2, \qquad \partial\varphi/\partial n = -Ta, \tag{1}$$

along ECF and $E'C'F'$:

$$\varphi = Ta^2(2 - \sin\theta), \qquad \partial\varphi/\partial n = -Ta\sin\theta.$$

Finally,

$$(\varphi_{xx} - \varphi_{yy})^2 + 4\varphi_{xy}^2 = 4k^2 \tag{2}$$

in each plastic region, k being the critical value of the stress intensity, while

$$\nabla^4\varphi = 0 \quad \text{and} \quad (\varphi_{xx} - \varphi_{yy})^2 + 4\varphi_{xy}^2 < 4k^2 \tag{3}$$

in the elastic region. Note that the curves separating elastic and plastic regions are not known beforehand but must be found from the condition that φ and its first and second derivatives must be continuous across these curves.

Obviously, boundary value problems of this kind are extremely difficult to solve. It is not surprising, therefore, that exact solutions have been obtained only in a few particular cases [3, 8]. The special methods which were successful in these cases are not applicable to the general problem outlined above. Some approximate numerical solutions have been obtained by Southwell and his collaborators [1, 5].

4. Bounds for the surface traction for impending plastic flow. In view of the fact that general methods of stress analysis in the range of contained plastic deformation are not available at present, it seems worth noting that bounds for the surface traction T_f for impending plastic flow can be obtained without difficulty as was shown recently by Greenberg, Drucker, and Prager [4].

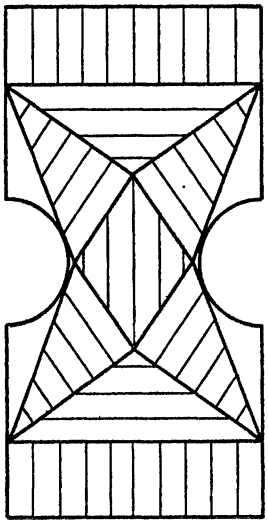

FIGURE 2

To obtain a lower bound for T_f, consider the following *relaxed* problem for the stress function $\varphi(x, y)$. Instead of requiring φ to be of class C^2 throughout the region which the specimen occupies in the x, y plane, we now require only that it be of class C^2 in each of a finite number of subregions, the function and its first derivatives being continuous across the lines separating adjacent subregions. In addition to satisfying the boundary conditions (1), the function φ of the relaxed problem needs to satisfy only the relation

$$(\varphi_{xx} - \varphi_{yy})^2 + 4\varphi_{xy}^2 \leqq 4k^2 \tag{4}$$

in each of the subregions under consideration. It can then be shown that for any value of T for which the relaxed problem has a solution, the *full* problem described in the preceding section also admits a solution.

From physical considerations, it seems likely that the solution of the full problem is unique (in this connection, see a recent paper by Galin [2]); the relaxed problem, however, admits an infinity of solutions, and some of these are readily obtained. Any value of T for which a solution of the relaxed problem can be constructed is then a lower bound for the surface traction T_f for impending plastic flow.

Fig. 2 indicates a simple solution of the relaxed problem. In each of the shaded subregions the function φ is a second order polynomial in x and y which satisfies (4) in the form of an equality; in the remainder of the specimen φ is linear in x and y and hence satisfies (4) as an inequality. Such fields of constant stress separated by straight lines of discontinuity were first discussed by Prager [9]; the particular pattern shown in Fig. 2 is due to Winzer and Carrier [11]. The corresponding value of T is found to be $1.26k$. Note that this *lower* bound for T_f slightly exceeds the value $1.22k$ given by Allen and Southwell. This may be due to the error introduced by replacing differential equations by difference equations. Alternatively, it may be possible that the bridge of plastic material which extends across the specimen must attain some finite width before the specimen begins to extend under constant surface traction.

An upper bound for T_f can be obtained from kinematic considerations. It can be shown that not only the applied surface traction T_f, but even all local stresses and, hence, the elastic energy stored in the specimen remain constant during the incipient plastic flow. Accordingly, all work done by the applied surface traction during the incipient plastic flow is dissipated in producing permanent strains.

To express this condition mathematically, consider a velocity field with components v_x, v_y defined throughout the region which the specimen occupies in the x, y plane and of class C^1 in each of a finite number of subregions. In each of these, the velocity components must satisfy the condition of incompressibility

$$\frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y} = 0; \tag{5}$$

across the line l separating two adjacent subregions, the velocity component v_n normal to l must be continuous, while the component v_t tangential to l may be discontinuous.

For the plastic material considered here, the rate of dissipation of energy associated with such a velocity field is

$$k\left[\iint\left\{\left(2\frac{\partial v_x}{\partial x}\right)^2 + \left(\frac{\partial v_y}{\partial x} + \frac{\partial v_x}{\partial y}\right)^2\right\} dx\, dy + \int [v_t]\, dl\right] \tag{6}$$

where the first integral includes all subregions and the second all lines l separating adjacent subregions, the absolute value of the jump in v_t across l being denoted

by $[v_i]$. If the considered velocity field is the actual velocity field during the incipient plastic flow, the rate at which energy is dissipated in plastic flow must equal the rate

$$T_f \int_{-a}^{a} \{v_y(x, 4a) - v_y(x, -4a)\}\, dx \tag{7}$$

at which the surface traction T_f does work. Greenberg, Drucker, and Prager [4] have shown that the value of T_f obtained by equating (6) and (7) for an *arbitrary* velocity field is an upper bound for the actual value of the surface traction for impending plastic flow.

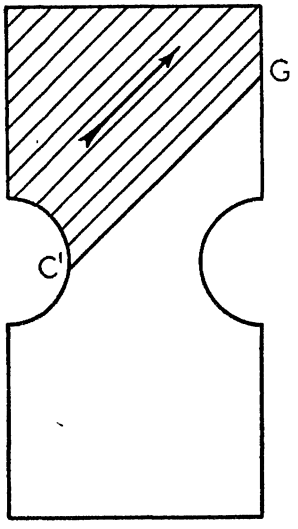

FIGURE 3

Fig. 3 shows a very simple discontinuous velocity field: the shaded portion moves as a rigid body in the direction indicated by the arrow, while the unshaded portion remains at rest. It is found that the upper bound for T_f obtained from this field assumes the minimum value $1.5k$ when the line of discontinuity $C'G$ is inclined under 45° against the x axis. Thus, even the very crude stress and velocity fields discussed in connection with Figs. 2 and 3 give the following bounds for the surface traction T_f:

$$1.26k \leqq T_f \leqq 1.50k. \tag{8}$$

By considering a more plausible velocity field, the author has been able to decrease the upper bound for T_f to $1.39k$. This value, together with the lower bound from (8), determine T_f to within about $\pm$ 5 per cent.

REFERENCES

1. D. N. DE G. ALLEN and SIR RICHARD SOUTHWELL, *Plastic straining in two-dimensional stress systems*, Philos. Trans. Roy. Soc. London Ser. A vol. 242 (1950) pp. 379–414.

2. L. A. GALIN, *On the existence of a solution of the elastic-plastic problem of torsion of prismatic bars* (Russian), Prikl. Mat. Mek. vol. 13 (1949) pp. 650–654.

3. ———, *Plane elastic-plastic problem: plastic zones in the vicinity of circular apertures* (Russian with English summary), Prikl. Mat. Mek. vol. 10 (1946) pp. 367–386.

4. H. J. GREENBERG, D. C. DRUCKER, and W. Prager, *The safety factor of an elastic-plastic body in plane strain*, Report A18-3 to Office of Naval Research, Brown University, 1950; to appear in Journal of Applied Mechanics.

5. J. A. JACOBS, *Relaxation methods applied to plastic flow*, The Philosophical Magazine (7) vol. 41 (1950) pp. 347–361, 458–467.

6. R. v. MISES, *Mechanik der festen Körper im plastisch deformablen Zustand*, Nachr. Ges. Wiss. Göttingen vol. 1913 (1913) pp. 582–592.

7. A. NADAI, *Plasticity, a mechanics of the plastic state of matter*, New York, McGraw-Hill, 1931.

8. O. S. PARASYUK, *An elastic-plastic problem with a non-biharmonic plastic state* (Russian), Prikl. Mat. Mek. vol. 13 (1948) pp. 367–370.

9. W. PRAGER, *Discontinuous solutions in the theory of plasticity*, Courant Anniversary Volume, New York, Interscience Publishers, 1948, pp. 289–299.

10. H. TRESCA, *Mémoire sur l'écoulement des corps solides*, Mémoires Prés. par Divers Savants vol. 18 (1868) pp. 733–799.

11. A. WINZER and G. F. CARRIER, *The interaction of discontinuity surfaces in plastic fields of stress*, Journal of Applied Mechanics vol. 15 (1948) pp. 261–264.

BROWN UNIVERSITY,
PROVIDENCE, R. I., U. S. A.

MATHEMATICAL THEORY OF WATER WAVES

J. J. STOKER

The mathematical theory of gravity waves in water begins with Laplace (1776), but the first complete derivation of the basic hydrodynamical theories was given by Lagrange during the period from 1781 to 1786. Lagrange developed not only the basic exact nonlinear theory, but also the two best known approximation theories: the linear theory which arises when the free surface amplitudes are considered to be sufficiently small and which leads to problems in potential theory with mixed boundary conditions; and the nonlinear shallow water theory in which the simplification arises through the assumption that the depth of the water is small compared with some significant horizontal dimension such as the wave length of the surface waves.

The general theory developed by Lagrange seems not to have been applied to concrete cases for some years. Indeed, even the problems posed by the linear theory seem to have been considered so difficult that the Paris Academy in 1815 proposed as its prize problem the problem of determining the motion of the water in a pond that results when a stone is tossed into it. This problem was solved by both Cauchy and Poisson by rather different methods on the basis of the approximate linear theory mentioned above.

After about 1840 interest in the field shifted from France to England, where a considerable number of the great English mathematical physicists of that day interested themselves in various problems concerning water waves. These include Stokes, Airy, Kelvin, and Rayleigh. Kelvin, for example, was the first to find a solution of the ship wave problem in its most primitive form in which the ship is considered to be a moving pressure point. Oversimplified though this model is, the difficulties in interpreting and discussing the integral representation of the solution are nevertheless quite formidable, and Kelvin was led to develop for this purpose the now widely used method of stationary phase—the prototype of the very important methods, such as the saddle point method, which are used to yield asymptotic developments for the solutions of many problems in wave propagation. Only recently [24][1] (1949) has the complete asymptotic development for the solution of Kelvin's ship wave problem (in which Kelvin's solution by stationary phase turns out to be the term of lowest order) been given.

Beginning about 1870 and continuing to about 1905 interest in water wave problems in France revived once more. The best known names associated with this development are those of de St. Venant, Boussinesq, and Poincaré. B. de St. Venant and Boussinesq represent a tendency which has been, and still continues to be, very strong in France; and that is the pursuit of what might be called

[1] Numbers in brackets refer to the bibliography at the end of the paper. It should be pointed out that this bibliography makes no claim to completeness, above all with regard to the older literature, since this latter task is well performed in the generally available book of Lamb.

mathematical hydraulics, or the study of flows and wave motions in rivers and other shallow open channels. The mathematical basis for this work is the second of the two approximate theories mentioned above, i.e., the nonlinear shallow water theory of Lagrange and extensions of it due to de St. Venant [4] and Boussinesq [1]. This theory has been used to study flood waves in rivers [2, 6], flood tides in estuaries, the development and propagation of bores, and regulation by dams in rivers to prevent floods. Poincaré in his *Mécanique céleste*, vol. 3 (1905) appears to have been the first to study the development of breakers from a smooth motion by using the theory of characteristics for a second order hyperbolic differential equation in the way that has become very familiar in gas dynamics in treating the development of a shock wave from a smooth flow. The work of Boussinesq [1] is very extensive and should be taken up once more and studied in detail, since it offers a wealth of problems that are equally interesting from the mathematical and the physical point of view.

After 1905 interest in the subject of water waves seems to have died out somewhat until the coming of World War II, with some notable exceptions. One such exception is the work of Levi-Cività [19] in 1925. Levi-Cività gave the first solution of the problem of the existence of irrotational waves of finite amplitude; he proved the existence of periodic progressing waves by proving that a series development of the solution with respect to the amplitude converges for sufficiently small amplitudes. The work of T. H. Havelock is another exception to the statement made above. Havelock's work has been largely concerned with the difficult and practically important problem of wave resistance to the motion of ships and its dependence on parameters which fix the shape of the hull of the ship. Such questions, and others in addition, have been treated by Havelock in a long series of papers published during the past forty years or more.

The onset of World War II brought with it a revival of interest in the subject of gravity waves, as it did with every other branch and sub-branch of science. One, but only one, reason for such an interest in water waves was caused by the unprecedented and large scale landing operations carried out on beaches in many parts of the world. The result of this interest has been a notable increase in our knowledge of water wave phenomena, in particular, from the side of the mathematical theory. For example, a number of problems involving waves of small amplitude in water of variable depth have been solved, in particular, the problem of progressing waves over a uniformly sloping beach has been completely solved [7, 11, 20, 22, 25, 27]. A number of problems involving the effect of barriers and obstacles on progressing waves have been solved [3, 8–10, 12, 17, 18, 21, 32, 33, 36]. The problem of breaking of waves, and the related problems of flood waves in channels and of roll waves moving down steep channels, have been studied extensively [2, 5, 6, 15, 16, 23, 26, 28–31]. These problems have been attacked on the basis of nonlinear shallow water theories. Some progress has been made in studying the motion of floating bodies, i.e., of cases in which the motion of a rigid body in water is determined by the pressure forces set up between it and the water [13]; in particular, the well-known criterion for

stability of a ship, which is based on the assumption that the motion of the water can be ignored, has been shown to be at least a sufficient condition for stability.

A considerable amount of attention has been given recently to the purely mathematical questions of existence and uniqueness of the solutions of water wave problems, which had previously been largely neglected even in the case of the linear problems. In treating the uniqueness questions for the linear problems, which are essentially problems requiring the determination of harmonic functions with a mixed boundary condition, a theorem of A. Weinstein [34] in 1927 furnished an important clue to the correct formulation of the conditions at ∞, i.e., that boundedness conditions alone suffice to fix the wave character of the motion at ∞. Weinstein's uniqueness theorem refers to the simplest case, i.e., that of two-dimensional progressing waves in water of uniform depth. Uniqueness theorems for progressing waves over sloping beaches [27, 35] and for waves in the presence of bounded fixed or floating obstacles have been given recently [3, 12–14, 18, 33].

All of the problems on water waves discussed here refer to a variety of partial differential equations: the Laplace equation with linear and nonlinear boundary conditions, the linear wave equation, nonlinear hyperbolic equations, equations of parabolic type; and, in the work of de St. Venant and Boussinesq, to nonlinear partial differential equations of third and higher order which are degenerate in type. It should therefore be clear that the subject of gravity waves in water has been, and continues to be, a happy hunting ground for problems in partial differential equations which are of equal interest from the mathematical and from the physical point of view.

Bibliography

1. J. Boussinesq, *Essai sur la théorie des eaux courantes*, Mémoirs par Divers Savants vol. 23 (1877).

2. A. Craya, *Calcul des régimes variables dans les canaux*, La Houille Blanche vol. 1 (1945).

3. W. R. Dean, *Note on waves on the surface of running water*, Proc. Cambridge Philos. Soc. vol. 43 (1947).

4. B. de St. Venant, *Théorie du mouvement non permanent des eaux*, C. R. Acad. Sci. Paris vol. 73 (1871).

5. R. F. Dressler, *Mathematical solution of the problem of roll waves in inclined open channels*, Communications on Pure and Applied Mathematics vol. 2 no. 2 (1949).

6. H. Favre, *Étude théorique et experimentale des ondes de translation dans les canaux découverts*, Paris, Dunad, 1935.

7. K. O. Friedrichs, *Water waves on a shallow sloping beach*, Communications on Pure and Applied Mathematics vol. 1 no. 2 (1948).

8. K. O. Friedrichs and H. Lewy, *The dock problem*, Communications on Pure and Applied Mathematics vol. 1 no. 2 (1948).

9. A. E. Heins, *Water waves over a channel of finite depth with a dock*, Amer. J. Math. vol. 70 (1948).

10. ———, *Water waves over a channel of finite depth*, Canadian Journal of Mathematics vol. 2 (1950).

11. E. ISAACSON, *Water waves over a sloping bottom*, Communications on Pure and Applied Mathematics vol. 3 no. 1 (1950).

12. F. JOHN, *Waves in the presence of an inclined barrier*, Communications on Pure and Applied Mathematics vol. 1 no. 2 (1948).

13. ———, *On the motion of floating bodies*, I, Communications on Pure and Applied Mathematics vol. 2 no. 1 (1949).

14. ———, *On the motion of floating bodies*, II: *Simple harmonic motions*, Communications on Pure and Applied Mathematics vol. 3 no. 1 (1950).

15. J. B. KELLER, *The solitary wave and periodic waves in shallow water*, Annals of the New York Academy of Science vol. 51 (1949).

16. G. H. KEULEGAN and G. W. PATTERSON, *Mathematical theory of irrotational translation waves*, Journal of Research, National Bureau of Standards U.S.A. vol. 24 (1940).

17. J. KREISEL, *Surface waves* I, II, III, Reports prepared for the British Admiralty, undated.

18. ———, *Surface waves*, Quarterly of Applied Mathematics vol. 7 (1949).

19. T. LEVI-CIVITÀ, *Détermination rigoureuse des ondes permanentes d'ampleur finie*, Math. Ann. vol. 93 (1925).

20. H. LEWY, *Water waves on sloping beaches*, Bull. Amer. Math. Soc. vol. 52 (1946).

21. S. C. LOWELL, *The propagation of waves in shallow water*, Communications on Pure and Applied Mathematics vol. 2 no. 3 (1949).

22. A. MICHE, *Mouvements ondulataires de la mer en profondeur constante ou décroissante*, Annales des Ponts et Chaussées, 1944.

23. W. H. MUNK, *The solitary wave theory and its application to surf problems*, Annals of the New York Academy of Science vol. 51 (1949).

24. A. S. PETERS, *A new treatment of the ship wave problem*, Communications on Pure and Applied Mathematics vol. 2 no. 2 (1949).

25. ———, *The effect of a floating mat on water waves*, Communications on Pure and Applied Mathematics vol. 3 no. 4 (1950).

26. R. RÉ, *Étude de lacher instantané d'une retenue d'eau dans un canal par la méthode graphique*, La Houille Blanche vol. 3 (1946).

27. J. J. STOKER, *Surface waves in water of variable depth*, Quarterly of Applied Mathematics vol. 5 (1947).

28. ———, *Formation of breakers and bores*, Communications on Pure and Applied Mathematics vol. 1, no. 1 (1948).

29. ———, *The breaking of waves in shallow water*, Annals of the New York Academy of Science vol. 51 (1949).

30. H. SVERDRUP and W. MUNK, *Breakers and surf*, Hydrographic Office, U. S. Navy Department, 1944.

31. H. A. THOMAS, *The hydraulics of flood movements in rivers*, Carnegie Institute of Technology, Pittsburgh, 1937.

32. F. URSELL, *The effect of a fixed vertical barrier on surface waves in deep water*, Proc. Cambridge Philos. Soc. vol. 43 (1947).

33. ———, *Surface waves on deep water in the presence of a submerged circular cylinder*, Proc. Cambridge Philos. Soc. vol. 46 (1949).

34. A. WEINSTEIN, *Sur un problème aux limites dans une bande indéfinie*, C. R. Acad. Sci. Paris vol. 184 (1927).

35. ———, *On surface waves*, Canadian Journal of Mathematics vol. 1 (1949).

36. M. WEITZ and J. B. KELLER, *Reflection of water waves from floating ice in water of finite depth*, Communications on Pure and Applied Mathematics vol. 3 no. 3 (1950).

NEW YORK UNIVERSITY,
NEW YORK, N. Y., U. S. A.

STATISTICAL MECHANICS

COMPREHENSIVE VIEW OF PREDICTION THEORY[1]

NORBERT WIENER

Some years ago a paper appeared[2] by Kolmogoroff in C. R. Acad. Sci. Paris on extrapolation and interpolation. From the point of view in which we are interested, the main contribution of this paper was a discussion of the greatest lower bound of the error of mean square prediction when applied to time series, and when the method of prediction was to be linear in the past. This paper led to a couple of papers in Russia with which Kolmogoroff's own name and that of M. Krein are jointly associated.

About a year after the first Kolmogoroff paper, the present author independently started a series of investigations in the same direction. His motivation was the problem of predicting the future position of an airplane on the basis of a general statistical knowledge of its mode of flight and of a more concrete knowledge of its immediate past. Thus there were from the beginning two points in which his emphasis was different from that of Kolmogoroff. Also at least one of these was actually covered in Kolmogoroff's work. The first difference is that while Kolmogoroff's explicit work is primarily concerned with the irreducible minimum error of prediction, the author's own work is concerned with the actual method of securing a prediction with this irreducible error, or at least prediction with error as near to this as we wish.

Next, Kolmogoroff's work is primarily concerned with a time consisting of discrete instants, whereas my work is concerned with the continuous time in which the flight of an airplane takes place. Associated with this is the fact that my work, unlike the explicitly published work of Kolmogoroff, concerns the instrumentation which is necessary to realize the theory of prediction in automatic apparatus for shooting ahead of an airplane. This engineering bias leads me to emphasize more than does Kolmogoroff the problem of prediction in terms of linear operators in the scale of frequency, rather than in similar operators on the scale of time.

This same engineering standpoint led me from the prediction problem to the related problem of filtering. In this problem a message, a noise, and their relations to one another are known statistically as well as the sum of the message and the noise, from minus infinity to a given point in time. The problem is to disentangle from this combination the message alone, or more generally, the message under lead or lag in time.

Since the harmonic analysis of a signal is not complete within any finite time, this separation must always involve the prediction of the future of the message

[1] This address was listed on the printed program under the title *The statistical mechanics in communication.*

[2] A. Kolmogoroff, *Interpolation und Extrapolation von stationären zufälligen Folgen*, Bull. Acad. Sci. URSS Sér. Math. vol. 5 (1941) pp. 3–14.

and noise, and so cannot be isolated from prediction theory in general. It turns out that the technique for the design of filters with a minimum mean square error is precisely parallel to that for the design of predictors.

In one point both Kolmogoroff and his school and I myself have developed the theory of prediction along similar lines. This point is that of multiple prediction, in which not merely one quantity varies with time, but a number of quantitites—even an infinite number of quantities. This is an extremely important problem for the meterorologists and in general for the statisticians. Until recently its theory has not been implemented by a corresponding communicable tool.

Both the Russian school and my own have published what is a fairly extensive literature concerning prediction. What is missing is a definitive paper to take up all the threads of the argument and to close them off in a single comprehensive discussion in which as many theorems as possible are given necessary and sufficient conditions. It is the purpose of the present paper to fill exactly this gap.

Let me first take up the simplest prediction problem, which is that concerning prediction in a simple discrete time series. All that follows this will consist in an extension or development of methods here given.

In the first place, we shall consider a time series to be, not a single sequence of of numbers, but a sequence of numbers with a parameter of distribution. Thus α appears merely for purposes of integration, and α may therefore be mapped on the segment of a line between 0 and 1, since the main properties of integration are independent of dimensionality. Let us then consider a function $f(\alpha)$ defined on (0, 1) and belonging to the Lebesgue parameter class L_2. Since the particular time series with which we are concerned are not attached to any particular point in time, let us introduce the transformation $T\alpha$, which preserves measure or probability, and which moves each instant of time into the next one ahead. Then $f(T^\nu\alpha)$ represents our time series. Here ν runs between $-\infty$ and ∞, if ν is positive, T^ν represents the transformation T iterated ν times. $T^{-\nu}\alpha$ represents the inverse transformation T^{-1} iterated ν times; and again ν is positive. We then shall see that $f(T^\nu\alpha)$, where ν runs over all negative and positive integral values including 0, is a time series in statistical equilibrium.

The functions $f(T^{-\nu}\alpha)$, where $\nu > 0$, are a denumerable set of functions of class L_2, and as such have a linear extension, containing all functions of L_2 which can be approximated in the L_2 sense by polynomials in the given functions. This set of functions will be known as the *past*. The *present* consists only of linear multiples of $f(\alpha)$ itself; while the *future* is the linear extension of all functions of the form $f(T^\nu\alpha)$, where ν is positive.

The problem of prediction is that of the projection of a function belonging to the present or future on the past. Every function belonging to L_2 consists of the sum of a function belonging to the past and a function orthogonal to every function belonging to the past. The first is called the projection on the past of a given function; whereas the mean square of the second with respect to α is the mean square error of prediction. In order to carry out this process, it is useful to orthogonalize the set of functions $f(T^{-\nu}\alpha)$.

Let us first take $f(T^{-1}\alpha)$. There are two cases possible. Either this is equivalent to 0 or it is not. If it is equivalent to zero, $f(\alpha)$ may be expressed in terms of $f(T^{-1}\alpha)$, and hence of the past. Similarly, $f(T\alpha)$ may be expressed in terms of $f(\alpha)$ and thus in terms of $f(T^{-1}\alpha)$, or again in terms of the past. Thus the whole present and future may be expressed in terms of the past and the prediction problem may be solved with 0 mean square error.

On the other hand, let us suppose that $f(T^{-1}\alpha)$ is not equivalent to 0. If it is not equivalent to zero, it can be normalized by the multiplication of an appropriate factor, so that the integral of the square of its modulus is one. Let us call this normalized function $g_1(\alpha)$. Now let us consider $f(T^{-2}\alpha)$. Either this is equivalent to a multiple of $f(T^{-1}\alpha)$; in which case a perfect prediction is possible; or it is not so equivalent. In the second case, we shall have the formula

$$f(T^{-2}\alpha) = g_1(\alpha)\int_0^1 f(T^{-2}\beta)\overline{g_1(\beta)}\,d\beta + \left(f(T^{-2}\alpha) - g_1(\alpha)\int_0^1 f(T^{-2}\beta)\overline{g_1(\beta)}\,d\beta\right)$$

where the term

$$f(T^{-2}\alpha) - g_1(\alpha)\int_0^1 f(T^{-2}\beta)\overline{g_1(\beta)}\,d\beta$$

is not equivalent to 0. This term again may be normalized; and we shall call the result of this normalization $g_2(\alpha)$. We repeat the process and express $f(T^{-3}\alpha)$ in terms of $g_1(\alpha)$ and $g_2(\alpha)$, with the remainder

$$f(T^{-3}\alpha) - g_1(\alpha)\int_0^1 f(T^{-3}\beta)\overline{g_1(\beta)}\,d\beta - g_2(\alpha)\int_0^1 f(T^{-3}\beta)\overline{g_2(\beta)}\,d\beta.$$

Either this remainder is equivalent to 0 or it is not. If it is equivalent to zero, we again can express $f(T^{-3}\alpha)$ in terms of its past; and if we cannot, we may introduce a third function $g_3(\alpha)$ by normalizing the remainder. This process can be continued until we have either orthogonalized the entire past of $f(\alpha)$ or until there are no more terms left to orthogonalize. If the process terminates at any stage, a perfect prediction is possible. In all other cases, we have a normal or orthogonal set $g_1(\alpha), g_2(\alpha), \cdots$, in terms of which we can express the past of $f(\alpha)$.

Now let us form the function $h(\alpha)$, in accordance with the formula

$$h(\alpha) = f(\alpha) - \sum_1^\infty g_n(\alpha)\int_0^1 f(\beta)\overline{g_n(\beta)}\,d\beta.$$

If this function $h(\alpha)$ should prove equivalent to 0, this means that $f(\alpha)$ can be expressed linearly in terms of its past, without any error whatever of prediction. If not, then the function $h(\alpha)$ can itself be normalized, and we obtain a function $H(\alpha)$, in accordance with the formula

$$H(\alpha) = \frac{h(\alpha)}{\left(\int_0^1 |\,h(\beta)\,|^2\,d\beta\right)^{1/2}}.$$

The function $H(\alpha)$ is linearly dependent on the present and past of f and orthogonal to that past. It is accordingly orthogonal to every function $H(T^{-\nu}\alpha)$; and since the transformation T is measure preserving, then the set $H(T^{\nu}\alpha)$, where ν varies from $-\infty$ to ∞, is a normal orthogonal set. There are now two cases possible.

Either $f(\alpha)$ can be represented in terms of this set according to the formula

$$f(\alpha) = \sum_0^\infty H(T^{-\nu}\alpha) \int_0^1 f(\beta)\overline{H(T^{-\nu}\beta)}\, d\beta,$$

or there is a remainder

$$f_2(\alpha) = f(\alpha) - \sum_0^\infty H(T^{-\nu}\alpha) \int_0^1 f(\beta)\overline{H(T^{-\nu}\beta)}\, d\beta$$

not equivalent to 0. In the second case we may write

$$f(\alpha) = f_1(\alpha) + f_2(\alpha)$$

where $f_1(\alpha)$ may be shown to generate the same function $H(\alpha)$ as does $f(\alpha)$. On the other hand, $f_2(\alpha)$ will be completely determined by its own past from any period of time back. In other words, $f_2(\alpha)$ will be linearly dependent on the set $f(T^{-\nu}\alpha), f(T^{-\nu-1}\alpha), \cdots$ no matter how large ν may be. The complete present, past, and future of $f_1(\alpha)$ is orthogonal to the complete present, past, and future of $f_2(\alpha)$. f_1 and f_2 have already been given in terms of f alone. We have thus reduced every case of the prediction problem to the perfectly predictable case on the one hand, and the case where the function $H(\alpha)$ exists and f can be expressed in terms of H and its past, on the other.

It is the second case with which we are chiefly concerned. Let us notice that

$$\int_0^1 f(T^{\nu}\alpha)\overline{f(\alpha)}\, d\alpha = \sum_{\substack{\mu>0 \\ \mu-\nu\geqq 0}} \int_0^1 f(\beta)H(T^{\nu-1}\beta)\, d\beta \int_0^1 \overline{f(\beta)}H(T^{-\mu}\beta)\, d\beta.$$

Let us also notice that

$$\sum_0^\infty \left| \int_0^1 f(\beta)H(T^{-n}\beta)\, d\beta \right|^2 < \infty.$$

Thus the sum of the squares of the moduli of the coefficients of

$$\sum_0^\infty e^{in\omega} \int_0^1 f(\beta)H(T^{-n}\beta)\, d\beta = \Psi(\omega)$$

converges and the function belongs to L_2. $\Phi(\omega)$, the square of the modulus of this, will belong to L, and will have the Fourier coefficients

$$\int_0^1 H(T^n\alpha)\overline{H(\alpha)}\, d\alpha.$$

This suggests that with suitable hypotheses, we may proceed directly from the auto-correlation coefficients

$$\int_0^1 H(T^n\alpha)\overline{H(\alpha)}\, d\alpha$$

and their related harmonic analysis function $\Phi(\omega)$ to the coefficients

$$\int_0^1 f(\beta)\overline{H(T^{-\nu}\beta)}\, d\beta,$$

by means of which we express $f(\alpha)$ in a series in terms of the functions $H(T^{-\nu}\alpha)$. The conditions under which this is possible may be proved to be

$$\int_{-\pi}^{\pi} |\log \Phi(\omega)|\, d\omega < \infty. \tag{1}$$

To see this, we express $(1/2)\log\Phi(\omega)$ by the corresponding series

$$\sum A_n e^{in\omega}.$$

We then form the corresponding series

$$\sum A_n \operatorname{sgn} n\, e^{in\omega},$$

which may be shown to determine a pure imaginary function by Cesàro summation. Let this function be $F(\omega)$. Then if we put

$$(\Phi(\omega))^{1/2} e^{F(\omega)} = \Psi(\omega),$$

we shall find that the Fourier series of $\Psi(\omega)$ will contain no negative frequencies.

Another function closely related to $\Psi(\omega)$ is

$$\Psi_1(z) = \frac{1}{2\pi}\int_0^{2\pi} \Psi(\omega)\, \frac{e^{i\omega}\, d\omega}{e^{i\omega} - z}.$$

It can be shown that as r tends to 0 from 1,

$$\lim_{r\to 1} \int_0^1 |\Psi(\omega) - \Psi_1(re^{i\omega})|^2\, d\omega = 0;$$

that $\Psi(\omega)$ is analytic within the unit circle; and what is more, it can be proved to have no zeros within the unit circle.

Let us now suppose that $\Phi(\omega)$ is any real function of class L whose logarithm fulfills our critical condition (1). It is then possible, by the use of ideas from Brownian motion theory, to give time series $f(\alpha)$ and a measure preserving function T such that

$$\Phi(\omega) \sim \sum_{-\infty}^{\infty} e^{i\nu\omega} \int_0^1 f(T^\nu\alpha)\overline{f(\alpha)}\, d\alpha.$$

It is also possible to prove that $f(\alpha)$ cannot be expressed completely in terms of its own past, and that it is orthogonal to its remote past, in the sense that the

projection of $f(\alpha)$ on the set $f_2(T^{\nu-1}\alpha)$, $f(T^{-\nu-2}\alpha)$, $\cdots$ tends to 0 as ν becomes infinite. We thus have a complete set of conditions for the factoring of $\Phi(\omega)$, and we have the basis for a theory of simple discrete prediction.

All the essential ideas in this theory of prediction may be extended to multiple prediction. Let us first take up the case of finite multiple prediction. Here we start with a set of functions $f_n(T^\nu\alpha)$, where n ranges from 1 to N. The past consists of the linear extension of the functions $f_n(T^\nu\alpha)$, where n ranges from 1 to N and ν is negative. T is, of course, as before, a measure preserving transformation. We now can proceed as before to determine the parts of $f_1(\alpha)$ to $f_n(\alpha)$ which are orthogonal to the complete past of all the f's. We carry this out by a procedure of orthogonalization exactly like that which we have already used. If this procedure of orthogonalization terminates before it gives N functions $H_n(\alpha)$ which are normal and orthogonal to one another and to the past, but linearly dependent on the past and present, then some one at least of the functions $f_n(\alpha)$ is completely determined by its own past and the past of the other functions. If that is not the case, we obtain a set of normal and orthogonal functions $H_n(T^\nu\alpha)$.

Either all the functions $f_k(\alpha)$ may be completely developed in terms of these, or there are remainders in the development which are not equivalent to 0. In the second case, just as in the corresponding simple case, we can separate a multiple time series into the sum of two multiple time series, such that the past, present, and future of one will be completely orthogonal to the past, present, and future of the other. One of these time series will be perfectly predictible, and the other will be expressible in terms of its own H functions. We now go through procedures exactly analogous to those through which we have gone in the simple case. Let us notice that

$$f_i(\alpha) \sim \sum_{\nu=0}^{\infty} \sum_{n=1}^{N} H_n(T^{-\nu}\alpha) \int_0^1 f_i(\beta)\overline{H_n(T^{-\nu}\beta)}\, d\beta. \tag{2}$$

Let us also notice that

$$\int_0^1 f_i(T^\nu\alpha) f_j(\alpha)\, d\alpha = \sum_{\substack{\mu>0 \\ \mu-\nu\geqq 0}} \sum_n \int_0^1 f_i(\beta)\overline{H_n(T^{\nu-\mu}\beta)}\, d\beta \int_0^1 \overline{f_j(\beta)} H_n(T^{-\mu}\beta)\, d\beta.$$

Then we obtain a matrix of functions

$$\Psi_{ij}(\omega) = \sum_{\nu=0}^{\infty} e^{i\nu\omega} \int_0^1 f_i(\beta)\overline{H_j(T^{-\nu}\beta)}\, d\beta,$$

belonging to L_2. We shall have

$$\mathbf{\Phi}(\omega) = \mathbf{\Psi}(\omega)\cdot\overline{\mathbf{\Psi}(\omega)},$$

where the matrix $\mathbf{\Phi}(\omega)$ has the Fourier coefficients

$$\int_0^1 f_i(T^\nu\beta)\overline{f_j(\beta)}\, d\beta.$$

As before, it is interesting to know what condition on the matrix

$$\Phi(\omega) \qquad (-\pi \leqq \omega \leqq \pi)$$

will make the existence of the set of functions $H_h(T\alpha)$ and their closure with respect to the corresponding f functions possible. Without going into proofs, it may be said that the condition is that $\Phi_{ij}(\omega)$ all belong to L_2, and that

$$\int_0^1 |\log|\text{Determinant } \Phi(\omega)\,||\, d\omega < \infty.$$

If this condition is fulfilled for an Hermitian matrix of positive definite Hermitian type, then this matrix may always be obtained as indicated from a set of functions $f_n(T^r\alpha)$.

In the simple prediction case, we have already given an algorithm which will enable us to carry out the computational work of the resolution of $\Phi(\omega)$ in the form

$$\Phi(\omega) = |\Psi(\omega)|^2.$$

The similar resolution of the matrix $\mathbf{\Phi}(\omega)$ in the form

$$\mathbf{\Phi}(\omega) = \mathbf{\Psi}(\omega)\cdot\overline{\mathbf{\Psi}(\omega)}$$

is complicated by the fact that matrix multiplication is not commutative, and therefore there is no easy use of the logarithm. However, there is a computational process which is not too difficult. This is a generalized form of the alternating process known in potential theory.

Let us suppose that we have two linear subspaces of Hilbert space, say S_1 and S_2, and that we have a vector in that space. This vector is to be projected on the smallest linear space containing these two subspaces S_1 and S_2. We then project the vector on S_1. The remainder we project on S_2. The remainder after this second projection we project on the space S_1, etc. We then add all the projections that we have obtained on the two spaces. It may be shown that this process is a convergent one, at least in the mean sense, which is the only relevant sense here, and that it ultimately yields the projection of the vector on the smallest linear extension of the two spaces.

For the moment let us consider only a prediction process of multiplicity 2. The problem which we are facing is to take some vector not necessarily exclusively belonging to the past and to project it on a past which represents the smallest linear extension of the space combining the past of one component with the past of the other. It is then possible to do this by a procedure of successive projections, which will turn out to have a computable algorithm.

While this method of carrying out the alternating process for purposes of prediction is available in the perfectly general multiple case, we shall illustrate it here in the case of multiplicity 2. Actually the best computational procedure for a case of higher multiplicity consists in a step by step use of a very similar

process, to embrace each time one new variable, after the process has already been completed for a number of variables.

To go to the case $n = 2$, let us start with two functions, $f_1(\alpha)$ and $f_2(\alpha)$. Using each of these functions alone we form the orthogonal set of functions $H_1(T^\nu\alpha)$ and $H_2(T^\nu\alpha)$ as before. We shall assume that f_1 can be completely expressed in terms of the orthogonal set belonging to H_1 and that f_2 can be completely expressed in terms of the orthogonal set belonging to H_2. This, as we have seen, is no great restriction for practical computation. Now we take $H_1(\alpha)$, and develop it in terms of the past of the orthogonal set belonging to the H_2's in the following form.

$$H_1(\alpha) = \sum_0^\infty H_2(T^{-\nu}\alpha) \int_0^1 H_1(\beta)\overline{H_2(T^{-\nu}\beta)}\, d\beta + r_1(\alpha).$$

We now develop $r_1(\alpha)$ in the form,

$$\begin{aligned} r_1(\alpha) &= \sum_0^\infty H_1(T^{-\nu}\alpha) \int_0^1 r_1(\beta)\overline{H_1(T^{-\nu}\beta)}\, d\beta + r_2(\alpha) \\ &= r_2(\alpha) - \sum_{\nu=0}^\infty H_1(T^{-\nu}\alpha) \sum \int_0^\infty H_2(T^{-\mu}\beta)\overline{H_1(T^{-\nu}\beta)}\, d\beta \\ &\qquad \times \int_0^\infty H_1(\beta)\overline{H_2(T^{-\mu}\beta)}\, d\beta. \end{aligned}$$

This process can be continued indefinitely, and $r_n(\alpha)$ will converge in the mean to a function orthogonal both to the past of $H_1(\alpha)$ and to that of $H_2(\alpha)$. What will be left apart from this remainder will be

$$\begin{aligned} \eta_1(\alpha) &= \sum_0^\infty H_2(T^{-\nu}\alpha) \int_0^1 H_1(\beta)\overline{H_2(T^{-\nu}\beta)}\, d\beta \\ &\quad - \sum_{\nu=0}^\infty \sum_{\mu=0}^\infty H_1(T^{-\nu}\alpha) \int_0^1 H_2(T^{-\mu}\beta)\overline{H_1(T^{-\nu}\beta)}\, d\beta \\ &\quad \times \int_0^1 H_1(\beta)\overline{H_2(T^{-\mu}\beta)}\, d\beta + \sum_{\nu=0}^\infty \sum_{\mu=0}^\infty \sum_{\lambda=0}^\infty H_2(T^{-\nu}\alpha) \int_0^1 H_1(T^{-\mu}\beta)\overline{H_2(T^{-\nu}\beta)}\, d\beta \\ &\quad \times \int_0^1 H_2(T^{-\lambda}\beta)\overline{H_1(T^{-\mu}\beta)}\, d\beta \int_0^1 H_1(\beta)\overline{H_2(T^{-\lambda}\beta)}\, d\beta - \cdots. \end{aligned}$$

It has been shown that this sum converges in the mean, and it is perfectly possible to show that it is expressible in terms of $H_1(\alpha)$ and the pasts of $f_1(\alpha)$ and $f_2(\alpha)$.

Similarly,

$$\begin{aligned} \eta_2(\alpha) &= \sum_0^\infty H_1(T^{-\nu}\alpha) \int_0^1 H_2(\beta)\overline{H_1(T^{-\nu}\beta)}\, d\beta \\ &\quad - \sum_{\nu=0}^\infty \sum_{\mu=0}^\infty H_2(T^{-\nu}\alpha) \int_0^1 H_1(T^{-\mu}\beta)\overline{H_2(T^{-\nu}\beta)}\, d\beta \int_0^1 H_2(\beta)\overline{H_1(T^{-\mu}\beta)}\, d\beta + \cdots \end{aligned}$$

converges in the mean, and may be expressed in terms of $H_2(\alpha)$ and the past of the f's. If $\eta_1(\alpha)$ and $\eta_2(\alpha)$ are not linearly independent, then either $H_1(\alpha)$ may be expressed in terms of $H_2(\alpha)$ and its past, or $H_2(\alpha)$ may be expressed in terms of $H_1(\alpha)$ and its past. In other words, only one of the two variables $H_1(\alpha)$ and $H_2(\alpha)$ really occurs in the fundamental prediction problem. Otherwise, we may orthogonalize $\eta_1(\alpha)$ and $\eta_2(\alpha)$ by the formulae:

$$X_1(\alpha) = \frac{\eta_1(\alpha)}{\left(\int_0^1 |\eta_1(\beta)|^2 d\beta\right)^{1/2}};$$

$$X_2(\alpha) = \frac{\eta_2(\alpha) - X_1(\alpha)\int_0^1 \eta_2(\beta)\overline{X_1(\beta)}\, d\beta}{\left(\int_0^1 |\eta_2(\beta)|^2 d\beta - \left|\int_0^1 \eta_2(\beta)\overline{X_1(\beta)}\, d\beta\right|^2\right)^{1/2}}.$$

We shall then have as a normal and orthogonal set:

$$X_1(T^\nu\alpha); \qquad X_2(T^\nu\alpha).$$

These will take the place of the $H_1(T^\nu\alpha)$ and $H_2(T^\nu\alpha)$ of formula (2).

In the problem of continuous prediction, we are now up against the fact that the set of orthogonal functions $H(T^\nu\alpha)$ which occurs in the problem of discrete prediction, and the similar set which occurs in the set of discrete multiple prediction, are replaced by functions which are no longer of the Lebesgue class L_2. This is not a finally forbidding difficulty, as it is possible to introduce the Hilbert theory of spectra to take the place of a theory of orthogonal functions. Still, the theory of spectra is much more detailed and inconvenient than that of orthogonal functions, and we must consider ourselves fortunate that there is a method to avoid introducing it directly. This depends on the fact that in the prediction theory which we have already developed, which makes use of measure-preserving point transformations, we may completely replace these measure-preserving point transformations by a unitary functional transformation. That is, wherever $f(T^\nu\alpha)$ appears, we may introduce an expression $T^\nu f(\alpha)$, where $T^\nu f(\alpha)$ is a linear transformation of Hilbert space into itself, and preserves all lengths and distances in Hilbert space.

Now, although there is a continuous group of measure-preserving functional transformations which plays the same role in continuous prediction theory that the discrete group of powers of a single measure-preserving transformation does in discrete prediction theory, and it is impossible to map out a continuous group on any such discrete group, there is a discrete group of functional transformations whose future is the same as the future of the continuous group. In order to introduce these functional transformations, let me introduce the Laguerre polynomials.

If I consider the expression $e^{i\omega\nu}$ which occurs in prediction theory, there is closely related to it the expression

$$\left(\frac{1 + i\vartheta/2}{1 - i\vartheta/2}\right)^\nu.$$

Similarly, if I replace $e^{i\omega\nu}(d\omega)^{1/2}$ by

$$\frac{(1 + i\vartheta/2)^n}{(1 - i\vartheta/2)^{n+1}} (d\vartheta)^{1/2},$$

I shall have found a way to transform the interval from $-\pi$ to π into the interval from $-\infty$ to ∞, if only I put $\vartheta = 2 \tan (\omega/2)$. If now I examine the functions

$$\frac{1}{(2\pi)^{1/2}} \frac{(1 + i\vartheta/2)^n}{(1 - i\vartheta/2)^{n+1}},$$

they will clearly prove to be a normal and orthogonal set if n runs through all integral values from $-\infty$ to ∞. We shall have

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{\left(1 + \frac{i\vartheta}{2}\right)^n}{\left(1 - \frac{i\vartheta}{2}\right)^{n+1}} e^{-i\vartheta t}\, d\vartheta$$

$$= \int_{-\infty}^{\infty} \sum_{0}^{n-1} A_k \frac{1}{\left(1 - \frac{i\vartheta}{2}\right)^{k+1}} e^{-i\vartheta t}\, d\vartheta \begin{cases} = \sum_{0}^{n} B t^k e^{-2t} \\ = 0 \ (t < 0), \end{cases} = p(t)e^{-2t} \ (t > 0)$$

where $p(t)$ is an appropriate polynomial of the nth degree. Similarly if n is 0 or negative, we shall have

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{\left(1 + \frac{i\vartheta}{2}\right)^n}{\left(1 - \frac{i\vartheta}{2}\right)^{n+1}} e^{-i\vartheta t} d\vartheta$$

$$= \int_{-\infty}^{\infty} \sum_{0}^{n-1} A_k \frac{1}{\left(1 - \frac{i\vartheta}{2}\right)^{k+1}} e^{-i\vartheta t}\, d\vartheta \begin{cases} = \sum_{0}^{-n} C_k t^k e^{2t} \\ = 0 \ (t > 0). \end{cases} = g(t)e^{2t} \ (t < 0)$$

In other words the Fourier transforms of

$$\frac{\left(1 + \frac{i\vartheta}{2}\right)^n}{\left(1 - \frac{i\vartheta}{2}\right)^{n+1}} \qquad (n > 0)$$

differ from 0 only on the positive half line, and those of

$$\frac{\left(1 + \frac{i\vartheta}{2}\right)^n}{\left(1 - \frac{i\vartheta}{2}\right)^{n+1}} \qquad (n \leqq 0)$$

differ from 0 only on the negative half line. The functions

$$p_n(t)e^{-2t} \ (t > 0); \qquad 0 \ (t < 0)$$

$$g_n(t)e^{2t} \ (t < 0); \qquad 0 \ (t > 0)$$

are known, with proper normalization with respect to dependent and independent variables, as the Laguerre functions. The transformation of each Laguerre function into the next later function is a unitary transformation whose powers, positive and negative, constitute a discrete group of all translations along the line. I repeat that up to the present we have used in our prediction theory no properties of our transformations which involve their being used as point transformations rather than general unitary valuations. Thus on the frequency scale, the change from $e^{in\omega}$ to

$$\frac{\left(1 + \frac{i\vartheta}{2}\right)^n}{\left(1 - \frac{i\vartheta}{2}\right)^n}$$

is one which involves no difficulty.

In view of this transformation we are now in a position to factor functions of frequency running from $-\infty$ to ∞, whether they are scalar or matrix functions, into a product in which one term is the transform of functions vanishing only for past time and the other is the transform of functions vanishing only for future time. The previous conditions of factorizability

$$\Phi(\omega) \in L \qquad (-\pi \leqq \omega \leqq \pi)$$

and

$$\int_{-\pi}^{\pi} |\log \Phi(\omega)| \, d\omega < \infty$$

are clearly replaced by the equivalent conditions

$$\Phi^*(\varphi) \in L, \qquad -\infty < \varphi < \infty$$

and

$$\int_{-\infty}^{\infty} \frac{1}{1 + \vartheta^2} |\log \Phi^*(\vartheta)| \, d\vartheta < \infty,$$

if only

$$\Phi(\omega) = \Phi(\vartheta);$$

$$\omega = 2 \tan^{-1} \frac{\vartheta}{2}.$$

Similar results and equivalences hold in the case of matrix factorization. If

$$\Psi^*(\vartheta) = \Psi(\omega),$$

we have

$$\Phi^*(\vartheta) = |\Psi^*(\omega)|^2,$$

and the Fourier transforms of the L_2 functions $\Psi^*(\vartheta)$ will contain no negative frequencies.

We now come to the actual mechanism of prediction. We have not time to take this up in more than the simplest case; namely, that of a one-dimensional discrete prediction, but the methods are valid with the most obvious changes both for the continuous and the multiple case. Let then

$$\sum A_n f(T^{-n}\alpha)$$

be any polynomial in the past which we desire to study as an approximation to $f(T^\nu\alpha)$. The mean square error of prediction will then be

$$\int_0^1 |\sum A_n f(T^{-n}\alpha) - f(T^\nu\alpha)|^2 d\alpha.$$

As we now wish to write this in terms of frequency rather than time, it will become

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} |\sum A_n \Psi(\omega) e^{in\omega} - \Psi(\omega)e^{-i\nu\omega}|^2 d\omega = \sum_{m=-\infty}^{\infty} \left|\sum_n \psi_{m-n-\nu} - \psi_m\right|^2.$$

It will then be seen that we have to reduce at the same time

$$\sum_n A_n \psi_{m-n-\nu}$$

as near to 0 as we can for $m < \nu$ and as near to ψ_m as we can for $n \geqq \nu$. If we have at our disposal not merely polynomials, but arbitrary combintations of the past, this will give us

$$\sum_{m=0}^{\infty} e^{im}(\sum_n A_n \psi_{m-n-\nu}) = \sum_p^{\infty} \psi_n e^{in\omega}$$

or, where ψ_m represents the Fourier coefficients of $\Psi(\omega)$,

$$\sum_0^{\infty} A_n e^{in\omega} = e^{i\nu\omega} \frac{\sum_p^{\infty} \psi_n e^{in\omega}}{\sum_0^{\infty} \psi_n e^{in\omega}}.$$

However, it is not difficult to prove that even if we have only polynomials at our disposal, we may reduce

$$\int_0^1 |\sum A_n f(T^{-n}\alpha) - f(T^\nu\alpha)|^2 d\alpha = \sum_0^p |A_n|^2$$

as near its absolute minimum as we wish.

We thus have solved the problem of prediction as nearly as we wish to optimum prediction in the case where the prediction is not perfect. Where the least mean

square error of prediction is 0, while there is in general no single optimum prediction, it is possible to approximate as nearly as we wish to perfect prediction by the following means:

In the first place, we blur the spectrum of the function which we wish to predict by taking its convolution with a narrow Gaussian distribution of unit area. Then we obtain the optimum prediction on the strength of this blurred spectrum. It may be shown that as the Gaussian distribution gets narrower and narrower the optimum prediction becomes more and more perfect.

To return to the case where there is not a perfect prediction, and in which the function

$$\sum_0^\infty e^{in\omega} \int_0^1 f(T^\nu\alpha)\overline{f(\alpha)}\, d\alpha,$$

can be factored. Let us note that formally, from the point of view of frequency rather than time, the optimum prediction operator for a lead μ amounts to multiplication by

$$\frac{1}{2\pi}\sum_{\nu=0}^\infty e^{-i\nu\omega}\int_{-\pi}^{\pi}\Psi(u)e^{-i\nu u}\, e^{i\mu u}\, du,$$

and that the mean square error of prediction is

$$\sum_{\nu=0}^{\mu}\left|\frac{1}{2\pi}\int_{-\pi}^{\pi}\Psi(u)e^{-i\nu u}\, du\right|^2.$$

From the original Kolmororoff point of view, this frequency treatment is not important, but from our point of view the frequency representation of operators is important just because it is the standard representation of alternating current engineering. It is in this form that we strive to realize operators through a network of coils, resistances, and condensers, and in fact the prediction operators which we have just obtained are very generally suitable for realization. This leads us to the problem of filtering.

I shall describe this in terms of the continuous case rather than the discrete, because filtering is commonly an electrical engineering operation, although indeed its precise analogue is useful in the statistical laboratory.

We start then with a message $f_1(\alpha)$ and a noise $f_2(\alpha)$ and we put formally

$$\phi_{11}(t) = \int_0^1 f_1(T^t\alpha)\,\overline{f_1(\alpha)}\, d\alpha; \qquad f(\alpha) = f_1(\alpha) + f_2(\alpha);$$

$$q(t) = \int_0^1 f_1(T^t\alpha)\,\overline{f(\alpha)}\, d\alpha; \qquad \phi(t) = \int_0^1 f(T^t\alpha)\,\overline{f(\alpha)}\, d\alpha;$$

$$\Phi_{11}(\omega) \sim \int_{-\infty}^{\infty}\phi_{11}(t)e^{it\omega}\, dt; \qquad \Phi(\omega) \sim \int_{-\infty}^{\infty}\phi(t)e^{i\omega t}\, dt;$$

$$k(\omega) \sim \int_0^\infty K(t)e^{i\omega t}\, dt; \qquad \Phi(\omega) = |\Psi(\omega)|^2$$

as before. The problem which we have is to minimize

$$\int_0^1 | f_1(T^{-\lambda}\alpha) - \int_0^\infty f(T^{-\tau}\alpha)K(\tau)\, d\tau\, |^2\, d\alpha.$$

Formally, this leads to the process of minimizing

$$\phi_{11}(0) - 2\mathrm{Re}\int_0^\infty q(\tau - \lambda)K(\tau)\, d\tau + \int_0^\infty K(\tau)\, d\tau \int_0^\infty \overline{K(\sigma)}\phi(\sigma - \tau)\, d\sigma$$

$$= \int_{-\infty}^\infty \left| \frac{\Phi(\omega)e^{-i\lambda\omega}}{\overline{\Psi(\omega)}} - k(\omega)\Psi(\omega) \right|^2 + \text{const.}$$

Thus again, formally, the optimum prediction is given by the frequency operator

$$k(\omega) = \frac{1}{2\pi\Psi(\omega)} \int_0^\infty e^{-i\omega t}\, dt \int_{-\infty}^\infty e^{itu} \frac{\Psi_1(u)}{\overline{\Psi(u)}} e^{-i\lambda u}\, du.$$

Let us notice that the technique of prediction may be carried over to multiple time series. From an engineering point of view this means that we have a number of messages linearly jumbled, but we can put them through an apparatus so that each message will come out of it in as pure a form as possible.

As I have said before this allows us to use interference to eliminate a message as well as simple attenuation available in the ordinary filter.

There are a number of other topics for which I have no more time available than enough simply to mention them. In the first place, the methods of multiple prediction make it possible to analyze the direction of causality in complicated situations. In the second place, the whole theory of prediction as given up to this point involves a perfect knowledge of statistical parameters of the past. This knowledge is in fact never available. It must be supplemented by some theory of extenuation from which we can obtain not merely the most probably values of our spectra, but also our distribution. We have made some headway in the problem of extenuation of parameters in the case where time series represents the impact on a resonator of a large number of randomly distributed phenomena. However, not even in this case have we brought the estimated theory to a point where it is yet suitable for practical computation, and secondly this is by no means the only significant case of linear time series.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY,
CAMBRIDGE, MASS., U. S. A.

SOME RECENT TRENDS IN THE MATHEMATICAL THEORY OF DIFFUSION

WILLIAM FELLER

1. Introduction. This lecture may seem to sail under a false flag since it appears under the general heading of Applied Mathematics. Only a small part of the mathematical diffusion theory is connected with immediate applications, and much of the recent literature on the subject has an abstract character. The situation is here as in the theory of harmonic functions which likewise stems from practical problems and has many applications, but which has spread into and influenced many fields: complex variable theory, boundary value problems for general differential equations, subharmonic functions, generalized potentials, etc.

Similarly, the diffusion theory has outgrown its origin as a special topic in partial differential equations. Nevertheless, not even this particular chapter is closed: recent applications to the theory of genetic evolution lead us to singular diffusion equations and confront us with a new type of boundary problems. We find in this connection unsolved problems, but of even greater interest is the fact that in diffusion theory we find for the first time an intimate interplay between differential equations and measure theory in function spaces. The latter throws new light even on the classical parts of diffusion theory; it leads to new types of solutions, and opens a new avenue of attack on problems connected with boundary conditions which were left open for a long time. It turns out that the adjoint of a differential equation is not necessarily itself a differential equation, and this new information is of value for the general theory of semigroups of operators. This is underlined by the fact that the probability approach leads us to consider diffusion equations which are not of local character and which can be treated by the methods developed, under different forms, by Bochner, M. Riesz, and L. Schwartz. In this way the theory of differential equations is tied in with other theories and we shall find unexpected connections between harmonic functions and diffusion processes.

It is the purpose of the present address to outline some of the new results and open problems. It is believed that the new methods promise to be fruitful also in other fields, in particular for other types of differential equations. [*Added in proof*: Since this was written, new results were obtained and some of the problems are no longer open; cf. [8].]

2. Classical diffusion and the Wiener space. For a first orientation consider the classical diffusion (or heat-conduction) equation

$$u_t = u_{xx} \tag{2.1}$$

over the interval $-\infty < x < \infty$. Its fundamental solution (or Green function) is given by the normal density function

$$u(y; t, x) = \frac{1}{2(\pi t)^{1/2}} \exp\{-(x - y)^2/4t\}. \tag{2.2}$$

Under very general conditions on $f(x)$ the function

$$u(t, x) = \int_{-\infty}^{+\infty} f(y)u(y; t, x)\, dy \tag{2.3}$$

represents the unique solution of (2.1) which, for fixed $t > 0$, is integrable with respect to x and which tends to $f(x)$ as $t \to 0$.

Now equation (2.1) is supposed to be connected with Brownian movement, the description of which involves a function space. In fact, the position of a Brownian particle is a function $X(t)$, and "observing a particle" means observing $X(t)$. Thus each $X(t)$ represents a sample point in our experiment, and the problem is to describe the properties of the possible paths $X(t)$ in probabilistic terms. This was first done by N. Wiener [25]. He starts from the assumption that (2.2) defines the transition probability density of our process, that is, the conditional probability density of the relation $X(t_0 + t) = x$ if it is known that $X(t_0) = y$; it is furthermore assumed that the increments $\Delta X(t) = X(t + h) - X(t)$ over nonoverlapping time intervals are statistically independent (or that the corresponding probabilities multiply). Wiener (and by other methods P. Lévy [20]) shows that with probability one the path function is continuous but of unbounded variation in every t-interval; the set of t-values for which $X(t) = 0$ has the structure of a Cantor set; with probability one $X(t)$ satisfies a Lipschitz condition of order $1/2$ but no Lipschitz condition of order $1/2 + \epsilon$ (more precise results are given by the local law of the iterated logarithm).

Obviously statements of this type describe the hypothetic diffusion process more directly than does the differential equation (2.1). It is less obvious that even the formal theory of (2.1) can profit from the function space approach. Consider, for example, the boundary value problem for (2.1) for a region bounded by the part $x > a$ of the x-axis and a curve $x = \varphi(t)$ with $\varphi(0) = a$. We are here confronted with a situation similar to that in the Dirichlet problem for $\Delta u = 0$. In general, an essentially unique solution exists, but it does not assume the prescribed boundary values at all points of the curve. As Fortet [10] noticed, the question whether a particular point x_0 of the boundary curve is an exceptional point can be solved by a comparision of the local properties of the boundary with those of the path curves $X(t)$. A final criterion is obtained from the local law of the iterated logarithm. Thus the measure-theoretical approach leads on one hand to an existence theorem which is more general than those obtained by purely analytical methods, and on the other hand we obtain a regularity criterion which admits of a direct interpretation in terms of path functions.

We want to show that even the simpler problem of diffusion in a finite fixed interval is intimately related to the theory of path functions and apparently cannot be treated in a satisfactory manner without the function space approach. It is well known that in the case of diffusion in a finite interval the solution is not determined by the initial values alone, but that different boundary conditions

can be imposed. Perhaps the most familiar boundary condition is that of an *absorbing barrier*, which is usually defined formally by the condition $u(t, x_0) = 0$. It must be emphasized that this condition is obtained only heuristically by a passage to the limit from a discrete random walk and is justified by the results. This could be considered satisfactory were it not that the heuristic method completely breaks down in other cases. In fact, a perusal of the literature shows that even in comparatively simple physical problems the appropriate boundary conditions remain unknown. We shall see later on that in a certain case the absorbing barrier is completely described by the condition $u(t, x_0) < \infty$, but again this was found only from the properties of the solutions to which this condition leads. In other cases we shall see that no boundary conditions can be imposed. In short, the appropriate analytical formulation of various boundary conditions is an unsolved problem. Apparently it is unavoidable to revert to the original definition of the various conditions, and this definition usually refers directly to the path functions. Thus, diffusion with absorbing barriers can be described by saying that the laws of free diffusion prevail in the interior but that the process stops when the particle for the first time reaches a boundary, that is, when $X(t)$ equals x_0 or x_1.

3. The adjoint diffusion equations. Consider now a more general diffusion process in the infinite line with transition probability $u(y; t, x)$. By this we mean that if at any time t_0 the position of the particle is $X(t_0) = y$, then

$$\Pr\{a < X(t + t_0) < b\} = \int_a^b u(y; t, x)\, dx. \tag{3.1}$$

If the initial position $X(0)$ of the particle is a random variable with probability density $f(x)$, then the probability density of $X(t)$ is given by the integral (2.3).

A general stochastic process cannot be described solely in terms of the initial distribution and the transition probabilities. A diffusion process, however, is of the Markov type which means, very roughly, that the future depends statistically on the present state, but not on the past history which led to it.[1] This implies in particular the fundamental identity

$$u(y; s + t, x) = \int_{-\infty}^{+\infty} u(y; s, \xi)u(\xi; t, x)\, d\xi, \tag{3.2}$$

known as the *Chapman-Kolmogorov equation*. By virtue of the Markov property the transition probability $u(y; t, x)$ determines all probability relation in the space of path functions $X(t)$. In particular, the joint probability density of $(X(t_1), X(t_2), \cdots, X(t_n))$ for $0 < t_1 < t_2 < \cdots < t_n$ is given by

$$u(t_1, x_1)u(x_1; t_2 - t_1, x_2) \cdots u(x_{n-1}; t_n - t_{n-1}, x_n), \tag{3.3}$$

where $u(t, x)$ is defined by (2.3). Thus the properties of the path functions $X(t)$ are implicitly given by $u(y; t, x)$, and could be derived exactly as Wiener ob-

[1] Cf. the discussion in §8.

tained the properties of the paths in Brownian motion from (2.2). Actually, Wiener's results have been only partially generalized; for example, general conditions which guarantee continuity of almost all path functions seem to be unknown.

Among all stochastic processes for which the Chapman-Kolmogorov equation (3.2) holds, the diffusion processes are characterized by the requirement that the probability of a change exceeding ϵ during a short time interval Δt is small as compared to Δt; more precisely, we assume[2] that

$$\frac{1}{\Delta t}\int_{|x-y|>\epsilon} u(y;\, \Delta t,\, x)\, dx \to 0 \tag{3.4}$$

for every fixed y.

If at some time t the position is $X(t) = y$, then the mean and the variance of the displacement $\Delta X(t)$ during the following time interval of duration Δt are

$$\int_{-\infty}^{+\infty} (x - y)u(y;\, \Delta t,\, x)\, dx \qquad \text{and} \qquad \int_{-\infty}^{+\infty} (x - y)^2 u(y;\, \Delta t,\, x)\, dx. \tag{3.5}$$

Actually these integrals may diverge,[3] and we introduce therefore the truncated moments

$$\int_{|x-y|<\epsilon} (x - y)u(y;\, \Delta t,\, x)\, dx = a(\Delta t;\, y), \tag{3.6}$$

$$\int_{|x-y|<\epsilon} (x - y)^2 u(y;\, \Delta t,\, x)\, dx = 2b(\Delta t;\, y). \tag{3.7}$$

In view of (3.4) the asymptotic behavior of $a(\Delta t;\, y)$ and $b(\Delta t;\, y)$ as $t \to 0$ is independent of ϵ, and the physical significance of these quantities is essentially the same as that of the moments (3.5), except that less weight is attributed to large displacements. It is therefore natural to suppose that the limits

$$a(y) = \lim \frac{a(\Delta t;\, y)}{\Delta t}, \qquad b(y) = \lim \frac{b(\Delta t;\, y)}{\Delta t} \tag{3.8}$$

exist.[4]

[2] This definition was given in [5] where diffusion processes are called Markov processes of the "purely continuous" type. It seems to be an open problem whether (3.4) implies that the path functions $X(t)$ are continuous with probability one.

[3] An (unpublished) example is given by the density $u(y;\, t,\, x) = \{2\pi^{1/2}t^{1/2}\}^{-1}\phi'(x)$ $\exp\{-[\phi(x) - \phi(y)]^2/4t\}$ where $\phi(x) = \log^{1/4}(100 + x/(1 + e^{-x}))$. This is a solution of the equations (3.12)–(3.13) with $b(x) = \{\phi'(x)\}^{-2}$, $a(x) = (1/2)\, b'(x)$.

[4] It will be seen from the following derivation (which follows [5]) that our assumptions actually imply the existence of the partial derivative $u_t(y;\, t,\, x)$. Conversely, if it is assumed that u is differentiable and that (3.4) holds, then it suffices to assume that *one* of the ratios occurring in (3.8) has a finite point of accumulation as $\Delta t \to 0$. Letting then Δt approach zero through an appropriate sequence of values it follows from (3.11) that for this particular sequence also the other limit in (3.8) exists, and this leads to (3.12). The first derivation of (3.12) is due to Kolmogorov [17] who, however, uses strong uniformity conditions and assumes the existence of the moments (3.5).

It is then easy to derive a differential equation for $u(y; t, x)$ as a function of t and y. From (3.2) and (3.4) we have

$$u(y; t + \Delta t, x) = \int_{-\infty}^{+\infty} u(y; \Delta t, \xi)u(\xi; t, x)\, d\xi \tag{3.9}$$
$$= \int_{|\xi - y| < \epsilon} u(y; \Delta t, \xi)u(\xi; t, x)\, d\xi + o(\Delta t),$$

provided $u(\xi; t, x)$ is uniformly bounded for every fixed $t > 0$. If we now assume that for fixed t, x the function $u(\xi; t, x)$ has two continuous derivatives with respect to ξ, then the Taylor formula leads to

$$u(y; t + \Delta t, x) = u(y; t, x) \int_{|\xi - y| < \epsilon} u(y; \Delta t, \xi)\, d\xi$$
$$+ u_y(y; t, x) \int_{|\xi - y| < \epsilon} (\xi - y)u(y; \Delta t, \xi)\, d\xi \tag{3.10}$$
$$+ \frac{1}{2} u_{yy}(y; t, x) \int_{|\xi - y| < \epsilon} (\xi - y)^2 u(y; \Delta t, \xi)\, d\xi + o(\Delta t).$$

Now the integral of $u(y; \Delta t, \xi)$ should be identically unity.[5] Using, then, (3.4) we can write (3.10) as

$$u(y; t + \Delta t, x) - u(y; t, x)$$
$$= b(\Delta t; y)u_{yy}(y; t, x) + a(\Delta t; y)u_y(y; t, x) + o(\Delta t). \tag{3.11}$$

Hence, as a function of t and y, *the transition probability* $u(y; t, x)$ *satisfies the first (or backward) diffusion equation*

$$z_t(t, y) = b(y)z_{yy}(t, y) + a(y)z_y(t, y). \tag{3.12}$$

The obvious initial conditions on transition probabilities show that $u(y; t, x)$ is, in fact, a fundamental solution of (3.12). One should therefore expect that for fixed y the function $u(y; t, x)$ should satisfy the adjoint equation

$$w_t(t, x) = \{b(x)w(t, x)\}_{xx} - \{a(x)w(t, x)\}_x, \tag{3.13}$$

known as *second (or forward) diffusion equation*, and also as *Fokker-Planck equation.*[6]

We shall see (in §5) that this surmise is true in general, but that there exist cases where $u(y; t, x)$ satisfies only the first diffusion equation; instead of (3.13) we get in these cases an inequality, and a positive operator is added to the right side. This phenomenon is connected with the existence of a new type of solutions which has not been encountered before. It changes our concept of adjoint operators and throws new light on the theory of semigroups in general.

[5] It is possible to relax this condition so as to take account of more general situations.

[6] It is possible to introduce additional terms in (3.12) and (3.13) which will account for the possibility of particles disappearing or being created.

4. Singular equations. The theory of the preceding section applies also if the infinite x-axis is replaced by a finite or semi-infinite interval, except that now boundary conditions have to be specified. As has been mentioned before, the derivation of the familiar boundary conditions (such as absorbing, reflecting, and elastic barriers) is essentially of a heuristic nature, and it would be desirable to derive them from the basic assumptions in the same way as the diffusion equation itself has been obtained. We want now to discuss some new aspects of the boundary problems for a type of equations which is of both theoretical and practical interest.

In physical applications the coefficient $b(y)$ is essentially positive. Now modern diffusion theory has found a new and interesting field of applications in biology, in particular in the mathematical theory of evolution (cf. [6]). Here the particle under diffusion is to be interpreted symbolically and stands for a population size, gene frequency, etc. For example, the frequency of a particular gene in a population is a random variable $X(t)$ which by definition is restricted to the interval $0 < x < 1$, and Sewall Wright's theory of evolution assumes essentially that $X(t)$ is subject to a diffusion process with

$$b(x) = \beta x(1 - x), \qquad a(x) = -\gamma x + \delta(1 - x) \tag{4.1}$$

where β, γ, δ are positive constants. Similarly, in the simplest growth process $X(t)$ stands for a population size and can assume all positive values; the underlying stochastic process is described by our diffusion equations with

$$b(x) = \beta x, \qquad a(x) = \alpha x + \gamma. \tag{4.2}$$

The fact that $b(x)$ vanishes on the boundaries represents a singularity and has important and curious consequences.

Consider, for example, the diffusion equations with the coefficients (4.2), which are discussed in [7]. If $\gamma \leqq 0$, then there exists a unique solution so that no boundary conditions can be imposed. For this solution the probability

$$\int_0^\infty u(y; t, x)\, dx = \pi(t, y) \tag{4.3}$$

decreases steadily in time. The difference $1 - \pi(t, y)$ represents the probability that a population of initial size y dies out before time t. (This is the absorption or extinction probability.)

The situation changes radically when $0 < \gamma < \beta$, and resembles the more familiar circumstances of ordinary diffusion. There exist infinitely many solutions, among which there is one for which $\pi(t, y) \equiv 1$; this corresponds to reflecting barriers. For all other solutions $\pi(t, y)$ decreases, and the solution for which the rate of decrease is greatest obviously corresponds to absorbing barriers. It is the only solution for which $u(y; t, 0) < \infty$, so that in this particular case an absorbing barrier is described by the boundary condition $u < \infty$.

For $\gamma > \beta$ the situation changes once more. In passing to the limit $\gamma \to \beta -$, the absorbing and the reflecting barrier solutions become identical, and for

$\gamma > \beta$ we are once more in the situation that no boundary conditions can be imposed and that our problem is completely determined by the differential equation itself.[7]

Accordingly, we can say that for $\gamma \leqq 0$ and $\gamma > \beta$ the origin acts as *natural boundary* where no artificial conditions can be imposed, while for $0 < \gamma < \beta$ we are confronted with conditions similar to those in physical diffusion theory.

It is possible (cf. [8]) to give criteria to determine whether for a particular singular equation the boundary acts as a natural one or whether there exist infinitely many solutions, but we are still unable to formulate the appropriate boundary conditions in all cases. Moreover, the biological applications lead us to a new type of boundary problems.

The classical concept of an elastic barrier at $x = 0$ may be interpreted by saying that whenever the particle reaches $x = 0$ it has a probability p to be absorbed and probability $q = 1 - p$ to be reflected. The limiting cases $p = 0$ and $p = 1$ represent reflecting and absorbing barriers, respectively. Now biological problems compel us to consider the more general case where the particle is only temporarily absorbed; that is, if the particle is absorbed, it remains fixed at the boundary for a finite time, and reverts then to the diffusion process. The sojourn time at the boundary is a random variable with an exponential distribution. In the familiar terminology of a probability mass spread over the x-axis this means that in addition to the solution $u(t, x)$ we have to consider a finite mass $m(t)$ concentrated at the origin. This mass flows at a rate proportional to $m(t)$ back into the interval $x > 0$, but it is partly replaced by new mass being absorbed at the origin. If the diffusion starts with all the initial mass distributed over $x > 0$, then $m(0) = 0$, and one would expect that $m(t)$ will increase to a certain saturation value.

Unfortunately the appropriate boundary conditions have not been formulated and the precise conditions on $a(y)$ and $b(y)$ under which such a process can take place are not known. For an interesting special case which is of some importance in mathematical genetics cf. [6].

5. Existence problems. Pathological solutions. We now return to the case of an infinite interval and the two diffusion equations (3.12) and (3.13). As was mentioned at the end of §3, the whole setup as well as the analogy with the classical diffusion equation $u_y = u_{xx}$ leads one to assume that the transition probability $u(y; t, x)$ will satisfy the two equations (3.12) and (3.13) and, conversely, that each of these equations should have a fundamental solution which is essentially uniquely determined and can serve as $u(y; t, x)$. However, already the first attempt to prove this conjecture showed that it cannot be true under

[7] However, for $\gamma > \beta$ there exist solutions for which $\pi(t, y)$ increases. These solutions have no probability significance and correspond to the solutions of the theory of heat with heat flowing into the medium. The existence of analogous solutions for the whole axis was discovered by Hille [13]; his "explosive" solutions are of this kind with $\pi(t, y) \to \infty$ as $t \to t_0 > 0$.

all circumstances. In fact, it was shown in [5] that whenever one of the two integrals

$$\int^{\infty} a^{-1/2}(x)\,dx, \qquad \int_{-\infty} a^{-1/2}(x)\,dx \tag{5.1}$$

converges, an appropriate $b(x)$ can be found such that (3.12) admits of a solution which is non-negative, integrable, and such that $w(0, x) = 0$. Thus in this case the equation (3.12) does not suffice to describe our process. However, assuming that both integrals (5.1) diverge, the construction of [5] shows that our conjecture is valid under very mild regularity restrictions on $a(x)$ and $b(x)$. In other words, in case of divergence of (5.1) the whole classical theory carries over to our diffusion equations.[8] The problem was taken up by Yosida [26] and Hille [11, 12] who obtained similar results under slightly different conditions. In particular, Hille showed that (3.13) has a unique, non-negative solution $u(t, x)$ satisfying the conditions

$$u(0, x) = f(x) \geqq 0, \qquad \int_{-\infty}^{+\infty} u(t, x)\,dx = \int_{-\infty}^{+\infty} f(x)\,dx \tag{5.2}$$

provided that

$$| a'(x) - b(x) | \leqq K\{ | x | + 1\} \tag{5.3}$$

and that the integral

$$\int \frac{x}{a(x)}\,dx \tag{5.4}$$

diverges both at $+\infty$ and $-\infty$; conversely, if (5.3) holds, the divergence of (5.4) is a necessary condition for the theorem.[9]

These unexpected results were at first rather disturbing since they seemed to indicate that the diffusion equations are, after all, an inadequate description of the actual process. Fortunately a satisfactory explanation can be found which makes the theory more harmonious than it would be if our diffusion equations always behaved as the classical equation $u_t = u_{xx}$. At the same time we shall be led to a new type of solution of the backward equation and to an interesting phenomenon concerning the forward equation.

The singular diffusion equations discussed in the preceding section point in

[8] In [5] the more general case is treated where the coefficients a and b depend on t. The later investigations depend on the theory of semigroups and apply therefore only to the temporally homogeneous case treated in the text.

[9] The apparent discrepancy between Hille's conditions and Feller's condition (5.1) is readily explained. If one of the integrals in (5.1) converges, then it is impossible that $| a'(x) | < K (| x | + 1)$, and hence (5.3) can hold only if the growth of $b(x)$ is adjusted in a special way to that of $a(x)$. On the other hand, if $| a'(x) | < K (| x | + 1)$, then $a(x) = O(x^2)$ and the divergence of (5.1) implies that of (5.4). [*Added in proof.* Unfortunately the text takes account only of Hille's paper [12] in which (3.13) is treated. The author is indebted to Hille for the manuscript and for many inspiring discussions. Subsequently Hille treated also (3.12) and obtained conclusive results concerning the uniqueness problems for the two equations; they are announced in [11].]

the proper direction. The first point to be noted is that there is no essential difference between diffusion in a finite or infinite interval. In fact, let $\gamma(s)$ be an arbitrary positive function and consider the transformation of variables

$$\xi = \int_0^x \gamma(s)\, ds, \qquad \eta = \int_0^y \gamma(s)\, ds. \tag{5.5}$$

If $u(y; t, x)$ satisfies the diffusion equations (3.12) and (3.13), then $u^*(\eta; t, \xi)/\gamma(\xi)$ (as a function of t, η and of t, ξ, respectively) satisfies similar equations with $a(y)$ and $b(y)$ replaced by $a(y)\cdot\gamma^2(y)$ and $a(y)\cdot\gamma'(y) + b(y)\cdot\gamma(y)$. Moreover, if $u(y; t, x)$ is, as a function of x, a probability density, then the same is true of u^* as a function of ξ. Thus we have a whole group of transformations which change the pair of diffusion equations into equivalent equations. We can make use of this fact either to simplify the equations or to change the interval. In particular, it is always possible to transform a finite interval into the entire real axis and vice versa. However, even if the coefficients are regular in the interior, the transformed equations will in general be of a singular type, with the coefficients unbounded at the endpoints, or $a(x)$ vanishing at an endpoint.

In particular, we may take the ordinary diffusion equation $u_t = u_{xx}$ for the finite interval $-1 < x < 1$, and transform the latter into $-\infty < \xi < \infty$. We obtain in this way a pair of diffusion equations for the infinite interval which exhibits all the anomalies which at first appeared so perplexing. There is no uniqueness, since we have the transforms of the classical absorbing and reflecting barrier solutions, etc. Furthermore, there exist positive solutions (well known from the theory of heat conduction) with zero initial values and so on.

In probabilistic language we can describe the situation as follows. The coefficients a and b can be such that with probability one the particle never reaches a boundary. In this case no boundary conditions need (or can) be imposed, and we have the phenomenon of *natural boundaries*. The typical example is the homogeneous diffusion on the entire axis, since Wiener's result that almost all path functions are continuous implies that the particle cannot reach infinity.

If the coefficients a and b are such that the particle has positive probability to reach a boundary, several contingencies may arise. For example, the drift toward the barrier can be so strong that the boundary automatically acts as an absorbing barrier, and again no boundary conditions can be imposed. A typical example is the origin in the case of the singular equation with coefficients given by (4.2) when $\gamma \leqq 0$ (cf. §4). Alternatively, the situation may be as with homogeneous diffusion in a finite interval, where various boundary conditions can be imposed.

In particular, it is possible that there exists a unique transition probability $u(y; t, x)$ which satisfies the two equations (3.12) and (3.13), but for which the total probability mass $\pi(t, y)$ (defined in (4.3)) is steadily decreasing. Here the boundaries are natural in the sense that no boundary conditions in the classical sense can be imposed. Nevertheless, we are led to a new type of solutions at least for the backward equation (3.12) and a curious phenomenon regarding the forward equation.

Following a remark of Doob[10] we can modify our diffusion process by stipulating that whenever the particle reaches a boundary, it is instantaneously transferred back to the interior. More precisely, let $\phi_i(x)$ $(i = 1, 2)$ be two density functions and let us assume that if the particle reaches the right or left boundary, it jumps to a position Y_1 or Y_2, where the Y_i are random variables with probability densities $\phi_i(x)$. It is easy to calculate the transition probability $u^*(y; t, x)$ of the so modified process, but for our purposes the explicit form is of no particular interest. It is rather obvious that $u^*(y; t, x)$ behaves asymptotically for $t \to 0$ as the transition probability $u(y; t, x)$ of the original process, so that according to the theory of §3 also $u^*(y; t, x)$ must satisfy the *backward equation* (3.12). It follows that $u^* - u$ is a non-negative solution of (3.12) with identically vanishing initial values, so that in this case *the initial value problem for* (3.12) *cannot be determined.* The interesting fact is that our phenomenon can occur even when the initial value problem for (3.13) is uniquely determined. In this case $u^*(y; t, x)$ cannot satisfy (3.13) and we have thus a diffusion process where *the transition probability satisfies the backward equation but not the forward equation.* Instead, the inequality sign $\geqq$ holds in (3.13) and the right side is to be modified by adding a positive operator.[11] We are thus led to the conclusion that the adjoint of the backward equation (3.12) is not always given by (3.13), and is in general not even a differential equation. This is a new phenomenon which seems of interest for the general theory of semigroups and of differential operators.

6. Ito's approach. Before passing to more general processes a word should be said on a new way of describing our diffusion processes which is due to Ito [14, 15]. He does not make direct use of the differential equations or the transition probabilities, but expresses (at least in certain cases) the path functions of a general process by means of those of the Wiener process.

Let $X(t)$ represent the path functions of the latter. We have seen that with probability one $X(t)$ is not of bounded variation so that the classical definition of the integral

$$\int_0^t f(\tau)\, dX(\tau) \tag{6.1}$$

[10] Doob discussed in detail a similar phenomenon for discrete Markov chains, cf. [2].

[11] For details cf. [8]. In the case of discrete Markov chains one is led to two infinite systems of ordinary differential equations playing the role of our diffusion equations. As Doob has shown, the analogue of (3.12) holds always, but the analogue of (3.13) holds only with the equality replaced by the sign $\geqq$. We now see why a direct derivation of (3.13) is impossible. In attempting a derivation along the lines which led to (3.12) one would introduce the function $F(t, y) = \int_{-\infty}^{+\infty} u(y; t, x) f(x)\, dx$ where $f(x)$ vanishes for $|x| > A$ and has, say, three continuous derivatives. Then (*) $F(t+\Delta t, y) = \int_{-\infty}^{+\infty} u(y; \Delta t, z) F(t, z)\, dz \geqq \int_{-\infty}^{+\infty} u(y; t, z)\, dz \cdot \int_{|x-z|<\epsilon} f(x) u(z; \Delta t, x)\, dx$. If one had here the equality sign, it would be easy to derive (3.13) under fairly mild conditions (essentially supposing that the limits in (3.4) and (3.5) are attained boundedly in every finite interval). However, in order to replace the inequality sign in (*) by the equality sign one has to introduce strong uniformity conditions which are actually not justified even in relatively simple cases.

breaks down, even if $f(t)$ is continuous: in fact, for almost every $X(t)$ we can find a sequence of subdivisions $t_k^{(n)}$ such that the corresponding Riemann sums will diverge. Nevertheless, it is possible to give a meaning to the integral (6.1). In fact, consider a fixed sequence of subdivisions, say $t_k^{(n)} = k/2^n$. The corresponding Riemann sum

$$\sum_k f(t_k^{(n)})\Delta X(t_k^{(n)}) \tag{6.2}$$

is a random variable, and it has been shown by various authors that it converges in probability to a random variable which has all the desired properties of an integral.

Using this notion Ito shows that under certain restrictions on a and b the stochastic integral equation

$$Y(t) = c + \int_0^t b(Y(\tau))\, d\tau + \int_0^t a(Y(\tau))\, dX(\tau) \tag{6.3}$$

admits of an essentially unique solution $Y(t)$. The latter is defined as a random variable on the Wiener space $\{X(t)\}$. It is fairly obvious from (6.3) that the infinitesimal transition probabilities of $Y(t)$ have the desired properties (3.4) and (3.8), and thus we have arrived at a representation of a stochastic process of the required type.

This approach has the advantage that it permits a direct study of the properties of the path functions $Y(t)$, such as their continuity, etc. In principle, we have here a possibility of proving existence theorems for the partial differential equations (3.12)–(3.13) directly from the properties of the path functions. However, the method is for the time being restricted to the infinite interval and the conditions on a and b are such as to guarantee the uniqueness of the solution. So far, therefore, we cannot obtain any new information concerning the "pathological" cases.

7. Diffusion in phase space. The method of §3 works also in two and more dimensions. In two dimensions the transition probability is of the form $u(y_1, y_2; t, x_1, x_2)$, and the equations corresponding to (3.12) and (3.13) are

$$u_t = \sum a_{ij} u_{y_i y_j} - \sum b_i u_{y_i}, \qquad i, j = 1, 2 \tag{7.1}$$

and

$$u_t = \sum (a_{ij} u)_{x_i x_j} + \sum (b_i u)_{x_i}, \qquad i, j = 1, 2 \tag{7.2}$$

where the coefficients may depend on the two space variables and the matrix (a_{ij}) is symmetric and positive definite. The meaning of the coefficients a_{ii} and b_i is the same as before, and the mixed coefficient $a_{1,2}$ gives the infinitesimal covariance:

$$a_{12}(y_1, y_2) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \iint_{|x_i - y_i| < \epsilon} (x_1 - y_1)(x_2 - y_2) u(y_1, y_2; \Delta t, x_1, x_2)\, dx_1\, dx_2. \tag{7.3}$$

These equations were first derived by Kolmogorov [17]. Feller's constructions of the transition probability was generalized to two dimensions by Dressel [4]. In the time homogeneous case the equations were integrated by Yosida [26] using the theory of semigroups. However, nothing is known for the case of singular equations or the case where no uniqueness exists, and practically nothing is known about the appropriate boundary conditions in various problems involving bounded domains.

We shall not pursue this line of investigation but shall be content to indicate how our equations lead to a refined model of diffusion in which the path functions have derivatives.

The fact that the original model of diffusion leads to the conclusion that the particles have no velocities has been pointed out by many authors. Actually this fact is neither perturbing nor surprising: the whole theory is based on the assumption that the process is Markovian, that is, that the particle has no memory. Now if the particle had a finite velocity, shocks in the direction of motion would be less probable, so that any change of velocity would affect the chances of further changes. In other words, finite velocities would imply that the position of the particle is a random variable of a stochastic process with aftereffects.

In a Section meeting at this Congress two proposals were made to modify the basic assumptions so as to endow the particles with finite velocities. In both cases the diffusion equations would be replaced by equations of the hyperbolic type. Now the general solution of the initial value problem for such equations does not preserve positivity or the integral mean. Moreover, it depends not only on the initial value but also on their derivatives, which have no apparent probabilistic meaning. In short, the hyperbolic equations in question cannot serve as appropriate descriptions of a stochastic process, and if a particular transition probability is a solution, this is a lucky coincidence.

However, a refined model of diffusion in which the particles have finite velocities (but no accelerations) is due to Ornstein and Uhlenbeck.[12] This theory assumes that the particle has a velocity $X_2(t)$ which is the subject of an ordinary diffusion process regulated by (2.1). The position $X_1(t)$ of the particle is then obtained from

$$X_1(t) = \int_0^t X_2(\tau)\, d\tau, \tag{7.4}$$

the integral having a meaning since $X_2(t)$ is continuous with probability one. As shown by Doob, $X_1(t)$ is the variable of a Gaussian (non-Markovian) process whose probability relations approach for large t those of the ordinary Brownian motion process.

An alternative way of reaching the same conclusion consists in studying the vector $(X_1(t), X_2(t))$ as the variable of a two-dimensional diffusion process (so

[12] Cf. [23]. A thorough discussion of this process which is more appropriate for our purposes is contained in Doob [3].

that the plane is really the phase space of a particle in one-dimensional motion). Under the above assumptions this two-dimensional process is Markovian, and the corresponding transition probability $u(y_1, y_2; t, x_1, x_2)$ should satisfy a pair of equations of the form (7.1), (7.2). Now according to Ornstein and Uhlenbeck the variable $X_2(t)$ is subject to a symmetric homogeneous diffusion process, so that $b_2 = 0$ and a_{22} is a constant which we can assume as unity. From (7.4) it follows that if at any time $X_1(t) = y_1$, $X_2(t) = y_2$, then

$$\begin{aligned} E(X_1(t + \Delta t) - X_1(t)) &= y_2\Delta t + o(\Delta t) \\ E((X_1(t + \Delta t) - X_1(t))^2) &= o(\Delta t), \end{aligned} \tag{7.5}$$

and hence $a_{11} = 0$, $b_1 = y_2$. From the second equation (7.5) and Schwarz' inequality it follows that also $a_{12} = 0$. Hence (7.1) reduces to

$$u_t = u_{y_2y_2} - y_2u_{y_1}. \tag{7.6}$$

This equation has first been given by Kolmogorov [19]. If the condition of homogeneity is dropped, the same consideration leads to the more general equation

$$u_t = b(y_1, y_2)u_{y_2y_2} + a(y_1, y_2)u_{y_2} - y_2u_{y_1}. \tag{7.7}$$

This equation is of a degenerate type and practically nothing is known about the appropriate boundary conditions for finite domains. A fundamental solution for the infinite plane was constructed by M. Weber [24].

To prove the equivalence of the two approaches one should prove that the solution of (7.7) satisfies the condition (7.4). This is intuitively rather obvious, but a satisfactory proof has not been given. Otherwise the same method could be applied to the calculation of various Wiener functionals, which were the object of investigation by Cameron and Martin, and by Kac (cf. [16]). Let again $X_2(t)$ be the random variable of the simplest diffusion process, and put

$$X_1(t) = \int_0^t V(X_2(\tau))\, d\tau, \tag{7.8}$$

where $V(x)$ is a given function, say x^2. Then the pair (X_1, X_2) is the variable of a two-dimensional diffusion process, and should satisfy an equation of the form (7.6) with the coefficient y_2 replaced by $V(y_2)$. The difficulty of this approach (as well as in Ito's approach) seems to lie in a direct verification that the transition probability of the pair (X_1, X_2) has all the properties assumed in the derivation of (7.6). Formally, however, one can start with our diffusion equation and derive Kac's results from it.

In the Ornstein-Uhlenbeck model of diffusion the position $X_1(t)$ is not the variable of a Markovian process, but the process becomes Markovian when the state of the particle is defined so as to include position and velocity. Thus the same physical process can be described as a Markovian or non-Markovian process depending on the parameters used for its description. For example, in

cosmic ray showers the most interesting variable is the total number of particles, but it is not the variable of a Markov process. However, we may use the conceptual simplicity of and the tools available for Markov processes at the expense of describing the state of the system by the positions, masses, and velocities of each particle. In this way we are led to a Markov process in a rather complicated phase space. If the process is continuous (which excludes instantaneous changes like splitting of particles), we shall again be led to partial differential equations of the diffusion type.

The most radical step in this direction was taken by Feynman [9] in his new approach to quantum physics. The starting point of Feynman's theory is the remark that the phase-time space can always be split into two parts involving the "past" and "future" in such a way that the physical process becomes what we would call a stochastic process of the Markov type. As a physicist Feynman does not mind generalizations and abstractions which would make mere mathematicians shudder, and thus his "past" is permitted even to include future events (thus introducing advance effects instead of aftereffects). Feynman then introduces an assumption equivalent to continuity and shows heuristically that the diffusion equation to which one is led is essentially the Schrödinger equation. It has complex coefficients since Feynman deals with complex probability amplitudes rather than with classical transition probabilities. It is not clear at present whether this is only an excellent analytical tool or whether we have here an essential generalization of the classical concept of a stochastic process.

8. Generalizations by means of M. Riesz' potentials. Usually partial differential equations appear to play a special role among functional equations, but from a probabilistic point of view the diffusion processes are only a particular type of a Markov process. Starting from the Chapman-Kolmogorov equation (3.2) we have derived a diffusion equation by imposing certain conditions on the asymptotic behavior of the transition probability $u(y; t, x)$ as $t \to 0$. Probability considerations suggest to us also more general conditions which then lead to a generalization of the diffusion equation in which certain integrals involving u are added to the right-hand member (cf. [5] and Yosida [27]). However, even the integro-differential equations thus obtained describe only a special class of Markov processes and represent from a certain point of view an unnatural type of functional equations. Our real problem is *to find a linear functional equation equivalent to the Chapman-Kolmogorov equation* (3.2). In the language of the theory of semigroups this amounts to saying that we desire to find *the infinitesimal generators to the most general semigroup* corresponding to (3.2).

Before proceeding it should be emphasized that the Chapman-Kolmogorov equation (3.2) is by no means restricted to probability problems, but that an analogous relation holds for all linear differential equations. In fact, suppose that the solution $u(t, x)$ of an equation (whose coefficients do not depend on t) is determined by its initial values. Then $u(t, x)$ will be given by an integral of the form (2.3), and the Chapman-Kolmogorov equation merely expresses the fact

that the solution $u(s + t, x)$ can be calculated either directly in terms of its initial values at time 0 or, alternatively, in terms of its values at time t which, in turn, are expressible in terms of the initial values. Equating the two expressions we obtain (3.2). It is clear that a similar reasoning holds, for example, for a hyperbolic differential equation except that a vector notation must be used since now the problem is determined by the initial values of the solution and its time derivative.

In short, a kernel $u(y; t, x)$ induces two associated linear transformations

$$T_t f(x) = \int f(y)u(y; t, x)\, dy,$$
$$S_t g(y) = \int g(x)u(y; t, x)\, dx \tag{8.1}$$

and the Chapman-Kolmogorov equation (3.2) merely expresses the fact that as t varies each of these transformations forms a semigroup. In operator theoretical language the equation $u_t = u_{xx}$ can be interpreted as stating that the infinitesimal generator (time derivative) of T_t is given by the operator d^2/dx^2, or symbolically

$$\frac{d}{dt} T_t = T \frac{d^2}{dx^2}. \tag{8.2}$$

A formal integration leads to the symbolic equation

$$T_t = e^{td^2/dx^2}. \tag{8.3}$$

If $f(x)$ is an entire function, then (8.3) gives

$$T_t f(x) = e^{td^2/dx^2} f(x) = \sum_{k=0}^{\infty} t^k/k!\, f^{(2k)}(x), \tag{8.4}$$

which agrees with the classical solution (2.2) and (2.3). The more general diffusion equations merely replace the operator d^2/dx^2 on the right by other differential operators. In the case of the integro-differential equations mentioned above we have on the right a combination of a differential and integral operator. However, for a general solution we shall have to consider much more general operators.

The stable distributions are the best known class of solutions of the Chapman-Kolmogorov equation which are not connected with either differential or integro-differential equations. Bochner [1] was the first to notice that the symmetric stable distributions can be interpreted as solutions of an operator equation

$$u_t = Au,$$

where the operator A is a certain fractional power of d^2/dx^2. A better understanding of this important observation may be derived from an application of a technique developed by M. Riesz [22] which also points to various generalizations.

A *fractional potential of order* α is defined, according to Riesz, by

$$I^{\alpha}f(x) = \frac{1}{2\Gamma(\alpha)\cos \pi\alpha/2}\int_{-\infty}^{+\infty} f(y)\,|\,y-x\,|^{\alpha-1}\,dy. \tag{8.5}$$

This integral diverges in general, but Riesz showed that a meaning can be attached to it for a large class of functions $f(x)$. In particular, $I^{0}f(x) = f(x)$, and $I^{-2n}f(x) = (-1)^{n}f^{(2n)}(x)$, whenever n is a positive integer. This relation is interesting for two reasons. First, it is clear that in general I^{α} is a global operator, that is, $I^{\alpha}f(x)$ depends on all values of $f(x)$. For the particular values $\alpha = 0$, -2, -4, $\cdots$, however, the operator takes on a purely local character. Secondly, since I^{-2} is the same as $-d^2/dx^2$, we can interpret I^{-1} as a square root of $-d^2/dx^2$, and similarly for other powers. This enables us to pursue in more detail Bochner's remark. We are thus led to consider the *general functional equation*

$$u_t = -I^{-\alpha}u. \tag{8.6}$$

If $\alpha = 2$, this reduces to the classical diffusion equation $u_t = u_{xx}$, but for other values of α the functional equation (8.6) has quite different a character, since the operator on the right is of a global character.

From Bochner's remark one should expect that for $0 < \alpha \leqq 1$ the equation (8.6) should lead to *symmetric stable distributions*. This is so, and we proceed to verify it in a purely formal fashion for $\alpha = 1$.

The symmetric stable distribution of order $\alpha = 1$ is characterized by the density

$$u(y;\,t,\,x) = \frac{t}{(x-y)^2 + t^2}, \tag{8.7}$$

which is known as *Cauchy distribution*. It plays for the Cauchy process the role of the normal density (2.2), and the solution (2.3) reduces in the present case to the function $u(t, x)$ *which is harmonic in the half plane* $t > 0$, *vanishes at infinity, and reduces for* $t = 0$ *to* $f(x)$.

If we integrate (8.6) for $\alpha = 1$ formally by analogy with (8.4), we are led to an expansion

$$u(t,\,x) = \sum_{k=0}^{\infty} \frac{(-t)^k}{k!}\,I^{-k}f(x). \tag{8.8}$$

Substituting from the definition (8.5) it is easily verified that the right side reduces to the anticipated solution

$$u(t,\,x) = \frac{1}{\pi}\int_{-\infty}^{+\infty} f(y)\,\frac{t}{(y-x)^2 + t^2}\,dy. \tag{8.9}$$

The formal steps leading to this formula require justification. However, by similar calculations all symmetric stable distributions can be obtained, and by a slight generalization of Riesz' definitions of potentials we can derive also the unsymmetric stable distribution. It appears, in fact, that all probabilistic solu-

tions of the Chapman-Kolmogorov equation can be obtained in this way, except that the parameters of the potentials on the right side will depend on x.

It is of particular interest that we have obtained the solution of a *Dirichlet problem* for the Laplace equation $\Delta u = 0$ as the solution of a generalized diffusion problem. The explanation lies in the semigroup property of the Green function (8.7), and a similar statement applies also to the Dirichlet problem for closed curves. The curve itself corresponds to the x-axis, and certain closed curves in the interior to the lines $t = \text{const}$. These contract to a single point which corresponds to $t = \infty$. From the values of u along the boundary curve the values along the curves in the interior can be calculated by the methods just described. The fundamental fact that as $t \to \infty$ all the values tend to the same limit follows then from the ergodic principle for Markov processes on closed manifolds. We have thus a direct connection of the Dirichlet problem for elliptic equations with certain generalized diffusion equations, which in turn are closely connected with measure in function spaces where almost no function is continuous.

References

1. S. Bochner, *Diffusion equation and stochastic processes*, Proc. Nat. Acad. Sci. U.S.A. vol. 35 (1949) pp. 368–370.

2. J. L. Doob, *Markoff chains—Denumerable case*, Trans. Amer. Math. Soc. vol. 58 (1945) pp. 455–473.

3. ———, *The Brownian movement and stochastic equations*, Ann. of Math. vol. 43 (1942) pp. 351–369.

4. F. G. Dressel, *The fundamental solution of the parabolic differential equation*, Duke Math. J. vol. 7 (1940) pp. 186–203.

5. W. Feller, *Zur Theorie der Stochastischen Prozesse (Existenz und Eindeutigkeitssätze)*, Math. Ann. vol. 113 (1936) pp. 113–160.

6. ———, *Diffusion processes in the mathematical theory of Evolution*, Proceedings of the Second Symposium on Probability and Mathematical Statistics, Berkeley [to appear].

7. ———, *Two singular diffusion problems*, Ann. of Math. vol. 52 (1951), to appear.

8. ———, *The parabolic differential equations and the associated semi-groups of transformations*, to appear, with a companion paper, in Ann. of Math.

9. R. P. Feynman, *Space-time approach to non-relativistic quantum mechanics*, Reviews of Modern Physics vol. 20 (1948) pp. 367–387.

10. R. Fortet, *Les fonctions aléatoires du type de Markoff associées à certaines équations linéaires aux derivées partielles du type parabolique*, J. Math. Pures Appl. vol. 22 (1943) pp. 177–243.

11. E. Hille, *Les probabilités continues en chaîne*, C. R. Acad. Sci. Paris vol. 230 (1950) pp. 34–35.

12. ———, *On the integration problem for Fokker-Planck's equation in the theory of stochastic processes*, [To appear in the proceedings of the Tenth Congress of Scandinavian Mathematicians in Trondheim, 1950].

13. ———, *"Explosive" solutions of Fokker-Planck's equation*, Proceedings of the International Congress of Mathematicians, Cambridge, Mass., 1950.

14. Kiyosi Ito, *On a stochastic integral equation*, Proceedings of the Japan Academy of Sciences vol. 22 (1946) pp. 32–35.

15. ———, *On stochastic differential equations*, Memoirs of the American Mathematical Society, no. 4 (1951).

16. M. Kac, *On the distribution of certain Wiener functionals*, Trans. Amer. Math. Soc. vol. 65 (1949) pp. 1–13.

17. A. KOLMOGOROFF, *Ueber die analytischen Methoden in der Wahrscheinlichkeitsrechnung*, Math. Ann. vol. 104 (1931) pp. 415–458.

18. ———, *Zur Theorie der stetigen zufälligen Prozesse*, Math. Ann. vol. 108 (1933) pp. 149–160.

19. ———, *Zufällige Bewegungen*, Ann. of Math. vol. 35 (1934) pp. 116–117.

20. P. LÉVY, *Théorie de l'addition des variables aléatoires*, Paris, Gauthier-Villars, 1937.

21. ———, *Processus stochastiques et mouvement Brownien*, Paris, Gauthier-Villars, 1948.

22. M. RIESZ, *L'intégrale de Riemann-Liouville et le problème de Cauchy*, Acta Math. vol. 81 (1948) pp. 1–223.

23. G. E. UHLENBECK and L. S. ORNSTEIN, *On the theory of the Brownian motion*, Physical Reviews vol. 36 (1930) pp. 823–841.

24. M. WEBER, *The fundamental solution of the diffusion equation in phase space* [To appear].

25. N. WIENER, *Generalized harmonic analysis*, Acta Math. vol. 55 (1930) pp. 117–258.

26. K. YOSIDA, *An operator-theoretical treatment of temporally homogeneous Markoff process*, Journal of the Mathematical Society of Japan vol. 1 (1949) pp. 244–253.

27. ———, *An extension of Fokker-Planck's equation*, Proceedings of the Japan Academy of Sciences vol. 25 (1949) pp. 1–3.

28. ———, *On the differentiability and the representation of one-parameter semi-group of linear operators*, Journal of the Mathematical Society of Japan vol. 1 (1948) pp. 15–21.

PRINCETON UNIVERSITY,
PRINCETON, N. J., U. S. A.

INTERNATIONAL CONGRESS

OF

MATHEMATICIANS

Cambridge, Massachusetts, U. S. A.

1950

CONFERENCE IN TOPOLOGY

Committee

Hassler Whitney (*Chairman*)

Deane Montgomery

N. E. Steenrod

Several international conferences in topology have been held. In particular: at Moscow, in 1935 (see Rec. Math. (Mat. Sbornik) N.S. vol. 1 (43) No. 5 (1936)); at Geneva, in 1935 (see l'Enseignment Mathématique vol. 35 (1936)); at Paris, in 1947 (see Colloques Internationaux du Centre National de la Recherche Scientifique, XII, Topologie Algébrique, Paris, 1949); and at Brussels, in 1950. In the United States, there have been conferences at the University of Michigan, Ann Arbor, Mich., in 1940 (see Lectures in Topology, The University of Michigan Press, 1941), and at Chicago in 1950.

The subject of algebraic topology and applications was chosen for one of the conferences of the Congress because of its great growth in recent years, and the increasingly large contact with other fields of mathematics, in geometry, algebra, and analysis. The subject of general topology has moved considerably into the domain of analysis. It was with great regret that the field of point set theory had to be omitted altogether.

There were four sessions of the conference. The first, on homology and homotopy theory, lies at the foundation of the subject. The second, on fiber bundles and obstructions, presents tools of basic importance for applications in many fields. The third, on differentiable manifolds, is of special importance in analysis, though it relates to other fields as well. The last session, on group theory, lies somewhat apart from the others.

HASSLER WHITNEY

HOMOLOGY AND HOMOTOPY THEORY

HOMOTOPY AND HOMOLOGY

W. HUREWICZ

The concept of homotopy is a mathematical formulation of the intuitive idea of a continuous transition between two geometrical configurations. The concept of homology gives a mathematical precision to the intuitive idea of a curve bounding an "area" or a surface bounding a "volume."

1. The first step toward connecting these two basic concepts of topology was taken by L. E. J. Brouwer in 1912 by demonstrating that two continuous mappings of a two-dimensional sphere into itself can be continuously deformed into each other if and only if they have the same degree (that is, if they are equivalent from the point of view of homology theory).

After having generalized Brouwer's result to an arbitrary number of dimensions, H. Hopf undertook a systematic study of the problem of classifying the continuous mappings of a polytope P into a polytope Q. Each mapping f induces homomorphisms of homology groups of P into the corresponding groups of Q. Two mappings f and g are said to belong to the same *homology class* if they induce identical homomorphisms of homology groups (for all dimensions and all coefficient domains). The mappings f and g are said to belong to the same *homotopy class* if they can be embedded into a common one-parameter continuous family of mappings. The homotopy class of a mapping determines its homology class, but not conversely, as shown by the example of the mappings of the sphere S_3 into S_2 which all belong to the same homology class although there is an infinite number of homotopy classes. The question arises: under what special conditions the homotopy classification of the mappings of P into Q coincides with their homology classification. The classical result of Hopf states that this is the case if P is a polytope of dimension n and Q the n-dimensional sphere S_n. Using cohomology groups instead of homology groups, H. Whitney gave the following elegant formulation to Hopf's theorem. Homotopy classes of mappings of an n-dimensional polytope P into the sphere S_n are in one to one correspondence with the elements of the n-dimensional cohomology group of P with integers as coefficients.

2. In 1934–1935 the author developed the concept and theory of higher dimensional homotopy groups. Given an arcwise connected topological space Y, the n-dimensional homotopy group $\pi_n(Y)$ is defined as follows: Let an arbitrary point $y_0 \in Y$ be singled out once and for all as the "reference point," and let also a fixed point x_0 be selected on the fixed n-sphere S_n. An element of $\pi_n(Y)$ is determined by a continuous mapping of S_n into Y, satisfying the condition $(x_0) = y_0$. Two mappings f and g determine the same element of $\pi_n(Y)$ if and

only if they can be continuously deformed into each other, in such a fashion that the image of x_0 remains at y_0 during the entire process of deformation (this condition can be dispensed with if Y is simply connected). In this case every continuous mapping of S_n into Y determines uniquely an element of $\pi_n(Y)$. The group composition law is defined in a fairly obvious way by identifying two n-spheres tangent at x_0 with the two hemispheres of a single sphere whose equator has been shrunk into x_0 .

An alternative way to introduce homotopy groups is to consider the topological space F of all continuous mappings of S_{n-1} into Y satisfying the condition $f(x_0) = y_0$, where x_0 is a fixed point of S_{n-1} . Although this functional space is, generally speaking, disconnected, it can be shown that its arcwise connected components have the same homotopy type (see below) and consequently have isomorphic fundamental groups. We can thus speak about the fundamental group of the functional space F, and this group turns out to be isomorphic to the n-dimensional homotopy group $\pi_n(Y)$ in the sense of the previous definition. The group $\pi_1(Y)$ is, of course, the fundamental group of Y. A simple geometric argument shows that for $n > 1$ the groups $\pi_n(Y)$ are *abelian*. In contrast with homology groups the homotopy groups of an n-dimensional space may be nontrival even in dimensions higher than n. For instance, $\pi_3(S_2)$ is an infinite cyclic group (Hopf's theorem) and for $n > 2$ the group $\pi_{n+1}(S_n)$ is of order 2 (Freudenthal-Pontrjagin theorem).

To determine the homotopy groups of a given space is, generally speaking, an extremely difficult problem (even for finite polytopes) which so far has been solved only in a few special cases. In this respect there is a significant difference between homotopy and homology. When a polytope P is broken up in two subpolytopes Q and R, there is a relatively simple relation (Meyer Vietoris theorem restated recently in terms of the so-called exact sequences) between the homology invariants of the polytopes P, Q, R and the intersection $Q \cap R$. No analogous relation exists for homotopy groups. This is tied up to the fact that a continuous image of the n-sphere in P cannot be decomposed into "small" spherical images, the way a simplicial chain can be decomposed into "small parts." Therefore the basic process of homology theory consisting in decomposing a space into smaller pieces with simpler homology structure has no counterpart in homotopy theory. The difficulty is illustrated by the fact that even in the case of a space P represented as the union of two subspaces Q and R with only one point in common, there is no simple relation between higher dimensional homotopy groups of P, Q, and R.

3. In certain "elementary" cases, homotopy groups can be reduced to homology groups. Let Y be an arcwise connected space and let $H_n(Y)$ be the n-dimensional homology group of Y based on singular chains, with integers as coefficients. A continuous image of S_n in Y can be regarded as a singular n-cycle. Since two homotopic spherical images determine homologous singular cycles, one obtains a "natural" homomorphism of $\pi_n(Y)$ into $H_n(Y)$. The fundamental *equivalence theorem* states:

If $n \geqq 2$ and the homotopy groups $\pi_i(Y)$ are trivial for $i < n$, the nth homotopy group $\pi_n(Y)$ is isomorphic to the nth homology group $H_n(Y)$ under the natural homomorphism.

For example $\pi_i(S_n)$ is trivial for $i < n$ and hence $\pi_n(S_n)$ is infinite cyclic. Arcwise connected spaces whose homotopy groups in dimensions less than or equal to n vanish are called *n-connected.* This property is equivalent to the condition that every continuous image of an arbitrary n-dimensional polytope in Y be homotopic to a single point. An obvious corollary to the equivalence theorem states that Y is n-connected if and only if the groups $\pi_1(Y), H_2(Y), \cdots, H_n(Y)$ are trivial. It follows that a polytope can be shrunk to a point in itself if and only if it is simply connected (= 1-connected) and has vanishing homology groups in all dimensions.

4. The equivalence theorem just stated can be formulated in the following way: If the arcwise connected space Y is $(n-1)$-connected ($n \geqq 2$), the homotopy classes of mappings of S_n into Y coincide with their homology classes. Comparing this result to Hopf's theorem mentioned above we find that the assertions in both theorems are of the same type. Hopf's theorem and the equivalence theorem are both contained in the following more general theorem:

If Y is an $(n-1)$-connected space ($n \geqq 2$) and P an n-dimensional polytope, the homotopy classification of P into Y agrees with their homology classification. More refined results in this direction can be obtained by using the concept of a *homotopy obstruction* developed by S. Eilenberg (implicitly this idea was used for the first time by H. Whitney in his revealing proof of Hopf's theorem). Let Y be a 1-connected space and P an arbitrary polytope. Let us denote by P^m the m-dimensional skeleton of P, that is, the union of all simplexes of P of dimensions less than or equal to m. Consider now two continuous mappings f and g of P into Y. An attempt to deform f continuously into g can be carried out stepwise, each step involving considerations in one dimension only. Suppose we have succeeded in deforming f into a mapping f' which agrees with g on the $(m-1)$-dimensional skeleton P^{m-1}. For each oriented simplex σ^m of P the images $f'(\sigma^m)$ and $g(\sigma^m)$ (which coincide on the boundary of σ^m) yield, in an obvious fashion, a continuous image of an m-sphere. Let us denote by $\varphi(\sigma^m)$ the element of the homotopy group $\pi_m = \pi_m(Y)$ determined by this spherical image. The function φ can be regarded as an m-dimensional cochain of P with coefficients in the group π_m. This cochain turns out to be a cocycle. Its cohomology class is called the homotopy obstruction for the couple (f', g). The notation is justified by the following theorem: *If the obstruction is zero (that is, if the cocycle φ is cobounding), the deformation process can be pushed one step further so as to deform f' into a mapping f'' which agrees with g on the m-dimensional skeleton P^m. Moreover the deformation can be carried out in such a way that the image of P^{m-2} (but not necessarily of P^{m-1}) remains unchanged.*

If the cohomology group $H(P, \pi_m)$ of P with coefficients in π_m is trivial (this will be the case, for instance, if $\pi_m = \pi_m(Y)$ vanishes), all obstructions in

dimension m are zero, and hence any two mappings which coincide on P^{m-1} are homotopic on P^m.

The author derived further results by connecting homotopy obstructions with so called *homology obstructions*. Let $H_m = H_m(Y)$ be the mth homology group of Y with integers as coefficients, in the sense of the singular homology theory, and let $H^m(P, H_m)$ be the mth cohomology group of P with coefficients in H_m. The natural homomorphism of π_m into H_m yields a homomorphism of $H(P, \pi_m)$ into $H(P, H_m)$. Under this homomorphism the homotopy obstruction of the couple (f', g) is sent into an element of $H(P, H_m)$ which is called the homology obstruction of the couple (f', g). The homology obstruction is zero if the mappings f' and g, or—what amounts to the same thing—the mappings f and g belong to the same homology class.

Under certain conditions, homotopy obstructions coincide with homology obstructions. This is, for instance, the case if $\pi_m(Y)$ is isomorphic to $H_m(Y)$ under the natural homomorphism. Under such circumstances the homotopy problem in dimension m is completely reducible to the corresponding homology problem.

5. The groups $\pi_n(Y)$ are a special case of more general invariants called *relative homotopy groups*, which are in many respects analogous to relative homology groups.

Let Y be a topological space and Z a subset of Y. Both Y and Z are assumed to be arcwise connected. For every integer $n \geqq 2$ we shall define the relative homotopy group $\pi_n(Y, Z)$. Let E_n be a fixed n-cell with the boundary S_{n-1}. Let us select a point x_0 of S_{n-1} and a point z_0 of Z. An element of $\pi_n(Y, Z)$ is determined by a continuous mapping of E_n into Y satisfying the boundary conditions

$$f(S_{n-1}) \subset Z, \qquad f(x_0) = z_0.$$

Two mappings determine the same element of $\pi_n(Y, Z)$ if they can be continuously deformed into each other in such a way that the boundary conditions are satisfied during the entire process of deformation. The composition law is defined by partitioning an n-cell into two cells with an $(n-1)$-cell in common and shrinking this $(n-1)$-cell into a single point x_0. It is evident that no reasonable composition can be defined when $n = 1$. An alternative definition describes relative homotopy groups as fundamental groups of suitably defined functional spaces.

For $n > 2$, $\pi_n(Y, Z)$ is abelian. The group $\pi_2(Y, Z)$ is in general nonabelian, and this accounts for some of the peculiar difficulties encountered in the homotopy theory of two-dimensional spaces.

In exactly the same way as in the case of absolute homotopy groups, one defines a natural homomorphism of the relative homotopy group $\pi_n(Y, Z)$ into the relative homology group $H_n(Y, Z)$ (with integer coefficients). We shall call the couple (Y, Z) *n-connected* ($n \geqq 2$) if (a) the group $\pi_1(Z)$ is isomorphic to $\pi_1(Y)$ under the natural homomorphism and (b) $\pi_m(Y, Z)$ vanishes for

$2 \leqq m \leqq n$. Without using homotopy groups, the definition can be formulated as follows: the couple (Y, Z) is n-connected if given any n-dimensional polytope P with a subpolytope Q and any continuous mapping f of P into Y satisfying $f(Q) \subset Z$, f can be deformed, without changing the image $f(Q)$, into a mapping g satisfying $g(P) \subset Z$. In analogy with the equivalence theorem for absolute homotopy groups we have:

If the couple (Y, Z) is $(n-1)$-connected, the homotopy group $\pi_n(Y, Z)$ is isomorphic to the homology group $H_n(Y, Z)$ under the natural homomorphism.

Relative homotopy groups play a basic role in the study of fibre spaces and fibre bundles.

An important generalization of relative homotopy groups has been developed recently by A. L. Blakers and W. S. Massey. They define homotopy groups of a so-called "triad," that is, of a space Y supplied with two closed sets U and V whose union is Y. Roughly speaking, the elements of the n-dimensional homotopy group of a triad are defined by mappings of an n-dimensional cell into Y such that one of the two hemispheres of the boundary of the cell is mapped into U and the other one into V. The theory of homotopy triads helps greatly to understand Freudenthal's so-called "suspension homomorphism" which is the basic tool in the discussion of homotopy groups of spheres.

6. The problem of classifying mappings of one space into another space is closely related to the problem of classifying spaces themselves according to their homotopy properties. Two spaces X and Y are said to have the same *homotopy type* if there exists a continuous mapping f of X into Y and a continuous mapping g of Y into X such that the combined mappings $f \circ g$ and $g \circ f$ are homotopic to identities. Two spaces which have the same homotopy types have isomorphic cohomology rings and isomorphic homotopy groups in all dimensions. As has been shown recently by J. H. C. Whitehead, a necessary and sufficient condition for X and Y to have the same homotopy type is the existence of a continuous mapping f of X into Y which induces isomorphic mappings of the fundamental group and the homology groups of X into the corresponding groups of Y.

J. H. C. Whitehead has succeeded in completely describing the homotopy types of simply connected four-dimensional polytopes in terms of their homology invariants. This description involves in addition to cohomology rings the so-called "Pontrjagin squares."

7. So far we have been concerned mainly with the problem of reducing homotopy properties of mappings and spaces to their homology properties. In certain cases, however, one is led to the converse problem of obtaining information about homology properties of a space from its known homotopy properties. A typical example is an *aspherical space*. By this is meant a space whose homotopy groups vanish in all dimensions $n \geqq 2$. It is known that the homotopy type and hence all homology invariants of an aspherical space are determined by its fundamental group. An analogous result holds for spaces which have only one nonvanishing homotopy group. The algebraical process by which in cases of

this type the homology invariants of the space are determined by its homotopy groups has been studied extensively by Eilenberg-MacLane. Their research resulted in a fruitful theory of homology invariants associated with abstract groups. This theory has interesting applications in algebra and in the theory of Lie groups.

8. At present the main effort in homotopy theory seems to concentrate on the problem of determining homotopy groups of spheres. The tools used in this research are predominantly of algebraical nature, like generalized Hopf invariants studied by G. Whitehead, or "cup products" introduced by N. Steenrod. Important advances have been made, most significant of which is the result established recently by G. Whitehead and Pontrjagin, to the effect that $\pi_{n+2}(S_n)$ is a group of order 2 for $n \geqq 3$. Nevertheless our knowledge of homotopy groups of spheres remains meager.

Perhaps the present trend of research does not put enough emphasis on tools that could be provided by the geometrical structure from the point of view of differential geometry, like properties of geodesic lines, study of critical points, etc. Recent work of E. Pitcher seems to indicate that some progress can be expected from this direction.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY,
CAMBRIDGE, MASS., U. S. A.

HOMOTOPY GROUPS AND ALGEBRAIC HOMOLOGY THEORIES

SAMUEL EILENBERG[1]

This note will present certain topological results obtained by Saunders MacLane and the author. Some of the algebraic aspects of these questions are presented by MacLane in another note in these Proceedings.

Let X be an arcwise connected topological space with base point x_0 and with vanishing homotopy groups $\pi_i(X)$ for $1 \leqq i < n$. The singular homology and cohomology groups of X may then be derived from the singular complex $S_n(X)$ consisting of the singular simplexes whose faces of dimension less than n all degenerate to the point x_0. A q-simplex T of $S_n(X)$ determines a system of "labels" consisting of elements of $\pi_n = \pi_n(X)$ attached to each n-dimensional face of T. The alternating sum of the labels on the faces of an $(n+1)$-face of T are zero. Such a system of labels may be regarded as an abstract simplex of a complex $K(\pi_n, n)$; this is a purely algebraic construction on the group $\Pi = \pi_n$ and the integer n. The function which to each simplex of $S_n(X)$ assigns its system of labels yields a simplicial mapping $\kappa: S_n(X) \to K(\pi_n, n)$. Each n-dimensional simplex of $K(\pi_n, n)$ consists of a single label; i.e., of an element of π_n. This yields the basic cohomology class $b^n \in H^n(\pi_n, n; \pi_n)$ of the complex $K(\pi_n, n)$ with coefficients in π_n, and the basic cohomology class $s^n = \kappa^* b^n \in H^n(X; \pi_n)$ of X with coefficients in π_n.

We shall further assume that the homotopy groups $\pi_i(X)$ vanish also for $n < i < q$. Then every simplex of $K(\pi_n, n)$ of dimension less than or equal to q can be realized geometrically in $S_n(X)$ and this yields an *inverse* simplicial mapping $\bar{\kappa}: K(\pi_n, n) \to S_n(X)$ defined in dimensions less than or equal to q. Using this map we have shown[2] that the homology and cohomology groups of X in dimensions less than q (and also partially in dimension q) are those of $K(\pi_n, n)$. In attempting to extend $\bar{\kappa}$ to the dimension $q+1$ one encounters an obstruction which is a cohomology class $k_n^{q+1} \in H^{q+1}(\pi_n, n; \pi_q)$ of the complex $K(\pi_n, n)$ with coefficients in $\pi_q = \pi_q(X)$.

Let K be a (possibly infinite) simplicial complex with ordered vertices and $f: K^n \to X$ a continuous mapping of the n-skeleton of K. Without loss of generality we may assume that $f(K^{n-1}) = x_0$. If the map f is extendable to a map $K^{n+1} \to X$, then the cohomology class $f^* s^n \in H^n(K^n; \pi_n)$ determines uniquely a cohomology class $f^{\blacktriangle} s^n \in H^n(K; \pi_n)$. If further $g: K^n \to X$ is another such map which agrees with f on a subcomplex L of K, then a relative cohomology class $(f-g)^{\blacktriangle} s^n \in H^n(K, L; \pi_n)$ is uniquely determined.

Let $f: K^n \cup L \to X$ be a map extendable to a map $K^{n+1} \cup L \to X$. For each

[1] John Simon Guggenheim Memorial Fellow.

[2] S. Eilenberg and S. MacLane, *Relations between homology and homotopy groups of spaces* I *and* II, Ann. of Math. vol. 46 (1945) pp. 480–509 and vol. 51 (1950) pp. 514–533.

simplex Δ of K and each n-face of Δ, f determines a label which is an element of π_n. These labels form a simplex of $K(\pi_n, n)$, thus yielding a simplicial map $\tilde{f}\colon K \to K(\pi_n, n)$. It can also be shown that f is extendable to a map $f'\colon K^q \cup L \to X$. The obstruction $c^{q+1}(f') \in H^{q+1}(K, L; \pi_q)$ is independent of the choice of f' and is called the secondary obstruction $z^{q+1}(f)$ of f.

THEOREM 1. *Let K be a simplicial complex, L a subcomplex, $f, g\colon K^n \cup L \to X$ maps such that $f(K^{n-1}) = g(K^{n-1}) = x_0$ and $f \mid L = g \mid L$. If both f and g are extendable to maps $K^{n+1} \cup L \to X$, then their secondary obstructions satisfy*

$$z^{q+1}(f) - z^{q+1}(g) = (\tilde{f} - \tilde{g})^* k_n^{q+1}.$$

Here $(\tilde{f} - \tilde{g})^*$ is the difference homomorphism $H^{q+1}(\pi_n, n; \pi_q) \to H^{q+1}(K, L, \pi_q)$ induced by the maps $\tilde{f}$ and $\tilde{g}$ which agree on L.

The above theorem yields interesting results only if one has more information about the element k_n^{q+1} and about the group $H^{q+1}(\pi_n, n; \pi_q)$ in which it lies. The latter is closely related with the homology theory of abelian groups based on the complexes $A(\Pi)$ and $A^1(\Pi)$ of the aforementioned note of MacLane.

We limit our attention here to the case $q = n + 1$. In this case we have natural isomorphisms

$$H^4(\Pi, n; G) \approx H^3(A^1(\Pi); G)$$

$$H^{n+2}(\Pi, n; G) \approx H^3(A(\Pi); G), \qquad n > 2.$$

Each element χ of $H^{n+2}(\Pi, n; g)$ yields a trace t which is a function defined on Π with values in G and which satisfies the conditions

(1) $t(x) - t(-x) = 0$

(2) $t(x + y + z) - t(y + z) - t(x + z) - t(x + y) + t(x) + t(y) + t(z) = 0$

for $n = 2$. For $n > 2$, (2) is replaced by the stronger condition

(2') $t(x + y) - t(y) - t(x) = 0$.

In particular, we consider the element $k_n^{n+2} \in H^{n+2}(\pi_n, n; \pi_{n+1})$, and prove that the trace of k_n^{n+2} is J. H. C. Whitehead's function $\eta\colon \pi_n \to \pi_{n+1}$ obtained by combining each map $S^n \to X$ with a map $S^{n+1} \to S^n$ that yields a generator of $\pi_{n+1}(S^n)$.[3]

The correspondence $\chi \to t$ yields an isomorphic mapping of the group $H^{n+2}(\Pi, n; G)$ onto the group of all functions satisfying conditions (1) and (2) (or (1) and (2') if $n > 2$). If the abelian group Π is finitely generated, then the inverse mapping may be described as follows.

Let K be a complex with sufficient simplicial structure, L a subcomplex, and let $t\colon \Pi \to G$ be a function satisfying (1) and (2). Then if Π is finitely generated, one may define a "Pontrjagin square"

$$P_t\colon H^q(K, L; \Pi) \to H^{2q}(K, L; G), \qquad q \text{ even},$$

[3] Cf. G. W. Whitehead, *On spaces with vanishing low-dimensional homotopy groups*, Proc. Nat. Acad. Sci. U. S. A. vol. 34 (1948) pp. 207–211; Theorem 5.

which itself satisfies (1) and (2). If t satisfies (1) and (2'), then one has a "Steenrod square"

$$Sq_t^2 : H^q(K, L; \Pi) \to H^{q+2}(K, L; G)$$

which itself satisfies (1) and (2'). If $b^n \in H^n(\Pi, n; \Pi)$ is the basic cohomology class of $K(\Pi, n)$, then the elements $P_t(b^2)$ and $Sq_t^2(b^n)$ for $n > 2$ of $H^{n+2}(\Pi, n; G)$ have precisely the trace t.

These results combined with some formal properties of P_t and Sq_t^2 and with Theorem 1 yield the following theorem.

THEOREM 2. *Let X be an arcwise connected topological space with $\pi_i(X) = 0$ for $i < n$ $(n > 1)$ and with $\pi_n(X)$ finitely generated. Let K be a simplicial complex, L a subcomplex, and $f, g : K^n \cup L \to X$ two maps extendable to maps $K^{n+1} \cup L \to X$ which agree on L. Then their secondary obstructions satisfy*

$$z^{n+2}(f) - z^{n+2}(g) = \begin{cases} P_\eta(\lambda^2) + g^\blacktriangle(s^2) \cup \lambda^2, & n = 2, \\ f^\blacktriangle(s^2) \cup \lambda^2 - P_\eta(\lambda^2), & n = 2, \\ Sq_\eta^2(\lambda^n), & n > 2, \end{cases}$$

where $s^n \in H^n(X; \pi_n)$ is the basic cohomology class of X, $\lambda^n = (f - g)^\blacktriangle s^n$, and $\cup$ denotes the ordinary cup product relative to the pairing $[x, y] = \eta(x + y) - \eta(x) - \eta(y)$.

In order to deduce from this theorem a classification theorem for maps $K^{n+1} \to X$ we need the "Postnikov square" which is a homomorphism

$$\overline{P}_t : H^q(K, L; \Pi) \to H^{2q+1}(K, L; G), \qquad q \text{ odd},$$

defined for each t satisfying conditions (1) and (2).

THEOREM 3. *Let X be as in Theorem 2 and let $f, g : K^{n+1} \to X$ be two maps which agree on $K^n \cup L$. Let $d^{n+1}(f, g) \in H^{n+1}(K, L; \pi_{n+1})$ be the cohomology class measuring the difference between f and g. Then f and g are homotopic relative to L if and only if there is a cohomology class $e^{n-1} \in H^{n-1}(K, L; \pi_n)$ such that*

$$d^{n+1}(f, g) = \begin{cases} \overline{P}_\eta(e^1) + f^*(s^2) \cup e^1, & n = 2, \\ Sq_\eta^2(e^{n-1}), & n > 2. \end{cases}$$

Theorems 2 and 3 constitute a generalization of Steenrod's results[4] for the case $X = S^n$. It should be noted that the method also is in a sense a generalization

[4] N. E. Steenrod, *Products of cocycles and extensions of mappings*, Ann. of Math. vol. 48 (1947) pp. 290–320.

of that of Steenrod: the complexes M^n are replaced here by the algebraic complexes $K(\pi_n, n)$. Our results are almost identical with results recently obtained by J. H. C. Whitehead by a different method and include earlier results of Whitney[5] and Postnikov.[6]

COLUMBIA UNIVERSITY,
NEW YORK, N. Y., U. S. A.

[5] H. Whitney, *Classification of the mappings of a 3-complex into a simply connected space*, Ann. of Math. vol. 50 (1949) pp. 270–284.

[6] M. M. Postnikov, *Classification of the continuous mappings of an arbitrary n-dimensional polyhedron into a connected topological space which is aspherical in dimensions greater than unity and less than n*, C. R. (Doklady) Acad. Sci. URSS N.S. vol. 67 (1949) pp. 427–430

ALGEBRAIC HOMOTOPY THEORY

J. H. C. WHITEHEAD

In homotopy theory, spaces are classified in terms of homotopy classes of maps, rather than individual maps of one space in another. Thus, using the word category in the sense of S. Eilenberg and Saunders MacLane [1],[1] a homotopy category of spaces is one in which the objects are topological spaces and the "mappings" are not individual maps but homotopy classes of ordinary maps. The equivalences are the classes with two-sided inverses, and two spaces are of the same homotopy type if and only if they are related by such an equivalence. The ultimate object of algebraic homotopy is to construct a purely algebraic theory, which is equivalent to homotopy theory in the same sort of way that "analytic" is equivalent to "pure" projective geometry. In discussing this project I shall confine myself to spaces which are covered by CW-complexes as defined in [4].

The following theorem is proved in [4]. Let π_n, π'_n be the nth homotopy groups of K, K', Let $\phi\colon K \to K'$ be a map of a complex K into a complex K', which induces an isomorphism $\phi_n\colon \pi_n \approx \pi'_n$ for every $n = 1, 2, \cdots$. Then $\phi\colon K \equiv K'$, which means that the homotopy class containing ϕ is an equivalence. Thus the homotopy groups constitute a system of algebraic invariants which, in a certain sense, are sufficiently powerful to characterize the homotopy type of a complex.

This does not mean that $K \equiv K'$ just because $\pi_n \approx \pi'_n$ for every $n = 1, 2, \cdots$. The crux of the matter is not merely that $\pi_n \approx \pi'_n$, but that a certain family of isomorphisms, $\phi_n\colon \pi_n \approx \pi'_n$, has a *geometrical realization*, $K \to K'$. That is to say, the latter induces the former. Therefore the emphasis is shifted to the realization problem, which is to find necessary and sufficient conditions in order that a given set of isomorphisms or, more generally, homomorphisms, $\phi_n\colon \pi_n \to \pi'_n$, have a geometrical realization $K \to K'$.

At this stage let me remark, in parentheses, that other sets of "sufficiently powerful" invariants may be defined in terms of the universal covering complex, $\tilde{K}$, of K. For the two preceding paragraphs may be restated in terms of the system of groups π_1, $H_n(\tilde{K})$, where $H_n(\tilde{K})$ is the nth integral homology group, defined in terms of finite chains of $\tilde{K}$. Thus, within the category of simply connected complexes, π_n could be replaced by $H_n(K)$. However, I shall continue the discussion in terms of the groups π_n.

When studying a complex K it is natural to consider in succession the sections $K^1, K^2, \cdots$, where K^n consists of all the cells in K of at most n dimensions. Now the homotopy type of K^n is not an invariant of K. Therefore consider the n-type, this being a homotopy invariant of K, which depends only on K^n. Two complexes K, K' are of the same n-type if and only if there are maps $\phi\colon K^n \to K'^n$,

[1] Numbers in brackets refer to the list of references at the end of this paper.

$\phi' : K'^n \to K^n$ such that

$$\phi'\phi \mid K^{n-1} \simeq 1, \qquad \phi\phi' \mid K'^{n-1} \simeq 1,$$

where $\simeq$ indicates homotopy in K^n, K'^n and each 1 stands for the appropriate identical map $K^{n-1} \to K^n$ or $K'^{n-1} \to K'^n$. Subject to this condition we write $\phi : K^n \equiv {}_{n-1}K'^n$. It is the fact that the above homotopies are in K^n, K'^n, not in K^{n-1}, K'^{n-1}, which makes the n-type a homotopy invariant. It is proved in [4], assuming K, K' to be connected, that $\phi : K^n \equiv {}_{n-1}K'^n$ if and only if $\phi_r : \pi_r \approx \pi'_r$ for $r = 1, \cdots, n-1$, where ϕ_r is induced by ϕ. Hence it follows that K, K' are of the same n-type if they are of the same m-type for any $m > n$, where $m \leqq \infty$ and the ∞-type means the homotopy type. Therefore the n-type is a homotopy invariant and a fortiori a topological invariant of the space covered by K. Within the category of at most $(n-1)$-dimensional complexes the n-type is the same as the homotopy type.

All connected complexes are of the same 1-type, so that the interest begins with $n = 2$. Any homomorphism $\pi_1 \to \pi'_1$ has a geometrical realization $K^2 \to K'^2$. Therefore it follows from the theorem quoted in the preceding paragraph (or from an easy ad hoc argument) that two complexes K, K' are of the same 2-type if and only if $\pi_1 \approx \pi'_1$. Moreover, given a group G, there is a complex K such that $\pi_1 \approx G$. Thus π_1 is an "algebraic equivalent" of the 2-type. It is natural to consider the problem of finding an algebraic equivalent, T^n, of the n-type. I shall explain what has been done in case $n = 3$ or, for simply connected complexes, in case $n = 4$. This will indicate what I mean by an algebraic equivalent of the n-type.

Let Q be an arbitrary (multiplicative) group and G an arbitrary (additive) group. Let $\eta : Q \to A(G)$ be a homomorphism of Q into the group, $A(G)$, of automorphisms of G, by means of which Q is expressed as a group of operators on G. Using these operators we define the nth cohomology group, $H^n(Q, G)$, of Q, with coefficients in G. Then an *algebraic 3-type* is defined as a quadruple, $T^3 = (Q, G, \eta, \mathbf{k})$, where $\mathbf{k}$ is an arbitrary element of $H^3(Q, G)$. Let $T'^3 = (Q', G', \eta', \mathbf{k}')$ be any algebraic 3-type. Then a homomorphism $F : T^3 \to T'^3$ consists of a pair of homomorphisms $f : Q \to Q'$, $h : G \to G'$ such that

$$h\eta(q) = \eta'(fq)h : G \to G'$$

$$f^*\mathbf{k}' = h_*\mathbf{k} \in H^3(Q, G'),$$

where $f^* : H^3(Q', G') \to H^3(Q, G')$ and $h_* : H^3(Q, G) \to H^3(Q, G')$ are induced by f, h. If f, h are isomorphisms (onto) we write $F : T^3 \approx T'^3$. It is proved in [3] that:

(1) *Any complex, K, determines an algebraic 3-type, $T^3(K)$, in which $Q = \pi_1$, $G = \pi_2$, η is the homomorphism determined by the way in which π_1 operates, as usual, on π_2 and $\mathbf{k}$ is the Eilenberg-MacLane invariant* (see [2] or [3]).

(2) *The homomorphisms $f : \pi_1 \to \pi'_1$, $h : \pi_2 \to \pi'_2$, which are induced by a map $K \to K'$, satisfy the above conditions for a homomorphism $T^3(K) \to T^3(K')$.*

(3) *Any homomorphism* $T^3(K) \to T^3(K')$ *has a geometrical realization* $K \to K'$.

(4) *If* T^3 *is given, there is a complex,* K, *such that* $T^3 \approx T^3(K)$.

In consequence of these conditions an algebraic 3-type may be described as an algebraic equivalent of the 3-type of a complex.

Now consider the category of simply connected complexes. Given any (additive) Abelian group, A, we define a group $\Gamma(A)$ as follows. It has a set of generators $\gamma(a)$, which is indexed to A, and these are subject to the (complete) set of relations:

$$\gamma(-a) = \gamma(a)$$

$$\gamma(a+b+c) - \gamma(b+c) - \gamma(c+a) - \gamma(a+b) + \gamma(a) + \gamma(b) + \gamma(c) = 0.$$

A homomorphism $f: A \to A'$, where A' is Abelian, induces a homomorphism $\Gamma f: \Gamma(A) \to \Gamma(A')$, which is given by $\Gamma f \gamma(a) = \gamma(fa)$. Then a (simply connected) algebraic 4-type is a triple $T_0^4 = (A, B, \mathfrak{i})$, where A, B are arbitrary Abelian groups and $\mathfrak{i}: \Gamma(A) \to B$ is an arbitrary homomorphism. A homomorphism $F: T_0^4 \to T_0'^4$, where $T_0'^4 = (A', B', \mathfrak{i}')$ consists of a pair of homomorphisms $f: A \to A'$, $g: B \to B'$, such that the diagram

$$\begin{array}{ccc} \Gamma(A) & \xrightarrow{\mathfrak{i}} & B \\ \downarrow{\scriptstyle \Gamma f} & & \downarrow{\scriptstyle g} \\ \Gamma(A') & \xrightarrow{\mathfrak{i}'} & B' \end{array}$$

is commutative.

Let K be a simply connected complex and let $\Gamma_3(K)$ be the image of $\pi_3(K^2)$ in the injection $\pi_3(K^2) \to \pi_3(K^3)$. Though apparently combinatorial in character, the group $\Gamma_3(K)$ is a topological invariant of K. In fact it is an invariant of the 3-type of K. Moreover it is the image of $\Gamma(\pi_2)$ in a natural isomorphism $\theta: \Gamma(\pi_2) \approx \Gamma_3(K)$. Let $\mathfrak{i}': \pi_3(K^3) \to \pi_3$ be the injection, let $\mathfrak{i}': \Gamma_3(K) \to \pi_3$ be the homomorphism determined by $\mathfrak{i}'$ and let

$$\mathfrak{i} = \mathfrak{i}'\theta: \Gamma(\pi_2) \to \pi_3 .$$

It is proved in [5] that, within the category of simply connected complexes, $T_0^4 = (\pi_2, \pi_3, \mathfrak{i})$ is an algebraic equivalent of the 4-type of K.

It should be stated that, in these cases, the relation between algebra and geometry is incomplete. For it cannot be stated, as it can in case $n = 2$, that maps $\phi_0, \phi_1 : K^n \to K'^n$ $(n = 3, 4)$, which induce the same homomorphism $T^n(K) \to T^n(K')$, are $(n-1)$-homomorphic, meaning that $\phi_0 \mid K^{n-1} \simeq \phi_1 \mid K^{n-1}$ in K'^n.

References

1. S. Eilenberg and Saunders MacLane, *General theory of natural equivalences*, Trans. Amer. Math. Soc. vol. 58 (1945), pp. 231–294.

2. ———, *Determination of the second homology and cohomology groups of a space by means of homotopy invariants*, Proc. Nat. Acad. Sci. U. S. A. vol. 32 (1946) pp. 277–280.

3. SAUNDERS MACLANE and J. H. C. WHITEHEAD, *On the 3-type of a complex*, Proc. Nat. Acad. Sci. U. S. A. vol. (1950) 36 pp. 41–48.

4. J. H. C. WHITEHEAD, *Combinatorial homotopy* I, Bull. Amer. Math. Soc. vol. 55 (1949) pp. 213–245.

5. ———, *A certain exact sequence*, Ann. of Math. vol. 52 (1950) pp. 51–110.

OXFORD UNIVERSITY,
OXFORD, ENGLAND.

HOMOTOPY GROUPS OF SPHERES

GEORGE W. WHITEHEAD

The homotopy groups of a topological space, discovered by Hurewicz[1] in 1935, have become an important tool in studying the classification and extension problems. The homotopy groups of a space have many properties in common with the homology groups; an important difference is that the homotopy groups do not have the excision property.

A special case of the excision problem is the following. Let S^n be an n-sphere, and let S^{n-1} be an equator dividing S^n into two "hemispheres", E^n_+ and E^n_-. Then the excision homomorphism maps $\pi_i(E^n_+, S^{n-1})$ into $\pi_i(S^n, E^n_-)$. Now the boundary operator maps $\pi_i(E^n_+, S^{n-1})$ isomorphically onto $\pi_{i-1}(S^{n-1})$, while $\pi_i(S^n)$ is mapped isomorphically onto $\pi_i(S^n, E^n_-)$ by the inclusion map. Hence the excision homomorphism induces a homomorphism $E\colon \pi_{i-1}(S^{n-1}) \to \pi_i(S^n)$. This homomorphism is the Freudenthal *suspension.*[2] This suspension operator has proved to be an important tool in studying $\pi_i(S^n)$.

The groups $\pi_i(S^n)$ have long been known for $i \leqq n$; for $i < n$, $\pi_i(S^n) = 0$, while $\pi_n(S^n)$ is infinite cyclic. It is also known that $\pi_i(S^1) = 0$ for $i > 1$.

The first attack on $\pi_i(S^n)$ for $i > n$ was made by Hopf,[3] who defined a homomorphism H_0 of $\pi_{2n-1}(S^n)$ into the additive group of integers and proved that $\pi_{2n-1}(S^n)$ is not zero, and in fact contains an element of infinite order, provided that n is even. It was further shown by Hurewicz[1] that $\pi_3(S^2)$ is infinite cyclic, and in fact the homomorphism H_0 is an isomorphism of $\pi_3(S^2)$ onto the group of integers. Hopf proved that if n is odd, then the *Hopf invariant* $H_0(\alpha)$ is zero for every $\alpha \in \pi_{2n-1}(S^n)$; while if n is even, there exists $\alpha \in \pi_{2n-1}(S^n)$ with $H_0(\alpha) = 2$. If $n = 2$, 4, or 8, there is an element of $\pi_{2n-1}(S^n)$ with Hopf invariant 1. Hopf raised the question of whether $\pi_{2n-1}(S^n)$ contains an element of Hopf invariant 1 for other values of n. This question was partially settled by the author,[4] who showed that no such element exists if $n \equiv 2 \pmod 4$ and $n > 2$.

J. H. C. Whitehead[5] has defined an operation of multiplication between the homotopy groups of a topological space X. This operation associates with each $\alpha \in \pi_p(X)$, $\beta \in \pi_q(X)$, an element $[\alpha, \beta] \in \pi_{p+q-1}(X)$. This operation is bilinear if $p > 1$ and $q > 1$; while if p or $q = 1$ it can be expressed in terms of the operations of the fundamental group on the higher homotopy groups.[6] The bracket product has the following additional properties:

[1] W. Hurewicz, Proceedings of the K. Akademie van Wetenschappen Amsterdam vol. 38 (1935) pp. 112–119, 521–528; vol. 39 (1936) pp. 117–126, 215–224.

[2] H. Freudenthal, Compositio Math. vol. 5 (1937) pp. 299–314.

[3] H. Hopf, Math. Ann. vol. 104 (1931) pp. 637–665; Fund. Math. vol. 25 (1935) pp. 427–440.

[4] G. W. Whitehead, Ann. of Math. vol. 51 (1950) pp. 192–237.

[5] J. H. C. Whitehead, Ann. of Math. vol. 42 (1941) pp. 409–428.

[6] S. Eilenberg, Fund. Math. vol. 32 (1939) pp. 167–175.

1) $[\beta, \alpha] = (-1)^{pq}[\alpha, \beta]$;

2) if ι_r is the element of $\pi_r(S^r)$ represented by the identity map, then

$$H_0([\iota_r, \iota_r]) = \begin{cases} 2 & (r \text{ even}) \\ 0 & (r \text{ odd}). \end{cases}$$

The groups $\pi_i(S^n)$ with $n < i < 2n - 1$ were studied by Freudenthal.[2] The Freudenthal theorems, with some improvements due to the author,[4] can be stated as follows:

(1) $E: \pi_i(S^n) \to \pi_{i+1}(S^{n+1})$ is onto if $i \leqq 2n - 1$;

(2) $E: \pi_i(S^n) \to \pi_{i+1}(S^{n+1})$ is an isomorphism if $i < 2n - 1$;

(3) the image of $E: \pi_{2n-2}(S^{n-1}) \to \pi_{2n-1}(S^n)$ is the kernel of H_0;

(4) the kernel K of $E: \pi_{2n-1}(S^n) \to \pi_{2n}(S^{n+1})$ is the subgroup generated by $[\iota_n, \iota_n]$; (a) if n is even, K is infinite cyclic; (b) if n is odd, K is zero or cyclic of order two, according as $\pi_{2n+1}(S^{n+1})$ contains an element of Hopf invariant 1 or not.

Any generalization of (3) would seem to require a generalization of the Hopf homomorphism H_0. A partial generalization of H_0 was introduced by the author,[4] who defined a homomorphism $H: \pi_n(S^r) \to \pi_n(S^{2r-1})$ for each $n < 3r - 3$ as follows.

Denote by $S^r \vee S^r$ the subset of $S^r \times y_0 \cup y_0 \times S^r$ of $S^r \times S^r$; $S^r \vee S^r$ is the union of two r-spheres with a point in common. Let $\phi: S^r \to S^r \vee S^r$ be a map which pinches S^{r-1} to the point $y_0 \times y_0$ and maps each open hemisphere topologically. Then ϕ induces a homomorphism $\phi_*: \pi_n(S^r) \to \pi_n(S^r \vee S^r)$.

In the homotopy sequence of the pair $(S^r \times S^r, S^r \vee S^r)$, the boundary homomorphism $\partial: \pi_{n+1}(S^r \times S^r, S^r \vee S^r) \to \pi_n(S^r \vee S^r)$ is an isomorphism into and the inclusion homomorphism $i_*: \pi_n(S^r \vee S^r) \to \pi_n(S^r \times S^r)$ is onto; moreover, $\pi_n(S^r \vee S^r)$ is in a natural way the direct sum of Image ∂ with a subgroup mapped isomorphically by i_*. Let $Q: \pi_n(S^r \vee S^r) \to$ Image ∂ be the projection associated with this direct sum decomposition.

Let j', j'' be the inclusion maps of $S^r \times y_0$ and $y_0 \times S^r$ into $S^r \vee S^r$. Then j'_* and j''_* are isomorphisms into, and $j'_*\pi_r(S^r \times y_0)$ and $j''_*\pi_r(y_0 \times S^r)$ are infinite cyclic groups generated by ι'_r, ι''_r.

It turns out that if $f: S^{2r-1} \to S^r \vee S^r$ is a map representing $[\iota'_r, \iota''_r]$, then f_* maps $\pi_n(S^{2r-1})$ isomorphically onto Image ∂ for $n \leqq 3r - 3$. We then define $H: \pi_n(S^r) \to \pi_n(S^{2r-1})$ by

$$H = f_*^{-1}\partial^{-1}Q\phi_*.$$

If $n = 2r - 1$, then H maps $\pi_{2r-1}(S^r)$ into the infinite cyclic group $\pi_{2r-1}(S^{2r-1})$ and $H(\alpha) = H_0(\alpha)\, \iota_{2r-1}$. Thus H is a generalization of H_0. The definition of H has been improved and extended by Blakers and Massey[7] and by Hilton.[8]

[7] A. L. Blakers and W. S. Massey, Proc. Nat. Acad. Sci. U. S. A. vol. 35 (1949) pp. 322-328. A full account will appear in Ann. of Math.

[8] Unpublished.

The attempt to generalize Freudenthal's theorems, based on the definition of H, has met with only partial success. It turns out that $H \cdot E = 0$, so that Kernel $H \supset$ Image E, but it is not known whether the opposite inclusion is true. Only a partial generalization of (4) has been obtained: if $\alpha \in \pi_n(S^r)$ and $E\alpha = 0$, then $H(\alpha) = 0$ if r is odd and $H(\alpha) \in 2\pi_n(S^{2r-1})$ if r is even. These facts are used by the author in the partial solution to Hopf's problem mentioned earlier, as well as in the proof[9] of the fact that $\pi_{n+2}(S^n)$ is of order two for $n \geqq 2$. (The latter theorem was proved independently by Pontrjagin[10] by entirely different methods.)

Another operation involving homotopy groups of spheres is that of composition. If X is a topological space, and if $f\colon S^n \to S^r$, $g\colon S^r \to X$ are maps, then $g \circ f\colon S^n \to X$, and the operation $(f, g) \to g \circ f$ induces an operation on homotopy classes, associating with each $\alpha \in \pi_n(S^r)$, $\beta \in \pi_r(X)$ an element $\beta \circ \alpha \in \pi_n(X)$. For this operation the left distributive law

$$\beta \circ (\alpha_1 + \alpha_2) = \beta \circ \alpha_1 + \beta \circ \alpha_2$$

is valid, but right distributivity does not in general hold. The deviation from right distributivity is expressed by the formula[4]

$$(\beta_1 + \beta_2) \circ \alpha = \beta_1 \circ \alpha + \beta_2 \circ \alpha + [\beta_1, \beta_2] \circ H(\alpha)$$

provided that $n < 3r - 2$. From this we can deduce the following identities:

$$(\beta_1 + \beta_2) \circ \alpha - \beta_1 \circ \alpha - \beta_2 \circ \alpha = 0 \qquad \text{if } n < 2r - 1;$$

$$(\beta_1 + \beta_2 + \beta_3) \circ \alpha - (\beta_1 + \beta_2) \circ \alpha - (\beta_1 + \beta_3) \circ \alpha$$
$$- (\beta_2 + \beta_3) \circ \alpha + \beta_1 \circ \alpha + \beta_2 \circ \alpha + \beta_3 \circ \alpha = 0 \qquad \text{if n } < 3r - 2.$$

It is not known whether the identities suggested by these formulas hold if $n \geqq 3r - 2$.

The homotopy groups of a space X were originally defined by Hurewicz[1] as the fundamental groups of certain function spaces over X. Let $y_0 \in S^n$ and let $F^n(X, x_0)$ be the space of all maps of (S^n, y_0) into (X, x_0), topologized by the compact-open topology. Let $\alpha \in \pi_n(X)$ and let $F^n_\alpha(X, x_0)$ be the path-component of $F^n(X, x_0)$ consisting of all maps which represent the element $\alpha \in \pi_n(X)$. Hurewicz proved that all the spaces $F^n_\alpha(X, x_0)$ have the same homotopy type, and that $\pi_k(F^n(X, x_0)) \approx \pi_{n+k}(X)$. One may therefore attempt to study the homotopy groups of X by examining the function spaces $F^n_\alpha(X, x_0)$.

The case $n = 1$, $\alpha = 0$, $X = S^p$ has recently been studied by Pitcher.[11] The space $F^1_0(S^p, x_0)$ is the space of closed paths in S^p starting and ending at x_0; this space has the same homotopy type as the space of paths joining two fixed

[9] G. W. Whitehead, Ann. of Math. vol. 52 (1950) pp. 245–247.
[10] L. Pontrjagin, C. R. Acad. Sci. URSS vol. 70 (1950) pp. 957–959.
[11] E. Pitcher, Bull. Amer. Math. Soc. Abstract 56-1-38.

points of S^p. Pitcher attacks the problem by means of the powerful tools of critical point theory in the calculus of variations.

Suppose that $n = p$, $\alpha = \iota_p$. Then the rotation group R_{p-1} of S^{p-1} is naturally imbedded in $F^p_{\iota_p}(S^p, x_0)$. Thus there is a natural homomorphism $J\colon \pi_q(R_{p-1}) \to \pi_{p+q}(S^p)$. This homomorphism was used by Hopf[3] in his study of $\pi_{2p-1}(S^p)$. The homomorphism J turns out to be an isomorphism onto if $q = 1$, but for $q = 2$, J is no longer onto since $\pi_2(R_{p-1}) = 0$ and $\pi_{p+2}(S^p) \neq 0$. However, the nonzero element of $\pi_{p+2}(S^p)$ can be obtained from the rotation group in another way.

Since R_{p-1} is a topological group, the set of homotopy classes of maps of $S^q \times S^r$ into R_{p-1} forms in a natural way a group $\pi_{q,r}(R_{p-1})$. There is a map J' of $\pi_{q,r}(R_{p-1})$ into $\pi_{p+q+r}(S^p)$, analogous to J, defined as follows. Let S^{p+q+r} be referred to coordinates (u, v, w) where $u \in E^{p-1}$, $v \in E^{q+1}$, $w \in E^{r+1}$ and $|u|^2 + |v|^2 + |w|^2 = 1$. The subset C of S^{p+q+r} defined by the equation

$$|v|^2 + |w|^2 = 8\,|v|^2\,|w|^2$$

is a homeomorphic copy of $S^{p-1} \times S^q \times S^r$, separating S^{p+q+r} into two subsets A, B. If $f\colon S^q \times S^r \to R_{p-1}$, then f defines a map $f'\colon S^{p-1} \times S^q \times S^r \to S^{p-1}$ by

$$f'(x\,y,\,z) = f(y,\,z)(x).$$

Then f' may be considered as a map of C into S^{r-1}, and can be extended to a map $f''\colon S^{p+q+r} \to S^r$ such that $f''(\bar{A}) \subset E^r_+$, $f''(\bar{B}) \subset E^r_-$. If f is homotopic to f_1, then f'' is homotopic to f_1'', and therefore the correspondence $f \to f''$ induces a map $J'\colon \pi_{q,r}(R_{p-1}) \to \pi_{p+q+r}(S^p)$.

The group $\pi_{q,r}(R_{p-1})$ has the following structure. The set of homotopy classes of maps of the form $f(x, y) = g(x)$, where $g\colon S^q \to R_{p-1}$, forms a group isomorphic with $\pi_q(R_{p-1})$. Similarly, $\pi_{q,r}(R_{p-1})$ contains an isomorphic copy of $\pi_r(R_{p-1})$. The set of homotopy classes of maps of the form $f\colon (S^q \times S^r, S^q \wedge S^r) \to (R_{p-1}, r_0)$ forms a third subgroup isomorphic with $\pi_{q+r}(R_{p-1})$. Then each element of $\pi_{q+r}(R_{p-1})$ is uniquely expressible in the form $\alpha\beta\gamma$, where $\alpha \in \pi_q(R_{p-1})$, $\beta \in \pi_r(R_{p-1})$, and $\gamma \in \pi_{q+r}(R_{p-1})$; the subgroup $\pi_{q+r}(R_{p-1})$ is contained in the center of $\pi_{q,r}(R_{p-1})$ and the factor group is isomorphic with $\pi_q(R_{p-1}) \times \pi_r(R_{p-1})$. In general, however, $\pi_{q,r}(R_{p-1})$ is not abelian.

The group R_1 is naturally homeomorphic with S^1; the group multiplication in R_1 defines a map $f\colon S^1 \times S^1 \to R_1$ representing an element $\alpha \in \pi_{1,1}(R_1)$. It turns out that $J'(\alpha)$ is the nonzero element of $\pi_4(S^2)$.

We may consider R_{p-1} as a subgroup of R_p; the inclusion map defines a homomorphism $\beta\colon \pi_{q,r}(R_{p-1}) \to \pi_{q,r}(R_p)$. Then the commutativity relation $J'\circ\beta = E\circ J'$, suggested by the diagram

$$\begin{array}{ccc} \pi_{q,r}(R_{p-1}) & \xrightarrow{\beta} & \pi_{q,r}(R_p) \\ \downarrow J' & & \downarrow J' \\ \pi_{p+q+r}(S^p) & \xrightarrow{E} & \pi_{p+q+r+1}(S^{p+1}) \end{array}$$

holds. Hence the nonzero element of $\pi_{p+2}(S^p)$ can be obtained from $\pi_{1,1}(R_{p-1})$ by means of the operation J'.

The operation J' has the following properties:

(1) $J' \mid \pi_{q+r}(R_{p-1}) = J$;

(2) $J'(\pi_q(R_{p-1})) = J'(\pi_r(R_{p-1})) = 0$;

(3) if $\alpha \in \pi_q(R_{p-1})$, $\beta \in \pi_r(R_{p-1})$, then the element $J'(\alpha \circ \beta)$ depends bilinearly on α and β.

In particular, J' is not in general a homomorphism.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY,
CAMBRIDGE, MASS., U. S. A.

FIBRE BUNDLES AND OBSTRUCTIONS

THE THEORY OF OBSTRUCTIONS

PAUL OLUM

1. Introduction. In the study of the mappings of one topological space into another, there are a number of rather different problems which seem to be most effectively studied by a certain common technique. Examples are the extension problem, the homotopy problem, the deformation problem (that is, to deform a single given mapping so that after the deformation it behaves in some prescribed fashion), and the cross-section or "lifting" problem in fibre bundles. The common technique in question is the method of "obstructions." An obstruction is simply the algebraic expression, in terms of cohomology theory, of the difficulty one runs into in trying to perform the required extension or homotopy or deformation or lifting. The first explicit formulation of the notion of an obstruction is due to Eilenberg [1], although the method had been used implicitly earlier by Whitney [6] in his restatement and proof of the Hopf mapping theorem in cohomology terms.

In view of the limitations of time, we shall confine ourselves here to a rather specific problem: Given two arcwise connected spaces X, Y with chosen base points x_0 , y_0 , we consider mappings $X \to Y$ which carry x_0 into y_0 and study their homotopy relative to x_0 . What I shall have to say about obstructions for this problem will be fairly representative also of the situation for free-homotopy or homotopy relative to an arbitrary connected subspace, and for the other problems mentioned earlier, although there are, of course, certain important differences.

We consider then two mappings f_0 , $f_1 : X \to Y$ with $f_0(x_0) = f_1(x_0) = y_0$. I propose first of all to define the nth obstruction to a homotopy of f_0 to f_1 relative to x_0 , which we shall denote $\mathcal{O}^n(f_0 , f_1)$ rel. x_0 , then to indicate a few of the most important properties of these obstructions; finally, in the remaining time, I shall describe some new results in an essentially classical vein—having to do with mappings of manifolds and the notion of degree—which are proved using the obstruction technique.

2. Cohomology over a local group. We begin by supposing there is given a local system of groups $\mathcal{G} = \{G_y\}$ in the space Y, as used by Steenrod in his paper on homology with local coefficients [4]. A mapping $f : X \to Y$ induces a local system $\mathcal{G}(f)$ in X, by assigning to each $x \in X$ the group $G_{f(x)}$ and to each path in X the isomorphism attached to the image path in Y. With this, we may then consider the cohomology groups $H^n(Y, \mathcal{G})$ and $H^n(X, \mathcal{G}(f))$.

We introduce one further definition here. By a *local group G at the point y_0* in Y we shall mean a group G together with an operation of the fundamental group $\pi_1(Y, y_0)$ on it. Clearly, in the local system $\mathcal{G} = \{G_y\}$, the group at the

base point G_{y_0} may be regarded as a local group at y_0. Now it is easily seen that the cohomology group $H^n(Y, \mathfrak{G})$ does not really depend on the full spread of the local system $\mathfrak{G}$ but is completely determined to within a unique isomorphism by the local group G_{y_0} carried by $\mathfrak{G}$. This leads (we are omitting the obvious formal details) to the definition of cohomology $H^n(Y, G_{y_0})$ over the local group G_{y_0}, and this cohomology group is isomorphic in a unique way to $H^n(Y, \mathfrak{G})$.

Now it is to be observed that the local group at x_0 in X carried by $\mathfrak{G}(f)$ does not depend on all of $\mathfrak{G}$ and on f itself, but only on the local group G_{y_0} and on the homomorphism $\theta: \pi_1(X, x_0) \to \pi_1(Y, y_0)$ determined by f. Let us then denote this local group at x_0 in X, induced by θ, by $\theta^* G_{y_0}$. Thus we are led to consider the group $H^n(X, \theta^* G_{y_0})$ which is related to $H^n(X, \mathfrak{G}(f))$ by a unique isomorphism.

As is well-known, the homotopy groups $\pi_n(Y, y)$ form a local system of groups in Y and this is the particular case of the local system $\mathfrak{G}$ which concerns us. Given any $\theta: \pi_1(X, x_0) \to \pi_1(Y, y_0)$, we have induced in X the local group $\theta^* \pi_n(Y, y_0)$, which we shall shorten simply to $\theta^* \pi_n$, and we may then consider the cohomology group $H^n(X, \theta^* \pi_n)$. It is this cohomology group which appears to be the appropriate algebraic vehicle for carrying the obstruction technique.

3. Definition of the obstruction. We now proceed to define the obstruction $\mathfrak{O}^n(f_0, f_1)$ rel. x_0 which, as we shall see, will turn out to be a subset (possibly a *void* subset) of the group $H^n(X, \theta_0^* \pi_n)$, where θ_0 is the homomorphism induced by f_0. (Details concerning the material of this and the following section will be found in [3].)

This obstruction can be defined completely invariantly and under the most general circumstances, with essentially no limitations on the spaces X and Y under consideration other than arcwise connectedness; see [3]. Here again, however, in the interests of simplicity and time, I shall suppose that X is triangulable and has been decomposed to form a simplicial complex K; it will be assumed also that the vertices of K are partially ordered so that the vertices of any simplex are simply ordered.

We consider then the cartesian product space $I \times X = \bar{X}$, where I is the unit interval, and suppose it also triangulated so that the triangulation (including the partial ordering of vertices) agrees with that already given on $0 \times X$ and $1 \times X$. We identify X with $0 \times X$, denote by $\bar{K}$ the complex of $\bar{X} = I \times X$ and by K_1 the complex of $1 \times X$. We define a mapping $\bar{f}_{01}: X \cup (I \times x_0) \cup (1 \times X) \to Y$ as follows: We map X by f_0, $1 \times X$ by f_1, and we map $I \times x_0$ all into y_0.

Now suppose $\bar{f}$ is an extension of this partial mapping $\bar{f}_{01}$ over $\bar{X}^n$, the space of the n-skeleton of $\bar{X}$, where $n \geqq 2$. (Of course, there may be no such $\bar{f}$). Let us consider any $(n+1)$-simplex $\bar{\sigma}_{n+1}$ of $\bar{X}$ with leading vertex $\bar{x}$. The mapping $\bar{f}$ confined to the boundary of $\bar{\sigma}_{n+1}$ then defines an element of $\pi_n(Y, \bar{f}(\bar{x}))$, and, in particular, for $\bar{\sigma}_{n+1} \in K$ or K_1, the element so defined is clearly the zero element of the group. Consequently, if $\pi_n(\bar{f})$ denotes the local system of groups in $\bar{X}$ induced by $\bar{f}$ from the local system $\pi_n(Y, y)$, then this means that $\bar{f}$ defines an

element, a cochain, of $C^{n+1}(\bar{K}, K \cup K_1, \pi_n(\bar{f}))$. (Actually, the local system $\pi_n(\bar{f})$ is not defined over all of $\bar{X}$, only on the subspace $\bar{X}^n$, but this is sufficient since for cohomology in $\bar{K}$ the local system need be given only on the 2-skeleton.)

We come now to the essential observation that this cochain is indeed a cocycle. This is, of course, analogous to the corresponding result in Eilenberg's original work; see [1]. It is basically a consequence of the definition of addition in the homotopy groups and of the isomorphisms of homotopy groups associated with paths in Y. In this way then $\bar{f}$ defines a cocycle and consequently an element $\bar{h}^{n+1}(\bar{f})$ of $H^{n+1}(\bar{K}, K \cup K_1, \pi_n(\bar{f}))$.

Next, with the object of "pulling" this cohomology element down into the base space X, we look at the homomorphism:

$$\Delta\colon H^n(K, \pi_n(f_0)) \to H^{n+1}(\bar{K}, K \cup K_1, \pi_n(\bar{f}))$$

which is defined by the coboundary operator; we have used here also the fact that $\pi_n(\bar{f})$ confined to X is just the local system $\pi_n(f_0)$ induced by the mapping f_0. This Δ is easily proved to be an isomorphism onto by imbedding it in an appropriate exact sequence in which the third term is clearly zero. This being so, we may apply Δ^{-1} to the element $\bar{h}^{n+1}(\bar{f})$ and get an element $\Delta^{-1}(\bar{h}^{n+1}(\bar{f})) \in H^n(K, \pi_n(f_0))$. As we observed at the outset, $H^n(K, \pi_n(f_0)) \approx H^n(K, \theta_0^* \pi_n)$ where $f_0\colon X \to Y$ induces θ_0. Thus $\Delta^{-1}(\bar{h}^{n+1}(\bar{f}))$ determines a unique element in $H^n(K, \theta_0^* \pi_n)$ which we denote by $h^n(\bar{f})$.

Now consider the totality $\{h^n(\bar{f})\}$ of *distinct* elements so defined for all possible extensions $\bar{f}$ over $\bar{X}^n$. These elements form a subset (possibly void) of $H^n(K, \theta_0^* \pi_n)$ which we denote $\mathfrak{O}^n(f_0, f_1)$ rel. x_0 and call *the nth obstruction to a homotopy of f_0 to f_1 relative to x_0*. Note that $n \geqq 2$ here; the role of the obstruction for the lowest significant dimension, $n = 1$, is played by the homomorphisms of fundamental groups induced by the mappings, as is made clear by Theorem I below.

4. Properties of obstructions. The principal geometric significance of these obstructions is contained in the following theorem which makes clear their precise relation to the homotopy problem and which justifies calling them "obstructions":

THEOREM I. *For fixed n the following statements are equivalent. Each statement applies only for the indicated range of n:*

(a) $(n = 1)\ \theta_0 = \theta_1$; $(n \geqq 2)\ \mathfrak{O}^n(f_0, f_1)$ rel. x_0 *is nonvoid and contains the zero element.*

(b) $(n \geqq 1)\ f_0 \cong f_1$ *rel.* x_0 (dim. n).[1]

(c) $(n \geqq 1)\ \mathfrak{O}^{n+1}(f_0, f_1)$ *rel.* x_0 *is nonvoid.*

The principal algebraic property of obstructions is given by the following theorem which I shall call the "addition theorem" for obstructions:

[1] This means "$f_0 \mid X^n$ is homotopic to $f_1 \mid X^n$ relative to x_0" where X^n is the space of the n-skeleton of K.

THEOREM II. *Given* f_0, f_1, f_2, *suppose* $\mathcal{O}^n(f_0, f_1)$ *rel.* x_0 *and* $\mathcal{O}^n(f_1, f_2)$ *rel.* x_0 *are nonvoid (so that* $\theta_0 = \theta_1 = \theta_2$*) and contain, respectively,* h_0^n *and* h_1^n. *Then*

$$\begin{aligned}\mathcal{O}^n(f_0, f_2) \text{ rel. } x_0 &= h_0^n + \mathcal{O}^n(f_1, f_2) \text{ rel. } x_0 \\ &= \mathcal{O}^n(f_0, f_1) \text{ rel. } x_0 + h_1^n \\ &= \mathcal{O}^n(f_0, f_1) \text{ rel. } x_0 + \mathcal{O}^n(f_1, f_2) \text{ rel. } x_0 .\end{aligned}$$

With one exception, to which we shall return shortly, essentially everything of a general character about these obstructions and their relevance for the homotopy problem is immediately contained in Theorems I and II. Note here, in Theorem II, the importance of our using cohomology over the local group $\theta^*\pi_n$ rather than over the local system $\pi_n(f)$ induced by f; this is necessary in order for the addition in Theorem II to be defined.

We mention a few particular consequences contained in Theorems I and II:

(1) $\mathcal{O}^n(f_0, f_0)$ rel. x_0 is a *subgroup* of $H^n(K, \theta_0^*\pi_n)$.

(2) $\mathcal{O}^n(f_0, f_1)$ rel. x_0 is (if nonvoid) a *coset* of this subgroup. (Thus $\mathcal{O}^n(f_0, f_1)$ rel. x_0 and $\mathcal{O}^n(f_0, f_2)$ rel. x_0 are either disjoint or identical.)

(3) Suppose f_0, f_1, f_2 homotopic in dimension $n-1$ rel. x_0. Then $f_1 \cong f_2$ rel. x_0 (dim. n) if and only if

$$\mathcal{O}^n(f_0, f_1) \text{ rel. } x_0 = \mathcal{O}^n(f_0, f_2) \text{ rel. } x_0 .$$

This last statement has the importance that, out of a set of mappings homotopic in dimension $n-1$, homotopy in dimension n can be studied by comparing each of the mappings with a fixed one, f_0.

Returning now to the exception referred to just above, the other piece of information we need about obstructions is an existence theorem:

THEOREM III. *Given* $f_0: X \to Y$ *inducing* θ_0 *and* $h_0^n \in H^n(K, \theta_0^*\pi_n)$, *there exists* $f_1: X^{n+1} \to Y$ *such that* $\mathcal{O}^n(f_0, f_1)$ *rel.* x_0 *contains* h_0^n.

The proof of this theorem is given by taking a cocycle $z_0^n \in Z^n(K, \pi_n(f_0))$ representing h_0^n and then modifying f_0, keeping it fixed on X^{n-1}, but "altering" it inside of each n-cell σ_n, in a certain well-defined fashion, by the homotopy element $z_0^n(\sigma_n)$. The resulting mapping f_1 is extendable over X^{n+1} and is easily proved to have the property stated in the theorem.

The corresponding existence theorem in the lowest dimension says that for any θ there is a mapping of the 2-skeleton inducing it.

Property (3) above and this existence theorem combine to yield what may be called the general "one-stage" homotopy enumeration theorem, namely: If one considers a class of mappings of X^{n+1} into Y, all homotopic in dimension $n-1$ rel. x_0, and singles one of them, f_0, out for reference, then the homotopy classes in dimension n are in 1–1 correspondence with the elements of the factor group $H^n(K, \theta_0^*\pi_n)/\mathcal{O}^n(f_0, f_0)$ rel. x_0. This indicates the central role played by the "self-obstruction" $\mathcal{O}^n(f_0, f_0)$ rel. x_0 in the study of homotopy classifications. As an example, Steenrod's classification of the maps of an $(n+1)$-complex into

the n-sphere [5] amounts in essence to the determination, for any such mapping, of this self-obstruction, which he showed can be expressed as a certain cup-sub-i product.

There is one particular case in which, by the theorems above (and a simple extension theorem corresponding to I), we have the complete homotopy enumeration result already at hand. This is the case where $X = X^n$ is an n-polyhedron and $\pi_r(Y, y_0) = 0$ for $1 < r < n$. Then

THEOREM IV. *The homotopy classes rel. x_0 of mappings $X \to Y$ are in* 1–1 *correspondence with the pairs (θ, h) where θ is any homomorphism of the fundamental groups and $h \in H^n(X, \theta^*\pi_n)$.*

The corresponding result for free-homotopy is a little more complicated. It can be written down equally explicitly, but involves an additional invariant, the Eilenberg-MacLane invariant k^{n+1} for the space Y.

5. Remarks. Theorem IV, as it stands, is simply an enumeration theorem; it tells "how many" homotopy classes there are. However, with an appropriate re-interpretation it can be much more than that. The key to this is provided by the existence Theorem III described earlier. In the proof of this theorem, which was indicated very briefly above, one has not only the existence of a mapping f_1 so that a prescribed coset is realized as an obstruction, but also a rather specific description of f_1 as an "alteration" of f_0 by certain homotopy elements. This, combined with Theorem IV, then gives what we might call a "homotopy representation theorem"; i.e., an enumeration of homotopy classes together with the presentation of a particular representative mapping from each homotopy class.

One can then hope to gain a deeper insight into the properties which characterize the mappings of a given homotopy class by a direct examination of these representative mappings. In this sense I think a theorem like Theorem IV may be regarded not only as a result but also, and even primarily, as a starting point for studying homotopy properties of mappings.

6. Mappings of manifolds and the twisted degree. I would like to leave the general discussion of obstructions now and try to illustrate these last remarks by describing in the remainder of this paper some results connected with mappings of manifolds and the notion of degree. I shall not refer to obstructions again except to say now that the results to be described are in large part consequences of this homotopy representation theorem. Proofs and further details are contained in a paper which will appear later.

We consider two connected n-manifolds M^n and Q^n (assumed triangulable) and we study the mappings $M^n \to Q^n$ for the case where $\pi_r(Q^n, y_0) = 0$ for $1 < r < n$. All mappings carry $x_0 \in M^n$ into $y_0 \in Q^n$.

Before stating the relevant theorems, I want to extend somewhat Brouwer's

original definition of degree to what I shall call the *twisted* degree. This will be done here in invariant fashion; there is a corresponding extremely simple geometric interpretation which we shall not take time for.

For this definition, we take M^n and Q^n to be compact and make no restriction on the homotopy groups of Q^n. Let I_0^t denote the *local group of twisted integers* at y_0 in Q^n; this means the integers I_0 together with the following operation of the fundamental group $\pi_1(Q^n, y_0)$ on them: an orientation-preserving element operates trivially, an orientation-reversing element operates nontrivially.

Let $f: M^n \to Q^n$ induce θ. Then we may consider the cohomology groups $H^n(Q^n, I_0^t)$ and $H^n(M^n, \theta^* I_0^t)$, and f induces, in obvious fashion, also a homomorphism:

$$m_f: H^n(Q^n, I_0^t) \to H^n(M^n, \theta^* I_0^t). \tag{6.1}$$

Now $H^n(Q^n, I_0^t)$ is easily seen to be infinite cyclic. Let us suppose for the moment that the homomorphism θ is *orientation-true*, that is, it carries orientation-preserving elements into orientation-preserving elements, orientation-reversing into orientation-reversing. Then $\theta^* I_0^t$ is clearly just the local group of twisted integers at x_0 in M^n and the group on the right is also infinite cyclic. Assuming that one of the two possible generators has been selected once and for all for these nth cohomology groups over the twisted integers, then the homomorphism m_f is completely described by an integer which we call the *twisted degree* of f and denote $\deg f$.

This twisted degree has the usual properties, transitivity, etc. One remark, however: If $f_0 \cong f_1$, then $\deg f_0 = \deg f_1$ or $-\deg f_1$ according as the image of x_0, during the homotopy, traces out an orientation-preserving or orientation-reversing path in Q^n.

If the homomorphism θ is not orientation-true, then the group on the right in (6.1) becomes cyclic order 2 and the twisted degree is an integer mod 2. In this case it is identical with the ordinary mod 2 degree.

7. Homotopy properties of mappings of manifolds. We now return to our original problem of the mappings $M^n \to Q^n$ with $\pi_r(Q^n, y_0) = 0$ for $1 < r < n$.

If either M^n or $\tilde{Q}^n$ (the universal covering space of Q^n) is noncompact, then the classification of these mappings is particularly simple: Two mappings are homotopic rel. x_0 [or, free-homotopic] if and only if they induce the same homomorphism [respectively, homomorphism-class][2] of the fundamental group.

We may as well suppose then M^n and $\tilde{Q}^n$ compact. This means, in particular, that the fundamental group $\pi_1(Q^n, y_0)$ has finite order; let us call it l. Now let $f_0, f_1: M^n \to Q^n$ induce θ_0, θ_1. There are two cases here:

Case 1. *Orientation-true* θ's. For this case we have the following theorem:

THEOREM V. *$f_0 \cong f_1$ rel. x_0 if and only if $\theta_0 = \theta_1$ and $\deg f_0 = \deg f_1$. Further-*

[2] Two homomorphisms belong to the same homomorphism-class if they differ by an inner automorphism of the image group.

more, for each θ, the associated twisted degrees are all of the integers of a single congruence class mod l.

The corresponding result for free-homotopy is an immediate consequence of Theorem V and the remark made earlier on the behavior of the twisted degree under a homotopy. It is given, in detailed fashion, by

THEOREM Va. *$f_0 \cong f_1$ with the image of x_0 moving through a path of $\omega \in \pi_1(Q^n, y_0)$ if and only if $\theta_0(\gamma) = \omega\theta_1(\gamma)\omega^{-1}$ for all $\gamma \in \pi_1(M^n, x_0)$ and*

(a) *(for ω orientation-preserving)*: $\deg f_0 = \deg f_1$.

(b) *(for ω orientation-reversing)*: $\deg f_0 = -\deg f_1$.

Case 2. *Non-orientation-true* θ's. If the homomorphisms are not orientation-true, then we no longer have so satisfactory a characterization for the homotopy class of a mapping. However, this is largely compensated for by a very great reduction in the number of distinct homotopy classes:

THEOREM VI. *For each θ, there are exactly two distinct homotopy classes rel. x_0 inducing θ.*

In general, for this non-orientation-true case, the twisted degree (or, what is the same thing, the degree mod 2) will not distinguish between the two classes. However, if l is odd (which means Q^n is orientable), then it will do so for this case also.

In the corresponding case of free-homotopy there are either one or two homotopy classes for each homomorphism-class[2] Θ of the fundamental groups. The result here is, however, not an immediate consequence of Theorem VI, unless l is odd, and a particular computation is required to settle the matter in any given circumstance. This computation is quite explicit and straightforward, but we shall not consider it here.

8. Homotopy-type. An important application of Theorem V is to the problem of homotopy-type for manifolds M^n, Q^n satisfying $\pi_r(M^n, x_0) = \pi_r(Q^n, y_0) = 0$ for $1 < r < n$. It is easily seen that a homotopy equivalence between two manifolds must induce an *orientation-true* isomorphism of the fundamental groups so that Theorem V is the relevant one.

A particular case worth noting in this connection is that of the higher dimensional lens spaces. Given two $(2n + 1)$-dimensional lens spaces $L = L(m; q_1, \cdots, q_n)$ and $L' = L(m'; q_1', \cdots, q_n')$, then, for any homomorphism θ of their fundamental groups, it is quite easy to write down one particular mapping of L into L' inducing θ and to determine its degree. An immediate consequence then is the homotopy-type classification of the higher dimensional lens spaces:

THEOREM VII. *L and L' have the same homotopy-type if and only if $m = m'$ and*

$$q_1 q_2 \cdots q_n \equiv \pm k^2 q_1' \cdots q_n' \pmod{m}$$

for some k relatively prime to m.

9. The group-ring degree. I want to make one final remark here concerning the notion of the degree of a mapping. It is possible to generalize the degree still further, and I think quite fruitfully, by using in the definition given above in §6, instead of the local group I_0^t, the group-ring of the fundamental group of Q^n over the integers.

If one does so, he finds that the homomorphism (6.1) can again be characterized by an integer (or an integer mod 2 for the special case that θ carries some orientation-reversing element into the unit element) even though the group on the right in (6.1) will no longer be infinite cyclic. This integer might then be called the "group-ring degree" of the mapping. It agrees with the twisted degree if θ is orientation-true and also includes, in the general case, the Hopf notion of the "Absolutgrad" of a mapping [2]; in fact, the "Absolutgrad" is just the absolute value of the group-ring degree.

This group-ring degree will help to distinguish between the two homotopy classes (Theorem VI) in some, but by no means all, of the cases where the twisted degree fails to do so.

BIBLIOGRAPHY

1. S. EILENBERG, *Cohomology and continuous mappings*, Ann. of Math. vol. 41 (1940) pp. 231–251.

2. H. HOPF, *Zur Topologie der Abbildungen von Mannigfaltigkeiten*, II, Math. Ann. vol. 102 (1929) pp. 562–623.

3. P. OLUM, *Obstructions to extensions and homotopies*, Ann. of Math. vol. 52 (1950) pp. 1–50.

4. N. STEENROD, *Homology with local coefficients*, Ann. of Math. vol. 44 (1943) pp. 610–627.

5. ———, *Products of cocycles and extensions of mappings*, Ann. of Math. vol. 48 (1947) pp. 290–320.

6. H. WHITNEY, *The maps of an n-complex into an n-sphere*, Duke Math. J. vol. 3 (1937) pp. 51–55.

CORNELL UNIVERSITY,
ITHACA, N. Y., U. S. A.

HOMOTOPY GROUPS OF TRIADS

W. S. MASSEY

1. Introduction. One of the most important problems of topology today is to devise methods for computing the homotopy groups of topological spaces. The present paper is an attack on one phase of this general problem.

One method of attacking the problem of determining the homotopy groups of a connected cell complex K is first to try to determine the relative homotopy groups $\pi_p(K^q, K^{q-1})$, where K^n denotes the n-skeleton of K. A slightly more general problem is to try to determine the relative homotopy groups $\pi_p(X^*, X)$, where X is a pathwise connected topological space, and X^* is obtained by the adjunction of q-cells, ε_1^q, ε_2^q, $\cdots$, to X; i.e., $X^* = X \cup \varepsilon_1^q \cup \varepsilon_2^q \cup \cdots$. This general problem will be the main concern of this paper.

The analogous problem from homology theory, namely, to compute the relative homology groups, $H_p(K^q, K^{q-1})$, or $H_p(X^*, X)$, can be solved completely, and when applied to the problem of determining the homology groups of the complex K, leads to the usual algorithm for computing homology groups by means of chains, cycles, bounding cycles, etc. This comparison naturally leads one to ask why the homology groups $H_p(X^*, X)$ can be determined, while the problem of determining the homotopy groups $\pi_p(X^*, X)$ remains unsolved. The answer to this question is that the so-called "excision property" (see Eilenberg and Steenrod [2]) holds for homology groups, but does not hold for homotopy groups.

The excision property may be formulated as follows. Let $X = A \cup B$; then under rather general conditions the homomorphism $i_*: H_p(A, A \cap B) \to H_p(X, B)$, induced by the inclusion map $i: (A, A \cap B) \to (X, B)$, is an isomorphism onto in all dimensions. Simple examples may be given to show that this property does not hold when homology groups are replaced by homotopy groups, even if X is a finite, connected simplicial complex, and A, B, and $A \cap B$ are connected subcomplexes. Although the excision property does not hold in general for homotopy groups, under certain special conditions the homomorphisms $i_*: \pi_p(A, A \cap B) \to \pi_p(X, B)$ are isomorphisms onto for special values of p, and even when these homomorphisms fail to be isomorphisms onto, it is advantageous to determine as much information as possible about them. The theory developed in this paper enables one to fit the homomorphisms

$$i_*: \pi_p(A, A \cap B) \to \pi_p(X, B),$$

for $p = 2, 3, 4 \cdots$, into an exact sequence of homomorphisms in which the remaining groups are the new homotopy groups of the "triad", $(X; A, B)$, denoted by $\pi_p(X; A, B)$. The exact sequence in question is as follows:

$$\cdots \to \pi_{p+1}(X; A, B) \to \pi_p(A, A \cap B) \xrightarrow{i_*} \pi_p(X, B) \to \pi_p(X; A, B) \to \cdots.$$

This enables one to translate theorems about the homomorphisms

$$i_*: \pi_p(A, A \cap B) \to \pi_p(X, B)$$

into theorems about the triad homotopy groups and vice-versa.

One important case where a homomorphism induced by an excision has been studied is the following. Let S^n be an n-sphere, E^n_+ and E^n_- the upper and lower "hemispheres," and S^{n-1} the "equatorial" $(n-1)$-sphere, $S^{n-1} = E^n_+ \cap E^n_-$. Then the homomorphism $i_*: \pi_p(E^n_+, S^{n-1}) \to \pi_p(S^n, E^n_-)$ is equivalent (in a way to be explained later) to the "Einhängung", or suspension, $E: \pi_{p-1}(S^{n-1}) \to \pi_p(S^n)$, first studied by Freudenthal [3]. It was the study of this special case which first suggested to the author the idea for the homotopy groups of a triad.

The author wishes to acknowledge that all the results in this paper were obtained in collaboration with A. L. Blakers; this is actually a report on joint work.

2. Definition of the homotopy groups of a triad. We shall use the following notation for certain subsets of Cartesian n-space, C^n. The coordinates of a point $x \in C^n$ are denoted by $(x_1, \cdots, x_n)$, and $|x| = (x_1^2 + \cdots + x_n^2)^{1/2}$.

$$E^n = \{x \in C^n \mid |x| \leqq 1\}, \qquad S^{n-1} = \{x \in E^n \mid |x| = 1\},$$
$$E^{n-1}_+ = \{x \in S^{n-1} \mid x_n \geqq 0\}, \qquad E^{n-1}_- = \{x \in S^{n-1} \mid x_n \leqq 0\},$$
$$S^{n-2} = \{x \in S^{n-1} \mid x_n = 0\}, \qquad E^n_1 = \{x \in E^n \mid x_2 \geqq 0\},$$
$$E^n_2 = \{x \in E^n \mid x_2 \leqq 0\}, \qquad p_0 = (1, 0, \cdots, 0).$$

Let $(X; A, B)$ be a *triad*; that is, A and B are subspaces of the topological space X, and the intersection, $A \cap B$, is nonvoid. Choose a base point $x_0 \in A \cap B$. A *map*

$$f: (E^n; E^{n-1}_+, E^{n-1}_-, p_0) \to (X; A, B, x_0)$$

is a continuous function $f: E^n \to X$ such that $f(E^n_+) \subset A$, $f(E^n_-) \subset B$, and $f(p_0) = x_0$. Two such maps f_0, f_1 are *homotopic* if they are connected by a continuous 1-parameter family of maps,

$$f_t: (E^n; E^{n-1}_+, E^{n-1}_-, p_0) \to (X; A, B, x_0),$$

where $0 \leqq t \leqq 1$. With these definitions, it is not difficult to prove the following:

LEMMA 1. *If $n \geqq 3$, and $f: (E^n; E^{n-1}_+, E^{n-1}_n, p_0) \to (X; A, B, x_0)$ is a given map, then there exist homotopic maps f', f'', such that $f'(E^n_1) = f''(E^n_2) = x_0$.*

Denote the set of all homotopy classes of maps $(E^n; E^{n-1}_+, E^{n-1}_-, p_0) \to (X; A, B, x_0)$ by $\pi_n(X; A, B, x_0)$. For $n \geqq 3$, we define an addition between any two elements $\alpha, \beta \in \pi_n(X; A, B, x_0)$ as follows: choose maps f, g belonging to the homotopy classes α, β, respectively, with $f(E^n_2) = g(E^n_1) = x_0$. The existence of representatives f and g satisfying these conditions is guaranteed by Lemma 1.

Define $h\colon (E^n; E_+^{n-1}, E_-^{n-1}, p_0) \to (X; A, B, x_0)$ by $h \mid E_1^n = f \mid E_1^n$ and $h \mid E_2^n = g \mid E_2^n$. Then $\alpha + \beta$ is defined to be the homotopy class of h. With this definition of addition, $\pi_n(X; A, B, x_0)$ becomes a group, called the nth *homotopy group of the triad* $(X; A, B)$ *at the base point* x_0.

The following are some of the simpler properties of the groups thus defined:

(a) $\pi_n(X; A, B, x_0)$ is abelian for $n > 3$; simple examples show that it need not be so for $n = 3$.

(b) The system of groups $\pi_n(X; A, B, x)$ for $x \in A \cap B$ forms a local system of groups in the space $A \cap B$ in the sense of Steenrod, and $\pi_1(A \cap B, x_0)$ is a group of operators on $\pi_n(X; A, B, x_0)$.

(c) If $A \supset B$, then $\pi_n(X; A, B)$ is isomorphic to $\pi_n(X, A)$.

(d) A continuous map $f\colon (X; A, B, x_0) \to (X'; A', B', x_0')$ induces homomorphisms of the corresponding triad homotopy groups. We denote these by $f_*\colon \pi_p(X, A, B, x_0) \to \pi_p(X'; A', B', x_0')$.

3. The exact sequences of a triad. Associated with a triad $(X; A, B)$ are two *boundary homomorphisms*,

$$\partial_+\colon \pi_n(X; A, B, x_0) \to \pi_{n-1}(A, A \cap B, x_0),$$

$$\partial_-\colon \pi_n(X; A, B, x_0) \to \pi_{n-1}(B, A \cap B, x_0),$$

defined as follows. If $f\colon (E^n; E_+^{n-1}, E_-^{n-1}, p_0) \to (X; A, B, x_0)$ represents

$$\alpha \in \pi_n(X; A, B, x_0) \qquad (n \geqq 3),$$

then the maps

$$f_+\colon (E_+^{n-1}, S^{n-2}, p_0) \to (A, A \cap B, x_0),$$

$$f_-\colon (E_-^{n-1}, S^{n-2}, p_0) \to (B, A \cap B, x_0),$$

defined by f are representatives of the elements $\partial_+(\alpha) \in \pi_{n-1}(A, A \cap B, x_0)$ and $\partial_-(\alpha) \in \pi_{n-1}(B, A \cap B, x_0)$ respectively (assuming, of course, that a suitable convention is adopted to determine the orientations of the cells E_+^{n-1} and E_-^{n-1}).

Now consider the following sequence of groups and homomorphisms:

$$\cdots \xrightarrow{\partial_+} \pi_n(A, A \cap B) \xrightarrow{i_*} \pi_n(X, B) \xrightarrow{j_*}$$

$$\pi_n(X; A, B) \xrightarrow{\partial_+} \pi_{n-1}(A, A \cap B) \longrightarrow \cdots.$$

The homomorphisms i_* and j_* are induced by the inclusion maps

$$i\colon (A, A \cap B) \to (X, B), \qquad j\colon (X; x_0, B) \to (X; A, B).$$

This sequence is one of the two homotopy sequences of the triad $(X; A, B)$; the other sequence is obtained by interchanging the roles of A and B throughout.

THEOREM 1. *The homotopy sequences of a triad are both exact.*

The proof of this theorem requires only a straightforward verification of exactness, using the definitions already given.

In case $X = A \cup B$, the inclusion map i: $(A, A \cap B) \to (X, B)$ is an excision, and the triad homotopy groups may be considered to be a measure of the amount by which the relative homotopy groups fail to satisfy the excision axiom, as was indicated in the introduction.

4. The homotopy groups of the triad $(X^*; \mathcal{E}^n, X)$. For the remainder of this paper we shall assume that X^* is a Hausdorff space and X is a closed, pathwise connected subset of X^* such that there exists a decomposition

$$X^* - X = U_1^n \cup U_2^n \cup \cdots \cup U_k^n$$

of $X^* - X$ into k disjoint open sets U_i^n having the following property. Denote the closure of U_i^n by $\mathcal{E}_i^n$, and let $\dot{\mathcal{E}}_i^n = X \cap \mathcal{E}_i^n$, then it is assumed that there exist continuous mappings

$$\psi_i: (E^n, S^{n-1}) \to (\mathcal{E}_i^n, \dot{\mathcal{E}}_i^n), \qquad i = 1, \cdots, k,$$

which map $E^n - S^{n-1}$ homeomorphically onto U_i^n. It follows that S^{n-1} is mapped onto $\dot{\mathcal{E}}_i^n$ by ψ_i, but not necessarily homeomorphically. Under these conditions we shall say that X^* is obtained from X by the adjunction of the cells $\mathcal{E}_i^n, \cdots, \mathcal{E}_k^n$. This is exactly the process by means of which a general cell complex is built. The following notation will be used consistently:

$$\mathcal{E}^n = \mathcal{E}_1^n \cup \mathcal{E}_2^n \cup \cdots \cup \mathcal{E}_k^n, \qquad \dot{\mathcal{E}}^n = \dot{\mathcal{E}}_1^n \cup \cdots \cup \dot{\mathcal{E}}_k^n = X \cap \mathcal{E}^n.$$

We shall always make the additional assumption that $\dot{\mathcal{E}}^n$ is arcwise connected. This last condition is not nearly as restrictive as it appears on first sight, since J. H. C. Whitehead has shown that the homotopy type of the space X^* depends only on the homotopy class of the maps $\psi_i' : S^{n-1} \to X$ defined by the maps ψ_i: $(E^n, S^{n-1}) \to (\mathcal{E}_i^n, \dot{\mathcal{E}}_i^n)$, $i = 1, 2, \cdots, k$ (see [7] and [8]).

As stated in the introduction, our main problem is to determine the relative homotopy groups $\pi_p(X^*, X)$. This problem leads naturally to consideration of the excision homomorphism $\pi_p(\mathcal{E}^n, \dot{\mathcal{E}}^n) \to \pi_p(X^*, X)$, and hence to consideration of the homotopy sequence of the triad $(X^*; \mathcal{E}^n, X)$. It is the purpose of this section to state two theorems of a comparatively elementary nature which often give considerable information about this sequence, and hence about the groups $\pi_p(X^*, X)$.

First, we need some definitions. A topological space Y is said to be *n-connected* $(n > 0)$ if it is pathwise connected and $\pi_p(Y) = 0$ for $1 \leqq p \leqq n$. A pair (Y, A) is said to be *n-connected* $(n \geqq 1)$ if both Y and A are pathwise connected, the injection $\pi_1(A) \to \pi_1(Y)$ is a homomorphism onto, and $\pi_p(Y, A) = 0$ for $1 < p \leqq n$. Similarly, we shall say a triad $(Y; A, B)$ is *n-connected* $(n \geqq 2)$ if both of the pairs $(A, A \cap B)$ and $(B, A \cap B)$ are 1-connected, the injections $\pi_2(A, A \cap B) \to \pi_2(Y, B)$ and $\pi_2(B, A \cap B) \to \pi_2(Y, A)$ are homomorphisms onto, and $\pi_p(Y; A, B) = 0$ for $2 < p \leqq n$.

THEOREM 2. *If the pair* $(X, \dot{\mathcal{E}}^n)$ *is* m*-connected,* $m \geqq 1$ *and* $n > 2$*, then the triad* $(X^*; \mathcal{E}^n, X)$ *is* $(m + n - 1)$*-connected. (In case* $m = 1$*, it is also necessary to assume that* $\pi_2(X, \dot{\mathcal{E}}^n)$ *is abelian.)*

THEOREM 3. *If the space* $\dot{\mathcal{E}}^n$ *is* m*-connected,* $m \geqq 1$*, then the boundary homomorphism*

$$\partial: \pi_{p+1}(X^*; \mathcal{E}^n, X) \to \pi_p(\mathcal{E}^n, \dot{\mathcal{E}}^n)$$

is the zero homomorphism for $2 \leqq p \leqq m + n - 1$.

The proof of these theorems is too long to be given here; for full details, see [1]. Theorem 2 is closely related to what J. H. C. Whitehead calls the "crude suspension theorem" (see [9]) and can be proved by his methods. These methods are highly geometrical, and go back to Freudenthal [3]. The author's proofs make more use of the techniques of algebraic topology, and are not so geometrical in character. In particular, use is made of the basic results of the theory of "obstructions" to extensions and deformations of continuous maps.

As an example of the application of these results, consider the triad $(S^n; E_+^n, E_-^n)$. By applying Theorem 2, we see that this triad is $(2n - 2)$-connected. It follows from the exactness of the homotopy sequence of this triad that the excision homomorphism $i_*: \pi_p(E_+^n, S^{n-1}) \to \pi_p(S^n, E_-^n)$ is an isomorphism onto if $p < 2n - 2$, and is a homomorphism onto if $p = 2n - 2$. In this case, the excision homomorphism is equivalent to the suspension or "Einhängung" homomorphism introduced by Freudenthal [3]; in fact, in the following diagram

$$\begin{array}{ccc} \pi_{p-1}(S^{n-1}) & \xrightarrow{E} & \pi_p(S^n) \\ \partial\uparrow & & \downarrow k_* \\ \pi_p(E_+^n, S^{n-1}) & \xrightarrow{i_*} & \pi_p(S^n, E_-^n) \end{array}$$

the boundary homomorphism ∂, and the homomorphism k_* induced by an inclusion map, are isomorphisms onto, and the commutativity relation $i_* = k_* E \partial$ holds, where E is the suspension homomorphism. The statements above are precisely the easier part of the first two theorems of Freudenthal [3].

5. The next problem. Assume that the pair $(X, \dot{\mathcal{E}}^n)$ is m-connected, $m \geqq 1$, as in Theorem 2. Then we know that $\pi_p(X^*; \mathcal{E}^n, X) = 0$ for $p < m + n$. The next problem that arises is to determine the group $\pi_{m+n}(X^*; \mathcal{E}^n, X)$. The dimension $m + n$ will be called the first nontrivial dimension for the triad $(X^*; \mathcal{E}^n, X)$.

A few results have already been obtained in this direction. The Freudenthal suspension theorems, together with some improvements due to G. W. Whitehead [5], show that the group $\pi_{2n-1}(S^n; E_+^n, E_-^n)$ is infinite cyclic if n is odd, and strongly suggest that the same thing is true if n is even. J. H. C. Whitehead [6]

has obtained some theorems which amount to describing the groups and homomorphisms in the following portion of the exact sequence of the pair (X^*, X):

$$\cdots \xrightarrow{\partial} \pi_n(X) \xrightarrow{i_*} \pi_n(X^*) \xrightarrow{j_*} \pi_n(X^*, X) \xrightarrow{\partial} \pi_{n-1}(X) \to \cdots .$$

Although he was not able to determine the group $\pi_{n+1}(X^*, X)$ in general, his results are closely related to our results on the following part of the exact sequence of the triad $(X^*; \mathcal{E}^n, X)$ in case $m = 1$:

$$\cdots \to \pi_{n+1}(\mathcal{E}^n, \dot{\mathcal{E}}^n) \to \pi_{n+1}(X^*, X) \to \pi_{n+1}(X^*; \mathcal{E}^n, X) \to \cdots .$$

The rest of this paper will be devoted to the description of a general theorem about the homotopy group in the first nontrivial dimension of the triad $(X^*; \mathcal{E}^n, X)$. The proof of the results obtained is much more complicated than the proofs of either Theorems 2 or 3, and cannot be given here. In order to be able to state the main result, it is necessary first to develop several auxiliary concepts.

6. The homotopy groups of a covering. We use the word "covering" in a very special sense: A *covering* of a space X is an *ordered* pair (A, B) of subspaces of X such that $A \cup B = X$ and $A \cap B$ is nonvoid. We denote such a covering by (A/B).

Let (A/B) be a covering of X in this restricted sense. Choose a base point $x_0 \in A \cap B$. By a *map*, $f\colon (E^n_+/E^n_-, p_0) \to (A/B, x_0)$, is meant a continuous function $f\colon S^n \to X$ such that $f(E^n_+) \subset A$, $f(E^n_-) \subset B$, and $f(p_0) = x_0$. Homotopy of two such maps is defined in an obvious fashion. Let $\pi_n(A/B, x_0)$ denote the set of all homotopy classes of maps $(E^n_+/E^n_-, p_0) \to (A/B, x_0)$. If $n > 1$, it is possible to define an addition between the elements of $\pi_n(A/B, x_0)$ in such a way that $\pi_n(A/B, x_0)$ becomes a group, the nth *homotopy group of the covering* (A/B) *at the base point* x_0. This group is abelian if $n > 2$, but need not be so if $n = 2$. It has many properties which are the analogs of properties of the homotopy groups of pairs and triads.

Associated with a covering (A/B) of X are two exact sequences:

$$\cdots \to \pi_n(A/B) \to \pi_n(A, A \cap B) \to \pi_{n-1}(B) \to \pi_{n-1}(A/B) \to \cdots ,$$

$$\cdots \to \pi_n(A/B) \to \pi_n(B, A \cap B) \to \pi_{n-1}(A) \to \pi_{n-1}(A/B) \to \cdots .$$

The homomorphisms in this sequence are defined by means of "boundary operators" and inclusion maps, similar to those in the exact sequence of a pair or a triad.

If $(X; A, B)$ is an arbitrary triad, then the subspaces A and B determine a covering (A/B) of $A \cup B$. The homotopy groups of this covering fit into an exact sequence with the homotopy groups of a triad, as follows:

$$\cdots \to \pi_{n+1}(X; A, B, x_0) \xrightarrow{\partial} \pi_n(A/B, x_0) \xrightarrow{i_*} \pi_n(X, x_0) \xrightarrow{j_*} \pi_n(X; A, B) \to \cdots .$$

7. Generalized Whitehead products. In [6], J. H. C. Whitehead has defined a "product" which associates with elements $\alpha \in \pi_p(x)$ and $\beta \in \pi_q(X)$ an element $[\alpha, \beta] \in \pi_r(X)$, where $r = p + q - 1$, and proved several properties of this product, the most important of which is that it is bilinear if $p > 1$ and $q > 1$. In an analogous fashion, we can define three new products which are closely related to each other and to the original Whitehead product. One of these products is used in the statement of our main theorem below, and the others are useful in making applications of this theorem.

These three new products are as follows:

(a) Let (X, A) be an arbitrary pair. Then given elements $\alpha \in \pi_p(A, x_0)$ and $\beta \in \pi_q(X, A, x_0)$, the products $[\alpha, \beta]$ and $[\beta, \alpha]$ are elements of $\pi_r(X, A, x_0)$, where $r = p + q - 1$, and $x_0 \in A$ is the base point.

(b) Let (A/B) be a covering of $X = A \cup B$. Then if $\alpha \in \pi_p(A \cap B)$ and $\beta \in \pi_q(A/B)$, the products $[\alpha, \beta]$ and $[\beta, \alpha]$ are defined, and belong to $\pi_r(A/B)$, $r = p + q - 1$.

(c) Let $(X; A, B)$ be an arbitrary triad, $\alpha \in \pi_p(X, A \cap B)$, and $\beta \in \pi_q(A/B)$. Then $[\alpha, \beta]$ and $[\beta, \alpha]$ are elements of $\pi_r(X; A, B)$, $r = p + q - 1$.

All three of these products are bilinear, provided the integers p and q are large enough so that all the groups involved are abelian. All three of them satisfy the following commutation rule:

$$[\alpha, \beta] = (-1)^{pq}[\beta, \alpha].$$

The product mentioned in (a) above is defined as follows. Let

$$f\colon (E^p, S^{p-1}) \to (A, x_0),$$

$$g\colon (E^q, S^{q-1}, E_+^{q-1}) \to (X, A, x_0)$$

represent $\alpha \in \pi_p(A, x_0)$ and $\beta \in \pi_q(X, A, x_0)$ respectively (assuming, of course, that appropriate orientations have been chosen for the cells E^p and E^q). Let A^r denote the following subset of the product space, $E^p \times E^q$:

$$A^r = (S^{p-1} \times E^q) \cup (E^p \times E_+^{q-1}).$$

Then A^r is a closed r-cell, and its bounding $(r-1)$-sphere, $\dot{A}^r$, is given by

$$\dot{A}^r = (S^{p-1} \times E_-^{q-1}) \cup (E^p \times S^{q-2}).$$

Define a map

$$\phi\colon (A^r, \dot{A}^r) \to (X, A)$$

by

$$\phi(x, y) = \begin{cases} f(x) \text{ if } (x, y) \in E^p \times E_+^{q-1}, \\ g(y) \text{ if } (x, y) \in S^{p-1} \times E^q. \end{cases}$$

The map ϕ thus defined is continuous, and its homotopy class depends only on the homotopy classes of f and g. The product $[\alpha, \beta]$ is now defined to be the

homotopy class of ϕ; to complete the definition, it is necessary to make some convention as to choice of a base point in $\dot{A}^r$ and of orientation for the cell A^r.

The products mentioned in (b) and (c) above are defined in a similar manner.

As examples of some of the properties of these products, we list the following without proof.

(a) Let $$\partial_n: \pi_n(X, A, x_0) \to \pi_{n-1}(A, x_0), \qquad n = 2, 3, \cdots,$$

denote the homotopy boundary operator of the pair (X, A), $\alpha \in \pi_p(A, x_0)$, $\beta \in \pi_q(X, A, x_0)$. Then

$$\partial_r[\alpha, \beta] = (-1)^{p-1}[\alpha, \partial_q\beta], \qquad \partial_r[\beta, \alpha] = -[\partial_q\beta, \alpha].$$

If $\alpha \in \pi_p(X, A, x_0)$ and $\beta \in \pi_q(X, A, x_0)$, then

$$[\partial_p\alpha, \beta] = (-1)^{p-1}[\alpha, \partial_q\beta].$$

(b) Let $f: (X, A, x_0) \to (Y, B, y_0)$ be a continuous map, and let

$$f_n: \pi_n(X, A, x_0) \to \pi_n(Y, B, y_0), \qquad n = 2, 3, \cdots,$$

$$g_m: \pi_m(A, x_0) \to \pi_m(B, y_0), \qquad m = 2, 3, \cdots,$$

be the homomorphisms induced by f. Then

$$f_r[\alpha, \beta] = [g_p\alpha, f_q\beta]$$

for $\alpha \in \pi_p(A, x_0)$, $\beta \in \pi_q(X, A, x_0)$.

(c) Let (X, A) be a pair, and

$$i_m: \pi_m(A, x_0) \to \pi_m(X, x_0), \qquad m = 1, 2, \cdots,$$

$$j_n: \pi_n(X, x_0) \to \pi_n(X, A, x_0), \qquad n = 1, 2, \cdots,$$

the injections. Then

$$j_r[i_p\alpha, \beta] = [\alpha, j_q\beta]$$

for $\alpha \in \pi_p(A, x_0)$, $\beta \in \pi_q(X, x_0)$.

(d) Let $(X; A, B)$ be an arbitrary triad, and let

$$i: \pi_p(X, A \cap B, x_0) \to \pi_p(X, A, x_0),$$

$$j: \pi_q(A, x_0) \to \pi_q(A/B, x_0),$$

$$j': \pi_r(X, A, x_0) \to \pi_r(X; A, B, x_0),$$

be homomorphisms induced by inclusion maps. Then

$$j'[i\alpha, \beta] = [\alpha, j\beta]$$

for $\alpha \in \pi_p(X, A \cap B, x_0)$, $\beta \in \pi_q(A, x_0)$.

8. The main theorem. Let $X^* = X \cup \varepsilon^n$ as in §4. In addition, we make the following assumptions:

(a) X and $\dot{\varepsilon}^n$ are compact absolute neighborhood retracts.
(b) X is simply connected.
(c) The maps ψ_i: $(E^n, S^{n-1}) \to (\varepsilon_i^n, \dot{\varepsilon}_i^n)$ map S^{n-1} *homeomorphically* onto $\dot{\varepsilon}_i^n$.
Define a homomorphism

$$\Phi\colon \pi_{m+1}(\varepsilon^n/X) \otimes \pi_n(\varepsilon^n, \dot{\varepsilon}^n) \to \pi_{m+n}(X^*; \varepsilon^n, X)$$

(where the symbol $\otimes$ signifies "tensor product") as follows: If $\alpha \in \pi_{m+1}(X/\varepsilon^n)$ and $\beta \in \pi_n(\varepsilon^n, \dot{\varepsilon}^n)$, then

$$\Phi(\alpha \otimes \beta) = [\alpha, i\beta]$$

where $i\colon \pi_n(\varepsilon^n, \dot{\varepsilon}^n) \to \pi_n(X^*, \dot{\varepsilon}^n)$ is the injection and the brackets denote a generalized Whitehead product. The fact that Φ is a homomorphism follows from the bilinearity of this Whitehead product. We can now state our main theorem, as follows:

THEOREM 4. *If the pair* $(X, \dot{\varepsilon}^n)$ *is* m*-connected,* $m \geqq 1$, *then* Φ *is an isomorphism onto.*

Note that under the assumptions we have made, $\pi_n(\varepsilon^n, \dot{\varepsilon}^n)$ and $\pi_{m+1}(\varepsilon^n/X)$ are isomorphic to the integral homology groups $H_n(\varepsilon^n, \dot{\varepsilon}^n)$ and $H_{m+1}(X, \dot{\varepsilon}^n)$ respectively, both of which are effectively calculable in case the pair $(X, \dot{\varepsilon}^n)$ is triangulable. Hence the group $\pi_{m+n}(X^*; \varepsilon^n, X)$ is also effectively calculable in this case.

The proof of this theorem is much more difficult than the proof of Theorems 2 and 3. The only proof that the author knows requires the use of the Steenrod functional cup product (see [4]), although the functional cup product is not used in the statement of the theorem itself. The functional cup product gives an effective method of determining whether or not two maps

$$(E^{m+n}; E_+^{m+n-1}, E_-^{m+n-1}) \to (X^*; \varepsilon^n, X)$$

represent the same element of $\pi_{m+n}(X^*; \varepsilon^n, X)$.

9. Applications of the main theorem. If Theorem 4 is applied to the triad $(S^n; E_+^n, E_-^n)$, we find that $\pi_{2n-1}(S^n; E_+^n, E_-^n)$ is infinite cyclic; a generator of this group is obtained by forming the generalized Whitehead product $[\alpha, \beta]$, where α is a generator of the infinite cyclic group $\pi_n(E_+^n/E_-^n)$, and β is one of the generators of the group $\pi_n(S^n, S^{n-1})$, corresponding either to a generator of $\pi_n(E_+^n, S^{n-1})$, or to a generator of $\pi_n(E_-^n, S^{n-1})$. This fact, together with the relations between the various generalized Whitehead products, and elementary properties of the Whitehead products for the homotopy groups of spheres, gives the Freudenthal theorems in the critical dimensions, together with the improvements due to G. W. Whitehead [5] referred to previously.

As a second application, assume that $X^* = X \cup \varepsilon^n$ as in §4, and that conditions (a) and (b) of §8 are satisfied; we do *not* assume that (c) of §8 is satis-

fied. We assume that X is m-connected, $m \geqq 1$, but $\pi_{m+1}(X) \neq 0$, and wish to determine as many of the groups $\pi_p(X^*, X)$ as possible. In case $m \geqq n - 1$, then all the maps $\psi_i' \colon S^{n-1} \to X$ (defined by the maps $\psi_i \colon (E^n, S^{n-1}) \to (\mathcal{E}_i^n, \dot{\mathcal{E}}_i^n)$) are inessential, and we may as well assume that $\dot{\mathcal{E}}^n$ consists of a single point. This case is best handled by other methods; hence we assume that $m < n - 1$. Furthermore, we shall assume that $\psi_i'(p_0) = x_0 \in X$, a fixed base point, for $i = 1, \cdots, k$.

Let $(E_1^n, S_1^{n-1}), \cdots, (E_k^n, S_k^{n-1})$ be k distinct homeomorphic copies of the pair (E^n, S^{n-1}), and let

$$\xi_i \colon (E_i^n, S_i^{n-1}) \to (E^n, S^{n-1}), \qquad i = 1, \cdots, k,$$

be homeomorphisms. Let $p_i = \xi_i^{-1}(p_0)$, $i = 1, \cdots, k$. Identify the points $p_1, p_2, \cdots, p_k$ to a single point $\bar{p}_0$, and denote the identification space thus obtained by E; then $E = E_1^n \cup \cdots \cup E_k^n$. Let $S = S_1^{n-1} \cup \cdots \cup S_k^{n-1}$. Define a map

$$\psi \colon (E, S) \to (X^*, X)$$

by $\psi \mid E_i^n = \psi_i \xi_i$. The map ψ induces homomorphisms

$$\psi_* \colon \pi_p(E, S) \to \pi_p(X^*, X).$$

We are now in a position to state our results on the groups $\pi_p(X^*, X)$. The problem breaks up into three cases.

(a) *Case where $m = n - 2$.* In this case the homomorphism $\psi_* \colon \pi_p(E, S) \to \pi_p(X^*, X)$ is an isomorphism onto for $p < m + n$, and is an isomorphism into for $p < 2n - 2$. Define a homomorphism

$$\zeta \colon \pi_{m+1}(X) \otimes \pi_n(X^*, X) \to \pi_{m+n}(X^*, X)$$

by $\zeta(\alpha \otimes \beta) = [\alpha, \beta]$ (Whitehead product). Then ζ is an isomorphism into, and $\pi_{m+n}(X^*, X)$ is the direct sum of the subgroups $\psi_*\pi_{m+n}(E, S)$ and $\zeta[\pi_{m+1}(X) \otimes \pi_n(X^*, X)]$.

(b) *Case where $m = n - 2$ (first possibility).* The map $\psi_i' \colon (S^{n-1}, p_0) \to (X, x_0)$ represents an element $\alpha_i \in \pi_{n-1}(X, x_0)$, $i = 1, \cdots, k$. Assume first that the elements $\alpha_1, \cdots, \alpha_k$ generate the entire group $\pi_{n-1}(X, x_0)$. Then the homomorphism $\psi_* \colon \pi_p(E, S) \to \pi_p(X^*, X)$ is an isomorphism onto for $p < 2n - 2$ and a homomorphism onto for $p = 2n - 2$. The kernel of the homomorphism $\psi_* \colon \pi_{2n-2}(E, S) \to \pi_{2n-2}(X^*, X)$ is the subgroup of $\pi_{2n-2}(E, S)$ generated by the elements $[\alpha, \beta]$, where α is an arbitrary element of $\pi_{n-1}(S)$, and β is an element of $\pi_n(E, S)$ which satisfies the condition $\partial\psi_*(\beta) = 0$, where $\partial \colon \pi_n(X^*, X) \to \pi_{n-1}(X)$ is the boundary operator of the pair (X^*, X).

(c) *Case where $m = n - 2$ (second possibility).* Finally, we assume that the elements $\alpha_1, \cdots, \alpha_k$ generate a *proper* subgroup of $\pi_{n-1}(X, x_0)$. Then the homomorphism $\psi_* \colon \pi_p(E, S) \to \pi_p(X^*, X)$ is an isomorphism onto for $p < 2n - 2$,

but need not be an isomorphism, nor onto for $p = 2n - 2$. One can assert, however, that the group $\pi_{2n-2}(X^*, X)$ is generated by the elements $\psi_*(\alpha)$ for $\alpha \in \pi_{2n-2}(E, S)$, and the Whitehead products $[\beta, \gamma]$, for $\beta \in \pi_{n-1}(X)$, $\gamma \in \pi_n(X^*, X)$.

These results are all a direct application of Theorem 4, together with the basic properties of the Whitehead products.

Another application of these results is to the study of the homotopy groups of a complex which is a union of spheres with a single point in common. Let $\mathfrak{S}_k^n$ be a cell complex which has a single vertex, e^0, and k n-dimensional cells, e_1^n, $\cdots$, e_k^n $(n > 1)$. The closure of the cell e_i^n is an n-sphere, denoted by S_i^n. The space $\mathfrak{S}_k^n$ is the union of the spheres S_1^n, $\cdots$, S_k^n, and $S_i^n \cap S_j^n = e^0$ if $i \neq j$. Let $\phi_i\colon S_i^n \to \mathfrak{S}_k^n$ be an inclusion map; ϕ_i represents an element $\alpha_i \in \pi_n(\mathfrak{S}_k^n)$; the elements $\alpha_1, \cdots, \alpha_k$ are a set of free generators of the group $\pi_n(\mathfrak{S}_k^n)$. Let $\phi_{ij}\colon S^{2n-1} \to \mathfrak{S}_k^n$ be a map representing the Whitehead product $[\alpha_i, \alpha_j]$. Then it can be shown that $\phi_{i*}\colon \pi_p(S_i^n) \to \pi_p(\mathfrak{S}_k^n)$ is an isomorphism into in all dimensions; and $\phi_{ij*}\colon \pi_p(S^{2n-1}) \to \pi_p(\mathfrak{S}_k^n)$ is an isomorphism into if $p < 4n - 4$ and $i \neq j$. If $p < 3n - 2$, then the group $\pi_p(\mathfrak{S}_k^n)$ is the direct sum of the subgroups $\phi_{i*}\pi_p(S_i^n)$, $i = 1, \cdots, k$, and $\phi_{ij*}\pi_p(S^{2n-1})$, $1 \leqq i < j \leqq k$. These results are most easily proved by applying a theorem of G. W. Whitehead ([5], Theorem 4.8) on the homotopy groups of the union of two spaces with a single point in common. This theorem enables one to prove these results on $\pi_p(\mathfrak{S}_k^n)$ by means of an induction on k, the number of spheres. It is also necessary to use Theorems 2 and 3 in the process.

By using this same method, and applying Theorem 4, it is possible to determine the structure of $\pi_{3n-2}(\mathfrak{S}_k^n)$, as follows:

THEOREM 5. *The group $\pi_{3n-2}(\mathfrak{S}_k^n)$ is the direct sum of the following subgroups*:
(1) $\phi_{i*}\pi_{3n-2}(S_i^n)$, $\quad 1 \leqq i \leqq k$,
(2) $\phi_{ij*}\pi_{3n-2}(S^{2n-1})$, $\quad 1 \leqq i < j \leqq k$,
(3) *a free abelian group on $\frac{1}{3}(k + 1)k(k - 1)$ generators, generated by the triple Whitehead products*

$$[\alpha_i, [\alpha_j, \alpha_l]], \qquad 1 \leqq j < l \leqq k, \qquad 1 \leqq i \leqq l.$$

Rather than apply Theorem 4 directly to prove Theorem 5, it is more convenient to use the result stated in (a) of this section.

10. Unsolved problems. The major unsolved problem in this subject at present is the determination of the homotopy groups of the triad $(X^*; \mathfrak{E}^n, X)$ in higher dimensions. The method used to determine the triad homotopy groups in the first nontrivial dimension breaks down completely when applied to higher dimensions. Apparently this problem is a major obstacle to the further progress of homotopy theory in general. Even the groups $\pi_{2n}(S^n; E_+^n, E_-^n)$ are unknown, except in the cases $n = 2$, 4, and 8, in which case $\pi_{2n}(S^n; E_+^n, E_-^n) \approx \pi_{2n}(S^{2n-1})$.

Bibliography

1. A. L. Blakers and W. S. Massey, *The homotopy groups of a triad.* I, Ann. of Math. vol. 53 (1951) pp. 161–205.

2. S. Eilenberg and N. E. Steenrod, *Axiomatic approach to homology theory*, Proc. Nat. Acad. Sci. vol. 31 (1945) pp. 117–120.

3. H. Freudenthal, *Über die Klassen der Sphärenabbildungen*, Compositio Math. vol. 5 (1937) pp. 299–314.

4. N. E. Steenrod, *Cohomology invariants of mappings*, Ann. of Math. vol. 50 (1949) pp. 954–988.

5. G. W. Whitehead, *A generalization of the Hopf invariant*, Ann. of Math. vol. 51 (1950) pp. 192–237.

6. J. H. C. Whitehead, *On adding relations to homotopy groups*, Ann. of Math. vol. 42 (1941) pp. 409–429.

7. ———, *On simply connected 4-dimensional polyhedra*, Comment. Math. Helv. vol. 22 (1949) pp. 48–92, Lemma 5.

8. ———, *Combinatorial homotopy* I, Bull. Amer. Math. Soc. vol. 55 (1949) pp. 213–245, Lemmas 3 and 4.

9. ———, *A note on suspension*, Quart. J. Math. (2) vol. 1 (1950) pp. 9–22.

Princeton University,
Princeton, N. J., U. S. A.

HOMOLOGY INVARIANTS AND FIBRE BUNDLES

GUY C. HIRSCH

1. Homology invariants[1] can be associated with fibre bundles (written f.b.)[2] in two main ways. Namely, (a) one can consider the relations between *homology properties* of the base space M, the bundle $\tilde{M}$, and the fibre $\check{F}$; or (b) one can also investigate a *classification* of f.b. based on homology properties only and, for instance, determine some of the classical invariants (characteristic class, obstructions, ...) by the consideration of homology properties, or vice-versa. We shall be concerned here mainly with the links between problems (a) and (b).

2. Problem (a) can be stated as follows:[3] What relations can be established between the cohomology rings (or even their additive structure) $\tilde{\mathfrak{H}}$, $\mathfrak{H}$, $\check{\mathfrak{H}}$ respectively of the bundle $\tilde{M}$, the base space M, and the fibre $\check{F}$, or even between their rings of cochains $\tilde{\mathfrak{C}}$, $\mathfrak{C}$, $\check{\mathfrak{C}}$?

Several contributions to the solution of this very general problem were given by Leray [18], particularly relations between the Poincaré polynomials of the three spaces and even part of the ring structure. But for our purposes it will be convenient to consider separately various aspects of this problem (a):

a1 (direct problem): Given M and $\check{F}$, what can be said about the homology of $\tilde{M}$?

a2 (inverse problem): Given $\tilde{M}$ and $\check{F}$, what can be said about the homology of M?

and even a3 (compatibility problem): Given $\tilde{M}$, what is the homology type of fibres $\check{F}$ compatible with the existence of a fibre structure in $\tilde{M}$?

We shall be concerned mainly with problem a1 and its implications (in §§6 ff.).

3. Problem a1 was solved in 1941 for the additive structure of $\tilde{\mathfrak{H}}$ and for particular spaces, firstly [8] when $\check{F}$ is an S^0 (2-sheeted covering) or a circle S^1, and [6] when (all spaces considered being manifolds) $\check{F}$ is a (homology) n-dimensional sphere S^n. In 1942, Samelson [19] determined $\tilde{\mathfrak{H}}$ when the base space M is a sphere S^n and a homogeneous space ($\tilde{M}$ and $\check{F}$ being group manifolds), and found two different types of spaces. Although as a rule I shall not consider in this paper properties which depend on more than homology properties, I mention Samelson's results because it can be shown[4] that they depend only on the nature of the cohomology rings of $\tilde{M}$ and $\check{F}$ (which are isomorphic to the rings

[1] Here I shall generally consider singular homology in locally compact spaces.

[2] No structure groups will be considered except in §15.

[3] The similar problems for U_i and the reduced powers of Steenrod are not considered here.

[4] This is a (yet unpublished) consequence of the general formulae mentioned in §8 and §10.

of topological products of spheres), M being only a homology sphere. When the ring of $\tilde{M}$ has not this property, there is a third type of spaces.[5]

The first general formulae were given in 1946 by Leray [18].[6]

4. Problem a2 was solved by Gysin [6] when $\tilde{M}$ and $\check{F}$ are (homology) spheres.

A particular case of a2 is given by *homogeneous spaces* ($\tilde{M}$ and $\check{F}$ being group manifolds). Important results were obtained by Leray [18], Koszul [17], H. Cartan [4]. These results do *not* hold if $\tilde{M}$ and $\check{F}$ have only the (additive) homology groups of group manifolds.

H. Cartan [3] solved a problem similar to a2, $\tilde{M}$ being a space in which a group operates.

5. Problem a3 is far from being solved in the general case. Some recent results are known when $\tilde{M}$ is a euclidean space [1]: the fibre (if different from a point) is never compact; or when $\tilde{M}$ is a sphere: $\check{F}$ cannot be a topological product of m circles ($m > 1$) [5]. Although the proof of the latter theorem uses homotopy methods, I believe that the theorem actually belongs to the theory of homology; it is, for instance, easy to prove (as a consequence of the general results of §8 and §10) that a homology sphere cannot be a f.b. where the fibre has the same cohomology ring as the product of two spheres (of arbitrary dimension). It might even be possible to prove that a (homology) sphere cannot be a f.b. where $\check{F}$ is neither a homology sphere nor a homology simplex. *Added in proof*: A. Borel proved recently (C. R. Acad. Sci. Paris vol. 231 (1950) pp. 943–945) that a homology sphere cannot be a f.b. where the fibre has the same cohomology ring as a product of m spheres ($m > 1$).

6. Although problem a1 finds an important application to *coverings* (given the base space M and the Poincaré group $\tilde{\pi}_1$ of the covering space $\tilde{M}$ as a subgroup of the Poincaré group π_1 of M), we shall restrict ourselves to the case where the domain of coefficients is a *field* and where π_1 operates trivially on the cohomology of $\check{F}$ (in the given field of coefficients).

Before entering into details about problem a1 and its relation to problem (b), I would like to mention two more problems which are closely related to problem (a), and might be called its "dynamic" aspect:

Problem (c): Let $\tilde{T}$ be a *mapping* of a f.b. $\tilde{M}$ in a f.b. $\tilde{M}'$, compatible with the fibre structure (this means that $\tilde{T}$ maps all the points belonging to the same fibre $\check{F}$ of $\tilde{M}$ into points belonging to the same fibre $\check{F}'$ of $\tilde{M}'$). The homology type of such a mapping is determined to a certain extent (as can be proved by applying the general theory of §8 and §10 to the induced mappings T of the

[5] In 1949, Wang [22] gave some results for f.b. where the base space is a homology sphere, under certain assumptions. The latter lead him back to the two types considered by Samelson.

[6] Leray considered even the more general problem of mappings instead of fibre bundles.

base space M into the base space M′, and $\check{T}$ of $\check{F}$ into $\check{F}'$—the latter being not unique, but having a constant homology type). Examples of this kind (with $\tilde{M} = \tilde{M}'$) are Borsuk's theorem [2] (1933) on "antipodentreue Abbildungen" (i.e., $\check{F}$ being a pair of antipodic points, $\tilde{M}$ the sphere and M the projective space, the degree of $\tilde{T}$ is *odd* when $\check{F}$ is mapped *onto* a fibre), or Hopf-Rueff's theorem [15] (1938), where $\tilde{M}$ is an S^{2k+1}, $\check{F}$ a circle, and M a k-dimensional complex projective space: the degree of $\tilde{T}$ is an exact $(k+1)$th power [the $(k+1)$th power of the degree of the mapping $\check{T}$ induced in $\check{F}$]. Generalisations of these theorems were given [9] when $\check{F}$ is discrete, or [8] when $\check{F}$ is a circle, $\tilde{M}$ being an arbitrary polyhedron (see also [7]).

Problem (d) is the following: Determine the homology type of a mapping T of a space M into a space M′, such that it induces a mapping $\tilde{T}$ of a f.b. $\tilde{M}$ over M as base space into a f.b. $\tilde{M}'$ over M′ as base space (for instance, $\tilde{M}$ being a covering of M). Example: M being a projective space and T a mapping of M into itself, the knowledge of the homomorphism (induced by T) in the Poincaré group π_1 determines the parity of the degree of T (or its absolute degree) [7].

7. In problem a1, it is of course not sufficient to know properties of M and $\check{F}$ to determine the cohomology ring $\tilde{\mathfrak{H}}$ of $\tilde{M}$, as the f.b. $\tilde{M}$ itself is in general not uniquely determined when M and $\check{F}$ are given. Some invariants will be needed to describe the type of f.b. considered. It will be convenient to express these invariants in terms of a relation between $\check{F}$ and M, which will allow the determination of $\tilde{\mathfrak{H}}$. As the classification of f.b. also rests usually on relations between $\check{F}$ and M (e.g., characteristic classes or obstructions), relations between both can be investigated. I showed in 1941 [8], that when $\check{F}$ is S^0 or S^1 (the result even applying to a fibre S^n, although this was not then considered in my paper), the invariant describing the homology structure of $\tilde{M}$ is equivalent to the characteristic class defined by Whitney [23]. This result is a special case of a theorem mentioned below (§9), connecting the "reduced characteristic isomorphism" and Steenrod's characteristic cocycle.

8. In Leray's solution [20] of problem a1, a sequence of differential operators is constructed, starting in the tensor product $\mathfrak{H} \otimes \check{\mathfrak{H}}$, and yielding, after a finite number of steps, a *filtration* of the cohomology ring $\tilde{\mathfrak{H}}$ (each ring being the "cohomology ring" of the preceding one; the sequence is a Leray-Koszul sequence [16]). Leray's operators use only homology properties, and are a topological invariant of the f.b. In my approach to the problem [10], a unique differential operator in the tensor product $\mathfrak{C} \otimes \check{\mathfrak{H}}$ ($\mathfrak{C}$ being the ring of cochains) yields the additive structure and part of the multiplicative structure of $\tilde{\mathfrak{H}}$. But this differential operator is, in general, not uniquely determined, as its definition rests on the "*characteristic isomorphism*" of $\check{\mathfrak{H}}$ in $\mathfrak{H}$ or $\mathfrak{H} \otimes \check{\mathfrak{H}}$, which depends on the choice of representatives, in the ring of cochains of $\tilde{M}$, of elements of cohomology classes of $\check{F}$.

It is possible to construct for $\check{\mathfrak{H}}$ a composition series without repeated terms,

$$0 \subset \check{\mathfrak{H}}_0 \subset \cdots \subset \check{\mathfrak{H}}_i \subset \cdots \subset \check{\mathfrak{H}}_s \subset \check{\mathfrak{H}}_{s+1} = \check{\mathfrak{H}};$$

$\check{\mathfrak{H}}_0$ is in $\check{\mathfrak{H}}$ the subgroup of images of $\tilde{\mathfrak{H}}$.

Let $\mathfrak{C}$ and $\tilde{\mathfrak{C}}$ be respectively the rings of cochains of M and $\tilde{\mathrm{M}}$; let P be the homomorphism of $\mathfrak{C}$ in $\tilde{\mathfrak{C}}$. Calling $\check{z}_i$ an element of $\check{\mathfrak{H}}_i/\check{\mathfrak{H}}_{i-1}$ and $\widetilde{\mathrm{ext}}(\check{z}_i)$ a cochain in $\tilde{\mathrm{M}}$ whose restriction to any fibre belongs to the class $\check{z}_i$, and defining a homomorphism P_i of $\mathfrak{C} \otimes \check{\mathfrak{H}}_i/\check{\mathfrak{H}}_{i-1}$ in $\tilde{\mathfrak{C}}$ by $P_i(\mathrm{C} \otimes \check{z}_i) = P(\mathrm{C}) \cup \widetilde{\mathrm{ext}}(\check{z}_i)$, $\check{\mathfrak{H}}_{i+1}$ is then defined by the property

$$\delta[\widetilde{\mathrm{ext}}(\check{z}_{i+1})] \in P_i(\mathfrak{C} \otimes \check{\mathfrak{H}}_i/\check{\mathfrak{H}}_{i-1}).$$

The choice of a representative $\widetilde{\mathrm{ext}}(\check{z}_{i+1})$ associates with each class $\check{z}_{i+1}$, by this definition, an element of $\mathfrak{C} \otimes \check{\mathfrak{H}}_i/\check{\mathfrak{H}}_{i-1}$, and even an element of $\mathfrak{H} \otimes \check{\mathfrak{H}}_i/\check{\mathfrak{H}}_{i-1}$. I call this correspondence the *characteristic isomorphism* W of $\check{\mathfrak{H}}_{i+1}$ in $\mathfrak{H} \otimes \check{\mathfrak{H}}_i/\check{\mathfrak{H}}_{i-1}$.[7] Its definition depends essentially on the choice of $\widetilde{\mathrm{ext}}$.[8]

However, defining $\check{\mathfrak{H}}_* \subset \check{\mathfrak{H}}$ and $\check{z}_* \in \check{\mathfrak{H}}_*$ by $\delta[\widetilde{\mathrm{ext}}(\check{z}_*)] \in P(\mathfrak{C})$, then this relation defines an *invariant* isomorphism W_* (called *reduced* characteristic isomorphism) of $\check{\mathfrak{H}}_*/\check{\mathfrak{H}}_0$ on a factor group of a subgroup of $\mathfrak{H}$.[9]

9. For the smallest nontrivial dimension h in $\check{\mathfrak{H}}$, the reduced characteristic isomorphism maps on a subgroup of $\mathfrak{H}$ and is equivalent[10] to Steenrod's characteristic cocycle [20], as can be shown by the values taken on an element of $\check{\mathfrak{H}}$ by the latter applied to a cycle of M.

10. The characteristic isomorphism is sufficient for determining the (additive) group $\tilde{\mathfrak{H}}$ (and even part of the multiplicative structure). P_i induces an isomorphism of $\mathfrak{F}_i^*/\mathfrak{N}_i^*$ (both subgroups of $\mathfrak{H} \otimes \check{\mathfrak{H}}_i/\check{\mathfrak{H}}_{i-1}$) on $\tilde{\mathfrak{H}}_i/\tilde{\mathfrak{H}}_{i-1}$, where $\tilde{\mathfrak{H}}_i$ is an element of a composition series

$$\tilde{\mathfrak{H}} = \tilde{\mathfrak{H}}_s \supset \tilde{\mathfrak{H}}_{s-1} \supset \cdots \supset \tilde{\mathfrak{H}}_i \supset \cdots \supset \tilde{\mathfrak{H}}_0 \supset 0.^{11}$$

Define a product $\smile$ between $\mathfrak{C}$ and $\mathfrak{C} \otimes \check{\mathfrak{H}}_i/\check{\mathfrak{H}}_{i-1}$ by the rule $\mathrm{C} \smile (\mathrm{C}' \otimes \check{z}_i) = (\mathrm{C} \cup \mathrm{C}') \otimes \check{z}_i$; define a homomorphism D_i mapping $\mathfrak{C} \otimes \check{\mathfrak{H}}_i/\check{\mathfrak{H}}_{i-1}$ into $\mathfrak{C} \otimes \check{\mathfrak{H}}_{i-1}/\check{\mathfrak{H}}_{i-2}$ by the rule $D_i(\mathrm{C} \otimes \check{z}_i) = (-1)^a \mathrm{C} \smile W(\check{z}_i)$ where $a = \dim \mathrm{C} \cdot \dim \check{z}_i$. Define $\mathfrak{Q}_i^* \subset \mathfrak{C} \otimes \check{\mathfrak{H}}_i/\check{\mathfrak{H}}_{i-1}$ by $\delta\mathfrak{Q}_i^* \subset D_{i+1}\mathfrak{Q}_{i+1}^*$, and $\mathfrak{M}_i^* \subset \mathfrak{C} \otimes \check{\mathfrak{H}}_i/\check{\mathfrak{H}}_{i-1}$ by $D_i\mathfrak{M}_i^* \subset \delta\mathfrak{M}_{i-1}^*$.

Then $\mathfrak{F}_i^*$ is the group of classes of $\mathfrak{Z} \otimes \check{\mathfrak{H}}_i/\check{\mathfrak{H}}_{i-1}$ contained in $\mathfrak{M}_i^*$ (where $\mathfrak{Z}$ is

[7] This is essentially the process defined in [10], where, however, I considered simultaneously homology and cohomology, which made the definitions more cumbersome.

[8] Even the length s of the series is not invariant.

[9] A similar notion has been introduced by Koszul [17] for homogeneous spaces, under the name of *transgression*.

[10] See the condition about π_1 in §6.

[11] For the intuitive aspect of this homomorphism, see [11].

the group of cocycles in M), and $\mathfrak{R}_i^*$ is the group of classes of $\mathfrak{Z} \otimes \check{\mathfrak{H}}_i/\check{\mathfrak{H}}_{i-1}$ containing $D_{i+1}\mathfrak{Q}_{i+1}^*$.[12]

11. Lack of space prevents me from enumerating properties of W; $\check{\mathfrak{H}}/\check{\mathfrak{H}}_0$ is similar in many respects to an exterior algebra, elements of $\check{\mathfrak{H}}/\check{\mathfrak{H}}_0$ being maximal in the ring and of odd dimension. But although the factor ring $\check{\mathfrak{H}}/\check{\mathfrak{H}}_0$ is generated by the elements appearing in the reduced characteristic isomorphism when M is a homogeneous space [17], [4], a similar result does *not* apply to general f.b. Also the elements $W_*(\check{\mathfrak{H}}_*/\check{\mathfrak{H}}_0)$ which are generators of $\mathfrak{H}$ for homogeneous spaces are not maximal, but only relatively maximal, in the general case.

12. The system of all characteristic isomorphisms (for all possible choices of representatives) is of course an invariant of the f.b. The relations between it and Leray's invariants might be investigated; it is probable that both are equivalent.

13. Neither the characteristic isomorphism nor Leray's sequence are sufficient to determine fully the multiplicative structure of $\check{\mathfrak{H}}$. It follows from §10 that the latter depends on the products of the extensions $\widetilde{\text{ext}}(\check{z}_i)$. These products define new invariants which attach to pairs of elements of $\check{\mathfrak{H}}$ elements of $\mathfrak{H} \otimes \check{\mathfrak{H}}$ when $\check{z}$ is a $\check{z}_0$ belonging to $\check{\mathfrak{H}}_0$, or homomorphisms in these tensor products for other $\check{z}_i$ [12].

14. When the fibre is an S^2 or, more generally, a k-dimensional complex projective space, these invariants are the same as those introduced by Hopf and relative to the second obstruction [14] [13]. This can be proved by the construction of a new f.b. $\check{\mathbf{M}}$ over M, replacing $\check{\mathbf{F}}$ by an S^3 (or an S^{2k+1}) which is itself a f.b. over $\check{\mathbf{F}}$, and using the definition of the characteristic class mentioned in §9 (as Hopf's second obstruction of the first f.b. becomes the characteristic class of the new f.b.). It is still an open problem to relate the invariants of the multiplicative structure to known invariants of a f.b. when $\check{\mathbf{F}}$ is not a projective space.

15. The latter considerations will lead to a new point of view. In this section, I shall consider a f.b. with a structure group $\mathfrak{G}$ which has an invariant subgroup $\mathfrak{G}_0$ and operates transitively on $\check{\mathbf{F}}$. $\check{\mathbf{F}}$ is itself a f.b. with a fibre $\check{\mathbf{F}}_0$ (defined by equivalence relative to $\mathfrak{G}_0$), and $\mathbf{F}$ as base space. In the given f.b. $\check{\mathbf{M}}$, consideration of the equivalence classes relative to $\mathfrak{G}_0$ leads to a new f.b. $\mathbf{M}$, which is

[12] An important special case ("reducible case") allows $\mathfrak{Q}_i^*$ to be replaced by $\mathfrak{Z} \otimes \check{\mathfrak{H}}_i/\check{\mathfrak{H}}_{i-1}$ and $\mathfrak{M}_i^*$ by $\mathfrak{C} \otimes \check{\mathfrak{H}}_i/\check{\mathfrak{H}}_{i-1}$, thus considering D as a derivation in the direct sum of $\mathfrak{H} \otimes \check{\mathfrak{H}}_i/\check{\mathfrak{H}}_{i-1}$ (graduated by the index i) for which the factor groups $\mathfrak{F}_i^*/\mathfrak{R}_i^*$ appear as homology groups (with D as boundary operator). It was, for instance, proved by Koszul that the reducible case applies to symmetric homogeneous spaces.

the base space for the f.b. $\tilde{M}$ (with the fibre $\tilde{F}_0$), and also a f.b. over M with **F** as fibre. (Example, similar to the last part of §14: $\tilde{F}$ being an S^{2k+1}, $\mathfrak{G}$ the unimodular group, and $\mathfrak{G}_0$ the subgroup leaving one point invariant; **F** is a complex projective space, $\tilde{F}_0$ a circle. In the problem of §14, it was **M** that was the given f.b., and $\tilde{M}$ the auxiliary one. A similar theory holds mod 2 with $\mathfrak{G}$ the orthogonal group, **F** being a real projective space.)

The invariants of the multiplicative structure of the auxiliary f.b. **M** (defined as in §13) are invariants of the f.b. M. For instance, when $\tilde{F}$ is a sphere and $\mathfrak{G}$ the orthogonal or the unimodular group, these invariants coincide with the characteristic classes of Stiefel-Whitney or Chern [13], which are thus defined without explicitly considering associated f.b.

16. In this respect, I might also mention here recent results of Thom [21], who determined these characteristic classes when $\tilde{F}$ is a sphere, by means of the reduced powers of Steenrod (thus proving their topological invariance and independence of the differential structure). However, these interesting results do not come quite in the scope of the present paper, as they would not apply if $\tilde{F}$ were only a *homology* sphere.

17. When the f.b. is the topological product of M and $\tilde{F}$, it is known that a *section* is equivalent to a *mapping* of M in $\tilde{F}$, thus defining a *homomorphism* (Hopf's homomorphism) of the cohomology group $\tilde{\mathfrak{H}}$ of $\tilde{\mathfrak{F}}$ in the cohomology group $\mathfrak{H}$ of M. It can be shown [13][13] that (when one section is given), a section in a general f.b. (which does not define a mapping) also determines a homomorphism of $\tilde{\mathfrak{H}}_0$ (defined in §8) in $\mathfrak{H}$; for elements of $\tilde{\mathfrak{H}}_i$ $(i \neq 0)$, there corresponds a cochain in M, or a homomorphism of $\mathfrak{F}_i^*$ (definition, §10) in $\mathfrak{H}$.

Bibliography

1. A. Borel and J. P. Serre, C. R. Acad. Sci. Paris vol. 230 (1950) pp. 2258–2260.
2. K. Borsuk, Fund. Math. vol. 20 (1933) p. 177; see also P. Alexandroff and H. Hopf, *Topologie* I, p. 483.
3. H. Cartan, C. R. Acad. Sci. Paris vol. 226 (1948) pp. 148–150, 303–305.
4. ———, Colloque de Topologie, Bruxelles, 1950 (Centre Belge de Recherches Mathématiques, 1950).
5. B. Eckmann, H. Samelson, and G. W. Whitehead, Bull. Amer. Math. Soc. vol. 55 (1949) pp. 433–438.
6. W. Gysin, Comment. Math. Helv. vol. 14 (1941) pp. 61–122.
7. G. Hirsch, Académie Royale de Belgique. Bulletin de la Classe des Sciences (5) vol. 23 (137) pp. 219–225; Comptes Rendus du Congrès Sc. Math. Liège, 1939, pp. 128–131.
8. ———, Bulletin de la Société Royale des Sciences de Liège (1941) pp. 246–260.
9. ———, Ann. École Norm. (3) vol. 60 (1943) pp. 113–142.
10. ———, C. R. Acad. Sci. Paris vol. 227 (1948) pp. 1328–1330; a summary appears in Colloque de Topologie, Bruxelles, 1950; an extensive paper will appear shortly.
11. ———, Bulletin de la Société Mathématique de Belgique (1947–1948) pp. 24–33.

[13] This was proved as a reply to a question by Hopf.

12. ———, C. R. Acad. Sci. Paris vol. 229 (1949) pp. 1297–1299; C. R. Acad. Sci. Paris vol. 230 (1950) pp. 46–48.
13. ———, Colloque de Topologie, Bruxelles, 1950 (Centre Belge de Recherches Mathématiques, 1950).
14. H. Hopf, Colloque de Topologie, Bruxelles, 1950 (Centre Belge de Recherches Mathématiques, 1950).
15. H. Hopf and M. Rueff, Comment. Math. Helv. vol. 11 (1938) pp. 49–61.
16. J. L. Koszul, C. R. Acad. Sci. Paris vol. 225 (1947) p. 217.
17. ———, Bull. Soc. Math. France vol. 78 (1950) pp. 65–127.
18. J. Leray, Several notes in C. R. Acad. Sci. Paris since 1946, and J. Math. Pures Appl. vol. 29 (1950) pp. 1–139.
19. H. Samelson, Ann. of Math. vol. 42 (1941) pp. 1091–1137.
20. N. E. Steenrod, Ann. of Math. vol. 43 (1942) pp. 116–131.
21. R. Thom, C. R. Acad. Sci. Paris vol. 230 (1950) pp. 427–429, 507–508.
22. Hsien-Chung Wang, Duke Math. J. vol. 16 (1949) pp. 33–38.
23. H. Whitney, Bull. Amer. Math. Soc. vol. 43 (1937) pp. 785–805.

University of Brussels,
Brussels, Belgium.

HOMOLOGY THEORY OF FIBER BUNDLES

E. H. SPANIER

1. The first comprehensive result on the homology properties of fiber bundles was obtained by Gysin [3] who studied sphere bundles over manifolds. Little more was done with this subject until fairly recently when several people began to get interested in it. Among these should be mentioned Leray [7, 8], Hirsch [4, 5, 6], Steenrod [12], Thom [13, 14], Wu [17], Borel and Serre [1], and Shapiro [9]. The chief tool in the approach of Leray and many of the others seems to be the "ring of the projection" of the bundle. This is an algebraic structure associated with the bundle which is obtained by considering cohomology of the base space with local coefficients in the cohomology ring of the fiber. In the general situation this algebraic structure is fairly complicated, and it is difficult to compute with it.

A method similar to that of Leray has been utilized by Chern and the author [2] to study the homology structure of fiber bundles where the base space is assumed to be a complex. The technique consists of a study of the homology properties of the inverse images of the skeletons of the base space under the projection. The resulting algebraic structure is essentially the same as the ring of the projection; however, because of the assumption that the base space is a complex, the situation is simpler than the general case and computations can be carried out more easily.

In this paper this algebraic machinery is discussed and some applications are given. In particular, it is shown how the characteristic class of the bundle can be obtained rather simply in this approach. By considering the graph of the projection as a bundle over the base space, some of the multiplicative properties of the cohomology ring of the bundle can be obtained. In this way a fairly simple theorem on sphere bundles is obtained which includes the main results of Gysin. The last application discussed is to the relative position of the fiber in the bundle. If the fiber is totally nonhomologous to zero, it is shown that the cohomology groups of the bundle are isomorphic to the corresponding groups of the product space of the base space by the fiber.

The algebraic structure which arises in this study seems to be of interest in other connections. Any simplicial map of one complex into another gives rise to a similar situation, and an investigation of this would perhaps lead to new invariants of mappings which could be applied to the homotopy classification problem. Another situation in which this algebraic structure is of interest is obtained by considering the homotopy groups of the skeletons of a complex. In this case the new sequence of J. H. C. Whitehead [16], can be obtained by this approach, and perhaps interesting generalizations of this can be found by deeper inspection.

2. Let $f: B \to X$ be a fibering with connected fiber F and structure group Γ [11]. We shall assume X is a finite connected cell complex of dimension n in which each cell is simply connected. We also assume that if σ is any cell of X, there is a homeomorphism

$$\phi_\sigma: |\sigma| \times F \to f^{-1}(|\sigma|)$$

such that

$$f\phi_\sigma(x, y) = x$$

for $x \in |\sigma|$ and $y \in F$. Moreover, if σ is a face of τ, the transformation

$$\gamma_{\sigma\tau}(x): F \to F$$

defined for $x \in |\sigma|$ by

$$\phi_\sigma(x, y) = \Phi_\tau(x, \gamma_{\sigma\tau}(x)(y))$$

is an element of Γ depending continuously on x in $|\sigma|$.

Let X^r denote the r-dimensional skeleton of X and set $B_r = f^{-1}(X^r)$. As coefficient system for cohomology in B we use a local system $\{G_b\}$ [10] which is simple over every fiber. Such a system is the inverse image under f of some local system $\{G_x\}$ on X. Since $|\sigma|$ is simply connected, $\{G_x\}$ can be replaced by a local system $\{G_\sigma\}$ over the cells of X.

From the fact that $f^{-1}(|\sigma|)$ is homeomorphic to $|\sigma| \times F$, an isomorphism

$$\lambda: H^p(B_q, B_{q-1}; G_b) \approx C^q(X; H^{p-q}(F_\sigma; G_\sigma))$$

can be constructed in a natural fashion. In the above the first group is the p-dimensional relative cohomology group of B_q modulo B_{q-1} with coefficients in $\{G_b\}$, while the second group is the group of q-dimensional cochains of X with coefficients in the $(p - q)$-dimensional cohomology group of F_σ with coefficients in G_σ. Furthermore, if

$$j: B_q \subset (B_q, B_{q-1})$$

denotes the inclusion map and δ^* denotes the coboundary operator for the pair (B_{q+1}, B_q), commutativity holds in the diagram

$$\begin{array}{ccccc} H^p(B_q, B_{q-1}; G_b) & \xrightarrow{j^*} & H^p(B_q; G_b) & \xrightarrow{\delta^*} & H^{p+1}(B_{q+1}, B_q; G_b) \\ \downarrow\lambda & & & & \downarrow\lambda \\ C^q(X; H^{p-q}(F_\sigma; G_\sigma)) & \xrightarrow{\delta} & C^{q+1}(X; H^{p-q}(F_\sigma; G_\sigma)) & & \end{array}$$

where δ denotes the coboundary map for cochains of X.

The following diagram will be called the *main diagram of the bundle.*

$$
\begin{array}{ccccccc}
 & & & & H^p(B_n, B_{n-1}; G_b) & \xrightarrow{j^*} & H^p(B_n = B; G_b) \\
 & & & & \downarrow{\scriptstyle i^*} & & \\
\cdots & \xrightarrow{\delta^*} & H^p(B_{n-1}, B_{n-2}; G_b) & \xrightarrow{j^*} & H^p(B_{n-1}; G_b) & \xrightarrow{\delta^*} & H^{p+1}(B_n, B_{n-1}; G_b) \xrightarrow{j^*} \cdots \\
 & & & & \downarrow{\scriptstyle i^*} & & \\
 & & & & \vdots & & \\
 & & & & \downarrow{\scriptstyle i^*} & & \\
\cdots & \xrightarrow{\delta^*} & H^p(B_p, B_{p-1}; G_b) & \xrightarrow{j^*} & H^p(B_p; G_b) & \xrightarrow{\delta^*} & H^{p+1}(B_{p+1}, B_p; G_b) \xrightarrow{j^*} \cdots \\
 & & & & \downarrow{\scriptstyle i^*} & & \\
 & & & & \vdots & & \\
 & & & & \downarrow{\scriptstyle i^*} & & \\
\cdots & \xrightarrow{\delta^*} & H^p(B_0, B_{-1}; G_b) & \xrightarrow{j^*} & H^p(B_0; G_b) & \xrightarrow{\delta^*} & B^{p+1}(B_1, B_0; G_b) \xrightarrow{j^*} \cdots \\
 & & & & \downarrow{\scriptstyle i^*} & & \\
 & & & & 0. & &
\end{array}
$$

Note that any sequence composed of j^*, i^*, δ^* in that cyclic order is exact.

Let $u \in H^p(B_q, B_{q-1}; G_b)$. We say that u *can be lifted* k*-stages to* $\bar{u} \in H^{p+1}(B_{q+k+1}, B_{q+k}; G_b)$ if there is $v \in H^p(B_{q+k}; G_b)$ such that $(i^*)^k v = j^* u$ and $\delta^* v = \bar{u}$. Naturally it can happen that a given $u \in H^p(B_q, B_{q-1}; G_b)$ can perhaps be lifted k-stages to different elements of $H^{p+1}(B_{q+k+1}, B_{q+k}; G_b)$. From the exactness properties of the main diagram it is clear that a given $u \in H^p(B_q, B_{q-1}; G_b)$ can be lifted 1-stage if and only if λu is a cocycle of $C^q(X; H^{p-q}(F_\sigma; G_\sigma))$.

3. Let G be an abstract group isomorphic with G_σ. Let $H_0^p(F; G)$ be the subgroup of $H^p(F; G)$ of elements invariant under all the operations of $\pi_1(X)$. It is well known that

$$Z^0(X; H^p(F_\sigma; G_\sigma)) \approx H_0^p(F; G).$$

Of particular interest are those elements of $H_0^p(F; G)$ which correspond to elements of $Z^0(X; H^p(F_\sigma; G_\sigma))$ which can be lifted p-stages.

Let d be the smallest positive integer for which $H^d(F; G)$ is nontrivial. This means that for $0 < q < d$, $H^d(B_q, B_{q-1}; G_b) = 0$ and $H^{d+1}(B_{q+1}, B_q; G_b) = 0$.

It follows that in the main diagram the sequence

$$\begin{array}{ccc} H^d(B_d, B_{d-1}; G_b) & \xrightarrow{j^*} & H^d(B_d; G_b) \\ & & \downarrow (i^*)^d \\ & & H^d(B_0; G_b) \xrightarrow{\delta^*} H^{d+1}(B_1, B_0; G_b) \end{array}$$

is exact. Hence, any $u \in H_0^d(F; G)$ corresponds to an element of $H^d(B_0; G_b)$ that can be lifted d-stages. Moreover, if u can be lifted d-stages to $\bar{u}_1$ and $\bar{u}_2$ it is clear that $\bar{u}_1 - \bar{u}_2 \in \delta^* j^* H^d(B_d, B_{d-1}; G_b)$. Hence, $\lambda\bar{u}_1$ and $\lambda\bar{u}_2$ are cohomologous cocycles of $C^{d+1}(X; H^0(F_\sigma; G_\sigma))$, which is isomorphic to $C^{d+1}(X; G_\sigma)$. In this way we have a well defined homomorphism

$$T: H_0^d(F; G) \to H^{d+1}(X; G_\sigma).$$

As an application of this homomorphism let $\pi_p(F) = 0$ for $p < d$ (where $d > 1$) and assume $\pi_d(F) \neq 0$. Then $\pi_d(F) \approx H_d(F)$. (Where no coefficients are indicated, integers are to be understood.) Let $G_b = H_d(F_b)$. In $H^d(F; H_d(F))$ consider the class ω whose value on any cycle z is the homology class of z. This class ω is the first obstruction to contracting F to a point and is invariant under any homeomorphism of F. Therefore, $\omega \in H_0^d(F; H_d(F))$ and so corresponds to an $\bar{\omega} \in H^d(B_0; H_d(F_\sigma))$ which can be lifted d-stages. For this element $T(\omega) = \Omega \in H^{d+1}(X; H_d(F_\sigma))$ can be shown to be minus the *characteristic class* of the bundle, a result originally proved by G. Whitehead [15].

4. We now show how some of the multiplicative relations between the groups of the main diagram for various coefficient systems can be found. If $f: B \to X$ is a bundle with fiber F and group Γ, then f induces in an obvious way a bundle $f': X \times B \to X \times X$ also with fiber F and group Γ.

Let $\{G_b\}$ be a local system on B simple over every fiber and let $\{G'_x\}$ be a local system on X. Then $\{G'_x \otimes G_b\}$ is a local system on $X \times B$ which is simple over every fiber. Let u be an element of $H^d(B_q, B_{q-1}; G_b)$ which can be lifted k-stages to $\bar{u} \in H^{d+1}(B_{q+k+1}, B_{q+k}; G_b)$. If $v \in H^p(X^p, X^{p-1}; G'_x)$ is a cocycle, then $v \times u \in H^{p+d}((X \times B)_{p+q}, (X \times B)_{p+q-1}; G'_x \otimes G_b)$ can be lifted k-stages to $v \times \bar{u}\, H^{p+d+1}((X \times B)_{p+q+k+1}, (X \times B)_{p+q+k}; G'_x \otimes G_b)$.

The main diagram associated with a bundle is natural in the sense that it maps properly under a cellular bundle map as is shown below. Let $f_1: B_1 \to X_1$ and $f_2: B_2 \to X_2$ be fiberings with the same fiber F and group Γ. Let $\phi: B_1 \to B_2$ be a cellular bundle map; that is, ϕ induces a cellular map $\phi_0 : X_1 \to X_2$ such that commutativity holds in the diagram

$$\begin{array}{ccc} B_1 & \xrightarrow{\phi} & B_2 \\ \downarrow f_1 & & \downarrow f_2 \\ X_1 & \xrightarrow{\phi_0} & X_2 \end{array}$$

If $\{G_{b_2}\}$ is a local system on B_2 simple over every fiber, $\phi^{-1}\{G_{b_2}\} = \{G_{b_1}\}$ is a local system on B_1 simple over every fiber. Since ϕ_0 is cellular, ϕ induces a homomorphism ϕ^* of the main diagram for B_2 with $\{G_{b_2}\}$ as coefficients into the main diagram of B_1 with $\{G_{b_1}\}$ as coefficients. This induced homomorphism ϕ^* clearly has the property that if u is an element of a group of the main diagram of B_2 that can be lifted k-stages to $\bar{u}$ then ϕ^*u can be lifted k-stages in B_1 to $\phi^*\bar{u}$.

5. Now we turn our attention to sphere bundles. Let $f: B \to X$ be a fibering with fiber S^d and arbitrary group Γ of homeomorphisms of S^d.

Let $\{G_b\}$ be a local system on B simple over every fiber. Since $H^p(F; G) = 0$ if $0 < p < d$ and $H^0(F; G) \approx H^d(F; G) \approx G$, we see that the main diagram reduces to the following:

$$\begin{array}{ccccccccc}
 & & & & 0 & \xrightarrow{j^*} & H^p(B; G_b) & & \\
 & & & & & & \downarrow i^* & & \\
\cdots & \xrightarrow{\delta^*} & H^p(B_p, B_{p-1}; G_b) & \xrightarrow{j^*} & H^p(B_p; G_b) & \xrightarrow{\delta^*} & H^{p+1}(B_{p+1}, B_p; G_b) & \xrightarrow{j^*} & \cdots \\
 & & & & & & \downarrow i^* & & \\
\cdots & \xrightarrow{\delta^*} & H^p(B_{p-d}, B_{p-d-1}; G_b) & \xrightarrow{j^*} & H^p(B_{p-d}; G_b) & \xrightarrow{\delta^*} & H^{p+1}(B_{p-d+1}; B_{p-d}; G_b) & & \\
 & & & & & & \downarrow i^* & & \\
 & & & & & & 0 & &
\end{array}$$

where again any sequence composed of j^*, i^*, δ^* in that cyclic order is exact.

Let $G_b' = H_d(F_b) \otimes G_b$. Then $\{G_b'\}$ is a local system on B simple over every fiber, and $H_d(F_b) \otimes G_b' \approx G_b$. It is clear that $H^0(F_b; G_b) \approx G_b$ and that $H^d(F_b; G_b) \approx G_b'$.

From the above diagram we see that any element of $H^p(B_{p-d}, B_{p-d-1}; G_b)$ in the kernel of $\delta^* j^*$ can be lifted d-stages to an element of $H^{p+1}(B_{p+1}, B_p; G_b)$. This lifting induces a homomorphism

$$\Phi: H^{p-d}(X; G_\sigma') \to H^{p+1}(X; G_\sigma).$$

Another homomorphism can be defined as follows. Let $u \in H^p(B; G_b)$. Then $\lambda j^{*-1}(i^*)^2(u)$ is a class of cohomologous cocycles in $C^{p-d}(X; G_\sigma')$. This class in X depends only on u so that we have a homomorphism

$$\Psi: H^p(B; G_b) \to H^{p-d}(X; G_\sigma').$$

The main theorem on sphere bundles is that the sequence

$$\cdots \xrightarrow{\Phi} H^p(X; G_\sigma) \xrightarrow{f^*} H^p(B; G_b) \xrightarrow{\Psi} H^{p-d}(X; G_\sigma') \xrightarrow{\Phi} H^{p+1}(X; G_\sigma) \cdots$$

is exact. This shows that the cohomology groups of B are determined up to a group extension by the cohomology groups of X and the homomorphism Φ.

To calculate the mapping Φ let D be a cellular bundle map of B into $X \times B$ such that D_0 is a cellular approximation to the diagonal map of X into $X \times X$. If $v \in H^p(B_{p-d}, B_{p-d-1}; G_b)$ such that $\delta^* j^* v = 0$, then λv is a cocycle of $C^{p-d}(X; G'_\sigma)$ and $D^*(\lambda v \times \bar{\omega}) = v$ where $\bar{\omega} \in H^d(B_0; H_d(F_\sigma))$ corresponds to $\omega \in H_0^d(F; H_d(F))$, which was defined in §3. Since $\bar{\omega}$ can be lifted d-stages to $\bar{u}$ where $\lambda \bar{u}$ is in the class $\Omega = T(\omega)$, it follows that v can be lifted d-stages to $D^*(\lambda v \times \bar{u})$. Now $\lambda D^*(\lambda v \times \bar{u}) = \lambda v \cup \lambda \bar{u}$. Therefore, we have

$$\Phi(v) = v \cup \Omega \quad \text{for} \quad v \in H^{p-d}(X; G'_\sigma).$$

If d is even, the same map D can be used to show that $2\Omega = 0$. These results contain those of Gysin [3] and are similar to some of the theorems of Thom [13] and Shapiro [9].

6. In this section we assume the coefficient system is simple and shall denote it by G. Let $l: F \to B$ be the injection of F into a fiber of B. We assume there is a homomorphism $\mu: H^p(F; G) \to H^p(B; G)$ such that $l^*\mu$ is the identity automorphism of $H^p(F; G)$ for all $p \geqq 0$. If $H^p(F; G)$ is without torsion, this condition is equivalent to assuming l^* maps $H^p(B; G)$ onto $H^p(F; G)$, but, in general, it is stronger than the latter.

Let $H_0^p(B_q; G)$ be the kernel of δ^* in $H^p(B_q; G)$. We shall prove by induction on q the existence of homomorphisms

$$\mu_q: H_0^p(B_q; G) \to H^p(B_n; G)$$

such that:

(a) $(i^*)^{n-p}\mu_q$ is the identity automorphism of $H_0^p(B_q; G)$.
(b) $\mu_{q+1}(i^*)^{n-q-1}\mu_q = \mu_q$.

Since X is connected, $H_0^p(B_0; G) \approx H^p(F; G)$, and the existence of μ_0 with the required properties follows from the existence of μ. Assume μ_q has been constructed for $q < k$. To construct μ_k note that (b) defines μ_k on the direct summand $(i^*)^{n-k}\mu_{k-1}H_0^p(B_{k-1}; G)$ of $H_0^p(B_k; G)$. The other summand of $H_0^p(B_k; G)$ is isomorphic to $H^q(X; H^{p-q}(F; G))$. By using a cellular bundle map $D: B \to X \times B$ as in §5, the homomorphism μ_k can be defined on this latter summand in such a way that conditions (a) and (b) hold.

Having constructed the homomorphisms μ_q for all $0 \leqq q \leqq n$, it follows by induction on m that $H_0^p(B_m; G) \approx \sum_{q=0}^{m} H^q(X; H^{p-q}(F; G))$. Therefore, since $H^p(B; G) \approx H_0^p(B_n; G)$, we see that

$$H^p(B; g) \approx H^p(X \times F; G).$$

This result was announced by Hirsch [5] for the case where G is a field.

Bibliography

1. A. Borel and J. P. Serre, *Impossibilité de fibrer un espace euclidien par des fibres compacts*, C. R. Acad. Sci. Paris vol. 230 (1950) pp. 2258–2260.

2. SHIING-SHEN CHERN and E. SPANIER, *The homology structure of fiber bundles*, Proc. Nat. Acad. Sci. vol. 36 (1950) pp. 248–255.

3. W. GYSIN, *Zur Homologietheorie der Abbildungen und Fäserungen von Mannigfaltigkeiten*, Comment. Math. Helv. vol. 14 (1942) pp. 61–122.

4. G. HIRSCH, *Un isomorphisme attaché aux structures fibrées*, C. R. Acad. Sci. Paris vol. 227 (1948) pp. 1328–1330.

5. ———, *L'anneau de cohomologie d'un espace fibré et les classes caracteristiques*, C. R. Acad. Sci. Paris vol. 229 (1949) pp. 1297–1299.

6. ———, *Sur la structure multiplicative de l'anneau de cohomologie d'un espace fibré*, C. R. Acad. Sci. Paris vol. 230 (1950) pp. 46–48.

7. J. LERAY, *Propriétés de l'anneau d'homologie de la projection d'un espace fibré sur sa base*, C. R. Acad. Sci. Paris vol. 223 (1946) pp. 395–397.

8. ———, *L'anneau spectral et l'anneau filtré d'homologie d'un espace localement compact et d'une application continue*, Journal de Mathématiques vol. 29 (1950) pp. 1–139.

9. A. SHAPIRO, *Cohomologie dans les espaces fibrés*, C. R. Acad. Sci. Paris vol. 231 (1950) pp. 206–207.

10. N. E. STEENROD, *Homology with local coefficients*, Ann. of Math. vol. 44 (1943) pp. 610–627.

11. ———, *The classification of sphere bundles*, Ann. of Math. vol. 45 (1944) pp. 294–311.

12. ———, *Cohomology invariants of mappings*, Ann. of Math. vol. 50 (1949) pp. 954–988.

13. R. THOM, *Classes caractéristiques et i-carrés*, C. R. Acad. Sci. Paris vol. 230 (1950) pp. 427–429.

14. ———, *Variétés plongées et i-carrés*, C. R. Acad. Sci. Paris vol. 230 (1950) pp. 507–508.

15. G. W. WHITEHEAD, *On the characteristic cohomology class of a fiber bundle*, Bull. Amer. Math. Soc. Abstract 55-5-261.

16. J. H. C. WHITEHEAD, *A certain exact sequence*, Ann. of Math. vol. 52 (1950) pp. 51–110.

17. WEN-TSUN WU, *Classes caractéristiques et i-carrés d'une variété*, C. R. Acad. Sci. Paris vol. 230 (1950) pp. 508–511.

UNIVERSITY OF CHICAGO,
CHICAGO, ILL., U. S. A.

DIFFERENTIABLE MANIFOLDS

DIFFERENTIAL GEOMETRY OF FIBER BUNDLES

SHIING-SHEN CHERN

The aim of this lecture is to give a discussion of the main results and ideas concerning a certain aspect of the so-called differential geometry in the large which has made some progress in recent years. Differential geometry in the large in its vaguest sense is concerned with relations between global and local properties of a differential-geometric object. In order that the methods of differential calculus may be applicable, the spaces under consideration are not only topological spaces but are differentiable manifolds. The existence of such a differentiable structure allows the introduction of notions as tangent vector, tangent space, differential forms, etc. In problems of differential geometry there is usually an additional structure such as: (1) a Riemann metric, that is, a positive definite symmetric covariant tensor field of the second order; (2) a system of paths with the property that through every point and tangent to every direction through the point there passes exactly one path of the system; (3) a system of cones of directions, one through each point, which correspond to the light cones in general relativity theory, etc. Among such so-called geometric objects the Riemann metric is perhaps the most important, both in view of its rôle in problems of analysis, mechanics, and geometry, and its richness in results. In 1917 Levi-Civita discovered his celebrated parallelism which is an infinitesimal transportation of tangent vectors preserving the scalar product and is the first example of a connection. The salient fact about the Levi-Civita parallelism is the result that it is the parallelism, and not the Riemann metric, which accounts for most of the properties concerning curvature.

The Levi-Civita parallelism can be regarded as an infinitesimal motion between two infinitely near tangent spaces of the Riemann manifold. It was Elie Cartan who recognized that this notion admits an important generalization, that the spaces for which the infinitesimal motion is defined need not be the tangent spaces of a Riemann manifold, and that the group which operates in the space plays a dominant rôle. In his theory of generalized spaces (*Espaces généralisés*) Cartan carried out in all essential aspects the local theory of what we shall call connections [1; 2]. With the development of the theory of fiber bundles in topology, begun by Whitney for the case of sphere bundles and developed by Ehresmann, Steenrod, Pontrjagin, and others, [8; 19], it is now possible to give a modern version of Cartan's theory of connections, as was first carried out by Ehresmann and Weil [7; 22].

Let F be a space acted on by a topological group G of homeomorphisms. A fiber bundle with the director space F and structural group G consists of topological spaces B, X and a mapping ψ of B onto X, together with the following:

(1) X is covered by a family of neighborhoods $\{U_\alpha\}$, called the coordinate

neighborhoods, and to each U_α there is a homeomorphism (a coordinate function) φ_α: $U_\alpha \times F \to \psi^{-1}(U_\alpha)$, with $\psi\varphi_\alpha(x, y) = x$, $x \in U_\alpha$, $y \in F$.

(2) As a consequence of (1), a point of $\psi^{-1}(U_\alpha)$ has the coordinates (x, y), and a point of $\psi^{-1}(U_\alpha \cap U_\beta)$ has two sets of coordinates (x, y) and (x, y'), satisfying $\varphi_\alpha(x, y) = \varphi_\beta(x, y')$. It is required that $g_{\alpha\beta}(x)$: $y' \to y$ be a continuous mapping of $U_\alpha \cap U_\beta$ into G.

The spaces X and B are called the base space and the bundle respectively. Each subset $\psi^{-1}(x) \subset B$ is called a fiber.

This definition of a fiber bundle is too narrow in the sense that the coordinate neighborhoods and coordinate functions form a part of the definition. An equivalence relation has thus to be introduced. Two bundles (B, X), (B', X) with the same base space X and the same F, G are called equivalent if, $\{U_\alpha, \varphi_\alpha\}$, $\{V_\beta, \theta_\beta\}$ being respectively their coordinate neighborhoods and coordinate functions, there is a fiber-preserving homeomorphism T: $B \to B'$ such that the mapping $h_{\alpha\beta}(x)$: $y \to y'$ defined by $\theta_\beta(x, y) = T\varphi_\alpha(x, y')$ is a continuous mapping of $U_\alpha \cap V_\beta$ into G.

An important operation on fiber bundles is the construction from a given bundle of other bundles with the same structural group, in particular, the principal fiber bundle which has G as director space acted upon by G itself as the group of left translations. The notion of the principal fiber bundle has been at the core of Cartan's method of moving frames, although its modern version was first introduced by Ehresmann. It can be defined as follows: For $x \in X$, let G_x be the totality of all maps $\varphi_{\alpha,y}(x)$: $F \to \psi^{-1}(x)$ defined by $y \to \varphi_\alpha(x, g(y))$, $y \in F$, $g \in G$, relative to a coordinate neighborhood U_α containing x. G_x depends only on x. Let $B^* = \bigcup_{x \in X} G_x$ and define the mapping ψ^*: $B^* \to X$ by $\psi^*(G_x) = x$ and the coordinate functions $\varphi^*_\alpha(x, g) = \varphi_{\alpha,g}(x)$. Topologize B^* such that the φ^*_α's define homeomorphisms of $U_\alpha \times G$ into B^*. The bundle (B^*, X) so obtained is called a principal fiber bundle. This construction is an operation on the equivalence classes of bundles in the sense that two fiber bundles are equivalent if and only if their principal fiber bundles are equivalent. Similarly, an inverse operation can be defined, which will permit us to construct bundles with a given principal bundle and having as director space a given space acted upon by the structural group G. An important property of the principal fiber bundle is that B^* is acted upon by G as right translations.

For the purpose of differential geometry we shall assume that all spaces under consideration are differentiable manifolds and that our mappings are differentiable with Jacobian matrices of the highest rank everywhere. In particular, the structural group G will be assumed to be a connected Lie group. For simplicity we suppose our base space X to be compact, although a large part of our discussions holds without this assumption.

The implications of these assumptions are very far-reaching indeed. First of all we can draw into consideration the Lie algebra $L(G)$ of G. $L(G)$ is invariant under the left translations of G, while the right translations and the inner automorphisms of G induce on $L(G)$ a group of linear endomorphisms $\mathrm{ad}(G)$, called

the adjoint group of G. Relative to a base of $L(G)$ there are the left-invariant linear differential forms ω^i and the right-invariant linear differential forms π^i, each set consisting of linearly independent forms whose number is equal to the dimension of G. A fundamental theorem on Lie groups asserts that their exterior derivatives are given by

$$(1)\qquad \begin{aligned} d\omega^i &= -\frac{1}{2}\sum_{j,k} c^i_{jk}\omega^j \wedge \omega^k, \\ d\pi^i &= +\frac{1}{2}\sum_{j,k} c^i_{jk}\pi^j \wedge \pi^k, \qquad i, j, k = 1, \cdots, \dim G, \end{aligned}$$

where c^i_{jk} are the so-called constants of structure which are antisymmetric in the lower indices and which satisfy the well-known Jacobi relations.

Returning to our fiber bundle, the dual mapping of the mapping $g_{\alpha\beta}$: $U_\alpha \cap U_\beta \to G$ carries ω^i and π^i into linear differential forms in $U_\alpha \cap U_\beta$, which we shall denote by $\omega^i_{\alpha\beta}$ and $\pi^i_{\alpha\beta}$ respectively. Since $g_{\alpha\gamma} = g_{\alpha\beta}g_{\beta\gamma}$ in $U_\alpha \cap U_\beta \cap U_\gamma$, we have

$$(2)\qquad \omega^i_{\alpha\gamma} = \sum_j \mathrm{ad}(g_{\beta\gamma})^i_j \omega^j_{\alpha\beta} + \omega^i_{\beta\gamma}.$$

We can also interpret $\omega^i_{\alpha\beta}$ as a vector-valued linear differential form in $U_\alpha \cap U_\beta$, with values in $L(G)$, and shall denote it simply by $\omega_{\alpha\beta}$ when so interpreted.

The generalization of the notion of a tensor field in classical differential geometry leads to the following situation: Let E be a vector space acted on by a representation $M(G)$ of G. A tensorial differential form of degree r and type $M(G)$ is an exterior differential form u_α of degree r in each coordinate neighborhood U_α, with values in E, such that, in $U_\alpha \cap U_\beta$, $u_\alpha = M(g_{\alpha\beta})u_\beta$. The exterior derivative du_α of u_α is in general not a tensorial differential form. It is in order to preserve the tensorial character of the derivative that an additional structure, a connection, is introduced into the fiber bundle.

A connection in the fiber bundle is a set of linear differential forms θ_α in U_α, with values in $L(G)$, such that

$$(3)\qquad \omega_{\alpha\beta} = -\mathrm{ad}(g_{\alpha\beta})\theta_\alpha + \theta_\beta, \qquad \text{in } U_\alpha \cap U_\beta.$$

It follows from (2) that such relations are consistent in $U_\alpha \cap U_\beta \cap U_\gamma$. As can be verified without difficulty, a connection defines in the principal fiber bundle a field of tangent subspaces transversal to the fibers, that is, tangent subspaces which, together with the tangent space of the fiber, span at every point the tangent space of the principal bundle. It follows from elementary extension theorems that in every fiber bundle there can be defined a connection. As there is great freedom in the choice of the connection, the question of deciding the relationship between the properties of the bundle and those of the connection will be our main concern in this paper.

Let us first define the process of so-called absolute differentiation. Let $\bar{M}(X)$, $X \in L(G)$, be the representation of the Lie algebra $L(G)$ induced by the repre-

sentation $M(G)$ of G. Then we have

$$dM(g_{\alpha\beta}) = M(g_{\alpha\beta})\bar{M}(\theta_\beta) - \bar{M}(\theta_\alpha)M(g_{\alpha\beta}). \tag{4}$$

It follows that if we put for our tensorial differential form u_α of degree r and type $M(G)$

$$Du_\alpha = du_\alpha + \bar{M}(\theta_\alpha) \wedge u_\alpha, \tag{5}$$

the form Du_α will be a tensorial differential form of degree $r + 1$ and the same type $M(G)$.

To study the local properties of the connection we again make use of a base of the Lie algebra, relative to which the form θ_α has the components θ_α^i. We put

$$\Theta_\alpha^i = d\theta_\alpha^i + \frac{1}{2}\sum_{j,k} c_{jk}^i \theta_\alpha^j \wedge \theta_\alpha^k, \qquad \text{in } U_\alpha. \tag{6}$$

The form Θ_α, whose components relative to the base are Θ_α^i, is then an exterior quadratic differential form of degree 2, with values in $L(G)$. It is easy to verify that $\Theta_\alpha = \mathrm{ad}(g_{\alpha\beta})\Theta_\beta$ in $U_\alpha \cap U_\beta$. The Θ_α's therefore define a tensorial differential form of degree 2 and type $\mathrm{ad}(G)$, called the curvature tensor of the connection. In a manner which we shall not attempt to describe here, the curvature tensor and tensors obtained from it by successive absolute differentiations give all the local properties of the connection. In particular, the condition $\Theta_\alpha = 0$ is a necessary and sufficient condition for the connection to be flat, that is, to be such that $\theta_\alpha = 0$ by a proper choice of the coordinate functions.

The following formulas for absolute differentiation can easily be verified:

$$\begin{aligned} \bar{M}(\Theta_\alpha) &= d\bar{M}(\theta_\alpha) + \bar{M}(\theta_\alpha)^2, \\ D\Theta_\alpha &= 0, \\ D^2u_\alpha &= \bar{M}(\Theta_\alpha)u_\alpha. \end{aligned} \tag{7}$$

Such relations are known in classical cases, the second as the Bianchi identity.

We now consider real-valued symmetric multilinear functions $P(Y_1, \cdots, Y_k)$, $Y_i \in L(G)$, $i = 1, \cdots, k$, which are invariant, that is, which are such that $P(\mathrm{ad}(a)Y_1, \cdots, \mathrm{ad}(a)Y_k) = P(Y_1, \cdots, Y_k)$ for all $a \in G$. For simplicity we shall call such a function an invariant polynomial, k being its degree. By the definition of addition,

$$(P + Q)(Y_1, \cdots, Y_k) = P(Y_1, \cdots, Y_k) + Q(Y_1, \cdots, Y_k), \tag{8}$$

all invariant polynomials of degree k form an abelian group. Let $I(G)$ be the direct sum of these abelian groups for all $k \geqq 0$. If P and Q are invariant polynomials of degrees k and l respectively, we define their product PQ to be an invariant polynomial of degree $k + l$ given by

$$(PQ)(Y_1, \cdots, Y_{k+l}) = \frac{1}{N}\sum P(Y_{i_1}, \cdots, Y_{i_k})Q(Y_{i_{k+1}}, \cdots, Y_{i_{k+l}}), \tag{9}$$

where the summation is extended over all permutations of the vectors Y_i, and N is the number of such permutations. This definition of multiplication, together with the distributive law, makes $I(G)$ into a commutative ring, the ring of invariant polynomials of G.

Let $P \in I(G)$, with degree k. For Y_i we substitute the curvature tensor Θ. Then $P(\Theta) = P(\Theta, \cdots, \Theta)$ is an exterior differential form of degree $2k$, which, because of the invariance property of P, is defined everywhere in the base space X. From the Bianchi identity (7_2) it follows that $P(\Theta)$ is closed. Therefore, by the de Rham theory, $P(\Theta)$ determines an element of the cohomology ring $H(X)$ of X having as coefficient ring the field of real numbers. This mapping is a ring homomorphism

$$h: I(G) \to H(X) \tag{10}$$

of the ring of invariant polynomials of G into the cohomology ring of X. It is defined with the help of a connection in the bundle.

Our first main result is the following theorem of Weil: h is independent of the choice of the connection [22]. In other words, two different connections in the fiber bundle give rise to the same homomorphism h. To prove this we notice that if θ_α and θ'_α are the linear differential forms defining these connections, their difference $u_\alpha = \theta'_\alpha - \theta_\alpha$ is a linear differential form of type ad(G), with values in $L(G)$. With the help of u_α Weil constructs a differential form whose exterior derivative is equal to the difference $P(\Theta') - P(\Theta)$, for a given invariant polynomial P. Another proof has been given recently by H. Cartan, by means of an invariant definition of the homomorphism h.

Our next step consists in setting up a relationship between this homomorphism h and a homomorphism which is defined in a purely topological manner. This requires the concepts of an induced fiber bundle and a universal fiber bundle.

Let a mapping $f: Y \to X$ be given. The neighborhoods $\{f^{-1}(U_\alpha)\}$ then form a covering of Y and coordinate functions $\varphi'_\alpha: f^{-1}(U_\alpha) \times F \to f^{-1}(U_\alpha) \times \psi^{-1}(U_\alpha)$ can be defined by $\varphi'_\alpha(\eta, y) = \eta \times \varphi_\alpha(f(\eta), y)$. This defines a fiber bundle $Y \times \psi^{-1}(f(Y))$ over Y, with the same director space F and the same group G. The new bundle is said to be induced by the mapping f. If the original bundle has a connection given by the differential form θ_α in U_α, the dual mapping f^* of f carries θ_α into $f^*\theta_\alpha$ in $f^{-1}(U_\alpha)$ for which the relation corresponding to (3) is valid. The forms $f^*\theta_\alpha$ therefore define an induced connection in the induced bundle.

This method of generating new fiber bundles from a given bundle is very useful. Its value is based on the fact that it provides a way for the enumeration of fiber bundles. In fact, let the director space and the structural group G be given and fixed for our present considerations. A bundle with the base space X_0 is called universal relative to a space X if every bundle over X is equivalent to a bundle induced by a mapping $X \to X_0$ and if two such induced bundles are equivalent when and only when the mappings are homotopic. If, for a space X, there exists a universal bundle with the base space X_0, then the classes of bundles over X are in one-one correspondence with the homotopy classes of mappings $X \to X_0$,

so that the enumeration of the bundles over X reduces to a homotopy classification problem.

It is therefore of interest to know the circumstances under which a universal bundle exists. A sufficient condition for the bundle over X_0 to be universal for all compact spaces X of dimension less than or equal to n is that the bundle B_0 of its principal fiber bundle have vanishing homotopy groups up to dimension n inclusive: $\pi_i(B_0) = 0$, $0 \leqq i \leqq n$, where the condition $\pi_0 = 0$ means connectedness.

Under our assumptions that X is compact and that G is a connected Lie group, bundles can be found such that these conditions are fulfilled. First of all, according to a theorem due to E. Cartan, Malcev, Iwasawa, and Mostow, [12; 14; 15], G contains a maximal compact subgroup G_1, and the homogeneous space G/G_1 is homeomorphic to a Euclidean space. This makes it possible to reduce problems of equivalence, classification, etc. of bundles with the group G to the corresponding problems for G_1. Since G_1 is a compact Lie group, it has a faithful orthogonal representation and can be considered as a subgroup of the rotation group $R(m)$ operating in an m-dimensional Euclidean space E^m. Imbed E^m in an $(m + n + 1)$-dimensional Euclidean space E^{m+n+1} and consider the homogeneous space $\tilde{B} = R(m + n + 1)/(I_m \times R(n + 1))$ as a bundle over $X_0 = R(m + n + 1)/(G_1 \times R(n + 1))$, where I_m is the identical automorphism of E^m, and $R(n + 1)$ is the rotation group of the space E^{n+1} perpendicular to E^m in E^{m+n+1}. This is a principal bundle with G_1 as its structural group. By the covering homotopy theorem we can prove that $\pi_i(\tilde{B}) = 0$, $0 \leqq i \leqq n$. In this way the existence of a universal bundle is proved by an explicit construction.

Suppose that a universal bundle exists, with the base space X_0. Let $H(X, R)$ be the cohomology ring of X, relative to the coefficient ring R. Since the classes of bundles over X are in one-one correspondence with the homotopy classes of mappings $X \to X_0$, the homomorphism $h'\colon H(X_0, R) \to H(X, R)$ is completely determined by the bundle. h' will be called the characteristic homomorphism, its image $h'(H(X_0, R)) \subset H(X, R)$ the characteristic ring, and an element of the characteristic ring a characteristic cohomology class. It will be understood that the coefficient ring R will be the field of real numbers whenever it is dropped in the notation.

The universal bundle is of course not unique. However, given any two bundles which are universal for compact base-spaces of dimension less than or equal to n, it is possible to establish between their base spaces X_0 and X_0' a chain transformation of the singular chains of dimension less than or equal to n which gives rise to a chain equivalence. From this it follows that up to the dimension n inclusive, the cohomology rings of X_0 and X_0' are in a natural isomorphism. The characteristic homomorphism is therefore independent of the choice of the universal bundle. Although this conclusion serves our purpose, it may be remarked that, in terms of homotopy theory, a stronger result holds between X_0 and X_0', namely, they have the same homotopy-n-type. From this the above assertion follows as a consequence.

A knowledge of $H(X_0, R)$ would be necessary for the description of the characteristic homomorphism. Since elements of dimension greater than n ($= \dim X$) of $H(X_0, R)$ are mapped into zero by dimensional considerations, $H(X_0, R)$ can be replaced by any ring which is isomorphic to it up to dimension n inclusive. On the other hand, it follows from the discussions of the last section that the choice of the universal bundle is immaterial, so that we can take the one whose base space is $X_0 = R(m + n + 1)/(G_1 \times R(n + 1))$. Using a connection in this universal bundle, we can, according to a process given above, define a homomorphism $h_0 = I(G_1) \to H(X_0)$ of the ring of invariant polynomials of G_1 into $H(X_0)$. X_0 being a homogeneous space, its cohomology ring $H(X_0)$ with real coefficients can be studied algebraically by methods initiated by E. Cartan and recently developed with success by H. Cartan, Chevalley, Kozsul, Leray, and Weil [13]. Thus it has been shown that, up to dimension n, h_0 is a one-one isomorphism. We may therefore replace $H(X_0)$ by $I(G_1)$ in the homomorphism h' and write the characteristic homomorphism as

$$h' \colon I(G_1) \to H(X). \tag{11}$$

This homomorphism h' is defined by the topological properties of the fiber bundle.

On the other hand, the homomorphism $h \colon I(G) \to H(X)$ defined above can be split into a product of two homomorphisms. Since an invariant polynomial under G is an invariant polynomial under G_1, there is a natural homomorphism

$$\sigma \colon I(G) \to I(G_1). \tag{12}$$

Since G_1 can be taken to be the structural group, the homomorphism

$$h_1 \colon I(G_1) \to H(X) \tag{13}$$

is defined. Now, a connection with the group G_1 can be considered as a connection with the group G. Using such a connection, we can easily prove

$$h = h_1 \sigma. \tag{14}$$

Our main result which seems to include practically all our present knowledge on the subject consists in the statement:

$$h' = h_1 . \tag{15}$$

Notice that h' is defined by the topological properties of the bundle and h_1 by the help of a connection, so that our theorem gives a relationship between a bundle and a connection defined in it, which is restrictive in one way or the other. In particular, when the structural group G is compact, we have $G_1 = G$ and σ is the identity, and the characteristic homomorphism is in a sense determined by the connection. For instance, it follows that the characteristic ring of the bundle has to be zero when a connection can be defined such that $h(I(G)) = 0$.

A proof for this theorem is obtained by first establishing it for the universal bundle. Under the mapping $f: X \to X_0$ it is then true for the induced bundle and the induced connection. Using the theorem of Weil that h is independent of the choice of the connection, we see that the relation is true for any connection in the bundle.

A great deal can be said about the rings of invariant polynomials $I(G)$, $I(G_1)$ and the homomorphism σ. When the structural group is compact, such statements can usually be proved more simply by topological considerations. In the other case we have to make use of the cohomology theory of Lie algebras. As we do not wish to discuss this, we shall restrict ourselves to the explanation of the corresponding topological notions. For this purpose we shall first discuss compact groups, that is, we begin by confining our attention to G_1.

We first recall some results on compact group manifolds. All the maximal abelian subgroups are conjugate and are isomorphic to a torus whose dimension is called the rank of the group. By an idea due essentially to Pontrjagin [16] we can define an operation of the homology classes of G_1 on the cohomology classes of G_1. In fact, $m: G_1 \times G_1 \to G_1$ being defined by the group multiplication, the image $m^*\gamma^k$ of a cohomology class of dimension k of G under the dual homomorphism m^* can be written $m^*\gamma^k = \sum u_i^r \times v_i^{k-r}$. The operation of a homology class c^s of dimension $s \leqq k$ on γ^k is then defined as $i(c^s)\gamma^k = \sum_i KI(c^s, u_i^s)v_i^{k-s}$. We call this operation an interior product. A cohomology class γ^k of G_1 is called primitive if its interior product by any homology class of dimension s, $1 \leqq s \leqq k-1$, is zero. The homology structure of compact group manifolds (with real coefficients) has a description given by the following theorem of Hopf and Samelson [11; 18]: (1) all primitive cohomology classes are of odd dimension; (2) the vector space of the primitive classes has as dimension the rank of G; (3) the cohomology ring of G_1 is isomorphic to the Grassmann algebra of the space of primitive classes.

The primitive classes play a rôle in the study of the universal principal fiber bundle $\psi: B_0 \to X_0$. Identify a fiber $\psi^{-1}(x)$ $(x \in X_0)$ with G_1, and let i be the inclusion mapping of G_1 into B_0. If γ^k is a cocycle of X_0, $\psi^*\gamma^k$ is a cocycle of B_0. Since B_0 is homologically trivial, there exists a cochain β^{k-1} having $\psi^*\gamma^k$ as coboundary. Then $i^*\beta^{k-1}$ is a cocycle in G_1 whose cohomology class depends only on that of γ^k. The resulting mapping of the cohomology classes is called a transgression. It is an additive homomorphism of the ring of invariant polynomials of G_1 into the cohomology ring of G_1 and it carries an invariant polynomial of degree k into a cohomology class of dimension $2k-1$. Chevalley and Weil proved that the image is precisely the space of the primitive classes.

When the group G is noncompact, the consideration of its Lie algebra allows us to generalize the above notions, at least under the assumption that G is semisimple. H. Cartan, Chevalley, and Koszul have developed a very comprehensive theory dealing with the situation, which can be considered in a sense as the algebraic counterpart of the above treatment. Among their consequences we mention the following which is interesting for our present purpose: The ring of in-

variant polynomials under G has a set of generators equal to the rank of G; these can be so chosen that their images under transgression span the space of primitive classes of G.

Using the fact that the cohomology theory of Lie algebras and transgression can be defined algebraically, and therefore for G, we have the following diagram

$$\begin{array}{ccc} H(G) & \xrightarrow{i^*} & H(G_1) \\ t\uparrow & & \uparrow t_1 \\ I(G) & \xrightarrow{\sigma} & I(G_1). \end{array}$$

It is not difficult to prove that commutativity holds in this diagram. Hence the image under σ depends on the image under i^* of $H(G)$, that is, on the "homological position" of G_1 in G. In general, $\sigma[I(G)] \neq I(G_1)$.

There are relations between the characteristic cohomology classes in our definition and the classes carrying the same name in the topological method of obstructions but we cannot discuss them in detail. The latter come into being when one attempts to define a cross-section in the fiber bundle (that is, a mapping f of X into B, such that ψf is the identity) by extension over the successive skeletons; they are cohomology classes over groups of coefficients which are the homotopy groups of the director space. As we shall see from examples, it is sometimes possible to identify them by identifying the coefficient groups. In general, however, our characteristic classes are based on homological considerations, while those of obstruction theory are based on homotopy considerations. Their rôles are complementary.

We shall devote the rest of this lecture to the consideration of examples. Although the main results will follow from the general theorems, special problems arise in individual cases which can be of considerable interest. To begin with, take for G the rotation group in m variables, and suppose that a connection is given in the bundle. This includes in particular the case of orientable Riemann manifolds with a positive definite metric, the bundle being the tangent bundle of the manifold and the connection being given by the parallelism of Levi-Civita; it also includes, among other things, the theory of orientable submanifolds imbedded in an orientable Riemann manifold.

By a proper choice of a base of the Lie algebra of $G = R(m)$, the space of the Lie algebra can be identified with the space of skew-symmetric matrices of order m. The connection can therefore be defined, in every coordinate neighborhood, by a skew-symmetric matrix of linear differential forms $\theta = (\theta_{ij})$, and its curvature tensor by a skew-symmetric matrix of quadratic differential forms $\Theta = (\Theta_{ij})$. The effect of the adjoint group is given by $\mathrm{ad}(a)\Theta = A\Theta\,{}^tA$, where A is a proper orthogonal matrix and ${}^tA$ is its transpose.

The first question is of course to determine a set of generators for the ring of invariant polynomials; using the fundamental theorem on invariants, it is easy to do this explicitly [23]. Instead of the invariant polynomials we write the

corresponding differential forms:

$$\Delta_s = \Theta_{i_1 i_2} \cdots \Theta_{i_s i_1}, \qquad s = 2, 4, \cdots, m+1, \qquad m \text{ odd}$$

$$\Delta_s = \Theta_{i_1 i_2} \cdots \Theta_{i_s i_1}, \qquad s = 2, 4, \cdots, m-2, \qquad m \text{ even} \tag{16}$$

$$\Delta_0 = \epsilon_{i_1 \cdots i_m} \Theta_{i_1 i_2} \cdots \Theta_{i_{m-1} i_m}, \qquad m \text{ even},$$

where repeated indices imply summation and where $\epsilon_{i_1 \cdots i_m}$ is the Kronecker tensor, equal to $+1$ or -1 according as $i_1, \cdots, i_m$ form an even or odd permutation of $1, \cdots, m$ and otherwise to 0. Since the rank of $R(m)$ is $(m+1)/2$ or $m/2$ according as m is odd or even, we verify here that the number of the above generators is equal to the rank. They form a complete set of generators, because they are obviously independent.

It follows that the cohomology classes determined by these differential forms or by polynomials in these differential forms depend only on the bundle and not on the connection. As a consequence, if all these differential forms are zero, the characteristic ring is trivial. The differential forms in (16) were first given by Pontrjagin [17].

For geometric applications it is useful to have a more explicit description of the base space of a universal bundle. This is all the more significant, since it would then allow us to study the characteristic homomorphisms with coefficient rings other than the field of real numbers. Our general theory gives as such a base space the Grassmann manifold

$$X_0 = R(m+n+1)/(R(m) \times R(n+1)),$$

which can be identified with the space of all oriented m-dimensional linear spaces through a point 0 of an $(m+n+1)$-dimensional Euclidean space E^{m+n+1}.

The homology structure of Grassmann manifolds has been studied by Ehresmann [9, 10]. A cellular decomposition can be constructed by the following process: Take a sequence of linear spaces

$$0 \subset E^1 \subset E^2 \subset \cdots \subset E^{m+n} \subset E^{m+n+1}.$$

Corresponding to a set of integers

$$0 \leqq a_1 \leqq a_2 \leqq \cdots \leqq a_m \leqq n+1,$$

denote by $(a_1 \cdots a_m)$ the set of all m-dimensional linear spaces $\xi \in X_0$ such that

$$\dim (\xi \cap E^{a_i+i}) \geqq i, \qquad i = 1, \cdots, m.$$

The interior points of $(a_1 \cdots a_m)$ form two open cells of dimension $a_1 + \cdots + a_m$. These open cells constitute a cellular decomposition of X_0, whose incidence relations can be determined. From this we can determine the homology and cohomology groups of X_0. In particular, it follows that the symbol $(a_1 \cdots a_m)^{\pm}$ can be used to denote a cochain, namely, the one which has the value $+1$ for the corresponding open cells and has otherwise the value zero. The characteristic homomorphism can then be described as a homomorphism of

combinations of such symbols into the cohomology ring $H(X, R)$ of X. When R is the field of real numbers, the result is particularly simple. In fact, a base for the cohomology groups of dimensions less than or equal to n consists of cocycles having as symbols those for which all a_i are even, together with the cocycle $(1 \cdots 1)$ when m is even.

This new description of the characteristic homomorphism allows us to give a geometric meaning to individual characteristic classes. In this respect the class $h'((1 \cdots 1))$, which exists only when m is even, deserves special attention. In fact, the bundle with the director space $S^{m-1} = R(m)/R(m-1)$ constructed from the principal bundle is a bundle of $(m-1)$-spheres in the sense of Whitney. For such a sphere bundle, Whitney introduced a characteristic cohomology class W^m with integer coefficients. It can be proved that W^m, when reduced to real coefficients, is precisely the class $h'((1 \cdots 1))$. On the other hand, the latter can be identified on the universal bundle with a numerical multiple of the class defined by the differential form Δ_0. Taking the values of these classes for the fundamental cycle of the base manifold, we can write the result in an integral formula

$$W^m \cdot X = c \int_X \Delta_0 , \tag{17}$$

where c is a numerical factor and X denotes a fundamental cycle of the base manifold. For a Riemann manifold, $W^m \cdot X$ is equal to the Euler-Poincaré characteristic of X and our formula reduces to the Gauss-Bonnet formula [3].

We introduce the notations

$$\begin{aligned} P^{4k} &= h'(0 \cdots 0\underbrace{2 \cdots 2}_{2k \text{ times}}) \\ \overline{P}^{4k} &= h'(0 \cdots 0\ 2k\ 2k) \\ \chi^m &= h'(1 \cdots 1), \qquad m \text{ even}, \end{aligned} \tag{18}$$

where the symbols denote also the cohomology classes to which the respective cocycles belong. By studying the multiplicative structure of the cohomology ring of X_0, we can prove that the characteristic homomorphism is determined by the classes P^{4k}, χ^m, $4k \leqq \dim X$ or the classes $\overline{P}^{4k}$, χ^m, $4k \leqq \dim X$.

We shall mention an application of the classes $\overline{P}^{4k}$. Restricting ourselves for simplicity to the tangent bundle of a compact differentiable manifold, the conditions $\overline{P}^{4k} = 0$, $2k \geqq n + 2$, are necessary for the manifold to be imbeddable into a Euclidean space of dimension $m + n + 1$. We get thus criteria on the impossibility of imbedding which can be expressed in terms of the curvature tensor of a Riemann metric on the manifold.

The second example we shall take up is the case that G is the unitary group. Such bundles occur as tangent bundles of complex analytic manifolds, and the introduction of an Hermitian metric in the manifold would give rise to a connection in the bundle.

The space of the Lie algebra of the unitary group $U(m)$ in m variables can be identified with the space of $m \times m$ Hermitian matrices A (${}^t\bar{A} = A$). A connection is therefore defined in each coordinate neighborhood by an Hermitian matrix of linear differential forms $\theta = (\theta_{ij})$ and its curvature tensor by an Hermitian matrix of quadratic differential forms $\Theta = (\Theta_{ij})$. Under the adjoint group the curvature tensor is transformed according to $\mathrm{ad}(a)\Theta = A\Theta{}^t\bar{A}$, A being a unitary matrix. Using this representation of the adjoint group, a set of invariant polynomials can be easily exhibited. We give their corresponding differential forms as

$$\Lambda_k = \Theta_{i_1 i_2} \cdots \Theta_{i_k i_1}, \qquad k = 1, \cdots, m. \tag{19}$$

Since they are clearly independent and their number is equal to the rank m of $U(m)$, they form a complete set of generators in the ring of invariant polynomials.

As in the case of the rotation group the complex Grassmann manifold $X_0 = U(m+n)/(U(m) \times U(n))$ is the base space of a universal bundle, whose study would be useful for some geometric problems. The results are simpler than the real case, but we shall not describe them here. A distinctive feature of the complex case is that a set of generators can be chosen in the ring of invariant polynomials whose corresponding differential forms are

$$\Psi_r = \frac{1}{(2\pi\,(-1)^{1/2})^{m-r+1}\,(m-r+1)!} \sum \delta(i_1 \cdots i_{m-r+1}; j_1 \cdots j_{m-r+1}) \cdot \Theta_{i_1 j_1} \cdots \Theta_{i_{m-r+1} j_{m-r+1}}, \qquad r = 1, \cdots, m, \tag{20}$$

where $\delta(i_1 \cdots i_{m-r+1}\,;\, j_1 \cdots j_{m-r+1})$ is zero except when $j_1, \cdots, j_{m-r+1}$ form a permutation of $i_1, \cdots, i_{m-r+1}$, in which case it is $+1$ or -1 according as the permutation is even or odd, and where the summation is extended over all indices $i_1, \cdots, i_{m-r+1}$ from 1 to m. This set of generators has the advantage that the cohomology classes determined by the differential forms have a simple geometrical meaning. In fact, they are the classes, analogous to the Stiefel-Whitney classes, for the bundle with the director space $U(m)/U(m-r)$. As such they are primary obstructions to the definition of a cross-section and are therefore more easily dealt with [4]. Substantially the same classes have been introduced by M. Eger and J. A. Todd in algebraic geometry, even before they first made their appearance in differential geometry [6; 20].

The situation is different for bundles with the rotation group, since the Stiefel-Whitney classes, except the highest-dimensional one, are essentially classes mod 2 and therefore do not enter into our picture. However, there is a close relationship between bundles with the group $R(m)$ and bundles with the group $U(m)$. In fact, given a bundle with the group $R(m)$, we can take its Whitney product with itself, which is a bundle with the same base space and the group $R(m) \times R(m)$. The latter can be imbedded into $U(m)$, so that we get a bundle with the group $U(m)$. Such a process is frequently useful in reducing problems on bundles with the rotation group to those on bundles with the unitary group.

We shall take as last example the case that the group is the component of the identity of the general linear group $GL(m)$ in m variables. A connection in the bundle is called an affine connection. An essential difference from the two previous examples is that the group is here noncompact.

The Lie algebra of the group $GL(m)$ can be identified with the space of all m-rowed square matrices, so that the curvature tensor in each coordinate neighborhood is given by such a matrix of exterior quadratic differential forms: $\Theta = (\Theta_i^j)$. The effect of the adjoint group being defined by $\mathrm{ad}(a)\Theta = A\ \Theta\ A^{-1}$, $a \in GL(m)$, it is easily seen that a set of generators of the ring of invariant polynomials can be so chosen that the corresponding differential forms are

$$M_s = \Theta_{i_1}^{i_2} \cdots \Theta_{i_s}^{i_1}, \qquad s = 1, \cdots, m-1. \tag{21}$$

According to the general theory it remains to determine the homomorphism of the ring of invariant polynomials under $GL(m)$ into the ring of invariant polynomials under its maximal compact subgroup, which is in this case the rotation group $R(m)$. It is seen that M_s, for even s, is mapped into Δ_s, and, for odd s, is mapped into zero. The class defined by Δ_0 does not belong to the image of the homomorphism. This fact leads to the interesting explanation that a formula analogous to the Gauss-Bonnet formula does not exist for an affine connection.

Perhaps the most important of the bundles is the tangent bundle of a differentiable manifold. We mentioned above the identification of a certain characteristic class with the Euler-Poincaré characteristic of the manifold, at least for the case that the manifold is orientable and of even dimension. Beyond this very little is known on the relations between topological invariants of the manifold and the characteristic homomorphism of its tangent bundle. Recently, contributions have been made by Thom and Wu which bear on this question [21; 25]. Although it is not known whether a topological manifold always has a differentiable structure, nor whether it can have two essentially different differentiable structures, Thom and Wu proved that the characteristic homomorphisms of the tangent bundle, with coefficients mod 2 and with coefficients mod 3, are independent of the choice of the differentiable structure, provided one exists. Briefly speaking, this means that such characteristic homomorphisms are topological invariants of differentiable manifolds. The proof for coefficients mod 3 is considerably more difficult than the case mod 2.

For bundles with other groups such questions have scarcely been asked. The next case of interest is perhaps the theory of projective connections derived from the geometry of paths. In this case the bundle with the projective group depends both on the tangent bundle and the family of paths. It would be of interest to know whether or what part of the characteristic homomorphism is a topological invariant of the manifold.

Before concluding we shall mention a concept which has no close relation with the above discussion, but which should be of importance in the theory of connections, namely, the notion of the group of holonomy. It can be defined as follows: if ω is the left-invariant differential form in G, with values in $L(G)$,

and if θ_α defines a connection, the equation

$$\theta_\alpha + \omega = 0 \tag{22}$$

is independent of the coordinate neighborhood. When a parametrized curve is given in the base manifold, this differential equation defines a family of integral curves in G invariant under left translations of the group. Let $x \in X$ and consider all closed parametrized curves in X having x as the initial point. To every such curve C let $a(C)$ be the endpoint of the integral curve which begins at the unit element e of G. All such points $a(C)$ form a subgroup H of G, the group of holonomy of the connection.

Added in proof: The details of some of the discussions in this article can be found in mimeographed notes of the author, *Topics in differential geometry*, Institute for Advanced Study, Princeton, 1951.

Bibliography

1. E. Cartan, *Les groupes d'holonomie des espaces généralisés*, Acta Math. vol. 48 (1926) pp. 1–42.

2. ———, *L'extension du calcul tensoriel aux géométries non-affines*, Ann. of Math. (2) vol. 38 (1947) pp. 1–13.

3. S. S. Chern, *A simple intrinsic proof of the Gauss-Bonnet formula for closed Riemannian manifolds*, Ann. of Math. (2) vol. 45 (1944) pp. 747–752.

4. ———, *Characteristic classes of Hermitian manifolds*, Ann. of Math. (2) vol. 47 (1946) pp. 85–121.

5. ———, *Some new viewpoints in differential geometry in the large*, Bull. Amer. Math. Soc. vol. 52 (1946) pp. 1–30.

6. M. Eger, *Sur les systèmes canoniques d'une variété algébrique*, C. R. Acad. Sci. Paris vol. 204 (1937) pp. 92–94, 217–219.

7. C. Ehresmann, *Les connexions infinitésimales dans un espace fibré différentiable*, Seminaire Bourbaki, March, 1950. This paper contains references to earlier notes of the author on the subject in C. R. Acad. Sci. Paris.

8. ———, *Sur la théorie des espaces fibrés*, Colloque de Topologie Algébrique, Paris, 1947, pp. 3–15.

9. ———, *Sur la topologie de certains espaces homogènes*, Ann. of Math. (2) vol. 35 (1934) pp. 396–443.

10. ———, *Sur la topologie des certaines variétés algébriques réelles*, Journal de Mathématiques vol. 104 (1939) pp. 69–100.

11. H. Hopf, *Über die Topologie der Gruppenmannigfaltigkeiten und ihre Verallgemeinerungen*, Ann. of Math. (2) vol. 42 (1941) pp. 22–52.

12. K. Iwasawa, *On some types of topological groups*, Ann. of Math. (2) vol. 50 (1949) pp. 507–558.

13. J. L. Koszul, *Homologie et cohomologie des algèbres de Lie*, Bull. Soc. Math. France vol. 78 (1950) pp. 65–127.

14. A. Malcev, *On the theory of Lie groups in the large*, Rec. Math. (Mat. Sbornik) N. S. vol. 16 (1945) pp. 163–189.

15. G. D. Mostow, *A new proof of E. Cartan's theorem on the topology of semi-simple groups*, Bull. Amer. Math. Soc. vol. 55 (1949) pp. 969–980.

16. L. Pontrjagin, *Homologies in compact Lie groups*, Rec. Math. (Mat. Sbornik) N. S. vol. 48 (1939) pp. 389–422.

17. ———, *On some topological invariants of Riemannian manifolds*, C. R. (Doklady) Acad. Sci. URSS vol. 43 (1944) pp. 91–94.

18. H. SAMELSON, *Beiträge zur Topologie der Gruppen-Mannigfaltigkeiten*, Ann. of Math. (2) vol. 42 (1941) pp. 1091–1137.

19. N. E. STEENROD, *Topology of fiber bundles*, Princeton, 1951. With few exceptions this will be our standard reference on fiber bundles.

20. J. A. TODD, *The geometrical invariants of algebraic loci*, Proc. London Math. Soc. (2) vol. 43 (1937) pp. 127–138.

21. R. THOM, *Variétés plongées et i-carrés*, C. R. Acad. Sci. Paris vol. 230 (1950) pp. 507–508.

22. A. WEIL, *Géométrie différentielle des espaces fibrés*, unpublished.

23. H. WEYL, *Classical groups*, Princeton, 1939.

24. H. WHITNEY, *On the topology of differentiable manifolds*, Lectures on Topology, Michigan, 1941, pp. 101–141.

25. WEN-TSÜN WU, *Topological invariance of Pontrjagin classes of differentiable manifolds*, to be published.

UNIVERSITY OF CHICAGO,
CHICAGO, ILL., U. S. A.

SUR LES VARIÉTÉS PRESQUE COMPLEXES

CHARLES EHRESMANN

1. Introduction. Etant donnée une variété topologique V_{2n}, de dimension $2n$, existe-t-il sur V_{2n} une *structure analytique complexe*? Plus abordable paraît la question suivante: Etant donnée une variété *différentiable* V_{2n}, existe-t-il sur V_{2n} une structure analytique complexe *subordonnée à sa structure différentiable*? Soit $T(V_{2n})$ l'espace des vecteurs tangents à V_{2n} et T_x l'espace vectoriel tangent en $x \in V_{2n}$. $T(V_{2n})$ est un espace fibré de base V_{2n} et de fibres T_x isomorphes à l'espace vectoriel R^{2n}. Une *structure presque complexe* sur V_{2n} sera définie par la donnée dans T_x d'une structure vectorielle complexe subordonnée à sa structure vectorielle réelle et dépendant d'une façon continue de x. Toute structure analytique complexe subordonnée à la structure différentiable de V_{2n} détermine sur V_{2n} une structure presque complexe, mais en général une structure presque complexe ne dérive pas d'une structure analytique complexe et on ignore si une *variété presque complexe* (c'est-à-dire, une variété munie d'une structure presque complexe) admet aussi une structure analytique complexe. La recherche des structures presque complexes sur V_{2n} est un problème de la théorie des espaces fibrés. Je rappellerai, en les complétant, les résultats que j'ai exposés au Colloque de Topologie Algébrique de Paris (1947) et j'indiquerai quelques résultats de Wen-tsün Wu; mais je ne pourrai pas exposer les méthodes de H. Hopf, qui a abordé la même question d'un point de vue un peu différent. Les nombres entre crochets renvoient à la bibliographie.

2. Structures fibrées subordonnées à une structure fibrée vectorielle [2]. Soit $E(B, F, G, H)$ un espace fibré de base B, de fibres isomorphes à F, de groupe structural topologique G. Nous supposons F muni d'une structure admettant G comme groupe d'automorphismes. Par les homéomorphismes distingués, dont l'ensemble est H, cette structure est transportée sur une structure bien déterminée dans chaque fibre F_x. H est alors l'ensemble des isomorphismes de F sur les fibres. Il est muni d'une structure fibrée $H(B, G, G_\gamma, \overline{H})$ et s'appelle espace *fibré principal associé*. Etant donné un sous-groupe G' de G, muni de la topologie induite, une structure fibrée $E(B, F, G', H')$ est dite *subordonnée* à $E(B, F, G, H)$ lorsque $H' \subset H$. Toute structure $E(B, F, G', H')$ détermine canoniquement une structure $E(B, F, G, H)$ à laquelle elle est subordonnée. Supposons que les classes sG' déterminent une structure fibrée sur G. Les structures $E(B, F, G', H')$ subordonnées à une structure donnée $E(B, F, G, H)$ correspondent alors d'une façon biunivoque aux *sections* de *l'espace fibré associé* à $E(B, F, G, H)$ par l'homomorphisme φ de G sur le groupe de transformations de l'espace homogène G/G', défini par $\varphi(s)(tG') = s(tG')$. C'est l'espace H/G' des classes hG', où $h \in G$; sa base est B et ses fibres sont isomorphes à G/G'. Les structures subordonnées $E(B, F, G', H')$ se répartissent en classes d'homotopie correspondant aux classes d'homotopie des sections; les structures d'une même classe sont iso-

morphes. Si $\mathfrak{S}'$ désigne une structure sur F admettant G' comme groupe d'automorphismes, l'espace H/G' peut s'appeler l'espace des structures isomorphes à $\mathfrak{S}'$ et subordonnées aux structures données sur les fibres F_x.

Soit R^n (resp. C^n, Q^n) l'espace numérique réel (resp. complexe, quaternionien), L_n (resp. L'_n, L''_n) le groupe linéaire homogène de R^n (resp. C^n, Q^n), O_n (resp. O'_n, O''_n) le groupe orthogonal dans R^n (resp. unitaire dans C^n, unitaire quaternionien dans Q^n), L_n^+ (resp. O_n^+) la composante connexe de l'unité de L_n (resp. O_n). Une structure $E(B, F, G, H)$ sera appelée structure fibrée *vectorielle réelle* (resp. *réelle orientée*, *complexe*, *quaternionienne*, *euclidienne*, *euclidienne orientée*, *hermitienne*, *hermitienne quaternionienne*) si G est L_n (resp. L_n^+, L'_n, L''_n, O_n, O_n^+, O'_n, O''_n), F étant suivant les cas R^n, C^n ou Q^n. Comme L_n/O_n (resp. L_n^+/O_n^+, L'_n/O'_n, L''_n/O''_n) est homéomorphe à un espace numérique, toute structure fibrée vectorielle réelle (resp. réelle orientée, complexe, quaternionienne) admet des structures euclidiennes (resp. euclidiennes orientées, hermitiennes, hermitiennes quaternioniennes) subordonnées et celles-ci appartiennent toutes à une même classe. En identifiant R^{2n} à C^n et C^n à Q^m, si $n = 2m$, on a $L'_n \subset L_{2n}^+$, $L''_m \subset L'_n$, $O'_n \subset O_{2n}^+$, $O''_m \subset O'_n$, et le problème d'existence de structures subordonnées se pose pour des structures du type suivant:

$$E(B, R^{2n}, L_{2n}, H), \qquad E(B, R^{2n}, O_{2n}, \cdot)$$
$$E(B, R^{2n}, L_{2n}^+, \cdot), \qquad E(B, R^{2n}, O_{2n}^+, \cdot)$$
$$E(B, C^n, L'_n, \cdot), \qquad E(B, C^n, O'_n, \cdot)$$
$$E(B, Q^m, L''_m, \cdot), \qquad E(B, Q^m, O''_m, \cdot).$$

D'après la remarque précédente, on peut toujours se ramener au cas de structures figurant dans la deuxième colonne ci-dessus. L'espace $T(V_{2n})$ associé à une variété différentiable V_{2n} est muni d'une structure fibrée vectorielle réelle, dont les structures subordonnées s'appellent respectivement: structure *vectorielle tangente orientée*, *presque complexe*, *presque quaternionienne*, *riemannienne*, *presque hermitienne*, *presque hermitienne quaternionienne*.

3. Les structures vectorielles complexes sur R_{2n}. Soit $(\epsilon_1, \cdots, \epsilon_n)$ la base canonique de C^n et identifions C^n à R^{2n} en identifiant $(\epsilon_1, i\epsilon_1, \cdots, \epsilon_n, i\epsilon_n)$ à la base canonique de R^{2n}. Alors L'_n est le sous-groupe de L_{2n} qui laisse invariante la transformation I_0 définie par $I_0 z = iz$, où $z \in C^n$. Une structure vectorielle complexe sur R^{2n}, subordonnée à la structure vectorielle réelle, est définie par une transformation linéaire I de R^{2n} telle que $I^2 = -1$, c'est-à-dire, $I(Ix) = -x$ pour $x \in R^{2n}$; le produit du nombre complexe $a + bi$ par x sera $(a + bI)x$. Les vecteurs x et Ix sont linéairement indépendants et déterminent un plan invariant par I. R^{2n} admet des bases de la forme $(e_1, Ie_1, \cdots, e_n, Ie_n)$; l'orientation correspondante de R^{2n} ne dépend que de I; c'est *l'orientation associée à* I. L'espace des structures vectorielles complexes sur R^{2n} est L_{2n}/L'_n, dont la composante connexe L_{2n}^+/L'_n est l'espace des structures complexes dont l'orientation associée est aussi associée à I_0.

Considérons R^{2n} comme l'espace des vecteurs réels de C^{2n}. La transformation I se prolonge à C^{2n} et admet les valeurs propres $\pm i$. L'ensemble des vecteurs propres correspondant à $-i$ forme un sous-espace X_n de dimension n. L'ensemble des vecteurs propres correspondant à $+i$ est $\bar{X}_n$, imaginaire conjugué de X_n, et l'on a $X_n \cap \bar{X}_n = 0$. Inversement tout sous-espace X_n de C^{2n} tel que $X_n \cap \bar{X}_n = 0$ détermine une transformation I. On peut définir X_n par n formes linéaires sur C^{2n} dont les restrictions à R^{2n} sont des formes linéaires à valeurs complexes ne s'annulant simultanément que pour le vecteur 0.

Posons $F(x, x) = x_1^2 + y_1^2 + \cdots + x_n^2 + y_n^2$, où $(x_1, y_1, \cdots, x_n, y_n)$ sont les coordonnées canoniques de $x \in R^{2n}$. O_{2n}/O'_n est l'espace des transformations I laissant invariante la forme quadratique $F(x, x)$, condition équivalente à $F(x, Ix) = 0$ ou $I \in O^+_{2n}$. L'espace X_n associé à I est alors une génératrice du cône défini dans C^{2n} par $F(x, x) = 0$. Donc O^+_{2n}/O'_n, que nous désignons par Γ_n, s'identifie à l'une des composantes connexes de l'espace des génératrices X_n de ce cône.

Comme $\lambda \in L_{2n}$ se prolonge à C^{2n}, l'espace fibré $T(V_{2n})$ admet un espace fibré associé $T^c(V_{2n})$ dont les fibres T^c_x sont isomorphes à C^{2n} et qui admet $T(V_{2n})$ comme sous-espace. Un élément de T^c_x s'appelle *vecteur complexe tangent* à V_{2n} en x, un sous-espace X_p de T^c_x s'appelle p-élément complexe tangent en x.

Une structure presque complexe sur V_{2n} est donc déterminée par un champ de transformations linéaires I, définies dans T_x et telles que $I^2 = -1$, ou par un champ de n-éléments complexes X_n tels qu'en chaque point $x \in V_{2n}$ on ait $X_n \cap \bar{X}_n = x$. Au voisinage de x elle est définie encore par n formes de Pfaff sur V_{2n}, à valeurs complexes et ne s'annulant simultanément pour aucun vecteur réel non nul. Il lui correspond une orientation bien déterminée de V_{2n}. Une structure presque hermitienne subordonnée à une structure riemannienne est définie par un champ de transformations orthogonales I ou par un champ de n-éléments isotropes.

4. Formes différentielles extérieures quadratiques sur V_{2n}. A une transformation I dans R^{2n} telle que $F(x, Ix) = 0$ est associée la forme bilinéaire alternée $\Psi(x, x') = F(Ix, x')$ de rang $2n$, et la forme $\Phi(x, x') = F(x, x') - i\Psi(x, x')$, qui est une forme d'Hermite définie positive par rapport à la structure complexe définie par I. A toute structure presque hermitienne sur V_{2n} est donc associée une forme différentielle extérieure quadratique Ω de rang $2n$; l'orientation associée est définie par Ω^n, forme de degré $2n$ non nulle en chaque point. Réciproquement toute forme différentielle extérieure quadratique Ω partout de rang $2n$ sur V_{2n} est associée de cette façon à des structures presque hermitiennes, qui sont toutes de même classe et qui correspondent à l'orientation définie par Ω^n. Ceci résulte du fait que $\tilde{L}_{2n}/L'_n$ est homéomorphe à un espace numérique, $\tilde{L}_{2n}$ désignant le sous-groupe de L_{2n} qui laisse invariante la forme $\Psi_0(x, x') = x_1 y'_1 - x'_1 y_1 + \cdots + x_n y'_n - x'_n y_n$.

Donc l'existence d'une structure presque complexe sur la variété orientée V_{2n} est équivalente à l'existence d'une forme différentielle extérieure quadratique

Ω telle que Ω^n soit non nulle partout et définisse l'orientation donnée. Appelons *variété presque kählerienne* une variété presque hermitienne dont la forme extérieure associée Ω est fermée, c'est-à-dire, $d\Omega = 0$. Appelons *variété symplectique* une variété V_{2n} munie d'une forme fermée Ω telle que $\Omega^n \neq 0$ en chaque point. Une variété symplectique admet toujours une structure presque kählerienne subordonnée et possède des propriétés topologiques plus particulières qu'une variété presque complexe quelconque. En particulier, si elle est compacte, ses nombres de Betti de dimension paire sont différents de 0, car $\Omega^k \sim 0$ pour $0 < k \leqq n$ (remarque que je dois à G. de Rham).

5. Topologie de l'espace $\Gamma_n = O^+_{2n}/O'_n$. Γ_1 est un point. Γ_2 est homéomorphe à S_2 . Γ_3 est homéomorphe à l'espace projectif complexe $P_3(C)$. Γ_4 est homéomorphe à la quadrique complexe $Q_6(C)$ à 6 dimensions complexes. Quel que soit $n > 1$, Γ_n admet une structure fibrée de base S_{2n-2} et de fibre Γ_{n-1} . On en déduit les premiers groupes d'homotopie de Γ_n . $\pi_2(\Gamma_n) \cong \pi_2(\Gamma_2)$, cyclique infini, pour $n \geqq 2$. Pour $i > 2$, on a $\pi_i(\Gamma_3) \cong \pi_i(S_7)$. Comme $Q_6(C)$ admet une structure fibrée [3] de base S_6 et de fibre $P_3(C)$, pour laquelle il existe une section[1], on a $\pi_i(\Gamma_4) \cong \pi_i(S_6) \times \pi_i(P_3(C))$.

Pour $n \geqq 4$, on a $\pi_i(\Gamma_n) = 0$, si $2 < i < 6$, et $\pi_6(\Gamma_n)$ est cyclique infini. Ces propriétés de Γ_n servent à démontrer les résultats du §6.

6. Conditions d'existence de structures presque complexes.[2] Etant donné $E(B,\ R^{2n},\ O^+_{2n}\ ,\ H)$, le premier obstacle à l'existence d'une structure fibrée hermitienne subordonnée c'est-à-dire d'une section de l'espace fibré associé de fibres isomorphes à Γ_n , est une classe de cohomologie W^3 de B à coefficients entiers. *Pour qu'il existe une section sur le squelette de dimension* 3 *de* B, qui est supposé être un complexe, *il faut et il suffit que* $W^3 = 0$. *La classe* W^3 *est identique à la classe caractéristique de Stiefel-Whitney*, premier obstacle à l'existence d'un champ associant à tout $x \in B$ une suite de $2n - 2$ vecteurs indépendants de la fibre R^{2n}_x. La variété de Stiefel $V_{2n,2n-2}$ est en effet un espace fibré de base Γ_n et de fibre $W_{n,n-1}$, variété des suites orthonormées de $n - 1$ vecteurs unitaires de l'espace hermitien C^n. Il en résulte un isomorphisme canonique de $\pi_2(\Gamma_n)$ sur $\pi_2(V_{2n,2n-2})$ et l'identification des deux premiers obstacles considérés.

Si $W^3 = 0$, *il y a un deuxième obstacle de dimension* 4 *pour* $n = 2$, *de dimension* 8 *pour* $n = 3$, *de dimension* 7 *pour* $n > 3$. Si $n \geqq 3$, $W^3 = 0$ entraîne donc $W^5 = 0$, où W^5 est la classe de Stiefel-Whitney de dimension 5. En tenant compte d'un isomorphisme canonique de $\pi_6(\Gamma_n)$ sur $\pi_6(V_{2n,2n-6})$, où $n > 3$, on voit que *le deuxième obstacle est identique à la classe* W^7 *de Stiefel-Whitney*, premier obstacle à l'existence d'un champ de $2n - 6$ vecteurs. Une condition nécessaire pour l'existence d'une structure fibrée hermitienne subordonnée est évidemment que toutes les classes W^{2k+1} de Stiefel-Whitney soient nulles.

[1] $Q_6(C)$ n'est pas homéomorphe à $S_6 \times P_3(C)$, contrairement à ce que j'ai affirmé dans [2].

[2] On tronvera des résultats concernant les structures presque quaternioniennes dans [2] et [9].

En particulier soit $E = T(V_{2n})$, muni de la structure fibrée tangente orientée de la variété orientée V_{2n}. La classe W^3 de V_{2n} est le premier obstacle à l'existence d'une structure presque complexe sur V_{2n}. *La condition* $W^3 = 0$ *est nécessaire et suffisante pour l'existence d'une structure presque complexe sur une variété orientée* V_6. Si de plus le groupe de cohomologie de dimension 2 de V_6 est nul, toutes les structures presque complexes sur V_6 et correspondant à une orientation donnée forment une seule classe. *A chaque orientation de la sphère* S_6 *correspondent ainsi des structures presque complexes, appartenant toutes à une même classe d'homotopie.* Mais il faudrait sans doute des méthodes nouvelles pour décider si S_6 admet aussi des structures analytiques complexes.

Chacun des parallélismes classiques dans l'espace projectif réel P_7 correspond [3] sur $Q_6(C)$ à une structure fibrée dont les fibres sont des génératrices isomorphes à $P_3(C)$. En supposant que S_6 soit la partie réelle de $Q_6(C)$, cette structure fibrée définit justement une structure presque hermitienne sur S_6. Celle-ci pourra aussi être définie à l'aide des octaves de Cayley. Elle ne dérive pas d'une structure analytique complexe.

Par une méthode s'appliquant aux sphères S_{2n}, j'ai montré que S_4 *n'admet aucune structure presque complexe*, résultat obtenu d'une manière différente par H. Hopf. *Si* S_{2n} *est presque complexe,* S_{2n+1} *est parallélisable (Kirchhoff,* [5]*), ce qui entraîne le fait que* S_{4n} *n'admet aucune structure presque complexe.*

D'après Whitney [7], la classe W^3 d'une variété orientée V_4 est nulle. Mais Wen-tsün Wu a montré que pour tout $n > 2$, il existe des variétés orientées V_{2n} dont la classe W^3 n'est pas nulle.

7. Quelques résultats de Wen-tsün Wu. Par l'étude approfondie des variétés de Grassmann réelles et complexes, Wen-tsün Wu a obtenu les relations suivantes entre les classes caractéristiques d'une structure fibrée vectorielle réelle orientée $\mathfrak{F}$ et celles d'une structure fibrée vectorielle complexe subordonnée $\mathfrak{F}'$: *Relations de Wu:* $W_2(\mathfrak{F}, t) = C_2(\mathfrak{F}', t)$; $\overline{W}_2(\mathfrak{F}, t) = \bar{C}_2(\mathfrak{F}', t)$; $P(\mathfrak{F}, t) = C(\mathfrak{F}', t) \cup C(\mathfrak{F}', it)$; $\overline{P}(\mathfrak{F}, t) = \bar{C}(\mathfrak{F}', t) \cup \bar{C}(\mathfrak{F}', it)$; $X^{2n}(\mathfrak{F}) = (-1)^n C^{2n}(\mathfrak{F}')$.

Dans ces formules on a posé: $W_2(\mathfrak{F}, t) = \sum W_2^k(\mathfrak{F}) t^k$, $\overline{W}_2(\mathfrak{F}, t) = \sum \overline{W}_2^k(\mathfrak{F}) t^k$, où $W_2^k(\mathfrak{F})$ (resp. $\overline{W}_2^k(\mathfrak{F})$) sont les classes (resp. classes duales) de Stiefel-Whitney réduites modulo 2. $P(\mathfrak{F}, t) = \sum (-1)^k P^{4k}(\mathfrak{F}) t^{4k}$, $\overline{P}(\mathfrak{F}, t) = \sum \overline{P}^{4k}(\mathfrak{F}) t^{4k}$, où $P^{4k}(\mathfrak{F})$ (resp. $\overline{P}^{4k}(\mathfrak{F})$) sont les classes (resp. classes duales) de Pontrjagin. $C(\mathfrak{F}', t) = \sum C^{2k}(\mathfrak{F}') t^{2k}$, $\bar{C}(\mathfrak{F}', t) = \sum \bar{C}^{2k}(\mathfrak{F}') t^{2k}$, où $C^{2k}(\mathfrak{F}')$ (resp. $\bar{C}^{2k}(\mathfrak{F}')$) sont les classes (resp. classes duales) de Chern de $\mathfrak{F}'$; ces deux polynomes réduits modulo 2 sont désignés par $C_2(\mathfrak{F}', t)$ et $\bar{C}_2(\mathfrak{F}', t)$. $X^{2n}(\mathfrak{F})$ est la classe caractéristique d'Euler-Poincaré. On trouvera la définition précise de ces classes dans la thèse de Wu.

Théorème de Wu. *Les classes d'isomorphie des structures presque complexes sur une variété orientée* V_4 *correspondent d'une façon biunivoque aux classes de cohomologie* C^2 *sur* V_4 *telles que*:

$$W_2^2(V_4) = C_2^2 \; ; \; P^4(V_4) + 2X^4(V_4) = C^2 \cup C^2,$$

où C_2^2 désigne la classe déduite de C^2 par réduction modulo 2, $P^4(V_4)$ la classe de Pontrjagin, et $X^4(V_4)$ la classe d'Euler-Poincaré. C^2 sera la classe de Chern de la structure presque complexe correspondante.

Les relations de Wu entraînent que S_{4k} n'admet pas de structure presque complexe. *Ce résultat est valable pour toute variété V_{4k} dont l'anneau de cohomologie est isomorphe à celui de S_{4k} et dont la classe de Pontrjagin P^{4k} est nulle.*

8. Les sous-variétés d'une variété presque complexe V_{2n}. Un p-élément X_p de V_{2n} sera dit *complexe* lorsqu'il est invariant par la transformation I definie dans l'espace tangent T_x qui contient X_p ; il sera dit *réel* lorsqu'il rencontre son ransformé par I au point x seulement. Une sous-variété V_p de V_{2n} sera dite *presque complexe* (resp. *réelle*) lorsque les p-éléments tangents à V_p sont complexes (resp. réels). Une sous-variété presque complexe est munie d'une structure presque complexe induite; celle-ci dérive d'une structure analytique complexe si V_{2n} est analytique complexe.

Soit V_n une sous-variété réelle de V_{2n}. L'espace fibré $T(V_n)$ admet un isomorphisme sur l'espace fibré $N(V_n)$ des vecteurs normaux à V_n, les points de V_n restant fixés. Réciproquement si cette condition est vérifiée pour une sous-variété V_n d'une variété quelconque V_{2n}, il existe dans un voisinage de V_n une structure presque complexe telle que V_n soit une sous-variété réelle. En particulier, *le voisinage de la diagonale Δ de $V_n \times V_n$ admet une structure presque complexe telle que Δ soit une sous-variété réelle. Il admet même une structure analytique complex telle que Δ soit une sous-variété analytique réelle.*

La position d'une sous-variété réelle V_n dépend de la structure de $T(V_n)$. L'espace fibré $T^c(V_n)$ admet un isomorphisme canonique dans $T(V_{2n})$ muni de la structure fibrée complexe. *Si V_n est déformable en un point de V_{2n}, $T^c(V_n)$ est isomorphe à $V_n \times C^n$.* Pour toute variété V_n, Wu a démontré la relation suivante: $C(V_n, t) = P(V_n, t)$, où $C(V_n, t)$ désigne le "polynome de Chern" de $T^c(V_n)$ et $P(V_n, t)$ le "polynome de Pontrjagin" de $T(V_n)$ définis au §7. Si V_n est une sous-variété réelle de V_{2n}, $C(V_n, t)$ est la trace sur V_n de $C(V_{2n}, t)$, polynome de Chern de $T(V_{2n})$ muni de la structure fibrée complexe.

Pour $n = 2$, la relation de Wu donne $C(V_2, t) = 1$, ce qui montre que *pour toute variété V_2 l'espace $T^c(V_2)$ est isomorphe à $V_2 \times C^2$.*

Soit V_n une sous-variété réelle de l'espace projectif complexe $P_n(C)$. D'après la relation de Wu, la trace sur V_n de $C^{4k+2}(P_n(C))$ est nulle; *il en résulte que les cycles de dimension $4k + 2$ de V_n sont ~ 0 dans $P_n(C)$. Par contre pour tout $k \leqq n$ les cycles de dimension $2k$ d'une sous-variété complexe ne sont pas tous ~ 0 dans $P_n(C)$; ce résultat est valable aussi pour toute sous-variété presque complexe d'une variété presque kählerienne.*

Les notions et les résultats précédents s'étendent aux variétés plongées (V_p, f) dans V_{2n}, où f est une application différentiable régulière de V_p dans V_{2n}. Si V_{2n} est muni d'une forme différentielle extérieure Ω telle que $\Omega^n \neq 0$ en chaque

point, *les variétés intégrales de Ω sont des variétés plongées réelles pour une certaine structure presque complexe de V_{2n}.*

9. Problème d'équivalence de deux structures presque complexes. Etant données deux variétés presque complexes V_{2n} et $\bar{V}_{2n}$, une *équivalence* de l'une à l'autre est un homéomorphisme différentiable de V_{2n} sur $\bar{V}_{2n}$ dont le prolongement à $T(V_{2n})$ est un isomorphisme de $T(V_{2n})$ sur $T(\bar{V}_{2n})$ par rapport aux structures fibrées vectorielles complexes. Si f est une application différentiable régulière d'une variété W_{2n} dans V_{2n}, à la structure presque complexe sur V_{2n} correspond une structure presque complexe sur W_{2n}, appelée image réciproque par f de la première. Si celle-ci est définie localement par n formes de Pfaff complexes $\omega_1, \cdots, \omega_n$, son image réciproque est définie par les formes $f^*(\omega_h)$. Le problème d'équivalence locale de deux structures presque complexes peut être traité par les méthodes de E. Cartan et il vient d'être étudié pour $n = 2$ par Paulette Libermann [6].

Pour qu'une structure presque complexe sur V_{2n} dérive d'une structure analytique complexe, il faut et il suffit qu'elle soit partout localement équivalente[3] à la structure complexe naturelle sur C^n, qui est définie par les formes $dz_1, dz_2, \cdots, dz_n$. Soit g une équivalence d'un ensemble ouvert de V_{2n} à un ensemble ouvert de C^n. Les formes $dg_h = g^*(dz_h)$ sont alors des combinaisons linéaires indépendantes des formes ω_h, d'où l'on déduit les relations: $d\omega_h = \sum \omega_h \wedge \omega_{hl}$, où ω_{hl} sont des formes de Pfaff complexes sur V_{2n}. Ces équations, indiquées par G. de Rham, sont des conditions nécessaires pour que la structure presque complexe dérive d'une structure complexe. Dans le cas où les composantes réelles et imaginaires des formes ω_h sont des formes de Pfaff analytiques sur V_{2n}, ces conditions sont aussi suffisantes.[4] En général, on aura:

$$d\omega_h = \sum \omega_h \wedge \omega_{hl} + \sum a_{hlm}\, \bar{\omega}_l \wedge \bar{\omega}_m .$$

On peut dire que les formes $\sum a_{hlm}\, \bar{\omega}_l \wedge \bar{\omega}_m$ définissent la torsion de la structure presque complexe; les a_{hlm} sont les composantes de son *tenseur hermitien de torsion.*

Remarquons que les formules intégrales de Chern [1] donnant les classes de Chern s'étendent au cas des structures presque complexes.

Bibliographie

1. S. S. Chern, *Characteristic classes of hermitian manifolds*, Ann. of Math. t. 47 (1946).

2. C. Ehresmann, *Sur la théorie des espaces fibrés*, Colloque de Topologie Algébrique, C.N.R.S., Paris, 1947.

3. ———, *Sur les congruences paratactiques et les parallélismes dans les espaces projectifs*, C. R. Acad. Sci. Paris, janvier, 1939.

[3] Ces équivalences locales définissent alors les systèmes de coordonnées locales analytiques complexes.

[4] On peut se ramener au théorème de Frobenius appliqué à des formes analytiques complexes.

4. H. Hopf, *Zur Topologie der komplexen Mannigfaltigkeiten*, Courant Anniversary vol., 1948.

5. A. Kirchhoff, *Sur l'existence de certains champs tensoriels sur les sphères*, C. R. Acad. Sci. Paris t. 225 (1947).

6. Paulette Libermann, *Problèmes d'équivalence relatifs à une structure presque complexe*, Académie Royale de Belgique. Bulletin de la Classe des Sciences, 1950.

7. H. Whitney, *On the topology of differentiable manifolds*, Lectures in Topology, Ann. Arbor, 1941.

8. Wen-tsün Wu, *Sur la structure presque complexe d'une variété différentiable réelle de dimensions* 4, C. R. Acad. Sci. Paris t. 227 (1948).

9. ———, *Sur les classes caractéristiques des structures fibrées sphériques*, Thèse, Strasbourg, 1949.

University of Strasbourg,
Strasbourg, France.

COMPLEX-ANALYTIC MANIFOLDS[1]

BENO ECKMANN

1. Generalities. 1.1. The concept of a *complex-analytic manifold* (in short *complex manifold*) is the natural generalization of the concept of a Riemann surface in the abstract sense. A $2m$-dimensional complex manifold M is a manifold of dimension $2m$ in the usual sense which is covered by a family of systems of local complex coordinates $z_1, \cdots, z_m$ (instead of the usual $2m$ real coordinates) in such a way that the relation between two such complex coordinate systems in the intersection of their existence domains is given by complex-analytic functions. Since the coordinate transformation is one-to-one, the Jacobian of these functions is not equal to 0, and the square of its absolute value is easily seen to be equal to the Jacobian of the corresponding real coordinate transformation. A complex manifold M is therefore orientable and has a natural orientation. We shall restrict ourselves throughout to closed manifolds. The set of all local complex coordinate systems which are admissible, i.e., which may be added to the given family in accordance with the analyticity condition, is called the *complex structure* of M. Concepts like analytic function, analytic map, etc. have an invariant meaning with respect to the given complex structure (i.e., are independent of special coordinate systems used to describe them).

1.2. Algebraic varieties in a complex projective space P_n (of n complex, i.e., $2n$ real, dimensions) have a natural complex structure, as well as P_n itself, and are therefore complex manifolds—provided that there are no "singularities". There exist, on the other hand, examples of complex manifolds which cannot be imbedded in a P_n (cf. 3.4). The concept of a complex manifold is therefore more general than that of an algebraic variety. It is probably also more general than that of a higher-dimensional Riemann surface in a concrete sense, defined as the space of all function elements of an algebroid function on some standard space S; but such a statement depends of course on the space S and on uniformization possibilities.

1.3. It is well-known that all orientable surfaces admit complex structures. For higher-dimensional manifolds (orientable, of even dimension) this is not the case. On the 4-dimensional sphere S^4, for example, it is not possible to define a complex structure (cf. Hopf [1][2], Ehresmann [2]). The general *existence problem* of a complex structure on a given manifold M, or the problem of what special properties of M are implied by such a structure, has several quite different aspects. Actually there is not much known about the complex structure itself; all consequences are deduced from assumptions which are weaker—the *almost complex* structure, or stronger—the existence of a *Kaehler metric* (or Hermitean metric without torsion). The reason for these two approaches is simple: Almost complex structure is an

[1] This address was listed on the printed program under the title *Topologie der komplexen Mannigfaltigkeiten.*

[2] Numbers in brackets refer to the bibliography at the end of the address.

assumption concerning the tangent bundle of M, and therefore suitable for fiber space methods; and Kaehler metric is an assumption on the Riemannian or Hermitean geometry of M, which can be investigated by the methods of harmonic differential forms and of differential geometry. In both cases powerful existing theories can be applied.

Concerning the first approach I shall make only a few remarks, which overlap to some extent with Ehresmann's address; after that I shall discuss in more detail the second approach.

2. Almost complex manifolds. 2.1. A manifold M is called *almost complex* if for each point x of M a linear map I_x of the tangent space T_x at x into itself is defined such that $I_x^2 = -1$ (1 = identity map) for all x, and such that I_x depends continuously upon x. This is possible only if M is of even dimension $2m$. A complex manifold is almost complex; for if we use in the tangent space T_x complex vector components corresponding to local complex coordinates, multiplication of the vector components by $i = (-1)^{1/2}$ is a linear map of the required type, independent of the special coordinate system. I do not know of an example of an almost complex manifold which does not admit a complex structure, but probably there are such.

2.2. Let M be almost complex and $v(x)$ a continuous nonzero vector field on M with one possible singularity; then $I_x v(x) = w(x)$ is a second field with one possible singularity at the same point and such that $w(x)$ and $v(x)$ are independent everywhere else since $I_x^2 = -1$. Therefore there exists on M a 2-field with one isolated singularity, which is obviously of a quite special nature. This fact was used by Hopf [1] to prove that certain manifolds (the spheres S^4 and S^8 and many other examples) do not admit an almost complex structure, i.e., a field of transformations I_x; a fortiori these are not complex manifolds. Another way (cf. Ehresmann [2]) to prove existence or nonexistence of an almost complex structure on a manifold M is to apply the obstruction theory of fiber bundles; the base space of the fibering to be used is M, and the fiber at a point x is the space of all tensors at x describing transformations I_x (or related to them). For the simplest manifolds, the *spheres*, both methods give only quite special results, and in this case the existence problem seems to be related in a peculiar way to other topological questions, as I would like to indicate by some remarks.

2.3. If on a *sphere* S^n ($n = 2m$) a field of transformations I_x is given, we may assume that for all tangent vectors v at all points x, $I_x v$ is orthogonal to v. We have then for all pairs of *orthogonal* unit vectors x, y in $(n+1)$-dimensional Euclidean space E^{n+1} a unit vector function $z = F(x, y)$, z orthogonal to both x and y, continuous in x and linear in y; this function is obtained by considering x as a point of S^n and y as a tangent vector at x, and defining $z = I_x y$. It can be extended in an obvious way to a function $\bar{F}(x, y) = z$ of two *arbitrary* vectors in E^{n+1} such that z is orthogonal to x and y and of length $(x^2y^2 - (x \cdot y)^2)^{1/2}$ ($x \cdot y$ is the scalar product of x and y, $x^2 = x \cdot x$); $\bar{F}$ may be called a "*vector product in*

E^{n+1}," (continuous in the first factor, linear in the second). Applying this to the usual complex structure on S^2, we obtain a vector product in E^3 which is the usual one, bilinear in x, y. By fiber space methods using special homotopy groups, it can be proved that *for $n = 4k$ there exists no such vector product in E^{n+1}* of the required type, not even in the more general sense of continuity in both factors. The proof is a simple extension of the method I used previously [3] for the same purpose, but only for $n = 4$ and 8; to extend it to all $n = 4k$, one has to compute more of the homotopy groups involved, which can be done by straightforward deformations. From the argument above it follows that *there is no almost complex structure on the spheres S^{4k}* (not even if one would admit fields of transformations I_x which are only continuous, not linear, in the tangent spaces T_x, but such that $v(x)$ and $I_x v(x)$ are always independent—this is, by the way, the concept of a "manifold of type I" considered by Hopf [1]).

2.4. It is well-known [3] that a vector product in E^{n+1} leads in a natural way to a *multiplication* in E^{n+2}, i.e., a rule associating with any two vectors X, $Y \in E^{n+2}$ a product $U = X \circ Y$, with a unit and with the "norm product rule" $U^2 = X^2 \cdot Y^2$. Namely, let $X \in E^{n+2}$ be given by a pair (x_0, x) of a real number x_0 and a vector $x \in E^{n+1}$; then $X \circ Y = U = (u_0, u)$ is defined by $u_0 = x_0 y_0 - x \cdot y$, $u = x_0 y + y_0 x + \bar{F}(x, y)$. The rule $U^2 = X^2 \cdot Y^2$ is easily checked, and $E = (1, 0)$ is the unit. For the usual vector product in E^3 (i.e., the complex structure on S^2), this gives exactly the *quaternion* multiplication in E^4. The Cayley numbers in E^8 are related in the same way to a (bilinear) vector product in E^7, hence to an almost complex structure on S^6. It is not known whether it corresponds to a complex structure on S^6, nor indeed whether there is at all a complex to complex structure on S^6. [*Added in proof*: It has been proved (cf. Eckmann and Frölicher, C. R. Acad. Sci. Paris vol. 232 (1951) p. 2284) that the almost complex structure on S^6, derived from the Cayley numbers does not belong to a complex structure on S^6 by studying the "integrability conditions" relating almost complex structures.]

Furthermore, if $U = X \circ Y$ is a multiplication in E^{n+2} as described before, continuous in X and linear in Y, we have for each X with $X^2 = 1$ an orthogonal transformation $U(Y) = X \circ Y$ of E^{n+2} such that for $Y = (1, 0)$, $U(Y) = X$; i.e., we have a map Φ of S^{n+1} into the orthogonal group Ω_{n+2} of $n + 2$ variables, such that under the natural projection of Ω_{n+2} onto S^{n+1} the map Φ becomes the identity of S^{n+1}. It is easy to see that this map Φ is nothing else than a global parallelism on S^{n+1} (a system of $n + 1$ pairwise orthogonal tangent unit vector fields). Combining this with 2.3, we obtain a theorem discovered and proved first by Kirchhoff [4] in a different way: *An almost complex structure on S^n induces a global parallelism on S^{n+1}*. From this it follows again that on S^{4k} there is no almost complex structure (same result as above, here only in the sense of linear I_x).

2.5. In all preceding remarks the existence problem of an almost complex or complex structure is considered in terms of a definite differentiable structure on the manifold, and it is not clear from the method whether for another differentiable structure the answer would be the same.

It might be worthwhile to note here that the complex structure on a (differentiable) manifold, if it exists, is in general not unique. There are manifolds which admit several different complex structures—different in the sense of analytic equivalence. The 2-dimensional torus, for example, has infinitely many different conformal structures (and the analogue holds for higher-dimensional tori); for all of them, however, the corresponding complex structure of the universal covering is the same. Hirzebruch [5] discovered the interesting result that there exist also *simply connected* manifolds with infinitely many different complex structures; the simplest example is the Cartesian product of S^2 by itself.

3. Complex structure and homology. 3.1. To obtain relations between homology properties and complex structure, it seems natural to study the influence of the transformations I_x belonging to that structure on *exterior differential forms*. A differential form α of degree r in the manifold M (in short, an r-form) may be considered, at each point x of M, as a skew-symmetric multilinear function of r vectors in the tangent space T_x; if these vectors are transformed by I_x, α becomes a new form denoted by $C\alpha$ (if not only real, but also complex forms, i.e., complex functions in T_x, are admitted, it is convenient to include in the definition of C the passage to the conjugate-complex form; an explicit expression of C is given in 3.2). Obviously the operator C verifies $CC\alpha = (-1)^r\alpha$ for all r-forms α, hence C is an isomorphi m of the linear group Φ^r of all r-forms in M onto itself. In general $dC\alpha \neq C\,d\alpha$, and $\tilde{d} = C^{-1}\,dC$ is a differential operator from Φ^r to Φ^{r+1} which is not equal to d, but has some of the same properties (e.g., $\tilde{d}\tilde{d} = 0$). $\tilde{d}$ leads to de Rham cohomology groups $\tilde{H}^r$ isomorphic to the usual ones H^r based on d. By Hodge's theorem ([6], [7]) these groups may be replaced by those of *harmonic* r-forms (with respect to a Riemann metric given on M), which are easier to handle than cohomology classes; i.e., by forms α with $\Delta\alpha = 0$ or $\tilde{\Delta}\alpha = 0$, where Δ is the generalized Laplace-Beltrami operator used by de Rham [6], and $\tilde{\Delta} = C^{-1}\Delta C$. It seems interesting to study the relation between Δ and $\tilde{\Delta}$; here we shall do this only in the case when we assume a complex structure in the strict sense, or even more than that. Before discussing this, we have to make some preliminary remarks.

3.2. On a $2m$-dimensional complex manifold M we can use, in the calculus of differential forms, instead of the $2m$ real differentials dx_ν (or vector components in T_x) the *complex differentials* dz_j, $j = 1, \cdots, m$, corresponding to admissible coordinate systems $z_1, \cdots, z_m$ *and their conjugates* $d\bar{z}_j$. The coordinate transformations, say from $z_1, \cdots, z_m$ to $\zeta_1, \cdots, \zeta_m$ being analytic, the dz_j are carried over to the $d\zeta_j$ and the $d\bar{z}_j$ to the $d\bar{\zeta}_j$ only, this splitting of the $2m$ differentials into two groups has an invariant meaning. A differential form which is homogeneous with respect to the $d\bar{z}_j$ will be called *pure*; and the corresponding degree in the $d\bar{z}_j$ only, its *type*. Any r-form α has a unique decomposition as a sum of pure forms $\alpha = \alpha_{(0)} + \cdots + \alpha_{(r)}$ of type $h = 0, 1, \cdots, r$. For a form $\alpha_{(h)}$ of type h, $C\alpha_{(h)}$ is given by $(-1)^h i^r \bar{\alpha}_{(h)}$, which is of type $r - h$; for pure forms, $\tilde{d}$ differs therefore from d only by a constant factor.

3.3. The Riemann metric, used in the definition of Δ, is called *Hermitean* if in

the given complex structure the line element ds^2 may be written as $\sum_{j,k} h_{jk} \cdot (dz_j\, d\bar{z}_k)$, $h_{jk} = \bar{h}_{kj}$, where $(\cdots)$ indicates the *ordinary* product of differentials. Such a metric is however not sufficient for our purposes. For it appears that, roughly speaking, one must, in the relations between harmonic forms and complex structure, consider permutations of the second covariant derivatives involved in Δ and therefore use the Riemann curvature tensor of the metric; and moreover, one should compute this tensor by the complex formalism, i.e., in the so-called Hermitean geometry. This is in general not possible: it is easy to see that the connexions (the Christoffel symbols) computed from the same Hermitean metric in Riemannian and in Hermitean geometry are *different*. There is a special case where they coincide, and this case seems of great importance: namely when the metric is a "*Kaehler metric*" [8]; i.e., has the special property that the exterior differential form $\omega = \sum_{j,k} h_{jk}\, dz_j\, d\bar{z}_k$ is *closed* ($d\omega \doteq 0$). Actually many of the homology properties we are going to describe can be deduced from assumptions which are stated in terms of real differential geometry, without using a complex structure; these are however less natural and intuitive, and for the purpose of this address we shall stay within the frame of the Kaehler metric.

It is well-known that on the complex projective space P_n there exists a Kaehler metric, and therefore on all algebraic varieties imbedded without singularities in P_n. More generally, if a complex manifold M is imbedded analytically in a Kaehler manifold, the induced metric on M fulfills also the Kaehler condition (cf. 3.6).

3.4. In the case of a Kaehler metric, the explicit expression of Δ yields the following results (cf.[10]) on pure forms and on C. (a) If a form α is harmonic, then all the "pure components" $\alpha_{(h)}$ of α (cf. 3.2) are also harmonic. In other words: H^r is the direct sum $H^r_{(0)} + H^r_{(1)} + \cdots + H^r_{(r)}$, where $H^r_{(h)}$ is the group of pure harmonic forms of type h. If $p^r_{(h)}$ denotes the rank of $H^r_{(h)}$, we have for the Betti number $p^r = \sum_{h=0}^{r} p^r_{(h)}$. Of course $p^r_{(h)}$ is also defined for a general Hermitean metric; but all one can prove without the Kaehler condition is the inequality $p^r \geqq \sum p^r_{(h)}$. (b) $\tilde{\Delta} = \Delta$. Let $\tilde{H}^r_{(h)}$ be the analogue of $H^r_{(h)}$, computed from $\tilde{\Delta}$ instead of Δ, $\tilde{p}^r_{(h)}$ its rank; it follows that $\tilde{H}^r_{(h)} = H^r_{(h)}$ and $\tilde{p}^r_{(h)} = p^r_{(h)}$. But since by definition $C\tilde{\Delta} = \Delta C$, C maps $\tilde{H}^r_{(h)}$ isomorphically onto $H^r_{(r-h)}$; hence $\tilde{p}^r_{(h)} = p^r_{(r-h)} = p^r_{(h)}$. Therefore, *if r is odd*, $p^r = 2\sum_{h=0}^{(r-1)/2} p^r_{(h)}$, i.e., $p^r \equiv 0 \pmod 2$; *if r is even*, $p^r \equiv p^r_{(r/2)} \pmod 2$. For algebraic varieties the fact that p^r is even in odd dimensions r is a well-known result already proved by Lefschetz and, using differential forms, by Hodge.

There exist complex manifolds which do not admit a Kaehler metric since their Betti numbers do not agree with the above conditions, namely the manifolds of topological structure $S^1 \times S^{2k+1}$ discovered by Hopf [1] (a complex structure on $S^1 \times S^3$, e.g., is obtained from the covering of $S^1 \times S^3$ by E^4 without the origin). From this it follows that these complex manifolds cannot be imbedded analytically in a complex projective space (nor in any Kaehler manifold).[3]

[3] *Added in proof*: E. Calabi and the author showed that the product of any two odd-

3.5. All this is only one, somewhat special, aspect of the Kaehler metric, and there are other surprisingly strong consequences.

There is one immediate fact, which does not even use harmonic forms, but just some simple properties of the closed 2-form ω (cf. 3.3), depending only weakly upon the relation of ω to the metric. Let $n = 2m$ be the dimension of the Kaehler manifold M. The mth power ω^m is equal, up to a constant factor not equal to 0, to $|\, h_{jk} \,| \, dz_1 d\bar{z}_1 \cdots dz_m d\bar{z}_m$; since the determinant $|\, h_{jk} \,|$ is everywhere positive, $\int_M \omega^m \neq 0$ (ω^m is actually, except for a constant nonzero factor, the volume element of the metric considered as a Riemann metric). Hence ω^m is not cohomologous to 0, and the same is true for all powers ω^k, $k = 0, 1, \cdots, m$; the corresponding Betti numbers p^{2k} are therefore greater than or equal to 1.

These nonvanishing cohomology classes in even dimensions may be carried over by the duality operator D of M to homology classes. Using the same symbol α for a closed form and for the corresponding cohomology class, the dual homology class $D\alpha$ is given by $\alpha \cap M$ (here M is the fundamental $2m$-cycle of the oriented manifold, and $\cap$ the Čech-Whitney cap-product). Let us call the classes $D\omega^k = \omega^k \cap M = Z_{2(m-k)}$ of dimension $2(m - k)$ the *principal homology classes* of M. Clearly, writing Z for $Z_{2(m-1)} = D\omega$, $Z_{2(m-2)}$ is the *intersection* $Z \otimes Z$, $Z_{2(m-3)} = Z \otimes Z \otimes Z$, etc. If M is the complex projective space P_m, the principal classes Z_{2q}, $q = 0, 1, \cdots, m$, are all represented, up to certain numerical factors, by the projective planes P_q in P_m.

3.6. Let L' be a complex manifold of dimension $2l < 2m$, and f an *analytic* and locally topological *imbedding of L' in the Kaehler manifold M*, $f(L') = L \subset M$. Considering the induced dual homomorphism f^*, the image $f^*\omega$ is the form ω' in L' corresponding to the induced metric in L', which obviously is also a Kaehler metric. Thus, by the same argument as above, $0 \neq \int_{L'} \omega'^l = \int_{L'} f^*\omega^l = \int_L \omega^l$; by Stokes theorem it follows *that $L \not\sim 0$ (not homologuous to 0) in M.*

Furthermore consider the principal homology classes $Z'_{2(l-k)}$ of L' ($k = 0, \cdots, l$), and their images $f(Z'_{2(l-k)})$ in M; they are equal to $f(\omega'^k \cap L') = f(f^*\omega^k \cap L') = \omega^k \cap f(L') = \omega^k \cap L = Z_{2(m-k)} \otimes L$, i.e., they are the intersections of L with the principal classes of M (for example, the principal classes of a manifold L imbedded in a complex projective space are given by the *plane sections of L*). *These classes are all $\not\sim 0$ in M.* For $\int_{\omega^k \cap L} \omega^{l-k} = \int_L \omega^l \neq 0$, hence by Stokes theorem $\omega^k \cap L \not\sim 0$.

Analogous remarks apply to "analytic cycles" in M, i.e., to the set of zeros of a distribution of local analytic functions. The problem of finding not only necessary, but also sufficient conditions for a homology class to contain analytic cycles seems to be difficult.

3.7. Very precise results on the role of ω and of principal classes in the homology structure of a Kaehler manifold M, including all of the known general homology structure of algebraic varieties, are obtained if, following the method of Hodge, harmonic forms in M are investigated in connection with special properties of

dimensional spheres $S^{2p+1} \times S^{2q+1}$ can be given a complex structure; for $p > 0$, $q > 0$ this is an example of a *simply connected* complex manifold which does not admit a Kaehler metric.

ω and of the differential geometry of the Kaehler metric. All these results are due to some simple formulas relating ω to the usual operations on differential forms: d, $*$ ($*\alpha$ is the adjoint of the r-form α, with respect to the metric, of degree $n - r = 2m - r$; cf. [6]), $\delta = \pm * d *$, $\Delta = d\delta + \delta d$, and the operator C, with $\bar{d}$, $\bar{\delta}$, etc. These formulas, most of which appear in Hodge's book [7] and in a note by A. Weil [9], have been completely established and discussed by Guggenheimer and myself [10]. They are of purely local character.

The main formula is a relation between $*$ and ω, replacing the operator $*$ for r-forms with $r \leqq m$ by C and by multiplication with ω^{m-r} (and with a certain numerical factor u; I omit here all details about such factors): $*\alpha = u\omega^{m-r}C\alpha$. This is however not true for all forms α. It holds (1) if $\omega * \alpha = 0$—such forms are called *effective*; and more generally (2) if $\alpha = \omega^k\beta$, where β is effective—we shall call such a form *simple, of class* k. The nonzero factor u depends upon m, r, and k. Under condition (2) we have for $r \leqq m - 2$, $\omega * (\omega\alpha) = u'\omega \cdot \omega^{m-r-2} C\omega\alpha = u'' * \alpha$, with nonzero factors u', u''. From this it follows that (a) $\omega\alpha = 0$ only if $\alpha = 0$. By an easy induction argument it can be proved that (b) *every* r*-form* α ($r \leqq m$) *has a unique decomposition into a sum of simple forms of class* $k = 0, 1, \cdots, q = [r/2]$, $\alpha = \beta_0 + \omega\beta_1 + \cdots + \omega^q\beta_q$. (a) holds therefore for *all* forms of degree $r \leqq m - 2$: (c) *Multiplication by* ω *is an isomorphism of* Φ^r, *the group of all* r*-forms, into* Φ^{r+2}.

Further relations give $\Delta(\omega^k\alpha) = \omega^k(\Delta\alpha)$ for all forms α, and commutation of Δ with all operators involved. Hence all results apply in particular to harmonic forms, i.e., to cohomology groups H^r. From (b) above it follows that H^r is for $r \leqq m$ the direct sum $H_0^r + \omega H_0^{r-2} + \cdots + \omega^q H_0^{r-2q}$, where $H_0^{r-2k} \subset H^{r-2k}$ is the group of effective harmonic $(r - 2k)$-forms; from (c), that multiplication by ω is, for $r \leqq m - 2$, an *isomorphism of* H^r *into* H^{r+2} (cf. [7], Chap. IV). For the Betti numbers of a Kaehler manifold M we obtain therefore $p^r \leqq p^{r+2}$ for $r \leqq m - 2$ (and, of course, the dual statements for higher dimensions). These are strong topological conditions for the existence of a Kaehler metric; they are not fulfilled by all complex manifolds (e.g., by the examples mentioned in 3.4).

3.8. To translate the results from cohomology to homology, we take the operator D, which maps H^r isomorphically onto the rth homology group H_r of M (with complex coefficients). Since $*\omega^k = u\omega^{m-k}$, $D*\omega^k$ is, except for a constant nonzero factor, the principal homology class Z_{2k} of M, (e.g., a plane section if M lies analytically in a complex projective space). Each homology class z_r of M of dimension $r \leqq m$ may be written as $z_r = D*\alpha = D*(\beta_0 + \omega\beta_1 + \cdots + \omega^q\beta_q)$, $q = [r/2]$, according to (b) in 3.7, with uniquely determined effective harmonic forms $\beta_0, \cdots, \beta_q$. The term $D*\omega^k\beta_k$ is, except for the nonzero factor u, equal to $D(C\omega^k\beta_k\omega^{m-r}) = C\omega^k\beta_k\omega^{m-r} \cap M$; that is, to $C\omega^k\beta_k \cap (\omega^{m-r} \cap M) = C\omega^k\beta_k \cap D\omega^{m-r} = C\omega^k\beta_k \cap Z_{2r} = DC\omega^k\beta_k \otimes Z_{2r}$, or to $C\beta_k \cap (\omega^{m-(r-k)} \cap M) = DC\beta_k \otimes Z_{2(r-k)}$. Therefore, (a) *all homology classes of dimension* $r \leqq m$ *lie on the principal class* Z_{2r}, in the sense that they are intersections of something with Z_{2r}, and (b) *they have a unique decomposition into a sum* $z_r = z_r^{[0]} + z_r^{[1]} + \cdots + z_r^{[q]}$, *where* $z_r^{[k]} = D*\omega^k\beta_k$ *lies on* $Z_{2(r-k)}$ (but not on Z_{2l}, $l < r - k$). Using

only the assumption of a Kaehler metric we obtain thus a geometric situation of the same nature as the homology structure of algebraic varieties discovered by Lefschetz.

3.9. I would like to conclude with a remark on the differential forms themselves. If we have a simple form $\alpha = \omega^k\beta$ (cf. 3.7.), and if α is closed, it follows easily from the general relations that $\tilde{\delta}\alpha = 0$; if in addition α is pure, this means $\delta\alpha = 0$, hence α is harmonic. A closed form α which is pure and simple is therefore always harmonic. If in particular $\alpha = d\gamma$, it must be equal to 0 (here we use a global result, based on the fact that M is closed), and we have the theorem: *If on a Kaehler manifold a form α is such that α and $d\alpha$ are both pure and simple, then $d\alpha = 0$, and α is harmonic.*

Let us take for example an *analytic* form α, i.e., a form in the dz_j only, with analytic coefficients. α is of type 0, and also $d\alpha$, since for all coefficients a of α we have $\partial a/\partial\bar{z}_j = 0$; it is easily seen that a form of type 0 is always effective. Therefore, by the above theorem, for *an analytic differential form α in a Kaehler manifold, $d\alpha$ is always equal to* 0, *and α harmonic.* Relations between the set of all analytic differential forms α in the manifold and its Betti numbers (cf. [7], [9], [10]) are therefore obtained without assuming in advance $d\alpha = 0$.

Bibliography

1. H. Hopf, *Studies and essays presented to R. Courant*, New York, 1948, pp. 167–185.
2. Ch. Ehresmann, *Topologie algébrique*, Paris, 1949, pp. 3–15.
3. B. Eckmann, Comment. Math. Helv. vol. 15 (1942–1943) pp. 318–339.
4. A. Kirchhoff, C. R. Acad. Sci. Paris vol. 225 (1947) pp. 1258–1260.
5. F. Hirzebruch, Doctoral thesis, University of Münster, 1950.
6. G. de Rham and P. Bidal, Comment. Math. Helv. vol. 19 (1946) pp. 1–49; G. de Rham, Annales de Grenoble vol. 22 (1946) pp. 135–152.
7. W. V. D. Hodge, *Harmonic integrals*, Cambridge, 1941.
8. E. Kaehler, Abh. Math. Sem. Hamburgishen Univ. vol. 9 (1933) pp. 173–186.
9. A. Weil, Comment. Math. Helv. vol. 20 (1947) pp. 110–116.
10. B. Eckmann and H. Guggenheimer, C. R. Acad. Sci. Paris vol. 229 (1949) pp. 464–466, 489–491, 503–505; B. Eckmann, C. R. Acad. Sci. Paris vol. 229 (1949) pp. 577–579.

Eidgenössische Technische Hochschule,
Zürich, Switzerland.

COHOMOLOGY ON REAL DIFFERENTIABLE MANIFOLDS

CARL B. ALLENDOERFER

1. Introduction. The theory of de Rham has established a relationship between the cohomology theory (real coefficients) of a real, closed differentiable manifold M_n and the theory of exterior differential forms on this manifold. The connection is expressed by the theorems:

FIRST THEOREM OF DE RHAM. *Let C^r be a fixed r-dimensional cohomology class (real coefficients) on M_n and let Z_r be an arbitrary homology class (real coefficients) on M_n. Then there exists a regular closed r-form Θ^r ($d\Theta^r = 0$) such that:*

$$C^r \cdot Z_r = \int_z \Theta^r$$

where z is any r-cycle representing Z_r.

SECOND THEOREM OF DE RHAM. *If for a closed r-form Θ^r*

$$\int_z \Theta^r = 0$$

for all r-cycles (real coefficients) z, then there exists a regular $r-1$ form Ψ^{r-1} on M_n such that:

$$d\Psi^{r-1} = \Theta^r;$$

i.e., Θ^r *is derived.*

This theory, however, is inadequate for a treatment of the relationship between differential forms and cohomology with coefficients which are integers or integers mod p (p a prime). It has been suggested by A. Weil [4] that such a theory can be constructed using singular r-forms. A specific example of such a theory was given by the author [1] in connection with the characteristic classes of a Riemannian manifold, and has served as motivation for the present paper which presents a general theory of this nature. In addition, the results are applied to harmonic forms and give a full cohomology theory for these forms which is an extension of the work of Hodge [2] and Kodaira [3]. The theory is also applied to derive an extension of de Rham's second theorem for cases in which the periods are not all zero (Theorem 6).

2. Elementary forms. Let M_n be a real, closed, differentiable manifold of class greater than or equal to 3. Consider a differentiable triangulation of M_n whose cells of dimension r are E_r^i ($i = 1, 2, 3, \cdots$), and whose r-skeleton is K_r. Also consider a dual, differentiable subdivision of M_n whose $n-r$ skeleton is ${}^*K^{n-r}$. Let C^r be a cochain on K_r whose values on E_r^i are α_i, where α_i are real

numbers. Associated with C^r is a dual chain $*C^{n-r}$ of $*K^{n-r}$ such that the intersection number $E_r^i \cdot *C^{n-r} = \alpha_i$. We prove the following theorem.

THEOREM 1. *There exist differential forms*

(1) θ^r *defined on* K_r *and regular of class greater than or equal to* 1 *on* K_r;

(2) φ^{r-1} *defined on* K_r *and regular of class greater than or equal to* 2 *on* K_{r-1} *with at most isolated point singularities on* $K_r - K_{r-1}$, *such that*

$$d\varphi^{r-1} = \theta^r; \tag{3}$$

$$\int_{E_r^i} \theta^r - \int_{\partial E_r^i} \varphi^{r-1} = \alpha_i. \tag{4}$$

PROOF. Choose a cell E_r^i. We seek to define forms θ_i^r and φ_i^{r-1} having properties (1), (2), and (3) of the theorem and such that

$$\int_{E_r^j} \theta_i^r - \int_{\partial E_i^j} \varphi_i^{r-1} = \delta_{ij}. \tag{5}$$

There are numerous ways of finding forms $*\theta_i^r$ and $*\varphi_i^{r-1}$ having properties (1), (2), and (3) and such that $*\varphi_i^{r-1}$ is regular on $[K_r \cap \text{Star } E_r^i - E_r^i]$ and

$$\int_{E_r^i} *\theta_i^r - \int_{\partial E_r^i} *\varphi_i^{r-1} = 1, \tag{6}$$

but which do not necessarily satisfy (5) for E_r^j $(j \neq i)$. Examples of such forms appear in the author's paper [1], but much simpler ones can be constructed.

Let U_n and V_n be neighborhoods of M_n such that $\text{Star } E_r^i \supset V_n \supset U_n \supset E_r^i$. Let f be a "local function", regular of class greater than or equal to 2 in M_n such that

$$f = \begin{cases} 1 \text{ in } U_n \\ 0 \text{ in } M_n - V_n. \end{cases}$$

Define $\varphi_i^{r-1} = f \cdot *\varphi_i^{r-1}$ on K_r;

$$\theta_i^r = \begin{cases} *\theta_i^r \text{ on } E_r^i \\ d\varphi_i^{r-1} \text{ on } K_r - E_r^i. \end{cases}$$

These forms have properties (1), (2), (3), and (5). Now define

$$\theta^r = \sum_i \alpha_i \theta_i^r; \qquad \varphi^{r-1} = \sum_i \alpha_i \varphi_i^{r-1}.$$

These forms have the required properties (1), (2), (3), and (4). Note that φ^{r-1} is regular on those cells for which $\alpha_i = 0$.

COROLLARY. *If* B_r *is any* r*-chain (real coefficients) of* K_r, *the forms of Theorem* 1 *are such that*

$$\int_{B_r} \theta^r - \int_{\partial B_r} \varphi^{r-1} = C^r \cdot B_r.$$

3. Extensions of elementary forms. In extending the forms θ^r and φ^{r-1} of Theorem 1 over the whole of M_n we shall proceed by induction from K_r to K_{r+1} to $\cdots$ to K_{n-1} to $K_n = M_n$. In extending over any cell E_{t+1} of K_{t+1} ($t = r, \cdots, n-1$) we shall use one of three methods listed below according to the circumstances involved. We call the extended forms: Θ^r and Φ^{r-1} respectively.

The methods described below make use of a special "polar" coordinate system in E_{t+1}. Let $x^1, \cdots, x^t$ be local coordinates on ∂E_{t+1}. Let L^{t-r} be a subcomplex of $\partial E_{t+1} \cap {}^*K^{n-r}$ which will be chosen to meet particular needs in the sequel. In E_{t+1} there is a vertex P of ${}^*K^{n-r}$ and a subcomplex of ${}^*K^{n-r}$, L^{t-r+1}, which contains P and is such that $L^{t-r+1} \cap \partial E_{t+1} = L^{t-r}$. Join P to ∂E_{t+1} with a simple differentiable family of curves such that the curves from P to L^{t-r} lie on L^{t-r+1}. Parametrize these curves with the parameter x^{t+1} so that $x^{t+1} = 0$ is at P and $x^{t+1} = 1$ is on ∂E_{t+1}. The set $(x^1, \cdots, x^{t+1})$ are our local "polar" coordinates for E_{t+1}.

METHOD 1. On ∂E_{t+1} assume Θ^r regular and closed, Φ^{r-1} regular, and $d\Phi^{r-1} = \Theta^r$. Further if $t = r$, assume $\int_{\partial E_{r+1}} \Theta^r = 0$.

Then by the second theorem of de Rham there exists a regular form ψ^{r-1} on ∂E_{t+1} such that $d\psi^{r-1} = \Theta^r$. Extend ψ^{r-1} differentiably of class greater than or equal to 2 over E_{t+1} so that in a neighborhood of ∂E_{t+1}

$$\begin{aligned} &\psi^{r-1}_{i_1 \cdots i_{r-1}}(x^1, \cdots, x^t, x^{t+1}) = \psi^{r-1}_{i_1 \cdots i_{r-1}}(x^1, \cdots, x^t, 1) \text{ for } j_1 \cdots j_{r-1} \neq t+1; \\ &\psi^{r-1}_{i_1 \cdots i_{r-2} t+1}(x^1, \cdots, x^{t+1}) = 0. \end{aligned} \tag{1}$$

This extension is trivial for there are no other restrictions on ψ^{r-1}. Now define $\Theta^r \mid E_{t+1} = d\psi^{r-1}$. Then Θ^r is regular and closed on E_{t+1} and reduces to the given form on ∂E_{t+1}.

To extend Φ^{r-1}, define

$$\omega^{r-1} = \psi^{r-1} - \Phi^{r-1} \quad \text{on} \quad \partial E_{t+1}. \tag{2}$$

Then ω^{r-1} is closed and regular on ∂E_{t+1} and can be extended regularly over E_{t+1} by the method just used for Θ^r. Finally define

$$\Phi^{r-1} = \psi^{r-1} - \omega^{r-1} \quad \text{on} \quad E_{t+1}. \tag{3}$$

It follows that Φ^{r-1} is regular on E_{t+1} and that $d\Phi^{r-1} = \Theta^r$.

METHOD 2. On ∂E_{t+1} assume Θ^r as in Method 1, Φ^{r-1} singular on L^{t-r}, and $d\Phi^{r-1} = \Theta^r$ on $\partial E_{t+1} - L^{t-r}$.

Extend Θ^r as in Method 1. Define ω^{r-1} on $\partial E_{t+1} - L^{t-r}$ by (2) above. In our "polar" coordinate system define ω^{r-1} on $E_{t+1} - L^{t-r+1}$ by equations (1) above with ω substituted for ψ. It is clear that we still have $d\omega = 0$. Finally define Φ^{r-1} by (3) above. Φ^{r-1} is now regular on $E_{t+1} - L^{t-r+1}$ and where it is regular: $d\Phi^{r-1} = \Theta^r$.

METHOD 3. On ∂E_{t+1} assume Θ^r regular on $\partial E_{t+1} - \Gamma^{t-r-1}$, Φ^{r-1} regular on $\partial E_{t+1} - L^{t-r}$, Γ^{t-r-1} a subcomplex of L^{t-r}, and where defined $d\Phi^{r-1} = \Theta^r$.

We use the "polar" coordinate system and call Γ^{t-r} the subcomplex of L^{t-r+1} consisting of curves joining P to points of Γ^{t-r-1}. Then we extend Θ^r and Φ^{r-1} by the equations:

$$\text{On } E_{t+1} - \Gamma^{t-r} \begin{cases} \Theta^r_{j_1 \cdots j_r}(x^1, \cdots, x^t, x^{t+1}) = \Theta^r_{j_1 \cdots j_r}(x^1, \cdots, x^t, 1) \\ \qquad\qquad\qquad\qquad\qquad\qquad \text{for } j_1 \cdots j_r \neq t+1; \\ \Theta^r_{j_1 \cdots j_{r-1} t+1}(x^1, \cdots, x^{t+1}) = 0. \end{cases}$$

$$\text{On } E_{t+1} - L^{t-r+1} \begin{cases} \Phi^{r-1}_{j_1 \cdots j_{r-1}}(x^1, \cdots, x^t, x^{t+1}) = \Phi(x^1, \cdots, x^t, 1) \\ \qquad\qquad\qquad\qquad\qquad\qquad \text{for } j_1 \cdots j_{r-1} \neq t+1; \\ \Phi^{r-1}_{j_1 \cdots j_{r-2} t+1}(x^1, \cdots, x^{t+1}) = 0. \end{cases}$$

Then Θ^r and Φ^{r-1} are regular where defined and $d\Phi^{r-1} = \Theta^r$.

4. Cochains. We now use these methods of extension to derive the following consequences of Theorem 1.

THEOREM 2. *Given an r-cochain C^r (real coefficients) of K_r and its associated dual chain $^*C^{n-r}$ of $^*K^{n-r}$. There exist differential forms Θ^r and Φ^{r-1} on M_n such that*

(1) *Θ^r is regular of class greater than or equal to 1 and closed on $M_n - \partial {}^*C^{n-r}$;*

(2) *Φ^{r-1} is regular of class greater than or equal to 2 on $M_n - {}^*C^{n-r}$;*

(3) *$d\Phi^{r-1} = \Theta^r$ on $M_n - {}^*C^{n-r}$;*

(4) *$\int_{B_r} \Theta^r - \int_{\partial B_r} \Phi^{r-1} = C^r \cdot B_r$*

where B_r is any chain (real coefficients) of K_r.

PROOF. On K_r construct the forms θ^r and φ^{r-1} of Theorem 1. θ^r is regular on K_r and φ^{r-1} has point singularities on those cells E^i_r for which $\alpha_i \neq 0$. We can place these singularities on $^*C^{n-r}$, for the intersection of $^*C^{n-r}$ and such a cell is exactly one point (with a coefficient α_i).

In extending from K_t to K_{t+1} we have three situations for a cell E_{t+1} :

(1) $\partial E_{t+1} \cap {}^*C^{n-r} = 0$. Thus Θ^r and Φ^{r-1} are regular on ∂E_{t+1} and when $t = r$, $\int_{\partial E_{r+1}} \Theta^r = 0$. Apply Method 1 above and obtain regular extensions of both forms over E_{t+1}.

(2) $\partial E_{t+1} \cap {}^*C^{n-r} \neq 0$, but $E_{t+1} \cap \partial {}^*C^{n-r} = 0$. Then Θ^r is regular on ∂E_{t+1} and when $t = r$, $\int_{\partial E_{r+1}} \Theta^r = 0$; but Φ^{r-1} has singularities on $\partial E_{t+1} \cap {}^*C^{n-r}$ which we now call L^{t-r}. Apply Method 2 above and obtain Θ^r regular over E_{t+1} and Φ^{r-1} regular over $E_{t+1} - L^{t-r+1}$. Note that the singular locus of φ^{r-1}, namely L^{t-r+1}, remains on $^*C^{n-r}$.

(3) $\partial E_{t+1} \cap {}^*C^{n-r} \neq 0$ and $E_{t+1} \cap \partial {}^*C^{n-r} \neq 0$. Then Θ^r is regular on ∂E_{t+1} except on $\partial E_{t+1} \cap \partial {}^*C^{n-r}$ which we now call Γ^{t-r-1}. When $t = r$, this intersection

is necessarily empty, but now $\int_{\partial E_{r+1}} \Theta^r \neq 0$. Also Φ^{r-1} is regular on ∂E_{t+1} except on L^{t-r} as in Case 2. Apply Method 3 above and obtain Θ^r regular over $E_{t+1} - \Gamma^{t-r}$ and Φ^{r-1} regular over $E_{t+1} - L^{t-r+1}$. When $t = r$, we define Γ^{t-r} to be the point P described above. Note that Γ^{t-r} lies on $\partial {}^*C^{n-r}$ and L^{t-r+1} on ${}^*C^{n-r}$.

By carrying out this process successively for t running from r to $n-1$, we establish the theorem.

COROLLARY. *Theorem 2 also applies to any chains B_r of M_n (not necessarily those of K_r) provided that $B_r \cap \partial {}^*C^{n-r} = 0$ and $\partial B_r \cap {}^*C^{n-r} = 0$.*

5. Cocycles. When C^r is a cocycle (real coefficients), $\partial {}^*C^{n-r} = 0$ and Θ^r is regular on M_n. This amounts to the following theorem.

THEOREM 3. *Given a cocycle C^r on M_n, there exist differential forms Θ^r and Φ^{r-1} on M_n such that*

(1) *Θ^r is closed and regular of class greater than or equal to 1 on M_n, and conditions* (2), (3), *and* (4) *of Theorem 2 and its corollary are satisfied.*

When B_r is an r-cycle (real coefficients), this theorem amounts to the first theorem of de Rham. Since an r-dimensional cohomology class (real coefficients) is determined by its periods over the r-cycles (real coefficients), there is nothing more to say. However, a cohomology class (integer coefficients) requires for its determination not only its periods on the integral cycles of M_n but also its values on those integral chains which are cycles mod p. Theorem 3 then gives as a supplement to the de Rham theorem for this case the following theorem.

THEOREM 4. *Let $\boldsymbol{C}^r_{(0)}$ be a fixed r-dimensional cohomology class (integer coefficients) on M_n and $\boldsymbol{Z}^{(0)}_r$ and $\boldsymbol{Z}^{(p)}_r$ arbitrary homology classes with integer and integer* mod p *coefficients respectively on M_n. Then there exists a regular closed form Θ^r and a form Φ^{r-1} with singularities on ${}^*C^{n-r}_{(0)}$ (the dual cycle associated with a particular cocycle $C^r_{(0)}$ representing $\boldsymbol{C}^r_{(0)}$) and with $d\Phi^{r-1} = \Theta^r$ on $M_n - {}^*C^{n-r}_{(0)}$ such that*

$$\boldsymbol{C}^r_{(0)} \cdot \boldsymbol{Z}^{(0)}_r = \int_z \Theta^r,$$

$$\boldsymbol{C}^r_{(0)} \cdot \boldsymbol{Z}^{(p)}_r \equiv \int_z \Theta^r - \int_{\partial z} \Phi^{r-1} \qquad \text{mod } p,$$

where z is a cycle which represents the corresponding homology class and on which the integrals are defined.

We can apply Theorem 2 to cochains $C^r_{(p)}$ with integers mod p as coefficients by selecting a particular residue mod p for the value of the cochain on each cell. Then we treat the cochain as an integral cochain. The conclusions of Theorem 2 remain valid except that (4) must be rewritten

$$(4') \qquad \int_{B_r^{(p)}} \Theta^r - \int_{\partial B_r^{(p)}} \Phi^{r-1} \equiv C^r_{(p)} \cdot B_r^{(p)} \qquad \text{mod } p,$$

where $B_r^{(p)}$ is any r-chain (coefficients integers mod p) on M_n on which the integrals are defined.

When $C^r_{(p)}$ is a cocycle mod p, we cannot show that Θ^r is regular over M_n (as was true for ordinary cocycles) for $\int_{\partial E_{r+1}} \Theta^r \equiv 0 \bmod p$ but is not necessarily equal to zero. However, it is true that

$$\int_{Z_r^{(0)}} \Theta^r \equiv 0 \qquad \text{mod } p$$

where $Z_r^{(0)}$ is a bounding integral cycle, for the values introduced by the singularities of Θ^r are all multiples of p.

Since a cohomology class with integers mod p for coefficients is determined by its values on those integral chains which are cycles mod p, we state the following analog of Theorem 4.

THEOREM 5. *Let $C^r_{(p)}$ be a fixed r-dimensional cohomology class (coefficients integers* mod p*) on M_n and $Z_r^{(0)}$ and $Z_r^{(p)}$ arbitrary homology classes with integer and integer* mod p *coefficients respectively. Then there exists a closed r-form Θ^r with singularities on $\partial {}^*C^{n-r}_{(p)}$ and an $r-1$ form Φ^{r-1} with singularities on ${}^*C^{n-r}_{(p)}$ and with $d\Phi^{r-1} = \Theta^r$ on $M_n - {}^*C^{n-r}_{(p)}$ such that*

$$C^r_{(p)} \cdot Z_r^{(0)} \equiv \int_z \Theta^r \bmod p, \qquad C^r_{(p)} \cdot Z_r^{(p)} \equiv \int_z \Theta^r - \int_{\partial z} \Phi^{r-1} \bmod p,$$

where z is a cycle which represents the corresponding homology class and on which the integrals are defined.

6. Extension of de Rham's second theorem. It is known that any closed form is locally derived and (from the second theorem of de Rham) that a closed form whose periods are all zero is derived in the large. Further, any closed form determines a cohomology class (real coefficients). Let C^r represent this class and let ${}^*C^{n-r}$ be the associated dual cycle. Then the following theorem is an extension of de Rham's theorem to the case where the periods are not all zero.

THEOREM 6. *Given any regular closed form ω^r on M_n, then there exists a form Ω^{r-1} regular except on ${}^*C^{n-r}$ such that on $M_n - {}^*C^{n-r}$*

$$\omega^r = d\Omega^{r-1}.$$

Let the forms corresponding to C^r be the Θ^r and Φ^{r-1} of Theorem 3. Define $\lambda^r = \omega^r - \Theta^r$. Then $d\lambda^r = 0$ and all the periods of λ^r are zero. Hence $\lambda^r = d\mu^{r-1}$ on M_n. Put $\Omega^{r-1} = \Phi^{r-1} + \mu^{r-1}$ on $M_n - {}^*C^{n-r}$. Then

$$d\Omega^{r-1} = d\Phi^{r-1} + \lambda^r \qquad = \Theta^r + \lambda^r = \omega^r.$$

7. Harmonic forms. The theorem of Hodge [2] states that among the numerous closed forms Θ^r which have the property described by de Rham's first theorem there is a unique harmonic form ω^r having this property. This suggests the following theorem.

THEOREM 7. *Theorems* 2, 3, 4, *and* 5 *remain valid if in their statements we replace* "Θ^r" *by* "*a harmonic form* ω^r" *and* "Φ^{r-1}" *by* "Ω^{r-1}". *In Theorems* 2, 3, *and* 4, ω^r *is unique, but in Theorem* 5 *it depends upon the choice of residues* mod p.

PROOF. The method of orthogonal projection [3] states that in M_n (if Θ^r is regular everywhere) or in $M_n - \partial^* C^{n-r}$ (if Θ^r is singular on $\partial^* C^{n-r}$) there exists a unique, regular, harmonic form ω^r and a regular form μ^{r-1} such that

$$\Theta^r = \omega^r + d\mu^{r-1}.$$

If we put $\Omega^{r-1} = \Phi^{r-1} - \mu^{r-1}$, we have that

$$\int_{E_r} \omega^r - \int_{\partial E_r} \Omega^{r-1} = \int_{E_r} \Theta^r - \int_{\partial E_r} \Phi^{r-1} + \left\{ \int_{\partial E_r} \mu^{r-1} - \int_{E_r} d\mu^{r-1} \right\}.$$

Since the expression in braces on the right vanishes, the forms ω^r, Ω^{r-1} have the required properties. When ω^r are singular, they are called "harmonic forms of the third kind" as defined by Kodaira [3, Theorem 17].

We can now construct the correspondence between cohomology with real coefficients or with integer coefficients and harmonic forms as follows below. In doing so we need the notion of an *integral harmonic form*. Let ω^r be harmonic, and let Ω^{r-1} be given as in Theorem 6 such that on $M_n - {}^*C^{n-r}$, $d\Omega^{r-1} = \omega^r$. Then ω^r is called *integral* if

$$\int_{E_r} \omega^r - \int_{\partial E_r} \Omega^{r-1}$$

is an integer for every cell E_r of K_r. With this understanding, we form the table of correspondences below:

Cohomology	*Unique harmonic forms* ω^r *and forms* Ω^{r-1} *such that* $d\Omega^{r-1} = \omega^r$
(1) Cochains (real coefficients)	(1) ω^r singular on $\partial^* C^{n-r}$ Ω^{r-1} singular on ${}^*C^{n-r}$
(2) Cochains (integer coefficients)	(2) As in (1) plus: ω^r is integral
(3) Cocycles (real or integer coefficients)	(3) As in (1) or (2) plus: ω^r regular on M_n
(4) Cobounding cocycles (real coefficients)	(4) As in (3) plus: $\omega^r = 0$; Ω^{r-1} unique harmonic
(5) Cobounding cocycles (integer coefficients)	(5) As in (4) plus: Ω^{r-1} integral
(6) Cobounding cocycles of order q—or torsion cocycles (integer coefficients)	(6) As in (4) plus: $q\Omega^{r-1}$ integral

A similar table can be readily constructed for cohomology with coefficients integers mod p.

References

1. C. B. Allendoerfer, *Characteristic cohomology classes in a Riemann manifold*, Ann. of Math. vol. 51 (1950) pp. 551–570.

2. W. V. D. Hodge, *The theory and applications of harmonic integrals*, Cambridge, Cambridge University Press, 1941.

3. K. Kodaira, *Harmonic fields in Riemannian manifolds*, Ann. of Math. vol. 50 (1949) pp. 587–665.

4. A. Weil, *The future of mathematics*, Amer. Math. Monthly vol. 57 (1950) pp. 295–306.

Haverford College,
Haverford, Pa., U. S. A.

TOPOLOGICAL GROUPS

SOME TOPOLOGICAL NOTIONS CONNECTED WITH A SET OF GENERATORS

P. A. SMITH

By a *local group*[1] we mean any system X which is like a group except that composition is not necessarily defined for all pairs of elements. The associative law takes the following form: if xy and yz are defined, then if one of the products $x(yz)$, $(xy)z$ is defined, so is the other and the two products are equal. It is assumed that each element of X has an inverse (in X) and that the products which are defined include $1x = x1 = x$, $xx^{-1} = 1$ for all $x \in X$. It is also assumed that $y^{-1}x^{-1}$ is defined whenever xy is. Evidently every group is a local group.

What is meant by a *local subgroup* of a local group Y is clear: X is a symmetric subset of Y which contains the identity and in which composition is simply the composition of Y cut down to X.

Which local groups can be imbedded in groups? It is easy to give examples of local groups which cannot be so imbedded. For example[2] let X consist of the distinct elements $1, x, y, z, u, v$ together with $a = ((x(yz))u)v$, $b = x(y(zu)v))$, the partial products of a and b, and the formal inverses of all these elements. Assume that no further compositions are defined except the trivial ones ($x1 = 1x = x, xx^{-1} = 1, \cdots$) and those required by the associative law (e.g., $z(z^{-1}y^{-1}) = y^{-1}$). X cannot be a local subgroup of a group since $a \neq b$.

Let us call a mapping f of one local group into another a homomorphism if $f(x)f(y)$ is defined and equal to $f(xy)$ whenever xy is defined. We shall call a homomorphism *strong* if $f(x)f(y)$ is defined *only* when xy is defined. An isomorphism (one-one homomorphism) between local groups is not a symmetric relation but a strong isomorphism is, and two local groups which are strongly isomorphic to each other are identical in structure.

The question of imbedding can now be stated as follows: which local groups are strongly isomorphic to local subgroups of groups. One can of course bring additional elements of structure into the picture by considering topological local groups and local Lie groups (their definitions need not be repeated here). The following remark however shows that so far as imbedding is concerned, it is sufficient to consider the purely algebraic case. Let X be a local subgroup of a group Q generated by X. If X is topological and possesses a symmetric nucleus Y such that YY is defined (in X), then Q can be converted into a topological group in such a way that Y is a nucleus of Q. If X is a local Lie group, Q will be a Lie group.

The well-known theorem of Cartan concerning the existence of Lie groups with given structure constants implies that every local Lie group Y admits a

[1] Called by Malcev [C. R. (Doklady) Acad. Sci. URSS. vol. 32] a symmetric groupoid.

[2] See Malcev, loc. cit.

local subgroup X, nucleus of Y, which is imbeddable in a group, i.e., is strongly isomorphic to a local subgroup of a group. The existing proofs of the Cartan theorem make use of the infinitesimal elements of Lie groups and seem to depend, more or less indirectly, on the vanishing of certain 2-dimensional homology groups. We shall presently state a result (Theorem 5) about purely algebraic local groups which concerns the imbedding of homomorphisms of local groups rather than the local groups themselves. Again the vanishing of certain 2-dimensional connectivities is involved, this time quite explicitly.

We begin by defining certain groups $\mathfrak{p}_1$, $\mathfrak{p}_2$ associated with a local group X. We denote by $G(X)$ the group of equivalence classes of words formed with the elements of $X - \{1\}$, two words being equivalent if one is obtainable from the other by insertions and deletions of syllables of the form xx^{-1}. ($G(X)$ is not necessarily a free group since x may equal x^{-1}.) Let w be a word formed with the elements of $X - \{1\}$. Call *allowable* any insertion of parentheses into w which would reduce w to an element of X if all the indicated products were defined. Call w a *local* word if the products indicated by any allowable insertion of parentheses are all defined. If w is local, any insertion of parentheses will reduce w to an element $\mu(w)$ of X. Using the associative law it can be shown that $\mu(w)$ is independent of the choice of parentheses. Now let $V(X)$ be the totality of equivalence classes of local words w such that $\mu(w) = 1$. $V(X)$ is in general not a group. Let $D(X)$ be the smallest normal subgroup of $G(X)$ containing $V(X)$. Let

$$\mathfrak{p}_1(X) = G(X)/D(X). \tag{1}$$

Assume now that X is a local subgroup of a group Q generated by X. The mapping which associates to each word formed with the elements of $X - \{1\}$ its group product in Q induces a homomorphism α of $G(X)$ onto Q. Let $R(Q, X)$ be the kernel of α. Evidently $V(X) \subset R(Q, X)$, hence $D(X) \subset R(Q, X)$. Let

$$\mathfrak{p}_1(Q, X) = R(Q, X)/D(X).$$

We shall say that Q is *simply connected* relative to X if $\mathfrak{p}_1(Q, X) = \{1\}$.

THEOREM 1. *Let X be a local subgroup of a group Q generated by X and let Q' be a group. If Q is simply connected relative to X, any homomorphism of X into Q' can be extended in a unique manner to a homomorphism of Q' into Q.*[3]

Let $\tilde{Q}$, Q be groups generated by local subgroups $\tilde{X}$, X and let τ be a homomorphism of $\tilde{Q}$ onto Q. We shall call $(\tilde{Q}, \tilde{X}, \tau)$ an *even covering* of (Q, X) if $\tau \mid \tilde{X}$ is a strong isomorphism of $\tilde{X}$ onto X.

THEOREM 2. *If X generates Q, there exists an even covering $(\tilde{Q}, \tilde{X}, \tau)$ of (Q, X) such that $\tilde{Q}$ is simply connected relative to X. This covering is essentially unique and $\tilde{Q} \cong \mathfrak{p}_1(X)$.* (See (1).)

[3] Proofs of the principal theorems here stated will appear elsewhere.

Theorems 1 and 2 are easily proved counterparts of well-known theorems about topological groups.

Let $\chi = \{X\}$ be a class of local subgroups of a group Q and assume that each X generates Q. Call Q simply connected relative to χ if it is s.c. with respect to each X. Suppose that the class χ is determined by Q. Then simple connectedness relative to χ is a group theoretic invariant of Q. For example, χ might consist (1) of all X's generating Q or (2) of all X's generating Q and such that $\mathfrak{p}_2(X) = \{1\}$ (see below). In the first case it seems likely that the only groups which are s.c. relative to χ are the cyclic groups of orders 1, 2, 3.

A relation between algebraic and topological simple connectedness is given by the following theorem.

THEOREM 3. *Let Q be a locally connected topological group and let χ consist of all symmetric nuclei of Q. Then Q is simply connected in the topological sense*[4] *if and only if Q is simply connected relative to χ.*

We turn to the definition of $\mathfrak{p}_2$. Let X be a local group and let $A(X)$ be the totality of formal expressions.

$$a = (g_1 * v_1) \cdots (g_n * v_n)$$

where $g_i \in G(X)$, $v_i \in V(X)$. We introduce an associative multiplication into A by juxtaposition, and a neutral element a_0. We define

$$a^{-1} = (g_1 * v_1^{-1}) \cdots (g_n * v^{-1}).$$

We allow $G(X)$ to operate on $A(X)$ according to the rule

$$g * a = (gg_1 * v_1) \cdots (gg_n * v_n).$$

Let β be the mapping $A \to D(X)$ defined by

$$\beta a = g_1 v_1 g_1^{-1} \cdots g_n v_n g_n^{-1}.$$

We introduce an equivalence $\sim$ into $A(X)$ by the relations

$$g * 1 \sim a_0 \tag{2}$$

$$(g * v_1)(g * v_2) \sim g * v_1 v_2 \tag{3}$$

$$gh * v \sim g * hvh^{-1} \tag{4}$$

$$a_1 a_2 a_1^{-1} \sim \beta a_1 * a_2 . \tag{5}$$

These formulas of course define relations only when they are meaningful. Thus (2) is a relation only when v_1, v_2, $v_1 v_2$ are in $V(X)$; (3) is a relation only when v and hvh^{-1} are in $V(X)$. In (4) on the other hand a_1, a_2 are arbitrary elements of $A(X)$.

The equivalence classes of $A(X)$ form a multiplicative system $\mathfrak{A}(X)$. It

[4] As defined by Chevalley [*The theory of Lie groups*].

follows from (2) and (5) that $aa^{-1} \sim a_0$. Hence $\mathfrak{A}(X)$ is a group. It is easy to see that β induces a homomorphism

$$\gamma\colon \mathfrak{A}(X) \to D(X).$$

Let $\mathfrak{p}_2(X)$ be the kernel of γ. The group $\mathfrak{p}_2$ lies in the center of $\mathfrak{A}$ and is therefore abelian. For suppose a is a representative in A of an element of $\mathfrak{p}_2(X)$. Then $\beta a = 1$ so that $aa_1a^{-1} \sim 1 * a_1 = a_1$ for each $a_1 \in A(X)$.

Now assume that X is a local subgroup of a group Q generated by X. Let $K(Q, X)$ be the simplicial complex whose n-simplexes are subsets $(q_0, \cdots, q_n)$ of Q of cardinal number $n + 1$ such that $q_i^{-1}q_j \in X$ $(i, j = 0, \cdots, n)$. It is easy to see that K is connected. It can be shown that

$$\mathfrak{p}_1(Q, X) = \pi_1(K), \qquad \mathfrak{p}_2(X) = \pi_2(K)$$

where $K = K(Q, X)$ and π denotes homotopy groups.[5]

If Q is a topological group satisfying certain conditions, $\mathfrak{p}_1$ and $\mathfrak{p}_2$ determine the first and second homotopy groups of Q. As concerns π_1, the situation is partially described by Theorem 3. We pass to the case of π_2. Let X be a local subgroup of a local group Y. The relation $X \subset Y$ gives rise in a natural manner to a homomorphism $i\colon \mathfrak{p}_2(X) \to \mathfrak{p}_2(Y)$. We denote the image $i\mathfrak{p}_2(X)$ by $\mathfrak{p}_2(X, Y)$. Now let Q be a topological group and consider the following conditions:

I. For every nucleus V there exists a nucleus $U \subset V$ such that singular 1-spheres in U are nullhomotopic in V.

II. There exists a nucleus W such that singular 2-spheres in W are nullhomotopic (in Q).

THEOREM 4. *Let Q be an arcwise connected topological group satisfying conditions* I *and* II. *Then there exists a nucleus Y_0 such that every symmetric nucleus X contained in Y_0 admits a symmetric subnucleus X such that $\mathfrak{p}_2(X, Y) = \pi_2(Q)$.*

It is known that if Q is a Lie group, then $\pi_2(Q)$ is trivial. This is not true however for topological groups in general.[6] We shall see that the problem of determining whether or not the second homotopy group of a topological group Q is trivial can be reduced to a purely algebraic study of the "extensions" of the local subgroups of Q. How the algebraic structure of Q determines the group $\pi_2(Q)$ when it is not trivial is not yet known.

Let Y be a local group. An *extension* of Y is a pair (E, ϕ) where E is a local group and ϕ a strong homomorphism of E onto Y. Note that if Y is a group, so is E. Let (E, ϕ) be an extension of Y and let X be a local subgroup of Y. It is

[5] Since K is not a space, π_1 and π_2 must be understood as being defined combinatorially. Thus π_2 is the second homology group, based on finite chains and integer coefficients, of the universal covering of K.

[6] An example of an arcwise topological group Q for which $\pi_2(Q)$ is infinite cyclic was communicated to me independently by Dr. H. C. Wang and Dr. Robert Taylor.

easy to see that by taking $E' = \phi^{-1}X$ and $\phi' = \phi \mid E'$, we obtain an extension (E', ϕ') of X. We write $(E', \phi') = (E, \phi) \mid X$.

Let Y, Z be local groups such that $Y \subset Z$. An extension (E, ϕ) of Y will be called *extendible over* Z if there exists an extension (E_1, ϕ_1) of Z such that $(E, \phi) = (E_1, \phi_1) \mid Y$. We shall say that (E, ϕ) is extendible over Z *from* X, where X is a local subgroup of Y, if $(E, \phi) \mid X$ is extendible over Z. We can now state the following theorem.

THEOREM 5. *Let X and Y be local subgroups of a group Q with X a local subgroup of Y. Assume that X generates Q, that Q is simply connected with respect to X, and that Y contains no elements of order 2. Then a necessary and sufficient condition that every extension of Y be extendible over Q from X is that $\mathfrak{p}_2(X, Y) = \{1\}$.*

Suppose Q is a simply connected topological group satisfying conditions I and II and that Q does not possess "arbitrarily small" elements of order 2. From Theorems 4 and 5 we see that there exists symmetric nuclei X and Y in Q with $X \subset Y$ such that the question whether or not $\pi_2(Q)$ is trivial is equivalent to the question whether or not all extensions of Y are extendible over Q from X.

By using Theorems 1 and 2 it is easy to see that an extension (E, ϕ) of a subgroup X of a group Q generated by X is extendible over Q if and only if E is imbeddable in a group. This suggests another form for Theorem 5; we state it only for the case $X = Y$.

THEOREM 5'. *Let X be a local subgroup of a group Q generated by X and assume that Q is simply connected relative to X and that X contains no elements of order 2. A necessary and sufficient condition that every local group E which can be mapped onto Q by a strong homomorphism be imbeddable in a group is that $\mathfrak{p}_2 X = \{1\}$.*

We now state in rough form an application of Theorem 5' to the problem of imbedding local groups in groups. Let Z be a local group and N a "normal" local subgroup of Z. Assume that the local groups N and $Y = Z/N$ are imbeddable in groups, that Y contains no elements of order 2, and that $\mathfrak{p}_2(Y) = \{1\}$. Then Z is imbeddable in a group.

In fact, it can be shown in a fairly straightforward manner that the "cosets" of N are contained in larger sets whose union E is a local group which contains Z and is mappable onto $Y = Z/N$ by a strong homomorphism (determined by the natural mapping $Z \to Z/N$). From Theorem 5', E is imbeddable in a group, and therefore so is Z.

An accurate statement of this theorem is complicated by the fact that the concepts "normal local subgroup" and "local quotient group" are somewhat nebulous. When accurately stated, the theorem asserts that a certain local subgroup Z' of Z, rather than Z itself, is imbeddable. When Z is topological, Z' can be taken to be a nucleus. The theorem can be given a formulation which

involves $\mathfrak{p}_2(X, Y)$ rather than $\mathfrak{p}_2(X)$. If the group Q in which Z/N is assumed to be imbeddable is topological and satisfies conditions I and II, the condition $\mathfrak{p}_2(X, Y) = \{1\}$ can be replaced by $\pi_2(Q) = \{1\}$ (see Theorem 4). It is not hard to see that in this form, the theorem furnishes a new proof of the fact that every local Lie group Z has a nucleus which is imbeddable in a group, provided one makes use of the proposition that the second homotopy group of every semi-simple group is trivial. One has only to take N to be a maximal solvable local subgroup of Z, so that N/Z is semi-simple.[7]

COLUMBIA UNIVERSITY,
NEW YORK, N. Y., U. S. A.

[7] That solvable and semi-simple local Lie groups are imbeddable in groups (or at least admit imbeddable symmetric nuclei) is little more than a triviality.

PROPERTIES OF FINITE-DIMENSIONAL GROUPS

DEANE MONTGOMERY

Topological groups and topological transformation groups are subjects without a very clear dividing line. To mention one connection, a subgroup H of a group G can be regarded as a transformation group acting on G by either left or right translations or by inner automorphisms. This paper recalls a few of the results and problems in both topics.

Let G be a topological group and M a topological space. A function $f(g; x) = g(x)$ defined and continuous on $G \times M$ with values in M is called a transformation group if

(a) for each fixed g, $g(x)$ is a homeomorphism of M onto itself,

(b) $$g_1[g_2(x)] = (g_1g_2)(x).$$

The transformation group is called effective if only the identity leaves all of M fixed. The set $G(x)$ for any x in M is called the orbit of x. The spaces G and M will always be locally compact and separable metric so that dimension theory may be used. When G and M coincide and G acts on itself by translation, the subject becomes the study of the topological group G.

A group which has a simple structure may offer difficult questions when operating as a transformation group. For example, the ways in which a cyclic group of order 2 can operate on a manifold, even on E^3, are far from completely known.

If h is any homeomorphism of M onto itself, then the above transformation group determines another in which $g(x)$ is replaced by $(hgh^{-1})(x)$. Two transformation groups related in this way are called equivalent and it is natural to try to determine conditions which imply the equivalence of two transformation groups. When equivalence is likely but unproved, it can be asked if some of the homology or other properties of the two transformation groups are the same.

It is often conjectured that many transformation groups are equivalent to known or comparatively simple ones. One question of this kind is the problem of Hilbert which asks, when G and M are both manifolds, whether coordinates may be so introduced in G and M that $f(g; x)$ becomes analytic in both variables. This is probably true in 1 below and at least in part in 2.

1. G is a manifold which acts on itself by group translation.

2. G is a compact manifold, and M is any manifold. In this case it follows from the work of von Neumann [13] that G is a Lie group, so that analytic parameters can be introduced into G; thus the problem 2 is that of also choosing x so that $f(g; x)$ is simultaneously analytic.

Case 2 suggests the following:

2′. If G is compact (not necessarily a manifold) and acts effectively on a manifold M, does this imply that G is a manifold, and hence a Lie group?

In connection with 2 and 2′, Zippin and the author have shown [10; 8; 16] that any compact connected group acting effectively on a three-dimensional manifold M must be a Lie group. If $M = E^3$, we showed further that G must be equivalent either to the group of all rigid motions about an axis or to the group of all rigid motions about a point. The case where G is a compact, zero-dimensional, effective group on three and higher dimensional manifolds remains open. Presumably such a G must be finite as suggested in 2′, but this has not been proved, and even if this is assumed and $M = E^3$, it is not known whether G is equivalent to a group of orthogonal transformations. This latter question has not been answered even in the differentiable case. It has been shown that a compact effective zero-dimensional group acting on a two-dimensional manifold is finite.

When M is an n-dimensional manifold and G is compact, effective, and has locally connected orbits, Zippin and the author have shown that G must be a Lie group [12]. This result is related to the fact that a pointwise periodic homeomorphism of a manifold must be periodic, that is, if every point has a period, then these periods are bounded [7]. If M is differentiable, G is compact and effective, and each element of G is of class C^1, then G is a Lie group and $f(g; x)$ is of class C^1 in both variables. Bochner and the author have shown [1] that if G is a Lie group and if $f(g; x)$ has certain properties of differentiability or analyticity with respect to x, then $f(g; x)$ has the same properties with respect to g. If M is a manifold and G is finite, many beautiful results have been obtained by Newman and Smith.

Even when G is a Lie group and M is a differentiable or analytic manifold and if $f(g; x)$ is assumed simultaneously differentiable or analytic, then many problems remain. At first glance this might seem unlikely after the great work of Lie and his followers. But much of their work is concerned with local questions and is not concerned with the topology of orbits or the interconnection of the topology of orbits, group, and space. Much progress has been made recently, especially by the French mathematicians, on the important case where a compact Lie group acts transitively on a compact manifold.

When a compact group acts intransitively, the orbits decompose the space in a manner reminiscent of a fibering, but there may be various kinds of singularities, and this suggests that it might be useful to consider fiberings with singularities to a greater extent than has been done so far.

For the remainder of the paper the topological group G will be considered for its own sake. As has been remarked this is a special case of the study of transformation groups. It is helpful to assume that G is n-dimensional and this will be done from this point on. Dimension is used in the sense of set theory and does not necessarily mean that G has local coordinates.

If H is a closed subgroup, it can be shown, using a theorem of Hurewicz, that $\dim G \leqq \dim H + \dim G/H$.

It is probable that

$$\dim G = \dim H + \dim G/H,$$

but this has been proved only in a few special cases. It has not been proved in general that $\dim G/H$ is finite, although the author has proved this when H is abelian. If H is n-dimensional, and G is connected, then $H = G$.

In low dimensions the structure of G can be analyzed completely, although not without considerable effort. If $\dim G = 1$, G connected, then G is either compact or isomorphic to the group of real numbers. If $\dim G = 2$, G connected, then G is either abelian and has a known structure or else it is the nonabelian group of transformations of the form $ax + b$, a positive. If G is locally connected and $\dim G = 1, 2$, or 3, then G must be locally euclidean. In these cases, that is, G locally connected and $\dim G = 1, 2, 3$, it is known that G is a Lie group [4].

For a general n, the local topological structure of G is known when G is compact or abelian by the work of von Neumann and Pontrjagin [1]. In this case there is an open set U including e with U a direct topological and group theoretical product

$$U = Z \times C$$

where Z is a compact zero-dimensional group and C is an n cell and a local Lie group. In the general case not nearly this much is known, but, using Gleason's result on the existence of an arc in a locally compact connected group [2], it has been shown that there is a neighborhood U of e with U a topological product

$$U = Z \times C$$

where Z is a compact zero-dimensional set and C is a connected, locally connected, invariant, n-dimensional local group [5]. By retopologizing the subgroup generated by C it is possible to prove the following theorem.

THEOREM 1. *Let G be a locally compact connected n-dimensional group. Then there exists a connected locally connected locally compact n-dimensional group L and a continuous one to one homomorphism of L into G such that $h(L)$ is everywhere dense in G.*

A familiar illustration occurs when G is a solenoid and L is the group of real numbers. This result reduces many questions about groups to the locally connected case.

Locally connected n-dimensional groups have been shown to have a few properties which make them resemble n-dimensional manifolds [6]. If such a group contains a closed n-dimensional subset E, then E contains an inner point. A similar fact is true for an n-dimensional homogeneous space of a locally connected group. In either case a sufficiently small compact set which carries an

essential $n - 1$ cycle must separate, and a closed $(n - 1)$-dimensional set must separate locally. Local homology connectedness can be proved in some dimensions. More generally, if a space M has a certain property called local homogeneity which, speaking roughly, requires that the space can be deformed, locally, rather freely, then M has the above-mentioned properties of a manifold. The property of local homogeneity is possessed by all manifolds but in how far it may characterize manifolds is not known and is probably worth study.

In another direction and by using some of the above results, Zippin and the author [9] have shown the following:

THEOREM 2. *Let G be a locally compact connected n-dimensional group, $n > 0$, which is not compact. Then G contains a closed subgroup isomorphic to the real numbers.*

It was shown previously by Gleason [3] that every n-dimensional group, $n > 0$, contained a one-parameter subgroup and by Zippin [15] that every two-ended group, whether finite-dimensional or not, contained a subgroup isomorphic to the real numbers.

One of the tools used in proving Theorem 1 is the device of considering G as acting on itself by inner automorphisms. A brief outline of the proof is now given. Assuming the theorem false implies the existence of a compact connected abelian subgroup of positive dimension. Using inner automorphisms shows then that every element of G is in such a compact abelian subgroup and this turns out to be a contradiction to the noncompactness of G.

Theorem 1 has as a corollary that G is a topological product

$$G = F \times R$$

where R is homeomorphic to a line. This follows from the fact that there is a cross section in the large of the cosets of a subgroup isomorphic to the reals. It is probable that a two-ended group is the direct product of a compact group and the group of real numbers, but this has not been shown.

Again by considering the group as acting on itself by inner automorphisms, Zippin and the author [11] have proved the following theorem.

THEOREM 3. *Let G be a locally compact, connected, n-dimensional group, $n > 1$, which is not compact. Then G contains a closed connected two-dimensional subgroup which is not compact.*

The corresponding fact is not true for compact groups since the proper orthogonal group in three variables and its covering group do not contain two-dimensional subgroups, but these two groups are the only exceptions to the statement that every locally compact n-dimensional group, $n > 1$, contains a closed two-dimensional subgroup.

Bibliography

1. S. Bochner and D. Montgomery, *Groups of differentiable and real or complex analytic transformations*, Ann. of Math. vol. 46 (1945) pp. 685–694.

2. A. M. Gleason, *On the existence of arcs in locally compact connected groups*, Bull. Amer. Math. Soc. vol. 55 (1949) p. 1079.

3. ———, *Arcs in locally compact groups*, Proc. Nat. Acad. Sci. U. S. A. vol. 56 (1950) pp. 663–667.

4. D. Montgomery, *Analytic parameters in three dimensional groups*, Ann. of Math. vol. 49 (1948) pp. 118–131.

5. ———, *Finite dimensional groups*, Ann. of Math. vol. 52 (1950) pp. 591–605.

6. ———, *Locally homogeneous spaces*, Ann. of Math. vol. 52 (1950) pp. 261–271.

7. ———, *Pointwise periodic homeomorphisms*, Amer. J. Math. vol. 59 (1937) pp. 118–120.

8. D. Montgomery and L. Zippin, *Compact abelian transformation groups*, Duke Math. J. vol. 4 (1938) pp. 363–373.

9. ———, *Existence of subgroups isomorphic to the real numbers*, Ann. of Math. vol. 53 (1951) pp. 298–326.

10. ———, *Non-abelian, compact, connected transformation groups of three space*, Amer. J. Math. vol. 61 (1939) pp. 375–387.

11. ———, *Two dimensional subgroups*, Proceedings of Amer. Math. Soc. vol. 2 (1951) pp. 822–838.

12. ———, *Topological transformation groups*, Ann. of Math. vol. 41 (1940) pp. 778–791.

13. J. von Neumann, *Die Einführung analytischer Parameter in topologische Gruppen*, Ann. of Math. vol. 34 (1933) pp. 170–190.

14. L. Pontrjagin, *Topological groups*, Princeton, Princeton University Press, 1939.

15. L. Zippin, *Two ended topological groups*, Proceedings of Amer. Math. Soc. vol. 1 (1950) pp. 309–315.

16. ———, *Transformation groups*, Michigan lectures in topology, edited by Wilder and Ayres, Ann Arbor, 1941, pp. 191–221.

Institute for Advanced Study,
Princeton, N. J., U. S. A.

SOME PROPERTIES OF (L)-GROUPS[1]

KENKICHI IWASAWA

One of the main problems in the theory of topological groups at the present time is the study of the structure of locally compact groups in both algebraical and topological aspects. In some particular cases, for example, in the case of Lie groups, we already know much about them. There are indeed still many difficult problems left unsolved concerning Lie groups, but we are in possession of powerful methods from algebra and analysis by which we can research deeply in these groups. The situation is similar for compact groups or locally compact abelian groups, for they are projective limits of Lie groups and we can study their structure by approximating them as precisely as we want by Lie groups. Thus the fifth problem of Hilbert is solved for these groups.

However if we leave these particular cases and consider locally compact groups in general, our knowledge about them is, it may be said, as yet insufficient. Indeed we have no general method by which we can analyze the structure of these groups as we do in the case of compact or abelian groups. Therefore it seems advisable to me to consider first the class of those locally compact groups which can be studied by known methods just as in the special cases mentioned above, and then ask the relation between these groups and general locally compact groups. Such are the class of (L)-groups, namely the class of those locally compact groups which are projective limits of Lie groups.[2]

As one expects from this definition, (L)-groups, in particular connected (L)-groups, have properties similar to those of Lie groups in various aspects. First, it is easy to see that subgroups and factor groups of (L)-groups are also (L)-groups.[3] Moreover we can prove the following theorem.

THEOREM 1.[4] *Let G be a connected locally compact group and N a closed normal subgroup of G. If N and G/N are both* (L)*-groups, then G itself is an* (L)*-group.*

The proof is essentially based upon the following lemma.

[1] This address was listed on the printed program under the title *Locally compact groups.*

[2] In my paper, *On some types of topological groups*, Ann. of Math. vol. 50 (1949), I defined the (L)-group as a locally compact group G containing a system of normal subgroups N_α such that G/N_α are Lie groups and such that the intersection of all N_α is e. If G is a projective limit of Lie groups, G surely contains such a system of normal subgroups, and the converse can also be proved to be true when we assume that G is connected. But we do not know whether these two definitions are completely equivalent.

[3] If we take the former definition of (L)-groups, we cannot yet decide whether a factor group of an (L)-group is always an (L)-group or not.

[4] For the proof of following theorems see the paper of K. Iwasawa, *On some types of topological groups*, Ann. of Math. vol. 50 (1949). Cf. also A. M. Gleason, *On the structure of locally compact groups*, Proc. Nat. Acad. Sci. U. S. A. vol. 35 (1949).

LEMMA. *Let G and N be as above. If N and G/N are both Lie groups, G itself is also a Lie group.*

Theorem 1 and the foregoing remark show us that the class of connected (L)-groups is closed under the group-theoretical operations, such as taking subgroups, forming factor groups or group extensions. Now Lie groups are of course (L)-groups, and compact groups and locally compact abelian groups are also known to be (L)-groups. Therefore those connected locally compact groups, which can be obtained from these "known groups" by successive group-theoretical operations as stated above, are always (L)-groups. In particular, connected locally compact solvable groups are (L)-groups.

A connected (L)-group has a characteristic local structure. We have namely the following theorem.

THEOREM 2. *A connected locally compact group G is an* (L)*-group if and only if G is locally the direct product of a local Lie group L and a compact normal subgroup K.*[5]

This theorem implies in particular that a locally Euclidean (L)-group is a Lie group. Therefore, as every connected locally compact solvable group is an (L)-group, every locally Euclidean solvable group is a solvable Lie group.[6]

As to the global structure of (L)-groups we have the following theorem.

THEOREM 3. *Let G be a connected* (L)*-group. Any compact subgroup of G is then contained in some maximal compact subgroup of G, whereas such maximal compact subgroups of G are all connected and conjugate to each other. Let K be one of them. G contains then subgroups $H_1, \cdots, H_r$, which are all isomorphic to the additive group of real numbers and are such that any element g in G can be decomposed uniquely and continuously in the form*

$$g = h_1 \cdots h_r k, \qquad h_i \in H_i, k \in K.$$

In particular, the space of G is the cartesian product of the compact space of K and that of $H_1 \cdots H_r$, which is homeomorphic to the r-dimensional Euclidean space.

This theorem is obviously a generalization of Cartan-Malcev's theorem on the topological structure of Lie groups.

By Theorem 3 we see that the topological structure of a connected (L)-group G is completely determined by that of its maximal compact subgroup K. For example, the study of the topological structure of a connected locally compact solvable group G is reduced to that of a compact abelian group, for we can prove in this case that the maximal compact subgroup K is an abelian group.

[5] For a simple proof cf. A. Borel, *Limites projectives de groupes de Lie*, C. R. Acad. Sci. Paris vol. 230 (1950).

[6] Chevalley's theorem.

Now, by making use of Theorem 3, we can also prove the following theorem.

THEOREM 4. *A connected locally compact group G contains a uniquely determined maximal connected solvable normal subgroup R, in which any other connected solvable normal subgroups of G are contained.*

We call such an R, just as in the case of Lie groups, the *radical* of G. We also say that G is semi-simple if the radical R of G is equal to e. In general a connected locally compact group G is an extension of the solvable radical R by the semi-simple group $\bar{G} = G/R$.

For (L)-groups we have moreover the following generalization of E. E. Levi's theorem.

THEOREM 5. *Let G be a connected* (L)*-group and R the radical of G. Then G contains a subgroup S, such that*

(i) *S is the homomorphic image of a connected semi-simple* (L)*-group S' in G,*

(ii) *$RS = G$ and $R \cap S$ is a totally disconnected normal subgroup of G.*[7]

Now it can be proved that a connected locally compact group G contains not only the radical R, but also a uniquely determined maximal connected compact normal subgroup N. By making use of these facts we have the following theorem.

THEOREM 6. *A connected locally compact group G contains a uniquely determined maximal normal subgroup Q of type (L), and G/Q contains no normal subgroup of type (L) other than e.*

On the other hand, it follows immediately from the definition of (L)-groups that a connected locally compact group G contains a uniquely determined minimal normal subgroup Q' of type (L). These facts show us some relations between the class of all connected (L)-groups and general connected locally compact groups. But we have as yet no essential result concerning the situation of connected (L)-groups in the set of all connected locally compact groups. However we have the following conjecture:[8]

(C_1) Any connected locally compact group is an (L)-group.

By theorem 6 this conjecture is easily seen to be equivalent to the following:

(C_2) A connected locally compact group which contains no arbitrary smal normal subgroup is a Lie group.

[7] Cf. A. Borel, *Limites projectives de groupes de Lie*, C. R. Acad. Sci. Paris vol. 230 (1950) and Y. Matsushima, *On the decomposition of an* (L)*-group*, Journal of the Mathematical Society of Japan vol. 1 (1950).

[8] Cf. K. Iwasawa, *On some types of topological groups*, Ann. of Math. vol. 50 (1949) and A. M. Gleason, *On the structure of locally compact groups*, Proc. Nat. Acad. Sci. U. S. A. vol. 35 (1949).

We can also prove that (C_1), (C_2) are equivalent to the following:

(C_3) A connected locally compact simple group is a simple Lie group.

If these conjectures could be proved to be true, we would have not only a complete solution of the fifth problem of Hilbert, but also decisive progress in the theory of locally compact groups.

INSTITUTE FOR ADVANCED STUDY,
PRINCETON, N. J., U. S. A.

ONE-PARAMETER SUBGROUPS AND HILBERT'S FIFTH PROBLEM

A. M. GLEASON

The affirmative solution of Hilbert's fifth problem requires that we bridge the gap between topologico-algebraic structure and analytic structure. In building this bridge we quite naturally seek an intermediate island on which to rest the piers. Such an island is provided by the one-parameter subgroups. One-parameter subgroups are themselves a topologico-algebraic concept and their existence can be demonstrated, in some cases at least, by the methods of topological algebra. On the other hand the one-parameter subgroups are perhaps the most striking feature of a Lie group and it is known that the analytic structure can be recovered from them.

A one-parameter subgroup of a group G is a subgroup which is a (continuous) homomorphic image of the additive group of real numbers R. We do not require that the subgroup be closed. The structure of such a subgroup can be quite complicated, even if the group G is locally compact, a condition which we shall assume throughout. If we consider only a part of the subgroup corresponding to a small segment of the reals including 0, the complications vanish and all such *local* one-parameter subgroups look the same. In a Lie group there is a neighborhood U of the identity e such that every element x of U is on a one-parameter subgroup. Furthermore this one-parameter subgroup is unique if we require that it go directly from e to x without leaving U. We shall say that a group has a canonical family of one-parameter subgroups if there exists a neighborhood U with these properties.

Our intermediate goal is to prove that every locally Euclidean group has a canonical family of one-parameter subgroups. Quite recently great strides have been made toward this objective.

An important class of groups is the class of those which contain no small subgroups; that is, those which have a neighborhood of the identity containing no entire subgroup except (e). It has long been known that a group with no small subgroups contains a one-parameter subgroup. An extension of this result was made by Chevalley and the author independently: A locally connected group of finite dimension which contains no small subgroups has a canonical family of one-parameter subgroups. Unfortunately little is known about the existence or nonexistence of small subgroups in locally connected groups of finite dimension. However, under the stronger hypothesis that the group be locally Euclidean, Newman has shown that there is a neighborhood of the identity containing no finite subgroup. Smith has extended his investigation and shown that if there are arbitrarily small subgroups H in a locally Euclidean group, then some of them must satisfy the implausible relation $\dim G/H > \dim G$.

Without the hypothesis concerning small subgroups we can say less, but still a great deal. The author proved that every n-dimensional group G $(n > 0)$

contains a one-parameter subgroup. Montgomery and Zippin have shown that, provided G is not compact, this subgoup can be chosen isomorphic to R. The method of proof, in both cases, is to prove the existence of subgroups of lower dimension, eventually winding up with a subgroup of dimension one. This method does not seem applicable to proving that there is a canonical family of one-parameter subgroups.

On the other side of our island the situation is not so bright. This is to be expected; it is here that we must make the transition from topological algebra to analysis. On the analytical side of the channel, the stringency of the conditions leading to analytic structure have gradually been relaxed from requiring three times differentiable coordinates to certain rather strong Lipschitz conditions; but all conditions have been truly analytic in character, and it seems safe to say that the first purely analytic result derived by the methods of topological algebra will prove decisive.

Consider the class Φ of homomorphisms of R into G. (There will of course be many distinct homomorphisms onto each one-parameter subgroup.) If G is a Lie group, every homomorphism of R into G has the form $t \to \exp tX$ where X is an element of the Lie algebra $\mathfrak{g}$ of G; hence, we may identify Φ with $\mathfrak{g}$. Evidently we should attempt to introduce the structure of a Lie algebra into Φ.

To introduce the additive structure into Φ, we again turn to the theory of Lie groups for our cue. We have the basic formula

$$\exp (X + Y) = \lim_{n\to\infty} (\exp X/n \exp Y/n)^n. \tag{1}$$

To carry this over to Φ we must prove that, for any two homomorphisms φ_1 and φ_2 of R into a locally Euclidean group, $\lim_{n\to\infty}(\varphi_1(1/n)\varphi_2(1/n))^n$ exists. If this is true, we can define addition in Φ quite simply by

$$(\varphi_1 + \varphi_2)(t) = \lim_{n\to\infty} (\varphi_1(t/n)\varphi_2(t/n))^n. \tag{2}$$

The commutativity of this operation is easily proved, but the associativity is in doubt. Scalar multiplication is defined, of course, by $(\alpha\varphi)(t) = \varphi(\alpha t)$, and it satisfies the distributive law with respect to addition.

If these ideas can be carried out, making Φ a linear vector space, it will follow that G is a Lie group. For if C is the center of G, then G/C will be represented faithfully by the linear transformations induced on Φ by the inner automorphisms of G. Hence G is a generalized Lie group, and, being locally Euclidean, it is a Lie group.

It may be noted that the program here outlined does not actually require that we have a canonical family of one-parameter subgroups. It is sufficient that the set of one-parameter subgroups generate G, and this can be shown to be no essential restriction.

HARVARD UNIVERSITY,
CAMBRIDGE, MASS., U. S. A.

RECENT DEVELOPMENT OF KNOT THEORY AT PRINCETON

R. H. Fox

I shall report[1] here on some of the principal results in knot theory obtained during the past five years by myself and my colleagues, and try to indicate the trend of these developments.[2]

Let us examine the question: What is knot theory and what is its proper domain? The subject matter of knot theory is usually considered to be the situation in euclidean 3-space of simple closed polygons or of systems of μ disjoint simple closed polygons, $1 \leqq \mu < \infty$. This description of what I may call classical knot theory tends, by its narrowness, to isolate the subject from the rest of topology. It is to be hoped that the various special theorems which make up classical knot theory will eventually turn out to be particular cases of general topological theorems. In working toward this end the following principles seem almost obvious:

(A) *The class of polygons should be replaced by a suitable topologically defined class of curves.* The minimal topologically invariant class is the class of *tame curves*,[3]—those which can be transformed into polygons by autohomeomorphisms of the underlying manifold. However an effective topological definition of this class is lacking. The maximal class, the class of *all simple closed curves*, has the disadvantage that it introduces into the subject the new complication of local pathology.[4] It would seem to be preferable to consider the intermediate class of *smooth curves*,—a curve is smooth if it can be transformed into a polygon by a homeomorphism of one of its neighborhoods onto a regular neighborhood of the polygon. It is natural to conjecture that every smooth curve is tame, but this has not been proved.

(B) *Euclidean 3-space should be replaced by other compact*[5] *3-manifolds.* Since the complement of a neighborhood of a smooth curve is a manifold with boundary, it would seem advantageous to admit these as well as the closed manifolds.

In accordance with these principles we see that *the natural domain of knot theory is the study of the types of ordered systems* $K = (K_1, K_2, \cdots, K_\mu)$ *of disjoint oriented smooth simple closed curves in the interior of a compact 3-manifold*

[1] In the interest of brevity several of the definitions are intentionally vague or even inaccurate.

[2] Since I am speaking about the work of the group with whom I have been associated at Princeton, I shall not be discussing the various recent results obtained in this field by Reidemeister and Seifert and those associated with them at Marburg and Heidelberg, although there has been a certain amount of overlapping. References to their results may be found in [24] and [20]. Numbers in brackets refer to the references at the end of the paper.

[3] Terminology introduced in [3].

[4] Instructive examples of such pathology have been studied by E. Artin and R. Fox [3], R. Fox [10], W. A. Blankinship [6], and W. A. Blankinship and R. H. Fox [7].

[5] In classical knot theory euclidean 3-space may be replaced with advantage by spherical 3-space S, which is a compact manifold.

M. (Such a system may be called a *knot* of $\mu \geqq 1$ components.) Two knots belong to the same *type* if there is an isotopy of M on itself which transforms the one knot into the other.

Knot theory really began in 1910 with the introduction [9] of *the group of the knot*. This is the fundamental group of the complement, $G = \pi(M - K) = \pi(M - V)$, where V is the union of sufficiently small open tubular neighborhoods V_i of the component knots K_i, $i = 1, \cdots, \mu$. It is known that this invariant is insufficient; examples to this effect have been given by Dehn,[6] Seifert,[6] and J. H. C. Whitehead.[6] However it is easy to replace the group by a more comprehensive invariant, *the group system of the knot*. This consists of the groups $\pi(\dot{V}_i)$ and $\pi(\bar{V}_i)$ together with a preferred generator for the infinite cyclic group $\pi(\bar{V}_i)$, $i = 1, \cdots, \mu$, the groups $\pi(M - V)$ and $\pi(M)$, and the indigenous homomorphisms. These latter may be represented by μ diagrams

$$\begin{array}{ccc} \pi(\dot{V}_i) & \longrightarrow & \pi(\bar{V}_i) \\ \downarrow & & \downarrow \\ \pi(M - V) & \longrightarrow & \pi(M) \end{array} \qquad i = 1, \cdots, \mu,$$

and there is the usual commutativity around the diagrams. Every known example of insufficiency of the group is resolved by the group system. For Dehn's example this was shown by Dehn;[6] for Whitehead's example it is quite trivial; for Seifert's example I have been able to accomplish this[7] by considering the representations of G in the symmetric group of order 5. Probably the known invariants of knot theory can all be expressed as invariants of the group system. The question of how strong an invariant the system really is may possibly depend on a resolution of the present unsatisfactory state of Dehn's lemma.

In 1928 Alexander [1] associated with each polygonal knot in 3-space an equivalence class of matrices $\{\boldsymbol{M}\}$; invariants of this class and hence of the given knot are the elementary divisors and in particular their product Δ. Alexander also showed how to derive an $\boldsymbol{M}$ from the group system; his algorithm requires one first to find a presentation of the group of a special kind called canonical. I have succeeded in finding an algorithm which is applicable to any presentation in which the number of generators is finite. This algorithm makes use of the free differential calculus,[8] and in fact it was the search for such an algorithm which led me to invent this calculus. With this "invariant" algorithm the matrices $\{\boldsymbol{M}\}$, their elementary divisors, and the Alexander polynomial Δ are defined for an arbitrary finitely generated group G and hence, in particular, for any knot in the interior of any compact 3-manifold. Δ is an element of the integral ring RB of the betti group B of G; it depends on a choice of basis for the free abelian group B and is determined up to a factor ϵ which is a unit of

[6] M. Dehn [8], H. Seifert [23], J. H. C. Whitehead [27]. In the examples of Dehn and Seifert, $\mu = 1$. In the example of Whitehead, $\mu = 2$, but his example has the additional feature that the complements not only have the same fundamental group but are even homeomorphic.

[7] Unpublished.

[8] [11] and [12]. A full account has not yet been published.

the ring RB. Because of the indeterminacy, Δ is essentially a polynomial in ν variables, where ν is the rank of the free abelian group B. For the case of a knot in 3-space S, $\nu = \mu$ and a basis $t_1, \cdots, t_\mu$ for B is uniquely determined by the group system.

Alexander conjectured and Seifert proved [22] that the polynomial of a tame connected knot in S is always symmetric: $\Delta(1/t) = t^\lambda \Delta(t)$. G. Torres [25] has recently proved that the polynomial of a tame knot of any multiplicity $\mu > 1$ is symmetric: $\Delta(1/t_1, \cdots, 1/t_\mu) = (-1)^\mu t_1^{\lambda_1} \cdots t_\mu^{\lambda_\mu} \Delta(t_1, \cdots, t_\mu)$. Whether symmetry is a property of Δ for knots in other 3-manifolds than S is not at present known; it is definitely not a property of Δ for an arbitrary finitely generated group.

The trivial property $\Delta(1) = 1$ of the Alexander polynomial of a tame connected knot in S has been generalized in several directions. For tame knots in S with $\mu = 2$, Torres[9] has shown that $\Delta(t_1, 1) = ((t_1^l - 1)/(t_1 - 1))\,\Delta(t_1)$, where $\Delta(t_1)$ is the polynomial of the first component knot and l is the linking number. Corollary: $\Delta(1, 1) = l$.

The property $\Delta(1) = 1$ has also been generalized[10] to the case of a tame connected knot in the interior of a compact 3-manifold M. Let N denote the kernel of the injection homomorphism $G \to \pi(M)$. I have shown that the group $G/[N]$ is an invariant of the injection $\pi(\bar{V}) \to \pi(M)$. (This is a refinement of that part of the Alexander duality theorem which states that $H = G/[G]$ is an invariant of this injection.) From a slight but important strengthening of this theorem it follows that the residue class[1] $\bar{\Delta}$ of Δ modulo the kernel ideal of the injection homomorphism $RH(M - K) \to RH(M)$ is also an invariant of this injection (at least for polygonal K). Thus, given M, there is invariantly associated to each element of $\pi(M)$ a residue class $\bar{\Delta}$. This invariant of M is sufficiently strong to duplicate Reidemeister's [18] combinatorial classification of the lens spaces, and there is some hope that, by this new method, the classification can be made topological.

The covering spaces of spherical 3-space S branched over a tame knot K are invariants of the knot, as was pointed out by Alexander [26, p. 158] in 1920. Alexander [2], Reidemeister [19], and Seifert [21], [22] gave useful algorithms for calculating the homology groups and linking invariants for the cyclic coverings S_g, $g = 1, 2, \cdots$, branched over a connected tame knot. It was not clear to me whether the results were topologically invariant since branched covering spaces had been defined only combinatorially. I have resolved[11] this difficulty as follows:[1] A pair (X, f) consisting of a space X and a continuous mapping f of X into a space Y is called a *covering* of $f(X)$ if the collection consisting of the components of the inverse images of the open sets of Y is a basis for the open sets of X. It may be shown that every covering (X, f) of $f(X)$ can be uniquely

[9] [25]. A similar result is proved for $\mu > 2$, and several other theorems of Seifert are generalized to the cases $\mu > 1$.

[10] Unpublished. This solves a problem that I proposed at the Princeton Bicentennial Conference. See Ann. of Math. vol. 50 (1949) p. 247.

[11] Unpublished.

extended to an "optimal" covering (X^*, f^*) of $\overline{f(X)}$. If we apply this to an "unbranched covering" of $M - K$ with $Y = M$, there results a unique "branched (over K) covering" of M which can be recognized intrinsically as a branched covering.[12] It is of interest to note that application to the case of a locally compact space X and the identity mapping f of X into the compact space Y which contains only one point not in X leads to the Freudenthal compactification [15]. Another application leads to Carathéodory's construction of prime ends.

The order of the homology group H_g of S_g for a tame connected knot K in S is equal [14] to the resultant of $t^g - 1$ and the Alexander polynomial $\Delta(t)$. To obtain the finer structure of H_g, one has recourse to the algorithm of Seifert [22], which involves finding the gth powers of certain $2h \times 2h$ matrices, where h is the genus of K. Using Fibonacci numbers this can be carried out explicitly[13] for $h = 1$. For the cloverleaf knot (or for any other knot of genus 1 with the same polynomial $1 - t + t^2$), the homology group H_g depends only on the residue class of g mod 6; S_g is a Poincaré space for $g \equiv \pm 1 \pmod 6$. Using p-adic numbers, Artin proved that among the cyclic coverings S_g $(g > 1)$ of knots of genus 1 these are the only Poincaré spaces.

The linking invariants of a closed oriented 3-manifold M were defined by Seifert [23]. His definition depends on a particular choice of basis for the torsion group of M. R. C. Blanchfield [4] developed an invariant definition for linking invariants (differing slightly from those of Seifert). Given a matrix $\boldsymbol{F}$ of boundary relations and a corresponding matrix $\boldsymbol{S}$ of intersection numbers these invariants are easily calculated [5] from certain "hybrid" determinants $(\boldsymbol{F} \,\backslash\!\!\!\backslash\, \boldsymbol{S})$ made up of some columns of $\boldsymbol{F}$ and the remaining columns from $\boldsymbol{S}$. These linking invariants have also been calculated explicitly for the knots of genus one.[13]

In euclidean 3-space the *total curvature* of a closed curve C of class C'' is the integral $\kappa(C) = \int_C |\, x''(s)\,|\, ds$ (s = arc length). This definition can be extended to arbitrary closed curves by defining, for any closed polygon P, $\kappa(P)$ to be the sum of the exterior angles of P, and, for any closed curve C, $\kappa(C)$ to be the least upper bound of $\kappa(P)$, P ranging over the polygons inscribed in C. J. W. Milnor [17] defines the *curvature* $\kappa(\mathfrak{C})$ of a type $\mathfrak{C}$ of knot to be the greatest lower bound of $\kappa(C)$, C ranging over $\mathfrak{C}$. He proved [17] that (I) $\kappa(\mathfrak{C}) < \infty$ if and only if $\mathfrak{C}$ is tame; (II) for tame $\mathfrak{C}$, $\kappa(\mathfrak{C})/2\pi$ is a positive integer; (III) for tame $\mathfrak{C}$ and $C \in \mathfrak{C}$, $\kappa(C) > \kappa(\mathfrak{C})$; (IV) $\mathfrak{C}$ is unknotted if and only if $\kappa(\mathfrak{C})/2\pi = 1$.[14] Using the number of elementary divisors of the Alexander matrix $\boldsymbol{M}$, I have shown [13] that there exist types $\mathfrak{C}$ for which $\kappa(\mathfrak{C})/2\pi$ is any prescribed positive integer. Using Morse's critical point theory, Milnor has generalized [16] some of these results to higher dimensions.

References

1. J. W. Alexander, *Topological invariants of knots and links*, Trans. Amer. Math. Soc. vol. 30 (1928) pp. 275–306.

[12] Address: *Covering spaces*. AMS, New York, October 29, 1949.

[13] Unpublished.

[14] This settles a conjecture of Borsuk.

2. J. W. Alexander and G. Briggs, *On types of knotted curves*, Ann. of Math. vol. 28 (1927) pp. 563–586.

3. E. Artin and R. Fox, *Some wild cells and spheres in three-dimensional space*, Ann. of Math. vol. 49 (1948) pp. 979–989.

4. R. C. Blanchfield, *The linking invariants*, Junior paper, Princeton, 1949.

5. R. C. Blanchfield and R. H. Fox, *Invariants of self-linking*, to appear in Ann. of Math.

6. W. A. Blankinship, *Generalization of a construction of Antoine*, (Doctoral thesis, Princeton University, 1949) Ann. of Math. vol. 53 (1951) pp. 276–297.

7. W. A. Blankinship and R. H. Fox, *Remarks on certain pathological open subsets of 3-space and their fundamental groups*, Proceedings of Amer. Math. Soc. vol. 1 (1950) pp. 618–624.

8. M. Dehn, *Die beiden Kleeblattschlingen*, Math. Ann. vol. 75 (1914) pp. 402–413.

9. ———, *Über die Topologie des dreidimensionalen Raumes*, Math. Ann. vol. 69 (1910) pp. 137–168.

10. R. H. Fox, *A remarkable simple closed curve*, Ann. of Math. vol. 50 (1949) pp. 264–265.

11. ———, *Free differential calculus and free groups*, Bull. Amer. Math. Soc. Abstract 55-5-228.

12. ———, *Free differential calculus and the Jacobian matrices of a group*, Bull. Amer. Math. Soc. Abstract 55-3-115.

13. ———, *On the total curvature of some tame knots*, Ann. of Math. vol. 52 (1950) pp. 258–260.

14. ———, *The order of the homology group of a cyclic covering of a knot*, Bull. Amer. Math. Soc. Abstract 55-7-395.

15. H. Freudenthal, *Über die Enden topologischer Raüme und Gruppen*, Math. Zeit. vol. 33 (1931) pp. 692–713.

16. J. W. Milnor, *On a relation between Betti numbers and Gaussian curvature*, Junior paper, Princeton, 1950.

17. ———, *On the total curvature of knots*, Ann. of Math. vol. 52 (1950) pp. 248–257.

18. K. Reidemeister, *Homotopieringe und Linsenräume*, Abh. Math. Sem. Hamburgischen Univ. vol. 11 (1935) pp. 102–109.

19. ———, *Knoten und Gruppen*, Abh. Math. Sem. Hamburgischen Univ. vol. 5 (1926) pp. 7–23.

20. ———, *Über Identitäten von Relationen*, Abh. Math. Sem. Hamburgischen Univ. vol. 16 (1949) pp. 114–118.

21. H. Seifert, *Die Verschlingungsinvarianten der zyklischen Knotenüberlagerung*, Abh. Math. Sem. Hamburgischen Univ. vol. 11 (1935) pp. 84–101.

22. ———, *Über das Geschlecht von Knoten*, Math. Ann. vol. 110 (1934) pp. 571–592.

23. ———, *Verschlingungsinvarianten*, Preuss. Akad. Wiss. Sitzungsber. vol. 26 (1933) pp. 811–828.

24. H. Seifert and W. Threlfall, *Old and new results on knots*, Canadian Journal of Mathematics vol. 2 (1950) pp. 1–15.

25. G. Torres, *Some properties of the Alexander polynomial*, Doctoral thesis, Princeton University, 1950.

26. O. Veblen, *Analysis situs*, Amer. Math. Soc. Colloquium Publications vol. 5_2, New York, 1931.

27. J. H. C. Whitehead, *On doubled knots*, J. London Math. Soc. vol. 12 (1937) pp. 63–71.

Princeton University,
Princeton, N. J., U. S. A.

TABLE OF CONTENTS

VOLUME II

INDEX OF AUTHORS

VOLUME II