Proceedings of the International
Congress of Mathematicians
August 16-24 1983
Warszawa

Proceedings of the International Congress of Mathematicians August 16-24 1983 Warszawa

Volume 2

PWN – Polish Scientific Publishers
Warszawa
North - Holland
Amsterdam • New York • Oxford
1984

Editors: Zbigniew Ciesielski and Czesław Olech

Cover design by Zygmunt Ziemka

Conference sign: Stefan Nargiełło

Library of Congress Cataloging in Publication Data

International Congress of Mathematicians (1983: Warsaw, Poland)
Proceedings of the International Congress of Mathematicians, August 16–24, 1983, Warszawa.

Papers in English and Russian.
Includes index.
1. Mathematics — Congresses. I. Olech, Czesław.
II. Ciesielski, Zbigniew, 1934– . III. Title.
QA1.I82 1983 510 84-13788
ISBN 0-444-86659-0 (v. 1)
ISBN 0-444-86660-4 (v. 2)
ISBN 0-444-86661-2 (set)

Published by
PWN — POLISH SCIENTIFIC PUBLISHERS — Warszawa
and
ELSEVIER SCIENCE PUBLISHERS B.V.
P.O. Box 1991, 1000 BZ Amsterdam
The Netherlands

Sole distributors for
the USA and Canada
ELSEVIER SCIENCE PUBLISHING COMPANY, Inc.
52 Vanderbilt Avenue
New York, NY 10017
USA

Albania, Bulgaria, Cuba, Czechoslovakia, Democratic People's Republic of Korea, German Democratic Republic, Hungary, Mongolia, People's Republic of China, Poland, Romania, the USSR, Vietnam and Yugoslavia
ARS POLONA
Krakowskie Przedmieście 7, 00-068 Warszawa

Printed in Poland by WDN

Contents

Volume 1

Section 6. Algebraic geometry

Section 7. Complex analysis

Volume 2

Section 8. Lie groups and representations

Section 9. Real and functional analysis

Section 10. Probability and mathematical statistics

Section 11. Partial differential equations

Section 12. Ordinary differential equations and dynamical systems

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

JAMES ARTHUR

The Trace Formula for Noncompact Quotient

1. In [12] and [13] Selberg introduced a trace formula for a compact, locally symmetric space of negative curvature. There is a natural algebra of operators on any such space which commute with the Laplacian. The Selberg trace formula gives the trace of these operators. Selberg also pointed out the importance of deriving such a formula when the symmetric space is assumed only to have finite volume. Then the Laplace operator will have continuous as well as discrete spectrum; it is the trace of the restriction of the operator to the discrete spectrum that is sought. Selberg gave such a formula for the quotient of the upper half plane by $\mathrm{SL}(2, \boldsymbol{Z})$. (See also [6] and [8].) Selberg also suggested how to extend the formula to any locally symmetric space of rank 1. Spaces of rank 1 are the easiest noncompact ones to handle for they can be compactified in a natural way by adding a finite number of points. I have recently obtained a trace formula for spaces of higher rank. In this article I shall illustrate the formula by looking at a typical example.

2. Let $\tilde{X}$ be the space of n by n symmetric positive definite matrices of determinant 1. The group $G = \mathrm{SL}(n, \boldsymbol{R})$ acts transitively on $\tilde{X}$ as isometries by

$$g: p \to gp^t g, \quad p \in \tilde{X}, \quad g \in G.$$

Since the isotropy subgroup of the identity matrix is $K = \mathrm{SO}(n, \boldsymbol{R})$, we can identify $\tilde{X}$ with the space of cosets G/K. Suppose that Γ is a discrete subgroup of G. Then the locally symmetric space

$$X = \Gamma \backslash \tilde{X}$$

can be identified with the space $\Gamma \backslash G/K$ of double cosets. We are interested in the spectrum of the Laplacian on $L^2(X)$. Let $\mathscr{H}_K$ be the space of smooth, compactly supported functions on G which are left and right

invariant under K. It is a commutative algebra under convolution,

$$(f_1 * f_2)(u) = \int_G f_1(y) f_2(y^{-1}u)\, dy, \quad u \in G.$$

For any $f \in \mathscr{H}_K$, define the operator $R(f)$ on $L^2(\Gamma\backslash G/K)$ by

$$\big(R(f)\phi\big)(x) = \int_G f(y)\phi(xy)\, dy, \quad \phi \in L^2(\Gamma\backslash G/K).$$

This gives a homomorphism of the algebra $\mathscr{H}_K$ into the algebra of bounded operators on $L^2(\Gamma\backslash G/K)$. The corresponding representation of $\mathscr{H}_K$ on $L^2(X)$ commutes with the Laplacian. Since the Laplacian can be approximated by operators $R(f)$, the problem of the spectral decomposition of the Laplacian on $L^2(X)$ is included in that of the spectral decomposition of $\mathscr{H}_K$ on $L^2(\Gamma\backslash G/K)$.

Suppose that $f \in \mathscr{H}_K$ and $\phi \in L^2(\Gamma\backslash G/K)$. Then

$$\begin{aligned}\big(R(f)\phi\big)(x) &= \int_G f(y)\phi(xy)\, dy = \int_G f(x^{-1}y)\phi(y)\, dy \\ &= \int_{\Gamma\backslash G} \sum_{\gamma\in\Gamma} f(x^{-1}\gamma y)\phi(y)\, dy,\end{aligned}$$

since G is unimodular and ϕ is left Γ invariant. Thus, $R(f)$ is an integral operator with a smooth kernel

$$K(x, y) = \sum_{\gamma\in\Gamma} f(x^{-1}\gamma y), \quad x, y \in \Gamma\backslash G.$$

If $\Gamma\backslash G$ is compact, the trace of the operator will be obtained by integrating the kernel over the diagonal

$$\int_{\Gamma\backslash G} \sum_{\gamma\in\Gamma} f(x^{-1}\gamma x)\, dx = \operatorname{tr} R(f).$$

Selberg's formula is obtained by grouping together those elements in Γ with the same eigenvalues and taking the integral separately of each such term. The result is a sum of G-invariant integrals over semisimple conjugacy classes of G.

3. From now on, we will take Γ to be the discrete subgroup $\mathrm{SL}(n, \boldsymbol{Z})$ of G. Then $\Gamma\backslash G$ has finite invariant volume, but is no longer compact. The integral of $K(x, y)$ over the diagonal does not converge.

However, it is possible to modify $K(x, x)$ by some functions on $\Gamma\backslash G/K$ which are supported near infinity and which reflect the various directions

in which the integral can diverge. The functions are parametrized by the standard parabolic subgroups

$$P_\pi = N_\pi M_\pi,$$

indexed by partitions

$$\pi = (n_1, \ldots, n_r), \quad n_1 + \ldots + n_r = n,$$

of n. The group M_π is the intersection of G with

$$\mathrm{GL}(n_1, \boldsymbol{R}) \times \ldots \times \mathrm{GL}(n_r, \boldsymbol{R}),$$

embedded diagonally in $\mathrm{GL}(n, \boldsymbol{R})$, while N_π is the group of matrices which differ from the identity by a matrix with entries only above the diagonal blocks of M_π. It is easy to show (using a variant of Gram–Schmidt orthogonalization, for example) that any $x \in G$ can be decomposed as

$$x = nmk,$$

with $k \in K$, $n \in N_\pi$ and the element

$$m = m_1 \ldots m_r, \quad m_i \in \mathrm{GL}(n_i, \boldsymbol{R}),$$

belonging to M_π. The decomposition is not unique, but the vector

$$H_\pi(x) = (\log|\det m_1|, \ldots, \log|\det m_r|),$$

which lies in the vector space

$$\mathfrak{a}_\pi = \left\{(u_1, \ldots, u_r) \in \boldsymbol{R}^r : \sum u_i = 0\right\},$$

is uniquely determined by x. Note that if π_0 is the partition $(1, \ldots, 1)$ corresponding to the minimal parabolic subgroup, there is a natural projection

$$(t_1, \ldots, t_n) = T \to T_\pi = (t_1 + \ldots + t_{n_1}, t_{n_1+1} + \ldots + t_{n_2}, \ldots.)$$

of $\mathfrak{a}_{\pi_0}$ onto $\mathfrak{a}_\pi$ such that $\left(H_{\pi_0}(x)\right)_\pi = H_\pi(x)$.

The modified kernel depends on a truncation parameter

$$T = (t_1, \ldots, t_n)$$

in $\mathfrak{a}_{\pi_0}$ such $t_j - t_{j+1}$ is large for each j. For any partition π let $\hat{\tau}_\pi$ be the characteristic function of the set of vectors $(u_1, \ldots, u_r)$ in $\mathfrak{a}_\pi$ such that

$$u_1 + \ldots + u_i > u_{i+1} + \ldots + u_r$$

for each $i = 1, \ldots, r-1$. The modified kernel is

$$\sum_{\pi} (-1)^{|\pi|+1} \sum_{\delta \in P_\pi \cap \Gamma \backslash G} \int_{N_\pi} \sum_{\gamma \in M_\pi} f(x^{-1}\delta^{-1}\gamma n \delta x) \hat{\tau}_\pi(H_\pi(\delta x) - T_\pi)\, dn,$$

where $|\pi| = r$ denotes the length of π. Note that if $\pi = (n)$, so that $M_\pi = G$, the function $\hat{\tau}_\pi$ is identically one and the group N_π is trivial. The corresponding summand is $K(x, x)$ itself. The other summands, as functions of x, are defined on $\Gamma \backslash G$ and are supported only near infinity.

Let $\mathcal{O}$ be the set of equivalence classes in $\Gamma = \mathrm{SL}(n, \mathbf{Z})$ of matrices with the same (complex) eigenvalues. The modified kernel can be written

$$\sum_{o \in \mathcal{O}} k_o^T(x, f),$$

where

$$k_o^T(x, f) = \sum_{\pi} (-1)^{|\pi|+1} \sum_{\delta \in P_\pi \cap \Gamma \backslash G} \int_{N_\pi} \sum_{\gamma \in M_\pi \cap o} f(x^{-1}\delta^{-1}\gamma n \delta x) \hat{\tau}_\pi(H_\pi(\delta x) - T_\pi)\, dn.$$

As we would hope, the function $k_o^T(x, f)$ is integrable. (One actually has to prove that $\sum_o \int_{\Gamma \backslash G} |k_o^T(x, f)|\, dx$ is finite [1, Theorem 7.1].) The integral

$$\int_{\Gamma \backslash G} k_o^T(x, f)\, dx,$$

defined a priori only if $t_i - t_{i+1}$ is large for each i, turns out to be a polynomial in T [3, Proposition 2.3]. We let $J_o(f)$ denote its value at $T = 0$. The left hand side of our trace formula will be

$$\sum_{o \in \mathcal{O}} J_o(f).$$

It is a generalization of the formula for compact quotient. For if the class o intersects no proper parabolic subgroup P_π, as is always the case when the quotient is compact, there are no correction terms and $J_o(f)$ is just a G invariant integral over a semisimple conjugacy class in G. In general, though, $J_o(f)$ is more complicated. If o contains only semisimple matrices, $J_o(f)$ will still be an integral over a semisimple conjugacy class, but sometimes with respect to a measure which is not G invariant. If o contains matrices which are not semisimple, $J_o(f)$ will be a sum of integrals over several conjugacy classes.

The proof of integrability requires some knowledge of the geometry $\Gamma \backslash G$ near infinity. If C is a compact fundamental domain for $N_{\pi_0} \cap \Gamma$ in N_{π_0},

the set

$$S = C\left\{\begin{bmatrix} a_1 & & \\ & \ddots & \\ & & a_n \end{bmatrix} : a_j > 0,\ a_j/a_{j+1} > \frac{2}{\sqrt{3}}\right\}K$$

is an approximate fundamental domain for Γ in G [7]. This means that $\Gamma S = G$, but only finitely many Γ translates of S intersect S. In particular, there are $(n-1)$ independent co-ordinates which can approach infinity. One studies the function $k_o^T(x, f)$ as x approaches infinity in the direction of each partition $\pi = (n_1, \ldots, n_r)$, in the sense that if

$$x = n\begin{bmatrix} a_1 & & \\ & \ddots & \\ & & a_n \end{bmatrix}k, \qquad n \in C, \qquad k \in K,$$

the co-ordinates a_{n_i}/a_{n_i+1} are each large, but all the other co-ordinates a_j/a_{j+1} remain within a compact set.

4. The other main difficulty in the noncompact case is the existence of continuous spectrum. This means that the right hand side of the formula for compact quotient has also to be seriously modified. The continuous spectrum has been completely characterized in terms of the discrete spectrum of spaces of lower dimension. It is handled by means of Eisenstein series, whose study was begun by Selberg, and completed by Langlands [9], [11]. If π is a partition of n, let M_π^1 be the subgroup of elements

$$m = m_1 \ldots m_r, \qquad m_i \in \mathrm{GL}(n_i, \boldsymbol{R}),$$

in M_π such that $|\det m_i| = 1$ for each i; let A_π be the subgroup of elements m such that each m_i is a positive multiple of the identity matrix. Then M_π is the direct product of M_π^1 and A_π. If $K_\pi = M_\pi \cap K$, we can define a convolution algebra $\mathscr{H}_{K_\pi}$ of functions on M_π^1 exactly as above. Eisenstein series are associated to eigenfunctions of $\mathscr{H}_{K_\pi}$ in $L^2(\Gamma \cap M_\pi \backslash M_\pi^1 / K_\pi)$. Suppose that ϕ is such an eigenfunction. Set

$$\phi_\pi(x) = \phi(m),$$

for any element

$$x = man k, \qquad m \in M_\pi^1, \qquad a \in A_\pi, \qquad n \in N_\pi, \qquad k \in K.$$

If λ belongs to $\mathfrak{a}_\pi^* \otimes C$, the space of complex linear functions on $\mathfrak{a}_\pi$, the Eisenstein series is defined by

$$E(x, \phi, \lambda) = \sum_{\delta \in \Gamma \cap P_\pi \backslash \Gamma} \phi_\pi(\delta x)\, e^{(\lambda + \varrho_\pi)(H_\pi(\delta x))},$$

where ϱ_π is the linear functional which maps any vector $u = (u_1, \ldots, u_r)$ in $\mathfrak{a}_\pi$ to the dot product

$$\left(-\frac{n-1}{2}, -\frac{n-3}{2}, \ldots, \frac{n-3}{2}, \frac{n-1}{2}\right)\cdot(\underbrace{u_1, \ldots, u_1}_{n_1}, \ldots, \underbrace{u_r, \ldots, u_r}_{n_r}).$$

The Eisenstein series actually converges only for certain λ, but Langlands shows that it can be analytically continued to all λ as a meromorphic function which has no poles when λ is purely imaginary. There is a functional equation which relates $E(x, \phi, \lambda)$ to the Eisenstein series in which the co-ordinates of λ are permuted by an element w in $S_{|\pi|}$ (the symmetric group on $|\pi|$ letters). For, then

$$w\pi = (n_{w(1)}, \ldots, n_{w(r)})$$

is another partition to which one can associate an eigenfunction $w\phi$ and a linear functional $w\lambda$. One can choose an orthonormal basis $\mathscr{B}_\pi$ of the subspace of $L^2(\Gamma\cap M_\pi\backslash M_\pi^1/K_\pi)$ spanned by the eigenfunctions such that $w\mathscr{B}_\pi = \mathscr{B}_{w\pi}$ for each w, and on which the functional equations are especially simple. For any $\phi \in \mathscr{B}_\pi$ the functional equation is just

$$E(x, \phi, \lambda) = m(w, \phi, \lambda)E(x, w\phi, w\lambda),$$

with $m(w, \phi, \lambda)$ a meromorphic function of λ. When λ is purely imaginary, $m(w, \phi, \lambda)$ has absolute value 1. It can be decomposed

$$m(w, \phi, \lambda) = \prod_{\{(i,j):\, i<j,\, w(i)>w(j)\}} m_\phi(\lambda_i - \lambda_j),$$

where $\lambda = (\lambda_1, \ldots, \lambda_r)$ and $m_\phi(z)$ is a meromorphic function of one complex variable. $m_\phi(z)$ equals the classical function

$$\frac{\pi^{\frac{z}{2}}\Gamma\left(\frac{z}{2}\right)\zeta(z)}{\pi^{\frac{z+1}{2}}\Gamma\left(\frac{z+1}{2}\right)\zeta(z+1)}$$

if $\pi = \pi_0$, but is obtained from a more general L-function for arbitrary π. (See [10].)

The importance of the function $E(\cdot, \phi, \lambda)$ is, of course, that it is an eigenfunction of $\mathscr{H}_K$. Indeed, it is not difficult to see from the definition that for any $f \in \mathscr{H}_K$,

$$R(f)E(\cdot, \phi, \lambda) = E\big(\cdot, R_\pi(\hat{f}_{\pi,\lambda})\phi, \lambda\big),$$

where R_π denotes the action of $\mathscr{H}_{K_\pi}$ on $L^2(\Gamma \cap M_\pi \backslash M_\pi^1 / K_\pi)$, and

$$\hat{f}_{\pi,\lambda}(m) = \int_{N_\pi} \int_{A_\pi} f(man)\, e^{(\lambda+\varrho_\pi)(H_\pi(a))} da\, dn, \qquad m \in M_\pi^1.$$

For each λ, $f \to \hat{f}_{\pi,\lambda}$ is in fact a homomorphism from $\mathscr{H}_K$ to $\mathscr{H}_{K_\pi}$. We are assuming that ϕ is an eigenfunction of $\mathscr{H}_{K_\pi}$; that is,

$$R_\pi(g)\phi = h_\phi(g)\phi, \qquad g \in \mathscr{H}_{K_\pi},$$

for a complex valued homomorphism h_ϕ of $\mathscr{H}_{K_\pi}$. It follows that

$$h_{\phi,\lambda} : f \to h_\phi(\hat{f}_{\pi,\lambda})$$

is a complex valued homomorphism of $\mathscr{H}_K$, and

$$R(f)E(\cdot, \phi, \lambda) = h_{\phi,\lambda}(f)E(\cdot, \phi, \lambda).$$

If $\pi = (n)$, $E(\cdot, \phi, \lambda)$ is just ϕ, which by assumption is square integrable. However if $\pi \neq (n)$, $E(\cdot, \phi, \lambda)$ will not be square integrable, and so will not lie in the discrete spectrum, Suppose that

$$\pi' = (n_1', \ldots, n_r')$$

is another partition of n, which equals $w\pi$ for some permutation w. Then if λ is purely imaginary and $\phi \in \mathscr{B}_\pi$ there is an asymptotic formula

$$e^{\varrho_{\pi'}(H_{\pi'}(x))} E(x, \phi, \lambda) \sim \sum_{\{w:\ w\pi = \pi'\}} m(w, \phi, \lambda) \cdot (w\phi)_{\pi'}(x)\, e^{(w\lambda)(H_{\pi'}(x))}$$

as x approaches infinity in the direction of π'. Since the function on the right is oscillatory and not square integrable in this direction, $E(\cdot, \phi, \lambda)$ cannot be square integrable. Incidentally, from this we recognize the functions $\{m(w, \phi, \lambda)\}$ as higher dimensional analogues of the classical scattering matrix.

5. Langlands shows that as π, $\phi \in \mathscr{B}_\pi$ and $\lambda \in i\mathfrak{a}_\pi^*$ vary, the Eisenstein series exhaust the spectrum. This gives a second formula

$$\sum_\pi \frac{1}{|\pi|} \sum_{\phi \in \mathscr{B}_\pi} \int_{i\mathfrak{a}_\pi^*} h_{\phi,\lambda}(f) E(x, \phi, \lambda) \overline{E(y, \phi, \lambda)}\, d\lambda$$

for the kernel $K(x, y)$. (It is convenient to take $d\lambda$ to be the measure on $i\mathfrak{a}_\pi^*$ which is dual to the Lebesgue measure associated to the basis

$$(1, -1, 0, \ldots, 0), \ldots, (0, \ldots, 0, 1, -1)$$

of $\mathfrak{a}_\pi$.) The summand with $\pi = (n)$ is just the kernel of the restriction of $R(f)$ to the discrete spectrum.

We have already discussed how to truncate $K(x, y)$ so that it can be integrated over the diagonal. The main result of [2] is that the second formula may be truncated in an apparently different way, more suitable to calculation, without changing the integral. The resulting integral is therefore a polynomial in the variable T of truncation. From its value at $T = 0$ we would hope to extract the trace of $R(f)$ on the discrete spectrum together with some terms. The answer turns out to be simpler than one has a right to expect. We will do no more than quote it.

Consider a partition

$$\pi = (n_1, \ldots, n_r)$$

of n. Let λ be a fixed point in $i\mathfrak{a}_\pi^*$ and let

$$\xi = (\xi_1, \ldots, \xi_r), \qquad \xi_i \in i\boldsymbol{R},$$

be a variable point in $i\mathfrak{a}_\pi^*$. (The co-ordinates ξ_i of ξ are uniquely determined modulo diagonal vectors $(\xi_0, \ldots, \xi_0)$.) Suppose for the moment that ϕ is any vector in $\mathscr{B}_\pi$. It is a simple exercise to show that

$$\sum_{w \in S_r} \frac{m(w, \phi, \lambda)^{-1} m(w, \phi, \lambda + \xi)}{(\xi_{w(1)} - \xi_{w(2)}) \ldots (\xi_{w(r-1)} - \xi_{w(r)})}$$

is a regular function of $\xi \in i\mathfrak{a}_\pi^*$—despite the apparent singularities from the denominator. Let $\mu_\pi(\phi, \lambda)$ be its value at $\xi = 0$. It is an interesting rational expression in the functions $m(w, \phi, \lambda)$ and their derivatives, which reduces to a logarithmic derivative if $r = 2$. More generally, suppose that π_1 is a partition of n which is finer than π. Then $i\mathfrak{a}_\pi^*$ is naturally embedded in $i\mathfrak{a}_{\pi_1}^*$ and $S_r = S_{|\pi|}$ represents certain cosets in $S_{|\pi_1|}$ modulo the subgroup of permutations in S_{π_1} which leave $i\mathfrak{a}_\pi^*$ pointwise fixed. Consequently, the expression above makes sense if ϕ is taken to be a vector in $\mathscr{B}_{\pi_1}$. It too is regular in $\xi \in i\mathfrak{a}_\pi^*$, so we continue to denote its value at $\xi = 0$ by $\mu_\pi(\phi, \lambda)$. Given π_1, let $\mathscr{B}_{\pi_1}(\pi)$ be the set of vectors ϕ in $\mathscr{B}_{\pi_1}$ such that $w\phi = \phi$ for each w in the subgroup of $S_{|\pi_1|}$ which leaves $i\mathfrak{a}_\pi^*$ pointwise fixed. For this set to be nonempty, π_1 must necessarily be of the form

$$\Big(\underbrace{\frac{n_1}{d_1}, \ldots, \frac{n_1}{d_1}}_{n_1}, \ldots, \underbrace{\frac{n_r}{d_r}, \ldots, \frac{n_r}{d_r}}_{n_r}\Big),$$

where each d_i is a divisor of n_i.

The formula for the integral of the second truncated kernel (at $T = 0$) ends up being

$$\sum_{\pi} \sum_{\pi_1} \sum_{\phi \in \mathscr{B}_{\pi_1}(\pi)} \frac{1}{r!\,(d_1 \dots d_r)^2} \int_{i\mathfrak{a}_\pi^*} \mu_\pi(\phi, \lambda) h_{\phi,\lambda}(f)\, d\lambda$$

with the numbers $r, d_1, \dots, d_r$ related to π and π_1 as above. (This formula is a special case of the main result, Theorem 8.2, of [4].) Actually, the terms must be grouped in a certain way to ensure convergence. This is because one does not know that $R(f)$ is of trace class on the discrete spectrum. Suppose for simplicity that the complication is not present. If π_1 equals (n) so does π, and the corresponding term is just the trace of $R(f)$ on the discrete spectrum. Our final formula then expresses this trace as

$$\sum_{o \in \mathscr{O}} J_o(f) - \sum_{\pi_1, \pi, \phi} \frac{1}{r!\,(d_1 \dots d_r)^2} \int_{i\mathfrak{a}_\pi^*} \mu_\pi(\phi, \lambda) h_{\phi,\lambda}(f)\, d\lambda,$$

where the sum is over partitions π_1 and π with $\pi_1 \neq (n)$, and vectors $\phi \in \mathscr{B}_{\pi_1}(\pi)$. We reiterate that the only terms left over from the case of compact quotient correspond to classes o which meet no proper parabolic subgroup. All the other terms are peculiar to the noncompact setting.

In general, though, we do not know that $R(f)$ has a trace on the discrete spectrum. The most that can be said at present is that $R(f)$ is of trace class on the space of cusp forms, a subspace of the discrete spectrum. Grouping the terms slightly differently will then give a formula for the trace of $R(f)$ on the cusp forms.

6. The main applications of the trace formula are actually to be found in a more general situation. We change notation slightly, writing $K_{\boldsymbol{R}}$ for $\mathrm{SO}(n, \boldsymbol{R})$ and $G(\boldsymbol{R})$ for $\mathrm{SL}(n, \boldsymbol{R})$, with G now standing for the algebraic group $\mathrm{SL}(n)$. The adèle group $G(\boldsymbol{A})$ is defined as the group of elements

$$(g_{\boldsymbol{R}}, g_2, g_3, \dots, g_p, \dots),$$

with $g_{\boldsymbol{R}} \in G(\boldsymbol{R})$ and $g_p \in G(\boldsymbol{Q}_p)$ for every prime number p, so that g_p actually belongs to the compact group $K_p = G(\boldsymbol{Z}_p)$ for almost all p. It is a locally compact group in which $G(\boldsymbol{Q})$ embeds diagonally as a discrete subgroup. It is not hard to show that natural embedding of $G(\boldsymbol{R})$ into $G(\boldsymbol{A})$ induces a diffeomorphism

$$G(\boldsymbol{Z}) \backslash G(\boldsymbol{R}) / K_{\boldsymbol{R}} \xrightarrow{\simeq} G(\boldsymbol{Q}) \backslash G(\boldsymbol{A}) / K,$$

where

$$K = K_{\boldsymbol{R}} \times K_2 \times K_3 \times \ldots \times K_p \times \ldots$$

The algebra $\mathscr{H}_{K_{\boldsymbol{R}}}$, whose action on $L^2\big(G(\boldsymbol{Z})\backslash G(\boldsymbol{R})/K_{\boldsymbol{R}}\big)$ we have been looking at, is now seen to be part of a larger algebra. Let $\mathscr{H}_K$ be the space of smooth, compactly supported functions on $G(\boldsymbol{A})$ which are left and right K invariant. It is also a commutative algebra under convolution. It acts on the space $L^2\big(G(\boldsymbol{Q})\backslash G(\boldsymbol{A})/K\big)$ (and hence also on $L^2\big(G(\boldsymbol{Z})\backslash G(\boldsymbol{R})/K_{\boldsymbol{R}}\big)$ and on $L^2(X)$). Thus, by introducing the adèles, we can see that the spectral decomposition of $L^2(X)$ comes with some rich extra structure that is not apparent at first glance. Everything we have discussed above extends and we obtain a trace formula for any function in $\mathscr{H}_K$. Note that an eigenvalue of $\mathscr{H}_K$ will be a formal product

$$h = h_{\boldsymbol{R}} \cdot h_2 \cdot h_3 \ldots h_p \ldots$$

of homomorphisms. It is the relationship of these local homomorphisms with each other that is expected to carry the interesting number theoretic information.

More generally, there is no reason to ask that functions be invariant under K. The associated convolution algebra will no longer be abelian, but that does not matter. Nor does G have to be $\mathrm{SL}(n)$. It can be any reductive algebraic group over $\boldsymbol{Q}$. With arguments that follow the general pattern outlined above one can establish a trace formula for any operator $R(f)$ on $L^2\big(G(\boldsymbol{Q})\backslash G(\boldsymbol{A})\big)$, with f a smooth, compactly supported function on $G(\boldsymbol{A})$. For more details, we refer the reader to the survey article [5].

References

[1] Arthur J., A Trace Formula for Reductive Groups I, *Duke Math. J.* **45** (1978), pp. 911–952.

[2] Arthur J., A Trace Formula for Reductive Groups II, *Comp. Math.* **40** (1980), pp. 87–121.

[3] Arthur J., The Trace Formula in Invariant Form, *Ann. of Math.* **113**, (1981), pp. 1–74.

[4] Arthur J. On a Family of Distributions Obtained from Eisenstein Series II, *Amer. J. Math.* **104** (1982), pp. 1289–1336.

[5] Arthur J., The Trace Formula for Reductive Groups. In: *Journeés Automorphes*, Publ. Math. de l'Université Paris VII, pp. 1–41.

[6] Duflo M. and Labesse J. P., Sur la formule des traces de Selberg, *Ann. Scient. Ec. Norm. Sup.* **4** (1971), pp. 193–284.

[7] Godement R., Domaines fondamentaux des groupes arithmétiques, *Séminaire Bourbaki*, 321 (1966).

[8] Jacquet H. and Langlands R. P., *Automorphic Forms on* GL (2), Lecture Notes in Math., 114 (1970).
[9] Langlands R. P., Eisenstein Series, *Proc. Sympos. Pure Math.*, vol. 9, Amer. Math. Soc. (1966), pp. 235–252.
[10] Langlands R. P., *Euler Products*, Yale University Press.
[11] Langlands R. P., *On the Functional Equations Satisfied by Eisenstein Series*, Lecture Notes in Math. 544 (1976).
[12] Selberg A., Harmonic Analysis and Discontinuous Groups in Weakly Symmetric Riemannian Spaces with Applications to Dirichlet Series, *J. Indian Math. Soc.* **20** (1956), pp. 47–87.
[13] Selberg A., Discontinuous Groups and Harmonic Analysis, *Proc. Int. Cong. Math.* 1962, pp. 177–189.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF TORONTO
TORONTO, M5S 1A1, CANADA

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

Р. С. ИСМАГИЛОВ

Бесконечномерные группы и их представления

Мы изложим некоторые результаты последних лет о представлениях „больших" групп. Кратко перечислим темы, которых коснёмся ниже.

В § 1 рассматривается группа $D^0(X)$ — связная компонента единицы в группе всех диффеоморфизмов гладкого многообразия X, тождественных вне компакта (если X не компактно). Даётся описание всех унитарных неприводимых представлений, ограничение которых на подгруппу диффеоморфизмов, сосредоточенных в фиксированной координатной окрестности, имеет отличный от нуля инвариантный вектор в пространстве представления. Затем рассматривается группа $D^0(X, \omega^n)$ — связная компонента единицы в группе диффеоморфизмов с компактным носителем, сохраняющих форму объёма ω^n. Наконец, в § 1 разбирается задача об индуктивном пределе семейства групп $D^0(U, \omega^n)$, $U \subset X$, $U \simeq \boldsymbol{R}^n$ (очевидным образом вложеных в $D^0(X, \omega^n)$) относительно вложений $D^0(U, \omega^n) \to D^0(V, \omega^n)$ при $U \subset V$. Решение этой задачи приводит к интересному центральному расширению группы $D^{00}(X, \omega^n)$, порождённой диффеоморфизмами с малыми носителями.

В § 2 мы отдельно рассмотрим группу $D^0(S^1)$. Она имеет представления, не охватываемые конструкцией из § 1. Кроме того, она имеет проективные представления; последние построены Г. Сигалом путём вложения в бесконечномерную симплектическую группу; мы приводим результаты, развивающие эту тему, в частности, недавние результаты Ю. А. Неретина.

В § 3 изучаются представления группы гладких отображений многообразия в компактную группу Ли.

§ 4 посвящён бесконечномерным аналогам классических метричных групп. Мы приведём некоторые результаты Г. И. Ольшанского, содержащие, в частности, описание обширного класса бесконечномерных групп типа I и конструкцию их представлений.

Наконец, § 5 посвящён вполне несвязным „большим" группам.

Здесь приводится описание сферических представлений группы $\mathrm{SL}_2(K)$ относительно $\mathrm{SL}_2(K_0)$, где K — поле с нетривиальным неархимедовым нормированием, K_0 — кольцо целых элементов, причём поле вычетов K_0/K_1 бесконечно (здесь K_1 — максимальный идеал в K_0). Рассматриваются также группы автоморфизмов деревьев, в вершинах которых сходится бесконечно много рёбер.

§1. Представления групп диффеоморфизмов

1. Представления группы $D^0(X)$. Пусть X — связное C^∞-многообразие, $\dim X = n < \infty$. Для любого преобразования $x \to x \cdot g$, $x \in X$, назовём его носителем множество $\operatorname{supp} g = \overline{\{x: x \neq x \cdot g\}}$. Через $D^0(X)$ обозначим группу всех диффеоморфизмов класса C^∞, для которых существует такое семейство g_t, непрерывно зависящее от $t \in [0, 1]$, что $g_0 = g$, $g_1 = e$ (здесь e — тождественный диффеоморфизм) и множество $\bigcup_t \operatorname{supp} g_t$ содержится в компакте. Для любого открытого $U \subset X$, группа $D^0(U)$ вложена в $D^0(X)$ (элементы из $D^0(U)$ продолжаются на X тождественно). Рассматриваются неприводимые унитарные представления $g \to T(g)$, $g \in D^0(X)$, удовлетворяющие следующему условию:

(N) *Для некоторого $U \subset X$, $U \simeq \boldsymbol{R}^n$, ограничение представления на подгруппу $D^0(U, \omega^n)$ содержит ненулевой инвариантный вектор.*

Укажем конструкцию таких представлений; наш основной результат состоит в том, что эта конструкция содержит в себе описание всех представлений со свойством (N) (см. об этом ниже).

Пусть Ω — множество всех замкнутых подмножеств $\omega \subset X$. Для каждого замкнутого $F \subset X$ положим $\Omega_F = \{\omega: \omega \subset F\}$; через B_Ω обозначим σ-алгебру подмножеств из Ω, порождённую всеми Ω_F. Группа $D^0(X)$ действует на B_Ω. Пусть μ — некоторая мера на B_Ω, квазиинвариантная относительно этого действия. Предположим, что $D^0(X)$ действует на B_Ω эргодично, $\mu(A \triangle A \cdot g) \to 0$ при $g \to e$ для любого $A \subset \Omega$ (здесь $\triangle$ — знак симметричной разности), $\int_\Omega |d\mu(\omega \cdot g)/d\mu - 1| d\mu \to 0$ при $g \to e$.

Пусть, далее, H_0 — фиксированное гильбертово пространство, $U(H_0)$ — группа его унитарных операторов со слабой топологией. Пусть для каждого $g \in D^0(X)$ задана функция $a(g, \omega)$, $\omega \in \Omega \bmod 0$ ($\Omega \bmod 0$ означает множество Ω с выброшенным подмножеством μ — меры нуль) со значениями в $U(h_0)$, удовлетворяющая следующим условиям: 1) $a(g_1 g_2, \omega) = a(g_1, \omega) a(g_2, \omega \cdot g_1)$, $\omega \in \Omega \bmod 0$ для любых g_1, g_2 из $D^0(X)$; 2) $a(g, \omega) = E$, $\omega \in \Omega_F \bmod 0$ при $g \in D^0(X \setminus F)$; 3) лю-

бая измеримая $U(H_0)$-значная функция $b(\omega)$, $\omega \in \Omega$, удовлетворяющая для каждого $g \in D^0(X)$ условию $b(\omega \cdot g)a(g,\omega) = a(g, \omega)b(\omega)$, $\omega \in \Omega \bmod 0$, имеет вид $b(\omega) = \lambda E$, $\omega \in \Omega \bmod 0$, где $\lambda \in \boldsymbol{C}$, E — единичный оператор. Теперь строим представление. Оно действует в пространстве H_0-значных вектор-функций $L_2(\Omega, H_0)$ по формуле

$$T(g)f(\omega) = \sqrt{\frac{d\mu(\omega \cdot g)}{d\mu(\omega)}}\, a(g, \omega)f(\omega \cdot g). \tag{1}$$

Для любого замкнутого $F \subset X$ обозначим через $D^0(X \setminus F)$ подгруппу тех g, которые можно соединить с e кривой g_t, где все g_t тождественны в окрестности F.

Теорема 1 [8]. *Представление* (1) *неприводимо и удовлетворяет условию* (N). *Для любого замкнутого $F \subset X$ подпространство векторов, неподвижных относительно $D^0(X \setminus F)$, совпадает с $H_F = \{f\colon f(\omega) = 0, \omega \in \Omega_F \bmod 0\}$.*

Следующая теорема содержит основной результат.

Теорема 2 [8]. *Каждое унитарное неприводимое представление со свойством* (N) *унитарно эквивалентно представлению* (1). *Мера μ определяется представлением с точностью до эквивалентности, а функция $a(g, \omega)$ — с точностью до замены на $b(\omega)a(g, \omega)b^*(\omega; g)$, где $b\colon \Omega \to U(H_0)$ — измеримое отображение.*

Итак, описание представлений свелось к мере μ и коциклу $a(g, \omega)$, удовлетворяющим перечисленным условиям. Наиболее простой пример таков: μ — это естественная мера на множестве Ω_n, состоящем из n-точечных подмножеств $\omega \subset X$, а коцикл порождён неприводимым унитарным представлением подгруппы $\mathrm{Stab}(\omega_0) = \{g\colon \omega_0 \cdot g = \omega_0\}$, тривиальным на $D^0(X \setminus \omega_0)$ (другими словами, представление зависит только от ростка элемента $g \in \mathrm{Stab}(\omega_0)$ в точке ω_0; этим обеспечивается приведённое выше условие 2) на коцикл $a(g, \omega)$). Соответствующее представление группы $D^0(X)$ индуцировано с $\mathrm{Stab}(\omega_0)$.

Приведём пример меры, сосредоточенной на множестве подмножеств с единственной предельной точкой; для упрощения ограничимся случаем $X = S^1$. Пусть S^1 реализована в виде $\boldsymbol{R}/\boldsymbol{Z}$; через $|x|$ обозначим ближайшее расстояние от $x \in S^1$ до целой точки. Возьмём последовательность независимых случайных точек $x_k \in S^1$, $k = 1, 2, \ldots$, распределённых с плотностью $p_1(x) = 1$, $p_k(x) = \delta_k |x|^{\delta_k - 1}$, $k = 2, 3, \ldots$, где $\delta_k > 0$ и $\sum \delta_k < \infty$. Положим $y_k = x_1 + \ldots + x_k$. На множестве Y

всех последовательностей $y = (y_1, y_2, \ldots)$ возникает вероятностная мера τ. Имеем отображение $Y \to \Omega$, ставящее в соответствие каждой последовательности $y = (y_1, y_2, \ldots)$ множество $\overline{\{y_k\}}$ входящих в неё точек. Образ меры τ в Ω обозначим через μ. Оказывается, мера μ квазиинвариантна и эргодична относительно $D^0(S^1)$. Представление

$$T(g)f(\omega) = \sqrt{d\mu(\omega \cdot g)/d\mu(\omega)}\, f(\omega \cdot g), \quad f \in L_2(\Omega)$$

неприводимо и обладает свойством (N). Здесь мы ограничились тривиальным коциклом. В случае $\dim X > 1$ имеется аналогичная, но гораздо более сложная конструкция меры ([8], [7]).

Аналоги теорем 1, 2 верны и для некомпактного X. В этом случае имеется большой класс мер на B_Ω, сосредоточенных на бесконечных локально-конечных конфигурациях из X. Примером является пуассоновская мера для $\boldsymbol{R}^n$; она замечательна тем, что решает задачу о представлениях группы $D^0(\boldsymbol{R}^n)$, сферических относительно $D^0(\boldsymbol{R}^n, \omega^n)$, где $\omega^n = dx_1 \wedge \ldots \wedge dx_n$ ([10]).

2. Представления группы $D^0(X, \omega^n)$. Пусть X компактно, ω^n — форма объёма на X, v — соответствующая мера на X. Для каждого измеримого $A \subset X$ обозначим через $M(A, v)$ группу измеримых преобразований множества A, сохраняющих меру v; на $X \setminus A$ преобразования продолжаются тождественно. На $M(A, v)$ рассматривается слабая сходимость: $g_k \to g$ (слабо), если $v(B \triangle B \cdot g_k) \to 0$ для любого измеримого B. Представление $g \to T(g)$, $g \in D^0(X, \omega^n)$, назовём локально-слабо непрерывным, если его ограничение на подгруппу вида $D^0(U, \omega^n)$, где $U \subset X$, $U \simeq \boldsymbol{R}^n$, непрерывно в смысле слабой сходимости (и, следовательно, продолжается до представления группы $M(U, v)$). Цель этого пункта — описание таких представлений. Для этого понадобится вложение i: $D^0(X, \omega^n) \to G$, которое мы опишем.

Пусть p: $\tilde{X} \to X$ — универсальное накрытие, Γ — группа накрывающих преобразований, $\tilde{D}^0$ — связная компонента единицы в группе всех диффеоморфизмов многообразия $\tilde{X}$, перестановочных с Γ и сохраняющих форму $p^*(\omega^n)$, $Z = \tilde{D}^0 \cap \Gamma$. Пусть, далее, Σ — группа всех преобразований $\tilde{x} \to \tilde{x} \cdot \sigma$, $\tilde{x} \in \tilde{X}$, сохраняющих меру $p^*(v)$, перестановочных с Γ и удовлетворяющих условию $\operatorname{supp} d(\tilde{x}, \tilde{x} \cdot \sigma) < \infty$, где $d(\tilde{x}, \tilde{y})$ — расстояние на $\tilde{X}$, порождающее его топологию. Положим $G = \Sigma / Z$. Имеем очевидное вложение i: $D^0(X, \omega^n) \to G$. Нетрудно проверить, что G изоморфна (не канонически) полупрямому произведению $M(X, v) \cdot \Pi^X$, где Π — фундаментальная группа многообразия

X, Π^X — группа измеримых функций $f\colon X\to\Pi$ с конечным числом значений (как обычно, множествами меры нуль пренебрегаем). На Π^X вводится слабая сходимость.

Для любого представления $\sigma\to T(\sigma)$, $\sigma\in\Sigma$, представление $g\to(T\circ i)(g)$, $g\in D^0(X,\omega^n)$, очевидным образом локально-слабо непрерывно. Нетривиальный факт состоит в том, что при $\dim X\geqslant 5$ верно и обратное; точнее, справедлива следующая теорема.

Теорема 3 [12]. *Если* $\dim X\geqslant 5$, *то любое локально-слабо непрерывное представление группы* $D^0(X,\omega^n)$ *представимо в виде* $g\to(T\circ i)(g)$, *где* $\sigma\to T(\sigma)$ *представление группы* G.

Опишем интересный класс локально-слабо непрерывных представлений, нумеруемых ненулевыми элементами пространства $H^1(X,\boldsymbol{R})$, ([12]). Обозначим через ϱ естественное представление группы $D^0(X,\omega^n)$ в пространстве $L_2^0=\{f\in L_2(X,v),\ \int f dv=0\}$. Пусть $a\in H^1(X,\boldsymbol{R})$, $a\neq 0$, a — соответствующая замкнутая 1-форма. Для любого $g\in D^0(X,\omega^n)$ форма $a-g\cdot a$ точна, так что $a-g\cdot a=d\varphi_g$; можно считать, что $\int\varphi_g dv=0$. Ясно, что $\varphi_{g_1g_2}=\varphi_{g_1}+\varrho(g_1)\varphi_{g_2}$. Имея представление ϱ и 1-коцикл φ_g, строим обычным способом представление в фоковском пространстве $\mathrm{Exp}L_2^0$ по формуле

$$T_a(g)\exp f=e^{-(\varrho(g)f,\varphi_g)-|\varphi_g|^2/2}\exp\left(\varrho(g)f+\varphi_g\right).$$

Представление T_a неприводимо при $a\neq 0$; $T_a\sim T_\beta$ лишь при $a=\beta$ либо $a=-\beta$. Оно локально-слабо непрерывно; если $U\subset X$, $U\simeq\boldsymbol{R}^n$, $v(X\setminus U)=0$, то его ограничение на $D^0(U,\omega^n)$ разлагается по формуле $\varrho^0\oplus\varrho^1\oplus\ldots$, где ϱ^k означает симметризованную k-ую степень представления ϱ. Повидимому, любое представление с последним свойством эквивалентно одному из T_a.

Задача об индуктивном пределе семейства $D^0(U,\omega^n)$. Напомним общее определение (см. [24]). Пусть дан набор групп $\{G_t\}$ и для некоторых t, s дан гомоморфизм $F_{st}\colon G_t\to G_s$, причём $F_{su}\circ F_{ts}=F_{tu}$, если определены F_{su}, F_{ts}. Индуктивный предел семейства $\{G_t\}$ состоит из группы G и гомоморфизмов $f_s\colon G_s\to G$, удовлетворяющих условиям 1) $f_s=f_t\circ F_{st}$, если определён F_{st}; 2) если группа G' и гомоморфизмы f_t' удовлетворяют аналогичным условиям, то $f_t'=p\circ f_t$, где $p\colon G\to G'$ — гомоморфизм.

Рассмотрим семейство групп $D^0(U,\omega^n)$, $U\subset X$, $U\simeq\boldsymbol{R}^n$. Имеем очевиднос вложение $D^0(U,\omega^n)\to D^0(V,\omega^n)$ при $U\subset V$. Наша цель — описать нидуктивный предел этого семейства.

Решение этой задачи приводит к новой бесконечномерной группе $E(X, \omega^n)$, являющейся центральным расширением группы $D^{00}(X, \omega^n)$, порождённой всеми $D^0(U, \omega^n)$. Наметим её построение.

Пусть E_k пространство внешних k-форм, d_k: $E_k \to E_{k+1}$ — оператор дифференцирования. Зафиксируем точку $x_0 \in X$, её окрестности $Y_0 \subset Y_1 \subset Y_2$, диффеоморфные $\boldsymbol{R}^n$, оператор I, обратный справа к d_{n-2} и удовлетворяющий условию: $\operatorname{supp} I(\alpha) \subset Y_2$, если $\operatorname{supp}\alpha \subset Y_1$. Для каждого элемента $g \in D^{00}(X, \omega^n)$, близкого к e, зафиксируем замкнутую $(n-1)$-форму ω_g^{n-1}, совпадающую с $\omega^{n-1} - g\cdot\omega^{n-1}$ в окрестности множества $X \setminus Y_0$, где $d\omega^{n-1} = \omega^n$ при $x \neq x_0$. Положим

$$a(h, g) = (I \circ h - h \circ I)\omega_g^{n-1}.$$

Легко видеть, что форма $a(h, g)$ замкнута. Через $a(h, g)$ обозначим её класс когомологий. Существует такая окрестность единицы $V \subset D^{00}(X, \omega^n)$, что $a(g_1g_2, g_3) + a(g_1, g_2) = a(g_1, g_2g_3) + a(g_2, g_3)$ для всех g_1, g_2, g_3 из V. Мы получили локальный 2-коцикл на со значениями в $H^{n-2}(X, \boldsymbol{R})$. Положим $L = H^{n-2}(X, \boldsymbol{R})/i\left(H^{n-2}(X, \boldsymbol{Z})\right)$, где гомоморфизм i порождён вложением $\boldsymbol{Z} \to \boldsymbol{R}$. Из функции $a(h, g)$ мы получаем 2-коцикл $\hat{a}(h, g)$ со значениями в L. Множество $V \times L$ превращается в локальную группу с операцией $(g_1, l_1)(g_2, l_2) = \left(g_1g_2, l_1 + l_2 + \hat{a}(g_1, g_2)\right)$. Она допускает продолжение до центрального расширения

$$0 \to L \to E_0(X, \omega^n) \to D^{00}(X, \omega^n) \to 0,$$

тривиального над каждой подгруппой вида $D^0(U, \omega^n)$, $U \simeq \boldsymbol{R}^n$. Наконец, возьмём наибольшую связную группу $E(X, \omega^n)$, накрывающую группу $E_0(X, \omega^n)$ и такую что над каждой подгруппой $D^0(U, \omega^n)$, $U \simeq \boldsymbol{R}^n$, соответствующее расширение тривиально. Группа $E(X, \omega^n)$ и есть искомый индуктивный предел семейства $\{D^0(U, \omega^n)\}$ относительно вложений $D^0(U, \omega^n) \to D^0(V, \omega^n)$, $(U \subset V)$ ([13]).

Представляет интерес аналогичная задача для других групп диффеоморфизмов (например, для группы автоморфизмов симплектического многообразия).

§ 2. Представления группы $D^0(S^1)$

Будут рассмотрены представления, не обладающие свойством (N), а также проективные представления.

Пусть H — вещественное гильбертово пространство; через $\mathrm{GL}_0(H)$ обозначим группу всех операторов вида $U+T$, где U ортогонален,

а T — оператор Гильберта–Шмидта. Для комплексного гильбертова пространства K обозначим через $\mathrm{Sp}_0(K)$ группу операторов, сохраняющих форму $\mathrm{Im}(\ ,\ )$ и представимых в виде $U+T$, где U унитарен, T — Гильберта–Шмидта. Аналогично, но с заменой $\mathrm{Im}(\ ,\ )$ на $\mathrm{Re}(\ ,\ )$, определяется группа $O_0(K)$. Напомним известные представления этих групп.

Группа $\mathrm{GL}_0(H)$ действует в ядерном расширении $H' \supset H$ преобразованиями, сохраняющими естественную гауссову меру μ. Это позволяет построить представления в пространстве $L_2(H', \mu)$ по формуле $f(\omega) \to \left(d\mu(\omega \cdot g)/d\mu(\omega)\right)^{1/2+is}(\omega \cdot g)$.

Группа $\mathrm{Sp}_0(K)$ имеет проективное представление в пространстве $\overline{S(K)}$ — пополнении симметрической алгебры $S(K)$ по естественной гильбертовой норме. Аналогично, группа $O_0(K)$ имеет проективное представление в $\overline{\Lambda(K)}$ — пополнении внешней алгебры над K. Эти представления хорошо известны из теории вторичного квантования (см., например, [1]).

Г. Сигал построил вложение группы $D^0(S^1)$ в $\mathrm{Sp}_0(K)$, что позволило построить проективные представления группы $D^0(S^1)$. Ю. А. Неретин указал недавно серию вложений группы $D^0(S^1)$ в $\mathrm{GL}_0(H)$ и исследовал соответствующие представления ([16]). Опишем кратко эти результаты. В вещественном пространстве $C^\infty(S^1)$ (S^1 реализуем в виде $\boldsymbol{R}/2\pi\boldsymbol{Z}$) введём форму

$$\langle f_1, f_2\rangle_\lambda = \int_0^{2\pi}\int_0^{2\pi} \left|\sin\left(\frac{\alpha-\beta}{2}\right)\right|^{-\lambda} f_1(\alpha) f_2(\beta)\, d\alpha\, d\beta,$$

где $0 < \lambda < 2$, $\lambda \neq 1$ (при $\lambda > 1$ интеграл определяется посредством аналитического продолжения). Пусть H_λ — пополнение $C^\infty(S^1)$ по норме $\sqrt{\langle f, f\rangle_\lambda}$. Группа $D^0(S^1)$ действует в H_λ по формуле $T_\lambda(g)f(\alpha) = f(\alpha \cdot g)\left(d(\alpha \cdot g)/d\alpha\right)^{1-(\lambda/2)}$.

Теорема 1 (Ю. А. Неретин [16]). *Оператор $T_\lambda(g)$ принадлежит $\mathrm{GL}_0(H_\lambda)$ для любого g.*

Используя предыдущую конструкцию, получаем серию унитарных представлений группы $D^0(S^1)$ в пространстве $L_2(H'_\lambda, \mu)$, $H'_\lambda \supset H_\lambda$. Оно имеет два очевидных инвариантных подпространства, соответствующие чётным и нечётным функциям $f(\omega)$, $\omega \in H'_\lambda$. Обозначим их (ограничиваясь случаем $s = 0$) через $\mathrm{Exp}(T_\lambda)_+$ и $\mathrm{Exp}(T_\lambda)_-$. В [16] описан спектр их ограничений на подгруппу $\mathrm{PSL}_2(\boldsymbol{R})$. Из этого описания следует, что эти представления имеют неприводимые компоненты;

такова, например, циклическая оболочка вакуумного вектора $1 \in$ $\in L_2(H', \mu)$. Являются ли они неприводимыми, неизвестно.

При $\lambda = 0$ аналогом формы $\langle\ ,\ \rangle_\lambda$ является

$$\langle f_1, f_2\rangle_0 = \int\int \ln\left|\sin\frac{\alpha-\beta}{2}\right| f_1(\alpha) f_2(\beta)\, d\alpha\, d\beta.$$

В этом случае представление T_0 имеет 1-коцикл $c(g, \alpha) = (\alpha \cdot g)'_\alpha - 1$. Каждый 1-коцикл $t \cdot c(g, \alpha)$ задаёт вложение группы $D^0(S^1)$ в полупрямое произведение $\mathrm{GL}_0(H_0) \cdot H_0$. Последняя группа действует в $L_2(H'_0, \mu)$; H_0 действует в H_0 сдвигами. В итоге получаем новое семейство представлений группы $D^0(S^1)$.

Соответствующая модификация конструкции Ю. А. Неретина позволяет также вложить $D^0(S^1)$ в $\mathrm{Sp}_0(K)$ и $O_0(K)$.

§ 3. Представления групп отображений в группу Ли

Пусть K — гильбертово пространство (комплексное либо вещественное), $A(K)$ — группа всех его движений, т.е. преобразований вида $x \to Ux + h$, где U — унитарный оператор, $h \in K$. Определим в пространстве $\mathrm{Exp} K$ операторы $B(U, h)$ формулой

$$\exp x \to e^{-(Ux,h)-|h|^2/2} \exp(Ux + h).$$

Они образуют унитарное представление группы $A(K)$, если K вещественно, и проективное представление для комплексного K. Если теперь $\tilde{G}$ — некоторая группа, $\tilde{g} \to \tilde{V}(\tilde{g})$, $\tilde{g} \in \tilde{G}$, её представление в K, $\tilde{\psi}(\tilde{g})$ — коцикл, то имеем гомоморфизм $\tilde{G} \to A(K)$, заданный формулой $i(\tilde{g})x = \tilde{V}(\tilde{g})x + \tilde{\psi}(\tilde{g})$, и, тем самым, представление $\tilde{g} \to (B \circ i)(\tilde{g})$, $\tilde{g} \in \tilde{G}$.

В работе Араки [26] этим методом были построены „факторизуемые" представления группы $\tilde{G}$ измеримых отображений пространства с X с мерой μ в группу Ли G. Они задаются измеримым семейством представлений $V_x(g)$, $g \in G$, действующих в гильбертовых пространствах H_x и коциклами $\psi_x(g)$. В этом случае $K = \int \oplus H_x d\mu$, $\tilde{V}(\tilde{g}) = \{V_x(\tilde{g}(x))\}$. Эти представления были затем подробно изучены в [2], [3]; в частности, были рассмотрены условия их неприводимости.

В работах [11], [27], [28] конструкция Араки была видоизменена применительно к группе $C_0^\infty(X, U)$ гладких отображений многообразия X в компактную полупростую группу Ли, имеющих компактный носитель. В этом случае K — это пространство 1-форм на X со значениями в алгебре Ли L группы; скалярное произведение в K задаётся римановой метрикой на X и инвариантным скалярным произведением в L. $C_0^\infty(X, U)$ действует в K посредством присоединённого

действия U в L. Наконец, коцикл задаётся формулой $R_g^{-1} \circ dg$, где R_g^{-1} левый сдвиг на g^{-1}.

Неприводимость полученного представления была доказана для $\dim X \geqslant 5$ в [11], для $\dim X \geqslant 4$ — в [27], для $\dim X = 3$ — в [25]. При $\dim X = 1$ оно приводимо. Случай $\dim X = 2$ разобран лишь частично в [25]: доказана неприводимость при некотором ограничении на риманову метрику.

§ 4. Представления безконечномерных классических групп

В последнее время было обнаружено, что ряд аналогов классических групп (бесконечномерных) допускает законченную теорию представлений. Мы изложим некоторые результаты по этой тематике, принадлежащие, главным образом, Г. И. Ольшанскому ([17]–[23]).

1. Группы ранга 0. Речь идёт о группах $\mathrm{SO}(\infty) = \bigcup_{n=1}^{\infty} \mathrm{SO}(n)$, $U(\infty) = \bigcup_{n=1}^{\infty} U(n)$, $\mathrm{Sp}(\infty) = \bigcup_{n=1}^{\infty} \mathrm{Sp}(n)$. В дальнейшем они будут играть ту же роль, что компактные группы в традиционной теории.

Ограничимся (с целью сократить изложение) группой $U(\infty)$. Рассмотрим её представления, непрерывные в равномерной операторной топологии; они допускают продолжение на группу унитарных операторов вида $E+T$, где E — единичный, T - компактный оператор. Имеется естественный набор неприводимых представлений, возникающих из разложения n-ой тензорной степени тождественного представления группы $U(\infty)$ в гильбертовом пространстве $H^{\otimes n} = H \otimes \ldots \otimes H$, $H = l_2$. Они нумеруются неприводимыми представлениями симметрической группы и обозначаются через ϱ_λ, $\lambda \in \hat{S}_n$. Положим $\varrho_{\lambda,\lambda'} = \varrho_\lambda \otimes \overline{\varrho}_{\lambda'}$. Следующий результат принадлежит А. А. Кириллову.

Теорема 1 ([15]). *Каждое неприводимое унитарное представление группы $U(\infty)$, непрерывное в равномерной топологии, эквивалентно одному из $\varrho_{\lambda,\lambda'}$.*

Кроме того, любое представление разлагается в дискретную сумму неприводимых. Представления, непрерывные в равномерной топологии, называются *ручными* (термин Г. И. Ольшанского [21]).

2. Группы конечного ранга $p = 1, 2, \ldots$ Так называются группы $\mathrm{SO}_0(p, \infty)$, $U(p, \infty) = \mathrm{Sp}(p, \infty)$, определяемые как индуктивные пределы $\mathrm{SO}_0(p, \infty) = \bigcup_{n=1}^{\infty} \mathrm{SO}_0(p, n)$ и т.д.

Определение (Г. И. Ольшанский [19]). Унитарное представление группы $SO_0(p, \infty)$, $U(p, \infty)$, $Sp(p, \infty)$ называется *допустимым*, если его ограничение на подгруппу $SO(\infty)$, $U(\infty)$, $Sp(\infty)$ является ручным.

Допустимые представления полностью описаны в [18]. Основной этап в их классификации — это описание голоморфных представлений группы $\widetilde{U(p, \infty)}$ ($\widetilde{\ }$ — накрытия группы $U(p, \infty)$, универсального над подгруппой $U(p)$); представления других групп получаются отсюда путём ограничения. Любое допустимое представление порождает алгебру фон Неймана типа I ([19]); таким образом, имеем весьма замечательный класс не локально-компактных групп типа I.

3. Полугрупповой подход. То обстоятельство, что в теории представлений „больших" групп возникают представления полугрупп, было обнаружено в [6], где рассматривались матричные группы с элементами из не локально-компактных полей. В [19] была исследована роль полугрупп в теории представлений групп $SO_0(p, \infty)$, $U(p, \infty)$, $Sp(p, \infty)$. Пусть G — одна из групп, G_n — подгруппа матриц $\begin{bmatrix} 1_n & 0 \\ 0 & * \end{bmatrix}$, где $n > p$, T — унитарное представление группы G в пространстве H, $H_n(T)$ — подпространство G_n-инвариантных векторов. Предположим, что T допустимо; в этом случае множество $\bigcup_{n=1}^{\infty} H_n(T)$ плотно в H. Обозначим через P_n проектор на $H_n(T)$. Тогда функция $A(g) = P_n T(g) P_n$ определена на пространстве $\Gamma(n) = G_n \backslash G / G_n$, которое можно отождествить с полугруппой матриц порядка $n \times n$ над $F = \boldsymbol{R}, \boldsymbol{C}, \boldsymbol{H}$, не увеличивающих индефинитный скалярный квадрат $-|x_1|^2 - \ldots - |x_p|^2 + \ldots + |x_n|^2$, $x \in F^m$. Оказывается, A является представлением полугруппы $\Gamma(n)$, причём $A(s^*) = (A(s))^*$. Это обстоятельство является ключевым при изучении допустимых представлений.

4. Группы бесконечного ранга некомпактного типа. Имеется 10 таких групп; в каждой из них выделяется подгруппа ранга 0 (либо подгруппа, являющаяся произведением подгрупп ранга 0). Примером служит группа $GL(\infty, C)$ состоящая из операторов вида $U+T$, где U унитарен, а T — оператор Гильберта–Шмидта; в неё вложена подгруппа $U(\infty)$. Допустимые представления определяются также, как для групп конечного ранга. Неизвестно, принадлежат ли эти группы типу I (в смысле, разъяснённом в п. 2). В [20] построено большое семейство неприводимых допустимых представлений. Это семейство

устойчиво относительно тензорного перемножения с последующим разложением на неприводимые компоненты. В [20] даны две реализации представлений: реализация, основанная на действии группы на пространстве с гауссовой мерой, и реализация посредством вложения группы в бесконечномерную метаплектическую группу и применения представления Вейля.

5. Группы бесконечного ранга компактного типа. Каждой паре (G, K) из п. 2 ставится в соответствие пара (G_0, K), где K — группа ранга 0, либо произведение двух таких групп. Как и в п. 2, представление группы G_0 называется *допустимым*, если его ограничение на K — ручное. Группа G_0 имеет существенно больше допустимых (по отношению к K) представлений, чем ручных. Большое семейство таких представлений построено в [23]. Конструкция напоминает вторую реализацию, описанную в п. 4; наряду с представлением Вейля бесконечномерной метаплектической группы используется спинорное представление бесконечномерной спинорной группы.

6. Представления класса I и фактор-представления типа II_1. Рассмотрим пару (G, K) компактного типа, в которой $G = K \times K$. Здесь K — это одна из групп $U(\infty)$, $\mathrm{SO}(\infty)$, $\mathrm{Sp}(\infty)$. В [23] замечено, что представления класса I такой пары совпадают с фактор-представлениями типа II_1 группы G. Это простое наблюдение позволяет связать допустимые представления с теорией II_1-представлений, развиваемой в [4], [29], [30]. В [4] получена классификация II_1-представлений.

7. С каждой парой (G, K) компактного типа можно связать пару (G^X, K^X), где X — пространство Лебега, а G^X и K^X — группы измеримых отображений из X в G и K. Представления класса I таких пар (и, в частности, II_1-представления группы G^X) описаны в [22].

§ 5. Представления вполне несвязных „больших" групп

1. Группа $\mathrm{SL}_2(K)$, K не локально-компактно. Пусть K — поле с неархимедовым нормированием $|x| = 2^{-w(x)}$ $(w(xy) = w(x) + w(y), w(x+y) \geqslant \geqslant \min\{w(x), w(y)\}$, множество значений функции w совпадает с $\boldsymbol{Z})$. Положим $K_0 = \{x\colon |x| \leqslant 1\}$, $K_1 = \{x\colon |x| < 1\}$. Поле вычетов K_0/K_1 предполагается бесконечным. Функция расстояния $|x-y|$ превращает K в не локально-компактное метрическое пространство. В работе [9] описаны все (не только унитарные) представления группы $G = \mathrm{SL}_2(K)$,

которые вполне неприводимы, унитарны на подгруппе $G_0 = \mathrm{SL}_2(K_0)$ и содержат единичное представление этой подгруппы. Эти представления строятся так. Пусть P — проективная прямая над K. Превратим P в ультраметрическое (в другой терминологии — неархимедово) пространство, объявив расстоянием между точками $x = (x_1 : x_2)$, $y = (y_1 : y_2)$ из P число

$$d(x, y) = |x_1 y_2 - x_2 y_1|, \quad \|x\| \equiv \max\{|x_1|, |x_2|\} = 1, \quad \|y\| = 1.$$

Пусть L — линейная оболочка характеристических функций всевозможных шаров этого пространства; существенно отметить, что эти функции линейно независимы. Возьмём последовательность $\{b_k\}_0^\infty$, и введём в L скалярное произведение, считая, что для любых двух шаров из L их характеристические функции ортогональны и норма характеристической функции шара радиуса 2^{-k} равна b_k. Пусть $H^{\{b_k\}}$ — пополнение L по норме $\sqrt{(f, f)}$.

Группа G действует в P очевидным образом. Важное (и не очевидное) обстоятельство состоит в том, что полученное таким образом действие группы G отображает пространство L в себя. Для каждого $g \in G$ рассмотрим функцию

$$a(g, x) = \|x \cdot g\|, \quad x = (x_1 : x_2) \in P, \quad \|x\| = 1.$$

Проверяется, что $a(g, \cdot) \in L$ и $a(g_1 g_2, x) = a(g_1, x) + a(g_2, x \cdot g_1)$ при $g_1, g_2 \in G$, $x \in P$. Отсюда следует, что для любого комплексного $\lambda \neq 0$ операторы

$$\left(T_\lambda(g) f\right)(x) = \lambda^{a(g, x)} f(x \cdot g), \quad f \in L,$$

задают линейное представление в пространстве L. Удобнее рассматривать представление $V_\lambda(g) = U_\lambda T_\lambda(g) U_\lambda^{-1}$, где оператор U_λ умножает на $\lambda^{[k/2]}$ характеристическую функцию шара радиуса 2^{-k}.

Теперь сформулируем основные утверждения ([5], [6], [9]).

1. $V_\lambda(g)$ — многочлен от λ для каждого g.

2. Если $0 < m < b_k/b_{k+1} < M < \infty$, то V_λ продолжается до ограниченного оператора в $H^{\{b_k\}}$.

3. Если $0 < \lambda < 1$, то V_λ унитарно в $H^{\{b_k\}}$ при $b_{2k} = (1-\lambda)^{-1}$, $b_{2k+1} = \lambda(1-\lambda)^{-1}$.

4. Если $\lambda \neq 0$, $\lambda \neq 1$, то V_λ вполне неприводимо.

5. Любое вполне неприводимое представление класса I (относительно G_0) группы G эквивалентно по М. А. Наймарку либо одному из V_λ, $\lambda \neq 0$, $\lambda \neq 1$, либо представлению $\tilde{V}_0$, индуцированному еди-

ничным одномерным представлением подгруппы G_0. (Разумеется, мы здесь говорим только о неодномерных представлениях группы G.)

6. Сферическая функция представления имеет вид

$$\varphi_\lambda(g) = \lambda^{-w(g)}, \quad w(g) = \min w(g_{ij}).$$

2. Группа $\mathrm{GL}_n(K)$, K не локально-компактно. Пусть K то же поле, что в п. 1. Положим $G = \mathrm{GL}_n(K)$, $G_0 = \mathrm{GL}_n(K_0)$. Рассмотрим случай $n > 2$. В [6] описаны все вполне неприводимые представления класса I группы G в линейном топологическом пространстве. Соответствующие сферические функции выражаются через гомоморфизмы аддитивной полугруппы векторов $t = (t_1, \ldots, t_n)$, где t_k — целые, $t_1 \geqslant \ldots \geqslant t_n$, в полугруппу C (с операцией умножения). Здесь впервые возникли полугруппы в теории представлений класса I; как уже говорилось в § 4, впоследствии полугрупповые соображения были применены в исследовании представлений бесконечномерных классических групп.

3. Группы автоморфизмов деревьев. Деревья I_n (в каждой вершине дерева сходится n рёбер, $n < \infty$) выполняют в теории p-адических групп роль симметрических пространств. Группы автоморфизмов деревьев I_n и их представления изучались в [17]. В [21] рассмотрены представления деревьев I_α (в каждой вершине сходится α рёбер, α — бесконечный кардинал). На этой группе G_α вводится слабая топология: фундаментальную систему окрестностей единицы составляют подгруппы, оставляющие на месте конечный набор вершин. В [21] доказано, что группа G_α принадлежит типу I; описаны все её унитарные неприводимые представления.

Литература

[1] Березин Ф. А., *Метод вторичного квантования*, Наука, Москва, 1965.

[2] Вершик А. М., Гельфанд И. М., Граев М. И., Представления группы $\mathrm{SL}_2(R)$, где R — кольцо функций, *Успехи математических наук* **28**, выпуск 3 (177), (1974), стр. 3–41.

[3] Вершик А. М., Гельфанд И. М., Граев М. И., Неприводимые представления группы G^X и когомологии, *Функциональный анализ и его приложения* **8**, выпуск 4 (1974), стр. 67–69.

[4] Вершик А. М., Керов С. В., Характеры и фактор-представления бесконечномерной унитарной группы, *Доклады АН СССР* **267**, № 2 (1982), стр. 272–276.

[5] Исмагилов Р. С., О линейных представлениях групп матриц с элементами из нормированного поля, *Изв. АН СССР, Сер. матем.* **33**, № 6 (1969), стр. 1296–1323.

[6] Исмагилов Р. С., Сферические функции над нормированным полем, поле вычетов которого бесконечно, *Функциональный анализ и его приложения* **4**, № 1 (1970), стр. 42–51.

[7] Исмагилов Р. С., Об унитарных представлениях группы диффеоморфизмов окружности, *Функциональный анализ и его приложения* **5**, вып. 3 (1971), стр. 45–54.

[8] Исмагилов Р. С., Об унитарных представлениях группы диффеоморфизмов компактного многообразия, *Изв. АН СССР, Сер. матем.* **36**, № 1 (1972), стр. 180–208.

[9] Исмагилов Р. С., О представлениях $SL_2(P)$, где P не локально компактно, *Функциональный анализ и его приложения* **7**, № 4 (1973), стр. 85–86.

[10] Исмагилов Р. С., Об унитарных представлениях группы диффеоморфизмов пространства R^n, $n \geqslant 2$, *Математический сборник* **98 (140)**, № 1 (9) (1975).

[11] Исмагилов Р. С., Об унитарных представлениях группы $C^\infty(X, G)$, $G = SU_2$, *Математический сборник* **100 (142)**, № 1 (5) (1976).

[12] Исмагилов Р. С., Вложение группы диффеоморфизмов, сохраняющих меру, в полупрямое произведение и её унитарные представления, *Математический сборник* **113 (155)**, № 1 (9) (1980).

[13] Исмагилов Р. С., О группе диффеоморфизмов, сохраняющих объём, *Изв. АН СССР, Сер. матем.* **44**, № 4 (1980), стр. 831–867.

[14] Исмагилов Р. С., О представлениях группы гладких отображений отрезка в компактную группу Ли, *Функциональный анализ и его приложения* **15**, вып. 2 (1981), стр. 73–74.

[15] Кириллов А. А., Представления бесконечномерной унитарной группы, *Доклады АН СССР* **212**, № 2 (1973), стр. 288–290.

[16] Неретин Ю. А., Дополнительная серия представлений группы диффеоморфизмов окружности, *Успехи матем. наук* **37**, вып. 2 (224) (1982), стр. 213–214.

[17] Ольшанский Г. И., Классификация неприводимых представлений групп автоморфизмов деревьев Брюа–Титса, *Функц. анализ и его приложения* **11**, вып. 1 (1977), стр. 32–44.

[18] Ольшанский Г. И., Унитарные представления бесконечномерных классических групп, $U(P, \infty)$, $SO_0(P, \infty)$, $Sp(P, \infty)$ и соответствующих групп движений, *Доклады АН СССР* **233**, № 6 (1978), стр. 1295–1298.

[19] Ольшанский Г. И., то же название, *Функциональный анализ и его приложения* **12**, № 3 (1978), стр. 32–44.

[20] Ольшанский Г. И., Конструкция унитарных представлений бесконечномерных классических групп, *Доклады АН СССР* **250**, № 2 (1980), стр. 284–288.

[21] Ольшанский Г. И., Новые „большие" группы типа I, в сборнике *Современные проблемы математики* **16**, ВИНИТИ, Москва, 1980, стр. 31–52.

[22] Ольшанский Г. И., Сферические функции и характеры на группе $U(\infty)^X$, *Успехи матем. наук* **37**, № 2 (1982), стр. 217–218.

[23] Ольшанский Г. И., Унитарные представления бесконечномерных пар (G, K) и формализм Р. Хау, *Доклады АН СССР* **269**, № 1 (1983), стр. 537–541.

[24] Серр Ж.-П., Амальгамы, деревья и SL_2, Математика **18**, № 1 (1974), стр. 3–51.

[25] Albeverio S. and Høegh-Krohn R., *The Energy Representations of Sobolev–Lie Group*, preprint, Univer. Bielefeld, 1976.

[26] Araki H., Factorisable Representations of Current Algebras, *Publ. RIMS, Kyoto Univ.* **A.5**, No. 3 (1970), pp. 361–342.

[27] Gelfand I. M., Graev M. I., and Versik A. M., Representations of the Group of

Function Taking Value in a Compact Lie Group, *Compositio mathematica* **42**, Fasc. 2 (1981), pp. 217–243.

[28] Parthasarathy K. R. and Schmidt K., A New Method for Constructing Factorisable Representations for Current Groups and Current Algebras, *Commun. Math. Phys.* **51**, No. 1 (1976), p. 1.

[29] Voiculescu D., Sur les representations factoriell finis de $U(\infty)$ et autres groups, *Compt. Rend. Acad. Sci. Paris* **A 28** (1975), pp. 945–946.

[30] Voiculescu D., Representations factoriell de thype II_1 de $U(\infty)$, *J. Math. Pures et Appl.* **55** (1976), pp. 1–20.

[31] Segal G., On Representations of Some Infinite-Dimensional Groups, *Commun. on Math. Phys.* **80** (1981), pp. 301–342.

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

GEORGE LUSZTIG

Characters of Reductive Groups over Finite Fields

Let G be a connected reductive algebraic group defined over a finite field F_q and let $G(F_q)$ be the group of its F_q- rational points. We shall report here on some recent results on the irreducible (complex) representations of $G(F_q)$.

1. Special conjugacy classes

Let H be a connected reductive algebraic group over C and let W be its Weyl group. The Springer correspondence allows us to parametrize the irreducible representations E of W as $E = E_{(u,\varphi)}$ where u is a unipotent element in G (up to conjugacy) and φ is an irreducible representation of the group of components $A_H(u) = Z_H(u)/Z_H^0(u)$. (However, not all φ arise in the parametrization.) For $E = E_{(u,\varphi)}$ let a'_E be the dimension of the variety $\mathscr{B}_u$ of Borel subgroups containing u and let a_E be the integer defined by the requirement: the irreducible representation of the Hecke algebra of W, corresponding to E, has formal degree of the form $\frac{1}{n_E} X^{a_E} +$ +higher powers of X (n_E = integer, > 0). One verifies that $a_E \leqslant a'_E$ for all $E \in W^v$.

Consider, for a unipotent element $u \in H$, the set

$$\{\varphi: \text{irreducible representation of } A_H(u) \text{ such that } E = E_{(u,\varphi)} \text{ is defined and } a_E = a_{E'}\}. \tag{1.1}$$

We say that u (or its class) is *special* if the set (1.1) is non-empty; in this case, we denote by $I_H(u)$ the intersection of kernels of all representations φ in the set (1.1). It is a normal subgroup of $A_H(u)$. More generally, an element $g \in H$ (or its conjugacy class) is said to be *special* if its unipotent part g_u is special with respect to $H_1 = Z_H^0(g_s)$, where g_s is the semisimple part of g. Let $A_H(g) = Z_H(g)/Z_H^0(g)$. If g is special, then both $I_{H_1}(g_u)$

and $A_{H_1}(g_u)$ are normal subgroups of $A_H(g)$ and we set $\bar{A}_H(g) = A_H(g)/I_{H_1}(g_u)$.

(For example, if $H = \mathrm{GL}_n$, all elements of H are special. If $H = \mathrm{Sp}_{2n}$, $(n \geqslant 2)$ a transvection in H is not special.)

We shall now fix an automorphism $j: H \to H$ of finite order which leaves stable some "épinglage" of G (in the sense of Bourbaki). Let $H_{(j,q)}$ be the set of special elements $g \in H$ such that g is conjugate in H to $j(g^q)$. Let $g \in H_{(j,q)}$. For each integer n, let $\tilde{Z}^{(n)}(g) = \{x \in H: xgx^{-1} = j^n(g^{q^n})\}$, and let $\tilde{Z}(g) = \coprod_n \tilde{Z}^{(n)}(g)$. Then $\tilde{Z}(g)$ is a group with multiplication defined by $x * x' = j^{n'}(x)x'$ for $x \in \tilde{Z}^n(g)$, $x' \in \tilde{Z}^{n'}(g)$. Then $\tilde{Z}^{(0)}(g)$ is a normal subgroup of $\tilde{Z}(g)$; it may be identified with $Z_H(g)$. Thus, we have a surjective map $\tilde{Z}(g) \to \mathbf{Z}$ with kernel $Z_H(g)$. Let $\tilde{A}(g) = \tilde{Z}(g)/Z_H^0(g)$. We have then a surjective map $\tilde{A}(g) \to \mathbf{Z}$ with kernel $A_H(g)$. It is easy to see that the kernel of the natural map $A_H(g) \to \bar{A}_H(g)$ is a normal subgroup of $\tilde{A}(g)$, hence the quotient $\bar{\tilde{A}}(g)$ of $\tilde{A}(g)$ by this kernel is defined and we have a surjective map $\bar{\tilde{A}}(g) \to \mathbf{Z}$ with kernel $\bar{A}_H(g)$. Let $\bar{\tilde{A}}^{(1)}(g) \subset \bar{\tilde{A}}(g)$ be the inverse image of $1 \in \mathbf{Z}$ under this map. We now define a set $\bar{\mathcal{M}}(g)$ as follows. It consists of pairs (x, σ) where x is an element of $\bar{\tilde{A}}^{(1)}(g)$ and σ is an irreducible representation of the centralizer of x in $\bar{A}_H(g)$ which can be extended to a representation of the centralizer of x in $\bar{\tilde{A}}(g)$; these pairs are taken up to the equivalence relation defined by conjugation by $\bar{\tilde{A}}(g)$. It is clear that the set $\bar{\mathcal{M}}(g)$ depends only on the conjugacy class of g: if g' is conjugate to g, there is a canonical bijection $\bar{\mathcal{M}}(g') \xrightarrow{\sim} \bar{\mathcal{M}}(g)$.

2. Classification of irreducible representations of $G(F_q)$

We shall take H in Section 1 to be the Langlands dual of G. Thus, given a maximal torus T and a Borel subgroup $B \supset T$ in G, stable under the Frobenius map $F: G \to G$, and given a maximal torus T' in H with a Borel subgroup B' containing T', we have a definite isomorphism of lattices $\iota: \mathrm{Hom}(T, \bar{F}_q^*) \xrightarrow{\sim} \mathrm{Hom}(C^*, T')$ under which the simple roots of G become the simple coroots of H and the simple coroots of G become the simple roots of H.

Let $a: T \to T$ be the automorphism defined by $F(t) = a(t^q)$ for all $t \in T$. Let $j: H \to H$ be an automorphism of finite order H with the following properties:

(a) j leaves stable an "épinglage" attached to T', B';

(b) the automorphism of $\mathrm{Hom}(C^*, T')$ defined by $j: T' \to T'$ corresponds under ι to the automorphism of $\mathrm{Hom}(T, \bar{F}_q^*)$ defined by $\alpha: T \to T$.

One of the main results of [5] may be interpreted as giving a bijection

$$\left\{\begin{array}{l}\text{isomorphism classes of}\\ \text{irreducible representations of } G(F_q)\end{array}\right\} \leftrightarrow \coprod_{\substack{g \in H(j,q) \\ \text{up to} \\ H\text{-conjugacy}}} \overline{\mathscr{M}}(g) \qquad (2.1)$$

at least under the assumption that G has connected centre. (However, the last assumption can be dropped.) The representation of $G(F_q)$ corresponding to $(x, \sigma) \in \overline{\mathscr{M}}(g)$, has dimension $\dfrac{\dim(\sigma)}{|Z_{\bar{A}(g)}(x)|} q^{\dim(C)/2} +$ lower powers of q. Moreover, the multiplicities of irreducible representations of $G(F_q)$ in the virtual representations R_T^θ of [2] where explicitly computed. By the results of [2], this implies explicit formulas for the character of any irreducible representation of $G(F_q)$ at any semisimple element.

3. Green functions

The Green functions are the values of the characters of R_T^θ (see [2]) at unipotent elements in $G(F_q)$. It is known from [2] that, once the Green functions are known, the character of R_T^θ will be automatically known at all elements. In the case where the characteristic of F_q is large enough, Springer [8] and Kazhdan [4] were able to express the Green functions as certain trigonometrical sums on the Lie algebra. Springer [8] has rewritten these trigonometrical sums in a geometric form, in terms of a certain Weyl group action (Springer's representation) on the l-adic cohomology of the variety $\mathscr{B}_u$. Recently, Shoji [6], [7] and Beynon and Spaltenstein [1] have computed the Green functions in all cases (in large characteristic); in the case of unitary groups, Kawanaka [3] was able to compute the Green functions without restriction on characteristic.

References

[1] Beynon W. M. and Spaltenstein N., *Green Functions of Finite Chevalley Groups of Type E_n ($n = 6, 7, 8$)*, to appear.

[2] Deligne P. and Lusztig G., Representations of Reductive Groups Over Finite Fields, *Ann. of Math.* **103** (1976), pp. 103–161.

[3] Kawanaka N., *Generalized Gelfand–Graev Representations and Ennola Duality*, to appear.

[4] Kazhdan D., Proof of Springer's Hypothesis, *Israel J. Math.* **28** (1977), pp. 272–286.
[5] Lusztig G., Characters of Reductive Groups Over a Finite field, *Ann. of Math. Studies* **107**, Princeton University Press, 1984.
[6] Shoji T., On the Green Polynomials of a Chevalley Group of Type F_4 *Comm. in Alg.* **10** (5), (1982), pp. 505–543.
[7] Shoji T., *On the Green Polynomials of Classical Algebraic Groups*, preprint.
[8] Springer T. A., Trigonometric Sums, Green Functions of Finite Groups and Representations of Weyl Groups, *Invent. Math.* **36** (1976), pp. 173–207.

DEPARTMENT OF MATHEMATICS
M.I.T.
CAMBRIDGE, MA 02139 U.S.A.

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

PIERRE VAN MOERBEKE*

Algebraic Complete Integrability of Hamiltonian Systems and Kac–Moody Lie Algebras

The discovery about a hundred years ago by Poincaré that most Hamiltonian systems are not completely integrable marked the end of a long and fruitful interaction between Hamiltonian mechanics and algebraic geometry; in fact many algebraic geometrical results have their origin in problems of mechanics.

The resolution of the Korteweg–de Vries equation some 15 years ago by spectral methods has led to unexpected connections between mechanics, spectral theory, Lie groups, algebraic geometry and even differential geometry, which have provided new insights into the old mechanical problems of last century, and many new ones as well. The study of specific systems and equations have led to general schemes, mainly in the realm of Lie algebras, which manufactures lots of completely integrable Hamiltonian systems; some of them can then be recognized to be of genuine mechanical or physical significance. However, given a Hamiltonian system, it is often hard to fit it into any of those general frameworks. But, luckily, most of the problems under consideration possess the much richer structure of algebraic complete integrability which in general is more restrictive than the real analytic one commonly used, although in many examples it appears that the notions of algebraic and analytic integrability coincide.

A Hamiltonian system

$$\dot{z} = J\frac{\partial H}{\partial z}, \quad z \in \boldsymbol{R}^n, \quad J = J(z) = \begin{cases}\text{antisymmetric matrix with polynomial entries in } z, \text{ satisfying the} \\ \text{Jacobi identities}\end{cases}$$

* Support of NSF grant 8102696 is gratefully acknowledged.

with polynomial right-hand side will be called *algebraically completely integrable* (a.c.i.) when

(1) the system possesses k polynomial trivial invariants $H_1, \ldots, H_k$ such that

$$J \frac{\partial H_i}{\partial z} = 0$$

and $m \equiv (n-k)/2$ invariants $H_{k+1}, \ldots, H_{k+m}$ (polynomial in z) in involution ($\{H_i, H_j\} = 0$) having the property that for most values of $c_i \in \boldsymbol{R}$, the invariant manifolds $\bigcap_{i=1}^{k+m} \{H_i = c_i\} \cap \boldsymbol{R}^n$ are compact, connected and therefore real tori by the Arnold–Liouville theorem;

(2) moreover, the real tori are part of abelian varieties (complex algebraic tori $T = \boldsymbol{C}^m/\text{Lattice}$); in the natural coordinates $(t_1, \ldots, t_m)$ of these tori, the flows (run with complex time) defined by the vector fields generated by the invariants $H_{k+1}, \ldots, H_{k+m}$ are straight lines and the coordinates $z_i = z_i(t_1, \ldots, t_m)$ are meromorphic in $(t_1, \ldots, t_m)$.

Mumford [41] has given a more general definition which coincides with the latter one when most invariant manifolds are compact, but which includes the noncompact case as well.

Two kinds of results will now be highlighted: in the first section we present a Lie algebra theoretical scheme leading to algebraic completely integrable systems, based on the Kostant–Kirillov coadjoint action. Many old and new problems discussed in Section 2 fit into this scheme, while others do not seem to or are not known to; clearly other types of reductions could also be envisaged. Therefore in Section 3 an analytic but more systematic approach has been developed for testing algebraic complete integrability, which is quite effective in small dimensions and becomes more complicated to implement in higher dimensions; it has the advantage of leading to global results, unlike the existing criteria for *real analytic integrability*, which, at this stage, are perturbation results (cf. Melnikov's method [31]). However, when a family of Hamiltonian systems is sufficiently homogeneous, then Melnikov's method can be extended to show nonanalytic integrability for a full family of Hamiltonian systems. These recent methods due to Ziglin and Holmes and Marsden will be explained in Section 4.

This lecture will focus chiefly on compact and discrete systems; many interesting methods have been devised for nonlinear partial differential equations, noncompact systems, etc., which I shall not consider.

1. Coadjoint orbits in Kac–Moody Lie algebras

We first state a theorem which is valid for any Lie algebra:

THEOREM 1 (Adler, Kostant, Symes [1, 2, 25, 47, 48]). *Let L be a Lie algebra paired with itself via a nondegenerate* ad*-invariant bilinear form $\langle\ ,\ \rangle$, L having a vector space decomposition $L = K + N$ with K and N Lie subalgebras. Then, with respect to $\langle\ ,\ \rangle$, we have the splitting $L = L^* = K^\perp + N^\perp$ and $N^* = K^\perp$ paired with N via the induced form $\langle\langle\ ,\ \rangle\rangle$ inherits the coadjoint symplectic structure of Kostant and Kirillov; its Poisson bracket between functions H_1 and H_2 on N^* reads*

$$\{H_1, H_2\}\ (a) = \langle\langle a, [V_{N^*}H_1, V_{N^*}H_2]\rangle\rangle, \qquad a \in N^*.$$

Let $\Gamma \subset N^*$ be a manifold invariant under the coadjoint action above and let $\mathscr{H}(\Gamma)$ be the algebra of functions on a neighborhood of Γ, invariant under the coadjoint action of L (which is distinct from the N-N^* action). Then the functions H in $\mathscr{H}(\Gamma)$ lead to commuting Hamiltonian vector fields of the Lax isospectral form[1]

$$\dot{a} = [a, P_K(VH)].$$

This theorem produces Hamiltonian systems having many commuting integrals; some precise results are known for interesting classes of orbits in the case of both finite- and infinite-dimensional Lie algebras.

Paradoxically, the finite-dimensional Lie algebras usually lead to noncompact systems, and the infinite-dimensional ones to compact systems. As announced, we shall merely concentrate on the latter situation and therefore, in particular, on the Kac–Moody Lie algebras.

Any finite-dimensional Lie algebra L with bracket $[\ ,\ ]$ and Killing form $\langle\ ,\ \rangle$ leads to an infinite-dimensional formal Laurent series extension

$$\mathscr{L} = \Big\{ \sum_{i=-\infty}^{N} A_i h^i \text{ with } N \text{ arbitrary} \in \boldsymbol{Z} \text{ and } A_i \in L \Big\}$$

with bracket

$$\Big[\sum_i A_i h^i, \sum_j B_j h^j\Big] = \sum_{i,j} [A_i, B_j] h^{i+j}$$

and ad-invariant, symmetric forms

$$\Big\langle \sum_i A_i h^i, \sum_j B_j h^j \Big\rangle_k = \sum_{i+j=-k} \langle A_i, B_j \rangle$$

---

[1] P_K and P_N denote projection onto K and N.

depending on $k \in \mathbf{Z}$. The forms $\langle\ ,\ \rangle_k$ are nondegenerate if so is $\langle\ ,\ \rangle$. Let $\mathscr{L}_{p,q}(p \leqslant q)$ be the vector subspace of $\mathscr{L}$, corresponding to powers of h between p and q.

The first interesting class of problems is obtained by taking $L = \mathrm{gl}(n, \boldsymbol{R})$ and by putting the form $\langle\ ,\ \rangle_1$ on the Kac–Moody extension $\mathscr{L}$; then we have the decomposition into Lie subalgebras

$$\mathscr{L} = \mathscr{L}_{0,\infty} + \mathscr{L}_{-\infty,-1} \equiv K + N \quad \text{with } K = K^{\perp},\ N = N^{\perp} \text{ and } K = N^{*}.$$

Consider the invariant manifold Γ_m, $m \geqslant 1$ in $K = N^{*}$, defined as

$$\Gamma_m = \Big\{ A = \sum_{i=0}^{m-1} A_i h^i + a h^m \,|\, a = \mathrm{diag}(a_1, \ldots, a_n) \text{ fixed} \Big\}$$

with $\mathrm{diag}(A_{m-1}) \equiv 0$.

THEOREM 2 (Adler, van Moerbeke [2]). *The manifold Γ_m has a natural symplectic structure (usually degenerate); the functions $H = \langle f(Ah^{-j}), h^k \rangle_1$ on Γ_m for good functions f lead to a.c.i. commuting Hamiltonian systems of the form*

$$\dot{A} = [A, P_K(f'(Ah^{-j})h^{k-j})], \qquad A = \sum_{i=0}^{m-1} A_i h^i + a h^m \tag{1}$$

and their trajectories are straight line motions on the Jacobian of the curve $\mathscr{C}$ of (generic) genus $(n-1)(nm-2)/2$ defined by $Q(z, h) \equiv \det(A - zI) = 0$. The coefficients of this polynomial provide the orbit invariants of Γ_m and an independent set of integrals of the motion.

Of particular interest are the flows where $j = m$, $k = m+1$, which have the following nice form:

$$\dot{A} = [A, \mathrm{ad}_{\beta}\mathrm{ad}_{a}^{-1} A_{m-1} + \beta h] \quad \text{with } \beta_i = f'(a_i); \tag{2}$$

the flow depends on f through the relation $\beta_i = f'(a_i)$ only. Reyman and Semenov-Tian-Shansky [47] have also integrated such equations, but in the real analytical sense.

Another class is obtained by choosing any semi-simple Lie algebra L; then the Kac–Moody extension $\mathscr{L}$ equipped with the form $\langle\ ,\ \rangle = \langle\ ,\ \rangle_0$ has the natural level decomposition

$$\mathscr{L} = \sum_{i \in \mathbf{Z}} \mathscr{L}_i \quad \text{with} \quad [\mathscr{L}_i, \mathscr{L}_j] \subset \mathscr{L}_{i+j},\ [\mathscr{L}_0, \mathscr{L}_0] = 0,\ \mathscr{L}_i^{*} = \mathscr{L}_{-i}.$$

Let $B^+ = \sum_{i\geqslant 0} \mathscr{L}_i$ and $B^- = \sum_{i<0} \mathscr{L}_i$; then the product Lie algebra $\mathscr{L}\times\mathscr{L}$ has the following bracket and pairing:

$$[(l_1, l_2), (l_1', l_2')] = ([l_1, l_1'], -[l_2, l_2']),$$
$$\langle (l_1, l_2), (l_1', l_2')\rangle = \langle l_1, l_1'\rangle + \langle l_2, l_2'\rangle.$$

It admits the decomposition into $K+N$, with (P_0 denotes the projection onto $\mathscr{L}_0$)

$$K = \{(l, -l) \mid l \in \mathscr{L}\}, \qquad N = \{(l, l') \in B^-\times B^+ \mid P_0(l) = P_0(l')\},$$
$$K^\perp = \{(l, l) \mid l \in \mathscr{L}\} \simeq \mathscr{L}, \qquad N^\perp = \{(l, l') \in B^-\times B^+ \mid P_0(l+l') = 0\}.$$

Then from Theorem 1 the orbits in $N^* = K^\perp$ possess a lot of commuting Hamiltonian vector fields of the Lax form:

THEOREM 3 (van Moerbeke and Mumford [37]; Adler and van Moerbeke [2]). *The N-invariant manifolds $\Gamma_{-j,k} = \sum_{-j\leqslant i\leqslant k} \mathscr{L}_i \subseteq \mathscr{L} \simeq K^\perp$ have a natural symplectic structure and the functions $H(l_1, l_2) = f(l_1)$ on $\Gamma_{-j,k}$ lead to commuting vector fields of the Lax form*

$$\dot{l} = [l, (P^+ - \tfrac{1}{2}P_0)\nabla H],$$

with P^+ the projection onto B^+; their trajectories are straight line motions on the Jacobian of a curve defined by the characteristic polynomial of elements in $\Gamma_{-j,k}$ (thought of as functions of h).

Finally a general and effective statement on linearization was given by Griffith in a beautiful recent paper [15] summarizing the situations discussed before. It is based on the observation that the tangent space to any deformation lies in a suitable cohomology group and that on algebraic curves "higher cohomology" can always be eliminated using duality theory. Given a Lax flow

$$\dot{A} = [A, B] \text{ with } A = \sum_{i=0}^{n} A_i h^i \text{ and } B = \sum_{i=0}^{N} B_i h^i,\ A_i \text{ and } B_i \text{ matrices},$$

Griffith defines the *Laurent tail* of B as follows: differentiating with regard to t the eigenvalue problem $Av = zv$ leads to $Bv = -\dot{v} + \lambda v$ for some meromorphic function λ depending on z, h and t. Then given the curve $\mathscr{C}$ defined by $\det(A - zI) = 0$ and $p \in \mathscr{C}$, he defines

$$[\text{Laurent tail } (B)]_p = \{\text{principal part of the Laurent expansion of } \lambda \text{ at } p\}.$$

THEOREM 4 (Griffith [15]). *The Law flow above linearizes on the Jacobian of the curve $\mathscr{C}$ if and only if the following conditions hold*[2] *for all* $p \in (h)_\infty$

$$[\text{Laurent tail } (B)]_p^{\cdot} \in \text{span of } \begin{cases} [\text{Laurent tail } (B)]_p, \\ \text{Laurent tails at } p \text{ of any meromorphic} \\ \text{function } f \text{ on } \mathscr{C} \text{ such that } (f) \geqslant -N(h)_\infty. \end{cases}$$

2. Distinguished examples of flows on Kac–Moody Lie algebras

(a) Example of Γ_1 in Theorem 2. The most noted one is to take $A = X + ah$, with $X \in \mathrm{so}(n)$; then the Hamiltonian flow (2), where α_j and β_j can be taken arbitrarily, is at the 0th order in h

$$\dot{X} = [X, \lambda \circ X] \quad \text{with } (\lambda \circ X)_{ij} = \lambda_{ij} X_{ij} \text{ and } \lambda_{ij} = \frac{\beta_i - \beta_j}{\alpha_i - \alpha_j}$$

and identity at the first order in h. Such a flow stays within $\mathrm{so}(n)$; introduced by Manakov [29] and extensively studied by Mishchenko, Fomenko, and Dikii [32, 33, 11], it expresses the *Euler–Arnold* [7] *equations for the geodesic flow on* $\mathrm{SO}(n)$ for a left-invariant diagonal metric $\sum_{i<j} \lambda_{ij} X_{ij}^2$ of the special form above (this is a restriction for $\mathrm{SO}(n)$, $n > 3$). The natural phase space for this motion is an orbit defined in $\mathrm{SO}(n)$ by $[n/2]$ orbit invariants. By Theorem 2, the problem is a.c.i. and the trajectories are straight lines on $\mathrm{Jac}(\mathscr{C})$ of dimension $(n-2)(n-1)/2$ and more specifically, on the Prym variety $\mathrm{Prym}(\mathscr{C}/\mathscr{C}_0) \subset \mathrm{Jac}(\mathscr{C})$ of dimension $\big(n(n-1)/2 - [n/2]\big)/2$ induced by the natural involution $(z, h) \curvearrowright (-z, -h)$ on $\mathscr{C}$ as a result of $X \in \mathrm{so}(n)$; $\mathscr{C}_0$ is the curve obtained by identifying (z, h) with $(-z, -h)$; the functions X_{ij}^2, $X_{12}X_{13}X_{23}$, etc., are abelian functions on $\mathrm{Prym}(\mathscr{C}/\mathscr{C}_0)$, but this is not true of the X_{ij} themselves; the reason is that the complex tori obtained by intersecting the constants of the motion relate to $\mathrm{Prym}(\mathscr{C})$ by doubling some periods; see Haine [18]. Finally, this is the only set of diagonal metrics for which the geodesic flow is a.c.i., as discussed in Section 3. Using Lie algebra techniques, Thimm [50] has also established the complete integrability of geodesic flow on some homogeneous spaces, which was further generalized by Guillemin and Sternberg [16].

---

[2] $(h)_\infty$ is the divisor of poles of h.

(b) Example of Γ_2 in Theorem 2. One of the most celebrated examples here is obtained by taking $A = ah^2 - hx\wedge y - y\otimes y$ for $x, y \in \boldsymbol{R}^n$, which can also be considered as a rank 2 perturbation of the diagonal matrix a; see Moser [38, 39] and [2].

The motion (2)

$$\dot{A} = [A, \Gamma + \beta h] \quad \text{with} \quad \Gamma = \mathrm{ad}_\beta \mathrm{ad}_a^{-1}(y\wedge x),\ \beta_i = f'(a_i)$$

decomposes into

$$\dot{x} = -\Gamma x - \beta y = -\frac{\partial H_\beta}{\partial y}, \quad \dot{y} = -\Gamma y = \frac{\partial H_\beta}{\partial x}, \tag{3}$$

where $H_\beta = \frac{1}{2}\sum_i \beta_i F_i(x, y)$, with

$$F_i(x, y) = y_i^2 + \sum_{j\neq i} \frac{(x_i y_j - x_j y_i)^2}{(a_i - a_j)},$$

which for $f(z) = \ln z$ (i.e. $\beta_i = a_i^{-1}$) is Jacobi's *geodesic flow on the ellipsoid* $\sum_{i=1}^n (x_i^2/a_i) = 1$, expressing the motion of the tangent line $\{x + sy \mid s \in \boldsymbol{R}\}$ to the ellipsoid in the direction y of the geodesic; for $f(z) = z^2/2$ (i.e $\beta_i = a_i$), it is *C. Neumann's motion of a point on the sphere* S^{n-1}, $|x| = 1$ *under the influence of the force* $-ax$. From Theorem 2, both motions are straight lines on $\mathrm{Jac}(\mathscr{C})$, where $\mathscr{C}$ turns out to be hyperelliptic of genus $n-1$ (much lower than the generic one) ramified at the following $2n$ points: some point at ∞, the n points a_i and $n-1$ other points λ_i of geometrical significance, based on the observation that generically a line in $\boldsymbol{R}^n$ touches $n-1$ confocal quadrics. To be precise, the set of all common tangent lines to $n-1$ confocal quadrics $Q_{\lambda_i}(x, x) + 1 = 0$, $i = 1, \ldots, n-1$, where $Q_z(x, y) = \langle (z-a)^{-1}x, y\rangle$ can be parametrized by the quotient of the Jacobian of the hyperelliptic curve $\mathscr{C}$ above by an abelian group Γ. The group is generated by the discrete action obtained by flipping the signs of x_k and y_k and some trivial one-dimensional action. Letting $h\to 0$ in the matrx A and excising the largest eigenvalue from this matrix leads to a new isospectral symmetric matrix $L = (I - P_y)(a - x\otimes x)(I - P_y)$ and a flow

$$\dot{L} = [\mathrm{ad}_\beta \mathrm{ad}_a^{-1} x\wedge y, L],$$

where the spectrum of L is given by the $n-1$ branch points λ_i above and zero. From these considerations, it follows that the tangent line $\{x + sy \mid s \in \boldsymbol{R}\}$

to the ellipsoid remains tangent to $n-2$ other confocal quadrics and the corresponding $n-1$ eigenfunctions of L provide the orthogonal set of normals to the $n-1$ quadrics at the points of tangency, hence recovering a theorem of Chasles. The close relationship between Jacobi's and Neumann's problems, which in fact live on the same orbits, was implemented by Knörrer [24], who showed that the normal vector to the ellipsoid moves according to the Neumann problem, when the point moves according to the geodesic. These facts, as investigated also by Knörrer [23], tie up with the following result of Reid [46] and Donagi [12]: the set of all $n-1$ dimensional linear subspaces in the intersection of two quadrics

$$x_1^2 + \ldots + x_n^2 - y_1^2 - \ldots - y_{n-1}^2 = 0,$$

$$\alpha_1 x_1^2 + \ldots + \alpha_n x_n^2 - \lambda_1 y_1^2 - \ldots - \lambda_{n-1} y_{n-1}^2 = x_0^2$$

in $P'_{2n-1}(C)$ is the Jacobian of the curve $\mathscr{C}$ defined above. This is done by observing that the set of linear subspaces in the above quadrics is the same as the set of $(n-2)$-dimensional linear subspaces tangent to $n-1$ quadrics

$$(\alpha_1 - \lambda_j)x_1^2 + \ldots + (\alpha_n - \lambda_j)x_n^2 = x_0^2, \quad j = 1, \ldots, n-1,$$

which is dual to the set of tangents to the confocal quadrics. The Neumann problem is also strikingly related to the KdV equation and various other nonlinear partial differential equations; see Deift, Lund, Trubowitz [10].

Finally, the symmetric top under gravity (symmetric about an axis through the fixed point) (*Lagrange top*) evolves on an orbit of type Γ_2; there $n = 3$, $A = \Gamma + Mh + ah^2$, $\Gamma \in \mathrm{so}(3) \simeq \boldsymbol{R}^3$ is the unit vector in the direction of gravity and $M \in \mathrm{so}(3) \simeq \boldsymbol{R}^3$ is the angular momentum in body coordinates with regard to the fixed point; moreover $\alpha = (\lambda+\mu)\chi$, where $\chi \in \mathrm{so}(3) \simeq \boldsymbol{R}^3$ expresses the coordinates of the center of mass and where $(\lambda+\mu, \lambda+\mu, 2\lambda)$ is the inertia tensor in diagonalized form. The situation then leads to a linear flow on an elliptic curve; see Ratiu and van Moerbeke [45] and, for higher-dimensional generalizations, Ratiu [44].

(c) Example of $\Gamma_{-j,k}$ in Theorem 3 (see [37] and [2]). Consider the periodic infinite band matrix M of period n, having $j+k+1$ diagonals; the *spectrum* of M is defined by the points $(z, h) \in C^2$ such that $Mv(h) = zv(h)$, where $v(h) = (\ldots, h^{-1}v, v, hv, \ldots)$, $v \in C^n$. Let M_h be the square matrix obtained from M in the way explained in Figure 1 and let $\mathscr{C}$ be the curve

defined by $\det(M_h - zI) = 0$. Then

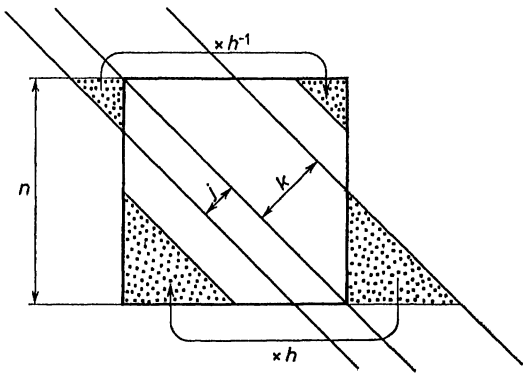


Fig. 1

$$\left\{\begin{array}{l}\text{infinite band matrices with}\\ j+k+1 \text{ diagonals of period } n\\ \text{with some regularity, with given}\\ \text{spectrum, modulo diagonal}\\ \text{periodic matrices}\end{array}\right\} = \text{Jac}(\mathscr{C})\setminus n-1 \text{ translates of the } \theta\text{-divisor.}$$

The coefficients of the polynomial $\det(M_h - zI)$ in (z, h, h^{-1}) provide the orbit invariants and a complete set of commuting vector fields. This is proved by applying Theorem 3 to $L = \mathrm{sl}(n)$ which confirms the symplectic structure first found in [37] and approached from the Lie algebra theoretical point of view in [2] and Symes [49]. Similar statements can be made for other semi-simple Lie algebras; see [2]. Of particular interest is the orbit $\Gamma_{-1,1}$ for different semi-simple Lie algebras: they lead to the motion of various periodic Toda lattices (particles connected with exponential springs), as introduced by Bogoyavlenski [8]. According to [34, 2] they can all be linearized on Jacobi varieties constructed in a similar fashion as above. One may wonder whether using different representations leads to new abelian varieties. In [2] it is shown that the higher-dimensional Jacobi varieties obtained by considering higher-dimensional representations all contain the fundamental Jacobi variety corresponding to the lowest-dimensional representation; this is done by using the theory of correspondences between curves and the results about the associated homomorphisms between their Jacobians.

The problems above also tie up with isospectral deformations of difference operators. A generic one-dimensional difference operator, given its periodic spectrum (in the sense above), can be deformed in as many ways as there are directions to go in the Jacobian of its spectral curve. In higher dimensions many partial results seem to lead to rigidity. In fact, Mumford (van Moerbeke [35]) has shown that a discrete 2-dimensional Laplacian cannot be deformed, given its periodic spectrum; the proof can be summarized by the observation that the Picard variety of most algebraic surfaces is trivial; the proof that the specific spectral surface defined by the 2-dimensional Laplacian has trivial Picard variety is based on the technique of toroidal embedding, which reduces cohomological computations to combinatorial questions.

Finally, inspired by the dynamical systems, Mumford [41] has given a beautiful description of hyperelliptic Jacobians of dimension g. Let $y^2 = R(z)$ be the monic polynomial of degree $2g+1$ defining the curve $\mathscr{C}$ and let θ be the θ-divisor. Then

$$\mathrm{Jac}(\mathscr{C}) \setminus \theta = \left\{ \begin{array}{l} \text{variety of polynomials } U, V \\ \text{with } \deg U = g, \ \deg V \leqslant g-1 \\ \text{and } U \text{ monic such that } U \mid R - V^2 \end{array} \right\}.$$

3. When is a system algebraically completely integrable?

Among the three-dimensional rigid body motions under gravity and with a fixed point, S. Kovalevski found only three a.c.i. cases: the Euler rigid body motion, the Lagrange top and the so-called "Kovalevski" top. Reminiscent of her method, we now state a necessary condition for systems to be a.c.i.:

THEOREM 5 [3, 4]. *If the Hamiltonian flow*

$$\dot{z} = J \frac{\partial H}{\partial z}, \qquad J = J(z) = \textit{polynomial in } z \in \boldsymbol{R}^n$$

is a.c.i., with generically irreducible abelian varieties, then this system of differential equations must admit Laurent expansion solutions in t such that

(1) *each z_i blows up for some value of t,*

(2) *the Laurent expansions of z_i around some places where z_i blows up admit $n-1$ parameters.*

This theorem is quite effective in pinning down the a.c.i. systems among a family of Hamiltonian systems; the existence of sufficiently

many parameters in the expansion translates into algebraic conditions on the parameters defining the family. Showing the sufficiency of the condition in Theorem 5 can be done as follows (although this has never been done rigorously in general): whenever the asymptotic solution

$$z = t^{-k}(z^{(0)} + z^{(1)}t + z^{(2)}t^2 + \ldots)$$

with $n-1$ free parameters $a_1, \ldots, a_{n-1}$ (which appear rationally), then there exist n rational functions $R_i(z)$ of z such that

$$s_i \equiv R_i(z), \quad 1 \leqslant i \leqslant n-1, \quad \tau = R_n(z)$$

admit Taylor expansions in t of the following nature

$$s_i = a_i + O(t), \quad 1 \leqslant i \leqslant n-1,$$
$$\tau = t + O(t^2);$$

with regard to these new variables $(s_1, \ldots, s_{n-1}, \tau_1)$ the differential equations (run with time τ) are nice and regular wherever the z_i blow up. This is equivalent to resolving the singularity of

$$A = \bigcap_{i=1}^{k+m} \{z \in \boldsymbol{P}^n \mid H_i = c_i z_0^2\}$$

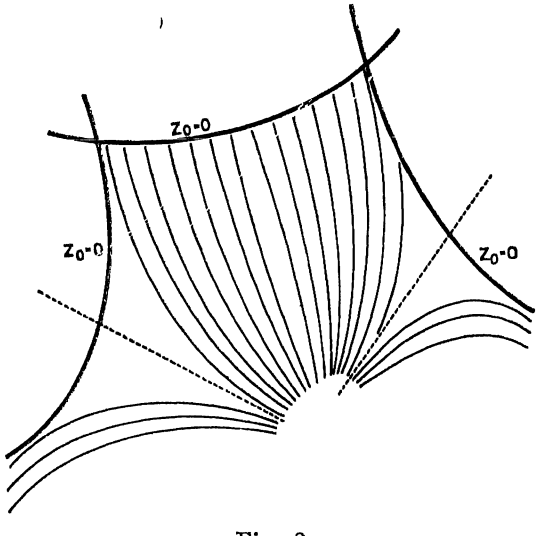


Fig. 2

along the locus $z_0 = 0$ reached by the flow. This is to say, the differential equations tell you to blow up the variety A along the part of the locus $z_0 = 0$ reached by the flow and to blow down A along the part not seen by the flow. It enables one to build and sew up the various affine pieces

defining the complex algebraic tori; this shows the algebraic integrability of the flow. This method is particularly useful whenever the flow does not have a Kac–Moody interpretation.

However, in some circumstances, the computations needed to verify condition (2) of Theorem 5 can be quite formidable and even impossible to carry out. It turns out that in some cases merely the fact that the general solutions of the differential equations are single-valued and analytic in t in the neighborhood of some special solutions suffices to single out the a.c.i. systems, as used by Haine in Theorem 8. All these ideas have now been applied to the following results.

THEOREM 6 (Adler, van Moerbeke [3]). *Consider the system of particles with non-nearest neighbor exponential interactions governed by the Hamiltonian*

$$H = \tfrac{1}{2}\sum_{i=1}^{l} y_i^2 + \sum_{i=1}^{l+1} \exp\Big(\sum_{j=1}^{l} N_{ij}x_j\Big)$$

where N_{ij} is a real $(l+1)\times l$ matrix of maximal rank. This system is algebraically completely integrable if and only if

$$\forall i \neq j \quad 2(NN^T)_{ij}(NN^T)_{ij}^{-1} \text{ is an integer} \leqslant 0,$$

i.e., if N is the Cartan matrix of a Kac–Moody Lie algebra (there are only a finite number of such matrices for each l, see Kac [22]). *Then these systems are the Toda lattices, given by the coadjoint orbit method on the corresponding Kac–Moody Lie algebra as explained in Section 2, Example* (c).

THEOREM 7 [4, 5]. *The geodesic flow on* SO(4)

$$\dot{z}' = \Big[z', \frac{\partial H}{\partial z'}\Big], \quad \dot{z}'' = \Big[z'', \frac{\partial H}{\partial z''}\Big],$$

$$z' \otimes z'' \in \mathrm{so}(3)\otimes \mathrm{so}(3) \cong \mathrm{so}(4), \quad z' = (z_1, z_2, z_3), \quad z'' = (z_4, z_5, z_6)$$

for the metric defined by the quadratic form

$$2H = \sum_{i=1}^{6} \lambda_i z_i^2 + 2\sum_{i=1}^{3} \lambda_{i,i+3} z_i z_{i+3}$$

is algebraically completely integrable if and only if one of the following conditions holds.

(1) *The quadratic form H is diagonal with regard to the customary coordinates of* so(4), *i.e.* $2H = \sum_{1\leqslant i<j\leqslant 4} \Lambda_{ij} X_{ij}^2$ *with* $\Lambda_{ij} = (\beta_i - \beta_j)/(\alpha_i - \alpha_j)$, $\beta_i, \alpha_i \in C$.

Then the system has besides two orbit invariants and the Hamiltonian $2H$ *one other quadratic invariant, which is diagonal as well. Then the flow is a straight line motion on the abelian variety* 1 *of Theorem* 9. (*A special case is the Clebsch case for the motion of a* 3*-dimensional rigid body in a fluid.*)

(2) *The quadratic form* H *satisfies the conditions (with* $\Delta_{ij} = \lambda_i - \lambda_j$*):*
$(\lambda_{14}^2, \lambda_{25}^2, \lambda_{36}^2)R^2 = \Delta_{21}\Delta_{54}\Delta_{32}\Delta_{65}\Delta_{13}\Delta_{46}\big(\Delta_{32}^{-1}\Delta_{65}^{-1}(\Delta_{25}-\Delta_{36})^2, \Delta_{13}^{-1}\Delta_{46}^{-1}(\Delta_{36}-\Delta_{14})^2, \Delta_{21}^{-1}\Delta_{54}^{-1}(\Delta_{14}-\Delta_{25})^2\big)$,
where $R = \Delta_{12}\Delta_{46} - \Delta_{13}\Delta_{45}$; *then the flow has one extra quadratic invariant and it linearizes on the abelian variety* 2 *of Theorem* 9. *The Steklov–Lyapunov case of the rigid body motion in an ideal fluid is a special case of this one.*

(3) *The quadratic form* H *satisfies the conditions:*

$$(\lambda_{14}^4, \lambda_{25}^4, \lambda_{36}^4) = \Delta_{13}\Delta_{46}\Delta_{21}\Delta_{54}\Delta_{32}\Delta_{65}(\Delta_{32}^{-1}\Delta_{65}^{-1}, \Delta_{13}^{-1}\Delta_{46}^{-1}, \Delta_{21}^{-1}\Delta_{54}^{-1})$$

with $\alpha^2 = \Delta_{46}/\Delta_{13}$, $\beta^2 = \Delta_{54}/\Delta_{21}$, $\gamma^2 = \Delta_{65}/\Delta_{32}$ *or their inverses satisfying the quadratic relations*

$$\alpha\beta+\beta\gamma+\gamma\alpha+1 = 0 \quad \text{and} \quad 3\beta\gamma+\gamma-\beta+1 = 0;$$

then the extra-invariant is quartic and the flow linearizes on an abelian variety.

Some examples of these metrics have been considered recently also by Bogoyavlenski [9].

THEOREM 8 (Haine [17]). *The geodesic flow on* $\mathrm{SO}(n)$ *for the metric defined by the diagonal quadratic form* $\sum \Lambda_{ij}X_{ij}^2$ *is algebraically completely integrable if and only if* $\Lambda_{ij} = (\beta_i - \beta_j)/(\alpha_i - \alpha_j)$, $\beta_i, \alpha_i \in \boldsymbol{C}$.

Theorem 7 is part of a wider statement of algebraic geometrical nature.

THEOREM 9 [5, 6]. *Consider four quadrics in* $\boldsymbol{C}^6$ *of the following nature (with some mild (checkable) nondegeneracy assumptions):*

$$Q_1 = \sum_{i=1}^{3} z_i^2, \quad Q_3 = \sum_{i=1}^{6} \lambda_i z_i^2 + 2\sum_{i=1}^{3} \lambda_{i,i+3} z_i z_{i+3},$$

$$Q_2 = \sum_{i=4}^{6} z_i^2, \quad Q_4 = \sum_{i=1}^{6} \mu_i z_i^2 + 2\sum_{i=1}^{3} \mu_{i,i+3} z_i z_{i+3};$$

Let $V_Q \cong \boldsymbol{P}^3(\boldsymbol{C})$ *be the linear span of the quadrics* Q_i *and let* Q_i *also denote the matrix of the quadratic form. Then the discriminant surface* Δ *has the following property:*

$$V_Q \supseteq \Delta = \{[t_1, t_2, t_3, t_4] \in \boldsymbol{P}^3(\boldsymbol{C}) \mid \det\big(\textstyle\sum t_i Q_i\big) = 0\} = K_1 \cup K_2 \cup K_3,$$

where the K_i are 3 quadratic cones. Then the following 3 statements are equivalent:

(i) *the three cones K_1, K_2 and K_3 intersect in a curve $\mathscr{C}_0'$ not in a hyperplane; this curve is then the locus of all rank 3 quadrics in V_Q;*

(ii) *the variety $\bigcap_{i=1}^{4} \{z \in \boldsymbol{P}^6 \mid Q_i = c_i z_0^2\}$ is singular along one or several curves at infinity (i.e. at $z_0 = 0$);*

(iii) *the affine variety $\bigcap_{i=1}^{4} \{z \in \boldsymbol{C}^6 \mid Q_i = c_i\}$ is equipped with two independent quadratic vector fields.*

Under any one of these 3 conditions the four quadrics intersect in the affine part of an abelian variety and $\mathscr{C}_0'$ can only be an elliptic curve or a curve isomorphic to $\boldsymbol{P}^1$.

1. *The curve $\mathscr{C}_0'$ is elliptic and nonsingular*; then $\mathscr{C}_0'$ contains 4 points $Q_1', \ldots, Q_4'$, still spanning V_Q, corresponding to quadrics which expressed in some new variables through a linear change have the following simple form:

$$
\begin{aligned}
Q_1'(x) &= \frac{x_4^2}{\alpha_1 - \alpha_4} + \frac{x_2^2}{\alpha_1 - \alpha_3} + \frac{x_3^2}{\alpha_1 - \alpha_2} = c_1' x_0^2, \\
Q_2'(x) &= \frac{x_1^2}{\alpha_2 - \alpha_3} + \frac{x_5^2}{\alpha_2 - \alpha_4} + \frac{x_3^2}{\alpha_2 - \alpha_1} = c_2' x_0^2, \\
Q_3'(x) &= \frac{x_1^2}{\alpha_3 - \alpha_2} + \frac{x_2^2}{\alpha_3 - \alpha_1} + \frac{x_6^2}{\alpha_3 - \alpha_4} = c_3' x_0^2, \\
Q_4'(x) &= x_1 x_4 + x_2 x_5 + x_3 x_6 = c_4' x_0^2 .
\end{aligned}
$$

Moreover, the singular locus at infinity of the intersection of these four quadrics is an elliptic curve $\mathscr{E}$, which is a 4-fold unramified cover of $\mathscr{C}_0'$. Let $\mathscr{C}$ be a double cover of $\mathscr{C}_0'$ ramified at the 4 points $\mathscr{C}_0' \cap \text{plane}\{\sum c_i' t_i = 0\}$, then according to Mumford [4] (see also Haine [18])

$$
\bigcap_{i=1}^{4} \{z \in \boldsymbol{C}^6 \mid Q_i = c_i\} = \operatorname{Prym}(\mathscr{C}'/\mathscr{C}_0') \setminus \left\{ \begin{array}{l} \text{a genus 9 curve, which is a ramified} \\ \text{cover of } \mathscr{E} \text{ with 16 branch points} \end{array} \right\}.
$$

This system of quadrics supports the geodesic flow equation of SO(4), which from example (a) (Section 2) linearizes on $\operatorname{Prym}(\mathscr{C}/\mathscr{C}_0)$, which is a dual abelian variety to $\operatorname{Prym}(\mathscr{C}'/\mathscr{C}_0')$ (Haine [18]); the functions X_{ij} are themselves meromorphic on $\operatorname{Prym}(\mathscr{C}_0'/\mathscr{C}_0')$, while only their squares are on $\operatorname{Prym}(\mathscr{C}/\mathscr{C}_0)$; in fact, $\operatorname{Prym}(\mathscr{C}/\mathscr{C}_0)$ is obtained from $\operatorname{Prym}(\mathscr{C}'/\mathscr{C}_0')$ by doubling half the periods (isogeny) and it is also isogeneous to the

Jacobian of a naturally arising hyperelliptic curve, as follows from Kötter's [26, 27] investigation of the Clebsch case for the motion of a rigid body in an ideal fluid.

2. *The curve $\mathscr{C}_0'$ is isomorphic to $\mathbf{P}^1$*; then $\mathscr{C}_0'$ contains 3 points Q_1', Q_2', Q_3', corresponding to rank 2 quadrics (which are also the vertices of the 3 cones K_i) and a fourth point Q_4'. They have the form:

$$\begin{aligned} Q_1' &= x_2^2 + (1-a)x_3^2, \\ Q_2' &= ax_1^2 - (1-a)x_6^2, \\ Q_3' &= ax_4^2 + x_5^2, \\ Q_4' &= (x_1-x_4)^2 + (x_2-x_5)^2 + (x_3-x_6)^2. \end{aligned}$$

In $\mathbf{P}^6$, these quadrics intersect at infinity in 8 curves isomorphic to $\mathbf{P}^1$; the variety is singular along 4 of them. Blowing up the variety along these four curves and blowing it down along the remaining four, as can be done in an elementary way by using the asymptotic solutions of the differential equations,one shows that,

$$\bigcap_{i=1}^{4} \{z \in \mathbf{C}^6 | Q_i = c_i\} = \text{Jac (hyperelliptic curve of genus 2)} \setminus \setminus \left\{\begin{array}{l}\text{4 genus 2 hyperelliptic curves pairwise trans-}\\ \text{versally intersecting in 2 points, each point}\\ \text{of intersection belonging to 3 of the 4 curves}\end{array}\right\},$$

completing the statement of Theorem 9.

4. The nonexistence of analytic integrals of Hamiltonian systems

Let $z = (x_1, x_2, I, \varphi) \in U \equiv D \times [I_0 - \varepsilon, I_0 + \varepsilon] \times S^1$, for a domain $D \subset \mathbf{R}^2$. Consider in U the Hamiltonian system

$$\dot{x}_1 = \frac{\partial H}{\partial x_2}, \quad \dot{x}_2 = -\frac{\partial H}{\partial x_1}, \quad \dot{\varphi} = \frac{\partial H}{\partial I}, \quad \dot{I} = -\frac{\partial H}{\partial \varphi},$$

where

$$H(z) = H_0(x, I) + \mu H_1(x, I, \varphi) + \ldots$$

Then I is a constant of the motion I_0 for the unperturbed system; assume that the unperturbed system has two hyperbolic fixed points in D, joined by separatrices; assume that the solution $\hat{x}(t)$ along these separatrices can be continued analytically to an appropriate complex strip $\Pi = \{0 \leqslant \operatorname{Im} t \leqslant b\}$, which contains a finite number of singular points.

THEOREM 10 (Ziglin [52]). *If the sum of the residues of* $\partial H_1/\partial\varphi\,(\hat{z}(t))$ *in* $t \in \Pi$, *where*

$$\hat{z}(t) = \left(\hat{x}(t), I_0, \int_{t_0}^{t} \frac{\partial H_0}{\partial I}(\hat{x}(\tau), I_0)\, d\tau\right),$$

is not zero, then in U *the perturbed system does not have an additional analytic constant of the motion, for any sufficiently small* $|\mu| \neq 0$.

Ziglin relates the nonvanishing of the sum of the residues above to the nonvanishing of the Melnikov integral. He uses Theorem 10 which appears as a local result to prove the global result below, by rescaling the variables so that the system can be viewed as a small perturbation of the Euler rigid body motion.

COROLLARY ([52], see also Holmes and Marsden [20]). *The motion of a nonsymmetric rigid body in the presence of gravity does not have any additional analytic integrals besides the known ones.*

The theorem does not apply whenever two principal momenta of inertia become equal, because of the fact that the hyperbolic points used before disappear. Ziglin [53] applies then successfully the following method: consider a Hamiltonian system and a special solution $X = \varphi(t)$, which can be continued analytically for $t \in C$; consider the linearized equations (normal component only) around this solution. To each meromorphic integral of the motion of the Hamiltonian system in a neighborhood of the curve $X = \varphi(t)$, there is a corresponding rational integral of the linearized equation; moreover, independence is preserved. Any closed path issuing from a given point of the Riemann surface $X = \varphi(t)$ produces a linear transformation mapping a vector into a new one obtained by solving the linearized equation around this loop. The set of such loops, and hence the corresponding maps, forms a group G of symplectic maps leaving invariant the rational integrals above.

THEOREM 11 (Ziglin [53]). *If the Hamiltonian system above is completely integrable with analytic integrals, then every transformation of the group of monodromy* G *must preserve the base point and the eigendirection of some nonresonant transformation of* G.

This theorem is applied to the symmetric rigid body in the presence of gravity around the special solution corresponding to the initial condition $(M, \Gamma) = (0, M_2, 0, \gamma_1, 0, \gamma_3)$ with M and Γ defined as in part (b) of Section 2. More precisely, we have

COROLLARY (Ziglin [53]). *The symmetric rigid body with gravity, whose center of mass belongs neither to the equatorial plane through the fixed point nor to the axis of symmetry cannot have analytic integrals besides the known ones.*

Arguing further, Ziglin finally shows that the rigid body motion can never be analytically integrable except for the three known cases: the Euler, Lagrange and Kovalevski cases. Melnikov integral arguments have also been used by V. V. Kozlov, D. A. Onishchenko and A. D. Veselov to show nonintegrability for rigid body motions in fluids.

References

[1] Adler M., On a Trace Functional for Pseudo-Differential Operators and the Symplectic Structure of the Korteweg-de Vries Equation, *Invent. Math.* **50** (1979), pp. 219–248.

[2] Adler M. and van Moerbeke P., Completely Integrable Systems, Euclidean Lie Algebras, and Curves. Linearization of Hamiltonian Systems, Jacobi Varieties, and Representation Theory, *Advances in Math.* **38** (1980), pp. 267–379.

[3] Adler M. and van Moerbeke P., Kowalewski's Asymptotic Method, Kac–Moody Lie Algebras and Regularization, *Comm. Math. Phys.* **83** (1982), pp. 83–106.

[4] Adler M. and van Moerbeke P., The Algebraic Integrability of Geodesic Flow on SO (4), *Invent. Math.* **67** (1982), pp. 297–326, with an appendix by D. Mumford.

[5] Adler M. and van Moerbeke P., Intersection of Quadrics and Hamiltonian Mechanics, *Proc. Nat. Acad. Sci. U.S.A.* (1983), to appear.

[6] Adler M. and van Moerbeke P., Geodesic flow on SO (4), to appear in Progress in Mathematics, Birkhäuser, Boston.

[7] Arnold V. I., *Mathematical Methods of Classical Mechanics*, Springer-Verlag, Berlin–Heidelberg–New York, 1978.

[8] Bogoyavlenski O. I., On Perturbations of the Periodic Toda Lattice, *Comm. Math. Phys.* **51** (1976), pp. 201–209.

[9] Bogoyavlenski O. I., Integrable Euler Equations on Six-Dimensional Lie Algebras, *Dokl. Akad. Nauk SSSR* **273** (1983), pp. 15–18.

[10] Deift P., Lund F., and Trubowitz E., Nonlinear Wave Equations and Constrained Harmonic Motion, *Comm. Math. Phys.* **74** (1980), pp. 141–188.

[11] Dikii L. A., Hamiltonian System Connected with the Rotation Group, *Funct. Anal. Appl.* **6** (1972), pp. 326–377.

[12] Donagi R., Group Law on Intersections Quadrics, *Ann. Scuola Norm. Sup. Pisa* (4) **7** (1980), pp. 217–239.

[13] Dubrovin B. A., Theta Functions and Nonlinear Equations, *Uspekhi Mat. Nauk* **36** (1981), pp. 11–80, *Russian Math. Surveys* **36** (1981), pp. 11–92.

[14] Dubrovin B. A., Matveev V. B., and Novikov S. P., Nonlinear Equations of Korteweg–de Vries Type, Finite-Zone Operators and Abelian Varieties, *Uspekhi Mat. Nauk* **31** (1976), pp. 55–136, *Russian Math. Surveys* **31** (1976), pp. 59–146.

[15] Griffith P. A., *Linearizing Flows and a Cohomological Interpretation of Lax Equations*, preprint, 1983.
[16] Guillemin V. and Sternberg S., On Collective Complete Integrability According to the Method of Thimm, *Ergodic Theory Dynamical Systems* **3** (1983), pp. 219–230.
[17] Haine L., The Algebraic Integrability of Geodesic Flow on SO(*n*), *Comm. Math. Phys.* (1984).
[18] Haine L., Geodesic Flow on SO(4) and Abelian Surfaces, *Math. Ann.* **263** (1983), pp. 435–472.
[19] Hitchin N. J., On the Construction of Monopoles, *Comm. Math. Phys.* **89** (1983), pp. 145–190.
[20] Holmes P. J. and Marsden J. E., Horseshoes and Arnold Diffusion for Hamiltonian Systems on Lie Groups, *Indiana Univ. Math. J.* **32** (1983), pp. 273–309.
[21] Iacob A. and Sternberg S., Coadjoint Structures, Solitons, and Integrability. In: *Lect. Notes in Phys.*, 120, Springer, 1980, pp. 52–84.
[22] Kac V. G., Simple Irreducible Graded Lie Algebras of Finite Growth, *Math. USSR–Izv.* **2** (1968), pp. 1271–1311.
[23] Knörrer H., Geodesics on the Ellipsoid, *Invent. Math.* **59** (1980), pp. 119–143.
[24] Knörrer H., Geodesics on Quadrics and a Mechanical Problem of C. Neumann, *J. Reine Angew. Math.* **334** (1982), pp. 69–78.
[25] Konstant B., The Solution to a Generalized Toda Lattice and Representation Theory, *Advances in Math.* **34** (1979), pp. 195–338.
[26] Kötter F., Über die Bewegung eines festen Körpers in einer Flüssigkeit I, II, *J. Reine Angew. Math.* **109** (1892), pp. 51–81, 89–111.
[27] Kötter F., Die von Steklow und Liapunow entdeckten integrablen Fälle der Bewegung eines starren Körpers in einer Flüssigkeit, *Sitzungsber. Königl. Preuss. Akad. d. Wiss. Berlin* **6** (1900), pp. 79–87.
[28] Kovalevski S., Sur le problème de la rotation d'un corps solide autour d'un point fixe, *Acta Math.* **12** (1889), pp. 177–232; Sur une propriété du système d'équations différentielles qui définit la rotation d'un corps solide autour d'un point fixe, *Acta Math.* **14** (1890/1891), pp. 81–93.
[29] Manakov S. V., Remarks on the Integrals of the Euler Equations of the *n*-Dimensional Heavy Top, *Funct. Anal. Appl.* **10** (4) (1976), pp. 93–94.
[30] McKean H. P., Integrable Systems and Algebraic Curves, In: *Lect. Notes in Math.*, 755, Springer, 1979, pp. 83–200.
[31] Melnikov V. K., On the Stability of the Center for Time-Periodic Perturbations, *Trans. Moscow Math. Soc.* **12** (1963), pp. 1–57.
[32] Miščenko A. S., Integral Geodesics of a Flow on Lie Groups, *Funct. Anal. Appl.* **4** (4) (1970), pp. 73–77.
[33] Miščenko A. S. and Fomenko A. T., Euler Equations on Finite-Dimensional Lie Groups, *Izv. Akad. Nauk. SSSR Ser. Mat.* **42** (1978), pp. 396–415.
[34] van Moerbeke P., The Spectrum of Jacobi Matrices, *Invent. Math.* **37** (1976), pp. 45–81.
[35] van Moerbeke P., About Isospectral Deformations of Discrete Laplacians. In: *Global Analysis*, Lect. Notes in Math., 755, Springer, 1979, pp. 313–370.
[36] van Moerbeke P., The Complete Integrability of Hamiltonian Systems In: *Equadiff 82, Proceedings, Würzburg 1982*, Lect. Notes in Math., 1017, Springer, 1983, pp. 462–475.

[37] van Moerbeke P. and Mumford D., The Spectrum of Difference Operators and Algebraic Curves, *Acta Math.* **143** (1979), pp. 93–154.

[38] Moser J., Various Aspects of Integrable Hamiltonian Systems. In: Guckenheimer J., Moser J., and Newhouse S. E., *Dynamical Systems, C.I.M.E. Lectures, Bressanone, Italy, June 1978*, Progr. Math., 8, Birkhäuser, Boston, 1980, pp. 233–289.

[39] Moser J., Geometry of Quadrics and Spectral Theory. In: *Symposium in Honor of S. S. Chern*, Berkeley, 1979, pp. 147–188.

[40] Mumford D., An Algebro-Geometrical Construction of Commuting Operators and of Solutions to the Toda Lattice Equation. In: *Proc. Kyoto Conf. in Alg. Geom.*, Publ. Math. Soc. Japan, 1977.

[41] Mumford D., *Tata Lectures on Theta II*, Progr. Math., **43**, Birkhäuser, Boston, 1984, Chapter III: Jacobian theta functions and differential equations.

[42] Nahm W., *All Self-Dual Monopoles for All Gauge Groups*, preprint, CERN, 1981.

[43] Perelomov A. M., Some Remarks on the Integrability of the Equations of Motion of a Rigid Body in an Ideal Fluid, *Funct. Anal. Appl.* **15** (1981), pp. 83–85, transl. pp. 144–146.

[44] Ratiu T., Euler–Poisson Equations on Lie Algebras and the N-dimensional Heavy Rigid Body, *Amer. J. Math.* **104** (1982), pp. 409–448.

[45] Ratiu T. and van Moerbeke P., The Lagrange Rigid Body Motion, *Ann. Inst. Fourier* **32** (1) (1982), pp. 211–234.

[46] Reid M., *The Complete Intersection of Two or More Quadrics*, Thesis, Cambridge, 1982.

[47] Reyman A. G. and Semenov-Tian-Shansky M. A., Reduction of Hamiltonian Systems, Affine Lie Algebras and Lax Equations I, II, *Invent. Math.* **54** (1979), pp. 81–100, *Soviet Math. Dokl.* **63** (1981), pp. 423–432.

[48] Reyman A. G., Semenov-Tian-Shansky M. A., and Frenkel I. B., Affine Lie Algebras and Completely Integrable Hamiltonian Systems, *Soviet Math. Dokl.* **247** (1979), pp. 802–805.

[49] Systems W., Systems of Toda Type, Inverse Spectral Problems, and Representation Theory, *Invent. Math.* **59** (1980), pp. 13–53.

[50] Thimm A., *Integrable Geodesic Flows on Homogeneous Spaces*, Doctoral Dissertation, Bonn, 1980, and *Ergodic Theory Dynamical Systems* **1** (1981), pp. 495–517.

[51] Verdier J. L., Algèbres de Lie, systèmes Hamiltoniens, courbes algébriques, *Sém. Bourbaki*, exp. 566 (1980/1981), Lect. Notes in Math., 901, Springer, 1981, pp. 85–94.

[52] Ziglin S. L., Splitting of Separatrices, Branching of Solutions and Nonexistence of an Integral in the Dynamics of a Solid Body, *Trans. Moscow Math. Soc.* **41** (1980), pp. 287–303, transl. **1** (1982), pp. 283–298.

[53] Ziglin S. L., Branching of Solutions and Nonexistence of First Integrals in Hamiltonian Mechanics, I, II, *Funct. Anal. Appl.* **16** (3) 1982, pp. 30–41, **17** (1) (1983), pp. 8–23.

[54] Ziglin S. L., *Funct. Anal. Appl.* **17** (1983), pp. 8–23.

DEPARTMENT OF MATHEMATICS, UNIVERSITY OF LOUVAIN, LOUVAIN-LA-NEUVE, BELGIUM
AND
BRANDEIS UNIVERSITY, WALTHAM, MASS., USA

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

TOSHIO OSHIMA

Discrete Series for Semisimple Symmetric Spaces

A homogeneous space $X = G/H$ of a connected Lie group G is called a *symmetric space* if there exists an involutive automorphism σ of G such that H lies between the fixed point group G^σ and its connected component G_0^σ containing the identity element.

For a connected Lie group G', put $G = G' \times G'$, $\sigma((g_1, g_2)) = (g_2, g_1)$ and $H = G^\sigma$. Then the homogeneous space $X = G/H$ is naturally isomorphic to G' by the map $(g_1, g_2) \mapsto g_1 g_2^{-1}$. Hence any connected Lie group is an example of a symmetric homogeneous space. Another typical example is a Riemannian symmetric space, which has been well studied from several viewpoints.

Now we restrict ourselves to the case where G is semisimple. In this case we call G/H a *semisimple symmetric space*. Berger [1] classifies all the pairs $(\mathfrak{g}, \mathfrak{h})$ of Lie algebras corresponding to semisimple symmetric pairs (G, H). For simplicity we assume that G is a real form of a complex Lie group G_c and that H is the identity component G_0^σ of G^σ. Then X admits an invariant measure and we have a unitary representation of G in the Hilbert space $L^2(G/H)$ of square integrable functions on X. Hence it is a fundamental problem to give an explicit decomposition of $L^2(G/H)$ into irreducible unitary representations of G.

Let $\boldsymbol{D}(G/H)$ be the ring of invariant differential operators on X. Then $\boldsymbol{D}(G/H)$ is a polynomial ring $\boldsymbol{C}[\Delta_1, \ldots, \Delta_r]$, where Δ_j are algebraically independent. The number r is called the *rank* of the symmetric space, which we will denote by $\operatorname{rank}(G/H)$. Here we may assume that Δ_j are self-adjoint operators on $L^2(G/H)$. Then the problem we have mentioned is almost equivalent to the following problem: give a simultaneous spectral decomposition of the self-adjoint operators.

In the decomposition of $L^2(G/H)$ there appear several different types of representations. Some of them correspond to most continuous spectra of $\boldsymbol{D}(G/H)$. They are called *most continuous unitary "principal series"* for

G/H. The projection operator onto the principal series is given by an integral transformation by the "Poisson kernel" for X, and the corresponding Plancherel measure is given by a c-function (cf. [5] for their definitions), which is calculated explicitly. Other important and fundamental representations are those which correspond to discrete spectra of $\boldsymbol{D}(G/H)$.

By the discrete series for G/H we mean the minimal closed G-invariant subspaces of $L^2(G/H)$. We fix a σ-invariant maximal compact subgroup K of G. Then the first fundamental result is

THEOREM 1. *The discrete series for G/H is non-empty if and only if*

$$\operatorname{rank}(G/H) = \operatorname{rank}(K/K\cap H). \tag{1}$$

Harish-Chandra [3] proves that a discrete series for a semisimple Lie group G exists if and only if $\operatorname{rank}(G) = \operatorname{rank}(K)$, which is a special case of the above theorem. In general, when the condition (1) holds, Flensted-Jensen [2] (cf. [7]) constructs infinitely many irreducible representations that belong to the discrete series. On the other hand it is proved in [6] that condition (1) is necessary for the existence of the discrete series.

Now to describe the discrete series we prepare some notation. Hereafter we assume condition (1). Let θ be the Cartan involution of G corresponding to K. We remark that $\sigma\theta = \theta\sigma$. The involutions of the Lie algebra $\mathfrak{g}$ of G induced by σ and θ, and the corresponding complex linear involutions of the Lie algebra $\mathfrak{g}_c$ of G_c are denoted by the same letters, respectively. Let $\mathfrak{g} = \mathfrak{h}+\mathfrak{q}$ (resp. $\mathfrak{g} = \mathfrak{k}+\mathfrak{p}$) be the decompositions of $\mathfrak{g}$ into the $+1$ and -1 eigenspaces for σ (resp. θ). Put $\mathfrak{k}^d = \mathfrak{k}\cap\mathfrak{h} + \sqrt{-1}(\mathfrak{p}\cap\mathfrak{h})$ and let K^d be the analytic subgroup of G_c corresponding to $\mathfrak{k}^d$. Then K^d is a compact real form of the complexification H_c of H in G_c. Let $\mathfrak{t}_\mathfrak{q}$ be a maximal Abelian subspace of $\mathfrak{k}\cap\mathfrak{g}$. Then $\mathfrak{t}_\mathfrak{q}$ is a maximal Abelian subspace of $\mathfrak{q}$, which is equivalent to condition (1). Let $(\mathfrak{t}_\mathfrak{q})_c^*$ denote the set of all linear maps of $\mathfrak{t}_\mathfrak{q}$ to $\boldsymbol{C}$. For an $\alpha \in (\mathfrak{t}_\mathfrak{q})_c^*$ we put $\mathfrak{g}_c(\mathfrak{t}_\mathfrak{q}; \alpha) = \{X \in \mathfrak{g}_c;$ $[Y, X] = \alpha(Y)X$ for all $Y \in \mathfrak{t}_\mathfrak{q}\}$ and moreover put $\Sigma(\mathfrak{t}_\mathfrak{q}) = \{\alpha \in (\mathfrak{t}_\mathfrak{q})_c^* \setminus \{0\};$ $\mathfrak{g}_c(\mathfrak{t}_\mathfrak{q}; \alpha) \neq \{0\}\}$. Fix a positive system $\Sigma(\mathfrak{t}_\mathfrak{q})^+$ of $\Sigma(\mathfrak{t}_\mathfrak{q})$ and put $\mathfrak{n}_\mathfrak{t}^+ = \Sigma_\alpha \mathfrak{g}_c(\mathfrak{t}_\mathfrak{q}; \alpha)$ where the sum is taken over all $\alpha \in \Sigma(\mathfrak{t}_\mathfrak{q})^+$ and put $\varrho(Y)$ $= \frac{1}{2}\operatorname{trace}_{\boldsymbol{C}}\left(\operatorname{ad}(Y)|_{\mathfrak{n}_\mathfrak{t}^+}\right)$ for $Y \in \mathfrak{t}_\mathfrak{q}$. Let W be the Weyl group of the root system $\Sigma(\mathfrak{t}_\mathfrak{q})$. Then W is naturally isomorphic to the quotient group of the normalizer $N_{K^d}(\mathfrak{t}_\mathfrak{q})$ of $\mathfrak{t}_\mathfrak{q}$ in K^d over the centralizer $Z_{K^d}(\mathfrak{t}_\mathfrak{q})$ of $\mathfrak{t}_\mathfrak{q}$ in K^d. Let $U(\mathfrak{g})$ (resp. $U(\mathfrak{t}_\mathfrak{q})$) be the universal enveloping algebras of the complexifications of $\mathfrak{g}$ (resp. $\mathfrak{t}_\mathfrak{q}$). Let $U(\mathfrak{g})^\mathfrak{h}$ (resp. $U(\mathfrak{t}_\mathfrak{q})^W$) be the subalgebras of $U(\mathfrak{g})$ (resp. $U(\mathfrak{t}_\mathfrak{q})$) consisting of $\mathfrak{h}$-invariant (resp. W-invariant) elements. Using the direct sum decomposition $U(\mathfrak{g}^d) = \left(\sigma(\mathfrak{n}_\mathfrak{t}^+)U(\mathfrak{g}) + U(\mathfrak{g})\mathfrak{h}\right)\oplus U(\mathfrak{t}_\mathfrak{q})$,

we define a projection p of $U(\mathfrak{g})$ onto $U(\mathfrak{t}_\mathfrak{q})$. Then it is known that the map $U(\mathfrak{g}) \ni D \mapsto e^{\varrho} \circ p(D) \circ e^{-\varrho} \in U(\mathfrak{t}_\mathfrak{q})$

$$U(\mathfrak{g})^{\mathfrak{h}} / U(\mathfrak{g})^{\mathfrak{h}} \cap U(\mathfrak{g})\mathfrak{h} \xrightarrow{\sim} U(\mathfrak{t}_\mathfrak{q})^{W}.$$

It is clear that the left-hand side is isomorphic to $\boldsymbol{D}(G/H)$. Then for a $\lambda \in (\mathfrak{t}_\mathfrak{q})_c^*$, we can define an algebra homomorphism χ_λ of $\boldsymbol{D}(G/H)$ to $\boldsymbol{C}$ by the above isomorphism. Here we note that $\chi_\lambda = \chi_\mu$ if and only if $\mu = w\lambda$ for a suitable $w \in W$.

For a $\lambda \in (\mathfrak{t}_\mathfrak{q})_c^*$, we define the following subspaces of $L^2(G/H)$:

$$L^2(G/H; \mathscr{M}_\lambda) = \{f \in L^2(G/H); Df = \chi_\lambda(D)f \text{ for all } D \in \boldsymbol{D}(G/H)\}$$

and

$$L^2_K(G/H; \mathscr{M}_\lambda) = \{f \in L^2(G/H; \mathscr{M}_\lambda); f \text{ is } K\text{-finite}\}.$$

Let V be a minimal closed G-invariant subspace in $L^2(G/H)$ and let V_K be the subspace of K-finite functions in V. Then it is clear that there exists a λ in $(\mathfrak{t}_\mathfrak{q})_c^*$ such that V is embedded in $L^2(G/H; \mathscr{M}_\lambda)$ and that

$$\operatorname{Re}\langle \lambda, \alpha \rangle \geqslant 0 \quad \text{for all } \alpha \in \Sigma(\mathfrak{t}_\mathfrak{q})^+. \tag{2}$$

Let $T_\mathfrak{q}$ be the analytic subgroup of G with the Lie algebra $\mathfrak{t}_\mathfrak{q}$, and let P_c^a be a parabolic subgroup of G_c with the Levi decomposition $P_c^a = M_c^a N_i^a$, where M_c^a is the centralizer of $T_\mathfrak{q}$ in G_c and N_i^a is the analytic subgroup of G_c corresponding to $\mathfrak{n}_i^+$. Define a lattice $L(\mathfrak{t}_\mathfrak{q})$ in $(\mathfrak{t}_\mathfrak{q})_c^*$ by

$$L(\mathfrak{t}_\mathfrak{q}) = \{\lambda \in (\mathfrak{t}_\mathfrak{q})_c^*; \lambda - \varrho \text{ can be extended to a character of } T_\mathfrak{q} \text{ and } \exp\langle \lambda, X \rangle = 1 \text{ for all } X \in \mathfrak{t}_\mathfrak{q} \text{ satisfying } \exp X \in H\}.$$

Then, for any $\lambda \in L(\mathfrak{t}_\mathfrak{q})$, we can define a holomorphic homomorphism τ_λ of P_c^a to $\boldsymbol{C}^\times$ so that $\tau_\lambda(\exp X) = \exp\langle \lambda - \varrho, X \rangle$ for all $X \in \mathfrak{t}_\mathfrak{q}$ and that $\tau_\lambda(\exp Y) = 1$ if $Y \in \{Y \in \mathfrak{p}_c^a; \langle X, Y \rangle = 0 \text{ for all } X \in \mathfrak{t}_\mathfrak{q}\}$. Here $\mathfrak{p}_c^a$ is the Lie algebra of P_c^a and $\langle\ ,\ \rangle$ is the Killing form of $\mathfrak{g}_c$ over $\boldsymbol{C}$. By this homomorphism τ_λ we can associate a holomorphic line bundle L_λ over G_c/P_c^a. In fact, for an open subset UP_c^a of G_c/P_c^a, any section f of L_λ over UP_c^a is a holomorphic function on UP_c^a and satisfies $f(gx) = f(g)\tau_\lambda(x)$ for all $g \in UP_c^a$ and $x \in P_c^a$.

Let K_c be the complexification of K in G_c. We define the following compact K_c-orbits in G_c/P_c^a, which have a strong connection with the discrete series. Let $W(K)$ be the quotient group of the normalizer of $\mathfrak{t}_\mathfrak{q}$ in $K \cap H$ over the centralizer of $\mathfrak{t}_\mathfrak{q}$ in $K \cap H$. Let $w_1, \ldots, w_m$ be representatives of the coset $W(K) \backslash W$ and let $\overline{w}_1, \ldots, \overline{w}_m \in K^a$ be representatives

of $w_1, \ldots, w_m$, respectively. Then $K_c\overline{w}_jP_c^d$ are compact K_c-orbits in G_c/P_c^d and if $i \neq j$, then $K_c\overline{w}_iP_c^d \neq K_c\overline{w}_jP_c^d$ (cf. [4]).

Then the following theorem is a slight modification of a result in [6]:

THEOREM 2. (i) *Assume the conditions* (1) *and* (2). *If* $L^2(G/H; \mathscr{M}_\lambda) \neq \{0\}$, *then*

$$\langle \lambda, \alpha \rangle \neq 0 \quad \textit{for all } \alpha \in \Sigma(\mathfrak{t}_\mathfrak{q}) \tag{3}$$

and

$$\lambda \in L(\mathfrak{t}_\mathfrak{q}). \tag{4}$$

(ii) *Under the conditions* (1), (2), (3) *and* (4) *we have the following* $(\mathfrak{g}, K)$-*isomorphism*

$$L^2_K(G/H; \mathscr{M}_\lambda) \simeq \bigoplus_{j=1}^{m} H^n_{K_c\overline{w}_jP_c^d}(G_c/P_c^d, \mathscr{O}_{\mathrm{alg}}(L_\lambda)), \tag{5}$$

where n *equals the complex codimension of the compact algebraic manifold* $K_c\overline{w}_jP_c^d$ *in* G_c/P_c^d (*which does not depend on* j) *and* $\mathscr{O}_{\mathrm{alg}}(L_\lambda)$ *means a sheaf of sections of* L_λ *in the sense of algebraic geometry.*

In the above theorem, if G is compact, then condition (1) is trivial and $m = 1$, $n = 0$, $K_c = G_c$ and the right-hand side of (5) equals $\Gamma(G_c/P_c^d; \mathscr{O}_{\mathrm{alg}}(L_\lambda))$. In this case, Theorem 2 (ii) is reduced to the well-known Borel–Weil theorem. So we want to call Theorem 2 (ii) the Borel–Weil theorem for discrete series for semisimple symmetric spaces.

References

[1] Berger M., Les espace symétriques noncompacts, *Ann. Sci. École Norm. Sup.* **74** (1957), pp. 85–177.

[2] Flensted-Jensen M., Discrete Series for Semisimple Symmetric Spaces, *Ann. of Math.* **111** (1980), pp. 253–311.

[3] Harish-Chandra, Discrete Series for Semisimple Lie Groups, I, II, *Acta Math.* **113** (1965), pp. 241–318; **116** (1966), pp. 1–111.

[4] Matsuki T., The Orbits of Affine Symmetric Spaces under the Action of Minimal Parabolic Subgroups, *J. Math. Soc. Japan* **31** (1979), pp. 331–357.

[5] Oshima T., Fourier Analysis on Semisimple Symmetric Spaces, *Non-Commutative Harmonic Analysis. Proceedings, 1980,* Lect. Notes in Math. **880**, pp. 357–369, Springer–Verlag, 1981.

[6] Oshima T. and Matsuki T., A Description of Discrete Series for Semisimple Symmetric Spaces, to appear in *Advanced Studies in Pure Math.*

[7] Schlichtkrull H., Applications of Hyperfunction Theory to Representations of Semi-simple Lie Groups, to appear in *Progress in Mathematics,* Birkhäuser.

DEPARTMENT OF MATHEMATICS, FACULTY OF SCIENCE,
UNIVERSITY OF TOKYO, TOKYO 113, JAPAN

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

R. PARTHASARATHY

Unitary Modules with Non-Vanishing Relative Lie Algebra Cohomology

For a real semisimple Lie algebra $\mathfrak{g}_0$ it is of interest to find the class of unitarizable, irreducible Harish–Chandra modules X such that $H^i(\mathfrak{g}_0, k_0, X \otimes F)$ is non-zero for some integer i. Here, k_0 is a maximal compactly imbedded subalgebra of $\mathfrak{g}_0$, F is a finite dimensional irreducible module for $\mathfrak{g}_0$ and $H^i(\ldots)$ are the relative Lie algebra cohomology spaces. It is known ([8]) that the class of such modules is a subclass of the class of modules $A_{\mathfrak{q},\lambda}$ constructed in [5] and [9]. Here, $\mathfrak{q}$ is a parabolic subalgebra of $\mathfrak{g}$ (dropping the subscript 0 means complexification) under a Cartan involution θ fixing k_0, and λ varies over a subset of t^* where t is a fundamental Cartan subalgebra of $\mathfrak{g}$ contained in $\mathfrak{q}$. There is a conjecture, as yet unproved, that $A_{\mathfrak{q},\lambda}$ belong to the subclass above. The only difficulty in proving this conjecture is the inability to prove that the $A_{\mathfrak{q},\lambda}$ in question are unitarizable. This problem has been solved in some special cases — to mention a few due to Speh ([7]) in the case of $\mathfrak{sl}(n, \boldsymbol{R})$, by Enright ([3]) in the complex case, by Baldoni-Silva and Barbash ([1]) in the case of real rank one groups and by the present author ([6]) in the case of highest weight modules. The general problem is often vaguely referred to as the problem of unitarizability of $A_{\mathfrak{q},\lambda}$. A solution has now been obtained by Vogan and more recently, also by Wallach.

For all $\mathfrak{g}$, when $\mathfrak{q}$ is of quasi-abelian type (see the definition below), it has been proved recently by Enright, Parthasarathy, Wallach, and Wolf ([4]) that $A_{\mathfrak{q},\lambda}$ are unitarizable. (We say $\mathfrak{q}$ is quasi-abelian if $[u \cap k, u \cap p] = 0$ where u is the nilradical of $\mathfrak{q}$).

In this paper we will assume that rank of $\mathfrak{g}_0$ = rank of k_0 and discuss the unitarizability of $A_{\mathfrak{q},\lambda}$ whenever $\mathfrak{q}$ contains the Borel subalgebra r corresponding to a Borel–de Siebenthal chamber P (Definition: P contains a unique non-compact simple root β and its coefficient in the highest root is 1 or 2).

We assume that $\mathfrak{g}_0$ is irreducible. Recall that rank of k_0 = rank of $\mathfrak{g}_0$. Let β be the unique noncompact simple root for P and, if $(\mathfrak{g}_0, k_0)$ is not a hermitian symmetric pair, let γ be the unique simple root for P_k which is not a simple root for P. Let u be the nilradical of $\mathfrak{q}$. Write

$$u \cap k = u'_k + u''_k,$$

where the coefficient of β in the roots in u'_k is zero and the coefficient of β in the roots in u''_k is two.

By looking at the construction in [5] one can infer that, if $\mathfrak{q}' \cap p = \mathfrak{q} \cap p$, then we may work with $\mathfrak{q}'$ instead of $\mathfrak{q}$. Thus we can assume that $\mathfrak{q}$ is maximal among all $\mathfrak{q}'$, such that $\mathfrak{q} \cap p = \mathfrak{q}' \cap p$.

By a verification case by case one can check that the following is true:

(*) *Either* $\mathfrak{q}$ *is of quasi-abelian type or, if* $\mathfrak{q}_1$, *with the Levi decomposition* $m_1 + u_1$ *is the maximal parabolic subalgebra obtainted by deleting* β, *then* $u_1 \cap k = u''$.

We only have to deal with the latter case since for parabolics of quasi-abelian type the result is proved in [4].

From the construction in [5] recall that the modules $A_{\mathfrak{q},\lambda}$ are obtained by a chain of "completions" of a $\mathfrak{g}$-Verma module $V_{\mathfrak{g},P,\mu}$ with respect to a reduced expression for $w \in W_k$, where $P_k \cap -wP_k$ = the roots in $u \cap k$. Here, the word "completion" is used in the sense of [2]. Thus, we can make

(*Step* 1) completion with respect to a reduced expression for w' where $P_k \cap -w'P_k$ = the roots in u'_k,

followed by

(*Step* 2) completion with respect to a reduced expression for w'' where $P_k \cap -w''P_k$ = the roots in u''_k.

Since β has coefficient zero in the roots in u'_k, the simple reflection s_γ does not occur in the reduced expression for w'. Thus, after completion of step 1, the original Verma module becomes another Verma module $V_{\mathfrak{g},P,\mu'}$. In view of (*), step 2 is now just like producing a module of type $A_{\mathfrak{g}_{1,*}}$.

The parabolic $\mathfrak{q}_1$ is of quasi-abelian type. However, the parameter μ' is in general off the list of parameters for which unitarizability has been established in [4].

The crucial "Dirac operator inequality" associated with unitarizability, which is often an aid in proving unitarizability, can be seen to hold for

the irreducible quotient of $V_{\theta,P,\mu}$. This inequality is preserved during completions. Thus a large part of $V_{\theta,P,\mu'}$ (considered as a k-module) which is adequate to analyse the module $A_{\mathfrak{q},\lambda}$ will satisfy the Dirac operator inequality. This circumstance enables one to employ the techniques familiar in the quasi-abelian case and leads to a proof of the unitarizability of $A_{\mathfrak{q},\lambda}$.

References

[1] Baldoni-Silva M. W. and Barbash D., The Unitary Spectrum for Real Rank One Groups, *Invent. Math.* **72** (1983), pp. 27–75.

[2] Enright T. J., On the Fundamental Series of a Real Semisimple Lie Algebra: Their Irreducibility, Resolution and Multiplicity Formulae, *Ann. of Math.* **110** (1979), pp. 1–82.

[3] Enright T. J., Relative Lie Algebra Cohomology and Unitary Representations of Complex Lie Groups, *Duke Math. J.* **46** (1979), pp. 513–525.

[4] Enright T. J., Parthasarathy R., Wallach N. and Wolf J., to appear.

[5] Parthasarathy R., A Generalization of the Enright–Varadarajan Modules, *Compositio Math.* **36** (1978), pp. 53–73.

[6] Parthasarathy R., Criteria for the Unitarizability of Some Highest Weight Modules, *Proc. Indian Acad. Sci. Sect. A.* **89** (1980), pp. 1–24.

[7] Speh B. In: *Non-Commutative Harmonic Analysis and Lie Groups, Lecture Notes in Mathematics* **880**, Springer-Verlag, 1981.

[8] Vogan D. and Zuckerman G., Unitary Representations with Continuous Cohomology, manuscript.

[9] Zuckerman G., handwritten notes on construction of representations by derived functors.

SCHOOL OF MATHEMATICS
TATA INSTITUTE OF FUNDAMENTAL RESEARCH
HOMI BHABHA ROAD
BOMBAY 400 005
INDIA

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

A. B. VENKOV

Spectral Theory of Automorphic Functions for Fuchsian Groups of the First Kind and Its Applications to Some Classical Problems of the Monodromy Theory

Introduction

In the first part of this address we give a description of the spectrum of the automorphic Laplacian A for the Fuchsian group of the first kind with noncompact fundamental domain. In the second part we establish a connection between the resolvent of the operator A and the harmonic Green function on the Riemann surface for the group Γ. As an application of this theory we obtain an expression for the Fourier coefficients of the Klein absolute invariant $\mathcal{J}$ for the group Γ with zero genus in terms of some series with Kloosterman sums. We find also certain explicit formulae for the accessory coefficients of the Fuchs equation with Γ as a monodromy group.

We begin by introducing some notations: H is the hyperbolic plane, ds^2 is the Poincaré metric, L is the Laplace operator in the ds^2-metric, $d\mu$ is the Riemannian measure associated with the ds^2-metric, Γ is an arbitrary Fuchsian group of the first kind with noncompact fundamental domain F (notation $\Gamma \in \mathfrak{M}$), $|F|$ is the $d\mu$-volume of F ($|F| < \infty$), χ is an arbitrary finite-dimensional unitary representation of Γ, $V = V(\chi)$ is the linear space over C for the representation χ, $\mathcal{H}(\Gamma; \chi) = L_2(F; d\mu; \chi)$ is the standard Hilbert space of functions $f: H \to V$ which are automorphic: $f(\gamma z) = \chi(\gamma) f(z)$, and which have a square integrable V norm on F with respect to $d\mu$, $A(\Gamma; \chi)$ is the self-adjoint non-negative operator generated by $-L$ in $\mathcal{H}(\Gamma; \chi)$ (the automorphic Laplacian).

Part 1

From the famous papers by A. Selberg and from the works of his successors in the 60's (R. Godement, L. Faddeev, R. Langlands, W. Roelcke) the following properties of the spectrum of the operator $A(\Gamma; \chi)$ are well known. The spectrum is spread over $[0, \infty)$ and contains a k-multiple continuous spectrum in $[1/4, \infty)$. The multiplicity $k = k(\Gamma; \chi)$ is the total degree of singularity of χ with respect to Γ. For example, if χ is trivial and one-dimensional ($\chi = 1$), then k is the number of all pairwise non-equivalent cusps of F. (If χ is regular, $\chi \in \mathfrak{N}_r(\Gamma)$, then the continuous spectrum is missing, $k = 0$. We consider here only the situation in which continuous spectrum does occur, $\chi \in \mathfrak{N}_s(\Gamma)$.) The eigenfunctions of the continuous spectrum arise from the Eisenstein–Maass series as a result of its analytical continuation. The general theory does not exclude also the existence of a discrete spectrum of $A(\Gamma; \chi)$ composed of isolated eigenvalues of finite multiplicity.

From the naïve point of view, one might suppose that the discrete spectrum of $A(\Gamma; \chi)$ is ample. Namely, one can expect that its distribution function $N(\lambda; \Gamma; \chi) = \{\text{the number of all } \lambda_j |\ \lambda_j \leqslant \lambda\}$ satisfies the Weyl formula

$$N(\lambda; \Gamma; \chi) \underset{\lambda \to \infty}{\sim} \frac{|F|}{4\pi} \dim V(\chi) \cdot \lambda \tag{1}$$

and the proof of this statement should not be too difficult. However, up to now the author does not know any proof in the general situation of $\Gamma \in \mathfrak{M}$, $\chi \in \mathfrak{N}_s(\Gamma)$, that there exists at least one eigenvalue of $A(\Gamma; \chi)$ which lies on the continuous spectrum.

A satisfactory solution to this problem might result from an investigation of certain Dirichlet series defined by A. Selberg in [16]. These series arise from the Fourier coefficients of the Eisenstein–Maass series. Investigation of this kind has been so far pursued in the very special situation of some congruence subgroup $\Gamma \subseteq \mathrm{PSL}(2; \boldsymbol{Z})$ (see [3]) (and for cycloidal subgroups) but it is very difficult in any other case (see [19]). Information about the discrete spectrum of $A(\Gamma; \chi)$ we get here is valid in a more general situation and is obtained by indirect methods only.

Let us now turn attention to theorems on the spectrum of $A(\Gamma; \chi)$. Let $I(s; \Gamma; \chi)$ be the determinant of the scatterring matrix defined by $A(\Gamma; \chi)$. It is a meromorphic function of $s \in \boldsymbol{C}$. From the Selberg trace formula the following formula (see [16]) results (called here the *Weyl–Selberg asymptotic formula*).

THEOREM 1. *For any* $\Gamma \in \mathfrak{M}$ *and* $\chi \in \mathfrak{N}_s(\Gamma)$

$$N(\lambda; \Gamma; \chi) - \frac{1}{4\pi} \int_{-T}^{T} \frac{I'}{I}\left(\frac{1}{2} + ir; \Gamma; \chi\right) dr \underset{\lambda\to\infty}{\sim} \frac{|F| \dim V}{4\pi} \lambda,$$

$\lambda = 1/4 + T^2$, $T > 0$.

The precision of this asymptotic formula may be increased, if necessary (see [20]).

COROLLARY 1. *For given* Γ *and* χ, *let* $I(s;\Gamma;\chi)$ *be a meromorphic function of order* 1; *then the Weyl formula* (1) *is valid for* $N(\lambda; \Gamma; \chi)$.

Let $Z(s; \Gamma; \chi) = \prod_{\{P\}} \prod_{k=0}^{\infty} \det\left(1_V - \chi(P)\mathcal{N}(P)^{-s-k}\right)$, $\mathrm{Re}\, s > 1$, be the Selberg zeta function (see [16]). It is known that it has a meromorphic continuation to the whole of $\boldsymbol{C}$ (as a function of s) and is a function of order 2 satisfying a certain functional equation. Some of its nontrivial zeroes lying on $\mathrm{Re}\, s = 1/2$ and $s \in [0, 1]$ correspond to eigenvalues λ_j of $A(\Gamma; \chi)$ and the other ones correspond to the poles of $I(s; \Gamma; \chi)$ in $\mathrm{Re}\, s < 1/2$. From the definition of $Z(s; \Gamma; \chi)$ and from the Selberg trace formula the following theorem is derived (the *Artin factorization formula* as I call it) (see [20], [21]):

THEOREM 2. *Let* $\Gamma \in \mathfrak{M}$ *and let* $\Gamma_1 \subset \Gamma$ *be an arbitrary subgroup of finite index,* $\chi \in \mathfrak{N}(\Gamma_1) = \mathfrak{N}_r(\Gamma_1) \cup \mathfrak{N}_s(\Gamma_1)$; *then*

(1) $Z(s; \Gamma_1; \chi) = Z(s; \Gamma; U^\chi)$;

(2) $I(s; \Gamma_1; \chi)\, \Omega(\Gamma_1; \chi)^{1-2s} = I(s; \Gamma; U^\chi)\, \Omega(\Gamma; U^\chi)^{1-2s}$; *here* U^χ *is the induced representation (from* χ*),* Ω *is a certain elementary function.*

COROLLARY 2. *If the Weyl formula* (1) *is valid for* $N(\lambda; \Gamma_1; 1)$ *then it is true for* $N(\lambda; \Gamma; 1)$, *whenever* Γ_1 *is a normal subgroup of finite index in* Γ.

From the results of D. Hejhal and Corollary 2 the next theorem follows:

THEOREM 3. *Let* Γ *be an arbitrary congruence subgroup (not necessarily principal)* $\Gamma \subseteq \mathrm{PSL}(2, \boldsymbol{Z})$; *then the Weyl formula* (1) *is valid for* $N(\lambda; \Gamma; 1)$.

We now turn to an inspection of those cases where one can only prove

$$N(\lambda; \Gamma; \chi) \underset{\lambda\to\infty}{\to} \infty; \tag{2}$$

this condition being called here the *Roelcke hypothesis*. I got to know this hypothesis from J. Elstrodt and L. Faddeev. It has a meaning for arbit-

rary $\Gamma \in \mathfrak{M}$, $\chi \in \mathfrak{R}_s(\Gamma)$. W. Roelcke proved (2) for Hecke groups $\Gamma = G(2\cos\pi/q)$ and $\chi = 1$ (see [14]).

From the well-known Fricke theorem on the algebraic structure of Γ and from Theorem 2 we get (see [20]):

THEOREM 4. (i) *Let Γ_1 be an arbitrary normal subgroup of finite index in $\Gamma \in \mathfrak{M}$. Then for any sufficiently large λ the inequality holds:*

$$N(\lambda; \Gamma_1; 1) \geqslant \Big(\sum_{\chi \in (\Gamma_1 \backslash \Gamma)^*_{\mathrm{reg}}} \dim^2 V(\chi) \Big) \frac{|F_1|}{4\pi[\Gamma : \Gamma_1]} \lambda.$$

Here $(\Gamma_1 \backslash \Gamma)^*_{\mathrm{reg}}$ is the set of all regular representations of the factor group $\Gamma_1 \backslash \Gamma$.

(ii) *For any $\Gamma \in \mathfrak{M}$ there is a subgroup Γ_1 as in* (i) *such that the set $(\Gamma_1 \backslash \Gamma)^*_{\mathrm{reg}}$ is not empty.*

From the formal point of view, the theory of general Hecke operators can be regarded as a different method of investigating the discrete spectrum of $A(\Gamma; \chi)$. These operators exist if Γ has a nontrivial commensurator and is generated by nontrivial translations $g: H \to H$ such that the double classes $\Gamma g \Gamma$ are discrete (we suppose here $\chi = 1$). Any Hecke operator $T(g)$ is linear, continuous in $\mathscr{H}(\Gamma; \chi)$, commutes with any bounded function of $A(\Gamma; \chi)$ and satisfies $T^*(g) = T(g^{-1})$, asterisk denoting the conjugate operator. These properties of $T(g)$ are useful for the investigation of the spectrum of $A(\Gamma; \chi)$. The following theorem is typical (see [20]).

THEOREM 5. *Let $\Gamma_1 \in \mathfrak{M}$ have a fundamental domain with one cusp only and let Γ_1 be a nontrivial subgroup $\Gamma_1 \subset \Gamma \in \mathfrak{M}$; then $N(\lambda; \Gamma_1; 1) \underset{\lambda \to \infty}{\to} \infty$.*

We now focus attention for a while on one important example of a group $\Gamma \in \mathfrak{M}$ of zero genus with a nontrivial commensurator (see [5]). For any such Γ the corresponding Riemann surface is symmetric and the spectral theory of the operator $A(\Gamma; 1)$ is reduced to the Dirichlet and Neumann boundary value problems. (This example will also be of importance in the second part of the address.)

Let M be a regular polygon in H, i.e., such that 1) M is limited by finitely many geodesics, 2) interior angles of M are of the form π/k where $k \in \mathbf{Z}$, $k > 1$, or $k = \infty$ if the corresponding angle is equal to zero. Let m be the number of zero interior angles of M. If $m \neq 0$, then M is not compact but $|M| < \infty$. Let Γ^0_M be the group generated by the reflections with respect to the sides of M. We define a subgroup Γ_M of index 2 in the

group Γ_M^0, consisting of words of even length with respect to the generators of Γ_M^0. If $m \neq 0$ then $\Gamma_M \in \mathfrak{M}$. Its fundamental domain F_M can be chosen to be $F_M = M \cup \mathscr{E}M$, where $\mathscr{E}$ is the reflection relative to some side of M. We have (see [20]):

LEMMA 1. (1) *The reflection $\mathscr{E}$ generates the Hecke operator $T(\mathscr{E})f(z) = f(\mathscr{E}z)$;*

(2) *$T(\mathscr{E})$ is self-adjoint;*

(3) *$P_1(\mathscr{E}) = \frac{1}{2}\big(I - T(\mathscr{E})\big), P_2(\mathscr{E}) = \frac{1}{2}\big(I + T(\mathscr{E})\big)$ are orthogonal projectors onto some subspaces $\mathscr{H}_1(\Gamma_M; 1)$, $\mathscr{H}_2(\Gamma_M; 1) \subset \mathscr{H}(\Gamma_M; 1)$, $\mathscr{H} = \mathscr{H}_1 \oplus \mathscr{H}_2$; here I is the identity operator in $\mathscr{H}$.*

THEOREM 6. (1) *The restriction of the operator $-A(\Gamma_M; 1)$ to $\mathscr{H}_1(\mathscr{H}_2)$ is isomorphic to the operator of Dirichlet (Neumann) boundary value problem in M for the differential operator L.*

(2) *The operator of the Dirichlet problem from* (1) *has a discrete spectrum only.*

(We remind that the Dirichlet (Neumann) condition is $f|_{\partial M} = 0$ $(\partial f/\partial n|_{\partial M} = 0)$, where ∂M is the boundary of M and $\partial/\partial n$ is the normal derivative.)

The general method of the Selberg trace formula with a construction of a Hecke operator (see [17]) admits a modification in the special case of the group Γ_M and the operator $T(\mathscr{E})$. Hence we obtain the Selberg trace formulae of Dirichlet and Neumann boundary value problems (see [20]). I omit here these formulae in all of their details and define the Selberg zeta-function of the Dirichlet problem by

$$Z_M(s) = Z^2(s; \Gamma_M; 1) \prod_{\{\mathscr{E}_\gamma\} \Gamma_M} \prod_{l=0}^{\infty} \left(\frac{1 + \mathscr{N}(\mathscr{E}_\gamma)^{-s-l}}{1 - \mathscr{N}(\mathscr{E}_\gamma)^{-s-l}} \right)^{\alpha(\mathscr{E}_\gamma)(-1)^l}, \quad \operatorname{Re} s > 1$$

where the product is taken over some set of conjugate element classes in $\mathscr{E}\Gamma_M$, $\mathscr{N}$ is the norm of the class, $\alpha(\mathscr{E}_\gamma)$ is equal to 1 or 2, depending on $\mathscr{E}_\gamma$ (for details see [20]), $Z(s; \Gamma_M; 1)$ is the ordinary Selberg zeta-function. The function $Z_M(s)$ has similar properties to those of $Z(s; \Gamma; 1)$, whenever Γ is a cocompact group. Let $N_M(\lambda)$ be the distribution function of the eigenvalues of the Dirichlet boundary value problem operator $P_1(\mathscr{E})A(\Gamma_M; 1)$ in M. According to the method of analytical number theory, modified by D. Hejhal (see [3]) and B. Randol (see [13]), of estimation of the Riemann zeta-function argument in the "critical strip" we have:

THEOREM 7. (See [20].) *The following formula is valid*:

$$N_M(1/4+T^2) = \frac{|M|}{4\pi}T^2 - \frac{m}{2\pi}T\ln T + \frac{1}{2\pi}(C_M+m)T + O(T/\ln T), \quad T\to\infty.$$

Here C_M is a constant (see [20]). (In this formula m may be equal to zero, i.e. M can be a regular compact polygon.)

COROLLARY 3. *For any sufficiently large λ the following inequality holds*:

$$N(\lambda;\ \Gamma_M;\ 1) \geqslant \frac{|F_M|}{8\pi}\lambda.$$

Along with examples of discrete spectrum lying on the continuous one there are also examples of groups $\Gamma\in\mathfrak{M}$ for which the operator $A(\Gamma;1)$ has a discrete spectrum lying outside the continuous one, i.e. on $[0, 1/4)$. Corresponding eigenfunctions are either the residues of Eisenstein–Maass series or cusp-functions. We do not discuss here this question, which is interesting for applications, and we refer the readers to papers [18], [23].

Part 2

One of the fruitful problems of the analytic theory of ordinary linear differential equations is the problem of reconstruction of the equation coefficients according to a given monodromy group. This is a classical problem (see [10]), which is not finally investigated even for a Fuchsian second order equation. An equation of this type is closely connected with a non-linear Schwarz equation. For example, an interesting Schwarz equation arises if one investigates the conformal mapping of the upper half-plane $P\subset C$ onto a simply connected curvilinear polygon M limited by a finite number of circular arcs. We consider this situation in detail.

If $z: P\to M$ is a given conformal mapping, then

$$\{z, \mathscr{I}\} = Q_M(\mathscr{I}) \tag{3}$$

where $\{z, \mathscr{I}\}$ is the Schwarzian derivative, $\{z, \mathscr{I}\} = z'^{-1}z''' - \frac{3}{2}z''^2z'^{-2}$, and $Q_M(\mathscr{I})$ a rational function uniquely defined by the polygon M. This function has second order poles at the points corresponding to the vertices b_k of M. The coefficients at these poles are evaluated in terms of the interior angles of the polygon M (see [15], [5], [4], [2], [8]). However, as far as I know, the explicit formulae are not known for all coefficients X_k (accessory parameters) (more exactly, such formulae are lacking for

$l-3$ of them, if l is the number of vertices of M) at the first order poles of $Q_M(\mathscr{I})$ situated at the points a_k, in terms of M or in terms of the monodromy group Γ_M corresponding to the Fuchsian equation

$$f''(\mathscr{I})+\tfrac{1}{2}Q_M(\mathscr{I})f(\mathscr{I}) = 0. \tag{4}$$

(As it will be seen later, if one knows the accessory coefficients then one knows the function $Q_M(\mathscr{I})$.)

In this part of the address we find (following [22]) some formulae for the desired accessory coefficients under the natural assumption that M is a regular polygon (see Part 1). Besides, we suppose that $m = m(M) \geqslant 1$. (The result is valid also in a more general situation (see the end of this part).) It is also convenient to assume: (i) the number of angles of M is equal to $l = n+1$, $n \geqslant 2$, (ii) $M \subset H = \{z \in \boldsymbol{C} | \, z = x+iy, y = \operatorname{Im} z > 0\}$, (iii) $b_j = z(a_j)$, $j = 1, \ldots, n+1$; $b_{n+1} = i\,\infty$, $a_{n+1} = \infty$, (iv) for any sufficiently large $a > 0$ we have $\{z \in M | \operatorname{Im} z \geqslant a\} = [0, 1/2] \times [a, \infty)$.

The formulae for the accessory parameters will be obtained from a certain exact connection between the coefficients of equation (4) and the coefficients of the expansion of a given conformal mapping. Earlier, a similar connection was useful for the investigation of the $\mathscr{I}$-invariants for triangle groups starting from the hypergeometric equation (see [6], [12]).

Now we pass to the results. The following statements are known: $a_j \in \boldsymbol{R}$, $j = 1, \ldots, n+1$. If f_1 and f_2 are two linearly independent solutions of equation (4) then $z(\mathscr{I}) = f_1(\mathscr{I})f_2(\mathscr{I})^{-1}$ is the solution of (3). The inverse mapping (absolute Klein invariant) $\mathscr{I}_M(z) = \mathscr{I}(z)$ exists and is a Γ_M-automorphic function analytic in H. Moreover, Γ_M is the group defined in Part 1. Γ_M has a cyclic subgroup Γ_∞ generated by the translation $z \to z+1$. The Fourier expansion

$$\mathscr{I}_M(z) = \sum_{k \in \boldsymbol{Z},\, k \geqslant -1} A_k \exp 2\pi i k z \tag{5}$$

holds and all $A_k \in \boldsymbol{R}$. From (3) we have

$$-\{\mathscr{I}_M, z\} = \mathscr{I}_M'^2(z) Q_M\big(\mathscr{I}_M(z)\big). \tag{6}$$

The following formulae are known

$$\begin{gathered} Q_M(\mathscr{I}) = \prod_{k=1}^{n} (\mathscr{I}-a_k)^{-1}\Big[E_{n-2}(\mathscr{I}) + \sum_{k=1}^{n} N_k(\mathscr{I}-a_k)^{-1}\Big], \\ N_k = \tfrac{1}{2}(1-a_k^2) \prod_{l=1}^{n}{}' (a_k - a_l). \end{gathered} \tag{7}$$

Here the prime means that the product is taken over all t except for $t = k$, $E_{n-2}(\mathcal{J}) = \sum_{k=0}^{n-2} X_k \mathcal{J}^k$, $X_{n-2} = 1/2$. The problem is to find the coefficients X_k, $k = 0, 1, \ldots, n-3$. (In the special situation of $n = 2$ equation (4) for 'triangle functions' does not have accessory parameters.)

We introduce the notation $\prod_{k=1}^{n} (\mathcal{J} - a_k) = \sum_{k=0}^{n} B_k \mathcal{J}^k$, $B_n = 1$, $\mathcal{J} = \mathcal{J}_M(z)$. We put

$$\Phi_p(r) = \sum_{\substack{t_1+t_2+\ldots+t_p+s_1+\ldots+s_4=r,\\ -1\leqslant t_1,\ldots,t_p,s_1,\ldots,s_4<\infty,}} A_{t_1}A_{t_2}\ldots A_{t_p}A_{s_1}A_{s_2}A_{s_3}A_{s_4}s_1s_2s_3s_4,$$

$$\Psi(r) = \sum_{k=0}^{n} B_k \sum_{\substack{l_1+l_2+\ldots+l_k+m_1+m_2=r\\ -1\leqslant l_1,\ldots,l_k,m_1,m_2<\infty}} A_{l_1}A_{l_2}\ldots A_{l_k}A_{m_1}A_{m_2}m_1^2m_2\left(\tfrac{3}{2}m_2 - m_1\right). \tag{8}$$

In (8) we assume that $4 \leqslant -r \leqslant n+2$, $0 \leqslant p \leqslant n-2$. The first principal statement of this part is

THEOREM 8. (1) *The matrix* $\{\Phi_p(r)\}$ *is invertible (moreover it is triangular with positive elements on the main diagonal); let* $\{\Phi_p(r)\}^{-1} = \eta_p(r)$.

(2) *For the coefficients* X_k, $0 \leqslant k \leqslant n-3$, *the following formula is valid*:

$$X_k = \sum_{r=-4}^{-n-2} \eta_k(r)\Psi(r). \tag{9}$$

The proof of this theorem is based on formulae (5)–(7) (for details see [22]).

The equality (9) makes sense if one knows the coefficient A_k from (5). Now we will obtain some formulae for $A_{-1}^{-1}A_k$, $k \geqslant 1$. It seems that the Ramanujan–Hardy–Littlewood–Rademacher–Lehner circle method (see [7]) might be adequate here; however, this method turns out to be not sufficiently exact for an arbitrary group Γ_M. Therefore we use another method which is based on the connection between the resolvent of the operator $A(\Gamma_M; 1)$ and the harmonic Green function. Furthermore this method gives some more information about the invariant $\mathcal{J}_M(z)$.

Let $G_M(z, z')$ be the Green function of the Dirichlet boundary value problem in M for the ordinary Laplace operator. From the conformal invariance of this function and from the Riemann theorem we derive

LEMMA 2. *The following formula is valid*:

$$G_M(z, z') = -\frac{1}{2\pi} \ln \frac{|\mathscr{S}_M(z) - \mathscr{S}_M(z')|}{|\mathscr{S}_M(z) - \overline{\mathscr{S}_M(z')}|},$$

where the bar above $\mathscr{S}_M(z')$ *denotes complex conjugation.*

Let $r(z, z'; s; \Gamma_M)$ be the kernel of the resolvent $\left(A(\Gamma_M; 1) - s(1-s)I\right)^{-1}$ at some regular point $s(1-s)$ (for example, $\operatorname{Re} s > 1$).

THEOREM 9. (see [22]). *We have*:

$$G_M(z, z') = \lim_{s \to 1+0} [r(z, z'; s; \Gamma_M) - r(-\bar{z}, z'; s; \Gamma_M)].$$

Now we introduce the *Siegel–Selberg series* as I call it (see [9]):

$$F_n(z, s; \Gamma_M) = \sum_{\gamma \in \Gamma_\infty \backslash \Gamma_M} \left(\exp 2\pi i n w(\gamma z)\right) \sqrt{y(\gamma z)}\ I_{s-1/2}\left(2\pi |n|\, y(\gamma z)\right);$$

this series converges absolutely in $\operatorname{Re} s > 1$. Here $n \in \boldsymbol{Z}$, $z = \operatorname{Re} z + i \operatorname{Im} z = w(z) + iy(z)$, $I_s(z)$ is the modified Bessel function of the first kind. The second principal statement of this part is

THEOREM 10. *The following formula is valid*:

$$\operatorname{Im} \mathscr{S}_M(z) A_1^{-1} = \pi i \lim_{s \to 1+0} \left(F_1(z, s; \Gamma_M) - F_{-1}(z, s; \Gamma_M)\right).$$

The proof of this theorem is based on Lemma 2 and Theorem 9. It requires also the use of an asymptotic expansion for $r(z, z'; s; \Gamma_M)$ when $\operatorname{Im} z \to \infty$ (see [22]).

In this way the problem of finding the Fourier coefficients of $\mathscr{S}_M(z)$ is reduced to that of finding the Fourier coefficients of the Siegel–Selberg series. These latter coefficients are well known from the papers by D. Niebur [9] and J. Fay [1]. Finally, consider the Kloosterman sums

$$S(m, n; c) = \sum_{0 \leqslant d < c} \exp 2\pi i (ma/c + nd/c)$$

where $\gamma z = (az+b)(cz+d)^{-1}$, $\gamma \in \Gamma_\infty \backslash \Gamma_M / \Gamma_\infty$, $\gamma \notin \Gamma_\infty$. By the Riemann–Roch theorem we have:

LEMMA 3. *For any group* Γ_M, *there are no cuspforms of weight* 2.

From Theorem 10, Lemma 3 and the results of papers [9], [1] we obtain (see [22]):

THEOREM 11. *For the Fourier coefficients of the Klein invariant* $\mathcal{J}_M(z)$ *the following formula is valid*:

$$A_k A_{-1}^{-1} = \frac{2\pi}{\sqrt{k}} \sum_{c>0} c^{-1} S(-k, 1; c) I_1\left(\frac{4\pi\sqrt{k}}{c}\right), \quad k \geqslant 1.$$

Let us remark here that the last formula extends the Rademacher formula which he obtained for the case of a modular group (see [11]).

In conclusion, let us point out certain possible generalizations of the results of this part. The basic results, i.e. the statements of Theorems 8, 11, are extended to the case of an arbitrary group $\Gamma \in \mathfrak{M}$ of zero genus. The proof of Theorem 11 is essentially different. It is connected with the conformal invariance of a certain modified Neumann function. For an arbitrary cocompact Fuchsian group of zero genus one should be able to create a similar theory. The accessory coefficients should be expressed in terms of the asymptotic expansion coefficients of the modified automorphic resolvent kernel in the neighbourhood of the pole of the conformal mapping $\mathcal{J}(z)$. All the theory is expected to be very complicated in the situation of a non-zero genus. I hope that in this case the spectral theory of automorphic functions might prove useful for the effective construction of conformal mappings of multiply-connected Riemann surfaces onto some special canonical domains.

References

[1] Fay J. D., Fourier coefficients of the resolvent for a Fuchsian group, *J. Reine Angew. Math.* **293/294** (1977), pp. 143–203.

[2] Голубев В. В., *Лекции по аналитической теории дифференциальных уравнений*, Москва–Ленинград, 1950.

[3] Hejhal D. A., The Selberg trace formula and the Riemann zeta function, *Duke Math. J.* **43.3** (1976), pp. 441–482.

[4] Hurwitz A. and Courant R., *Funktionentheorie*, Springer, 1939.

[5] Klein F. and Fricke R., *Vorlesungen über die Theorie der Automorphen Funktionen*, B. 2, Teubner, 1901.

[6] Lehner J., Note on the Schwarz Triangle Functions, *Pacific Journ. of Math.* **4** (1954), pp. 243–249.

[7] Lehner J., *Discontinuous groups and automorphic functions*, A. M. S., Providence, 1964.

[8] Nehari Z., *Conformal mappings*, McGraw-Hill, 1952.

[9] Niebur D., A class of nonanalytic automorphic functions, *Nagoya Math. J.* **52** (1973), pp. 133–145.

[10] Poincaré H., Sur les groupes des équations linéaires, *Acta Math.* **4** (1884), pp. 201–312.
[11] Rademacher H., The Fourier coefficients of the modular invariant $J(\tau)$, *Amer. J. Math.* **60** (1938), pp. 501–512.
[12] Raleigh J., On the Fourier coefficients of triangle functions, *Acta Arith.* **8** (1962), pp. 107–111.
[13] Randol B., The Riemann hypothesis for Selberg's zeta-function and the asymptotic behaviour of eigenvalues of the Laplace operator, *Trans, AMS* **236**, No. **513** (1978), pp. 209–224.
[14] Roelcke W., Über die Wellengleichung bei Grenzkreisgruppen erster Art, *Sitz. Heidelberg Akad. Wiss. Math.-natur. Klasse*, Bd. **4**, 1956.
[15] Schwarz H. A., *Gesammelte Mathematische Abhandlungen*, B. 2, Springer, 1890.
[16] Selberg A., *Harmonic analysis, Teil* **2**, *Vorlesungsniederschrift*, Göttingen, 1954.
[17] Selberg A., Harmonic analysis and discontinuous groups in weakly symmetric Riemannian spaces with applications to Dirichlet series, *J. Indian Math. Soc.* **20** (1956), pp. 47–87.
[18] Selberg A., On the estimation of Fourier coefficients of modular forms, *Proc. Symp. Pure Math. AMS* **8** (1965), pp. 1–15.
[19] Венков А. Б., Об ассоциированных с определяющими уравнениями и непрерывными дробями рядах Дирихле в теории автоморфных функций, *Труды МИАН им. Стеклова* **158** (1981), pp. 31–44.
[20] Венков А. Б., Спектральная теория автоморфных функций, *Труды МИАН им. Стеклова* **153** (1981), pp. 1–171.
[21] Венков А. Б., Зограф П. Г., Об аналогах формул факторизации Артина в спектральной теории автоморфных функций связанных с индуцированными представлениями фуксовых групп, *ДАН СССР* **259**, № 3 (1981), pp. 523–526.
[22] Венков А. Б., О точных формулах для акцессорных коэффициентов в уравнении Шварца, *Функциональный анализ и его приложения* **17**, № 3 (1983), pp. 1–9.
[23] Зограф П. Г., О спектре автоморфных лапласианов в пространствах параболических функций, *ДАН СССР* **269**, № 4 (1983), pp. 802–805.

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

MICHÈLE VERGNE

Formule de Kirillov et indice de l'opérateur de Dirac

Dans cet article nous annonçons une formule pour l'indice équivariant de l'opérateur de Dirac et montrons son analogie avec la formule universelle proposée par Kirillov pour le caractère des représentations d'un groupe de Lie. Cette analogie suggère une généralisation de la formule de Kirillov au cas des orbites non génériques de la représentation coadjointe.

Ceci est un travail commun avec Nicole Berline.

Soit G un groupe de Lie d'algèbre de Lie $\mathfrak{g}$. On considère la fonction analytique, G-invariante, définie sur $\mathfrak{g}$ par $j(X) = \det\left(\frac{e^{\mathrm{ad}X/2} - e^{-\mathrm{ad}X/2}}{\mathrm{ad}X}\right)$. Elle admet, au moins dans un voisinage de 0, une racine carrée analytique $j^{1/2}$, telle que $j^{1/2}(0) = 1$.

Supposons G unimodulaire et de type I. Soit $\hat{G}$ l'ensemble des classes de représentations unitaires irréductibles de G. Kirillov a conjecturé la formule suivante [16]: pour presque tout $T \in \hat{G}$ (au sens de la mesure de Plancherel) il existe une orbite $\mathcal{O}$ de G dans le dual $\mathfrak{g}^*$ de $\mathfrak{g}$ telle que, dans un voisinage de 0 dans $\mathfrak{g}$, on ait l'égalité de fonctions généralisées

$$\mathrm{tr}\, T(\exp X) j^{1/2}(X) = \int_{\mathcal{O}} e^{i\langle f, X\rangle} d\beta_{\mathcal{O}}(f), \tag{1}$$

en notant $\beta_{\mathcal{O}}$ une mesure invariante sur l'orbite $\mathcal{O}$, que nous préciserons dans la suite. La validité de cette conjecture a été établie dans de nombreux cas [10, 16, 14, 15].

Cette conjecture est un des aspects de la méthode des orbites qui consiste à mettre en correspondance, génériquement tout au moins, représen-

tations unitaires irréductibles d'un groupe de Lie et orbites de la représentation coadjointe. Dans cette correspondance, à une représentation de la série discrète est associée une orbite $\mathcal{O} = G \cdot f$ admissible dont le stabilisateur est compact [11]. Réciproquement, supposons que $\mathcal{O}$ admette une structure riemannienne G-invariante, et soit $\mathcal{V} = \mathcal{V}_\tau$ un module de Clifford admissible sur $\mathcal{O}$ (voir 2.12). Considérons la représentation virtuelle $T_{\mathcal{V}}$ dans la différence des noyaux L^2 des opérateurs de Dirac $D^{\pm}$. Dans de nombreux cas, cette représentation est la représentation $T_{f,\tau}$ de Duflo. La formule universelle de Kirillov dicte une formule pour la trace (au sens fonctions généralisées) de la représentation $T_{\mathcal{V}}$, c'est-à-dire une formule pour l'indice équivariant de l'opérateur de Dirac D.

Il parait donc naturel de rechercher une formule similaire à celle de Kirillov pour l'indice équivariant d'un complexe elliptique G-équivariant sur une variété M. Nous avons montré dans [3], [6], comment la notion de formes G-équivariantes permet de proposer une telle formule universelle pour certains complexes classiques.

Nous montrons ici comment écrire la formule de Kirillov en termes de formes équivariantes. Si l'orbite $\mathcal{O}$ satisfait certaines conditions, on peut en effet écrire la formule de Kirillov sous la forme:

$$\operatorname{tr} T(\exp X) = \int_{\mathcal{O}} \operatorname{Ch}(X, \tau) \mathcal{J}^{-1/2}(X, T\mathcal{O}), \tag{2}$$

où $\operatorname{Ch}(X, \tau)$ et $\mathcal{J}^{-1/2}(X, T\mathcal{O})$ sont des classes de cohomologie équivariante par rapport à $X \in \mathfrak{g}$ définies dans la section 1. Dans le cas où l'orbite $\mathcal{O}$ est de dimension maximale, le terme $\mathcal{J}^{-1/2}(X, T\mathcal{O})$ se réduit à la constante $j^{-1/2}(X)$ et $\int_{\mathcal{O}} \operatorname{Ch}(X, \tau)$ coïncide avec $\int_{\mathcal{O}} e^{i\langle f, X\rangle} d\beta_{\mathcal{O}}(f)$. Un exemple de Khalgui [14] montre que pour une orbite non générique la formule (1) est fausse, même si on remplace $j^{1/2}$ par une autre fonction analytique G-invariante.

L'écriture ci-dessus suggère que, pour une orbite non générique, la formule de Kirillov doit être modifiée pour tenir compte du terme $\mathcal{J}^{-1/2}(X, T\mathcal{O})$.

Malheureusement, nous devons souligner les défauts de notre interprétation dans son état actuel:

(1) Nous ne définissons le terme $\mathcal{J}^{-1/2}(X, T\mathcal{O})$ que sous des conditions probablement trop restrictives sur l'orbite $\mathcal{O}$.

(2) Dans le cas où $\mathcal{J}^{-1/2}(X, T\mathcal{O})$ est défini, nous ne pouvons assurer que le membre de droite de (2) définisse une fonction généralisée.

1. Cohomologie et classes caractéristiques équivariantes

1.1. Soit M une variété différentielle. On note $\mathscr{A}(M) = \oplus \mathscr{A}^r(M)$ l'algèbre sur C des formes différentielles; on note $\mathscr{A}^+(M)$ la sous-algèbre, commutative, des formes paires; on note d la dérivation extérieure. Si ξ est un champ de vecteurs, on note $c(\xi)$: $\mathscr{A}(M) \to \mathscr{A}(M)$ la contraction, $\mathscr{L}(\xi)$ la dérivation de Lie. On a $\mathscr{L}(\xi) = d \cdot c(\xi) + c(\xi) \cdot d$.

Soit G un groupe de Lie agissant sur M à gauche. Soit $\mathfrak{g}$ l'algèbre de Lie de G. Pour $X \in \mathfrak{g}$ on note $X^* = X^*_M$ le champ de vecteurs sur M défini par $(X^* f)(m) = (d/dt) f(\exp tX \cdot m)|_{t=0}$. Soit $\mathscr{A}_X$ la sous-algèbre des formes $\mu \in \mathscr{A}(M)$ telles que $\mathscr{L}(X^*)\mu = 0$.

On rappelle les définitions de [3]. L'opérateur (non homogène) $d_X = d - 2i\pi c(X^*)$ sur $\mathscr{A}(M)$ est une antidérivation qui inverse la parité des formes et vérifie $(d - 2i\pi c(X^*))^2 = -2i\pi \mathscr{L}(X^*)$. Il est donc de carré nul sur $\mathscr{A}_X$. On pose

$$Z(M, d_X) = \ker (d - 2i\pi c(X^*)),$$
$$B(M, d_X) = (d - 2i\pi c(X^*)) \mathscr{A}_X.$$

On a donc $B(M, d_X) \subset Z(M, d_X) \subset \mathscr{A}_X$. On pose

$$H^*(M, d_X) = Z(M, d_X)/B(M, d_X).$$

Il est clair que si $X^*_M = 0$, l'anneau $H^*(M, d_X)$ est l'anneau de cohomologie ordinaire de M.

1.2. La cohomologie de d_X est particulièrement simple à décrire lorsque M est compacte et que le groupe à un paramètre $\exp tX$ est relativement compact [6]. Dans ce cas les zéros de X^*_M forment une sous-variété M_0 de M et on a la proposition suivante:

1.3. PROPOSITION. *L'application i^*: $H^*(M, d_X) \to H^*(M_0)$ est un isomorphisme.*

La formule de localisation [3], [6] généralise un résultat de R. Bott [7], [8]. Nous l'énonçons ici dans un cas particulier.

Si $\mu = \sum \mu^{[r]} \in \mathscr{A}(M)$ et si N est une sous-variété compacte orientée de M, on écrit $\int_N \mu$ pour $\int_N \mu^{[\dim N]}$. Si m est un point de M, on pose $\mu(m) = \mu^{[0]}(m)$. Si N est une sous-variété compacte orientée invariante par le groupe à un paramètre $\exp tX$, l'application $\mu \to \int_N \mu$ est bien définie sur $H^*(M, d_X)$.

Si $m \in M$ est un zéro du champ de vecteurs X^*, la dérivation de Lie $\mathscr{L}(X^*)$ induit un endomorphisme $L_m(X)$ de l'espace tangent T_mM.

Supposons que le groupe G soit compact. Si $X \in \mathfrak{g}$ et si m est un zéro de X^* tel que $L_m(X)$ soit inversible, il existe une base $e_1, e_2, \ldots, e_{2n-1}, e_{2n}$ de T_mM telle que:

$$L_m(X)e_{2j-1} = \lambda_j e_{2j},$$

$$L_m(X)e_{2j} = -\lambda_j e_{2j-1}.$$

On suppose cette base d'orientation positive. Alors, le produit $\lambda_1 \ldots \lambda_n$ ne dépend pas du choix d'une telle base. On pose:

$$\chi\big(L_m(X)\big) = i^{-n}\lambda_1\lambda_2 \ldots \lambda_n.$$

On dit que X^* est *non dégénéré* lorsque $L_m(X)$ est inversible pour tout zéro m de X^*. Rappelons alors la

1.4. Proposition. *Soit G un groupe de Lie compact agissant sur une variété compacte orientée de dimension $2n$. Soit $X \in \mathfrak{g}$ tel que le champ de vecteurs X_M^* soit non dégénéré. Soit $\mu \in H^*(M, d_X)$. Alors*

$$\int_M \mu = \sum_{\text{zéros de } X^*} \frac{\mu(m)}{\chi\big(L_m(X)\big)}.$$

1.5. *Remarque.* La proposition 1.4 est aussi obtenue dans [1] par les méthodes de la cohomologie T-équivariante ([19], voir aussi [23]).

1.6. Revenons au cas où G est un groupe de Lie quelconque. Une application $X \mapsto \mu_X$ de $\mathfrak{g}$ dans $\mathscr{A}(M)$ sera appelée une *forme équivariante* si

$$\mu_X \in Z(M, d_X),$$

$$\mu_{g \cdot X} = g \cdot \mu_X \quad \text{pour } g \in G, X \in \mathfrak{g}.$$

Donnons dès maintenant un exemple:

1.7. Soit (M, σ) une variété symplectique, munie d'une action hamiltonienne d'un groupe de Lie G. Notons f_X le moment de $X \in \mathfrak{g}$. Par définition on a $c(X^*)\sigma + df_X = 0$ et comme $d\sigma = 0$, il en résulte que $f_X - (\sigma/2i\pi)$ est un élément de $Z(M, d_X)$.

En particulier, soit $\mathcal{O} \subset \mathfrak{g}^*$ une orbite de la représentation coadjointe de G. Pour la structure symplectique canonique de $\mathcal{O}$, le moment de $X \in \mathfrak{g}$ est donné par $f_X(l) = -\langle l, X\rangle$ pour $l \in \mathcal{O}$. L'élément $\exp\left[-i\left(f_X - (\sigma/2i\pi)\right)\right] \in Z(M, d_X)$ jouera un rôle important dans la suite.

1.8. Rappelons comment l'analogue équivariant de la construction de Chern–Weil fournit des éléments de $H^*(M, d_X)$ ([5]).

Soit H un groupe de Lie d'algèbre de Lie $\mathfrak{h}$. Soit $P \to M$ un fibré principal de fibre H sur lequel H opère à droite. On suppose P muni d'une action à gauche du groupe G. On fait l'hypothèse que P admet une 1-forme de connexion α invariante par G. (Cette hypothèse est toujours satisfaite si G est compact, mais il serait préférable de l'éviter dans la situation générale). Pour $X \in \mathfrak{g}$ on définit le moment de X par $J_X = c(X_P^*)\alpha$. On note Ω la courbure de α. Soit Φ une fonction polynômiale H-invariante sur $\mathfrak{h}$. On définit par multilinéarité la forme $\Phi\left(J_X - (\Omega/2i\pi)\right)$ sur P. Cette forme se projette en une forme sur M notée $\Phi(X, \alpha)$, qui appartient à $Z(M, d_X)$ et dont la classe dans $H^*(M, d_X)$ ne dépend pas de la connexion G-invariante α choisie. On note cette classe $\Phi(X, P)$. On peut encore définir $\Phi(X, P)$ lorsque Φ est un germe de fonction analytique en 0 sur $\mathfrak{h}$: si Φ est entière, $\Phi(X, P)$ est une forme sur M, dont les coefficients dépendent analytiquement de X; si le rayon de convergence de Φ est fini, on peut définir $\Phi(X, P)$ sur tout ouvert relativement compact de M, pour X assez petit. Si on note $\hat{I}(\mathfrak{h})$ l'algèbre des fonctions entières H-invariantes sur $\mathfrak{h}$, l'application $\Phi \mapsto \Phi(X, P)$ est un homomorphisme d'algèbres de $\hat{I}(\mathfrak{h})$ dans $H^*(M, d_X)$.

1.9. Fibré trivial. Supposons que P soit le fibré trivial $M \times H$, l'action de G étant donnée par $g(m, h) = (gm, \gamma(g)h)$, pour un homomorphisme γ de G dans H. Notons aussi γ l'homomorphisme de $\mathfrak{g}$ dans $\mathfrak{h}$ qui s'en déduit. Soit α la connexion (plate) sur P image réciproque de la forme de Maurer–Cartan sur H. Alors α est G-invariante et on vérifie immédiatement que $\Phi(X, \alpha)$ est la fonction constante sur M égale à $\Phi\left(\gamma(X)\right)$.

1.10. Homomorphisme de fibrés. Un homomorphisme $H \to H'$ de groupes de Lie définit de manière naturelle un homomorphisme du fibré P dans le fibré principal $P' = P \times_H H'$ de fibre H'. L'action de G sur P se transporte naturellement à P'. Si P' admet une connexion G-invariante, on identifie, par abus de notation, les formes associées (1-forme de connexion, courbure, moment, etc.) et les formes sur P qui sont leurs images réciproques par l'homomorphisme $P \to P'$.

1.11. Fibrés vectoriels. Soient V un espace vectoriel, réel ou complexe, et $\mathscr{V}\to M$ un fibré vectoriel de fibre-type V. Le fibré principal associé a pour fibre $H = \mathrm{GL}(V)$. Si G agit sur $\mathscr{V}$ en préservant une connexion linéaire, à toute fonction Φ H-invariante sur $\mathfrak{h} = gl(V)$ est associée par 1.8 une classe $\Phi(X, \mathscr{V}) \in H^*(M, d_X)$. En particulier, on notera $\mathrm{Ch}(X, \mathscr{V})$ (*caractère de Chern*) la classe associée à la fonction $A \mapsto \mathrm{tr} e^A$.

Soit $Q(z) = \sum a_n z^n$ une fonction analytique d'une variable z. Pour tout espace vectoriel V, la fonction $A \mapsto \det Q(A)$ sur $gl(V)$ est invariante. Soient $\mathscr{V}_i \to M$ $(i = 1, 2)$ deux fibrés G-équivariants et admettant des connexions G-invariantes. Soit $\mathscr{V}_1 \oplus \mathscr{V}_2$ leur somme de Whitney. Il est clair qu'on a:

1.12. Lemme.

$$\det Q(X, \mathscr{V}_1 \oplus \mathscr{V}_2) = \det Q(X, \mathscr{V}_1) \det Q(X, \mathscr{V}_2).$$

1.13. La fonction $j(z) = (e^{z/2} - e^{-z/2})/z$ et sa racine carrée $j^{1/2}$ définie (au voisinage de $z = 0$) par $j^{1/2}(0) = 1$ apparaissent dans la formule du caractère et de l'indice de l'opérateur de Dirac. On pose

$$\mathscr{J}^{1/2}(X, \mathscr{V}) = \det j^{1/2}(X, \mathscr{V}),$$
$$\mathscr{J}^{-1/2}(X, \mathscr{V}) = \det j^{-1/2}(X, \mathscr{V}).$$

Sur les algèbres de Lie $so(n) \subset gl(n)$ et $sp(n) \subset gl(n)$ la fonction $A \mapsto \det j(A)$ admet une racine carrée analytique entière. Si donc $\mathscr{V}$ est associé à un fibré principal G-équivariant de fibre $\mathrm{SO}(n)$ ou $\mathrm{Sp}(n)$ et admettant une connexion G-invariante, la forme $\mathscr{J}^{1/2}(X, \mathscr{V})$ est définie sur M toute entière et dépend analytiquement de X.

1.14. Notons $\mathrm{DL}(n, C)$ le groupe $\mathrm{GL}(n, C)/\pm 1$. Un fibré principal $\mathscr{W}$ de groupe $\mathrm{DL}(n, C)$ sera appelé *pseudo-fibré vectoriel.* Si $\mathscr{W}$ est G-équivariant et admet une connexion G-invariante, on peut encore définir le caractère de Chern $\mathrm{Ch}(X, \mathscr{W})$.

Nous pouvons maintenant énoncer la formule pour l'indice équivariant de l'opérateur de Dirac obtenue dans [6].

Soient M une variété riemanienne compacte orientée de dimension $2l$ et G un groupe connexe d'isométries de M préservant l'orientation. Soit $\mathscr{V} = \mathscr{V}^0 \oplus \mathscr{V}^1$ un fibré gradué de Clifford [2], G-équivariant, sur M. Pour chaque $m \in M$, la fibre $\mathscr{V}_m$ est donc un module gradué pour l'algèbre de Clifford $C(T_m M)$. Notons S_m^+ (resp. S_m^-) l'espace des spineurs pairs (resp.

impairs). Pour chaque $m \in M$, on a des décompositions

$$V_m^0 = S_m^+ \otimes W_m \quad \text{et} \quad V_m^1 = S_m^- \otimes W_m.$$

Les espaces $W_m/\pm 1$ définissent un pseudo-fibré vectoriel G-équivariant $\mathscr{W}$ [6].

Soit $D_{\mathscr{V}}^0 : \Gamma(\mathscr{V}^0) \to \Gamma(\mathscr{V}^1)$ l'opérateur de Dirac associé au module de Clifford $\mathscr{V}$, défini à l'aide d'une connexion G-invariante. Alors $\operatorname{Ker} D_{\mathscr{V}}^0$ et $\operatorname{Coker} D_{\mathscr{V}}^0$ sont des G-modules de dimension finie et on a:

1.15. Théorème. *Si $X \in \mathfrak{g}$ est suffisamment petit,*

$$\operatorname{tr}_{\operatorname{Ker} D_{\mathscr{V}}^0}(\exp X) - \operatorname{tr}_{\operatorname{Coker} D_{\mathscr{V}}^0}(\exp X) = \int_M \operatorname{Ch}(X, \mathscr{W}) \mathscr{J}^{-1/2}(X, TM).$$

1.16. Fibrés homogènes. On suppose désormais que M est un espace homogène G/H. Une connexion G-invariante sur le fibré principal $G \to G/H$ est définie par une décomposition $\mathfrak{g} = \mathfrak{h} \oplus \mathfrak{m}$ où $\mathfrak{m}$ est un supplémentaire H-invariant de $\mathfrak{h}$.

1.17. Soit τ une représentation de H dans un espace vectoriel de dimension finie V_τ. Une connexion linéaire G-invariante sur le fibré vectoriel $G \times_H V_\tau$ est une connexion sur le fibré principal $G \times_H \mathrm{GL}(V_\tau)$, image du fibré principal $G \to M$ par l'homomorphisme $H \to \mathrm{GL}(V_\tau)$. La 1-forme sur G à valeurs dans $gl(V_\tau)$ qui correspond à cette connexion (conventions de 1.10) est définie par une application linéaire $\tilde{\tau} : T_e G = \mathfrak{g} \to gl(V_\tau)$ qui vérifie:

1.18.
$$\tilde{\tau}|_{\mathfrak{h}} = \tau,$$
$$\tilde{\tau}(\operatorname{ad} h \cdot X) = \tau(h) \tilde{\tau}(X) \tau(h)^{-1} \text{ pour tout } X \in \mathfrak{g}, h \in H.$$

Si le fibré principal $G \to G/H$ admet une connexion G-invariante, on obtient une telle application $\tilde{\tau}$ en posant $\tilde{\tau}|_{\mathfrak{m}} = 0$.

Toujours avec les conventions de 1.10, le moment de $X \in \mathfrak{g}$ et la courbure de la connexion $\tilde{\tau}$ sont les formes différentielles sur G données par:

1.19.
$$J_X(g) = \tilde{\tau}(\operatorname{ad} g^{-1} \cdot X) \quad \text{pour } g \in G.$$

1.20. Ω est G-invariante par les translations à gauche et

$$\Omega(Y, Z) = [\tilde{\tau}(Y), \tilde{\tau}(Z)] - \tilde{\tau}[Y, Z] \quad \text{pour } Y, Z \in \mathfrak{g}.$$

1.21. L'étude des orbites admissibles de la représentation coadjointe demande de considérer le cas où τ est une représentation non pas du groupe H lui-même, mais d'un revêtement à deux feuillets $\tilde{H}$, telle que $\tau(\tilde{e})$ soit égal à ± 1 pour tout élément $\tilde{e} \in \tilde{H}$ qui se projette sur l'élément neutre de H. Alors τ définit un homomorphisme de H dans le quotient $\mathrm{GL}(V_\tau)/(\pm 1)$ de $\mathrm{GL}(V_\tau)$. Une application $\tilde{\tau}: \mathfrak{g} \to gl(V_\tau)$ qui vérifie les conditions 1.18, avec $\tilde{H}$ à la place de H, définit une connexion G-invariante sur le fibré principal $G \times_H (\mathrm{GL}(V_\tau)/(\pm 1))$. À toute fonction $\Phi \in \hat{I}(gl(V_\tau))$ est donc associée, pour $X \in \mathfrak{g}$, une classe dans $H^*(M, d_X)$ notée $\Phi(X, \tau)$. Cette classe provient de la forme $\Phi(J_X - (\Omega/2i\pi))$ sur G, où J_X et Ω sont données par 1.19 et 1.20.

2. Classes caractéristiques pour les orbites de la représentation coadjointe et formule du caractère

2.1. Soit $\mathcal{O} \subset \mathfrak{g}^*$ une orbite de la représentation coadjointe de G. On munit $\mathcal{O}$ de sa structure symplectique canonique σ. Soit $f \in \mathcal{O}$. Grâce à l'homomorphisme $G(f) \to \mathrm{Sp}(\mathfrak{g}/\mathfrak{g}(f)) \hookrightarrow \mathrm{SL}(\mathfrak{g}/\mathfrak{g}(f))$, le revêtement à 2 feuillets $\widetilde{\mathrm{SL}}(\mathfrak{g}/\mathfrak{g}(f)) \to \mathrm{SL}(\mathfrak{g}/\mathfrak{g}(f))$ induit un revêtement à 2 feuillets $\tilde{G}(f) \to G(f)$, qui coïncide avec le revêtement défini par M. Duflo [11]. On définit, en suivant [11], l'ensemble $\mathcal{X}(f)$ des [classes de] représentations unitaires irréductibles τ de $\tilde{G}(f)$ telles que $\tau(\exp X) = e^{i\langle f, X\rangle} \mathrm{Id}_{V_\tau}$ pour $X \in \mathfrak{g}(f)$ et $\tau(\varepsilon) = -\mathrm{Id}_{V_\tau}$, en notant ε l'élément non neutre de $\tilde{G}(f)$ au-dessus de l'élément neutre de $G(f)$. L'orbite est dite *admissible* lorsque l'ensemble $\mathcal{X}(f)$ n'est pas vide. Si G est algébrique, $\mathcal{X}(f)$ consiste en un nombre fini de représentations de dimension finie. Plaçons-nous dans ce cas. Notons $P_\tau = G \times_{G(f)} (\mathrm{GL}(V_\tau)/(\pm 1))$ le fibré principal défini par $\tau \in \mathcal{X}(f)$.

2.2. Rappelons que P_τ admet une connexion G-invariante canonique, introduite par B. Kostant [17], obtenue en posant $\tilde{\tau}(X) = i\langle f, X\rangle \mathrm{Id}_{V_\tau}$ pour $X \in \mathfrak{g}$ (notations de 1.18). Comme $\mathrm{ad}\, g \cdot f = f$ si $g \in G(f)$, la courbure Ω et le moment J_X de cette connexion sont en fait déjà des formes sur $\mathcal{O} = G/G(f)$, données par

$$J_X(l) = i\langle l, X\rangle \mathrm{Id}_{V_\tau} \quad \text{pour } l \in \mathcal{O},^1$$

$$\Omega = -i\sigma \mathrm{Id}_{V_\tau}.$$

---

[1] Le moment de X relativement à cette connexion diffère donc par un facteur $-i$ du moment défini en 1.7 relativement à l'action hamiltonienne de G sur $\mathcal{O}$.

Le caractère de Chern du fibré P_τ est donc la forme sur $\mathcal{O}$ donnée par

2.3. $$\mathrm{Ch}(X,\tau) = \mathrm{tr}(e^{Jx-\Omega/2i\pi}) = \dim V_\tau e^{i\langle l,X\rangle} e^{\sigma/2\pi}.$$

On voit donc apparaître le premier terme de la formule (2) de l'introduction. Pour donner un sens au terme $\mathcal{J}^{-1/2}(X, T\mathcal{O})$, nous allons faire momentanément l'hypothèse suivante.

2.4. Le fibré $G\to\mathcal{O} = G/G(f)$ admet une connexion G-invariante, donnée par une décomposition $G(f)$-invariante $\mathfrak{g} = \mathfrak{g}(f)\oplus\mathfrak{m}$.

Considérons la suite exacte de $G(f)$-modules:

$$0\to(\mathrm{ad}\,\mathfrak{g})\cdot f\to\mathfrak{g}^*\to\mathfrak{g}(f)^*\to 0.$$

La suite exacte de fibrés vectoriels associée définit le fibré normal $N\mathcal{O}$ à $\mathcal{O}$ dans $\mathfrak{g}^*$:

$$0\to T\mathcal{O}\to\mathcal{O}\times\mathfrak{g}^*\to N\mathcal{O}\to 0.$$

L'action de G sur $\mathcal{O}\times\mathfrak{g}^*$ est donnée par $g(l,l') = (\mathrm{ad}^*g\cdot l, \mathrm{ad}^*g\cdot l')$. La forme $\mathcal{J}^{-1/2}(X, \mathcal{O}\times\mathfrak{g}^*)$ est donc la fonction constante sur $\mathcal{O}$ donnée par:

$$\det_{\mathfrak{g}^*}\left(\frac{e^{\mathrm{ad}^*X/2} - e^{-\mathrm{ad}^*X/2}}{\mathrm{ad}^*X}\right)^{-1/2} = j^{-1/2}(X).$$

Grâce à l'hypothèse 2.4 on a, en utilisant 1.12:

2.5. $$j^{1/2}(X) = \mathcal{J}^{1/2}(X, T\mathcal{O})\mathcal{J}^{1/2}(X, N\mathcal{O}).$$

Si $f\in\mathcal{O}$ et $X\in\mathfrak{g}(f)$, i.e. si f est un zéro de X^*, on a

$$\mathcal{J}^{1/2}(X, N\mathcal{O})^{[0]}(f) = \det_{\mathfrak{g}(f)}\left(\frac{e^{\mathrm{ad}X/2} - e^{-\mathrm{ad}X/2}}{\mathrm{ad}X}\right)^{1/2}.$$

2.6. Considérons le cas d'une orbite de dimension maximale $2d$. Alors $\mathfrak{g}(f)$ est commutative [12]. Il en résulte, d'après 1.17, que le fibré normal $N\mathcal{O} = G\times_{G(f)}\mathfrak{g}(f)^*$ admet une connexion G-invariante plate, et on a $\mathcal{J}^{1/2}(X, N\mathcal{O}) = 1$.

Dans ce cas, sous l'hypothèse 2.4, la forme $\mathcal{J}^{-1/2}(X, T\mathcal{O})$ se réduit donc à la fonction constante $j^{-1/2}(X)$.

La formule de Kirillov:

$$\mathrm{tr}\,T(\exp X)j(X)^{1/2} = \dim\tau\int_{\mathcal{O}} e^{i\langle\cdot,X\rangle}\beta_{\mathcal{O}},$$

où

$$\beta_{\mathcal{O}} = \frac{\sigma^d}{(2\pi)^d d!},$$

s'écrit donc aussi en termes de formes G-équivariantes:

$$\operatorname{tr} T(\exp X) = \int_{\mathcal{O}} \operatorname{Ch}(X, \tau)\mathcal{J}^{-1/2}(X, T\mathcal{O}).$$

Ce qui précède nous conduit à proposer une généralisation de la conjecture de Kirillov (sous l'hypothèse 2.4). Nous la proposons sous deux formes:

2.7. CONJECTURE. *Soit $\mathcal{O}$ une orbite admissible, fermée, de G dans $\mathfrak{g}^*$, $\tau \in \mathcal{X}(f)$. Il existe une représentation unitaire irréductible $T_{f,\tau}$ de G telle que:*

$$\operatorname{tr} T_{f,\tau}(\exp X) = \int_{\mathcal{O}} \operatorname{Ch}(X, \tau)\mathcal{J}^{-1/2}(X, T\mathcal{O}) \tag{C.1}$$

ou plutôt

$$\operatorname{tr} T_{f,\tau}(\exp X) j^{1/2}(X) = \int_{\mathcal{O}} \operatorname{Ch}(X, \tau)\mathcal{J}^{1/2}(X, N\mathcal{O}). \tag{C.2}$$

Indiquons maintenant des situations où cette formule est justifiée.

2.8. Tout d'abord si $\mathcal{O}$ est fermée de dimension maximale, (C.2) est la formule de Kirillov, et sa validité a été établie dans [15] par Khalgui pour les représentations $T_{f,\tau}$ de Duflo.

2.9. Si $\mathcal{O} = \{0\}$, on a $\mathcal{J}^{1/2}(X, N\mathcal{O}) = j(X)^{1/2}$ et la formule (C.2) est donc vérifiée si on associe à l'orbite $\{0\}$ la représentation triviale.

2.10. Si $\mathfrak{g}(f)$ est réductive, l'hypothèse 2.4 est satisfaite et la forme $\mathcal{J}^{1/2}(X, N\mathcal{O})$ est définie sur $\mathcal{O}$ toute entière et dépend analytiquement de $X \in \mathfrak{g}$. Malheureusement, nous ne pouvons assurer que l'expression

$$\int_{\mathcal{O}} \operatorname{Ch}(X, \tau)\mathcal{J}^{1/2}(X, N\mathcal{O})$$

définisse une fonction généralisée sur un voisinage de 0 dans $\mathfrak{g}$.

Cependant, supposons que le sous-groupe à un paramètre $\mathrm{ad}(\exp tX)$ soit relativement compact dans $\mathrm{ad}G$, et que le champ de vecteurs $X^*_{\mathcal{O}}$ n'ait qu'un nombre fini de zéros. Soient $f \in \mathcal{O}$ et $X \in \mathfrak{g}(f)$. Soit W_f un sous-espace de $(\mathfrak{g}/\mathfrak{g}(f))_C$, stable par $\exp tX$, totalement isotrope positif pour la 2-forme canonique. La formule de localisation 1.4 s'écrit alors formellement:

$$\int_{\mathcal{O}} \mathrm{Ch}(X, \tau)\mathscr{J}^{1/2}(X, N\mathcal{O})j(X)^{-1/2} = \sum_{f \in \{\text{zéros de } X^*_{\mathcal{O}}\}} \dim V_\tau \frac{e^{i\langle f, X\rangle}}{\det_{W_f}(e^{\mathrm{ad}X/2} - e^{-\mathrm{ad}X/2})}.$$

Supposons que G soit semi-simple, et soit λ un élément elliptique (non nécessairement régulier) de $\mathfrak{g}^*$. Associons à l'orbite $G\cdot\lambda$ la représentation T_λ de Zuckerman. (Il n'est pas démontré que T_λ soit unitarisable). Alors la formule du caractère [22] pour T_λ coïncide sur l'ouvert des éléments elliptiques réguliers de $\mathfrak{g}$ avec l'expression précédente.

En particulier, lorsque G est compact connexe, notre conjecture est bien vérifiée: si $\mathfrak{t}$ est une sous-algèbre de Cartan telle que $\lambda \in \mathfrak{t}^*$, la représentation T_λ associée au fibré de Clifford canonique sur l'orbite $G\cdot\lambda$ a pour poids extrême $i\lambda - \varrho_\lambda$, où ϱ_λ est la demi-somme des racines de $\mathfrak{t}_C$ dans $\mathfrak{g}_C$ telles que $\langle i\lambda, \alpha\rangle > 0$, et la formule (C.2) coïncide avec la formule du théorème 2.8.

2.11. Si le fibré normal n'admet pas de connexion G-invariante, le terme $\mathscr{J}^{1/2}(X, N\mathcal{O})$ n'est même pas défini. Il est vraisemblable qu'on peut améliorer la conjecture (C.2) en cherchant, en guise de $\mathscr{J}^{1/2}(X, N\mathcal{O})$, une fonction analytique G-invariante $X \to J^\#(X)$, à valeurs dans l'espace des formes différentielles sur $\mathcal{O}$, qui vérifie:

(a) $J^\#(X) \in Z(M, d_X)$,

(b) $[J^\#(X)]^{[0]}(f) = \left[\det_{\mathfrak{g}(f)} \frac{e^{\mathrm{ad}X/2} - e^{-\mathrm{ad}X/2}}{\mathrm{ad}X}\right]^{1/2}$ pour $f \in \mathcal{O}$, $X \in \mathfrak{g}(f)$.

Si G est compact, d'après la proposition 1.3, ces conditions déterminent la classe de $J^\#(X)$ dans $H^*(M, d_X)$ pour les éléments X dont les zéros sont isolés.

S'il existe une fonction G-invariante $j_{\mathcal{O}}$ sur $\mathfrak{g}$ telle que

$$j_{\mathcal{O}}(X) = \det_{\mathfrak{g}(f)}\left(\frac{e^{\mathrm{ad}X/2} - e^{-\mathrm{ad}X/2}}{\mathrm{ad}X}\right) \quad \text{pour } f \in \mathcal{O},\ X \in \mathfrak{g}(f),$$

un choix naturel pour $J^{\#}(X)$ sera la forme de degré 0, égale à la constante $j_{\mathcal{O}}(X)^{1/2}$. La formule des caractères obtenue est alors celle de Duflo [10]. Dans le cas contraire, $J^{\#}(X)$ devra nécessairement comporter des termes de degré supérieur.

2.12. Revenons enfin à notre motivation initiale. Soit $\mathcal{O}$ une orbite de dimension maximale de la représentation coadjointe. Supposons que $\mathcal{O}$ admette une structure riemannienne G-invariante. Il existe donc une forme euclidienne Q sur $\mathfrak{g}/\mathfrak{g}(f)$ invariante par $G(f)$. Le revêtement $\tilde{G}(f)$ de $G(f)$ (2.1) est aussi défini par le diagramme

$$\begin{array}{ccc} \tilde{G}(f) & \to & \mathrm{Spin}(Q) \\ \downarrow & & \downarrow \\ G(f) & \to & \mathrm{SO}(Q). \end{array}$$

Soit $\varrho = \varrho^+ \oplus \varrho^-$ la représentation de $\mathrm{Spin}(Q)$ dans l'espace des spineurs. Soit $\tau \in \mathcal{X}(f)$; les représentations $\tau \otimes \varrho^{\pm}$ sont alors des représentations de $G(f)$. On peut définir le module de Clifford

$$\mathscr{V}_\tau = G \times_{G(f)} (\tau \otimes \varrho)$$

(un tel module de Clifford sera dit *admissible*).

Le fibré $\mathscr{V}_\tau$ est muni d'une connexion G-invariante déduite de la connexion de Kostant et de la connexion de Levi-Cività. On peut alors définir l'opérateur de Dirac

$$D_\tau^{\pm} : \Gamma(\mathscr{V}_\tau^{\pm}) \to \Gamma(\mathscr{V}_\tau^{\mp}).$$

Notons $\mathrm{Ker}\, D_\tau^{\pm}$ le noyau L^2 de $D_\tau^{\pm}$. La représentation de G dans $\mathrm{Ker}\, D_\tau^{\pm}$ est une somme finie de représentations de la série discrète ([21], [18], [9]). De plus, si G est nilpotent ou semi-simple, [9] montre que la représentation $\mathrm{Ker}\, D_\tau^+ - \mathrm{Ker}\, D_\tau^-$ coïncide avec la représentation irréductible $T_{f,\tau}$ dont le caractère est donné par la formule de Kirillov ([20], voir aussi [4]).

Dans ce cas, la "formule universelle" donne donc une formule intégrale pour la fonction de Lefschetz

$$X \mapsto \mathrm{tr}_{\mathrm{Ker} D_\tau^+}(\exp X) - \mathrm{tr}_{\mathrm{Ker} D_\tau^-}(\exp X)$$

au sens fonctions généralisées.

Soit $\mathscr{V}_\tau$ un module de Clifford admissible sur un espace homogène $M = G/H$ tel que H soit compact. A. Connes et H. Moscovici ont obtenu

pour l'indice L^2 de l'opérateur de Dirac D_τ une formule utilisant la classe $\mathrm{Ch}(\tau)\mathscr{J}^{-1/2}(TM)$ [9]. Nous pensons que le théorème 1.15 est encore valable dans ce cas pour exprimer la fonction de Lefschetz de l'opérateur D_τ.

Bibliographie

[1] Atiyah M. and Bott R., The Moment Map and Equivariant Cohomology, *Topology* **23** (1984), pp. 1–28.

[2] Atiyah M., Bott R., and Shapiro A., Clifford Modules, *Topology* **3**, Suppl. 1 (1964). pp. 3–38.

[3] Berline N. et Vergne M., Classes caractéristiques équivariantes, formule de localisation en cohomologie équivariante, *C. R. Acad. Sci. Paris* **295** (1982), pp. 539–541.

[4] Berline N. et Vergne M., Fourier Transforms of Orbits of the Coadjoint Representation. In: *Representation Theory of Reductive Groups, Proceedings of the University of Utah Conference 1982*, Progress in Mathematics, 40, Birkhäuser, Boston, 1983, pp. 53–67.

[5] Berline N. et Vergne M., Zéros d'un champ de vecteurs et classes caractéristiques équivariantes, *Duke Math. J.* **50** (1983), pp. 539–549.

[6] Berline N. et Vergne M., The Equivariant Index and Kirillov's Character Formula, *Amer. J. Math.*, to appear.

[7] Bott R., Vector Fields and Characteristic Numbers, *Michigan Math. J.* **14** (1967), pp. 231–244.

[8] Bott R., A Residue Formula for Holomorphic Vector Fields, *J. Diff. Geometry* **4** (1967), pp. 311–332.

[9] Connes A. et Moscovici H., The L^2-Index Theorem for Homogeneous Spaces of Lie Groups, *Ann. of Math.* **115** (1982), pp. 291–330.

[10] Duflo M., Caractères des groupes et des algèbres de Lie résolubles, *Ann. Sci. École Norm. Sup.* **3** (1980), pp. 23–74.

[11] Duflo M., *Construction de représentations unitaires d'un groupe de Lie*, C.I.M.E. II ciclo 1980, Liguori editore, Napoli, 1982.

[12] Duflo M. et Vergne M., Une propriété de la représentation coadjointe d'une algèbre de Lie, *C. R. Acad. Sci. Paris* **268** (1969), pp. 583–585.

[13] Duistermaat J. and Heckman G., On the Variation of the Cohomology of the Symplectic Form on the Reduced Phase Space, *Invent. Math.* **69** (1982), pp. 259–268; addendum **72** (1983), pp. 153–158.

[14] Khalgui M. S., Sur les caractères des groupes de Lie à radical cocompact, *Bull. Soc. Math. France* **109** (1981), pp. 331–372.

[15] Khalgui M. S., Caractères des groupes de Lie, *J. Funct. Anal.* **47** (1982), pp. 64–77.

[16] Kirillov A. A., Characters of Unitary Representations of Lie Groups, *Funct. Anal. Appl.* **2** (2) (1967), pp. 40–55.

[17] Kostant B., Quantization and Unitary Representations. In: *Modern Analysis and Applications*, Lecture Notes in Math., 170, Springer, 1970, pp. 87–207.

[18] Parthasarathy R., Dirac Operator and the Discrete Series, *Ann. of Math.* **96** (1972), pp. 1–30.

[19] Quillen D., Spectrum of a Cohomology Ring I, II, *Ann. of Math.* **94** (1971), pp. 549–602.

[20] Rossmann W., Kirillov's Character Formula for Reductive Groups, *Invent. Math.* **48** (1973), pp. 207–220.
[21] Schmid W., On a Conjecture of Langlands, *Ann. of Math.* **93** (1971), pp. 1–42.
[22] Vogan D. and Zuckerman G., *Unitary Representations with Non Zero Cohomology*, preprint, 1982.
[23] Witten E., *Supersymmetry and Morse Theory*, preprint, Princeton University, 1982.

UNIVERSITÉ DE RENNES I, U.E.R. MATHÉMATIQUES ET INFORMATIQUE, BÂTIMENT DU 1er CYCLE, AVENUE DU GÉNÉRAL LECLERC, RENNES BEAULIEU, 35042 RENNES CÉDEX, FRANCE

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

RICHARD ASKEY*

Orthogonal Polynomials and Some Definite Integrals

The connection between sets of orthogonal polynomials that can be found explicitly, their recurrence relations, and some specific definite integrals and series that can be evaluated is very close. Some examples are given. The integrals are extensions of the beta integral, and a pair of series are the Rogers–Ramanujan identities.

1. Introduction

At the last International Congress, E. M. Nikishin [8] said it would be very useful to expand the list of weight functions whose orthogonal polynomials can be found explicitly. The specific weight function he suggested is $[e^{2\pi\sqrt{t}}-1]^{-1}dt$, $0<t<\infty$, whose moments are

$$\int_0^\infty t^n[e^{2\pi\sqrt{t}}-1]^{-1}dt = (2n+1)!\zeta(2n+2)/(2\pi)^{2n+2}$$
$$= (-1)^n B_{2n+2}/(4n+4). \tag{1.1}$$

I do not know how to find these polynomials, but polynomials orthogonal with respect to $e^{\pi\sqrt{t}}[e^{2\pi\sqrt{t}}-1]^{-1}dt$ can be found, and they are a special case of a much more general class of orthogonal polynomials that can be given explicitly as generalized hypergeometric series. There is a further extension, where the polynomials are basic hypergeometric series. Contained in this class is a set of polynomials introduced by L. J. Rogers in 1894 [10] and 1895 [11], and used by him to discover the Rogers–Ramanujan identities.

* Research supported in part by grants from the National Science Foundation of the United States.

2. Orthogonal polynomials

A set of polynomials $\{p_n(x)\}$ is orthogonal if

$$\int_{-\infty}^{\infty} p_n(x)p_m(x)d\alpha(x) = 0, \quad m \neq n \tag{2.1}$$

for a positive measure $d\alpha(x)$. Any set of orthogonal polynomials satisfies a three term recurrence relation

$$\begin{aligned} xp_n(x) &= A_n p_{n+1}(x) + B_n p_n(x) + C_n p_{n-1}(x), \\ p_{-1}(x) &= 0, \quad p_0(x) = 1, \end{aligned} \tag{2.2}$$

with $A_n C_{n+1} > 0$, A_n, B_n, C_{n+1} real, $n = 0, 1, \ldots$ Conversely any set of polynomials that satisfies (2.2) is orthogonal with respect to a positive measure (which may not be unique). Historical comments will not be given here. See [13, vol. 3, Comment to [68–1], pp. 866-869] for some.

3. Hypergeometric series and orthogonal polynomials

A series $\sum c_k$ is a hypergeometric series if c_{k+1}/c_k is a rational function of k. If

$$\frac{c_{k+1}}{c_k} = \frac{(k+a_1)\ldots(k+a_p)x}{(k+b_1)\ldots(k+b_q)(k+1)}, \quad c_0 = 1$$

then

$${}_pF_q\left[\begin{matrix} a_1, \ldots, a_p \\ b_1, \ldots, b_q \end{matrix}; x\right] = \sum_{k=0}^{\infty} \frac{(a_1)_k \ldots (a_p)_k}{(b_1)_k \ldots (b_q)_k} \frac{x^k}{k!}, \tag{3.1}$$

where the shifted factorial is defined by

$$(a)_k = \Gamma(k+a)/\Gamma(a). \tag{3.2}$$

Two hypergeometric series are contiguous if they differ by one in one parameter, and the other parameters are the same. Gauss showed that ${}_2F_1(a, b; c; x)$ and any two ${}_2F_1$ series contiguous to it are linearly related. Some of the linear relations can be interpreted as three term recurrence relations like (2.2), so there are orthogonal polynomials that can be represented as hypergeometric series. It is also possible to iterate the Gauss relations and obtain three term recurrence relations for other

orthogonal polynomials. Three examples follow:

$$K_n(x; p, N) = {}_2F_1(-n, -x; -N; 1-p^{-1}), \quad n = 0, 1, \ldots, N, \tag{3.3}$$

$$P_n^{(\alpha,\beta)}(x) = \frac{(\alpha+1)_n}{n!}\, {}_2F_1\left(\begin{matrix}-n, n+\alpha+\beta+1 \\ \alpha+1\end{matrix}; \frac{1-x}{2}\right), \tag{3.4}$$

$$S_n(x; a, b) = \left[\frac{x(a+1)-\zeta}{2a}\right]^n {}_2F_1\left[\begin{matrix}-n, b[\zeta-x(a+1)]/2a\zeta \\ b/a\end{matrix}; \frac{-2\zeta}{x(a+1)-\zeta}\right], \tag{3.5}$$

$$\zeta = [(a+1)^2x^2 - 4a]^{1/2}.$$

Notice that $K_n(x)$ and $S_n(x)$ have essentially the same form, ${}_2F_1(-n, a; c; t)$, but the parameters and the power series variable have completely different forms, and the orthogonality relations are very different. The orthogonality relations are

$$\sum_{x=0}^{N} K_n(x; p, N) K_m(x; p, N) \binom{N}{x} p^x (1-p)^{N-x} = 0, \qquad 0 \leqslant m \neq n \leqslant N, \quad 0 < p < 1, \tag{3.6}$$

$$\int_{-1}^{1} P_n^{(\alpha,\beta)}(x) P_m^{(\alpha,\beta)}(x)(1-x)^\alpha (1+x)^\beta dx = 0, \quad m \neq n, \alpha, \beta > -1. \tag{3.7}$$

The orthogonality relation for $S_n(x; a, b)$ is too complicated to state here. There is an absolutely continuous part of the measure supported on $(-c, c)$, $c = 2\sqrt{a}/(1+a)$. When $a > 1$, $b > 0$ that is the complete measure, and the orthogonality relation was found by Pollaczek and Szegö. When $0 < a < 1$, $b > 0$ there are also infinitely many discrete masses, at $x_k = \pm(b+2ak)[b+k(a+1)]^{-1/2}[b+ak(a+1)]^{-1/2}$, $k = 0, 1, \ldots$ When $a \to 0$ the absolutely continuous part of the measure disappears, and the resulting orthogonality was discovered independently by Carlitz, and Karlin and McGregor over twenty years ago. The general case was done by Askey and Ismail [3], using the recurrence relation satisfied by $S_n(x; a, b)$, Darboux's method to find asymptotic values for two linearly independent solutions to this second order difference equation, and the theorem of Markoff stated by Nikishin. See [3] for the explicit formulas and references. The sum in the case $a = 0$ was found by Euler as a limiting case of an identity of Lambert, and can be used to count the number of rooted labeled trees.

The first two of these polynomials contain special cases that arise as spherical functions on some two point homogeneous spaces, the compact connected rank one symmetric spaces in the first case, the space of N-tuples with $0, 1, \ldots, b-1$ as entries, using the Hamming distance in the second case. The third one arises as the random walk polynomials associated with a birth and death process whose parameters are linear functions of n. The recurrence relation in the third case is

$$x[(a+1)n+b)]S_n(x) = (an+b)S_{n+1}(x)+nS_{n-1}(x). \tag{3.8}$$

Chebyshev was the first to find a set of orthogonal polynomials that needs a higher hypergeometric series to represent them. He showed that

$$Q_n(x;\alpha,\beta,N) = {}_3F_2\left[\begin{matrix}-n, n+\alpha+\beta+1, -x\\ \alpha+1, -N\end{matrix};1\right] \tag{3.9}$$

satisfy

$$\sum_{x=0}^{N} Q_n(x)Q_m(x)\binom{x+\alpha}{x}\binom{N-x+\beta}{N-x} = 0, \quad 0 \leqslant m \neq n \leqslant N. \tag{3.10}$$

This orthogonality relation is a common extension of (3.6) and (3.7). The most general hypergeometric orthogonal polynomial of this type is a balanced ${}_4F_3$. The absolutely continuous case is

$$\frac{W_n(x^2;a,b,c,d)}{(a+b)_n(a+c)_n(a+d)_n} = {}_4F_3\left(\begin{matrix}-n, n+a+c+d-1, a+ix, a-ix\\ a+b, a+c, a+d\end{matrix};1\right). \tag{3.11}$$

Wilson [15] showed that

$$\int_0^\infty W_n(x^2)W_m(x^2)\left|\frac{\Gamma(a+ix)\Gamma(b+ix)\Gamma(c+ix)\Gamma(d+ix)}{\Gamma(2ix)}\right|^2 dx = 0, \quad m \neq n, \tag{3.12}$$

when $a, b, c, d > 0$. There is a discrete orthogonality as well, when $a+d = -N$. It is equivalent to Racah's orthogonality for the Racah coefficients (or $6-j$ symbols) of quantum angular momentum theory. The orthogonality relation (3.12) is an easy consequence of the evaluation of

$$\frac{1}{2\pi}\int_0^\infty \left|\frac{\Gamma(a+ix)\Gamma(b+ix)\Gamma(c+ix)\Gamma(d+ix)}{\Gamma(2ix)}\right|^2 dx = \frac{\Gamma(a+b)\Gamma(a+c)\Gamma(a+d)\Gamma(b+c)\Gamma(b+d)\Gamma(c+d)}{\Gamma(a+b+c+d)}, \tag{3.13}$$

which was found by Wilson [15]. This is the most general hypergeometric type extension of the beta integral that I know.

The recurrence relation for

$$b_n = \sum_{k=0}^{n} \binom{n}{k}^2 \binom{n+k}{k}^2 = {}_4F_3\left[\begin{matrix} -n, & -n, & n+1, & n+1 \\ & 1, & 1, & 1 \end{matrix}; 1\right]$$

that was found by Apéry and proven at the last Congress by H. Cohen, see [14], is an easy consequence of the three term recurrence relations for the ${}_4F_3$ polynomials. These three term relations go back to Racah in the early 1940's.

4. Basic hypergeometric series and orthogonal polynomials

A basic hypergeometric series $\sum c_k$ has c_{k+1}/c_k a rational function of q^k for a fixed value of q. We will take $|q| < 1$, or the limiting case when q is a root of unity. Two examples are the q-binomial theorem

$$\sum_{k=0}^{\infty} \frac{(a;q)_k}{(q;q)_k} x^k = \frac{(ax;q)_\infty}{(x;q)_\infty} \tag{4.1}$$

and the theta function

$$\sum_{-\infty}^{\infty} (-1)^k q^{k^2} x^k = (q^2;q^2)_\infty (xq;q^2)_\infty (q/x;q^2)_\infty, \tag{4.2}$$

where

$$(a;q)_\infty = \prod_{n=0}^{\infty} (1-aq^n), \qquad (a;q)_k = (a;q)_\infty/(aq^k;q)_\infty. \tag{4.3}$$

The most easily motivated set of orthogonal polynomials that can be represented as basic hypergeometric functions arise in the following way. Fejér defined generalized Legendre polynomials $p_n(x) = p_n(\cos\theta)$ by

$$|f(re^{i\theta})|^2 = \sum_{n=0}^{\infty} p_n(\cos\theta)\, r^n, \tag{4.4}$$

when a_k are real and

$$f(z) = \sum_{k=0}^{\infty} a_k z^k$$

converges in a neighborhood of $z = 0$. $f(z) = (1-z)^{-1/2}$ gives the Legendre polynomials, and $f(z) = (1-z)^{-\lambda}$ the ultraspherical polynomials. Since these polynomials are orthogonal with respect to $(1-x^2)^{\lambda-1/2}$ when $\lambda > -1/2$, it is natural to ask if any other polynomials of this type are orthogonal. Feldheim and Lanzewizky independently asked and answered this question. The most general such polynomials come when $f(z)$ is the series in the q-binomial theorem. The three term recurrence relation is

$$2x(1-\beta q^n)C_n(x;\beta|q) = (1-q^{n+1})C_{n+1}(x;\beta|q)+(1-\beta^2q^{n-1})C_{n-1}(x;\beta|q). \tag{4.5}$$

They were unable to find the explicit orthogonality relation. When $-1 < \beta, q < 1$ it is

$$\int_{-1}^{1} C_n(x;\beta|q)C_m(x;\beta|q)\prod_{k=0}^{\infty}\left[\frac{1-2(2x^2-1)q^k+q^{2k}}{1-2(2x^2-1)\beta q^k+\beta^2q^{2k}}\right]\frac{dx}{\sqrt{1-x^2}} = 0, \quad m \neq n. \tag{4.6}$$

These polynomials were discovered by Rogers. In [10] he used the special case when $\beta = 0$ and his solution to the connection coefficient problem between these polynomials and Chebychev polynomials to derive the Rogers–Ramanujan identities:

$$\sum_{n=0}^{\infty}\frac{q^{n^2}}{(q;q)_n} = \frac{1}{(q;q^5)_\infty(q^4;q^5)_\infty}, \tag{4.7}$$

$$\sum_{n=0}^{\infty}\frac{q^{n^2+n}}{(q;q)_n} = \frac{1}{(q^2;q^5)_\infty(q^3;q^5)_\infty}. \tag{4.8}$$

In [11] he introduced the general polynomials and proved a number of incredible formulas. These include the linearization result, which would be rediscovered in the case $q = 1$ over twenty years later and in the case $q = -1$ over sixty years later; the connection coefficient problem, and a q-extension of Mehler's bilinear generating function for Hermite polynomials.

The most general set of orthogonal polynomials of this type is

$$\frac{a^nW_n(x;a,b,c,d|q)}{(ab;q)_n(ac;q)_n(ad;q)_n} = {}_4\varphi_3\left[\begin{matrix}q^{-n}, q^{n-1}abcd, ae^{i\theta}, ae^{-i\theta}\\ ab, ac, ad\end{matrix};q,q\right], \tag{4.9}$$

where $x = \cos 0$ and

$$ {}_{p+1}\varphi_p\left[\begin{matrix} a_0, \ldots, a_p \\ b_1, \ldots, b_p \end{matrix}; q, t\right] = \sum_{k=0}^{\infty} \frac{(a_0; q)_k \cdots (a_p; q)_k t^k}{(b_1; q)_k \cdots (b_p; q)_k (q; q)_k}. \tag{4.10} $$

When $-1 < q, a, b, c, d < 1$ the orthogonality relation is

$$ \int_0^{\pi} \frac{W_n(\cos\theta)\, W_m(\cos\theta) |(e^{2i\theta}; q)_\infty|^2 d\theta}{|(ae^{i\theta}; q)_\infty (be^{i\theta}; q)_\infty (ce^{i\theta}; q)_\infty (de^{i\theta}; q)_\infty|^2} = 0, \quad m \neq n. \tag{4.11} $$

See [4] for this orthogonality relation, the recurrence relation, a second order divided difference equation, a Rodrigues type formula using divided differences, and other facts, and [6] for a generating function and asymptotics. There are many very interesting special cases. Those of Rogers are treated directly and in some detail in [2], and some special cases that extend Jacobi polynomials are studied in [9]. Also see [5]. A q extension of polynomials of Pollaczek and Szegö is given in [3]. Some general results of Nevai [7] were used in this case.

Recently we have realized that $q \to e^{2\pi i/k}$ leads to some very interesting results. Take $\beta = s^{\lambda k+1} e^{2\pi i/k}$, $q = s e^{2\pi i/k}$ and let $s \to 1$ in (4.5) dividing first by $(1-s)$ if the limit would vanish. The recurrence relation is

$$ \begin{aligned} 2x p_n(x) &= p_{n+1}(x) + p_{n-1}(x), \quad n+1 \neq mk, \\ 2(m+\lambda) x p_n(x) &= m p_{n+1}(x) + (m + 2\lambda - 2) p_{n-1}(x), \\ & n = mk - 1. \end{aligned} \tag{4.12} $$

The generating function (4.4) is

$$ (1 - 2xr + r^2)^{-1} (1 - 2T_k(x) r^k + r^{2k})^{-\lambda} = \sum_{n=0}^{\infty} p_n(x) r^n \tag{4.13} $$

with $T_k(\cos\theta) = \cos k\theta$. It is surprising that a generating function that is so nice was not found long ago, but it seems not to have been. The orthogonality relation is

$$ \int_{-1}^{1} p_n(x) p_m(x) \prod_{j=0}^{k-1} |x^2 - \cos^2 \pi j/k|^{\lambda} (1 - x^2)^{1/2} dx = 0, \tag{4.14} $$

$$ m \neq n, \quad \lambda > -1/2. $$

Further facts about these polynomials, and a related set whose weight

function is $(1-x^2)^{-1}$ times the above one, are given in [1]. There are more general polynomials that come from the q-Wilson polynomials and correspond to Jacobi polynomials. Until these polynomials were found the only explicit polynomials whose weight function had a zero inside an interval supporting an absolutely continuous measure came from the weight function $(1-x^2)^a|x|^b$ on $[-1, 1]$. The Pollaczek–Szegö random walk polynomials are interesting because the weight function can vanish so rapidly at the ends of the interval that the measure is not in the Szegö class ($a > 1$, $b > 0$), and because the weight function can have an absolutely continuous part and infinitely many discrete masses ($0 < a < 1$, $b > 0$). The corresponding limit when $q \to e^{2\pi i/k}$ in the q-version will lead to a very interesting orthogonality relation, but it has not been worked out yet.

There are many other interesting orthogonal polynomials that can be represented as basic hypergeometric series. See Stanton [12] for some of the geometric settings discovered by Delsarte, Dunkl and Stanton.

References

[1] Al-Sala m W., Allaway Wm. R. and Askey R., *Sieved Ultraspherical Polynomials, Trans. Amer. Math. Soc.*

[2] Askey R. and Ismail M., A Generalization of Ultraspherical Polynomials. In: P. Erdös (ed.), *Studies in Pure Mathematics*, Birkhäuser, 1983, pp. 55–78.

[3] Askey R. and Ismail M., Recurrence Relations, Continued Fractions and Orthogonal Polynomials, *Mem. Amer. Math. Soc.* **300** (1984).

[4] Askey R. and Wilson J., Some Basic Hypergeometric Orthogonal Polynomials that Generalize Jacobi Polynomials, *Mem. Amer. Math. Soc.*

[5] Gasper G. and Rahman M., Positivity of the Poisson Kernel for the Continuous q-Ultraspherical Polynomials, *SIAM J. Math. Anal.* **14** (1983), pp. 409–420.

[6] Ismail M. and Wilson J., Asymptotic and Generating Relations for the q-Jacobi and ${}_4\varphi_3$ Polynomials, *J. Approx. Th.* **36** (1982), pp. 43–54.

[7] Nevai P., Orthogonal Polynomials, *Mem. Amer. Math. Soc.* **18** (1979), No. 213.

[8] Nikishin E. M., The Padé Approximants, In: *Proceedings of the International Congress of Mathematicians, Helsinki*, **1978**, Vol. 2, Helsinki, 1980, pp. 623–630.

[9] Rahman M., The Linearization of the Product of Continuous q-Jacobi polynomials, *Can. J. Math.* **33** (1981), pp. 961–987.

[10] Rogers L. J., Second Memoir on the Expansion of Certain Infinite Products, *Proc. London Math. Soc.* **25** (1894), pp. 318–343.

[11] Rogers L. J., Third Memoir on the Expansion of Certain Infinite Products, *Proc. London Math. Soc.*, **26** (1895), pp. 15–32.

[12] Stanton D., Some q-Krawtchouk Polynomials on Chevalley Groups, *Amer. J. Math.* **102** (1980), pp. 625–662.

[13] Szegö G., *Collected Papers*, 3 volumes, Birkhäuser, Boston, 1982.
[14] van der Poorten A., A Proof that Euler Missed, ..., Apéry's Proof of the Irrationality of $\zeta(3)$, *Math. Intelligencer*, **1** (1979), pp. 195–203.
[15] Wilson J., Some Hypergeometric Orthogonal Polynomials, *SIAM J. Math. Anal.* **11** (1980), pp. 690–701.

UNIVERSITY OF WISCONSIN-MADISON
MADISON, WI 53706, U.S.A.

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

J. BOURGAIN

New Banach Space Properties of Certain Spaces of Analytic Functions

This report describes some recent results in the isomorphic theory of certain spaces of analytic functions, mainly the disc algebra $A = A(D)$. One of the purposes was to investigate to what extent these spaces are similar to the so-called classical Banach spaces, namely to the space C of continuous functions and the Lebesgue space L^1, as far as operators, bases, complemented subspaces, etc., are concerned. At this point, many questions can be answered and the remaining open problems offer new perspectives of investigation. Already some time ago, natural connections between analytic function theory and general Banach space theory were noticed, motivating a more systematic study. It turned out that this research is closely tied up with classical analysis.

The starting point is the remarkable equality $B(l^1, l^2) = \Pi_1(l^1, l^2)$ obtained by A. Grothendieck in [9], stating that each bounded linear operator from l^1 into l^2 is 1-summing. Recall that an operator u between Banach spaces X, Y is p-summing $(0 < p < \infty)$ provided $\pi_p(u) < \infty$, where

$$\pi_p(u) = \sup \left(\sum \|u(x_i)\|^p \right)^{1/p}$$

and the supremum is taken over all finite sequences (x_i) in X with $\sum |\langle x_i, x^* \rangle|^p \leqslant 1$ for each $x^* \in X^*$, $\|x^*\| \leqslant 1$. Let us say that X has the Grothendieck property (GP) provided $B(X, l^2) = \Pi_1(X, l^2)$. For some time, the only known examples of such spaces were the $\mathscr{L}^1$-spaces, i.e., spaces with the same local structure as l^1. The class was enlarged by independent work of S. Kisliakov [15] and G. Pisier [24], who proved that the quotient of an $\mathscr{L}^1$-space by a reflexive subspace remains a GP-space. Denote the circle by Π and let

$$H_0^1 = \{f \in L^1(\Pi); \hat{f}(n) = 0 \text{ for } n \leqslant 0\}.$$

Our first main result, solving affirmatively a problem raised separately by A. Pełczyński and N. Varopoulos, can be stated as follows

THEOREM 1 [2]. *The space L^1/H^1_0 and hence the dual A^* of the disc algebra, have* (GP).

There are many ways to formulate this fact. It means for instance that each bounded bilinear form on A can be extended to a bounded bilinear form on $C(\Pi)$, or that the projective tensor algebra $A \hat{\otimes} A$ is a closed subalgebra of $C(\Pi) \hat{\otimes} C(\Pi)$. Similar results for C^*-algebras were proved in successive work of G. Pisier [25] and U. Haagerup [10]. Let us point out that the embedding of $A \hat{\otimes} A$ in $C(\Pi) \hat{\otimes} C(\Pi)$ is not isometric, as observed by S. Kajser in [14]. A consequence of Theorem 1 is the cotype 2 property of A^*, thus

COROLLARY 2. *There is a fixed constant $C > 0$ such that*

$$\int \Big\| \sum \varepsilon_i x_i \Big\|_{A^*} d\varepsilon \geqslant C \Big(\sum \|x_i\|^2 \Big)^{1/2}$$

holds for each finite sequence (x_i) in A^*, (ε_i) being the usual Rademacher sequence.

Theorem 1 can be proved by different methods. It can be established by the so-called extrapolation technique, depending on the following interpolation inequality for p-summing norms of operators on A.

PROPOSITION 3. *Assume that $1 < p < \infty$ and that u is a p-summing operator from the disc algebra in an arbitrary Banach space. Let $p < q < \infty$ and $1/q' = \theta + (1-\theta)/p'$. Then for all $0 < \varphi < \theta$, the q-summing norm of u can be estimated as follows:*

$$\pi_q(u) \leqslant \frac{c(p)}{\theta - \varphi} \|u\|^{\varphi} \pi_p(u)^{1-\varphi}.$$

A byproduct of Proposition 3 is the fact that each operator u of rank n defined on A admits an extension $\tilde{u}$ to $C(\Pi)$ with $\|\tilde{u}\| \leqslant C \log n \|u\|$. This result is clearly the best possible.

The proof of Proposition 3 relies on the weak-type property of the Hilbert transform and the log modularity of A. More precisely, the density Δ, corresponding to $\pi_p(u)$, obtained by A. Pietsch's factorization and F. M. Riesz's decomposition theorems, is in general not a weight in the sense of B. Muckenhoupt [19] and the usual Riesz projection is not bounded, as an operator acting on $L^2(\Delta)$. This difficulty is avoided by con-

structing new "analytic projections", using the log modular property of A. The following fact has been shown:

PROPOSITION 4 [3]. *Assume Δ in $L^1(\Pi)$ and $\Delta \geqslant 1$. There exist $\bar{\Delta}$ in $L^1(\Pi)$, $\bar{\Delta} \geqslant \Delta$, $\|\bar{\Delta}\|_1 \leqslant \text{const}\,\|\Delta\|_1$, and a projection P from $L^2(\bar{\Delta})$ onto $H^2(\bar{\Delta})$ which is $L^p(\bar{\Delta}) - L^p(\bar{\Delta})$ $(1 < p < \infty)$ and $L^1(\bar{\Delta}) - L^{1,\infty}(\bar{\Delta})$ bounded.*

Several of the results described above extend formally to H^∞ and its dual by arguments of local reflexivity. However, it is not known if each bounded operator from H^∞ to $(H^\infty)^*$ factors through a Hilbert space.[1] The approximation problem for H^∞ is still unsolved. Further extensions to the algebras lying between H^∞ and L^∞ follow from the Douglas property [16]. There are also finite-dimensional models, obtained by considering the polynomial spaces $[1, e^{i\theta}, \ldots, e^{in\theta}]$ equipped with L^∞-norm.

Nothing is known about the local structure of algebras of bounded analytic functions in more variables. However, our methods permit us to solve certain interpolation problems for the bi-disc algebra. For instance [3] the following proposition generalizes a well-known result due to C. Paley [22]:

PROPOSITION 5. *Let $S \subset \boldsymbol{Z}_+$ be a Λ_2-set and let $(\varphi_k)_{k\in S}$ be a weakly 2-summable sequence of H^∞-functions, i.e.,*

$$\sup_{|z|<1} \sum_{k\in S} |\varphi_k(z)|^2 < \infty.$$

Then there exists a Φ in $H^\infty(D \times D)$ satisfying for each $k \in S$

$$\int \Phi(\theta, \psi) e^{-ik\theta} d\theta = \varphi_k(\psi).$$

Denote by $U = U(\Pi)$ the space of functions $f \in A(D)$ such that

$$f = \lim_{n\to\infty} (D_n * f) \text{ uniformly}$$

$D_n = \sum_{p=0}^{n} e^{ip\theta}$) equipped with the norm

$$\|f\|_U = \sup_n \|D_n * f\|_\infty.$$

In [20], D. Oberlin derived from A. Vinogradov's inequality [27]

$$\|C_+[\mu]\|_{1,\infty} \leqslant \text{const}\,\|\mu\|_{U^*} \qquad (\mu \in M(\Pi))$$

[1] This fact was more recently proved in the affirmative.

that compact subsets of measure 0 of Π are the peak sets for U. As a consequence, the dual space U^* admits a decomposition which is very similar to the classical theorem of F. and M. Riesz. By using related ideas, a version of Havin's lemma [11] was derived in [7], implying in particular

PROPOSITION 6.1. *The space U^* is weakly complete, i.e., weakly Cauchy sequences in U^* are weakly convergent.*

2. *Each reflexive subspace of U^* can be linearly embedded in L^p for some $1 < p < \infty$.*

Of course the usual character sequence $(e^{in\theta})_{n=0,1,2,\ldots}$ is a Besselian basis for U, since clearly $\|\sum a_n e^{in\theta}\|_U \geqslant \left(\sum |a_n|^2\right)^{1/2}$. On the other hand, it was proved in [5] that the results of A. Olevski [21], S. Bockarev [1] and S. Szarek [26] on Lebesgue functions remain valid for the disc algebra.

PROPOSITION 7. *Let n be a positive integer and $M > 0$. Suppose that $(\varphi_k)_{1\leqslant k\leqslant n}$, $(x_k)_{1\leqslant k\leqslant n}$ are bi-orthogonal sequences in A and A^* respectively, such that $\|\varphi_k\|_\infty \leqslant M$ $(1 \leqslant k \leqslant n)$ and $\left\|\sum a_k x_k\right\| \leqslant M\left(\sum |a_k|^2\right)^{1/2}$ for all scalar sequences (a_k). Then for some $z \in C$, $|z| < 1$, we have*

$$\frac{1}{n}\sum_{1\leqslant m\leqslant n}\left\|\sum_{1\leqslant k\leqslant m}\varphi_k(z)x_k\right\| \geqslant \varrho(M)\log n.$$

COROLLARY 8. *The disc algebra has no Besselian basis. Consequently, there is no linear isomorphism between the spaces A and U.*

The proof of Proposition 7 relies on Proposition 4. It is not known whether or not U^* has the Grothendieck property. The results stated above indicate a restriction in extending the methods of disc algebra to the space U.

The three-space problem for the Banach space L^1 is one of the still open questions in classical Banach space theory: Assume that Z is a subspace of L^1. It is not known if L^1 embeds in either Z or L^1/Z. It was shown in [23] that the space L^1 does not admit a complemented embedding in L^1/H^1. For this reason the pair $L^1 = L^1(\Pi), Z = H_0^1$ is a natural candidate for a counterexample. By using the results of [23], it can be proved that the complex L^1-space cannot be isometrically embedded in L^1/H_0^1. However, [4],

THEOREM 9. *L^1 embeds isomorphically in L^1/H_0^1. More precisely, there exists an increasing sequence of integers (n_k) such that if $\mathfrak{S}$ is the σ-algebra*

on Π generated by the functions $\sigma_k(\theta) = \operatorname{sign} \cos n_k\theta$, then the restriction of the quotient map $q: L^1 \to L^1/H^1_0$ to $L^1(\mathfrak{S})$ is an isomorphism. Consequently, each bounded $\mathfrak{S}$-measurable function on Π can be obtained as an conditional expectation of an H^∞-function.

In fact, in the above we may take any sequence (n_k) which increases rapidly enough. Our method of proof consists in studying the behaviour of certain H^1-valued martingales.

I end this report with a summary of the known results on the linear isomorphism and non-isomorphism of algebras of analytic functions and H^1-spaces in a distinct number of variables. It was shown by B. Mitiagin and A. Pełczyński [18] that the disc algebra $A(D)$ and the bi-disc algebra $A(D^2)$ are not isomorphic. Their invariant, the so-called $(i_p - \pi_p)$ ratio, is of local nature. The dual of $A(D^2)$ can be identified with the space

$$L^1(\Pi^2)/_R \oplus (A \hat{\otimes} C)^* \oplus C^*,$$

where $R = \{f \in L^1(\Pi^2); \hat{f}(m, n) = 0 \text{ if } m \leqslant 0,\ n < 0\}$, allowing us to establish the non-existence of an isomorphism between $A(D^2)$ and $A(D^3)$ for topological reasons. This method is similar to Henkin's proof, using the theory of analytic measures ([12], [13]), that the polydisc algebras $A(D^m)$ and ball algebras $A(B_n)$ are not isomorphic except for $m = n = 1$.

The investigation of H^1-spaces started with B. Maurey's result [17] on the isomorphism of the Hardy space $H^1(D)$ and the diadic martingale space $H^1(\delta)$. It solved in particular the open problem whether or not $H^1(D)$ has an unconditional basis.

Explicit bases were obtained shortly later by L. Carleson [8] and P. Wojtaszczyk [28]. The isomorphic classification of H^1-spaces in several variables is now essentially understood and it turns out that isomorphism depends on the nature of the singularity appearing in the reproducing kernel.

THEOREM 10. 1. *$H^1(D^m)$ and $H^1(D^n)$ are not isomorphic for $m \neq n$* (see [6]).

2. *$H^1(B_m)$ is isomorphic to $H^1(\delta)$ in each dimension m* (see [29]).

References

[1] Bockarev S. V., Logarithmic Growth of Arithmetic Means of Lebesgue Functions of Bounded Orthonormal Systems, *Dokl. Acad. Nauk SSSR* **223** (1) (1975), pp. 799–802.

[2] Bourgain J., New Banach Space Properties of the Disc Algebra and H^∞, to apper in *Acta Math.*

[3] Bourgain J., *Extension of H^∞-Valued Operators and Bounded Bianalytic Functions*, Inst. Mittag-Leffler, Rep. No. 6 (1983).

[4] Bourgain J., Embedding L^1 in L^1/H^1, to appear in *Trans. AMS.*

[5] Bourgain J., *On Bases in the Disc Algebra*, preprint, Vrije Universiteit, Brussels.

[6] Bourgain J., Non-Isomorphism of H^1-Spaces in One and Several Variables, *J. Funct. Anal.* **46** (1) (1982), pp. 45–57.

[7] Bourgain J., Quelques propriétés linéaires de l'espace des séries de Fourier uniformément convergentes, *C.R.A. Sc. Paris.* Ser. I, **295** (1982), pp. 623–625.

[8] Carleson L., *An Explicit Unconditional Basis in H^1*, Inst. Mittag-Leffler, Rep. No. 2 (1980).

[9] Grothendieck A., Résumé de la théorie métrique des produits tensoriels topologiques, *Bol. Soc. Matem. Sao Paulo* **8** (1956), pp. 1–79.

[10] Haagerup U., *The Grothendieck Ineguality for Bilinear Forms on C*-Algebras*, Matem. Inst. Odense Universitet, No. 1 (1981).

[11] Havin V. P., Weak Completeness of the Space L^1/H^1_0, *Vestnik Leningrad. Univ.* **13** (1973), pp. 77–81.

[12] Henkin G. M., Non Isomorphism of Some Spaces of Functions of Different Numbers of Variables, *Funkt. Analiz i Priloz.* **1** (4) (1967), pp. 57–68.

[13] Henkin G. M., Banach Spaces of Analytic Functions on the Ball and on the Bicylinder Are not Isomorphic, *Funkt. Analiz i Priloz.* **2** (4) (1968), pp. 82–91.

[14] Kajser S., Some Results in the Metric Theory of Tensor Products, *Studia Math.* **63** (1978), pp. 157–170.

[15] Kisliakov S. V., On Spaces with "Small" Annihilators, *Zap. Nauk Sem. Leningrad. Otdel. Mat. Inst. Steklova* (*LOMI*) **65** (1976), pp. 192–195.

[16] Marshall D., Subalgebras of L^∞ Containing H^∞, *Acta Math.* **137** (1976), pp. 91–98.

[17] Maurey B., Isomorphismes entre espaces H^1, *Acta Math.* **145** (1980), pp. 79–120.

[18] Mitiagin S. S. and Pełczyński A., On the Nonexistence of Linear Isomorphisms between Banach Spaces of Analytic Functions of One and Several Complex Variables, *Studia Math.* **56** (1975), pp. 85–96.

[19] Muckenhoupt B., Weighted Norm Inequalities for the Hardy Maximal Function *Trans. A.M.S.* **265** (1972), pp. 207–226.

[20] Oberlin D., A Rudin–Carleson Theorem for Uniformly Convergent Taylor Series, *Michigan Math. J.* **27** (3) (1980), pp. 309–324.

[21] Olevski A. M., Fourier Series of Continuous Functions Relative to Bounded Orthonormal Systems, *Izv. Akad. Nauk SSSR, Ser. Mat.* **30** (1966), pp. 387–432.

[22] Paley R. E., On the Lacunary Coefficients of Power Series, *Annals Math.* **34** (1933), pp. 615–616.

[23] Pełczyński A., *Banach Spaces of Analytic Functions and Absolutely Summing Operators*, Conf. board of Math., SC, Regional Conf. Series in Math., 30 (1976).

[24] Pisier G., Une nouvelle classe d'espaces de Banach vérifiant le théorème de Grothendieck, *Annales de l'Institut Fourier* **28** (1978), pp. 69–90.

[25] Pisier G., Grothendieck's Theorem for Non-Commutative C^*-Algebras with an Appendix on Grothendieck's Constant, *J. Funct. Anal.* **29** (1978), pp. 397–415.

[26] Szarek S., Bases and Biorthogonal Systems in the Spaces C and L^1, *Arkiv Matematik* **17** (2) (1979), pp. 255–271.

[27] Vinogradov S. A., Convergence almost Everywhere of Fourier Series of Functions in L^2 and the Behaviour of the Coefficients of Uniformly Convergent Fourier Series, *Soviet Mat. Dokl.* **17** (5) (1976), pp. 1323–1327.

[28] Wojtaszczyk P., The Franklin System is an Unconditional Basis in H^1, to appear in *Arkiv för Math.*

[29] Wojtaszczyk P., *Hardy Spaces on the Complex Ball Are Isomorphic to Hardy Spaces on the Disc*, $1 \leqslant p < \infty$, preprint.

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

BJÖRN E. J. DAHLBERG

Real Analysis and Potential Theory

The basic aim of this paper is to give a survey of some recent results on the boundary behavior of harmonic functions and the solvability of the Dirichlet problem for general domains. Many of the results have been established by combining potential theoretic ideas with recent results from real analysis. Many of the results will also hold for solutions of more general partial differential equations. We start by recalling some classical results.

The Fatou theorem says that if u is harmonic and bounded from below in the unit disc then u has nontangential limits almost everywhere on the unit circle. Privalov [24] showed the local analogue of this, namely that if u is harmonic in the unit disc and if for every point $e^{i\theta}$ of a measurable subset E of the unit circle there is an $\alpha > 1$ such that u is bounded from below on $\Gamma_\alpha(e^{i\theta}) = \{z: |z| < 1, |z - e^{i\theta}| < \alpha(1 - |z|)\}$ then u has a nontangential limit at almost every $e^{i\theta} \in E$, that is, u restricted to $\Gamma_\beta(e^{i\theta})$ has a limit as $z \to e^{i\theta}$ for all $\beta > 1$. This result was proved by using conformal mappings. The extension to higher dimensions therefore needed new methods. This was first done in 1950 by Calderon [3]. He showed that if u is a harmonic function on $H_n = \{x \in \boldsymbol{R}^n: x_1 > 0\}$, nontangentially bounded at every point Q of a measurable set $E \subset \partial H_n$ (i.e., for every $Q \in E$ there exist $\alpha > 1$, $h > 0$ and M such that $|u| \leqslant M$ in $\Gamma_{\alpha,h}(Q) = \{P \in H_n: |P - Q| < \alpha \operatorname{dist}\{P, \partial H_n\}, |P - Q| \leqslant h\}$), then u has a nontangential limit at almost every $Q \in E$. In 1962 Carleson [6] showed that the same conclusion follows under the weaker hypothesis of nontangential boundedness from below. Both Calderon and Carleson studied the behavior of the harmonic function u in "saw-tooth" regions, that is, domains of the form $\bigcup \Gamma_{\alpha,h}(Q)$, $F \subset \partial H_n$. A crucial part of Carleson's proof consisted of very precise estimates of harmonic measures on saw-tooth regions. The harmonic measure is defined as follows. For a domain D and a continuous function f on D let H_f denote the solution of the Dirichlet problem with bound-

ary values f, i.e., H_f is a continuous function on $\bar{D}$ which agrees with f on ∂D and is harmonic in D. (The Dirichlet problem is not always solvable in the sense we have stated here but the class of domains for which this is possible has been characterized by Wienner [27].) For every $Q \in D$ the mapping $f \to H_f(Q)$ is a positive continuous linear functional on the space of continuous functions on ∂D. Therefore, by the Riesz representation theorem there is a unique probability measure ω_Q on ∂D such that

$$H_f(Q) = \int_{\partial D} f d\omega_Q.$$

The measure ω_Q is called the *harmonic measure* for D evaluated at Q. We remark that the harmonic measures are mutually absolutely continuous; hence they have the same null-sets.

In 1968 and 1970 Hunt and Wheeden proved that if $D \subset \boldsymbol{R}^n$ is a bounded Lipschitz domain and u is a harmonic function that is nontangentially bounded from below at every $Q \in E$, E being a measurable subset of ∂D, then u has nontangential limits at almost every point $Q \in E$ relative to harmonic measure. Their result implies both the results of Calderon and Carleson. Here we recall that a bounded domain D is called a *Lipschitz domain* if ∂D can be described locally as graphs of Lipschitz functions, i.e., functions φ satisfying the estimate $|\varphi(x) - \varphi(y)| \leqslant M|x-y|$. The significance of Lipschitz domains lies in the fact that any domain Ω given as a union of, say, convex circular cones is a countable union of Lipschitz domains. Therefore one can, in principle, reduce questions about nontangential behavior to the case of Lipschitz domains.

The results mentioned above lead to the question of characterizing the null-sets for the harmonic measure of Lipschitz domains. If the domain is sufficiently smooth, e.g., if the boundary is $C^{1,\alpha}$ (i.e., is described locally as graphs of functions with Hölder continuous gradients), then the harmonic measure is bounded from above and below by a constant multiplied by the surface measure. However, this is not the case even for C^1-domains. The problem of characterizing sets of harmonic measure zero for Lipschitz domains was solved by Dahlberg in [9]. For a simple proof see Jerrison and Kenig [18].

THEOREM 1. *If D is a Lipschitz domain then the surface measure σ and the harmonic measure are mutually absolutely continuous. Moreover, the harmonic measure belongs to the Muckenhoupt class A_∞. Furthermore, if k denotes the density of the harmonic measure then for all balls B with center*

on ∂D one has the estimate

$$\left(\frac{1}{\sigma(S)}\int_S k^2 d\sigma\right)^{1/2} \leqslant \text{Const}\,\frac{1}{\sigma(S)}\int_S k d\sigma, \qquad S = B\cap\partial D. \tag{1}$$

We recall here that a positive measure μ belongs to the Muckenhoupt class A_∞ if there exist $\alpha, \beta \in (0,1)$ such that for all balls B with center on ∂D and all Borel sets $E \subset S = B\cap\partial D$ with the property that $\sigma(E) < \alpha\sigma(S)$ we have $\mu(E) < \beta\mu(S)$. We remark that from (1) follows that (see Coifman and Fefferman [7]) to each Lipschitz domain there is an $\varepsilon = \varepsilon(D) > 0$ such that $k \in L^{2+\varepsilon}(\sigma)$. On the other hand for each $q > 2$ there is a Lipschitz domain such that

$$\int_{\partial D} k^q d\sigma = \infty.$$

However, if D is a C^1-domain, i.e., the boundary is locally represented as graphs of continuously differentiable functions, then for all $q \in (1, \infty)$ we have the estimate (Dahlberg [10])

$$\left(\frac{1}{\sigma(S)}\int_S k^q d\sigma\right)^{1/q} \leqslant C(q)\int_S k d\sigma. \tag{1'}$$

In fact the function $\log k$ is of vanishing mean oscillation (Jerrison and Kenig [20]).

The upshot of the reversed Hölder inequality (1) is that for all $f \in L^2(\sigma)$, f is integrable with respect to the harmonic measure. Therefore the Dirichlet problem is solvable when the data are in $L^2(\sigma)$ and the solution H_f has the following properties (Dahlberg [10]):

THEOREM 2. *Let D be a Lipschitz domain and $2 \leqslant p \leqslant \infty$. If $f \in L^p(\sigma)$ then $u = H_f$ is harmonic in D and has the nontangential limit f almost everywhere on ∂D. Furthermore, if $\alpha > 1$ and $N_\alpha u(P) = \sup\{|u(Q)|: Q \in \Gamma_\alpha(P)\}$, $P \in \partial D$,* then

$$\|N_\alpha u\|_{L^p(\sigma)} \leqslant C(\alpha)\|f\|_{L^p(\sigma)}. \tag{2}$$

If in addition D is a C^1-domain then (2) holds for all $p > 1$.

We now describe another way of analyzing the boundary behavior of harmonic functions, namely by the so-called *area integral*. For u being a function in $D \subset \boldsymbol{R}^n$ and $\alpha > 1$, $P \in \partial D$ the area integral of u at P is

defined as

$$A_\alpha u(P) = \Big(\int_{\Gamma_\alpha(P)} |\operatorname{grad} u|^2 \operatorname{dist}\{\varrho, \partial D\}^{2-n} dm(P) \Big)^{1/2},$$

where m is the Lebesgue measure on $\boldsymbol{R}^n$. The area integral was introduced by Marcinkiewicz and Zygmund [23], who showed that if u is harmonic in the unit disc U then $u(z)$ has a nontangential limit almost everywhere on a set $E \subset \partial u$ if and only if the area integral is finite almost everywhere in E. Calderon [4] showed that if u is harmonic in H_n and is nontangentially bounded on $E \subset \partial H_n$ then the truncated area integral

$$A_{\alpha,h} u(P) = \Big(\int_{\Gamma_{\alpha,h}(P)} |\operatorname{grad} u|^2 \operatorname{dist}(Q, \partial H_n)^{2-n} dm(Q) \Big)^{1/2}$$

is finite almost everywhere on E. Later Stein [25] showed the converse.

Later versions of this result have involved inequalities comparing the area integral and the nontangential maximal function. Such inequalities have been proved by many authors in increasing generality, see [1, 14, 15]. By using the reversed Hölder inequality (1) in a crucial step of proof of the so-called "good λ" inequalities, the following result was established in Dahlberg [11].

THEOREM 3. *Let $D \subset \boldsymbol{R}^n$ be a Lipschitz domain and suppose μ is a positive measure on ∂D that satisfies the A_∞-condition. Let M denote the class of harmonic functions that vanish at $Q_0 \in D$. Then for all α and β greater than 1 the L^p-norms of $A_\alpha u$ and $N_\beta u$ are equivalent on M for all $p \in (0, \infty)$.*

It is straightforward to verify that

$$\int_{\partial D} (A_\alpha u)^2 d\sigma \sim \int_D |\operatorname{grad} u|^2 \operatorname{dist}\{Q, \partial D\} dm.$$

Therefore we have, by combining Theorem 2 with Theorem 3, that for $u \in M$

$$\int |\operatorname{grad} u|^2 \operatorname{dist}\{Q, \partial D\} dm \sim \int_{\partial D} u^2 d\sigma.$$

Using the results above one can prove local results about nontangential convergence for harmonic functions in general domains. For $P \in \boldsymbol{R}^n$ let $\Gamma(P)$ denote the class of all possibly truncated open cones with vertex at P. For D a domain and $P \in \partial D$ let $\Gamma(P, D) = \{\gamma \in \Gamma(P) : \gamma \subset D\}$. We can now formulate a general result concerning the existence of nontangential limits.

THEOREM 4. *Let $D \subset \boldsymbol{R}^n$ be a domain and suppose $E \subset \partial D$ is Borel measurable with the property that $\Gamma(P, D) \neq \emptyset$ and $\Gamma(P, \boldsymbol{R}^n - \bar{D}) \neq \emptyset$ for all $P \in E$. If u is harmonic in D then the following statements are equivalent:*

(a) *u has a nontangential limit at almost every point $P \in E$.*

(b) *for almost every $P \in E$ there is a $\gamma \in \Gamma(P, D)$ such that u is bounded from bellow on γ.*

(c) *for almost every $P \in E$ there is a $\gamma \in \Gamma(P, D)$ such that*

$$\int_\gamma |\operatorname{grad} u|^2 |Q-P|^{2-n} dm(Q) < \infty.$$

Here "almost every" is taken with respect to the $(n-1)$-dimensional Hausdorff measure.

For a very interesting extension of this result we refer to Jerrison and Kenig [21].

The crucial ingredient in the proofs of the above results has been the reversed Hölder inequality (1). Another approach to these problems is also possible: by solving the Dirichlet problem by the classical layer potential method. For a Lipschitz domain $D \subset \boldsymbol{R}^n$ and $f \in L^2(\sigma)$ we recall that the double layer potential of f is defined by

$$W(P) = \frac{1}{\omega_n} \int_{\partial D} \frac{f(Q)\langle Q-P, N(Q)\rangle}{(P-Q)^n} d\sigma(Q), \quad P \in D,$$

where $N(Q)$ is the unit outward normal to ∂D at $Q \in \partial D$ and ω_n is the area of the unit sphere in $\boldsymbol{R}^n$. It follows from the recent works of Calderon [5] and Coifman, Mc Intosh and Meyer [8], that w has a nontangential limit $Tf(P)$ for almost every $P \in \partial D$ and for all $\alpha > 1$ we have

$$\|Tf\|_{L^2(\sigma)} \leqslant \|N_\alpha w\|_{L^2(\sigma)} \leqslant \text{Const } \|f\|_{L^2(\sigma)}.$$

In order to show the solvability of the Dirichlet problem with data in $L^2(\sigma)$, it is therefore enough to show that the operator T has a bounded inverse on $L^2(\sigma)$. This has recently been proved to be the case by Verchota [26]. For the case of C^1-domains Fabes, Jodeit and Riviere proved that T was invertible on $L^p(\sigma)$ for all p, $1 < p < \infty$. Clearly, for the case of Lipschitz domains, one cannot in general claim that T is invertible for all $p > 1$. The method of layer potentials can also be used to treat other boundary value problems like e.g. the Neumann problem. For a discussion of this see the above papers. A direct treatment of Neumann problem has been found by Jerrison and Kenig [19].

We shall next briefly indicate some anologues of the above results for solutions of more general equations. Let us consider uniformly elliptic equations of $Lu = 0$ where $L = \sum \frac{\partial}{\partial x_j} a_{ij} \frac{\partial}{\partial x_j}$ where the a_{ij}'s are bounded and measurable. In this case it is possible to define the analogue of the harmonic measure, see Littman, Stampacchia and Weinberger [22], and it is called the *L-harmonic measure*. It is also known ([22]) that a domain D has the property that the solutions of the equation $Lu = 0$, $u = f$ on ∂D are continuous on $\bar{D}$ for all $f \in C(\partial D)$ if and only if D has the same property relative to the Dirichlet problem for harmonic functions.

It is therefore natural to ask for the relation between the L-harmonic measure and the surface measure. Even for smooth domains it may happen that the surface measure and the L-harmonic measure are mutually singular (see Caffarelli, Fabes and Kenig [2]). However, if the coefficients are assumed to be continuous and the modulus of continuity $\omega(\delta)$ satisfies the condition

$$\int_0 \omega^2(h) h^{-1} d\lambda < \infty$$

then the methods developed in [12] yield that the surface measure and the L-harmonic measure are mutually equivalent.

References

[1] Burkholder D. and Gundy R., Distribution Function Inequalities for the Area Integral, *Studia Math.* **44** (1972), pp. 527–544.

[2] Caffarelli L., Fabes E. and Kenig C. E., Completely Singular Elliptic-Harmonic Measures, *Indiana J. Math.* **30** (1981), pp. 917–924.

[3] Calderon A. P., On the Behavior of Harmonic Functions Near the Boundary, *Trans. Amer. Math. Soc.* **68** (1950), pp. 47–54.

[4] Calderon A. P., On a Theorem of Marcinkiewicz and Zygmund, *Trans. Amer. Math. Soc.* **68** (1950), pp. 55–61.

[5] Calderon A. P., Cauchy Integrals on Lipschitz Curves and Related Operators *Proc. Nat. Acad. Sc. USA* **74** (1977), pp. 1324–1327.

[6] Carleson L., On the Existence of Boundary Values for Harmonic Functions in Several Variables, *Ark. Mat.* **4** (1962), pp. 393–399.

[7] Coifman R. R. and Fefferman C., Weighted Norm Inequalities for Maximal Functions and Singular Integrals, *Studia Math.* **51** (1974), pp. 269–274.

[8] Coifman R. R., McIntosh A. and Meyer Y., L'integrale de Cauchy définit un opérateur borné sur L^2 pour les courbes Lipschitziennes, *Annals of Math.* **116** (1982), pp. 361–388.

[9] Dahlberg B. E. J., On Estimates of Harmonic Measure, *Arch. Rat. Mech. Anal.* **65** (1977), pp. 272–288.
[10] Dahlberg B. E. J., On the Poisson Integral for Lipschitz and C^1-Domains, *Studia Math.* **66** (1979), pp. 13–24.
[11] Dahlberg B. E. J., Weighted Norm Inequalities for the Lusin Area Integral and Nontangential Maximal Functions for Functions Harmonic in a Lipschitz Domain, *Studia Math.* **67** (1980), pp. 297–314.
[12] Fabes E. B., Jerrison D. S. and Kenig C. E., Multilinear Littlewood-Paley Estimates with Applications to Partial Differential Equations. *Proc. Nat. Acad. Sci. USA* **79** (1982), pp. 5746–5750.
[13] Fabes E. B., Jodeit M. and Riviere N. M., Potential Techniques on C^1-Domains, *Acta Math.* **141** (1978), pp. 165–186.
[14] Fefferman C. and Stein E. M., H^p-Spaces of Several Variables, *Acta Math.* **129** (1972), pp. 137–193.
[15] Gundy R. and Wheeden R., Weighted Integral Inequalities for the Nontangential Maximal Function, Lusin Area Integral and Walsh–Paley Series, *Studia Math.* **49** (1974), pp. 107–124.
[16] Hunt R. R. and Wheeden R. L., On the Boundary Values of Harmonic Functions, *Trans. Amer. Math. Soc.* **132** (1968), pp. 307–322.
[17] Hunt R. R. and Wheeden R. L., Positive Harmonic Functions on Lipschitz Domains, *Trans. Amer. Math. Soc.* **147** (1970), pp. 507–527.
[18] Jerrison D. S. and Kenig C. E., The Dirichlet Problem in Non-Smooth Domains, *Annals of Math.* **113** (1981), pp. 367–382.
[19] Jerrison D. S. and Kenig C. E., The Neumann Problem on Lipschitz Domains, *Bull. Amer. Math. Soc.* **4** (1981), pp. 203–207.
[20] Jerrison D. S. and Kenig C. E., The Logarithm of the Poisson Kernel of a C^1-Domain has Vanishing Mean Oscillation, *Trans. Amer. Math. Soc.* **273** (1982), pp. 781–794.
[21] Jerrison D. S. and Kenig C. E., Boundary Behavior of Harmonic Functions in Nontangentially Accessible Domains, *Advances in Math.* **46** (1982), pp. 80–147.
[22] Littman W., Stampacchia G. and Weinberger H., Regular Points for Elliptic Equations with Discontinuous Coefficients, *Annali della Scuola Normale Superiore di Pisa*, **XVIII** (1963), pp. 45–79.
[23] Marcinkiewicz J. and Zygmund A., A Theorem of Lusin, *Duke Math. J.* **4** (1938), pp. 473–485.
[24] Privalov I., *Randergenschaften analytischer Funktionen*, Deutscher Verlag der Wissenschaften, Berlin, 1956.
[25] Stein E. M., On the Theory of Harmonic Functions of Several Variables. II. Behavior Near the Boundary, *Acta Math.* **106** (1961), pp. 137–174.
[26] Verchota G. C., *Layer Potentials and Boundary Value Problems for Laplace's Equation on Lipschitz Domains*, Thesis, University of Minnesota, 1982.
[27] Wiener N., The Dirichlet Problem, *J. Math. Phys.* **3** (1924), pp. 127–146.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF GÖTEBORG AND
CHALMERS UNIVERSITY OF TECHNOLOGY
S-412 96 GÖTEBORG
SWEDEN

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

TADEUSZ FIGIEL

Local Theory of Banach Spaces and Some Operator Ideals

The local theory of Banach spaces is concerned with the structure of finite-dimensional Banach spaces and the relation between an infinite-dimensional Banach space and its finite-dimensional subspaces.

In this theory two isomorphic Banach spaces X, Y are considered to be close to each other if the *Banach–Mazur distance coefficient*, $d(X, Y)$, defined as

$$\inf\{\|T\|\cdot\|T^{-1}\| \mid T \text{ is a linear isomorphism of } X \text{ onto } Y\}$$

is close to 1. One puts $d(X, Y) = \infty$ if X, Y are not isomorphic.

We shall denote by $\mathscr{S}X$ the set of all finite-dimensional linear subspaces of X and, for any Banach space F, we put

$$d(F, \mathscr{S}X) = \inf\{d(F, E) \mid E \in \mathscr{S}X\}.$$

The space Y is said to be *finitely a-represented in X* provided that $d(F, \mathscr{S}X) \leqslant a$ for $F \in \mathscr{S}Y$. We write Y f.r. X if this is true for $a = 1$.

The concept of finite representability is due to A. Grothendieck ([23]), who also conjectured that *l_2 is finitely 1-represented in every infinite-dimensional Banach space*. This conjecture was proved (in the case of real scalars) by A. Dvoretzky ([12]). That result, which strengthens considerably the so-called Dvoretzky–Rogers Lemma, marks the beginning of the local theory of Banach spaces.

A further incentive was the work of R. C. James, which led to the introduction of *super-reflexive* Banach spaces ([25]). If $\mathscr{P}$ is a class of Banach spaces, the class called *super-$\mathscr{P}$* consists of those $X \in \mathscr{P}$ for which Y f.r. X implies $Y \in \mathscr{P}$. The property defining the class $\mathscr{P}$ is called a *super-property* if $\mathscr{P} =$ *super-$\mathscr{P}$*.

The local theory studies super-properties and some of them are very important in the theory, but the notion of a local property of a Banach

space is wider. It includes also, e.g., the local $\mathscr{F}$-structures which we define in Section 3. In fact, many concepts studied in functional analysis and pertaining to finite or finite-dimensional objects become "local" if suitable bounds are imposed on integer-valued parameters (cf. [48]). Here we mention only the concept of the uniform approximation property.

It is not possible to cover in this report all aspects of the local theory of Banach spaces. For more information we refer especially to the monograph [40], the reports of A. Pełczyński and G. Pisier in these Proceedings and the recent survey [43].

1

Let us exemplify a natural family of super-properties. To every bounded linear operator $A: E_1 \to E_2$ where E_i is a linear subspace of $L_{p_i}(\mu_i)$, $0 < p_i \leqslant \infty$ for $i = 1, 2$, there corresponds a super-property which we call $\mathscr{P}(A)$. (A special case of this construction appears also in the address of A. Pełczyński in these Proceedings.) Namely, if X is a Banach space, we say that $X \in \mathscr{P}(A)$ if the tensor product

$$A \otimes \mathrm{id}_X: E_1 \otimes X \to E_2 \otimes X$$

is bounded when $E_i \otimes X$ is regarded as a linear subspace of the space $L_{p_i}(\mu_i; X)$ for $i = 1, 2$. If $X \in \mathscr{P}(A)$, in particular if $\dim X < \infty$, then the number $\|A \otimes \mathrm{id}_X\|$ is a parameter characterizing the $\mathscr{P}(A)$ property of X.

We shall mention some special cases. If $\mathscr{F}: L_2(\boldsymbol{R}) \to L_2(\boldsymbol{R})$ is the Fourier transform on the real line, then $X \in \mathscr{P}(\mathscr{F})$ iff X is isomorphic to a Hilbert space ([36]).

Now, let $L_2 = L_2([0, 1], dx)$ and let $\mathrm{Rad} \subset L_2$ be the subspace spanned by the Rademacher functions $r_1, r_2, \ldots$ Consider first the operator $\mathscr{R}: L_2 \to L_2$, $\mathscr{R}f = \sum_{i \geqslant 1} (f, r_i) r_i$. The property $X \in \mathscr{P}(\mathscr{R})$ was introduced in [44] as *K-convexity*. The important result of [52] says that

$$X \in \mathscr{P}(\mathscr{R}) \text{ iff } l_1 \text{ is not f.r. } X.$$

Let $\mathscr{C}_q: \mathrm{Rad} \to l_q$, $q \geqslant 2$, $\mathscr{C}_q\big(\sum_{i \geqslant 1} a_i r_i\big) = (a_i)_{i \geqslant 1}$ and $\mathscr{T}_p: l_p \to L_2$, $1 \leqslant p \leqslant 2$, $\mathscr{T}_p\big((a_i)\big) = \sum_{i \geqslant 1} a_i r_i$. The spaces in $\mathscr{P}(\mathscr{C}_q)$ (resp., in $\mathscr{P}(\mathscr{T}_p)$) are said to be of *cotype* q (resp., of *type* p). The equalities

$$C_q(X) = \|\mathscr{C}_q \otimes \mathrm{id}_X\|, \quad T_p(X) = \|\mathscr{T}_p \otimes \mathrm{id}_X\|$$

are consistent with the usual definition of the *cotype q* and *type p constants of X*.

The notions of type (resp., cotype) appeared first in probabilistic considerations (cf. [24]) as sufficient (resp., necessary) conditions for the almost sure convergence of the random series $\sum_i \pm x_i$, where $x_i \in X$ for $i = 1, 2, \ldots$ Later they were found to be important in many other contexts, also outside the local theory of Banach spaces. A deep study of their properties was made in [44]. It culminated in the following result.

Given an infinite-dimensional Banach space X, let

$$p(X) = \sup\{p \leqslant 2 |\ T_p(X) < \infty\},$$
$$q(X) = \inf\{q \geqslant 2 |\ C_q(X) < \infty\}.$$

Then $l_{p(X)}$ f.r. X and $l_{q(X)}$ f.r. X.

As a corollary of the above result, X can have no super-property which fails for $l_{p(X)}$ or for $l_{q(X)}$. In particular, since $p(l_r) = \min(2, r)$, $q(l_r) = \max(2, r)$, we see *why* X is of no better type than $p(X)$ and of no better cotype than $q(X)$.

The proof in [44] involved a detailed study of sequences of vectors which X must contain if $p(X) = p < 2$ (resp., $q(X) = p > 2$); if $p = 2$ this reduces to the Dvoretzky–Rogers Lemma. Applying to those sequences the technique developed in [6], which depends on the Ramsey Theorem, they were in a position to use the deep result of [35], which gives a sufficient condition for l_p to be finitely represented in a Banach lattice.

Some simplifications and strengthenings have been found in [55], [46] and [37].

Those combinatorial arguments give no useful estimates for the dimension of l_p^k-subspaces one can obtain starting from a sequence of n vectors. Since the study of finite-dimensional Banach spaces requires quantitative estimates, this leads to a general question: *Can the values of basic parameters characterizing "geometric" properties of the space X be explained in terms of some simple objects in X?*

Few results are known in this direction (cf. [59], [33]). In this report we shall consider more specific questions concerning the quantity

$$k_{\varepsilon,p}(X) = \sup\{k |\ d(l_p^k, \mathscr{S}X) \leqslant 1+\varepsilon\}.$$

2

We shall first discuss the case of nearly Euclidean subspaces. Following [18], we write $k_\varepsilon(X)$ instead of $k_{\varepsilon,2}(X)$.

The proofs of Dvoretzky's theorem given in [45] and [57], besides being simpler than the original argument, cover also the case of complex Banach spaces. They yield the estimates (stronger than those in [12])

$$k_\varepsilon(X) \geqslant c\varepsilon^2\big(\log(2/\varepsilon)\big)^{-1}k_1(X), \qquad 0<\varepsilon<1,$$

$$k_1(X) \geqslant c\,\max\big(\log n,\, nd(X, l_2^n)^{-2}\big).$$

where $n = \dim X$ and $c>0$ is an absolute constant.

Since $k_1(l_\infty^n) \leqslant C\log n$ (cf. [45], [18]), the second estimate cannot be improved in the general case. The first estimate is almost exact. The logarithmic term may be just a consequence of the methods.[1] There are, however, examples where $k_1(X) = \dim X = m$ and $k_\varepsilon(X) \leqslant C\varepsilon^2 m$ for $1 \geqslant \varepsilon \geqslant \sqrt{(\log m)/m}$. This estimate can be viewed as a quantitative analogue of the distortion problem for l_2 (cf. [40]).

In [18] a subtler estimate of $k_1(X)$ was found. Namely, *there is a* $c>0$ *such that, if* $\dim X = n$ *and* $q \geqslant 2$, *then*

$$k_1(X) \geqslant cC_q(X)^{-2}n^{2/q}.$$

In particular, if $X \in \mathscr{S}L_p$, $1 \leqslant p < \infty$, then $k_1(X) \geqslant c_1 n^r$, $r = \min(1, 2/p)$. (In the case $X = l_1^n$ this was also obtained in [30].)

Conversely, if $k_1(X) \geqslant c(\dim X)^\alpha$, $c, \alpha > 0$, for every $X \in \mathscr{S}Y$, then $C_q(Y) < \infty$ for $q > 2/\alpha$ (cf. [18]).

The key ingredient in [45] and [18] was the inequality discovered by P. Lévy ([38]), which follows from the solution of the isoperimetric problem for subsets of the sphere $S^{n-1} \subset \boldsymbol{R}^n$.

If $f: S^{n-1} \to \boldsymbol{R}$ *satisfies* $|f(x)-f(y)| \leqslant \|x-y\|_2$ *for* $x, y \in S^{n-1}$, *then there exists an* M_f *such that for* $\delta > 0$

$$\lambda(\{x \in S^{n-1} \mid |f(x)-M_f| \geqslant \delta\}) \leqslant 4e^{-n\delta^2/2}.$$

(Here λ denotes the normalized Lebesgue measure on S^{n-1}.)

One considers functions of the form $f(x) = \|Tx\|$ where $T: l_2^n \to X$ is a linear operator of norm 1. If T is suitably chosen, since f is almost constant off a set of a small measure, it is possible to show that f is also almost constant on $E \cap S^{n-1}$, where E is a "typical" linear subspace of

[1] *Added in proof*: This follows from Corollary 2.6 in [21].

dimension k, provided that $k \leqslant anM_f^2$. The number a depends on the expected bound for $\sup\{f(x)/f(y)|\ x, y \in E \cap S^{n-1}\}$.

This leads to the estimate for $k_s(X)$ fairly easily. To estimate $k_1(X)$ one can use the operator T which results from the Dvoretzky–Rogers Lemma.

Considering a T for which $d(X, l_2^n)$ is attained, one obtains the estimate

$$k_1(X)k_1(X^*) \geqslant cn^2 d(X, l_2^n)^{-2} \geqslant cn.$$

In particular, $\max(k_1(X), k_1(X^*)) \geqslant \sqrt{cn}$. This estimate was used in order to solve the finite-dimensional version of the complemented subspace problem.

It is also possible to obtain a formula for $k_1(X)$. We refer to [49] for the definition of the $U_{(2,2)}$-*norm* of operators (also called the γ-*summing norm*).

There exists a $c > 0$ *such that for every* X

$$cU_{(2,2)}(\mathrm{id}_X)^2 \leqslant k_1(X) \leqslant 4U_{(2,2)}(\mathrm{id}_X)^2.$$

This equivalence allows one to obtain the estimates of $k_1(X)$ from rather formal results concerning the $U_{(2,2)}$-norm. On the other hand, it is possible to characterize the $U_{(2,2)}$-norm of arbitrary operators in terms of their behaviour on some Euclidean subspaces of the domain.

Let us mention a consequence of the estimate $k_1(X) \geqslant c \dim X$ for $X \in \mathscr{S}L_1$. It implies that, whenever $k_1(E)/\dim E$ is sufficiently small, the space E has "many" subspaces with rather bad properties. Namely (cf. [15]), *there is a* $c > 0$ *such that, if* $k > k_1(E)$, *then every* $F \in \mathscr{S}E$ *with* $\dim F \geqslant k/c$ *has a subspace* G *such that* $\dim G = k$ *and* $\mathrm{gl}(G) \geqslant c(k/k_1(E))^{1/2}$. The parameter $\mathrm{gl}(G)$, introduced in [22], estimates from below $d(G, U)$ for any space U with a 1-unconditional basis. In fact, $\mathrm{gl}(G) \leqslant \lambda d(G, Z)$ if Z is λ-complemented in a Banach lattice.

Recall that a linearly independent sequence $x_1, \ldots, x_n \in X$ is called *a-unconditional* (resp., *a-symmetric*) provided that, for every choice of $\varepsilon_i \in \{+1, -1\}$ and scalars $c_1, \ldots, c_n$, one has

$$\Big\| \sum_{i=1}^n \varepsilon_i c_{\pi(i)} x_i \Big\| \leqslant a \Big\| \sum_{i=1}^n c_i x_i \Big\|$$

where $\pi(i) = i$ for $i = 1, \ldots, n$ (resp., for every permutation π of $\{1, \ldots, n\}$).

The study of such sequences (and their infinite analogues) plays an important role in the investigation of the structure of Banach spaces. Only recently, however, strong quantitative results have been found (cf. [2], [42]) concerning the following general question:

Let $x_1, \ldots, x_n \in X$ be a (linearly independent) sequence with a property $\mathscr{P}$. Consider another condition, say $\mathscr{Q}$. For what values of k is it possible to find $y_1, \ldots, y_k \in X$ of the form $y_j = \sum_{i \in A_j} a_i x_i$, where $A_j \cap A_l = \emptyset$ for $j \neq l$, so that the sequence $y_1, \ldots, y_k$ satisfies $\mathscr{Q}$?

It turns out that in several important cases one can obtain estimates of the form $k \geqslant cn^a$ where $c, a > 0$ depend on $\mathscr{P}$, $\mathscr{Q}$ and sometimes also on the geometric properties of X. In [2] and [42] such problems are studied with the technique involving "isoperimetric inequalities" for some finite groups and for products of spheres. This leads to strong estimates, e.g., in the case: $\mathscr{P}$—arbitrary, $\mathscr{Q}$—$(1+\varepsilon)$-unconditional, and also to some quantitative versions of theorems from [44] and [35].

The classical isometric embedding J_p: $l_p \to L_1$, $1 < p \leqslant 2$, maps the unit vectors into independent identically distributed p-stable random variables $(\theta_i)_{i=1}^{\infty}$. One can show (cf. [16]) that for any $a > 1$, if $X \in \mathscr{S}L_1$, $d(X, l_1^n) \leqslant a$ and $\theta_1, \ldots, \theta_m \in X$, then $m \leqslant c(\log n)^{p/(2p-2)}$, $c = c_{a,p} < \infty$.

The surprising result of [29] yields, in particular, *for every* $1 < p < 2$ *and* $\varepsilon > 0$, *the estimate*

$$k_{\varepsilon,p}(l_1^n) \geqslant cn$$

for $n = 1, 2, \ldots$, *where* $c = c_{\varepsilon,p} > 0$. The embedding is accomplished by means of a random matrix which somehow simulates the J_p.

In the case $1 < s < p < 2$, it is known only that, for some $a = a(s,p) < \infty$, $k_{a,p}(l_s^n) \geqslant c_{s,p} n$; this is deduced from the above result by using Maurey's factorization theorem.

The more general result obtained subsequently in [54] makes use of the *stable type* p *constant* $ST_p(X)$, which is equal to $\|J_p \otimes \mathrm{id}_X\|$ (cf. Section 1) (if $p = 1$, the 1-stable variables are not elements of L_1, but one may use, e.g., the operator $J_1: l_1 \to L_{1/2}$). Let us mention that $T_p(X) \leqslant c_p ST_p(X)$.

It is proved in [54] that, *for each* $\varepsilon > 0$ *and* $1 \leqslant p < 2$, *there is a* $\delta = \delta_{p,\varepsilon} > 0$ *such that for every* X

$$k_{\varepsilon,1}(X) \geqslant \exp\bigl(\delta ST_1(X)\bigr),$$

$$k_{\varepsilon,p}(X) \geqslant \delta\bigl(ST_p(X)\bigr)^{p/(p-1)}.$$

In particular, if $ST_p(X) = \infty$, then l_p f.r. X. Conversely, if l_p f.r. X, then $ST_p(X) \geqslant ST_p(l_p) = \infty$. In fact, one has either $ST_2(X) < \infty$ or

$$\{p \in [1, 2] |\ l_p \text{ f.r. } X\} = \{p |\ ST_p(X) = \infty\} = [p(X), 2].$$

This corollary, obtained in [54], recaptures the essential part of the results in [44] concerning the type, without making use of the most difficult parts of the proof.

Since $ST_p(l_1^n) \geqslant n^{1-1/p}$, also the result of [29] is a corollary. However, those lower estimates are not always sharp, e.g., if $1 < p < 2$, one obtains $k_{\varepsilon,p}(l_p^n) \geqslant \delta(\log n)^{1/(p-1)}$.

There is only one case of $p \neq 2$ where a formula for $k_{\varepsilon,p}$ is known, at least for some $\varepsilon > 0$. For real scalars this was proved in [13], using probabilistic and combinatorial techniques.

For each $\delta > 0$ there exist $\beta, c > 0$ such that, if $x_1, \ldots, x_n \in X$, $\|x_i\| \leqslant 1$, $i = 1, \ldots, n$ and

$$\int_0^1 \Big\| \sum_{i \leqslant n} r_i(t) x_i \Big\| dt \geqslant \delta n,$$

then there is an $A \subset \{1, \ldots, n\}$ with $\mathrm{Card}(A) = k \geqslant cn$ which is β^{-1}-equivalent to the unit vector basis in l_1^k.

Very recently I have learned that the complex case was solved in [47]. These results generalize some theorems on Sidon sets obtained in [53].

Analogous results for $p > 2$ are less satisfactory. We mention only the estimate found in [1]

$$\max\big(k_{\varepsilon,2}(X), k_{\varepsilon,\infty}(X)\big) \geqslant \exp(c\sqrt{\log n})$$

where $c = c_\varepsilon > 0$ does not depend on X. Easy examples show that this estimate cannot be improved. The proof uses [18] and some combinatorial methods.

These results concerning finite-dimensional normed spaces should be compared with what is known about the corresponding infinite-dimensional problems.

It is a still open major question whether every infinite-dimensional Banach space contains an infinite a-unconditional sequence. On the other hand, there are infinite-dimensional Banach spaces with an unconditional basis which have neither a subspace nor a quotient space with a symmetric (or even a subsymmetric) basis and, moreover, one can assume in addition

that X is of type 2 and cotype $2+\varepsilon$ for each $\varepsilon > 0$. Similarly, a space with a symmetric basis may contain no isomorphic copy of any l_p space (cf. [14], [40]).

3

Now we present some cases where a local approach was used in the solution of problems concerning the global structure of Banach spaces.

The well-known problem whether a Banach space $L_1(\mu)$ can contain an uncomplemented subspace X which is isomorphic to $L_1(\nu)$ was solved in [3].

The solution follows from the local analogue, which says that *for some* $D < \infty$ *one can find, for* $n = 1, 2, \ldots$, *subspaces* $X_n \subset l_1^{m(n)}$ *with* $d(X_n, l_1^n) \leqslant D$ *such that* $\|P\| \geqslant a_n$, *for any projection* P *of* $l_1^{m(n)}$ *onto* X_n, *where* $\lim a_n = \infty$.

On the other hand, in the results of [5] and [28], which we present below, there is no local version which is a priori equivalent to the global one.

First we recall the notion of a *local* $\mathscr{F}$*-structure* (cf. [48]). Let $\mathscr{F}$ be a family of finite-dimensional Banach spaces. A Banach space X is said to have an $\mathscr{F}$*-structure* if X has an $\mathscr{F}_\lambda$-structure for some $\lambda \in [1, \infty)$. This means that for each $E \in \mathscr{S}X$ there are $F \in \mathscr{F}$ and $G \in \mathscr{S}X$ such that $G \supseteq E$ and $d(F, G) \leqslant \lambda$.

Two well-known special cases are obtained if $\mathscr{F} = \mathscr{L}_p = \{l_p^n | \ n = 1, 2, \ldots\}$ for some $p \in [1, \infty]$ or if $\mathscr{F} = \mathscr{U} = \{E \in \mathscr{S}l_\infty | \ E$ has a 1-unconditional basis$\}$. Space with $\mathscr{L}_p$-structure coincide with the $\mathscr{L}_p$*-spaces* introduced in [39]. The $\mathscr{U}$-structure is called the *local unconditional structure* in the sense of [11] (abbreviated to l.u.st.).

The above class of l.u.st.'s is larger than (but closely related to, cf. [17]) the class of Banach lattices. It contains most Banach spaces encountered in Analysis (some spaces of analytic functions and differentiable functions being notable exceptions). Certain pathologies which are possible in the general case do not occur within this class, e.g., if X has a l.u.st. and is not super-reflexive then l_1 f.r. X ([26]).

It is not known whether or not l.u.st. coincides with the more general notion introduced in [22]. This is related to the problem whether a λ-complemented subspace of a space X with a c-unconditional basis has an $f(c, \lambda)$-unconditional basis. The problem is still open even in the important special case of $X = L_p(\mu)$, $p \neq 2$.

The $\mathscr{L}_p$-spaces are important in the isomorphic theory of Banach spaces, because they seem to be the right generalization of the classical L_p-spaces in this context. In the past few years the theory of those spaces has been considerably enriched, especially due to the work of J. Bourgain (cf. [4]). Here we shall describe only the examples found in [5], whose construction is "finite-dimensional".

Each of the spaces called $X_{a,b}$ is obtained in [5] as the direct limit of a sequence $(E_n)_{n\geqslant 1}$ of λ-isomorphs of $l_\infty^{k(n)}$. The inductive construction of the E_n's forces some projections on $X_{a,b}$ to satisfy an estimate involving the parameters a, b. As a consequence, *the $\mathscr{L}_{\infty,\lambda}$-spaces $X_{a,b}$ have the Radon–Nikodym property*, in particular, *they have no subspace isomorphic to c_0*. This disproves a conjecture supported by the result for spaces which are $\mathscr{L}_{\infty,\lambda}$ for each $\lambda > 1$ ([61]). Moreover, *for any $\lambda > 1$, among the $X_{a,b}$'s there are $2^{\aleph_0}$ mutually non-isomorphic $\mathscr{L}_{\infty,\lambda}$-spaces*. For other striking features of those examples we refer to [4] or [5].

Even more essential uses of local methods in global problems as well as many purely local results can be found in [28] (cf. also [40]). Two natural classes of separable Banach spaces with a global symmetric structure are the *rearrangement invariant* (r.i.) *lattices of measurable functions on* $[0, 1]$ *or* $[0, \infty)$. Those spaces obviously have, for $\lambda > 1$, an *$\mathscr{S}_\lambda$-structure* where $\mathscr{S} = \{E \in \mathscr{S}l_\infty \mid E \text{ has a 1-symmetric basis}\}$. The $\mathscr{S}$-structure of those spaces is a very efficient tool in their investigation. Many results in [28] are new, even for the spaces $L_p[0, 1]$, $1 < p < \infty$, $p \neq 2$.

A partial case of such a result shows that *if X is an r.i. function space on* $[0, 1]$ *and* $d(X, L_p) < \infty$, *for some* $p \in [1, \infty)$, *then the linear spaces X and L_p are equal and the norms are equivalent*. This means that $L_p[0, 1]$ has a *unique representation* as a r.i. function space on $[0, 1]$. (Not every r.i. space Y on $[0, 1]$ has the above property.)

The proofs of that and other results in [28] often depend on quantitative finite-dimensional analogues, which are also interesting in themselves. We quote one of them in the stronger form obtained, by another method, in [56].

If $(e_i)_{i=1}^n$ and $(f_i)_{i=1}^n$ are normalized 1-symmetric bases in the Banach spaces E and F, $d(E, l_2^n) \geqslant n^r$, $r > 0$ and $T \in L(E, F)$, $Te_i = f_i$ for $i = 1, \ldots, n$, then $\max\{\|T\|, \|T^{-1}\|\} \leqslant C < \infty$ *where C depends only on r and $d(E, F)$.*

In this sense the symmetric structure of E is unique provided that E is not "close to the Hilbert space". It is not known whether or not C has to depend on r.

Two recent results, also concerning n-dimensional symmetric spaces, show in two different ways that, for large n, those spaces form a small subclass of the class $\mathscr{M}_n$ of n-dimensional Banach spaces.

It is proved in [59] that, *if* $E, F \in \mathscr{M}_n$ *have* 1*-symmetric bases, then* $d(E, F) \leqslant 2^{18} n^{1/2}$.

This should be compared with the remarkable result

$$\inf_n n^{-1} \sup \{d(X, Y) |\ X, Y \in \mathscr{M}_n\} > 0,$$

obtained in [20]. The method introduced by E. D. Gluskin was later used in [41] to solve the problem of large asymmetry constants (cf. [19]).

A space $E \in \mathscr{M}_n$ is said *to have enough symmetries*, in short $s(E) = 1$, provided that

$$\dim \{T \in L(E, E) |\ TU = UT \text{ for } U \in G(X)\} = 1$$

where $G(X) = \{U \in L(X, X) |\ \|U\| = \|U^{-1}\| = 1\}$. (Many facts proved for spaces with a 1-symmetric basis can be obtained by assuming only that $s(X) = 1$.) Setting for $X \in \mathscr{M}_n$

$$s(X) = \inf \{d(X, E) |\ E \in \mathscr{M}_n, s(E) = 1\},$$

we obtain the *asymmetry constant of* X. The result of [41] says that

$$\sup \{s(X) |\ X \in \mathscr{M}_n\} \geqslant c n^{1/2},$$

where $c > 0$ does not depend on n. (Note that $s(X) \leqslant d(X, l_2^n) s(l_2^n) \leqslant n^{1/2}$ for $X \in \mathscr{M}_n$.)

4

In the last part of this report I present some recent results from the theory of operator ideals. In many situations in Banach space theory, and especially in the local theory, certain operator ideals play a very important role, but the scope of their applications is much wider. Particularly important in those applications are the ideals of p-summing operators, $0 < p < \infty$, and some related ideals (cf. [39], [22], [58], [33]).

Let us recall the definition of the (p, q)*-summing operators*, $0 < q \leqslant p < \infty$. Given Banach spaces X, Y, the space $\Pi_{p,q}(X, Y)$ consists of such operators $T \in L(X, Y)$ that $\pi_{p,q}(T) < \infty$, where

$$\pi_{p,q}(T) = \sup \Big(\sum_{i \leqslant n} \|Tx_i\|^p \Big)^{1/p},$$

the supremum being taken over all finite sequences $x_1, \ldots, x_n \in X$ such that $\sum_{i \leqslant n} |x^*(x_i)|^q \leqslant \|x^*\|^q$ for $x^* \in X^*$. The *p-summing operators* correspond to the case $q = p$, in this case one writes simply $\Pi_p(X, Y)$ and $\pi_p(T)$.

We refer to the treatise [49] for comprehensive information about operator ideals. Here we just mention that, if a class $\mathfrak{A}$ of operators is an operator ideal, then for any Banach spaces X, Y the intersection $\mathfrak{A} \cap L(X, Y)$ is a linear space denoted by $\mathfrak{A}(X, Y)$, and if $A \in \mathfrak{A}(X, Y)$, $T \in L(W, X)$ and $S \in L(Y, Z)$ then $SAT \in \mathfrak{A}(W, Z)$.

I shall discuss the problem of the distribution of eigenvalues of bounded linear operators. Although the problem may seem to be far from the local theory of Banach spaces, it has been solved by using basically finite-dimensional methods, many of them being analogous to those used in the local theory. Also the possibility of applying general results concerning classes of linear operators in a given space often depends strongly on the local geometry of the space (cf., e.g., [8])

For the sake of simplicity, we consider here complex Banach spaces. If $T \in L(X, X)$ and $T - \lambda \mathrm{id}_X$ is a Fredholm operator for $\lambda \neq 0$, we denote by $(\lambda_j(T))_{j \geqslant 1}$ the sequence of eigenvalues of T counted according to their multiplicities and ordered so that $|\lambda_1(T)| \geqslant |\lambda_2(T)| \geqslant \ldots$ We put $\lambda_n(T) = 0$ if T has fewer than n eigenvalues.

The classical result of H. Weyl ([60]) shows that, if H is a Hilbert space and $0 < p < \infty$, then the Schatten class $S_p(H)$ is an ideal in the algebra $L(H, H)$ such that $(\lambda_j(T)) \in l_p$ for $T \in S_p(H)$. Hence the degree of compactness of T predetermines the rate of decreasing of the eigenvalues.

A generalization of this fact would be to find an operator ideal $\mathfrak{A}_p$ such that (i) $\mathfrak{A}_p(H, H) \supseteq S_p(H)$ and (ii) for any X, if $T \in \mathfrak{A}_p(X, X)$, then $(\lambda_j(T)) \in l_p$.

This problem is a kind of equation in which the unknown quantity is an operator ideal and one looks for a maximal solution which, hopefully, has some useful properties and can be related to more familiar objects. It is easy to prove, cf. [49], that for $p = 2$ the operator ideal Π_2 is a solution. Since $\Pi_p(H, H) = S_2(H)$, for $0 < p < \infty$, the p-summing operators may have just 2-summable eigenvalues if $p < 2$. However, if $p > 2$, then *the ideal Π_p satisfies* (ii). This non-trivial fact was proved in [27].

One has $\Pi_{p,2}(H, H) = S_p(H)$ for $p > 2$, but the ideal $\Pi_{p,2}$ lacks property (ii). In fact, it was shown in [34] that, *if* $T \in \Pi_{p,2}(X, X)$, *then* $|\lambda_j(T)| \leqslant c_p \pi_{p,2}(T) j^{-1/p}$ *for* $j = 1, 2, \ldots$ and that no stronger conclusion can be obtained.

It was also shown in [27] that, if $T \in L(X, X)$ can be factored as

$T = T_n \circ \ldots \circ T_1$, with $\pi_{p_i}(T_i) < \infty$, $p_i \geqslant 2$, then the eigenvalues of T are p-summable where $p^{-1} = \sum_{i \leqslant n} p_i^{-1}$. Analogous results were obtained in [34] for compositions of $(p_i, 2)$-summing operators.

While those results were rather strong and had interesting applications, they did not generalize (except for the case of Π_2) the inequality of Weyl, which says that, for $T \in L(H, H)$, $0 < p < \infty$ and $n = 1, 2, \ldots, \infty$,

$$\sum_{j=1}^{n} |\lambda_j(T)|^p \leqslant c_p \sum_{j=1}^{n} s_j(T)^p$$

where $s_j(T) = \lambda_j\big((T^*T)^{1/2}\big)$ are the so-called *singular numbers* of the operator T. (In this case $c_p = 1$.)

It seemed natural to look for a solution of the form $\mathfrak{A}_p = \mathscr{L}_p^{(s)}$ where s is a generalization of the "singular numbers" to all operators and

$$\mathscr{L}_p^{(s)}(X, Y) = \{T \in L(X, Y) \mid \big(s_j(T)\big) \in l_p\},$$

or, more generally, if $0 < u \leqslant \infty$,

$$\mathscr{L}_{p,u}^{(s)}(X, Y) = \{T \in L(X, Y) \mid \big(s_j(T)\big) \in l_{p,u}\}$$

where $l_{p,u}$ is a Lorentz sequence space.

The largest among the reasonable generalizations of the s_j's (cf. [49]) are the *approximation numbers* defined, for $T \in L(X, Y)$ and $n = 1, 2, \ldots$, by

$$a_n(T) = \inf\{\|T - S\| \mid S \in L(X, Y), \operatorname{rank} S < n\}.$$

If $T \in L(H, H)$, then $a_n(T) = s_n(T)$ for each n. The Weyl-type inequality with s_j replaced by a_j was obtained in [31] (where $c_p > 1$), hence the ideals $\mathscr{L}_p^{(a)}$ solve the above problem for $0 < p < \infty$.

Even better solutions $\mathscr{L}_p^{(c)}$ and $\mathscr{L}_p^{(d)}$ were found in [27]. They correspond to the so-called Gelfand numbers and Kolmogorov numbers (cf. [49]).

The approach of [32] and [34] led A. Pietsch to the notion of *Weyl numbers*

$$x_n(T) = \sup\{a_n(TU) \mid U \in L(l_2, X), \|U\| \leqslant 1\},$$

$n = 1, 2, \ldots$, and to remarkable simplifications of the method (cf. [50]). He proved that, for $0 < p, u \leqslant \infty$,

$$T \in \mathscr{L}_{p,u}^{(x)}(X, X) \quad \textit{implies} \quad \big(\lambda_n(T)\big) \in l_{p,u}.$$

Since, for $p \geqslant 2$, $\Pi_{p,2} \subset \mathscr{L}^{(x)}_{p,\infty}$, and the ideals $\mathscr{L}^{(x)}_{p,u}$ satisfy a nice composition formula (in the indices p and u), these results improve those of [34] and clarify them as well. The improvement is rather subtle, since $\Pi_{p,2} \supset \mathscr{L}^{(x)}_{p,p}$ for $p > 2$ and $\Pi_2 \supset \mathscr{L}^{(x)}_{2,1}$. Also, for no $p \geqslant 2$ does one have $\Pi_p \subseteq \mathscr{L}^{(x)}_{p,p}$.

An alternative solution improving those of [27], was found in [7]. The operator ideals $\mathscr{L}^{(e)}_{p,u}$ are defined by using the *entropy numbers* $(e_n(T))_{n\geqslant 1}$ (cf. [49]), which are related to the entropy of $T(\mathrm{Ball}_X)$ in Y. One has a very simple relation

$$|\lambda_n(T)| \leqslant \sqrt{2}\, e_n(T)$$

for $T \in L(X, X)$ and $n = 1, 2, \ldots$, (cf. [7]).

The ideals $\mathscr{L}^{(x)}_{p,u}$ and $\mathscr{L}^{(e)}_{p,u}$ have similar properties, but for some operators T the sequences $(e_n(T))$ and $(x_n(T))$ behave very differently. The eigenvalues of many natural operators in Analysis can be studied by using one of those methods. For examples we refer, e.g., to the papers we have quoted. Those applications include some integral operators either weakly singular or with a smooth kernel (cf. [51]) and, more generally, operators in Sobolev or Besov spaces which "improve smoothness".

For instance, if M is a compact smooth manifold and T maps the Besov space $B^q_{p,u}(M)$ into a smaller space $B^\sigma_{q,v}(M)$, then the properties of the embedding $I\colon B^\sigma_{q,v} \to B^q_{p,u}$ reflect on the eigenvalues of IT. If $p \geqslant q$ one uses $e_n(I)$, if $p \leqslant q$ then either $x_n(I)$ or their dual analogue is needed. The study of the operator I can be made simply (cf. [51]) in terms of the spline bases on M constructed in [9] and [10]. In each case the methods yield optimal indices of summability under given assumptions.

References

[1] Alon N. and Milman V. D., Embedding of l^k_∞ in Finite Dimensional Banach Spaces, *Israel J. Math.* **45** (1983), pp. 265–280.

[2] Amir D. and Milman V. D., Unconditional and Symmetric Sets in n-Dimensional Normed Spaces, *Israel J. Math.* **37** (1980), pp. 3–20.

[3] Bourgain J., A Counterexample to a Complementation Problem, *Compositio Math.* **43** (1981), pp. 133–144.

[4] Bourgain J., New Classes of $\mathscr{L}^p$-Spaces, *Lecture Notes in Math.* **889**, Springer-Verlag, Berlin–Heidelberg–New York 1981.

[5] Bourgain J. and Delbaen F., A Class of Special $\mathscr{L}_\infty$-Spaces, *Acta Math.* **145** (1980), pp. 155–176.

[6] Brunel A. and Sucheston L., On B-Convex Banach Spaces, *Math. Systems Theory* **7** (1974), pp. 294–299.

[7] Carl B., Entropy Numbers, s-Numbers, and Eingenvalue Problems, *J. Funct. Anal.* **41** (1981), pp. 290–306.
[8] Carl B., On a Characterization of Operators from l_q into a Banach Space of Type p with Some Applications to Eingenvalue Problems, *J. Funct. Anal.* **48** (1982), pp. 394–407.
[9] Ciesielski Z., Constructive Function Theory and Spline Systems, *Studia Math.* **53** (1975), pp. 277–302.
[10] Ciesielski Z. and Figiel T., Spline Bases in Classical Function Spaces on Compact C^∞ Manifolds, *Studia Math.* **76** (1983), pp. 1–58, 95–136.
[11] Dubinsky Ed., Pełczyński A., and Rosenthal H. P., On Banach Spaces X for which $\Pi_2(\mathscr{L}_\infty, X) = B(\mathscr{L}_\infty, X)$, *Studia Math.* **44** (1972), pp. 617–648.
[12] Dvoretzky A., Some Results on Convex Bodies and Banach Spaces. In: *Proc. Symp. on Linear Spaces*, Jerusalem 1961, pp. 123–160.
[13] Elton J., Sign-Embedding of l_1^n, *Trans. Amer. Math. Soc.* **279** (1983), pp. 113–124.
[14] Figiel T. and Johnson W. B., A Uniformly Convex Banach Space which Contains no l_p, *Compositio Math.* **29** (1974), pp. 179–190.
[15] Figiel T. and Johnson W. B., Large Subspaces of l_∞^n and Estimates of the Gordon-Lewis Constant, *Israel J. Math.* **37** (1980), pp. 92–112.
[16] Figiel T., Johnson W. B., and Schechtman G., *Factorization of Natural Embeddings of l_p^n into L_r*, to appear.
[17] Figiel T., Johnson W. B., and Tzafriri L., On Banach Lattices and Spaces Having Local Unconditional Structure, with Applications to Lorentz Function Spaces, *J. Approx. Theory* **13** (1975), pp. 395–412.
[18] Figiel T., Lindenstrauss J., and Milman V. D., The Dimension of Almost Spherical Sections of Convex Bodies, *Acta Math.* **139** (1977), pp. 53–94.
[19] Garling D. J. H. and Gordon Y., Relations Between Some Constants Associated with Finite Dimensional Banach Spaces, *Israel J. Math.* **9** (1971), pp. 346–361.
[20] Gluskin E. D., The Diameter of the Minkowski Compactum is Roughly Equal to n, *Funkcional. Anal. i Priložen.* **15** (1981), pp. 72–73 (in Russian).
[21] Gordon Y., *Some Inequalities for Gaussian Processes and Applications*, to appear.
[22] Gordon Y. and Lewis D. R., Absolutely Summing Operators and Local Unconditional Structure, *Acta Math.* **133** (1974), pp. 27–48.
[23] Grothendieck A., Sur certaines classes de suites dans les espaces de Banach et le théorème de Dvoretzky–Rogers, *Bol. Soc. Mat. Sao Paulo* **8** (1953), pp. 83–110.
[24] Hoffmann-Jørgensen J., Probability in Banach Spaces, *Lecture Notes in Math.* **598**, Springer-Verlag, Berlin–Heidelberg–New York 1977.
[25] James R. C., Super-Reflexive Banach Spaces, *Canad. J. Math.* **24** (1972), pp. 896–904.
[26] Johnson W. B., On Finite Dimensional Subspaces of Banach Spaces with Local Unconditional Structure, *Studia Math.* **51** (1974), pp. 223–238.
[27] Johnson W. B., König H., Maurey B., and Retherford J. R., Eigenvalues of p-Summing and l_p-Type Operators in Banach Spaces, *J. Funct. Anal.* **32** (1979), pp. 353–380.
[28] Johnson W. B., Maurey B., Schechtman G., and Tzafriri L., Symmetric Structures in Banach Spaces, *Mem. Amer. Math. Soc.* **217** (1979).
[29] Johnson W. B. and Schechtman G., Embedding l_p^m into l_1^n, *Acta Math.* **149** (1982), pp. 71–85.

[30] Kashin B. S., Diameters of Some Finite Dimensional Sets and Classes of Smooth Functions, *Izv. Akad. Nauk SSSR Ser. Mat.* **41** (1977), pp. 334–351 (Russian).
[31] König H., Interpolation of Operator Ideals with an Application to Eigenvalue Distribution Problems, *Math. Ann.* **233** (1978), pp. 35–48.
[32] König H., Weyl-Type Inequalities for Operators in Banach Spaces. In: *Proc. Conf. Functional Analysis*, Paderborn 1979, North-Holland, Amsterdam–New York–Oxford 1980, pp. 297–317.
[33] König H., Type Constants and $(q, 2)$-Summing Norms Defined by n Vectors, *Israel J. Math.* **37** (1980), pp. 130–138.
[34] König H., Retherford J. R., and Tomczak-Jaegermann N., On the Eigenvalues of $(p, 2)$-Summing Operators and Constants Associated with Normed Spaces, *J. Funct. Anal.* **37** (1980), pp. 88–126.
[35] Krivine J.-L., Sous-espaces de dimension finie des espaces de Banach reticules, *Ann. of Math.* **104** (1976), pp. 1–29.
[36] Kwapień S., Isomorphic Characterizations of Inner Product Spaces by Orthogonal Series with Vector Valued Coefficients, *Studia Math.* **44** (1972), pp. 583–595.
[37] Lemberg H., Nouvelle demonstration d'un théorème de J. L. Krivine sur la finie representation de l_p dans un espace de Banach, *Israel J. Math.* **39** (1981), pp. 341–348.
[38] Lévy P., *Problèmes concrets d'analyse fonctionnelle*, Gauthier-Villars, Paris 1951.
[39] Lindenstrauss J. and Pełczyński A., Absolutely Summing Operators in $\mathscr{L}_p$-Spaces and their Applications, *Studia Math.* **29** (1968), pp. 275–326.
[40] Lindenstrauss J. and Tzafriri L., *Classical Banach Spaces*, Springer-Verlag, Berlin–Heidelberg–New York, vol. I: *Sequence Spaces* — 1977, vol. II: *Function Spaces* — 1979.
[41] Mankiewicz P., Finite Dimensional Banach Spaces with Symmetry Constant of Order $\sqrt{n}$, *Studia Math.* **79** (1984), to appear.
[42] Maurey B., Constructions de suites symétriques, *C. R. Acad. Sci. Paris, Ser. A–B* **288** (1979), pp. 679–681.
[43] Maurey B., Sous-espaces l^p des espaces de Banach, Seminaire Bourbaki, 1982/83, no. 608.
[44] Maurey B. and Pisier G., Series de variables aléatoires vectorielles indépendantes et propriétés géométriques des espaces de Banach, *Studia Math.* **58** (1976) pp. 45–90.
[45] Milman V. D., A New Proof of the Theorem of A. Dvoretzky on Sections of Convex Bodies, *Funkcional. Anal. i Prilożen.* **5** (1971), pp. 28–37 (Russian).
[46] Milman V. D. and Sharir M., A New Proof of the Maurey–Pisier Theorem, *Israel. J. Math.* **33** (1979), pp. 73–87.
[47] Pajor A., Plongement de l_n^1 dans les espaces de Banach complexes, *C.R. Acad. Sci. Paris, Ser. A–B* **296** (1983), pp. 741–743.
[48] Pełczyński A. and Rosenthal H. P., Localization Techniques in L^p-Spaces, *Studia Math.* **52** (1975), pp. 263–289.
[49] Pietsch A., *Operator Ideals*, Verlag der Wissenschaften and North-Holland, Berlin 1978.
[50] Pietsch A., Weyl Numbers and Eigenvalues of Operators in Banach Spaces, *Math. Ann.* **247** (1980), pp. 149–168.
[51] Pietsch A., Eigenvalues of Integral Operators, I, *Math. Ann.* **247** (1980), pp. 169–178.

[52] Pisier G., Holomorphic Semi-Groups and the Geometry of Banach Spaces, *Ann. of Math.* **15** (1982), pp. 375–392.

[53] Pisier G., De nouvelles caractérisations des ensembles de Sidon, *Adv. in Math. Suppl. Stud.*, vol. 7B, Academic Press, New York 1981.

[54] Pisier G., On the Dimension of the l_p^n-Subspaces of Banach Spaces, for $1 \leqslant p < 2$, *Trans. Amer. Math. Soc.* **276** (1983), pp. 201–211.

[55] Rosenthal H. P., On a Theorem of J. L. Krivine Concerning Block Finite Representability of l^p in General Banach Spaces, *J. Math. Anal. Appl.* **28** (1978), pp. 197–225.

[56] Schütt C., On the Uniqueness of Symmetric Bases in Finite Dimensional Banach Spaces, *Israel J. Math.* **40** (1981), pp. 97–117.

[57] Szankowski A., On Dvoretzky's Theorem on Almost Spherical Sections of Convex Bodies, *Israel J. Math.* **17** (1974), pp. 325–338.

[58] Tomczak-Jaegermann N., Computing 2-Summing Norm with Few Vectors, *Ark. Mat.* **17** (1979), pp. 273–277.

[59] Tomczak-Jaegermann N., The Banach–Mazur Distance between Symmetric Spaces, *Israel J. Math.* **46** (1983), pp. 40–66.

[60] Weyl H., Inequalities between the Two Kinds of Eigenvalues of a Linear Transformation, *Proc. Nat. Acad. Sci. U.S.A.* **35** (1949), pp. 408–411.

[61] Zippin M., On Some Subspaces of Banach Spaces whose Duals are L_1-Spaces, *Proc. Amer. Math. Soc.* **23** (1969), pp. 378–385.

INSTITUTE OF MATHEMATICS
POLISH ACADEMY OF SCIENCES
81-825 SOPOT
POLAND

Proceedings of the International Congress of Mathematicians
August 16-24, 1983, Warszawa

Б. С. КАШИН

Некоторые результаты об оценках поперечников

Статья посвящена оценкам поперечников классов гладких функций и конечномерных множеств. Рассматриваемая тематика, как по постановкам задач, так и по методам их решения, принадлежит к области пограничной между теорией приближения, теорией ортогональных рядов и геометрией нормированных пространств. В последние годы оценки поперечников начинают находить применение в прикладной математике, при анализе качества вычислительных алгоритмов и в функциональном анализе, при изучении распределения собственных чисел операторов. Задачи о поперечниках в последние десять лет интенсивно изучались и наша цель — дать обзор некоторых, полученных здесь результатов.

Пусть X — банахово пространство, K — компакт в X, $L \subset X$ — подпространство. Положим

$$\Delta_X(K, L) = \sup_{y \in K} \inf_{z \in L} \|y - z\|_X. \tag{1}$$

Определение 1 (А. Н. Колмогоров [10]). *n-поперечником* компакта K в X называется величина

$$d_n(K, X) = \inf_{L} \Delta_X(K, L),$$

где inf берется по всем подпространствам $L \subset X$ размерности $\leqslant n$.

Определение 2. *Линейным n-поперечником* компакта K в X называется величина

$$\delta_n(K, X) = \inf_{T} \sup_{x \in K} \|x - Tx\|_X,$$

где inf берется по всем, действующим в X линейным операторам T ранга $\leqslant n$ ($T\colon X \to L$, $L \subset X$, $\dim L \leqslant n$).

Исторически первая и наиболее известная задача о поперечниках — задача об оценке поперечников классов Соболева W_p^r в пространствах L^q, $1 \leqslant q \leqslant \infty$ ($L^\infty = C$). Напомним, что для $r > 0$

$$W_p^r(-\pi, \pi) = \{f\colon \|f\|_{L^p(-\pi,\pi)} + \|f^{(r)}\|_{L^p(-\pi,\pi)} \leqslant 1\},$$

где

$$f^{(r)}(x) = \sum_{k=-\infty}^{\infty} (ik)^r c_k(f) \cdot e^{ikx} \quad \text{и} \quad c_k(f) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) \cdot e^{-ikt}\, dt$$

— коэффициенты Фурье функции $f(x)$. Аналогично определяется класс $W_p^r(T^s)$-функций s переменных, заданных на торе $T^s = (-\pi, \pi)^s$.

Ниже пойдет речь о скорости убывания при $n \to \infty$ поперечников

$$d_n(W_p^r, L^q),^1 \tag{2}$$

при этом мы ограничимся случаем $s = 1$, так как рассмотрение классов функций многих переменных не вносит здесь существенных дополнительных трудностей.

Порядки поперечника (2) при $p = q = 2$ были определены еще А. Н. Колмогоровым [10]. В 50-х годах эти вопросы рассматривались Рудиным [15] и С. Б. Стечкиным [16]. В [15] показано, что

$$d_n(W_1^1, L^2) \asymp n^{-1/2},^2 \tag{3}$$

а в [16] найдены порядки величины (2) при $p = q = \infty$.

В 60-х годах были определены порядки поперечников (2) для любых p и q с $p \geqslant q$, а затем и при $1 \leqslant p < q \leqslant 2$; были вычислены и точные значения некоторых поперечников (см. подробнее [5], [7], [18]). Во всех указанных случаях показывалось, что

$$d_n(W_p^r, L^q) > c \cdot \Delta_{Lq}(W_p^r, T_n); \quad c = c(p, q) > 0, \tag{4}$$

где $T_n = \left\{P(x)\colon P(x) = \sum_{k=-n}^{n} a_k e^{ikx}\right\}$ — пространство тригонометрических полиномов. Неравенство (4) означает, что при $q \leqslant \max(2, p)$ пространство тригонометрических полиномов наилучшим (по порядку)

---

[1] При рассмотрении поперечников (2) конечно всегда предполагают, что $W_p^r(T^s)$ — компакт в $L^q(T^s)$, для этого необходимо и достаточно, чтобы $r/s - 1/p + 1/q > 0$ (см. [1]).

[2] Запись $a_n \asymp b_n$ означает, что при $n = 1, 2, \ldots$, $0 < A < a_n/b_n < B < \infty$, где A и B — абсолютные постоянные.

возможным образом приближает класс W_p^r в метрике L^q. С тридцатых годов известно, что

$$\mathcal{A}_{L^q}(W_p^r, T_n) \asymp \begin{cases} n^{-r}, & \text{если } p \geqslant q, \\ n^{-r+1/p-1/q}, & \text{если } p < q, \end{cases} \tag{5}$$

поэтому из соотношения (4) сразу находились порядки поперечников (2) при $q \leqslant \max(2, p)$.

Р. С. Исмагиловым [5] было показано, что при $q > \max(2, p)$ поведение величин (2) резко меняется. В частности, в [5] получено неравенство

$$d_n(W_1^2, C) < c \cdot n^{-6/5} \cdot \ln n < c' \cdot n^{-1/5} \cdot \ln n \cdot \mathcal{A}_{L^\infty}(W_1^2, T_n),$$

из которого следует, что обычные тригонометрические полиномы не являются хорошим аппаратом приближения класса W_1^2 в метрике C. В этом случае, в качестве „хорошо приближающего" n-мерного подпространства, Р. С. Исмагиловым было взято подпространство вида:

$$T_n' = \left\{P(x)\colon P(x) = \sum_{j=1}^{n} a_j e^{ik_j x}\right\}, \tag{6}$$

где $\{k_j\}_{j=1}^n$ — некоторый, специально выбранный, набор целых чисел. Забегая вперед, отметим, что с помощью подпространств вида (6) можно хорошо аппроксимировать классы W_1^r $(r > \frac{3}{2})$ и показать (см. [13]), что при $q > 2$

$$\inf_{T_n'} \mathcal{A}_{L^q}(W_1^r, T_n') \asymp d_n(W_1^r, L^q) \asymp n^{-r+1/2} \tag{7}$$

(в (7) inf берется по всем подпространствам вида (6)). В связи с этими результатами возникло понятие тригонометрического поперечника класса функций K в пространстве L^q:

$$\inf_{T_n'} \mathcal{A}_{L^q}(K, T_n'),$$

однако для $K = W_p^r, 1 < p < \infty = q$ точно оценить эту величину не удается и нет оснований считать, что по порядку она совпадает с поперечником (2).

Уже в первых работах, посвященных поперечникам классов гладких функций (см. например [16]) производили дискретизацию задачи, т.е. сведение ее к задаче о поперечниках некоторого множества в R^m. При изучении классов Соболева таким конечномерным аналогом

является задача об оценке величины

$$d_n(B_p^m, l_q^m), \tag{8}$$

где B_p^m — единичный шар в пространстве l_p^m и для $x = \{x_j\}_{j=1}^m \in R^m$

$$\|x\|_{l_p^m} = \begin{cases} \left(\sum_{j=1}^m |x_j|^p\right)^{1/p}, & \text{если } 1 \leqslant p < \infty, \\ \max\limits_{1 \leqslant j \leqslant m} |x_j|, & \text{если } p = \infty. \end{cases}$$

Достаточно точные способы дискретизации были предложены в работах [3] и [13], однако отсутствие хороших оценок величины (8) не давало возможности продвинуться в решении задачи о скорости убывания поперечников классов Соболева. Автором в работе [7] (см. также [6]) был предложен метод оценки поперечников (8), а следовательно и (2), основанный на вероятностных соображениях. Кратко этот метод можно описать так: рассматривается достаточно широкое множество $\Omega_{m,n}$ n-мерных подпространств $L \subset R^m$, наделенное мерой μ (например множество всех n-мерных подпространств в R^m с мерой Хаара или множество подпространств $L \subset R^m$, $\dim L = n$, которые в стандартном базисе задаются матрицей с элементами ± 1 и $\mu L = 2^{-mn}$) и оценивается величина

$$I_{p,q} = I_{p,q}(m, n) = \int\limits_{\Omega_{m,n}} A_{l_q^m}(B_p^m, L)\, d\mu.$$

Очевидно, что

$$d_n(B_p^m, l_q^m) = \inf_{L,\, \dim L \leqslant n} A_{l_q^m}(B_p^m, L) \leqslant I_{p,q}. \tag{9}$$

Оказывается, что в ряде случаев разница между левой и правой частью в (9) невелика и точно оценив интеграл $I_{p,q}$, мы получим хорошую оценку поперечника (8). К примеру, указанным методом автором было показано (см. [7], а также статью автора в сборнике [14]), что при любых p и q, $1 \leqslant p, q \leqslant \infty$

$$d_n(B_p^{2n}, l_q^{2n}) \asymp I_{p,q} \asymp \begin{cases} 1 & \text{если } 1 \leqslant p < q \leqslant 2, \\ n^{-1/2+1/q}, & \text{если } p < 2 < q, \\ n^{-1/p+1/q}, & \text{если } p \geqslant \min(2, q). \end{cases} \tag{10}$$

Из (10) следует, что случайное n-мерное подпространство при любых p и q хорошо аппроксимирует шар B_p^{2n} в метрике l_q^{2n}, в то время как пространство дискретных тригонометрических полиномов той-же раз-

мерности дает хорошее приближение шара B_p^{2n} в l_q^{2n} только при $q \leqslant \max(2, p)$.

С помощью описанного подхода в [7] доказана и

Теорема 1. *При* $m > n$

$$\tfrac{1}{2} n^{-1/2} < d_n(B_2^m, l_\infty^m) < C n^{-1/2} \left(1 + \ln \frac{m}{n}\right)^{3/2}.$$

Отметим, что доказать соотношение (10) заметно проще, чем получить достаточно точные (см. теоремы 1 и 4) оценки поперечников $d_n(B_p^m, l_q^m)$, $1 \leqslant p \leqslant 2 < q \leqslant \infty$ при m много больше, чем n. В последнем случае требуется подробнее изучить геометрию шаров B_p^m.

Для того, чтобы дать представление о том, какого типа геометрические вопросы возникают при оценках поперечников (8), приведем один результат, полученный в [7] попутно.

Теорема 2. *Для каждого* $m = 1, 2, \ldots$ *найдется такое ортогональное преобразование* T'_m *пространства* R^m, *что*

$$m^{-1/2} \cdot B_2^m \subset B_1^m \cap T'_m(B_1^m) \subset K m^{-1/2} B_2^m,$$

где K — *абсолютная постоянная, а* $a \cdot B_2^m$ — *шар радиуса* a.

Оценки поперечника $d_n(B_2^m, l_\infty^m)$ оказались наиболее важными для приложений. Опираясь на теорему 1 и используя отмеченные выше способы дискретизации, в [7] автор завершил определение порядков поперечников $d_n(W_p^r, L^q)$, $r > 1$. В итоге оказалось, что имеет место

Теорема 3. *Пусть* $1 \leqslant p, q \leqslant \infty, r \cdot p > 1$. *Тогда*

$$d_n(W_p^r, L^q) \asymp \begin{cases} n^{-r}, & \text{если } p \geqslant q \text{ или } 2 < p < q, \\ n^{-r-1/2+1/p}, & \text{если } p \leqslant 2 < q, \\ n^{-r-1/q+1/p}, & \text{если } 1 \leqslant p < q \leqslant 2. \end{cases}$$

Как видно из теоремы 3 и (5) при $q > \max(2, p)$ поперечник (2) существенно меньше, чем $\Delta_{Lq}(W_p^r, T_n)$.

Для полноты рассмотрения вопроса оставалось оценить поперечник (2) при $rp \leqslant 1$, т.е. когда класс W_p^r не компактен в $C(-\pi, \pi)$. Оказалось (см. [9]), что в этом случае поведение поперечника (2) резко меняется и для его точной оценки необходимо знать точные, равномерные по n и m оценки величин (8) при $q > \max(2, p)$. Такие оценки были получены в частном случае $p = 1$ автором [8] и в общем случае Е. Д. Глускиным [4], который, используя описанный выше вероятностный метод, доказал, что справедлива

Теорема 4. *Пусть* $1 \leqslant p < q < \infty$, $n < m$. *Тогда*

$$0 < c_1(p, q) \leqslant \frac{d_n(B_p^m, l_q^m)}{\Phi(m, n, p, q)} \leqslant c_2(p, q),$$

где $c_1(p, q)$ *и* $c_2(p, q)$ *— константы, зависящие только от* p *и* q, *а*

$$\Phi(m, n, p, q) = \begin{cases} (\min\{1, m^{1/q} \cdot n^{-1/2}\})^{\frac{1/p-1/q}{1/2-1/q}}, & 2 \leqslant p < q \leqslant \infty, \\ \max\left\{m^{1/q-1/p}, \left(1 - \frac{n}{m}\right)^{1/2} \cdot \min\{1, m^{1/q} \cdot n^{-1/2}\}\right\}, & \\ \qquad\qquad 1 \leqslant p < 2 \leqslant q \leqslant \infty, & \\ \max\left\{m^{1/q-1/p}, \left(1 - \frac{n}{m}\right)^{\frac{1}{2}\frac{1/p-1/q}{1/p-1/2}}\right\}, & 1 \leqslant p < q \leqslant 2. \end{cases} \tag{11}$$

Таким образом в настоящее время при $q < \infty$ порядки поперечников (8) полностью определены. Вопросы остаются только в случае $q = \infty$; неизвестно, в частности, каков порядок (при $n \to \infty$) величин $d_n(B_p^{n^2}, l_\infty^{n^2})$, $p > 1$.

С помощью теоремы 4 Е. Д. Куланин [11] показал, что справедлива

Теорема 5. *Пусть* $-1/q < r - 1/p \leqslant 0$, $q > \max(2, p)$. *При* $n \to \infty$ *справедливы соотношения*

$$d_n(W_p^r, L^q) \asymp \begin{cases} n^{-r}, & \text{если } p > 2 \text{ и } r > \frac{1}{2} \cdot \frac{1/p - 1/q}{1/2 - 1/q}, \\ n^{-\frac{q}{2}\left(r - \frac{1}{p} + \frac{1}{q}\right)}, & \text{если } p > 2 \text{ и } r < \frac{1}{2} \cdot \frac{1/p - 1/q}{1/2 - 1/q}, \\ n^{-\frac{q}{2}\left(r - \frac{1}{p} + \frac{1}{q}\right)}, & \text{если } p \leqslant 2 \text{ и } rp < 1. \end{cases}$$

Утверждение теоремы 5 при $p = 1$ было получено ранее в [9]. Обратим внимание на то, что в теореме 5 не рассмотрены случаи, когда $rp = 1$, $1 \leqslant p \leqslant 2$ и $r = \frac{1}{2} \cdot \frac{1/p - 1/q}{1/2 - 1/q}$, $2 < p < \infty$. Здесь знание порядков величин $d_n(B_p^m, l_q^m)$ не достаточно для точной оценки поперечников классов Соболева. Более того, Е. Д. Куланиным было показано, что для любого $\varepsilon > 0$ и $q \in (2, \infty)$ найдутся такие постоянные $c_{q,\varepsilon}$ и C_q', что

$$0 < c_{q,\varepsilon} \cdot n^{-1/2} \cdot \ln^{1/2-\varepsilon} n < d_n(W_1^1, L^q) \leqslant C_q' \cdot n^{-1/2} \cdot \ln n, \quad n = 2, 3, \ldots, \tag{12}$$

т.е. в оценках для поперечника (2) могут по существу возникать множители вида $\ln^{\gamma} n$. Интересно сравнить оценки (3) и (12).

Коснемся теперь вопроса о линейных поперечниках. Вероятностный подход оказался применимым и в задаче о порядках при $n \to \infty$ величин $\delta_n(W_p^r, L^q)$ и $\delta_n(B_p^m, l_q^m)$. Е. Д. Глускин [4], усиливая более ранний результат Хеллига [19], показал, что справедлива

Теорема 6. *Пусть* $1 \leqslant p < q \leqslant \infty$, $(p, q) \neq (1, \infty)$, $n < m$. *Тогда*

$$0 < c_3(p, q) < \frac{\delta_n(B_p^m, l_q^m)}{\Psi(m, n, p, q)} \leqslant c_4(p, q),$$

где $c_3(p, q)$ *и* $c_4(p, q)$ — *постоянные зависящие только от* p *и* q, *и* (см. (11))

$$\Psi(m, n, p, q) = \begin{cases} \Phi(m, n, p, q), & \text{если } 1 \leqslant p < q \leqslant p', \\ \Phi(m, n, q', p'), & \text{если } \max\{p, p'\} < q, \end{cases}$$

$$\frac{1}{p} + \frac{1}{p'} = \frac{1}{q} + \frac{1}{q'} = 1.$$

Теорема 6 выясняет порядки поперечников $\delta_n(B_p^m, l_q^m)$ во всех случаях за исключением случая $p = 1$, $q = \infty$ (отметим, что $\delta_n(B_1^m, l_\infty^m) = d_n(B_1^m, l_\infty^m)$). Окончательный результат о порядках линейных поперечников классов W_p^r, $r > 1$ был получен В. Е. Майоровым [13] и Хеллигом [19]:

Теорема 7. *Пусть* $r > 1$ *и* $p < 2 < q$. *Тогда*

$$\delta_n(W_p^r, L^q) \asymp n^{-r+1/p-1/q-\beta}, \quad \text{где}$$

$$\beta = \beta(p, q) = \min\left(\frac{1}{2} - \frac{1}{q}, \frac{1}{p} - \frac{1}{2}\right).$$

Ранее было известно (см. [5]), что при $p > 2$ или $q < 2$ $\delta_n(W_p^r, L^q) \asymp$ $\asymp \Delta_{Lq}(W_p^r, T_n)$. Отметим (см. теоремы 3 и 7), что линейные поперечники классов W_p^r, $p > 1$ в пространстве $C(-\pi, \pi)$ существенно больше, чем поперечники по А. Н. Колмогорову.

В настоящее время ясно, что использование случайных подпространств дает возможность определить порядки поперечников для широкого класса компактов. Некоторым недостатком вероятностного подхода является невозможность явного указания подпространства, хорошо аппроксимирующего компакт. Поэтому естественно возникает

вопрос (см. [14], стр. 320, а также [2]) о конструктивном доказательстве теоремы 3.

В заключение отметим, что рассмотренная нами задача о поперечниках классов Соболева имеет отчасти модельный характер и ставится конечно и для других классов функций. При этом может возникнуть потребность в изучении поперечников конечномерных множеств более сложного, чем шары B_p^m вида. Вместе с тем, применения оценок величин (8) не ограничиваются случаем классов W_p^r. Так, используя теорему 1 и ее следствия, В. Н. Темляков [17] определил порядки поперечников в пространствах $L^q(T^s)$, $1 < q < \infty$ классов $H_{p,\bar{a}}^{\bar{r}}$ $(p > 1)$ — функций s переменных представимым в виде свертки:

$$H_{p,\bar{a}}^{\bar{r}} = \left\{ f(x) \colon f(x) = \frac{1}{(2\pi)^s} \int\limits_{T^s} \varphi(y) \cdot F_{\bar{r}}(x-y, \bar{a})\, dy,\ \|\varphi\|_{L^q(T^s)} \leqslant 1 \right\},$$

где $\bar{r} = (r_1, \ldots, r_s)$, $\min\limits_{1 \leqslant j \leqslant s} r_j > 1$, $\bar{a} = (a_1, \ldots, a_s)$ и ядро определяется равенством

$$F_r(x, \bar{a}) = 2^s \sum_k \prod_{j=1}^{s} k_j^{-r_j} \cdot \cos\left(k_j \cdot x_j + \tfrac{1}{2} a_j \cdot \pi\right),$$

в котором суммирование производится по всем наборам $k = (k_1, \ldots, k_s)$, $k_j > 0$, $1 \leqslant j \leqslant s$. Открытой здесь остается, возникшая в начале 60-х годов, задача о порядках поперечников $d_n\left(H_{p,\bar{a}}^{\bar{r}}, C(T^s)\right)$.

Литература

[1] Бесов О. В., Ильин В. П., Никольский С. М., *Интегральные представления функций и теоремы вложения*, изд-во „Наука", Москва, 1975.

[2] de Boor C., DeVore R., and Höllig K., Mixed norm n-width, *Proc. Amer. Math. Soc.* **80** (1980), стр. 577–583.

[3] Глускин Е. Д., Об одной задаче о поперечниках, *Доклады АН СССР* **219**, № 3 (1974), стр. 527–530.

[4] Глускин Е. Д., Нормы случайных матриц и поперечники конечномерных множеств, *Матем. сборник* **120**, № 2 (1983), стр. 180–189.

[5] Исмагилов Р. С., Поперечники множеств в линейных нормированных пространствах и приближение функций тригонометрическими многочленами, *Успехи матем. наук* **29**, № 3 (1974), стр. 161–178.

[6] Кашин Б. С., О колмогоровских поперечниках октаэдров, *Доклады АН СССР* **214**, № 5 (1974), стр. 1024–1026.

[7] Кашин Б. С., Поперечники некоторых конечномерных множеств и классов

гладких функций, *Известия АН СССР* (*сер. матем.*) **41**, № 2 (1977), стр. 334–351.

[8] Кашин Б. С., О некоторых свойствах матриц ограниченных операторов из пространства l_2^m в l_2^m, *Известия АН Арм. ССР* (*сер. матем.*) **15**, № 5 (1980), стр. 379–394.

[9] Кашин Б. С., О поперечниках классов Соболева малой гладкости, *Вестник МГУ* (*сер. матем., механ.*) № **5** (1981), стр. 50–54.

[10] Kolmogoroff A., Über die beste Annäherung von Funktionen einer gegebenen Funktionenklasse, *Annals of Math.* **37**, № 1 (1936), стр. 107–111.

[11] Куланин Е. Д., Оценки поперечников классов Соболева малой гладкости, *Вестник МГУ* (*Сер. матем., механ.*) № 2 (1983), стр. 24–30.

[12] Майоров В. Е., Дискретизация задачи о поперечниках, *Успехи матем. наук* **30**, № **6** (1975), стр. 179–180.

[13] Майоров В. Е., О линейных поперечниках соболевских классов, *Доклады АН СССР* **243**, № 5 (1978), стр. 1127–1130.

[14] *Quantitative approximation* (Proceedings of Bonn symposium, August 20–24, 1979), Academic Press, New York, 1980.

[15] Rudin W., L^2-approximation by partial sums of orthogonal developments, *Duke Math. J.* **19**, № 1 (1952), стр. 1–4.

[16] Стечкин С. Б., О наилучших приближениях заданных классов функций любыми полиномами, *Успехи матем. наук* **9**, № 1 (1954), стр. 133–134.

[17] Темляков В. Н., Поперечники некоторых классов функций нескольких переменных, *Доклады АН СССР* **267**, № 2 (1982), стр. 314–317.

[18] Тихомиров В. М., *Некоторые вопросы теории приближений*, Изд-во „МГУ", Москва, 1976.

[19] Höllig K., Approximationszahlen von Sobolev-Einbettungen, *Math. Annalen* **242** (1979), стр. 273–281.

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

G. G. KASPAROV

Operator K-theory and Its Applications: Elliptic Operators, Group Representations, Higher Signatures, C^*-extensions

1. Introduction

K-theory is a wide area of mathematics embracing substantial parts of algebra, topology and in recent years — analysis. I will be talking about *operator* K-theory, the K-theory of C^*-algebras. Its origins are in topological K-theory which may be regarded as a part of operator K-theory.

Recall that Grothendieck's group $K_0(A)$ for an associative ring with unit A is defined by introducing formal differences in a semigroup of isomorphism classes of finitely generated projective A-modules ([1]). This group is covariant in A. In topology $A = C(X)$, the algebra of continuous complex-valued functions on a compact space X. In this case projective A-modules correspond to complex vector bundles and the group $K_0(A)$ is denoted by $K^0(X)$. (Upper and lower position of indices reflects contra- and covariance properties respectively.) The group $K^0(X)$ gives rise to a $\mathbf{Z}_2$-graded cohomology theory K^* ([1]).

An independent development of operator K-theory began with the discovery of deep connections between elliptic operators and the K_*-homology theory dual to K^* ([2]). It appeared that by axiomatizing the notion of an elliptic operator one can define, for a C^*-algebra A, a group $K^0(A)$

dual to Grothendieck's $K_0(A)$ ([14], [15]). On the other hand, Grothendieck's group $K_0(A)$ itself appeared to be closely connected with index of elliptic operators on regular coverings of closed manifolds invariant under the discrete group of transformations of the covering (see [15], § 8). This led to an operator-theoretic definition of $K_0(A)$ in terms of Fredholm operators over C^*-algebra A (see [15], § 8; [18], § 6; [25], §§ 1, 2).

But the most important thing for the development of operator K-theory was its connection with C^*-extensions. Observed for a special case of extensions in [6], this connection appeared to be general and prompted a way of unifying both K_* and K^*-functors into the K-bifunctor $KK_*(A, B)$ ([16], [18]). We now come to formal definitions, but before that it is worth pointing out that unified homology-cohomology theories may be defined also in a purely topological situation (cf. [18], § 6, Def. 4). Very likely, this will give new methods not only in K-theory.

2. Definition of the K-bifunctor

We begin with generalizing the notion of Hilbert space (see [27], and [17], [18]). Let B be a C^*-algebra. A right B-module E is called a *Hilbert B-module* if there is a "B-scalar product" on E: $(x, y) \in B, \forall x, y \in E$ which is B-linear in y $\big((x, yb) = (x, y)b, \forall b \in B\big)$, selfadjoint $((y, x) = (x, y)^*)$, positive $((x, x) \geqslant 0)$ and non-degenerate (if $(x, x) = 0$, then $x = 0$). Moreover, E must be complete in the norm: $\|x\| = \|(x, x)\|^{1/2}$. We shall consider only Hilbert B-modules having a countable set $\{x_i\}$ with $\bigcup_{n=1}^{\infty} \big(\sum_{i=1}^{n} x_i B\big)$ dense in E.

A map of Hilbert B-modules $T: E_1 \to E_2$ is called an *operator* if there is an adjoint map $T^*: E_2 \to E_1$, such that $(Tx, y) = (x, T^*y)$, $\forall x \in E_1$, $y \in E_2$. The algebra of operators $\mathscr{L}(E)$ is a C^*-algebra. An ideal of compact operators $\mathscr{K}(E)$ is a closure of a subspace spanned by "rank 1" operators $\theta_{x,y}$: $\theta_{x,y}(z) = x(y, z)$, $x, y, z \in E$. For $B = \boldsymbol{C}$ we get the usual Hilbert space H, algebra $\mathscr{L}(H)$ and ideal $\mathscr{K}(H)$ (which is denoted by $\mathscr{K}$ when H is separable and infinite-dimensional).

Given a homomorphism of C^*-algebras $f: B \to D$, a Hilbert D-module $E \otimes_B D$ is defined as a completion of $E \otimes D$ in the norm corresponding to the D-scalar product: $(x_1 \otimes d_1, x_2 \otimes d_2) = d_1^* f\big((x_1, x_2)\big) d_2$. Finally, we need the notion of *graded* Hilbert B-module E. This is simply a sum of two Hilbert B-modules: $E = E^{(0)} \oplus E^{(1)}$.

Definition 1 ([18]). Let A and B be C^*-algebras, with A *separable*. Consider pairs (E, F) where E is a graded Hilbert B-module with A acting

on E from the left via a (diagonal) $*$-homomorphism $\varphi^{(0)} \oplus \varphi^{(1)}: A \to \mathscr{L}(E^{(0)} \oplus \oplus E^{(1)})$, and $F: E^{(0)} \to E^{(1)}$ is such an operator that $\forall a \in A$:

$$aF - Fa, \quad a(FF^* - 1), \quad a(F^*F - 1) \tag{1}$$

are compact operators. We shall identify isometrically isomorphic pairs and denote the set of all pairs by $\mathscr{E}_0(A, B)$. The direct sum operation makes this set a semigroup. A pair with all entries in (1) zero will be called *degenerate*. An addition of a degenerate pair will be called a *stabilization*. Pairs (E_0, F_0) and (E_1, F_1) are *homotopy equivalent* if they are restrictions of some $(\tilde{E}, \tilde{F}) \in \mathscr{E}_0(A, B \otimes C[0,1])$ to the end-points of $[0,1]$ (i.e., $E_t = \tilde{E} \otimes_{B \otimes C[0,1]} B$, $F_t = \tilde{F} \otimes 1$ for $t = 0, 1$). We define a semigroup $KK_0(A, B)$ by identifying in $\mathscr{E}_0(A, B)$ pairs which become homotopy equivalent after stabilization. Quite similarly, a semigroup $KK_1(A, B)$ is defined starting with pairs (E, S) where E is a Hilbert B-module (non-graded), $S \in \mathscr{L}(E)$ and, instead of (1), operators

$$aS - Sa, \quad a(S^2 - 1), \quad a(S^* - S) \tag{2}$$

are compact.

$KK_i(A, B)$ are in fact abelian groups, covariant in B, contravariant in A, and *homotopy invariant* in A and B. This means that if a path of homomorphisms $f_t: B_1 \to B_2$ is (pointwise) continuous in $t \in [0,1]$ then $(f_t)_*: KK_i(A, B_1) \to KK_i(A, B_2)$ does not depend on t (and similarly for A).

In the case of $A = \boldsymbol{C}$ we can identify $KK_0(A, B)$ with $K_0(B)$. Let B have a unit 1. Taking any $(E, F) \in \mathscr{E}_0(\boldsymbol{C}, B)$ we can find a $(\tilde{E}, \tilde{F}) \in \mathscr{E}_0(\boldsymbol{C}, B)$ differing from (E, F) by stabilization and a compact perturbation of F, so that $1 - \tilde{F}\tilde{F}^*$ and $1 - \tilde{F}^*\tilde{F}$ are compact projections ([18], § 6). But then the B-modules $\operatorname{Ker} \tilde{F}$ and $\operatorname{Ker} \tilde{F}^*$ are projective and finitely generated ([18], § 6). Considering their difference as an element of $K_0(B)$, we get an isomorphism: $KK_0(\boldsymbol{C}, B) \simeq K_0(B)$.

We shall denote $KK_i(\boldsymbol{C}, B)$ by $K_i(B)$ and $KK_i(A, \boldsymbol{C})$ by $K^i(A)$. For a locally compact space X we put $K_i(X) = K^i\big(C(X)\big)$, $K^i(X) = K_i\big(C(X)\big)$ where $C(X)$ is the algebra of continuous functions on X tending to 0 at ∞. Moreover, we put $KK^i(X, Y) = KK_i\big(C(X), C(Y)\big)$.

Now a few words about C^*-extensions: $0 \to B \to D \xrightarrow{q} A \to 0$ (cf. [18], § 7). They are closely related to the group $KK_1(A, B)$. Indeed, the left action of D on B defines a homomorphism $D \to \mathscr{L}(B)$ and (after dividing by B) a Busby homomorphism $\beta: A \to \mathscr{L}(B)/B$. If q admits a completely positive lifting, then applying a standard Stinespring construction ([27], [17]) we get a Hilbert B-module E and a homomorphism $\varphi: A \to \mathscr{L}(B \oplus E)$ such that $P\varphi(a)P \pmod{B} = \beta(a)$ where P is a projection $B \oplus E \to B$.

A pair $(B\oplus E, 2P-1)$ represents an element of $KK_1(A, B)$ corresponding to the initial extension.

The main techniques to deal with K-bifunctors are provided by the bilinear pairing ([18])

$$KK_i(A_1, B_1\otimes D)\otimes KK_j(D\otimes A_2, B_2)\to KK_{i+j}(A_1\otimes A_2, B_1\otimes B_2) \quad (3)$$

$(i, j\in \mathbf{Z}_2)$, which generalizes all the known cup and cap products in K_* and K^*-functors (cf. [15], §§ 4, 6). For this reason (3) may be called a *cup-cap product* (or a product intersection). To indicate that the intersection (cap) is in D we shall write a cup-cap product of elements x and y as $x\otimes_D y$. The idea of the construction of (3) is taken from elliptic operators, but the construction itself is rather difficult (even after simplifications given in [11] and [33]), and we omit it.

The cup-cap product is functorial in all its arguments and, what is most important, *associative*. This last means that $(x_1\otimes_{D_1}x_2)\otimes_{D_2}x_3 = x_1\otimes_{D_1}(x_2\otimes_{D_2}x_3)$ for $x_1\in KK_i(A_1, B_1\otimes D_1)$, $x_2\in KK_j(D_1\otimes A_2, B_2\otimes D_2)$, $x_3\in KK_r(D_2\otimes A_3, B_3)$. In particular, $KK_0(A, A)$ is an associative ring with unit $1_A = (E, F)$, where $E = A\oplus 0, F = 0$.

An immediate corollary of the existence of a cup-cap product is the general periodicity. We shall call element $\alpha\in KK_j(D_1, D_2)$ *D_2-invertible* (*D_1-invertible*, or simply *invertible*) if $\exists\beta\in KK_j(D_2, D_1)$ such that $\beta\otimes_{D_1}\alpha = 1_{D_2}(\alpha\otimes_{D_2}\beta = 1_{D_1}$, or both, respectively). A cup-cap product with invertible elements α and β gives isomorphisms: $KK_i(A, B\otimes D_1)\simeq KK_{i+j}(A, B\otimes D_2)$, $KK_i(A\otimes D_1, B)\simeq KK_{i+j}(A\otimes D_2, B)$. In this way one obtains the usual Bott periodicity, as well as stability: changing A to $A\otimes C(R^2)$ or $A\otimes\mathscr{K}$ we do not change the KK-groups (and the same for B).

There is also an equivariant version of the K-bifunctor ([19]). Let G be a separable locally compact group. Consider a category of C^*-algebras with G-action (and G-equivariant morphisms). A C^*-algebra B with a G-action $G\times B\to B$ continuous in the sense of the norm will be called a G-algebra. G-action on a Hilbert B-module E must satisfy the additional conditions $g(xb) = g(x)g(b)$, $g\big(\langle x, y\rangle\big) = \big\langle g(x), g(y)\big\rangle$ and must be continuous (in norm). Note that the induced G-action on $\mathscr{L}(E)$ is usually not continuous. We define the group $KK_0^G(A, B)$ for G-algebras A and B just as before, but adding to the list (1) operators $a\big(g(F)-F\big)$, $\forall g\in G$, and demanding that function $g\to g(F)$ should be continuous in norm. The group $KK_1^G(A, B)$ is defined in the same way.

Product (3) generalizes to the KK^G-bifunctor with all properties listed

above. In particular, $R(G) = KK_0^G(C, C)$ is a commutative ring with 1, and all groups $KK_i^G(A, B)$ are $R(G)$-modules. For a compact group G, $R(G)$ is the conventional representation ring of G, whence the notation.

3. Methods of computation

DEFINITION 2. Let X be a complete Riemannian manifold and G a separable locally compact group acting on X by isometries. A cotangent manifold T^*X has a natural almost complex structure. Associated with it, there is a Dolbeault operator on smooth forms with compact support: $D = \bar{\partial} + \bar{\partial}^*$: $\Omega_c^{0,*}(T^*X) \to \Omega_c^{0,*}(T^*X)$. Since D is essentially selfadjoint, on the L^2-completion H of $\Omega_c^{0,*}(T^*X)$ we can consider the operator $T = (1+D^2)^{-1/2} D \in \mathscr{L}(H)$. A decomposition of H into forms of even and odd dimension gives a grading: $H = H^{(0)} \oplus H^{(1)}$, with T taking the form: $T = \begin{bmatrix} 0 & T_1^* \\ T_1 & 0 \end{bmatrix}$. The pair (H, T_1) represents a canonical "Dolbeault element" $[\bar{\partial}_X] \in K_0^G(T^*X)$.

THEOREM 1. *Let X be simply connected and have a non-positive sectional curvature. Then the element $[\partial_X] \in K_0^G(T^*X)$ is $C(T^*X)$-invertible, i.e., there exists a $\delta_X \in K_G^0(T^*X)$ such that $[\bar{\partial}_X] \otimes_C \delta_X = 1_{C(T^*X)} \in KK_G^0(T^*X, T^*X)$.*

Note that the construction of δ_X (see [19], § 5) is based on the idea of using the radial geodesic covector field ([23], § 6). The proof of Theorem 1 also makes use of a "rotation" of $X \times X$ ([19], § 5, Lemma 3).

THEOREM 2 ([19], § 5). *Let group G be connected, let G_K be its maximal compact subgroup and let $X = G/G_K$. Then the element $[\bar{\partial}_X]$ is $C(T^*X)$-invertible. Denote its inverse by δ_X and the element $\delta_X \otimes_{C(T^*X)} [\bar{\partial}_X] \in R(G)$ by γ_G. The element γ_G is idempotent and depends only on G and not on the choice of G_K and δ_X. For any G-algebras A and B the restriction r: $KK_i^G(A, B) \to KK_i^{G_K}(A, B)$ maps the subgroup $\gamma_G \cdot KK_i^G(A, B)$ isomorphically onto $KK_i^{G_K}(A, B)$ with* $\operatorname{Ker} r = (1-\gamma_G) \cdot KK_i^G(A, B)$.

Note that $\gamma_G \neq 1$ for groups G with the Kazhdan property (T). On the other hand, $\gamma_G = 1$ for amenable groups and Lorentz groups $\mathrm{SO}_0(n, 1)$. More precisely, we have

THEOREM 3. *Let G be a connected group having an amenable normal subgroup N with G/N locally isomorphic to a product of a compact group and a finite number of Lorentz groups. Then $\gamma_G = 1$.*

The proof is based on the construction, for the group $G = \mathrm{SO}_0(2n+1, 1)$, of an element $a \in K_0^G(S^{2n})$ mapped into $1 \in R(G)$

under the homomorphism induced by mapping S^{2n} into a point. It is very likely that $\gamma_G = 1$ also for all groups $\mathrm{SU}(n, 1)$.

We now come to the computation of the K-functor for the group C^*-algebras $C^*(G)$ and crossed products $C^*(G, A)$ ([28]). Most useful here is the natural homomorphism ([19], § 6):

$$j_G: KK_i^G(A, B) \to KK_i\big(C^*(G, A), C^*(G, B)\big),$$

which maps 1_A into $1_{C^*(G,A)}$ and transfers the product $x \otimes_D y$ of elements $x \in KK_i^G(A, D)$ and $y \in KK_j^G(D, B)$ into $j_G(x) \otimes_{C^*(G,D)} j_G(y)$.

THEOREM 4 ([19], § 6). *Let G be a connected group, G_K its maximal compact subgroup and V a cotangent space to G/G_K at the point (G_K). Then there are $C^*\big(G_K, A \otimes C(V)\big)$-invertible elements in the groups $KK_0\big(C^*(G, A), C^*\big(G_K, A \otimes C(V)\big)\big)$ and $KK_0\big(C^*_{\mathrm{red}}(G, A), C^*\big(G_K, A \otimes C(V)\big)\big)$. If $\gamma_G = 1$ then both these elements are invertible. In particular, for K-groups, if the action of G_K on V is spinor, then, in general, both $K_i\big(C^*(G, A)\big)$ and $K_i\big(C^*_{\mathrm{red}}(G, A)\big)$ contain as a direct summand the group $K_{i+\dim V}^{G_K}(A)$, and if $\gamma_G = 1$, then all these three K-groups are isomorphic.*

4. Index of elliptic operators

Let X be a closed smooth manifold (with a fixed Riemannian metric) and $p: T^*X \to X$ a projection. A pseudo-differential operator (cf. [13]) of order 0 acting from sections of the vector bundle $\eta^{(0)}$ to those of $\eta^{(1)}$ can be considered as an operator $F: L^2(\eta^{(0)}) \to L^2(\eta^{(1)})$. (We fix Hermitian metrics on $\eta^{(0)}$ and $\eta^{(1)}$). The symbol σ_F of operator F is a homomorphism of vector bundles: $p^*(\eta^{(0)}) \to p^*(\eta^{(1)})$. We shall call F *elliptic* if $\|\sigma_F \sigma_F^* - 1\|_{x,\xi} \to 0$ and $\|\sigma_F \sigma_F^* - 1\|_{(x,\zeta)} \to 0$ uniformly in $x \in X$ as $\zeta \to \infty$ in T_x^*X. (Operators that are elliptic in the ordinary sense can clearly be normalized to satisfy this condition.)

We shall consider $\eta = \eta^{(0)} \oplus \eta^{(1)}$ as a graded bundle and $L^2(\eta)$ as a graded Hilbert space. The space of continuous sections of the vector bundle $p^*(\eta)$ over T^*X, tending to 0 at ∞, will be denoted by $C\big(p^*(\eta)\big)$. The algebra $C(X)$ acts on $L^2(\eta)$ and on $C\big(p^*(\eta)\big)$ from the left by multiplication. Thus we get elements $[F] = (L^2(\eta), F) \in K_0(X)$ and $[\sigma_F] = \big(C\big(p^*(\eta)\big), \sigma_F\big) \in KK^0(X, T^*X)$. Similarly, for a selfadjoint elliptic operator S we get elements $[S] \in K_1(X)$ and $[\sigma_S] \in KK^1(X, T^*X)$.

THEOREM 5. *If $[\bar{\partial}_X] \in K_0(T^*X)$ is the Dolbeault element then*

$$[F] = [\sigma_F] \otimes_{C(T^*X)} [\bar{\partial}_X], \qquad [S] = [\sigma_S] \otimes_{C(T^*X)} [\bar{\partial}_X]. \tag{4}$$

This theorem was established for the first time in [14] and [15]. But the theorem itself (cf. [4], [5]) and its proof (cf. [10], [11]) are probably natural enough to be discovered again and again. (However some references to the previous work in the new attempts would not seem to be superfluous.)

From Theorem 5 one easily obtains a "cohomological form" of the index theorem ([3], 2.12). For this, one must take the Chern character of both sides of (4) and notice that $\mathrm{ch}([\bar{\partial}_X])$ is Poincaré dual in T^*X to the Todd class of the complexified cotangent bundle of X.

A number of index theorems can be obtained following the above scheme (for example, the index theorem for foliations [11]). Another example (see [25], [32]) is the index of elliptic operators in locally trivial bundles over X with fibres being finitely generated projective (right) modules over a C^*-algebra A with 1. (Such bundles are called *A-bundles*.) Here $[F] \in KK_0\big(C(X), A\big)$, $[\sigma_F] \in KK_0\big(C(X), A \otimes C(T^*X)\big)$ and the index theorem for A-operators is again just (4).

We will now generalize this situation to the case of group action (cf. [9], [20]). Let X be a complete Riemannian manifold and G a separable locally compact group acting on X by isometries. We shall assume that all stability subgroups, as well as the orbit space X/G, are compact. A will be a separable G-algebra with 1. We will consider G-invariant elliptic A-operators on X belonging to the class $L^0_{\varrho,\delta}$ ([13], § 2.3) and assume that all operators and symbols have regular supports ([13], § 2.1). In this case there are elements $[F] \in KK^G_0\big(C(X), A\big)$, $[\sigma_F] \in KK^G_0\big(C(X), A \otimes C(T^*X)\big)$ (and similarly for selfadjoint S) and relations (4) are valid. But most interesting here is the index with values in $K_*\big(C^*(G, A)\big)$.

To define it we introduce on the space $C^\infty_c(\eta)$ of smooth, compactly supported sections of a $G-A$-bundle η a structure of right $C_c(G, A)$-module by the formula:

$$(e \cdot a)(x) = \int_G g(e)(x) \cdot g\big(a(g^{-1})\big) \cdot \mu(g)^{-1/2} dg \in C^\infty_c(\eta),$$

where $e \in C^\infty_c(\eta)$, $a \in C_c(G, A)$, μ is the modular function of G. Moreover, if $\langle\ ,\ \rangle$ is a G-invariant Hermitian A-metric on η, we can define a $G_c(G, A)$-scalar product of any two $e_1, e_2 \in C^\infty_c(\eta)$ as

$$(e_1, e_2)(g) = \mu(g)^{-1/2} \cdot \int_X \langle e_1(x), g(e_2)(x) \rangle\, dx \in C_c(G, A).$$

Denote by E_η the completion of $C^\infty_c(\eta)$ in the norm $\|e\| = \|(e, e)\|^{1/2}_{C^*(G,A)}$. By the method of ([13], § 2.2) one can check that a G-invariant elliptic

operator F of order 0 defines a "Fredholm" operator $F: E_{\eta(0)} \to E_{\eta(1)}$, i.e., an element $\mathrm{ind}_a(F) \in K_0\big(C^*(G, A)\big)$. Selfadjoint S defines $\mathrm{ind}_a(S) \in K_1\big(C^*(G, A)\big)$.

Recall now that there is a canonical element $[c] \in K_0\Big(C^*\big(G, C(X)\big)\Big)$ defined by a "cut off" function $c(x)$ on X (see [20]). Put $\mathrm{ind}_t(F) = [c] \otimes \otimes_{C^*(G, C(X))} j_G([F]) \in K_0\big(C^*(G, A)\big)$, and similarly for S.

THEOREM 6. $\mathrm{ind}_a(F) = \mathrm{ind}_t(F)$, $\mathrm{ind}_a(S) = \mathrm{ind}_t(S)$.

For $A = C$ this is just Theorem 3 of [20]. The case of an arbitrary A and $X = G = R^n$ was treated in [9] (Theorem 10).

Example 1. In the notation of Theorem 4 assume that $X = G/G_K$ is even-dimensional and the action of G_K on V is spinor. Denote by $S^{\pm}$ the half-spin $\mathrm{Spin}(V)$-modules. For any finitely generated projective $G_K - A$-module P define $G - A$-bundles $\xi_P^{\pm} = G \times_{G_K} (S^{\pm} \otimes P)$ over X and consider the (G-invariant) Dirac operator $D_P^{\pm}: C_c^{\infty}(\xi_P^{\pm}) \to C_c^{\infty}(\xi_P^{\mp})$. Multiplying $D_P^{\pm}$ on the left by a G-invariant operator R with symbol $\sigma(x, \zeta) = (1 + \|\zeta\|^2)^{-1/2}$, we get an operator $F_P^{\pm} = R \cdot D_P^{\pm}$ of order 0. Thus we obtain a homomorphism $\partial_A: K_0^{G_K}(A) \to K_0\big(C^*(G, A)\big)$: $\partial_A([P]) = \mathrm{ind}_a(F_P^{\pm})$, which just coincides with the inclusion described in Theorem 4. Note that when $A = C$ one can take $R = (1 + D_P^{\pm} D_P^{-})^{-1/2}$ (see [20], Theorem 2).

5. K-theory of representations

At present, this is a non-existent mathematical region. However, it may really come into existence in a near future. Consider a simple example.

Example 2. Studying irreducible unitary representations of the group $G = \mathrm{SL}_2(R)$ entering the regular representation, one can "compute" explicitly the algebra $C_{\mathrm{red}}^*(G)$. Each discrete series representation gives a direct summand $\mathscr{K}$ to $C_{\mathrm{red}}^*(G)$, the even principal series adds a direct summand $A_0 = C\big([0, \infty)\big) \otimes \mathscr{K}$, and to the odd principal series there corresponds a direct summand A_1 isomorphic to a subalgebra of A_0 consisting of those continuous functions on $[0, \infty)$ with values in $\mathscr{K}$ which are reduced at the point 0 by some decomposition $H = H' \oplus H''$. It is easily verified that $K_0(\mathscr{K}) = \boldsymbol{Z}$, $K_0(A_0) = 0$, $K_0(A_1) = \boldsymbol{Z}$, and therefore, $K_0\big(C_{\mathrm{red}}^*(G)\big) \simeq (\bigoplus_{n \neq 0} \boldsymbol{Z}) \oplus \boldsymbol{Z}$.

Let us compare this with the isomorphism $\partial_C: R(S^1) \xrightarrow{\sim} K_0\big(C^*(G)\big) \xrightarrow{\sim} K_0\big(C_{\mathrm{red}}^*(G)\big)$ given in Example 1. Denote by $D_n^{\pm}$ the Dirac operators on $X = G/S^1$ corresponding to one-dimensional S^1-modules $P_n(e^{i\varphi} \to e^{in\varphi})$.

Put $F_n^+ = (1+D_n^+D_n^-)^{-1/2}D_n^+$. Note that for $n \neq 0$ the spectrum of the operator $(D_n^+ \oplus D_n^-)^2$ on L^2-sections contains 0 as an isolated point. According to [20], it follows that $\operatorname{Ker} D_n^{\pm}$, as Hilbert $C^*_{\mathrm{red}}(G)$-modules, are isomorphic to the ranges of some projections $q_n^{\pm} \in \mathscr{K}(E_{\xi\pm})$ and that $\operatorname{ind}(F_n^+) = [q_n^+] - [q_n^-]$ in $K_0(C^*_{\mathrm{red}}(G))$. In fact, for $n > 0$, $\operatorname{Ker} D_n^+$ is a holomorphic discrete series representation and $\operatorname{Ker} D_n^- = 0$. For $n < 0$, $\operatorname{Ker} D_n^-$ is an anti-holomorphic discrete series representation and $\operatorname{Ker} D_n^+ = 0$. The remaining generator $[P_0] \in R(S^1)$ maps into a generator of $K_0(A_1) = \boldsymbol{Z}$.

Suppose that while computing $C^*_{\mathrm{red}}(G)$ we have "forgotten" that an odd principal series representation at $t = 0$ is reducible ($H = H' \oplus H''$). Then instead of a subalgebra $A_1 \subset C^*_{\mathrm{red}}(G)$ we get A_0 and the direct summand $\boldsymbol{Z} = K_0(A_1)$ in $K_0(C^*_{\mathrm{red}}(G))$ disappears. The element $\partial_{\boldsymbol{C}}([P_0])$ becomes 0, in contradiction to Theorem 4. This shows how one "missing" point of the unitary spectrum $\hat{G}$ can be detected by K-theory. This method may become useful, for instance, in the problem of unitarity for representations of semisimple Lie groups.

However, to make such applications possible one should compute $K_*(C^*(G))$ for a much larger class of groups than in Theorem 3. Closely connected with this is the problem of computing representation rings $R(G)$. Note that groups $K_*(C^*_{\mathrm{red}}(G))$ have already been computed for all complex semisimple groups and Lie groups of real rank 1 ([29], [35]). Unfortunately, this computation is based just on a complete knowledge of algebras $C^*_{\mathrm{red}}(G)$.

A reasonable conjecture on the structure of $K_*(C^*_{\mathrm{red}}(G))$ is Connes's conjecture ([31]): $K_i(C^*_{\mathrm{red}}(G)) \simeq K_{GK}^{i+\dim G/GK}(\mathrm{point})$. However, as can be seen from Example 2, in order to make this conjecture much more valuable for applications one should specify a concrete isomorphism (which was not done in the original form of the conjecture).

Conjecture 1. *The homomorphism* $\partial_A\colon K_i^{GK}(A \otimes C(V)) \to K_i(C^*_{\mathrm{red}}(G, A))$ *described in Theorem* 4 *and Example* 1 *is an isomorphism.*

6. Higher signatures

Let M^n be a smooth oriented closed manifold, $[M^n]$ its fundamental class in $H_n(M^n)$, and $L_*(M^n)$ the Pontriagin–Hirzebruch characteristic class of M^n. Fix a discrete group π and denote by $B\pi$ its classifying space. For any continuous map $f\colon M^n \to B\pi$ and any $x \in H^*(B\pi) \otimes \boldsymbol{Q}$ one can construct a real number $\langle L_*(M^n) \cdot f^*(x), [M^n] \rangle$. Novikov's higher signature conjecture (for group π) is the assertion that all these numbers for

any x and f depend only on a homotopy type of M^n (for any M^n). If we denote the Poincaré duality by D, the problem is to prove that $f_*\big(D\big(L_*(M^n)\big)\big) \in H_*(B\pi)\otimes \boldsymbol{Q}$ depends only on the homotopy type of M^n (for any M^n and f).

In a series of papers [14], [15], [22], [23] the present author and A. S. Miščenko independently suggested a method of reducing this problem to a problem of operator K-theory. Here is this reduction (following [15]). Without loss of generality one can assume that n is even. Let $(d+\delta)$ be the *signature operator* on M^n (see [3], § 6) and $[d+\delta] \in K_0(M^n)$ the corresponding element. From a cohomological form of the Atiyah–Singer theorem it easily follows that the problem is to prove the homotopy invariance of $f_*([d+\delta]) \in RK_0(B\pi)\otimes \boldsymbol{Q}$, where $RK_0(B\pi)$ means $\varinjlim_{X\subset B\pi} K_0(X)$, the inductive limit along the directed set of all compact subsets $X \subset B\pi$.

Now we will define a natural homomorphism

$$\beta\colon RK_0(B\pi)\to K_0\big(C^*(\pi)\big).$$

Consider a compact subset $X \subset B\pi$. Let $\tilde{X}$ be its regular covering (with group π) induced by inclusion $X \subset B\pi$. Then $\tilde{X}\times_\pi C^*(\pi)$ is a $C^*(\pi)$-bundle over X. It defines an element $[\beta_X] \in K_0\big(C(X)\otimes C^*(\pi)\big)$. The cap-product $[\beta_X]\otimes_{C(X)}\colon K^0\big(C(X)\big)\to K_0\big(C^*(\pi)\big)$ after passing to $\varinjlim$ gives the homomorphism β.

It appears that the element $\beta\big(f_*([d+\delta])\big) \in K_0\big(C^*(\pi)\big)\otimes \boldsymbol{Q}$ always depends only on the homotopy type of M^n (see [24], [26] and [15], [19] — — even without tensoring by $\boldsymbol{Q}$). The proof exploits the close relationship between this element and Miščenko's homotopy invariant element $\sigma(M^n)$ in Wall's group $L_n(\pi)$ ([21]). Now we can formulate a conjecture [15] (called in [32] the *strong Novikov conjecture*), from which the conventional one follows:

CONJECTURE 2 (SNC). *$\beta\otimes\boldsymbol{Q}\colon RK_0(B\pi)\otimes\boldsymbol{Q}\to K_0\big(C^*(\pi)\big)\otimes\boldsymbol{Q}$ is a monomorphism.*

Note that SNC was proved for the case of $B\pi$ admitting the structure of a closed manifold with a non-positive sectional curvature in [23]. This includes the case of *uniform* subgroups without torsion of semisimple Lie groups. In his publications later on (cf. [24]) Miščenko usually omitted the condition of compactness for $B\pi$, thus giving the impression that non-compact $B\pi$ admitting a non-positive curvature is also treated in [23]. It should be pointed out that no proof for non-compact $B\pi$ was provided by Miščenko either in [23] or in his other publications.

All known proofs of SNC in the assumptions of a non-positive curvature of $B\pi$ or inclusion of π into a Lie group are based on some analogues of Theorems 1 and 2, the main difficulty lying in these theorems (see [19], § 9 for π a closed subgroup of a connected Lie group). A direct proof for $\pi \subset \mathrm{GL}(n, C)$ announced in [15] was technically difficult just for the lack of a cup-cap product technique.

THEOREM 7. SNC *is valid for π if $B\pi$ is a complete Riemannian manifold of a non-positive sectional curvature. In fact, in this case even β is monomorphic.*

THEOREM 8. SNC *is valid for closed discrete subgroups of connected Lie groups. For subgroups π without torsion β is monomorphic.*

Here are short proofs. In the case of Theorem 8 let us assume first that $\pi(\subset G)$ has no torsion. Then we can take $B\pi = \pi \backslash G/G_K$. In any case, denoting a universal covering of $B\pi$ by X, we have a $C(T^*X)$-invertible element $[\bar{\partial}_X]$ in $K_0^\pi(T^*X)$ (Theorems 1 and 2). Hence, $j_\pi([\bar{\partial}_X])$ is $C^*(\pi, C(T^*X))$-invertible. But $C^*(\pi, C(T^*X)) \simeq C(T^*B\pi) \otimes \mathscr{K}$ (see [12]), and so we have a $C(T^*B\pi)$-invertible element $j_\pi([\bar{\partial}_X]) \in KK_0(C(T^*B\pi), C^*(\pi))$. In particular, the cap-product $\otimes_{C(T^*B\pi)} j_\pi([\bar{\partial}_X])$: $K^0(T^*B\pi) \to K_0(C^*(\pi))$ is monomorphic. But this homomorphism can easily be identified with β via the Poincaré duality: $K^0(T^*B\pi) \simeq RK_0(B\pi)$ (see [19], § 8), which gives the required result. For the general case of π with torsion (in Theorem 8) the assertion easily reduces to finitely generated π. Then applying Theorem 6.13 of [30], we see that (possibly after passing to a factor group of G by some finite central normal subgroup) π contains a subgroup of finite index without torsion. In view of Propositions 2.7 and 2.8 of [32] we come to the case where π is without torsion, which ends the proof.

Note that at present Novikov's conjecture seems unlikely to be true in general. It is more likely that there must be some restrictions on π of the type of nuclearity (injectivity, amenability) of some "enveloping" algebras for π (for example, $C^*(G)$ in the case of Theorem 8). From this point of view it could be useful to analyse the class of groups π considered in [7] and, in particular, to prove SNC for this class.

7. C^*-extensions

Earlier, discussing special cases of the K-bifunctor, we already associated some element $e_D \in KK_1(A, B)$ with an extension

$$0 \to B \to D \xrightarrow{q} A \to 0. \tag{5}$$

It was done under the assumption that q admits a completely positive lifting. We shall call extensions with this property *admissible*. (Recall that in the case of nuclear A all extensions are admissible [8].) If q has a cross-section homomorphism, the extension is called *split*. For such an extension $e_D = 0$.

An interesting class of extensions is associated with group C^*-algebras. If Γ is a (non-trivial) normal subgroup in G, we have an extension

$$0 \to B \to C^*(G) \xrightarrow{q} C^*(G/\Gamma) \to 0, \tag{6}$$

where $B = \mathrm{Ker} q$. Since for split extensions the homomorphism q_* of K_*-groups is always epimorphic, we obtain from Theorem 4 in particular

THEOREM 9. *If G is connected and Γ simply connected and solvable, then extension* (6) *does not split.*

In fact, for a large class of simply connected nilpotent Lie groups G and one-dimensional Γ the element $e_{C^*(G)}$ corresponding to (6) can be computed explicitly by using only some simple geometric observations ([19], § 6).

We have not yet described the exact sequences for the K-bifunctor. These 6-periodic exact sequences are

$$\ldots \to KK_i(A, B) \to KK_i(I, B) \xrightarrow{\partial} KK_{i+1}(A/I, B) \to KK_{i+1}(A, B) \to \ldots, \tag{7}$$

$$\ldots \to KK_i(A, B) \to KK_i(A, B/J) \xrightarrow{\delta} KK_{i+1}(A, J) \to KK_{i+1}(A, B) \to \ldots, \tag{8}$$

where I and J are ideals in A and B respectively. Originally established in [18] on the assumption of nuclearity for A, they are now obtained on better assumptions ([34]). Namely, only the admissibility of extensions $0 \to I \to A \to A/I \to 0$ and $0 \to J \to B \to B/J \to 0$ (respectively) is required. (For (7) this is probably the best possible assumption.) The homomorphisms ∂ and δ can be described here just as in [18]: $\partial(x) = e_A \otimes_I x$, $\delta(x) = x \otimes_{B/J} e_B$ where $e_A \in KK_1(A/I, I)$, $e_B \in KK_1(B/J, J)$.

And finally there is, of course, the question of recovering the original extension (5) from the corresponding element $e_D \in KK_1(A, B)$. Assume that A is nuclear. Then $\forall a \in KK_1(A, B)$ there exists an *absorbing* extension of the type $0 \to \mathscr{K} \otimes B \to D_a \to A \to 0$, which is unique up to unitary equivalence ([18], § 7). If $e_D = a$ then the algebra D can be included in D_a in such a way that B becomes a *full corner* in $\mathscr{K} \otimes B$ and $D + (\mathscr{K} \otimes B) = D_a$, $D \cap (\mathscr{K} \otimes B) = B$ (see [18], § 7, Corollary 2). It is an open question how to classify such subalgebras $D \subset D_a$.

References

[1] Atiyah M. F., K-theory, Benjamin, New York and Amsterdam, 1967.

[2] Atiyah M. F., Global Theory of Elliptic Operators, In: *Proc. Intern. Conf. on Functional Analysis and Related Topics*, Univ. of Tokyo Press, Tokyo 1970, pp. 21–30.

[3] Atiyah M. F. and Singer I. M., The Index of Elliptic Operators, III, *Ann. of Math.* **87** (1968), pp. 546–604.

[4] Baum P. and Douglas R. G., Index Theory, Bordism and K-Homology. In: *Operator Algebras and K-Theory*, Contemporary Mathematics, vol. 10, Amer. Math. Soc., 1982, pp. 1–31.

[5] Baum P. and Douglas R. G., K-Homology and Index Theory. In: *Operator Algebras and Applications*, Proc. Symp. Pure Math., vol. 38, part 1, Amer. Math. Soc., 1982, pp. 117–173.

[6] Brown L. G., Douglas R. G. and Fillmore P. A., Extensions of C^*-Algebras and K-Homology, *Ann. of Math.* **105** (1977), pp. 265–324.

[7] Cappell S. E., On Homotopy Invariance of Higher Signatures, *Invent. Math.* **33** (1976), pp. 171–179.

[8] Choi M.-D. and Effros E. G., The Completely Positive Lifting Problem for C^*-Algebras, *Ann. of Math.* **104** (1976), pp. 585–609.

[9] Connes A., C^*-algèbres et géométrie différentielle, *C. R. Acad. Sci. Paris Sér. A.* **290** (1980), pp. 599–604.

[10] Connes A. and Skandalis G., Théorème de l'indice pour les feuilletages, *C. R. Acad. Sci. Paris Sér. A.* **292** (1981), pp. 871–876.

[11] Connes A. and Skandalis G., The Longitudinal Index Theorem for Foliations, preprint, I.H.E.S., 1982.

[12] Green P., C^*-Algebras of Transformation Groups with Smooth Orbit Space, *Pacific J. Math.* **72** (1977), pp. 71–97.

[13] Hörmander L., Fourier Integral Operators, I, *Acta Math.* **127** (1971), pp. 79–183.

[14] Каспаров Г. Г., Обобщенный индекс эллиптических операторов, *Функц. анализ* **7**, № 3 (1973), pp. 82–83.

[15] Каспаров Г. Г., Топологические инварианты эллиптических операторов, I, K-Гомологии, *Изв. АН СССР Серия матем.* **39** (1975), pp. 796–838.

[16] Каспаров Г. Г., K-функтор в теории расширений C^*-алгебр, *Функц. анализ* **13**, № 4 (1979), pp. 73–74.

[17] Kasparov G. G., Hilbert C^*-Modules: Theorems of Stinespring and Voiculescu, *J. Operator Theory* **4** (1980), pp. 133–150.

[18] Каспаров Г. Г., Операторный K-функтор и расширения C^*-алгебр, *Изв. АН СССР Серия матем.* **44** (1980), 571–636.

[19] Kasparov G. G., K-Theory, Group C^*-Algebras, and Higher Signatures (conspectus), parts 1, 2, preprint, Chernogolovka, 1981.

[20] Каспаров Г. Г., Индекс инвариантных эллиптических операторов, K-теория и представления групп Ли, *Докл. АН СССР* **268** (1983), pp. 533–537.

[21] Мищенко А. С., Гомотопические инварианты неодносвязных многообразий. 1. Рациональные инварианты, *Изв. АН СССР Серия матем.* **34** (1970), pp. 501–514.

[22] Мищенко А. С., Бесконечномерные представления дискретных групп и го-

мотопические инварианты неодносвязных многообразий, *Успехи матем. наук* **28**, № 4 (1973), pp. 239–240.

[23] Мищенко А. С., Бесконечномерные представления дискретных групп и высшие сигнатуры, *Изв. АН СССР Серия матем.* **38** (1974), pp. 81–106.

[24] Mishchenko A. S., *C**-Algebras and K-Theory, *Lecture Notes in Math.* **763**, 1979, pp. 262–274.

[25] Мищенко А. С., Фоменко А. Т., Индекс эллиптических операторов над *C**-алгебрами, *Изв. АН СССР Серия матем.* **43** (1979), pp. 831–859.

[26] Мищенко А. С., Соловьев Ю. П., Представления банаховых алгебр и формулы типа Хирцебруха, *Мат. сборник* **111** (1980), pp. 209–226.

[27] Paschke W. L., Inner Product Modules over *B**-Algebras, *Trans. Amer. Math. Soc.* **182** (1973), pp. 443–468.

[28] Pedersen G. K., *C*-Algebras and Their Automorphism Groups*, Academic Press, New York and London 1979.

[29] Plymen R. J., *Proof of Connes' Conjecture for Complex Semisimple Groups*, preprint, Univ. of Manchester, 1982.

[30] Raghunathan M. S., *Discrete subgroups of Lie Groups*, Springer-Verlag, Berlin–Heidelberg–New York 1972.

[31] Rosenberg J., *Group C*-Algebras and Topological Invariants*, preprint, Univ. of Pennsylvania, 1981.

[32] Rosenberg J., *C*-Algebras, Positive Scalar Curvature, and the Novikov Conjecture*, preprint, Univ. of Maryland, 1982.

[33] Skandalis G., *Some Remarks on Kasparov Groups*, preprint, Queen's Univ., 1982.

[34] Skandalis G., *Exact Sequences for the Kasparov Groups of Graded Algebras*, preprint, Queen's Univ., 1983.

[35] Valette A., *K-Theory for the Reduced C*-Algebra of a Semi-Simple Lie Group with Real Rank* 1 *and Finite Centre*, preprint, Univ. Libre de Bruxelles, 1982.

[36] Wolf J. A., *Essential Self-Adjointness for the Dirac Operator and Its Square, Indiana Univ. Math. J.* **22** (1973), pp. 611–640.

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

YVES MEYER

Intégrales singulières, opérateurs multilinéaires, analyse complexe et équations aux derivées partielles

1. La continuité L^2 des opérateurs définis par des intégrales singulières

Soit, avec les notations usuelles, $T: \mathscr{D}(\boldsymbol{R}^n) \to \mathscr{D}'(\boldsymbol{R}^n)$ un opérateur linéaire continu. Appelons $K(x, y) \in \mathscr{D}'(\boldsymbol{R}^n \times \boldsymbol{R}^n)$ le noyau-distribution de T, T^* le transposé de T dont le noyau-distribution est $K(y, x)$, $\langle \cdot, \cdot \rangle$ la forme bilinéaire de dualité entre $\mathscr{D}(\boldsymbol{R}^n)$ et $\mathscr{D}'(\boldsymbol{R}^n)$ et $\Omega \subset \boldsymbol{R}^n \times \boldsymbol{R}^n$ l'ouvert défini par $y \neq x$, $x \in \boldsymbol{R}^n$, $y \in \boldsymbol{R}^n$. Nous dirons que le noyau $K(x, y)$ vérifie les *estimations de Calderón–Zygmund* s'il existe un exposant $\varepsilon > 0$, $\varepsilon \leqslant 1$, et une constante $C \geqslant 0$ tels que la restriction de $K(x, y)$ à l'ouvert Ω soit une fonction continue ayant les trois propriétés suivantes:

$$|K(x, y)| \leqslant C|x-y|^{-n} \quad \text{pour tout } (x, y) \in \Omega, \tag{1}$$

$$|K(x', y)-K(x, y| \leqslant C|x-x'|^{\varepsilon}|x-y|^{-n-\varepsilon} \quad \text{si } (x, y) \in \Omega \text{ et } |x-x'| \leqslant \tfrac{1}{2}|x-y| \tag{2}$$

et finalement

$$|K(x, y)-K(x, y')| \leqslant C|y-y'|^{\varepsilon}|x-y|^{-n-\varepsilon} \quad \text{si } (x, y) \in \Omega \text{ et } |y-y'| \leqslant \tfrac{1}{2}|x-y|. \tag{3}$$

Nous désignerons par $\mathscr{E}$ l'ensemble des opérateurs $T: \mathscr{D}(\boldsymbol{R}^n) \to \mathscr{D}'(\boldsymbol{R}^n)$ dont le noyau vérifie (1), (2) *et* (3).

Un opérateur linéaire continu $T: \mathscr{D}(\boldsymbol{R}^n) \to \mathscr{D}'(\boldsymbol{R}^n)$ est appelé dans cet exposé (ceci n'est pas la terminologie usuelle) un *opérateur de Calderón–Zygmund* s'il existe une constante $C \geqslant 0$ telle que pour toute fonction $f \in \mathscr{D}(\boldsymbol{R}^n)$, $T(f)$ appartienne à $L^2(\boldsymbol{R}^n; dx)$ et vérifie $\|Tf\|_2 \leqslant C\|f\|_2$, et si, en outre, le noyau-distribution $K(x, y)$ de T vérifie les estimations de Calderón–Zygmund. Les célèbres méthodes de variable réelle de Calderón et

Zygmund s'appliquent alors et l'on obtient $\|Tf\|_p \leqslant C_p\|f\|_p$ si $1 < p < +\infty$ tandis que T envoie continûment $L^\infty(\boldsymbol{R}^n)$ dans l'espace $\mathrm{BMO}(\boldsymbol{R}^n)$ de John et Nirenberg ([20]). Enfin pour $1 < p < +\infty$ et tout poids $\omega \in A_p$ de la classe de Muckenhoupt, tout opérateur de Calderón–Zygmund se prolonge, par continuité, en un opérateur linéaire continu de $L^p(\boldsymbol{R}^n; \omega dx)$ dans lui-même ([4]).

Nous désignerons par $E \subset \mathscr{E}$ l'espace vectoriel des opérateurs de Calderón–Zygmund. *Le problème fondamental de cette théorie est de trouver un critère commode permettant de déterminer si un opérateur $T \in \mathscr{E}$ appartient, en fait, à E.* Le critère que nous allons donner porte sur l'objet $T(1)$ que nous allons maintenant définir si $T \in \mathscr{E}$; 1 désigne la fonction identiquement égale à 1.

Soit $\mathscr{D}_0(\boldsymbol{R}^n) \subset \mathscr{D}(\boldsymbol{R}^n)$ le sous-espace des fonctions φ telles que $\int_{\boldsymbol{R}^n} \varphi(x)dx = 0$. Si $\varphi \in \mathscr{D}_0$ et $T \in \mathscr{E}$, la distribution $T^*(\varphi)$ est, en fait, continue hors du support de φ et est $O(|x|^{-n-\varepsilon})$ à l'infini. Alors $\langle T^*(\varphi), 1\rangle$ a un sens et $T(1) = S$ est une forme linéaire continue sur $\mathscr{D}_0$ définie par $\langle S, \varphi\rangle = \langle T^*(\varphi), 1\rangle$. De même $T^*(1)$ est une forme linéaire continue sur $\mathscr{D}_0(\boldsymbol{R}^n)$.

Nous désignerons par $H^1(\boldsymbol{R}^n)$ l'espace de Stein et Weiss dont le dual est $\mathrm{BMO}(\boldsymbol{R}^n)$. Alors $\mathscr{D}_0$ est dense dans H^1 et nous écrirons $T(1) \in \mathrm{BMO}$ pour exprimer que la forme linéaire continue $T(1)$, définie sur $\mathscr{D}_0$, se prolonge à $H^1(\boldsymbol{R}^n)$. De même $T^*(1) \in \mathrm{BMO}$ a un sens.

Une dernière définition nous sera utile. Soient $\varphi \in \mathscr{D}(\boldsymbol{R}^n)$, $u \in \boldsymbol{R}^n$, $\delta > 0$. Alors on pose $\varphi^{(u,\delta)}(x) = \varphi\big((x-u)/\delta\big)$ et l'on dit que $T: \mathscr{D}(\boldsymbol{R}^n) \to \mathscr{D}'(\boldsymbol{R}^n)$ est un *opérateur d'ordre* 0 si, pour toute partie bornée $\mathscr{B} \subset \mathscr{D}(\boldsymbol{R}^n)$, il existe une constante $C = C(\mathscr{B})$ telle que $|\langle T\varphi_1^{(u,\delta)}, \varphi_2^{(u,\delta)}\rangle| \leqslant C\delta^n$ pour tout $\varphi_1 \in \mathscr{B}$, tout $\varphi_2 \in \mathscr{B}$, tout $u \in \boldsymbol{R}^n$ et tout $\delta > 0$.

THÉORÈME 1. *Soit $T: \mathscr{D}(\boldsymbol{R}^n) \to \mathscr{D}'(\boldsymbol{R}^n)$ un opérateur linéaire et continu dont le noyau-distribution $K(x, y)$ vérifie les estimations de Calderón–Zygmund. Alors les deux conditions suivantes sont équivalentes:*

$$T \text{ se prolonge en un opérateur continu sur } L^2(\boldsymbol{R}^n), \tag{4}$$

$$T(1) \in \mathrm{BMO},\ T^*(1) \in \mathrm{BMO} \text{ et } T \text{ est d'ordre } 0. \tag{5}$$

Soit P_t, $t \geqslant 0$, le semi-groupe de Poisson. Pour toute fonction $\beta \in \mathrm{BMO}(\boldsymbol{R}^n)$, désignons par L_β l'opérateur défini (formellement) par $4\int_0^\infty Q_t\{(Q_t\beta)(P_tf)\}\frac{dt}{t} = L_\beta(f)$ où $Q_t = -t\frac{\partial}{\partial t}P_t$. Alors L_β est un opérateur de Calderón–Zygmund tel que $L_\beta(1) = \beta$ et $L_\beta^*(1) = 0$. Si β et γ appartiennent à $\mathrm{BMO}(\boldsymbol{R}^n)$, $L = L_\beta + L_\gamma^*$ est aussi un opérateur de Cal-

derón–Zygmund tel que $L(1) = \beta$, $L^*(1) = \gamma$. Cela montre que les deux fonctions de BMO intervenant dans (5) sont arbitraires.

L'ensemble $A \subset \mathscr{L}\big(L^2(\boldsymbol{R}^n), L^2(\boldsymbol{R}^n)\big)$ des opérateurs $T \in \mathscr{E}$ tels que $T(1) = T^*(1) = 0$ est, en fait, une algèbre ([25]) et le théorème 1 peut être précisé en

$$E \simeq A \oplus \mathrm{BMO}(\boldsymbol{R}^n) \oplus \mathrm{BMO}(\boldsymbol{R}^n). \tag{6}$$

L'isomorphisme est $T \to \big(R, T(1), T^*(1)\big)$ où $R = T - L_\beta - L_\gamma^*$ si $\beta = T(1)$, $\gamma = T^*(1)$.

Pour démontrer le théorème 1, on se ramène, grâce à l'isomorphisme précédent, au cas où le noyau-distribution $R(x, y)$ de R vérifie les estimations de Calderón–Zygmund et où $R(1) = R^*(1) = 0$. Alors la continuité de R sur L^2 se démontre grâce au lemme de Cotlar ([15], [19], [26]).

Si $K: \Omega \to \boldsymbol{C}$ vérifie les estimations de Calderón–Zygmund et si $K(y, x) = -K(x, y)$, on pose $K_\varepsilon(x, y) = K(x, y)$ si $|x-y| \geqslant \varepsilon$, $K_\varepsilon(x, y) = 0$ sinon. Alors $\lim_{\varepsilon \downarrow 0} K_\varepsilon(x, y)$ existe dans $\mathscr{D}'(\boldsymbol{R}^n \times \boldsymbol{R}^n)$ et définit une distribution notée v.p.$K(x, y)$ et un opérateur $T: \mathscr{D}(\boldsymbol{R}^n) \to \mathscr{D}'(\boldsymbol{R}^n)$ d'ordre 0. La continuité de T sur L^2 est alors équivalente à $T(1) \in \mathrm{BMO}$.

Ces remarques fournissent une démonstration particulièrement simple des résultats de [3].

On appelle T_k, $k \in \boldsymbol{N}$, les opérateurs dont les noyaux-distribution sont $\big(A(x) - A(y)\big)^k (x-y)^{-k-1}$ lorsque $A: \boldsymbol{R} \to \boldsymbol{C}$ est lipschitzienne. Pour obtenir la continuité de T_k sur $L^2(\boldsymbol{R})$ il suffit de vérifier que $T_k(1) \in \mathrm{BMO}(\boldsymbol{R})$. Or $T_k(1) = T_{k-1}(A')$. Puisque $A' \in L^\infty(\boldsymbol{R})$, un raisonnement par récurrence et le théorème 1 donnent immédiatement $\|T_k\| \leqslant C^{k+1} \|A'\|_\infty^k$. On sait aujourd'hui que $\|T_k\| \leqslant 6400(1+k)^4 \|A'\|_\infty^k$ ([11]).

2. Applications à l'analyse complexe

Soit $\varphi: \boldsymbol{R} \to \boldsymbol{R}$ une fonction lipschitzienne: $|\varphi(x) - \varphi(y)| \leqslant M|x-y|$ pour une certaine constante $M \geqslant 0$, tout $x \in \boldsymbol{R}$ et tout $y \in \boldsymbol{R}$. Considérons le noyau-distribution v.p. $\big(x - y + i\big(\varphi(x) - \varphi(y)\big)\big)^{-1}$ et l'opérateur $T_\varphi: \mathscr{D}(\boldsymbol{R}) \to \mathscr{D}'(\boldsymbol{R})$ associé.

Théorème 2. *L'opérateur T_φ est borné sur $L^2(\boldsymbol{R})$ et sa norme ne dépasse pas $C(1+M)^8$.*

Il existe, à l'heure actuelle, deux démonstrations du théorème 2. La première (A. P. Calderón, G. David) consiste à démontrer d'abord la con-

tinuité de T_φ sur $L^2(\boldsymbol{R})$ en appliquant soit le théorème 1, soit le théorème de Calderón présenté au congrès d'Helsinki ([3], [14]).

Une utilisation ingénieuse des inégalités aux bons λ de Burkholder et Gundy permet alors de passer au cas général. La seconde ([11]) consiste à étudier les opérateurs T_k définis ci-dessus par de nouveaux algorithmes dus à A. McIntosh. La signification géométrique du théorème 2 est la suivante. Soit $\Gamma \subset \boldsymbol{R}^2$ le graphe de la fonction φ, Ω_1 l'ouvert situé au-dessus de Γ, Ω_2 celui au-dessous de Γ. On appelle $H^2(\Omega_1) \subset L^2(\Gamma) = L^2(\Gamma; ds)$ l'espace de Hardy défini comme la fermeture dans $L^2(\Gamma)$ des fractions rationnelles $P(z)/Q(z)$ nulles à l'infini et dont les pôles appartiennent à Ω_2. On définit de même $H^2(\Omega_2)$ et le théorème 2 signifie que $L^2(\Gamma)$ est la somme directe des sous-espaces $H^2(\Omega_1)$ et $H^2(\Omega_2)$. Observons que T_φ est un opérateur de Calderón–Zygmund.

Nous pouvons généraliser l'étude précédente.

Soit Γ une courbe de Jordan fermée et rectifiable du plan complexe, limitant le domaine Ω_1. Nous désignerons par Ω_2 l'extérieur de Γ et par $s \in [0, l]$ la longueur d'arc sur Γ (l étant la longueur totale de Γ).

Suivant Keldysh, Lavrentiev et Smirnov, on définit, pour $1 \leqslant p < +\infty$, deux sous-espaces fermés $H^p(\Omega_1)$ et $\mathscr{H}^p(\Omega_1)$ de $L^p(\Gamma; ds)$. L'espace de Hardy $H^p(\Omega_1)$ est la fermeture dans $L^p(\Gamma; ds)$ de l'ensemble des polynômes $P(z)$. D'après le théorème de Runge, on peut également définir $H^p(\Omega_1)$ comme la fermeture des fractions rationnelles $P(z)/Q(z)$ dont les pôles appartiennent à Ω_2. De même on définira ensuite $H^p(\Omega_2)$ comme la fermeture dans $L^p(\Gamma; ds)$ des fractions rationnelles $P(z)/Q(z)$ nulles à l'infini et dont les pôles appartiennent à Ω_1.

Le second espace $\mathscr{H}^p(\Omega_1)$ est défini, de façon indirecte, comme l'ensemble des $f \in L^p(\Gamma; ds)$ tels que $\int_\Gamma z^k f(z)\,dz = 0$ pour tout $k \in \boldsymbol{N}$. Lavrentiev a démontré que ces deux espaces sont, en général, distincts. Leur égalité ne dépend pas de p et, si elle a lieu, on dit que Ω_1 est un *domaine de Smirnov*.

Pour définir $\mathscr{H}^p(\Omega_2)$, on remplace z^k par $(z-a)^{-k}$, $a \in \Omega_1$, $k \geqslant 1$. On dit que Γ est une *courbe de Lavrentiev* si en désignant par $z(s)$, $s \in \boldsymbol{R}/l\boldsymbol{Z}$, le paramétrage de Γ par la longueur d'arc, on a $|s'-s| \leqslant C|z(s')-z(s)|$ pour tout s et tout s'. On dit que Γ est lipschitzienne si Γ est localement le graphe d'une fonction lipschitzienne.

Enfin Γ est une *courbe régulière d'Ahlfors* si, en désignant par $|E|$ la mesure de Lebesgue d'un borélien $E \subset \boldsymbol{R}/l\boldsymbol{Z}$, il existe une constante $C \geqslant 2$ telle que, pour tout nombre complexe $z_0 \in \boldsymbol{C}$ et tout $r > 0$, on a $|\{s \in \boldsymbol{R}/l\boldsymbol{Z};\ |z(s)-z_0| \leqslant r\}| \leqslant Cr$.

Avec ces notations on a ([14])

THÉORÈME 3. *L'espace $L^2(\Gamma; ds)$ est la somme directe de $H^2(\Omega_1)$ et de $H^2(\Omega_2)$ si et seulement si Γ est une courbe régulière d'Ahlfors. Alors Ω_1 et Ω_2 sont des domaines de Smirnov et pour $1 < p < +\infty$, $L^p(\Gamma; ds)$ est la somme directe de $H^p(\Omega_1)$ et de $H^p(\Omega_2)$.*

Donnons quelques indications sur la preuve du théorème 3. On appelle T l'opérateur défini (formellement) par le noyau de Cauchy $(z(s)-z(t))^{-1}$, $s \in \boldsymbol{R}/l\boldsymbol{Z}$, $t \in \boldsymbol{R}/l\boldsymbol{Z}$, et tout se ramène à l'étude de la continuité de T sur $L^2(\boldsymbol{R}/l\boldsymbol{Z})$. On utilise, dans cette étude, trois ingrédients:

— le fait que T soit borné sur L^2 lorsque Γ est lipschitzienne (théorème 2);

— une décomposition de tout intervalle d'une courbe régulière d'Ahlfors en deux parties: la première, après rotation des axes, est contenue dans le graphe d'une fonction lipschitzienne et la seconde a une petite mesure relative;

— l'utilisation des inégalités aux bons λ de Burkholder et Gundy pour passer du cas local au cas global.

Le succès de cette application à l'analyse complexe vient de la décision de traiter le problème à l'aide des méthodes de l'analyse réelle. On peut d'ailleurs remplacer le noyau de Cauchy $1/(z-w)$ par n'importe quel noyau $K(z-w)$ où $K: \boldsymbol{R}^2 \setminus \{0\} \to \boldsymbol{C}$ est impaire, homogène de degré -1 et indéfiniment dérivable.

Nous allons poursuivre l'étude de l'opérateur défini par le noyau de Cauchy dans le cas des courbes de Lavrentiev ouvertes; elles sont paramétrées par la longueur d'arc $s \in \boldsymbol{R}$ et l'on a $|s-t| \leqslant C|z(s)-z(t)|$ pour tout $s \in \boldsymbol{R}$ et tout $t \in \boldsymbol{R}$.

Ces courbes de Lavrentiev sont donc caractérisées par le fait que l'opérateur T_Γ défini par le noyau de Cauchy soit un opérateur de Calderón–Zygmund.

Nous allons définir la variété $\mathscr{V}$ des courbes de Lavrentiev, la paramétrer à l'aide d'un ouvert V de $\mathrm{BMO}(\boldsymbol{R})$ et enfin démontrer que l'application qui à Γ associe T_Γ (ou la représentation conforme) est réelle-analytique sur V.

La variété $\mathscr{V}$ des courbes de Lavrentiev orientées sera d'abord décrite comme une ensemble. On part des couples (Γ, z_0) d'une courbe de Lavrentiev orientée Γ et d'un point $z_0 \in \Gamma$. Ensuite on considère que deux tels couples sont équivalents si l'on peut trouver un déplacement plan g ($g(z) = e^{i\theta}z + \gamma$) tel que $g(\Gamma) = \Gamma'$ et $g(z_0) = g(z_0')$.

Désignons par $\mathrm{BMO}(\boldsymbol{R})$ l'espace de Banach des fonctions réelles $b: \boldsymbol{R} \to \boldsymbol{R}$ appartenant à l'espace de John et Nirenberg. Il existe alors une

partie ouverte $V \subset \mathrm{BMO}(\boldsymbol{R})$ telle que, pour tout couple (Γ, z_0), il existe $b \in V$ de sorte que $z(s) = z_0 + \int_0^s \exp ib(t)\,dt$. De plus b est la détermination "naturelle" de $\arg z'(s)$. En fait, V est une carte globale de la variété $\mathscr{V}$ des courbes de Lavrentiev, munie de la relation d'équivalence ci-dessus.

Pour $b \in V$, on forme l'opérateur T_b dont le noyau est $K(t,s) = 1/\pi i$ v.p. $\big(z(s)-z(t)\big)^{-1} dz(s)$; b et z sont reliés comme il vient d'être dit. Alors on a ([8])

THÉORÈME 4. *L'application* $T\colon V \to \mathscr{L}\big(L^2(\boldsymbol{R}), L^2(\boldsymbol{R})\big)$ *qui à* $b \in V$ *associe l'opérateur de Cauchy* T_b *est une fonction réelle-analytique sur* V.

Cela signifie que pour tout $b_0 \in V$, il existe $\varepsilon > 0$ et $C > 0$ de sorte que si $\|b - b_0\|_{\mathrm{BMO}} < \varepsilon$, on ait

$$T_b = \sum_0^\infty T_{b_0}^{(k)}(b-b_0, \ldots, b-b_0)$$

où les $T_{b_0}^{(k)}\colon \mathrm{BMO} \times \ldots \times \mathrm{BMO} \to \mathscr{L}\big(L^2(\boldsymbol{R}), L^2(\boldsymbol{R})\big)$ sont des opérateurs multilinéaires vérifiant $\|T_{b_0}^{(k)}(f_1, \ldots, f_k)\|_{\mathscr{L}(L^2,L^2)} \leqslant C^{k+1}\|f_1\|_{\mathrm{BMO}} \cdots \|f_k\|_{\mathrm{BMO}}$.

Pour démontrer le théorème 4, on construit explicitement le prolongement analytique en remplaçant $b_0 \in V$ par $b_0 + \beta$ où $\beta\colon \boldsymbol{R} \to \boldsymbol{C}$ vérifie $\|\beta\|_{\mathrm{BMO}} < \varepsilon$. On utilise alors le fait que, si $\|\gamma\|_{\mathrm{BMO}} < \varepsilon$ (γ est reliée à la partie imaginaire de β), $\omega(s) = \exp\gamma(s)$ est un poids vérifiant la condition A_2 de Muckenhoupt. On sait par ailleurs que les opérateurs de Calderón–Zygmund restent continus lorsque $L^2(\boldsymbol{R}; dx)$ est remplacé par $L^2\big(\boldsymbol{R}; \omega(x)\,dx\big)$ et que $\omega \in A_2$.

On peut préciser le théorème 4 en appelant $X \subset \mathscr{L}\big(L^2(\boldsymbol{R}), L^2(\boldsymbol{R})\big)$ l'ensemble de tous les opérateurs de Cauchy T_b, $b \in V$. On munit X de la métrique définie par la norme d'opérateur. Alors l'application $T\colon V \to X$ est un homéomorphisme ([13]). En d'autres termes V est une carte globale de X.

Soient $\boldsymbol{R}^2_+$ le demi-plan supérieur ouvert et $\Phi\colon \boldsymbol{R}^2_+ \to \Omega_1$ une représentation conforme qui se prolonge en un homéomorphisme croissant de $\boldsymbol{R}$ sur Γ orientée. Définissons l'homéomorphisme réciproque $h\colon \boldsymbol{R} \to \boldsymbol{R}$ par $\Phi\big(h(s)\big) = z(s)$, s étant la longueur d'arc sur Γ. Alors $h'(x) = \omega(x)$ est un poids appartenant à la classe A_∞ de Muckenhoupt (Lavrentiev). Il en résulte que $\beta = \log h'(x)$ appartient à $\mathrm{BMO}(\boldsymbol{R})$.

THÉORÈME 5. *L'application qui à la fonction* $b \in V$ *associe la fonction* $\beta \in \mathrm{BMO}(\boldsymbol{R})$ *est réelle-analytique.*

Pour le voir on utilise la formule de Kerzman–Stein ([24]) permettant de passer du noyau de Cauchy au noyau de Szegö et l'on relie ce dernier à la représentation conforme ([8]).

3. Opérateurs multilinéaires et applications aux equations aux dérivées partielles

Ces applications spectaculaires, prévues par A. P. Calderón, ont été obtenues par E. Fabes, D. Jerison, C. Kenig et leurs élèves. Nous commençons par décrire de nouvelles actions multilinéaires de $L^\infty(\boldsymbol{R}^n)$ sur $L^2(\boldsymbol{R}^n)$ généralisant le produit ponctuel usuel. Ces actions ont des propriétés très remarquables que nous allons d'abord énoncer.

Désignons par $\mathscr{T}$ la topologie forte sur l'algèbre $\mathscr{L}\big(L^2(\boldsymbol{R}^n), L^2(\boldsymbol{R}^n)\big)$ notée $\mathscr{A}$. Soit $\mathscr{B} \subset \mathscr{A}$ la sous-algèbre des opérateurs de multiplication ponctuelle par les fonctions $b(x) \in L^\infty(\boldsymbol{R}^n)$, sous-algèbre que l'on munira encore de la topologie forte $\mathscr{T}$. Enfin le groupe $\mathscr{G}$ se compose des automorphismes de $\mathscr{A}$ de la forme particulière $T \to STS^{-1}$ où S agit sur $L^2(\boldsymbol{R}^n)$ par $Sf(x) = f(\delta x + x_0)$, $\delta > 0$, $x_0 \in \boldsymbol{R}^n$.

Les opérateurs multilinéaires T_k: $\big(L^\infty(\boldsymbol{R}^n)\big)^k \to \mathscr{A}$ que nous allons construire auront les deux propriétés suivantes:

si $b_{j,m}$, $1 \leqslant j \leqslant k$, $m \geqslant 1$, est une suite de fonctions de $L^\infty(\boldsymbol{R}^n)$ et si les opérateurs de multiplication correspondants $B_{j,m}$ convergent fortement vers l'opérateur B_j de multiplication par b_j, alors $T_k(b_{1,m}, \ldots, b_{k,m}) \to T_k(b_1, \ldots, b_k)$ au sens de la topologie $\mathscr{T}$. (7)

T_k commute avec l'action de $\mathscr{G}$ au sens que $T_k(Sb_1, \ldots, Sb_k) = ST_k(b_1, \ldots, b_k)S^{-1}$ pour tout S défini comme ci-dessus. (8)

Voici maintenant une recette pour construire de telles actions. Soit $\psi \in L^1(\boldsymbol{R}^n)$ une fonction vérifiant $\int_{\boldsymbol{R}^n} \psi(x)\,dx = 0$, $|\psi(x)| \leqslant C|x|^{-n+1}$ et $|\nabla\psi(x)| \leqslant C|x|^{-n}$ si $|x| \leqslant 1$, $|\psi(x)| \leqslant C_m|x|^{-m}$ et $|\nabla\psi(x)| \leqslant C_m|x|^{-m}$ pour tout $m \geqslant 1$ si $|x| \geqslant 1$.

Nous posons alors, pour tout $t > 0$, $\psi_t(x) = t^{-n}\psi(x/t)$ et appelons Q_t l'opérateur de convolution avec ψ_t. Nous appellerons $\varphi \in L^2(\boldsymbol{R}^n)$ une fonction ayant toutes les propriétés de ψ à l'exception de $\int_{\boldsymbol{R}^n} \varphi(x)\,dx = 1$ et $P_t: L^2(\boldsymbol{R}^n) \to L^2(\boldsymbol{R}^n)$ l'opérateur de convolution avec φ_t. Soit $m(\xi) \in C^\infty(\boldsymbol{R}^n \setminus \{0\})$ une fonction homogène de degré 0 et $M: L^2(\boldsymbol{R}^n) \to L^2(\boldsymbol{R}^n)$ l'opérateur de convolution associé. Posons, pour tout $t > 0$, $M_t = (1 - P_t)M$. Supposons l'existence d'une fonction $h \in \mathscr{S}(\boldsymbol{R}^n)$ telle que $M_t = H_t + R_t$ où H_t est l'opérateur de convolution avec $h_t(x) = t^{-n}h(x/t)$ et où le noyau-distribution de R_t est porté par $|x-y| \leqslant t$.

On appelle $b_j(x) \in L^\infty(\boldsymbol{R}^n)$, $1 \leqslant j \leqslant k$, des fonctions vérifiant $\|b_j\|_\infty \leqslant 1$; on désigne par $B_j: L^2(\boldsymbol{R}^n) \to L^2(\boldsymbol{R}^n)$ les opérateurs de multiplication correspondants et enfin $\mu(t) \in L^\infty(0, +\infty)$. Avec toutes ces notations, on a

THÉORÈME 6. *Il existe une constante C, ne dépendant que des fonctions φ, ψ, m et h (et de la dimension n) telle que pour tout $k \geqslant 1$, tout choix des $b_j \in L^\infty(\boldsymbol{R}^n)$ et de $\mu \in L^\infty(0, +\infty)$, l'opérateur*

$$L_k = \int_0^\infty Q_t B_1 M_t B_2 \dots M_t B_k M_t \mu(t) \frac{dt}{t} \tag{9}$$

soit continu sur $L^2(\boldsymbol{R}^n)$ et que $\|L_k\| \leqslant C^{k+1}\|\mu\|_\infty$. De plus le noyau-distribution $L_k(x, y)$ de L_k vérifie

$$\int_{|x-y| \geqslant 2|x-x'|} |L_k(x, y) - L_k(x', y)|\, dy \leqslant C^{k+1}\|\mu\|_\infty \tag{10}$$

pour tout $x \in \boldsymbol{R}^n$ et tout $x' \in \boldsymbol{R}^n$.

On a également l'estimation quadratique correspondante: si $f \in L^2(\boldsymbol{R}^n)$, alors $Q_t B_1 M_t \dots B_k M_t f$ appartient à $L^2(\boldsymbol{R}^{n+1}_+; dx\,dt/t)$ avec une norme $\leqslant C^{k+1}\|f\|_2$.

L'estimation (10) se prouve directement tandis que la continuité des opérateurs L_k sur $L^2(\boldsymbol{R}^n; dx)$ s'obtient par récurrence sur $k \geqslant 0$. On définit à cet effet L'_k et L''_k en remplaçant respectivement le *dernier* M_t intervenant dans (9) par P_t et par Γ_t (opérateur de convolution de symbole $\exp(-t^2|\xi|^2)$. Puisque $M_t = (1-P_t)M$, il vient $L_k = L_{k-1}B_k - L'_k M$. Par ailleurs $P_t - \Gamma_t = \tilde{Q}_t$ a les mêmes propriétés que Q_t ce qui rend immédiate la continuité de $L'_k - L''_k$. Pour terminer, on applique le théorème 1 à L''_k et la seule vérification non triviale est $L''_k(1) \in \mathrm{BMO}$. Or $L''_k(1) = L_{k-1}(b_k)$. Il suffit d'utiliser la continuité de L_{k-1} sur $L^2(\boldsymbol{R}^n)$ et (10) pour conclure que L_{k-1} envoie $L^\infty(\boldsymbol{R}^n)$ dans $\mathrm{BMO}(\boldsymbol{R}^n)$.

Pour terminer, le théorème 6 sera appliqué à la conjecture de Kato dont nous rappelons l'énoncé. Soit $A(x) = \big(a_{j,k}(x)\big)_{1 \leqslant j,k \leqslant n}$ une matrice à coefficients dans $L^\infty(\boldsymbol{R}^n)$. Posons, si $\xi \in \boldsymbol{C}^n$ et $\eta \in \boldsymbol{C}^n$, $\langle \xi, \eta \rangle = \sum_1^n \xi_j \overline{\eta}_j$ et supposons qu'il existe une constante $c > 0$ telle que, pour tout $\xi \in \boldsymbol{C}^n$ on ait $\mathrm{Re}\langle A(x)\xi, \xi \rangle \geqslant c|\xi|^2$.

À l'aide de $A(x)$, on construit suivant Kato [22] l'opérateur accrétif-maximal $T_A: V_A \to L^2(\boldsymbol{R}^n)$, défini formellement sur le sous-espace dense $V_A \subset L^2$ par $T_A f = -\mathrm{div}\big(A(x)\mathrm{Grad} f\big)$. Le domaine V_A dépend, de façon non linéaire, de A et l'on a $\mathrm{Re}(T_A u, u) \geqslant c\|\mathrm{Grad}\, u\|_2^2$ en posant (u, v)

$= \int_{R^n} u\bar{v}dx$. Kato a conjecturé que, dans ces conditions, le domaine de la racine carrée accrétive maximale $\sqrt{T_A}$ de T_A est $H^1(R^n)$, l'espace de Sobolev usuel.

Nous ne savons pas encore démontrer ce fait en toute généralité. Le théorème 6 fournit cependant l'existence d'une constante $\varepsilon_n > 0$ telle que la conjecture de Kato soit vraie dès que $\|A(x)-1\|_\infty < \varepsilon_n$. Pour le voir, on écrit $A(x) = 1+B(x)$ et l'on développe $\sqrt{T_A}$ en une série d'opérateurs multilinéaires en B que l'on traite par le théorème 6. La contrainte $\|B\|_\infty < \varepsilon_n$ permet de sommer la série écrite ([10], [16], [17], [18]).

Bibliographie

[1] Calderón A. P., Algebras of Singular Integral Operators, *Proc. Symp. Pure Math.* **10** (1965), pp. 18–55.

[2] Calderón A. P., Cauchy Integrals on Lipschitz Curves and Related Operators, *Proc. Nat. Acad. Sci. USA* **74** (1977), pp. 1324–1327.

[3] Calderón A. P., Commutators, Singular Integrals on Lipschitz Curves and Applications, *Proc. Internat. Congress Math., Helsinki* **1** (1978), pp. 85–96.

[4] Coifman R. R. and Fefferman Ch., Weighted Norm Inequalities for Maximal Functions and Singular Integrals, *Studia Math.* **51** (1974), pp. 241–250.

[5] Coifman R. R. et Meyer Y., Commutateurs d'intégrales singulières et opérateurs multilinéaires, *Ann. Inst. Fourier* **28** (3) (1978), pp. 177–202.

[6] Coifman R. R. et Meyer Y., Au delà des opérateurs pseudo-différentiels, *Astérisque* **57** (1978), Soc. Math. France.

[7] Coifman R. R. et Meyer Y., Une généralisation du théorème de Calderón, Proc. Sem. held at El Escorial, June 1979.

[8] Coifman R. R. et Meyer Y., Lavrentiev's Curves and Conformal Mappings, Inst. Mittag-Leffler, Report no. 5 (1983).

[9] Coifman R. R., David G. et Meyer Y., La solution des conjectures de Calderón, *Advances in Math.* **48** (1983), pp. 144–148.

[10] Coifman R. R., Deng D. G. et Meyer Y., Domaine de la racine carrée de certains opérateurs différentiels accrétifs, *Ann. Inst. Fourier* **33** (2) (1983), pp. 123–134.

[11] Coifman R. R., McIntosh A. et Meyer Y., L'intégrale de Cauchy définit un opérateur borné sur les courbes lipschitziennes, *Ann. of Math.* **116** (1982), pp. 361–387.

[12] Coifman R. R., Meyer Y. et Stein E. M., Un nouvel espace fonctionnel adapté à l'étude des opérateurs définis par des intégrales singulières, *Proc. Cortona Meeting*, July 1982, *Lecture Notes* à paraître.

[13] David G., Courbes corde–arc et espaces de Hardy généralisés, *Ann. Inst. Fourier* **32** (3) (1982), pp. 227–239.

[14] David G., *Opérateurs intégraux singuliers sur certaines courbes du plan complexe*, Éc. Polytechnique, Centre de Mathématique, mai 1983.

[15] David G. et Journé J.-L., Une caractérisation des opérateurs intégraux singuliers bornés sur $L^2(R^n)$, *C. R. Acad. Sci. Paris Sér. I* **296** (1983), pp. 761–764.

[16] Fabes E., Jerison D. et Kenig C., Multilinear Littlewood–Paley Estimates with

Applications to Partial Differential Equations, *Proc. Nat. Acad. Sci. USA* **79** (1982), pp. 5746–5750.

[17] Fabes E., Jerison D. et Kenig C., Multilinear Square Functions and Partial Differential Equations, Univ. Minnesota, Math. Report 82–167.

[18] Fabes E., Jerison D. et Kenig C., Necessary and Sufficient Conditions for Absolute Continuity of Elliptic-Harmonic Measure, *Ann. of Math.* **119** (1984), pp. 121–141.

[19] Fefferman Ch., Recent Progress in Classical Fourier Analysis, *ICM Vancouver* **1** (1974), pp. 95–118.

[20] Fefferman Ch. and Stein E. M., H^p Spaces of Several Variables, *Acta Math.* **129** (1972), pp. 137–193.

[21] Jones P. et Zinsmeister M., Sur la représentation conforme des domaines de Lavrentiev, *C. R. Acad. Sci. Paris* **295** (1982), pp. 563–566.

[22] Kato T., *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1966.

[23] Kenig C. and Meyer Y., *Kato's Square Roots of Accretive Operators and Cauchy Kernels on Lipschitz Curves Are the Same*, Institut Mittag-Leffler, Report no. 4 (1983).

[24] Kerzman N., Singular Integrals in Complex Analysis, *Proc. Symp. Pure Math.* **35**, part 2 (1979), pp. 3–41.

[25] Lemarie P. G., Communication orale.

[26] Meyer Y., *Lemme de Cotlar, opérateurs définis par des intégrales singulières et applications aux équations aux dérivées partielles*, Cours donné à l'Univ. Autonome de Madrid, avril 1983 (à paraître).

[27] Meyer Y., Intégrales singulières, opérateurs multilinéaires et équations aux dérivées partielles, *Sém. Goulaouic–Schwartz*, École Polytechnique, 3 mai 1983.

[28] Verchota G. C., *Layer Potentials and Boundary Value Problems for Laplace Equation in Lipschitz Domains*, University of Minnesota, Minneapolis (June 1982).

UNIVERSITÉ DE PARIS-SUD, ÉQUIPE DE RECHERCHE ASSOCIÉE AU CNRS (296), ANALYSE HARMONIQUE, MATHÉMATIQUE (BÂT. 425), 91405 ORSAY CÉDEX
ET
CENTRE DE MATHÉMATIQUES, ÉCOLE POLYTECHNIQUE, 91128 PALAISEAU CÉDEX, FRANCE

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

B. S. PAVLOV

Spectral Theory of Nonselfadjoint Differential Operators

As recently as twenty years ago the spectral analysis of singular, nonselfadjoint operators still appeared to be, to a large extent, "terra incognita". By that time, owing to the papers of Gelfand [3], Najmark [20, 21, 22] and Martirosyan [18], the rough features of spectrum structure of one-dimensional and three-dimensional Schrödinger operators with rapidly decreasing potential were already known. The main difficulties appearing in the study of these operators and, in general, any operators of that type were also recognized. It was the above two examples which made it clear (Pavlov [25, 26]) that even a very small (one-dimensional) nonselfadjoint perturbation could produce a point spectrum with an extremely rich structure. Their spectral function appears to be generalized (Marchenko [17]) and any attempt to consider, in the expansion theorem, continuous and point spectra separately is, generally speaking, doomed to failure. Problems of completeness and expansions with respect to eigenfunctions were, by then, solved simultaneously on the basis of a technique not beyond the limits of the Riesz integral. In the case of the point spectrum this led to the necessity of, analytically very delicate but practically hardly effective, summation "with brackets" (Lidskij [16]) and in the case of the continuous spectrum — to isolating so-called spectral singularities and computating them by the expansion theorem separately (Lyantse [14, 15]).

I will discuss here the results of spectral analysis of singular differential operators, mainly of Schrödinger type, as well as some results concerning abstract dissipative operators, which appeared after 1970. The analytic basis to most of these results is given by the Nagy–Foias functional model and Lax–Phillips scattering theory. We employ them for spectral analysis of differential operators of Schrödinger type, and of some operators which appear when considering resonance scattering.

Naturally, the concrete problems also forced us into some additional abstract considerations. I will discuss

1. Localization of the spectra of one-dimensional Schrödinger operators.

2. Separation of spectral components of abstract dissipative and Schrödinger-type operators.

3. Completeness of the family of eigenfunctions of an operator which has a discrete spectrum. "Joint completeness" for resonance scattering.

4. Biorthogonal eigenfunction-expansion theorems for abstract dissipative operators and operators of Schrödinger type with absolutely continuous spectrum. Summability of the eigenfunction expansions for operators which have absolutely continuous spectra and spectral singularities.

5. Some special problems: order (serial) structure of eigenvalues, comparative study of the spectral properties of selfadjoint and dissipative operators by means of scattering theory.

Progress made in the above-mentioned directions by various mathematicians has always involved some shift in neighbouring mathematical branches. In this sense a nonselfadjoint singular operator and particularly the Schrödinger operator serve as a unique testing ground for a great variety of theories and methods. In particular, we may say that it has already, to a large extent, come true what M. G. Krein (Ist Congress of Mathematicians, Moscow, 1966) predicted, namely that in the investigation of nonselfadjoint operators facts and methods of analytic functions theory will play a more and more important role.

1. Indeed, localization of the spectrum of a nonselfadjoint operator and the study of its structure was based on then known uniqueness theorems for analytic functions of various classes of smoothness. On the other hand, the needs of the spectral theory of such operators stimulated investigations of uniqueness theorems for the Gevrey classes (Khrushchev [9]). These investigations have recently produced the final results concerning the spectrum structure and localization of the Schrödinger operator with a rapidly decreasing potential, $\int_0^\infty |q(x)|x^n dx = C_n < \infty$, $n = 1, 2, \ldots$ (Khrushchev [11]).

Working in the spirit of Pavlov [25, 26] S. V. Khrushchev intruduced a class N_α of functions meromorphic in the upper half-plane C_+, holomorphic in the first quadrant C_{++} and satisfying the conditions:

1. $\sup_{k \in C_{++}} |g(k)| = 1$,

2. $\sup_{k\in C_{++}} |g^{(n)}(k)| \leqslant C_g Q_g^n n!\, n^{n/a}$, $n = 0, 1, \ldots$,
3. $|g(iy)| = 1$, $y > 0$,
4. $|g(x)| < 1$, $0 < x < \infty$,
5. $g'|(0)| \neq 0$,
6. $\lim_{k\to\infty} k[1-g(k)] = a > 0$.

Functions of class N_a turn out to be related to the Weyl functions of differential operators of the form $l_h = -y''_+ q(x)y$ with real potentials of class B_a:

$$|q(x)| \leqslant C_q \exp[-C_q x^{a/(1+a)}], \qquad -\infty < x < \infty,\ a > 0,$$

and complex boundary condition $y'(0) - hy(0) = 0$: they coincide with the characteristic functions (c.f.) of operators which can be expressed in terms of the Weyl function $m_\infty(\lambda)$ by the formula

$$S_h(\lambda) = \frac{m_\infty(\lambda)h}{m_\infty(\lambda)\bar{h}}.$$

THEOREM 1 (Khrushchev [11]). *Let $S(k)$ be an analytic function holomorphic at infinity. Then the function $S(\sqrt{\lambda})$, $\lambda \in C_+$ is the characteristic function of a Schrödinger operator l_h with $h \in C_+$, $q \in B_a$, $\operatorname{Im} q = 0$ iff $S \in N_a$.*

One of the most important problems in spectral analysis is to describe the null set of a characteristic function — i.e., the set of eigenvalues and spectral singularities of an operator. The following theorem describes fully all the possible null sets of characteristic functions for operators of type l_h with $h \in C_+$, $q \in B_a$, $\operatorname{Im} q = 0$:

THEOREM 2 (Khrushchev [11]). *A closed set E, $E \subset \bar{C}_+$, is a null set of the characteristic function of an operator of the above type iff*

1. *E is compact in $\bar{C}_+$,*
2. *$E \cap R \subset (0, \infty)$*
3. *E is a set of non-uniqueness for the Gevrey class, i.e., $E = f(0)$, $f \in G_a$.*

The last assertion links the description of null sets with the description of sets of uniqueness for functions of the Gevrey class G_a (Khrushchev [9]).

The progress made by various mathematicians in solving problems 2, 3, 4 is essentially related to the development of the theory of functional models for nonselfadjoint operators. For dissipative operators such a model was introduced and thoroughly investigated by Szökefalvi-Nagy, Foias, Lax, Phillips, Adamian and Arov. Owing to it one can formulate

many questions of operator theory in the language of function theory. The question about classification of invariant (spectral) subspaces reduces to the question about factorization of matrix-valued analytic functions (Szökefalvi-Nagy and Foias [41], Ginzburg [4] and later).

2. The problem of separation of spectral components of a dissipative operator can also be reduced to a question of analytic function theory (Pavlov [28], Vasyunin [42]).

Let B be a dissipative operator in Hilbert space H and S the characteristic function of B. Suppose now that S has a "scalar multiple" and can be factorized in the forms

$$S = S_i \tilde{S}_e = S_e \tilde{S}_i,$$

where S_i and $\tilde{S}_i$ are inner functions in C_+ and S_e and $\tilde{S}_e$ are outer functions; $S_i = \theta_s \tilde{\Pi}_d = \Pi_d \tilde{\theta}_s$ where θ_s and $\tilde{\theta}_s$ are singular inner functions and Π_d and $\tilde{\Pi}_d$ are Blaschke–Potapov products. We define absolutely continuous, discrete, and singular spectral invariant subspaces N_a, N_d, N_s, where the parts of the operator B have characteristic functions S_e, Π_d, θ_s respectively. The spectral invariant subspace N_i, where the part $B|N_i = B_i$ has characteristic function S_i we call the inner subspace of B. Generally, $N_i = \overline{N_d + N_s}$. The corresponding subspaces of the adjoint operator B^* we denote by N_a^*, N_d^*, N_s^*.

The analogous subspaces of a selfadjoint operator are of course mutually orthogonal. This is not true in general for nonselfadjoint dissipative operators. Before one can construct the spectral decomposition of a nonselfadjoint B, one must investigate the separation conditions for the parts of B in the subspaces N_a, N_d, N_s.

The key to the separability of the spectral components $B_a = B|N_a$, $B_d = B|N_d$, $B_s = B|N_s$ is the following simple fact. Let $f(k) = \{(k - k_a)^{-1} e_a, 0\}$, $e_a \in \operatorname{Ker} \tilde{S}_i(k_a)$, $\|e_a\| = 1$. Then f is an eigenvector of B,

$$Bf = k_a f, \quad \operatorname{Im} k_a > 0,$$

written in terms of the Nagy–Foias model. An easy computation shows that

$$\sin(f, N_a) = \|\tilde{S}_e(k_a) e_a\|.$$

So we see that $\sin(f, N_a)$ is small if $\tilde{S}_e(k_a) e_a$ is small. Roughly speaking, we can say that subspaces N_d and N_a will be at a non-zero angle if the discrete spectrum $\sigma_d = \sigma(B_d)$ is disjoint from the set of all real zeros of S_e, i.e., "spectral singularities". Now let us be more precise.

DEFINITION 1. Let us call the real point k_0 a *spectral singularity of the dissipative operator* B if it is a zero of its c.f. in the following weak sense:

$$\sup_{|k-k_0|<\varepsilon, \operatorname{Im} k>0} |\tilde{S}_\varepsilon^{-1}(k)| = \infty, \quad \forall \varepsilon.$$

The set of all spectral singularities of B we denote by $\sigma_0(B)$.

DEFINITION 2. Let us call a contour $\gamma \in \bar{C}_+$ a *Carleson contour* if it is a deformation of the real axis and the Lebesgue measure $|d\gamma|$ fulfils the Carleson condition

$$\sup_{-\infty<\lambda<\infty} \sup_{\varepsilon} Y_\varepsilon \int_{|\lambda-k|<\varepsilon} |d\gamma(k)| = C_0(\gamma) < \infty.$$

We say that the Carleson contour γ *separates* σ_0 from $\gamma_i = \sigma_d \cup \sigma_s$ if Ω_γ does not contain any point of σ_0 and ω_γ does not contain any point of σ_i.

Here Ω_γ is the open in $\bar{C}_+$ subset lying above γ, and ω_γ is the open set below γ; $\Omega_\gamma \cup \gamma \cup \omega_\gamma = \bar{C}_+$.

THEOREM 3 (Pavlov [28]). *If there exists a Carleson contour* γ, $\gamma \subset \bar{C}_+$, *which separates* σ_0 *from* σ_i, *and if*

$$\operatorname{ess\,sup}_{k\in\gamma} \|S^{-1}(k)\| \leqslant C < \infty,$$

then the subspaces N_i *and* N_a *form a non-zero angle. In particular, if* σ_i *and* σ_0 *are disjoint,*

$$\operatorname{dist}(\sigma_i, \sigma_0) > 0,$$

and if σ_0 *is compact, then* $\sin(N_i, N_a) > 0$. *In the other direction, if the c.f.* S *is scalar and continuous,* $S \in \operatorname{Lip} a$, $a > 0$, *and* $\sigma_i \cap \sigma_0 \neq \emptyset$, *then* $\sin(N_a, N_i) = 0$.

Another form of separability conditions for the spectral components depends on explicit calculation of the angle between N_a and N_i.

THEOREM 4 (Pavlov [28]). *If* S *has a scalar multiple* s *and* s_e *is its outer factor, then*

$$\sin(N_a, N_i) \geqslant \|\{s_e(B_i)\}^{-1}\|^{-1}.$$

THEOREM 5. (Pavlov [28], Vasyunin [42]). *Let $b, \tilde{b}$ be the Blaschke factors of $s_i, \tilde{s}_i, B_s = B|N_s, \tilde{B}_s = -B^*|N_s^*$. Then*

$$\sin(N_d, N_s) \geqslant \|\{\tilde{b}(\tilde{B}_s)\}^{-1}\|^{-1},$$

$$\sin(N_d^*, N_s^*) \geqslant \|\{b(B_s)\}^{-1}\|^{-1}.$$

In scalar case the equalities take place (Vasyunin [42]). If both right sides here are positive, then $\sin(N_d, N_s) > 0$ and $N_d + N_s = N_i$ is a direct sum (as is $N_d^* + N_s^* = N_i^*$).

THEOREM 6 (Pavlov [28]). *The spectral components N_i and N_d of a dissipative differential operator l_h are separable if $V \in B_\alpha$, $\alpha > 1$. On the other hand, for every α, $\alpha < 1$, there exists an operator l_h with a real potential $V \in B_\alpha$, $\operatorname{Im} h > 0$, such that its spectral components are not separable.*

Since the number of eigenvalues of l_h is finite if $\alpha > 1$, and moreover its c.f. does not have any real zeros of infinite order, the first assertion is clear. The second is based on some results on spectral inverse problems.

3. Among the various problems concerning completeness of the system of eigenfunctions of a dissipative differential operator, two are of special interest:

(α) When is $N_a + N_d = H$? (That is, when is $N_s = 0$?)

(β) Let N_a be 0. When is $N_d + N_d^* = H$?

The first is the usual problem of completeness. It is equivalent to asking when the singular factor θ_s of the characteristic function is trivial. This can be checked with the help of Helson's theorem [5]:

If S has a scalar multiple s, $s = s_e \cdot s_i$, and s_i is a Blaschke product, then $S = S_e \cdot \Pi$, $N_s = 0$. Here Π is a Blaschke–Potapov product.

Unfortunately, Helson's theorem holds only for operators generated by ordinary differential expressions (see Pavlov [27]). These occur in one-dimensional problems of resonance scattering. (See also Ivanov and Pavlov [8].) The problem of completeness of the family of eigenfunctions in many-dimensional problems of resonance scattering remains open. Lack of new function-theoretic criteria for triviality of singular factors restricts the study of many interesting operators we meet in mathematical physics (Lax and Phillips [12], Pavlov [33]).

Interesting analytic problems arise from the question of simultaneous completeness of eigenfunction systems of dissipative operators B, $-B^*$ with discrete spectra. This question arises, when one tries to show that solutions of the so-called Regge problem form a complete system

in $L_2(\varrho, (0, a))$. The Regge problem is one of the simpler Sturm–Liouville problem with impedance boundary condition, for instance:

$$-y'' = \varkappa^2\varrho^2 y, \quad y'(0) = 0, \quad y'(a) + i\varkappa y(a) = 0 \qquad (*)$$

(cf. Nikolskii, Hruscev, and Pavlov [23]). This can be reduced to the following more general problem in Hilbert space:

Let B be a dissipative operator in Hilbert space. Assume that B has an inner characteristic function: $S = \Pi\tilde{\theta}_s = \theta_s\tilde{\Pi}$. Then what are sufficient conditions for $N_d + N_d^* = H$?

There is as yet no solution of this problem in terms of the spectra of the singular function θ_s and the Blaschke product Π. In the scalar case, a less precise result can be obtained from the Helson–Szegö–Devinatz–Widom theorem, which allows one to characterize the case when $\sin(N_s^*, N_s) > 0$ (Nikol'skii, Pavlov, and Khrushchev [23]). Let us note that

$$N_s^* = H\theta N_d, \quad N_s = H\theta N_d^*.$$

THEOREM 7 (Nikol'skii, Pavlov, and Khrushchev [23]). *Let $S = \pi\theta$ be an inner scalar characteristic function defined on C_+,*

$$K_\pi = H_+^2 \ominus \pi H_+^2, \quad K_\theta = H_+^2 \ominus \theta H_+^2;$$

let P_π and P_θ be the orthogonal projectors onto K_π and K_θ. Then the following conditions are equivalent:

1. *$P_\theta: K_\pi \to K_\theta$, $P_\pi: K_\theta \to K_\pi$ are isomorphisms.*
2. *For the corresponding model operator B on $K = H_+^2 \ominus SH_+^2$ we have*

$$\sin(N_d, N_d^*) > 0, \quad N_d + N_d^* = K.$$

3. $\operatorname{dist}_{L_\infty}(\bar{\pi}\theta, H^\infty) < 1$, $\operatorname{dist}_{L_\infty}(\pi\bar{\theta}, H^\infty) < 1$.

This theorem gives the following result for the mentioned above Regge problem.

THEOREM 8 (Nikol'skii, Pavlov, and Khrushchev [23]). *For a smooth positive function p on $[0, a_\varrho] \cup (a_\varrho, a)$ such that $\varrho(a_\varrho) \neq 1$, $\varrho(x) = 1$, $x > a_\varrho$, we have*

(α) If $a_\varrho \leqslant a < a_\varrho + \int_0^{a_\varrho} \varrho\, ds$, then the family of eigenfunctions of the Regge problem (or eigenfunctions and associated functions if there are multiple eigenvalues) is complete in $L_{2,\varrho}(0, a)$.

(β) *If* $a = a_\varrho + \int_0^{a_\varrho} \varrho\, ds$, *then this family comprises a Riesz basis in* $L_{2,\varrho}(0, a)$.

The proof of the second assertion is parallel to the proof of the basis property for exponentials (Nikol'skii, Pavlov, and Khrushchev [23]).

Up to now only a few joint completn essproblem have been studied. Thus the analogous Regge problem with continuous spectrum is not yet solved. One anticipates that in the scalar case one will always get completeness.

4. In solving questions about expansions with respect to eigenfunctions it is an important step to choose a suitable system of eigenfunctions corresponding to the continuous spectrum. This problem is nontrivial even in the case of a selfadjoint operator with a simple spectrum and for operators with partial derivatives it plays in fact a crucial role. There is no general way to construct canonical eigenfunction system for selfadjoint differential operators. However, in the case of a sufficiently rapidly decreasing potential, the Schrödinger operator has such a system — scattered waves (Povzner [39]).

Surprisingly, a dissipative operator always has a canonical eigenfunction system for the continuous spectrum (Pavlov [29]). This system can be constructively written in terms of functional models, and, in fact, is fully analogous to Povzner's scattered waves since it is obtained by projecting the solutions of the scattering problem of a dilation of a given operator B onto the original space in which B acts. Further, the existence of such a canonical system permits us to obtain biorthogonal formulae for spectral projectors of operators both in the abstract case (Pavlov [29]) and for concrete differential operators (Pavlov [30]).

Let B be some dissipative operator in Hilbert space K which has outer c.f. $S: E \to E$, $I - S(k) \subset G_\infty(E)$, $K \in \bar{C}_+$; let Z be the selfadjoint dilation of B, $\exp iBt = P_K \exp iZt|K$, $t > 0$. To construct the system in question, one has only to project orthogonally a certain orthogonal basis of eigenfunctions of the dilation Z (of B) from the dilation space H onto K. These basis functions are also, for other reasons, called the "solution of the scattering problem". They can be expressed as linear combinations of eigenfunctions of Z associated with outgoing and incoming spectral representations $T_\pm$. On the other hand, the latter systems $\{\Phi_+\}$, $\{\Phi_-\}$ are determined up to uintary equivalence in an auxiliary Hilbert space E:

$$\Phi_+(\lambda, \nu) \xrightarrow{T_+} \begin{bmatrix} \delta(k-\lambda)\eta \\ 0 \end{bmatrix}, \quad \Phi_-(\lambda, \nu) \xrightarrow{T_-} \begin{bmatrix} 0 \\ \delta(k-\lambda)\nu \end{bmatrix}.$$

It is easy to see that these functions are not orthogonal, but their linear span is H. The complete (in H) orthogonal system of eigenfunctions of the dilation Z is composed of two parts:

$$\{\Phi_-(\lambda, \nu)\}, \quad -\infty < \lambda < \infty, \ \nu \in E,$$

$$\{\Phi^<(\lambda, \eta)\}, \quad \Phi^<(\lambda, \eta) = \delta_\eta^{-1}\{\Phi_+(\lambda, \eta) - S(\lambda)\Phi_-(\lambda, \eta)\},$$

where $\Delta = I - S^*S(\lambda) > 0$, η are the eigenvectors of Δ, $\Delta\eta = \delta_\eta\eta$, $\delta > 0$. The linear span of $\{\Phi^<\}$ is just the "additional component" in the sense of Nagy–Foias.

THEOREM 9 (Pavlov [29]). *A system of eigenfunctions of B complete in N_a is obtained by projecting $\Phi^<(\lambda, \eta)$ orthogonally onto K; this can be written in the distribution sense as follows:*

$$\varphi^<(\lambda, \eta) = \delta_\eta^{-1}(\lambda)P_k[\Phi_+(\lambda, \eta) - S(\lambda)\Phi(\lambda, \eta)].$$

The biorthogonal system consists of functions:

$$\psi^>(\lambda, \eta) = \delta_\eta^{-1}(\lambda)P_k[\Phi_-(\lambda, \eta) - S^*(\lambda)\Phi_+(\lambda, \nu)],$$

$$\nu = S(\lambda)\eta/s_\eta(\lambda), \quad s_\eta(\lambda) = \sqrt{1-\delta_\eta},$$

which are eigenfunctions of B^*:

$$\langle \psi^<_{\lambda,\eta}, \psi^>_{\lambda',\nu'} \rangle = -\delta(\lambda - \lambda')\frac{\langle S(\eta)\eta, \nu' \rangle}{\delta_\eta}.$$

The spectral projectors onto the subspaces N_a^ω of N_a corresponding to intervals ω of the absolutely continuous spectrum, have biorthogonal representations as integral operators with distribution kernels:

$$\varepsilon_\omega(k, k') = -\int_\omega \sum_{\{\eta_\lambda\}} \frac{\delta_\eta(\lambda)}{s_\eta(\lambda)} \varphi^<_{\lambda,\eta}(k)\psi^>_{\lambda,\nu}(k')\,d\lambda$$

provided there are no spectral singularities on ω ($\inf_{\lambda\in\omega} s_\eta(\lambda) > 0$).

In the case of a dissipative Schrödinger operator $B = -\Delta + V$, $V = q + i\alpha^2$, $\alpha \geqslant 0$, the role of auxiliary space E is played by $L_2(\operatorname{supp}\alpha)$. If we construct the minimal selfadjoint dilation Z of B in $L_2(R^-, E) \oplus L_2(R^3) \oplus$

$\oplus L_2(R^+, E)$ (Pavlov [30])

$$Z\begin{bmatrix} v_- \\ u \\ v_+ \end{bmatrix} = \begin{bmatrix} i\dfrac{dv_-}{d\xi} \\ Bu + \dfrac{a}{2}[V_-(0) + V_+(0)] \\ i\dfrac{dv_+}{d\xi} \end{bmatrix},$$

$$v_\pm \in W_2^1(R^\pm, E), \quad v_+(0) - v_-(0) = ia P_E u,$$

then the Nagy–Foias functional model of B can be constructed by expressing B in some special (say, incoming-outgoing) spectral representation of Z. Then, using the preceding theorem, one gets a canonical system of eigenfunctions $\{\varphi^<_{\lambda,\eta}\}$, $\{\psi^>_{\lambda,\nu}\}$, of the absolutely continuous spectrum of B and B^*. These functions satisfy the equations

$$L\varphi^<_{\lambda,\eta} = \lambda\varphi^<_{\lambda,\eta}, \quad ia^<_{\lambda,\eta} = \frac{1}{\sqrt{a\pi}}\eta,$$

$$L^*\psi^>_{\lambda,\nu} = \lambda\psi^>_{\lambda,\nu}, \quad ia\psi^>_{\lambda,\nu} = -\frac{1}{\sqrt{a\pi}}\nu;$$

here $\Delta\eta = \delta_\eta\eta$, $\delta_\eta > 0$; $\nu = S\eta/s_\eta$. Then we can easily construct spectral projectors and state the expansion theorem (Pavlov [30]).

The quality of the convergence or summability of the eigenfunction expansion depends on the "smoothness" (relative to B) of the function to be expanded and on localization of the spectral singularities of B.

Let $\{\sigma_\varepsilon\}$ be some family of subsets of R^1 such that $\|S^{-1}(\lambda)\| < 1/\varepsilon$, $\lambda \in \sigma_\varepsilon$, $\sigma_\varepsilon \to R$, when $\varepsilon \to 0$. Let $\hat{N}_a$ be the linear set, dense in $\hat{N}_a$, consisting of all $f \in K$ which are orthogonal to $D_-: f = P_K g$. We then set $\|f\|_s = \|g\|$ and call the norm $\|\ \|_s$ the strong norm in $\hat{N}_a$.

THEOREM 10 (Pavlov [29], [31]). *If* $f \in \hat{N}_a$, *then* $\operatorname{s-lim}_{\varepsilon\to 0} P_{\sigma_\varepsilon} f = f$.

By applying the above biorthogonal construction of projectors, it is possible to obtain formulae of summation, in the original norm, for spectral expansions with respect to the absolutely continuous spectrum. Namely the following is true:

THEOREM 11 (Pavlov [31]). *If the scalar multiple s_e of the c.f. of the dissipative operator B is a smooth function, $s_e \in \operatorname{Lip}\alpha(\bar{C}_+)$, $\alpha > 0$, then there exists a family $\{s_e^\delta\}$ of outer functions such that*

(α) $|s_e^\delta(k)s_e^{-1}(k)| \leqslant C(\delta)$, $k > C_+$, $\delta > 0$,
(β) $s_e^\delta(k) \underset{\delta\to 0}{\Longrightarrow} 1\{k\colon \operatorname{dist}(k, \sigma_0) \geqslant \delta > 0\}$.

We also have $s_e^\delta(B)f \subset \hat{N}_a$, $\delta > 0$, and

$$\text{s-}\lim_{\delta\to 0} s_e^\delta(B)f = f.$$

We observe that the method of summation of eigenfunction expansions indicated here shows clearly that expansion theorem need not involve separate terms corresponding to spectral singularities (see Lyantse [14]).

5. The effectiveness of function-theoretic methods in the study of spectral problems for nonselfadjoint differential operators has stimulated the search for an analogue of the functional model for general (nondissipative) nonselfadjoint operators. One of the first versions of such a model was presented by Davis and Foias. Unfortunately its realization required the spectral analysis of a J-selfadjoint operator on a space with an indefinite metric which present near the same difficulties. Naboko [19] introduced another functional model for a nonselfadjoint operator of the general form $A+iV$ which requires only the spectral analysis of a selfadjoint operator — a dilation of the dissipative operator $A+iV$. This model has already proved to be useful in numerous problems of perturbation theory, particularly in constructing a smooth theory of scattering for operators of the form $A+iV$ and clarifying questions about similarity to a selfadjoint operator (Naboko). The possibilities of application of this model are still far from being exhausted.

We will now concentrate on two special questions of the spectral theory of dissipative operators.

A very interesting question of spectral analysis is that of order structure of eigenvalues and eigenfunctions of a nonselfadjoint operator. Nikol'skii and Pavlov [24] found explicit order (serial) conditions which guarantee the basis property of the eigenfunctions of a contraction; these apply also for dissipative operators. Later Pavlov [27] showed that indeed these order conditions hold for systems of ordinary differential equations. In [40] Shubova found the serial basis in a problem of resonant scattering of acoustic waves by an "almost spherically symmetric" obstacle; and Popov in [38] did the same for the problem of resonant

scattering of electron waves by a system of zero-radius potentials. The order approach to bases was used by Avdonin [1] and Ivanov [7] in a problem in control theory.

After a fully description of bases of exponentials in $L_2(0, a)$, $a < \infty$, had been given in papers of Pavlov and Khrushchev [40], Avdonin and Ivanov started the investigation of the corresponding vector bases. They discovered that vector bases of exponentials appearing in a class of control problems also have order (serial) nature.

A basis of exponentials $\{e^{ik_j t}e_j\}$, $e_j \in E$, in $L_2\big((0, a), E\big)$, $\dim E < \infty$, is called serial if the set of all frequencies $\{k_j\}$ splits into a finite number of "series" $\{k_j^p\}_{j=1}^{\infty} = \Lambda^p$ such that unit vectors e_j^p corresponding to each series converge to specified limits as $j \to \infty$: $e_j^p \xrightarrow{E} e^p$. Applying simple serial bases of exponentials, Avdonin [1] and Ivanov [7] gave a full solution of the control problem for a system of connected strings.

A system of strings can be represented by a graph Γ with the control forces attached at the vertices. Such a system is called controllable if for any initial state U_0 there exist controls $\{f_j(t)\}$ ensuring that the solution $U(t, x)$, $x \in \Gamma$ of the system of equations describing the oscillation of the graph become zero at $t = T$: $U(x, T) = 0$, $x \in \Gamma$.

To each string l_j connecting nodes (r), (s) we assign two sequences of vector exponentials with frequencies equal to the eigenvalues of the string:

$$e^{i\frac{2n}{|e_j|}t}\left\{\begin{matrix}0\\0\\\vdots\\0\\1\\0\\\vdots\\0\\1\\0\\0\end{matrix}\right.\begin{matrix}\\\\\\\\(r)\\\\\\\\(s)\\\\\end{matrix}\Bigg\}_{n=0,\pm1,\pm2,\ldots}, \qquad e^{i\frac{2n+1}{|e_j|}t}\left\{\begin{matrix}0\\0\\\vdots\\0\\1\\0\\\vdots\\0\\-1\\0\\0\end{matrix}\right.\begin{matrix}\\\\\\\\(r)\\\\\\\\(s)\\\\\end{matrix}\Bigg\}_{n=0,\pm1,\pm2,\ldots}.$$

The joint family of these sequences is a basis in its linear span over the interval $(0, T)$ iff the given system of strings is controllable. In particular, a "tree" of strings is always controllable. The problem of controllability for a cycle is more complicated. For instance, if the strings are commensurable, $|l_j| = m_j a$, then the cycle is controllable iff the number of strings is odd and their joint "length" $\sum m_j$ is even.

All that has been said suggests that bases of exponentials which correspond to differential operators and bases of the eigenfunctions of those operators have in a number of interesting cases an exceptional inner structure whose extrenal expression is the serial structure exploited in various problems. The study of convergence of expansions with respect to such systems should obviously utilize this structure, which itself deserves some deep investigation. A differential operator is the quantum analogue of a "classical" dynamic system and, it seems, the serial structure of the operator's eigenvalues is determined by that system in both nonselfadjoint and selfadjoint (Lazutkin [13]) cases.

A comparative study of the spectral properties of dissipative and selfadjoint operators appearing in the problems of resonance scattering in systems with weakly related "outer" and "inner" components is another interesting question.

So far we have investigated very few problems of this sort. We may say that the experimental evidence is still accumulating and, perhaps, the time for generalizations has not yet arrived. We come across systems of that kind in studying the scattering of acoustic waves on trap-type domains (Petras [37]), the scattering of electron waves by a crystal or a fine film (one-dimansional model: Pavlov and Smirnov [36]), the scattering on a resonator with a small opening. As regards the last example, the theory of extensions has been used to study only a simple model problem (Pavlov, Faddeev). Analogous approach has been used for scattering of electron waves by a complex molecule (Pavlov, Faddeev, a model in terms of the theory of extensions). In all these problems pairs of operators come to light — a selfadjoint and a dissipative one. Their spectral properties are in a natural correspondence analogous to that which was studied in a corresponding model situation for operators with discrete spectra (D. Clark [2]) and latter for operators with general spectra (Pavlov and Faddeev [35], by means of scattering theory). From the point of view of applications one of the most important of such problems is that of the distance estimate between corresponding eigenvalues — real λ_n of a selfadjoint operator and complex k_n — of related dissipative operator; $\{k_n\}$ can be interpreted as resonances and $(\operatorname{Im} k_n)^{-1}$ as the lifetime of the resonance.

Another attractive prospect is opened by the fact, mentioned above (see 4), that eigenfunction expansion for dissipative operators is, in a way, even more constructive than for selfadjoint. In fact one has the following result, of a rather "experimental" character (Pavlov and Faddeev [35]).

Let U_0 be the shift operator in $L_2(C) = H_0$

$$U_0\colon f(\theta)\to\exp(i\theta)f(\theta), \quad 0<\theta<2\pi,$$

and let U be a unitary operator with absolutely continuous spectrum of multiplicity 1 in an abstract Hilbert space K. By means of extension theory one can construct a compound operator U_V, which acts on $H_0 \oplus K$ and differs from $U_0 \oplus U$ only by two-dimensional perturbation, which can be related by the coupling operator V. This pair can be studied by the methods of scattering theory.

THEOREM 12. *For coupling V tending to zero, the eigenfunctions of the absolutely continuous spectrum of the contraction operator $T = P_K U_V|_K$ converge in the distribution sense to eigenfunctions for the absolutely continuous spectrum of U.*

In this lecture I have not touched upon many interesting subjects (some of them close to my own research). Also I have only mentioned a small number of mathematicians working in the area chosen for this survey. I hope that the theory of nonselfadjoint differential operators of mathematical physics, so conveniently situated in the "triple point" between the theory of differential equations, functional analysis and function theory will in time multiply the number of its followers and then the time will come for more detailed surveys.

I wish to explain my cordial gratitude to my friends G. Bjork and C. Davis for their helpful support and constructive discussions of the matters, described in this lecture.

References

[1] Авдонин С. А., *Школа по теории операторов в функциональных пространствах*, тезисы докладов, Минск, 1982, стр. 3–4.
[2] Clark D., *J. Anal. Math.* **25** (1972), pp. 169–191.
[3] Гельфанд И. М., *Успехи математических наук* **7** (6) (1952).
[4] Гинзбург Ю. П., *Доклады АН СССР* **159** (3) (1964), стр. 489–492.
[5] Helson H., *Lectures on Invariant Subspaces*, 1964.
[6] Helson H. and Szegö G., *Ann. Mat. Pura Appl.* **51** (1960), pp. 107–138.
[7] Иванов С. А., *Школа по теории операторов в функциональных пространствах*, тезисы докладов, Минск, 1982, стр. 75.
[8] Иванов С. А., Павлов В. С., *Изв. АН СССР* **42** (1) (1978), стр. 26–55.
[9] Хрущев С. В., *Arkiv för Math.* **15** (1977), pp. 253–304.
[10] Хрущев С. В., *Доклады АН СССР* **247** (1) (1979), стр. 44–48.
[11] Хрущев С. В., Автореф. докторс. диссертации, Ленинград, 1982.
[12] Lax P. and Phillips R., *Comm. on Pure and Appl. Math.* **30** (1977), pp. 193–233.
[13] Лазуткин В. Ф., *Выпуклый биллиард и собственные функции оператора Лапласа*, Изд. ЛГУ, 1981.
[14] Лянце В. Э., *Математический сборник* **64** (4) (1964), стр. 521–561.

[15] Лянце В. Э., *Математический сборник* **65** (1) (1964), стр. 47–103.
[16] Лидский В. Б., *Труды моск. мат. общ.* **11** (1962), стр. 3–35.
[17] Марченко В. А., *Математический сборник* **52** (2) (1960).
[18] Мартиросян Р. М., *Изв. АН СССР, сер. мат.* **10** (1) (1957).
[19] Набоко С. Н., *Труды МИАН* **147** (10) (1980), стр. 86–114.
[20] Наймарк М. А., *Доклады АН СССР* **85** (1952), стр. 41–44.
[21] Наймарк М. А., *Доклады АН СССР* **89** (1953), стр. 213–216.
[22] Наймарк М. А., *Труды моск. мат. общ.* **3** (1954), стр. 181–270.
[23] Nicolskij N. K., Hruscev S. V., and Pavlov B. S., *Complex Analysis and Spectral Theory*, Lecture Notes in Mathematics **864**, Springer-Verlag, Berlin–Heidelberg–New York, 1981.
[24] Никольский Н. К., Павлов Б. С., *Изв. АН СССР, сер. мат.* **34** (1970), стр. 90–133.
[25] Павлов Б. С., *Доклады АН СССР* **141** (1961), стр. 807–810.
[26] Павлов Б. С., *Доклады АН СССР* **146** (1962), стр. 1267–1270.
[27] Павлов Б. С., *Изв. АН СССР, сер. матем.* **37** (1973), стр. 217–246.
[28] Павлов Б. С., *Изв. АН СССР, сер. матем.* **39** (1975), стр. 123–148.
[29] Павлов Б. С., *Вестник ЛГУ, сер. матем. мех.* **1** (1975), стр. 130–137.
[30] Павлов Б. С., *Математический сборник* **102** (144) (4) (1977).
[31] Павлов Б. С., *Проблемы матем. физики*, ИЗД. ЛГУ **9** (1979), стр. 113–121.
[32] Павлов Б. С., *Докл. АН СССР* **247** (1979), стр. 37–40.
[33] Павлов Б. С., *Проблемы матем. физ.*, ИЗД. ЛГУ **10** (1982), стр. 183–208.
[34] Павлов Б. С., Фаддеев М. Д., *Записки научн. семин. ЛОМИ* **115** (1982), стр. 215–227.
[35] Павлов Б. С., Фаддеев М. Д., *Записки научн. семин. ЛОМИ* **126** (1983), стр. 113–126.
[36] Павлов Б. С., Смирнов Н. В., *Вестник ЛГУ, сер. мат. мех.* **13** (1977), стр. 71–80.
[37] Петрас С. В., *Функциональный анализ* **9** (2) (1975), стр. 89–90.
[38] Попов И. Ю., *Проблемы матем. физ.* ИЗД. ЛГУ **10** (1982), стр. 209–241.
[39] Повзнер А. Я., *Мат. сборник* **32** (74) (1952), стр. 109–156.
[40] Шубова М. А., *Проблемы матем. физ.* ИЗД. ЛГУ **10** (1979), стр. 145–179.
[41] Szökefalvi-Nagy B. and Foias C., *Analyse harmonique des operateurs de l'espace de Hilbert*, Masson, Paris, and Akad. Kiado, Budapest, 1967.
[42] Васюнин В. И., Кандидатская диссертация, Ленинград, 1976.

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

GILLES PISIER

Finite Rank Projections on Banach Spaces and a Conjecture of Grothendieck

In this report we discuss several recent results concerning the existence or non-existence of well behaved finite rank projections on a Banach space. We will be interested in projections with large ranks and norms as small as possible.

0. Standard notations

We remind the definition of the Banach–Mazur distance $d(E, F)$ between two Banach spaces:

$$d(E, F) = \inf\{\|T\| \ \|T^{-1}\|\},$$

where the infimum is over all isomorphisms T from E onto F. If E and F are not isomorphic, we set $d(E, F) = +\infty$. As usual, we will denote by l_p^n the space R^n equipped with the norm

$$\|(\alpha_i)\| = \Big(\sum_1^n |\alpha_i|^p\Big)^{1/p}.$$

1. The finite-dimensional basis problem

Since Enflo's example [4], we know that there are Banach spaces which fail the approximation property and, *a fortiori*, fail to have a basis.

By definition, a Banach space X has the *approximation property* (in short A.P.) if the identity is approximable by finite rank operators uniformly on every compact subset of X.

The space X has a basis if (and only if) there is a sequence $\{P_n\}$ of finite

rank projections on X such that

$$P_{n-1}(X) \subset P_n(X) \text{ for all } n > 1; \tag{1}$$

$$\operatorname{rank} P_n = n; \tag{2}$$

$$\overline{\bigcup P_n(X)} = X; \tag{3}$$

$$\sup \|P_n\| < \infty. \tag{4}$$

If we then select a sequence $\{x_n\}$ such that $x_n \neq 0$ and $x_n \in P_n(X) \cap \cap \operatorname{Ker} P_{n-1}$, we obtain a basis of X in the usual sense.

One defines the *basis constant* of X as $b(X) = \inf\{\sup \|P_n\|\}$ where the infimum runs over all possible sequences $\{P_n\}$ as above. Clearly, this makes sense also when X is of finite dimension d, by restricting (1) and (2) to all $n \leqslant d$ (and $P_n = \mathrm{Id}_X$ for all $n \geqslant d$).

We now examine $b(X)$ for a finite-dimensional space X. Although this is surprising at first glance, the result of Enflo does *not* imply (and is *not* implied by !) the existence of a sequence $\{X_n\}$ of finite-dimensional spaces with basis constants $b(X_n)$ tending to infinity with n. In fact, until recently, the following question was still open:

Question 1. Is there a universal bound for $b(X)$ when X runs over all finite-dimensional (in short f.d.) spaces ?

The only known upper bound for $b(X)$ is $b(X) \leqslant (\dim X)^{1/2}$, which follows immediately from a classical result of F. John: on every n-dimensional space there is an inner product norm which is $\sqrt{n}$ equivalent to the original one (cf. [14]). Recently, Gluskin and Szarek gave the expected negative answer to Question 1, (cf. [8], [32]). More precisely, we have

THEOREM 1 ([32]). *There is an absolute constant $\delta > 0$ with the following property: for each integer n, there is an n-dimensional space X_n such that, for every projection $P: X_n \to X_n$ with rank $\left[\frac{n}{2}\right]$, we have $\|P\| \geqslant \delta \sqrt{n}$. In particular, we have $b(X_n) \geqslant \delta \sqrt{n}$.*

We refer to [32] for the rather long and delicate proof of this result. It should be mentioned that Szarek's construction relies on probabilistic ideas: the spaces X_n are selected "at random" in a clever way so that the property in Theorem 1 occurs actually with large probability. Szarek's methods were directly inspired by a previous (quite remarkable) paper of Gluskin [7], where the latter proved the existence (with positive probability, in some sense) for each n of two n-dimensional spaces X_n, Y_n

satisfying

$$\inf_n n^{-1} d(X_n, Y_n) > 0.$$

Later, Gluskin pushed his methods to answer Question 1 with a weaker form of Theorem 1 (cf. [8]), while Szarek proved independently the essentially sharp version of Theorem 1 quoted above.

2. The complemented l_p^n problem

Enflo's example tells us that there exist Banach spaces which do not have "enough" finite rank projections of uniformly bounded norms to approximate the identity operator in the pointwise topology.

This leaves open the following question, which can be roughly formulated as follows: are there non-trivial finite rank projections *at all* on a general space?

Question 2. Let X be an infinite-dimensional Banach space. Does there exist a sequence $\{P_n\}$ of finite rank projections on X with uniformly bounded norms and unbounded ranks?

Of course, this is much weaker than saying that X has a basis, since $\{P_n\}$ is not required to satisfy (1) or (3); and, quite obviously, there are spaces without the A.P. which possess the above property. In [20], Lindenstrauss reformulated the preceding question, in a stronger formulation:

Question 3. Let $k_n = \operatorname{rank} P_n$. Can one find P_n's as in Question 2 with the additional property that, for some p in $[1, \infty]$, we have $\sup_n d(P_n(X), l_p^{k_n}) < \infty$?

In that case, we can as well assume that $k_n = n$ and we then say that X contains uniformly complemented l_p^n's. Moreover, it is easy to see that the problem reduces to the cases $p = 1, 2$ and ∞.

In several special cases, positive answers were given in [34], [15] and in [35]. To motivate Question 3, let us recall a fundamental theorem of Dvoretzky (cf. [5]): for every infinite-dimensional space X, for every n and $\varepsilon > 0$, we can find a subspace X_n of X such that $d(X_n, l_2^n) < 1 + \varepsilon$. Roughly, this means that X reproduces somewhere the structure of Euclidean spaces almost isometrically. For various reasons, mainly in operator theoretic considerations, it is of interest to decide when we can find subspaces X_n as above together with projections P_n: $X \to X_n$ such that $\sup \|P_n\| < \infty$. It is rather simple to check that L^p spaces or l_p spaces have this property for $1 < p < \infty$ and not for $p = 1$ or ∞. In L^p, we even

have complemented *infinite*-dimensional Hilbertian subspaces (take the span of Gaussian variables, or Rademacher functions); but, in l_p, this is no longer true, so that, in general, we must restrict our attention to the finite-dimensional subspaces.

Unfortunately, in general, the answer to Question 2 is negative:

THEOREM 2 ([30]). *There is an infinite-dimensional Banach space X and a constant $\delta > 0$ such that all finite rank projections $P: X \to X$ satisfy*

$$\|P\| \geqslant \delta \ (\mathrm{rank} P)^{1/2}.$$

In particular, X is a counterexample to the property in Question 1.

Remark. Let E be an n-dimensional subspace of a general space X. Then there is a projection $P: X \to E$ with $\|P\| \leqslant \sqrt{n}$. This is a classical result (originally due to Kadeč–Snobar). Therefore, in the above space X, this general upper bound cannot be improved (at least asymptotically)-regardless of how E is chosen in X.

The space constructed for Theorem 2 also fails the A. P. We will rel turn to this in the next section.

There is however a major difference here from the approximation problem. Indeed, now we know that there are extremely "nice" spaces, e.g. uniformly convex spaces, which fail the A.P. (see [31] for examples of subspaces of l_p, $p \neq 2$, failing the A.P.). However, it turns out that, in all uniformly convex spaces, Question 3 (and a fortiori Question 2) has a positive answer, so that the "optimistic" conjecture is correct for these spaces. (See the corollary of Theorem 5). To state this in full generality, we will need some terminology. We will say that a Banach space X contains l_p^n's *uniformly* if, for some $\lambda > 1$, there is a sequence of subspaces X_n of X such that $d(X_n, l_p^n) \leqslant \lambda$. It is known (cf. [13] [17]) that, if this property holds for some $\lambda > 1$, it also holds for all $\lambda > 1$. In particular, a uniformly convex space cannot contain l_1^n's uniformly (consider the case $n = 2$ and let λ tend to 1).

With this terminology, Dvoretzky's theorem says that any infinite-dimensional space contains l_2^n's uniformly.

The study of the l_p^n-subspaces of a Banach space is intimately connected with the notions of type and cotype, which are defined as follows. Let $D = \{-1, 1\}^{\mathbf{N}}$, let μ be the uniform probability on D and let $\varepsilon_n: D \to \{-1, +1\}$ be the n-th coordinate on D. We will denote the space $L^2(D, \mu; X)$ simply by $L^2(X)$.

DEFINITION. Let $1 \leqslant p \leqslant 2 \leqslant q \leqslant \infty$. A space X *is of type* p (resp. *of cotype* q) if there is a constant C such that, for all finite sequences $(x_1, \ldots, x_n)$ in X, we have

$$\Big\| \sum_1^n \varepsilon_i x_i \Big\|_{L^2(X)} \leqslant C \Big(\sum \|x_i\|^p \Big)^{1/p}$$

$$\Big(\text{resp. } \Big\| \sum_1^n \varepsilon_i x_i \Big\|_{L^2(X)} \geqslant \frac{1}{C} \Big(\sum \|x_i\|^q \Big)^{1/q} \Big).$$

We will denote by $T_p(X)$ (resp. $C_q(X)$) the smallest constant C satisfying this. Every normed space is of type 1 and of cotype ∞. We refer to [22] for more details. We should mention that these notions are used frequently in the current study of probability on Banach spaces (cf. e.g. [12]). In the latter area, the spaces which do not contain l_1^n's uniformly are called *B-convex*; this class of spaces was introduced by A. Beck, in the early sixties, to investigate the strong law of large numbers for vector valued random variables.

The results of [22] and [17] combined together, yield the following theorem, which relates these analytic notions with the more geometric concept of "containing l_p^n's".

THEOREM 3 ([22] [17]). *For an infinite-dimensional space* X, *let*

$$p(X) = \sup \{p | X \text{ is of type } p\}$$

and

$$q(X) = \inf \{q | X \text{ is of cotype } q\}.$$

Then, X *contains* l_p^n*'s uniformly for* $p = p(X)$ *and also for* $p = q(X)$.

This implies that $p(X)$ (resp. $q(X)$) coincides with the smallest (resp. largest) p such that X contains l_p^n's uniformly.

In particular, $p(X)$ is non-trivial, i.e. $p(X) > 1$, iff X does not contain l_1^n's uniformly; while $q(X)$ is non-trivial, i.e. $q(X) < \infty$, iff X does not contain l_∞^n's uniformly.

In the concluding remarks of [22], it was asked whether there is a "duality" between $p(X)$ and $q(X^*)$, when $p(X) > 1$.

The key to solve this problem is the notion of "K-convexity"; a Banach space X is called K-convex if the orthogonal projection R from $L^2(D, \mu)$ onto the span of the sequence $\{\varepsilon_n\}$ induces a bounded operator, denoted

by $\tilde{R}$, on $L^2(X)$. Whenever it is bounded, $\tilde{R}$ is a projection onto the space of all series $\sum_{n=1}^{\infty} \varepsilon_n x_n$, with x_n in X, which converge in $L^2(X)$.

It is rather striking that the boundedness of this single projection on $L^2(X)$ implies the boundedness of many projections on X, as the following result of Figiel and Tomczak–Jaegermann shows:

THEOREM 4 ([6]). *Any K-convex space X is locally π-Euclidean, which means that there is a constant C and, for each n and $\varepsilon > 0$, there is an integer $N = N(\varepsilon, n)$ satisfying the following property:*

for every subspace $E \subset X$ with $\dim E \geqslant N$, there is a subspace $F \subset E$ of dimension n and a projection $P: X \to F$ such that $d(F, l_2^n) < 1+\varepsilon$ and

$$\|P\| \leqslant C.$$

The proof uses the same isoperimetric inequality as in [5]; the space F and the projection P are obtained by a suitable random choice. For a different approach using random matrices, see [1].

Fortunately, it turns out that K-convexity admits a simple "geometric" characterization:

THEOREM 5 ([28]). *A Banach space is K-convex iff it does not contain l_1^n's uniformly.*

It is the "if" part which is non-trivial. This shows that the converse to Theorem 4 is true:

COROLLARY. *The properties "X does not contain l_1^n's uniformly", "$p(X) > 1$", "X is K-convex" and "X is locally π-Euclidean" are all equivalent.*

Several special cases were already known, in particular, for Banach lattices, cf. [34] [15]. Moreover, if X is K-convex, then X is of type p iff X^* is of cotype p' (cf. [22], remark 2.10); hence if $p(X) > 1$, we have:

$$\frac{1}{p(X)} + \frac{1}{q(X^*)} = \frac{1}{p(X^*)} + \frac{1}{q(X)} = 1.$$

We should mention that the proof of Theorem 5 relies heavily on some results from the theory of holomorphic semi-groups.

We refer to [28] for more details.

Remark. The results of [29] suggest the following conjecture.

CONJECTURE. *In any space X in which Question 2 has an affirmative answer, the same is true for Question 3.*

Indeed, this is verified in [29] if $q(X) = 2$, and there is still some hope that the approach of [29] will prove the conjecture in general (cf. [29], p. 143).

3. A conjecture of Grothendieck

Let X, Y be Banach spaces and let $u = \sum_{i=1}^{n} x_i \otimes y_i$ be an element of the algebraic tensor product $X \otimes Y$ ($x_i \in X$, $y_i \in Y$). Let B_X be the unit ball of X. Grothendieck defined the injective and projective norms as

$$\|u\|_\vee = \sup\left\{\sum_1^n x^*(x_i)y^*(y_i) \mid x^* \in B_{X^*};\ y^* \in B_{Y^*}\right\}$$

and

$$\|u\|_\wedge = \inf\left\{\sum_1^n \|x_i\|\|y_i\|\right\},$$

where the infimum runs over all possible representations of u. He denoted by $X \check{\otimes} Y$ and $X \hat{\otimes} Y$ the completions of $X \otimes Y$ with respect to the corresponding norms (cf. [10] [11]).

Obviously, $\|u\|_\vee \leqslant \|u\|_\wedge$, so that there is a natural norm decreasing map from $X \hat{\otimes} Y$ into $X \check{\otimes} Y$. At the end of [10], Grothendieck listed six open questions, which are now (essentially) all solved. The first (and main) one was the A.P. problem. The last one was the following:

Question 4 If $X \hat{\otimes} Y = X \check{\otimes} Y$, is it true that either X or Y must be finite-dimensional?

In view of the fact that $\|\ \|_\vee$ and $\|\ \|_\wedge$ are, respectively, the smallest and the greatest reasonable tensor norm, it is natural to ask if they can happen to be equivalent on $X \otimes Y$ in any other case than the trivial one when one of the dimensions is finite. This is precisely the content of Question 4. Let us consider the case when X and Y are in duality. Then $X^* \check{\otimes} X$ can be identified with the closure of the finite rank operators in the space $\mathscr{L}(X, X)$ of all bounded operators on X.

On the other hand, the elements u in $X^* \check{\otimes} X$ which are in the image of the natural map $J\colon X^* \hat{\otimes} X \to X \check{\otimes} X$ are exactly those which can be written as

$$u(x) = \sum_{n=1}^{\infty} x_n^*(x)x_n \tag{5}$$

for all x in X, with x_n^* in X^* and x_n in X, such that $\sum_1^\infty \|x_n^*\|\,\|x_n\| < \infty$. These are called *nuclear operators* and their "nuclear norm" is defined as $N(u) = \inf \sum_1^\infty \|x_n^*\|\,\|x_n\|$ where the infimum runs over all representations satisfying (5). Grothendieck showed that X has the A. P. iff $J: X^* \hat{\otimes} X \to X \check{\otimes} X$ is *injective*; in that case the trace of a nuclear operator is well defined, and it is then easy to prove that J is an *isomorphism* only if the dimension of X is finite. However, until recently it was not known whether X must be f.d. when J is merely assumed to be *surjective*. This question belongs to the same family as Question 2. Roughly formulated, it reads: if the dimension of X is infinite, is there any non-trivial operator at all in $X^* \check{\otimes} X$? Indeed, the nuclear operators are trivial in the sense that they are just absolutely convergent series of *rank one* operators.

We should mention that a positive answer to each of the preceding questions was given in [3] if X does not contain l_1^n's uniformly (this can now be derived easily from the more recent Theorems 4 and 5). Moreover, the following finite-dimensional version of Grothendieck's conjecture was proved in [25].

THEOREM 6. *Let $\{X_n\}$ be a sequence of f.d. Banach spaces and let Y be a Banach space. Assume that for some constant C, we have $\|u\|_\wedge \leqslant C\|u\|_\vee$, for all u in $X_n \otimes Y$ and for all n. Then either* $\sup_n \dim X_n < \infty$ *or* $\dim Y < \infty$.

In particular, if either X or Y has a basis (or merely the property in Question 2), then the answer to Question 4 in "yes". However, in general, the answer to both of these questions is negative:

THEOREM 7. *Let E be any Banach space of cotype 2. (For instance $E = l_1$ or $E = l_2$.) Then there exists a Banach space X which contains E isometrically and is such that:*

(i) $X \hat{\otimes} X = X \check{\otimes} X$

(ii) *The map J from $X^* \hat{\otimes} X$ into $X^* \check{\otimes} X$ is surjective. Equivalently, there is a constant C such that every finite rank operator u on X satisfies $N(u) \leqslant C\|u\|$.*

Moreover, if E is separable, we can obtain a separable space X as above.

We do not know, however, if there is a reflexive space X (or merely not containing l_1) which possesses any of the properties (i) and (ii). Similarly, Question 2 is still open for reflexive spaces. Also, we could not construct a space X such that every *compact* operator on X is nuclear.

Note that if $P: X \to X$ is a finite rank projection, then it is well known that $N(P) \geqslant (\operatorname{rank} P)^{1/2}$, therefore Theorem 2 is a corollary of Theorem 7 with $\delta = \frac{1}{C}$. To describe the proof of Theorem 7, we focus on (i) ((ii) is obtained as a consequence of (i) by the results of [25]).

The basic idea is to construct a sequence of Banach spaces $E_0 \subset E_1 \subset E_2 \subset \ldots E_n \subset E_{n+1} \subset \ldots$, with E_n isometrically embedded in E_{n+1} with $E_0 = E$, and such that, for some constant K, we have for all n and all u in $E_n \otimes E_n$

$$\|u\|_{E_{n+1}\hat{\otimes}E_{n+1}} \leqslant K \|u\|_{E_n \check{\otimes} E_n}. \tag{6}$$

Once the sequence $\{E_n\}$ is obtained, it is quite easy to check that $X = \overline{\bigcup E_n}$ satisfies the above property (i).

The difficulty in the construction of the sequence $\{E_n\}$ lies in the fact that (6) can hold for some E_{n+1} containing E_n, *only if* E_n satisfies a certain restrictive condition; therefore, to carry on the construction, we must make sure, at each step, that E_{n+1} satisfies not only (6) but also this condition, which we now make more explicit.

Let $u: E \to F$ be an operator between Banach spaces, we say that u factors through a Hilbert space H if there are operators $A: E \to H$ and $B: H \to F$ such that $u = BA$; this property is "controlled" by the following norm: $\gamma_2(u) = \inf(\|B\| \, \|A\|)$, where the infimum is over all possible factorizations of u. If u is an element of $E \otimes E$, we will denote by $\gamma_2(u)$ the above norm computed for the operator from E^* into E associated to u. It is then easy to see that $\gamma_2(u) \leqslant \|u\|_{E\hat{\otimes}E}$ for any u in $E \otimes E$. Therefore, if (6) holds, then the space E_n must satisfy

$$\forall u \in E_n \otimes E_n, \quad \gamma_2(u) \leqslant K \|u\|_\vee. \tag{7}$$

This strongly indicates that, in order to prove Theorem 7, we must first investigate this condition (7). This was done in [25].

THEOREM 8. [25] *Let E and F be Banach spaces such that both E^* and F are of cotype 2. Then there exists a constant K (depending only on the cotype 2 constants of E^* and F) such that every finite rank operator $u: E \to F$ satisfies*

$$\gamma_2(u) \leqslant K \|u\|.$$

COROLLARY. *If moreover E or F has the A.P., then any bounded operator $u: E \to F$ factors through a Hilbert space.*

Applied to the identity operator, this yields

COROLLARY. *If a Banach space E and its dual E^* are of cotype 2, and if E possesses the A.P., then E is isomorphic to a Hilbert space.*

These results were conjectured in [21] (without the A.P.). Up to now, they cover all the known couples of Banach spaces E and F such that every bounded operator $u: E \to F$ factors through a Hilbert space. They can be viewed as an "abstract" form of a classical theorem of Grothendieck, who proved this for $E = L^\infty$ and $F = L^1$. His result was extended in many ways. Maurey (cf. [21]) discovered the relation with the notion of cotype and proved this result for $E = L_\infty$ and F any space of cotype 2. The main examples of cotype 2 spaces are L^1 spaces and their subspaces. More generally, the dual or the predual of a C^*-algebra is of cotype 2 [33], as well as the quotients L^1/R when R is a reflexive subspace of L^1 ([16], [24]). Recently, Bourgain [2] proved that L^1/H^1 is of cotype 2. Actually, the last two examples play an important role in the proof of Theorem 7. It is conceivable that the assumptions of Theorem 8 are necessary if neither E nor F is isomorphic to a Hilbert space (see [25], remark 2.4). However, the A.P. cannot be removed from the preceding two corollaries. Indeed, in the proof of Theorem 7, we actually construct a sequence $\{E_n\}$ verifying (6) and also such that

$$\sup_n C_2(E_n) < \infty. \tag{8}$$

This last property implies by Theorem 8 that for some constant K (independent of n) we have (7) and this enables us to carry on the inductive process. Finally, the space X constructed for Theorem 7 is of cotype 2, as well as its dual, but it cannot be isomorphic to a Hilbert space; in fact, this space X fails the A.P. and this shows that both corollaries would be false without the A.P.

4. Upper bounds for the projection constants

Let X be a Banach space.

In this section we estimate the projection constant of an n-dimensional subspace E of X when n tends to infinity. We can define

$$\lambda_X(E) = \inf \|P\| \quad \text{and} \quad \mu_X(E) = \inf \{\gamma_2(P)\}$$

where the infimum runs over all possible projections $P: X \to E$. $\lambda_X(E)$ is called the *projection constant* of E relative to X. We have clearly $\lambda_X(E) \leqslant \mu_X(E)$.

We then let

$$e_n(X) = \sup \mu_X(E)$$

where the supremum runs over all n-dimensional subspaces $E \subset X$. For such an E, we have $d(E, l_2^n) \leqslant e_n(X)$ and there exists a projection $P: X \to E$ such that $\|P\| \leqslant e_n(X)$. The asymptotic behaviour of $e_n(X)$ when $n \to \infty$ has attracted a lot of attention in recent years. For a general space, we have $e_n(X) \leqslant \sqrt{n}$, and the "worst" cases are attained (at least asymptotically) for $X = L^1$ or $X = L^\infty$. But if a space is "far" from these extreme cases, this can be improved. In [19], Lewis proved that

$$e_n(L^p) \leqslant n^{\left|\frac{1}{p}-\frac{1}{2}\right|}$$

Following Lewis, this was generalized by many authors (Lewis, Tomczak-Jaegermann, ...). For instance, it was proved in [18] (cf. also [37] for a better proof and other results) that if X is of type $p > 1$ and of cotype $q < \infty$, then $e_n(X) \leqslant Cn^\alpha$ for some constant C and $\alpha = 1/p - 1/q$. The question whether this can be improved to $\alpha = \max(1/p - \frac{1}{2}, \frac{1}{2} - 1/q)$ (or any $\alpha < \frac{1}{2}$, when $1/p - 1/q \geqslant \frac{1}{2}$) is still open. It was (essentially) verified for Banach lattices in [27].

Although the "right" exponent is still in doubt, we *do* know that $n^{-1/2} e_n(X) \to 0$ when $n \to \infty$ iff X does not contain l_1^n's uniformly (cf. [23] and [26]), which means that $p(X) > 1$ and $q(X) < \infty$.

5. Open problems

In this section, we mention two important open questions. First, the infinite-dimensional analogue of Theorem 1 or 2 is not known:

Problem 1. Let X be an arbitrary infinite-dimensional space. Is there a bounded projection $P: X \to X$ such that both P and $I - P$ have infinite-dimensional ranges? In other words, can any X be split into a non-trivial direct sum?

For an interesting particular case, see [9], page 226. More generally, although there are spaces with few finite rank operators (cf. Section 3), it is not known whether there is a space which admits few bounded operators. Precisely, the following is open:

Problem 2. Is there an infinite-dimensional space X such that every bounded operator $u: X \to X$ is of the form $\lambda Id_X + v$ with λ scalar and v nuclear?

Actually, this is unknown even if we only ask for a compact v. A related example (X non-separable and v's of separable ranges) is constructed in [36], using special axioms.

Of course, a positive answer to Problem 2 implies a negative one to Problem 1. Moreover, a separable space X as in Problem 2 would be the first example of a separable Banach space on which every bounded operator has a non-trivial invariant subspace.

References

[1] Benyamini Y. and Gordon Y., Random factorizations of operators between Banach spaces, *J. d'Analyse Jerusalem* **39** (1981), pp. 45–75.

[2] Bourgain J., New Banach Space Properties of the Disc Algebra and H^∞, to appear in *Acta Math.*

[3] Davis W. and Johnson W., Compact, Non-Nuclear Operators, *Studia Math.* **51** (1974), pp. 81–85.

[4] Enflo P., A Counterexample to the Approximation Problem in Banach Spaces, *Acta Math.* **130** (1973), pp. 309–317.

[5] Figiel T., Lindenstrauss J. and Milman V., The Dimensions of Almost Spherical Sections of Convex Bodies, *Acta Math.* **139** (1977), pp. 53–94.

[6] Figiel T. and Tomczak-Jaegermann N., Projections onto Hilbertian Subspaces of Banach Spaces, *Israel J. Math.* **33** (1979), pp. 155–171.

[7] Gluskin E., The Diameter of the Minkowski Compactum is Roughly Equal to n, *Funct. Anal. Appl.* **15** (1981), pp. 72–73.

[8] Gluskin E., to appear.

[9] Graham C. and Mc. Gehee C., *Essays in Commutative Harmonic Analysis*, Springer Verlag, 1979.

[10] Grothendieck A., Résumé de la théorie métrique des produits tensoriels topologiques, *Bol. Soc. Matem.* Sao-Paulo **8** (1956), pp. 1–79.

[11] Grothendieck A., *Produits tensoriels topologiques et espaces nucléaires*, Memoirs of the A.M.S. **16** (1955).

[12] Hoffmann-Jørgensen J., *Probability in Banach Spaces*, Ecole d'Eté de St. Flour. VI – 1976, Springer Lecture Notes **598**.

[13] James R. C., Uniformly Non Square Banach Spaces, *Annals of Maths.* **80** (1964), pp. 542–550.

[14] John F., Extremum Problems with Inequalities as Subsidiary Conditions, *Courant Anniversary Volume*, pp. 187–204, Interscience, New-York 1948.

[15] Johnson W. and Tzafriri L., On the Local Structure of Subspaces of Banach Lattices., *Israel J. Math.* **20** (1975), pp. 292–299.

[16] Kisliakov S. V., On Spaces with "Small" Annihilators, Zap. Naucn. Sem. Leningrad. Otdel. Math. Institute Steklov (LOMI) 65 (1976), pp. 192–195 (in Russian).

[17] Krivine J. L., Sous-espaces de dimension finie des espaces de Banach réticulés, *Annals of Math.* **104** (1976), pp. 1–29.

[18] König H., Retherford J. and Tomczak-Jaegermann N., On the Eigenvalues of $(p, 2)$-Summing Operators and Constants Associated to Normed Spaces, *Journal of Funct. Analysis* **37** (1980), pp. 88–126.

[19] Lewis D., Finite Dimensional Subspaces of L_p, *Studia Math.* **63** (1978), pp. 207–212.

[20] Lindenstrauss J., The Geometric Theory of the Classical Banach Spaces, *Actes Congr. Internat. Math.* Nice, 1970, Vol. 2, pp. 365–372.

[21] Maurey B., Quelques problémes de factorisation opérateurs linéaires, *Actes. Congrès Intern. Math.* Vancouver, 1974, Vol. 2, p. 75.

[22] Maurey B. and Pisier G., Séries de variables aléatoires vectorielles indépendantes et propriétés géométriques des espaces de Banach, *Studia Math.* **58** (1976), pp. 45–90.

[23] Milman V. and Wolfson H., Minkowski Spaces with Extremal Distance from the Euclidian Space, *Israel J. Math.* **29** (1978), pp. 113–131.

[24] Pisier G., Une nouvelle classe d'espaces vérifiant le théorème de Grothendieck, *Annales de l'Inst. Fourier*, **28** (1978), pp. 69–80.

[25] Pisier G., Un théorème sur les opérateurs linéaires entre espaces de Banach qui se factorisent par un espace de Hilbert, *Annales de l'E.N.S.*, **13** (1980), pp. 23–43.

[26] Pisier G., *Sur les espaces de Banach de dimension finie à distance extrémale d'un espace euclidien, d'après V. D. Milman et H. Wolfson*, Exposé n° 16, Séminaire d'Analyse Fonctionnelle 78/79, Ecole Polytechnique–Palaiseau.

[27] Pisier G., Some applications of the complex interpolation method to Banach lattices, *Journal d'Anal. Math. Jerusalem* **35** (1979), pp. 264–281.

[28] Pisier G., Holomorphic Semi-Groups and the Geometry of Banach Spaces, *Annals of Math.* **115**, (1982), pp. 375–392.

[29] Pisier G., On the Duality Between Type and Cotype, Proceedings of a conference, *Martingale Theory in Harmonic Analysis and Banach Spaces*, Cleveland, July 1981, Lecture Notes 939, Springer.

[30] Pisier G., Counterexamples to a Conjecture of Grothendieck, *Acta Math.* **151** (1983), pp. 181–208.

[31] Szankowski S., Subspaces Without the Approximation Property, *Israel J. Math.* **30** (1978), pp. 123–129.

[32] Szarek S., The Finite Demensional Basis Problem with an Appendix on Nets of Grassmann Manifold. *Acta Math.* **151** (1983), pp. 153–180.

[33] Tomczak-Jaegermann N., On the Moduli of Smoothness and Convexity and the Rademacher Averages of the Trace Classes S_p $(1 \leqslant p < \infty)$, *Studia Math.* **50** (1974), pp. 163–182.

[34] Tzafriri L., On Banach Spaces with Unconditional Basis, *Israel J. Math.* **17** (1974), pp. 84–93.

[35] Bellenot S., Uniformly Complemented l_p^n's in Quasi-Reflexive Spaces, *Israel J. Math.* **39** (1981), pp. 234–246.

[36] Shelah S., A Banach Space with Few Operators, *Israel J. Math.* **30** (1978), pp. 181–191.

[37] Tomczak-Jaegermann N., Computing 2-Summing Norms with few Vectors, *Ark. Math.* **17** (1979), pp. 173–177.

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

DAN VOICULESCU

Hilbert Space Operators Modulo Normed Ideals

1. The developments in the K-theory of operator algebras have led to important progress in our understanding of Hilbert space operators modulo the ideal of compact operators (see [10]). It appears that further developments of the K-theory methods and of the functional analysis technique that emerged in this context may be used in the study of the more refined properties of Hilbert space operators modulo normed ideals smaller than the compacts. I will survey some of the work and problems which fit into this perspective.

2. $\mathscr{H}$ will denote a separable complex Hilbert space of infinite dimension, $\mathscr{L}(\mathscr{H})$ and $\mathscr{K}(\mathscr{H})$ the bounded and the compact operators on $\mathscr{H}$ respectively. We consider general normed ideals of compact operators $\mathfrak{S}_\Phi^{(0)}$ with the norm $|T|_\Phi = \Phi$ (eigenvalues $(T^*T)^{1/2}$), as in [12]. For $\Phi = \Phi_p(\xi_1, \xi_2, \ldots) = \left(\sum|\xi_j|^p\right)^{1/p}$ $(1 \leqslant p < \infty)$ we get the Schatten–von Neumann classes $(\mathscr{C}_p, |\cdot|_p)$ and for $\Phi = \Phi_p^-$ $(\xi_1, \xi_2, \ldots,) = \sum \xi_j^* j^{-1+1/p}$ where (ξ_j^*) is the decreasing rearrangement of $(|\xi|_j)$ we get certain Lorentz-type ideals, which we shall denote by $(\mathscr{C}_p^-, |\cdot|_p^-)$.

3. First, we shall consider questions about trivial extensions in the sense of ([5]).

Recall the classical facts about perturbations of self-adjoint operators. Let $X_1, X_2 \in \mathscr{L}(\mathscr{H})$ be self-adjoint operators without isolated eigenvalues of finite multiplicity with equal spectra $\sigma(X_1) = \sigma(X_2) = K$. Then by the theorem of Weyl and von Neumann there is a unitary U such that $UX_1U^* - X_2 \in \mathscr{K}(\mathscr{H})$. By Kuroda's theorem, $\mathscr{K}(\mathscr{H})$ may be replaced by any $\mathfrak{S}_\Phi^{(0)} \neq \mathscr{C}_1$. Further $\mathscr{K}(\mathscr{H})$ may be replaced by $\mathscr{C}_1$ if and only if the absolutely continuous parts of X_1, X_2 are unitarily equivalent (by the Kato–Rosenblum theorem).

Now, giving X_1, X_2 as above is equivalent to giving faithful $*$-homomorphisms $\varrho_j: C(K) \to \mathscr{L}(\mathscr{H})$ with $\varrho_j(C(K)) \cap \mathscr{K}(\mathscr{H}) = 0$ $(j = 1, 2)$, and

thus for such homomorphisms there is a unitary U, so that $U\varrho_1(f)U^* - \varrho_2(f) \in \mathscr{K}(\mathscr{H})$ for all $f \in C(K)$. The non-commutative Weyl–von Neumann type theorem of [20] asserts that in the preceding statement $C(K)$ may be replaced by any unital separable C^*-algebra.

For normed ideals other than $\mathscr{K}(\mathscr{H})$, note that $UX_1U^* - X_2 \in \mathfrak{S}_\Phi^{(0)}$ does not imply that $f(UX_1U^*) - f(X_2) \in \mathfrak{S}_\Phi^{(0)}$ for all $f \in C(K)$ but only for f in a dense $*$-subalgebra of $C(K)$ containing the restrictions of polynomial functions to K. Thus, in general, consider a dense $*$-subalgebra $B \subset A$, $B \ni 1$, countably generated as an algebra. The non-commutative Weyl–von Neumann type theorem extends as follows ([21]): the conclusion is that $U\varrho_1(b)U^* - \varrho_2(b) \in \mathfrak{S}_\Phi^{(0)}$ for $b \in B$ and one assumes additionally for $j = 1, 2$ that

(*) there are finite-rank R_m, $0 \leqslant R_m \leqslant I$, $R_m \uparrow I$ so that $\|[R_m, \varrho_j(b)]\|_\Phi \to 0$ as $m \to \infty$ for $b \in B$.

Condition (*) always holds when $\mathfrak{S}_\Phi^{(0)} = \mathscr{K}(\mathscr{H})$ and the role this fact plays in the proof of the non-commutative Weyl–von Neumann type theorem for $\mathscr{K}(\mathscr{H})$ was pointed out in the improved exposition of [20] given in [1].

It may seem that the above result of [21] involving condition (*) is not of much use. However, it provided the means for proving that

for every n-tuple τ with $n \geqslant 2$ of commuting self-adjoint operators there is a diagonal n-tuple δ of commuting self-adjoints such that $\tau - \delta \in \mathscr{C}_n$.

The case $n = 2$ solves a problem, attributed to Halmos, refining one of his well-known ten problems [13].

To any A, B given as above, $\varrho: A \to \mathscr{L}(\mathscr{H})$ and $\mathfrak{S}_\Phi^{(0)}$, there corresponds a central projection $E_\Phi^0(\varrho) \in \big(\varrho(A)\big)'$ which is the greatest projection of $\big(\varrho(A)\big)'$ on which ϱ satisfies (*). Then for $E_\Phi(\varrho) = I - E_\Phi^0(\varrho)$ we have:

(**) $(T_n)_1^\infty \subset \mathfrak{S}_\Phi^{(0)}$, $\|T_n\| < C$, and $\lim_{n\to\infty} \|[T_n, \varrho(b)]\|_\Phi = 0$ for all $b \in B$ $\Rightarrow$ s-$\lim_{n\to\infty} T_n E_\Phi(\varrho) = 0$.

In [21] it is shown that in the case $A = C(K)$, $K \subset \boldsymbol{R}^n$, $B =$ polynomial functions, and $\mathfrak{S}_\Phi^{(0)} = \mathscr{C}_n^-$, the decomposition $I = E_\Phi(\varrho) + E_\Phi^0(\varrho)$ is precisely the decomposition into singular and absolutely continuous parts. The proof of this depends on the asymptotic of the Fourier coefficients of $(z_1 - 1)/\big(\sum_{j=1}^n |z_j - 1|^2\big)$, viewed as a function on the n-torus. Thus $\mathscr{C}_n^-$ seems to be the right generalization to the case of n-tuples of the ideal $\mathscr{C}_1$ in the theory of perturbations of a single self-adjoint operators.

By using (**) some abstract theorems about generalized wave-operators were proved in [21]. In particular, these results have the following corollaries:

(a) For n-tuples τ, τ' of commuting self-adjoint operators with $n \geqslant 2$ such that $\tau - \tau' \in \mathscr{C}_n^-$, their absolutely continuous parts are unitarily equivalent.

(b) For n-tuples τ, τ' of commuting self-adjoint operators, with $n \geqslant 3$, $\tau - \tau' \in \mathscr{C}_n^-$ and $(f_m)_1^\infty$ $\mathscr{C}^\infty$-functions in a neighbourhood of $\sigma(\tau) \cup \sigma(\tau')$ such that $|f_m| = 1$ and $\text{w-}\lim_{m\to\infty} f_m = 0$ in $L^2(\sigma(\tau) \cup \sigma(\tau'), d\lambda)$, it follows that

$$W = \text{s-}\lim_{m\to\infty} (f_m(\tau'))^* f_m(\tau) E_{ac}(\tau)$$

exists and is independent of $(f_m)_1^\infty$.

These results are stronger than those obtained in [23] by other methods.

4. In connection with the above results let me mention the following problems:

(a) Though higher-dimensional generalizations of the Kato–Rosenblum theorem have been obtained by these techniques, it is an open problem to give a proof of the original Kato–Rosenblum theorem within this framework.

(b) For a measure μ on $\boldsymbol{R}^n$ with compact support K, $A = C(K)$, $B =$ polynomial functions, and ϱ — the representation of A on $L^2(\mu)$, the decomposition into E_Φ^0 and E_Φ corresponds to $\mu = \mu_1 + \mu_2$ with μ_1, μ_2 concentrated on disjoint Borel sets. What is this decomposition for $\mathfrak{S}_\Phi^{(0)} = \mathscr{C}_p^-$, $1 \leqslant p < n$?

5. The next situation corresponds to non-trivial extensions in te case of subsets of the plane.

There is an important work about pairs of self-adjoint operators (X_1, X_2) with $[X_1, X_2] \in \mathscr{C}_1$, due especially to J. D. Pincus, R. V. Carey, J. W. Helton, R. Howe, C. Berger and others. The Helton–Howe theorem ([14]) asserts the existence of a measure P on $\boldsymbol{R}^2$ such that

$$\text{Tr}\,[p(X_1, X_2), q(X_1, X_2)] = \int_{\boldsymbol{R}^2} \left(\frac{\partial p}{\partial x} \frac{\partial q}{\partial y} - \frac{\partial q}{\partial x} \frac{\partial p}{\partial y} \right) dP,$$

where p, q are polynomials in two variables, and $p(X_1, X_2)$, $q(X_1, X_2)$ are defined modulo $\mathscr{C}_1$. In case $X_1+iX_2-z_0I$ is Fredholm for some z_0, dP is proportional to the Lebesgue measure multiplied by the index (X_1+iX_2-zI) in some neighbourhood of z_0.

Actually, as shown in [6], $dP = g\,dx\,dy$, where $g \in L^1$ is a function introduced in another context in [18].

Note that the Helton–Howe formula provides representations of index data by traces of commutators. Even more, g seems to generalize the function $z \to \mathrm{index}(T-zI)$, which is the basic invariant in the classification of essentially normal operators. This suggests the following problem ([22]):

Is there a $\mathscr{C}_2$-analogue of the Brown–Douglas–Fillmore theorem? Explicitly, if $[X_1, X_2] \in \mathscr{C}_1$, $[X_1', X_2'] \in \mathscr{C}_1$ and $dP_{(X_1,X_2)} = dP_{(X_1',X_2')}$, do there exist N normal and U unitary so that

$$(X_1+iX_2)\oplus N - U\big((X_1'+iX_2')\oplus N\big)U^* \in \mathscr{C}_2?$$

In particular, this would imply that X_1+iX_2 is unitarily equivalent mod $\mathscr{C}_2$ with $Q\tilde{N}|Q\mathscr{H}$, where $\tilde{N}$ is normal and Q is a projection such that $[Q, \tilde{N}] \in \mathscr{C}_2$ — i.e., a Choi–Effros type result relative to $\mathscr{C}_2$.

The earliest attempt to relate pairs (X_1, X_2) as above to K-theory is perhaps [3].

6. Attempts at generalizing the Helton–Howe theorem from the lowest even dimension to higher even dimensions have led the same authors [15] to consider the more general trace-form $\mathrm{Tr}[X_1, \ldots, X_{2n}] = \mathrm{Tr}\big(\sum_{\sigma\in S_{2n}} X_{\sigma(1)}\ldots X_{\sigma(2n)}\big)$. However, in higher dimensions the abstract theory for systems $(X_1, \ldots, X_{2n})$ satisfying some conditions on the commutators has encountered serious difficulties, though in the case of pseudo-differential operators Helton and Howe obtained the beautiful formula:

$$\mathrm{Tr}[A_1, \ldots, A_{2n}] = \mathrm{const} \int_{S^*(M)} f_1 df_2 \wedge \ldots \wedge df_{2n},$$

where f_j is the symbol of A_j, M is a compact manifold and $S^*(M)$ is the co-sphere bundle.

On the abstract side, it was shown in [11] that, in the case of $(2n-1)$-dimensional sphere-extensions with commutators in $\mathscr{C}_n$, the trace-form represents the index-data associated with the extension, but it seems that in order to get a non-zero index one should allow the commutators to be in an ideal bigger than $\mathscr{C}_n$.

7. Modulo the compacts, extensions can always be replaced by "abstract elliptic operators" in the sense of Atiyah–Kasparov ([2], [16]). This means, roughly, replacing for instance an essentially commuting system of self-adjoints by a pair consisting of a commuting system and a certain element in its essential commutant. The connection between the two points of view depends on the Choi–Effros theorem.

There is an obvious refinement to normed ideals of the Atiyah–Kasparov framework and there is a general approach to trace-forms representing index-data in the context to a recent theory of Connes ([8], [9]). Connes's approach works for both K-homology groups, and for a glimpse in this direction it is convenient to switch from one K-homology group to the other. In the commutative case it corresponds to the passage from even to odd dimensions.

Let A be a C^*-algebra, $\varrho: A \to \mathscr{L}(\mathscr{H})$ a $*$-homomorphism, $\varepsilon^2 = I$ a grading of $\mathscr{H}$, i.e., $\mathscr{H} = \mathscr{H}^+ \oplus \mathscr{H}^-$, such that ϱ is of degree 0, i.e., $\varrho = \varrho^+ \oplus \oplus \varrho^-$, $F^2 = I$, $F = F^* \in \mathscr{L}(\mathscr{H})$, $\deg F = 1$, i.e., $F = \begin{bmatrix} 0 & U \\ U^* & 0 \end{bmatrix}$ with U unitary and suppose $[\varrho(a), F] \in \mathscr{K}(\mathscr{H})$ for $a \in A$. Further, let $B = \{a \in A \mid [F, \varrho(a)] \in \mathscr{C}_p\}$ and assume $\bar{B} = A$ (which implies $K_0(A) = K_0(B)$). Then there is a homomorphism $K_0(A) \to \boldsymbol{Z}$ given by

$$[P] \to \operatorname{index}_{\varrho^+(P)\mathscr{H}^+} |\varrho^+(P)\, U \varrho^-(P)|_{\varrho^-(P)\mathscr{H}^-},$$

where P is an idempotent.

If $2m \geqslant p-1$ and $b_0, \ldots, b_{2m} \in B$ let $\tau_{2m} = \text{const} \cdot \operatorname{Tr}\left(\varepsilon F [F, \varrho(b_0)] \ldots \right.$ $\left. \ldots [F, \varrho(b_{2m})]\right)$; then we have

$$\operatorname{index}_{\varrho^+(P)\mathscr{H}^+} |\varrho^+(P)\, U \varrho^-(P)|_{\varrho^-(P)\mathscr{H}^-} = \tau_{2m}(\underbrace{P, \ldots, P}_{2m+1 \text{ times}}).$$

Now for τ_{2m} one has $\tau_{2m}(b_0 b_1, b_2, \ldots, b_{2m+1}) - \tau_{2m}(b_0, b_1 b_2, b_3, \ldots) + + \tau_{2m}(b_0, b_1, b_2 b_3, \ldots) + \ldots + \tau_{2m}(b_0, \ldots, b_{2m} b_{2m+1}) - \tau_{2m}(b_{2m+1} b_0, b_1, \ldots$ $\ldots, b_{2m}) = 0$ and $\tau_{2m}(b_1, \ldots, b_{2m}, b_0) = \tau_{2m}(b_0, b_1, \ldots, b_{2m})$. Then τ_{2m} is actually a cocycle for a cohomology theory $H^*_\lambda(B)$ ([9]) having as cochains $(n+1)$-linear maps $\varphi: B \times \ldots \times B \to \boldsymbol{C}$ such that

$$\varphi(b_1, \ldots, b_n, b_0) = (-1)^n \varphi(b_0, b_1, \ldots, b_n)$$

and the coboundary map is given by $(b\varphi)(b_0, \ldots, b_{n+1}) = \varphi(b_0 b_1, b_2, \ldots$ $\ldots, b_{n+1}) - \varphi(b_0, b_1 b_2, b_3, \ldots) + \ldots + (-1)^n \varphi(b_0, \ldots, b_n b_{n+1}) + (-1)^{n+1} \times$ $\times \varphi(b_{n+1} b_0, b_1, \ldots, b_n)$. Now the cohomology theory H^*_λ is very large. This corresponds to redundancies reflecting the fact that an index expressible by a trace-form for $\mathscr{C}$ can also be expressed by a trace form for $\mathscr{C}_{n+1}, \mathscr{C}_{n+2}, \ldots$ Disposing of cup-products, it is possible to eliminate

such redundancies and it is shown in [9] that after elimination of redundancies Connes's cohomology theory for the algebra of C^∞-functions on a compact manifold corresponds to the usual de Rham homology of V.

8. Concerning the appearance of $\mathscr{C}_p$-commutators in concrete situations in analysis we should mention the recent progress which started with the work of Peller on $\mathscr{C}_p$-Hankel operators [19].

9. Among the problems appearing in the general situation we mention the problem of finding some analogue of the Choi–Effros theorem and the question originating, in the work of R. G. Douglas, of determining the smallest ideal such that a given K-homology class may be realized with commutators in that ideal.

References

[1] Arveson W. B., Notes on Extensions of C^*-Algebras, *Duke Math. J.* **44** (1977), pp. 329–355.

[2] Atiyah M. F., Global Theory of Elliptic Operators. In: *Proc. Internat. Conf. on Functional Analysis and Related Topics*, Univ. of Tokyo Press, Tokyo, 1970.

[3] Brown L.G., The Determinant Invariant for Operators with Trace Class Self-Commutators. In: *Proc. of Conf. on Operator Theory*, Lecture Notes in Mathematics **345**, Springer-Verlag, 1973, pp. 210–228.

[4] Brown L. G., Douglas R. G., and Fillmore P. A., Unitary Equivalence modulo the Compact Operators and Extensions of C^*-Algebras. In: *Proc. Conf. on Operator Theory*, Lecture Notes in Mathematics **345**, Springer-Verlag, 1973.

[5] Brown L. G., Douglas R. G., and Fillmore P. A., Extensions of C^*-Algebras and K-Homology, *Ann. of Math.* **105** (1977), pp. 265–324.

[6] Carey R. V. and Pincus J. D., Commutators, Symbols and Determining Function, *J. Functional Analysis* **19** (1975), pp. 50–80.

[7] Choi M. D. and Effros E. G., The Completely Positive Lifting Problem, *Ann. of Math.* **104** (1976), pp. 585–609.

[8] Connes A., *The Chern Character in K-Homology*, preprint IHES, 1982.

[9] Connes A., *De Rham Homology and Non-Commutative Algebra*, preprint IHES, 1983.

[10] Douglas R. G., Extensions of C^*-Algebras and Algebraic Topology. In: *Proc. Internat. Congress of Mathematicians, Helsinki, 1978.*

[11] Douglas R. G. and Voiculescu D., On the Smoothness of Sphere Extensions, *J. Operator Theory* **6** (1981), p. 103.

[12] Gohberg I. T. and Krein M. G., *Introduction to the Theory of Nonselfadjoint Operators*, Moscow, 1965 (in Russian).

[13] Halmos P. R., Ten Problems in Hilbert Space, *Bull. Amer. Math. Soc.* **76** (1970), pp. 887–933.

[14] Helton J. W. and Howe R., Integral Operators, Commutator Traces, Index and Homology, Springer Lecture Notes in Math. **345** (1973), pp. 141–209.

[15] Helton J. W. and Howe R., Traces of Commutators of Integral Operators, *Acta Math.* **136** (1976), pp. 271–365.
[16] Kasparov G. G., K-Functor and Extensions of C^*-Algebras, *Izv. Akad. Nauk SSSR, Ser. Mat.* **44** (1980), pp. 571–636.
[17] Pimsner M., Popa S., and Voiculescu D., Homogeneous C^*-Extensions of $C(X)\otimes K(H)$ *I*, *J. Operator Theory* **1** (1979), pp. 55–108; ibidem II, *J. Operator Theory* **4** (1980), pp. 210–248.
[18] Pincus J. D., Commutators and Systems of Integral Equations I, *Acta Math.* **121** (1968), pp. 219–249.
[19] Peller V. V., Hankel Operators of the Class γ_p and Their Applications, *Mat. Sbornik* **113** (1980), pp. 538–581 (in Russian).
[20] Voiculescu D., A Non-Commutative Weyl–von Neumann Theorem, *Rev. Roum. Math. Pures Appl.* **21** (1976), pp. 97–113.
[21] Voiculescu D., Some Results on Norm-Ideal Perturbations of Hilbert Space Operators, *J. Operator Theory* **2** (1979), pp. 3–37; ibidem II, *J. Operator Theory* **5** (1981), pp. 77–100.
[22] Voiculescu D., Remarks on Hilbert–Schmidt Perturbations of almost Normal Operators. In: *Topics in Modern Operator Theory*, Birkhäuser, 1981.
[23] Voigt J., Perturbation Theory for Commutative m-Tuples of Self-Adjoint Operators, *J. Functional Analysis* **25** (1977), pp. 317–334.

DEPARTMENT OF MATHEMATICS
INCREST
BD. PACII 220
BUCHAREST 79622
ROMANIA

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

DAVID R. BRILLINGER

Statistical Inference for Random Processes*

1. Introduction

Statistics is concerned with data collection, data analysis, data reduction, data modelling and inference. Its primitive concept is that of data. Statistics is part of the methodology of science — pure and applied. It is pertinent to the various goals of science proper: explanation and understanding, prediction and control, discovery and application, justification-classification. Two things at the heart of science are observation and inference. Inference may be deductive, arguing from the premises to conclusions, or what is the major process in science, inductive, intuiting from the specific to the more general.

Statistical inference is concerned with making statements that go beyond the data collected. Its traditional paradigm is that of from the sample to the population or parameter. The strength of statements made depends on the situation at hand. There are several schools of statistical inference. The schools are often in conflict; however, these days, their chosen principles are fairly clear.

By now statistics has amassed quite a collection of procedures for drawing inferences from data; however, with the passage of time, the data of concern has gotten steadily more complex. This essay is concerned with statistical inference in general and for random process data in particular. In barest detail a random process is an indexed family of random variables (or chance quantities). In operational use a random process is a random function, or random measure, or random generalized function with domain that is temporal or spatial or spatial–temporal. Its values have coordinates. Its realizations are: curves, surfaces, shapes, figures,

* Prepared with the partial support of the National Science Foundation, Grant CEE-7901642 and while the author was a Guggenheim Fellow.

sequences and the like. It relates to situations where things move and change.

We begin with an example of statistical inference for random processes taken from our own experience. The example is one with a precise experimental setup yet, apparently, inferences may not be drawn from direct examination of the data or after the realization of new experiments. Rather, a statistical concept of some subtlety is required to unravel the situation. We remark that the statistician is concerned with the probabilistic conceptualization of natural processes. At the same time he is a guardian of a collection of tools that bring order to complex data sets, tools which have had real successes. The remaining sections of the paper reflect these two aspects. Scientific investigation and modelling are discussed in general terms. Process data analysis and its aims are discussed in particular terms.

Though it is not brought out specifically in the paper, mathematics is always present for the statistician. Sometimes, especially in the theory of random processes, his work is indistinguishable from mathematics. At other times mathematics is a potent heuristic aid for planning data collection and analyzing data at hand.

2. An example

A sequence of nerve impulses, or spike train, is a common form of neurophysiological data. The times of the pulses correspond to the times at which a particular neuron fires off. The heights of the pulses are nearly constant and, provided the experimental conditions are reasonably fixed and the experiment is not continued too long, the character of the spike train is not seen to be evolving with time. It appears that this kind of data may be reasonably modelled as a piece of a realization of a stationary point process on the real line. Such a process may be defined as a random process whose realizations $N(\cdot)$ are non-negative integer-valued Borel measures on R with the (stationarity) property that the probability that $N(I_1+t) = n_1, \ldots, N(I_K+t) = n_K$ does not depend on t for I_K a Borel subset of R and $K = 1, 2, \ldots$ Suppose that the observed times of consecutive pulses, for a given spike train, are $t_1, \ldots, t_n$. Then a key role is played in the example by the empirical Fourier transform

$$d(\lambda) = \sum_{j=1}^{n} \exp\{-i\lambda t_j\} = \int_T \exp\{-i\lambda t\} N(dt),$$

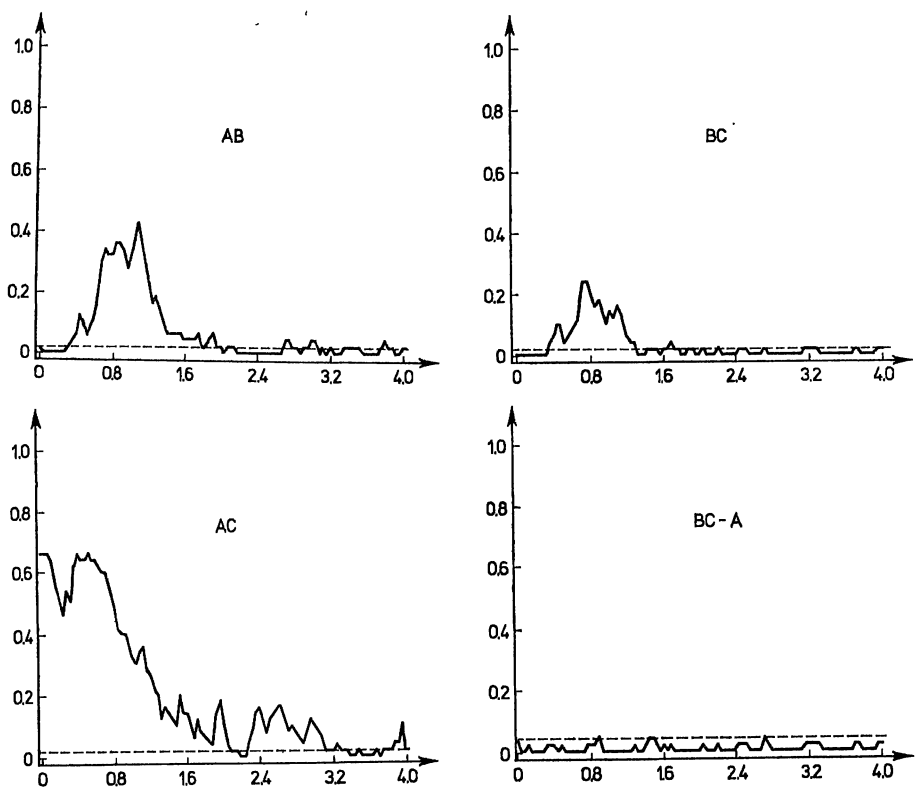
where $\lambda \in R$ and T is the observation domain.

In our example, spike trains could be recorded simultaneously for three neurons A, B, C of *Aplysia californica*. It was "known" that neuron A was driving neurons B and C. It was not known whether there was some separate connection between neurons B and C and this was the scientific question of interest. (Details may be found in Brillinger *et al.*, *Biol. Cybernetics* **22** (1976), 213–228.)

A useful statistic for measuring the degree of association (at frequency λ) of two empirical spike trains, A and B, is the sample coherency

$$\hat{R}_{AB}(\lambda) = \hat{f}_{AB}(\lambda) / \sqrt{\hat{f}_{AA}(\lambda)\hat{f}_{BB}(\lambda)}$$

where $\hat{f}_{AB}(\lambda)$ is obtained by averaging values of $d_A(\mu)\overline{d_B(\mu)}$ for μ in a neighborhood of λ. Provided the same averaging is employed in forming $\hat{f}_{AA}(\lambda)$, $\hat{f}_{BB}(\lambda)$ one has $|\hat{R}_{AB}(\lambda)|^2 \leqslant 1$, with values near 1 corresponding to strong association. The Figure shows the functions $|\hat{R}_{AB}|^2$, $|\hat{R}_{BC}|^2$, $|\hat{R}_{AC}|^2$ for one particular set of experimental data and the spike trains are indeed "found"



to be associated in pairs. The issue is whether the association of neurons B and C results totally from their both being driven by neuron A, or whether they have some association (connection) beyond that. To the extent that relationships involved are well enough captured by quadratic statistics, one can address such questions by partial coherencies.

The sample partial coherency of trains B and C given train A is

$$\hat{R}_{BC\cdot A} = (\hat{R}_{BC} - \hat{R}_{BA}\hat{R}_{AC}) / \sqrt{(1-|\hat{R}_{BA}|^2)(1-|\hat{R}_{CA}|^2)}.$$

One has $|\hat{R}_{BC\cdot A}|^2 \leqslant 1$, with values near 0 corresponding to weak association of trains B and C having "removed" the effects of train A. The Figure presents this function for the given data. There is the strongest suggestion of no direct connection between neurons B and C.

To formalize this "strongest suggestion" the 5 per cent significance line is given in each plot, as the horizontal dashed line. Were there no separate connection of B and C, the probability of this line being exceeded at a given frequency would be (approximately) 0.05.

The situation now reached is typical of what happens in science and what statistical inference has to offer. The hypothesis (of no direct connection) cannot be verified absolutely; hence it is given an opportunity to show itself false. What has happened is that the data have shown themselves compatible with the hypothesis up to the limits of the inherent variation present. Probability has been used to formalize this last.

3. Scientific investigation

In an earlier paper on our topic, (*J. Royal Statistical Society* **A 130** (1967), pp. 457–477) M. S. Bartlett sets up a "ladder diagram" of scientific enquiry of the following form:

(Theory)		(Practice)
model	↔	planning/design
deduction	→	data collection
induction	←	data analysis
new model	↔	new planning/design

Things are initiated by some idea, question or problem. Then one moves down and across the steps as work progresses. (Similar schemata have been given by G. E. P. Box, *J. American Statistical Association* **71** (1976), pp. 791–799, and H. Mohr, *Structure and Significance of Science*, Springer-Verlag (1977).) Deductions from the model play a broad role and a narrow

one. Broadly they may be predictions that science and technology use to make progress. Narrowly, they may be used just to validate the model with extant data. (Statisticians have been much concerned with this last.)

An essential feature of the whole investigative procedure is its cyclic/iterative character: ... deduction to induction to deduction to ...

4. Process data

Commonly the term process has referred to a phenomenon which showed a continuous change with time. However, the idea has been substantially abstracted with the time parameter allowed to be discrete, multidimensional, set-valued and function-valued amongst other things. Further, any requirement of continuity has been directly adapted to the situation at hand.

Process data refers to information that has been derived by observation of the process at some collection of "time" values. The information will often have numerical form; however, its values can lie in some general structured space. We shall write process data as $\{Y(t), t \in T\}$, T denoting the observation domain.

In using the term we have in mind things like: the recorded arrival times of individual photons collected by a telescope aimed in some direction, stereoscopic photographs from a distance of some land or sea surface, the collection of time series recorded at an array of sensors after a pulse of energy is input to the earth, measurements of X-ray absorption by the head as a function of the direction of a submitted X-ray beam, the distribution of earthquakes through space and time. In discussions of process data it is usual to work in situations for which the number of realizations, n, of the process $Y(t)$ is much less than the dimension, p, of the observation domain T. Multivariate data analysis, in contrast, concentrates on the case $n \gg p$.

Thanks to the dramatic advances in equipment and instrumentation during the past 30 years, researchers have effective tools for dealing with the collection of many sorts of process data, e.g. ultrafast phenomena and spatial-temporal fields. Issues arising are: data selection (auxiliary variates?), data storage (device, structure), data retrieval, data display, data auditing and flagging. Particular aspects of the process of interest affecting how this is done are: data type, data frequency content/dynamic range/information content and whether one is working in real-time

or off-line. It is clear that digital computers are important. Optical computers are now beginning to play an important role as well, e.g. in data smoothing and Fourier transforming.

As indicated in the ladder diagram, the model, deductions from the model and the design of the investigation affect data collection. We shall return to these stages later.

5. Aims of process data analysis

A time series, $Y(t)$, is a particular type of process for which t and $Y(t)$ are real-valued. J. W. Tukey, *Directions in Time Series* (Eds. D. R. Brillinger and G. C. Tiao), Institute of Mathematical Statistics (1980), has listed the following aims of time series analysis:

1. discovery of phenomena,
2. modelling,
3. preparation for further inquiry,
4. reaching conclusions in statistical terms,
5. assessment of predictability,
6. description of variability.

These apply to the general process case as well. Having in mind the great variety of process data, we may also mention: control, classification, establishing causation, description of relationship, summarization, removal of concommitant variation, measuring degree of association, signal reconstruction and enhancement, questioning conformity of theory to data, focusing information, precise measurement of constants, comparative analysis.

The neurophysiological example that we presented earlier was concerned with reaching conclusions; however, the technique employed, Fourier analysis, is well-suited to discovering unsuspected phenomena.

We have available today a broad collection of methods for meeting the aims above. Various factors enter into the choice of method for an intended analysis. One of the most important is the degree of urgency involved in the situation at hand. A second is the computing facilities available.

6. Methods for process data analysis

At the operational level the methods available for process data analysis depend upon the type of process of concern; however, there do exist a number of techniques of quite broad applicability. We shall concentrate

on these. Further any technique employed will depend intimately on the aim of the analysis.

Manipulations possible for process data depend upon the particular character of the process under study as well as the computational and instrumental facilities available. Linear forms in the data are by far the most common. They may be real-valued or function-valued. Included are Fourier and other transforms and least squares projections. In many cases they are chosen to have high information content.

It is clear that one can contemplate working with quadratic and other polynomial forms in the data. This has proved to be successful on many occasions. Great advantages of such forms are that they may be manipulated directly and that computational devices for their evaluation are often available.

The step away from polynomial forms is a long one. Experience and insight have sometimes suggested particular statistics to work with. Alternatively, models of the situation of concern have proved a rich source. We will return to the concept of model shortly.

Things computed and displayed are located at several levels.Some things are the primary goals of the work. Other things are intended to indicate the uncertainty (or instability) of those primaries. Yet other quantities are evaluated to examine and challenge assumptions (the model) that drove the analysis.

Among specific methods applicable to process data are: spectrum analysis, smoothing, inversion, likelihood, Kalman–Bucy, clustering, re-expression, dimensional reduction, contingency, analysis of variance, least squares, simulation. Specific algorithms exist for their application to many types of data. However, there are continual difficulties that arise in practice and complicate the use of the algorithms. These include: missing data, out-of-line data values, measurement error, concomitant variation, extra structure in the data, artifacts, heterogeneous data, censored data, biased collection procedure, jitter, discretization error. A broad variety of procedures now exist for dealing with these difficulties.

7. One important method

In a surprisingly large number of situations, the Fourier transform provides a meaningful method for handling process data. It is broadly defined, flexible and has useful mathematical, statistical and computational properties. We have already indicated the form of the Fourier transform

of some point process data. If instead we had planar data on a continuous process, it would take the form

$$d(\lambda_1, \lambda_2) = \iint_T Y(t_1, t_2) \exp\{-i(\lambda_1 t_1 + \lambda_2 t_2)\} dt_1 dt_2,$$

T denoting the domain of observation. In many situations it turns out to be helpful, and sometimes even crucial, to insert a convergence factor, ψ, forming for example

$$\iint \psi(t_1, t_2) Y(t_1, t_2) \exp\{-i(\lambda_1 t_1 + \lambda_2 t_2)\} dt_1 dt_2,$$

the support of ψ being contained in T, ψ being approximately 1, but tapering off to 0 as it approaches the boundary of T. The last expression extends quite directly to the case of a generalized (Schwartz–Bruhat) process over an abelian locally compact group.

It should be no surprise that the Fourier transform of process data is useful for handling convolutional relationships. (Indeed, this was one reason for its use in the example of Section 2.) It is also useful for examining a process for phenomena at "frequency" λ. One way this is done is via the periodogram, $|d(\lambda)|^2$, or some smoothed form of this last. The field of seismology provides two pertinent examples. Consider the suite of time series recorded by an array of seismometers. Following an earthquake a seismic signal may move across the array. A periodogram type analysis of this data can be used to estimate the direction of the source of the seismic energy and the velocity with which it is travelling (and this may be done for individual temporal frequency bands). By doing this analysis for successive time periods, changes in the energy source may be noted and associated phenomena viewed. Aki and Chouet, *J. Geophysics Res.* **80** (1975), pp. 3322–3342, provide an example wherein, following an explosion, Fourier analysis first shows energy coming from the appropriate direction with the expected velocity, this is then followed by energy arriving from all directions with various velocities — apparently the result of backscattering. Bolt *et al.*, *Earthquake Engineering and Struct. Dynam.* **10** (1982), pp. 561–573, provide another example of this sort of analysis. In their case, records from a nearby earthquake were processed. The apparent direction of the source of seismic energy was seen to shift with time. This may have been the first experimental measurement of a seismic dislocation moving along a rupturing fault. In each case, Fourier analysis allowed one to "discover" the presence of suspected scientific phenomena.

One tremendous statistical advantage of employing Fourier analysis

is that, in the case of a stationary process, the problem is turned into one involving independent identically distributed random variates.

8. Modelling

An ubiquitous concept in the work of statisticians (and indeed of all researchers) is that of model. A variety of meanings are attached to the word. (Some of these are reviewed by P. Suppes, *Synthese* **12** (1960), pp. 287–301.) It is often taken to mean a theory. With a model at hand, much of a researcher's work becomes deductive and manipulative. The greatest difficulties lie in creating pertinent models. Statisticians end up with a schizophrenic attitude to them. This is well illustrated by two statements of G. E. P. Box: "Statistics is or should be the art and science of building scientific models which (necessarily) involve probability.", "Since all models are wrong the scientist cannot obtain a 'correct' one by excessive elaboration."

Workers have developed a number of methods for assessing, impartially, the strength of evidence for or against a particular model, (i.e. for model validation) and for estimating the values of quantities characterizing a given model (parameters). Much work with models is concerned with investigating them theoretically and examining their goodness-of-fit empirically.

The vast majority of statistical analyses rest on a probability model of a process under investigation. Consideration of a random entity allows all of probability theory to be brought to bear on problems—in particular, for example, results concerning special random processes. In the case of a system (that is, a structure consisting of possible inputs, an operation and corresponding outputs) there now exists an immense literature concerning identification given data consisting of pieces of (process) input and corresponding pieces of (process) output. An essential practical distinction arises between situations in which the scientist can select (some of) the inputs and those where they are outside his control. Another distinction is whether the model is mechanistic (based on specific description of the natural components involved) or empirical (based on regularities that caught the researchers eye). The former is the fundamental one.

9. Statistical inference

A statistical inference is a map from data to an uncertain conclusion. The logic involved is multi-valued. The procedure is inductive. Statements made are correct only in some average sense. The statistician usually pro-

ceeds by building a chance model for the situation. Questions that arise include: is the constructed model adequate for the data? how should subjective information be incorporated? in what form should the conclusions be stated and what is then their meaning? are there important unmeasured variables? what is the goal of the work? on what should probability statements be conditional? how is a better model to be discovered? what parallel models should be considered? how should fact of preliminary analysis be incorporated? how should costs be included?

Uncertain conclusions drawn after a data analysis have various forms and levels. At one extreme one has what Mosteller and Tukey, *Data Analysis and Regression*, Addison–Wesley (1977), call a "concealed inference" wherein the data are so strong that no formalism or arithmetic are required to come to a solid conclusion. Indeed, the very goal of experimentation is to end up with such certain conclusions. At another extreme, a conclusion involves but an elementary indication of the suspected variability (stability) of some primary entity derived from the data at hand. In between one has a broad collection of inference forms and tools. We mention: tests of significance, confidence regions, likelihood graphs, posterior distributions, tolerance regions, standard errors, distance measures, prob-values, fiducial probabilities, sensitivity analyses, simulations.

One of the major contemporary works on statistical inference for random processes is that of U. Grenander, *Abstract Inference*, J. Wiley (1981). It is worth indicating some of the distinctions he recognizes and problems and procedures that he highlights. By his choice of the term "abstract inference" he deliberately leaves ambiguous whether he means the sample space (set of possible observations) or parameter space (values for quantities characterizing the probability distribution at hand) or both to be "abstract". In the work he discusses each case. For inference he employs: linear methods, likelihood based estimates and direct methods (the latter being based on common sense estimates). Classical statistical inference falls from the first two, once the appropriate structure is set up. To deal with the fact that classical procedures sometimes fail if the parameter space is too large, Grenander introduces the "method of sieves"—employing the classical procedure over a subset of the parameter space. The method is like Tihonov regularization and, for example, leads to splines in the case of nonparametric regression. Related circles of ideas include: pénalized maximum likelihood, Courant regularization, Bayesian estimation, ridge regression, and Stein estimates.

In the analysis of process data three situations, requiring different

statistical techniques, occur in practice: the signal-like situation, the noise-like situation and the mixture of signal and noise situation. In the signal-like case records for the same circumstances differ chiefly by measurement noise, e.g. images under the same conditions, identical utterances by one individual. In the noise-like case realizations have quite different appearances, e.g. the roughness of two pieces of road surface, turbulent fields generated in repetitions of an experiment. The third case is a hybrid, e.g. an earthquake recorded near a sea storm. In the signal-like case interest often is to estimate the signal. Smoothing or deconvolution operations, including regularization, may be invoked. In the noise-like case interest lies in the population from which the realization came and, for example, what may be sought is a description of the variability present or of other underlying characteristics. Difficulties arise if one uses a technique developed for one case, with another. Comparison of signals requires generalization of classical ANOVA.

So-called inverse problems fall into the signal-like case. These include the problems of computerized tomography, image reconstruction and earth modelling. They may often be formulated as: $y = X\theta + \varepsilon$, with y, 0, ε lying in abstract spaces, with X a known operator and with y also given. The problem is to estimate the signal θ. Difficulties arise because of the presence of the noise ε and because X is often unbounded. The Tihonov regularization approach chooses as estimate the value of θ minimizing $\|y - X\theta\|^2 + \alpha\|\theta\|_0^2$ for some scalar α and θ lying in some normed space. In a number of cases the estimate may be written $\theta = (X'X + \alpha A)^{-1}X'y$, for A an operator.

Photon correlation spectroscopy provides an example of a noise case where one is interested in describing the variability present. In one application, similar particles suspended in a liquid are in motion with differing velocities. It is desired to estimate the distribution of velocities. To do this, the liquid is illuminated by a laser beam. The motion of the particles induces Doppler shifts of the laser frequency, specifically the autocovariance function of the scattered light is proportional to $1 + a|b(u)|^2$ at lag u where a is a constant and $b(u) = \int [(\sin uqv)/uqv] f(v)\,dv$, $f(v)$ being the desired velocity distribution and q a known constant. The autocovariance may be estimated from a photo-multiplier record of the fluctuating light. The function $f(v)$ may be estimated by regularizaton. One reference is Frost and Cummins, *Science* **212** (1981), pp. 1520–1522. They measure sperm motility.

It seems fair to say that once a stochastic model has been set down much of the work of statistical inference proceeds in a regular manner.

The book by I. V. Basawa and B.L.S. Prakasa Rao, *Statistical Inference for Stochastic Processes*, Academic Press (1980) contains many results for a broad array of random processes. Difficulties arise on two fronts. First, many of the results are based on approximations, so they need study in any particular situation. Second, and more importantly, there is the problem of obtaining a reasonable model. In seeking a model the researcher typically turns to substantive theory and exploratory data analysis (using J. W. Tukey's term). At some point the researcher has to have an insight. This is a subconcious act and there is little likelihood that it can ever be made mechanical, but with today's marvellous visual display devices and growing collection of exploratory data tools, environments for insight can be set. Process data typically involves an element of change or movement, making visual displays especially appropriate.

10. Planning and experimental design

We conclude with a few comments on planning/design issues for process data. The distinction between experimental and observational data is crucial. (In the system case—the distinction between chosen and natural input.) The quality of inferences that may be drawn depends dramatically on which type of data is at hand. With observational data one has always to be concerned that some unsuspected or "hidden" variable was controlling the situation, not the variables that showed themselves. Through the choice of factors to vary, through the design of input, through the use of randomization a researcher can validate his statistical inferences and make efficient use of resources.

Once again many situations may be studied via the model $y = X\theta + \varepsilon$, provided one is flexible in definitions. Taking X such that $X'X$ is the identity has long been known to be an effective plan in elementary experimental design. In the case of a process system, this leads to taking as input things like: Gaussian white noise, a homogeneous Poisson, pseudorandom binary noise and a train of chirp signals. A noteworthy phenomenon is that stimuli developed for experiments in one substantive field find use in other substantive fields. We mention the chirp signal moving from radar to exploration seismology, the sinusoid moving from power engineering to laser spectroscopy, white noise moving from mechanical engineering to nuclear magnetic resonance spectroscopy. An additional benefit of employing random stimuli is that hidden variables are neutralized, as in traditional statistical experiments.

In the case of a nonlinear system, only a few input processes have been

studied extensively. N. Wiener argued for the use of Gaussian white noise in the case of polynomial systems. It has led to satisfactory results in a number of physical situations.

11. Epilogue

Taking note of the site of this Congress *and* the sitè of the next, it would be remiss not to make specific mention of Jerzy Neyman. His following words are as true today as they were some twenty years ago: "Currently, in the period of dynamic indeterminism in science, there is hardly a serious piece of research which, if treated realistically, does not involve operations on stochastic processes. The time has arrived for the theory of stochastic processes to become an item of usual equipment of every applied statistician." *J. Amer. Statist. Assoc.* **55** (1960), pp. 625–639.

THE UNIVERSITY OF CALIFORNIA
BERKELEY; U.S.A.

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

D. M. CHIBISOV

Asymptotic Expansions and Deficiencies of Tests

Introduction

In this paper a brief survey of the asymptotic theory of hypotheses testing is given and some of the author's recent results are presented. The survey is not intended to be complete; it contains mainly results related to the author's interests. A detailed review of this field can be found in Pfanzagl [30].

We adopt the approach with the probabilities of errors of first and second kind being bounded away from zero and therefore we study the power of tests against local alternatives. Special attention is paid to asymptotically efficient tests for testing a simple hypothesis about a univariate parameter. We shall consider only "regular" families for which local alternatives approach the hypothesis at a rate of $n^{-1/2}$.

1. First order asymptotic theory

In this section we present some results based on asymptotic normality, which are closely related to the subsequent higher order theory. We shall classify the results according to the following four directions:

Distributions of test statistics under the null hypothesis.

Distributions under alternatives; hence, asymptotic power, efficiency, deficiency.

Asymptotic behaviour of the likelihood ratio.

Asymptotic optimality, most powerful tests, complete classes.

1.1. There is a vast literature on asymptotic distributions under the hypothesis. A general method of proving the asymptotic normality of a statistic was to approximate it by a sum of independent random variables to which the Central Limit Theorem could be applied.

1.2. An approach which was widely used consisted in the study of the asymptotic behaviour of a test statistic for an arbitrary underlying distribution, which may correspond either to the hypothesis or to the alternative. E.g., let the underlying distribution depend on θ and test statistics T_n be used to test H_0: $\theta = \theta_0$. Suppose that, for every θ, one can prove that T_n is asymptotically normal $\mathcal{N}(\mu_n(\theta), \sigma_n^2(\theta))$. Then, for a sequence of local alternatives, $\theta_n \to \theta_0$, say, one gets $\mathcal{N}(\mu_n(\theta_n), \sigma_n^2(\theta_n))$ as an asymptotic distribution of T_n, provided the convergence is uniform in θ.

This "direct" approach may be suitable when the hypothesis plays no special role in the whole family of distributions. But even then, a more appropriate method would be to take into account the local nature of the alternatives at the very beginning. This is done in the theory based on the concept of contiguity developed by L. LeCam [23], see, e.g. the monograph by G. Roussas [32].

In particular, if Λ_n is the logarithm of the likelihood ratio (LR) of the distributions of the sample under θ_n and θ_0 and the joint distribution of (Λ_n, T_n) under $\theta = \theta_0$ converges to a limit, then the distribution of (Λ_n, T_n) under $\theta = \theta_n$ also has a limit, which is readily determined (if the limiting distributions have densities, $p_0(x, y)$ and $p_1(x, y)$, then $p_1(x, y) = e^x p_0(x, y)$).

This method proved to be particularly effective in case of rank statistics, see Hájek and Šidak [19], where the distribution of the vector of ranks under the alternative is much more complicated than under the hypothesis.

1.3. The asymptotic behaviour of the LR (or its logarithm) plays a fundamental role in this theory. In particular, it is a basis for checking the conditions of contiguity. An important class of models where this theory works are locally asymptotically normal (LAN) families of distributions.

To be specific, let $\{P_\theta, \theta \in \Theta \subset R\}$ be a family of distributions on a measurable space $(\mathcal{X}, \mathcal{A})$ having densities, p_θ, w.r.t. a σ-finite measure ν. Assuming without loss of generality that $\theta_0 = 0$, consider testing H_0: $\theta = 0$ against H_1: $\theta > 0$ based on a sample $X_1, \ldots, X_n$. Let $P_{n,\theta} = P_\theta \times \times \ldots \times P_\theta$ (n times),

$$\Lambda_{n,t} = \log \prod_{i=1}^{n} \frac{p_{tn^{-1/2}}(X_i)}{p_0(X_i)}, \quad t > 0. \tag{1.1}$$

The family $\{P_{n,\theta}, \theta \in \Theta\}$ is called LAN if there exist r.v.'s L_n and a constant $I > 0$ such that

$$\Lambda_{n,t} - (tL_n - \tfrac{1}{2}t^2 I) \xrightarrow{P_{n,0}} 0 \quad \text{for any } t > 0 \tag{1.2}$$

and L_n is asymptotically normal $\mathcal{N}(0, I)$.

In this case the distribution of $\Lambda_{n,t}$ is asymptotically normal

$$\mathcal{N}(\mp \tfrac{1}{2}t^2 I, t^2 I) \text{ under } P_{n,0} \text{ and } P_{n,tn^{-1/2}} \text{ resp.} \tag{1.3}$$

and $\{P_{n,0}\}$ and $\{P_{n,\theta_n}\}$ are contiguous for any sequence $\theta_n > 0$ such that $n^{1/2}\theta_n$ is bounded.

In case of one-parameter family, a simple sufficient condition for LAN was obtained by J. Hájek [18]: p_θ should be differentiable w.r.t. θ (see Section 3.2 for a precise formulation) and

$$I(\theta) \to I(0) > 0 \quad \text{as } \theta \downarrow 0. \tag{1.4}$$

Then (1.2) holds with $L_n = L_{n1}$ and $I = I(0)$ where $L_{n1} = n^{-1/2} \sum l_0^{(1)}(X_i)$, $I(\theta) = E_0 l_\theta^{(1)}$, $l_\theta^{(j)} = (\partial/\partial\theta)^j l_\theta$, $l_\theta = \log p_\theta$.

1.4. When (1.2) holds, L_n is an asymptotically sufficient statistic and a test based on

$$T_n = L_n + \eta_n \quad \text{with } \eta_n \xrightarrow{P_{n,0}} 0 \tag{1.5}$$

is asymptotically most powerful (AMP) against H_1: $\theta > 0$ (more precisely, it is locally AMP, see Roussas [32]).

2. Results on higher order asymptotics

Denote by $\Phi(\cdot \mid \mu, \sigma^2)$ and $\varphi(\cdot \mid \mu, \sigma^2)$ the d.f. and the density of $\mathcal{N}(\mu, \sigma^2)$, let $\Phi(\cdot) = \Phi(\cdot \mid 0, 1)$ and $\varphi(\cdot) = \varphi(\cdot \mid 0, 1)$. A sequence of d.f.'s F_n is said to admit an Edgeworth expansion of order k if

$$\sup_{x \in R} |F_n(x) - \Phi_{n,k}(x)| = o(n^{-k/2}) \tag{2.1}$$

with $\Phi_{n,k}$ of the form

$$\Phi_{n,k}(x) = \Phi(x) + \sum_{j=1}^{k} n^{-j/2} Q_j(x)\varphi(x), \tag{2.2}$$

where the Q_j are polynomials.

We classify the results according to the same four directions as in the preceding section.

2.1. The validity of Edgeworth expansions was first proved for distributions of sums of independent r.v.'s by H. Cramér, see Cramér [16]. For the further development of this theory see V. V. Petrov [26], R. N. Bhattacharya and R. R. Rao [3].

However, statistics typically arising in hypothesis testing are not exactly sums of independent r.v.'s. A possibility of approximating them by such sums cannot be used directly because the error of approximation influences the higher order terms of the expansion.

Edgeworth expansions for particular kinds of statistics were obtained, e.g., by P. L. Hsu [22] for the sample variance, by K. L. Chung [15] for Student's t, by Yu. V. Linnik and N. M. Mitrofanova [24] for maximum likelihood estimates. Behind these results for special cases lies a general method, which can be explained by the following example.

Example 2.1. Let $X_1, \ldots, X_n$ be i.i.d. r.v.'s; consider the distribution of Student's statistic,

$$t_n = n^{1/2}(\bar{X}-\mu)/s, \tag{2.3}$$

where

$$\mu = EX_1, \quad \bar{X} = n^{-1}\sum X_i, \quad s^2 = (n-1)^{-1}\sum (X_i-\bar{X})^2.$$

Without loss of generality, assume that $\mu = 0$, $\operatorname{var} X_1 = 1$. Let

$$S_{n0} = n^{-1/2}\sum X_i, \quad S_{n1} = n^{-1/2}\sum (X_i^2-1). \tag{2.4}$$

Then

$$t_n = \left(\frac{n-1}{n}\right)^{1/2} S_{n0}(1+n^{-1/2}S_{n1}-n^{-1}S_{n0}^2)^{-1/2} \tag{2.5}$$

and applying Taylor's formula we get

$$\left(\frac{n}{n-1}\right)^{1/2} t_n = S_{n0} - \tfrac{1}{2}n^{-1/2}S_{n0}S_{n1} + \tfrac{1}{8}n^{-1}(4S_{n0}^3+3S_{n0}S_{n1}^2)+\ldots \tag{2.6}$$

Now if it is required, e.g., to obtain an Edgeworth expansion of order 1, the problem reduces to obtaining this expansion for the distribution of

$$Z_n = S_{n0} - \tfrac{1}{2}n^{-1/2}S_{n0}S_{n1}. \tag{2.7}$$

Of course, one has to prove that the influence of the remaining terms on the distribution is $o(n^{-1/2})$.

This example suggests considering the following general problem. Let

$$(Y_{0i}, \mathbf{Y}_i) = (Y_{0i}, Y_{1i}, \ldots, Y_{pi}), \quad i = 1, \ldots, n, \tag{2.8}$$

be i.i.d. random vectors in R^{p+1}, and

$$S_{nl} = n^{-1/2} \sum Y_{li}, \quad l = 0, 1, \ldots, p, \quad \mathbf{S}_n = (S_{n1}, \ldots, S_{np}). \tag{2.9}$$

Given functions h_j: $R^{p+1} \to R^1$, $j = 1, \ldots, J$, it is required to obtain an Edgeworth expansion for the distribution of

$$Z_n = S_{n0} + \sum_{j=1}^{J} n^{-j/2} h_j(S_{n0}, \mathbf{S}_n). \tag{2.10}$$

Usually the h_j are polynomials and J equals k, the required order of the expansion. However, sometimes some additional terms have to be considered in order to justify the approximation of the original statistic by Z_n, which is called a stochastic expansion.

The following theorem was proved by Chibisov [7] with a stronger condition (ii). In the present form it was obtained by Pfanzagl [27] (where in fact a more general result is proved in which, in particular, Z_n is multivariate).

THEOREM 2.1 *Suppose that the h_j in* (2.10) *are polynomials, and*

(i) $EY_{l1} = 0$, $EY_{01}^2 = 1$, $E|Y_{l1}|^{k+2} < \infty$, $l = 0, 1, \ldots, p$, *for an integer* $k \geqslant 1$.

(ii) *Cramér's condition* C: *For any* $a > 0$ *there exists* $0 < \eta < 1$ *such that*

$$\sup_{\|s,t\| > a} |f(s, t)| \leqslant \eta \tag{2.11}$$

where $f(s, t)$, $t = (t_1, \ldots, t_p)$ *is the characteristic function (ch.f.) of* $(Y_{01}, \mathbf{Y}_1)$

Then the d.f. of Z_n *admits an Edgeworth expansion of order* k.

In Chibisov's paper [7] a convenient formal rule was also given for obtaining the explicit form of the polynomials Q_j in (2.2) from the cumulants of Y's and the coefficients of h's in (2.10).

Theorem 2.1 gave a method of obtaining Edgeworth expansions for a large class of statistics arising in parametric problems of hypotheses testing and estimation, see Pfanzagl [30]. In that paper one can also find a brief review of results obtainable by other methods. We mention here some papers where Edgeworth expansions were obtained: Albers, Bickel, Van Zwet [1]; Bickel, Van Zwet [4], Does [17] — for linear rank

statistics; Helmers [20] — for linear combinations of order statistics, and Callaert, Janssen and Veraverbeke [6] — for U-statistics.

2.2. In a manner very similar to the "direct" approach described in Section 1.2, the same methods were applied to obtain expansions under alternatives. Besides providing better approximations to be used for numerical purposes, these expansions were applied to obtaining asymptotic expansions for power functions of tests. These expansions are of particular interest in case of AMP tests which have the same limiting power and may be distinguished from one another by higher order terms.

In the problem considered in Section 1.3, let $\varphi_{n,t} = \varphi_{n,t}(X_1, \ldots, X_n)$ be the critical function of a size α LR test based on $\Lambda_{n,t}$, $t > 0$, i.e. for a suitable constant $c_{n,t}$:

$$\varphi_{n,t} = \begin{cases} 1, & \Lambda_{n,t} > c_{n,t}, \\ 0, & \Lambda_{n,t} < c_{n,t}, \end{cases} \qquad E_{n,0}\varphi_{n,t} = \alpha. \tag{2.12}$$

Let

$$\beta_{n,t}(\theta) = E_{n,\theta}\varphi_{n,t}, \qquad \bar{\beta}_n(\theta) = \beta_{n,\theta n^{1/2}}(\theta); \tag{2.13}$$

then $\beta_{n,t}(\theta)$ is the power of $\varphi_{n,t}$ for an arbitrary $\theta > 0$ and $\bar{\beta}_n(tn^{-1/2})$ is the power of $\varphi_{n,t}$ for the corresponding alternative $\theta = tn^{-1/2}$. Thus, by the Neyman–Pearson lemma, we see that for each $\theta > 0$, $\bar{\beta}_n(\theta)$ is the maximal power which can be attained by size α tests (envelope power function).

For $\bar{\beta}_n$ an asymptotic expansion was obtained,

$$\bar{\beta}_n(tn^{-1/2}) = \beta_t + n^{-1/2}\,\bar{R}_{1t} + n^{-1}\,\bar{R}_{2t} + o(n^{-1}) \tag{2.14}$$

where β_t is determined by (1.3) with $I = I(0)$; hence

$$\beta_t = 1 - \Phi(c_t \mid \tfrac{1}{2}t^2 I, t^2 I) = \Phi(tI^{1/2} - u_\alpha) \tag{2.15}$$

with

$$c_t = \lim_{n\to\infty} c_{n,t} = -\tfrac{1}{2}t^2 I + u_\alpha t I^{1/2}, \qquad u_\alpha = \Phi^{-1}(1-\alpha), \tag{2.16}$$

and $\bar{R}_{mt}$ are products of $\varphi(c_t \mid \tfrac{1}{2}t^2 I, t^2 I) = (tI^{1/2})^{-1}\varphi(tI^{1/2} - u_\alpha)$ and polynomials in t whose coefficients depend on the moments of $l_0^{(j)}$, $j = 1, \ldots, m$.

Moreover, expansions for the powers, $\beta_n(\theta)$, of various AMP tests were obtained,

$$\beta_n(tn^{-1/2}) = \beta_t + n^{-1/2}R_{1t} + n^{-1}R_{2t} + o(n^{-1}). \tag{2.17}$$

It was found that typically, under general regularity conditions, $R_{1t} = \bar{R}_{1t}$, i.e. "first order efficiency implies second order efficiency", see Pfanzagl [29]. It was shown in Bickel, Chibisov, Van Zwet [5] that this property holds under very general conditions.

Thus, when the expansions (2.14), (2.17) hold, then the power, β_n, of an AMP test differs from the largest possible one by an amount of order n^{-1}. This fact is connected with the concept of deficiency describing the performance of a test in terms of the number of observations needed to attain prescribed errors of 1st and 2nd kind, α and ω.

Namely, given an $\omega > 0$ such that $1-\omega > \alpha$, define θ_n by $\beta_n(\theta_n) = 1-\omega$ and m_n by

$$\beta_{m_n}(\theta_n) \leqslant 1-\omega \leqslant \beta_{m_n+1}(\theta_n);$$

let

$$\gamma_n = \frac{1-\omega-\beta_{m_n}(\theta_n)}{\beta_{m_n+1}(\theta_n)-\beta_{m_n}(\theta_n)}.$$

Then the randomized test using m_n or m_n+1 observations with probabilities $1-\gamma_n$ and γ_n has the power of $1-\omega$. Let us say that the test requires $\overline{m}_n = m_n+\gamma_n$ observations. AMP tests are asymptotically Pitman efficient, i.e. $\overline{m}_n/n \to 1$ as $n \to \infty$.

It was proposed by Hodges, Lehmann [21] to consider in this case the difference $d_n = \overline{m}_n - n$, which was called the deficiency of the test; its limit, $d = \lim_{n\to\infty} d_n$, if it exists, is called the asymptotic deficiency.

If (2.14) and (2.17) hold then the asymptotic deficiency is finite and equals

$$d = \frac{2r_{t(\alpha,\omega)}}{(u_\alpha+u_\omega)^2\varphi(u_\omega)}, \tag{2.18}$$

where $r_t = \bar{R}_{2t} - R_{2t}$ and $t(\alpha,\omega) = (u_\alpha+u_\omega)/I^{1/2}$ (this is the value of t for which $\beta_t = 1-\omega$).

Example 2.2. For $t_0 > 0$, consider the test φ_{n,t_0}. Denote by $R_{2,t_0,t}$ the coefficient at n^{-1} in the expansion (2.17) of $\beta_{n,t_0}(tn^{-1/2})$. The difference $r_{t_0,t} = \bar{R}_{2,t} - R_{2,t_0,t}$ was written out in Chibisov's paper [9]. This difference and the asymptotic deficiency, d_{t_0}, of φ_{n,t_0} are

$$r_{t_0,t} = \frac{D_{t_0,t}}{2tI^{1/2}}\varphi(tI^{1/2}-u_\alpha), \qquad d_{t_0} = \frac{D_{t_0,t(\alpha,\omega)}}{(u_\alpha+u_\omega)^2}, \tag{2.19}$$

where

$$D_{t_0,t} = \tfrac{1}{4}t^2(t-t_0)^2[\mathrm{var}_0 l_0^{(2)} - \mathrm{cov}_0^2(l_0^{(1)}, l_0^{(2)})/I]. \tag{2.20}$$

These expressions with $t_0 = 0$ correspond to the test based on L_{n1} which will be called $\varphi_{n,0}$. In a certain sense it is a limit of φ_{n,t_0} as $t_0 \to 0$.

As a rule, in the papers cited in Section 2.1, where Edgeworth expansions for distributions of test statistics were obtained, they were used to obtain asymptotic expansions for the power and asymptotic deficiencies.

The case where nuisance parameters are present was studied by Pfanzagl, Wefelmeyer [31], see also Pfanzagl [30].

2.3. In the papers cited above $\Lambda_{n,t}$ was approximated by a stochastic expansion by writing it as $\Lambda_{n,t} = \sum [l_{tn^{-1/2}}(X_i) - l_0(X_i)]$ and applying Taylor's formula in terms of $tn^{-1/2}$. The resulting stochastic expansion contains sums

$$L_{nj} = n^{-1/2} \sum [l_0^{(j)}(X_i) - a_j], \qquad a_j = E_0 l_0^{(j)}, \tag{2.21}$$

with factors $n^{-j/2}$; then Theorem 2.1 is applied, see, e.g. Pfanzagl [29].

Some other methods were used by Chibisov [9] and Albers, Bickel, Van Zwet [1] to obtain an Edgeworth expansion for the d.f. of $\Lambda_{n,t}$ under weaker conditions.

2.4. Now, AMP tests can be compared to each other by considering the higher order terms of their power or, equivalently, their deficiencies and one may ask whether there exists an asymptotically (up to $o(n^{-1})$) most powerful test and, if not, whether one can find a sufficiently small asymptotically complete class. An answer was given by Pfanzagl [28] who showed that, under suitable regularity conditions, the family of tests, $\{\varphi_{n,t_0}, t_0 \geqslant 0\}$ forms an asymptotically complete class. This means that for any sequence of size α tests having powers $\beta_n(\theta)$, there exists a sequence $t_n \geqslant 0$ such that

$$\beta_n(tn^{-1/2}) \leqslant \beta_{n,t_n}(tn^{-1/2}) + o(n^{-1}) \quad \text{for all } t > 0. \tag{2.22}$$

The formulas (2.19) and (2.20) show that the LR tests for different t_0 do not dominate each other and their powers differ by terms of order n^{-1}, unless the expression in brackets in (2.20) vanishes.

It does so when $\{l_\theta\}$ is an exponential family because then a uniformly most powerful test exists and $\beta_{n,t_0}(\theta)$ does not depend on $t_0 \geqslant 0$. On the other hand, it can be shown that if

$$\operatorname{var}_\theta l_\theta^{(2)} - \operatorname{cov}_\theta^2(l_\theta^{(1)}, l_\theta^{(2)})/I(\theta) \equiv 0$$

for θ running over a nondegenerate interval then $\{P_\theta\}$ is an exponential family.

In the case where nuisance parameters are present a theorem describing an asymptotically complete class was obtained by Pfanzagl and Wefelmeyer [31]. Such a class again consists of LR tests. However, its completeness is asserted only within the class of tests based on statistics admitting a stochastic expansion (see also Pfanzagl [30] for a detailed discussion).

3. Recent results

3.1. In the papers quoted in Section 2.1 the stochastic expansion (2.10) was treated as a transform of the vector of normalized sums, (S_{n_0}, S_n), whose distribution admits a multivariate Edgeworth expansion. In order to obtain this expansion one has to impose the same moment conditions on all components of the summands (see condition (i) of Theorem 2.1). Since the components of S_n enter into terms having factors $n^{-j/2}$, $j \geqslant 1$, one may expect that the conditions on the corresponding summands can be reduced. The following result taking this into account was obtained by Chibisov [10].

Let h's in (2.10) be polynomials and denote by M_j, $j = 1, 2, \ldots, J$, the set of those $\boldsymbol{m} = (m_1, \ldots, m_p)$ for which there exists a monomial $x_0^{m_0} x_1^{m_1} \ldots x_p^{m_p}$ entering $h_j(x_0, \boldsymbol{x})$ with a non-zero coefficient.

THEOREM 3.1 *Let the following conditions be fulfilled*:

(i) $E\,Y_{01} = 0$, $E\,Y_{01}^2 = 1$, $E\,|Y_{01}|^{k+2} < \infty$.

(ii) *There exist* $r_1, \ldots, r_p$ *such that* $0 < r_l \leqslant k+2$, $E\,|Y_{l1}|^{r_l} < \infty$ *for* $l = 1, \ldots, p$, *and*

$$\sum_{l=1}^{p} m_l \left(\frac{k+2}{r_l} - 1 \right) \leqslant j \quad \text{for all } \boldsymbol{m} \in M_j,\ j = 1, \ldots, k. \tag{3.1}$$

(iii) *Cramér's condition* C, *see* (2.11).

Then for the d.f. of Z_n *an Edgeworth expansion of order* k *holds.*

Example 3.1. Suppose that we want to obtain an Edgeworth expansion of order 1 for the d.f. of Student's t-statistic, see example 2.1. In Theorem 2.1 the existence of 3rd absolute moments of all summands in (2.4) is required which results in assuming that $EX_1^6 < \infty$. To apply Theorem 3.1 to Z_n in (2.7), it is sufficient to require that $E\,|X_1|^3 < \infty$ and Cramér's condition C on (X_1, X_1^2) is satisfied. Condition (ii) is then fulfilled with $r_1 = 3/2$, $m_1 = j = 1$. The approximation of t_n by Z_n may be justified

under the same conditions, i.e. the expansion given by Theorem 3.1 holds for the d.f. of t_n itself.

3.2. It will be convenient for us to interchange the order of presentation and to consider the results on the LR in this section. These results were obtained by the author jointly with W. R. van Zwet.

Their short formulations are contained in the communications by Chibisov, Van Zwet [14], Chibisov [11], [12]. These results provide Edgeworth expansions for the d.f. of $\Lambda_{n,t}$, see (1.1), under $P_{n,0}$ and $P_{n,tn^{-1/2}}$. They allow one to reduce the regularity conditions required to obtain these expansions and the expansion (2.14) for β_n as compared to the papers mentioned in Section 2.3. In particular, the set $\{x: p_\theta(x) > 0\}$ may depend on θ and no Cramér type condition on the joint distribution of $(l_0^{(1)}, l_0^{(2)}, \ldots)$ is imposed.

CONDITION B_r ($r \geqslant 2$, integer). For any $x \in X$, $p_\theta(x)$ is absolutely continuous in $\theta \in U$ for a neighbourhood $U = [0, \varepsilon]$; for any $\theta \in U$, the derivative $p_\theta^{(1)}(x) = (\partial/\partial\theta)p_\theta(x)$ exists for ν-almost all $x \in X$;

$$0 < \limsup_{\theta \downarrow 0} E_\theta |l_\theta^{(1)}|^r = E_0 |l_0^{(1)}|^r < \infty. \tag{3.2}$$

Note that B_2 is exactly Hájek's condition, see (1.4).

Let $h_\theta = \theta^{-1}(l_\theta - l_0)$, $\theta > 0$. Then (see (1.1))

$$\Lambda_{n,t} = tn^{-1/2} \sum h_{tn^{-1/2}}(X_i)$$

is a normalized sum of i.i.d. r.v.'s whose distribution depends on n. Define l_θ to be $-\infty$ when $p_\theta = 0$. Then under the condition B_r h_θ may be infinite with a positive probability, and any moment conditions can be satisfied only for appropriate truncations of h_θ.

THEOREM 3.2 *Under the condition* B_r *there exists* $\tilde{h}_\theta$ *such that* $P_0\{\tilde{h}_\theta \neq h_\theta\} = o(\theta^r)$ *and* $\tilde{h}_{\theta_n}^r$ *is uniformly integrable w.r.t.* P_0 *for any sequence* $\theta_n \downarrow 0$.

The following condition may be regarded as an asymptotic Cramér type condition. Denote the distribution of h_θ under H_0 by G_θ, i.e. $G_\theta(A) = P_0(h_\theta \in A)$ for Borel sets $A \in R$.

CONDITION AC. There exist $-\infty < b_1 < b_2 < \infty$, positive ε, c, M, $0 \leqslant \gamma < 2$ and, for each $\theta \in U = [0, \varepsilon]$, a number $0 < \alpha_\theta < 1$, a distribution $G_{\theta,1}$ and a measure $G_{\theta,2}$ on R ($G_{\theta,2}(R) \leqslant 1$) such that

$$G_\theta = \alpha_\theta G_{\theta,1} + (1 - \alpha_\theta) G_{\theta,2}, \qquad \alpha_\theta \geqslant c\theta^\gamma,$$

$G_{0,1}$ has a density $g_{0,1}$ w.r.t. the Lebesgue measure such that $g_{0,1} \leqslant M$ and $g_{0,1} = 0$ outside $[b_1, b_2]$.

Example 3.2. Let $p_0(x) = \frac{1}{2}e^{-|x-0|}$, $x \in R$. Then the Condition AC is satisfied with $\alpha_\theta \sim \theta/2$ as $\theta \downarrow 0$.

Denote by $\Phi_{n,k}(x, \{\chi_j\})$ the Edgeworth expansion of order k for the d.f. of a normalized sum of i.i.d. r.v.'s having cumulants $\{\chi_j\}$; it is of form (2.1) with $Q_j\varphi = P_j(-\Phi: \{\chi_j\})$, see Bhattacharya, Rao [3].

For $\tilde{h}_0$ of Theorem 3.2 let $\tilde{\mu}_j(\theta) = E_0\tilde{h}_\theta^j$, $j = 1, 2, \ldots$, and $\tilde{\chi}_j(\theta)$ be corresponding cumulants. Write $\tilde{\sigma}^2$ for $\tilde{\chi}_2$. Theorem 3.2 ensures that $\tilde{\mu}_j(\theta)$, $\tilde{\chi}_j(\theta)$ exist for $j = 1, \ldots, r$ under the condition B_r.

The following theorem follows immediately from Theorem 1 of Chapter VI of Petrov [26] and Theorem 3.2.

THEOREM 3.3. *If the conditions* B_{k+2} *and* AC *are fulfilled then*

$$\sup_{x \in R} \left| P_{n,0}\left(\frac{\Lambda_{n,t} - \tilde{\mu}_1(tn^{-1/2})n^{1/2}}{\tilde{\sigma}(tn^{-1/2})}\right) - \Phi_{n,k}\left(x, \{\tilde{\chi}_j(tn^{-1/2})\}\right)\right| = o(n^{-k/2}). \tag{3.3}$$

Using the relation

$$dP_{n,tn^{-1/2}}(\Lambda_{n,t} < x) = e^x dP_{n,0}(\Lambda_{n,t} < x)$$

one can deduce from (3.3) an asymptotic expansion for the d.f. of $\Lambda_{n,t}$ under $P_{n,tn^{-1/2}}$. In this context the following theorem is useful.

THEOREM 3.4. *Let* $E_\theta \in \mathcal{A}$, $\theta > 0$, *be sets such that* $P_0(E_\theta) = o(\theta^r)$ *and the condition* B_r *be satisfied. Then* $P_\theta(E_\theta) = o(\theta^r)$.

Theorem 3.2 gives no constructive description of $\tilde{h}_\theta$, hence, no method of calculating the $\tilde{\chi}$'s entering into (3.3). The following condition allows us to calculate the moments of $\tilde{h}_{tn^{-1/2}}$ by formally using the Taylor expansion

$$\tilde{h}_{tn^{-1/2}} \approx l_0^{(1)} + \tfrac{1}{2}tn^{-1/2}l_0^{(2)} + \ldots$$

CONDITION $B^{(r)}$. For any $x \in X$, $p_\theta(x)$ is $r-1$ times continuously differentiable in $\theta \in U$ and $p_\theta^{(r-1)}(x)$ is absolutely continuous in $\theta \in U$; for any $\theta \in U$, $p_\theta^{(r)}(x) = (\partial/\partial\theta)p_\theta^{(r-1)}(x)$ exists for almost all $x \in X$;

$$E_0|p_0^{(j)}/p_0|^{r/j} < \infty \quad \text{for } j = 1, \ldots, r;$$

$$\limsup_{\theta \downarrow 0} E_0|p_\theta^{(r)}/p_\theta| \leqslant E_0|p_0^{(r)}/p_0|.$$

The following Theorem 3.5 holds true also under a similar condition

with the last relations replaced by

$$E_0|l_0^{(j)}|^{r/j} < \infty \quad \text{for } j = 1, \ldots, r;$$

$$\limsup_{\theta \downarrow 0} E_\theta |l_\theta^{(r)}| \leqslant E_0 |l_0^{(r)}|.$$

Let

$$\bar{h}_\theta = \sum_{j=1}^{r} \frac{\theta^{j-1}}{j!} l_0^{(j)}, \quad \bar{\mu}_j(\theta) = E_0[\bar{h}_\theta^j]_{r-j}$$

where for a polynomial $H(\theta) = \sum c_l \theta^l$ and $a > 0$ we write $[H(\theta)]_a = \sum_{l \leqslant a} c_l \theta^l$. By Hölder's inequality $\bar{\mu}_j(\theta)$ exist for $j = 1, \ldots, r$ under $\mathrm{B}^{(r)}$.

THEOREM 3.5. *Under the Condition* $\mathrm{B}^{(r)}$ *there exists* $\tilde{h}_\theta$ *for which the assertion of Theorem* 3.2 *holds and*

$$\tilde{\mu}_j(\theta) = \bar{\mu}_j(\theta) + o(\theta^{r-j}), \quad j = 1, \ldots, r. \tag{3.4}$$

Denote by $\bar{\chi}_j(\theta)$ the "cumulants" corresponding to the "moments" $\bar{\mu}_j(\theta)$, i.e. the expressions which are obtained from $\{\bar{\mu}_j(\theta)\}$ by formal relations between moments and cumulants. It follows from the structure of Edgeworth expansions that (3.4) with $r = k+2$ implies

$$\sup_{x \in R} |\Phi_{n,k}(x, \{\tilde{\chi}_j(tn^{-1/2})\}) - \Phi_{n,k}(x, \{\bar{\chi}_j(tn^{-1/2})\})| = o(n^{-k/2}).$$

Therefore under the conditions $\mathrm{B}^{(k+2)}$ and AC, (3.3) holds with $\bar{\chi}$'s instead of $\tilde{\chi}$'s.

The following theorem provides a sufficient condition for AC when $\mathscr{X} = R$ and ν is the Lebesgue measure.

THEOREM 3.6. *Suppose there exist* $-\infty < a_1 < a_2 < \infty$, $\delta > 0$, $\varepsilon > 0$ *and* $C > 0$ *such that*

(I) $p_0(x) > 0$ *for* $x \in [a_1, a_2]$;

(II) $\liminf_{\theta \downarrow 0} p_\theta(x) \geqslant p_0(x)$ *for* ν*-almost all* $x \in [a_1, a_2]$;

(III) *The derivatives* $p_0^{(1)}(a_i)$, $i = 1, 2$, *exist and* $l_0^{(1)}(a_1) \neq l_0^{(1)}(a_2)$;

(IV) *For any* $\theta \in U = [0, \varepsilon]$, $p_\theta(x)$ *is absolutely continuous in* $x \in [a_1, a_2]$; *let* $p'_\theta = \partial p_\theta / \partial x$;

(V) $\int_{a_1}^{a_2} |p'_\theta(x)|^{2+\delta} dx \leqslant C$ *for* $\theta \in U$.

Then the Condition AC *is fulfilled.*

3.3. Turning attention to asymptotic powers and deficiencies of tests we present in this section a result which describes directly the difference, $\bar{\beta}_n - \beta_n$, between the envelope power function and the power of an AMP test without obtaining the expansions (2.14) and (2.17) for them. Namely, it provides a formula for

$$r_t = \lim_{n\to\infty} n(\bar{\beta}_n(tn^{-1/2}) - \beta_n(tn^{-1/2})). \tag{3.5}$$

Using this result, the asymptotic expansion for β_n can be obtained by writing out, first, the expansion for $\bar{\beta}_n$ (see Section 3.2) which is completely determined by the family $\{P_\theta\}$ and then introducing the correction (3.5), which depends on the test under consideration. Moreover, this result immediately provides the asymptotic deficiency by (2.18).

Given $t > 0$, the test statistic, say, T_n, of an AMP test may be typically transformed by a linear function into a r.v. $Z_{n,t}$ which is close to $\Delta_{n,t}$, so that

$$Z_{n,t} = \Delta_{n,t} + \Delta_{n,t}, \quad \text{where } \Delta_{n,t} \xrightarrow{P_{n,0}} 0. \tag{3.6}$$

In case of (1.5) one may set $Z_{n,t} = tT_n - \frac{1}{2}t^2 I$ (cf. (1.2)). Note that this transformation does not influence the test function, and hence, the power. In connection with Example 2.2 consider

Example 3.3. Writing out the Taylor expansion of $\Delta_{n,t}$ as described in Section 2.3 and noting that $a_2 = -I$ in "regular" cases (see (2.21) for the notation) we have

$$\Delta_{n,t} = tL_{n1} - \tfrac{1}{2}t^2 I + n^{-1/2}(\tfrac{1}{2}t^2 L_{n2} + \tfrac{1}{6}t^3 a_3) + \dots \tag{3.7}$$

In order to consider the power of φ_{n,t_0} at $\theta = tn^{-1/2}$, we introduce

$$Z_{n,t_0,t} = \frac{t}{t_0}\Delta_{n,t_0} + \tfrac{1}{2}t(t_0 - t)I. \tag{3.8}$$

Then (3.6) holds with

$$\Delta_{n,t_0,t} = Z_{n,t_0,t} - \Delta_{n,t} = n^{-1/2}[\tfrac{1}{2}t(t_0 - t)L_{n2} + \tfrac{1}{6}t(t_0^2 - t^2)a_3] + \dots \tag{3.9}$$

In the theorem to be stated we consider r_t given by (3.5) with $\bar{\beta}_n$ defined by (2.14) and $\beta_n(tn^{-1/2}) = E_{n,tn^{-1/2}}\psi_{n,t}$, the power of a test $\psi_{n,t}$ such that

$$\psi_{n,t} = \begin{cases} 1, & Z_{n,t} > b_{n,t}, \\ 0, & Z_{n,t} < b_{n,t}, \end{cases} \qquad E_{n,0}\psi_{n,t} = \alpha. \tag{3.10}$$

Assume that there exist functions $\boldsymbol{y} = (y_1, \ldots, y_p)$: $\mathscr{X} \to R^p$, H_t: $R^{q+1} \to R^1$, $q \leqslant p$, and $K_{n,t}$: $R^{p+1} \to R$ such that

$$\Delta_{n,t} = n^{-1/2} H_t(\Delta_{n,t}, \boldsymbol{S}_n) + K_{n,t}(\Delta_{n,t}, \boldsymbol{S}_{n1}), \tag{3.11}$$

where $\boldsymbol{S}_n = (S_{n1}, \ldots, S_{nq})$, $\boldsymbol{S}_{n1} = (S_{n1}, \ldots, S_{np})$, $S_{nl} = n^{-1/2} \sum y_l(X_i)$, $l = 1, \ldots, p$.

Under the conditions of the theorem, the distribution of $(\Delta_{n,t}, \boldsymbol{S}_n)$ converges to a normal one; denote by $(\Delta_t, \boldsymbol{S})$ a random vector in R^{q+1} having this limiting distribution.

CONDITION M. There exist polynomials $h_{t,1}(x, \boldsymbol{u})$, $(x, \boldsymbol{u}) \in R^{q+1}$, and $h_{t,j}(x, \boldsymbol{w})$, $(x, \boldsymbol{w}) \in R^{p+1}$, $j = 2, \ldots, J$, such that

$$|H_t(x, \boldsymbol{u})| \leqslant h_{t,1}(|x|, |\boldsymbol{u}|), \qquad |K_{n,t}(x, \boldsymbol{w})| \leqslant \sum_{j=2}^{J} n^{-j/2} h_{t,j}(|x|, |\boldsymbol{w}|),$$

where $|\boldsymbol{u}| = (|u_1|, \ldots, |u_q|)$, $|\boldsymbol{w}| = (|w_1|, \ldots, |w_p|)$.

Denote by M_j, $j = 1, 2, \ldots, J$, the set of those $\boldsymbol{m} = (m_1, \ldots, m_p)$ ($\boldsymbol{m} = (m_1, \ldots, m_q)$ for $j = 1$) for which there is a monomial $x^{m_0} w_1^{m_1} \ldots w_p^{m_p}$ ($x^{m_0} u_1^{m_1} \ldots u_q^{m_q}$ for $j = 1$) in $h_{t,j}(x, \boldsymbol{m})$, $j = 2, \ldots, J$ ($h_{t1}(x, \boldsymbol{u})$), with a non-zero coefficient.

THEOREM 3.7. *Let the conditions* B_4, AC *and* M *hold; let* $H_t(\cdot)$ *be continuous on* R^{q+1}; *there exist* $r_1, \ldots, r_p$, $0 < r_l \leqslant 4$, *such that* $E_0 |y_l(X_1)|^{r_l} < \infty$, $l = 1, \ldots, p$, *and*

$$\sum_{l=1}^{p} m_l \left(\frac{4}{r_l} - 1 \right) \leqslant j \quad \text{for all } \boldsymbol{m} \in M_j,\ j = 1, 2, \ldots, J.$$

Then

$$r_t = \tfrac{1}{2} \varphi(c_t \mid \tfrac{1}{2} t^2 I, t^2 I) D_t, \tag{3.12}$$

where c_t *is defined by* (2.16) *and*

$$D_t = \operatorname{var}[H_t(\Delta_t, \boldsymbol{S}) \mid \Delta_t = c_t]. \tag{3.13}$$

By (2.18) and (3.12) we immediately get

$$d = \frac{D_{t(a,\omega)}}{(u_a + u_\omega)^2}. \tag{3.14}$$

Applying this result to Examples 2.2 and 3.3, we see that (2.19), (2.20) could be obtained at once from (3.9) and (3.12)–(3.14).

The formulation of Theorem 3.7 has been published in Chibisov's notes [11]–[13], in Chibisov's note [12] a sketch of the proof is also given.

The method of the proof of Theorem 3.7 carries over LeCam's approach described in Section 1.2 to higher order asymptotics. The proof does not involve asymptotic expansions, which makes the method applicable to various other problems, including those where no methods of obtaining asymptotic expansions are known. The result (3.12), (3.14) then holds with

$$D_t = \operatorname{var}[H_t \mid \Delta_t = c_t], \tag{3.15}$$

where (Δ_t, H_t) is a random vector whose distribution is the limiting one for $(\Delta_{n,t}, n^{1/2}\Delta_{n,t})$. The results of this kind were obtained by V. E. Bening [2] for one-sample rank tests and linear combinations of order statistics and by V. K. Malinovskii [25] for some AMP tests in case of observations which form a Markov chain.

3.4. The result by Pfanzagl on the asymptotically complete class quoted in Section 2.4 has a final form; one can only reduce the regularity conditions using the results of Section 3.2. However it is easy to show that the class $\{\varphi_{n,t_0}, t_0 \geqslant 0\}$ is asymptotically complete within the class of AMP tests based on statistics (1.4) for which (3.12) with D_t given by (3.15) holds. This may serve as an explanation of Pfanzagl's result.

Let a test be based on T_n from (1.4) and assume that the joint distribution of $(L_{n1}, L_{n2}, n^{1/2}\eta_n)$ converges under $P_{n,0}$ to that of (L_1, L_2, H). Letting $Z_{n,t} = tT_n - \frac{1}{2}t^2 I$ we have (see (3.6), (3.7))

$$n^{1/2}\Delta_{n,t} = tn^{1/2}\eta_n - (\tfrac{1}{2}t^2 L_{n2} + \tfrac{1}{6}t^3 a_3)$$

and the joint distribution of $(\Delta_{n,t}, n^{1/2}\Delta_{n,t})$ converges to that of (Δ_t, H_t) with

$$\Delta_t = tL_1 - \tfrac{1}{2}t^2 I, \qquad H_t = tH - \tfrac{1}{2}t^2 L_2 - \tfrac{1}{6}t^3 a_3.$$

Then

$$D_t = \operatorname{var}[tH - \tfrac{1}{2}t^2 L_2 \mid L_1 = u_\alpha I^{1/2}], \tag{3.16}$$

$$D_{t_0,t} = \operatorname{var}[\tfrac{1}{2}t(t_0 - t)L_2 \mid L_1 = u_\alpha I^{1/2}] \tag{3.17}$$

(see (2.20), (3.9)) and we need to prove that there exists $t_0 \geqslant 0$ such that

$$D_{t_0,t} \leqslant D_t \quad \text{for all } t > 0. \tag{3.18}$$

Let b be the coefficient of regression of H on L_2 given $L_1 = u_\alpha I^{1/2}$, i.e.

$$H = bL_2 + U, \quad \text{where} \quad \operatorname{cov}(L_2, U \mid L_1 = u_\alpha I^{1/2}) = 0.$$

Then

$$D_t = t^2 \operatorname{var}[(b - \tfrac{1}{2}t)L_2 \mid L_1] + t^2 \operatorname{var}[U \mid L_1 = u_a I^{1/2}]. \qquad (3.19)$$

On comparing (3.19) to (3.17), we see that (3.18) holds with $t_0 = 2b$ if $b \geqslant 0$ and with $t_0 = 0$ if $b < 0$.

This argument does not prove Pfanzagl's complete class theorem (see (2.22)) because this theorem holds for an arbitrary sequence of tests without any regularity of its asymptotic behaviour. However, a proper modification of Theorem 3.7 may be used to simplify the proofs of Pfanzagl and Wefelmeyer [31] quoted at the end of Section 2.4.

References

[1] Albers W., Bickel P. J. and Van Zwet W. R., Asymptotic Expansion for the Power of Distribution Free Tests in the One-Sample Problem, *Ann. Statist.* **4** (1976), pp. 108–156; correction **6** (1978), pp. 1170–1171.

[2] Bening V. E., Asymptotic Expansions under Local Alternatives. In: *IV USSR–Japan Symposium on Probability Theory and Mathematical Statistics. Abstracts of Communications*, Vol. I, Tbilisi, 1982, pp. 120–121.

[3] Bhattcharya R. N. and Rao R. R., *Normal Approximation and Asymptotic Expansions*, Wiley, New York, 1976.

[4] Bickel P. J. and Van Zwet W. R., Asymptotic Expansions for the Power of Distribution Free Tests in the Two-Sample Problem, *Ann. Statist.* **6** (1978), pp. 937–1004.

[5] Bickel P. J., Chibisov D. M. and Van Zwet W. R., On Efficiency of First and Second Order, *Internat. Statist. Rev.* **49** (1981), pp. 169–175.

[6] Callaert H., Janssen P. and Veraverbeke N., An Edgeworth Expansion for U-Statistics, *Ann. Statist.* **8** (1980), pp. 299–312.

[7] Chibisov D. M., An Asymptotic Expansion for the Distribution of a Statistic Admitting an Asymptotic Expansion (in Russian), *Teor. Veroyatnost. i Primenen.* **17** (1972), pp. 658–668; *Theor. Probability Appl.* **17** (1972), pp. 620–630.

[8] Chibisov D. M., An Asymptotic Expansion for the Distribution of Sums of a Special Form with an Application to Minimum-Contrast Estimates (in Russian), *Teor. Veroyatnost. i Primenen.* **18** (1973), pp. 689–702; *Theor. Probability Appl.* **18** (1973), pp. 649–661.

[9] Chibisov D. M., Asymptotic Expansions for some Asymptotically Optimal Tests. In: *Proceedings of the Prague Symposium on Asymptotic Statistics*, Vol. 2, (J. Hájek, ed.), pp. 37–68, Charles University, Prague, 1974.

[10] Chibisov D. M., Asymptotic Expansion for the Distribution of a Statistic Admitting a Stochastic Expansion I, II (in Russian), *Teor. Veroyatnost. i Primenen.* **25** (1980), pp. 745–756; **26** (1981), pp. 3–14; *Theor. Probability Appl.* **25** (1980), pp. 732–744; **26** (1981), pp. 1–12.

[11] Chibisov D. M., Higher Order Properties of Asymptotically Optimal Tests in One-Parameter family. In: *IV USSR–Japan Symposium on Probability Theory and Mathematical Statistics, August 23-29, 1982, Tbilisi. Abstracts of Communications*, Vol. I, pp. 171–173, Tbilisi, 1982.

[12] Chibisov D. M., Asymptotic Expansions in Problems of Testing Hypotheses,

I, II, (in Russian), *Izv. Akad. Nauk. UzSSR Ser. Fiz.-Mat. Nauk* **5** (1982), pp. 18–26; **6** (1982), pp. 23–30.

[13] Chibisov D. M., Power and Deficiency of Asymptotically Optimal Tests (in Russian), *Teor. Veroyatnost. i Primenen.* **27** (1982), pp. 812–813.

[14] Chibisov D. M. and Van Zwet W. R., On Asymptotic Expansion for the Distribution of the Logarithm of the Likelihood Ratio. In: *Third Vilnius Conference on Probability Theory and Mathematical Statistics, June 22–27, 1981, Vilnius. Abstracts of Communications*, Vol. III, pp. 55–56.

[15] Chung K. L., The Approximate Distribution of Student's Statistics, *Ann. Math. Statist.* **17** (1946), pp. 447–465.

[16] Cramér H., *Random Variables and Probability Distributions*, Cambridge University Press, 1937.

[17] Does R., *Higher Order Asymptotics for Simple Linear Rank Statistics*, Mathematisch Centrum, Amsterdam, 1982.

[18] Hájek J., Local Asymptotic Minimax and Admissibility in Estimation. In: *Proc. 6th Berkeley Sympos. Math. Statist. and Probab.*, Vol. 1, pp. 175–194, Univ. of California Press, Berkeley 1972.

[19] Hájek J. and Šidak Z., *Theory of Rank Tests*, Academia, Prague 1967.

[20] Helmers R., *Edgeworth Expansions for Linear Combinations of Order Statistics*, Mathematical Centre Tracts, Mathematisch Centrum, Amsterdam 1979.

[21] Hodges J. L., Jr., and Lehmann E. L., Deficiency, *Ann. Math. Statist.* **41** (1970), pp. 783–801.

[22] Hsu P. L., The Approximate Distributions of the Mean and the Variance of a Sample of Independent Variables, *Ann. Math. Statist.* **16** (1945), pp. 1–29.

[23] LeCam L., Locally Asymptotically Normal Families of Distributions, *Univ. of California Publ. in Statist.* **3** (1960), pp. 37–98.

[24] Linnik Yu. V. and Mitrofanova N. M., Some Asymptotic Expansions for the Distribution of the Maximum Likelihood Estimate, *Sankhyā Ser. A* **27** (1965), pp. 73–82.

[25] Malinovskii V. K., On Calculating the Deficiency of an Asymptotically Efficient Test in the Case of Markov Observations (in Russian), *Dokl. Akad. Nauk SSSR* **267** (1982), pp. 1053–1057.

[26] Petrov V. V., *Sums of Independent Random Variables* (in Russian), Nauka, Moscow 1972.

[27] Pfanzagl J., Asymptotically Optimum Estimation and Test Procedures. In: *Proceedings of the Prague Symposium on Asymptotic Statistics*, Vol. 1 (J. Hájek, ed.), pp. 201–272, Charles University, Prague 1974.

[28] Pfanzagl J., On Asymptotically Complete Classes; in: *Statistical Inference and Related Topics*, Vol. 2 (M. L. Puri, ed.), pp. 1–43, Academic Press, New York 1975.

[29] Pfanzagl J., First Order Efficiency Implies Second Order Efficiency. In: *Contributions to Statistics. Jaroslav Hájek Memorial Volume* (J. Jurečková, ed.), pp. 167–196, Academia, Prague 1979.

[30] Pfanzagl J., Asymptotic Expansions in Parametric Statistical Theory; in: *Developments in Statistics*, Vol. 3, pp. 1–97, Academic Press, New York 1980.

[31] Pfanzagl J. and Wefelmeyer W., An Asymptotically Complete Class of Tests, *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **45** (1978), pp. 49–72.

[32] Roussas G. G., *Contiguity of Probability Measures: Some Applications in Statistics*, University Press, Cambridge 1972.

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

HARRY KESTEN

Percolation Theory and Resistance of Random Electrical Networks

1. Percolation

The simplest and most classical percolation model deals with bond-percolation on $\boldsymbol{Z}^2$. This was introduced by Broadbent [1] and Broadbent and Hammersley [2] as a model for the spread of a gas or fluid through a random porous medium. The medium here is represented by the *bonds* (also called *edges*) between neighboring points of $\boldsymbol{Z}^2$. Through some of these bonds the gas or fluid may pass; these bonds are called *passable* or *open*. The other bonds are called *blocked* or *closed*. It is assumed that the characters of all the bonds of $\boldsymbol{Z}^2$ are independent random variables such that for each bond e

$$P\{e \text{ is open}\} = p, \quad P\{e \text{ is closed}\} = q := 1-p. \tag{1}$$

The corresponding probability measure on the configurations of open and closed bonds is denoted by P_p. A *path* on $\boldsymbol{Z}^2$ is an alternating sequence of vertices and bonds of the form $(v_0, e_1, v_1, \ldots, v_{n-1}, e_n, v_n)$ with v_{i-1} and v_i adjacent vertices of $\boldsymbol{Z}^2$, and e_i the bond connecting v_{i-1} and v_i, $1 \leqslant i \leqslant n$. A path is called *open* if all its edges are open. The *open cluster of a vertex* v, denoted by $W(v)$, is the collection of all vertices and bonds which belong to an open path starting at v. Most questions in percolation theory deal with the distribution of $\# W(v) :=$ the number of edges in $W(v)$, and in particular with the dependence of this distribution on p. What makes percolation theory interesting for statistical physicists is the occurrence of a "critical phenomenon", that is the existence of a *critical value*, p_H say, such that in the two domains $p < p_H$ and $p > p_H$ the system shows markedly different behavior. Specifically, define the *per-*

colation probability $\theta(p)$ by

$$\theta(p) = P_p\{\#\, W(v) = \infty\}, \tag{2}$$

and

$$p_H = \sup\{p\colon\ \theta(p) = 0\}. \tag{3}$$

Then there will exist infinite open clusters only for $p > p_H$ (see Theorem 1 below).

Note that in the present case the distribution of $\#W(v)$ is independent of v, and therefore the same is true for $\theta(p)$ and p_H. Broadbent and Hammersley [2] and Hammersley [12], [13] proved that $0 < p_H < 1$. Harris [16] showed that $p_H \geqslant 1/2$ and Sykes and Essam [28] made it highly plausible that $p_H = 1/2$. After important contributions by Russo [22] and Seymour and Welsh [24] Kesten [18] showed that p_H is indeed $\frac{1}{2}$. (A new proof of this relation is in the interesting recent article [23] of Russo). The full result is as follows.

THEOREM 1. (*For bond-percolation on* $\boldsymbol{Z}^2$).
If $p \leqslant \frac{1}{2}$ *then there exists no infinite open clusters a.e.,* $[P_p]$. (4)
If $p > \frac{1}{2}$ *then there exists exactly one infinite open cluster a.e.,* $[P_p]$. (5)

The transition from a regime with only finite clusters to one with an infinite cluster is reminiscent of the phase transition in the Ising model between "no long range order" (or no spontaneous magnetization) and "long range order" (or spontaneous magnetization). There is actually a rigorous relation between percolation and the Ising model (see [17]) but we shall not discuss this here. Other possible interpretations of the model deal with the spread of a blight in an orchard, binary alloys and the onset of gelation, etc. There even seem to be applications to questions about petroleum recovery. The reader is referred to one of the large number of recent surveys and popularizations [3], [6], [7], [9], [14], [15], [19], [25], [27], [29], [30] for more information.

In order to prove Theorem 1 use was made of two other critical probabilities p_T and p_S, which were defined in [22] and [24]

$$p_T := \sup\{p\colon\ E_p\{\#\, W\} < \infty\},$$

where E_p denotes expectation with respect to P_p. For our present purposes p_S is more important. It is the separation point between the p-values for which the "crossing probabilities" of large squares tend to zero and the p-values for which these "crossing probabilities" have a strictly

positive lim sup. (In the above model the crossing probabilities actually tend to 1 for $p > p_S$.) Formally set

$\sigma(n, p) = P_p\{\exists$ open path in $[0, n] \times [0, n]$ which connects a point on the left edge and a point on the right edge of this square$\}$.

and[1]

$$p_S = \sup\{p: \lim_{n\to\infty} \sigma(n, p) = 0\}.$$

Part of the proof of Theorem 1 consists of showing that

$$p_T = p_S = p_H. \tag{6}$$

Since $p_H = \frac{1}{2}$ this implies that for $p < \frac{1}{2}$ the P_p-probability of an open connection between the left and right edge of $[0, n] \times [0, n]$ goes to zero. This fact provides the connection with questions about resistances of a random network we shall discuss now.

2. Resistance problems

Again, we restrict ourselves in this section to the simplest case. To an edge e between two adjacent vertices of $\mathbf{Z}^2$ a resistance $R(e)$ will be assigned. It is assumed that all the $R(e)$ are independent random variables and that $R(e)$ can take on only the values 1 ohm or ∞ ohm. Again we have a family of probability measures $\{P_p\}_{0\leqslant p\leqslant 1}$ on the configurations of resistances. This time

$$P_p\{R(e) = 1\} = p, \quad P_p\{R(e) = \infty\} = q = 1 - p \tag{7}$$

for each edge e. e is called a *conductor* (*insulator*) if $R(e) = 1$ ($R(e) = \infty$). Insulators cannot conduct electricity and for resistance calculations we can therefore construct a network equivalent to the above one by removing each edge, independently of all others, with probability q, and giving resistance 1 ohm to each edge that remains. Alternatively we can use the percolation construction of Section 1 and identify open (closed) edges with conductors (insulators). Now restrict the network to the square $[0, n] \times [0, n]$ and connect all vertices on the left edge of this square, i.e., $\{0\} \times [0, n]$, by some superconducting material of zero resistance.

[1] This is the definition of Seymour and Welsh [24], which is the more intuitive one. For technical reasons it is better to define p_S by means of open crossings from left to right in the rectangle $[0, n] \times [0, 3n]$, as done in [19]. In the present model these two definitions are equivalent.

Also connect all vertices on the right edge $\{n\}\times[0,n]$ by material of zero resistance. Denote by R_n the (random) resistance between the left and right edge in $[0,n]\times[0,n]$ after these superconducting connections have been made. In principle R_n is determined by the standard rules for combining resistances in series and in parallel which can be found in any physics book (see for instance [8], Section I. 25.5, II. 22.3). However, explicit calculations are not feasible for large n, and it is precisely in the asymptotic behavior of R_n that we are interested. Early on it was realized that this asymptotic behavior exhibits another critical phenomenon. (See [20] and its references). In particular $R_n=\infty$ if and only if there exists no conducting path in $[0,n]\times[0,n]$ connecting its left and right edge. Thus the result $p_S=\frac{1}{2}$ can be rephrased as follows:

$$\text{If } p<\tfrac{1}{2} \text{ then } P_p\{R_n=\infty\}\to 1 \ (n\to\infty). \tag{8}$$

$$\text{If } p>\tfrac{1}{2} \text{ then } \limsup_{n\to\infty} P_p\{R_n<\infty\}>0. \tag{9}$$

Note that (9) still allows $R_n\to\infty$ with $n\to\infty$. For $p=1$, that is when every edge is a conductor, an easy calculation gives $R_n=1+n^{-1}$ so that we actually expect R_n to remain bounded, in some sense, for large p. The following results confirm this.

THEOREM 2 ([19], Ch. 11).

$$\textit{If } p<\tfrac{1}{2}, \textit{ then } P_p\{R_n=\infty \textit{ eventually}\}=1. \tag{10}$$

$$\textit{If } p=\tfrac{1}{2}, \textit{ then } P_p\{R_n\to\infty\}=1. \tag{11}$$

$$\begin{aligned}&\textit{If } p>\tfrac{1}{2}, \textit{ then}\\ &P_p\{C_1(p-\tfrac{1}{2})^{-\delta_1}\leqslant \liminf R_n\leqslant \limsup R_n\leqslant C_2(p-\tfrac{1}{2})^{-\delta_2}\}=1\end{aligned} \tag{12}$$

for some constants $0<C_1, C_2, \delta_1, \delta_2<\infty$ *which are independent of* p.

The proof of Theorem 2 makes use of one more critical probability, viz.

$$\hat{p}_R := \inf\{p\colon \exists C(p)>0 \text{ such that } P_p\{\exists\ C(p)n \text{ edge-disjoint conducting connections between the left and right edge in } [0,n]\times[0,n] \text{ for all large } n\}=1\}.$$

One shows that also $\hat{p}_R=\frac{1}{2}$. Thus, once p exceeds $p_S=\frac{1}{2}$ there is not just a positive probability for the existence of a single conducting connection between the left and right edge in $[0,n]\times[0,n]$, but there is even a probability close to 1 that there are $C(p)n$ distinct conducting connections,

for some $C(p) > 0$. The maximal number of such connections seems of some interest by itself. Define

$k(n)$ = maximal number of edge-disjoint conducting paths in $[0, n] \times [0, n]$ which connect a point on $\{0\} \times [0, n]$ with a point on $\{n\} \times [0, n]$.

The asymptotic behavior of $k(n)$ (including large deviation estimates) is described by the following theorem, which can be proved by slight modifications of Theorems 2.1 and 3.2 in Grimmett and Kesten [11].

THEOREM 3. *Let $p > \frac{1}{2}$. Then there exists a constant $\mu(p) > 0$ such that*

$$\frac{k(n)}{n} \to \mu(p) > 0 \quad a.e., [P_p].$$

Moreover, for $0 < \varepsilon < \mu(p)$ there exist constants $0 < C_j = C_j(\varepsilon, p) < \infty$, $j = 1, 2, 3$, such that

$$P_p\left\{\frac{k(n)}{n} \leqslant \mu(p) - \varepsilon\right\} = e^{-n(C_1 + o_n(1))}$$

(with $o_n(1) \to 0$ as $n \to \infty$), and

$$P_p\left\{\frac{k(n)}{n} \geqslant \mu(p) + \varepsilon\right\} \leqslant C_2 e^{-C_3 n^2}$$

(note the n^2 in the exponent on the right).

Remark. $\mu(p)$ equals the so-called time-constant of first-passage percolation when the time-coordinate $t(e)$ of the edge e has the Bernoulli distribution

$$P\{t(e) = 1\} = p, \quad P\{t(e) = 0\} = q = 1 - p$$

for each e (see Smythe and Wierman [26] for terminology).

3. Generalizations and open problems

An obvious generalization — which is appropriate for Section 1 as well as Section 2 — is to replace $\boldsymbol{Z}^2$ by $\boldsymbol{Z}^d$, or any other regular graph. The formulation of the percolation problem for such graphs is fairly obvious (see for instance [19], Chapters 2,3). However, only a few exact results are known. p_H has been determined for only a few planar graphs (basi-

cally the triangular lattice, the honeycomb lattice and the simple quadratic lattice), but p_H is not even known for $\mathbf{Z}^d$, $d \geqslant 3$.

Going still further, one can in the percolation problem allow the probability for a bond to be open to be different for different bonds. E.g., on $\mathbf{Z}^2$, one can give one value to $P\{e \text{ is open}\}$ for all horizontal edges e, and another value for all vertical edges e. Or one can consider site-problems, in which the vertices (instead of the bonds) are passable or blocked. (cf. [19] again). Recently one has even considered mixed problems in which both the bonds and the vertices can be blocked ([14]). Many other variations and generalizations can be found in the literature. For instance considerable progress has been made recently [4], [5] on "oriented-percolation" problems in which a path can traverse the bonds only in one direction (up and to the right, say).

Turning now to the resistance problem of Section 2 a generalization of another kind immediately comes to mind. One can allow an arbitrary distribution on $[0, \infty]$ for $R(e)$, rather than the Bernoulli distribution concentrated on $\{1, \infty\}$ of (7). In this generalization the $R(e)$ are i.i.d., random variables such that

$$P\{R(e) = \infty\} = p(\infty),$$

$$P\{R(e) \leqslant x\} = (1-p(\infty))\ F(x), \quad x < \infty,$$

for some distribution F on $[0, \infty)$. Results analogous to Theorem 2 for this situation can be found in [19], Chapter 11.

Whether one deals with the original resistance problem of Section 2 or with the above generalization, in either case one would like to show that on $\mathbf{Z}^2$ R_n has a limit in some sense as $n \to \infty$. Golden and Papanicolaou [10] and R. Künnemann [21] have proved that R_n has an L^2 limit, provided

$$P\{a \leqslant R(e) \leqslant b\} = 1 \quad \text{for some } 0 < a \leqslant b < \infty.$$

It is not clear whether their proof can be extended to cases where $R(e)$ can take the values 0 and/or ∞ (as in the models considered here).

When considering the resistance problem on $\mathbf{Z}^d$ with $d \geqslant 3$ we take for R_n the resistance between the faces $\{0\} \times [0, n]^{d-1}$ and $\{n\} \times [0, n]^{d-1}$ in the restriction of the network to the cube $[0, n]^d$. In this case the asymptotic order of R_n for large p should be n^{2-d} (since $R_n = n(n+1)^{1-d}$ when *all* edges have a resistance of 1 ohm). Thus one expects at least that $n^{d-2}R_n$ is bounded when $R(e)$ has the distribution (7) for a $p > p_{\vec{H}}(\mathbf{Z}^d)$: $:=$ critical probability for bond-percolation on $\mathbf{Z}^d$. However, we have

been able to prove this result only for $p > \frac{1}{2}$ (see [19], Theorem 11.3), even though it is known that $p_H(\boldsymbol{Z}^d) < \frac{1}{2}$ for $d \geqslant 3$ ([19], Example 10.2 (iii)).

On the basis of analogy with other critical phenomena and Monte Carlo evidence one expects that R_n satisfies a powerlaw near the critical probability. E.g., in the special case of $\boldsymbol{Z}^2$ and (7) one expects (see [20], [27]) that there exists a constant $\mu > 0$ such that

$$\lim_{n\to\infty} R_n = (p - \tfrac{1}{2})^{-\mu + o_p(1)}, \quad \text{where } o_p(1) \to 0 \text{ when } p \downarrow \tfrac{1}{2}. \tag{13}$$

Clearly (12) is only a poor approximation to (13). However, to the best of our knowledge no powerlaws such as (13) have been established for the percolation problems discussed here.

References

[1] Broadbent S. R., In Discussion of Symposium on Monte Carlo Methods, *J. Roy. Stat. Soc.* (B), **16** (1954), 68.

[2] Broadbent S. R. and Hammersley J. M., Percolation Processes, *Proc. Cambr. Phil. Soc.* **53** (1957), pp. 629–641 and pp. 642–645.

[3] de Gennes P. G., La percolation: un concept unificateur, *La Recherche* **7** (1976), pp. 919–927.

[4] Durrett R., Oriented Percolation in Two Dimensions, *Ann. Probability* **12** (1984).

[5] Durrett R. and Griffeath D., Supercritical Contact Processes on $\boldsymbol{Z}$, *Ann. Probability* **11** (1983), pp. 1–15.

[6] Efros A., What is the Theory of Percolation (in Russian), *Kvant* **2** (1982), pp. 2–9.

[7] Essam J. W., Percolation Theory, *Reports on Progress in Physics*, **43** (1980), pp. 833–912.

[8] Feynman R. P., Leighton R. B. and Sands M., *The Feynman Lectures in Physics, Vol. I-III*, Addison-Wesley Publ. Co., 1963–1965.

[9] Frisch H. L. and Hammersley J. M., Percolation Processes and Related Topics, *J. Soc. Indust. Appl. Math.* **11** (1963), pp. 894–918.

[10] Golden K. and Papanicolaou G., Bounds for Effective Parameters of Heterogeneous Media by Analytic Continuation, *Comm. Pure Appl. Math.* **90** (1983), pp. 473–491.

[11] Grimmett G. R. and Kesten H., *First-passage Percolation, Network Flows and Electrical Resistances*, to appear in *Z. Wahrsch. verw. Geb.*

[12] Hammersley J. M., Percolation Processes. Lower Bounds for the Critical Probability, *Ann. Math. Statist.* **28** (1957), pp. 790–795.

[13] Hammersley J. M., Bornes supérieurs de la probabilité critique dans un processus de filtration, In: *Le calcul des probabilités et ses applications*, CNRS Paris, 1959, pp. 17–37.

[14] Hammersley J. M., Percolation, In: *Biological Growth and Spread*, Springer Lecture Notes in Biomathematics, Vol. 38, 1980, pp. 484–494.

[15] Hammersley J. M. and Welsh D. J. A., Percolation Theory and its Ramifications, *Contemp. Phys.* **21** (1980), pp. 593–605.
[16] Harris T. E., A Lower Bound for the Critical Probability in a Certain Percolation Process, *Proc. Cambr. Phil. Soc.*, **56** (1960), pp. 13–20.
[17] Kasteleyn P. W. and Fortuin C. M., Phase Transition in Lattice Systems with Random Local Properties, *Proc. Intern. Conf. Stat. Mech.* Kyoto, 1968, *J. Phys. Soc. Japan* **26**, Supplement (1969), pp. 11–14.
[18] Kesten H., The Critical Probability of Bond Percolation on the Square Lattice Equals $\frac{1}{2}$, *Comm. Math. Phys.* **74** (1980), pp. 41–59.
[19] Kesten H., *Percolation Theory for Mathematicians, Progress in Prob. and Statistics*, Vol. 2, Birkhäuser, 1982.
[20] Kirkpatrick S., Percolation and Conduction, *Rev. Modern Phys.*, **45** (1973), pp. 574–588.
[21] Künnemann R., *Effective Conductivity on a Lattice as a Limit of Box-conductivities*, to appear.
[22] Russo L., A note on Percolation, *Z. Wahrsch. verw. Geb.* **43** (1978), pp. 39–48.
[23] Russo L., An Approximate Zero-One Law, *Z. Wahrsch. verw. Geb.* **61** (1982), pp. 129–139.
[24] Seymour P. D. and Welsh D. J. A., Percolation Probabilities on the Square Lattice, *Ann. Discrete Math.* **3** (1978), pp. 227–245.
[25] Smythe R. T., A Mathematical Approach to Percolation in Percolation Structures and Processes, *Ann. Israel Phys. Soc.* **5** (1983), pp. 477–498.
[26] Smythe R. T. and Wierman J. C., *First-Passage Percolation on the Square Lattice*, Springer Lecture Notes in Math. **671** (1978).
[27] Stauffer D., Scaling Theory of Percolation Clusters, *Phys. Reports*, **54**, No. 1 (1979), pp. 1–74.
[28] Sykes M. F. and Essam J. W., Exact Critical Percolation Probabilities for Site and Bond Problems in Two Dimensions, *J. Math. Phys.* **5** (1964), pp. 1117–1127.
[29] Thouless D. J., Percolation and localization. In: R. Balian, R. Maynard and G. Toulouse, (eds.), *La matiere mal condensée — III condensed matter*, North-Holland, 1979, pp. 5–62.
[30] Wierman J. C., Percolation Theory, *Ann. Probability* **10**, (1982), pp. 509–524.

CORNELL UNIVERSITY
ITHACA, NEW YORK 14853
U.S.A.

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

PAUL MALLIAVIN

Analyse différentielle sur l'espace de Wiener

N. Wiener en développant en 1923 les fonctionnelles de carré intégrable du mouvement brownien suivant les polynomes d'Hermite a commencé l'étude de fonctionnelles *non linéaires* du mouvement brownien. K. Itô en introduisant en 1943 la théorie des équations différentielles stochastiques, aboutissait à des fonctionnelles non linéaires de Wiener, permettant de calculer *effectivement* les solutions d'équations paraboliques sur $\boldsymbol{R}^n$ à coefficients variables. L'analyse différentielle sur l'espace de Wiener se propose d'étudier des fonctionnelles non linéaires "régulières" en mélangeant les ressources du calcul intégral de Wiener–Itô avec les techniques d'un calcul différentiel "ad hoc" défini de telle sorte que les fonctionnelles de Itô soient différentiables. Il est bien connu que l'intégrale stochastique n'est pas un algorithme *robuste*: les fonctionnelles de Itô ne sont continues pour aucune norme de Banach; a fortiori la théorie du calcul différentiel classique dans les espaces de Banach ne saurait leur être appliquée. On souhaite pouvoir appliquer aux fonctionnelles de Itô des techniques classiques en dimension finie (phase stationnaire, désintégration suivant une famille d'hypersurfaces, etc. ...).

(I) *Différentes approches à la théorie de la dérivation et leurs équivalences.*

(II) *Géométrie différentielle sur l'espace de Wiener.*

(III) *Application à des estimations de fonctionnelles.*

I. Notion de différentiabilité

a. Dérivation vectorielle. Nous notons par X l'espace de Banach des applications continues de $[0,1]$ dans $\boldsymbol{R}^m$, nulles en zéro. Le mouvement brownien sur $\boldsymbol{R}^m$ permet de définir sur X une mesure gaussienne μ, appelée mesure de Wiener. On note par H l'espace de Hilbert des fonctions $x \in X$ dont la dérivée $dx/d\tau$ est de carré intégrable et par H_0 le sous-espace

dense de H constitué des fonctions telles que $d^2x/d\tau^2$ soit une mesure. Alors si $h_0 \in H_0$ le produit scalaire $h \to (h \mid h_0)_H$ se prolonge à X en une forme linéaire continue $\langle x, h_0 \rangle$.

Les vecteurs de H sont des *vecteurs de dérivation* pour la mesure μ. Notant τ_h: $x \mapsto x+h$, on a le résultat classique de Cameron–Martin:

$$(\tau_h)_* \mu = \exp\left(\int_0^1 h' dx - \tfrac{1}{2}\int_0^1 |h'|^2 d\tau\right) \mu.$$

En particulier définissant la divergence par la formule

$$\mathrm{div}_\varrho(h)\, d\varrho = \lim_{\varepsilon \to 0} \varepsilon^{-1}[(\tau_{\varepsilon h})_* \varrho - \varrho]$$

on obtient

$$\big(\mathrm{div}_\mu(h)\big)(x) = \int_0^1 h'(\tau)\, dx(t)$$

En analyse de Fourier sur $\boldsymbol{R}^n$, un procédé classique pour introduire les espaces de Sobolev est de considérer les opérateurs de dérivation à valeur vectorielle obtenus sur $L^p(\boldsymbol{R}^n)$ comme limite d'opérateurs de translation. Dans cette approche le fait que la mesure de Haar de $\boldsymbol{R}^n$ soit invariante joue un rôle déterminant. Ici la mesure μ est quasi-invariante. De plus

$$\mathrm{div}_\mu(h) \in \mathscr{W}_0(X) \quad \text{où on a défini} \quad \mathscr{W}_0(X) = \bigcap_p L^p(X).$$

Sur $\mathscr{W}_0(X)$ on prendra la topologie d'espace de Fréchet. On note $\mathscr{W}_1(X) = \{f \in \mathscr{W}_0(X);\ \forall h \in H,\ D_h f \in \mathscr{W}_0(X)\}$ où encore

$$D_h f = \lim_{\varepsilon \to 0} \varepsilon^{-1}[f(x+\varepsilon h) - f(x)] \tag{1}$$

la limite étant prise au sens de la topologie de $\mathscr{W}_0(X)$. Alors l'application $h \mapsto D_h f$ est une application linéaire continue de H à valeurs dans $\mathscr{W}_0(X)$. Par suite il existe une application ∇f définie sur X à valeurs dans H telle que

$$(\nabla f \mid h) = D_h f.$$

Si G est un espace de Hilbert abstrait, on note par $L^p(X; G)$ l'espace des fonctions L^p sur X à valeurs dans G. Alors si $f \in \mathscr{W}_1(X)$, on a $\nabla f \in \mathscr{W}_0(X; H)$. D'une manière plus générale si G est un espace de Hilbert abstrait, on définit $H \otimes G$ l'espace de Hilbert des applications linéaires de H dans G muni de la norme de Hilbert–Schmidt. Si $u \in \mathscr{W}_0(X; G)$, $D_h u$ est encore défini par (i) d'où la définition de $\mathscr{W}_1(X; G)$. Pour $v \in \mathscr{W}_1(X; G)$ on a $\nabla v \in \mathscr{W}_0(X; H \otimes G)$. En définissant H_p comme le produit tensoriel symétrique $H \otimes \ldots \otimes H$ muni de la norme de Hilbert–Schmidt on définit alors $\mathscr{W}_p(X; G)$ et l'opérateur

$$\nabla^p\colon \mathscr{W}_p(X; G) \to \mathscr{W}_0(X; H_p \otimes G).$$

Les opérateurs ∇^r sont fermés et les espaces $\mathscr{W}_r(X;G)$ sont complets. (On trouve dans [25] une approche alternative utilisant les fonctions cylindriques. D'autre part la notion de vecteur de Cameron–Martin permet de commencer une théorie de la dérivation [16] lorsque la mesure de base n'est plus gaussienne.)

b. Domaines de l'opérateur de Ornstein–Uhlenbeck. Considérons la décomposition donnée par le chaos de Wiener $L^2(X) = \oplus \mathscr{H}_n$ où $\mathscr{H}_n$ note l'espace des fonctionnelles de Wiener de degré n, linéairement engendré par les polynomes de Hermite de degré total n. L'opérateur de Ornstein–Uhlenbeck est alors défini sur $\mathscr{H}_n$ par

$$\mathscr{L}|_{\mathscr{H}_n} = -n \text{ Identité sur } \mathscr{H}_n. \tag{i}$$

L'opérateur de Cauchy $\mathscr{C}$ est défini par $-\sqrt{-\mathscr{L}}$. Sur les fonctions cylindriques $\mathscr{L}$ s'exprime comme un opérateur différentiel de la forme

$$\mathscr{L} = \tfrac{1}{2}\left\{\sum \partial^2_{\xi_i} - \xi_i \partial_{\xi_i}\right\}. \tag{ii}$$

Sous la forme (i) la fermeture de $\mathscr{L}$ ou de $\mathscr{C}$ dans $L^2(X)$ s'écrit immédiatement. Dans [28] on introduit $\mathscr{D}^p(\mathscr{L};G)$, le domaine de la fermeture de $\mathscr{L}$ dans $L^p(X;G)$. De même dans [23] le domaine $\mathscr{D}^p(\mathscr{L};\mathscr{C})$ est défini. On pose $\bigcap \mathscr{D}^p(\mathscr{L};G) = \mathscr{D}^\infty(\mathscr{L};G)$. Alors $\int u\mathscr{L}v d\mu = \int v\mathscr{L}u\, d\mu$ si u, $v \in \mathscr{D}^\infty$. Enfin, pour $c > 0$, $(\mathscr{L} - c)$ est un opérateur *inversible* de $\mathscr{D}^p(\mathscr{L}, G) \to L^p(X, G)$.

Stabilité par calcul symbolique. Soit $f \in \mathscr{D}^\infty(\mathscr{L};\boldsymbol{R})$ et soit φ est une fonction définie sur $\boldsymbol{R}$, à croissance polynomiale ainsi que ses dérivées d'ordre $\leqslant 2$, alors $\varphi \circ f \in \mathscr{D}^\infty(\mathscr{L};\boldsymbol{R})$.

Inégalité d'intégrales singulières [23]. *On a l'isomorphisme topologique*

$$\mathscr{D}^\infty(\mathscr{C};G) \simeq \mathscr{W}_1(X,G),$$

et plus précisément

$$\frac{1}{C(p)}\|\mathscr{C}u\|_{L^p(X,G)} \leqslant \|\nabla u\|_{L^p(X,H\otimes G)} \leqslant C(p)\|\mathscr{C}u\|_{L^p(X,G)} \quad (1 < p < \infty).$$

Cet isomorphisme entraîne évidemment le même isomorphisme entre $\mathscr{D}^\infty(\mathscr{C}^r;G)$ et $\mathscr{W}_r(X;G)$ (cf. également [32]).

c. Calcul stochastique le long du processus d'Ornstein–Uhlenbeck. Le processus d'Ornstein–Uhlenbeck (ou *processus O. U.*) en dimension finie est

déterminé par ses probabilités de transition, $p_t(x_0, dx)$, qui peuvent être écrites explicitement par la formule de Mehler:

$$\int p_t(x_0, dx)\varphi(x) = \int \varphi\left(x_0 e^{-t/2} + (1-e^{-t})^{1/2} y\right) d\mu(y).$$

Cette formule peut s'écrire aussi bien en dimension infinie et définit le processus O.U. sur X. Un autre procédé est de considérer une base orthonormée de H composée d'éléments de H_0, par exemple la base $(k\pi)^{-1}\sin(k\pi\tau) = e_k$, $k > 1$, et $e_0 = 1$ *et de poser*

$$x_w(t) = \sum_{k=0}^{+\infty} e_k G^k_{w_k}(t) \tag{i}$$

où $G^k_{w_k}$ sont des échantillons indépendants du processus O.U. sur $\boldsymbol{R}^m$, $w = (w_0, \ldots, w_k, \ldots)$.

Le processus (i), laisse la mesure μ invariante, possède une p.s. version continue sur X [18]. On peut d'autre part le réaliser à partir d'un bruit blanc à deux paramètres [18].

Une fonction f sera dite une fois régulière si $M_w(t) = f\left(x_w(t)\right)$ est *une semimartingale.* Alors $\mathscr{L}f$ sera défini [9], [18], [27] comme réalisant la partie à variation bornée de $M_w(t)$. D'autre part, $\|\nabla f\|^2$ sera défini de même comme le processus croissant associé à $M_w(t)$. Ces définitions ont été inspirées des travaux [7] sur la différentiabilité des fonctions finement holomorphes sur C.

Enfin notant $\varphi_k(x) = \langle x, e_k\rangle$, on pourra suivant [25], [28] définir ∇f par

$$\nabla f = \sum_k e_k \nabla f \cdot \nabla \varphi_k$$

où $\nabla f \cdot \nabla \varphi_k$ est la polarisation de la forme quadratique $\|\nabla f\|^2$.

Cette technique de localisation sur les trajectoires du processus O.U. permet de définir la régularité de fonction f sur un ouvert fin de X (cf. [22]). Pour l'équivalence globale de cette approche avec a. et b. voir [27].

d. Un exemple de fonctions régulières. Donnons-nous des fonctions C^∞ à dérivées bornées sur $\boldsymbol{R}^q$, $a^i_k(y)$, $c^i(y)$. Considérons le système différentiel stochastique $dy^k_x(\tau) = a^k_i(y)\, d_\tau x^k(\tau) + c^k(y)\, d\tau$, $y^k(0) = 0$. Alors $g(x) = y_x(1)$ définit [9], [18], [27] une application $g \in \mathscr{W}_\infty(X; \boldsymbol{R}^q)$. De plus la matrice de covariance $\nabla g^i \cdot \nabla g^j$ a une expression explicite ([1], [2], [18], [27]) qui permet de démontrer des théorèmes d'hypoellipticité C^∞ avec dégénérescence ([18], [19]). Le calcul de ∇g fait intervenir du point de vue

de a. une variation par une courbe H, au sens de [14], [26]. Cette variation peut être exprimée par une formule de Girsanov locale [1]. Du point de vue de c. intervient un calcul stochastique sur le processus de O.U. qui obtient les solutions de S.D.E. comme *limite* fonctionnelle de solution d'équation qu'un principe de transfert différentiel ordinaire permet de ramener (cf. [6], [20]) ce calcul stochastique au calcul des variations sur des équations différentielles ordinaires, suivi ensuite d'un passage à la limite. Si on utilise l'approche b. un lemme fondamental [27] est la *commutation* de $\mathscr{L}$ avec le foncteur de l'intégrale stochastique.

II. Géométrie différentielle sur l'espace de Wiener

Donnons quelques exemples:

(a) *Remontée d'un champ de vecteur sur* X [19]. Soit $g \in \mathscr{W}_0(X; \boldsymbol{R}^q)$. Supposons $\left(\det(\nabla g^i \cdot \nabla g^j)\right)^{-1} \in \mathscr{W}_0(X)$. Alors si A est champ de vecteur C^∞ sur $\boldsymbol{R}^q$, il existe un champ de vecteur $\tilde{A} \in \mathscr{W}_\infty(X; H)$ tel que

$$g'(x)\cdot\tilde{A} = A_{g(x)}.$$

Notant $\varrho = g_*(\mu)$ on en déduit [19] que

$$\int_{\boldsymbol{R}^q} |\mathrm{div}_\varrho(Z)|\, d\varrho \leqslant \int |\mathrm{div}_\mu(\tilde{Z})|\, d\mu.$$

Comme on verra en (c) que $\mathrm{div}_\mu(\tilde{Z}) \in L^1_\mu$ si $\tilde{Z} \in \mathscr{W}^2_2(X)$ on en déduit [15] une majoration de

$$\|\nabla(u^\beta)\|_{L^p(\boldsymbol{R}^n)} \quad \text{où} \quad d\varrho = u dy \text{ et où } \beta > 0.$$

(b) *Cobord stochastique.* Si π est une forme différentielle de degré 1 sur $\boldsymbol{R}^q$ alors $\int \pi \circ dy_x$ est définie de façon intrinsèque [8], [20] et définit une fonctionnelle $g \in \mathscr{W}_\infty(X)$. Le cobord de π permet de calculer la différentielle de g ([10], [21]).

(c) *Complexe de de Rham–Hodge.* Considérant l'adjoint de l'opérateur ∇ pour la mesure μ, on peut définir le complexe de de Rham–Hodge sur X. En degré p le laplacien de de Rham–Hodge s'écrit $-\mathscr{L} - p$ Identité ([26]). Par suite le complexe est acyclique. D'autre part si $Z \in \mathscr{W}_2(X; H)$, alors $\mathrm{div}_\mu(Z) \in \mathscr{W}_1(X)$.

(d) *Formules générales de Cameron–Martin.* Soit $Z \in \mathscr{W}_\infty(X; G)$ satisfaisant de plus à des conditions d'intégrabilité exponentielle convenable, alors le flot $U^Z_t(x_0)$ est défini pour presque tout x_0 et $(U^Z_t)_*\mu$ se calcule par une intégrale de la divergence sur les trajectoires du flot [5].

(e) *Intégrales stochastiques anticipantes.* L'opérateur $\mathrm{div}_\mu(Z)$ permet de retrouver l'intégrale stochastique de Skohorod de fonctions anticipantes [13], [32].

(f) *Fonctions implicites.* A chaque norme de $\mathscr{W}_\infty(X)$ est associé une *capacité* $c_{p,r}$. On dit que $A \subset X$ est *mince* si pour tout (p, r, ε) on peut trouver un ouvert 0 de X, $0 \supset A$ tel que $c_{p,r}(0) < \varepsilon$. La projection orthogonale de corang fini d'un ensemble mince est mince; les fonctions $f \in \mathscr{W}_\infty(X)$ possèdent une redéfinition à l'extérieur d'un ensemble mince; de cette redéfinition résulte un théorème des fonctions implicites en corang fini [22] (cf. [3] pour l'étude des grandes déviations).

(g) *Phase stationnaire.* Les méthodes de phase stationnaire peuvent être appliquées à des intégrales oscillantes sur X [12].

(h) *Equations aux dérivées partielles stochastiques hyperboliques.* Le formalisme du calcul différentiel sur X, transporté à des espaces de bruit blanc à plusieurs paramètres permet de donner une définition de l'intégrale stochastique adapté à la résolution d'équations non linéaires, hyperboliques, aux dérivées partielles stochastiques [11].

(*i*) *Remontée des courants à l'espace de Wiener* [33].

III. Applications

(a) *Estimées elliptiques.* Sur une variété riemannienne une estimée de $\nabla(\log p_t)$ a été obtenue [3] uniquement en termes du tenseur de Ricci, intégré sur les trajectoires du mouvement brownien.

(b) *Estimées elliptiques en dimension infinie.* Un exemple de telle estimée est obtenu dans [15], estimée qui contrôle l'énergie libre d'un système de Ising à spins continus.

(c) *Estimées hypoelliptiques d'équations dégénérées* du type de Hörmander [2], [17], [19], [29] et cf. l'article de D. Stroock dans ce volume.

(d) *Régularité de semi-groupes de processus réfléchis* [2]. Un calcul des variations sur les excursions est construit. Les semi-groupes obtenus sont *progressivement* régularisants.

(e) *Régularité en filtrage non linéaire* ([4], [24]) obtenu à l'aide d'un calcul des variations partiel sur les degrés de liberté restant disponibles une fois l'observation effective.

Bibliographie

[1] Bismut J.-M., Martingales, The Malliavin Calculus and Hypoellipticity under General Hörmander's Conditions, *Z. Wahrsch.* **56** (1981), pp. 469–505.

[2] Bismut J.-M., Calcul des variations stochastique et processus de sauts, *Z. Wahrsch.* **63** (1983), pp. 147–263.

[3] Bismut J.-M., *Large deviations and Malliavin's calculus*, Birkhäuser, 1984 et *CRAS* **296** (1984), p. 1009.

[4] Bismut J.-M. et Michel D., Diffusions conditionnelles, *Journal of Functional Analysis* **44** (1981), pp. 174–212 et **45** (1982), pp. 274–293 (Application au filtrage).

[5] Cruzeiro A.-B., Flots sur l'espace de Wiener et formules de Cameron–Martin, *Journal of Functional Analysis* **54** (1983), pp. 206–227.

[6] Belopol'skaya Ya. L. and Daleskii L., Itô equations and differential geometry, *Usp. Mat. Nauk* **37** (3) (1982), pp. 95–142.

[7] Debiard A. et Gaveau B., Différentiabilité des fonctions finement harmoniques, *Inventiones Mathematicae* **29** (1975), pp. 111–123, et T. J. Lyons, Finely holomorphic functions, *Journal of Functional Analysis* **37** (1980), pp. 1–18.

[8] Ikeda N. and Manabe S., Integral of differential forms along the path of diffusion processes, *Publ. R. I. M. S. Kyoto University* **15** (1979), pp. 827–852.

[9] Ikeda N. and Watanabe S., *Stochastic Differential Equations and Diffusion Processes*, North-Holland, 1981, pp. 322–352.

[10] Gaveau B., Différentielle extérieure stochastique, *Compte Rendus Académie des Sciences* (*CRAS*) **286** (1978), pp. 381–384.

[11] Gaveau B. et Moulinier J., Géométrie infinitesimale gaussienne et intégrales stochastiques de bruits blancs à plusieurs paramètres, *Compte Rendus Académie des Sciences* **296** (1983), pp. 43–46.

[12] Gaveau B. et Moulinier J., Méthode de la phase stationnaire pour l'évaluation de fonctionnelles de Wiener, *Journal of Functional Analysis* **54** (1983), pp. 161–177.

[13] Gaveau B. et Trauber Ph., Opérateur de divergence sur l'espace de Wiener, *Journal of Functional Analysis* **42** (1981), pp. 356–368.

[14] Gihman I. L. and Skorohod A. V., *Stochastic Differential Equations*, Ergeb. der Mathematik, Springer, 1972.

[15] Holley R. and Stroock D., Diffusions on an infinite dimensional torus, *Journal of Functional Analysis* **42**, (1981), pp. 29–63.

[16] Kusuoka S., Dirichlet forms and diffusion processes on Banach spaces, *Journal of the Faculty of Science, The University of Tokio* **29**, Section I. A. (1982), pp. 79–95.

[17] Kusuoka S., *Analytic Properties of Wiener Functional*, Lec. Notes **923**, pp. 1–47.

[18] Malliavin P., Stochastic calculus of variation and hypoelliptic operators, *Proc. Int. Conference on S. D. E., Kyoto, 1976*, pp. 195–264. Kinokuniya, Tokyo, 1978.

[19] Malliavin P., C^k-hypoellipticity with degeneracy, in: *Stochastic Analysis*, A. Friedmann et Marc Pinsky, Editors, Acad. Press, 1978, pp. 199–214, 327–340.

[20] Malliavin P., *Géométrie différentielle stochastique*, Cours Univ. de Montréal, 1977, Presses de l'Université de Montréal, 1978, 128 pp.

[21] Malliavin P., Sur certaines intégrales stochastiques oscillantes, *Compte Rendus Acad. des Sciences* **295** (1982), pp. 295–300.

[22] Malliavin P., Implicit function in finite corank on the Wiener space, *Symp. SDE Kyoto 1982,* pp. 353–370, Kinokuniya, 1984.

[23] Meyer P. A., Inégalités d'intégrales singulières sur l'espace du mouvement brownien, *Proceedings of conference of Bangalore,* 1982, Lecture Notes in Control Theory, pp. 201–214.

[24] Michel D., Régularité des lois conditionnelles en théorie du filtrage non linéaire et calcul des variations stochastique, *Journal of Functional Analysis* **41** (1981), pp. 8–36, et *Proc. of Inter. Symp. SDE Kyoto 1982.*

[25] Shigekawa I., Derivatives of Wiener functionals and absolute continuity of induced measures, *J. Math. Kyoto University* **20** (1980), pp. 263–289.

[26] Shigekawa I., De Rham–Hodge complex on the Wiener Space, *Séminaire Ecole Normale,* Février 1983.

[27] Stroock D., The Malliavin calculus and its application to second order parabolic differential equations, *Math. Systems,* **14** (1981), pp. 25–65.

[28] Stroock D., The Malliavin Calculus, a Functional Analytic Approach, *Journal of Functional Analysis* **44** (1981), pp. 212–257.

Ajouté à la correction des épreuves:

[29] Bismut J.-M., Calculus of boundary processes, to appear in *Annales de l'Ecole Normale* 1984.

[30] Bismut J.-M., The Atiyah-Singer Theorems: A probabilistic approach I. The Index Theorem, *Journal of Funct. Analysis* **57** (1984), pp. 56–99.

[31] Cruzeiro A.B., Unicité de solutions d'équations différentielles sur l'espace de Wiener, *Journal of Functional Analysis* **58** (1984), September.

[32] Krée M., Propriétés de trace en dimension infinie, *Bull. Soc. Math. France* **105** (1977), pp. 141–163 et avec Krée P., Continuité de la divergence dans les espaces de Sobolev sur l'espace de Wiener, *Comptes Rendus* **296** (1983), pp. 833–836.

[33] Watanabe S., Malliavin's calculus in terms of generalized functionnals de Bangalore, 1982, *Lecture Notes in Control Theory.*

10, RUE SAINT LOUIS EN L'ILE
75004 – PARIS

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

PETR MANDL

Self-Optimizing Control of Markov Processes and Markov Potential Theory

The topic of the paper will be exposed on two families of processes: the controlled one-dimensional diffusion processes and the controlled Markov processes with a countable state space. It is hoped that the extensions to other types of processes will be apparent.

The former family includes random processes $X = \{X_t, t \geqslant 0\}$ having stochastic differential

$$dX_t = a(X_t, U_t)\,dt + b(X_t)\,dW_t, \qquad t \geqslant 0. \tag{1}$$

The first term on the right-hand side expresses the drift of the process. The drift coefficient $a(x, u)$ depends on the state variable $x \in (-\infty, \infty)$ and on the control parameter $u \in \mathcal{U}$. We assume that $a(x, u)$ is continuous and $\mathcal{U}$ is a metric space. The function $b(x)$ determines the magnitude of the random disturbance or the local diffusion. Let $b(x)$ be Lipschitz continuous, $b(x) > 0$. The control $U = \{U_t, t \geqslant 0\}$ in (1) is a random process nonanticipative with respect to X. More explicitly, U is progressively measurable with respect to the nondecreasing family of σ-algebras $\mathcal{F}^X = \{\mathcal{F}_t^X = \sigma a(X_s, s \in [0, t]), t \geqslant 0\}$. The controller chooses the parametr value according to the past trajectory of X. To speak in a more intuitive manner, we suppose that $a(x, u)$ and $b(x)$ express the properties of a physical system S in the way that X satisfying (1) is the trajectory of S under control U. In other words, S is a system with generating operator

$$A(u) = \tfrac{1}{2} b(x)^2 \frac{d^2}{dx^2} + a(x, u) \frac{d}{dx}.$$

To guarantee the existence of a week solution of (1) for each $U = \{U_t, t \geqslant 0\}$, additional conditions are to be imposed. Let us present a set of such conditions using Girsanov's theorem.

PROPOSITION 1. *Let*

$$a^+(x) = \sup_{u \in \mathcal{U}} |a(x, u)|, \quad x \in (-\infty, \infty),$$

and let $Y = \{Y_t, t \geqslant 0\}$ *be the solution of*

$$dY_t = b(Y_t)\,dW_t, \quad t \geqslant 0, \; Y_0 = J.$$

If

$$E_y \exp\left\{\tfrac{1}{2} \int_0^t a^+(Y_s)^2\, b(Y_s)^{-2}\, ds\right\} < \infty, \quad t \geqslant 0, \tag{2}$$

then (1) *has a week solution for any nonanticipative control* U *and for initial position* $X_0 = y$.

The subscript y in (2) is to stress the initial state of Y.

A controlled Markov process with a countable state space I, say $I = \{1, 2, \ldots\}$, is most frequently characterized by means of its transition rates. The meaning of the rates can be condensed in the formula

$$P(X_{t+dt} = j \mid \mathcal{F}_t^X) = a(X_t, j; U_t)\,dt \quad \text{for } j \neq X_t. \tag{3}$$

We call $a(i, j; u)$, $i \neq j$, the *transition rate* from state i into state j under control parameter u. We assume $a(i, j; u)$ to be continuous in u. The generating operator for S is, in this case, the matrix of transition rates

$$A(u) = \|a(i, j; u)\|_{i,j \in I}.$$

We let

$$0 < a(i, u) = \sum_{j \neq i} a(i, j; u), \quad a(i, i, u) = -a(i, u), \quad i \in I.$$

A formal treatment of (3) is as follows. We say that random process X is the trajectory of S under control U, if for any function $w(j)$, $j \in I$, with a finite support,

$$M_t = w(X_t) - w(X_0) - \int_0^t A(U_s)\, w(X_s)\, ds, \quad t \geqslant 0, \tag{4}$$

is a local martingale. The statement remains valid for $w(j)$ satisfying an appropriate restriction on the growth at infinity.

For diffusion processes, (4) is the consequence of the Itô formula for $w(x)$, $x \in (-\infty, \infty)$, twice continuously differentiable. We then have

$$M_t = \int_0^t b(X_s) \frac{d}{dx} w(X_s)\, dW_s, \quad t \geqslant 0. \tag{5}$$

The formulation of a control problem involves the introduction of a criterion to compare the controls. To this purpose suppose we are given a continuous function $c(x, u)$, $x \in (-\infty, \infty)$, $u \in \mathcal{U}$. Set

$$C_t = \int_0^t c(X_s, U_s)\, ds, \qquad t \geqslant 0.$$

C_t is the evaluation of the trajectory. We shall interpret it as the cost incurred up to time t. We shall concentrate on the optimization problem

$$\lim_{t \to \infty} t^{-1} C_t = \min \qquad \text{a.s.} \tag{6}$$

I.e., we consider infinite planning horizon and the average cost per unit time as criterion. Time averaging is given priority to taking mathematical expectation.

In cases of interest the minimum on the right-hand side of (6) is a constant. Thus, take any constant θ, and introduce it into (4) in the following way

$$C_t - t\theta + w(X_t) - w(X_0) = M_t + \int_0^t \big(A(U_s)\, w(X_s) + c(X_s, U_s) - \theta\big)\, ds, \qquad t \geqslant 0. \tag{7}$$

Denoting the left-hand side by R_t and writing

$$\varphi(x, u) = A(u)\, w(x) + c(x, u) - \theta, \qquad x \in (-\infty, \infty),\ u \in \mathcal{U},$$

we can rewrite (7) as

$$R_t = M_t + \int_0^t \varphi(X_s, U_s)\, ds = M_t + \Phi_t, \qquad t \geqslant 0. \tag{8}$$

If we succeed to find θ, $w(x)$ such that

$$\min_{u \in \mathcal{U}} \varphi(x, u) = 0, \qquad x \in (-\infty, \infty), \tag{9}$$

then R is a submartingale for each U, and (8) is its Doob–Meyer decomposition.

Let (9) hold. To see the connection to (6) write

$$t^{-1} C_t - \theta = t^{-1} M_t + t^{-1}\big(w(X_0) - w(X_t)\big) + t^{-1} \Phi_t, \qquad t \geqslant 0. \tag{10}$$

The law of large numbers for martingales applies to the first term on the right under rather general conditions. In the case of the diffusion

process, M has quadratic variation

$$\langle M\rangle_t = \int_0^t \left(b(X_s)\frac{d}{dx}w(X_s)\right)^2 ds, \quad t \geqslant 0.$$

A local martingale with continuous quadratic variation arises from a Wiener process by time scale transformation

$$M_t = \mathcal{W}_{\langle M\rangle_t}, \quad t \geqslant 0.$$

From the law of the iterated logarithm for the Wiener procees the next proposition is obtained.

PROPOSITION 2. *Let*

$$\lim_{t\to\infty} t^{-2}\langle M\rangle_t \log\log t = 0 \quad a.s. \ (in\ probability).$$

Then

$$\lim_{t\to\infty} t^{-1}M_t = 0 \quad a.s. \ (in\ probability).$$

The second term on the right in (10) is negligible provided that certain stability conditions are fulfilled. The third term is nonnegative if (9) holds. We thus get

$$\overline{\lim_{t\to\infty}}\, t^{-1}C_t \geqslant \theta \quad \text{a.s.}$$

under any control U.

On the other hand, denoting by $\hat{u}(x)$ the minimizer in (9), we have

$$\varphi\big(x, \hat{u}(x)\big) = 0, \quad x \in (-\infty, \infty).$$

Hence, under the feedback control

$$\hat{U} = \{\hat{u}(X_t),\ t \geqslant 0\},$$

we have

$$C_t - t\theta = M_t + w(X_0) - w(X_t), \quad t \geqslant 0.$$

This leads us to

$$\lim_{t\to\infty} t^{-1}C_t = \theta \quad \text{a.s.}, \tag{11}$$

$$\int_0^\infty E_x\big(c\big(X_t, \hat{u}(X_t)\big) - \theta\big)\,dt = \lim_{t\to\infty}(E_x C_t - t\theta) = w(x) - \lim_{t\to\infty} E_x w(X_t)$$

$$= w(x) + \text{const}, \tag{12}$$

assuming that under $\hat{U}$ the distribution of X_t stabilizes as $t\to\infty$.

The heuristic conclusion which we have obtained is that in (9) θ is the minimal average cost and $w(x)$ is the cost potential associated with the optimal feedback control. Abbreviate

$$A\big(\hat{u}(x)\big) = \hat{A}, \quad c\big(x, \hat{u}(x)\big) = \hat{c}(x).$$

Under $\hat{U}$ we have

$$dX_t = \hat{a}(X_t)\,dt + b(X_t)\,dW_t, \quad t \geqslant 0.$$

Consequently, X is Markovian, and the integral on the left of (12) is the potential associated with charge $\hat{c}(x) - \theta$.

Because of its relationship to other dynamic programming equations, we call (9) the *Bellman equation* for minimal average cost.

The feedback control $\hat{U}$ solves problem (6) for a particular system S. In situations when the mathematical description of S is not completely known to the controller, he has to seek controls U such that (6) holds for some class of systems, e.g., a parametrized family. Controls having this property are called *self-optimizing*.

Decomposition (10) is a convenient tool for investigation of self-optimizing controls. This is seen from the papers listed in the references. The directions for the use of (10) are summarized in the following points:

(i) Apply the limit theorems for martingales to M, and by estimating Φ examine the speed of convergence of $t^{-1}C_t$ to θ as $t \to \infty$.

(ii) Use the conclusions from (i) to classify the controls satisfying (6).

(iii) Consider the cases, when the specification of S involves parameters whose values are to be estimated from the trajectory observed. The control is accomplished by inserting the parameter estimate into the optimal feedback control. Give conditions for the self-optimizing property, and determine the speed of convergence for various estimation techniques like the maximum likelihood, the minimum contrast, and the recursive methods.

Let us first mention some points pertaining to the determination of cost potentials. Since $\hat{u}(x)$ is the minimizer of (9), we have

$$\hat{A}w(x) + \hat{c}(x) - \theta = 0, \quad x \in (-\infty, \infty). \tag{13}$$

(13) is a second order differential equation which has a two-dimensional set of solutions for each θ. Boundary conditions in $\pm\infty$ were introduced in [22] to pick up the right solution.

PROPOSITION 3. *Set*

$$Q(x) = \exp\Big\{2\int_0^x \hat{a}(y)\,b(y)^{-2}\,dy\Big\}, \quad x \in (-\infty, \infty).$$

Let

$$\int_{-\infty}^{\infty} b(x)^{-2}Q(x)\,dx < \infty, \qquad \int_{-\infty}^{\infty} b(x)^{-2}Q(x)\,|\hat{c}(x)|\,dx < \infty.$$

The average cost θ is the unique constant such that (13) *has a solution $w(x)$ satisfying*

$$\lim_{x\to\pm\infty} Q(x)\frac{d}{dx}w(x) = 0. \tag{14}$$

In the case of a countable state space, the cost potentials for controlled Markov chains were treated in [9], under a Liapunov type assumption. An analogous approach is applicable to continuous time processes with countably many states. To formulate the Liapunov condition we again use the cap on the symbols, referring to the chosen feedback control. Let $k \in I$ be a particular state, say $k = 1$. From $\hat{A}$ we construct the matrix

$$\hat{P} = \|\hat{p}(i,j)\|_{i,j\in I} = \begin{bmatrix} 0 & \hat{a}(1,2)/\hat{a}(1) & \hat{a}(1,3)/\hat{a}(1) & \dots \\ 0 & 0 & \hat{a}(2,3)/\hat{a}(2) & \dots \\ 0 & \hat{a}(3,2)/\hat{a}(3) & 0 & \dots \\ \dots & \dots & \dots & \dots \end{bmatrix}.$$

$\hat{P}$ is the transition probability matrix of the Markov chain imbedded in X until its first jump into state 1.

ASSUMPTION 1. *There exists a $y(j) \geqslant 0$, $j \in I$, such that*

$$|\hat{c}(i)| + 1 + \hat{a}(i)\Big(\sum_j \hat{p}(i,j)y(j) - y(i)\Big) \leqslant 0, \qquad i \in I, \tag{15}$$

$$\lim_{n\to\infty} \hat{P}^n y = 0. \tag{16}$$

PROPOSITION 4. *Let Assumption* 1 *hold. Then the average cost θ is the unique number such that there exists a $w(j)$, $j \in I$, satisfying*

$$\hat{A}w(i) + \hat{c}(i) - \theta = 0, \qquad i \in I,$$

together with

$$|w(j)| \leqslant \text{const}\, y(j), \qquad j \in I. \tag{17}$$

A corresponding statement about the Bellman equation is also valid.

PROPOSITION 5. *Let $\mathcal{U}$ be compact, let* (15) *and* (16) *hold for any feedback control, and let the left-hand side of* (15) *depend continuously on the control.*

Then the minimal average cost θ is the unique number for which the equation

$$\min_{u\in\mathcal{U}} \{A(u)w(i)+c(i,u)-\theta\} = 0, \quad i\in I, \tag{18}$$

has a solution fulfilling (17).

The origin of (18) is in the paper [1] by R. Bellman.

Returning to the diffusion processes, consider an example. Let the equation for the trajectory of S be

$$dX_t = -(U_t+aX_t)dt+dW_t, \quad t\geqslant 0, \tag{19}$$

and let

$$C_t = \int_0^t (X_s^2+U_s^2)ds, \quad t\geqslant 0.$$

Thus, we have the simplest case of a linear controlled process with a quadratic cost function. The Bellman equation is

$$\min_u \left\{\frac{1}{2}\frac{d^2}{dx^2}w(x)-(u+ax)\frac{d}{dx}w(x)+x^2+u^2-\theta\right\} = 0, \quad x\in(-\infty,\infty).$$

Solving it with regard to (14), one obtains

$$\theta = \sqrt{a^2+1}-a, \quad w(x) = \left(\sqrt{a^2+1}-a\right)x^2.$$

The optimal feedback control is

$$\hat{U}_t = \left(\sqrt{a^2+1}-a\right)X_t, \quad t\geqslant 0. \tag{20}$$

Under $\hat{U}$ the limit relation holds,

$$\lim_{t\to\infty} t^{-1}C_t = \theta \quad \text{a.s.} \tag{21}$$

Assume that the controller knows only that a is some positive number. He is thus unable to insert the value of a into (20). He still wants to have (21), and therefore he compiles a self-optimizing control replacing a in (20) by its estimate from the past trajectory. An estimate is

$$a_t^* = -\left(\int_0^{t-h} X_s dX_s+\int_0^{t-h} U_s X_s ds\right)\Big/\int_0^{t-h} X_s^2 ds, \quad t>h,$$

where h is a time-lag. Substituting from (19) we get

$$a_t^* = a-\int_0^{t-h} X_s dW_s \Big/ \int_0^{t-h} X_s^2 ds = a-\mathcal{W}\left(\int_0^{t-h} X_s^2 ds\right)\Big/\int_0^{t-h} X_s^2 ds, \quad t>h,$$

where $\{\mathscr{W}(z),\ z \geqslant 0\}$ is a Wiener process. From the strong law of large numbers for the Wiener process follows

$$\lim_{t\to\infty} a_t^* = a \quad \text{a.s.}$$

Hence, the control

$$U_t = \left(\sqrt{a_t^{*2}+1} - a_t^*\right) X_t, \quad t > h, \tag{22}$$

approaches to (20) as $t\to\infty$. A stronger statement than (21) can be proved.

PROPOSITION 6. *$(C_t - t\theta)/\sqrt{t}$ has, under either of the controls* (20) *and* (22) *asymptotically normal distribution $N(0, \sigma^2)$ as $t\to\infty$, with*

$$\sigma^2 = \left(\sqrt{a^2+1} - a\right)^2 / \sqrt{a^2+1}.$$

The following relation holds,

$$\overline{\lim_{t\to\infty}} \pm (C_t - t\theta)/\sqrt{2t \ln\ln t} = \sigma \quad a.s.$$

We see that under the self-optimizing control (22) the cost is subjected to the central limit theorem and to the law of the iterated logarithm with the same parameters as under the optimal feedback control (20).

Proposition 6 shows the pattern of more general theorems based on the decomposition (10). In proving them one has to verify the corresponding limit theorem for the martingale M, and then to show that the control tends to the optimal stationary one sufficiently quickly to make the remaining terms negligible. The quadratic variation $\langle M\rangle_t$ is a functional of the same type as the cost, formula (5) confirms it. Thus, to prove

$$\lim_{t\to\infty} t^{-1} \langle M\rangle_t = \sigma^2 \quad \text{a.s.}, \tag{23}$$

a decomposition analogous to (10) is used. For the one-dimensional diffusion process, (23) implies

$$M_t = \mathscr{W}_{\langle M\rangle_t} \sim \mathscr{W}_{\sigma^2 t} \quad \text{as } t\to\infty.$$

Consequently, the limit theorems for the Wiener process imply those for the martingale M. The procedure is more complex for a countable state system. The applicability of the limit theorems to the discrete parameter martingale $\{M_n,\ n = 0, 1, \ldots\}$ is to be examined. Theorems like the next proposition are at hand.

PROPOSITION 7. *Set $Y_k = M_{k+1} - M_k$, $k = 0, 1 \ldots$ Let*

$$\lim_{n\to\infty} n^{-1} \sum_{k=0}^{n-1} E\{Y_k^2 \mid \mathscr{F}_k^X\} = \sigma^2 \tag{24}$$

in probability, where σ^2 *is a constant, and let*

$$\overline{\lim_{n\to\infty}}\, n^{-1} \sum_{k=0}^{n-1} EY_k^4 < \infty.$$

Then $M_n/\sqrt{n}$ *has asymptotically normal distribution* $N(0, \sigma^2)$ *as* $n\to\infty$. *If, moreover,* (24) *holds almost surely, then*

$$\overline{\lim_{n\to\infty}}\, \pm M_n/\sqrt{(2n\ln\ln n)} = \sigma.$$

The verification of (24) is reduced to the investigation of a cost functional in virtue of the equality

$$E\{Y_k^2 \mid \mathscr{F}_k^X\} = E\Big\{\int_k^{k+1} c_2(X_t, U_t)\,dt \mid \mathscr{F}_k^X\Big\}, \quad k = 0, 1, \ldots,$$

where

$$c_2(i, u) = \sum_j a(i, j; u)\big(w(j) - w(i)\big)^2, \quad i \in I, \quad u \in \mathscr{U}.$$

The left-hand side of (24) can be replaced by $n^{-1}\int_0^n c_2(X_s, U_s)\,ds$.

The contents of the papers in the references can be seen from their titles. The results of [7] and of [23] were widely used in the present survey.

References

[1] Bellman R., A Markovian Decision Process, *J. of Math. and Mech.* **6** (1957), pp. 679–684.

[2] Borkar V. and Varaiya P., Adaptive Control of Markov Chains I: Finite Parameter Set, *IEEE Trans. Autom. Control* **24** (1979), pp. 953–957.

[3] Borkar V. and Varaiya P., *Adaptive Control of Markov Chains,* Lecture Notes Control and Inform. Sc. **16**, pp. 294–296, Springer-Verlag, 1979.

[4] Brychta J., *Recursive Parameter Estimates in Controlled Markov Chains,* Thesis, Charles University, Prague, 1977 (in Czech).

[5] Doshi B. T., Adaptive Control of a Production Inventory System. *J. Appl. Probability* **18** (1981), pp. 204–215.

[6] Doshi B. T. and Shreve S. E., Strong Consistency of a Modified Maximum Likelihood Estimator for Controlled Markov Chains, *J. Appl. Probability* **17** (1980), pp. 726–734.

[7] Dufková-Lánská V., On Controlled One-Dimensional Diffusion Processes with Unknown Parameter, *Adv. Appl. Probability* **9** (1977), pp. 105–124.

[8] Georgin J. P., *Estimation et contrôle des chaînes de Markov sur des espaces arbitraires,* Lecture Notes Math. 636, pp. 71–113, Springer-Verlag, 1978.

[9] Hordijk A., *Dynamic Programming and Markov Potential Theory,* Math. Centrum, Amsterdam, 1974.

[10] Kolonko M., *Dynamische Optimierung unter Unsicherheit in einem Semi-Markoff-Modell mit abzählbarem Zustandsraum*, Thesis, Friedrich–Wilhelm University, Bonn, 1980.

[11] Kolonko M., Strongly Consistent Estimation in a Controlled Markov Renewal Model. *J. Appl. Probability* **19** (1982), pp. 532–545.

[12] Kolonko M. and Schäl M., Optimal Control of Semi-Markov Chains Under Uncertainty with Application to Queueing Models, *Proc. Operations. Res.* **9** (1980), pp. 430–435.

[13] Kunderová P., On Limit Properties of the Reward From a Markov Replacement Process, *Acta Univ. Pal. Olomucensis f. rer. nat.* **69** (1981), pp. 133–146.

[14] Kurano M., Discrete Time Markovian Decision Processes With an Unknown Parameter-Average Return Criterion, *J. Operations Res. Soc. Japan* **15** (1972), pp. 67–76.

[15] Mandl P., On the Control of a Markov Chain in the Presence of Unknown Parameters, *Trans. 6th Prague Conf. Inform. Theory etc.* 1971, pp. 601–612, Academia, Prague, 1973.

[16] Mandl P., *An Application of Itô's Formula to Stochastic Control Systems*, Lecture Notes Math. **294**, pp. 8–13, Springer-Verlag, 1972.

[17] Mandl P., Asymptotically Optimal Controls of Markov Processes Involving Unknown Parameters, *Proc. Prague Symp. Asymptotic Stat. 1973*, Vol. II, pp. 247–256, Charles University, Prague, 1974.

[18] Mandl P., Estimation and Control in Markov Chains, *Adv. Appl. Probability* **6** (1974), pp. 40–60.

[19] Mandl P., On the Adaptive Control of Countable Markov Chains, *Probability Theory*, Banach Center Publ. **5**, pp. 159–173, PWN, Warsaw, 1979.

[20] Mandl P., *On Self-Optimizing Control of Markov Processes*, Banach Center Publ., to appear.

[21] Menyhértová-Skřivánková V., On Adaptive Replacement Policies, *Kybernetika* (Prague) **16** (1980), pp. 512–525.

[22] Morton R., On the Optimal Control of Stationary Diffusion Processes with Inaccessible Boundaries and No Discounting, *J. Appl. Probability* **8** (1971), pp. 551–560.

[23] Romera R., *Adaptive Control of Countable Markov Processes*, Thesis, Polytechnical University, Madrid, in preparation.

[24] Sagalovsky B., Adaptive Control and Parameter Estimation in Markov Chains: A Linear Case, *IEEE Trans. Autom. Control* **27** (1982), pp. 414–419.

CHARLES UNIVERSITY
PRAGUE
CZECHOSLOVAKIA

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

DANIEL W. STROOCK*

Stochastic Analysis and Regularity Properties of Certain Partial Differential Operators

0. Introduction

Let $V_0, \ldots, V_d \in C^\infty_\uparrow(R^N; R^N)$ ($C^\infty_\uparrow$ stands for the class of infinitely differentiable functions all of whose derivatives have, at most, polynomial growth) having uniformly bounded first order derivatives. Thinking of V_k as being the directional derivative

$$\sum_{i=1}^{N} V_k^i(y)\,\frac{\partial}{\partial y_i},$$

define L on $C^2(R^N)$ by

$$L = \tfrac{1}{2}\sum_{k=1}^{d} V_k^2 + V_0. \tag{0.1}$$

The research about which this note is a report studies various properties of the operator L by means of the diffusion process associated with L.

To be more precise, denote by Θ the space of continuous maps $\theta\colon [0, \infty)\to R^d$ which start at the origin. Give Θ the topology of uniform convergence on compacts, denote by $\mathscr{B}$ the Borel field over Θ, and let $\mathscr{W}$ be the standard Wiener measure on $(\Theta, \mathscr{B})$, i.e., for all $0 \leqslant s < t$ and $\Gamma \in \mathscr{B}_{R^N}$:

$$\mathscr{W}\big(\theta(t) \in \Gamma \mid \mathscr{B}_s\big) = \big(2\pi(t-s)\big)^{-d/2} \int_\Gamma \exp\big(-|y-\theta(s)|^2/2(t-s)\big)\,dy$$

(a.s., $\mathscr{W}$), where $\mathscr{B}_s$ is the σ-algebra over Θ generated by the maps $\theta \in \Theta\to\theta(u) \in R^d$ for $0 \leqslant u \leqslant s$. Following Itô, we construct a diffusion

* This research was sponsored in part by N.S.F. Grant MCS 80-07300.

associated with L by solving the stochastic integral equation:

$$X(T, x, \theta) = x + \sum_{k=1}^{d} \int_0^T V_k\big(X(t, x, \theta)\big) \circ d\theta_k(t) + \int_0^T V_0\big(X(t, x, \theta)\big)\, dt, \quad T \geqslant 0, \tag{0.2}$$

for each $x \in R^N$. (Loosely speaking, (0.2) can be thought of as an ordinary integral equation into which $\theta(\cdot)$ enters as a "control".) Of course, since $\mathscr{W}$-almost no $\theta(\cdot)$ is anywhere of bounded variation, the meaning of $d\theta_k(t)$ integration has to be specified. The presence of "$\circ$" in our $d\theta_k(t)$ integral means that we have adopted the Stratanovič convention so that the solution of (0.2) is, in a weak sense, the limit of solutions to the equations obtained from (0.2) by replacing the path $\theta(\cdot)$ with its own mollification. It was shown by Itô that for each $x \in R^N$, (0.2) uniquely determines, up to a $\mathscr{W}$-null set, a map $X(\cdot, x)\colon \Theta \to C\big([0, \infty); R^N\big)$ with the property that for each $T \geqslant 0$ the map $X(T, x)\colon \Theta \to R^N$ is measurable with respect to the $\mathscr{W}$-completion $\overline{\mathscr{B}}_T$ of $\mathscr{B}_T$. Moreover, it is now well-known (cf. [6]) that a careful selection of solutions leads to a version of X such that $(t, x) \to X(t, x)$ is $\mathscr{W}$-almost surely an element of $C^{0,\infty}\big([0, \infty) \times R^N; R^N\big)$ (i.e. $\partial^{|\alpha|} X / \partial x^\alpha$ exists and is continuous on $[0, \infty) \times R^N$ for all multi-indices α). In fact, at the same time one can show that for all $n \geqslant 0$, $T > 0$, $R > 0$, and $p \in [0, \infty)$ there is an estimate of the form:

$$E^{\mathscr{W}}\left[\sup_{0 \leqslant t \leqslant T} \sup_{|x| \leqslant R} \max_{\alpha \leqslant n} \left| \frac{\partial^{|\alpha|} X}{\partial x^\alpha}(t, x) \right|_{R^N}^p \right]^{1/p} \leqslant C(n, p)(1 + R)^{\lambda(n,p)} e^{\mu(n,p)T} \tag{0.3}$$

(cf. [15]).

The basic connection between the operator L in (0.1) and the equation (0.2) is the one discovered by Itô. Namely, if the probability measure $P(T, x, *)$ on R^N is defined by

$$P(T, x, *) = \mathscr{W} \circ \big(X(T, x)\big)^{-1} \tag{0.4}$$

(i.e., $\int \varphi(y) P(T, x, dy) = E^{\mathscr{W}}\big[\varphi\big(X(T, x)\big)\big]$) then

$$\int \varphi(y) P(T, x, dy) - \varphi(x) = \int_0^T \left(\int L\varphi(y)\, P(t, x, dy) \right) dt \tag{0.5}$$

for all $(T, x) \in [0, \infty) \times R^N$ and $\varphi \in C_b^2(R^N)$ (C_b^n means bounded continuous derivatives up to nth order). In other words, for each $x \in R^N$, $P(\cdot, x, *)$

is a generalized solution of the Fokker–Planck equation

$$\frac{\partial P}{\partial t}(t, x, *) = L^* P(t, x, *), \quad t > 0, \tag{0.6}$$

$$P(0, x, *) = \delta_x(*),$$

where L^* is the formal adjoint of L and in this equation acts on $P(t, x, *)$ as a function of "$*$". Moreover, as a consequence of the uniqueness of solutions of (0.2), one can show that $P(T, x, *)$ satisfies the Chapman–Kolmogorov equation

$$P(s+t, x, \Gamma) = \int P(t, y, \Gamma) P(s, x, dy). \tag{0.7}$$

Combining (0.3) with (0.7) and defining

$$P_t\varphi(x) = \int \varphi(y) P(t, x, dy), \tag{0.8}$$

we conclude that $\{P_t\colon t > 0\}$ is a weakly continuous semi-group on $C_b(R^N)$ and that $C_\uparrow^n(R^N)$ is $\{P_t\colon t > 0\}$ invariant for each $n \geqslant 0$. Finally using (0.5), we see that for any $\varphi \in C_\uparrow^2(R^N)$, the function $u(t, x) = P_t\varphi(x)$ is an element of $C^{1,2}\big([0, \infty) \times R^N\big)$ and solves the initial value Cauchy problem:

$$\begin{aligned} &\frac{\partial u}{\partial t} = Lu, \quad t > 0, \\ &u(0, \cdot) = \varphi(\cdot). \end{aligned} \tag{0.9}$$

Up to this point, everything that we have discussed is the natural extension of what is well-known about the deterministic case when $L = V_0$ and $X(\cdot, x)$ is simply the integral curve of V_0 starting at x. In particular, so far there is nothing more that we can say in the case where L is elliptic (i.e., span $\{V_1(y), \ldots, V_d(y)\} = R^N$ for all $y \in R^N$) than we can when $L = V_0$. The traditional way in which probabilists have incorporated elliptic regularity theory into their field has been to use "uniqueness" to identify their probabilistically constructed quantities with analytically constructed smooth solutions to equations like (0.9). (The point here is that from (0.5) one can easily show that any reasonable solution u of (0.9), must be given by $u(t, x) = P_t\varphi(x)$.) In this way, quantities like $P(t, x, *)$ are seen to have whatever regularity properties the theory of partial differential equations predicts for the fundamental solution of (0.9). Obviously, this identification procedure yields no probabilistic insight into the origins of elliptic regularity.

1. A probabilistic approach to elliptic regularity

In order to see how one might study elliptic regularity with probabilistic machinery, consider the problem of finding a condition which will guarantee that the measure $P(T, x, *)$ in (0.4) admits a smooth density. Since (0.4) explicitly displays $P(T, x, *)$ as the image of $\mathscr{W}$ under a map, it is tempting to reason as follows:

(i) $\mathscr{W}$ is a "smooth" measure in the sense that it has quasi-invariance properties when translated by paths having square integrable derivatives. This is the famous theorem of Cameron and Martin (cf. [2]).

(ii) Since $X(T, x, \theta)$ is described by an equation in which $\theta(\cdot)$ enters as a "differential", $X(T, x, \theta)$ should be "smooth" under translations by paths having square integrable derivatives.

(iii) If some appropriate notion of the "Jacobian" of $\theta \to X(T, x, \theta)$ is sufficiently non-degenerate, then one should be able to "lift" translation in R^N to translation in Θ.

Combining (i), (ii), and (iii), one is led to guess that as "the image of a smooth measure under a non-degenerate smooth map" $P(T, x, *)$ ought to admit a smooth density. This, in broad outline, is the idea proposed by P. Malliavin [9] and [10] for studying the regularity properties of $P(T, x, *)$. Not only is the basic idea Malliavin's but also Malliavin is the person responsible for suggesting a method of transforming the basic idea into rigorous mathematics. Since Malliavin's groundbreaking articles, there have been various schemes worked out for carrying out his program. A sampling of the work in this direction may be found in: [14], [13], [5], [1], [16] and [7].

It is not surprising that the difficult part of Malliavin's program is that of relating the nondegeneracy condition of (iii) to a nondegeneracy statement about the vector fields $V_0, \ldots, V_d$. Indeed, the smoothness demanded in (ii) is a relatively simple consequence of the smoothness of these vector fields and has nothing to do with ellipticity. Thus, elliptic regularity results are obtained by this methodology as a consequence of finding out how non-degeneracy of the V_k's is manifested in non-degeneracy of the Jacobian mentioned in (iii).

2. Malliavin's covariance matrix

After carrying out the program outlined in Section 1, one finds that the nondegeneracy required in (iii) is measured by the *Malliavin covariance*

matrix $\tilde{A}(T, x)$ given by the formula

$$\tilde{A}(T, x) = \sum_{k=1}^{d} \int_0^T J^{-1}(t, x) V_k\big(X(t, x)\big) \otimes J^{-1}(t, x) V_k\big(X(t, x)\big)\, dt. \quad (2.1)$$

In (2.1), $J(t, x)$ is the Jacobian $\partial X(t, x)/\partial x$ of the map $x \to X(T, x)$ and $J^{-1}(t, x)$ is its inverse. (Since $J(t, x)$ satisfies a simple linear equation which is the stochastic analogue of the familiar one associated with the flow generated by a vector field, there is no difficulty showing that $J(t, x)$ is invertible.) The fundamental theorem connecting $\tilde{A}(T, x)$ with regularity results about $P(T, x, *)$ is that

$$P(T, x, dy) = p(T, x, y)\, dy \quad (2.2)$$

with $p(T, x, \cdot) \in \mathscr{S}(R^N)$ (the Schwartz class) if

$$1/\det \tilde{A}(T, x) \in \bigcap_{q \in [1, \infty)} L^q(\mathscr{W}). \quad (2.3)$$

A simple argument shows that (2.3) is equivalent to

$$\sup_{\eta \in S^{N-1}} \big\| 1/\big(\eta, \tilde{A}(T, x)\eta\big)_{R^N} \big\|_{L^q(\mathscr{W})} < \infty, \qquad q \in [1, \infty). \quad (2.4)$$

Since

$$\big(\eta, \tilde{A}(T, x)\eta\big)_{R^N} = \sum_{k=1}^{d} \int_0^T \big(J^{-1}(t, x) V_k\big(X(t, x)\big), \eta\big)_{R^N}^2 dt, \quad (2.5)$$

the facts that $X(0, x) = x$ and $J(0, x) = I$ make it clear that (2.3), and therefore (2.2), holds when span $\{V_1(x), \ldots, V_d(x)\} = R^N$. Of course, span $\{V_1(x), \ldots, V_d(x)\} = R^N$ is equivalent to saying that L is strictly elliptic at x, and so one should expect regularity under this condition.

Getting away from ellipticity requires more work, and the idea here again originates with Malliavin. What one does is expand each $J^{-1}(t, x) \times$ $\times V_k\big(X(t, x)\big)$ in a "stochastic Taylor series" around $t = 0$. The expression at which one thereby arrives is

$$J^{-1}(t, x) V_k\big(X(t, x)\big) = \sum_{\|\nu\| \leqslant l-1} V_k^{(\nu)}(x) \theta^{(\nu)}(t) + R_l(t, x; V_k). \quad (2.6)$$

To explain (2.6) requires the introduction of a little notation. First, let $\mathscr{A}$ be the index set $\{\emptyset\} \cup \bigcup_{l=1}^{\infty} (\{0, \ldots, d\})^l$. Given $\nu \in \mathscr{A}$, let $|\nu| = 0$ if $\nu = \emptyset$ and $|\nu| = l$ if $\nu \in (\{0, \ldots, d\})^l$. Also, if $|\nu| \geqslant 1$, set $[\nu] = \{1 \leqslant j \leqslant l;\ \nu_j = 0\}$ and $|[\nu]| = \text{card}\,([\nu])$. The norm $\|\nu\|$ is then defined by

$$\|\nu\| = |\nu| + |[\nu]|.$$

Next, if $V \in C_{\uparrow}^{\infty}(R^N; R^N)$ and $\nu \in \mathscr{A}$, then $V^{(\nu)}$ is defined inductively by: $V^{(\emptyset)} = V$ and, for $\nu = (\nu_1, \ldots, \nu_l)$ with $l \geqslant 1$, $V^{(\nu)} = [V_{\nu_l}, V^{(\nu')}]$ where

$$\nu' = \begin{cases} \emptyset & \text{if } l = 1, \\ (\nu_1, \ldots, \nu_{l-1}) & \text{if } l \geqslant 2. \end{cases}$$

(The notation $[X, Y]$ denotes the commutator or Lie product of the vector fields X and Y.) Finally, set $\theta_0(t) = t$ and define $\theta^{(\nu)}(t)$ inductively by: $\theta^{(\emptyset)}(t) = 1$; $\theta^{(\nu)}(t) = \theta_{\nu_1}(t)$ if $|\nu| = 1$; and $\theta^{(\nu)}(t) = \int_0^t \theta^{(\nu')}(s) \circ d\theta_{\nu_l}(s)$ for $\nu = (\nu_1, \ldots, \nu_l)$ with $l \geqslant 2$.

Simple time-scaling considerations show that (2.6) is a good Taylor expansion in the sense that for any $\varepsilon > 0$, $\theta^{(\nu)}(\varepsilon t)$, $t \geqslant 0$, has the same distribution under $\mathscr{W}$ as $\varepsilon^{\|\nu\|/2}\theta^{(\nu)}(t)$, $t \geqslant 0$. Moreover, standard martingale estimates allow one to show that $\int_0^T |R_l(t, x; V_k)|^2 dt$ is on the order of T^{l+1}. Less standard, and in fact at the heart of this analysis, is the fact that

$$\frac{\sum_{k=1}^{d} \int_0^T \Big(\sum_{\|\nu\| \leqslant l-1} \big(V_k^{(\nu)}(x), \eta\big)_{R^N} \theta^{(\nu)}(t) \Big)^2 dt}{\sum_{k=1}^{d} \sum_{\|\nu\| \leqslant l-1} \big(V_k^{(\nu)}(x), \eta\big)_{R^N}^2}$$

is on the order of T^l. Accepting these two estimates, plugging them into (2.6), and using the resulting expression to estimate the right-hand side of (2.5), one concludes that (2.3) holds so long as span $\{V_k^{(\nu)}(x)\colon 1 \leqslant k \leqslant d$ and $\|\nu\| \leqslant l-1\} = R^N$ for some $l \geqslant 0$. In particular, one shows in this way that (2.2) is true whenever span $\{V_k^{(\nu)}(x)\colon 1 \leqslant k \leqslant d$ and $\nu \in \mathscr{A}\} = R^N$.

The theorem just discussed is the culmination of efforts by several authors. It all began with Malliavin in [9] and [10]. Many important details were added by Watanabe [5]. Both these authors restricted their attention to the case when the Lie algebra generated by $V_1, \ldots, V_d$ has full rank at x. The first attempt to take V_0 into account appears in [11], where a special case is handled. Bismut [1] was the first to try for the full result, although he settled for the existence of a density without further regularity. Finally, Kusuoka and I completed the program. One version of our work appears in Section 8 of [17]. A second version, more along the lines just described, will be forthcoming in our article [8], Part II.

Before moving on, it should be remarked that although the preceding regularity result about $P(T, x, *)$ might be thought to follow directly from the renowned theorem of Hörmander [4], in fact Hörmander's

theorem predicts slightly less. To be precise, Hörmander's theorem says that $P(T, x, *)$ will admit a smooth density on the set U consisting of those $y \in R^N$ satisfying span $\{V_k^{(\nu)}(y)\colon 1 \leqslant k \leqslant d \text{ and } \nu \in \mathscr{A}\} = R^N$; it says nothing about $P(T, x, *)$ outside of U. Although this may seem to be a rather trivial quibble, it exemplifies an important distinction between the present approach to regularity theory and the more traditional analytic ones. Namely, this stochastic analytic approach is inherently global, whereas most analytic theories are based on local considerations. The global nature of the present approach has allowed us to show that the regularity due to non-degeneracy at x propagates to the whole of R^N. Section 3 will be devoted to another example in which this global nature of the stochastic analysis is exploited.

There are several directions in which one can go, starting from our basic result about $\tilde{A}(t, x)$. For example, one can try to recover Hörmander's theorem about the hypoellipticity of L. What one finds is that if $u \in \mathscr{D}(R^N)$ then the wave front set W.F. (u) of u is contained in the union of W.F. (Lu) with $\{(y, \xi)\colon \xi \perp V_k^{(\nu)}(y)$ for all $1 \leqslant k \leqslant d$ and $\nu \in \mathscr{A}\}$. Of course, if span $\{V_k^{(\nu)}(y)\colon 1 \leqslant k \leqslant d \text{ and } \nu \in \mathscr{A}\} = R^N$ for each y in the open set U, then it follows that $u \in C^\infty(U)$ whenever $Lu \in C^\infty(U)$. In order to get the full Hörmander theorem from this statement, one can use the fact that if the Lie algebra generated by $\{V_0, \ldots, V_d\}$ has full rank at x^0, then there is a $\varrho \in C_0^\infty(R^N)$ such that $\varrho(x^0) > 0$ and span $\{\tilde{V}_k^{(\nu)}(x^0)\colon 1 \leqslant k \leqslant d \text{ and } \nu \in \mathscr{A}\} = R^N$ when $\tilde{V}_0 = \varrho^2 V_0 - \frac{1}{2}\sum_{k=1}^{d} \varrho\big(V_k(\varrho)\big) V_k$, $\tilde{V}_k = \varrho V_k$ for $1 \leqslant k \leqslant d$, and $\tilde{V}_k^{(\nu)}$ is defined accordingly. Since $\varrho^2 L = \frac{1}{2}\sum_1^d \tilde{V}_k^2 + \tilde{V}_0$, it is now clear how to deduce that L is hypoelliptic on the set of y at which the Lie algebra generated by $\{V_0, \ldots, V_d\}$ has full rank. This is the theorem of Hörmander.

In connection with questions about the hypoellipticity of L, it is interesting to note that one can use stochastic analysis to obtain certain criteria guaranteeing hypoellipticity even when Hörmander's criterion fails. Without going into details, suffice it to say that these criteria are stated in terms of the rate at which $\mathscr{V}_l(x) \equiv \inf_{\eta \in S^{N-1}} \sum_{k=1}^{d} \sum_{\|\nu\| \leqslant l} \big(V_k^{(\nu)}(x), \eta\big)^2$ becomes positive as x moves away from a "thin set" on which $\mathscr{V}_l$ vanishes. (Of course, such results are interesting only for non-analytic V_k's.) The results alluded to here are related to some of the criteria found by Oleinik and Radekevich in [11].

The theorems on which the preceding two paragraphs are based will be presented in the forthcoming article [8], Part II. In the same article,

we will also present certain subelliptic estimates related to but slightly different from those obtained by Rothschild and Stein [12].

3. Some global considerations

For the most part, Section 2 was devoted to applications of stochastic analysis to local regularity questions about L. However, as was mentioned there, the stochastic analytic approach is inherently global in nature and can be made to yield interesting non-local results. The purpose of this section is to explain one such application.

Let M be a compact C^∞-submanifold of R^N and assume that for each $0 \leqslant k \leqslant d$, $V_k|_M$ is a vector field on M. Then it is quite easy to see that, for each $x \in M$, $\mathscr{W}\big(X(t,x) \in M,\ t \geqslant 0\big) = 1$. Thus, we can think of $\{P_t\colon t > 0\}$ as a strongly continuous semi-group on $C(M)$. Suppose that one wants to find conditions which guarantee that $\{P_t\colon t > 0\}$ is uniformly ergodic in the sense that there exists a probability measure μ on M and an $\varepsilon > 0$ such that $\|P_t - \mu I\|_{\mathrm{op}} \leqslant Ce^{-\varepsilon t}$, $t \geqslant 0$, for some $C < \infty$. One condition which one needs is that for each non-negative $\varphi \in C(M) \setminus \{0\}$ and each $x \in M$ there exists a $T > 0$ such that $P_T\varphi(x) > 0$. This is a *recurrence condition* and is equivalent to the condition that for each open $U \neq \emptyset$ and $x \in M$, $\mathscr{W}\big((\exists t > 0)X(t,x) \in U\big) > 0$. In addition to recurrence, one needs to know that $\{P_t\colon t > 0\}$ satisfies a mild regularity condition. One such regularity condition is the Kakutani–Yoshida version of *Doeblin's condition* (cf. [3]), namely: there is a $T > 0$ and a compact $K_T\colon C(M) \to C(M)$ such that $\|P_T - K_T\|_{\mathrm{op}} < 1$. Recurrence together with Doeblin's condition guarantees uniform ergodicity. To understand what the stochastic analysis discussed in Section 1 has to say about Doeblin's condition, recall that regularity properties of P_T follow from non-degeneracy of the Malliavin covariance matrix $\tilde{A}(T,x)$. It turns out that the mild regularity required by Doeblin is satisfied as soon as one knows that for all $x \in M$ and each $\eta \in S^{N-1}$ which is tangent to M at x there exists a $T > 0$ such that $\mathscr{W}\big(\langle \eta, \tilde{A}(T,x)\eta\rangle_{R^N} > 0\big) > 0$. In order to phrase these conditions without reference to Wiener measure, we use the "support theorem" in [18]. To be precise, for each $\psi \in C\big([0,\infty);\ R^d\big)$ and $x \in M$, define $\Phi(\cdot, x; \psi) \in C\big([0,\infty);\ M\big)$ by:

$$\Phi(T,x;\psi) = x + \sum_{k=1}^{d} \int_0^T V_k\big(\Phi(t,x;\psi)\big)\psi_k(t)\,dt + \int_0^T V_0\big(\Phi(t,x;\psi)\big)\,dt, \qquad T \geqslant 0.$$

Then the recurrence condition is equivalent to the existence, for each open $U \neq \emptyset$ in M and each $x \in M$, of a $\psi \in C([0, \infty); R^d)$ such that $\Phi(\cdot, x; \psi)$ enters U. The condition on $\tilde{A}(\cdot, x)$ is equivalent to the existence, for each $x \in M$ and each $\eta \in S^{N-1}$ tangent to M at x, of a $\psi \in C([0, \infty); R^d)$ such that

$$\sum_{k=1}^{d} \int_0^\infty \left(\left(\frac{\partial \Phi(t, x; \psi)}{\partial x} \right)^{-1} V_k(\Phi(t, x; \psi)), \eta \right)_{R^N}^2 dt > 0.$$

Obviously, the preceding criteria are based on global properties of the V_k's. Thus, this application demonstrates the power of stochastic analysis to handle global regularity questions. A related global question, and one to which so satisfactory a solution has not yet been found, is the problem of global hypoellipticity. A joint paper with Kusuoka about these and other global questions will be forthcoming [8], Part III.

References

[1] Bismut J. M., Martingales, the Malliavin Calculus, and Hörmander's Theorem. In: D. Williams (ed.), *Stochastic Integrals*, Lecture Notes in Math. **851**, Springer-Verlag, 1980, pp. 85–109.

[2] Cameron R. H. and Martin W. T., Transformations of Wiener Integrals under General Class Transformations, *Trans. Amer. Math. Soc.* **58** (1945), pp. 184–219.

[3] Dunford N. and Schwartz J., *Linear Operators*, vol. I, Interscience Publ. Co., J. Wiley, N.Y., 1957.

[4] Hörmander L., Hypoelliptic Second Order Differential Equations, *Acta Math.* **119** (1967), pp. 147–171.

[5] Ikeda N. and Watanabe S., *Stochastic Differential Equations and Diffusion Processes*, North Holland Math. Library, Amsterdam–Oxford–N.Y., 1981.

[6] Kunita H., On the Decomposition of Solution to Stochastic Differential Equations. In: D. Williams (ed.), *Stochastic Integrals*, Lecture Notes in Math. **851**, Springer-Verlag, 1980, pp. 213–255.

[7] Kusuoka S. and Stroock D., Applications of the Malliavin Calculus, Part I, to appear in *Proc. Taniguchi Conf. at Katata, 1982*, Kinokuniya Publ. Co., Tokyo.

[8] Kusuoka S. and Stroock D., *Applications of the Malliavin Calculus*. Parts II and III, in preparation.

[9] Malliavin P., Stochastic Calculus of Variations and Hypoelliptic Operators. In: *Proc. Intnall. Conf. on Stoch. Differential Eqs. in Kyoto (1976)*, Kinokuniya Publ. Co., Tokyo (Wiley, N.Y.), pp. 195–263.

[10] Malliavin P., C^k-hypoellipticity with Degeneracy. In: A. Friedman and M. Pinsky (eds.), *Stochastic Analysis*, Academic Press, N. Y., 1978.

[11] Oleinik O. A. and Radekevich, *Second Order Equations with Non-negative Characteristic Form*, Amer. Math. Soc., Providence R. I., and Plenum Press, N. Y. 1973.

[12] Rothschild L. and Stein E., Hypoelliptic Differential Operators and Nilpotent Groups, *Acta Math.* **137** (1976), pp. 247–320.

[13] Shigekawa I., Derivatives of Wiener Functionals and Absolute Continuity of Induced Measures, *J. Math. Kyoto Univ.* **20** (2) (1980), pp. 263–289.

[14] Stroock D., The Malliavin Calculus and its Applications to Second Order Parabolic Differential Equations, I, II, *Math. Systems Theory* **14** (1981), pp. 25–65 and 141–171.

[15] Stroock D., *Topics in Stochastic Differential Equations*, Tata Inst. Lec. Notes in Math. (distributed by Springer-Verlag), 1982.

[16] Stroock D., The Malliavin Calculus, a Functional Analytic Approach, *J. Functional Analysis* **44** (2) (1981), pp. 212–257.

[17] Stroock D., Some Applications of Stochastic Analysis to Partial Differential Equations. In: P. L. Hennequin (ed.), *École d'Été de Probabilitiés de San Flour 1981*, Lecture Notes in Math., Springer-Verlag.

[18] Stroock D. and Varadhan S. R. S., On the Support of Diffusion Processes, with Applications to the Strong Maximum Principle. In: *Proc. 6th Berkeley Symp. on Math. Stat. and Prob.*, vol. III, 1970, pp. 333–360.

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

S. WATANABE

Excursion Point Processes and Diffusions

1. Introduction

About forty years have passed since K. Itô introduced stochastic integrals and stochastic differential equations. Since then, the theory of differential-integral calculus for sample paths of stochastic processes, often called *stochastic analysis* or *stochastic calculus*, has been extensively developed and has proved to be one of the most useful methods in the theory of stochastic processes and its applications.

In this theory, point processes and stochastic integrals based on them play an important role in describing the discontinuities of semimartingales. But the theory of point processes can also be applied to continuous semimartingales. A typical example is the case of Brownian motions in which the collection of all excursions of Brownian sample paths, from the origin to the origin, say forms a Poisson point process on a function space, called a Poisson point process of Brownian excursions (Itô [2]). This point process is useful in the study of fine structure of Brownian sample paths; a typical example of this is a simple proof of Lévy's down-crossing theorem as was given in [1] and independently by Maisonneuve [5], (cf. e.g. Williams [8], Rogers [6], Kasahara [4] and [1] for related topics).

The purpose of this report is to present some general results on constructing a "whole" semimartingale starting from "pieces" of the semimartingale to be constructed. Such results, applied to the case of well-known semimartingales like Brownian motions, yield a decomposition of the process into pieces like excursions, which, as we mentioned, provides a useful method in studying the properties of sample paths.

Also these results can be applied to construction problems. In [7] (cf. also [1], Chap. IV, Sec. 7) we have constructed diffusion processes corresponding to Wentzell's boundary conditions by means of Poisson

point processes of Brownian excursions. There we constructed some stochastic processes but the question as to how they are related to the given analytical data remains unanswered. By our results obtained here, we can write down stochastic differential equations governing these processes explicitly and thus show that they are diffusions corresponding actually to the analytical data.

2. Construction of Brownian motion from pieces

Here we discuss a simple case to see the idea. Suppose we are given a σ-finite but infinite measure space $(W, \mathscr{B}_W, n)$ and also a right-continuous increasing family $(\mathscr{B}_t)_{t>0}$ of sub σ-fields of $\mathscr{B}_W$. Though we are dealing with an infinite measure space, the notion of stochastic processes can be defined similarly as in the case of probability measure spaces. So let $F(u) = F(u, w)$, $u \in [0, \infty)$, $w \in W$, be a real continuous stochastic process adapted to $(\mathscr{B}_t)$ with the following properties:

(F. 1) For a.a. $w\,(n(dw))$, $F(0) = 0$ and there exists a $(\mathscr{B}_t)$-stopping time $\sigma = \sigma(w)$ such that $0 < \sigma < \infty$ and $F(u) = F(u \wedge \sigma)$.

(F. 2) For every $u > 0$, $\int_W u \wedge \sigma(w)\ n(dw) < \infty$ and $F(u, w) \in L^2(W, n)$.

Furthermore, for every $u_2 > u_1 > 0$ and $H \in L^\infty(W, n) \cap L^2(W, n)$ which is $\mathscr{B}_{u_1}$-measurable, we have

$$\int_W [F(u_2, w) - F(u_1, w)]\ H(w)\, n(dw) = 0 \tag{2.1}$$

and

$$\int_W [F(u_2, w) - F(u_1, w)]^2 H(w)\, n(dw) = \int_W [u_2 \wedge \sigma(w) - u_1 \wedge \sigma(w)]\, H(w)\, n(dw). \tag{2.2}$$

Example 1 (Brownian excursions). Let $W^+ = \{w \in C([0, \infty) \to \boldsymbol{R}^1);$ $w(0) = 0$, $\exists \sigma = \sigma(w)$, $0 < \sigma(w) < \infty$ and $w(t) > 0$ for $t \in (0, \sigma(w))$, $w(t) = 0$ for $t \geqslant \sigma(w)\}$, $W^- = \{w \in C([0, \infty) \to \boldsymbol{R}^1);\ -w \in W^+\}$ and $W = W^+ \cup \cup W^-$. Let the σ-fields $\mathscr{B}_{W^+}$, $\mathscr{B}_{W^-}$ and $\mathscr{B}_W$ be generated by Borel cylinder sets. Let n^+ be the σ-finite measure on $(W^+, \mathscr{B}_{W^+})$ defined by

$$n^+(B) = \int_0^\infty (2\pi t)^{-3/2} P_{0,0}^{0,t}(B \cap \{\sigma = t\})\, dt,$$

where $P_{0,0}^{0,t}$ is the Brownian excursion law on $W^+ \cap \{w;\ \sigma(w) = t\}$, ([3] or [1], Ex. IV. 8.4. Cf. also [6] and [8] for other interesting descriptions

of n^+). n^- on $(W^-, \mathscr{B}_{W^-})$ is defined similarly. Then the measure $n = c^+ n^+ + c^- n^-$ on $(W, \mathscr{B})$, where c^+ and c^- are non-negative constants such that $c^+ + c^- = 1$, together with $F(u, w) = w(u)$ and the natural filtration on W, satisfies the above properties.

Example 2 (Square root boundaries). Let Γ be the domain in the space-time plane $[0, \infty) \times \boldsymbol{R}^1$ defined by $\Gamma = \{(x, t);\ 0 < t < \infty,\ c_1\sqrt{t} < x < c_2\sqrt{t}\}$ where $c_1 \leqq 0 \leqq c_2$, $-\infty \leqq c_1 < c_2 \leqq \infty$ and $\min(|c_1|, |c_2|) < \infty$. Let $W = \{w \in C([0, \infty) \to \boldsymbol{R}^1),\ w(0) = 0,\ \exists\, \sigma = \sigma(w)$ such that $0 < \sigma(w) < \infty$ and $(t, w(t)) \in \Gamma$ for $0 < t < \sigma(w)$ and $(\sigma(w), w(t)) = (\sigma(w), w(\sigma(w)) \in \partial\Gamma$ for $t \geqq \sigma(w)\}$, let $\mathscr{B}_W$ be generated by Borel cylinder sets and $(\mathscr{B}^t)$ be the natural filtration. Let $\{P_x\}_{x \in \boldsymbol{R}^1}$ be the one-dimensional Brownian motion given canonically on $C([0, \infty) \to \boldsymbol{R}^1)$ and $m_t = m_t(w) = \inf\{u \geqq 0;\ (w(u), t+u) \in \partial\Gamma\}$, $w \in C([0, \infty) \to \boldsymbol{R}^1)$. Then there exists a unique σ-finite measure n on $(W, \mathscr{B}_W)$ such that, for every $t > 0$ and $E \in \mathscr{B}_{C([0,\infty)\to\boldsymbol{R}^1)}$,

$$n\{w \in W;\ w_t^+ \in E,\ \sigma(w) > t\} = \int_{c_1\sqrt{t}}^{c_2\sqrt{t}} \mu(t, x)\, P_x(w:\ w_{m_t}^- \in E)\, dx.$$

$(w_t^+, w_t^- \in C([0, \infty) \to \boldsymbol{R}^1)$ for $w \in C([0, \infty) \to \boldsymbol{R}^1)$ are defined as usual by $w_t^+(s) = w(t+s)$ and $w_t^-(s) = w(t \wedge s))$, where $\mu(t, x) = t^{-(\lambda_0 + 1/2)} \psi_0\left(\dfrac{x}{\sqrt{t}}\right) \times$ $\times e^{-x^2/2t}$, $[\lambda_0, \psi_0]$ being the first eigenvalue and the first normalized, positive eigenfunction of the following eigenvalue problem on $L^2[(c_1, c_2), e^{-x^2/2}\, dx]$:

$$\tfrac{1}{2}(\psi'' - x\psi') + \lambda\psi = 0 \text{ in } (c_1, c_2) \text{ with } \psi(c_i) = 0 \text{ if } |c_i| < \infty.$$

Since $\int_{c_1\sqrt{t}}^{c_2\sqrt{t}} \mu(t, x)\, dx = \text{const} \cdot t^{-\lambda_0}$, $\int_W [\sigma(w) \wedge u] n(dw) < \infty$ for $u > 0$ if and only if $\lambda_0 < 1$. Therefore this n and $F(u, w) = w(u)$ satisfy the above assumptions if and only if $\lambda_0 < 1$.

Now we construct, on a suitable probability space $(\Omega, \mathscr{F}, P)$ with a filtration $(\mathscr{F}_t)$, a stationary $(\mathscr{F}_t)$-Poisson point process $N(dt, dw)$ $(= N(dt, dw)(\omega))$ over W with the characteristic measure n (cf. [1]). It is a Poisson random point measure on $(0, \infty) \times W$ such that $E(N(dt, dw)) = dt\, n(dw)$ and, for each $\omega \in \Omega$, $N(\{t\} \times W) = 0$ or 1 for every t. We set, for each $\omega \in \Omega$, $\mathbf{D}_p$ $(= \mathbf{D}_{p(\omega)}) = \{t;\ N(\{t\} \times W) = 1\}$ and define $p(t)$ $(= p(t)(\omega))$ for $t \in \mathbf{D}_p$ as the unique element in W such that $N(\{t, p(t)\}) = 1$. If $f(t, w, \omega)$ on $[0, \infty) \times W \times \Omega$ is $(\mathscr{F}_t)$-predictable and

$\int_0^t ds \int_W f^2(s, w, \omega)\, n(dw) < \infty$ for every $t > 0$, for a.a. $\omega(P)$, then the stochastic integral

$$M_t = \int_0^{t+} \int_W f(s, w, \cdot)\, \tilde{N}(ds, dw), \qquad \tilde{N}(ds, dw) = N(ds, dw) - ds\, n(dw),$$

is defined as an $(\mathscr{F}_t)$-locally square integrable martingale such that $\langle M \rangle_t = \int_0^t ds \int_W f(s, w, \cdot)^2 n(dw)$, cf. [1] for details.

Set

$$A(t) = \int_0^{t+} \int_W \sigma(w)\, N(ds, dw) = \Sigma_{s \leqslant t,\, s \in D_p} \sigma[p(s)]. \tag{2.5}$$

Then $A(t)$ is a process with stationary independent increments with Lévy measure $\nu(dx) = n\{\sigma(w) \in dx\}$. Since $\int_W [\sigma(w) \wedge 1]\, n(dw) < \infty$, $A(t)$ is well defined as a right-continuous strictly increasing function in t such that $\lim_{t \to \infty} A(t) = \infty$ a.s. Therefore, for every $t \geqslant 0$ there exists a unique $s \geqslant 0$ such that $A(s-) \leqslant t \leqslant A(s)$, which we denote by $s = \varphi_t$. We set, for every fixed $t > 0$,

$$f_t(s, w, \omega) = \begin{cases} F\big(t - A(s-), w\big) & \text{if } t > A(s-) \\ 0 & \text{if } t \leqslant A(s-) \end{cases}$$

which is $(\mathscr{F}_t)$-predictable and satisfies

$$E\Big[\int_0^T ds \int_W f_t(s, w, \cdot)^2 n(dw)\Big] = E\Big[\int_0^{T \wedge \varphi_t} ds \int_W f_t(s, w, \cdot)^2 n(dw)\Big]$$

$$\leqslant E\Big[\int_0^{\varphi_t} ds \int_W f_t(s, w, \cdot)^2 n(dw)\Big] = t \quad \text{for every } T > 0.$$

THEOREM 1. *For every fixed $t > 0$, set*

$$B(t) = \int_0^{\varphi_t +} \int_W f_t(s, w, \cdot)\, \tilde{N}(ds, dw) = \int_0^{\varphi_t +} \int_W F\big(t - A(s-), w\big)\, \tilde{N}(ds, dw).$$

Then $t \to B(t)$ has a continuous modification which is a one-dimensional Brownian motion.

This theorem, applied to Example 1, yields that the process defined by

$$X(t) = \begin{cases} p(s)[t - A(s-)] & \text{if } A(s-) \leqslant t \leqslant A(s),\ s \in \boldsymbol{D}_p \\ 0 & \text{otherwise} \end{cases}$$

is a continuous process such that $X(t) - (c^+ - c^-)\varphi_t = B(t)$ is a one-dimensional Brownian motion. Also the theorem applied to Example 2 yields the following results of P. Greenwood and E. Perkins for one-dimensional Brownian motion $B(t)$:

$P\big(\exists t > 0,\ \exists \varepsilon > 0$ *such that* $c_1\sqrt{h} < B(t+h) - B(t) < c_2\sqrt{h}$ *for all* $h \in (0, \varepsilon)\big) = 1$ *if* $\lambda_0 = \lambda_0(c_1, c_2) < 1$.

3. Construction of martingale from pieces

Now we generalize the result of the previous section. Let $(W, \mathscr{B}_W, n)$ and the $(\mathscr{F}_t)$-Poisson point process $p(t)$ on $(\Omega, \mathscr{F}, P)$ be the same as above. But we allow here $F(u) = F(u, w)$, $\sigma = \sigma(w)$ and $(\mathscr{B}_u)$ to depend on $(t, \omega) \in [0, \infty) \times \Omega$ and $F(u)$ be multi-dimensional. To be precise, suppose we are given, for each $(t, \omega) \in [0, \infty) \times \Omega$, the following objects:

$F^{t,\omega}(u) = \big(F_i^{t,\omega}(u, w)\big)_{i=1}^d$

$\sigma^{t,\omega} = \sigma^{t,\omega}(w) \geqslant 0$ (here we allow the possibility that $\sigma^{t,\omega}(w) = 0$)

$(\mathscr{B}_u^{t,\omega})$: a right-continuous increasing family of sub σ-fields of $\mathscr{B}_W$

and suppose that they *depend on* (t, ω) $(\mathscr{F}_t)$-*predictably* in the following sense. Both $(t, \omega) \to \sigma^{t,\omega} \in L^0\big(W \to [0, \infty)\big)$ and $(t, \omega) \to F^{t,\omega} \in L^0\big(W \to C\big([0, \infty) \to \boldsymbol{R}^d\big)\big)$ are $(\mathscr{F}_t)$-predictable ($L^0(W \to E)$, in general, is the space of E-valued n-measurable function on W) and, for $(\mathscr{B}_u^{t,\omega})$, we assume the following: there exists a Polish space S and an S-valued function $\xi = \xi^{t,\omega}(u, w)$ measurable in (t, ω, u, w), $(\mathscr{F}_t)$-predictable in (t, ω) and continuous in $u \in [0, \infty)$ such that, for fixed (t, ω),

$$\mathscr{B}_u^{t,\omega} = \bigcap_{\varepsilon > 0} \sigma[\xi^{t,\omega}(v, \cdot);\ v \leqslant u + \varepsilon]. \tag{3.1}$$

We assume conditions similar to (F. 1) and (F. 2) for $F_i^{t,\omega}$ and $\sigma^{t,\omega}$, for every (t, ω) and $i = 1, 2, \ldots, d$, except for (2.2). Instead, we assume that, for every (t, ω), $i, j = 1, 2, \ldots, d$, $u_2 > u_1 > 0$, and every $H \in L^\infty(W, n)$ which is $\mathscr{B}_{u_1}^{t,\omega}$-measurable, we have

$$\begin{aligned} &\int_W [F_i^{t,\omega}(u_2, w) - F_i^{t,\omega}(u_1, w)]\,[F_j^{t,\omega}(u_2, w) - F_j^{t,\omega}(u_1, w)]\,H(w) \otimes n(dw) \\ &\quad = \int_W [\langle F_i^{t,\omega}, F_j^{t,\omega}\rangle(u_2, w) - \langle F_i^{t,\omega}, F_j^{t,\omega}\rangle(u_1, w)]\,H(w)\ n(dw) \end{aligned} \tag{3.2}$$

where $\langle F_i^{t,\omega}, F_j^{t,\omega}\rangle(u, w)$ is real-valued, measurable in (t, ω, u, w), $(\mathscr{F}_t)$-predictable in (t, ω) and for each (t, ω) and a.a. $w(n)$, $u \to \langle F_i^{t,\omega}, F_j^{t,\omega}\rangle(u, w)$ is continuous with finite total variation on each bounded interval, $\langle F_i^{t,\omega}, F_j^{t,\omega}\rangle(0, w) = 0$ and $\langle F_i^{t,\omega}, F_j^{t,\omega}\rangle(\sigma^{t,\omega} \wedge u, w) = \langle F_i^{t,\omega}, F_j^{t,\omega}\rangle(u, w)$; furthermore, it is $(\mathscr{B}_u^{t,\omega})$-adapted for each (t, ω). We assume also that, for every $t > 0$,

$$\int_0^t ds \int_W [1 \wedge \sigma^{s,\omega}(w)]\, n(dw) < \infty \quad \text{a.s.} \tag{3.3}$$

Let $\varrho(s)$ be an $(\mathscr{F}_t)$-adapted, nonnegative, measurable and locally integrable process and $A(0)$ be an $\mathscr{F}_0$-measurable nonnegative random variable. Set

$$A(t) = A(0) + \int_0^{t+} \int_W \sigma^{s,\omega}(w)\, N(ds, dw) + \int_0^t \varrho(s)\, ds \tag{3.4}$$

and assume that $t \to A(t)$ is strictly increasing and $\lim_{t\to\infty} A(t) = \infty$ a.s. Then for every $t \geqslant 0$ there exists a unique s such that $A(s-) \leqslant t \leqslant A(s)$ $(A(0-) = 0)$, which we denote by $s = \varphi_t$. Finally we assume that, for every $t > 0$ and $i = 1, 2, \ldots, d$,

$$\int_0^{t+} \int_W \langle F_i^{s,\omega}, F_i^{s,\omega}\rangle(\sigma^{s,\omega}, w)\, N(ds, dw) < \infty \quad \text{a.s.} \tag{3.5}$$

For each fixed $t > 0$, we define a σ-field $\mathscr{H}_t$ on Ω by

$$\mathscr{H}_t = \mathscr{F}_{\varphi(t)-} \vee \sigma\left[\xi^{\varphi_t,\omega}\big(u - A(\varphi_t-),\ p(\varphi_t)\big);\ u \leqslant t\right], \tag{3.6}$$

in which we used the following convention. Let Δ and a be extrapoints attached to W and S, respectively, and we set $p(t) = \Delta$ if $t \notin \mathbf{D}_p$, $\xi^{t,\omega}(u, \Delta) \equiv a$, $F_i^{t,\omega}(u, \Delta) \equiv 0$, $\xi^{t,\omega}(u, w) = a$ and $F^{t,\omega}(u, w) = 0$ if $u < 0$.

THEOREM 2. *Set, for each $t \in [0, \infty)$ and $i = 1, 2, \ldots, d$,*

$$M_i(t) = \int_0^{\varphi_t+} \int_W F_i^{s,\omega}\big(t - A(s-), w\big)\, \tilde{N}(ds, dw). \tag{3.7}$$

Then $M_i(t)$ has a continuous modification which is an $(\mathscr{H}_t)$-locally square integrable martingale and

$$\langle M_i, M_j\rangle(t) = \int_0^{\varphi_t+} \int_W \langle F_i^{s,\omega}, F_j^{s,\omega}\rangle\big(t - A(s-), w\big)\, N(ds, dw). \tag{3.8}$$

We denote this system $M = \{M_i\}$ of martingales by $M^F = \{M_i^F\}$. Let $\Phi^{t,\omega}(u, w)$, $(t, \omega) \in [0, \infty) \times \Omega$, $(u, w) \in [0, \infty) \times (W \cup \{\Delta\})$ be given; suppose it is real-valued, measurable in (t, ω, u, w), $(\mathscr{F}_t)$-predictable in (t, ω), $(\mathscr{B}_u^{t,\omega})$-predictable in (u, w) and $\Phi^{t,\omega}(u, \Delta) = 0$. Assume, for simplicity, that $\Phi^{t,\omega}(u, w)$ is bounded. Then for each fixed (t, ω) and $i = 1, 2, \ldots, d$, we can define the stochastic integral

$$(\Phi \cdot F)_i^{t,\omega}(u, w) = \int_0^{\sigma^{t,\omega} \wedge u} \Phi^{t,\omega}(v, w)\, d_v F_i(v, w)$$

on the measure space $\{W, \mathscr{B}_W, n, (\mathscr{B}_u^{t,\omega})\}$ in the same way as the usual Itô integrals, so that $(\Phi \cdot F)^{t,\omega} = \{(\Phi \cdot F)_i^{t,\omega}\}$ has the same properties as $F^{t,\omega}$. Define an $(\mathscr{H}_t)$-predictable process $\tilde{\Phi}$ by

$$\tilde{\Phi}(t, \omega) = \Phi^{\varphi_t, \omega}\big(t - A(\varphi_t -),\ p(\varphi_t)\big). \tag{3.9}$$

THEOREM 3.

$$\int_0^t \tilde{\Phi}(s, \omega)\, dM_i^F(s) = M_i^{\Phi \cdot F}(t), \qquad i = 1, 2, \ldots, d.$$

Example 3. A typical example of such $F^{t,\omega}(u) = \big(F_i^{t,\omega}(u, w)\big)$ and $\sigma^{t,\omega} = \sigma^{t,\omega}(w)$ is furnished by those $\Phi\big(\xi(t-), w\big)(u)$ and $\sigma\big[\Phi\big(\xi(t-), w\big)\big]$ defined in [7] (cf. also [1], p. 217); then $\mathscr{B}_u^{t,\omega}$ is generated by $\Phi\big(\xi(t-), w\big)(v)$, $v \leqslant u$. In this case, they depend on (t, ω) through the $(\mathscr{F}_t)$-predictable process $\xi(t-)$. Applying the above Theorems 2 and 3, we can obtain the stochastic differential equation governing the process $X(t)$ constructed there and we see that $X(t)$ is actually a diffusion process corresponding to the given differential operator and Wentzell's boundary condition.

References

[1] Ikeda N. and Watanabe S., *Stochastic Differential Equations and Diffusion Processes*, North-Holland, Kodansha, 1981.

[2] Itô K., Poisson Point Processes Attached to Markov Processes, *Proc. Sixth Berkeley Symp.* III, pp. 225–239. Univ. California Press, 1972.

[3] Itô K. and McKean H. P., *Diffusion Processes and Their Sample Paths*, Springer, 1965.

[4] Kasahara Y., *Limit Theorems for Lévy Processes and Poisson Point Processes and Their Applications to Brownian Excursions*, to appear in J. Math. Kyoto Univ. **24** (1984).

[5] Maisonneuve B., Temps local et dénombrements d'excursions, *Z. Wahrsch. verw. Geb.* **52** (1980), pp. 109–113.

[6] Rogers L. C. G., Williams' Characterization of the Brownian Excursion Law: Proof and Applications, *Séminaire de Prob.* XV 1979/80, pp. 227–250, Lect. Notes in Math. 850, Springer, 1981.

[7] Watanabe S., Construction of Diffusion Processes with Wentzell's Boundary Conditions by Means of Poisson Point Processes of Brownian Excursions, *Probability Theory*, Banach Center Publ., Vol. 5, pp. 255–271, PWN — Polish Scientific Publishers, 1979.

[8] Williams D., *Diffusions, Markov Processes, & Martingales*, Vol. 1, *Foundations*, John Wiley & Sons, 1979.

DEPARTMENT OF MATHEMATICS
FACULTY OF SCIENCE
KYOTO UNIVERSITY
KYOTO, 606 JAPAN

Proceedings of the International Congress of Mathematicians
August 16-24, 1983, Warszawa

ANTONIO AMBROSETTI

Existence and Multiplicity Results for Some Classes of Nonlinear Problems

This survey is concerned with the existence of multiple solutions of nonlinear problems as elliptic boundary value problems and Hamiltonian systems. The main tools usually employed to get results of that type are: topological degree, bifurcation theory, critical point theory, global inversion theorems in presence of singularities, etc.

Here the discussion will be restricted to the following cases: §1 — study of singularities of mappings and connections with multiplicity results, §2 — variational problems and critical point theory.

§ 1

Let X, Y be Banach spaces and $\Phi \in C^1(X, Y)$. A *singular* point is defined to be any $u \in X$ such that the Fréchet derivative $d\Phi(u) \in \mathcal{L}(X, Y)$ is not invertible. Denote by Σ the set of all singular points of Φ. If $\Sigma = \emptyset$ then Φ is a homeomorphism from X onto Y, provided Φ is proper (i.e., $\Phi^{-1}(K)$ is compact for all compact sets $K \subset Y$). In contrast to this, the presence of singularities gives rise to the existence of multiple solutions.

In order to describe the range of Φ, it is convenient to study both Σ and $\Phi(\Sigma)$. Precisely, let us suppose that $\Phi \in C^2(X, Y)$ and, given $u \in \Sigma$, that the following holds:

$$\exists\, v \in X,\ v \neq 0 \quad \text{and} \quad \psi \in Y^* - \{0\} \quad \text{such that:} \quad \operatorname{Ker}\big(d\Phi(u)\big) = \boldsymbol{R}v \ \text{and Range}\ \big(d\Phi(u)\big) = \operatorname{Ker}(\psi). \tag{Φ1}$$

$$\exists\, w \in X \ \text{such that} \ \langle d^2\Phi(u)[v][w],\ \psi \rangle \neq 0. \tag{Φ2}$$

Under conditions (Φ1), (Φ2) it is possible to show that $\exists\, \varepsilon > 0$ such that $\Sigma \cap B_\varepsilon(u)$ is a C^1-manifold of codimension 1 in X, where $B_\varepsilon(u) := \{z \in X:\ \|z - u\| < \varepsilon\}$. Moreover, in the case of $\Phi =$ Identity $-$ Compact,

it is also possible to establish how the Leray–Schauder index of the possible non-singular solutions of $\Phi(u) = 0$ changes, depending on their position with respect to Σ. This analysis is already sufficient for evaluating the number of solutions of certain problems [5]:

1. *Example.* Let E be a Hilbert space, let L be a linear, compact, self-adjoint positive operator in E, let $H \in C^2(E, E)$ be compact, homogeneous of degree a, a odd, with $dH(u)$ self-adjoint strictly positive, $(\forall\, u \neq 0)$. Consider the equation:

$$u - \lambda Lu + H(u) = 0. \tag{1_λ}$$

Denote by μ_i, $0 < \mu_1 \leqslant \mu_2 \leqslant \ldots$, the characteristic values of L. The following holds concerning the nontrivial (i.e., $\not\equiv 0$) solutions of (1_λ):

(i) if μ_1 is simple and $\mu_1 < \lambda \leqslant \mu_2$ then (1_λ) has precisely 2 nontrivial solutions;

(ii) if μ_2 is also simple and $\mu_2 < \lambda < \mu_2 + \varepsilon$, ε small enough, then (1_λ) has precisely 4 nontrivial solutions.

For example, the Van Karman equations of nonlinear elasticity can be put in the form (1_λ).

In order to study $\Phi(\Sigma)$, the assumption $(\Phi 2)$ has to be strengthened:

$$\langle d^2\Phi(u)[v][v],\ \psi\rangle \neq 0. \tag{$\Phi 3$}$$

Under conditions $(\Phi 1)$, $(\Phi 3)$ it is possible to prove that $\exists\ \varepsilon > 0$ such that $\Phi\big(\Sigma \cap B_\varepsilon(u)\big)$ is a C^1-manifold of codimension 1 in Y. This, in turn, allows one to show [9]:

2. THEOREM. *Let $\Phi \in C^2(X, Y)$ be proper and let $(\Phi 1)$, $(\Phi 3)$ hold for all $u \in \Sigma$. Moreover, suppose that Σ is connected and that $\Phi(u) = h$ has a unique solution $\forall\, h \in \Phi(\Sigma)$. Then $\Phi(\Sigma)$ is a C^1-manifold of codimension 1 in Y and there are two open sets Y_0 and Y_2 such that $Y = Y_0 \cup \Phi(\Sigma) \cup Y_2$ and $\Phi(u) = y$ has precisely 0, 1 or 2 solutions, according as $y \in Y_0$, $\Phi(\Sigma)$ or Y_2.*

3. *Example.* Consider the boundary value problem

$$-\Delta u = f(u) + h \text{ in } \Omega, \quad u = 0 \text{ on } \partial\Omega, \tag{2}$$

where Ω is a bounded domain in $\boldsymbol{R}^n$ with smooth boundary $\partial\Omega$. Let X, Y be the Hölder spaces $C_0^{2+\alpha}(\Omega)$ and $C^{0+\alpha}(\Omega)$, respectively, and denote by $0 < \lambda_1 < \lambda_2 \leqslant \ldots$ the eigenvalues of $-\Delta$ with zero Dirichlet boundary conditions. If $f \in C^2(\boldsymbol{R})$ is such that $f''(s) > 0$, $\forall\, s \in \boldsymbol{R}$, and

$$-\infty < \lim_{s \to -\infty} f'(s) < \lambda_1 < \lim_{s \to \infty} f'(s) < \lambda_2,$$

then $\Phi(u) := -\Delta u - f(u)$ turns out to verify the assumptions of Theorem 2. Hence the number of solutions of (2) is completely determined by the position of h in Y, according to the theorem.

The study of (2) has been carried over. Assuming $f \in C^1(\boldsymbol{R})$ is such that $\lim_{s\to-\infty} s^{-1}f(s) < \lambda_1 < \lim_{s\to\infty} s^{-1}f(s) < \infty$ and taking $h = t\varphi + h_1$, φ being the eigenfunction associated to λ_1, $\int \varphi h_1 = 0$, it is shown in [1] that there exists $t^* = t^*(h_1)$ such that (2) has at least 2 solutions for $t < t^*$, at least one solution for $t = t^*$, and no solutions for $t > t^*$.

It would be interesting to obtain a kind of global inversion theorem in the case where (Φ3) is not satisfied, or else, to establish the precise number of solutions of (2) when $f'(s)$ crosses more than one eigenvalue.

§ 2

If $\Phi = 0$ is the Euler equation of some functional $J\colon X \to \boldsymbol{R}$, the critical point theory supplies a method for finding multiplicity results, mainly when J is invariant under some group action (as the S^1-action given by the time-translation in the case of autonomous systems of ordinary differential equations, or the $\boldsymbol{Z}_2$-symmetry in the case of boundary value problems with odd nonlinearities, see below).

However, regarding for the moment the general case, we consider the rather typical situation in which $u = 0$ is a stationary point of J and one looks for nontrivial solutions of $J'(u) = 0$. When J is bounded neither from above nor from below, these stationary points have to be found via min-max procedures. A result of this kind is the following [10]:

4. THEOREM. *Let $J \in C^1(X, \boldsymbol{R})$ satisfy the* (PS) *condition, i.e., every u_n with $J(u_n)$ bounded and $J'(u_n)\to 0$ has a converging subsequence. Moreover, suppose $J(0) = 0$ and $\exists\ r, \varrho > 0$ and $\hat{u} \in X$ such that*: (i) $J(u) \geqslant \varrho$ *for all* $\|u\| = r$; (ii) $\|\hat{u}\| > r$ *and* $J(\hat{u}) = 0$. *Then J has at least one stationary point $v \neq 0$ such that*

$$J(v) = \inf_{p\in\Gamma} \max \{J(p(t))\colon 0 \leqslant t \leqslant 1\}$$

where Γ denotes the class of paths from $u = 0$ to $u = \hat{u}$.

Theorem 4 can be applied to find nontrivial solutions of several classes of differential equations as, for example, elliptic superlinear boundary value problems [10], periodic solutions of the vibrating string [15], etc.

For a generalization to handle the case where J is indefinite at $u = 0$, see [13], [22].

In order to find further critical points of J, Theorem 4 has to be somewhat sharpened. A result in this direction has been proved in [3]: if the stationary point v found as inf-max in Theorem 4 is nondegenerate, then it has Morse index equal to 1. As an application one has:

5. *Example.* Consider the boundary value problem (2) with $h = t\varphi + h_1$. The solutions of (2) are the stationary points of

$$J(u) := \tfrac{1}{2}\|u\|^2_{W_0^{1,2}(\Omega)} - \int_\Omega F(u) - \int_\Omega hu \quad \text{on } X = W_0^{1,2}(\Omega),$$

where $F(u) := \int_0^u f(s)\,ds$. Assuming $\lim_{s\to-\infty} f'(s) < \lambda_1$ and $\lambda_2 < \lim_{s\to\infty} f'(s) \neq \lambda_i$ and finite, one shows that $\exists t_0\, \forall t < t_0$ J has a local minimum (then with Morse index 0) at some $u_1 < 0$, and another stationary point $u_2 > 0$ which is nondegenerate and with Morse index $\geqslant 2$. Therefore J has to possess at least a third critical point, for any such t.

Another situation which occurs in applications is that one can need to find stationary points with additional properties, see the examples below. To this purpose, we will describe a device which allows to obtain stationary points of J as critical points of J constrained on a suitable manifold.

Set $h(u) := \langle J'(u), u\rangle$ and $M := \{u \in X - \{0\}\colon h(u) = 0\}$. Suppose that $h \in C^1$ and that $\langle h'(u), u\rangle \neq 0$, $\forall u \in M$ and let $u^* \in M$ be a critical point of J on M. Then $\exists\lambda$ such that $J'(u^*) = \lambda h'(u^*)$. But $\langle J'(u^*), u^*\rangle = h(u^*) = 0$ and $\langle h'(u^*), u^*\rangle \neq 0$ implies $\lambda = 0$ and thus u^* is a stationary point of J. Of particular interest is the case where J has a minimum on M.

6. *Example.* Consider the system of ordinary differential equations

$$-\ddot{x} = Ax + \nabla V(x) \tag{3}$$

where $x \in \boldsymbol{R}^n$, A is a symmetric matrix with eigenvalues $0 < \omega_1^2 \leqslant \omega_2^2 \leqslant \ldots$ $\ldots \leqslant \omega_n^2$ and $V \in C^2(\boldsymbol{R}^n, \boldsymbol{R})$ is convex and β-homogeneous, $\beta > 2$. Let $X := L^a(0, 2\pi; \boldsymbol{R}^n)$, $a = \beta(\beta-1)^{-1}$. If $\sigma > \omega_n$ then the densely defined operator $x \mapsto -\sigma^2 x - \ddot{x} - Ax$ is invertible with compact inverse L_σ. Let V^* denote the Legendre transform of V, and define J on X by setting

$$J(u) := -\tfrac{1}{2}\int_0^{2\pi} u \cdot L_\sigma u + \int_0^{2\pi} V^*(u).$$

One verifies that the stationary points u of J correspond, through the relation $x = \nabla V^*(u)$, to the $2\pi/\sigma$-periodic solutions of (3) (the Dual Action Principle [16]). Using the above procedure, one shows that J has on M a minimum which corresponds to solutions of (3) with *minimal* period $2\pi/\sigma$, $\forall \sigma > \omega_n$.

For other results concerning solutions of Hamiltonian systems with a prescribed minimal period, see [17] for subquadratic Hamiltonians, [6] for superquadratic ones and [19] for some results in the case of non-convex Hamiltonians.

Other applications concern the existence of steady vortex rings in an ideal fluid [7] and a proof of a remarkable result by Ekeland and Lasry [18] which we are now going to expose.

Let $H \in C^1(\boldsymbol{R}^{2n}, \boldsymbol{R})$ and consider the Hamiltonian system

$$\dot{z} = \mathcal{J} \nabla H(z) \tag{HS}$$

where $z = (p, q) \in \boldsymbol{R}^{2n}$, $\cdot = d/dt$ and $\mathcal{J}(p, q) = (-q, p)$. Suppose $H^{-1}(1) = \partial\Omega$, where Ω is a bounded convex domain in $\boldsymbol{R}^{2n}$, $0 \in \operatorname{int} \Omega$, and that $\nabla H(z) \neq 0$, $\forall z \in \partial\Omega$. Denote by B_r the ball in $\boldsymbol{R}^{2n}$ of radius r and center 0.

7. THEOREM. *Let H satisfy the above assumptions. Moreover, suppose there are r, $R > 0$ with $R^2 < 2r^2$ such that $B_r \subset \Omega \subset B_R$. Then* (HS) *has at least n distinct periodic orbits on $H^{-1}(1)$.*

Following [8], the proof can be carried out as follows:

(i) Without loss of generality H can be taken homogeneous, of degree 4, say, so that G, the Legendre transform of H, turns out to be homogeneous of degree 4/3. Let $E = \left\{u \in L^{4/3}(0, 2\pi; \boldsymbol{R}^{2n}) : \int_0^{2\pi} u = 0\right\}$ and denote by L the operator defined on E by setting $Lu = z$ iff $-\mathcal{J}\dot{z} = u$. If $u \in E$ is a critical point of $J := -\frac{1}{2}\int_0^{2\pi} u \cdot Lu + \int_0^{2\pi} G(u)$, with $k = G(u) \neq 0$, then $z := k^{-1/4}\, \nabla G\left(u(k^{-1/2}t)\right)$ is a periodic solution of (HS) on $H^{-1}(1)$.

(ii) One examines J constrained on $M = \{u \in E : \langle J'(u), u\rangle = 0\}$. In the present case one has $J_{|M}(u) = \frac{1}{2}\int_0^{2\pi} G(u)$. Both J and M are invariant under the S^1-action $u(t) \mapsto u(t+\theta)$ and the Lusternik–Schnirelman critical point theory provides the existence of critical points of $J_{|M}$ via inf-max. Using the specific assumption of Theorem 7 one proves that at least n of those critical points correspond to elements of E with *minimal* period 2π, and hence give rise to different periodic orbits of (HS) on $H^{-1}(1)$.

An improvement of Theorem 7 was recently obtained; see [14].

We end the report with a brief discussion of a question concerning superlinear boundary value problems.

Let Ω be a bounded domain in $\boldsymbol{R}^n$ with smooth boundary $\partial\Omega$ and consider

$$-\Delta u = u|u|^{p-1} + h(x) \quad \text{in } \Omega, \quad u = 0 \text{ on } \partial\Omega. \tag{4}$$

If $h = 0$ and $1 < p < (n+2)/(n-2)$ then (4) has infinitely many solutions which can be found as stationary points on $W_0^{1,2}(\Omega)$ of

$$J(u) = \tfrac{1}{2}\|u\|^2_{W_0^{1,2}} - \frac{1}{p+1}\int_\Omega |u|^{p+1}.$$

Of course, the $\boldsymbol{Z}_2$-symmetry of J has to be used here.

By means of replacing this by the equivalent critical point problem with constraint, we have proved in [2] a perturbation result: roughly, it states that $\forall k \in \boldsymbol{N}$, $\exists \varepsilon(k)$ such that (4) has at least k solutions provided $\|h\|_{L^\infty} < \varepsilon(k)$. Using essentially the same device jointly with some asymptotic estimates, the existence of infinitely many solutions of (4) has been proved in [12], [24], but for a smaller range of p. For example, among others, one has

8. THEOREM. *Let l be the greatest root of $(2n-2)s^2-(n+2)s-n = 0$. Then if $1 < p < l$, (4) has infinitely many solutions.*

For another proof, see [23]. The problem to see whether or not Theorem 8 holds for all $p < (n+2)/(n-2)$ is open. For a result in this direction, see [11]. For other perturbation results using the Morse Theory and applications to nonlinear eigenvalue problems, see [21].

Lastly, a few words on the problem

$$-\Delta u = \lambda u - g(u) \quad \text{in } \Omega, \quad u = 0 \text{ on } \partial\Omega, \tag{5}$$

where $g(u) = o(|u|)$ at $u = 0$ and $u^{-1}g(u) \to \infty$ as $|u| \to \infty$. While for g odd (5) has been exhaustively investigated, much less is known in the general case, see [4], [20], [25]. To obtain a complete multiplicity result for (5) seems to be an interesting goal to pursue.

References

[1] Amann H. and Hess P., A multiplicity results for a class of elliptic boundary value problems, *Proc. Roy. Soc. Edinb.* **84-A** (1979), pp. 145–151.

[2] Ambrosetti A., A perturbation theorem for superlinear boundary value problems, *M. R. C. Univ. of Wisconsin Tech. Summ. Report* N. **1446**, 1974.

[3] Ambrosetti A., Elliptic equations with jumping nonlinearities, *J. Math. Phys. Sci.*, to appear.

[4] Ambrosetti A. and Lupo D., On a class of nonlinear Dirichlet problems with multiple solutions, *Nonlin. Anal. TMA*, to appear.
[5] Ambrosetti A. and Mancini G., Sharp nonuniqueness results for some nonlinear problems, *Nonlin. Anal. TMA* **3-5** (1979), pp. 635–645.
[6] Ambrosetti A. and Mancini G., Solutions of minimal period for a class of convex Hamiltonian systems, *Math. Annalen* **255** (1981), pp. 405–421.
[7] Ambrosetti A. and Mancini G., On some free boundary problems, in: *Recent contributions to nonlinear partial differential equations*, ed. H. Berestycki and H. Brézis, Pitman, 1981, pp. 24–36.
[8] Ambrosetti A. and Mancini G., On a theorem by Ekeland and Lasry concerning the number of periodic Hamiltonian trajectories, *J. Diff. Equat.* **43** (1981), pp. 1–6.
[9] Ambrosetti A. and Prodi G., On the inversion of some differentiable mappings with singularities between Banach spaces, *Ann. Mat. Pura Appl.* **93** (1972), pp. 231–247.
[10] Ambrosetti A. and Rabinowitz P. H., Dual variational methods in critical point theory and applications, *J. Funct. Anal.* **14** (1973), pp. 349–381.
[11] Bahri A., Topological results on a certain class of functionals and applications, *J. Funct. Anal.* **41** (1981), pp. 397–427.
[12] Bahri A. and Berestycki H., A perturbation method in critical point theory and applications, *Trans. A. M. S.* **267-1** (1981), pp. 1–32.
[13] Benci V. and Rabinowitz P. H., Critical point theorems for indefinite functionals, *Inv. Math.* **52** (1979), pp. 241–273.
[14] Berestycki H., Lasry J. M., Mancini G., and Ruf B., *Existence of multiple periodic orbits on star-shaped Hamiltonian surfaces*, preprint.
[15] Brézis H., Coron J.-M., and Nirenberg L., Free vibrations for a nonlinear wave equation and a theorem of P. Rabinowitz, *Comm. P. A. M.* **33** (1980), pp. 667–684.
[16] Clarke F., Solution périodique des équations hamiltoniennes, *C. R. Acad. Sci. Paris* **287** (1978), pp. 951–952.
[17] Clarke F. and Ekeland I., Hamiltonian trajectories having prescribed minimal period, *Comm. P. A. M.* **33** (1980), pp. 103–116.
[18] Ekeland I. and Lasry J. M., On the number of periodic trajectories for a Hamiltonian flow on a convex energy surface, *Ann. Math.* **112** (1980), pp. 283–319.
[19] Girardi M. and Matzeu M., Some results on solutions of minimal period to superquadratic Hamiltonian systems, *Nonlin. Anal. TMA* **7-5** (1983), pp. 475–482.
[20] Hofer H., Variational and topological methods in partially ordered Hilbert spaces, *Math. Annalen* **261** (1982), pp. 493–514.
[21] Marino A. and Prodi G., Metodi perturbativi nella teoria di Morse, *Boll. U. M. I.* **11** (1975), pp. 1–32.
[22] Ni W. M., Some minimax principles and their applications in nonlinear elliptic equations, *J. d'Anal. Math.* **37** (1980), pp. 248–275.
[23] Rabinowitz P. H., *Multiple critical points of perturbed symmetric functionals*, preprint.
[24] Struwe M., Infinitely many critical points for functionals which are not even and applications to superlinear boundary value problems, *Manus. Math.* **32** (1980), pp. 335–364.
[25] Struwe M., A note on a result of Ambrosetti and Mancini, *Ann. Mat. Pura Appl.* **131** (1982), pp. 107–115.

Proceedings of the International Congress of Mathematicians
August 16-24, 1983, Warszawa

JEAN-MICHEL BONY

Propagation et interaction des singularités pour les solutions des équations aux dérivées partielles non-linéaires

1. Introduction

Nous nous intéresserons aux équations non linéaires générales d'ordre m:

$$F\big(x, c(x), u(x), \ldots, \partial^\beta u(x), \ldots\big)_{|\beta|\leqslant m} = 0 \tag{1}$$

et au cas particulier des équations semi-linéaires:

$$\sum_{|\alpha|=m} a_\alpha(x)\,\partial^\alpha u(x) + G\big(x, c(x), \ldots, \partial^\beta u(x), \ldots\big)_{|\beta|\leqslant m-1} = 0 \tag{2}$$

où $x \in \boldsymbol{R}^n$, où F (ou G) est une fonction C^∞ à valeurs réelles de ses arguments, où $c(x)$ (le contrôle) est une fonction à valeurs vectorielles donnée (pas nécessairement C^∞), et où $u(x)$ est une fonction inconnue à valeurs réelles.

Depuis 1978, un certain nombre de travaux ont été consacrés à la propagation des singularités pour ces équations. Il est remarquable que des résultats généraux, indépendants de la nature de F, puissent être obtenus, alors que les problèmes de l'existence de solutions, ou de l'apparition d'ondes de choc, dépendent fortement de la nature (croissance, convexité ...) de F. Il est également remarquable que les méthodes d'analyse microlocale, surtout développées jusqu'ici pour les équations linéaires, s'appliquent à ce problème.

Pour fixer les idées, nous décrirons notre problématique dans le cas où l'équation (1), munie d'une solution u, est strictement hyperbolique dans $\boldsymbol{R}^n$ entier par rapport à la direction x_n, c'est-à-dire que le polynôme en ξ_n: $\sum_{|\alpha|=m} (\partial F/\partial u_\alpha)\big(x, c(x), u(x) \ldots\big)\,\xi^\alpha$ possède m racines réelles distinctes

pour $\xi' = (\xi_1, \ldots, \xi_{n-1})$ non nul (bien entendu, pour (2), cette condition ne dépend que de l'équation et pas de u).

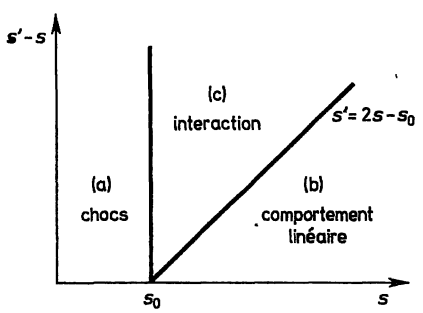
Nous poserons le problème de la propagation des singularités de la façon suivante (le problème de l'existence de solutions est évacué par la formulation même).

PROBLÈME 1. Soit u une solution de (1) appartenant à l'espace de Sobolev H^s dans $\boldsymbol{R}^n$ entier. On suppose connues les singularités de $c(x)$ dans $\boldsymbol{R}^n$, et celles de $u(x)$ dans $\boldsymbol{R}^n_- = \{x \in \boldsymbol{R}^n, x_n < 0\}$.

Pour $s' > s$, peut-on déterminer si u appartient à $H^{s'}$ au voisinage d'un point $\tilde{x}$ avec $\tilde{x}_n > 0$?

PROBLÈME 1'. Même énoncé, mais on suppose connues les singularités de $c(x)$, et des données de Cauchy $(\partial/\partial x_n)^j u(x', 0)$, pour $j = 0, \ldots, m-1$.

La situation est schématisée dans le tableau suivant:



(a) Si s est inférieur à un indice critique s_0, on ne peut pas prédire en général le comportement des singularités dans l'avenir en ne connaissant que les singularités de u dans le passé. Il est a priori possible (cela dépend de la nature de F) que u soit C^∞ dans le passé et qu'une singularité (choc) apparaisse brusquement. Cela est par contre impossible pour $s > s_0$ (seuls des chocs "violents" peuvent apparaître). On peut prendre $s_0 = n/2 + m + 1$ pour l'équation (1), $s_0 = n/2 + m - 1$ pour (2), et des valeurs plus faibles pour des équations moins méchamment non linéaires.

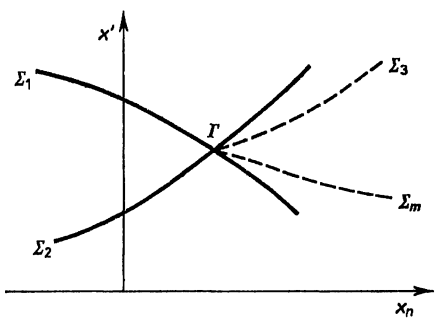
(b) Pour $s > s_0$, si l'on ne recherche qu'une régularité limitée $s' < 2s - s_0$ sur u, les résultats sont les mêmes que pour une équation linéaire. Supposons que pour chaque bicaractéristique (nulle) rétrograde B issue de $\tilde{x}$, on ait $c(x) \in H^{s'-m+1}$ le long de B, et qu'il existe un point de $B \cap \boldsymbol{R}^n_-$ où u appartienne à $H^{s'}$, alors u appartient à $H^{s'}$ au voisinage de $\tilde{x}$.

On trouvera au § 2 les résultats microlocaux et plus précis (théorèmes 1, 2 et 3). La méthode de démonstration développée dans [6] consiste

à ramener l'équation (1) à une équation "paradifférentielle" linéaire, à laquelle les techniques pseudodifférentielles s'appliquent avec relativement peu de modifications. Cela est résumé au § 3.

(c) Pour $s' > 2s - s_0$ (en particulier, si l'on désire déterminer si u appartient à C^∞ en $\tilde{x}$), le problème est encore largement ouvert. Il est encore possible, au moins dans certains cas, de prédire le comportement des singularités dans l'avenir, mais un phénomène nouveau, typiquement non linéaire, apparaît: l'interaction des singularités. Nous voudrions d'abord donner une explication heuristique de ce phénomène.

Considérons par exemple une solution u de (2), avec $c = 0$ pour simplifier. Supposons que, dans le passé, u soit C^∞ hors de deux hypersurfaces caractéristiques Σ_1 et Σ_2, et même que son spectre singulier (ou front d'onde) $\mathrm{SS}(u)$ soit la réunion $T^*_{\Sigma_1} \cap T^*_{\Sigma_2}$ des variétés conormales à Σ_1 et à Σ_2. Supposons que, pour $x_n > 0$, Σ_1 et Σ_2 se coupent le long de Γ, de codimension 2.



Si l'équation était linéaire, on aurait également $\mathrm{SS}(u) = T^*_{\Sigma_1} \cup T^*_{\Sigma_2}$ dans l'avenir, et il est raisonnable de penser que, dans notre cas, on a encore "en général" $\mathrm{SS}(u) \supset T^*_{\Sigma_1} \cup T^*_{\Sigma_2}$.

Le point-clef est que, même si $\mathrm{SS}(u)$ était réduit à $T^*_{\Sigma_1} \cup T^*_{\Sigma_2}$, le spectre singulier de $g(x) = G(x, u, \ldots, \partial^\beta u, \ldots)$ serait plus gros au dessus de Γ (dans les meilleurs cas — u distribution de Fourier — ce serait $T^*_{\Sigma_1} \cup T^*_{\Sigma_2} \cup T^*_\Gamma$). Les opérations non linéaires introduisent des singularités microlocales en des points (x_0, ξ_0) où u n'en a pas, sauf si on se limite à des singularités d'ordre $s' \leqslant 2s - s_0$ (cf. corollaire 5) ce qui "explique" la validité et les limites des résultats (b).

Si effectivement $\mathrm{SS}(g)$ contient T^*_Γ, on a

$$\sum a_\alpha(x)\, \partial^\alpha u = -g$$

et d'après les propriétés des équations linéaires, SS(u) va contenir dans l'avenir les conormaux des autres hypersurfaces caractéristiques $\Sigma_3, \ldots$ $\ldots, \Sigma_m$ passant par Γ. D'après (b), les singularités apparues sur $\Sigma_3, \ldots, \Sigma_m$ issues de l'interaction seront moins violentes ($2s - s_0$) que les singularités sur Σ_1 et $\Sigma_2(s)$.

On dispose maintenant de toute une série d'exemples de solutions u mettant en évidence ce phénomène d'interaction (voir [1], [13], [17] et [20]). Nous décrivons aux §§ 4 et 5 l'essentiel des résultats positifs (concluant à la régularité de u dans certaines régions) connus à ce jour.

2. Localisation et propagation des singularités ([4], [5], [6])

Rappelons d'abord quelques définitions sur la régularité microlocale des distributions. Soit $(x_0, \xi_0) \in \boldsymbol{R}^n \times (\boldsymbol{R}^n \setminus \{0\})$.

— On dit que $u(x)$ appartient *microlocalement à l'espace de Sobolev H^s au point* (x_0, ξ_0), si pour toute fonction $\varphi(x) \in C_0^\infty$, à support assez proche de x_0, on a

$$\widehat{\varphi u}(\xi)(1+|\xi|^2)^{s/2} \in L^2(\Gamma),$$

où Γ est un cône ouvert de $\boldsymbol{R}^n$, contenant ξ_0. Il est équivalent de dire qu'il existe un opérateur pseudo-différentiel classique E, d'ordre 0, de symbole principal non nul en (x_0, ξ_0), tel que $Eu \in H^s$.

— On dit de même que $u(x)$ appartient *microlocalement à l'espace de Hölder C^α*,[1] s'il existe un opérateur E comme ci-dessus tel que $Eu \in C^\alpha$.

L'ensemble des (x, ξ) tels que u n'y soit pas microlocalement C^∞ n'est autre que le spectre singulier (ou front d'onde) de u.

Revenons à l'équation (1) munie d'une solution u:

$$F\big(x, c(x), u(x), \ldots, \partial^\beta u(x), \ldots\big) = 0$$

(aucune hypothèse d'hyperbolicité n'étant faite). Si c est continue, et si u est m fois continûment différentiable, on définit le "symbole principal" de (1) (au facteur $(i)^m$ près) par

$$p_m(x,\xi) = \sum_{|\alpha|=m} \frac{\partial F}{\partial u_\alpha}\big(x, \ldots, \partial^\beta u(x), \ldots\big)\xi^\alpha.$$

---

[1] Pour $\alpha = k+\beta$, $k \in \boldsymbol{N}$, $0 < \beta < 1$, on dit que u appartient à C^α si $|\partial^\lambda u(x) - \partial^\lambda u(y)| \leqslant c|y-x|^\beta$ pour $|\lambda| \leqslant k$. Pour α entier, les espaces C^α doivent être remplacés par les classes de Zygmund correspondantes.

— Nous dirons que (x_0, ξ_0) est *caractéristique* pour (1) si on a $p_m(x_0, \xi_0) = 0$.

— Nous appellerons *bicaractéristiques* les courbes intégrales du champ hamiltonien de p_m (en supposant p_m à gradient lipschitzien). La bicaractéristique issue de (x_0, ξ_0) est la solution $x(s)$, $\xi(s)$ de

$$\frac{dx_i}{ds} = \frac{\partial p_m}{\partial \xi_i}(x(s), \xi(s)); \quad \frac{d\xi_i}{ds} = -\frac{\partial p_m}{\partial x_i}; \quad x(0) = x_0; \quad \xi(0) = \xi_0.$$

Les résultats fondamentaux de ce paragraphe sont les suivants:

THÉORÈME 1. *Soit u une solution de* (1) *appartenant à $C^{m+\varrho}$ (resp. $H^{n/2+m+\varrho}$) avec $\varrho > 0$, et supposons que le contrôle c appartienne à C^ϱ (resp. $H^{n/2+\varrho}$). Soit (x_0, ξ_0) un point non caractéristique tel que c appartienne microlocalement à $C^{\varrho'}$ (resp. $H^{n/2+\varrho'}$) en (x_0, ξ_0). Alors u appartient microlocalement à $C^{m+\varrho''}$ (resp. $H^{n/2+m+\varrho''}$) en (x_0, ξ_0) avec $\varrho'' = \mathrm{Min}(\varrho + \varrho', 2\varrho)$.*

En l'absence de singularités du contrôle, on obtient un gain de régularité ϱ en tous les points non-caractéristiques.

THÉORÈME 2. *Soit u une solution de* (1) *appartenant à $H^{n/2+m+1+\varrho}$, $\varrho > 0$, et telle que $p_m(x, \xi)$ soit à gradient lipschitzien. On suppose que c appartient à $H^{n/2+1+\varrho}$.*

Soit (x_0, ξ_0) un point caractéristique, et Γ un arc de bicaractéristique issu de (x_0, ξ_0). Soit $\sigma \leqslant 2\varrho$, et supposons que u appartienne à $H^{n/2+m+1+\sigma}$ microlocalement en (x_0, ξ_0), et que c appartienne à $H^{n/2+1+\sigma}$ microlocalement en tout point de Γ. Alors u appartient à $H^{n/2+m+1+\sigma}$ microlocalement en tout point de Γ.

Remarques. (a) Pour des équations moins méchamment non-linéaires, la régularité minimale exigée pour u s'affaiblit. On trouvera dans [6] les résultats précis, ainsi que la manière de les étendre à des équations qui ne sont pas sous la forme (1) [Par exemple, $\sum\sum \frac{\partial}{\partial x_i}\left(A_{ij}(x, u)\frac{\partial u}{\partial x_j}\right) = 0$.]

(b) Pour des équations linéaires, en considérant le second membre comme terme de contrôle, on ne retrouve que des résultats classiques. On obtient par contre des résultats non triviaux sur les équations linéaires à coefficients peu réguliers en prenant comme terme de contrôle l'ensemble des coefficients (voir aussi [2]).

(c) Revenons au cas de l'introduction où l'équation est strictement

hyperbolique par rapport à la direction x_n (on peut remplacer $\boldsymbol{R}^n$ et $\boldsymbol{R}^n_-$ par un ouvert Ω et $\Omega^- = \Omega \cap [x_n < 0]$ à condition que Ω soit dans le domaine d'influence de Ω^-: les bicaractéristiques nulles rétrograde issues d'un point de Ω entrent dans Ω^- avant de sortir de Ω.

Siune solution u appartient à H^s $(s = n/2 + m + 1 + \varrho)$ et est de classe C^∞ dans Ω^-, il résulte des théorèmes 1 et 2 que u appartient microlocalement à $H^{s+\varrho+1}$ aux points non caractéristiques, et à $H^{s+\varrho}$ aux points caractéristiques. La fonction u appartient donc à $H^{s+\varrho}$ dans Ω tout entier. Par récurrence, on en déduit que $u \in C^\infty(\Omega)$. Il ne peut pas apparaître de chocs dès que la régularité de u dépasse $H^{n/2+m+1}$.

Pour le problème de Cauchy, on a le résultat suivant.

THÉORÈME 3. *Soient* $s = n/2 + m + 1 + \varrho$; $\varrho > 0$; $0 \leqslant \theta \leqslant \varrho$; $t < s + \theta$. *Soit* u *une solution de* (1) *avec* $u \in H^s(\Omega)$ *et* $c \in H^{s-m}(\Omega)$. *Soit* $\tilde{x}', \tilde{\xi}' \in \boldsymbol{R}^{n-1} \times (\boldsymbol{R}^{n-1} \setminus \{0\})$ *et soit* $\tilde{\xi}_n \in \boldsymbol{R}$ *tel que le point* $(\tilde{x}', 0; \tilde{\xi}', \tilde{\xi}_n)$ *soit caractéristique. Soit* Γ *un arc de bicaractéristique issu de ce point. Sous les trois hypothèses suivantes, on a* $u \in H^t$ *microlocalement en tout point de* Γ.

(a) $\left(\dfrac{\partial}{\partial x_n}\right)^j u(x', 0) \in H^{s+\theta+1/2-j}$ *microlocalement en* $(\tilde{x}', \tilde{\xi}')$, *pour* $j = 0, \ldots, m-1$.

(b) $c(x) \in H^{s-m+\theta+1}$ *microlocalement en tout point de* Γ.

(c) *Pour* $\varphi(x) \in C_0^\infty(\boldsymbol{R}^n)$, *à support près de* $(\tilde{x}', 0)$, *et pour* G *voisinage cônique de* $\tilde{\xi}'$ *dans* $\boldsymbol{R}^{n-1}$, *on a*

$$\widehat{\varphi c}(\xi)(1 + |\xi|^2)^{(s-m)/2}(1 + |\xi'|^2)^{(\theta+1)/2} \in L^2(G \times \boldsymbol{R})$$

(*appartenance microlocale aux espaces* $H^{s-m,\theta+1}$ *de* [11]).

On construit d'abord, grâce à un théorème de relèvement des traces, une fonction $v(x)$ vérifiant $(\partial/\partial x_n)^k v(x', 0) = (\partial/\partial x_n)^k u(x', 0)$ pour $0 \leqslant k \leqslant [s - 1/2]$, et appartenant microlocalement à l'espace $H^{s,1+\theta}$, puis on applique le théorème 2 à l'équation vérifiée par la fonction $w(x)$ égale à $u(x) - v(x)$ pour $x_n > 0$, et à 0 sinon.

En utilisant également les techniques paradifférentielles, P. Godin [10] a étudié le problème des dérivées obliques non linéaires (non nécessairement elliptique):

$$F(x, u, \nabla u, \nabla^2 u) = 0 \quad \text{dans } \Omega,$$

$$f(x, u, \nabla u) = 0 \quad \text{sur } \partial\Omega$$

en supposant que le problème linéarisé est sous-elliptique. Il montre alors que les solutions assez régulières de ce problème sont en fait C^∞.

3. Calcul paradifférentiel et linéarisation ([4], [5], [6])

3.1. Paramultiplication. Soit $a(x)$ une foction de classe C^ϱ ($\varrho > 0$ non entier). Nous lui associerons l'opérateur suivant T_a (paramultiplication par a) défini par

$$\widehat{T_a u}(\xi) = \int \chi(\xi-\eta, \eta)\, \hat{a}(\xi-\eta)\, \hat{u}(\eta)\, d\eta/(2\pi)^n$$

où χ est une fonction indéfiniment dérivable hors de 0, homogène de degré 0, égale à 1 pour $|\xi-\eta| \leqslant \varepsilon_1|\eta|$ et à 0 pour $|\xi-\eta| \geqslant \varepsilon_2|\eta|$ ($0 < \varepsilon_1 < \varepsilon_2 < 1$).

Contrairement à la multiplication par a, l'opérateur T_a applique H^s (resp. C^σ) dans lui-même quels que soient s et σ. Sa définition fait intervenir une fonction arbitraire χ, mais un changement de fonction χ ne modifie T_a que par l'addition d'un opérateur ϱ-régularisant appliquant H^s (resp. C^σ) dans $H^{s+\varrho}$ (resp. $C^{\sigma+\varrho}$) quels que soient s et σ.

Nous renvoyons à [6] pour la définition (beaucoup plus parlante) et l'étude de T_a à partir de la décomposition de Littlewood–Paley et des techniques de [9]. Les propriétés de composition ($T_a \circ T_b - T_{ab}$ est ϱ-régularisant) et de commutation permettent d'englober dans un même calcul les paramultiplications et les opérateurs pseudo-différentiels classiques.

3.2. Symboles et opérateurs paradifférentiels. On dit que $p(x, \xi)$ appartient à la classe de symboles Σ^m_ϱ si

$$p(x, \xi) = p_m(x, \xi) + p_{m-1}(x, \xi) + \ldots + p_{m-[\varrho]}(x, \xi),$$

où p_{m-j} est de classe $C^{\varrho-j}$ par rapport à x, de classe C^∞ en ξ, et homogène de degré $m-j$ en ξ.

A chaque classe de symboles correspond une classe d'opérateurs $\mathrm{Op}(\Sigma^m_\varrho)$, et on a les propriétés suivantes.

(a) Si $P \in \mathrm{Op}(\Sigma^m_\varrho)$, P applique H^s dans H^{s-m} (et C^δ dans $C^{\delta-m}$). Si $u \in H^s$ et si u appartient à $H^{s'}$ microlocalement en (x_0, ξ_0) alors $Pu \in H^t$ microlocalement en (x_0, ξ_0) avec $t = \mathrm{Min}(s'-m,\ s-m+\varrho)$.

(b) A chaque $P \in \mathrm{Op}(\Sigma^m_\varrho)$ correspond son symbole $\sigma(P) \in \Sigma^m_\varrho$. Les formules donnant les symboles du composé $\sigma(P \circ Q)$ et de l'adjoint $\sigma(P^*)$ sont les formules classiques en calcul pseudo-différentiel, mais arrêtées aux $[\varrho]$ premiers termes.

(c) Si $a \in C^\varrho$, on a $T_a \in \mathrm{Op}\ \Sigma^0_\varrho$ et $\sigma(T_a) = a(x)$. Si P est un opérateur pseudo-différentiel classique d'ordre m, on a $P \in \mathrm{Op}\ \Sigma^m_\varrho$ pour tout ϱ, et $\sigma(P)$ est son symbole classique, arrêté aux $[\varrho]$ premiers termes.

(d) Si $P \in \mathrm{Op}(\Sigma^m_\varrho)$, on a $\sigma(P) = 0$ si et seulement si P applique H^s dans $H^{s-m+\varrho}$.

(e) Si le symbole principal de P est non nul, P est inversible modulo un opérateur ϱ-régularisant.

Au vu de ces propriétés, il est clair qu'un grand nombre des démonstrations classiques utilisant les opérateurs pseudo-différentiels s'étendront aux opérateurs paradifférentiels, la différence essentielle étant que les résultats seront obtenus "modulo un opérateur ϱ-régularisant" au lieu de "modulo un opérateur infiniment régularisant".

Des extensions du calcul paradifférentiel au cas de symboles non homogènes sont dûes à Y. Meyer [14] d'une part, et à P. Godin [10] d'autre part.

3.3. Fonctions composées et paramultiplication. En comparant l'expression de $\widehat{au}$ à l'expression du n° 3.1, on voit que $au = T_a u + T_u a + r$, où r correspond à l'intégration sur le domaine $\varepsilon_1|\eta| \leqslant |\xi - \eta| \leqslant (1/\varepsilon_1)|\eta|$. Le terme $T_a u$ a la régularité locale et microlocale de u, le terme $T_u a$ a la régularité locale et microlocale de a, quant à r, il appartient à $H^{s+t-n/2}$ si $a \in H^s$ et $u \in H^t$. Le résultat clef de la théorie est la généralisation suivante.

THÉORÈME 4. *Soient $u_1, \ldots, u_N$ appartenant à C^ϱ [resp. $H^{n/2+\varrho}$], $\varrho > 0$, et $F(x, u_1, \ldots, u_N)$ une fonction C^∞ de ses arguments. On a*

$$F\big(x, u_1(x), \ldots, u_N(x)\big) = \sum_1^N T_{\partial F/\partial u_i} \cdot u_i + r$$

avec $r \in C^{2\varrho}$ [resp. $H^{n/2+2\varrho}$].

Nous avons démontré ce théorème dans [6], avec le résultat plus faible $r \in H^{n/2+2\varrho-\varepsilon}$ pour tout $\varepsilon > 0$. Le résultat précis est dû à Y. Meyer [14].

En appliquant 3.2 (a) et le théorème précédent au produit uv, on retrouve le résultat suivant.

COROLLAIRE 5. (a) *L'espace des fonctions u appartenant à C^ϱ et appartenant à $C^{\varrho'}$ microlocalement en (x_0, ξ_0) est une algèbre dès que $\varrho > 0$ et $\varrho' \leqslant 2\varrho$.*

(b) (J.-M. Bony, B. Lascar, J. Rauch). *L'espace des fonctions appartenant à H^s et appartenant microlocalement à $H^{s'}$ est une algèbre dès que $s > n/2$ et $s' \leqslant 2s - n/2$.*

Remarque. Historiquement c'est ce résultat qui a été le point de départ en 1978 des travaux sur la propagation des singularités pour les équations non linéaires. La démonstration directe que nous en avions donnée (voir [3]) contenait déjà en germe l'idée de la paramultiplication. Indépendamment et simultanément, B. Lascar [13] et J. Rauch [16], démontraient également la partie (b). Ces trois auteurs en déduisaient divers cas particuliers des théorèmes 1 et 2.

Ultérieurement, plusieurs auteurs ont retrouvé d'autres cas des théorèmes 1 et 2 en démontrant, sans recourir au calcul symbolique, la régularité microlocale de solutions d'équations aux dérivées partielles à coefficients peu réguliers (voir [2], [8]).

3.4. Linéarisation des équations non linéaires.

THÉORÈME 6. *Soit u une solution de* (1) *et appartenant à $C^{m+\varrho}$ [resp. $H^{n/2+m+\varrho}$], $\varrho > 0$, avec $c \in C^{\varrho}$ [resp. $H^{n/2+\varrho}$]. Alors u vérifie l'équation paradifférentielle suivante:*

$$Pu = Ec + r$$

où $P \in \mathrm{Op}(\Sigma^m_{\varrho})$; $E \in \mathrm{Op}(\Sigma^0_{\varrho})$; $r \in C^{2\varrho}$ [resp. $H^{n/2+2\varrho}$]. Le symbole principal de P est (au facteur i^m près)

$$p_m(x, \xi) = \sum_{|\alpha|=m} \frac{\partial F}{\partial u_\alpha}(x, c, u, \ldots, \partial^\beta u, \ldots)\xi^\alpha.$$

C'est une conséquence immédiate du théorème 4

$$F(x, c, u, \ldots) = \sum T_{\partial F/\partial u_\beta} \cdot \partial^\beta u + T_{\partial F/\partial c} \cdot c + r' = 0$$

en posant $P = \sum_{|\beta| \geq m-[\varrho]} T_{\partial F/\partial u_\beta} \cdot \partial^\beta$ (les autres termes entreront dans le reste) et $E = -T_{\partial F/\partial c}$.

Remarque. Pour l'équation (2), sous l'hypothèse $u \in C^{m-1+\varrho}$, le second membre est du même type, et on a $P \in \mathrm{Op}(\sum^m_{\varrho+1})$, le symbole principal de P étant celui de l'opérateur linéaire $\sum a_\alpha \partial^\alpha$. On trouvera dans [6] le calcul de la classe de P et de la régularité du second membre selon le type de non linéarité de l'équation.

Les théorèmes 1 et 2 sont donc ramenés à l'étude de la localisation et de la propagation des singularités pour des équations paradifférentielles linéaires. On peut alors "recopier", avec un certain nombre de modifications pour le théorème 2, les démonstrations classiques dans le

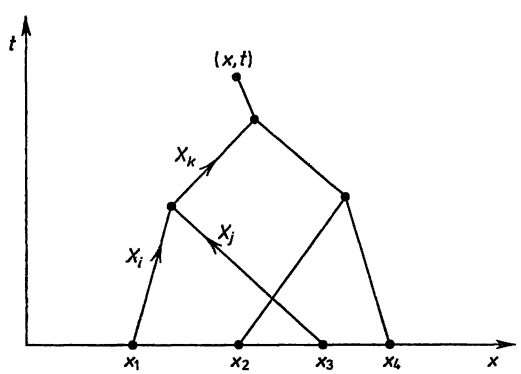
cas pseudo-différentiel: inversion microlocale des opérateurs non caractéristiques pour le théorème 1, estimations d'énergie de Hörmander [12] pour le théorème 2.

4. Interaction des singularités (Dimension 1 d'espace)

J. Rauch et M. Reed ont consacré une série d'articles ([17], [18], [19]) au cas des équations et systèmes semi-linéaires strictement hyperboliques en dimension 1 d'espace. Nous décrivons ci-dessous le résultat principal de [19] qui est essentiellement optimal. Soit

$$\frac{\partial u_i}{\partial t} + c_i(x,t)\frac{\partial u_i}{\partial x} = f_i(t, x, u_1, \ldots, u_m) \tag{3}$$

un système de m équations, où les f_i et les c_i sont réelles et C^∞, et où on a $c_i(x,t) \neq c_j(x,t)$ pour $i \neq j$. On note U le vecteur de composantes u_i, et X_i le champ de vecteurs $\partial/\partial t + c_i(x,t)\partial/\partial x$.



Nous appellerons arbre caractéristique un arbre du type ci-dessus dont les lignes sont des arcs caractéristiques (courbes intégrales de l'un des champs X_i) et tels qu'en chaque sommet, il arrive deux courbes intégrales de X_i et X_j ($i \neq j$) et il parte une courbe intégrale de X_k ($k \neq i$ et $k \neq j$). Pour un arbre A, nous noterons $\Pi(A)$ l'ensemble des "pieds" de l'arbre ($\Pi(A) = \{x_1, x_2, x_3, x_4\}$ sur le schéma).

THÉORÈME 7 (J. Rauch et M. Reed). *Soit U une solution bornée de* (3). *On suppose que les données de Cauchy $U_0(x) = U(x, 0)$ appartiennent à H^s pour $s > 1/2$. Soit $S(x)$ une fonction telle que, pour tout x, $U_0(x)$ appartienne à $H^{S(x)}$ au voisinage de x. Alors $U(x,t)$ appartient à $H^{\sigma(x,t)}$ au voi-*

sinage de chaque (x, t), *avec*

$$\sigma(x, t) = \operatorname*{Inf}_{A} \Big(\sum_{x_i \in \Pi(A)} S(x_i) \Big)$$

l'infimum étant pris sur tous les arbres caractéristiques aboutissant en (x, t).

En fait, les résultats de Rauch et Reed sont plus précis et sont exprimés en termes d'espaces $\mathscr{A}(\varrho; r_1, \ldots, r_m)$ que nous allons décrire.

On définit d'abord l'espace $(H^s)^k_{X_i}$ comme l'espace des u appartenant à H^s, telles que $X_i^l u$ appartienne à H^s pour $l \leqslant k$ pour k entier; et par interpolation (par exemple) pour k réel positif.

Si ϱ, $r_1, \ldots, r_m$ sont des réels vérifiant $0 \leqslant r_j \leqslant \varrho$, on dit que $u \in \mathscr{A}(\varrho; r_1, \ldots, r_m)$ si

(a) en chaque point $(x_0, t_0, \xi_0, \tau_0)$ non caractéristique $(\tau_0 + c_i(x_0, t_0)\,\xi_0 \neq 0$ pour tout $i)$ on a $u \in H^\varrho$ microlocalement en $(x_0, t_0, \xi_0, \tau_0)$,

(b) en chaque point $(x_0, t_0, \xi_0, \tau_0)$ caractéristique pour X_i, on a $u \in (H^{r_i})^{\varrho - r_i}_{X_i}$ microlocalement en $(x_0, t_0, \xi_0, \tau_0)$.

Le résultat important est que $\mathscr{A}(\varrho; r_1, \ldots, r_m)$ est une algèbre si on a $r_j > 1/2$, $\varrho > 1$, $\varrho \leqslant \operatorname{Min}_{i \neq j} (r_i + r_j)$. Contrairement aux espaces définis uniquement en terme d'appartenance microlocale aux espaces de Sobolev (cf. corollaire 5), on dispose ici d'espaces stables par les opérations non linéaires, où la régularité microlocale contrôlée (ϱ) peut dépasser de beaucoup le double de la régularité locale $s = \operatorname{Min}_i (r_i)$. C'est ce qui permet (entre autres) l'étude de l'interaction.

Dans [18], Rauch et Reed ont étudié le même problème, dans le cas où les données de Cauchy et donc la solution sont indéfiniment dérivables par morceaux. L'un des résultats assure que le phénomène d'interaction se produit "en général". Plus précisément, lors du croisement de deux singularités portées par des courbes intégrales de X_i et X_j respectivement, il naît une singularité portée par la courbe intégrale de X_k dès que $(\partial^2 f_k / \partial x_i \partial x_j)\,(x, t, U) \neq 0$.

Le problème de la réflexion des singularités pour de tels systèmes est étudié par M. Oberguggenberger dans [15].

5. Interaction des singularités (dimension quelconque)

5.1. Distributions conormales. Nous allons brièvement décrire les résultats de [7]. Nous allons d'abord associer, à des configurations géométriques simples, des espaces qui (a) sont des algèbres, (b) sont stables par l'action

des opérateurs pseudo-différentiels, (c) permettent un contrôle de la régularité microlocale au-delà du double de la régularité locale.

DÉFINITION. (a) Soit V une sous-variété de codimension d de $\boldsymbol{R}^n$, on note $H^{s,k}_V$, $s \in \boldsymbol{R}$, $k \in \boldsymbol{N}$, l'espace des u appartenant à H^s telles que l'on ait $Z_{i_1} Z_{i_2} \dots Z_{i_l} u \in H^s$ pour $l \leqslant k$ lorsque les Z_i sont des champs de vecteurs tangents à V.

(b) Soit Σ la réunion de m hypersurfaces $\Sigma_1, \dots, \Sigma_m$ se coupant deux à deux transversalement le long d'une variété Γ de codimension 2. On note $H^{s,k}_\Sigma$ l'espace des $u \in H^s$ tels que $M_{i_1} M_{i_2} \dots M_{i_l} u \in H^s$ pour $l \leqslant k$, lorsque les M_i sont des opérateurs pseudo-différentiels d'ordre 1 dont le symbole principal s'annule sur la réunion des variétés conormales à Γ et aux Σ_i.

Dans le cas (a), et pour $k = \infty$, les éléments de $H^{s,k}_V$ sont effectivement des distributions de Fourier associées à la variété lagrangienne conormale à V.

Dans le cas (b), si $u \in H^{s,k}_\Sigma$, on a $u \in H^{s+k}$ au voisinage de tout point de $\boldsymbol{R}^n \setminus \Sigma$, et microlocalement en dehors des conormaux à Γ et aux Σ_i, et $u \in H^{s,k}_{\Sigma_i}$ au voisinage de tout point de $\Sigma_i \setminus \Gamma$.

Ces espaces sont des algèbres pour $s > n/2$, ce qui est évident dans le cas (a), mais non dans le cas (b) (on ne peut pas remplacer les M_i par des champs de vecteurs).

5.2. Interaction de deux singularités [7]. Revenons à l'équation semi-linéaire (2) supposée strictement hyperbolique (avec $c = 0$ pour simplifier).

THÉORÈME 9. *Soient Σ_1 et Σ_2 deux hypersurfaces caractéristiques, disjointes pour $x_n < 0$, et se coupant en Γ. Soient $\Sigma_3, \dots, \Sigma_m$ les autres hypersurfaces caractéristiques issues de Γ et $\Sigma = \bigcup_1^m \Sigma_j$.*

Soit u une solution de (2) appartenant à H^s $(s > n/2 + m)$ et telle que, pour $x_m < 0$, on ait $u \in H^{s+k}$ hors de $\Sigma_1 \cup \Sigma_2$ et $u \in H^{s,k}_{\Sigma_i}$ près de Σ_i $(i = 1, 2)$.

Alors on a $u \in H^{s,k}_\Sigma$, et plus précisément pour $x_n > 0$:

- *Hors de Σ on a $u \in H^{s+k}$.*
- *Près de $\Sigma_i \setminus \Gamma$; $i = 1, 2$; on a $u \in H^{s,k}_{\Sigma_i}$.*
- *Près de $\Sigma_j \setminus \Gamma$; $j = 3, \dots, m$; on a en posant $\varrho = s + 1 - n/2 - m$ $u \in H^{s+\varrho,[k-\varrho]}_{\Sigma_j}$ si $k > \varrho$, $u \in H^{s+k}$ sinon.*

Bien entendu ce résultat est local. Il n'est valable globalement dans $\boldsymbol{R}^n$ que si les Σ_j ne se recoupent pas et restent régulières. On peut toutefois réappliquer le théorème tant que les Σ_j ne se recoupent que deux à deux.

5.3. Problème de Cauchy. Dans le cas le plus simple de données de Cauchy singulières (distributions conormales associées à une hypersurface de $x_n = 0$), on a le résultat suivant, sous les hypothèses du n° 5.2.

THÉORÈME 10 [7]. *Soit H l'hyperplan $x_n = 0$ et Γ une hypersurface de H. Soient Σ_j, $j = 1, \ldots, m$ les hypersurfaces caractéristiques passant par Γ et $\Sigma = \bigcup \Sigma_j$. Soit u une solution de* (2) *appartenant à H^s, $s > n/2 + m$, telles que les données de Cauchy $\gamma_j u(x') = (\partial/\partial x_n)^j u(x', 0)$ appartiennent à $H_\Gamma^{s-j+1/2,k}$ pour $j = 0, \ldots, m-1$.*

Alors $u \in H_\Sigma^{s,k}$, et en particulier, $u \in H^{s+k}$ hors de Σ et $u \in H_{\Sigma_j}^{s,k}$ près de $\Sigma_i \setminus \Gamma$.

Il peut sembler restrictif de se limiter à des distributions conormales, mais un résultat remarquable de M. Beals [1] montre que des hypothèses de ce type ne peuvent être totalement évitées.

Pour une solution assez régulière de l'équation des ondes non linéaires

$$\Box u = \frac{\partial^2 u}{\partial x_n^2} - \sum_1^{n-1} \frac{\partial^2 u}{\partial x_j^2} = f(u)$$

il n'est pas très difficile de montrer que, si les données de Cauchy sont des distributions conormales associées à l'origine ($\gamma_j u \in H_0^{s-j+1/2,k}$ pour $j = 0, 1$), alors la solution u appartient à H^{s+k} en dehors de la surface G du cône d'onde et à $H_G^{s,k}$ près de $G \setminus \{0\}$. Par contre, M. Beals montre qu'il existe des données de Cauchy C^∞ en dehors de l'origine, telles que la solution u soit singulière non seulement sur G, mais à l'intérieur de ce cône d'onde.

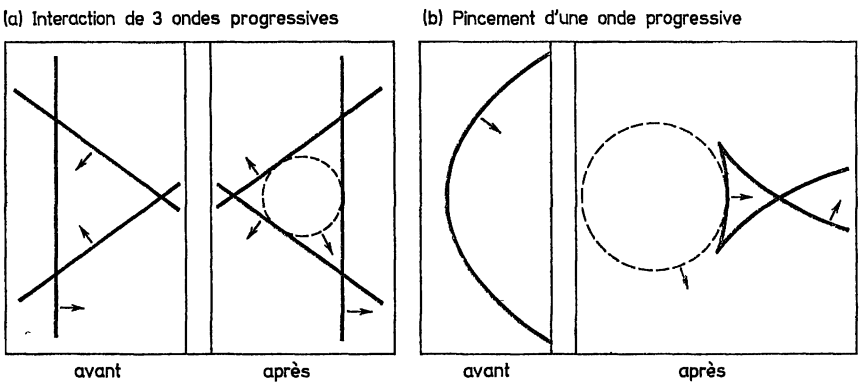
THÉORÈME 11 (M. Beals [1]). (a) *Soit u une solution appartenant à H^s, $s > n/2$, de $\Box u = f(u)$. Supposons que toutes les bicaractéristiques nulles issues de $\tilde{w}$ rencontrent l'hyperplan $x_n = 0$ en des points où $\gamma_0 u$ et $\gamma_1 u$ sont C^∞. Alors $u \in H^{3s-n+2-\varepsilon}$ près de $\tilde{w}$.*

(b) *Pour $n > 2$ et $s > n/2$ il existe une solution u appartenant à H^s, de $\Box u = \beta(x) u^3$ avec $\beta \in C^\infty$, dont les données de Cauchy sont C^∞ hors de l'origine, et dont le support singulier est l'intérieur du cône d'onde: $|x'| \leqslant |x_n|$.*

5.4. Conjectures. On peut raisonnablement conjecturer le résultat suivant: Soient $\Sigma \subset \hat{\Sigma}$ deux sous-ensembles de codimension 1, stratifiés en sous-variétés C^∞, et désingularisables. On suppose que Σ et $\hat{\Sigma}$ sont caractéristiques aux points lisses et que la réunion des variétés conormales aux strates

de $\hat{\Sigma}$ est stable par le flot hamiltonien de p_m. Alors, si u est une solution de (1) appartenant à H^s, $s > n/2 + m$, et appartenant à $H^{s,\infty}_{\Sigma}$ pour $x_n < 0$, on a $u \in C^\infty$ hors de $\hat{\Sigma}$, et $u \in H^{s,\infty}_{\hat{\Sigma}}$ aux points lisses de $\hat{\Sigma}$.

Pour l'équation des ondes $\Box u = F(x, u, \nabla u)$ en dimension 2 d'espace, les deux cas les plus simples non encore résolus sont les suivants:



On a figuré en pointillé les singularités prévues par la conjecture et moins violentes que les singularités initiales.

Rauch et Reed ont donné récemment [20] un exemple de solution ayant le comportement (a).

Ajouté sur épreuves: Nous avons démontré [21] la validité de la conjecture dans le cas (a) ci-dessus. Voir aussi [22] pour un résultat voisin.

Bibliographie

[1] Beals M., Self spreading and strength of singularities for solutions to semilinear wave equations, *Ann. of Math.* **118** (1983), pp. 187–214.

[2] Beals M. and Reed M., Propagation of singularities for hyperbolic pseudodifferential operators with nonsmooth coefficients, *Comm. Pure Appl. Math.* **XXXV** (1982), pp. 169–184.

[3] Bony J.-M., Localisation et propagation des singularités pour les équations non linéaires, *Journées des Equations aux Dérivées Partielles*, St Jean de Monts (1978).

[4] Bony J.-M., Calcul symbolique et singularités des solutions des équations aux dérivées partielles non linéaires, *Journées des Equations Aux Dérivées Partielles*, St Cast (1979).

[5] Bony J.-M., Singularités des solutions des équations aux dérivées partielles non linéaires, *Proceedings, Les Houches 1979*, Lect. Notes in Physics **126**, pp. 123–129.

[6] Bony J.-M., Calcul symbolique et propagation des singularités pour les équations aux dérivées partielles non linéaires, *Ann. Sci. Ec. Norm. Sup. 4ème série* **14** (1981), pp. 209–246.

[7] Bony J.-M., Interaction des singularités pour les équations aux dérivées partielles non linéaires, *Sém. Goulaouic-Meyer-Schwartz* 1979/80, n° **22** et 1981/82, n° **2**.
[8] Chen S., *Pseudo-differential operators with finitely smooth symbol and their applications to quasilinear equations*, (preprint).
[9] Coifman R. et Meyer Y., Au-delà des opérateurs pseudo-différentiels, *Astérisque* **57** (1978).
[10] Godin P., *Subelliptic non linear oblique derivative problems*, (preprint).
[11] Hörmander L., *Linear Partial Differential Operators*, Springer-Verlag, 1963.
[12] Hörmander L., On the existence and the regularity of solutions of linear pseudo-differential equations, *L'enseignement Math.* **17** (1971), pp. 99–163.
[13] Lascar B., Singularités des solutions d'équations aux dérivées partielles non linéaires, *C. R. Acad. Sci. Paris* **287** A (1978), pp. 521–529.
[14] Meyer Y., Régularité des solutions équations aux dérivées partielles non linéaires, *Séminaire Bourbaki*, 1979/1980, n° **560**.
[15] Oberguggenberger M., *Propagation of singularities for semilinear mixed hyperbolic systems in two variables*, (preprint).
[16] Rauch J., Singularities of solutions of semilinear wave equations, *J. Math. Pures et Appl.* **58** (1979), pp. 299–308.
[17] Rauch J. and Reed M., Propagation of singularities for semilinear hyperbolic equations in one space variable, *Annals of Math.* **111** (1980), pp. 531–552.
[18] Rauch J. and Reed M., Jump discontinuities of semilinear strictly hyperbolic systems in two variables: creation and propagation, *Comm. Math. Phys.* **81** (1981), pp. 203–227.
[19] Rauch J. and Reed M., Non-linear microlocal analysis of semilinear hyperbolic systems in one space dimension, *Duke Math. Journ.* **49**, 2 (1982), pp. 397–475.
[20] Rauch J. and Reed M., Singularities produced by the nonlinear interaction of three progressing waves; examples, *Comm. P. D. E.* **7** (1982), pp. 1117–1133.

Ajouté sur épreuves:

[21] Bony J.-M., Interaction des singularités pour les équations de Klein–Gordon non linéaires, *Sém. Goulaouic–Meyer–Schwartz*, 1983-84, No. **10**.
[22] Melrose R. and Ritter N., *Interaction of non linear progressing waves*, (preprint).

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

V. S. BUSLAEV

Regularization of Many-Particle Scattering

For a long time the Schrödinger operator which describes the motion of a quantum particle in a field of a potential decreasing at infinity remained the basic model of the mathematical scattering theory. After the well-known works of L. D. Faddeev dedicated to the investigation of a scattering pattern in the 3-particle system, the Schrödinger operator of this system (and generally of the many-particle system) also became one of the main models in scattering theory. The peculiarity of the many-particle Schrödinger operator H manifests itself most expressively in the behaviour of the corresponding dynamical system

$$i\psi_t = H\psi$$

for large t (time). In such a system qualitatively new asymptotic types of behaviour can be formed which are absent in the case of non-interacting particles. According to the heuristic conceptions of scattering theory the appearance of such asymptotical types of behaviour is equivalent to the existence of "junior branches of the spectrum" of the operator H, which can be characterized in terms of the discrete spectrum of subsystems. However, the difference between many-particle scattering and the scattering of a particle by the potential is not limited by this phenomenon. Even in the case where new asymptotical types of behaviour do not arise the scattering differs essentially from the scattering of a particle by the potential. Such differences can be called, a little indefinitely, *singularities of many-particle scattering*. Although the singularities of many-particle scattering cannot be ignored, their role in different constructions varies greatly. There is a set of important formula statements in the scattering theory of a particle by the potential whose propagation in the region of the many-particle scattering depends critically on the concrete structure of these singularities. Another essential moment is that the singularities of many-particle scattering can be described ex-

plicitly on the level of heuristic constructions. Owing to this fact it is well to suggest new methods of justification of the heuristic conceptions of scattering theory.

All the results given in this report are connected with the singularities of many-particle scattering. For simplification of the text we assume that the spectrum of the operator H does not contain the junior branches. Besides, as a rule, we shall suppose that the number of particles n is equal to 3. All the motivations and statements of problems are meaningful for any n. Due to the limited space of this article it is impossible to list explicitly even the basic formulas; therefore we shall confine ourselves only to characterizing them descriptively. Some general formulas will be given explicitly and used as a support in the presentation.

1. Singularities of many-particle scattering

For the system of n d-dimensional particles with a pair interaction the Schrödinger operator is a differential operator of the form $\psi \mapsto -\Delta\psi + V\psi$, where $\psi = \psi(x) \in \boldsymbol{C}$, $x = (x_1, \ldots, x_n) \in \boldsymbol{R}^N$, $x_i \in \boldsymbol{R}^d$, $N = nd$, Δ is the Laplacian on $\boldsymbol{R}^N$, $V = \sum_{i<j} v(x_i - x_j)$, $v(\xi) \in \boldsymbol{R}$. The masses of the particles and the pair potentials are assumed to be identical. In those parts of this report where a specific character of one-dimensional particles will be discussed this assumption will be essential. To exclude the trivial effects produced by the motion of the centre of inertia the operator described by the same formula is to be restricted to functions ψ defined over the linear subspace $\Lambda = \{x\colon x_1 + \ldots + x_n = 0\} \subset \boldsymbol{R}^N$, rather than over $\boldsymbol{R}^N$; in this case Δ is simply the Laplacian on Λ. Under certain conditions this construction leads to a self-adjoint operator in $L_2(\Lambda)$ which we denote by H, and by H_0 in the case $v = 0$. To simplify the presentation we shall consider $v(\xi)$ to be a smooth function, decreasing at infinity with the speed of rather great degree of $|\xi|^{-1}$. At $n = 2$ the operator H describes the scattering by the potential decreasing at infinity.

Given these assumptions the continuous spectrum of the operator H coincides with $[0, \infty)$. A basis of eigenfunctions of the continuous spectrum can be constructed from functions $\psi(x, p)$, $x, p \in \Lambda$, constituting the solutions of the equation $-\Delta\psi + V\psi = p^2\psi$, which describe the scattering of plane waves $\exp(ipx)$ by the potential V. Solutions ψ are well known for $n = 2$; their existence for $n = 3$ follows from Faddeev's investigations. Under special assumptions, the existence of such solutions was

proved also for $n > 3$. The Fourier transform

$$W(p, p') = F^{-1}_{x \to p} \psi(x, p')$$

of ψ has a special structure

$$W(p, p') = \delta(p - p') - T(p, p')[E(p) - E(p') - io]^{-1}, \quad E(p) = p^2. \quad (1)$$

The kernel $W(p, p')$ generates the operator W, which can be characterized as a wave operator in the momentum representation

$$W = \underset{t \to -\infty}{\text{s.lim}} \exp(itH) \exp(-itH_0).^1$$

This formula can be taken as a definition of W and therefore of the solution ψ. The solution ψ can be characterized also by its asymptotic behaviour (in the distribution sense), as $x \to \infty$:

$$\psi(x, p) \sim \exp(ipx) + C_{N-d}(|p||x|)^{(1+d-N)/2} f(\hat{x}, p) \exp(i|p||x|).$$

Here $\hat{x} = x|x|^{-1}$, C_K depends only on K, and f is an indefinite coefficient, called the *scattering amplitude*. The following equality is valid:

$$f(\hat{x}, p) = \tfrac{1}{2}|p|^{(N-d-2)/2} T(\hat{x}|p|, p). \quad (2)$$

Finally, let us introduce an S-matrix, dependent on E, $E > 0$, in operator $S(E)$ which acts on functions $\Lambda_1 \to C$, where Λ_1 is the unit sphere in space Λ. The kernel of operator $S(E)$ is defined by the formula

$$S(\hat{p}, \hat{p}'; E) = \delta(\hat{p}, \hat{p}') - 2\pi i f(\hat{p}, \hat{p}' \sqrt{E}), \quad (3)$$

$\hat{p}, \hat{p}' \in \Lambda_1$. According to the heuristic conceptions of scattering theory the operator W in $L_2(\Lambda)$ must be isometric with a finite index and the operator $S(E)$ in $L_2(\Lambda_1)$ must be unitary.

In the problem of scattering by the decreasing potential the kernel T is a smooth function. On the contrary, for $n > 2$ the kernel T is a distribution with some characteristic singularities. We call them *singularities of many-particle scattering*. In terms of the kernels T for the subsystems a kernel $T_a(p, p')$ can be constructed such that $T - T_a$ will have the prescribed smoothness. The kernel T_a can be extracted from first iterations of the Faddeev equations (if $n = 3$) and of the Faddeev–Yakubovsky equations (if $n > 3$). Knowing T_a, we can construct f_a and S_a as well as W_a and ψ_a, i.e., the terms of the corresponding objects, involving the

[1] In various places of the text the same operator will be considered in different representations, and this will not be indicated by symbols.

singularities of many-particle scattering. The function ψ_a can be characterized independently as a smooth function for which the difference $\psi - \psi_a$ for $x \to \infty$ is a diverging wave with an amplitude, having the prescribed smoothness. It is important to note that the amplitude $f(\hat{x}, p)$ has singularities not only on the subspaces $\Lambda_{ij} = \{x: x_i = x_j\}$. To find the set on which f has singularities outside Λ_{ij} the rays connected with the eikonal $x \mapsto px$ of the plane wave have to be introduced and their (possibly multiple) scattering on the subspaces Λ_{ij} have to be considered. The directions of the scattered rays sweep out a set on Λ_1 which is the support of the singularities of $f(\cdot, p)$. This construction can easily be extended to describe the types of these singularities. If $d > 1$, the dimension of the support increases with the multiplicity of scattering, and the strength of the singularities, correspondingly, falls. After a certain number of scatterings the singularities disappear.

It is not difficult to describe the singularities somehow. Different tasks impose different requirements on the description of the singularities. Usually it is desirable to find such a representation for S_a as would immediately show that this operator is almost unitary in $L_2(\Lambda_1)$. Besides, this representation is desired to describe effectively the singularities of the value $\varphi(S(E))$ where φ is an arbitrary function. The same questions can be formulated also about the operator W_a.

The first strict formulas with a decisive role of the singularities of many-particle scattering seem to have been obtained in [6, 7]. These works dealt with trace formulas which we shall consider later. A series of subsequent works was dedicated to studying explicit formulas for the function ψ_a. As a starting point of the construction a form of the Faddeev equation was chosen. The formulas obtained contain some special functions and show a close connection between this problem and diffraction problems [13, 14].

The questions touched upon in this part have something in common with some problems arising in the theory of scattering of a particle by a potential decreasing may be slowly. We mean the problem concerning an explicit construction of the function $f_a(\hat{x}, p)$ that the difference $f - f_a$ has an assigned smoothness on the beam $\hat{x} = \hat{p}$ (forward scattering) [10, 17]. The singular term f_a is responsible for the asymptotic behaviour of the S-matrix spectrum [1, 2].

2. Method of unitary regularization

It has already been mentioned that the possibility of an explicit description of the singularities of many-particle scattering on the level of probable constructions gives possibility of a new approach to a strict justification

of the heuristic conceptions of scattering theory and of the structure of the singularities themselves. In this way the basis of an investigation of n-particle systems is as usual sufficiently exact information about the properties of kernels T of the subsystems.

In this part it will be convenient to use a representation which can be characterized as the spectral representation of the operator H_0. To pass to this representation from the momentum representation is to introduce spherical coordinates $p \leftrightarrow (E, a)$, $E = p^2$, $a = \hat{p}$ in the momentum space Λ and to consider the function $f(p)$ as a function on a semi-axis $E > 0$ whose values are in a space of function $\Lambda_1 \to C$. In this representation linear operators A will be described by kernels $A(E, E')$ whose values at fixed E and E' are integral operators in the space of functions $\Lambda_1 \to C$. The kernel $A(E, E')$ is naturally connected with the kernel $A(p, p')$ of the operator A in the momentum representation.

Suppose we have all reasons to expect that the singularities of the kernel T coincide with the singularities of some "known" kernel T_a. Let us use (1) to construct the operator W_a in terms of T_a. The problem we are going to discuss is, in its essence, a unitarization of the operator W_a by means of adding a smooth component to the kernel T_a. In course of the construction we shall make several assumptions about the structure of the kernel T_a, which can be regarded as indications of the fact that the kernel T_a indeed properly reflects the singularities of the kernel T.

Using formulas (2), (3), we can construct, in terms of T_a, the operator $S_a(E)$ in $L_2(\Lambda_1)$. A real operator $S(E)$ must be a continuous function of E which is factorizable at $E \to \infty$ into a product of unitary S-matrices of the subsystems. Formulas describing the singularities of T usually correspond to the limiting transition at $E \to \infty$. Let us introduce the operators which characterize the deviation of the operator S_a from a unitary one: $\Delta_1(E) = S_a^*(E) S_a(E) - I$, $\Delta_2(E) = S_a(E) S_a^*(E) - I$. Let us assume that:

(A) the operator $S_a(E)$ is a continuous function of E, $E > 0$, in the uniform operator topology, which becomes close to a unitary operator at $E \to \infty$, and operators Δ_1, Δ_2 have smooth kernels.

Under this condition the operator $S_a(E)$ has a finite index $\operatorname{ind} S_a(E)$. Since $\operatorname{ind} S_a(E)$ does not depend on E, and the index of a unitary operator is equal to zero, the index $\operatorname{ind} S_a(E)$ is also equal to zero. Therefore it is clear that operator $\tilde{S}_a(E) = S_a(E) \left(I + \Delta_1(E)\right)^{-1/2}$, isometric in accordance with the construction, will be unitary. The kernel of the operator $\tilde{S}_a(E)$ differs from that of the operator $S_a(E)$ only by a smooth addend and it is possible to introduce in the kernel T_a such a smooth addend that the corresponding operator $S_a(E)$ will be unitary. Let us consider the kernel

T_a to be chosen from the very beginning in such a way that the operator $S_a(E)$ is unitary.

As a second step let us introduce operators which characterize a deviation of the operator W_a from a unitary one: $D_1 = W_a^* W_a - I$, $D_2 = W_a W_a^* - I$. Simple calculations show that the kernels of operators D_1 and D_2 contain the addends $\Delta_1(E)(E-E'-io)^{-1}$ and $\Delta_2(E)(E-E'-io)^{-1}$ respectively. If the kernel T_a was chosen as a smooth kernel, the operators D_1 and D_2 would be the sums of smooth kernels and the indicated addends. In our case $\Delta_1 = 0$, $\Delta_2 = 0$; therefore it is to be assumed that:

(B) the operators D_1 and D_2 have smooth kernels, quickly decreasing at E, $E' \to \infty$.

Let us introduce the isometric operator $\tilde{W}_a = W_a(I+D_1)^{-1/2}$. As before, this operator has form (1) and its kernel $\tilde{T}_a$ differs from original kernel T_a only by a smooth, quickly decreasing addend. The index of the operator $\tilde{W}_a$ is finite: let us denote it by $\varkappa$. There are no reasons to expect that $\varkappa = 0$. Let R be an operator which has form (1) with a smooth quickly decreasing kernel T_R. It is not difficult to construct the kernel T_R in explicit form so that R will be an isometric operator with the same index $\varkappa$ and the product $R^{-1}\tilde{W}_a$ will be a unitary operator. In this case the operator $R^{-1}\tilde{W}_a$ will have the previous form (1), and the difference $\tilde{T}_a - T_{a,R}$, where $T_{a,R}$ corresponds to $R^{-1}\tilde{W}_a$, will be a smooth, quickly decreasing function.

So, the construction of a unitary operator W_a which has form (1) and whose kernel T_a differs from the original kernel only by a smooth, quickly decreasing addend has been completed.

So far the concordance of the singularities of the kernel T_a with the structure of the operator H has not been used. Let us consider the operator $K = HW_a - W_a H_0$. If, in this definition, the real operator W was in place of W_a, the operator K would be zero. Let us assume that

(C) the operator $\tilde{V} = W_a^* K$ has a smooth, quickly decreasing kernel.

Let us introduce the operator $H_a = W_a^* H W_a$. As $H_a = H_0 + \tilde{V}$, from (C) it follows that the pair of operators H_0 and H_a can be investigated by means of any method which is suitable for the operator describing the scattering by the decreasing potential. Particularly, it can be asserted that the isometric wave operator with the finite index exists:

$$U = \operatorname*{s.lim}_{t \to -\infty} \exp(itH_a)\exp(-itH_0).$$

Hence the existence of the wave operator W follows. Also the equality $W = W_a U$ is true, which indicates that the operator W_a reflects rightly the singularities of many-particle scattering.

The application of the above scheme to a concrete operator H demands some calculations. They are necessary for the construction of the kernel T_a and for testing its properties expressed by the conditions (A)–(C). These calculations have been made for $n = 3$, $d = 1, 3$ and also in frames of the so-called Friedrichs model of many-particle scattering for any number n of particles [11, 12, 18, 19]. In the latter case an inductive character was imparted to the given scheme.

In the conclusion of the discussion of the unitary regularization method we should note that essentially it can be considered as a variant of the integral equations method. All its merits and demerits are of course connected with this circumstance. Particularly, one of the advantages is the possibility of obtaining very detailed information about the kernel T if the properties of kernels T for the subsystems are known sufficiently well. The given method naturally can be used not only in many-particle scattering problems but also in any case where the kernel of the wave operator W has form (1) and T possesses some characteristic singularities.

Of course, the method of unitary regularization is not the only scheme which makes it possible to use heuristic information about the singularities of many-particle scattering to justify the formal conceptions of the scattering theory and to investigate strictly the structure of those singularities themselves. Another approach is connected with studying the asymptotic behaviour of the kernel $R(x, y; z)$ of the resolvent $(H - zI)^{-1}$ in the coordinate representation. Having constructed the singular asymptotic terms, one can obtain for $R(x, y; z)$ an integral equation with a smooth, quickly decreasing kernel. This method seems to be somewhat complicated but it has been used successfully for two important models. One of them is the Schrödinger operator with the slowly decreasing potential [17]. The second model is the Schrödinger operator of the 3-particle system with the pair potentials whose asymptotic behaviour admits the following description: $v(\xi) = v_0|\xi|^{-1} + v_1(\xi)$, $\xi \to \infty$, where v_0 is a constant and v_1 is a quickly decreasing function [15, 16]. Unfortunately we cannot discuss these problems in more detail.

3. One-dimensional particles

In the systems of one-dimensional particles the singularities of many-particle scattering manifest themselves in a rather specific way. This is connected with the fact that, in general, in the case of one-dimensional particles the force of the singularities of the multiple scattering does not depend on the multiplicity. The system of the particles with equal masses

possesses an additional specificity. In such an n-particle system the singularities dependent on a sequence of double scatterings manifest themselves with the same force at multiplicity $r \leqslant \frac{1}{2}n(n-1)$ and disappear at greater r. Particularly, if $n = 3$, the singularities manifest themselves at $r \leqslant 3$. In this case the subspace Λ is two-dimensional. In such a system a ray parallel to the vector p leaves the system of the screens Λ_{ij} after a triplet of scatterings. Six directions of the rays leaving the system of the screens can be obtained from the vector $p = (p_1, p_2, p_3)$ by transpositions of its coordinates. The procedure of description of the singularities of many-particle scattering given in Section 2 leads to some linear combinations of plane waves in each angle between the screens. The coefficients of these combinations can be explicitly expressed in terms of the elements of the 2-particle scattering matrix. In general, the above-mentioned linear combinations are discontinuous along the indicated six rays. In the case of identical pair potentials the discontinuity remains only on the rays directed along the vectors (p_3, p_1, p_2) and (p_2, p_3, p_1). On these two rays the amplitude of the plane waves with wave vectors parallel to the same rays has a discontinuity which is equal to

$$d(p) = r(k_3)r(k_2)s(k_1) + r(k_1)r(k_2)s(k_3) - r(k_1)r(k_3)s(k_2).$$

In the above expression s is the transmission coefficient and r is the reflection coefficient; they both correspond to the one-particle scattering by the potential v. Besides, $k_1 = \frac{1}{2}(p_3 - p_2)$, $k_2 = \frac{1}{2}(p_1 - p_3)$, $k_3 = \frac{1}{2}(p_2 - p_1)$. Let us replace the discontinuous plane waves by the Fresnel waves, i.e., the smooth solutions of free equation $-\Delta\psi = p^2\psi$, well known in the theory of diffraction of the plane wave on the screen. Outside an angular vicinity of some ray a Fresnel wave asymptotically reduces to a plane wave with a discontinuous amplitude. After the above-mentioned replacement, the solution ψ_a of the free equation appears inside each angle. Asymptotically this solution is equal to the solution ψ, i.e., their difference is a diverging wave with a smooth scattering amplitude.

Following these constructions we can describe the singularities of the whole scattering amplitude f. Now it is convenient to introduce a new representation in which a function $\Lambda \to C$ is considered as a function $F \to C^6$, where F is a fundamental domain of symmetrical group S_3, acting naturally on Λ. As F we can choose one of the angles between the screens. Let λ be the angular coordinate of p in the domain F. In this representation the scattering matrix $S(E)$ turns into an integral operator on arc $C = \Lambda_1 \cap F$ whose kernel $S(\lambda, \lambda'; E)$ is a matrix of sixth order. Outside the screens Λ_{ij} the singularities of this kernel coincide with the singularities of the

following kernel:

$$S_a(\lambda, \lambda'; E) = \hat{S}_1\hat{S}_2\hat{S}_3 P_+(\lambda, \lambda') + \hat{S}_3\hat{S}_2\hat{S}_1 P_-(\lambda, \lambda'), \tag{4}$$

where $P_\pm(\lambda, \lambda') = \pm(2\pi i)^{-1}(\lambda - \lambda' \pm io)^{-1}$ and $\hat{S}_1, \hat{S}_2, \hat{S}_3$ are unitary matrices-functions on C. Part of their matrix elements are equal to the matrix elements of the one-particle S-matrix, others (most of them) are equal to zero. In general, the matrices $\hat{S}_1, \hat{S}_2, \hat{S}_3$ do no commute. Since $P_\pm$ are the kernels of two additional orthogonal projectors in $L_2(\boldsymbol{R})$, the suggested description of the singularities reflects very expressively the unitary property of the operator $S(E)$. Besides, formula (4) leads immediately to an explicit description of the singularities of the kernel of a function of the operator $S(E)$.

If $d = 0$, the Fresnel waves fall out of the function ψ_a and it becomes a linear combination of plane waves. Equations of type $d = 0$ can be met with in another range of problems in connection with the problem of factorization of the S-matrix, see e.g. [20]. One of the terms used for them in this sphere is the Yang–Baxter relation. If $d = 0$, the operators $\hat{S}_1\hat{S}_2\hat{S}_3$ and $\hat{S}_3\hat{S}_2\hat{S}_1$ become equal and the operator S_a turns into a factorizable operator in the following sense:

$$S_a(\lambda, \lambda'; E) = \hat{S}_1\hat{S}_2\hat{S}_3\,\delta(\lambda - \lambda').$$

It appears that the relation $d = 0$, if it is regarded as an equation for functions s and r, can be solved explicitly in the class of functions possessing the properties of the transition and reflection coefficients. With the help of the apparatus of the inverse scattering problem the whole class of potentials which obey the equation $d = 0$ can be described completely. It is known that this class contains the function $v(\xi) = 2\Omega\delta(\xi)$ and the so-called soliton potentials characterized by the property $r = 0$. The total class of the potentials of this type is a non-trivial composition of the δ-function and the solitons.

There exists one more mechanism leading to the falling out of the Fresnel waves from ψ_a and to the factorization of the singular part S_a of the S-matrix. The same effects take place also if the problem is restricted to the subspace of functions symmetrical or asymmetrical with respect to the action of the group S_3. These last facts remain if n is arbitrary. The considerations of this part are based on the results of [5, 8, 9].

4. Trace formulas

In the problem of the scattering of a particle by a decreasing potential a trace formula is known, see e.g. [3], which gives the function

$$\omega(E) = 2i \lim_{\varepsilon \downarrow 0} \operatorname{Im} \operatorname{tr}[(H - zI)^{-1} - (H_0 - zI)^{-1}], \qquad z = E + i\varepsilon,$$

in terms of the S-matrix: $\omega(E) = \frac{d}{dE} \operatorname{tr} \ln S(E)$. In the case of the Schrödinger operator H of an n-particle system the difference of the resolvents which figure in the definition of function ω has to be replaced by the so-called connected part of the resolvent of H. If $n = 3$, one has to consider the function

$$\Omega(E) = 2i \lim \operatorname{Im} \operatorname{tr} \Big\{(H - zI)^{-1} - (H_0 - zI)^{-1} - \sum_a [(H_a - zI)^{-1} - (H_0 - zI)^{-1}]\Big\},$$

where H_a is an operator of type H containing, however, only one of the three potentials whose sum forms the operator V. Indeed, we should note that the right side of the last formula has a meaning only if $d = 1$, but with some unimportant modification it can be generalized to the common case. We shall not discuss here the sense of the function Ω but we note only that it is an essential object in some relations of statistical physics.

The following natural direct generalization of the trace formula:

$$\Omega(E) = \frac{d}{dE} \xi_0, \quad \text{where } \xi_0 = \operatorname{tr} \Big[\ln S - \sum_a \ln S_a\Big],$$

cannot be valid. This follows already from the fact that the trace ξ_0 does not exist, see (4). In general, the trace ξ_0 cannot be defined as an invariant one. Its satisfactory regularization seems to have been carried out so far only for one-dimensional particles and for the Friederichs' model. This regularization is based on formula (4). To regularize the trace ξ_0 we should calculate the trace first as the matrix trace and then as the trace of a scalar kernel. However, the formula stated remains wrong after regularization. It remains wrong even for the special class of potentials indicated in Section 3 although for this class regularization is not necessary.

After the above-mentioned regularization of the trace ξ_0 the accurate formula has the form $\Omega(E) = \frac{d}{dE}(\xi_0 - \tilde{\xi})$, where $\tilde{\xi}$ can be expressed explicitly in terms of the one-particle S-matrices. We have no opportunity to write this expression here: although its general structure is quite clear, it cannot be written down briefly. As in the problem of a description of the singularities of many-particle scattering, the situation becomes

essentially simplified for the special potentials discussed in Section 3 and also after passing to the Bose or Fermi statistics. In these cases with a proper term ξ_0 the additional term $\tilde{\xi}$ becomes a quadratic functional on the logarithms of the matrix elements of the one-particle S-matrix.

A whole series of delusions were inherent in the initial investigations of the problem touched upon here. The first strict result based on the Faddeev equations was obtained in [6, 7] for the system of three-dimensional particles. This result gave an explicit description of the function Ω in terms of the total S-matrix and the S-matrices of the two-particle subsystems. The structure of the formula obtained turned out to be rather complicated. The above results for the system of one-dimensional particles and for the Friedrichs model were obtained in [4, 5].

References

[1] Birman M. Sh. and Jafaev D. R., *Zapiski nauchn. semin. LOMI* **110** (1981), pp. 3–29.
[2] Birman M. Sh. and Jafaev D. R., *Theor. Math. Phys.* **51** (1982), pp. 44–53.
[3] Buslaev V. S., *Probl. math. phys. Leningr. Un.* **1** (1966), pp. 82–101.
[4] Buslaev V. S., *Zapiski nauchn. semin. LOMI* **27** (1972), pp. 47–66.
[5] Buslaev V. S., *Theor. Math. Phys.* **16** (1973), pp. 247–259.
[6] Buslaev V. S. and Merkuriev S. P., *Dokl. Akad. Nauk USSR* **189** (1969), pp. 269–272.
[7] Buslaev V. S. and Merkuriev S. P., *Theor. Math. Phys.* **5** (1970), pp. 372–389.
[8] Buslaev V. S., Merkuriev S. P., and Salikov S. P., *Probl. math. phys. Leningr. Un.* **9** (1979), pp. 14–30.
[9] Buslaev V. S., Merkuriev S. P., and Salikov S. P., *Zapiski nauchn. semin. LOMI* **84** (1979), pp. 16–22.
[10] Buslaev V. S. and Skriganov M. M., *Theor. Math. Phys.* **19** (1974), pp. 217–232.
[11] Buslaev V. S. and Vakulenko A. F., *Vestnik Leningr. Un.* **13** (1977), pp. 22–30.
[12] Buslaev V. S. and Vakulenko A. F., *Trudy konf. po uravn. v chastn. proizv., Moscow – 1976,* Moscow Un. (1978), pp. 56–58.
[13] Merkuriev S. P., *Theor. Math. Phys.* **8** (1971), pp. 235–250.
[14] Merkuriev S. P., *Nucl. Phys.* **A233** (1974), pp. 395–408.
[15] Merkuriev S. P., *Theor. Math. Phys.* **32** (1977), pp. 187–207.
[16] Merkuriev S. P., *Ann. of Phys.* **130** (1980), pp. 395–426.
[17] Skriganov M. M., *Zapiski nauchn. semin. LOMI* **69** (1977), pp. 171–199.
[18] Vakulenko A. F., *Dokl. Akad. Nauk USSR* **245** (1979), pp. 1348–1352.
[19] Vakulenko A. F., *Dokl. Akad. Nauk USSR* **249** (1979), pp. 825–828.
[20] Zamolodchikov A. B., *Ann. of Phys.* **120** (1979), pp. 253–261.

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

LUIS A. CAFFARELLI

Variational Problems with Free Boundaries

A free boundary problem is the problem that arises when one attempts to describe a discontinuous change of behaviour in a physical, or biological quantity, an optimal strategy, etc.

The evolution of an ice–water mixture, the behaviour of an elasto-plastic material, an elastic membrane constrained to stay within a given region, are typical examples.

In some of the simplest examples one can find weak solutions to the problem by variational methods or methods from non-linear P.D.E.

For instance, if one considers, in a 3-space, an elastic membrane described by the graph of a function $z = u_1(x, y)$ in a domain $D \subset R^2$, attached at the boundary of $D(u_1|_{\partial D} = f)$ and constrained to stay above the (x, y)-plane $(u \geqslant 0)$, upon which a constant downward pressure is exerted, the equilibrium configuration should be a minimum of the energy functional

$$J_1(v) = \int\int [(\text{grad } v)^2 + 2v]\,dx \tag{i}$$

among all admisible functions $\{v; v|_{\partial D} = f,\ v \geqslant 0\}$ (see [1]).

Another variational free boundary problem of interest is the study of local minimizers u_2 of

$$J_2(v) = \int\int (\text{grad } v)^2 + \chi_{v>0}\,dx. \tag{ii}$$

A similar, but non-variational, free boundary problem is that of studying solutions u_3 of the equation

$$\Delta u_3 = D_{x_n}(\chi_{u>0}). \tag{iii}$$

Note that in all these problems u (or more precisely its derivatives) have a discontinuous behaviour when u becomes zero.

In fact, if one tries to deduce Euler equations for u, they are not satisfied across such a surface.

Heuristically, from the Hadamard variational formulas, what u should satisfy is the following:

for (i) $u_1 \geqslant 0$,

$\Delta u_1 = \chi_{u_1>0}$;

for (ii) $\Delta u_2 = 0$ on $\Omega(u_2) = \{x\colon u_2 \neq 0\}$,

$(u_2^+)_\nu^2 - (u_2^-)^2 = 1$;

for (iii) $\Delta u_3 = 0$ on $\Omega(u_3) = \{x\colon u_3 \neq 0\}$,

$(u_3^+)_\nu - (u_3^-)_\nu = \nu \cdot e_n$.

In problems (ii) and (iii), note that u may be negative, problem (ii) describes the flow of two perfect jets, problem (iii) that of two flows in a porous medium. If u is non-negative, we have a one-phase flow.

The study of one-phase problems and two-phase problems is essentially different.

In one-phase problems, one has an a priori optimal regularity, which looks like a Harnack inequality. That is, if u_1 is a minimizer of (i) in $B_1(0)$ (the unit ball in R^n) and $u(0) = 0$, then the $C^{1,1}$ norm of $u(\sup|u| + \sup|D_{ij}u|)$ in $B_{1/2}(0)$ is bounded by a universal constant (Frehse).

If u_2(resp. u_3) is a non-negative minimizer of (ii) or a weak solution of (iii) and $u(0) = 0$, then the Lipschitz norm of $u(\sup|u| + \sup|D_i u|)$ is bounded in $B_{1/2}(0)$ by a universal constant (Alt and Caffarelli; Alt).

Simple one-dimensional examples show that one cannot expect such a behavior if u changes sign (for instance in (ii)) since a solution can be built from very steep lines.

How does one prove regularity in such a problem?

The theory of minimal surfaces suggests that a monotonicity-type formula may be the answer (Alt, Caffarelli, Friedman).

Let $u^{(i)}$ ($i = 1$ and 2) be two $C^\alpha \cap H^1$ functions in $B_1(0)$ (of R^n). Assume that $u^{(1)}u^{(2)} \equiv 0$ (disjoint support) and that $\Delta u^{(i)} \equiv 0$ when $u^{(i)} \neq 0$.

Then, if $u^{(i)}(0) = u^{(2)}(0) = 0$, the function

$$f(R) = \frac{\int_{\Sigma_1}\int_0^R |\nabla u^{(1)}|^2 r\, dr\, d\sigma \int_{\Sigma_1}\int_0^R |\nabla u^{(2)}|^2 r\, dr\, d\sigma}{R^4}$$

is monotone increasing in R (r, σ: polar coordinates in R^n).

Since, heuristically,

$$f(0) = [u_\nu^{(1)}]^2 [u_\nu^{(2)}]^2,$$

this product must be finite.

If one applies this result to the solutions of (ii) of (iii) along the free boundary, one obtains (always heuristically)

for (ii) $\quad (u_\nu^+)^2 - (u_\nu^-)^2 = 1,\ (u_\nu^+)^2 (u_\nu^-)^2 \leqslant C$

and for (iii) $\quad |u_\nu^+ - u_\nu^-| \leqslant 1,\ (u_\nu^+)^2 (u_\nu^-)^2 \leqslant C,$

from where Lipschitz continuity follows after some careful analysis (Alt, Caffarelli, Friedman).

As regards free boundary regularity, again, one-phase problems are well studied (see [1]).

For two-phase problems (ii) and (iii) we have at present only a two-dimensional analysis (Alt, Cafferelli and Friedman, to appear).

Reference

[1] Friedman, Avner, *Variational Principles and Free-Boundary Problems*, Wiley, 1982.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF CHICAGO

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

G. ESKIN*

Initial-Boundary Value Problem for Hyperbolic Equations

1. Introduction

This report will be on initial-boundary value problems for strictly hyperbolic equations. The particular form of the hyperbolic equation will not be important and without loss of generality one can consider a second order hyperbolic equation

$$A(x, D)u = 0 \tag{1}$$

in a cylindrical domain $\Omega = (-\infty, +\infty) \times G \subset \boldsymbol{R}^{n+1}$, where $x_0 \in (-\infty, +\infty)$ is a time variable and $(x_1, \ldots, x_n) \in G$ are space variables.

Solutions of (1) are subject to zero initial conditions

$$u = 0 \quad \text{for } x_0 < 0,\ x \in \Omega, \tag{2}$$

and some boundary condition

$$B(x, D)u|_{\partial\Omega} = h(x'), \quad x' \in \partial\Omega, \tag{3}$$

where $h = 0$ for $x_0 < 0$.

The problem is to find necessary and sufficient conditions on $B(x, D)$ such that the initial-boundary value problem (1), (2), (3) is well-posed. Note that all theorems that are formulated below will also apply to the case when (1) is a general hyperbolic equation or a hyperbolic system of equations of arbitrary order provided that all components of the characteristic cone are strictly convex (cf. [3]). Besides the Dirichlet and the Neumann conditions there are many boundary conditions that are of interest in mathematical physics, for example, (a) the impedance boundary condition

$$\frac{\partial u}{\partial \nu} - a(x)\frac{\partial u}{\partial x_0}\bigg|_{\partial\Omega} = h, \tag{4}$$

* Partially supported by Grant MCS 81-01656 from the National Science Foundation of the United States.

where $\partial/\partial\nu$ is the conormal derivative, (b) the boundary conditions in the linearized water wave theory, (c) the boundary conditions in elastodynamics for a solid with a free boundary. The last two examples describe interesting phenomena in wave propagation: supersonic boundary waves in the linearized water wave theory and Rayleigh's waves (subsonic boundary waves) in elastodynamics.

2. Weak and strong Lopatinsky condition

In the theory of general boundary value problems for elliptic equations and initial-boundary value problems for parabolic equations, the following condition is necessary and sufficient for well-posedness:

Let $\tilde{x}$ be an arbitrary point of the boundary $\partial\Omega$. Freeze coefficients of $A(x, D)$ and $B(x, D)$ at the point $\tilde{x}$ and consider the constant coefficient boundary value problem for the principal parts of A and B in the half-space formed by the tangent plane to $\partial\Omega$ at the point $\tilde{x}$. This constant coefficient problem in the half-space can be solved explicitly using the Fourier transform. The algebraic condition of the well-posedness of the constant coefficient problem is called the *Shapiro-Lopatinsky condition.* If the Shapiro–Lopatinsky condition is satisfied for any (x', ξ') in the cotangent bundle $T_0^*(\partial\Omega)$, where $x' \in \partial\Omega$ and $\xi' \neq 0$, then the boundary problem in Ω is well-posed. One can try the same criterion for the initial-boundary problem for hyperbolic equations. Since we know that h and u are zero for $x_0 < 0$, the Fourier transform with respect to x_0 will indeed be the Laplace transform, so that the variable dual to x_0 will be $\xi_0 + i\tau$ where $\tau > 0$. It can be shown (see [14]) that for the well-posedness of the initial-boundary value problem for hyperbolic equations, it is necessary that the Shapiro–Lopatinsky condition will be satisfied for any $(x', \xi') \in T^*(\Omega)$ and any $\tau > 0$. This condition is called the *weak Lopatinsky condition.* It is necessary and sufficient for the well-posedness of IBVP (initial-boundary value problem) for hyperbolic equations with constant coefficients in the half-space. This result was established first by R. Hersch. As we shall see later, the weak Lopatinsky condition is not sufficient, in general, for the well-posedness of IBVP for hyperbolic equations. The reason is that there are boundary conditions that are sensitive to the local geometry of boundary (to convexity or concavity for instance). Therefore the tangent plane model is not good enough in these cases. Note that among such sensitive boundary conditions are the boundary conditions that produce boundary waves.

A general sufficient condition for the well-posedness of hyperbolic IBVP was found by Kreiss [15] and Sakamoto [21] in 1970. They proved that if the Shapiro–Lopatinsky condition is satisfied not only for all $(x', \xi') \in T_0^*(\Omega)$ and $\tau > 0$ but also when $\tau = 0$ then the hyperbolic IBVP is well-posed. Such a condition appeared first in the work of S. Agmon [1] and it is now called the *Agmon–Kreiss–Sakamoto condition* or the *strong* (or *uniform*) *Lopatinsky condition*. The strong Lopatinsky condition is satisfied in many problems of interest and it is independent of the shape of the boundary but there are important boundary conditions that do not belong to this class such as the Neumann boundary condition or any of the boundary conditions where boundary waves are present. An important class of boundary conditions for second order hyperbolic equations that implies an estimate of the solution in the energy norm was studied by S. Miyatake [19], L. Garding [9], and extended by R. Melrose and J. Sjostrand [18].

3. Microlocalization

For simplicity consider a boundary condition of the form

$$\frac{\partial u}{\partial \nu} + \lambda(x', D')u|_{\partial\Omega} = h(x'), \tag{5}$$

where $\partial/\partial\nu$ is the conormal derivative and $\lambda(x', D')$ is a first order differential or pseudodifferential operator in tangential variables. Although the operator λ is not pseudodifferential in physical applications for second order hyperbolic equations, pseudodifferential λ arises when one considers a hyperbolic equation of higher order or a system of hyperbolic equations. Then after a microlocalization the problem is reduced to a pseudodifferential equation $A(x, D)$ of the second order that is differential in the normal variable and the boundary operator has the form (5) with a pseudodifferential $\lambda(x', D')$. There is a natural partition of $T_0^*(\partial\Omega)$ into three regions: (1) the elliptic region where the principal symbol of $A(x, D)$ has no real zeros with respect to a variable dual to the normal, (2) the hyperbolic region where there are two distinct real zeros, and (3) the diffraction region where there is one double real zero.

In the elliptic and hyperbolic region the investigation of the IBVP can be reduced to a study of a pseudodifferential equation on the boundary. This reduction was done first by P. D. Lax and L. Nirenberg (see [20]).

4. Case of a strictly concave boundary

The most difficult part of the problem is the study of the neighborhood of the diffraction region. We shall consider this problem under the additional restriction that the boundary $\partial\Omega$ is either strictly convex or strictly concave with respect to the bicharacteristics of the hyperbolic operator $A(x, D)$ that are tangent to $\partial\Omega$. Indeed one needs the concavity or convexity conditions only on the intersection of the set where the strong Lopatinsky condition does not hold with the diffraction region because in the region of $T_0^*(\partial\Omega)$ where the strong Lopatinsky condition holds one can use the Kreiss method.

Important examples of the IBVP with a convex or a concave boundary are the initial-boundary value problem for the wave equation in the interior or in the exterior of a strictly convex bounded domain in $\boldsymbol{R}^n$.

The following theorem describes conditions of the well-posedness of IBVP in the case of concave boundary.

THEOREM 1 (see [6]). *Assume, for simplicity, that the boundary condition has form* (5) *and it is of principal type with respect to* x_0 *in the elliptic and hyperbolic region. Then the weak Lopatinsky condition is necessary and sufficient for the well-posedness of IBVP with a concave boundary.*

5. Hypoelliptic boundary conditions

There is an important class of boundary conditions (5) which are called *hypoelliptic* (see [18]).

The hypoellipticity of the boundary condition means that for any distribution solution u of (1), (2), (3) the following inclusion holds:

$$WF(u|_{\partial\Omega}) \subset WF(h), \tag{6}$$

where $u|_{\partial\Omega}$ is the restriction of u on $\partial\Omega$ and $WF(u|_{\partial\Omega}) \subset T_0^*(\partial\Omega)$ and $WF(h) \subset T_0^*(\partial\Omega)$ are the wave front sets.

For example, the Neumann problem is hypoelliptic in the case of a concave boundary (see [16] and [22]). R. Melrose and J. Sjostrand (see [18], Part II) found a sufficient condition for the hypoellipticity of the IBVP and they formulated the following general conjecture that was proven in [8]:

THEOREM 2. *The boundary condition* (5) *in the case of concave boundary is hypoelliptic if the strong Lopatinsky condition holds in the elliptic and hyperbolic regions and for any point* $(\tilde{x}, \tilde{\xi})$ *in the diffraction region where*

the strong Lopatinsky condition does not hold there is a neighborhood U_0 *in the diffraction region such that*

$$-\frac{\pi}{2}+\varepsilon_0 \leqslant \arg \lambda_1 \leqslant \pi-\varepsilon_0, \qquad \varepsilon_0>0, \tag{7}$$

for any $(x', \xi') \in U_0$, *where* λ_1 *is the restriction of the principal part of* $\lambda(x', \xi)$ *to the diffraction region.*

6. Case of a convex boundary

The case of a convex boundary is more complicated than the case of a concave boundary because of multiple reflections of waves. The following general result was proven in [8]:

THEOREM 3. *Assume that the boundary is convex and the weak Lopatinsky condition is satisfied. Let, for simplicity, the boundary operator* (5) *be of principal type with respect to* x_0 *in the elliptic and hyperbolic regions. Assume that for any point* $(\tilde{x}, \tilde{\xi})$ *in the diffraction region where the strong Lopatinsky condition does not hold there is a neighborhood* U_0 *in the diffraction region such that*

$$-\operatorname{Re}\lambda_1 \leqslant C(\operatorname{Im}\lambda_1)^2|\xi'|^{-1}\ln\frac{1}{|\operatorname{Re}\lambda_1|\,\xi'|^{-1}|}, \tag{8}$$

for all $(x', \xi') \in U_0$, *where* λ_1 *is the same as in Theorem 2. Then the IBVP* (1), (2), (5) *is well-posed.*

It was shown in [7] that the conditions of Theorem 3 are necessary and sufficient for the well-posedness of a model problem with a convex boundary. For a general hyperbolic equation the situation is more complicated. Consider, for simplicity, the case when λ_1 is a real-valued symbol. Then the condition (8) simply means that $\lambda_1 \geqslant 0$ in U_0.

THEOREM 4 (see [8]). *Let* λ_1 *be real and assume that the condition* (8) *is not satisfied at the point* $(\tilde{x}, \tilde{\xi})$ *in the diffraction region. Let the Poisson bracket*

$$\{\lambda_1, \mu\} = 0 \quad \text{at the point } (\tilde{x}, \tilde{\xi}), \tag{9}$$

where $\mu = 0$ *is the equation of the diffraction region. If the boundary is strictly convex at the point* $(\tilde{x}, \tilde{\xi})$ *then the IBVP is ill-posed.*

Therefore the conditions for well-posedness of IBVP with a convex boundary are very restrictive when the strong Lopatinsky condition fails. Nevertheless it was shown in [7] for a model problem with a convex boundary that if the condition (8) is not satisfied but the Poisson bracket (9) is not equal to zero then the IBVP is still well-posed.

7. Examples

Let (1) be the wave equation and the boundary condition has the form (4) (the impedance boundary condition). Then the weak Lopatinsky condition has the form

$$a(x') > -1 \quad \text{on } \partial\Omega. \tag{10}$$

It follows from Theorem 1 that for the wave equation in the exterior of a strictly convex domain with the boundary condition (4) the condition (10) is necessary and sufficient for the well-posedness of IBVP (see also M. Ikawa [10]). Consider the same problem in the interior of a convex domain. If $a(\tilde{x}) = 0$ and $\dfrac{\partial}{\partial x_0} a(\tilde{x}) = 0$ for some point $\tilde{x} \in \partial\Omega$ and if $a(x')$ is not a nonnegative function in a neighborhood of $\tilde{x}$ (i.e. the condition (8) is not satisfied) then Theorem 4 implies that IBVP is ill-posed.

As a second example, consider the case of the oblique derivative boundary condition

$$\frac{\partial u}{\partial \nu} + \tau u|_{\partial\Omega} = h, \tag{11}$$

where $\partial/\partial\nu$ is the interior conormal derivative and τ is a tangential vector field on $\partial\Omega$.

The exterior problem with the boundary condition (11) is always well-posed. This fact was proved by M. Ikawa (see [10]). But the interior problem is ill-posed when τ is not identically zero and has a degenerate critical point. Analogous results hold for the transmission problem

$$\begin{aligned} A_1(x, D)u_1 &= 0 \quad \text{in } \Omega, \\ A_2(x, D)u_2 &= 0 \quad \text{in } C\Omega, \end{aligned} \tag{12}$$

with the transmission conditions

$$u_1|_{\partial\Omega} = u_2|_{\partial\Omega}, \tag{13}$$

and

$$\frac{\partial u_1}{\partial \nu_1} + \tau_1 u_1|_{\partial\Omega} = -\frac{\partial u_2}{\partial \nu_2} + \tau_2 u_2|_{\partial\Omega}, \tag{14}$$

where $C\Omega$ is the complement to Ω in R^{n+1}, $\partial/\partial\nu_1$ and $\partial/\partial\nu_2$ are the interior conormal derivatives with respect to A_1 and A_2, τ_1 and τ_2 are tangential vector fields to $\partial\Omega$ and we assume that $\partial\Omega$ is strictly convex with respect to the tangential bicharacteristics of A_1 and is strictly concave with respect to the tangential bicharacteristics of A_2. Then the transmission problem (12), (13), (14) with zero initial conditions is ill-posed if $\tau_1-\tau_2$ is not equal to zero identically and there exists a degenerate critical point of $\tau_1-\tau_2$.

8. Propagation of singularities

There is a close relation between the well-posedness of IBVP and the propagation of singularities. R. Melrose [16] and M. Taylor [22], first completely described the singularities of IBVP with concave boundaries for the cases of the strong Lopatinsky boundary conditions and the Neumann boundary condition.

The case of IBVP with convex boundaries for the Dirichlet and the Neumann boundary conditions was done independently by K. G. Andersson and R. Melrose [2], G. Eskin [4] and V. Ia. Ivrii [12]. It was shown in these works that the singularities propagate along broken bicharacteristics that are undergoing multiple reflections on the boundary and along the gliding rays. Further important progress was made by R. Melrose and J. Sjostrand in [18] where the propagation of singularities for general domains were studied without restriction on convexity or concavity of the boundary.

Under the restriction that the boundary is concave, the propagation of singularities for an arbitrary boundary operator with a real-valued symbol satisfying the weak Lopatinsky condition was studied in [6]. For such a general boundary condition the boundary waves may appear but they do not represent a threat to the well-posedness of IBVP.

In the case of convex boundaries there is no hypoelliptic boundary condition in the sense of the definition (6) because there is always a propagation of singularities along the gliding rays. When the conditions of Theorem 3 are satisfied and the strong Lopatinsky condition holds in the elliptic and the hyperbolic regions, then the only singularities of $u|_{\partial\Omega}$ are contained in the union of all gliding rays and broken bicharacteristics that start at the points of $WF(h)$.

If the condition (8) is not satisfied then the picture of the propagation of singularities is more complicated. The singularities of $u|_{\partial\Omega}$ come from bonudary waves propagating along the boundary, and waves propaga-

ting in Ω and undergoing multiple reflections. In general (when the condition (9) holds) that leads to singularities so strong that the solution of IBVP ceases to be a distribution. And this is a reason for the ill-posedness of IBVP under the conditions of Theorem 4. Indeed the solution becomes an ultradistribution, i.e. a functional over the space of C^∞ functions that belong to a certain Gevrey class.

References

[1] Agmon S., *Problèmes mixtes pour les équations hyperbolique d'ordre superieurs, Les Equations aux derivées partielles*, Paris 1962, pp. 13–18.

[2] Andersson K. G. and Melrose R., The Propagation of Singularities Along Gliding Rays, *Invent. Math.* **41** (1977), pp. 197–232.

[3] Eskin G., A Parametrix for Mixed Problems for Strictly Hyperbolic Equations of an Arbitrary Order, *Comm. P.D.E.* **1** (1976), pp. 521–560.

[4] Eskin G., Parametrix and Propagation of Singularities for the Interior Mixed Hyperbolic Problem, *Journ. d'Analyse Math.* **32** (1977), pp. 17–62.

[5] Eskin G., *Well-Posedness and Propagation of Singularities for Initial-Boundary Value Problem for Second Order Hyperbolic Equation with General Boundary Condition*, Sem. Goulaouic–Schwartz, 1979–1980, expose No. 2.

[6] Eskin G., Initial-Boundary Value Problem for Second Order Hyperbolic Equation with General Boundary Condition, I, *Journ. d'Analyse Math.* **40** (1981), pp. 43–89.

[7] Eskin G., General Initial-Boundary Problems for Second Order Hyperbolic Equations, in: *Singularities in Boundary Value Problems*, H. G. Garnir (ed.), Nato Advanced Study Institute Series, D. Reidel Publ. Comp., 1980, pp. 17–54.

[8] Eskin G., *Initial-Boundary Value Problems for Hyperbolic Equations, II*, to appear in *Comm. P.D.E.*

[9] Garding L., Le problème de la derivée oblique pour l'équation des ondes, *C. R. Acad. Sci. Paris.* **285** (1977), pp. 773–775; *C. R. Acad. Sci. Paris*, **286** (1978), p. 1199.

[10] Ikawa M., Mixed Problems for the Wave Equation, IV, The Existence and Exponential Decay of Solutions, *Journ. of Math. Kyoto Univ.*, **19** (3) (1979), pp. 375–411.

[11] Ikawa M., On the Mixed Problem for the Wave Equation in an Interior Domain, *Comm. P. D. E.* **3** (3) (1979), pp. 249–295 (see also preprint (1979)).

[12] Ivrii V. Ia., Propagation of Singularities of Solutions of the Wave Equations Along the Boundary of Domain, *Russ. Math. Survey* **32** (5) (1977), pp. 185–186.

[13] Ivrii V. Ia., The Nonclassical Propagation of Singularities of a Solution of a Wave Equation Near a Boundary, *Soviet Math. Dokl.* **19** (4) (1978), pp. 947–949.

[14] Kajitani K., A Necessary Condition for the Well-Posed Hyperbolic Mixed Problem With Variable Coefficients, *J. Math. Kyoto Univ.* **14** (1974).

[15] Kreiss H. O., Initial-Boundary Value Problems for Hyperbolic Systems, *Comm. Pure Appl. Math.* **23** (1970), pp. 277–298.

[16] Melrose R., Microlocal Parametrices for Diffractive Boundary Value Problems, *Duke Math. J.* **42** (1975), pp. 605–635.

[17] Melrose R., Transformation of Boundary Problems, *Acta Math.* **147** (1981), pp. 149–236.
[18] Melrose R. and Sjostrand J., Singularities of Boundary Value Problems, I, *Comm. Pure Appl. Math.* **31** (1978), pp. 593–617; II, *Comm. Pure Appl. Math.* **35** (1982), pp. 129–168.
[19] Miyatake S., A Sharp Form of the Existence Theorem for Hyperbolic Mixed Problems of Second Order, *J. Math. Kyoto Univ.* **17** (2) (1977), pp. 199–223.
[20] Nirenberg L., *Lectures on Linear Partial Differential Equations*, Regional Conference Series in Math. **20**, AMS, Providence, RI, 1973.
[21] Sakamoto R., Mixed Problems for Hyperbolic Equations, I, *J. Math. Kyoto Univ.* **10** (1970), pp. 349–373; II, **10** (1970), pp. 403–417.
[22] Taylor M., Grazing Rays and Reflection of Singularities of Solutions to Wave Equations, *Comm. Pure Appl. Math.* **29** (1976), pp. 1–37.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF CALIFORNIA AT LOS ANGELES
LOS ANGELES, CA 90024, USA

Proceedings of the International Congress of Mathematicians
August 16-24, 1983, Warszawa

ENNIO DE GIORGI

G-Operators and Γ-Convergence

Avant de commencer mon exposé je veux remercier le comité organisateur du I. C. M. pour l'invitation que j'ai bien appréciée pour plusieurs raisons. En premier lieu je suis heureux de pouvoir manifester mon admiration pour la contribution fondamentale des mathématiciens polonais à notre science et aussi pour tout ce que la Pologne a donné à l'humanité. Je pense aussi que l'I. C. M. est une occasion importante pour la réflexion commune des mathématiciens sur la valeur et la signification de notre science qui est un facteur fondamental de toute civilisation humaine, une oeuvre millenaire de l'humanité, un signe remarquable de la dignité de l'homme, de sa soif de conaissance que je crois signe d'un désir secret de voir quelques rayons de la gloire de Dieu.

Je crois que la conscience de travailler à cette oeuvre commune, la conscience de la valeur et de la dignité de la mathématique doit conduire à une amitié profonde et sincère entre tous les mathématiciens du monde. Cette amitié est la base de notre échange d'idées et des informations, nécessaires au progrès de la science, elle peut contribuer à la compréhension et à l'amitié entre tous les hommes et tous les peuples, fondement de la paix, nécessaire pour que les découvertes scientifiques donnent à l'humanité progrès et non destruction.

Dans cet esprit même l'exposition d'un sujet assez particulier comme les opérateurs de type G et la Γ-convergence veut être un signe d'amitié.

In the last years there has been a wide study of many limit cases in problems in Differential Equations, Calculus of Variations, Control Theory, Probability Theory and so on, often motivated by physical situations (see [55]) like the mathematical theory of homogenization. Therefore it was natural to look for some general ideas in order to place the various results already known in an abstract setting, and also to find new methods for the questions still open.

A possible answer in this direction is the theory of G-operators and Γ-convergence; this theory represents a substantial generalization of earlier work on quadratic functionals (see [29]) related with limit problems for second-order elliptic partial differential equations.

An example

A simple and well-known example in which Γ-convergence theory applies is the following (see [59]). Consider a sequence of one-dimensional boundary value problems for the second-order differential equations

$$-\frac{d}{dt}\left(\varphi_h(t)\frac{du}{dt}\right)=f_h(t); \quad u(a)=u(b)=0 \quad (h=1,2,\ldots), \tag{1}$$

where φ_h are measurable real functions on the interval $[a,b]$ of $\boldsymbol{R}$ such that

$$0<\Lambda_1\leqslant\varphi_h(t)\leqslant\Lambda_2<+\infty, \quad \forall\, t\in[a,b], \quad \forall\, h\in\boldsymbol{N}$$

and f_h belongs to $L^2(a,b)$ for every $h\in\boldsymbol{N}$. If the sequence $(1/\varphi_h)$ converges weakly in $L^1(a,b)$ to a function $1/\varphi_\infty$ and the sequence (f_h) converges in $L^2(a,b)$ to a function f_∞, then the sequence (u_h) of solutions of (1) converges uniformly on $[a,b]$ to the solution u_∞ of

$$-\frac{d}{dt}\left(\varphi_\infty\frac{du}{dt}\right)=f_\infty, \quad u(a)=u(b)=0.$$

If we regard (1) as the Euler–Lagrange equation of the integral functional

$$F_h(u)=\int_a^b\left[\varphi_h\left(\frac{du}{dt}\right)^2-2f_hu\right]dt,$$

we have

$$\lim_{h\to+\infty}\min\{F_h(u):\ u(a)=u(b)=0\}=\min\{F_\infty(u):\ u(a)=u(b)=0\}, \tag{2}$$

where F_∞ is obtained from F_h by substituting φ_∞ and f_∞ in place of φ_h and f_h.

It may happen that the sequence (φ_h) converges weakly in $L^1(a,b)$ to a function ψ_∞ different from φ_∞, in which case the sequence of functionals (F_h) converges pointwise in $L^2(a,b)$ to a functional G_∞ different from F_∞. On the other hand, it is obvious that in general pointwise convergence does not guarantee the convergence of the minimum values as in (2), even for real smooth functions (consider $F_h(x)=hx\exp(-h^2x^2)$). A type of convergence in which (F_h) converges to F_∞ and which ensures, under suitable hypotheses, the convergence of the minimum values is just Γ-convergence (see Theorem 1 below).

Before giving the definition of Γ-convergence, we conclude the illustration of the previous example by remarking that, with the help of general results on Γ-convergence theory (Theorems 2 and 3 below), (2) may be generalized as follows:

$$\lim_{h\to+\infty} \min\left\{F_h(u)+\int_a^b \psi\big(t, u(t)\big)\,dt:\ u(a)=u(b)=0\right\}$$

$$=\min\left\{F_\infty(u)+\int_a^b \psi\big(t, u(t)\big)\,dt:\ u(a)=u(b)=0\right\} \tag{2'}$$

whenever $\psi\colon [a,b]\times\boldsymbol{R}\to\boldsymbol{R}$ is a continuous bounded function. Observe that if ψ is not smooth, we cannot write the Euler–Lagrange equations to obtain (2'); thus the affinity between Γ-convergence methods and the classical direct methods in Calculus of Variations is already clear from this first elementary example.

The definition of Γ-convergence

We now give the definition of Γ-convergence (see [26]). Let X be a topological space, and let (F_h) be a sequence of functions defined on a subset E of X with values in $\overline{\boldsymbol{R}}=\boldsymbol{R}\cup\{+\infty,-\infty\}$. For every point x_0 in the closure $\bar{E}$ of E, define

$$\big(\Gamma(X^-)\liminf_{h\to+\infty} F_h\big)(x_0)=\sup_{U\in J(x_0)}\liminf_{h\to+\infty}\inf_{x\in U} F_h(x), \tag{3'}$$

$$\big(\Gamma(X^-)\limsup_{h\to+\infty} F_h\big)(x_0)=\sup_{U\in J(x_0)}\limsup_{h\to+\infty}\inf_{x\in U} F_h(x), \tag{3''}$$

where $J(x_0)$ denotes the family of all neighborhoods of x_0 in X. Moreover, if

$$\Gamma(X^-)\liminf_{h\to+\infty} F_h(x_0)=\Gamma(X^-)\limsup_{h\to+\infty} F_h(x_0)=\lambda,$$

then we say that the sequence (F_h) $\Gamma(X^-)$ *converges* at x_0 to λ and we write

$$\lambda=\Gamma(X^-)\lim_{h\to+\infty} F_h(x_0). \tag{3'''}$$

The following properties of the Γ-limits (3) are straightforward:

(i) the Γ-limits (3') and (3'') are lower semicontinuous functions of x_0 in $\bar{E}$;

(ii) if the sequence (F_h) is constant, that is $F_h = F_1$ for every $h \in \boldsymbol{N}$, then (F_h) $\Gamma(X^-)$ converges to the lower semicontinuous envelope of F_1 on E;

(iii) if the functions F_h have real values and are equicontinuous at $x_0 \in E$, then the Γ-limits (3′) and (3″) reduce to the ordinary $\liminf_{h\to+\infty} F_h(x_0)$ and $\limsup_{h\to+\infty} F_h(x_0)$.

Further generalizations

Note that the Γ-limits (3′) and (3″) may be rewritten as

$$\Gamma(X^-)\liminf_{h\to+\infty} F_h(x_0) = \sup_{U\in J(x_0)} \sup_{k\in\boldsymbol{N}} \inf_{h\geqslant k} \inf_{x\in U} F_h(x),$$

$$\Gamma(X^-)\limsup_{h\to+\infty} F_h(x_0) = \sup_{U\in J(x_0)} \inf_{k\in\boldsymbol{N}} \sup_{h\geqslant k} \inf_{x\in U} F_h(x),$$

so it is natural to give the following definition. If X_1, X_2 are topological spaces, $E_1 \subseteq X_1$, $E_2 \subseteq X_2$ and $f\colon E_1\times E_2\to\overline{\boldsymbol{R}}$ is a function, we define the functions $\Gamma(X_1^+, X_2^+)f$, $\Gamma(X_1^+, X_2^-)f$, $\Gamma(X_1^-, X_2^+)f$, $\Gamma(X_1^-, X_2^-)f$ on $\bar{E}_1\times\bar{E}_2$ by the formulae

$$\begin{aligned}
\big(\Gamma(X_1^+, X_2^+)f\big)(\xi_1,\xi_2) &= \inf_{U_2\in J(\xi_2)} \inf_{U_1\in J(\xi_1)} \sup_{x_1\in U_1} \sup_{x_2\in U_2} f(x_1,x_2),\\
\big(\Gamma(X_1^+, X_2^-)f\big)(\xi_1,\xi_2) &= \sup_{U_2\in J(\xi_2)} \inf_{U_1\in J(\xi_1)} \sup_{x_1\in U_1} \inf_{x_2\in U_2} f(x_1,x_2),\\
\big(\Gamma(X_1^-, X_2^+)f\big)(\xi_1,\xi_2) &= \inf_{U_2\in J(\xi_2)} \sup_{U_1\in J(\xi_1)} \inf_{x_1\in U_1} \sup_{x_2\in U_2} f(x_1,x_2),\\
\big(\Gamma(X_1^-, X_2^-)f\big)(\xi_1,\xi_2) &= \sup_{U_2\in J(\xi_2)} \sup_{U_1\in J(\xi_1)} \inf_{x_1\in U_1} \inf_{x_2\in U_2} f(x_1,x_2).
\end{aligned} \tag{4}$$

Then, if $X_1 = \overline{\boldsymbol{N}}$, $E_1 = \boldsymbol{N}$, $\xi_1 = +\infty$, $f(h, x_2) = F_h(x_2)$, we find that

$$\big(\Gamma(\overline{\boldsymbol{N}}^+, X_2^-)f\big)(+\infty,\xi_2) = \Gamma(X_2^-)\limsup_{h\to+\infty} F_h(\xi_2),$$

$$\big(\Gamma(\overline{\boldsymbol{N}}^-, X_2^-)f\big)(+\infty,\xi_2) = \Gamma(X_2^-)\liminf_{h\to+\infty} F_h(\xi_2).$$

One may also define the $\Gamma(X^+)$-limits in a symmetric way as

$$\Gamma(X_2^+)\liminf_{h\to+\infty} F_h(\xi_2) = \big(\Gamma(\overline{\boldsymbol{N}}^-, X_2^+)f\big)(+\infty,\xi_2),$$

$$\Gamma(X_2^+)\limsup_{h\to+\infty} F_h(\xi_2) = \big(\Gamma(\overline{\boldsymbol{N}}^+, X_2^+)f\big)(+\infty,\xi_2).$$

It is not difficult to extend (4) to Γ-operators for functions depending on three or more variables and so to obtain the definition of Γ-limits of sequences of functions of two or more variables. For example, in some minimax problems the following Γ-limit is considered (see [18]):

$$\Gamma(X_1^+, X_2^-) \liminf_{h\to+\infty} F_h(\xi_1, \xi_2) = \big(\Gamma(\overline{N}^-, X_1^+, X_2^-)f\big)(+\infty, \xi_1, \xi_2)$$
$$= \sup_{U_2\in J(\xi_2)} \inf_{U_1\in J(\xi_1)} \liminf_{h\to+\infty} \sup_{x_1\in U_1} \inf_{x_2\in U_2} F_h(x_1, x_2).$$

I shall not compare here the definition of Γ-limits with that of other limits used in topology, differentiation and tangency theory, convex-concave functions and minimax theory, variational inequalities etc. (for this, see for example [4], [12], [19], [31], [32], [36], [37], [40], [47], [48], [51], [53], [54], [61]). However, I want to point out that the main properties of Γ-limits depend essentially on the fact that they are elements of a more general class of operators, the G-operators, which we now define.

Let A be an arbitrary set, and let $\mathscr{L}$ be an arbitrary family of complete lattices. For every $l \in \mathscr{L}$, we denote by $\mathrm{adm}(A, l)$ the set of all functions f such that $\mathrm{dom}\, f \subseteq A$, $\mathrm{range}\, f \subseteq l$. We do not exclude the empty function, that is the function f such that $\mathrm{dom}\, f = \mathrm{range}\, f = \emptyset$. Now, we say that an operator g is a $G(A, \mathscr{L})$-*operator* if the following four properties hold:

(i) $\mathrm{dom}\, g = \{(f, l)\colon l \in \mathscr{L},\ f \in \mathrm{adm}\,(A, l)\}$,

$$g(f, l) \in \mathrm{adm}\,(A, l) \quad \forall\, (f, l) \in \mathrm{dom}\, g;$$

(ii) $\forall\, l_1, l_2 \in \mathscr{L}$, $\forall\, f_1 \in \mathrm{adm}\,(A, l_1)$, $\forall\, f_2 \in \mathrm{adm}\,(A, l_2)$

$$\mathrm{dom}\, f_1 \subseteq \mathrm{dom}\, f_2 \Rightarrow \mathrm{dom}\, g(f_1, l_1) \subseteq \mathrm{dom}\, g(f_2, l_2),$$
$$\mathrm{dom}\, f_1 = \emptyset \Rightarrow \mathrm{dom}\, g(f_1, l_1) = \emptyset;$$

(iii) $\forall\, l \in \mathscr{L}$, $\forall f_1, f_2 \in \mathrm{adm}\,(A, l)$

$$\mathrm{dom}\, f_1 = \mathrm{dom}\, f_2, \quad f_1 \leqslant_l f_2 \Rightarrow g(f_1, l) \leqslant_l g(f_2, l);$$

(iv) if $l_1, l_2 \in \mathscr{L}$ and $\varphi\colon l_1 \to l_2$ is a complete lattice morphism (i.e., for every $E_1 \subseteq l_1$, $E_1 \neq \emptyset$ one has $\varphi(l_1 - \sup E_1) = l_2 - \sup \varphi(E_1)$ and $\varphi(l_1 - \inf E_1) = l_2 - \inf \varphi(E_1)$) then

$$\varphi \circ g(f, l_1) = g(\varphi \circ f, l_2) \quad \forall f \in \mathrm{adm}\,(A, l_1).$$

From (iv) we immediately obtain the following condition:

(iv′) if $f(x) = c$ for every $x \in \mathrm{dom}\, f$, then $g(f, l)(y) = c$ for every $y \in \mathrm{dom}\, g(f, l)$.

We shall call the operators satisfying (i), (ii), (iii) and the weaker condition (iv′) $G'(A, \mathscr{L})$-*operators*.

Naturally, the case of the $\Gamma(X_1^{\pm}, X_2^{\pm})$ operators corresponds to $A = X_1 \times X_2$ and $\mathscr{L} = \{\bar{\boldsymbol{R}}\}$, and they are actually $G(A, \mathscr{L})$-operators.

Moreover, with each pair (g_1, g_2) of $G(A, \mathscr{L})$-operators we may associate the (g_1, g_2)-convergence by considering the pairs (f, l) such that $g_1(f, l) = g_2(f, l)$. An example of convergence in this sense is the $\Gamma(X^-)$ convergence defined in (3‴).

Finally, we remark that among G-operators there are also other interesting operators different from the Γ-operators: for example the monotone rearrangements of Hardy and Littlewood are G-operators. An example of a G'-operator is the operator T_λ considered in Theorem 3.

Γ-convergence and Calculus of Variations

Let us return to the $\Gamma(X^-)$ limits. We begin by stating the general results that provide the link between Γ-convergence theory and Calculus of Variations (see [26]).

THEOREM 1. *Let X be a topological space and (F_h) a sequence of functions on a dense subset E of X with values in $\bar{\boldsymbol{R}}$. Suppose that for every $x \in X$*

$$F_\infty(x) = \Gamma(X^-) \lim_{h \to +\infty} F_h(x)$$

exists. If (x_h) is a sequence of points of E such that

$$\lim_{h \to +\infty} x_h = x_\infty \quad \text{and} \quad \lim_{h \to +\infty} F_h(x_h) = \lim_{h \to +\infty} \left(\inf_{x \in E} F_h(x)\right),$$

then

$$F_\infty(x_\infty) = \min_{x \in X} F_\infty(x) = \lim_{h \to +\infty} \inf_{x \in E} F_h(x).$$

If $F_h(x_h) = \min_{x \in E} F_h(x)$, then Theorem 1 gives the convergence of minima and minimal points of F_h to minimum and minimal points of F_∞.

THEOREM 2. *Let X, (F_h) be as in Theorem 1. Suppose that for every $x \in X$*

$$F_\infty(x) = \Gamma(X^-) \lim_{h \to +\infty} F_h(x)$$

exists. If (G_h) is a sequence of functions on X with values in $\boldsymbol{R}$ converging uniformly on X to a continuous function G_∞, then for every $x \in X$

$$F_\infty(x) + G_\infty(x) = \Gamma(X^-) \lim_{h \to +\infty} \left(F_h(x) + G_h(x)\right).$$

Theorem 2 gives the stability of $\Gamma(X^-)$ convergence with respect to continuous perturbations.

It frequently happens in applications that E may be embedded in many topological spaces X; in this case Theorems 1 and 2 play an opposite role, because the topology of X must be chosen weak enough to ensure the compactness of the minimizing sequences and strong enough to ensure a larger class of continuous perturbations.

THEOREM 3. *Let (X,d) be a metric space and (F_h) be a sequence of functions on a dense subset E of X with values in $\overline{\boldsymbol{R}}$. Let $\alpha > 0$ be fixed and define for every $\lambda \geqslant 0$, $h \in \boldsymbol{N}$ and $x \in X$*

$$(T_\lambda F_h)(x) = \inf_{y \in E} \left[F_h(y) + \lambda \left(d(x, y)\right)^\alpha\right].$$

Suppose that for every $x \in X$

$$\lim_{\lambda \to +\infty} \lim_{h \to +\infty} (T_\lambda F_h)(x) = F_\infty(x)$$

exists and that $F_\infty(x) > -\infty$. Then

$$F_\infty(x) = \Gamma(X^-) \lim_{h \to +\infty} F_h(x), \qquad \forall\, x \in X.$$

Theorem 3 reduces the calculation of a Γ-limit in a metric space to a limit of the minimum values of $F_h + \psi$ with ψ varying in a restricted class of continuous functions. In a certain sense, Theorem 3 is the converse of Theorems 1 and 2.

By applying Theorems 3, 2 and 1 successively, it is possible to treat the example at the beginning of this paper. Similarly, these theorems can be applied in more complicated cases, e.g. to the sequences of quadratic functionals

$$F_h(u) = \int_\Omega \left[\sum_{i,j=1}^n a_{ij}^{(h)} D_i u D_j u - 2 f_h u\right] dx + \int_\Omega \psi\left(x, u(x)\right) dx \qquad (h = 1, 2, \ldots),$$

by relating the $\Gamma\left(L^2(\Omega)^-\right)$ convergence of (F_h) with the asymptotic behavior as $h \to +\infty$ of the equations

$$-\sum_{i,j=1}^n D_i(a_{ij}^{(h)} D_i u) = f_h.$$

More generally, the $\Gamma(X^-)$ convergence of sequences of functionals such as

$$F_h(u) = \int_\Omega f_h\left(x, u(x), Du(x)\right) dx + \int_\Omega \psi\left(x, u(x)\right) dx \qquad (5)$$

may be studied by examining the asymptotic behavior as $h \to +\infty$ of the Euler–Lagrange equations of the first integral in (5). This subject has been extensively studied (we refer the reader to [5], [8], [29], [38], [41], [44], [49], [58], [59], [60]).

Besides this approach, which one might call "indirect", there is a different one, related to the direct methods in Calculus of Variations and to measure theory. We describe them now.

Direct methods in Γ-convergence

In order to illustrate the results in Γ-convergence obtained by direct methods, let us begin with a result concerning area-like functionals (see [24]). For $h \in \boldsymbol{N}$, A an open bounded subset of $\boldsymbol{R}^n$ and $u \in C^1(A)$, define

$$F_h(u, A) = \int_A f_h\big(x, u(x), Du(x)\big)\, dx, \tag{6}$$

where f_h: $\boldsymbol{R}^n \times \boldsymbol{R} \times \boldsymbol{R}^n \to \boldsymbol{R}$ are measurable functions such that

$$|z| \leqslant f_h(x, y, z) \leqslant c(1 + |y| + |z|), \tag{7}$$

$$|f_h(x, y_1, z_1) - f_h(x, y_2, z_2)| \leqslant c(|y_1 - y_2| + |z_1 - z_2|) \tag{8}$$

with c a real constant independent of h. Then it is possible to select a subsequence $(F_{\sigma(h)})$ such that

$$\Gamma\big(L^1(A)^-\big) \lim_{h \to +\infty} F_{\sigma(h)}(u, A) = F_\infty(u, A)$$

exists for every open bounded set A and for every $u \in L^1(A)$, and the functional $F_\infty(u, A)$ is represented on $W^{1,1}(A)$ by an integral like (6). Note that $F_\infty(u, A)$ takes finite values on $W^{1,1}(A)$ and also, in contrast to F_h, on the space $BV(A)$ of integrable functions on A whose first order distributional derivatives are measures with finite total variation on A. However, under the general hypotheses (7) and (8), we do not know a meaningful integral representation of $F_\infty(u, A)$ when $u \in BV(A) \setminus W^{1,1}(A)$. This representation problem on $BV(A)$ is also nontrivial if $f_h(x, y, z) = f(x, y, z)$, and is solved only in some particular cases. For example, if $f(x, y, z) = |z|$ we have, for every $u \in BV(A)$, the formula

$$F_\infty(u, A) = \int_A |Du| = \sup\Big\{ \int_A u \operatorname{div} \varphi \, dx \colon \varphi \in [C_0^\infty(A)]^n, |\varphi| \leqslant 1 \Big\},$$

which illustrates the link between Γ-convergence and nonparametric minimal surface theory (see [45]). Analogous formulae for more general f_h have been found recently (see [20]).

The previous result is essentially a compactness theorem with respect to the Γ-convergence of a class of integral functionals. Similar results for other classes of integral functionals have been proved by several authors (see for example [11], [15], [42], [56]).

There are also interesting examples in which the Γ-limit of a sequence of functionals exists but has a form very different from the approximating functionals. Recall for example (see [46]) that if Ω has a smooth boundary, then for every $u \in BV(\Omega)$

$$\Gamma\big(L^1(\Omega)^-\big) \lim_{h\to+\infty} \int_\Omega \left[\frac{|Du|^2}{h} + h \sin^2(\pi u)\right] dx = \begin{cases} \dfrac{4}{\pi} \displaystyle\int_\Omega |Du| & \text{if } \sin(\pi u) = 0 \text{ a.e.,} \\ +\infty & \text{otherwise.} \end{cases}$$

Another interesting example in which a sequence of obstacles converges to a smooth perturbation was found in [16]. There, a sequence (ψ_h) of real functions on $\boldsymbol{R}^n$ was constructed such that, setting

$$J_{\psi_h}(u, A) = \begin{cases} 0 & \text{if } u \geqslant \psi_h \text{ a.e. on } A, \\ +\infty & \text{otherwise} \end{cases}$$

for A an open subset of $\boldsymbol{R}^n$ and $u \in L^2(A)$, we have

$$\Gamma\big(L^2(A)^-\big) \lim_{h\to+\infty} \left[\int_A |Du|^2 dx + J_{\psi_h}(u, A)\right] = \int_A |Du|^2 dx + \int_A (|u| - u)^2 dx$$

for every bounded open subset A of $\boldsymbol{R}^n$ with smooth boundary and for every $u \in H^1(A)$.

More generally, the Γ-convergence of integral functionals constrained by unilateral or bilateral obstacles has been studied by several authors (see [3], [22]); we mention here a compactness result (see [21]).

Let $f\colon \boldsymbol{R}^n \times \boldsymbol{R} \times \boldsymbol{R}^n \to \boldsymbol{R}$ be a function such that $f(x, y, z)$ is continuous in y and convex in z. Suppose that for a fixed $p > 1$

$$|z|^p / \leqslant f(x, y, z) \leqslant C(1 + |y|^p + |z|^p)$$

with C a real constant. Then, for every bounded open subset A of $\boldsymbol{R}^n$ with smooth boundary and for every sequence (ψ_h) of uniformly bounded functions on $\boldsymbol{R}^n$, there exist a subsequence $(\psi_{\sigma(h)})$, a function $\varphi\colon \boldsymbol{R}^n \times \times \boldsymbol{R} \to \bar{\boldsymbol{R}}$ and a positive measure μ such that

$$\Gamma\big(L^p(A)^-\big) \lim_{h\to+\infty} \left[\int_A f\big(x, u(x), Du(x)\big) dx + J_{\psi_{\sigma(h)}}(u, A)\right]$$
$$= \int_A f\big(x, u(x), Du(x)\big) dx + \int_A \varphi\big(x, u(x)\big) d\mu(x)$$

for every $u \in H^{1,p}(A)$.

Besides these obstacle problems, we can consider also the $\Gamma(L^p(\Omega)^-)$ convergence of integral functionals subjected to varying boundary Dirichlet conditions. For example, let (F_h) be the sequence of functionals on $C^1(\overline{\Omega})$ defined by

$$F_h(u) = \begin{cases} \int_\Omega \sqrt{1+|Du|^2}\,dx & \text{if } u = \psi_h \text{ on } \partial\Omega, \\ +\infty & \text{otherwise,} \end{cases}$$

where Ω is a bounded open subset of $\boldsymbol{R}^n$ with smooth boundary and (ψ_h) is a given sequence in $C^1(\partial\Omega)$.

First, assume that (ψ_h) converges to ψ_∞ in $L^1(\partial\Omega)$. Then, for every $u \in BV(\Omega)$, we have

$$\Gamma(L^1(\Omega)^-) \lim_{h\to+\infty} F_h(u) = \int_\Omega \sqrt{1+|Du|^2} + \int_{\partial\Omega} |\overline{u} - \psi_\infty|\,d\mathscr{H}_{n-1}, \tag{9}$$

where

$$\int_\Omega \sqrt{1+|Du|^2} = \sup\Big\{ \int_\Omega \Big[\varphi_0 + u\sum_{i=1}^n D_i\varphi_i\Big]dx\colon \varphi_0, \varphi_1, \dots, \varphi_n \in C_0^\infty(\Omega), \sum_{i=0}^n \varphi_i^2 \leqslant 1\Big\},$$

$$\overline{u}(x) = \lim_{\varrho\to 0} \frac{1}{\operatorname{meas}(\Omega\cap B_\varrho(x))} \int_{\Omega\cap B_\varrho(x)} u(y)\,dy.$$

Here $B_\varrho(x)$ is the ball centered at x with radius ϱ and $\mathscr{H}_{n-1}$ is the Hausdorff surface measure on $\partial\Omega$.

Observe that in this case the Γ-limit (9) corresponds to the modern weak formulation of the nonparametric Plateau problem (see [45]). The weak convergence in $L^1(\partial\Omega)$ of (ψ_h) does not guarantee the existence of the Γ-limit (9). Nevertheless, if we suppose that the sequence (ψ_h) is bounded in $L^1(\partial\Omega)$, then there exists a subsequence $(F_{\sigma(h)})$ such that for every $u \in BV(\Omega)$

$$\Gamma(L^1(\Omega)^-) \lim_{h\to+\infty} F_{\sigma(h)}(u) = \int_\Omega \sqrt{1+|Du|^2} + \int_{\partial\Omega} w\big(x, \overline{u}(x)\big)\,d\mathscr{H}_{n-1}(x),$$

where $w(x,t)$ is convex and Lipschitz continuous in t, but in general $w(x,t)$ does not have the form $|g(x)-t|$ for any $g \in L^1(\partial\Omega)$. This last result has been used in [28] to construct an example of an everywhere discontinuous function ψ_∞ on the boundary of the unit circle in $\boldsymbol{R}^2$ such that the functional (9), and hence the weak Plateau problem on the circle, have infinitely many minimum points.

In all the results quoted above and in many other problems of Γ-convergence solved by direct methods, one deals with a sequence $(F_h(u, A))$ of functionals which are measures with respect to A. By compactness theorems, it is relatively easy to obtain the Γ-convergence of a subsequence $(F_{\sigma(h)}(u, A))$ to a functional $F_\infty(u, A)$, whilst it is more difficult to prove that $F_\infty(u, A)$ is a measure with respect to A and then to represent $F(u_\infty, A)$ by an integral over A. A recent paper (see [23]) gives some general criteria and methods for answering the question of whether a Γ-limit is a measure.

Very recently, the direct methods of Γ-convergence have been applied to functionals

$$F(u, A) = \int_\Omega f(x, u(x), Du(x))\,dx$$

depending on vector-valued functions u. Their case is rather different from the scalar one, mainly because the lower semicontinuity of $F(u, A)$ does not imply the convexity of the integrand f in the gradient variables, as in the scalar case (see [34], [42]). However, some results, such as the one on area-like functionals satisfying (7) and (8), may be extended to the vector case: in particular, by taking $F_h = F$ all equal, one obtains some interesting relaxation results (see [1], [7]).

We conclude this section with two open problems of the same type: to find the weakest assumptions on the integrands ensuring the convergence of a sequence (F_h) of integral functionals (Γ-convergence of the whole sequence, not only of a subsequence).

The first problem is related with homogenization: the functionals F_h have the form

$$F_h(u) = \int_\Omega f(hx, x, u, Du, D^2u, \ldots, D^mu)\,dx \tag{10}$$

or

$$F_h(u) = \int_\Omega f(x, hu, Du, \ldots, D^mu)\,dx, \tag{10'}$$

where Ω is an open bounded subset of $\boldsymbol{R}^n$ and f is periodic in the (vector) replacement variable hx in (10) or in the (scalar) replacement variable hu in (10'). Here one considers the $\Gamma\left(L^p(\Omega)^-\right)$ limit ($1 \leqslant p < +\infty$): for (10) there are many results (for the direct methods, see [8], [9], [52], [57], [59], [60], for the indirect methods [2], [11], [29], [42], and for a counterexample see [33]). There are fewer results for (10') (see [10]), but in both cases we do not know of results or counterexamples which show that the optimality of the assumptions on f is attained.

The second problem is related with singular perturbations: the functionals F_h have the form

$$F_h(u) = \int_\Omega f(x, u, h^{-1}Du, h^{-2}D^2u, \ldots, h^{-m}D^mu)\,dx.$$

In this case it is the $\Gamma\left(L^p_{\text{weak}}(\Omega)^-\right)$ limit $(1 \leqslant p < +\infty)$ that is interesting (see [13]), and again we should like to obtain information about the optimal assumptions on f.

Other problems and applications

In this last section we mention, without any claim to precision or completeness, some other problems in which Γ-convergence can be applied and which merit further study. In fact, the results already obtained show that Γ-convergence is a flexible instrument which could become useful in many branches of mathematics.

(a) *Γ-convergence of optimal control problems.* This is a direction of research that seems to be promising (see [12]). From the point of view of Γ-convergence, it is useful to formulate an optimal control problem as

$$\min[F(u, v) + K(u, v)].$$

Here F is usually an integral functional depending upon the two functions u and v, and K is the characteristic function of the state equation, that is, $K(u, v) = 0$ if (u, v) satisfies the state equation and $K(u, v) = +\infty$ otherwise. In this kind of problem the extension of the direct methods of Calculus of Variations encounters both the difficulties connected with obstacles and with functionals depending on vector-valued functions, whose components often have to vary in different function spaces.

(b) *Γ-convergence of differential equations.* The Γ-limit of characteristic functions appears also in problems of convergence of differential equations not related to control problems. Suppose we have a sequence

$$E_h(u) = f \tag{11}$$

of differential equations, and let $K_h\colon X \times Y \to \bar{\boldsymbol{R}}$ be the characteristic function such that $K_h(f, u) = 0$ if $E_h(u) = f$ and $K_h(f, u) = +\infty$ otherwise. First one should ask whether the limits

$$K'(f, u) = \Gamma(X^-, Y^-) \lim_{h \to +\infty} K_h(f, u),$$

$$K''(f, u) = \Gamma(X^+, Y^-) \lim_{h \to +\infty} K_h(f, u)$$

exist. If so, the pairs (f, u) such that $K'(f, u) = K''(f, u) = 0$ are "stable" limits of solutions of (11), whereas the pairs (f, u) such that $K'(f, u) = 0$ and $K''(f, u) = +\infty$ are "unstable" limits. Another problem is to establish whether there exists a differential operator E_∞ such that $E_\infty(u) = f$ if and only if $K'(f, u) = 0$ (or $K''(f, u) = 0$). One example of the above situation is the bounce problem (see [14], [30]). Other limits of differential equations have been studied by methods of Γ-convergence (see [52]), and many other results are obtained in the papers on indirect methods in Γ-convergence quoted above.

(c) *Convergence of stationary points, steepest descent curves.* Suppose that (F_h) is a sequence of functionals defined on a Hilbert space X which $\Gamma(X^-)$ converges to a functional F_∞. Then it is interesting to find out when the solutions u_h of the equations

$$f = -\operatorname{grad} F_h(u), \qquad \frac{du}{dt} = -\operatorname{grad} F_h(u), \qquad \frac{d^2u}{dt^2} = -\operatorname{grad} F_h(u) \quad (12)$$

converge to the solution u_∞ of the respective equations with F_∞ instead of F_h.

If F_h are quadratic integrals the conditions (12) may be equivalent to some linear partial differential equations (elliptic, parabolic or hyperbolic; see [59]); in the general case the functional equations (12) may be strongly nonlinear and it may be convenient to replace the condition $\operatorname{grad} F(u_0) = a$ with the weaker subdifferentiability condition

$$\liminf_{u \to u_0} \frac{F(u) - F(u_0) - \langle a, u - u_0 \rangle}{\|u - u_0\|} \geqslant 0,$$

in which $\| \ \|$ and $\langle \ , \ \rangle$ obviously denote norm and scalar product in X. Results are known for the "elliptic" case $f = -\operatorname{grad} F_h(u)$ where F_h is the sum of a convex function and of a smooth function (see [2]) and for the "parabolic" case $du/dt = -\operatorname{grad} F_h(u)$ studied by generalized steepest descent curves (see [27]). The first application of this theory has been the study of geodesic curves with obstacles (see [43]).

(d) *Convergence of minimax and of Pareto minima.* Here we only want to quote the recent works [4], [18], [51].

(e) *Random homogenization.* Besides the classical "deterministic" homogenization, which essentially concerns periodic structures, several authors have studied similar problems for structures only statistically periodic (see [39], [50], [62]). A study has been begun of random variables

on spaces of integral functionals and their convergence (see [35]). Here it is natural and useful to consider Γ-convergence as a "topological" structure on spaces of integral functionals.

Acknowledgment

I wish to thank C. Sbordone, G. Dal Maso, L. Modica, G. Buttazzo, D. Mundici, and S. Salamon for their help in the preparation of this lecture.

References

For a wider bibliography, we refer the interested reader to E. De Giorgi [25].

[1] Acerbi E. and Fusco N., Semicontinuity Problems in Calculus of Variations, *Arch. Rational Mech. Anal.* (to appear).

[2] Ambrosetti A. and Sbordone C., Γ-convergenza e G-convergenza per problemi non lineari di tipo ellittico, *Boll. Un. Mat. Ital.* (5) **13**-A (1976), pp. 352–362.

[3] Attouch H. and Picard C., *Asymptotic Analysis of Variational Problems with Constraints of Obstacle Type*, Publications Mathématiques d'Orsay, preprint.

[4] Attouch H. and Wets R. J., A Convergence Theory for Saddle Functions, *Trans. Amer. Math. Soc.* (to appear).

[5] Babuška I., Solution of Interface Problems by Homogenization, *SIAM J. Math. Anal.* **7** (1976), pp. 603–645; **8** (1977), pp. 923–937.

[6] Bahvalov N. S., Averaging of Partial Differential Equations with Rapidly Oscillating Coefficients, *Soviet Math. Dokl.* **16** (1975), pp. 351–355; transl. from *Dokl. Akad. Nauk SSSR* **221** (1975), pp. 516–519.

[7] Ball J. M., Convexity Conditions and Existence Theorems in Nonlinear Elasticity, *Arch. Rational Mech. Anal.* **63** (1977), pp. 337–403.

[8] Bensoussan A., Lions J. L., and Papanicolau G., *Asymptotic Methods in Periodic Structures*, North-Holland, Amsterdam, 1978.

[9] Boccardo L. and Murat F., *Homogénéisation de problèmes quasi-linéaires*, Publ. I.R.M.A., Lille, **3** (7) (1981).

[10] Buttazzo G. and Dal Maso G., Γ-Limit of a Sequence of Nonconvex and Non-Equilipschitz Integral Functionals, *Ricerche Mat.* **27** (1978), pp. 235–251.

[11] Buttazzo G. and Dal Maso G., Γ-Limits of Integral Functionals, *J. Analyse Math.* **37** (1980), pp. 145–185.

[12] Buttazzo G. and Dal Maso G., Γ-Convergence and Optimal Control Problems, *J. Optimization Theory Appl.* **38** (1982), pp. 385–407.

[13] Buttazzo G. and Dal Maso G., Γ-convergence et problèmes de perturbation singulière, *C. R. Acad. Sci. Paris Sér. I* **296** (1983), pp. 649–651.

[14] Buttazzo G. and Percivale D., On the Approximation of the Elastic Bounce Problem on Riemannian Manifolds, *J. Differential Equations* **47** (1983), pp. 227–245.

[15] Carbone L., Sur la Γ-convergence des intégrales du type de l'énergie à gradient borné, *J. Math. Pures Appl.* (9) **56** (1977), pp. 79–84.

[16] Carbone L. and Colombini F., On Convergence of Functionals with Unilateral Constraints, *J. Math. Pures Appl.* (9) **59** (1980), pp. 465–500.

[17] Carriero M. and Pascali E., Il problema del rimbalzo unidimensionale e sue approssimazioni con penalizzazioni non convesse, *Rend. Mat.* (4) **13** (1980), pp. 541–554.

[18] Cavazzuti E., Γ-convergenza multipla, convergenza di punti di sella e di max-min, *Boll. Un. Mat. Ital.* (6) **1**-B (1982), pp. 251–274.

[19] Choquet G., Convergences, *Ann. de l'Univ. de Grenoble* **23** (1947-1948), pp. 55–112.

[20] Dal Maso G., Integral Representation on $BV(\Omega)$ of Γ-Limits of Variational Integrals, *Manuscripta Math.* **30** (1980), pp. 387–416.

[21] Dal Maso G., Limits of Minimum Problems for General Integral Functionals with Unilateral Obstacles, *Atti Accad. Naz. Lincei* (to appear).

[22] Dal Maso G. and Longo P., Γ-Limits of Obstacles, *Ann. Mat. Pura Appl.* (4) **128** (1980), pp. 1–50.

[23] Dal Maso G. and Modica L., A General Theory of Variational Functionals. In: *Topics in functional analysis 1980-1981*, Scuola Normale Superiore, Pisa, 1981, pp. 149–221.

[24] De Giorgi E., Sulla convergenza di alcune successioni di integrali del tipo dell'area, *Rend. Mat.* (4) **8** (1975), pp. 277–294.

[25] De Giorgi E., Generalized Limits in Calculus of Variations. In: *Topics in functional analysis 1980-1981*, Scuola Normale Superiore, Pisa, 1981, pp. 117–148.

[26] De Giorgi E. and Franzoni T., Su un tipo di convergenza variazionale, *Rend. Sem. Mat. Brescia* **3** (1979), pp. 63–101.

[27] De Giorgi E., Marino A., and Tosques M., Problemi di evoluzione in spazi metrici e curve di massima pendenza, *Atti Accad. Naz. Lincei, Rend. Cl. Sci. Fis. Mat. Natur.* (8) **68** (1980), pp. 180–187.

[28] De Giorgi E. and Modica L., *Γ-convergenza e superfici minime*, Scuola Normale Superiore, Pisa, 1979, preprint.

[29] De Giorgi E. and Spagnolo S., Sulla convergenza degli integrali dell'energia per operatori ellittici del II ordine, *Boll. Un. Mat. Ital.* (4) **8** (1973), pp. 391–411.

[30] Degiovanni M., Multiplicity of Solutions for the Bounce Problems, *J. Differential Equations* (to appear).

[31] Denkowski Z., The Convergence of Generalized Sequences of Sets and Functions in Locally Convex Spaces I, *Zeszyty Naukowe UJ* **22** (1980), pp. 37–58.

[32] Dolecki S., Salinetti G., and Wets R. J., Convergence of Functions: Equi-Semicontinuity, *Trans. Amer. Math. Soc.* **276** (1983), pp. 409–429.

[33] Donato P., *Una stima per la differenza di H-limiti e qualche applicazione a problemi di omogeneizzazione*, Univ. Salerno, preprint.

[34] Ekeland I. and Temam R., *Convex Analysis and Variational Problems*, North-Holland, Amsterdam, 1978.

[35] Facchinetti G. and Russo L., Un caso unidimensionale di omogeneizzazione stocastica, *Boll. Un. Mat. Ital.* (6) **2**-C, pp. 159–170.

[36] Greco G. H., Limites et fonctions d'ensembles, *Rend. Sem. Mat. Padova* (to appear).

[37] Joly J. L. and de Thelin F., Convergence of Convex Integrals in L^p Space, *J. Math. Anal. Appl.* **54** (1976), pp. 230–244.

[38] Kha T'en Ngoan, Kozlov S. M., Oleinik O. A., and Zhikov V. V., Averaging and G-Convergence of Differential Operators, *Uspekhi Mat. Nauk* **34** (1979), pp. 65–133; transl. in *Russian Math. Surveys* **34** (1979), pp. 69–147.

[39] Kozlov S. M., Averaging of Random Operators, *Math. USSR-Sb.* **109** (1979), pp. 188–192..

[40] Kuratowski K., *Topology*, Academic Press, New York, 1966.

[41] Lions J. L., *Some Methods in the Mathematical Analysis of Systems and Their Control*, Science Press, Beijing, China; Gordon and Breach Inc., New York, 1981.

[42] Marcellini P. and Sbordone C., Homogenization of Non Uniformly Elliptic Operators, *Applicable Anal.* **68** (1978), pp. 101–114.

[43] Marino A. and Scolozzi D., Geodetiche con ostacolo, *Boll. Un. Mat. Ital.* (6) **2**-B (1983), pp. 1–31.

[44] Marino A. and Spagnolo S., Un tipo di approssimazione dell'operatore $\Sigma_{ij}D_i(a_{ij}D_j)$ con operatori $\Sigma_i D_i(\beta D_i)$, *Ann. Scuola Norm. Sup. Pisa, Cl. Sci. Fis. Mat.* (3) **23** (1969), pp. 657–673.

[45] Miranda M., Distribuzioni aventi derivate misure. Insiemi di perimetro finito, *Ann. Scuola Norm. Sup. Pisa, Cl. Sci. Fis. Mat.* (3) **18** (1964), pp. 27–56.

[46] Modica L. and Mortola S., Un esempio di Γ-convergenza, *Boll. Un. Mat. Ital.* (5) **14**-B (1977), pp. 285–299.

[47] Moscariello G., Γ-convergenza negli spazi sequenziali, *Rend. Accad. Sci. Fis. Mat. Napoli* (4) **43** (1976), pp. 333–350.

[48] Mosco U., Convergence of Convex Sets and of Solutions of Variational Inequalities, *Advances in Math.* **3** (1969), pp. 510–585.

[49] Murat F., Compacité par compensation, *Ann. Scuola Norm. Sup. Pisa, Cl. Sci.* (4) **5** (1978), pp. 489–507.

[50] Papanicolaou G. C. and Varadhan S. R. S., Boundary Value Problems with Rapidly Oscillating Random Coefficients. In: *Random Fields II, Colloquia Mathematica Janos Bolyai* **27**, North-Holland, Amsterdam, 1981, pp. 835–873.

[51] Peirone R., Γ-limiti e minimi di Pareto, *Atti Accad. Naz. Lincei* (to appear).

[52] Piccinini L. C., Homogenization for Ordinary Differential Equations, *Rend. Cir. Mat. Palermo* (2) **27** (1978), pp. 95–112.

[53] Rockafellar R. T., Generalized Directional Derivatives and Subgradients of Nonconvex Functions, *Canad. J. Math.* **32** (1980), pp. 257–280.

[54] Salinetti G. and Wets R. J., On the Convergence of Closed-Valued Measurable Multifunctions, *Trans. Amer. Math. Soc.* **266** (1981), pp. 275–289.

[55] Sanchez-Palencia E., *Nonhomogeneus Media and Vibration Theory*, Lecture Notes in Physics, **127**, Springer, 1980.

[56] Sbordone C., Su alcune applicazioni di un tipo di convergenza variazionale, *Ann. Scuola Norm. Sup. Pisa, Cl. Sci.* (4) **2** (1975), pp. 617–638.

[57] Senatorov P. K., The Stability of a Solution of Dirichlet's Problems for an Elliptic Equation with Respect to Perturbation in Measure of Its Coefficients, *Differentsial'nye Uravneniya* **6** (1970), pp. 1725–1726; transl. in *Differential Equations* **6** (1970), pp. 1312–1313.

[58] Simon L., On G-Convergence of Elliptic Operators, *Indiana Univ. Math. J.* **28** (1979), pp. 587–594.

[59] Spagnolo S., Sulla convergenza delle soluzioni di equazioni paraboliche ed ellittiche, *Ann. Scuola Norm. Sup. Pisa, Cl. Sci. Fis. Mat.* (3) **22** (1968), pp. 575–597.

[60] Tartar L., Homogénéisation en hydrodynamique, *Lecture Notes in Mathematics*, **594**, Springer, (1977), pp. 474–481.

[61] Wijsman R., Convergence of Sequences of Convex Sets, Cones and Functions II, *Trans. Amer. Math. Soc.* **123** (1966), pp. 32–45.

[62] Yurinskij V. V., Averaging Elliptic Equations with Random Coefficients, *Siberian Math. J.* **20** (1980), pp. 611–623; transl. from *Sibirsk. Mat. Zh.* **21** (1980), pp. 209–223.

[63] Zolezzi T., Characterization of Some Variational Perturbations of the Abstract Linear-Quadratic Problems, *SIAM J. Control Optim.* **16** (1978), pp. 106–121.

SCUOLA NORMALE SUPERIORE,
PIAZZA DEI CAVALIERI, 7,
I-56100 PISA,
ITALY

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

TADEUSZ IWANIEC

Some Aspects of Partial Differential Equations and Quasiregular Mappings

This article grew out of my dual interest in partial differential equations and quasiconformal mappings. It can be considered as complementary to the lecture of Jussi Väisälä [35] at the Helsinki Congress where direct geometric methods of quasiregular theory were stressed. As a whole, quasiconformal theory develops into a branch of geometric multidimensional analysis with rather broad connections.

1. Elliptic systems in two variables

Quasiconformal mappings in the plane were introduced as early as 1928 by Herbert Grötzsch. However, the general concept of quasiconformality was clarified in the pioneering papers on nonlinear hydrodynamics by M. A. Lavrentiev. In his approach prominence was given to fundamental connections of quasiconformal and quasiregular mappings with partial differential equations of elliptic type. Quasiregular mappings may be considered as weak solutions of the Beltrami equation

$$f_{\bar{z}} = \mu f_z \tag{1}$$

with arbitrary complex-valued measurable coefficient $\mu = \mu(z)$ satisfying the uniform ellipticity condition $|\mu(z)| \leqslant q_0 < 1$ a.e. This equation is the most important case of the general quasilinear elliptic system

$$f_{\bar{z}} = q_1(z,f) f_z + q_2(z,f)\overline{f_z}, \qquad |q_1(z,f)| + |q_2(z,f)| \leqslant q_0 < 1 \tag{2}$$

as well as the nonlinear system, strongly elliptic in the geometric sense [2]:

$$\begin{gathered} f_{\bar{z}} = h(z, f, f_z), \\ |h(z,f,\xi) - h(z,f,\zeta)| \leqslant q_0 |\xi - \zeta|, \qquad h(z,f,0) \equiv 0. \end{gathered} \tag{3}$$

The latter arose from the studies of the complex gradient $f(z) = u_x - iu_y$ of the solution of a second-order nonlinear elliptic equation $F(x, y, u_x, u_y, u_{xx}, u_{xy}, u_{yy}) = 0$, in particular, of the gas dynamics equation.

The list of differential equations related to quasiconformal theory is longer but we end it here for simplicity. The modern state of the theory owes much to L. Ahlfors, L. Bers, B. Bojarski, C. Morrey, L. Nirenberg, I. N. Vekua and many others.

2. Quasiregular mappings in $\boldsymbol{R}^n$

Let Ω and Ω' be domains of $\boldsymbol{R}^n$ and let $f\colon \Omega \to \Omega'$ be a mapping of the Sobolev class $W^1_{n,\mathrm{loc}}(\Omega)$. Then f is said to be *K-quasiregular*, briefly *K-q.r.*, if and only if

$$|Df(x)|^n \leqslant KJ(x, f) \quad \text{a.e. in } \Omega, \quad 1 \leqslant K < \infty, \tag{4}$$

where $Df(x)$ denotes the Jacobian matrix or the linear tangent map $Df(x)\colon T_x\Omega \to T_{f(x)}\Omega'$ of the tangent spaces $T_x\Omega$ and $T_{f(x)}\Omega'$ identified with $\boldsymbol{R}^n$. The norm and the Jacobian determinant of $Df(x)$ are denoted by $|Df(x)|$ and $J(x, f)$ respectively. The smallest K for which (4) holds is called the *outer dilatation* of f. A K-q.r.m. which is also homeomorphic is called a *K-quasiconformal mapping*, briefly K-q.c.m.

Geometrically, inequality (4) means that f takes infinitesimal spheres onto infinitesimal ellipsoids of uniformly bounded eccentricity.

Systematic studies of higher-dimensional q.c.m. began with F. W. Gehring [10] and J. Väisälä [33], while the q.r.m. were introduced in the sixties by Yu. Rešetniak [25] and by Finnish mathematicians: O. Martio, S. Rickman, and J. Väisälä [22], [23].

Allied with holomorphic functions, the quasiregular mappings behave similarly. They are open, discrete and have the branch sets of topological dimension $\leqslant n-2$.

In the development of q.r.m. the methods of p.d.e. proved to be the most flexible and universal tools. On the other hand, geometric and topological methods in q.r.m. gave inspiration and new ideas for pure p.d.e., differential geometry and classical analysis. The true value of quasiconformal theory lies in its interdisciplinary role ([1], [3], [4], [35]).

3. Basic first-order differential systems

These are derived from (4) as a consequence of uniform estimates of the tangent map $Df(x)$: $T_x\Omega \to T_{f(x)}\Omega'$. A point x in Ω will be called *regular* if $Df(x)$ exists and $J(x,f) \neq 0$. Define the symmetric matrix $G(x)$ as

$$G(x) = \begin{cases} \big(J(x,f)\big)^{-2/n} D^*f(x)\,Df(x) & \text{if } x \text{ is regular,} \\ I \text{ (the unit matrix)} & \text{otherwise.} \end{cases}$$

Hereafter D^*f stands for the transposed Jacobian matrix. It follows from (4) that G is uniformly bounded and uniformly positive, precisely

$$K^{(2-2n)/n}\,\xi^2 \leqslant \langle G(x)\,\xi,\,\xi\rangle \leqslant K^{2/n}\,\xi^2, \qquad (x,\,\xi) \in \Omega \times \boldsymbol{R}^n. \tag{5}$$

The domain Ω becomes a Riemannian manifold with the positive-definite element of arc length $ds^2 = \sum G_{ij}(x)\,dx_i\,dx_j$. In general, the metric tensor $G = G(x)$ need not be continuous. By the definition of G, the map f satisfies the matrix equation

$$D^*f(x)\,Df(x) = J(x,f)^{2/n} G(x) \quad \text{a.e. in } \Omega. \tag{6}$$

This extends the two-dimensional Beltrami equation (1). For $G(x) = I$ (6) reduces to an n-dimensional analogue of the Cauchy–Riemann system

$$D^*f(x)\,Df(x) = J(x,f)^{2/n} I, \qquad f \in W^1_{n,\mathrm{loc}}(\Omega). \tag{7}$$

Now suppose that $H = H(y)$ is a Riemann tensor defined a.e. in $\Omega' \subset \boldsymbol{R}^n$. In the classical differential geometry the system

$$D^*f(x)\,H\big(f(x)\big)\,Df(x) = J(x,f)^{2/n} G(x), \qquad f \in W^1_{n,\mathrm{loc}}(\Omega) \tag{8}$$

expresses the condition for f to be a conformal equivalence between Ω and Ω'.

A slight generalization of (8) leads to the system

$$D^*f(x)\,H(x,f)\,Df(x) = J(x,f)^{2/n} G(x,f), \qquad f \in W^1_{n,\mathrm{loc}}(\Omega) \tag{9}$$

with symmetric and positive matrices G and H defined on the product domain $\Omega \times \Omega'$ on which they are uniformly bounded and normalized by $\det G(x,f) = \det H(x,f) = 1$. This extends the general quasilinear system (2) for which we have

$$q_1 = \frac{G_{11} - G_{22} - 2iG_{12}}{G_{11} + G_{22} + H_{11} + H_{22}}, \qquad q_2 = -\frac{H_{11} - H_{22} + 2iH_{12}}{G_{11} + G_{22} + H_{11} + H_{22}}.$$

The matrices G and H are referred to as the *matrix characteristics* of f. In spite of many similarities, the elliptic systems in the plane differ very much from their higher-dimensional counterparts. The latter are overdetermined and nonlinear. More importantly, their type classification breaks down at the points x_0 where $Df(x_0) = 0$, and the system (9) "degenerates". Nevertheless, on the basis of such systems various fundamental problems of q.r.m. are more deeply understood and new prospects for their solutions appear.

4. Degenerate second-order elliptic equations

The case of $K = 1$ in (4) is extremal and we have $\mathrm{Tr}^{n/2}(D^*fDf) = n^{n/2}J(x, f)$. This identity is also the extremal case of Hadamard's inequality $n^{n/2}J(x, h) \leqslant \mathrm{Tr}^{n/2}(D^*hDh)$ which is in the direction opposite to (4). The above facts apply in the following variational problem: given a 1-q.r.m. f: $\Omega \to \boldsymbol{R}^n$ and a subdomain $U \subset \Omega$, find a mapping h: $U \to \boldsymbol{R}^n$ such that $h - f \in \mathring{W}^1_n(U)$ and h minimizes the integral

$$I[h] = \int_U \mathrm{Tr}^{n/2}[D^*h(x)\, Dh(x)]\,dx. \tag{10}$$

The direct method applies and by the convexity of the integrand one sees that the minimizer exists and is unique. However, if f and h agree on the boundary of U, then one has the identity $\int_U J(x, f)\,dx = \int_U J(x, h)\,dx$ (this is a property of so-called null-Lagrangians). Hence

$$\int_U \mathrm{Tr}^{n/2}(D^*fDf) = n^{n/2} \int_U J(x, f)\,dx = n^{n/2} \int_U J(x, h)\,dx$$

$$\leqslant \int_U \mathrm{Tr}^{n/2}(D^*hDh).$$

This shows that f minimizers I.

If f is a K-q.r.m., one finds in this way that f is a quasiminimizer of I, precisely

$$\int_U \mathrm{Tr}^{n/2}(D^*fDf) \leqslant K \int_U \mathrm{Tr}^{n/2}(D^*hDh)$$

for every h such that $h - f \in \mathring{W}^1_n(U)$. Generalizing, the solutions of (6) minimize the integral

$$I_{G^{-1}}[h] = \int_U \mathrm{Tr}^{n/2}[D^*h(x)G^{-1}(x)Dh(x)]\,dx. \tag{11}$$

Variational problems for vector-valued functions are studied in nonlinear elastostatics, but rather without discussing their connections with first-order differential systems. It is of great significance for quasiconformal theory that the unknowns $f = (f^1, \ldots, f^n)$ which minimize the functional (11) uncouple the Euler–Lagrange system corresponding to (11), providing they solve (6). Therefore each component $u = f^i$, $i = 1, \ldots, n$, of the solution of the Beltrami system (6) satisfies the single equation

$$\operatorname{div} A(x, \nabla u) = 0, \qquad A(x, \xi) = \langle G^{-1}(x)\,\xi, \xi\rangle^{(n-2)/2} G^{-1}(x)\,\xi. \tag{12}$$

At the same time this is the Euler–Lagrange equation corresponding to the variational integral

$$J[u] = \int_U F(x, \nabla u)\,dx, \qquad F(x, \xi) = \langle G^{-1}(x)\,\xi, \xi\rangle^{n/2}. \tag{13}$$

For conformal mappings, (12) turns out to be the n-harmonic equation $\operatorname{div} |\nabla u|^{n-2}\nabla u = 0$, whose solutions $u \in W^1_{n,\mathrm{loc}}(U)$ are called n-harmonic functions [7].

Equation (9) leads to the second-order divergence system

$$\sum_{j=1}^{n} \frac{\partial}{\partial x_j} A_{ij}(x, f, Df) = 0, \qquad i = 1, \ldots, n, \tag{14}$$

where $A_{ij}(x, f, L)$: $\Omega \times \Omega' \times \boldsymbol{R}^{n\times n} \to \boldsymbol{R}$ are given $(n-1)$-homogeneous forms in the matrix variable $L \in \boldsymbol{R}^{n\times n}$ [16]. In general, the unknowns are coupled, and system (14) cannot be reduced to single equations.

The functional (13) is a typical example of a quasiregular variational integral for which $C^{1+\varepsilon}$ regularity is the best possible. Even n-harmonic functions need not be C^2 smooth [7]. Nevertheless, some ideas of the classical regular variational problems may be applied to (13). The integrand $F(x, \xi)$ for $|\xi| \to \infty$ behaves like $|\xi|^n$ with $n = \dim\ \Omega$. Therefore, the natural solution space to consider (12) is $W^1_{n,\mathrm{loc}}(\Omega)$, which is the borderline space in Sobolev's imbedding theorems. Notice that the class of equations (12) is quasiconformally invariant, the n-harmonic equation is conformally invariant. These and other special facts concerning equations (12) are responsible for basic geometric and topological properties of their solutions. This is particularly apparent in the works of Yu. Rešetniak (see [27], [30]).

THEOREM 1. *Let $B \subsetneq \Omega$ be a relatively closed subset of Ω and let u be a solution of* (12) *in* $W^1_{n,\mathrm{loc}}(\Omega \setminus B)$. *Suppose that every* $x_0 \in B$ *is a limit point of* $\Omega \setminus B$ *and also* $\lim_{x \to x_0} u(x) = +\infty$. *Then B is a set of zero conformal capacity.*

Now, let $f\colon \Omega \to \boldsymbol{R}^n$ be a q.r.m. and let $f_0 \in \boldsymbol{R}^n$. Denote $B = \{x \in \Omega;\ f(x) = f_0\}$, the preimage of f_0. It can be verified that the function $u(x) = -\log|f(x) - f_0|$ is also a solution of (12) on $\Omega \setminus B$. In view of Theorem 1, B is a set of zero conformal capacity which implies that the linear Hausdorff measure of B is equal to zero. This property of f allows us to define the local topological degree of f. By general geometric arguments (see [32]) the discreteness and openness of f follow ([6], [30]).

5. Smooth characteristics

We assume here that G and H are twice differentiable and that the solution f of (9) has continuous derivatives up to the third order. By analogy with the classical calculations of Weyl and Schouten on the conformal curvature, the following second-order equations are derived [3]:

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = F_{ij}\big(x, f, Df, \operatorname{grad} J(x,f)\big), \qquad i,j = 1, \ldots, n. \tag{15}$$

For the conformal case we write them explicitly as

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{1}{nJ}\left(J_{x_i} f_{x_j} + J_{x_j} f_{x_i} - \delta_{ij} \sum_a J_{x_a} f_a\right), \qquad J = J(x,f). \tag{16}$$

One more differentiation of (15) leads to an algebraic linear system for the third derivatives of f. When $n > 2$ this system can be resolved:

$$\frac{\partial^3 f}{\partial x_i \partial x_j \partial x_k} = F_{ijk}\big(x, f, Df, \operatorname{grad} J(x,f)\big), \qquad i,j,k = 1, \ldots, n. \tag{17}$$

In the case of the n-dimensional Beltrami system (6), these formulas contain a linear and uniformly elliptic equation for the scalar function $v = J(x,f)^{(n-2)/2n}$,

$$\sum_{i,j} G^{ij}(x) \frac{\partial^2 v}{\partial x_i \partial x_j} - \sum_{i,j,k} G^{ij}(x)\, \Gamma^k_{ij}(x) \frac{\partial v}{\partial x_k} + \frac{n-2}{4(n-1)} R(x) v = 0, \tag{18}$$

where $\Gamma^k_{ij}(x)$ are the Christoffel symbols and $R(x)$ is the scalar curvature of the Riemann space (Ω, G). This equation is in fact the linear part of the equation which appears in the Yamabe problem. From (18) we see

that v has rather special properties. In the conformal case ($K = 1$) v is harmonic. Thus (18) provides us with some information about the Jacobians of quasiregular mappings.

The system (9) behaves like a system of ordinary differential equations, so that the solutions are uniquely determined by their derivatives of order $\leqslant 2$ at a given point. The following theorem is an immediate consequence of (15) and (17) [3].

THEOREM 2 (the point Cauchy problem). *Let* $G, H \in C^2(\Omega \times \Omega')$ *and let* $x_0 \in \Omega$, $f_0 \in \Omega'$, $\xi \in \boldsymbol{R}^n$, $L \in \boldsymbol{R}^{n\times n}$, $\det L > 0$, $L^* H(x_0, f_0) L = (\det L)^{2/n} \cdot G(x_0, f_0)$. *Then the system* (9) *has at most one local solution of class* C^3 *such that* $f(x_0) = f_0$, $Df(x_0) = L$, $\operatorname{grad} J(x_0, f) = \xi$.

In other words, the family of local solutions of (9) depends on $\frac{1}{2}(n+1) \times (n+2)$ real parameters. The number of the parameters is exactly equal to the dimension of the general Möbius group acting in $\boldsymbol{R}^n$.

6. The Liouville theorem

By using finite difference methods various requlority properties of 1-q.r.m. may be derived from (7). In particular, $[J(x,f)]^2 \in \operatorname{Lip}(\Omega)$, hence $J(x,f)$ is Hölder continuous in Ω. Moreover, $J(x,f)$ is C^∞ on the open set $\Omega_+ = \{x \in \Omega;\ J(x,f) > 0\}$ since $v = \big(J(x,f)\big)^{(n-2)/2n}$ is harmonic (see (18)). From (16) it follows that f is also C^∞ on Ω_+. Therefore we are justified in using the uniqueness theorem in the point Cauchy problem for the system (7) in Ω_+. On the other hand, there exists a (global) solution of (7) with arbitrary Cauchy data, namely a Möbius transformation. Hence, f is Möbius in each connected component of Ω_+. Explicit expressions for the Möbius transformations show that $J(x,f)$ is positive on $\Omega \cap \overline{\Omega}_+$. This means that Ω_+ is relatively closed in Ω. Summarizing, if Ω_+ is empty then f is constant. If this is not the case, then $\Omega_+ = \Omega$ and so f is Möbius in Ω.

These quite elementary arguments were used for the following general version of the classical Liouville theorem [5].

THEOREM 3. *Every* 1-*q.r.m. in a domain of* $\boldsymbol{R}^n$, $n \geqslant 3$, *is either constant or a sense-preserving Möbius transformation.*

In such a form this theorem plays important role in higher-dimensional quasiconformal theory. The Liouville theorem without any a priori regularity assumptions was first proved by F. W. Gehring [10] for locally

1-1 mappings. The case of an arbitrary 1-q.r.m. was treated by Yu. Rešetniak [26]. Both authors referred to non-standard and difficult results of p.d.e., geometry and general quasiconformal theory.

The first important question that arises as a consequence of the Liouville theorem is whether the injectivity remains valid for K-q.r.m. with dilatation K sufficiently close to 1.

INJECTIVITY THEOREM 4. *There exists an* $\varepsilon = \varepsilon(n) > 0$, $n > 2$ *such that every* K*-q.r.m. with* $1 \leqslant K \leqslant 1+\varepsilon$ *is either constant or a local homeomorphism* [14], [23].

7. The regularity theorem

A version of Theorem 4 for mappings of Riemannian manifolds reads as follows:

REGULARITY THEOREM 5. *Let* G *and* H *be of class* $C^k(\Omega \times \Omega')$, $k > 0$. *Then every nonconstant solution of* (9) *is a local diffeomorphism of class* $C^{k+1}(\Omega)$.

Other cases of this theorem were treated with various tools by Yu. Rešetniak [29], J. Lelong-Ferrand [20], and the author [16]. In [16] the second-order system (14) was exploited. Although we have no satisfactory regularity results for general solutions of such systems, in our special case of quasiregular mappings it is possible to use the methods relevant for a single elliptic equation. By the technique of truncated functions we first show that the Jacobian $J(x, f)$ is locally bounded. Now fix $x_0 \in \Omega$ and denote $f_0 = f(x_0) \in \Omega'$. After an affine change of the variables x and f it can be assumed that $G(x_0, f_0) = H(x_0, f_0) = I$. Therefore f becomes a $(1+\varepsilon)$-q.r.m. in a neighbourhood of x_0. Thus, f is a locally 1-1 mapping (or constant). On the other hand, in a neighbourhood of $y = f_0 \in \Omega'$ the inverse map $g = f^{-1}(y)$ satisfies the system $D^*g(y)G(g, y) \cdot Dg(y) = J(y, g)^{2/n}H(g, y)$ which looks like (9). Hence $J(y, g)$ is also locally bounded. In view of the identity $J(x, f)J(y, g) = 1$ it follows that $J(x, f)$ is bounded away from zero and thus the system (14) becomes uniformly elliptic on the solution f. Finally, classical regularity results complete the proof of the theorem.

The above illustrates that the regularity problems for general elliptic systems are more deeply seen if the solutions (or some special classes of the solutions) are interpreted as mappings of Euclidean domains or Riemannian manifolds. Then the analytical methods of study are enriched with additional new tools, of geometrical and topological character.

8. Self-improving property of reverse inequalities

A priori estimates in p.d.e. are in principle the reverse of the classical imbedding inequalities. Restricted to "regular" neighbourhoods, these estimates preserve their structural constants. As many examples show, such estimates improve themselves automatically. To clarify these heuristic remarks we will give some examples that arose in p.d.e. and q.c.m.

Let Ω be a domain of $\boldsymbol{R}^n$. Our "regular" basis of neighbourhoods at a given point $x \in \Omega$ will be the family $\mathscr{F}(x)$ of parallel n-cubes Q such that $x \in Q \Subset \Omega$. Let ω be a measurable function defined on Ω. We will work, under suitable assumptions on ω, with the following local averages over the cubes $Q \in \mathscr{F}(\Omega) = \bigcup_{x \in \Omega} \mathscr{F}(x)$:

$$\omega_Q = \frac{1}{|Q|} \int_Q \omega(y)\, dy, \qquad \|\omega\|_{p,Q} = \left((|\omega|^p)_Q\right)^{1/p} \quad \text{if } p \neq -\infty, 0, +\infty,$$

$$\|\omega\|_{-\infty,Q} = \operatorname*{ess\,inf}_Q |\omega|, \qquad \|\omega\|_{\infty,Q} = \operatorname*{ess\,sup}_Q |\omega|,$$

$$\|\omega\|^{\#}_{p,Q} = \|\omega - \omega_Q\|_{p,Q} \quad \text{if } p \geqslant 1,$$

$$\|\omega\|^{(1)}_{p,Q} = |Q|^{1/n} \|\nabla \omega\|_{p,Q} \quad \text{for } \omega \in W^1_p(Q) \text{ and } 1 \leqslant p \leqslant \infty.$$

By a *reverse Hölder inequality* we mean that for some $0 \neq r < p \neq 0$ and a constant $A = A(n, r, p) \geqslant 1$ the following is true:

$$\|\omega\|_{p,Q} \leqslant A \|\omega\|_{r,Q} \quad \text{for each } Q \in \mathscr{F}(\Omega). \tag{19}$$

Let us recall some examples. The limit case of $p = \infty$ and $r = -\infty$ reduces to the classical Harnack principle. When $p = 1$, $r = 1/(1-q)$, $1 < q < \infty$, then (19) defines the Muckenhoupt class $\mathscr{A}_q$. If $\ln|\omega| \in \mathrm{BMO}(\Omega)$, i.e., $\sup\{\|\ln|\omega|\|^{\#}_Q;\ Q \in \mathscr{F}(\Omega)\} < \infty$, then (19) holds with $p = -r = \varepsilon$ for some positive ε (the John–Nirenberg lemma). The crucial result for q.c.m. was the inequality of F. W. Gehring

$$\|Df\|_{n,Q} \leqslant A(n, K) \|Df\|_{1,Q},$$

which holds for every K-q.c.m. f: $\Omega \to \boldsymbol{R}^n$ [12].

Unfortunately, estimates of the above type are too strong to hold for solutions (or derivatives of solutions) of the classical elliptic differential equations. The point is that (19) implies "the unique continuation property": $\omega \equiv 0$ on a set of positive measure implies $\omega \equiv 0$ a.e. The well-known examples of elliptic equations with the nonunique continuation

property due to A. Pliś and K. Miller illustrate that (19) is not true in general. For solutions of differential equations we can only succeed in proving *weak reverse Hölder inequalities*

$$\|\omega\|_{p,\sigma Q} \leqslant A(n,p,r,\sigma)\|\omega\|_{r,Q}, \qquad 0<r<p\leqslant\infty, \tag{21}$$

where σ is a parameter from the open interval $(0,1)$ and $Q\in\mathscr{F}(\Omega)$. Henceforth σQ stands for the cube of the same center as Q but of diameter σ times smaller.

The simplest example is a subharmonic function ω, for which $\|\omega\|_{\infty,\sigma Q} \leqslant (1-\sigma)^{-n}\|\omega\|_{1,Q}$. Usually, (21) is applied to the derivatives of solutions of elliptic equations or systems. The exponent p is then naturally determined by the leading coefficients. Essentially, (21) follows from the Caccioppoli type estimation: $\|u\|^{(1)}_{p,\sigma Q}\leqslant C(1-\sigma)^{-1}\|u\|_{p,Q}$ regarded as a weak reverse Poincaré inequality. An analogue for q.r.m. is the inequality

$$\|Df\|_{n,\sigma Q}\leqslant A(n,\sigma,K)\|Df\|_{1,Q}, \qquad Q\in\mathscr{F}(\Omega). \tag{22}$$

The self-improving property of (19) with exponents of different signs, $r<0<p$, first appeared in Muckenhoupt's proof that $\mathscr{A}_q=\bigcup_{q'<q}\mathscr{A}_{q'}$ [24]. For positive exponents, it was discovered by F. W. Gehring [12]. We formulate his result in a somewhat greater generality.

LEMMA 1. *Suppose* $\omega\in L^p_{\mathrm{loc}}(\Omega)$ *and* (21) *holds. Then there exists an* $\varepsilon=\varepsilon(n,p,r,A)>0$ *so that* $\omega\in L^{p+\varepsilon}_{\mathrm{loc}}(\Omega)$. *Moreover, for every* $p',r'\in(0,p+\varepsilon)$ *there exists an* $A'=A'(n,r,p,r',p',\sigma,A)$ *such that*

$$\|\omega\|_{p',\sigma Q}\leqslant A'\|\omega\|_{r',Q} \quad \textit{for each } Q\in\mathscr{F}(\Omega). \tag{23}$$

The following comments must be made here: the fact that p in (21) can be increased follows from the Hardy–Littlewood maximal theorem and the following interpolation inequality [6]: for $\omega\in L^{p'}(\Omega)$, $p'\geqslant p\geqslant 1$,

$$\|\omega\|^{p'}_{p',\Omega}\leqslant\|\omega\|^{p'}_{p,\Omega}+2^n\left(1-\frac{p}{p'}\right)\|M_p\omega\|^{p'}_{p',\Omega}.$$

Here Ω is a cube in $\boldsymbol{R}^n$ and $(M_p\omega)(x)=\sup\{\|\omega\|_{p,Q};\ Q\in\mathscr{F}(x)\}$. The right-hand exponent r in (21) can be arbitrarily decreased by an iteration process and a Whitney type subdivision of Q [19].

Discovered in quasiconformal theory, this lemma soon became an important tool for L^p estimates in elliptic equations and systems [9], [13].

Let us make some comments connected with the role of the exponent r. If ω is the positive solution of an elliptic second-order divergence equation then $u=\log\omega$ is a function of class BMO(Ω). Hence by the John–

Nirenberg lemma, $\|\omega\|_{r',Q} \leqslant C\|\omega\|_{-r',Q}$ for sufficiently small $r' > 0$, and by (23) we get the Muckenhoupt type estimation $\|\omega\|_{p',\sigma Q} \leqslant CA'\|\omega\|_{-r',Q}$. In practice, this inequality acts as a substitute of Harnack's inequality [6].

Another useful property of a function ω which satisfies a weak reverse Hölder inequality is that almost every zero of ω is of infinite order. Taking into account that zeros of infinite order for the Jacobian of a nonconstant q.r.m. cannot occur [6], it follows that the Jacobian is positive a.e. Notice that (22) holds in fact for $\sigma = 1$ but with the constant A depending additionally on the topological degree of f [21].

The above remarks on reverse Hölder inequalities reflect a rather general principle on geometrical stability of properties of function classes satisfying natural double type inequalities which appear on several occasions in analysis and geometry [4].

9. L^p-integrability with large exponents

LEMMA 2. *Suppose* $\omega \in L^r_{1\infty}(\Omega)$, $r \geqslant 1$, $0 < \sigma \leqslant 1$, $0 < \theta \leqslant 10^{-8nr}$ *and suppose that for each* $Q \in \mathscr{F}(\Omega)$

$$\|\omega\|^{\#}_{r,\sigma Q} \leqslant \theta \|\omega\|_{r,Q}. \tag{24}$$

Then $\omega \in L^p_{1\infty}(\Omega)$ *for* $p = \dfrac{1}{8n} \log_{10} \dfrac{1}{\theta}$. *Moreover,*

$$\|\omega\|_{p,\varrho Q} \leqslant A_p(n, \varrho)\sigma^{-n/r}\|\omega\|_{r,Q} \tag{25}$$

for each $Q \in \mathscr{F}(\Omega)$ *and* $\varrho \in (0,1)$.

Notice that p does not depend on σ.

The lemma is a consequence of the following maximal inequality: given Ω, which in this case means a cube in $\boldsymbol{R}^n$, and given $f \in L^p(\Omega)$, $p \geqslant r \geqslant 1$, we have

$$\|M_r f\|_{p,\Omega} \leqslant 10^{6np}\|M^{\#}_r f\|_{p,\Omega} + 10^{2n}\|f\|_{r,\Omega},$$

where $(M^{\#}_r f)(x) = \sup\{\|f\|^{\#}_{r,Q};\ Q \in \mathscr{F}(x)\}$.

The latter is in fact a local version of the Fefferman–Stein inequality (see [18]). The case of $\sigma = 1$ has been considered by L. G. Gurov and Yu. Rešetniak [15]. However, in various applications the case $\sigma < 1$ seems to be more useful. For example, this arises typically in elliptic differential equations [17], [18].

Through stability estimates the L^p-problems may be treated by use of Lemma 2, the stronger the estimates, the larger the exponent p. This will be exemplified below.

10. Stability

When speaking of stability in Liouville's theorem, we usually mean an estimation of the distance between a given K-q.r.m. and the Möbius group, which is expressed in terms of the dilatation K. Perhaps the simplest example is the following *"weak stability"* result which directly follows by compactness arguments.

PROPOSITION 1. *There is an increasing function* $\nu\colon [1,\infty)\to[0,2]$, $\lim_{K\to 1}\nu(K)=0$ *with the following property: given a cube* $Q\in\mathscr{F}(\Omega)$ *and a* K*-q.r.m.* $f\in W^1_n(Q)$, $n\geqslant 3$, *there exists a Möbius map* $\varphi\in W^1_n(Q)$ *such that*

$$\|Df-D\varphi\|_{n,\frac{1}{2}Q}\leqslant \nu(K)\,\|Df\|_{n,\Omega}. \tag{26}$$

This result is far from optimal. Deeper studies, due mainly to M. A. Lavrentiev, P. P. Belinskiĭ and Yu. Rešetniak have led to stability in the large as for example [31]:

PROPOSITION 2. *Let* $Q\in\mathscr{F}(\Omega)$ *and let* $f\colon \Omega\to\mathbf{R}^n$ *be a nonconstant* K*-q.r.m. Then there exists a Möbius map* φ *such that*

$$\|\varphi^{-1}\circ f-\mathrm{id}\|_{\infty,Q}\leqslant C(n)(K-1)\operatorname{diam}Q. \tag{27}$$

We emphasize here the universality of the factor $C(n)(K-1)$ in this estimation. It is surprising that the infinite-dimensional objects (q.r.m.) are "uniformly approximated" by objects of fixed finite dimension (Möbius mappings). This estimation can be extended to wider classes of domains, as for instance John's domains in place of Q [31]. Extended in this way, the stability estimates became an important apparatus dealing with global injectivity problems in quasiconformal theory.

Passing to the example of stability estimates in p.d.e. we consider the stationary Schrödinger equation

$$\Delta\omega+V\omega=0,\qquad \omega\in W^1_2(\Omega),\ \Omega\subset\mathbf{R}^n \tag{28}$$

with the potential $V\in L^{n/2}(\Omega)$. The aim is to estimate the distance between ω and harmonic functions by means of local $L^{n/2}$-norms of the potential V.

For each cube $Q\in\mathscr{F}(\Omega)$ one can solve the Dirichlet problem

$$\begin{cases}\Delta h=0, & h\in W^1_2(Q),\\ h-\omega\in \overset{\circ}{W}{}^1_2(Q).\end{cases}$$

Then the standard method of test functions and the Sobolev imbedding inequality yield

$$\|\omega - h\|_{\frac{2n}{n-2},Q} \leqslant \varepsilon(n, \Omega)\|\omega\|_{\frac{2n}{n-2},Q}, \tag{29}$$

where $\varepsilon(n, \Omega) = C(n)\left(\int_{\Omega} |V(x)|^{n/2}dx\right)^{2/n}$.

A distinction between (26) and (29) has to be made here. On the left-hand side of (29) we are concerned with the entire cube Q while in (26) with the subcube $\frac{1}{2}Q$, hence the name "weak stability".

We come now to one of our purposes.

11. L^p-integrability from stability estimates

Since h in (29) depends on Q, this is an uncontrolled term. The following simple trick eliminates h from this inequality. We have

$$\|h\|_{\frac{2n}{n-1},Q} \leqslant \left(1 + \varepsilon(n, \Omega)\right)\|\omega\|_{\frac{2n}{n-2},Q}.$$

And for each $\sigma \in (0, 1]$, (29) implies

$$\|\omega - h\|_{\frac{2n}{n-2},\sigma Q} \leqslant \sigma^{(2-n)/2}\varepsilon(n, \Omega)\|\omega\|_{\frac{2n}{n-2},Q}. \tag{30}$$

On the other hand, harmonic functions are regular in that we have the following universal estimate:

$$\|h\|^{\#}_{\frac{2n}{n-2},\sigma Q} \leqslant C(n)\sigma\|h\|_{\frac{2n}{n-2},Q} \leqslant C(n)\sigma\left(1 + \varepsilon(n, \Omega)\right)\|\omega\|_{\frac{2n}{n-2},Q}. \tag{31}$$

Combining (30) with (31) shows easily that

$$\|\omega\|^{\#}_{\frac{2n}{n-2},\sigma Q} \leqslant \theta\|\omega\|_{\frac{2n}{n-2},Q} \quad \text{for } \sigma \in (0, 1], \tag{32}$$

where $\theta = 2C(n)\sigma + 2\left[C(n)\sigma + \sigma^{(2-n)/2}\right]\varepsilon(n, \Omega)$.

With an appropriate choice of σ and small size of Ω (the latter is needed to make $\varepsilon(n, \Omega)$ small enough) the factor θ will be as small as one likes. Lemma 2 is then applicable. It follows that the solutions ω of the Schrödinger equation (28) are in $L^p_{\mathrm{loc}}(\Omega)$ with arbitrary $p \geqslant 1$.

Similar arguments may be used for other cases of L^p estimates [17], [18], [28]. For q.r.m. the weak stability (26) and the regularity estimation

$$\|D\varphi\|^{\#}_{n,\sigma Q} \leqslant C(n)\sigma\|D\varphi\|_{n,Q}, \quad 0 < \sigma < 1$$

of a Möbius mapping $\varphi \in W^1_n(Q)$ are sufficient to prove that every K-q.r.m. belongs to $W^1_{p,\mathrm{loc}}(\Omega)$ with $p = p(n, K) \to \infty$ as K approaches 1 [17]. Stronger stability estimates imply $p(n, K) \geqslant C(n)/(K-1)$ as $K \to 1$ [28].

12. Existence

The fact that differential systems for isometries and conformal mappings are overdetermined makes the problem of solvability rather complicated. However, geometric analysis of examples in q.c. theory suggested some new ideas. These came from an interpretation of the metric tensor $G(x)$ [8]. We explain these ideas in the case of isometries between n-manifolds in $\boldsymbol{R}^n$. Similarities to the Yamabe problem can be seen.

Any metric tensor $G = G(x)$, $x \in \Omega \subset \boldsymbol{R}^n$, defines a distribution of Euclidean ellipsoids $\mathscr{E}(x) = \{\xi \in T_x\Omega;\ \langle G(x)\xi, \xi\rangle = 1\}$ in the tangent spaces, and vice versa. A map $f\colon (\Omega, G) \to (\Omega', H)$ is isometric if and only if its tangent map $Df\colon T\Omega \to T\Omega'$ is a transformation of the underlying distributions of ellipsoids, i.e.

$$D^*f(x)H(y)Df(x) = G(x), \qquad y = f(x) \in \Omega'. \tag{30}$$

This system becomes well-determined when we introduce $\frac{1}{2}n(n-1)$ additional unknowns. For this purpose we allow free motions of $\mathscr{E}(x)$ in $T_x\Omega$ under the action of a $\frac{1}{2}n(n-1)$-dimensional group on $T_x\Omega$.

In the example below the action is that of the orthogonal group $\mathcal{O}(n)$ in $\boldsymbol{R}^n$ (which applies to quasi-isometries). Then system (30) takes the form:

$$D^*f(x)H(y)Df(x) = O^*(x)G(x)O(x), \qquad y \in f(x), \tag{31}$$

where $O(x) \in \mathcal{O}(n)$ play the role of additional unknowns. These unknowns are independent if and only if the eigenvalues of $G(x)$ are mutually distinct. With this assumption and with G and H real analytic we have:

THEOREM 6. *The system* (31) *has local analytic solutions.*

Thus we observe that the introduction of additional (algebraic) variables into the (nonsolvable!) system (30) makes possible the removal of the geometric obstructions to local solvability (the curvature tensor).

References

[1] Ahlfors L., Conditions for Quasiconformal Deformations in Several Variables. In: *Contributions to Analysis, a collection of papers dedicated to Lipman Bers*, Academic Press, New York, 1974, pp. 19–25.

[2] Bojarski B. and Iwaniec T., Quasiconformal Mappings and Non-Linear Elliptic Systems in Two Variables, Part I, II, *Bull. Acad. Polon. Sci.* **22** (1974), 473–484.

[3] Bojarski B. and Iwaniec T., Topics in Quasiconformal Theory in Several Variables. In: *Proc. First Finnish–Polish Summer School, Podlesice 1977*, Part II, pp. 21–44.

[4] Bojarski B. and Iwaniec T., *Some New Concepts in the Analytical Theory of QC-maps in* $\boldsymbol{R}^n$, $n \geqslant 3$, *and Differential Geometry*, the lecture to the Conference on Global Analysis, Garwitz (DDR), October, 1981.

[5] Bojarski B. and Iwaniec T., Another Approach to Liouville Theorem, *Math. Nachr.* **107** (1982), pp. 253–262.

[6] Bojarski B. and Iwaniec T., Analytical Foundations of the Theory of Quasiconformal Mappings in $\boldsymbol{R}^n$, *Ann. Acad. Sci. Fenn.* **8** (1983), pp. 257–324.

[7] Bojarski B. and Iwaniec T., *p-Harmonic Equation and Quasiregular Mappings*, Bonn, preprint no. 617, 1983.

[8] Bojarski B., Iwaniec T., and Kopiecki R., Riemannian Manifolds with Non-Smooth Metric Tensors and QC-maps. In: *Proc. Seminar on Monge–Ampère Equations and Related Topics, Firenze 1980*, Roma 1982, pp. 123–165.

[9] Elcrat A. and Meyers N. G., Some Results on Regularity for Solutions of Non-Linear Elliptic Systems and Quasiregular Functions, *Duke Math. J.* **42** (1975), pp. 121–136.

[10] Gehring F. W., *Ring Domains in Space, Quasiconformal Mappings in Space and The Liouville Theorem in Space*, the lecture to the 1960 Summer Meeting of AMS in Lansing, Michigan.

[11] Gehring F. W., Rings and Quasiconformal Mappings in Space, *Proc. Nat. Acad. Sci. USA* **47** (1961), pp. 98–105.

[12] Gehring F. W., The L^p-Integrability of the Partial Derivatives of a Quasiconformal Mapping, *Acta Math.* **130** (1973), pp. 265–277.

[13] Giaquinta M. and Modica G., Regularity Results for Some Classes of Higher Order Non-Linear Elliptic Systems, *J. Reine Angew. Math.* **311/312** (1979), pp. 145–169.

[14] Goldstein V. M., On the Behaviour of Mappings with Bounded Distortion with Distortion Coefficient Close to Unity, *Sibirsk. Mat. Zh.* **12**, No 6 (1971), pp. 1250–1258 (in Russian).

[15] Gurov L. G. and Rešetniak Yu., An Analogue of Functions with Bounded Mean Oscillation, *Sibirsk. Mat. Zh.* **17**, No 3 (1976), pp. 540–546 (in Russian).

[16] Iwaniec T., *Regularity Theorems for Solutions of Partial Differential Equations for QC-Maps in Several Dimensions*, Preprint No 153, Pol. Acad. Sci., 1978. and *Dissertationes Math.* **198** (1982).

[17] Iwaniec T., Projections onto Gradient Fields and L^p-Estimates for Degenerated Elliptic Operators, *Studia Math.* **75** (1983), pp. 293–312.

[18] Iwaniec T., On L^p-Integrability in PDEs and Quasiregular Mappings for Large Exponents, *Ann. Acad. Sci. Fenn.* **7** (1982), pp. 301–322.

[19] Iwaniec T. and Nolder C., The Hardy–Littlewood Inequality for Quasiregular

Mappings in Certain Domains in R^n, *Ann. Acad. Sci. Fenn.*, the volume in honour of Professor Olli Lehto on his sixtieth birthday, to appear.

[20] Lelong-Ferrand J., Regularity of Conformal Mappings of Riemannian Manifolds. In: *Proc. Romanian–Finnish Seminar, Bucharest 1976*, Lect. Notes in Math. 743, Berlin, 1979.

[21] Martio O., On the Integrability of the Derivatives of a Quasiregular Mapping, *Math. Scand.* **35** (1974), pp. 43–48.

[22] Martio O., Rickman S., and Väisälä J., Definitions for Quasiregular Mappings, *Ann. Acad. Sci. Fenn.* **448** (1969), pp. 1–40.

[23] Martio O., Rickman S., and Väisälä J., Topological and Metric Properties of Quasiregular Mappings, *Ann. Acad. Sci. Fenn.* **488** (1971), pp. 1–31.

[24] Muckenhoupt B., Weighted Norm Inequalities for the Hardy Maximal Function, *Trans. Amer. Math. Soc.* **165** (1972), pp. 207–226.

[25] Rešetniak Yu., Estimates for Moduli of Continuity of Some Mappings, *Sibirsk. Mat. Zh.* **7**, No 5 (1966), pp. 1106–1114 (in Russian).

[26] Rešetniak Yu., Liouville Conformal Mappings Theorem under Minimal Regularity Hypothesis, *Sibirsk. Mat. Zh.* **8**, No 4 (1967), pp. 835–840 (in Russian).

[27] Rešetniak Yu., On the Set of Singular Points of Solutions of Some Non-Linear Elliptic Equations. *Sibirsk. Mat. Zh.* **9**, No 2 (1968), pp. 354–367 (in Russian).

[28] Rešetniak Yu., Stability Estimates in Liouville Theorem in the Class W_p^1 for Closed Domains, *Sibirsk. Mat. Zh.* **17**, No 6 (1976), pp. 1382–1394 (in Russian).

[29] Rešetniak Yu., Differentiability Properties of Quasiconformal and Conformal Mappings of Riemann Spaces, *Sibirsk. Mat. Zh.* **19**, No 5 (1978), pp. 1166–1183 (in Russian).

[30] Rešetniak Yu., *Mappings with Bounded Distortion in Space*, Nauka, Novosibirsk, 1982 (in Russian).

[31] Rešetniak Yu., *Stability Theorems in Geometry and Analysis*, Nauka, Novosibirsk, 1982 (in Russian).

[32] Titus C. J. and Young G. S., The Extension of Interiority, with Some Applications, *Trans. Amer. Math. Soc.* **103** (1962), pp. 329–340.

[33] Väisälä J., On Quasiconformal Mappings in Space, *Ann. Acad. Sci. Fenn.* **298** (1961), pp. 1–36.

[34] Väisälä J., *Lectures on n-Dimensional Quasiconformal Mappings*, Lect. Notes in Math. 229, 1971.

[35] Väisälä J., A Survey of Quasiregular Maps in R^n. In: *Proceedings of the International Congress of Mathematicians, Helsinki, 1978*, pp. 685–691.

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

SERGIU KLAINERMAN

Long Time Behaviour of Solutions to Nonlinear Wave Equations

Most basic equations of both physics and geometry have the form of nonlinear, second order, autonomous systems

$$G(u, u', u'') = 0, \tag{1}$$

where $u = u(x^1, x^2, \ldots, x^{n+1})$, and u', u'' denote all the first and second partial derivatives of u. For simplicity we will assume here that both u and G are scalars and denote by u_a, u_{ab}, the partial derivatives $\partial_a u$ and respectively $\partial^2_{ab} u$; $a, b = 1, 2, \ldots, n+1$. Let $u^0(x)$ be a given solution of (1). Our equation is said to be elliptic or hyperbolic at $u^0(x)$ according to whether the $(n+1)\times(n+1)$ matrix, whose entries are $G_{ab} = \dfrac{\partial G}{\partial u_{ab}}(u^0, u^{0\prime}, u^{0\prime\prime})$, is nondegenerate and has signature $(1, \ldots, 1, 1)$ or $(1, \ldots, 1, -1)$. Nonlinear elliptic equations and systems have received a lot of attention in the past forty or fifty years and in this period a lot of progress was made and powerful methods were developed. By comparison, the field of nonlinear hyperbolic equations is wide open. In what follows I will try to point out some recent developments concerning long-time behaviour of smooth solutions to a large class of such equations.

Let us assume that $G(0, 0, 0) = 0$ and that (1) is hyperbolic around the trivial solution $u^0 \equiv 0$. Typically, the operator obtained by linearizing (1) around $u^0 \equiv 0$ contains only second derivatives. Without further loss of generality we may assume it to be the wave operator $\partial_1^2 + \ldots + \partial_n^2 - \partial_t^2 = -\Box$, where we have ascribed to x_{n+1} the role of the time variable t. The equation (1) takes the form (1′)

$$\Box u = F(u, u', u'') \tag{1'}$$

with F a smooth function of (u, u', u''), independent of u_{tt}, vanishing together with all its first derivatives at $(0, 0, 0)$.

Associate to (1′) the pure initial value problem

$$u(x, 0) = \varepsilon f(x), \quad u_t(x, 0) = \varepsilon g(x) \qquad (1'a)$$

with f, g, C^∞-functions, decaying sufficiently fast at infinity (for simplicity, say $f, g \in C_0^\infty(\boldsymbol{R}^n)$) and ε is a parameter which measures the amplitude of the data. Given f, g and F we define the life span $T_* = T_*(\varepsilon)$ as the supremum over all $T \geqslant 0$ such that a C^∞-solution of (1′), (1′a) exists for all $x \in \boldsymbol{R}^n$, $0 \leqslant t < T$. The fundamental *local existence theorem* ([3], [4], [14], [24], [27]), asserts that, if ε is sufficiently small, so that the initial data lies in a neighborhood of hyperbolicity of the zero solution, then $T_*(\varepsilon) > 0$. Moreover, a simple analysis of the proof shows that $T_*(\varepsilon) \geqslant A\dfrac{1}{\varepsilon}$ where A is some small constant, depending only on a finite number of derivatives of F, f, g. This lower bound for T_* is in general sharp if the number of *space dimensions n is equal to one.* Indeed let our variables in (1′) be x and t and $F = \sigma(u_x)u_{xx}$ with σ a smooth function, $\sigma(0) = 0$. An old result of P. Lax [22], extended to systems of wave equations by F. John [10], shows that, under the assumption of "genuine nonlinearity", $\sigma'(0) \neq 0$, all solutions to the corresponding initial value problem (1′a) blow up by the time $O\left(\dfrac{1}{\varepsilon}\right)$. Recently, in [20], it was proved that $T_* < \infty$ even if the genuine nonlinearity condition is violated. More precisely, assume that $\sigma'(0) = \ldots = \sigma^{(P)}(0) = 0$, $\sigma^{(P+1)}(0) \neq 0$ then the corresponding solutions blow up by the time $T = 0(1/\varepsilon^{P+1})$. In both situations the blow-up occurs in the second derivatives of u i.e. u_{xx} becomes infinite while u_t, u_x remain bounded. Such blow-ups are typical of shock formations and are observable phenomena of physical reality. If the original equation, or system can be written in conservation form i.e., in our case, $F(u, u', u'') = \sum_{a=1}^{n+1} \partial_a f^a(u, u')$, one can try to extend the solutions past these breakdown points by introducing the concept of weak solutions. This was successfully accomplished for very general first order systems of conservation laws, in one space dimension, by the fundamental work of Oleinik, P. Lax and J. Glimm (see [24] for a bibliography). In this lecture I will restrict myself, however, to classical, i.e. C^∞-solutions.

Suprisingly, the situation looks better in higher dimensions. In 1976 F. John [9] proved that, under the assumption $F = F(u', u'')$, and $n \geqslant 3$, $T_*(\varepsilon)$ can be significantly improved and, in 1980, S. Klainerman [15] was able to push $T_*(\varepsilon)$ to infinity, and thus obtain global solutions, provided that $n \geqslant 6$. More generally, see ([17], [21], [28]),

THEOREM 1. *Assume that* $F = F(u', u'') = O(|u'| + |u''|)^{p+1}$ *for small* u', u'' *and that* $\frac{1}{p}\left(1+\frac{1}{p}\right) < \frac{n-1}{2}$, *then there exists an* ε_0 *sufficiently small such that for all* $\varepsilon \leqslant \varepsilon_0$, (1′), (1′a) *has a unique smooth solution for all* $x \in \boldsymbol{R}^n$, $t \geqslant 0$.[1]

The reason for this improved behaviour of solutions of (1′) in higher dimensions was beautifully illustrated by F. John [9] with the following quotation from Shakespeare, Henry VI:

"Glory is like a circle in the water,
Which never ceaseth to enlarge itself,
Till by broad spreading it disperse to naught".

Indeed, the higher the dimension the more room for waves to disperse and thus decay. Accordingly, the key to [9], [15], [17], [21] and [28] is to use decay estimates for solutions to the classical wave equation, $\Box u = 0$ (see [25], [26], [31]), and to combine them with energy estimates for higher derivatives of solutions to the original, nonlinear equations.

The dimension $n = 3$, which nature gives preference, is not only the most important but also the most challenging. In [7] F. John exhibited an example for which $T_* < \infty$. More precisely consider $F = u_t \cdot u_{tt}$ and the corresponding equation (1′) in three space dimensions. Imposing only one mild restriction on the data, $\int g(x)\,dx \geqslant 0$, F. John showed that there are no C^2-solutions defined for all $x \in \boldsymbol{R}^3$, $t \geqslant 0$. However, for sufficiently small ε, the solutions remain smooth for an extremely long-time before a breakdown occurs. We have in fact the following very general.

THEOREM 2 (F. John, S. Klainerman). *Assume that* F *verifies one of the following hypothesis*:

(H_1) F *does not depend explicitly on* u *i.e.,* $F = F(u', u'')$

(H_2) F *can be written in conservation form,*

$$F(u, u', u'') = \sum_{a=1}^{4} \partial_a f^a(u, u'),$$

$$(H_2') \quad F(u, u', u'') = \sum_{a=1}^{4} \partial_a f^a(u, u') + O(|u| + |u'| + |u''|)^3$$

for small u, u', u''

[1] The author was recently able to improve this result [32]. The sharp condition which assures global existence is $p > 2/(n-1)$.

Then, there exist three small, positive constants ε_0, A, *depending only on a finite number of derivatives of* F, f, g, *such that for every* $0 \leqslant \varepsilon < \varepsilon_0$, $T_*(\varepsilon) \geqslant \exp(A/\varepsilon)$.

Previously a weaker, polynomially long time existence result, was proved by F. John in [12] using an asymptotic expansion in powers of ε, for u (see also [9]). The exponential long-time existence result was first proved, for spherically symmetric solutions (in the semilinear case $F = F(u')$, by F. John [7] and T. Sideris [29], and for $F = F(u', u'')$ by S. Klainerman [18]).

The result of Theorem 2 is in general sharp. Indeed, F. John [11] proved recently that this is the case in the context of his previous example, $F = u_t u_{tt}$. There is, however, quite a rich class of nonlinearities F for which global existence holds. The following can be regarded as a generalization of Theorem 1 in dimension $n = 3$.

THEOREM 3. *Assume that* F *verifies the following "Null"-condition*

$$\text{(i)} \sum_{a,b=1}^{4} F''_{u_a u_b}(u, u', u'') X^a X^b = O(|u| + |u'| + |u''|)^3,$$

$$\text{(ii)} \sum_{a,b,c=1}^{4} F''_{u_a u_{bc}}(u, u', u'') X^a X^b X^c = O(|u| + |u'| + |u''|)^3,$$

$$\text{(iii)} \sum_{a,b,c,d=1}^{4} F''_{u_{ab} u_{cd}}(u, u', u'') X^a X^b X^c X^d = O(|u| + |u'| + |u''|)^3 \qquad \text{(N)}$$

for every sufficiently small u, u', u'' *and any fixed null space-time vector* (X^1, X^2, X^3, X^4) *i.e.,* $(X^1)^2 + (X^2)^2 + (X^3)^2 - (X^4)^2 = 0$.

Then, if ε *is sufficiently small, a global smooth solution of* (1'), (1'a) *exists.*

To illustrate the content of Theorem 2 note that either of the following examples verifies the condition (N)

Example 1 $F = u_a u_{bc} - u_b u_{ac}$, for any three indices $a, b, c = 1, 2, 3, 4$

Example 2 $F = \partial_a(u_t^2 - u_1^2 - u_2^2 - u_3^2)$, for any index $a = 1, 2, 3, 4$.

The proof of both Theorems 2 and 3 depends on some recent [19] weighted L^∞ and L^1 estimates for solutions to the classical, inhomogeneous, wave equation in dimension $n = 3$. They were first used, in the spherical symmetric case, in [18] and then extended to the general case by introducing the angular momentum operators $\Omega_1 = x_3 \partial_2 - x_2 \partial_3$, $\Omega_2 = x_1 \partial_3 - x_3 \partial_1$, $\Omega_3 = x_3 \partial_1 - x_1 \partial_3$. Their key property is that they commute with the wave operator $\Box$ and thus can be treated as the usual partial deriva-

tives ∂_1, ∂_2, ∂_3. In particular, this allows us to extend the energy estimates used in [9], [15], [17], [21] and [28], to any combination of the derivatives ∂_1, ∂_2, ∂_3, Ω_1, Ω_2, Ω_3. The operators Ω_1, Ω_2, Ω_3 are closely connected to the "radiation operators" L_1, L_2, L_3 which played a fundamental role in [12].

A different, and very interesting proof of Theorem 3, based on some conformal mapping methods, was given by D. Christodoulou [1]. (See also his previous joint work with Y. Choquet-Bruhat [2].)

Both Theorems 2 and 3 have straightforward extensions to systems, in particular to those of the type arising in Nonlinear Elasticity and General Relativity. There are important problems, like that of stability of the Minkowski Space as a solution of the Einstein equations in vacuum, for which we hope that Theorems 2 and 3 could be relevent. In the scalar case we believe that the picture provided by these theorems, together with the nonexistence results of F. John [7], [11] can be completed. In other words, we conjecture that if one of the hypothesis (H1), (H2), (H2′) holds and (N) fails, then the lower bound on $T_*(\varepsilon)$ given by Theorem 2 is sharp, i.e. singularities must develop by that time, for any choice of f or g and ε small. An important open question is to describe the type of blow-up which occurs in that case. If F is quasilinear and verifies H1, we expect that, as for $n = 1$, the breakdown occurs when the second derivatives of u become infinite while the first derivatives remain bounded. The recent work of F. John [28] points in this direction, but completely satisfactory results are still missing. Another open question is to derive results similar to Theorems 2 and 3 for the dimensions $n = 2$ and $n = 4$. We suspect that the corresponding, optimal lower bound for T_* for $n = 2$ must be $O\left(\frac{1}{\varepsilon^2}\right)$ while for $n = 4$ one should be able to prove global existence.[2] In this respect we hope to find decay estimates similar to those of [19] for $n \neq 3$. The same type of questions can be asked for equations (1′) where the wave operator $\Box$ is replaced by the Klein–Gordon operator, $\Box + m^2$ or the Schrödinger operator $-i\partial_t + \Delta$. General results of the type of Theorem 1 were derived in [17], [21], [28], and for nonlinearities depending only on u in [30], (see also the reference there). The methods used to derive Theorems 2 and 3 might be used to substantially improve these results.

In the end, I like to apologize for not mentioning the work of many

[2] See footnote, p. 1211.

other authors. In particular I have left out a lot of interesting results concerning semilinear equations i.e. $F = F(u)$ in (1'). For an up to date bibliography concerning such results I refer to the recent papers of R. Glassey [5], [6].

References

[1] Christodoulou D., *Global Existence to Nonlinear Wave Equations*, in preparation.

[2] Choquet-Bruhat Y. and Christodoulou D., Existence of Global Solutions of the Yang–Mills, Higgs Fields in 4-Dimensional Minkowski Space, I, II, *Comm. Math. Physics* **83** (1982), pp. 171–191, pp. 193–212.

[3] Courant R., Friederichs K. O. and Lewy H., Über die partiellen Differentialgleichungen der mathematischen Physik. *Math. Annalen* **100** (1928), pp. 32–74.

[4] Friederichs K. O., Symmetric Hyperbolic Linear Differential Equations, *Comm. Pure Appl. Math.* **7** (1954), pp. 345–392.

[5] Glassey R., Existence in the Large for $\Box u - F(u)$ in Two Space Dimensions *Math., Z.* **178** (1981), pp. 233–261.

[6] Glassey R., Finite Time Blow-Up for Solutions of Nonlinear Wave Equations, *Math. Z.* **177** (1981), pp. 323–340.

[7] John F., Blow-up for Quasilinear Wave Equations in Three Space Dimensions, *Comm. Pure Appl. Math.* **34** (1981), pp. 29–51.

[8] John F., *Blow-up of Radial Solutions of* $\Box u = \dfrac{\partial F(u_t)}{\partial u_t}$, in preparation.

[9] John F., Delayed Singularity Formation in Solutions of Nonlinear Wave Equations in Higher Dimensions, *Comm. Pure Appl. Math.* **29** (1976), pp. 649–681.

[10] John F., Formation of Singularities in One-Dimensional Nonlinear Wave Propagation, *Comm. Pure Appl. Math.* **27** (1974), pp. 377–405.

[11] John F., *Improved Estimates for Blow-up for Solutions of Strictly Hyperbolic Equations in Three Space Dimensions*, in preparation.

[12] John F., *Lower Bounds for the Life-Span of Solutions of Nonlinear Wave Equations in Three Space Dimensions*, to appear in Comm. Pure Appl. Math., 1983.

[13] John F., Klainerman S., Almost Global Existence to Nonlinear Wave Equations in Three Space Dimensions, *C.P.A.M.*, 1984.

[14] Kato T., Linear and Quasilinear Equations of Evolution of Hyperbolic Type, C.I.M.E. II CICLO, 1976.

[15] Klainerman S., Global Existence for Nonlinear Wave Equations, *Comm. Pure Appl. Math.* **33** (1980), pp. 43–101.

[16] Klainerman S., *Global Existence to Nonlinear Wave Equations in Three Space Dimensions*, in preparation.

[17] Klainerman S., Long-Time Behaviour of Solutions to Nonlinear Evolution Equations, *Arch. Rat. Mech. and Anal.* **78** (1982), pp. 73–98.

[18] Klainerman S. On Almost Global Existence to Nonlinear Wave Equations in Three Space Dimensions, to appear in *Comm. Pure Appl. Math.*, 1983.

[19] Klainerman S., Weighted L^∞ and L^1 Estimates for Solutions to the Equation in Three Dimensions, to appear in *Comm. Pure Appl. Math.*, 1984.

[20] Klainerman S. and Majda A., Formation of Singularities for Wave Equations Including the Nonlinear Vibrating String, *Comm. Pure Appl. Math.* **33** (1980), pp. 241–263.

[21] Klainerman S. and Ponce G., Global Small Amplitude Solutions to Nonlinear Evolution Equations, *Comm. Pure Appl. Math.* **36** (1983), pp. 133–141.

[22] Lax P. D., Development of Singularities of Solutions of Nonlinear Hyperbolic Partial Differential Equations, *J. Math. Phys.* **5** (1964), pp. 611–613.

[23] Lax P., *Hyperbolic Systems of Conservation Laws and the Mathematical Theory of Shock Waves*, C. B. M. S., Monograph No. 11, SIAM, 1973.

[24] Leray J., *Hyperbolic Differential Equations*, Princeton, 1952.

[25] Marshall B., Strauss W. and Wainger S., $L^p - L^q$ Estimates for the Klein–Gordon Equation, *J. Math. Pure Appl.* **59** (1980), pp. 417–440.

[26] Pecher H., L^p-Abschätzungen und klassische Lösungen für nichtlineare Wellengleichungen, *Math. Z.* **150** (1976), pp. 159–183.

[27] Schauder J., Das Anfangswertproblem einer quasilinearen hyperbolischen Differentialgleichung zweiter Ordnung in beliebiger Anzahl von unabhängigen Veränderlichen, *Fundamenta Mathematicae* **24** (1935), pp. 213–246.

[28] Shatah J., *Global Existence of Small Solutions to Nonlinear Evolution Equations*, preprint.

[29] Sideris T., *Global Behaviour of Solutions to Nonlinear Wave Equations in Three Space Dimensions*, preprint, 1982.

[30] Strauss W., Nonlinear Scattering Theory of low energy, *J. Funct. Anal.* **41** (1981), pp. 110–133.

[31] Von Wahl W., L^p-Decay Rates for Homogeneous Wave Equations, *Math. A.*, **120** (1971), pp. 93–106.

Added in proof:

[32] Klainerman S., *Apriori Uniform Decay Estimates for Solutions of the Classical Wave Equation and Applications to Nonlinear Problems*, preprint.

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

ANDREW MAJDA*

Systems of Conservation Laws in Several Space Variables

We describe some recent progress in the short-time existence of discontinuous solutions for the Cauchy problem for an $m \times m$ system of hyperbolic conservation laws in n-space variables,

$$\begin{aligned} &\frac{\partial u}{\partial t} + \sum_{j=1}^{n} \frac{\partial}{\partial x_j} F_j(u) = 0, \quad t > 0, \\ &u(x, 0) = u^0(x), \end{aligned} \tag{1}$$

where $x = (x_1, \ldots, x_n) \in \boldsymbol{R}^n$, $u = {}^t(u_1, \ldots, u_m)$ is an m-vector, and the $F_j(u)$ are smooth nonlinear mappings of $\boldsymbol{R}^m$ to $\boldsymbol{R}^m$ with $A_j(u) = \partial F_j / \partial u$ the corresponding $m \times m$ Jacobian matrices for $1 \leqslant j \leqslant n$. The prototypical example of a system of conservation laws is given by the compressible Euler equations of fluid dynamics, (where $m = n+2$)

$$\begin{aligned} &\frac{\partial \varrho}{\partial t} + \operatorname{div}(\boldsymbol{m}) = 0, \\ &\frac{\partial m_i}{\partial t} + \operatorname{div}\left(\frac{\boldsymbol{m} m_i}{\varrho}\right) + \frac{\partial p}{\partial x_i} = 0, \quad i = 1, \ldots, n, \\ &\frac{\partial E}{\partial t} + \operatorname{div}\left(\boldsymbol{m}\left(\frac{E}{\varrho} + \frac{p}{\varrho}\right)\right) = 0, \end{aligned} \tag{2}$$

expressing conservation of mass, momentum, and total energy. In (2), ϱ is the density with $1/\varrho = \tau$ the specific volume, $\boldsymbol{v} = {}^t(v_1, \ldots, v_n)$ is the fluid velocity with $\varrho \boldsymbol{v} = \boldsymbol{m}$ the momentum vector, p is the scalar pressure,

* Partially supported by N.S.F. Grant #MCS-81-02360 and A.R.O. Grant #483964-25530.

and $E = \frac{1}{2}(\boldsymbol{m}\cdot\boldsymbol{m})/\varrho + \varrho e(\tau, p)$ is the total energy with e the internal energy, a given function of (τ, p) defined through thermodynamic considerations. Other hyperbolic conservation laws often occur in classical physics and engineering — in particular, in describing magneto-fluid dynamics, combusion in certain regimes, shallow water waves, and petroleum reservoir engineering ([1], [10]).

Despite the abundance of concrete problems associated with systems of hyperbolic conservation laws, the rigorous mathematical theory, especially in several space dimensions ($n > 1$), is only beginning. Here we describe the recent theoretical progress regarding discontinuous solutions in several space variables ([8], [9]), we contrast the new phenomena for $n > 1$ with those when $n = 1$, and also we mention some of the interactions of this theory with more concrete applied problems ([11], [12]). To emphasize this concrete point of view, we mostly state these results in the context of the compressible Euler equations in (2). The reader can consult the above papers for the general framework as well as the author's recent lectures ([10]) for a leisurely discussion of many of the topics mentioned briefly below.

For linear hyperbolic equations, jump discontinuities in solutions always follow characteristic hypersurfaces for short enough times before focusing occurs. The most interesting manifestation of the strong nonlinearity in (1) is that unlike the linear case, *jump discontinuities* of (1) *do not* typically *follow characteristic surfaces*. To motivate this, we comment that a piecewise smooth weak solution of (1) which is smooth except for a jump across the space-time hypersurface $S(t)$ with respective sides G_+, G_- and space-time normal $(n_t, n_1, \ldots, n_n)$ necessarily satisfies the quasilinear equations

$$\frac{\partial u_\pm}{\partial t} + \sum_{j=1}^{n} A_j(u_\pm)\frac{\partial u_\pm}{\partial x_j} = 0 \tag{3}$$

in the respective smooth regions $G_\pm$ for u as well as the nonlinear boundary conditions across the hypersurface, S, given by

$$n_t[u]\big|_S + \sum_{j=1}^{n} n_j[F_j(u)]\big|_S = 0, \tag{4}$$

where [ ] denotes the jump across S, i.e., $[u]|_S = (u_+ - u_-)|_S$. In general the highly nonlinear boundary conditions in (4) force the weak solution to jump across noncharacteristic surfaces, $S(t)$, called *shock fronts* (see [1]

for plane wave solutions of (2) and (4) for the elementary theory of genuinely nonlinear plane waves where $S(t)$ is always noncharacteristic).

Next, we describe initial data for (1) where intuitively one would expect that a shock front solution of (1) with the qualitative structure described in (3), (4) would be generated by this initial data for sufficiently short times. We take discontinuous initial data so that there is a smooth initial hypersurface, M, parametrized by α, with two sides Ω_+, Ω_- so that

$$u^0(x) = \begin{cases} u^0_+(x) & \text{for} \quad x \text{ in } \Omega_+, \\ u^0_-(x) & \text{for} \quad x \text{ in } \Omega_-, \end{cases} \tag{5}$$

where $u^0_\pm$ are smooth functions. Given the qualitative structure in (4) anticipated for all small times $t > 0$, it is natural that we also require for this initial data that there is a scalar function, $\sigma(\alpha)$, so that

$$-\sigma(\alpha)[u^0] + \sum_{j=1}^{n} n_j [F_j(u^0)](\alpha) = 0 \tag{6}$$

for all $\alpha \in M$ where $\boldsymbol{n}' = (n_1, \ldots, n_n)$ is the normal to M. Initial data satisfying (5) and (6) are called *shock front initial data.* What conditions are needed in several space variables to guarantee the existence and structural stability of solutions satisfying (3) and (4) with the initial data from (5), (6)? We state the least technical version of the main theorem in [9] specialized to the compressible Euler equations in (2). Before doing this, we introduce three important physical parameters associated with the shock front initial data for (2).

The *normal Mach numbers,*

$$M_\pm(\alpha) = \frac{|v_\pm \cdot n - \sigma|}{C_\pm}(\alpha)$$

with C the speed of sound,

The *compression ratio,* $\mu(\alpha) = \left(\frac{\varrho_-}{\varrho_+}\right)(\alpha)$.

The *Gruneisen coefficient,* Γ_-, measuring the equation of state,

$$\Gamma_- = \left(\varrho_- e_p(\tau_-, p_-)\right)^{-1} > 0.$$

We have

THEOREM. *Assume that the shock front initial data for the compressible Euler equations belong to the Sobolev space,* $H^s(\Omega_\pm)$, $s \geqslant [n/2]+7$ *and satisfy* (6) *as well as the related compatibility conditions up to order* $s-1$

(see [9]). *Asume that the normal Mach numbers satisfy*

$$M_+^2(a) > 1 > M_-^2(a) \quad \text{for all } a \in M \tag{A}$$

and for $n \geqslant 2$ *that the Gruneisen coefficient, compression ratio, and normal Mach numbers also satisfy*

$$(\mu(a)-1)M_-^2(a) < 1/\Gamma(a)+1 \tag{B}$$

for all $a \in M$. *Then for sufficiently short times, there is a* C^2 *hypersurface* $S(t)$ *and* C^1 *functions* $u_\pm$ *defined on the respective sides of this hypersurface satisfying* (3), (4), *and defining a shock front solution of* (2) *with the given initial data. Furthermore, any compressive shock front initial data (i.e.* $\mu(a) > 1$) *for ideal polytropic gases where* $e = p\tau(r-1)^{-1}$, $r > 1$ *automatically satisfies* (A), (B) *and therefore has a shock-front solution.*

We remark here that for $n = 1$, under assumption (A), sharper results are known and a complete theory of the perturbed Riemann problem has been developed in [6], [7]. Also, with the above rigorous theorem, some of the formal calculations in [13] can be justified. The condition in (A) is very natural and corresponds to Lax's geometric entropy conditions in the general case ([14]). The additional condition in (B) for $n \geqslant 2$ might seem at the moment to be a technical restriction; however, the evidence, both physical and mathematical, is overwhelming that when (B) is violated for $n \geqslant 2$, more complex inherently multi-D wave patterns occur rather than shock fronts ([11], [12]). In [5], [15], interesting general geometric entropy conditions always implying (A) are developed. A natural question arises: Does every jump discontinuity for (2) satisfying these general entropy conditions automatically satisfy the inequalities in (B)? The answer is no and the corresponding examples are constructed in [8], [10].

The shock front problem described in (3), (4) can be viewed as a highly nonlinear free boundary value problem for a quasi-linear hyperbolic system since $S(t)$ is non-characteristic and must be determined as part of the solution of the problem. There are three main steps in the proof of the above theorem.

(I) Map to a fixed domain.

(II) Linearization of perturbed shock fronts.

(III) Construction of the shock front solution via a classical iteration scheme.

Without giving any details, we illustrate some of the main points of this proof in the very special case of perturbed steady planar shock fronts

in two space variables. We assume the shock front initial data has the special form

$$u^0(x, y) = \begin{cases} u_+^0 + v_+^0(x, y), & x > 0, \\ u_-^0 + v_-^0(x, y), & x < 0, \end{cases}$$

where $v_\pm^0 \in C_0^\infty(\overline{R^2})$ and $F_1(u_+^0) = F_1(u_-^0)$. For small positive times, the anticipated shock front emanating from $x = 0$ should have the form of a graph, i.e., $S(t)$ can be described by the equation $x = \varphi(y, t)$. We carry out step (I) in this special case by mapping the unknown shock surface $x = \varphi$ onto $x = 0$. Thus, to construct the shock front solution, we need to find functions $u_\pm(x, y, t)$, $\varphi(y, t)$ in a new co-ordinate system still denoted by (x, y, t) satisfying the interior equations,

$$\frac{\partial u_\pm}{\partial t} + \big(A_1(u_\pm) - \varphi_t I - \varphi_y A_2(u_\pm)\big)\frac{\partial u_\pm}{\partial x} + A_2(u_\pm)\frac{\partial u_\pm}{\partial y} = 0$$

$$\text{for} \quad \begin{matrix} x > 0 \\ x < 0 \end{matrix}, \ t > 0, \tag{8}$$

$$u_\pm(x, y, 0) = u_\pm^0 + v_\pm^0, \qquad \varphi(y, 0) = 0$$

and the nonlinear boundary conditions,

$$\varphi_t[u] + \varphi_y[F_2(u)] - [F_1] = 0 \ \text{ on } x = 0, \quad t > 0. \tag{9}$$

Step (II) in the outline of the proof involves linearizing (8), (9) at a typical varying perturbed state and analyzing the associated linear problem. At the special unperturbed state $u_\pm \equiv u_\pm^0$, $\varphi \equiv 0$, the linearized problem for the unknowns $(\tilde{v}_\pm, \tilde{\varphi})$ becomes the constant coefficient boundary value problem,

$$\frac{\partial \tilde{v}_\pm}{\partial t} + A_1(u_\pm^0)\frac{\partial \tilde{v}_\pm}{\partial x} + A_2(u_\pm^0)\frac{\partial \tilde{v}_\pm}{\partial y} = F_\pm \quad \text{for} \quad \begin{matrix} x > 0 \\ x < 0 \end{matrix}, \quad t > 0,$$

$$\tilde{\varphi}_t[u^0] + \tilde{\varphi}_y[F_2(u^0)] - A_1(u_+^0)\tilde{v}_+ - A_1(u_-^0)\tilde{v}_- = g \tag{10}$$

$$\text{for } x = 0, \quad t > 0$$

together with appropriate initial conditions. The boundary conditions in (10) should be regarded as an over-determined evolution equation for the perturbed shock front boundary, $\tilde{\varphi}$, coupled to the boundary values of solutions of hyperbolic equations. A general variable coefficient theory for the problems in (10) is developed in [8]. What estimates define the well-posedness for the mixed problem in (10)? Looking back at the full

nonlinear problem in (8), (9), we see that it is crucial to *gain* a derivative of $\tilde{\varphi}$ beyond the regularity of $\tilde{v}_{\pm}$ to avoid "loss of derivatives" in the nonlinear iteration scheme. Such shock fronts with an associated linearized problem allowing for both this gain in regularity and also a well-posed interior mixed problem are called *uniformly stable* in [8] and admit an algebraic characterization analogous to the uniform Lopatinski condition for standard mixed problems ([3], [14]). Returning to the physical example of linearized shock fronts for the compressible Euler equations, we have the following facts:

PROPOSITION. *For the compressible Euler equations*

I. *Shock fronts are uniformly stable for* $n=1$ *iff* (A) *from the theorem is satisfied.*

II. *Assuming* (A) *shock fronts are uniformly stable for* $n \geqslant 2$ *iff* (B) *from the theorem is satisfied.*

III. *When the inequality*

$$(\mu-1)M_-^2 > 1+M_-/\Gamma_-$$

is satisfied, shock fronts are violently unstable for $n \geqslant 2$ (see [8], [11]).

IV. *In the transition regime between the inequalities from* (B) *of the theorem and* III, *i.e., when*

$$\frac{1}{\Gamma+1} < (\mu-1)M_-^2 < 1+M_-/\Gamma_-$$

causal radiating boundary wave solutions $(\tilde{v}_+, \tilde{v}_-, \tilde{\varphi})$ *for* (10) *exist* (see [11]).

In [11], [12], the special linearized solutions mentioned in IV are used as the starting point of a formal asymptotic expansion which also incorporates nonlinear effects and leads to a theory which predicts the experimentally observed formation of Mach stems in reacting shock fronts — thus, the theorem on stable shock fronts requiring condition (B) for $n \geqslant 2$ is sharp. As regards the general linear problem from (10) in multi-D for the general system (1), we have the following fact:

PROPOSITION. *A necessary condition for any shock front for a system of conservation laws in* $\mathbf{R}^n$ *to be uniformly stable is that the number of equations,* m, *satisfies* $m \geqslant n$.

In particular, in contrast to the case of a single space variable, shock fronts for the scalar conservation law in two space variables,

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}(\tfrac{1}{2}u^2) + \frac{\partial}{\partial y}(\tfrac{1}{2}u^2) = 0$$

are less stable than those for polytropic gases in 2-*D* (see the extended discussion in [10]).

The final main step in the proof of the theorem is the convergence of a nonlinear iteration scheme based on the linearized problems analyzed in part (B). Here the strategy follows that used in the Cauchy problem ([2], [10]) but the technical details are more complex due to both the strong nonlinearity in the boundary conditions and also to the use of square integrable weighted norms in space-time as opposed to maximum norms in time where the linearized problem is well-posed. It is desirable to find a simpler and sharper proof of the convergence than the one given in [9].[1]

The results described here are the only rigorous ones known to the author regarding discontinuous solutions of Multi-*D* conservation laws. Obviously, this is a field in its mathematical infancy and a large number of very interesting open problems remain. The author hopes that this lecture stimulates the interest of other mathematicians in this important subject.

Bibliography

[1] Courant R. and Friedrichs K. O., *Supersonic Flow and Shock Waves*, Wiley-Interscience, New York, 1949.

[2] Kato T., The Cauchy problem for quasi-linear symmetric systems, *Arch. Rat. Mech. Anal.* **58** (1975), pp. 181–205.

[3] Kriess H. O., Initial boundary value problems for hyperbolic systems, *Comm. Pure Appl. Math.* **23** (1970), pp. 277–298.

[4] Lax P. D., Hyperbolic Systems of Conservation Laws and the Mathematical Theory of Shock Waves, *S.I.A.M. Regional Conf. Ser. Appl. Math.* #**11** (1973), Philadelphia.

[5] Liu T. P., The Riemann problem for general systems of conservation laws, *J. Diff. Equations* **18** (1975), pp. 218–234.

[6] Li Da-qian and Yu Wen-ci, Some existence theorems for quasi-linear hyperbolic systems of partial differential equations in two independent variables, II, *Scientia Sinica* **13** (1964), pp. 551–564.

[7] Li Da-qian and Yu Wen-ci, The local solvability of boundary value problems for quasilinear hyperbolic systems, *Scienta Sinica* **23** (1980), pp. 1357–1367.

[8] Majda A., The stability of multi-dimensional shock fronts, *Memoirs A.M.S.* #**275**, January 1983.

[9] Majda A., The existence of multi-dimensional shock fronts, *Memoirs A.MS.* #**281** May 1983.

[1] A quite different but much simpler proof for the shock-front theorem for 2nd order wave equations has been given in [16],

[10] Majda A., *C. I. M. E. Lectures on Compressible Fluid Flow and Systems of Conservation Laws in Several Space Variables* (to appear in Springer-Verlag Applied Math. Science Series).

[11] Majda A. and Rosales R., A theory for spontaneous Mach stem formation in reacting shock fronts, I: the basic perturbation analysis, (to appear in *S.I.A.M. J. Appl. Math.* **43** (1983), pp. 1310–1334.

[12] Majda A. and Rosales R., A theory for spontaneous Mach stem formation in reacting shock fronts, II: the evidence for breakdown, (to appear in *Studies in Appl. Math* in 1983).

[13] Maslov V. P., Propagation of shock waves in an isentropic nonviscous gas, *J. Sov. Math.* **13** (1980), pp. 119–163.

[14] Sakamoto R., Mixed problems for hyperbolic equations, I, II, *J. Math. Kyoto Univ.* **10** (1970), pp. 349–373 and pp. 403–417.

[15] Wendorff B., The Riemann problem for materials with nonconvex equations of state II. General flow, *J. Math. Anal. Appl.* **38** (1972), pp. 649–658.

Added in proof:

[16] Majda A. and Thomann E., *Multidimensional shock fronts for second order wave equations* (to appear),

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CA 94720 USA

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

V. E. ZAKHAROV

Multidimensional Integrable Systems

Introduction

The discovery and successful development of exact solving methods for certain nonlinear partial differential equations (known as the Inverse Scattering Problem methods or ISP method) resulted in a change of our point of view on the subject of mathematical physics. Instead of equations of general form, specific, exeptional equations with strong intrinsic symmetry began to play the main role. The most known examples of such equations are the following:

1. The KdV equation

$$U_t + 6UU_x + U_{xxx} = 0. \tag{I.1}$$

2. The nonlinear Schrödinger equation

$$iU_t + U_{xx} \pm |U|^2 U = 0. \tag{I.2}$$

3. The Sine–Gordon equation

$$U_{tt} - U_{xx} + \sin U = 0. \tag{I.3}$$

These equations arise naturally in many physical problems and are of universal nature. They are Hamiltonian systems and their exceptional character lies, in particular, in the fact that they are completely integrable in the sense of classical mechanics, i.e., they have a full set of motion integrals I_n ($n = 1, \ldots, \infty$) whose Poisson brackets are equal to zero.

All known equations to which the ISP method can be applied are Hamiltonian systems and have an infinite set of commuting motion integrals, therefore we shall call them integrable. It is not rigorous because the completeness of sets of integrals which is necessary for integrability in a strict sense is proved only in rare cases (in particular for equations (I.1)–(I.3)).

All nonlinear equations to which the ISP method can be applied are the compatibility conditions for some overdetermined systems of linear equations. In most cases such systems can be reduced to the form (see [41], [42])

$$\Psi_x = U\Psi, \quad \Psi_t = V\Psi. \tag{I.4}$$

Here $\Psi = \Psi(x, t, \lambda)$ is a nondegenerated $(N \times N)$-matrix function, and U and V are $(N \times N)$-matrix functions of variables x and t which are rationally dependent on the complex parameter λ. The compatibility condition for overdetermined system (I.4) is of the form

$$U_t - V_x + [U, V] = 0. \tag{I.5}$$

Fixing the positions of the poles of functions U, V and specifying a certain number of reductions (some a priori relations for the coefficients of these functions) we obtain a special system of nonlinear equations, i.e., equations (I.1)–(I.4). Within infinite set of such systems a few dozen physically significant ones have been studied in some more detail.

The most general method to investigate systems of type (I.5) was developed in [41], [42]. It is based on some algebraic properties of solutions of the matrix Riemann problem on the complex plane and it enables us (as a minimum) to construct infinite sets of exact solutions of systems of type (I.5). This method is called "the dressing method".

If the variable t has the meaning of time, systems like (I.5) describe one-dimensional dynamical processes. It is important from physical point of view to study such processes in two, three or more dimensions. All the systems for which space dimension d is greater then or equal to 2, we shall call multidimensional. This report is devoted to multidimensional integrable systems.

Each of such systems, when restricted to a straight line, generates a one-dimensional integrable system; therefore multidimensional systems can be regarded as a result of multidimensional generalization of one-dimensional ones. This point of view dominates here.

The first, fairly effective method of multidimensional generalization, was formulated in 1974 ([40]). In particular, it was shown that the Kadomtzev–Petviqshvily equation (K–P)

$$\frac{\partial}{\partial x}(U_t + 6UU_x + U_{xxx}) = \pm \frac{\partial^2 U}{\partial y^2} \tag{I.6}$$

(which had been known since 1970) is an integrable multidimensional generalization of the KdV equation (see [1], [12]–[14], [18]–[21], [23],

[25], [26], [32], [38]). A few papers ([10], [16], [17], [33]), were devoted to "three waves" system

$$\begin{aligned}
&\frac{\partial U_1}{\partial t} + (\vec{V}_1, \nabla U_1) = iU_2 U_3,\\
&\frac{\partial U_2}{\partial t} + (\vec{V}_2, \nabla U_2) = iU_1 U_3^*,\\
&\frac{\partial U_3}{\partial t} + (\vec{V}_3, \nabla U_3) = iU_1 U_2^*
\end{aligned} \tag{I.7}$$

(here $\vec{V}_1$, $\vec{V}_2$, $\vec{V}_3$ are arbitrary two-dimensional vectors) which is a multidimensional generalization of the case where the vectors are parallel. This case corresponds to a well-known system of type (I.5) (see [15], [36]). A multidimensional generalization of equation (I.2), namely the system

$$\begin{aligned}
&iU_t + U_{xx} + \alpha U_{yy} \pm 2|U|^2 U + VU = 0,\\
&\left(\frac{\partial^2}{\partial x^2} - \alpha \frac{\partial^2}{\partial y^2}\right) V = \pm 2 \frac{\partial^2}{\partial x^2} |U|^2,
\end{aligned} \tag{I.8}$$

is known as the Davey–Stewardson system ([11]). There are other examples of multidimensional generalizations. One of the main systems of type (I.5) is the equation of a "principal chiral field"

$$\frac{\partial}{\partial t}(g^{-1} g_t) = \frac{\partial}{\partial x}(g^{-1} g_x). \tag{I.9}$$

Here g is an element of an arbitrary Lie group G. The equation (I.9) admits the following integrable multidimension generalization:

$$\frac{\partial}{\partial \bar{w}}(g^{-1} g_t) = \frac{\partial}{\partial y}(g^{-1} g_x) \tag{I.10}$$

where y and $\bar{w}$ are new independent variables.

In the particular case where $\bar{w} = -\bar{t}$, $Y = \bar{x}$ and g can be represented in the form $g = f^+ f$, the equation (I.10) is equivalent to the known self-duality equation ([5])

$$F_{\mu\nu} = \pm \tfrac{1}{2} \varepsilon_{\mu\nu\alpha\beta} F_{\alpha\beta}. \tag{I.11}$$

Here $F_{\mu\nu}$ is the Yang–Mills field with the unitary group. In the other particular case ([27]) where $\bar{w} = t$, $y = \bar{x}$, the equation (I.10) is the only known relativistically invariant classical model of field theory on the plane. Some other interesting integrable systems are also known.

The purpose of this report is not to describe the properties of particular systems but to clarify the principles of multidimensional generalization of one-dimensional systems of type (I.5).

There are two essentially different ways of such generalization. Equations (I.6)–(I.8) are of the first type; equation (I.10) of the second type. We discuss also the rule of obtaining of large classes of exact solutions of the multidimensional systems under consideration.

1. The Riemann problem and the "dressing method"

Let us describe briefly the "dressing method" for the systems of the type (I.4), (I.5).

$$U = \frac{L_1(\lambda, x, t)}{m_1(\lambda, x)}, \qquad V = \frac{L_2(\lambda, x, t)}{m_2(\lambda, t)}. \tag{1.1}$$

Here

$$\begin{aligned} L_1 &= l_1^0(t, x)\lambda^{p_1} + l_1^1(t, x)\lambda^{p_1-1} + \ldots, \\ L_2 &= l_2^0(t, x)\lambda^{p_2} + l_2^1(t, x)\lambda^{p_2-1} + \ldots \end{aligned} \tag{1.2}$$

are the polynomials of degrees p_1, p_2 with matrix coefficients $l_{1,2}^i(x, t)$ which are unknown functions, and m_1, m_2 — are the polynomials of degrees q_1, q_2 ($q_1 \leqslant p_1$, $q_2 \leqslant p_2$) with given scalar coefficients.

Substituting (1.1) into (I.4), (I.5), grouping like terms and putting the coefficients for all degrees of λ equal to zero, we obtain a system of p_1+p_2+1 nonlinear equations for p_1+p_2+2 coefficients of polynomials $L_{1,2}$. One equation is lacking and this is explained by the invariance of the system under gauge transformations of the form

$$\begin{aligned} \Psi &\to \tilde{\Psi} = f\Psi, \\ U &\to \tilde{U} = fUf^{-1} + f_x f^{-1} \\ V &\to \tilde{V} = fVf^{-1} + f_t f^{-1}, \end{aligned} \tag{1.3}$$

where f is an arbitrary and λ-independent matrix function with $\det f \neq 0$.

The main idea of the "dressing" method is to enlarge the gauge freedom due to λ-dependent functions f. Let functions f, f^{-1} be analytic out of the contour Γ. Let f_1, f_2 be the boundary values of the functions f at the opposite sides of the contour and let

$$f_1^{-1} f_2 = F.$$

Functions $\tilde{U}$, $\tilde{V}$ (as compared with U, V) have in general a jump on the contour Γ. One can easily prove that there is no jump if the transition function F satisfies the equations

$$F_x = [U, F], \quad F_t = [V, F]. \tag{1.4}$$

The common solution of these equations is of the form

$$F = \Psi F_0(\lambda)\Psi^{-1}.$$

Here $F_0(\lambda)$ is an arbitrary matrix function defined on Γ, and Ψ is some compatible solution of (I.4) given on Γ. The variation of Ψ is reduced to re-defining of F. Knowing F; one can find a gauge function f by solving the Riemann problem and determine a function with values on the sides that meet each other related by the equation

$$f_2 = f_1 F. \tag{1.5}$$

Freedom in determination of f (any given f can be replaced by $g(x,t)f$ where g is an arbitrary non-degenerated matrix) is connected with the gauge invariance. If we additionally require $f|_{\lambda=\infty} = 1$, then the problem (1.5) can be reduced to the singular integral equation

$$\varrho(\lambda) = \left\{1 + \frac{1}{2\pi i}\int_\Gamma \frac{\varrho(\xi)}{\lambda - \xi + i0}\, d\xi\right\} T(\lambda), \tag{1.6}$$

where

$$T = F - 1, \quad \varrho = f_1 - f_2.$$

So the method described above permits using the Riemann problem to proliferate the solutions of system (I.4). The solution $\tilde{U}$, $\tilde{V}$ is called the "dressing" of solutions U, V by means of function F.

For the initial (background) solutions U, V, which are supposed to be known, one can take, for example

$$U = U(\lambda, x), \quad V = V(\lambda, t), \quad [U, V] = 0. \tag{1.7}$$

This "dressing" method is especially effective if the gauge function f is a rational function of λ. In this case, instead of the Riemann problem there arises a finite system of linear algebraic equations and the result of dressing (of the trivial background (1.7)) is called the "soliton solution".

2. The second method of multidimensional generalization

The simplest way to increase the number of space variables in the systems discussed above is called here the second method only for historical reasons. In a clear form it was formulated in [6], [7] in 1977, though the particular

cases had been known earlier ([8], [9], [37]). The "second method" is based on the fact that the relation between the number of connections between the functions $l^i_{1,2}(x, t)$ and the number of those functions does not change if we transform L_1, L_2 into the first order differential operators

$$l^i_{1,2}(x, t) \to l^i_{1,2}(x, t) + \partial^i_{1,2}. \tag{2.1}$$

Here $\partial^i_{1,2}$ means differentiation with respect to new independent variables. The whole Riemann problem scheme extends this case. Let us note that, after transformation (2.1), the operators $\partial^i_{1,2}$, ∂_x and ∂_t became equivalent. Multiplying equations (I.4) by $m_1(\lambda)$, $m_2(\lambda)$ respectively, one can reduce them to the form

$$D_1\Psi = 0, \quad D_2\Psi = 0, \tag{2.2}$$

$$D_1 = \sum_{k=0}^{p_1} (\partial_{k,1} + \tilde{l}_1^{(k)})\lambda^k,$$

$$D_2 = \sum_{k=0}^{p_1} (\partial_{k,2} + \tilde{l}_2^{(k)})\lambda^k.$$

The compatibility condition (I.5) takes the form

$$[D_1, D_2] = 0.$$

Generally speaking, all the differentials $\partial_{k,1}$; $\partial_{k,2}$ are independent and their total number is equal to $p_1 + p_2 + 2$. In the simplest case $p_1 = p_2 = 1$ this number is equal to 4, but there may be linear relations between differentials expressed by basic differentials $\tilde{\partial}_l$:

$$\partial_{k,1} = \sum_{l=1}^{N} A^l_k \tilde{\partial}_l, \quad \partial_{k,2} = \sum_{l=1}^{N} B^l_k \tilde{\partial}_l. \tag{2.3}$$

If $N = 2$, we have a two-dimensional problem. If we put

$$\begin{aligned} D_1 &= (\partial_1 + B_1)\lambda + (-\bar{\partial}_2 + B_2^+), \\ D_2 &= (\partial_2 + B_2)\lambda - (-\bar{\partial}_1 + B_1^+) \end{aligned} \tag{2.4}$$

where

$$\partial_1 = \frac{\partial}{\partial x_1} + i\frac{\partial}{\partial x_2}, \quad B_1 = A_1 + iA_2,$$

$$\partial_2 = \frac{\partial}{\partial x_3} + i\frac{\partial}{\partial x_4}, \quad B_2 = A_3 + iA_4$$

(x_i — are the coordinates in R^4, A_i — Yang–Mills potentials in antihermitian Lie algebra), then equations (2.3) coincide with self-duality equations (I.11) (see [5]–[7]).

The self-duality equations can be considered only in a space of an even signature.

The elliptic case (0.4) is investigated best of all, in this case the formulation of the problem is quite different from that given in the evolution case. Representation (2.4) permits easy construction of local solutions of self-dual equations with strong singularities. The construction of global regular (multiinstanton) solutions was solved in principle (though not very effectively) in the well-known papers [3], [4] by means of a method different from the ISP method. Using ISP method, we can obtain for the present only a part of the multiinstanton solutions, but in an explicit form. The case of a signature (2.2) has no direct physical meaning, but if we suppose that there is no dependence on any of the coordinates, we obtain the only hither to known relativistically invariant integrable system in the plane [27]. In this case it is convenient to write down the operators $D_{1,2}$ in the form

$$D_1 = \partial_t + \lambda\partial_{\bar{z}} + A, \tag{2.5}$$

$$D_2 = \lambda^2\partial_{\bar{z}} + \lambda A - \partial_z - B. \tag{2.6}$$

Here $A = g^{-1}g_t$, $B = g^{-1}g_z$,

$$\frac{\partial}{\partial t}\left(g^{-1}g_t\right) = \frac{\partial}{\partial \bar{z}}\left(g^{-1}g_z\right).$$

It is evident that this system is invariant with respect to rotations in a plane $z = (x_1, x_2)$. When performing the Lorentz transforms

$$\partial_t = \gamma(\partial_t + \beta\partial_1), \qquad \partial_1 = \gamma(\partial_1 + \beta\partial_t)$$

where

$$-1 < \beta < 1, \qquad \gamma = \frac{1}{\sqrt{1-\beta^2}},$$

we substitute

$$g \to \chi^+ g\chi$$

where

$$\chi(t, z, \bar{z}) = \Psi(t, z, \bar{z}, \lambda)|_{\lambda=(1-\gamma)/\beta\gamma}. \tag{2.7}$$

Formulas (2.7) and (2.8) specify an action of the Lorentz group on (I.10) which cannot be described in terms of linear representations. It is inter-

esting to note that the scattering problem for equation (2.5) coincides with the non-abelian Radon transformation the specifying the integrals of ordered exponential functions of the type $Te^{\int A dS}$ along all lines in the plane (x_1, x_2). Using the Riemann problem, we obtain also the solution of the inversion of this transform.

On the whole, the second method for multidimensional generalization is not sufficiently developed. It can easily be extended to supersymmetric systems and the activity in this direction seems promising (see the recent papers [30], [31]).

3. Polynomial reductions and general covariance

Even for small p_1, p_2 the system (I.5) has an extremely high number of unknown functions, therefore it is important to reduce this number. One can do this by posing additional relations compatible with the system. Those algebraic or differential relations are known as reductions.

Progress in this field was achieved in [28] where it was shown that important classes of reductions form a finite group.

There are also other reductions, which cannot be reduced to those described in [28].

Let U and V in (1.1) be of the form

$$U = M_1^{-1}(\lambda)L_1(\lambda), \qquad V = M_2^{-1}(\lambda)L_2(\lambda). \tag{3.1}$$

Here M_1, M_2 are the polynomials of λ, and their coefficients are the matrix $N \times N$ functions of x and t. U and V can be represented in the form (1.1) $M_{1,2}$ ($\overline{m}_{1,2} = \det M_{1,2}$) but the degree of $L_{1,2}$ increases N times and this means a reduction. Now equations (I.4) take the form

$$M_1 \frac{\partial \Psi}{\partial x_1} = L_1 \Psi, \qquad M_2 \frac{\partial \Psi}{\partial t} = L_2 \Psi. \tag{3.2}$$

Let

$$\begin{aligned} M_1(\lambda) &= \lambda^{q_1} + m_1^1(t, x)\lambda^{q_1-1} + \ldots, \\ M_2(\lambda) &= \lambda^{q_2} + m_2^1(t, x)\lambda^{q_2-1} + \ldots \end{aligned} \tag{3.3}$$

(as before, $q_1 \leqslant p_1$ and $q_2 \leqslant p_2$).

Consider the equation

$$R_2 M_1 - R_1 M_2 = 0. \tag{3.4}$$

If we look for $R_{1,2}$ in the form (2.5) the coefficients of the polynomials can be determined uniquely.

Applying to the first of equations (3.2) the operator $R_2\frac{\partial}{\partial t}$ and to the second — the operators $R_1\frac{\partial}{\partial x}$ and subtracting the first one to the second we obtain, due to (3.4), the first order operator. From (3.2) we then obtain

$$R_2\frac{\partial M_1}{\partial t}+R_1L_2=AM_1,\quad R_1\frac{\partial M_2}{\partial x}+R_2L_1=BM_2 \tag{3.6}$$

where we look for polynomials A, B in the form

$$\begin{aligned}A&=a_0\lambda^{p_2}+a_1\lambda^{p_2-1}+\dots,\\ B&=b_0\lambda^{p_1}+b_1\lambda^{p_1-1}+\dots\end{aligned} \tag{3.7}$$

Simple calculations show that in (3.4) and (3.6) the total number of the relations, which does not depend on λ, is equal to $2(p_1+p_2)+q_1+q_2+3$ and hence to the total number of unknown functions minus one. So system (3.4)–(3.6) is determined up to gauge transformations, which are now of the form

$$\tilde{M}_{1,2}=fM_{1,2}f^{-1},\quad \tilde{L}_1=fL_1f^{-1}-fM_1(f^{-1})_x,\quad \tilde{L}_2=fL_2f^{-1}-fM_2(f^{-1})_t$$

where f is an arbitrary matrix function. We can extend the "dressing" method to systems (3.4)–(3.6). As before, it will be a generalization of gauge transformation. When the "dressing" transformation $\Psi\to f(x,t,\lambda)\Psi$ takes place equations (3.2) must be multiplied from the left by the functions $g(x,t,\lambda)$ and $h(x,t,\lambda)$.

Let functions $f, f^{-1}, g, g^{-1}, h, h^{-1}$ be analytic except on the contour Γ, with the following boundary conditions

$$f_2=f_1F,\quad g_2=g_1G,\quad h_2=h_1H, \tag{3.8}$$

and the transition functions F, G, H satisfying

$$\begin{aligned}M_1F&=GM_1, & M_2F&=HM_2,\\ M_1F_x&=GL_1-L_1F, & M_2F_t&=HL_2-L_2F.\end{aligned} \tag{3.9}$$

Then the functions

$$\begin{aligned}\tilde{M}_1&=gM_1f^{-1}, & \tilde{M}_2&=hM_2f^{-1},\\ \tilde{L}_1&=gL_1f^{-1}-gM_1(f^{-1})_x & \tilde{L}_2&=hL_2f^{-1}-hM_2(f^{-1})_t\end{aligned} \tag{3.10}$$

are polynomials of and they are of the same form as $M_{1,2}$, $L_{1,2}$ and generate the compatible system (3.2) for the function Ψ. The polynomials $R_{1,2}$,

A, B are transformed according to the formulae similar to (3.7). Note that equations (3.6) can be solved uniquely by means of a single arbitrary function $F_0(\lambda)$ specified at the contour

$$F = \Psi F_0 \Psi^{-1}, \quad G = M_1 \Psi F_0 (M_1 \Psi)^{-1},$$
$$H = M_2 \Psi F_0 (M_2 \Psi)^{-1}. \tag{3.11}$$

Thus, dressing the system (3.4)–(3.2) can be generalized to the multidimensional case of k variables $x_1, \ldots, x_k$ by taking into account the following over determined system

$$M_{ij} \frac{\partial \Psi}{\partial x_j} = L_i \Psi. \tag{3.12}$$

Here M_{ij}, L_i are polynomials of λ.

The construction of compatibility conditions for the system (3.11) is a straightforward generalization of the procedure described above. At the first stage we solve the system of equations for the polynomial matrix R_{ik} of λ

$$R_{ik} M_{kj} + R_{jk} M_{ki} = 0. \tag{3.15}$$

Then, by applying the differential operator $R_{ik} \frac{\partial}{\partial x_i}$ to (3.9) we equate the resulting differential first order operator to zero. We have

$$R_{ik} \frac{\partial M_{kj}}{\partial x_i} - R_{jk} L_k = T_k M_{kj}, \tag{3.16}$$

$$R_{ik} \frac{\partial L_k}{\partial x_i} + T_k L_k = 0. \tag{3.17}$$

The dressing of the system (3.10)–(3.12) is done by means of the generalized gauge transformation $\Psi \to f\Psi$ and by multiplication of (3.12) from the left by the matrix $S_{ik}(x)$. The function $f(\lambda)$ and the matrix $S_{ik}(\lambda)$ are obtained from the solution of the Riemann problems

$$S_{ik}^{(2)} = S_{il}^{(1)} G_{lk} \tag{3.18}$$

at the contour Γ, where the transition functions G_{lk} and F satisfy the linear equations

$$M_{ik} F = G_{il} M_{lk},$$
$$M_{ik} \frac{\partial F}{\partial x_k} = G_{il} L_l - L_i F. \tag{3.19}$$

These equations are satisfied if we take F and G_{ik} in the form

$$F = \Psi F_0 \Psi^{-1}, \quad G_{ik} = M_{il}\Psi F_0 (M_{lk}\Psi)^{-1}.$$

As before, the dressing is performed by means of a single matrix function $F(\lambda)$.

Let L_i be polynomials of the same degree P, and M_{ij} — polynomials of the same degree $q \leqslant p$. Then system (3.9) (being in general covariant), is also invariant with respect to any diffeomorphism group in R: $x_i = x_i(y_1, \ldots, y_k)$. As yet, we do not know any physical applications of systems of types (3.4), (3.6), (3.17), (3.19) but we hope they will be found soon. Let us also note that system (3.6) is greatly simplified if $M_2 = 1$. Then $R_2 = 1$, $R_1 = M$, $B = L_1$ and for M_1 and L we have

$$\begin{aligned} &\frac{\partial M_1}{\partial t} + M_1 L_2 = A M_1, \\ &\frac{\partial L_1}{\partial t} - M_1 L_{2x} = A L_1 - L_1 L_2. \end{aligned} \tag{3.20}$$

4. The first multidimensional generalization

Consider the system (I.4) assuming that U and V are polynomials of λ. By performing the Fourier transformation with respect to λ (replacing $\lambda \to i\frac{\partial}{\partial S}$) equation (I.5) changes into a relation between differential operators with S-independent coefficients. However, there would be no essential difference if we assumed that the coefficients of the operators U and V depend on S. The number of equations and unknown functions would be as in the previous case. This fact, which was clarified in 1974 ([40]), became the basis of the method which we call the first method for multidimensional generalization. Equations (I.6)–(I.8) are of type (I.5), where U and V are differential operators. In particular, the K–P, equation (I.6) becomes a compatibility condition for the following system of linear equation

$$\begin{aligned} &\alpha^2 \frac{\partial \Psi}{\partial y} = \frac{\partial^2 \Psi}{\partial x^2} + U\Psi, \quad \alpha^2 = \pm 1, \\ &\frac{\partial \Psi}{\partial t} = \frac{\partial^3 \Psi}{\partial x^3} + \frac{3}{2} U \frac{\partial \Psi}{\partial x} + W\Psi. \end{aligned} \tag{4.1}$$

The question arises under what conditions the overdetermined system of linear partial differential equations

$$p_i\left(\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_k}\right)\Psi = 0, \quad i = 1, \ldots, N, \tag{4.2}$$

generates (as a compatibility condition) an integrable system of nonlinear, physically significant equations. The polynomial reductions obtained above answer this question partially.

Consider (instead of (3.2)) equations of the form

$$M_1\left(i\frac{\partial}{\partial S}\right)\frac{\partial \Psi}{\partial x} = L_1\left(i\frac{\partial}{\partial S}\right)\Psi,$$

$$M_2\left(i\frac{\partial}{\partial S}\right)\frac{\partial \Psi}{\partial t} = L_2\left(i\frac{\partial}{\partial S}\right)\Psi. \tag{4.3}$$

One can see that the procedure of deriving compatibility conditions would not differ from that outlined in Section 3. As before, the compatibility conditions would be of the form (3.4), (3.6) where M_i, L_i, A and B would be linear differential operators.

The nonlinear systems arising as a compatibility condition for the linear equations (4.3) can be viewed as the first multidimensional generalization of general rational systems (I.4). In the particular case where $M_2 = 1$, the system (4.3) can be rewritten in the form

$$\mathscr{L}\Psi = M_1\frac{\partial \Psi}{\partial x} - L_1\Psi = 0, \quad \Psi_t = A\Psi. \tag{4.4}$$

Now, its compatibility conditions are of form (3.2), and can be written as

$$\frac{\partial \mathscr{L}}{\partial t} + [\mathscr{L}, A] = B\mathscr{L} \tag{4.5}$$

where

$$B = L_1 - A$$

One can say that equations of type (4.4) have an "L–A–B triad". Equations of this type occurred and were studied in [34], and general equations of type (4.3) — in [35]. Some examples of equations of type (4.5) have been known since 1976 ([22]).

We do not give the explicit form of these equations because we do not know yet any examples of their physical applications. Equations of type (4.3) can be a source of new reductions for one-dimensional systems. For

example, we can assume that the coefficients $M_{1,2}$, $L_{1,2}$ do not depend on x. Then, after replacing $\frac{\partial \Psi}{\partial x} \to i\lambda\Psi$, we have a non-trivial reduction of the scheme (I.4).

In contrast to the one-dimensional situation, equations of type (4.3) have the meaning even in the case of scalar coefficients of the operators $M_{1,2}$, $L_{1,2}$. Let those coefficients be constant. Then the Fourier transformation $\Psi \simeq e^{i(px+qt+\lambda S)}$ leads to the equations

$$p = \frac{l_1(\lambda)}{m_1(\lambda)}, \quad q = \frac{l_2(\lambda)}{m_2(\lambda)} \tag{4.6}$$

where $l_{1,2}$, $m_{1,2}$ are the symbols of the operators $L_{1,2}$, $M_{1,2}$. Relations (4.6) are a rational parametrization of some algebraic curve of genus zero. On view of this, we call equations (4.3) elementary rational equations (also in the case of variable coefficients). Let us consider a system composed of a finite number of elementary rational equations

$$M_i \frac{\partial \Psi}{\partial x_i} = L_i \Psi. \tag{4.7}$$

By applying to system (4.7) a finite number of differential and algebraic operations, one can eliminate from it all the derivatives of Ψ with respect to S. This results in a system of type (4.2), its coefficients being uniquely expressed by the coefficients of system (4.7) and a finite number of their derivatives with respect to x_i and S. Then, denote their derivatives of all orders with respect to S by new symbols. We say that system (4.8) obtained in this way admits partial rational parametrization. If there are no derivatives with respect to S among the coefficients of this system, we say that it admits total (full) rational parametrization. The compatibility conditions for overdetermined systems admitting the total rational parametrization are well determined non-linear integrable systems. The overdetermined but compatible solutions of a parametrizing system of equations (4.7) are the compatibility conditions of systems admitting partial parametrization. Only very special systems of type (4.2) admit even partial parametrization. System (4.2) of the general form with variable coefficients has no compatible solutions. If a rationally parametrized (or even partially parametrized) system of type (4.2) has constant coefficients, the Fourier transformation $\Psi \sim e^{i(p_i x_i)}$ changes it to a system of polynomials describing an algebraic curve in the space C^k, with at least

one rational component. The Fourier transformation of system (4.7) determines its rational parametrization.

Recently a number of papers have appeared where systems of type (I.4) with functions U and V rational on an algebraic curve of genus one ([24], [39]) and also of an arbitrary genus are considered. The "dressing" of such systems is performed by means of a solution of the Riemann problem on a corresponding curve. One can assume that such systems admit also the first multidimensional generalization. It should be formulated in terms of systems of type (4.2), which do not admit even partial rational parametrization, but generate compatible (though underdetermined) non-linear integrable systems. To classify such systems is a problem for the future.

The Fourier transformation of equations with constant coefficients should generate an algebraic curve of a finite genus.

Systems (3.15)–(3.17) also admit the first multidimensional generalization. To see this it is enough to assume that M_{ij} and L_i are differential operators with respect to S. All the above statements concerning generally covariant systems remain valid in this case, expect for the situation where operators M_{ij} and L_i have scalar coefficients. This case is physically significant. The simplest generally covariant system of type (3.13) is of the form

$$\frac{\partial^2 \Psi}{\partial x^i \partial S} + U_i^j \frac{\partial \Psi}{\partial x_j} = \frac{\partial \Phi}{\partial x^i} \Psi \tag{4.8}$$

with an unknown second order tensor of mixed covariance U_i^j and an unknown scalar Φ. By differentiation (4.8) with respect to x_k and rearrangement we obtain the following system of type (4.2):

$$\frac{\partial}{\partial x^k}\left(U_k^j \frac{\partial \Psi}{\partial x^j}\right) - \frac{\partial}{\partial x^i}\left(U_k^j \frac{\partial \Psi}{\partial x^j}\right) = \frac{\partial \Phi}{\partial x^i}\frac{\partial \Psi}{\partial x^k} - \frac{\partial \Phi}{\partial x^k}\frac{\partial \Psi}{\partial x^i} \tag{4.9}$$

which has no derivatives with respect to S and admits a full rational parametrization. As yet, nobody has managed to obtain a compact form of the equation for U_j^i and Φ or to find any physical interpretation.

5. The "dressing" procedure and the non-local Riemann problem

The "dressing" procedure for the first multidimensional generalizations of type (I.4) was developed in [40] for the particular case of V and U being polynomials. In [34], [35] this was extended to the case of general rational

systems of type (I.4). Let us describe this procedure for systems of type (4.7). Let M_i, L_i be some differential operators for which system (4.7) is compatible and its compatible solution Ψ depends on the parameter λ. Then the system

$$\frac{\partial}{\partial x_i}\tilde{\Psi} = -\tilde{\Psi}M_i^{+(-1)}L_i^+ \tag{5.1}$$

is also compatible. Here M_i^+, L_i^+ are operators conjugated to M_i, L_i. If $M_i(\lambda)$ and $L(\lambda)$ are polynomials then $\tilde{\Psi} = \Psi^{-1}$. We construct the operators F and Q_i. Integral with respect to S and having kernels $F(S, S', x_i)$ and $Q_i(S, S', x_i)$ of the form

$$\begin{gathered} F = \int \Psi(S, x_i, \lambda)F_0(\lambda, \lambda')\tilde{\Psi}(S', x_i, \lambda')\,d\lambda\,d\lambda', \\ Q_i = \int M_i\Psi(S, x_i, \lambda)F_0(\lambda, \lambda')g_i(S', x_i, \lambda')\,d\lambda\,d\lambda', \\ M_i g_i = \tilde{\Psi}. \end{gathered} \tag{5.2}$$

It is easy to verify in a straightforward way that the kernels F, Q_i in (5.2) satisfy the differential equations

$$\begin{gathered} M_i\frac{\partial F}{\partial x_i} = L_iF - Q_iL_i^+, \\ M_iF = Q_iM_i^+, \end{gathered} \tag{5.3}$$

which are equivalent to the operator relation

$$(1+\hat{Q}_i)\left(M_i\frac{\partial}{\partial x} - L_i\right) = \left(M_i\frac{\partial}{\partial x} - L_i\right)(1+\hat{F}). \tag{5.4}$$

Let us consider triangle factorizations of the operators $1+\hat{F}$, $1+\hat{Q}_c$:

$$\begin{gathered} (1+\hat{K}^+)^{-1}(1+\hat{K}^-) = 1+\hat{F}, \\ (1+\hat{P}_i^+)^{-1}(1+\hat{P}^-) = 1+\hat{Q}_i, \end{gathered} \tag{5.5}$$

which are equivalent because of their kernel, to the Marchenko's equations

$$F(S, S', x_i) + K^+(S, S', x_i) + \int_S^\infty K^+(S, S'', x_i)F(S'', S', x_i)\,dS' = 0, \tag{5.6}$$

$$Q_i(S, S', x_i) + P_i^+(S, S', x_i) + \int_S^\infty P_i^+(S, S'', x_i)Q_i(S'', S', x_i)\,dS' = 0.$$

It follows from (5.4) and (5.5) that

$$(1+P_i^-)\left(M_i\frac{\partial}{\partial x}-L_i\right)(1+K^-)^{-1}$$

$$=(1+P_i^+)\left(M_i\frac{\partial}{\partial x}-L_i\right)(1+K^+)^{-1}=\tilde{M}_i\frac{\partial}{\partial x_i}-\tilde{L}_i. \quad (5.9)$$

This means that the operators M_i, L_i are differential.

Their coefficients can be expressed in terms of the kernels P_i^+, K_i by the relation

$$(1+P_i^+)\left(M_i\frac{\partial}{\partial x_i}-L_i\right)=\left(\tilde{M}_i\frac{\partial}{\partial x_i}-\tilde{L}_i\right)(1+K^+). \quad (5.8)$$

If $M_i=\dfrac{\partial}{\partial x_i}$, then

$$\tilde{M}_i=\frac{\partial}{\partial x_i}-[K^+(S,S)+P_i^+(S,S)]. \quad (5.9)$$

Let Ψ_0 be a solution of system (4.7). It is clear from (5.7) that the functions

$$\Phi^\pm=(1+\hat{K}^\pm)\Psi_0 \quad (5.10)$$

satisfy the equations

$$\left(\tilde{M}_i\frac{\partial}{\partial x_i}-\tilde{L}_i\right)\Phi^\pm=0, \quad (5.11)$$

i.e., this is a solution of a "dressed" system. Assume that the coefficients of the operators M_i, L_i do not depend on S. Among the solutions of system (5.3), (5.4) there are solutions depending on the difference $S-S'$ only; hence their factorization results in $K^\pm$ and $P^\pm$ which are also difference kernels. By performing the Fourier transformation of (5.10) we have

$$\Phi^\pm(\lambda)=(1+f^\pm(\lambda))\Psi_0(\lambda), \quad (5.12)$$

where the function f^+ is analytic in the upper half-plane and the function f^- in the lower one. Formula (5.12) shows that in this case the "dressing" is equivalent to the gauge transformation described in Section 1. The real axis plays the role of the contour Γ. Substitute the kernel F in Marchenko's equation (5.6) of the form

$$F(S,S',x_i)=\frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}G(\lambda',\lambda,x_i)e^{i(\lambda'S-\lambda S')}d\lambda d\lambda' \quad (5.13)$$

and find K of the form

$$K^+(S, S', \omega_i) = -\frac{1}{2\pi}\int_{-\infty}^{\infty} f(\lambda, S, \omega_i)e^{-i\lambda S'}d\lambda. \tag{5.14}$$

We obtain

$$f(\lambda, S) = \int_\Gamma G(\lambda', \lambda)e^{i(\lambda' S - \lambda S')}d\lambda' + \frac{1}{2\pi i}\int_\Gamma \frac{f(\xi, S)G(\lambda', \lambda)e^{i\lambda' S}}{\lambda' - \xi + i0}\,d\lambda'\,d\xi. \tag{5.15}$$

Here the real axis is a contour Γ. Consider the analytic function $\Phi = \Phi(\lambda, S, \omega_i)$ with respect to λ

$$\Phi = 1 + \frac{1}{2\pi i}\int \frac{f(\lambda', S, \omega_i)}{\lambda - \lambda'}d\lambda'$$

having a cut at the real axis. It is easy to verify that its limit values at the cut $\Phi_{1,2}$ satisfy the relation

$$\Phi_2(\lambda) = \Phi_1(\lambda) + \int_\Gamma \Phi_1(\lambda')\tilde{G}(\lambda', \lambda)d\lambda' \tag{5.16}$$

where

$$\tilde{G}(\lambda', \lambda) = G(\lambda', \lambda)e^{i(\lambda' - \lambda)S}. \tag{5.17}$$

Hence, the Marchenko's equation (5.6) at the real axis gives a solution of the non-local Riemann problem (5.16), parametrically dependent on the variable S. This fact was first noted in [38]. In the particular case where $G(\lambda, \lambda') = G(\lambda)\delta(\lambda - \lambda')$ we obtain the local Riemann problem described in Section 1. In this case there is no dependence on δ.

The possibility of interpreting the outlined "dressing" scheme as the non-local Riemann problem allows us to hope that one can essentially generalize the dressing method to the non-local case without any restriction on the contour Γ.

This programme is not yet sufficiently developed but it seems the prospective way to obtain a general solution of the systems under consideration.

As a conclusion one can say that we now have fairly general methods to obtain both multidimensional integrable systems and vast sets of their solutions. There is no doubt that the "dressing" procedure of the Riemann problems should permit studying in detail the equations listed in the introduction, which have important physical applications. Only the future will show whether this list will be extended owing to the possibilities I have outlined.

References

[1] Ablovitz M. J., Van Yaacov D., and Fokas A. S., On the IST for KP II, INS N 21, November 1982, *Stud. Appl. Math.*, to appear.
[2] Anker O. and Freeman N. S., *Proc. Roy. Soc. London Ser. A* **360** (1976), p. 101.
[3] Atiyah M. F., Hitchin H. I., Drinfield V. G., and Manin Yu. I., *Phys. Rev. Lett.* **65** (1978), p. 185.
[4] Atiyah M. F. and Ward R. S., *Comm. Math. Phys.* **55** (1977), p. 117.
[5] Belavin A. A., Polyakov A. M., Shwartz A. S., and Tiupkin S. V., *Phys. Lett.* **B59** (1975), p. 87.
[6] Belavin A. A. and Zakharov V. E., *JETF Lett.* **25** (1977), p. 603.
[7] Belavin A. A. and Zakharov V. E., *Phys. Lett. B* **73** (1978), p. 53.
[8] Calogero F., *Lett. Nuovo Cimento (2)* **14** (1975), p. 133.
[9] Calogero F. and Degasfesis A., *Lett. Nuovo Cimento (2)* **16** (1976), p. 425.
[10] Cornill H., *J. Math. Phys.* **20** (1979), p. 1653.
[11] Davey A. and Stewardson K., *Proc. Roy. Soc. London Ser. A* **333** (1971), p. 101.
[12] Dubrovin B. A., Matveev V. B., and Novikov S. P., *Uspekhi Mat. Nauk* **31**, p. 55.
[13] Fokas A. S. and Ablovitz M. J., On the Inverse Scattering and Direct Linearizing Transforms for KP INS N9, March 1982, *Phys. Lett. A*, to appear.
[14] Fokas A. S. and Ablovitz M. J., On the Inverse Scattering of the Time Dependent Schrödinger Equation and the Associated KPI Equation, INS N22, November 1982, *Stud. Appl. Math.*, to appear.
[15] Kaup D. J., *Stud. Appl. Math.* **55** (1976), p. 9.
[16] Kaup D. J., *Phys. D.* **1** (1980), p. 45.
[17] Kaup D. J., *Stud. Appl. Math.* **62** (1980), p. 73.
[18] Krichever I. M., *Functional Anal. Appl.* **11** (1977), p. 15.
[19] Krichever I. M. and Novikov S. P., *Functional Anal. Appl.* **12** (1978).
[20] Krichever I. M. and Novikov S. P., *Dokl. Akad. Nauk SSSR* **247** (1979), p. 33.
[21] Lectures on the Inverse Scattering Transform for Multidimensional 2+1 Problems by A. S. Fokas and M. J. Ablovitz, INS N28, Mexico, December 1982.
[22] Manakov S. V., *Uspekhi Fiz. Nauk* **5** (1976), p. 245.
[23] Manakov S. V., *Phys. D* **3** (1981), p. 420.
[24] Manakov S. V., Dissertation, Moskow 1983.
[25] Manakov S. V., Santini P. and Takhtajan L. A., *Phys. Lett. A* **74** (1980), p. 451.
[26] Manakov S. V. Zakharov V. E., Bordag L., Ifg A. R., and Matveev V. B., *Phys. Lett. A* **63** (1979), p. 205.
[27] Manakov S. V. and Zakharov V. E., *Lett. Math. Phys.* **2** (1981), p. 247.
[28] Michailov A. V., *Phys. D.* **3** (1981), p. 73.
[29] Polmeyer R., *Comm. Math. Phys.* **6** (1980), p. 86.
[30] Volovich I. V., *Dokl. Akad. Nauk Ukrain. SSR Ser. A* **269** (1983).
[31] Volovich I. V., *Theoret. and Math. Phys.* **55** (1983), p. 39.
[32] Zakharov V. E., *JETF Lett.* **22** (1975), p. 364.
[33] Zakharov V. E., *Dokl. Akad. Nauk Ukrain. SSR Ser. A* **229** (1976), p. 1314.
[34] Zakharov V. E. In: R. K. Bullough and P. K. Candry (eds.), *Current Topics in Physics*, Springer-Verlag, Berlin 1980.
[35] Zakharov V. E., Integrable Systems in Multidimensional Spaces, *Lecture Notes in Phys.* vol. 153, Springer-Verlag, Berlin 1982, p. 190.
[36] Zakharov V. E. and Manakov S. V., *JETF* **69** (1975), p. 1654.

[37] Zakharov V. E. and Manakov S. V., *Theoret. and Math. Phys.* **27** (1976), p. 283.
[38] Zakharov V. E. and Manakov S. V., The Theory of Solitons, *Soviet Sci. Rev. Sect. C: Math. Phys. Rev.* **1** (1979), p. 133.
[39] Zakharov V. E. and Mikhailov A. V., *Functional Anal. Appl.* 1983, in print.
[40] Zakharov V. E. and Shabat A. B., *Functional Anal. Appl.* **8** (1974), p. 43.
[41] Zakharov V. E. and Shabat A. B., *JETF* **74** (1978), p. 1953.
[42] Zakharov V. E. and Shabat A. B., *Functional Anal. Appl.* **13** (1979), p. 13.

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

ANATOLE KATOK

Nonuniform Hyperbolicity and Stru cture o Smooth Dynamical Systems

§ 1. Introduction

A substantial part of recent progress in the theory of smooth dynamical systems is based on better and more systematic understanding, than before, of the role played by "hyperbolic" behavior and more specifically by nonuniform hyperbolicity and Lyapunov characteristic exponents. One and probably the most important aspect of this development concerns ergodic properties of smooth dynamical systems with respect to absolutely continuous invariant measures or other measures naturally connected with the smooth structure. The main work in that area in the last decade was done by Pesin [10], [11], [12], [13], and is now often referred to as the Pesin theory. Both the methods employed by Pesin and his results are essential for the subsequent development. He discovered the crucial role of nonuniform hyperbolicity and Lyapunov characteristic exponents and using these tools developed an ergodic theory for smooth dynamical systems with respect to an absolutely continuous invariant measure. His results include the celebrated entropy formula which shows that the entropy comes exclusively from the exponential expansion, the description of π-partition and a complete classification of systems with nonzero exponents.

Among the developments that appeared after Pesin's work I would like to point out Mañé's proof of the entropy formula [8], which contains a fundamental simplification of the original approach, the recent works of Ledrappier [6] and Ledrappier and L.-S. Young [7] on the characterization of measures satisfying the entropy formula and a work on ergodic theory of geodesic flows on manifolds of nonpositive curvature by Bal-

lmann and Brin [1]. The lack of space does not allow me to discuss here an extensive work by various authors on absolutely continuous invariant measures for one-dimensional maps and on various special, primarily 2-dimensional examples, including both conservative transformations and maps with nonuniformly hyperbolic attractors.

In this talk I am going to discuss another aspect of the development based on the concept of nonuniform hyperbolicity, namely, how certain global "exponential" properties of a dynamical system produce certain types of orbits including the abundance of periodic orbits and large hyperbolic sets. The structure of a dynamical system on a locally maximal hyperbolic set is well understood. It includes such ingredients as stable and unstable manifolds, local product structure, shadowing property, closing lemma, local stability of the set, density of periodic orbits among the recurrent orbits, Markov partitions, existence and uniqueness of measure with maximal entropy on basic sets, the uniform distribution of periodic orbits according to that measure and the growth of the number of periodic orbits with the exponential rate given by the topological entropy. Thus, the existence of an infinite locally maximal hyperbolic set for a given dynamical system provides considerable information about the orbit structure of the system and all effects obtained that way persist under small perturbations of the system.

Most of the results discussed below are contained in my papers [2], [3], [4], [5], although in several cases I will formulate theorems in slightly stronger or more general form than they were written.

Before proceeding to a more technical discussion let me outline the strategy of the approach. We begin with a certain "global" property which indicates that some kind of exponential growth is present. Here are some examples of global exponential properties.

(i) Positive topological entropy, i.e., the exponential growth rate ot the number of different orbits distinguishable with an arbitrary fine buf fixed precision.

(ii) Exponential behavior of the iterates of the map f_* induced by a diffeomorphism f: $M \to M$ on the fundamental group $\pi_1(M)$, i.e., the exponential growth of the word-length norm of the iterates $f_*^n \gamma$ for all (or some) $\gamma \in \pi_1(M) \setminus \{\mathrm{id}\}$.

(iii) Similar exponential behavior of the maps induced on homology groups.

(iv) Exponential growth of the volume of a ball on the universal covering of a compact Riemann manifold M. This property appears when

the dynamical system under consideration is the geodesic flow generated by the metric.

(v) In the same situation as in (iv), the exponential growth of the fundamental group $\pi_1(M)$ is another exponential type property.

We will derive from a global property the existence of invariant measures for the dynamical system such that orbits typical with respect to such a measure possess a weaker type of hyperbolicity than the orbits belonging to a hyperbolic set. The linearized system along such an orbit allows an exponential dychotomy but the coefficients in front of the exponential terms may oscillate as the initial point moves along the orbit. This is the reason for calling those orbits nonuniformly hyperbolic. However, in our case the oscillations of the coefficients are not too big, they are essentially subexponential. The existence of many such regular nonuniformly hyperbolic orbits follows from Oseledec's Multiplicative Ergodic Theorem [9]. A neighborhood of a regular nonuniformly hyperbolic orbit possesses certain properties similar to a neighborhood of a hyperbolic set. Using proper variations of closing and shadowing arguments one can catch many orbits which never leave a (noninvariant) neighborhood with uniform hyperbolic estimates and thus possess a uniform hyperbolic structure. This construction may be supplemented with the estimates on the number of different orbits found and on the quality of hyperbolic estimates along those orbits.

Let us discuss the last notion in detail. Let x be a hyperbolic periodic point of period n. The degree of hyperbolicity of x is measured by the number

$$m(x) = \frac{1}{n} \min_{\lambda \in \operatorname{sp} Df_x^n} |\log|\lambda||. \tag{1}$$

Our standard set-up in the discrete time case is to consider a diffeomorphic embedding $f\colon U \to M$ of an open neighborhood U of a compact invariant set Γ; here M is an ambient smooth manifold. Let for an open set $V \supset \Gamma$ and for $\chi > 0$, $n \in \mathbf{Z}_+$, $P_{n,\chi}^V(f)$ be the number of hyperbolic points $x \in V$ of period n with $m(x) \geqslant \chi$.

Furthermore, let

$$p_\chi^V(f) = \overline{\lim_{n\to\infty}} \frac{\log P_{n,\chi}^V(f)}{n}$$

and

$$p_\chi^\Gamma(f) = \inf_{V \supset \Gamma} p_\chi^V(f).$$

If $\Gamma = M$ we will write $p_\chi(f)$ instead of $p_\chi^\Gamma(f)$.

Similar definitions can be made for a continuous time dynamical system in a similar set-up. In the definition of $m(x)$, the eigenvalue 1 corresponding to the direction of the vector field should be excluded; instead of periodic points, periodic orbits would be counted; instead of orbits of a fixed period, one should count all orbits of period $\leqslant T$.

§ 2. Main results and applications

We will assume the standard set-up described above. All maps and flows are assumed of class $C^{1+\delta}$ for some $\delta > 0$. In the continuous time case we also assume that the flow does not have fixed points on Γ (added in proof: I have recently been able to remove this assumption). In both cases, h_Γ will denote the topological entropy of the dynamical system restricted to Γ. We assume $h_\Gamma > 0$.

THEOREM 1. *Let* $f\colon U \to M$ *and* $\dim M = 2$. *Then for every* $\varepsilon > 0$

$$p^\Gamma_{h_\Gamma - \varepsilon}(f) \geqslant h_\Gamma.$$

THEOREM 2. *Let* $f_t\colon U \to M$ *be a flow and* $\dim M = 3$. *Then for every* $\varepsilon > 0$

$$p^\Gamma_{h_\Gamma - \varepsilon}(f) \geqslant h_\Gamma.$$

THEOREM 3. *Under the assumptions of Theorem* 1, *for every* $\varepsilon > 0$ *and every open set* $V \supset \Gamma$ *there exists an invariant locally maximal hyperbolic set* $\Lambda_\varepsilon \subset V$ *such that* $f|_{\Lambda_\varepsilon}$ *is topologically conjugate to a subshift of finite type and*

$$h(f|_{\Lambda_\varepsilon}) > h_\Gamma - \varepsilon.$$

THEOREM 4. *Under the assumptions of Theorem* 2, *for every* $\varepsilon > 0$ *and every open set* $V \supset \Gamma$ *there exists an invariant locally maximal hyperbolic set* $\Lambda_\varepsilon \subset V$ *such that* $f_t|_{\Lambda_\varepsilon}$ *is topologically conjugate to a suspension over a subshift of finite type and*

$$h(f_t|_{\Lambda_\varepsilon}) > h_\Gamma - \varepsilon.$$

COROLLARY 1. *The topological entropy* $h(f)$ *of any* $C^{1+\delta}$ *diffeomorphism* $f\colon M \to M$ *is upper-semicontinuous as a function of* f *in* C^0 *topology.*

Proof. Follows immediately from Theorem 3 applied to $\Gamma = M$ and from the topological stability of hyperbolic sets.

THEOREM 5. *Let* $f\colon M \to M$ *be an area-preserving diffeomorphism of a compact surface. Then* f *has a hyperbolic periodic point iff*

$$\overline{\lim_{n\to\infty}} \frac{\log \|Df^n\|}{n} > 0. \tag{2}$$

Here we assume that a Riemannian metric is fixed on M so that $\|Df\| = \max_{x \in M} \sup_{v \in T_x M \setminus \{0\}} \|Df v\| / \|v\|$. However, the quantity in the left-hand part of (2) does not depend on the choice of Riemannian metric.

All results stated above about the existence of many periodic points and nontrivial invariant sets depend on smoothness. M. Rees [14] constructed an example of a minimal homeomorphism of the 2-torus with positive topological entropy. It is not clear, however, whether the $C^{1+\delta}$ assumption can be replaced by C^1.

The next group of results deals with the situations where the existence of many periodic orbits has been established by topological or variational methods. Such methods, however, usually say nothing about the hyperbolicity of those orbits. By applying the above-stated theorems one can ensure the existence of many hyperbolic orbits.

Let $f: T^2 \to T^2$ be a diffeomorphism of the two-dimensional torus which acts on the first homology group hyperbolically. This action is determined by an integer matrix $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ such that $\det A = \pm 1$ and $|\operatorname{tr} A| > 2$. Let λ be the eigenvalue of A of absolute value greater than 1 and $\alpha = \log |\lambda|$. Then $h(f) \geqslant \alpha$.

COROLLARY 2. *For every* $\varepsilon > 0$

$$p_{\alpha-\varepsilon}(f) \geqslant \alpha.$$

If, in addition, f is an Anosov diffeomorphism then

$$p_{\alpha}(f) \geqslant \alpha.$$

Let M be a compact surface of genus greater than one and $f: M \to M$ be a diffeomorphism homotopic to a pseudo-Anosov map f_0. Then $h(f) \geqslant \alpha$ where $\alpha = h(f_0)$ and α is also equal to the exponential growth rate of the word-length norm for the iterates $f_*^n \gamma$ where γ is an arbitrary element of $\pi_1(M)$ different from identity. Nielsen's theorem implies that the exponential growth rate of the number of periodic orbits for f is $\geqslant \alpha$.

COROLLARY 3. *For every* $\varepsilon > 0$

$$p_{\alpha-\varepsilon}(f) \geqslant \alpha.$$

The next example is more interesting. Let σ be a Riemannian metric of class $C^{2+\delta}$ on a compact surface M with negative Euler characteristic E such that the total area of M is equal to v. Let φ_t^σ be the geodesic flow generated by that metric. The exponential growth rate $p_{\sigma,\chi}$ for the num-

ber of hyperbolic closed geodesics with the positive Lyapunov exponent $\geqslant \chi$ coincides with what we denote by $p_\chi(\varphi_t^\sigma)$. Let $K(E, v) = (-2\pi E/v)^{1/2}$. If σ is a metric of constant negative curvature then this number represents the common value of the topological entropy, entropy with respect to Liouville (smooth) measure and the positive Lyapunov exponent along any orbit.

THEOREM 6 [4].

$$p_{\sigma,K(E,v)} \geqslant K(E, v)$$

and this inequality is strict unless σ is a metric of constant negative curvature. Moreover, for every metric of nonconstant curvature there exists $\varepsilon_\sigma > 0$ such that

$$p_{\sigma,K(E,v)+\varepsilon_\sigma} > K(E, v).$$

Thus, any metric of nonconstant curvature has more closed geodesics with stronger hyperbolic properties than any metric of constant curvature on the same surface with the same total area.

This theorem follows from Theorem 2 and an entropy estimate. The metric σ is conformally equivalent to a metric σ_0 of constant negative curvature and the same total area. Let ϱ be the conformal coefficient. Its average is equal to one. Therefore, the average of $\varrho^{1/2}$, which we will denote by ϱ_σ, is less than 1 unless $\varrho \equiv 1$.

Let h_σ be the topological entropy of the geodesic flow φ_t^σ. Here is the desired entropy estimate.

THEOREM 7[4].

$$h_\sigma \geqslant \varrho_\sigma^{-1} K(E, v).$$

§ 3. Hyperbolic measures

Let μ be a Borel probability measure supported by Γ and invariant and ergodic with respect to a map or a flow under consideration. Let $\chi_1^\mu < \chi_2^\mu < \ldots < \chi_r^\mu$ be the Lyapunov characteristic exponents of the dynamical system with respect to μ. The multiplicative ergodic theorem implies that for μ-almost every point $x \in \Gamma$ there exists a measurable invariant decomposition of the tangent space $T_xM = E_1(x) \oplus \ldots \oplus E_r(x)$ such that for $v \in E_i(x)$

$$\lim_{t \to \pm\infty} \frac{\log \|Df_t v\|}{t} = \pm\chi_i^\mu.$$

By the ergodicity, $\dim E_i(x)$ must be constant μ-almost everywhere. We will denote this dimension by k_i^μ and call it the *multiplicity* of the exponent χ_i^μ.

DEFINITION 1. A measure μ is called *hyperbolic* if

(i) in the discrete time case, all χ_i^μ are different from 0,

(ii) in the continuous time case, the zero exponent has multiplicity one.

Sometimes we will also call a nonergodic invariant measure hyperbolic if almost all its ergodic components are hyperbolic measures.

For a hyperbolic invariant measure μ let

$$m(\mu) = \min_{i:\chi_i^\mu \neq 0} (\chi_i^\mu).$$

This definition agrees with (1) for a measure concentrated on a single hyperbolic periodic orbit. Naturally, $m(\mu)$ characterizes the minimal rate of exponential behavior typical for the system.

THEOREM 8. *Let μ be an invariant ergodic hyperbolic measure for a map or a flow. Let $x \in \operatorname{supp}\mu$. Then for any $\delta > 0$, any neighborhoods $V \ni x$ and $W \supset \operatorname{supp}\mu$, and any collection of continuous functions $\varphi_1, \ldots, \varphi_k$ there exists a hyperbolic periodic point $z \in V$ such that the orbit of z is contained in W and*

$$m(z) > m(\mu) - \delta.$$

Moreover, in the diffeomorphism case

$$\left|(\operatorname{per} z)^{-1} \sum_{k=0}^{\operatorname{per} z-1} \varphi_i(f^k z) - \int \varphi_i d\mu\right| < \delta$$

for $i = 1, \ldots, k$. A similar property holds for flows.

The last statement means that the orbit of the point z is almost uniformly distributed with respect to μ.

Theorem 5 follows easily from Theorem 8 since (2) implies the existence of an f-invariant measure whose largest exponent is positive and the preservation of area ensures that the second exponent for that measure is negative. Another corollary is "weak stability" of hyperbolic measures in C^1 topology.

COROLLARY 4. *Let μ be an invariant ergodic hyperbolic measure for a diffeomorphism f or a flow f_t. If f_n converges to f (correspondingly $f_t^{(n)}$ converges to f_t) in C^1 topology, then f_n ($f_t^{(n)}$) has an invariant hyperbolic measure μ_n such that μ_n converges to μ weakly.*

THEOREM 9. *If, under the assumptions of Theorem 8, μ is not concentrated on a single periodic orbit, then z has a transversal homoclinic orbit.*

COROLLARY 5. *If a diffeomorphism or a flow has a hyperbolic ergodic invariant measure whose support is an infinite set then its topological entropy is positive.*

THEOREM 10. *Under the assumptions of Theorem* 8, *let* $\operatorname{supp}\mu = \Gamma$ *and* $h_\mu(f)$ (*corr.* $h_\mu(f_t)$) *be equal to* $h > 0$. *Then for any* $\varepsilon > 0$

$$p^\Gamma_{h-\varepsilon}(f) \geqslant h \quad (\text{corr. } p^\Gamma_{h-\varepsilon}(f_t) \geqslant h).$$

Theorems 1 and 2 follow easily from Theorem 10, variational principle, and Ruelle's entropy inequality [15].

THEOREM 11. *Under the assumptions as in the previous theorem, there exists an f-invariant, locally maximal hyperbolic set* Λ_ε *such that the restriction* $f|_{\Lambda_\varepsilon}$ *is topologically conjugate to a subshift of finite type and*

$$h(f|_{\Lambda_\varepsilon}) > h_\mu(f) - \varepsilon.$$

Moreover, any orbit on Λ_ε *is almost uniformly distributed with respect to* μ (cf. Theorem 8).

Theorem 11 and its counterpart for flows which we do not formulate explicitly imply Theorems 3 and 4 in the same fashion as Theorem 10 implies Theorems 1 and 2.

It also allows us to strengthen weak stability of Corollary 4 to "entropy stability".

COROLLARY 6. *Under the assumptions of Corollary* 4, *the sequence of measures* μ_n *can be chosen with the additional property* $h_{\mu_n}(f_n) \to h_\mu(f)$ (*corr.* $h_{\mu_n}(f_t^{(n)}) \to h_\mu(f_t)$).

References

[1] Ballmann W. and Brin M., On the Ergodicity of Geodesic Flows, *Ergod. Th. and Dyn. Syst.* **2** (1982), pp. 311–315.

[2] Katok A., Lyapunov Exponents, Entropy and Periodic Orbits for Diffeomorphisms, *Publ. Math. IHES* **51** (1980), pp. 137–173.

[3] Katok A., Hyperbolicity, Entropy and Minimality for Smooth Dynamical Systems. In: *Atas do Décimo Segundo,* Colóquio Brasileiro de Mathematica, vol. 2, Rio de Janeiro, 1981, pp. 571–581.

[4] Katok A., Entropy and Closed Geodesics, *Ergod. Th. and Dyn. Syst.* **2** (1982), pp. 339–365.

[5] Katok A., *Lyapunov Exponents, Entropy, Hyperbolic Sets and ε-orbits,* to appear.

[6] Ledrappier F., Propriétés ergodiques des mesures de Sinai, *Publ. Math. IHES* (1984).

[7] Ledrappier F. and Young L.-S., *Metric Entropy of Diffeomorphisms,* to appear.

[8] Mañé R., A proof of Pesin's Formula, *Ergod. Th. and Dyn. Syst.* **1** (1981), pp. 95–102; errata **3** (1983), pp. 159–160.

[9] Oseledec V. I., Multiplicative Ergodic Theorem. Lyapunov Characteristic Numbers for Dynamical Systems, *Trans. Moscow Math. Soc.* **19** (1968), pp. 197–221; translated from Russian.

[10] Pesin Ja. B., Families of Invariant Manifolds Corresponding to Nonzero Characteristic Exponents, *Math. USSR–Izv.* **10** (6) (1976), pp. 1261–1305; translated from Russian.

[11] Pesin Ja. B., Characteristic Lyapunov Exponents and Smooth Ergodic Theory, *Russian Math. Surveys* **32** (4) (1977), pp. 55–114; translated from Russian.

[12] Pesin Ja. B., Description of π-Partition of a Diffeomorphism with Invariant Measure, *Math. Notes* **22** (1) (1977), pp. 506–514; translated from Russian.

[13] Pesin Ja. B., Geodesic Flows on Closed Riemannian Surfaces without Focal Points, *Math. USSR–Izv.* **11** (6) (1977), pp. 1195–1228; translated from Russian.

[14] Rees M., A Minimal Positive Entropy Homeomorphism of the 2-Torus, *J. London Math. Soc.* **23** (1981), pp. 537–550.

[15] Ruelle D., An Inequality for the Entropy of Differentiable Maps, *Bol. Soc. Brasil. Mat.* **9** (1978), pp. 83–87.

DEPARTMENT OF MATHEMATICS,
UNIVERSITY OF MARYLAND,
COLLEGE PARK, MD 20742, USA

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

A. LASOTA

Asymptotic Behaviour of Solutions: Statistical Stability and Chaos

Introduction

In the history of attempts to describe the behaviour of complicated dynamical systems, the work of Boltzmann some hundred years ago marked a turning point. With the current intense interest in the properties of chaotic systems, Boltzmann's original idea of treating the evolution of densities under the action of a dynamical system is even more attractive.

In the past few years a few simple sufficient conditions for the asymptotical stability of sequences of densities have been discovered. These criteria, which are related to a spectral decomposition theorem for positive operators, are quite powerful and, potentially, of great utility in practical situations. This paper will present these stability criteria and examine several systems to which they are applicable. It is especially significant that the same technique may be used to: (1) demonstrate the statistical stability of the dynamical systems generated by piecewise expanding transformations on intervals, on the real line and on manifolds; and (2) the asymptotic stability of densities which are solutions of the linear Boltzmann equation, of integral equations and partial differential equations of diffusion type.

Throughout the paper no attempt will be made to present the results in their most general form, as our primary concern is to illustrate the variety of problems which can be solved by the same method.

1. Stochastic semigroups and dynamical systems

Let $(X, \mathscr{A}, \mu)$ be a measure space with a nonnegative σ-finite measure μ. A linear mapping $P\colon L^1 \to L^1$ $\left(L^1 = L^1(X, \mathscr{A}, \mu)\right)$ will be called a *Markov*

operator (cf. [6]) if it satisfies the following two conditions:

$$Pf \geqslant 0 \quad \text{for } f \geqslant 0,\ f \in L^1, \tag{1.1}$$

$$\|Pf\| = \|f\| \quad \text{for } f \geqslant 0,\ f \in L^1, \tag{1.2}$$

where $\|\cdot\|$ stands for the norm in L^1. Conditions (1.1) and (1.2) imply that

$$|Pf| \leqslant P|f| \quad \text{and} \quad \|Pf\| \leqslant \|f\| \quad \text{for } f \in L^1. \tag{1.3}$$

Let T be a nontrivial semigroup of real nonnegative numbers, i.e., $T \neq \{0\}$ and $t_1 \pm t_2 \in T$ for every $t_1, t_2 \in T$ $(t_1 \geqslant t_2)$. A family of Markov operators $\{P^t\}_{t\in T}$ will be called a *stochastic semigroup* if

$$P^{t_1+t_2} = P^{t_1}P^{t_2} \quad \text{for } t_1, t_2 \in T. \tag{1.4}$$

By $D = D(X, \mathscr{A}, \mu)$ we denote the set of all nonnegative elements of L^1 with norm equal to one. The elements of D will be called *densities*. Since every element of L^1 can be written as a linear combination of two densities, in studying the asymptotic properties of $\{P^t\}_{t\in T}$ it is sufficient to consider $\{P^t f\}_{t\in T}$ for $f \in D$.

A density f_0 is called *stationary* if $P^t f_0 = f_0$ for $t \in T$. From (1.3) it follows that every stationary density is stable. In fact, if $f_0 \in D$ is stationary, then for every other $f \in D$

$$\|P^t f - f_0\| = \|P^t f - P^t f_0\| \leqslant \|f - f_0\|.$$

A stationary density f_0 will be called *asymptotically stable* if

$$\lim_{t\to\infty} \|P^t f - f_0\| = 0 \quad \text{for } f \in D. \tag{1.5}$$

Of course for a given semigroup $\{P^t\}_{t\in T}$ there is at most one asymptotically stable density. If such a density exists, then $\{P^t\}_{t\in T}$ will also be called asymptotically stable.

Stochastic semigroups usually arise from pure probabilistic problems such as random walks, stochastic differential equations and many others. It is of great importance that they can also be generated by "deterministic" semidynamical systems. A family of transformations $S_t\colon X \to X$ $(t \in T)$ will be called a *semidynamical system* if it satisfies the following two conditions:

$$S_t^{-1}(A) \in \mathscr{A} \quad \text{for } A \in \mathscr{A},\ t \in T, \tag{1.6}$$

$$S_{t_1+t_2} = S_{t_1} \circ S_{t_2} \quad \text{for } t_1, t_2 \in T. \tag{1.7}$$

A semidynamical system will be called *nonsingular* if, in addition,

$$\mu\big(S_t^{-1}(A)\big) = 0 \quad \text{for } t \in T \text{ whenever } \mu(A) = 0. \tag{1.8}$$

Given a nonsingular semidynamical system $\{S_t\}$, we may define a family of operators P_S^t: $L^1 \to L^1$ by setting

$$\int_A P_S^t f(x)\mu(dx) = \int_{S_t^{-1}(A)} f(x)\mu(dx) \quad \text{for } f \in L^1,$$

$$A \in \mathscr{A} \quad \text{and} \quad t \in T. \tag{1.9}$$

Due to the nonsingularity of $\{S_t\}_{t\in T}$, condition (1.9) uniquely defines $\{P_S^t\}_{t\in T}$ via the Radon–Nikodym theorem. It is also easy to verify that $\{P_S^t\}_{t\in T}$ is a stochastic semigroup.

The semigroup $\{P_S^t\}_{t\in T}$ has an interesting physical interpretation. Assume namely that a large number of particles move independently in the space X and that the trajectory of a particle starting from the point x at time $t = 0$ is given by $\{S_t(x)\}_{t\in T}$. Assume moreover that at $t = 0$ the positions of the particles are distributed according to a density f. Then this density evolves in time and its evolution is described by $\{P_S^t f\}_{t\in T}$.

The behaviour of $\{P_S^t\}_{t\in T}$ allows us to determine many properties of the semidynamical system $\{S_t\}_{t\in T}$ such as preservation of measure, ergodicity, mixing and exactness. Here we will not concentrate on these problems. Instead, we make the following definition. A nonsingular system $\{S_t\}_{t\in T}$ will be called *statistically stable* if the corresponding stochastic semigroup $\{P_S^t\}_{t\in T}$ is asymptotically stable.

To conclude this section consider the special discrete time case of stochastic semigroups and dynamical systems, $T = N = \{0, 1, \ldots\}$ (see [5]). In this case P^n is the n-th power of the operator $P = P^1$ and S_n is the n-th iterate of the mapping $S = S_1$. Given a nonsingular S, we define the corresponding Markov operator P_S by the simple formula

$$\int_A P_S f(x)\mu(dx) = \int_{S^{-1}(A)} f(x)\mu(dx) \quad \text{for } A \in \mathscr{A}, \tag{1.13}$$

and we have $P_S^n = P_{S_n}$. The operator P_S is called the *Frobenius–Perron operator* corresponding to S.

It is easy to prove the following proposition, which emphasizes the role of discrete time semigroups.

PROPOSITION 1.1. *Let $\{P_S^t\}_{t\in T}$ be a stochastic semigroup and let $t_0 \in T$ be a fixed positive number. Then the asymptotical stability of $\{P_S^t\}_{t\in T}$ is equivalent to the asymptotical stability of the discrete time semigroup $\{P^{t_0 n}\}_{n\in N}$.*

The notion of statistical stability for discrete time semigroups is closely related to the concept of exactness in the sense of Rochlin [29]. In particu-

lar, when $S\colon X\to X$ is measure preserving and $\mu(X)=1$, these notions are equivalent [23].

2. A spectral decomposition theorem for Markov operators

It is well known that positive contractions of Banach lattices have some special spectral properties [30]. These properties were used by G. Keller to study the asymptotic behaviour of stochastic semigroups corresponding to picewise monotonic transformations of the interval and piecewise analytic mappings on the plane [10], [11]. Here we present a general theory applicable to every nonsingular mapping on a σ-finite measure space.

Let a σ-finite measure space $(X, \mathscr{A}, \mu)$ be given and let P be a Markov operator. We say that P is strongly (weakly) *constrictive* if there exists a strongly (weakly) compact set $F\subset L^1$ such that

$$\lim_{n\to\infty} d(P^n f, F) = 0 \quad \text{for } f\in D. \tag{2.1}$$

Here $d(g, F)$ denotes the distance between g and F, that is, the infimum of $\|g-f\|$ for $f\in F$.

The following two theorems summarize the properties of constrictive operators [17], [20].

THEOREM 2.1. *Let $S\colon X\to X$ be a nonsingular mapping and P_S the corresponding Frobenius–Perron operator. If P_S is weakly constrictive, then P_S is also strongly constrictive.*

THEOREM 2.2. *Let P be a strongly constrictive Markov operator. Then there exist a sequence of densities $\{g_i\}$ $(i=1,\dots,r)$ with mutually disjoint supports $(g_i g_j = 0$ for $i\neq j)$ and a sequence of linear functionals $\{\lambda_i\}$ $(\lambda_i\in L^{1*})$ such that*

$$\lim_{n\to\infty}\Big\|P^n\Big(f-\sum_{i=1}^{r}\lambda_i(f)g_i\Big)\Big\| = 0 \quad \text{for } f\in L^1 \tag{2.2}$$

and

$$Pg_i = g_{a(i)} \quad \text{for } i=1,\dots,r. \tag{2.3}$$

where a is a permutation of the integers $1,\dots,r$.

From Theorem 2.2 it follows immediately that the n-th power P^n of P can be written in the form

$$P^n f = \sum_{i=1}^{r}\lambda_i(f)g_{a^n(i)} + R_n f \quad \text{for } f\in L^1, \tag{2.4}$$

where α^n denotes the n-th itegrate of the permutation α and the remainder R_n converges strongly to zero as $n \to \infty$. Thus every sequence $\{P^n f\}_{n \in N}$ is asymptotically periodic with a period which does not exceed $r!$.

It is easy to find an estimation of the integer r in (2.4). In fact, assume that there exists a function $g \in L^1$ such that

$$\lim_{n \to \infty} \|(P^n f - g)^+\| = 0 \quad \text{for } f \in D, \tag{2.5}$$

where $(u)^+ = \max(0, u)$. Then according to (2.4) and (2.5)

$$Lf = \lim_{n \to \infty} P^{r!n} f = \sum_{i=1}^{r} \lambda_i(f) g_i \leqslant g \quad \text{for } f \in D.$$

In particular, setting $f = g_k$, we have $Lf = g_k \leqslant g$. Integrating over X and bearing in mind that the supports of g_i are disjoint, we obtain

$$r = \sum_{i=1}^{r} \int_X g_i(x) \mu(dx) \leqslant \int_X g(x) \mu(dx),$$

which means that $r \leqslant \|g\|$.

Now observe that the set $F = \{f\colon 0 \leqslant f \leqslant g\}$ is weakly compact and that $r = 1$ implies the asymptotic stability of $\{P^n\}_{n \in N}$. Taking all this into account, we obtain from Theorems 2.1 and 2.2 the following

COROLLARY 2.1. *Let $S\colon X \to X$ be a nonsingular transformation and let P_S be the Frobenius–Perron operator corresponding to S. If P_S satisfies* (2.5) *and $\|g\| < 2$, then $\{S_n\}_{n \in N}$ is statistically stable.*

From the point of view of applications the condition $\|g\| < 2$ is quite restrictive. However, it may be replaced by an estimation of $P^n f$ from below. Namely, using Theorem 2.2, it is easy to deduce the following

COROLLARY 2.2. *Let P be a strongly constrictive Markov operator. Assume there is a set $A \in \mathscr{A}$ of positive measure with the property that for every $f \in D$ there is an integer $n_0(f)$ such that*

$$P^n f(x) > 0 \quad \text{for } x \in A,\ n \geqslant n_0(f). \tag{2.6}$$

Then $\{P^n\}_{n \in N}$ is asymptotically stable.

Again in order to prove this corollary it is sufficient to show that $r = 1$ in formula (2.4). Indeed, if r were bigger than 1, then there would exist an integer k such that A would not be contained in the support of g_k and the sequence $\{P^n f\}$ with $f = g_k$ would not satisfy (2.6). Thus $r = 1$.

Finally we may completely eliminate the assumption that P is constrictive if we replace (2.6) by the stronger condition that the semigroup $\{P^n\}$ has a so-called "lower bound function". Because of a later application to differential equations we formulate this part of the theory for general (not necessarily discrete time) semigroups.

Let $\{P^t\}_{t\in T}$ be a stochastic semigroup. A function $h \in L^1$ is called a *lower bound function* for $\{P^t\}_{t\in T}$ if

$$\lim_{t\to\infty} \|(h - P^t f)^+\| = 0 \quad \text{for } f \in D. \tag{2.7}$$

A lower bound function h is called *nontrivial* if $h \geqslant 0$ and $\|h\| > 0$. By using arguments similar to those in the classical work of A. Markov [24] t is easy to prove the following [16]

THEOREM 2.3. *A stochastic semigroup $\{P^t\}_{t\in T}$ is asymptotically stable if and only if it has a nontrivial lower bound function.*

To close this section observe that is is not necessary to verify conditions (2.1), (2.5) and (2.7) for all possible $f \in D$. Due to the fact that the Markov operators are linear and contractive it is sufficient to verify these conditions for all f belonging to an arbitrary dense subset of D.

3. Dynamical systems generated by expanding mappings

The asymptotic behaviour of discrete time stochastic semigroups generated by expanding transformations is now well understood. We will show how this behaviour may be explained by using the theory developed in the previous section.

Our first application refers to piecewise monotonic transformations of a finite interval. This class of mappings was introduced by A. A. Kosyakin and E. A. Sandler [13] and was intensively studied, among others, by T. Y. Li and J. A. Yorke [22], S. Wong [35], G. Wagner [32], R. Bowen [3], G. Pianigiani [27], Z. S. Kowalski [12], A. Boyarsky [4] and G. Keller [11]. A precise description of the statistical properties of piecewise monotonic transformations may be given by a spectral decomposition of P_S. The possibility of such a decomposition was observed by G. Keller [11] and is the subject of our Proposition 3.1.

Consider a piecewise monotonic transformation S: $[0,1] \to [0,1]$ which satisfies the following conditions:

(M1) There is a partition $0 = a_0 < \ldots < a_m = 1$ such that for each integer i the restriction of S to the interval (a_{i-1}, a_i) is a C^2 function.

(M2) $\inf|S'(x)| > 1$ $(x \neq a_i)$.

(M3) $\sup\left(|S''(x)|/(S'(x))^2\right) < \infty$ $(x \neq a_i)$.

As usual, denote by P_S the Frobenius–Perron operator corresponding to S. It is quite easy to estimate the total variation of functions $P_S^n f$ for large n. Namely if S satisfies conditions (M1)–(M3), then there exists a constant K, independent of f, such that

$$\lim_{n\to\infty} \bigvee_0^1 P_S^n f \leqslant K \tag{3.1}$$

for every $f \in D$ of bounded variation [21]. The set

$$F = \{g \in D: \bigvee_0^1 g \leqslant K\}$$

is strongly compact in L^1 and by (3.1) the sequence $\{P^n f\}$ converges to F. Therefore we have established the following result:

PROPOSITION 3.1. *If* S: $[0,1] \to [0,1]$ *satisfies conditions* (M1)–(M3), *then the operator* P_S *is strongly constrictive. In particular, for every* $f \in L^1$ *the sequence* $\{P^n f\}$ *is asymptotically periodic.*

Analogous results can be obtained for piecewise analytic and expanding transformations of the unit square [10]. It is also possible to prove the constrictiveness of P_S for some piecewise convex transformations of the n-dimensional cube considered by M. Jabłoński [9], and some transformations of the unit interval with a negative Schwarzian derivative studied by M. Misiurewicz [25]. However, these results are technically difficult and we omit the details here.

As a second example consider a piecewise convex transformation S: $[0,1] \to [0,1]$ which satisfies the following conditions:

(C1) There is a partition $0 = a_0 < \ldots < a_m = 1$ such that for each integer i the restriction S_i of S to the interval $[a_{i-1}, a_i)$ is a convex function.

(C2) $S_i(a_{i-1}) = 0$ and $S_i'(a_{i-1}) > 0$ for $i = 1, \ldots, m$.

(C3) $S_1(0) > 1$.

In this case the set D_0 of all densities of the form

$$f(x) = \sum_{i=1}^{k} \lambda_i 1_{[c_i, d_i)}(x)$$

with

$$c_i, d_i \in \bigcup_{n=0}^{\infty} S^{-n}\{a_0, \ldots, a_m\}$$

is dense in D. It can also be verified [19] that the sequences $\{P_S^n f\}$ with $f \in D_0$ have the following two properties: (1) for every f there is an integer $n_0(f)$ such that the functions $\{P_S^n f\}$ are non-increasing for $n \geqslant n_0$; and (2) there is a constant K independent of f such

$$\limsup P_S^n f < K.$$

From (1) and (2) it follows immediately that

$$P_S^n f \geqslant \varepsilon 1_{[0,\varepsilon]} \qquad (\varepsilon = (2K)^{-1})$$

for n sufficiently large. Thus $h = \varepsilon 1_{[0,\varepsilon]}$ is a lower bound function for P_S and as a consequence of Theorem 2.3 we have the following:

PROPOSITION 3.2. *If a transformation* $S\colon [0,1] \to [0,1]$ *satisfies conditions* (C1)–(C3), *then the system* $\{S_n\}_{n \in N}$ *is statistically stable.*

Conditions (C1)–(C3) are automatically satisfied by the mappings $S^\beta(x) = \beta x (\mathrm{mod}\, 1)$ with $\beta > 1$. This particular class of transformations was studied by A. Rényi [28], A. O. Gelfond [7], W. Parry [26], V. A. Rochlin [29], R. L. Adler [1], M. Smorodinsky [31] and P. Walters [33], who proved a stability result stronger than the convergence in L^1.

As the last and most spectacular example of applications of the theory of statistical stability to deterministic systems we consider expanding mappings on manifolds.

Let M be a compact connected smooth (C^∞) manifold equipped with a Riemannian metric $|\cdot|$ and let μ be the corresponding Borel measure. Consider a C^2 *expanding* mapping $S\colon M \to M$. Thus we assume the existence of a constant $\lambda > 1$ such that

$$|dS(x)\,\xi| \geqslant \lambda |\xi| \qquad \text{for } x \in M,\ \xi \in T_x(M). \tag{3.2}$$

It is easy to estimate the logarithmic derivatives of $P_S^n f$. Namely there exists a constant K, independent of f, such that [16]

$$\limsup_{n \to \infty} \left\{ \max \left(\frac{1}{P_S^n f} |\mathrm{grad}\, P_S^n f| \right) \right\} < K \tag{3.3}$$

for every C^1 positive density f. From (3.3) it follows immediately that

$$P_S^n f \geqslant \frac{1}{\mu(M)}\, e^{-KL} \qquad (L = \mathrm{diam}\, M)$$

for sufficiently large n. Thus the constant function $h = e^{-KL}/\mu(M)$ is a lower bound function for P_S and from Theorem 2.3 we have the following

PROPOSITION 3.3. *If* $S: M \to M$ *is a* C^2 *transformation satisfying* (3.2), *then the system* $\{S_n\}_{n\in N}$ *is statistically stable.*

The ergodic properties of expanding mappings on manifolds were studied by A. Avez [2], W. Szlenk [15] and K. Krzyżewski [14] who, in particular, proved that they are exact. The proof of statistical stability based on different ideas was given by P. Walters [34].

Observe that in all our examples the fact that the properties of the functions $P_S^n f$ were "improving" for large n played a crucial role. Thus in the first case the variation of $P_S^n f$ was shrinking, in the second $P_S^n f$ became non-increasing and bounded, and in the third the logarithmic derivatives of $P_S^n f$ became bounded. The speed of this "improvement" in general depends on the choice of the initial density f. This is exactly what is required by the constrictivity condition.

4. Integral and differential equations

The technique developed in Section 2 can also be used for examining the behaviour of stochastic semigroups generated by some differential and integral equations. We shall demonstrate this by three examples in which our methods work equally as well as in the case of stochastic semigroups generated by expanding transformations. This shows once again a deep similarity of the so-called chaotic dynamical systems to the classical Markov processes.

Example 1. In the theory of cell proliferation there appears [18] a recurrence of the form

$$f_{n+1}(x) = Pf_n(x) \equiv \int_0^{2x} k(x, y) f_n(y)\, dy, \qquad x \geqslant 0, \tag{4.1}$$

where

$$k(x, y) = -\frac{d}{dx} \exp\left\{-\int_y^{2x} q(z)\, dz\right\}, \qquad x, y \geqslant 0 \tag{4.2}$$

and q is a continuous nonnegative function defined on $[0, \infty)$ satisfying $\limsup_{x\to\infty} q(x) > 0$. An easy calculation shows [18] that there exists a constant $c > 0$, independent of f, such that

$$\limsup_{n\to\infty} \int_0^{\infty} xP^n f(x)\, dx < c \tag{4.3}$$

for every $f \in D([0, \infty))$ with compact support. From (4.3) and the Chebyshev inequality it follows that

$$\int_0^{2c} P^n f(x)\,dx > \tfrac{1}{2} \tag{4.4}$$

for sufficiently large n $(n \geqslant n_0(f))$. Thus

$$P^{n+1}f(x) \geqslant \tfrac{1}{2} \inf_{y \leqslant c} k(x, y) = q(2x) \exp\Big\{-\int_0^{2x} q(z)\,dz\Big\}$$

$$\text{for } x \geqslant x, \quad n \geqslant n_0(f),$$

which shows that P has a non-trivial lower bound function. Thus, by Theorem 2.3, the semigroup $\{P^n\}_{n \in N}$ defined by (4.1), (4.2) is asymptotically stable, and for every initial density f_0 the sequence $\{P^n f_0\}$ converges to a unique (in D) solution f_* of the equation

$$f(x) = 2q(2x) \int_0^{2x} \exp\Big\{-\int_y^{2x} q(z)\,dz\Big\} f(y)\,dy.$$

The function f_* has interesting analytical properties. It satisfies a differential equation with an advanced argument and is flat at $x = 0$, i.e.,

$$\lim_{x \to 0} \big(x^{-n} f_*(x)\big) = 0 \quad \text{for } n \geqslant 0.$$

Example 2. Consider a linear version of the Boltzmann equation (see [16])

$$\frac{\partial u(t, x)}{\partial t} + u(t, x) = \int_0^{\infty} b(x, y) u(t, y)\,dy, \quad t, x \geqslant 0 \tag{4.5}$$

where $b(x, y)$ is a measurable non-negative kernel satisfying

$$\int_0^{\infty} b(x, y)\,dx = 1, \quad \int_0^{\infty} x b(x, y)\,dx \leqslant \alpha y + \beta, \quad y \geqslant 0, \tag{4.6}$$

where α, β are constants and $\alpha < 1$. We further assume that

$$\int_0^{\infty} \inf_{0 \leqslant y \leqslant r} b(x, y)\,dx > 0 \quad \text{for } r > 0. \tag{4.7}$$

All these conditions are automatically satisfied when b is given by (see [16])

$$b(x, y) = \begin{cases} -e^y Ei(-y) & \text{for } x \leqslant y, \\ -e^y Ei(-x) & \text{for } y > x. \end{cases}$$

Equation (4.5) with the initial condition

$$u(0, x) = f(x), \quad x \geqslant 0 \tag{4.8}$$

has exactly one solution u: $[0, \infty) \to L^1([0, \infty))$ for every $f \in L^1([0, \infty))$, and the formula $P^t f(x) = u(t, x)$ defines a stochastic semigroup. By using (4.6) it is easy to verify [16] that $\{P^t f\}$ satisfies an inequality similar to (4.3), namely

$$\limsup_{t\to\infty} \int_0^{2c} x P^t f(x)\, dx < c, \quad c = \frac{1+\beta}{1-\alpha}$$

for every $f \in D$ with compact support. Thus again by the Chebyshev inequality we have

$$\int_0^{2c} P^t f(x)\, dx \geqslant \tfrac{1}{2}$$

or sufficiently large t $(t \geqslant t_0(f))$. Now from this and (4.5) it follows that

$$P^t f(x) \geqslant \frac{1}{c} \int_0^{\infty} b(x, y) P^t f(y)\, dy \geqslant \frac{1}{2c} \inf_{y<2c} b(x, y) \quad \text{for } t \geqslant t_0(f)$$

which, according to (4.7), shows that $\{P^t\}_{t\geqslant 0}$ has a non-trivial lower function. Hence, by Theorem 2.3 the semigroup $\{P^t\}_{t\geqslant 0}$ is asymptotically stable.

Example 3. Consider a partial differential equation of the parabolic type

$$\frac{\partial u}{\partial t} = \sum_{ij=1}^{m} \frac{\partial}{\partial x_i}\left\{a_{ij}(x) \frac{\partial u}{\partial x_j}\right\} - \sum_{i=1}^{m} \frac{\partial}{\partial x_i} [b_i(x) u] \tag{4.9}$$

in the half space $t \geqslant 0$, $x \in R^m$. Assume that the coefficients are sufficiently smooth (for example a_{ij}, $\partial a_{ij}/\partial x_k$, b_i, $\partial b_i/\partial x_k$ bounded and uniformly Hölderian) and assume that the form $\sum a_{ij} \xi_i \xi_j$ is uniformly elliptic. Under these assumptions equation (4.9) with the initial condition

$$u(0, x) = f(x), \quad x \in R^m \tag{4.10}$$

has exactly one solution u: $[0, \infty) \to L^1(R^m)$ for every $f \in L^1(R^m)$, and the formula $P^t f(x) = u(t, x)$ defines a stochastic semigroup. This semigroup may or may not be asymptotically stable (see [8]). However, if there exist a ball $B \subset R^m$ and a constant $\varepsilon > 0$ such that the inequality

$$\liminf_{t\to\infty} \int_B u(t, x)\, dx \geqslant \varepsilon \tag{4.11}$$

is satisfied for every $f \in D$, then Theorem 2.3 implies the statistical stability of $\{P^t\}_{t\geqslant 0}$. In fact, under condition (4.11) the lower function for $\{P^t\}_{t\geqslant 0}$ is given by

$$h(x) = \varepsilon \inf_{y\in B} G(t_0, x, y),$$

where G is the Green's function corresponding to the Cauchy problem (4.9), (4.10) and t_0 is an arbitrary positive number.

References

[1] Adler R. L., F-expansions Revisited, *Lecture Notes Math.* **318** (1973), pp. 1–5.

[2] Avez A., Propriétés ergodiques des endomorphismes dilatantes des variétés compacts, *C. R. Acad. Sci. Paris Sér. A* **266** (1968), pp. 610–612.

[3] Bowen R., Invariant Measures for Markov Maps of the Interval, *Comm. Math. Phys.* **69** (1979), pp. 1–17.

[4] Boyarsky A., Randomness Implies Order, *J. Math. Anal. Appl.* **76** (1980), pp. 483–497.

[5] Collet P. and Eckmann J. P., Iterated Maps on the Interval as Dynamical Systems, *Progress in Physics* **1**, Birkhäuser, 1980.

[6] Foguel S. R., *The Ergodic Theory of Markov Processes*, Van Nostrand Reinhold, 1969.

[7] Gelfond A. O., On a General Property of Numbers Systems (in Russian), *Izv. Akad. Nauk SSSR* **23** (1959), pp. 809–814.

[8] Hasminskii R. Z., *Stochastic Stability of Differential Equations*, Sijhoff and Noordhoff, 1980.

[9] Jabłoński M., On Invariant Measures for Piecewise Convex Transformations, *Ann. Polon. Math.* **32** (1976), pp. 207–214.

[10] Keller G., Ergodicité et measures invariantes pour les transformations dilatantes par morceaux d'une région bornée du plan, *C. R. Acad. Sci. Paris. Sér A*, **289** (1979), pp. 635–627.

[11] Keller G., *Stochastic Stability in Some Chaotic Dynamical Systems*, preprint, Universität Heidelberg.

[12] Kowalski Z. S., Piecewise Monotonic Transformations and Their Invariant Measure, *Bull. Acad. Polon. Sci.* **27** (1979), pp. 63–69.

[13] Kosyaikin A. A. and Sandler E. A., Ergodic Properties of a Certain Class of Piecewise Smooth Transformations of a Segment, *Izv. Vyssh. Uchebn. Zaved. Matematika* **118** (1972), pp. 32–40.

[14] Krzyżewski K., Some Results on Expanding Mappings, *Astérisque* **50** (1977), pp. 205–218.

[15] Krzyżewski K. and Szlenk W., On Invariant Measures for Expanding Differentiable Mappings, *Studia Math.* **33** (1969), pp. 83–92.

[16] Lasota A., Statistical Stability of Deterministic Systems, *Lect. Notes Math.* **1107** (1983), pp. 386–419.

[17] Lasota A., Li T. Y., and Yorke J. A., *Asymptotic Periodicity of the Iterates of Markov Operators*, to appear in *Trans. AMS*.

[18] Lasota A. and Mackey M. C., Globally Asymptotic Properties of Proliferating Cell Populations, *J. Math. Biology* **19** (1984), pp. 43–62.
[19] Lasota A. and Yorke J. A., Exact Dynamical Systems and the Frobenius–Perron Operator, *Trans. AMS* **273** (1982), pp. 375–384.
[20] Lasota A. and Yorke J. A., *Statistical Periodicity of Deterministic Systems*, preprint, University of Maryland.
[21] Lasota A. and Yorke J. A., On the Existence of Invariant Measures for Piecewise Monotonic Transformations, *Trans. AMS* **186** (1973), pp. 481–488.
[22] Li T. Y. and Yorke J. A., Ergodic Transformations from an Interval into Itself, *Trans. AMS* **235** (1978), pp. 183–192.
[23] Lin M., Mixing for Markov Operators, *Z. Wahrsch. verw. Gebiete* **19** (1971), pp. 231–242.
[24] Markov A. A., *Wahrscheinlichkeitsrechnung*, B. G. Teubner, 1912.
[25] Misiurewicz M., Absolutely Continuous Measures for Certain Maps of an Interval, *Publ. Math. IHES* **53** (1981), pp. 17–51.
[26] Parry W., On the β-Expansion of Real Numbers, *Acta Math. Acad. Sci. Hungar.* **11** (1960), pp. 401–416.
[27] Pianigiani G., Existence of Invariant Measures for Piecewise Continuous Transformations, *Ann. Polon. Math.* **40** (1981), pp. 39–45.
[28] Rényi A., Representation of Real Numbers and Their Ergodic Properties, *Acta Math. Acad. Sci. Hungar.* **8** (1957), pp. 477–493.
[29] Rochlin V. A., Exact Endomorphisms of Lebesgue Spaces, *Izv. Acad. Nauk SSSR, Ser. Math.* **25** (1961), pp. 499–530.
[30] Schaefer H. H., On Positive Contractions in L^P Spaces, *Trans. AMS* **257** (1980), pp. 261–268.
[31] Smorodinsky M., β-Automorphisms Are Bernoulli Shifts, *Acta Math. Acad Sci. Hungar.* **24** (1973), pp. 272–278.
[32] Wagner G., The Ergodic Behavior of Piecewise Monotonic Transformations, *Z. Wahrsch. verw. Gebiete* **46** (1979), pp. 217–324.
[33] Walters P., Equilibrium States for β-Transformations and Related Transformations, *Math. Z.* **159** (1978), pp. 65–88.
[34] Walters P., Invariant Measures and Equilibrium States for Some Mappings with Expanding Distances, *Trans. AMS* **236** (1978), pp. 121–153.
[35] Wong S., Some Metric Properties of Piecewise Monotonic Mappings of the Unit Interval, *Trans. AMS* **136** (1978), pp. 493–500.

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

RICARDO MAÑÉ

Oseledec's Theorem from the Generic Viewpoint

Let f be a C^1 diffeomorphism of a compact boundaryless manifold M. A point $x \in M$ is said to be *regular* if the sequence of derivatives at x of the iterates of f admits the following description: there exists a splitting $T_xM = \bigoplus_{i=1}^{m} E_i(x)$ (the Lyapunov splitting at x) and numbers $\lambda_1(x) > \ldots > \lambda_m(x)$ (the Lyapunov exponents at x) such that $\lim_{n\to\pm\infty} n^{-1}\log\|(D_xf^n)v\| = \lambda_i(x)$ for every $1 \leqslant i \leqslant m$ and $0 \neq v \in E_i(x)$. The dimension of $E_i(x)$ is called the *multiplicity* of x. It is easy to check that the Lyapunov splitting and exponents are unique and that if x is regular so is $f(x)$ with Lyapunov splitting $T_{f(x)}M = \bigoplus_{i=1}^{m} (D_xf)E_i(x)$ and exponents $\lambda_i(f(x)) = \lambda_i(x), 1 \leqslant i \leqslant m$

Denote by $\Lambda(f)$ the set of regular points. In the late sixties Oseledec proved that the set $\Lambda(f)$ has *total probability*, i.e. $\mu(\Lambda(f)) = 1$ for every f-invariant probability measure μ on the Borel σ-algebra of M [4]. Since then this theorem evolved into one of the central tools in smooth ergodic theory. Recently Oseledec's theorem was extended to compact maps of infinite dimensional Banach manifolds (Ruelle [6], Mañé [3]). This extension can be useful for the analysis of the dynamical systems generated by retarded functional differential equation or semilinear parabolic P.D.E's.

Oseledec's theorem is essentially a measure theoretical result and therefore the information it provides holds only in that category. For instance, the Lyapunov splitting is just a measurable function of the point and the limits defining the Lyapunov exponents are not uniform. It is clear that this is not a defficiency of the theorem but the natural counterweight to its remarkable generality. However, one can pose the problem, and this is the purpose of this exposition, of whether these aspects can be substantially improved by working under generic conditions. There are two approaches to this problem. One is to study Oseledec's theorem for

generic invariant measures of generic diffeomorphisms. The other, to study Oseledec's theorem with respect to Lebesgue measure and generic volume preserving diffeomorphisms. Let us introduce some notation and terminology. Denote $\mathrm{Diff}^1(M)$ the space of C^1 diffeomorphisms of M endowed with the C^1-topology and let $\mathscr{M}$ be the space of probability measures on the Borel σ-algebra of M endowed with the weak topology. Let $d(\cdot, \cdot)$ be a metric in $\mathscr{M}$ associated to this topology. The support of a measure μ will be denoted by $s(\mu)$. Define $\mathscr{M}(f)$ to be the set of f-invariant elements of $\mathscr{M}$ and $\mathscr{M}_e(f)$ to be the set of ergodic measures in $\mathscr{M}(f)$. If $f \in \mathrm{Diff}^1(M)$ and $\gamma = \{x = f^m(x), f(x), \ldots, f^{m-1}(x)\}$ is a periodic orbit of f, denote by μ_γ the measure $\mu_\gamma = m^{-1} \sum_{j=0}^{m-1} \delta_{f^j(x)}$. Recall that a subset of a topological space is residual (or generic) if it contains the intersection of a countable family of open dense subsets. A Baire space is a topological space such that every residual subset is dense. $\mathscr{M}$ and $\mathscr{M}(f)$ are compact. $\mathscr{M}_e(f)$ is a Baire space (in fact, a residual subset of its closure).

Our main tool in developing the first approach described above is the following theorem; in the next section we shall show that it follows from the ergodic version of Pugh's Closing Lemma which we proved in [2]:

THEOREM A. *Given $f \in \mathrm{Diff}^1(M)$, $\varepsilon > 0$ and a neighborhood $\mathscr{U}$ of f, there exist $g \in \mathscr{U}$ and a g-periodic orbit γ such that $d(\mu, \mu_\gamma) < \varepsilon$.*

Applying standard methods involving upper semicontinuous set valued functions together with Theorem A, it is easy to prove the following corollary:

COROLLARY. *If $\mathscr{M}_p(f)$ denotes the set of measures μ_γ, where γ is a periodic orbit of $f \in \mathrm{Diff}^1(M)$, then $\mathscr{M}_p(f)$ is dense in $\mathscr{M}_e(f)$ for a residual set of diffeomorphisms f.*

If $f \in \mathrm{Diff}^1(M)$ and Λ is a compact invariant set we say that a splitting $TM/\Lambda = E_1 \oplus \ldots \oplus E_m$ is a *dominated splitting* if it is continuous, f-invariant (i.e. invariant under the derivative of f) and there exists $C > 0$, $0 < \lambda < 1$ such that, if $1 \leqslant i \leqslant m$, setting $E^+ = \bigoplus_{j \leqslant i} E_j$ and $E^- = \bigoplus_{j > i} E_j$ we have

$$\|(D_x f^n)/E^+_x\| \cdot \|(D_{f^n(x)} f^{-n})/E^-_{f^n(x)}\| \leqslant C\lambda^n$$

for all $x \in \Lambda$, $n \geqslant 0$.

In the next section we shall prove the following theorem:

THEOREM B. *For generic diffeomorphisms* $f \in \mathrm{Diff}^1(M)$, *there exists a residual set* $\mathscr{D}(f) \subset \mathscr{M}_e(f)$ *such that, if* $\mu \in \mathscr{D}(f)$, *there exists a dominated splitting* $TM/s(\mu) = \bigoplus_{i=1}^{m} E_i$ *which coincides with the Lyapunov splitting at* μ*-a.e. point of* $s(\mu)$.

Using the domination property it is possible to prove the existence of stable and unstable manifolds for a.e. point with respect to measures in $\mathscr{D}(f)$, even if f is only C^1 and not $C^{1+\varepsilon}$, as required in general. This follows from the results of Hirsch, Shub and Pugh [1].

Theorem B has a serious defficiency, namely, the fact that for a C^1 generic diffeomorphism, the entropy is zero with respect to generic invariant measures. This property follows easily from the methods used by Sigmund in [7] to prove the same statement for Axiom A, replacing the specification property by the corollary of Theorem A. This means that generic elements of $\mathscr{M}_e(f)$ fail to reflect the dynamic complexity of f. To avoid this problem one should work in the space $\mathscr{M}_e^c(f) = \{\mu \in \mathscr{M}_e(f) | \; |h_\mu(f) > c\}$, proving that generic measures in $\mathscr{M}_e^c(f)$ satisfy a strong form of Oseledec's theorem. So far we have obtained no results in this direction.

Now let us consider the second approach we proposed: Oseledec's theory for generic volume preserving diffeomorphisms. For technical reasons which we discuss later we shall work with sympletic instead of volume preserving diffeomorphism. Given $f \in \mathrm{Diff}^1(M)$ and $x \in \Lambda(f)$, define $E^s(x) = \oplus \{E_i(x)\colon \lambda_i(x) < 0\}$, $E^u(x) = \oplus \{E_i(x)\colon \lambda_i(x) > 0\}$, $E^c(x) = E_j(x)$ if $\lambda_j(x) = 0$. When f is sympletic its Lyapunov exponents have the following symmetry property: If $\lambda_i(x)$ is a Lyapunov exponent then $-\lambda_i(x)$ is also a Lyapunov exponent with the same multiplicity. Thus $\dim E^s(x) = \dim E^u(x)$. Define the elliptic, hyperbolic and partially hyperbolic regions respectively by $\Lambda_E(f) = \{x \in \Lambda(f)\colon E^c(x) = T_xM\}$, $\Lambda_H(f) = \{x \in \Lambda(f)\colon T_xM = E^s(x) \oplus E^u(x)\}$, $\Lambda_P(f) = \{x \in \Lambda(f)\colon 0 < \dim E^c(x) < \dim M\}$. Let $\mathrm{Symp}^1(M, \omega_0)$ be the space of symplectic diffeomorphisms of the symplectic compact manifold (M, ω_0) and let λ be the Lebesgue measure (associated to the volume form $\omega = \omega_0 \wedge \overset{k}{\ldots} \wedge \omega_0$, $k = \dim M/2$).

THEOREM C. *There exists a residual set of* $\mathrm{Symp}^1(M, \omega_0)$ *whose elements satisfy one of the following properties:*

(a) $\lambda(\Lambda_E(f)) = 1$,

(b) f *if Anosov (and then* $\Lambda_E(f) = \Lambda_P(f) = \emptyset$),

(c) $\lambda(\Lambda_H(f)) = 0$, $\lambda(\Lambda_P(f)) > 0$ *and for all* $\varepsilon > 0$ there *exists a Borel* f*-invariant set* $\Lambda \subset \Lambda_P(f)$ *with* $\lambda(\Lambda_P(f) - \Lambda) < \varepsilon$ *and such that there exists*

a dominated splitting of $TM/\overline{\Lambda}$ which coincides with the Lyapunov splitting at a.e. point of Λ.

Moreover, Λ is a uniformly partially hyperbolic set, i.e. there exist $C>0$ and $0<\lambda<1$ such that

(1) $\|(D_x f^n)/E^s(x)\| \leqslant C\lambda^n$,

(2) $\|(D_x f^{-n})/E^u(x)\| \leqslant C\lambda^n$

for all $x \in \Lambda$, $n \geqslant 0$.

The following corollary is interesting:

COROLLARY. *C^1-generically, area preserving diffeomorphisms of compact two-dimensional manifolds are either Anosov or satisfy $\lim_{n\to\pm\infty} n^{-1}\log\|(D_x f^n)v\| = 0$ for λ-a.e. x and every $0 \neq v \in T_xM$.*

We do not know (but this seems likely) if C holds for volume preserving diffeomorphisms. The symplectic property has two important features in its proof. First, it makes possible to deduce hyperbolicity conditions ((1) and (2) in (c)) from the domination property, that is what we can actually prove. Second, during the proof we need to approximate a C^1-symplectic diffeomorphism by a C^2-symplectic diffeomorphism. This is possible in the symplectic case (Zehnder [8]) but is unknown in the volume preserving case. The proof of Theorem C, much longer and more complicated than that of B, will appear elsewhere.

Proofs of Theorems A *and* B. If $f \in \mathrm{Diff}^1(M)$, denote by $\Sigma(f)$ the set of points $x \in M$ such that for every neighborhood $\mathscr{U}$ of x and all $\varepsilon > 0$ there exist $g \in \mathscr{U}$ and a g-periodic point y such that $d(f^j(x), g^j(y)) < \varepsilon$ for all $0 \leqslant j \leqslant m$ where m is the g-period of y. It is conjectured that $\Sigma(f)$ coincides with the set of recurrent points of f. For our purpose, it will be sufficient, combined with the result proved in [2], that $\Sigma(f)$ is a total probability set. Now suppose that $\mu \in \mathscr{M}_e(f)$ and that $\varphi\colon M \to R$ is a continuous function. Since $\mu(\Sigma(f)) = 1$, we can take $x \in \Sigma(f)$ such that

$$\lim_{n\to\pm\infty} \frac{1}{n}\sum_{j=0}^{n-1} \varphi(f^j(x)) = \int_M \varphi\, d\mu.$$

Given $\varepsilon > 0$, we can choose $N > 0$ such that

$$\left|\frac{1}{n}\sum_{j=0}^{n-1} \varphi(f^j(x)) - \int_M \varphi\, d\mu\right| < \varepsilon/2 \tag{1}$$

if $n \geqslant N$. If $\mathscr{U}$ is a neighborhood of f, there exists, by the definition of

$\Sigma(f)$, a diffeomorphism $g \in \mathscr{U}$ and a g-periodic point y such that $d\big(f^j(x), g^j(x)\big)$ is so small for all $0 \leqslant j \leqslant m$ (where m is the g-period of y) that

$$\left| \frac{1}{m} \sum_{j=0}^{m-1} \varphi\big(g^j(y)\big) - \frac{1}{m} \sum_{j=0}^{m-1} \varphi\big(f^j(x)\big) \right| < \varepsilon/2. \tag{2}$$

Moreover, observe that if x is f-periodic, there is nothing to prove, and that if it is not, then m is very large if g and y are close enough to f and x. Therefore we can suppose that $m \geqslant N$. Then (1) and (2) imply (denoting $\gamma = \{x, g(x), \ldots, g^{m-1}(x)\}$)

$$\left| \int_M \varphi \, d\mu_\gamma - \int_M \varphi \, d\mu \right| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

This concludes the proof of Theorem A. To prove B we shall use the following elementary lemma:

LEMMA I. *Let F be a Baire space, K a compact space and $S \subset F \times K$ a subset such that the function $F \ni x \to (\{x\} \times K) \cap S$ has a residual set of points of lower semicontinuity. Then, if $S_0 \subset S$ is a residual subset of S (in the relative topology of S), $S_0 \cap (\{x\} \times K)$ is a residual subset of $S \cap (\{x\} \times K)$ for a residual set of values of $x \in F$.*

We shall apply this lemma to $\mathrm{Diff}^1(M) \times \mathscr{M}$ and the subset $S = \{(f, \mu)$: $f \in \mathrm{Diff}^1(M)$, $\mu \in \mathscr{M}_e(f)\}$. Observe that at points $f \in \mathrm{Diff}^1(M)$ where every periodic point is hyperbolic and $\mathscr{M}_p(f)$ is dense in $\mathscr{M}_e(f)$, the function $\mathrm{Diff}^1(M) \ni f \to S \cap (\{f\} \times \mathscr{M}) = \mathscr{M}_e(f)$ is lower semicontinuous. This set of points is residual in $\mathrm{Diff}^1(M)$ by the corollary of Theorem A. Then we can apply the lemma and the proof of B is reduced to finding a residual subset $S_0 \subset S$ such that, if $(f, \mu) \in S_0$ then μ satisfies the properties in Theorem B. To find S_0, we define a function Φ: $S \to \boldsymbol{R}^l$, where $l = \dim M$, which associates with each $(f, \mu) \in S$ the vector $\big(\lambda_1(f, \mu), \ldots, \lambda_l(f, \mu)\big)$, where $\lambda_1(f, \mu) \geqslant \ldots \geqslant \lambda_l(f, \mu)$ are the Lyapunov exponents of f with respect to μ, repeated according to their multiplicities. Denote by $D^k f$: $TM \wedge \overset{k}{\ldots} \wedge TM$ the k-th exterior power of Df and define

$$\Phi_k(f, \mu) = \inf_{n \geqslant 1} \frac{1}{n} \int_M \log \|(D^k f)\| \, d\mu, \qquad k = 1, 2, \ldots$$

It is known that

$$\Phi_k(f, \mu) = \sum_{j=1}^{k} \lambda_j(f, \mu).$$

Then

$$\Phi(f,\mu) = \big(\Phi_1(f,\mu), \Phi_2(f,\mu)-\Phi_1(f,\mu), \ldots, \Phi_l(f,\mu)-\Phi_{l-1}(f,\mu)\big). \quad (3)$$

Moreover, the functions Φ_k are upper-semicontinuous because they are lower bounds of continuous functions. Consequently there exists a residual set $S_k \subset S$ of points of continuity of Φ_k. Write $S_0 = \bigcap_{k\geqslant 1} S_k$. By (3) Φ is continuous at every point of S_0. Now we need the following lemma:

LEMMA II. *Let Λ be a compact invariant set of $f \in \mathrm{Diff}^1(M)$ such that there exists a sequence of diffeomorphisms $\{f_n\colon n \in \mathbb{Z}^+\}$ and periodic orbits γ_n of f_n, with period m_n, satisfying the following properties:*

(a) *For every $x \in \Lambda$ there exist $x_n \in \gamma_n$, $n = 1, 2, \ldots$ such that $\lim x_n = x$;*

(b) *There exist $\beta > 0$ and a neighborhood $\mathcal{U}$ of f such that if $g \in \mathcal{U}$ and, for some n, γ_n is a periodic orbit of g, then the eigenvalues of the Poincaré map of g at γ_n have moduli $\neq \beta^{m_n}$;*

(c) *There exists an integer $k(\beta)$ such that, for all n, the dimension of the subspace associated with the eigenvalues of modulus $< \beta^{m_n}$ is $k(\beta)$.*

Then there exists a dominated splitting $TM/\Lambda = E^+ \oplus E^-$ such that $\dim E^+_x = k(\beta)$ for all $x \in \Lambda$.

The proof of this lemma is a direct application of the results of [2], Section 3. In fact, in [2] we worked with 1 playing the role of the constant β in Lemma II, but it is clear that the method of construction of dominated splittings developed in [2] works also in this case.

Now suppose that $(f,\mu) \in S_0$. Let $\lambda_1 > \ldots > \lambda_m$ be the Lyapunov exponents of f with respect to μ and let $k_1, \ldots, k_m$ be their multiplicities. Take numbers $\beta_1 > \ldots > \beta_{m-1}$ separating the Lyapunov exponents of f with respect to μ (i.e. $\lambda_1 > \beta_1 > \lambda_2 > \beta_2 > \ldots > \lambda_{m-1} > \beta_{m-1} > \lambda_m$). Let $f_n \to f$ be a sequence of diffeomorphisms with periodic orbits γ_n such that $\mu_{\gamma_n} \to \mu$. Such sequences exist by Theorem A. Then $\Phi(f_n, \mu_{\gamma_n})$ is near to $\Phi(f,\mu)$ for large values of n. But it is easy to see that $\Phi(f_n, \mu_{\gamma_n}) = (|\alpha_1|^{1/m_n}, \ldots, |\alpha_l|^{1/m_n})$ where m_n is the f_n-period of γ_n and $\alpha_1, \ldots, \alpha_l$ are the eigenvalues of the Poincaré map of f_n at γ_n repeated according to their multiplicities and ordered so that $|\alpha_1| \geqslant \ldots \geqslant |\alpha_l|$. This means that $\beta_1^{m_n} > \beta_2^{m_n} > \ldots > \beta_{m-1}^{m_n}$ are not moduli of eigenvalues of the Poincaré map of f_n at γ_n, and moreover, the dimension of the subspace associated with the eigenvalues of modulus $< \beta_i^{m_n}$ of the Poincaré map of f_n at γ_n is $\sum_{j>i} k_j$. Since this holds for any sequence of diffeomorphisms $f_n \to f$ and periodic orbits γ_n satisfying $\mu_{\gamma_n} \to \mu$, it follows that if we fix one such sequence then this sequence satisfies the hypothesis of Lemma II for

$\beta = \beta_i$ and $k(\beta) = \sum_{j>i} k_j$ for every $1 \leqslant i \leqslant m-1$. Thus fix i and apply Lemma II to $\beta = \beta_i$ and $k(\beta) = \sum_{j>i} k_j$. We obtain a dominated splitting $TM/s(\mu) = E^+ \oplus E^-$ with $\dim E_x^+ = k(\beta)$ for all $x \in s(\mu)$. We claim that $E_x^+ = E_{i+1}(x) \oplus \ldots \oplus E_m(x)$ for μ-a.e. $x \in s(\mu)$. To simplify the notation set $S(x) = E_{i+1}(x) \oplus \ldots \oplus E_m(x)$, $G(x) = E_1(x) \oplus \ldots \oplus E_i(x)$. Take a Borel set $K \subset \Lambda(f)$ such that the subspaces $S(x)$ and $G(x)$ are continuous functions of x when x varies in K and for all $x \in K$ there exists a sequence $n_j \to +\infty$ such that $f^{n_j}(x) \in K$ for all j and $\lim_{j \to +\infty} f^{n_j}(x) = x$. Since there exist sets with this property and μ-measure arbitrarily close to 1, it is sufficient, in order to prove the claim, to show that $S(x) = E_x^+$ and $G(x) = E_x^-$ when $x \in K$. Write $S(x) = S^+ \oplus S_0$, $G(x) = G^- \oplus G_0$ where $S^+ \subset E_x^+$, $G^- \subset E_x^-$, $S_0 \cap E_x^+ = \{0\}$, $G_0 \cap E_x^- = \{0\}$. We know that

$$\lim_{n \to +\infty} \|(D_x f^n)/E_x^+\| \cdot \|(D_{f^n(x)} f^{-n})/E_{f^n(x)}^-\| = 0 \tag{1}$$

$$\lim_{n \to -\infty} \|(D_x f^{-n})/E_x^-\| \cdot \|(D_{f^{-n}(x)} f^n)/E_{f^{-n}(x)}^+\| = 0. \tag{2}$$

Choose a sequence n_j such that $f^{n_j}(x) \to x$ if $j \to +\infty$ and $f^{n_j}(x) \in K$ for all j. From (1) it follows that $(D_x f^{n_j}) S_0$ converges to a subspace $S_\infty \subset E_x^-$. By the continuous dependence of $S(x)$ when x varies in K, this implies that

$$S(x) = S^+ \oplus S_\infty.$$

In a similar way we prove that

$$G(x) = G^- \oplus G_\infty$$

with $G_\infty \subset E_x^+$. But using the definition of $S(x)$, $G(x)$ and the Lyapunov exponents we obtain

$$\lim_{n \to +\infty} \|(D_x f^n)/S(x)\| \cdot \|(D_{f^n(x)} f^{-n})/G(x)\| = 0$$

or

$$\lim_{n \to +\infty} \frac{\|(D_x f^n) v\|}{\|(D_x f^n) w\|} = 0$$

if $0 \neq v \in S(x)$, $0 \neq w \in G(x)$. However, if we take $v \in S_\infty \subset E_x^-$ and $w \in G_\infty \subset E_x^+$, it follows from (1) that this limit is ∞. This contradiction shows that we must have $S_\infty = \{0\}$, $G_\infty = \{0\}$ and this will only happen if $S_0 = \{0\}$, $G_0 = \{0\}$. This means that $S(x) \subset E_x^+$, $G(x) \subset E_x^-$ and, because of the dimensions of these subspaces, we conclude $S(x) = E_x^+$, $G(x) = E_x^-$. This proves the claim. Applying this property to all the β_i's from $i = 1$ to $i = m-1$, we prove that μ satisfies the assertion of the theorem.

References

[1] Hirsch M., Shub M. and Pugh C., *Invariant Manifolds*, Lecture Notes in Mathematics 583, Springer-Verlag, 1977.

[2] Mañé R., An Ergodic Closing Lemma, *Ann. of Math.* **116** (1982), pp. 503–541.

[3] Mañé R., Lyapounov Exponents and Stable Manifolds for Compact Transformations, to appear in the *Proceedings of the Symposium on Dynamical Systems, Rio de Janeiro 1981*, Lecture Notes in Mathematics, Springer-Verlag.

[4] Oseledec V. I., A Multiplicative Ergodic Theorem, *Trans. Moscow Math. Soc.* **19** (1968), pp. 197–231.

[5] Pesin Ya., Invariant Manifolds Associated to Non Vanishing Exponents, *Math. USSR* **10** (1976).

[6] Ruelle D., Characteristic Exponent and Invariant Manifolds in Hilbert Spaces, *Ann. of Math.* **115** (1982).

[7] Sigmund K., Generic Properties for Axiom A Diffeomorphisms, *Inventiones Mathematicae* **11** (1970), pp. 99–109.

[8] Zehnder E., Note on Smoothing Symplectic and Volume Preserving Diffeomorphisms, *Geometry and Topology*, Lecture Notes in Mathematics 597, pp. 828–855. Springer-Verlag, 1977.

INSTITUTO DE MATEMÁTICA PURA E APLICADA (IMPA)
RIO DE JANEIRO,
RJ, BRAZIL

Proceedings of the International Congress of Mathematicians
August 16-24, 1983, Warszawa

MICHAŁ MISIUREWICZ

One-Dimensional Dynamical Systems

By a *one-dimensional dynamical system* I mean a continuous map $f\colon X \to X$ where X is either an interval or a circle. We study its iterates f^n and the asymptotic behaviour of orbits as n tends to infinity. Such systems can display a surprisingly wide variety of types of behaviour. Most of the interesting effects arising for flows and homeomorphisms in higher dimensions are present in the case of maps in one dimension. On the other hand, the existence of an ordering of the real line allows us to use some specific methods in one dimension.

We can look at dynamical systems from different points of view. One of them I would call *topological*. We try to determine the structure of the map, look at periodic points, non-wandering points and all possible asymptotic behaviours of the orbits. The other point of view can be called *physical*. We are interested only in those types of behaviour which are present for a set of orbits of positive Lebesgue measure. A periodic orbit is important only if it is attracting. Of course, there are many problems equally interesting from both points of view.

In this paper, to illustrate the types of problems and results which arise for one-dimensional dynamical systems, I shall concentrate on the case of unimodal maps of an interval. For a broader treatment of this and other subjects, see e.g. [2], [19], [17]. Let $I = [0, 1]$, let $f\colon I \to I$ be a continuous map such that $f(0) = f(1) = 0$ and, for some $c \in (0, 1)$, f is increasing on $[0, c]$ and decreasing on $[c, 1]$. We shall call such a map *unimodal*. The natural way to study the topological properties of such a system is to use symbolic dynamics. This can be done by coding. To every $x \in I$ we assign a sequence of symbols $(\varepsilon_i)_{i=0}^{\infty}$ by setting $\varepsilon_i = +1$ if $f^i(x) \in [0, c]$ and $\varepsilon_i = -1$ if $f^i(x) \in [c, 1]$. (There is an ambiguity if $f^i(x) = c$, but we shall omit here some problems caused by this.) The set of points corresponding to a given sequence can either be empty or consist of one point or be an interval (such an interval is called a *homterval*, since all

iterates of f restricted to it are homeomorphisms). The shift in the space of sequences corresponds to the map f. Therefore, up to homtervals, we can study the properties of our system by studying the symbolic system. The power of this fairly general method in the case of unimodal maps is based on the fact that the sequence corresponding to c (the so-called *kneading sequence* of f, denoted by $K(f)$) determines the whole symbolic system. These ideas, already present in Myrberg's paper [18], were developed by Milnor and Thurston [14] and then by other authors to form the so-called kneading theory. It describes all possible sequences of the symbolic system if a kneading sequence is given, all possible kneading sequences, and the dependence of a sequence on a point and of a kneading sequence on a parameter in one-parameter families of maps.

Here I have to remark that the method described above is not the only possible way of coding. F. Hofbauer [7] introduced another method (the intervals corresponding to the symbols have intersections with non-empty interiors, there are infinitely many of them, but one gets a subshift of finite type) and, in subsequent papers, obtained very interesting and general results.

The next step in the kneading theory is to decompose some kneading sequences [10]. A situation where a map f has a decomposable $K(f)$ has the following geometric interpretation [6]. There is an interval I_1 containing c in its interior such that, for some n, the intervals $I_1, f(I_1), \ldots, f^{n-1}(I_1)$ are pairwise disjoint, $f^n(I_1) \subset I_1$ and $f_1 = \varphi^{-1} \circ f^n|_{I_1} \circ \varphi$ (where φ is an affine map from I onto I_1) is again unimodal. Now $K(f_1)$ can perhaps be also decomposed (we then get I_2 and f_2), etc. Thus, we obtain 3 types of kneading sequences:

1. periodic,
2. aperiodic, finitely decomposable,
3. aperiodic, infinitely decomposable.

To avoid complications caused by homtervals we shall assume additionally that f is of class C^3, the second derivative of f is negative and the Schwarzian derivative of f, $Sf = f'''/f' - \frac{3}{2}(f''/f')^2$, is negative (the condition of a negative Schwarzian derivative was introduced for interval maps by Singer [21]). Such a function is called *S-unimodal*. The best known examples of such maps are quadratic maps $f_a(x) = ax(1-x)$. Two S-unimodal maps with the same kneading sequence are topologically conjugate (except the casew hen an attracting periodic orbit exists; it can be attracting from one side or from both sides). The kneading theory allows us to describe completely the structure of an S-unimodal

map f: $I \to I$. If $K(f)$ is periodic, then f has one periodic attracting orbit. This orbit attracts almost all (in the sense of Lebesgue measure) points of I. If $K(f)$ is aperiodic, then there are no periodic attracting orbits and no homtervals. If $K(f)$ is aperiodic and non-decomposable, then the set of non-wandering points $\Omega(f)$ consists of a fixed point 0 and an interval $[f^2(c), f(c)]$. If $K(f)$ is decomposable, then every step of decomposition gives some invariant totally disconnected closed set as a subset of $\Omega(f)$. What remains in $\Omega(f)$ depends on the length of the decomposition. If it is finite, then for some n we have f_n with $K(f_n)$ non-decomposable, and we use the previous description. If it is infinite, we are left with some set Ω_∞ homeomorphic to the Cantor set, on which f is a homeomorphism conjugate to some generalized adding machine. There is a conjecture that the Lebesgue measure of Ω_∞ is zero. An affirmative answer to this conjecture in some specific situations follows from the Feigenbaum theory [4], [5], [11], [12] (which deals in particular with the lengths of the intervals I_k). Without the assumption of $f''(c) < 0$ and $Sf < 0$, the measure of Ω_∞ can be positive.

The difference between the two points of view may be illustrated by the example of notions of chaotic behaviour. From the topological point of view, chaos is the existence of uncountably many different asymptotic behaviours of orbits [12]. This is equivalent to the existence of periodic points of periods different from powers of 2, and to the positivity of topological entropy [12], [20], [15]. For a piecewise monotone map f of an interval, topological entropy $h(f)$ can be defined as $\lim_n \frac{1}{n} \log$ (the number of pieces of monotonicity of f^n). In particular, if $K(f)$ is periodic with period different from powers of 2, f is chaotic in the above sense. Nevertheless, almost every trajectory is attracted by the same periodic orbit. Therefore, from the physical point of view it is not chaotic. From this point of view, chaos means the existence of an ergodic invariant probabilistic measure, absolutely continuous with respect to the Lebesgue measure. If such a measure exists, then, by Birkhoff's ergodic theorem, almost every trajectory is distributed uniformly with respect to the density of that measure.

Since topological entropy depends only on $K(f)$, so does the chaotic behaviour in the topological sense. Assume that f is S-unimodal. It is not known whether the existence of absolutely continuous invariant measure depends only on the kneading sequence. There is a conjecture that such a measure cannot exist for f with $K(f)$ infinitely decomposable

If c does not belong to the closure of $\{f^n(c)\}_{n=1}^{\infty}$, then such a measure exists [16]. For one-parameter families of maps of the form $f_\lambda(x) = \lambda \cdot f(x)$, the set of parameters λ for which such a measure exists, has a positive Lebesgue measure [9], [1].

From among many other problems concerning S-unimodal maps, let me mention two more.

Intuitively, it seems almost obvious that in every C^r-neighbourhood of an S-unimodal map without a periodic attractor there is a map with a different kneading sequence. However, this is known only for $r \leqslant 1$ [8]. For larger r, attempts at the proof meet obstructions similar to those for C^r closing lemma.

The second problem is whether $h(f_\lambda)$ is a non-decreasing function of λ for $f_\lambda(x) = \lambda \cdot f(x)$. All computer experiments show that it is. However, in the general case only some partial results have been obtained [13], [22]. A complete solution exists only for the case of quadratic maps. It follows from the results obtained for the family of quadratic maps of C, which depend stronlgy on the properties of these complex maps [3]. If this method could be applied to other families, it would fit very nice into the kneading theory. For complex quadratic maps one can also define a kneading sequence. It turns out that those sequences which cannot occur as kneading sequences in the real case can nevertheless appear in the complex case.

References

[1] Benedicks M. and Carleson L., *On Iterations of* $1-ax^2$ *on* $(-1, 1)$, preprint.

[2] Collet P. and Eckmann J.-P., Iterated Maps on the Interval as Dynamical Systems, *Progr. Phys.* **1**, Birkhäuser, Boston, 1980.

[3] Douady A., Systèmes dynamiques holomorphes, *Semin. Bourbaki* **599** (1982).

[4] Feigenbaum M., Quantitative Universality for a Class of Non-Linear Transformations, *J. Stat. Phys.* **19** (1978), pp. 25–52.

[5] Feigenbaum M., The Universal Metric Properties of Non-Linear Transformations, *J. Stat. Phys.* **21** (1979), pp. 669–706.

[6] Guckenheimer J., Sensitive Dependence on Initial Conditions for One Dimensional Maps, *Commun. Math. Phys.* **70** (1979), pp. 133–160.

[7] Hofbauer F., On Intrinsic Ergodicity of Piecewise Monotonic Transformations with Positive Entropy, *Israel J. Math.* **34** (1979), pp. 213–237.

[8] Jakobson M., On Smooth Mappings of the Circle into Itself, *Mat. Sbornik* **85** (1971), pp. 163–188.

[9] Jakobson M., Absolutely Continuous Invariant Measures for One Parameter Families of One-Dimensional Maps, *Commun. Math. Phys.* **81** (1981), pp. 39–88.

[10] Jonker L. and Rand D., Bifurcations in One Dimension, I: The Nonwandering Set, *Invent. Math.* **62** (1981), pp. 347–365.

[11] Lanford O. III, Smooth Transformations of Intervals, *Semin. Bourbaki* **563** (1980).
[12] Li T.-Y. and Yorke J., Period Three Implies Chaos, *Amer. Math. Monthly* **82** (1975), pp. 985–992.
[13] Matsumoto S., *On the Bifurcation of Periodic Points of One Dimensional Dynamical, Systems of a Certain Kind*, preprint.
[14] Milnor J. and Thurston W., *On Interated Maps of the Interval*, preprint, Princeton, 1977.
[15] Misiurewicz M. and Szlenk W., Entropy of Piecewise Monotone Maps, *Studia Math.* **67** (1980), pp. 45–63 (short version — *Astérisque* **50** (1977), pp. 299–310).
[16] Misiurewicz M., Absolutely Continuous Measures for Certain Maps of an Interval, *Publ. Math. IHES* **53** (1981), pp. 17–51.
[17] Misiurewicz M., Maps of an Interval. In: *Les Houches 1981 — Chaotic Behaviour of Deterministic Systems*, North-Holland, Amsterdam, 1983.
[18] Myrberg P. J., Iteration der reelen Polynome zweiten Grades III, *Ann. Acad. Sci. Fenn.* **336/3** (1963), pp. 1–18.
[19] Nitecki Z., Topological Dynamics on the Interval. In: *Ergodic Theory and Dynamical Systems II, Maryland 1979–80*, Birkhäuser, Boston, 1982, pp. 1–73.
[20] Sharkovskii A. N., Coexistence of Cycles of a Continuous Map of the Line into Itself (in Russian), *Ukr. Mat. Zh.* **16** (1964), pp. 61–71.
[21] Singer D., Stable Orbits and Bifurcations of Maps of the Interval *SIAM J. Appl. Math.* **35** (1978) pp. 260–267.
[22] Zdunik A., *Entropy of Transformations of the Unit Interval*, preprint.

Proceedings of the International Congress of Mathematicians
August 16-24, 1983, Warszawa

GEORGE R. SELL

Linearization and Global Dynamics*

In this paper we show how the spectral theory of linear skew-product flows may be used to study the following three questions in the qualitative theory of dynamical systems: (1) When is an ω-limit set or an attractor a manifold? (2) Under which conditions will a dynamical system undergo a Hopf–Landau bifurcation from a k-dimensional torus to a $(k+1)$-dimensional torus? (3) When is a vector field in the vicinity of a compact invariant manifold smoothly conjugate to the linearized vector field and how smooth is the conjugacy?

I. Introduction

Much of the current research into the qualitative behavior of dynamical systems is concerned with two fundamental problems involving the asymptotic behavior of the motions. The first of these problems is to describe the attractors or, more generally, the ω-limit sets of the motions. If one knows the structure of the ω-limit set, then one has essentially complete information about the given motion.

The second problem arises when one is studying dynamical systems which depend on a parameter. Once again one is interested in the attractors, but now one wants to study their dependence on the underlying parameter. In this study one encounters two correlated theories. First, there is a perturbation theory in which the goal is to find sufficient conditions for the attractors to appear to be unchanged. Secondly, there is a bifurcation theory where the objective is to describe such phenomena as period-doubling, Hopf–Landau bifurcations and the occurrence of "strange" attractors.

Our main objective in this lecture is to illustrate how the classical techniques of linearization can be used to address these problems of global

* This research was supported in part by a grant from the National Science Foundation.

dynamics. Specifically we are concerned with the question of linearization in the vicinity of a compact invariant manifold or more generally, near a bounded motion in a dynamical system. The linearization theory we require is primarily a theory of linearization near a time-varying solution.

As an illustration of the power of these linearization techniques, we will address here three specific problems. The first of these, which we study in Section III, is the question of determining when an attractor or an ω-limit set is a manifold.

In Section IV we study the second of these problems by illustrating how the technique of linearization can be used to develop a bifurcation theory for invariant manifolds. Specifically we will describe conditions under which a k-dimensional torus may undergo a Hopf–Landau bifurcation to a $(k+1)$-dimensional torus. This bifurcation theory is not simply a linear theory, but it also depends on the occurrence of certain irremovable nonlinear terms. Such nonlinearities give rise to "normal forms" for differential equations, which in turn form the basis for developing various bifurcation theories. The study of these normal forms in the vicinity of a smooth invariant manifold is an important chapter in the development of a qualitative description of dynamical systems.

The first step in a study of normal forms is the question of smooth linearization near an invariant manifold. This theory of smooth linearization, which completes our triad of problems, is described in Section V.

The theory we will describe here is valid for all compact invariant manifolds M or, more generally, for all bounded solutions φ of the underlying differential equation. For the most part the new contributions of our theory occur when M (or φ) is not a fixed point or a periodic orbit. One noteworthy exception occurs in the linearization theory in Section V. As a corollary of our methods, we are able to give an answer to the question of determining whether there is a C^N-linearization ($1 \leqslant N < \infty$) of a (nonlinear) vector field in the vicinity of a hyperbolic fixed point or periodic orbit.

Before turning to the mathematical details, we wish to express our sincere gratitude to Robert Sacker for the essential role he played in the development of these theories. Many of the ideas described below find their origins in our collaborations with Dr. Sacker.

II. The spectrum

We want to study the concept of a linear skew-product flow π defined on a vector bundle $\mathscr{E}$ over a compact base space M, Sacker and Sell [32]

and Selgrade [35]. Let us begin with a specific example which will be of interest later in the lecture.

Consider a smooth vector field or ordinary differential equation

$$X' = f(X) \tag{1}$$

defined in some smooth Riemannian manifold W, which we assume (for simplicity) to be an open set in a fixed Euclidean space R^n. Let M denote a given compact invariant set for (1). For example, one may have

$$M = \text{Hull}(\varphi) = \text{Closure}\{\varphi(t)\colon\ t \in R\}$$

where φ is a solution of (1) with range in a compact subset of W.

For $\theta \in M$ we let $\theta \cdot t$ denote the solution of (1) that satisfies $\theta \cdot 0 = \theta$. Let $A = Df$ denote the linear part of f (i.e. the Jacobian matrix) and let $\Phi(\theta, t)$ denote the fundamental solution matrix of

$$x' = A(\theta \cdot t)x, \tag{2}$$

with $\Phi(\theta, 0) = I$. Then

$$\pi(x, \theta, t) = \big(\Phi(\theta, t)x,\ \theta \cdot t\big) \tag{3}$$

is a linear skew-product flow on $R^n \times M$, or equivalently, Φ is a co-cycle on M, Ellis and Johnson [8]. The *shifted flow* associated with Eq. (3) is

$$\pi_\lambda(x, \theta, t) = \big(\Phi_\lambda(\theta, t)x,\ \theta \cdot t\big)$$

where $\Phi_\lambda(\theta, t) = e^{-\lambda t}\Phi(\theta, t)$ and λ is a real parameter. In other words, $\Phi_\lambda(\theta, t)$ is the fundamental matrix solution of

$$x' = [A(\theta \cdot t) - \lambda I]x.$$

The concept of a linear skew-product flow extends directly to a vector bundle $\mathscr{E}$ over M where $\theta \cdot t$ is a flow on M, see Sacker and Sell [32]. In this case $\Phi(\theta, t)$ is a linear mapping from the fibre $\mathscr{E}(\theta)$ over $0 \in M$ to the fibre $\mathscr{E}(\theta \cdot t)$. For example, if M is a compact invariant manifold and one restricts the vectors x in Eq. (2) to be tangent vectors to M, then π becomes the induced linearized flow on the tangent bundle TM $(= \mathscr{E})$ generated by (1).

We say that a linear skew-product flow π admits an *exponential dichotomy over* M if there is a projector $\hat{P}\colon \mathscr{E} \to \mathscr{E}$ and constants $K \geqslant 1$, $\alpha > 0$

such that

$$|\Phi(\theta, t)P(\theta)\Phi^{-1}(\theta, s)| \leqslant Ke^{-\alpha(t-s)}, \quad s \leqslant t,$$

$$|\Phi(\theta, t)[I-P(\theta)]\Phi^{-1}(\theta, s)| \leqslant Ke^{-\alpha(s-t)}, \quad t \leqslant s.$$

Recall that a *projector* is a continuous mapping $\hat{P}$: $\mathscr{E} \to \mathscr{E}$ that satisfies $\hat{P}(x, \theta) = (P(\theta)x, \theta)$ where $P(\theta)$ is a linear projection on the fibre $\mathscr{E}(\theta)$.

The (continuous) *spectrum* $\Sigma = \Sigma(M)$ of π over M is defined as the collection of all $\lambda \in R$ for which the shifted flow π_λ fails to admit an exponential dichotomy over M. The complement $\varrho(M) = R - \Sigma(M)$ is the *resolvent set.* If $\lambda \in \varrho(M)$ we let $\mathscr{S}_\lambda$ and $\mathscr{U}_\lambda$ denote the range and null space of the projector associated with the exponential dichotomy for π_λ. These are invariant subbundles for π and one has $\mathscr{E} = \mathscr{S}_\lambda + \mathscr{U}_\lambda$ as a Whitney sum. Also we define

$$N_\lambda = \text{fibre-dim}\, \mathscr{S}_\lambda,$$

where the fibre-dimension of any subbundle $\mathscr{V}$ of $\mathscr{E}$ is defined by

$$\text{fibre-dim}\, \mathscr{V} = \dim \mathscr{V}(\theta).$$

Note that N_λ is monotone nondecreasing in λ for $\lambda \in \varrho(M)$.

If M is connected then the Spectral Theorem assures us that $\Sigma(M)$ is the union of k nonoverlapping compact intervals $I_1, \ldots, I_k$, where $1 \leqslant k \leqslant n$ and $n = \text{fibre-dim}\, \mathscr{E}$. Also associated with each spectral interval I_i there is an invariant subbundle $\mathscr{V}_i$ of $\mathscr{E}$, where $n_i = \text{fibre-dim}\, \mathscr{V}_i$ satisfies $n_i \geqslant 1$, $1 \leqslant i \leqslant k$, with $n = n_1 + \ldots + n_k$. Furthermore if $\mu, \lambda \in \varrho(M)$ with $\mu < \lambda$ and the open interval (μ, λ) contains precisely one spectral interval I_i, then one has

$$N_\lambda - N_\mu = \text{fibre-dim}\, \mathscr{V}_i. \tag{4}$$

See Sacker and Sell [33] for more details.

If M is a smooth compact invariant submanifold for the flow generated by (1) on W, then there are three spectra (Σ, Σ_T and Σ_N) which we wish to study. First there is the *full* spectrum Σ, which is the spectrum of π on the full bundle $R^n \times M$. The tangent bundle TM is an invariant subbundle for the linearized flow π. By restricting π to the tangent bundle TM one obtains the *tangential* spectrum Σ_T. Next let $\mathscr{N}$ denote any subbundle in $R^n \times M$ which is complementary to TM. The linearized flow π then induces an associated flow π_N on $\mathscr{N}$ and the spectrum of π_N, is the *normal* spectrum Σ_N. As shown in Sacker and Sell [34], the normal spectrum Σ_N is independent of the choice of the normal subbundle $\mathscr{N}$. One can

compute Σ_T and Σ_N by using the projections of the Jacobian matrix A in the tangential and normal directions, respectively. Some of the properties of the three spectra are the following (cf. Sacker and Sell [34]):

(1) $\Sigma_T \cup \Sigma_N \subseteq \Sigma$.

(2) Usually (but not always) one has $\Sigma = \Sigma_T \cup \Sigma_N$.

(3) If $M \neq$ point, then $0 \in \Sigma_T$ and $0 \in \Sigma$.

Remarks. 1. The theory we describe here extends readily to the study of homeomorphisms, diffeomorphisms and, in general, linear skew-product flows with discrete time t. We will not develop the discrete version of this theory in this report. Instead we invite the reader to consult the references cited above.

2. This notion of the continuous spectrum is very closely related to the ergodic concept of the measurable spectrum, which is based on the theory of Lyapunov exponents, see Oseledec [23] and Ruelle [28]. The connections between these two concepts are described in Johnson, Palmer, and Sell [19]. By exploiting these interconnections Perry [24] has developed numerical algorithms for approximating the spectral intervals and the associated spectral subbundles. These numerical methods can then be used to study various bifurcation phenomena.

III. Hyperbolic almost periodic motions

We can now address the first of the three questions posed above. A somewhat more general version of this question is to ask when does the ω-limit set of a given trajectory lie on an invariant submanifold. A classical answer to this question is given in terms of the first integrals of the differential system. However, even in the presence of first integrals one can rephrase the question by restricting to submanifolds of low dimension. In this rather general form, it seems overly optimistic to expect that there is any situation where this rather subtle problem can be resolved by studying the linearized equations above. Nevertheless this does occur in study of almost periodic solutions.

Let $\varphi(t)$ be an almost periodic solution of (1) and let $M = \mathrm{Hull}(\varphi)$. The Pontryagin Duality Theorem assures us that the topological dimension of M agrees with the algebraic dimension of M. (The latter is the dimension of the Fourier–Bohr frequency module.) Let k denote this dimension. Let Σ denote the spectrum of the linearized flow on $R^n \times M$.

First we note that if $k \geqslant 1$, then $\lambda = 0 \in \Sigma$. In this case, let I_0 denote the spectral subinterval that contains $\lambda = 0$ and let $\mathscr{V}_0$ denote the associ-

ated spectral subbundle. It is shown in Sell [36] that fibre-dim $\mathscr{V}_0 \geqslant k$. Furthermore if fibre-dim $\mathscr{V}_0 = k$, then M is Lipschitz homeomorphic to the k-dimensional torus T^k and $\varphi(t)$ is a quasi-periodic solution. These conditions on k can be checked by using (4).

The proof of these assertions, which can be found in Sell [36], is based on ideas developed by Pliss [25].

IV. Perturbation and bifurcation of manifolds

Consider next the dynamical system

$$X' = f(X, \alpha) \tag{5.α}$$

on W, where f depends smoothly on X and a parameter α. Assume that for a fixed value, say $\alpha = \alpha_0$, there exists a compact invariant submanifold M_0 for (5.α_0). We want to study the behavior of M_0 as α varies in a neighborhood of α_0. The perturbation theories of Sacker [31], Fenichel [9] and Hirsch, Pugh and Shub [18] describe sufficient conditions under which M_0 can be imbedded in a smooth family of invariant manifolds M_α, for α near α_0, with $M_{\alpha_0} = M_0$. These theories can be summarized in terms of the spectra.

The manifold M_0 is said to be *normally hyperbolic of order* r, where r is a positive integer, if there exist real numbers a, b with $0 \leqslant a \leqslant ra < b$ and such that

$$\begin{aligned} &1)\ \lambda \in \Sigma_T \Rightarrow |\lambda| \leqslant a, \\ &2)\ \lambda \in \Sigma_N \Rightarrow |\lambda| \geqslant b. \end{aligned} \tag{6}$$

If M_0 is normally hyperbolic of order r, then there is a smooth family of invariant manifolds M_α of class C^r defined for α near α_0 with $M_{\alpha_0} = M_0$.

If the assumption of normal hyperbolicity breaks down at α_0, then the behavior of the flow generated by (5.α) near M_0, for α near α_0, can be very complicated, see Chenciner [5], Meyer [21], Sell [38] and Smale [41]. A full understanding of this behavior, even from a generic point of view, still eludes us. However there are a number of situations where one can obtain some insight. One very interesting case arises in the study of the Hopf–Landau bifurcation of a k-dimensional torus into a $(k+1)$-dimensional torus.

Assume that the parameter α is real and that for $\alpha \in I$, where I is an open interval with $0 \in I$, Eq. (5.α) has a family of k-dimensional invariant tori $\tau(\alpha)$ which varies smoothly in α. (Smooth variation means of class

C^N, for N sufficiently large.) Next we shall assume that the tori satisfy Hypotheses I and II of Sell [37], which means that one can find smooth local coordinates $Z \in R^{n-k-2}$, $x \in R^2$, $\varphi \in T^k$ near $\tau(a)$ so that Eq. (5.a) can be written in the form

$$\begin{aligned} x' &= A_{11}(\varphi, a)x + aA_{12}(\varphi, a)z + F(x, z, \varphi, a), \\ z' &= aA_{21}(\varphi, a)x + [B(\varphi) + aA_{22}(\varphi, a)]z + H(x, z, \varphi, a), \\ \varphi' &= G(x, z, \varphi, a). \end{aligned} \tag{7}$$

The terms B and A_{ij} denote matrices of the appropriate dimensions and F and H contain higher order terms in x and z. Furthermore one has $(F, H) = (0, 0)$ when $(x, z) = (0, 0)$. Also the differential equation $\varphi' = G(0, 0, \varphi, a)$ denotes the restriction of the flow generated by (5.a) to the torus $\tau(a)$. The system (7) is a Hopf–Landau dynamical system.

Let (ϱ, θ_0) denote polar coordinates in the x-plane and let $\theta = (\theta_0, \varphi) = (\theta_0, \theta_1, \ldots, \theta_k)$ denote a typical point in T^{k+1}, where $\varphi \in T^k$. For any continuous function $u = u(\theta)$ on T^{k+1} we let $M_0[u]$ denote the mean value of u.

The main hypotheses concern the (2×2) matrix A_{11} and the function G. First let us expand A_{11} in terms of a, that is, let $A_{11} = \Omega + aW$ where $\Omega = A_{11}(\varphi, 0)$. Let $w_{ij}(\varphi, a)$ denote the entries of $W(\varphi, a)$. Consider the following hypotheses:

H1. *The* (2×2) *matrix* Ω *satisfies*

$$\Omega = \begin{bmatrix} 0 & -\omega_0 \\ \omega_0 & 0 \end{bmatrix}$$

where ω_0 *is a nonzero constant.*

H2. *The mean value* $W = M_0[w_{11}\cos^2 2\pi\theta_0 + w_{22}\sin^2 2\pi\theta_0]$ (*at* $a = 0$) *is nonzero.*

H3. *There is a vector* $\omega = (\omega_1, \ldots, \omega_k)$ *and a smooth function* $L(x, z, \varphi, a)$ *such that* $G = \omega + aL$, *and the* $(k+1)$*-dimensional vector* $\omega = (\omega_0, \omega_1, \ldots, \omega_k)$ *where* ω_0 *is given by* H1 *above, satisfies the nonresonance condition* $|n \cdot \omega| \geqslant c|n|^{-\delta}$ *for integral vectors* $n = (n_0, n_1, \ldots, n_k) \neq 0$. (Here c and δ are positive constants that do not depend on n.)

In the theorem we state next, reference is made to a constant K. This constant, which is expressed as a mean value, depends upon the low-order terms (i.e. order $\leqslant 3$) in the Taylor series expansion of (7). The formula

for K appears in Sell ([37], Eq. (4.10)). The proof of this theorem relies on an invariant manifold theorem due to Hale [14].

THEOREM 1. *If Hypotheses* H1–H3 *are satisfied and the constant K is non-zero, then there is a unique family of $(k+1)$-dimensional invariant tori $\hat{\tau}(a)$ defined for $a \cdot \operatorname{sgn}(WK) < 0$ and one has $\hat{\tau}(a) \to \tau(0)$ as $a \to 0$. Furthermore if $W > 0$, $K < 0$ and the tori $\tau(a)$ are asymptotically stable for $a < 0$, then the bifurcating tori $\hat{\tau}(\beta)$ are asymptotically stable for $\beta > 0$.*

Remarks. 3. For $k = 1$ this result is essentially due to Sacker [30] who uses a weaker form of the nonresonance condition H3. Also see Marsden and McCracken [20] and Ruelle and Takens [29].

4. Basically the same theorem, formulated for mappings instead of differential equations, appears in Chenciner and Iooss [6]. Earlier versions of this result are cited in Haken [13].

5. A recent paper of Flockerzi [10] shows that the conclusion of Theorem 1 remains valid in some cases when both W and K vanish. In these cases Eq. (7) admits different normal forms.

V. Linearization near a compact invariant manifold

We shall begin this section by studying a nonlinear vector field

$$x' = Ax + F(x) \tag{8}$$

in the vicinity of a fixed point $x = 0$. We seek sufficient conditions for the existence of a smooth curvilinear coordinate system with the property that the vector field is linear when written in terms of the new coordinate system. Given such a linearization theory, a natural question then is to determine the smoothness of the new curvilinear coordinate system. Also, if the new coordinate system is lacking in smoothness, we want to determine the obstacles to smooth linearization. As we will now show, we can give a satisfactory and definitive resolution of this problem, when $x = 0$ is hyperbolic.

The differential equation (8) is said to admit a *C^N-linearization* near $x = 0$ if there is a C^N-diffeomorphism $H\colon V_1 \to V_2$, where V_1 and V_2 are neighborhoods of $x = 0$, that satisfies the following two properties:

(i) $H(0) = 0$.

(ii) Whenever $x(t)$ is a solution of (8) with $x(t) \in V_1$ for t in some interval

I, then $y(t) = H\big(x(t)\big)$ is a solution of

$$y' = Ay \tag{9}$$

for $t \in I$. Similarly, whenever $y(t)$ is a solution of (9) with $y(t) \in V_2$ for $t \in I$, then $x(t) = H^{-1}\big(y(t)\big)$ is a solution of (8) for $t \in I$. (The mapping $y = H(x)$ above is referred to as a *C^N-conjugation* between (8) and (9).)

Let A be an $(n \times n)$ matrix with eigenvalues $\lambda_1, \ldots, \lambda_n$ repeated with multiplicities and let $\Sigma(A) = \{\lambda_1, \ldots, \lambda_n\}$. Let $m = (m_1, \ldots, m_n)$ be a vector with nonnegative integer entries $m_1, \ldots, m_n$, and define $\gamma(\lambda, m)$ by

$$\gamma(\lambda, m) = \lambda - (m_1\lambda_1 + \ldots + m_n\lambda_n),$$

where λ is a complex number. Let $|m| = m_1 + \ldots + m_n$.

We shall say that A is *hyperbolic* if $\operatorname{Re}\lambda \neq 0$ for all $\lambda \in \Sigma(A)$. A is said to be *stable* if $\operatorname{Re}\lambda < 0$ for all $\lambda \in \Sigma(A)$. A is said to satisfy the *Sternberg condition of order N*, $N \geqslant 2$, if $\gamma(\lambda, m) \neq 0$ for all $\lambda \in \Sigma(A)$ and all m satisfying $2 \leqslant |m| \leqslant N$. We shall say that A satisfies the *strong Sternberg condition of order N*, if A satisfies the Sternberg condition of order N and

$$\operatorname{Re}\gamma(\lambda, m) \neq 0 \tag{10}$$

for all $\lambda \in \Sigma(A)$ and all m with $|m| = N$. It is easy to see that if A satisfies the strong Sternberg condition of order $N \geqslant 2$, then A is hyperbolic.

Let A be hyperbolic and let $\Sigma^+(A)$ or $\Sigma^-(A)$ denote, respectively, those eigenvalues $\lambda \in \Sigma(A)$ with $\operatorname{Re}\lambda > 0$ or $\operatorname{Re}\lambda < 0$. We shall say that A is *strictly hyperbolic* if A is hyperbolic and both $\Sigma^+(A)$ and $\Sigma^-(A)$ are nonempty. If A is hyperbolic and $\Sigma^i(A) \neq \emptyset$, we define the *spectral spread* ϱ^i by

$$\varrho^i = \frac{\max\{|\operatorname{Re}\lambda|\colon \lambda \in \Sigma^i(A)\}}{\min\{|\operatorname{Re}\lambda|\colon \lambda \in \Sigma^i(A)\}}$$

where $i = +$ or $-$.

Let Q be a positive integer and let A be hyperbolic. We define the *Q-smoothness* of A to be the largest integer $K \geqslant 0$ such that:

(1) $Q - K\varrho^- \geqslant 0$, if $\Sigma^+(A) = \emptyset$.

(2) $Q - K\varrho^+ \geqslant 0$, if $\Sigma^-(A) = \emptyset$.

(3) There exist positive integers M, N with $Q = M + N$, $M - K\varrho^+ \geqslant 0$, $N - K\varrho^- \geqslant 0$, when A is strictly hyperbolic.

Since the spectral spreads are $\geqslant 1$, we see that the Q-smoothness of A satisfies $K \leqslant \min(M, N)$ when A is strictly hyperbolic.

The following two theorems are proved in Sell [39].

THEOREM 2. *Let $Q \geqslant 2$ be an integer, and assume that F is of class C^{3Q} on $U \subseteq X$ with $0 \in U$ where $D^P F(0) = 0$ for $P = 0, 1$. Let A be strictly hyperbolic, and consider one of the following two assumptions:*

(A) *Assume that A satisfies the strong Sternberg condition of order Q.*

(B) *Assume that $D^P F(0) = 0$ for $0 \leqslant P \leqslant Q-1$ and that*

$$\mathrm{Re}\gamma(\lambda, m) \neq 0$$

for all $\lambda \in \Sigma(A)$ and all m with $|m| = Q$.

Under either assumption (A) *or* (B), *Eq.* (1.1) *admits a C^K-linearization, where K is the Q-smoothness of A.*

If A is stable, then one can weaken the assumption on the smoothness of F. In particular, we will prove the following result:

THEOREM 3. *If A is stable, then Theorem* 1 *remains valid when F is of class C^{2Q}.*

Remarks. 6. Sternberg ([42], [43]) studies the question of finding sufficient conditions that Eq. (8) admits a C^s-linearization. He showed that there is a function $V(s, \lambda_1, \ldots, \lambda_n) \geqslant 0$ with the property that if A is hyperbolic and satisfies the Sternberg condition of order N where $N \geqslant s + V$, then Eq. (8) admits a C^s-linearization. While there are several alternate proofs of Sternberg's Theorem (cf. Chen [4], Hartman [17], Nelson [22], and Takens [44]), the implicit formulae for N and V are very complicated. See Hartman ([17], p. 257), for example. Our theorems assert that under the stronger assumption that (10) is valid, we can give sharper and simpler estimates on the order of smoothness of the conjugation to the linear system. The homeomorphic version of these theorems (i.e., $N = 0$) appears in Grobman ([11], [12]) and Hartman ([15], [16]).

7. Our methods extend easily to the question of smooth linearization for diffeomorphisms in the vicinity of a fixed point. Our theory for the hyperbolic case where $N = 1$ is similar to, but not as strong as, a theorem of Bileckii ([1], [2]).

The assumption that $\gamma(\lambda, m) \neq 0$ for $2 \leqslant |m| \leqslant Q$ allows one to introduce a polynomial change of variables to eliminate the terms in Taylor series expansion of F with order between 2 and Q. The stronger assumption that $\mathrm{Re}\gamma(\lambda, m) \neq 0$ for $|m| = Q$ allows us to eliminate the remainder term in the Taylor series expansion of F.

The argument which we use to prove Theorems 2 and 3 is based on the theory of nonlinear perturbations of linear equations with exponential dichotomies, cf. Coppel [7]. The change of variables we introduce gives rise to a related nonlinear differential equation on a different finite dimensional Banach space. The quantities $\gamma(\lambda, m)$, for $\lambda \in \Sigma(A)$ and $|m| = Q$, arise as the eigenvalues of the associated linear equation, and Ineq. (10) ensures that this linear equation has an exponential dichotomy, Sell [39].

It is especially noteworthy that the methods we used to prove Theorems 2 and 3 extend readily to the study of smooth linearization near a compact invariant manifold M. In order to simplify the following discussion, we will assume that M, which is smoothly imbedded in W, has a trivial normal bundle. (The general problem can easily be reduced to this case.) It then follows that one can introduce smooth curvilinear local coordinates so that in the vicinity of M the vector field (1) becomes

$$x' = A(\theta)x + F(x, \theta), \qquad \theta' = g(\theta) + G(x, \theta), \tag{11}$$

where θ represents local coordinates on M and $x \in R^k$ represents a normal vector to M. Furthermore F and G satisfy

$$(F, D_1F, G)(0, \theta) = (0, 0, 0)$$

where $D_1 = \partial/\partial x$. Here $A(\theta)$ denotes the linear part of F projected in the normal x-direction at the point $\theta \in M$. The equation $\theta' = g(\theta)$ describes the flow on the manifold M.

The linearized vector field near M is defined as the vector field

$$y' = A(\varphi)y, \qquad \varphi' = g(\varphi) \tag{12}$$

where $\varphi \in M$ and $y \in R^k$. The linearized flow in the tangent bundle TM is given (in these coordinates) by

$$v' = B(\theta)v, \qquad \theta' = g(\theta) \tag{13}$$

where $B = D_2g$, $D_2 = \partial/\partial\theta$ and $v \in R^p$ where $p = \dim M$. The normal spectrum Σ_N and the tangential spectrum Σ_T are the spectra of the linear skew-product flows generated by (12) and (13), respectively.

We seek sufficient conditions in terms of the matrices $A(\theta)$ and $B(\theta)$ in order that there exists a C^N-conjugacy H of the form

$$y = x + u(x, \theta), \qquad \varphi = \theta + v(x, \theta)$$

which maps Eq. (11) to Eq. (12) in the vicinity of M.

Since the dimension of the normal bundle is k, it follows from the Spectral Theorem that Σ_N is the union of q nonoverlapping compact intervals, $I_1, \ldots, I_q$, where $1 \leqslant q \leqslant k$. Let $n_i = \text{fibre-dim}\, \mathscr{V}_i$, where $\mathscr{V}_i$ is the spectral bundle associated with I_i. Then $n_i \geqslant 1$ and $n_1 + \ldots + n_q = k$. We shall say that a k-tuple $(\lambda_1, \ldots, \lambda_k)$ from the spectrum Σ_N is *admissible* provided

(i) the mapping $j \to \lambda_j$ from $\{1, \ldots, k\}$ to R has its range in Σ_N, and
(ii) $\text{Card}\{j\colon \lambda_j \in I_i\} = n_i$, $1 \leqslant i \leqslant q$.

If the matrix A in (12) is independent of θ and has only real eigenvalues, then an admissible k-tuple is a listing of the eigenvalues of A repeated with their multiplicities.

The definitions of strict hyperbolicity, spectral spreads and Q-smoothness of (12) are made as in the constant coefficient case. One simplification to note is that the spectrum for (12) is real.

THEOREM 4. *Consider the equation* (11) *near* M *where the coefficients are of class* $3Q$ *and* M *is normally hyperbolic of order* Q. *Let* a *and* b *be defined so that* (6) *holds with* $r = K$, *and assume that one has*

$$\begin{aligned} &|\lambda - (m_1\lambda_1 + \ldots + m_k\lambda_k)| > Ka, \\ &|m_1\lambda_1 + \ldots + m_k\lambda_k| > (K+1)a \end{aligned} \tag{14}$$

for all $\lambda \in \Sigma_N$, *and all admissible* k*-tuples* $(\lambda_1, \ldots, \lambda_k)$ *and nonnegative integers* $m_1, \ldots, m_k$ *that satisfy*

$$2 \leqslant (m_1 + \ldots + m_k) \leqslant Q,$$

where K *is the* Q*-smoothness of* (12). *Then there is a* C^K*-conjugacy between* (11) *and* (12).

Remarks. 8. The homeomorphic version of the last theorem for a stable manifold appears in Pugh and Shub [26].

9. A result of Robinson [27] can also be used to study smooth conjugacies between (11) and (12). However, his hypotheses concern the Taylor series expansion of the nonlinear terms F and G instead of the spectral properties of M.

10. Our results are somewhat stronger than similar theorems developed in Bogoljubov, Mitropolskii and Samoilenko [3].

Bibliography

[1] Belickii G. R., Functional equations and the conjugacy of diffeomorphisms of finite smoothness class, *Functional Anal. Appl.* **7** (1973), pp. 268–277.

[2] Belickii G. R., Equivalence and normal forms of germs of smooth mappings, *Russian Math. Surveys* **33** (1978), pp. 107–177.

[3] Bogoljubov N. N., Mitropolskii I. A., and Samoilenko A. M., *Methods of Accelerated convergence in Nonlinear Mechanics*, Springer, Berlin, Heidelberg, New York, 1976.

[4] Chen K. T., Equivalence and decomposition of vector fields about an elementary critical point, *Amer. J. Math.* **85** (1963), pp. 693–722.

[5] Chenciner A., *Courbes fermees invariant es non normalement hyperboliques au voisinage d'une bifurcation de Hopf degeneree diffeomorphismes de* (R^2, 0), Preprint, 1983.

[6] Chenciner A. and Iooss G., Bifurcations de torus invariante, *Arch. Rational Mech. Anal.* **69** (1979), pp. 109–198.

[7] Coppel W. A., *Stability and Asymptotic Behavior of Differential Equations*, Heath, Boston, 1965.

[8] Ellis R. and Johnson R. A., Topological dynamics and linear differential systems, *J. Diff. Eqns.* **44** (1982), pp. 21–39.

[9] Fenichel N., Persistence and smoothness of invariant manifolds for flows, *Indiana Univ. Math. J.* **21** (1971/72), pp. 193–226.

[10] Flockerzi D., *Generalized bifurcation of higher dimensional tori*, preprint, 1983.

[11] Grobman D. M., Homeomorphisms of systems of differential equation, *Dokl. Akad. Nauk SSR* **128** (1959), pp. 880–881.

[12] Grobman D. M., Topological classification of the neighborhood of a singular point in n-dimensional space, *Mat. Sb.* (*N.S.*). **56** (**98**) (1962), pp. 77–94.

[13] Haken H., Chaos and order in nature, in: *Chaos and Order in Nature*, Proceedings of International Symposium on Synergetics, Springer-Verlag, New York, 1981, pp. 2–11.

[14] Hale J. K., Integral manifolds of perturbed differential systems, *Ann. Math.* **73** (1961), pp. 496–531.

[15] Hartman P., A lemma in the theory of structural stability of differential equations, *Proc. Amer. Math. Soc.* **11** (1960), pp. 610–620.

[16] Hartman P., On the local linearization of differential equations, *Proc. Amer. Math. Soc.* **14** (1963), pp. 568–573.

[17] Hartman P., *Ordinary Differential Equations*, Wiley, 1964.

[18] Hirsch M. W., Pugh C. C., and Shub M., *Invariant Manifolds*, Lecture Notes in Mathematics No. **583** Springer-Verlag, New York, 1977.

[19] Johnson R., Palmer K. and Sell G. R., *Ergodic properties of linear dynamical systems*, IMA Preprint No. 65, University of Minnesota, 1984.

[20] Marsden J. E., and McCraken M., *The Hopf Bifurcation and its Applications*, Applied Math. Sciences, Vol. **19**, Springer-Verlag, Berlin, Heidelberg, New York, 1976.

[21] Meyer K., *Tori in resonance*, IMA Preprint No. 13, University of Minnesota, 1983.

[22] Nelson E., *Topics in Dynamics I. Flows*, Princeton University Press, 1969.

[23] Oseledec V. I., A multiplicative ergodic theorem. Characteristic Lyapunov exponents of dynamical systems, *Trudy Moskov. Mat. Obshch.* **19** (1968), pp. 179–210.

[24] Perry D., *A numerical investigation of the Sacker–Sell spectrum with applications to bifurcation theory*, University of Minnesota, Ph.D. Thesis (to appear) (1984).
[25] Pliss V. A., A reduction principle in the theory of stability of motion, *Izv. Akad. Nauk SSSR Ser. Mat.* **28** (1964), pp. 1297–1324.
[26] Pugh C. C. and Shub M., Linearization of normally hyperbolic diffeomorphisms and flows, *Invent. Math.* **10** (1970), pp. 187–198.
[27] Robinson C., Differentiable conjugacy near compact invariant manifolds, *Bol. Soc. Brasil. Mat.* **2** (1971), pp. 33–44.
[28] Ruelle D., Ergodic theory of differentiable dynamical systems, *Publ. Math. IHES* **50** (1979), pp. 27–58.
[29] Ruelle D. and Takens F., On the nature of turbulence, *Comm. Math. Phys.* **20** (1971), pp. 167–192, and **23** (1971), pp. 343–344.
[30] Sacker R. J., *On invariant surfaces and bifurcation of periodic solutions of ordinary differential equations*, New York Univ. Tech. Report IMM-NYU, 1964.
[31] Sacker R. J., A perturbation theorem for invariant manifolds and Holder continuity, *J. Math. Mech.* **18** (1969), pp. 705–762.
[32] Sacker R. J. and Sell G. R., Existence of dichotomies and invariant splittings for linear differential systems I, *J. Diff. Eqns.* **15** (1974), pp. 429–458. Part II, *loc. cit.* **22** (1976), pp. 478–496. Part III, *loc. cit.* **22** (1976), pp. 497–522.
[33] Sacker R. J. and Sell G. R., A spectral theory for linear differential systems, *J. Diff. Eqns.* **27** (1978), pp. 320–358.
[34] Sacker R. J., and Sell G. R., The spectrum of an invariant submanifold, *J. Diff. Eqns.* **38** (1980), pp. 135–160.
[35] Selgrade J. F., Isolated invariant sets for flows on vector bundles, *Trans. Amer. Math. Soc.* **203** (1975), pp. 359–390.
[36] Sell G. R., The structure of a flow in the vicinity of an almost periodic motion, *J. Diff. Eqns.* **27** (1978), pp. 359–393.
[37] Sell G. R., Bifurcation of higher dimensional tori, *Arch. Rational Mech. Anal.* **69** (1979), pp. 199–230.
[38] Sell, G. R., Resonance and bifurcation in Hopf-Landau dynamical systems, in: *Turbulence and Nonlinear Dynamics*, Pitman, Boston, 1983, pp. 305–313.
[39] Sell G. R., *Smooth linearization near a fixed point*, IMA Preprint No. 16, University of Minnesota, 1983.
[40] Sell G. R., *Smooth linearization in the vicinity of a compact invariant manifold* (to appear), 1984.
[41] Smale S., Differentiable dynamical Systems, *Bull. Amer. Math. Soc.* **73** (1967), pp. 747–817.
[42] Sternberg S., Local contractions and a theorem of Poincaré, *Amer. J. Math.* **79** (1957), pp. 809–824.
[43] Sternberg S., On the structure of local homeomorphisms of Euclidean n-space, *Amer. J. Math.* **80** (1958), pp. 623–631.
[44] Takens F., Partially hyperbolic fixed points, *Topology* **10** (1971), pp. 133–147.

SCHOOL OF MATHEMATICS
INSTITUTE FOR MATHEMATICS AND ITS APPLICATIONS
UNIVERSITY OF MINNESOTA
MINNEAPOLIS, MINNESOTA 55455

Proceedings of the International Congress of Mathematicians
August 16-24, 1983, Warszawa

MICHAEL AIZENMAN*

Stochastic Geometry in Statistical Mechanics and Quantum Field Theory

1. Introduction

An early success of statistical mechanics has been the provision of a solid framework which encompasses both the laws of mechanics and the basic principles of thermodynamics. Among the spectacular results of this approach are the explanations, from basic principles, of the various phase transitions which are observed in nature. A not less challenging goal is to reach a complete explanation of the critical behavior of bulk systems. The universality of the critical exponents, which has been observed experimentally, is an intrinsically significant effect. It both calls for, and raises, the possibility of a mathematical elucidation of the subject.

Great advances towards the formation of a global picture, and the approximate calculation of critical exponents, have been made by renormalization-group related methods. By their nature, these methods do not offer exact solutions to any specific statistical-mechanical model. (Although they do produce sharp predictions above, and *at*, the upper-critical dimension.) Instead, the analysis incorporates the notion of universality, which from the point of view of physics may indeed be better founded than any given model. However, in view of the mathematical intractability of these treatments, the problem of the critical behavior is still well worth further attention.

In recent years progress has been made in the rigorous analysis of some of the most important models of statistical mechanics and quantum field theory. Instrumental for this advance has been the identification

---

* A. P. Sloan Foundation Fellow. Research supported in part by NSF Grant PHY-8301493.

of some stochastic-geometrical effects which play a key role in their phase transitions, and the corresponding critical behavior.

The results for statistical mechanics include the following:

(1) Proof that the critical behavior of the "susceptibility", in Ising-type models, stabilizes when the model's dimension reaches the value $d = 4$. Above that "upper-critical dimension" the corresponding critical exponent, γ, takes exactly the value 1 — which is suggested by an extremely simple (mean-field) approximation.

(2) A simple explanation of certain aspects of the critical behavior which are manifested in low dimensions. Specifically, "hyperscaling" — which fails for $d > 4$, was proven to be "universally" valid in $d = 2$ dimensions.

(3) A partial result for percolation models, which lends certain tenuous support to the physicists' "educated guess" — that these models are also endowed with an "upper-critical dimension", and that its value is $d = 6$.

I shall briefly describe here the stochastic-geometric aspects of these models and of the above results. The basic phenomenon which is discussed is the existence of the "upper-critical dimensions", above which the critical behavior in statistic mechanical systems is quite simple, and below which it is rather non-trivial.

This situation carries somewhat opposite implications for the constructive quantum field theory, where the level of difficulties of the goals is found to be in an inverse relation to those posed by the challenges of statistical mechanics. However, since this is also the subject of the talk of K. Osterwalder, only a few words would be said here on the quantum field theory.

The contribution to the Congress Proceedings contains only a brief summary of the talk. A somewhat more detailed report on this subject, and a more complete reference list, are given in Aizenman (1983).

2. Stochastic geometry

We use this term when refering to properties of random geometrical objects. An instructive example is obtained by considering two random lines, in $\boldsymbol{R}^d$, which are obtained as the trajectories of two independent Brownian motions (Wiener processes). The question of how typical is it for these paths to intersect, has been dealt with in the papers of Kakutani

(1944) and Dvoretsky Erdös, and Kakutani (1950). The answer can be summarized as follows.

THEOREM 2.1. *Let $b_1, b_2 \in \boldsymbol{R}^d$ be the sets of sites visited during times $t \in (0, \infty)$ by two independent, d-dimensional, Brownian paths which start at $x_1, x_2 \in \boldsymbol{R}^d$. Then*

$$\text{Prob}(b_1 \cap b_2 \neq \emptyset) = \begin{cases} 0 & \text{for } d \geqslant 4, \\ 1 & \text{for } d < 4. \end{cases} \tag{1}$$

An analogous (but significantly different) statement holds for random walks.

THEOREM 2.2. *Let $\omega_1, \omega_2 \subset \boldsymbol{Z}^d$ be the sets of lattice sites visited by a pair of independent simple random walks, with some specified starting points, at times $t \in \boldsymbol{Z}_+ \setminus \{0\}$. Then*

$$\text{Prob}(\omega_1 \cap \omega_2 = \emptyset) \text{ is } \begin{cases} > 0 & \text{for } d > 4, \\ = 0 & \text{for } d \leqslant 4. \end{cases} \tag{2}$$

(Notice that the LHS in (2) corresponds to 1—LHS of (1).)

These results exhibit two striking features. First, Theorem 2.1 may be counter-intuitive if one thinks about b_i as *lines*. (After all — b_i are the ranges of a pair of continuous mappings of $\boldsymbol{R}_+$ into $\boldsymbol{R}^d$.) Random rays miss each other already for $d > 1+1$. This puzzle is resolved by the celebrated observation that the Hausdorff dimension of the paths is 2 rather than 1. The other curious effect is that as a result of the discretization the behavior *at* the critical dimension, $d = 4$, is changed from being high-dimensional in (1) to the low-dimensional in (2). This is due to the fact, which is not reflected in (1), that with probability 1 the two paths b_1, b_2 come arbitrarily close to each other, even in $d = 4$ dimensions.

3. Percolation model

In statistical mechanics one typically deals not with isolated components, as in the above example, but rather with infinite arrays of such elements. This is the case in the Bernoulli percolation model, which is based on a collection of independent random variables $\{n_b\}$ which are associated with the bonds (i.e. unit segments joining pairs of neighboring sites) of the lattice $\boldsymbol{Z}^d$. The variables take the values 1, 0 with the homogeneous probabilities p, $1-p$. For each given configuration of values of $\{n_b\}$, the

lattice is decomposed into connected clusters — by regarding a bond as connecting if $n_b = 1$. Let $C(x)$ be the cluster which contains the site $x \in \mathbf{Z}^d$, and $|C(x)|$ the number of its points.

The model exhibits a phase transition, when p is varied, associated with the spontaneous formation of infinite clusters. The transition is manifested in the behavior of the quantities

$$\tau(x, y) = \text{Prob}\,(y \in C(x)),$$

$$P_\infty = \text{Prob}\,(|C(0)| = \infty)$$

and

$$\varkappa = \langle |C(0)| \rangle = \sum_x \tau(0, x), \tag{3}$$

which are monotone functions of p (with $P_\infty(p) \equiv 0$ for p small enough and $\lim_{p\to\infty} P_\infty(p) = 1$, at $d \geqslant 2$). At some critical point, p_c, $\varkappa(p)$ diverges (for $d \geqslant 2$). The expected power behavior

$$\varkappa \cong |p_c - p|_+^{-\gamma}$$

is the basis for the definition of the critical exponent γ, whose values is one of the questions we shall address.

When pressed for some quick guess for the critical behavior, one could try a "tree approximation", in which $\mathbf{Z}^d$ is replaced by a Bethe lattice (Cayley tree) preserving the local number of neighbors. The calculation is then trivial, yielding

$$\varkappa_{\text{B.L.}}(p) = (1+p) \sum_{k=0}^{\infty} (2d-1)^k p^k = \frac{1+p}{2d-1}\, |p_c^{\text{B.L.}} - p|_+^{-\gamma} \tag{4}$$

with

$$p_c^{\text{B.L.}} = (2d-1)^{-1}, \tag{5}$$

and

$$\gamma^{\text{B.L.}} = 1, \quad \text{independently of } d\ (!). \tag{6}$$

This of course is a very unreliable estimate of the actual values of p_c and γ. Furthermore, one can prove that $p_c \neq p_c^{\text{B.L.}}$, which leaves the second "prediction" even more suspicious. (Indeed, it is contradicted by numerical results in low dimensions.) Yet interestingly enough, these values do offer general bounds.

THEOREM 3.1. *For a general d,*

(i) $$p_c > p_c^{\text{B.L.}} \quad (\neq) \tag{7}$$

and

(ii) $$\gamma \geqslant 1 \ (= \gamma^{\text{B.L.}}). \tag{8}$$

In its weak form, (i) is a commonly made observation which is very elementary; (ii) is more recent (Aizenman and Newman, to appear) although its proof is surprisingly easy. The interesting point for us is that both can be proven using the observation that $d\varkappa^{-1}/dp$ is related to the probability that two *neighboring* sites, 0 and 1, belong to a pair of very large clusters which *do not intersect.* More specifically:

$$\frac{1-p}{\varkappa^2}\frac{d\varkappa}{dp} = \overline{\text{Prob}\big(C(0)\cap C(1) = \emptyset \,\big|\, C(0) \ni x,\ C(1) \ni y\big)}, \tag{9}$$

where the probability is conditioned on the event on the right side, and is averaged over (x, y) with the normalized weights $\tau(0, x)\ \tau(1, y)/\varkappa^2$.

The above representation is quite telling. First, since $\text{Prob}(-) \leqslant 1$, it leads to a bound on $|d\varkappa^{-1}/dp|$, which when integrated down from p_c, where $\varkappa^{-1}(p_c) = 0$, directly leads to (ii). (We skip here some details on the how does one settle the question of the continuity of $\varkappa^{-1}$ at p_c.)

Furthermore, (9) demonstrates that in order for γ to take some other value than 1, the probability of mutual avoidance of a pair of "incipient clusters" should necessarily vanish. As we saw, for random walks, this happens only in the low dimensions $d < 4$.

While random clusters are significantly different from random walks, these considerations suggest that there may be an "upper critical dimension" above which the critical exponent γ takes *exactly* the value which it has in the simple "tree approximation". We shall later mention a criterion which lends some support to the physicists' claim that the upper critical dimension is $d = 6$.

The percolation model was presented first only because its geometric features are plainly manifest. However, the basic approach discussed here was first developed in the analysis of ferromagnetic spin systems, for which the results are more conclusive.

4. Ferromagnetic spin systems

Statistic-mechanical models of ferromagnetism consist of lattice arrays of variables ("spins") σ_x, $x \in \mathbb{Z}^d$, with probability measures of the form:

$$\varrho(d\sigma) = \lim_{\Lambda \nearrow \mathbb{Z}^d} \left\{ \exp \beta \left[\sum_{|x-y|=1} \sigma_x \sigma_y + h \sum_{x \in \Lambda} \sigma_x \right] \right\} \prod_{x \in \mathbb{Z}^d} \varrho_0(d\sigma_x) / Z(\Lambda), \tag{10}$$

where $\varrho_0(d)$ is some non-interacting *even* single-spin measure, and $Z(\Lambda)$ is the normalizing factor. The physical parameters of the model are the temperature $-\beta^{-1}$, and the applied magnetic field $-h\beta^{-1}$.

A quantity of special interest is the magnetization, defined as:

$$M(\beta, h) = \int \sigma_0 \varrho(d\sigma). \tag{11}$$

For sufficiently large values of β, $M(\beta, h)$ — as a function of h is discontinuous at $h = 0$. A critical value of β is defined by the divergence of the susceptibility:

$$\chi(\beta) = \left. \frac{\partial M}{\partial h} \right|_{h=0}. \tag{12}$$

It is expected that the critical exponent γ defined as

$$\gamma = \lim_{\beta \nearrow \beta_c} \frac{\log \chi(\beta)}{\log 1/(\beta_c - \beta)} \tag{13}$$

is (along with the other critical exponents) independent of many of the details of the model.

The phase transition is a cooperative effect in a system with an infinite number of degrees of freedom. Consequently, $\chi(\beta)$ is not expected to be derivable from any *a priori* suggestive, finite, system of differential equations. However, a description of this sort is obtained within the very simplistic mean-field approximation, in which $M = M(\beta, h)$ is the solution of:

$$M = f(2d\beta M + \beta h), \tag{14}$$

with

$$f(\lambda) = \int \sigma e^{\lambda\sigma} \varrho_0(d\sigma) / \int e^{\lambda\sigma} \varrho_0(d\sigma).$$

For a general class of distributions, which satisfy the Griffiths–Hurst–Sherman inequality (Griffiths, Hurst and Sherman, 1970; Ellis, Monroe and Newman, 1976; Ellis and Newman, 1978), the function $f(\lambda)$ is antisymmetric and concave on $[0, \infty]$. In such cases (14) leads to the following "predictions":

(1) $$\beta_c = [2d\langle\sigma^2\rangle_0]^{-1}, \quad \text{where } \langle\sigma^2\rangle_0 = f'(0). \tag{15}$$

(2) For $\beta < \beta_c$:

$$\chi(\beta) = 1/[2d(\beta_c - \beta)] \tag{16}$$

and in particular

$$\gamma = 1. \tag{17}$$

The mean-field equation (14) is certainly not correct. Nevertheless, it is not entirely useless. The values obtained from it for β_c, $M(\beta, h)$, and even γ, form in fact rigorous upper bounds.

The earliest such a result is in the works of Fisher (1967) and Griffiths (1967) which show that for the Ising model (with $\varrho_0(d\sigma) = \delta(\sigma^2-1)d\sigma$):

$$\beta \geqslant \beta_c^{\text{M.F.}}. \tag{18}$$

Since then, the inequality has been redrived by a great variety of arguments, listed in Aizenman (1983), for a wide class of a priori distributions.

Furthermore, there is the following result which bears more on the critical behavior.

THEOREM 3.1 (Glimm and Jaffe, 1974). *For systems in which the Lebowitz inequality* (Griffiths, Hurst and Sherman, 1970; Ellis, Monroe and Newman, 1976; Ellis and Newman, 1978; Lebowitz, 1974) *is satisfied* (*including the Ising model with* $\varrho_0(d\sigma) = \delta(\sigma^2-1)d\sigma$, *and the* φ^4 *lattice systems with* $\varrho(d\varphi) = e^{-\lambda\varphi^4+B\varphi^2}d\varphi$)

$$\gamma \geqslant 1 \ (= \gamma^{\text{M.F.}}). \tag{19}$$

5. Vindication of the mean field approximation above the upper critical dimension

There is an important difference between the two "predictions" of equations (15) and (17). The value of β_c is clearly model (i.e. ϱ_0) dependent. It should not be expected to be, and is not, exactly equal to $\beta_c^{\text{M.F.}}$. On the other hand, the critical exponents are expected to be universal. Indeed,

the mean field result exhibits universality. (In fact too much of that — since it is false in low dimensions.) It may sound remarkable that even though $\beta_c \neq \beta_c^{\text{M.F.}}$, the power law of χ as a function of $\beta - \beta_c$ could be predicted correctly by this approximation. That however is the case in the "high dimensions" — $d > 4$, — a fact which was proven by arguments inspired by a stochastic-geometric picture.

The stochastic geometric formulation of the problem is achieved by a representation of ferromagnetic systems by means of "dual" systems of "random currents". Omitting the details (given in Aizenman, 1982), let us just say that the currents are described by random flux numbers which are somewhat similar to the variables $\{n_b\}$ which were described in our discussion of the bond percolation model. A key difference is that the fluxes are constrained to be "sourceless" (i.e. $\sum_{b \ni x} n_b \equiv 0 \mod 2$), except at specified sites. Such a representation can be developed for each system with $\varrho_0(d\sigma)$ in the Griffiths–Simon class (Griffiths, 1969; Simon and Griffiths, 1973) which includes the Ising and φ^4 cases. One of the results is that the quantity $\frac{1}{2d}\left|\frac{\partial \chi^{-1}}{\partial \beta}\right|$ reduces exactly to the *probability of intersection of two long current clusters*. The resulting expression is strikingly analogous to the expression (9) for $\left|\frac{d\varkappa^{-1}}{dp}\right|$.

The consequences of such a relation were discussed above (Section 3). These include simple proofs of the relations (18) and (19), and a novel characterization of the upper-critical dimension, above which $\gamma = 1$. The property which characterizes the high dimensions (here: $d > 4$), is that the probability of two long current clusters, which contain a given pair of sites, to be totally disjoint is uniformly positive.

This framework permits to approach the problem in a way which is somewhat analogous to the analysis of the random walk's property (2). Such arguments have led to the following result (Aizenman, 1982; Aizenman and Graham, 1983):

THEOREM 4.1. *In systems with ϱ_0 in the Griffiths–Simon class (which includes the φ^4 and the Ising spins) the function $\chi(\beta)$, for $\beta < \beta_c$, satisfies:*

$$(2d)^{-1}(\beta_c - \beta)^{-1} \leqslant \chi(\beta) \leqslant \begin{cases} c(d, \varrho_0)(\beta_c - \beta)^{-1}, & d > 4, \\ \bar{c}(\varrho_0)(\beta_c - \beta)^{-1}[1 + \log 1/(\beta_c - \beta)], & d = 4 \end{cases} \tag{20}$$

with some c, $\bar{c} < \infty$.

The upper bounds in (20) are consequences of the geometrically inspired inequality

$$\frac{1}{2d}\left|\frac{\partial\chi^{-1}}{\partial\beta}\right| \geqslant [1+(2d\beta)^2\sum_x \langle\sigma_0\sigma_x\rangle^2]^{-1} \tag{21}$$

which shows that a sufficient condition for γ to attain its mean-field value is the uniform boundedness (for $\beta < \beta_c$) of:

$$B = \beta^2\sum_x \langle\sigma_0\sigma_x\rangle^2 = \frac{\beta^2}{(2\pi)^d}\int_{[-\pi,\pi]^d} dp_1\ldots dp_d\, G^2(p), \tag{22}$$

where $G(p) = \sum_x e^{ipx}\langle\sigma_0\sigma_x\rangle$.

In Aizenman and Newman (to appear) it is shown that for percolation models the above condition is replaced by the finiteness, at $p = p_c$, of

$$V = \sum_{x,y}\tau(0,x)\tau(x,y)\tau(y,0) = \frac{1}{(2\pi)^d}\int_{[-\pi,\pi]^d} dp_1\ldots dp_d\,\hat{\tau}(p)^3. \tag{23}$$

The proof of Theorem 4.1, for which $d = 4$ is the critical dimension, is completed by the Gaussian bound of Fröhlich, Simon and Spencer (1976) which implies that for $\beta < \beta_c$ (and any d):

$$\beta G(p) \leqslant \frac{c}{p^2+\chi^{-1}}. \tag{24}$$

In the above analysis, $\langle\sigma_0\sigma_x\rangle$ plays a role which is somewhat similar to the hitting probability for random walks. Another proof of (20), for $d > 4$, has been given in the work of Fröhlich (1982) by means of a random walk expansion.

Had the analog of (24) been known for $\hat{\tau}(p)$, at least for $d > 6$, it would have implied that for percolation $\gamma = 1$ in $d > 6$ dimension. However, such a bound has not been derived, and is not expected to hold there as a dimension-independent inequality.

6. Few comments on the quantum field theory

Above the upper-critical dimension not only do the critical exponents simplify, but also the correlation functions acquire a simple form — the random fields defined by their scaling limits are Gaussian. This fact has somewhat opposite implications for the statistic mechanician and the quantum field theorist. The former's task involves finding the critical behavior — which, as explained above, is simple above d_c and quite non-trivial in low dimensions. The goals of the latter require setting models of interacting — i.e. non-Gaussian, fields. This goal is made more difficult by the fact that at least the simplest candidate for such a construction fails — due to the attraction to a "Gaussian fixed point". Conversely, below the critical dimension there is a rigorizable perturbation theory which leads to non trivial fields — which presumably also describe the critical regime of statistic mechanical models.

Here too, stochastic-geometric represetnations offer frameworks for the proofs of the above assertions — for which an important insight is derived from Theorem 1. Recent results in this vane can be found in references (Aragao de Carvalho, Caraciolo and Frölich, 1983; Aizenman and Graham, 1983; Brydges, Fröhlich and Sokal, 1983).

References

Aizenman M., 1981, *Phys. Rev. Lett.* **47**, p. 1; and 1982, *Commun. Math. Phys.* **86**, p. 1.

Aizenman M., 1983, Rigorous Results on the Critical Behavior in Statistical Mechanics. In: J. Fröhlich (ed.), *Scaling and Self-Similarity (Renormalization in Statistical Mechanics and Dynamics)*, Birkhäuser, Boston–Basel–Stuttgart.

Aizenman M. and Graham R., 1983, *Nucl. Phys.* **B225** [FS9], p. 261.

Aizenman M. and Newman C. M., Tree Diagram Bounds and the Critical Beahvior in Percolation Models, to appear in *J. Stat. Phys.*

Argao de Carvalho C., Caraciolo S. and Fröhlich J., 1983, *Nucl. Phys.* **B215** [FS7], p. 209.

Brydges D. C., Fröhlich J. and Sokal A. D., 1983, *Commun. Math. Phys.* **91**, p. 117.

Dvoretsky A., Erdös P. and Kakutani S., 1950, *Acta Sci. Math. (Szeged)* **12**, p. 75.

Ellis R. S., Monroe J. L. and Newman C. M., 1976, *Commun. Math. Phys.* **46**, p. 167.

Ellis R. S. and Newman C. M., 1978, *Trans. Am. Math. Soc.* **237**, p. 83.

Fisher M. E., 1967, *Phys. Rev.* **162**, p. 480.

Fröhlich J., Simon B. and Spencer T., 1976, *Commun. Math. Phys.* **50**, p. 79.

Fröhlich J., 1982, *Nucl. Phys.* **B200** [FS4], p. 281.

Glimm J. and Jaffe A., 1974, *Phys. Rev.* **D10**, 536.

Griffiths R. B., 1967, *Commun. Math. Phys.* **6**, p. 121.

Griffiths R. B., 1969, *J. Math. Phys.* **10**, p. 1559.
Griffiths R., Hurst C. and Sherman S., 1970, *J. Math. Phys.* **11**, p. 790.
Kakutani S., 1944, *Proc. Japan Acad.* **20**, p. 648.
Lebowitz J. L., 1974, *Commun. Math. Phys*, **35**, p. 87.
Simon B. and Griffiths R., 1973, *Commun. Math. Phys.* **33**, p. 145.

DEPARTMENTS OF MATHEMATICS AND PHYSICS
RUTGERS UNIVERSITY
NEW BRUNSWICK, N.J. 08903, USA

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

J. M. BALL

Energy-Minimizing Configurations in Nonlinear Elasticity

We discuss applications of the calculus of variations to nonlinear elasticity, and certain related issues. We confine attention to problems in $n > 1$ space dimensions. (A comprehensive account of one-dimensional problems has been given by Antman [3].)

Consider an elastic body occupying in a reference configuration a bounded domain $\Omega \subset \boldsymbol{R}^n$. We assume for ease of exposition that the body is *homogeneous*; i.e. it is composed of the same material at each point $x \in \Omega$. In a typical deformation $X\colon \Omega \to \boldsymbol{R}^n$ the *total stored-energy* of the body is given by the functional

$$E(x) = \int_{\Omega} W\big(x\nabla(X)\big)\,dX \tag{1}$$

where W denotes the *stored-energy function* of the material. Let $M^{n\times n}$ denote the space of real $n \times n$ matrices. We suppose that $W\colon M^{n\times n} \to \overline{\boldsymbol{R}}$ is continuous and bounded below, and that $W(A) = \infty$ if and only if $\det A \leqslant 0$. (The last requirement is imposed with the intention of making it energetically impossible to compress part of the body to zero volume or to change its orientation.) We suppose that the body is subjected to external body forces with potential $\Psi(X, x)$ per unit volume and for simplicity we consider the case when $\Psi\colon \boldsymbol{R}^n \times \boldsymbol{R}^n \to \boldsymbol{R}$ is continuous and bounded below. We consider a mixed displacement zero traction boundary value problem, in which it is required that

$$x(X) = \bar{x}(X), \quad \text{a.e. } X \in \partial\Omega_1, \tag{2}$$

while the remainder $x(\partial\Omega \setminus \partial\Omega_1)$ of the boundary is traction-free. In (2) $\partial\Omega_1$ denotes a measurable subset of the boundary $\partial\Omega$ of Ω (which we assume to be strongly Lipschitz) with positive (n −1)-dimensional measure,

and $\bar{x}$: $\partial\Omega_1 \to \boldsymbol{R}^n$ is a given measurable function. (More general conservative boundary problems are considered in Ball [4].) We define the total energy functional $I(x)$ by

$$I(x) \overset{\text{def}}{=} E(x) + \int_\Omega \Psi\big(X, x(X)\big)\,dX. \tag{3}$$

Corresponding to (2) we consider the set

$$\mathscr{A} = \{x \in W^{1,1}(\Omega; \boldsymbol{R}^n)\colon x \text{ satisfies (2)}, I(x) < \infty\} \tag{4}$$

and pose the following problem.

PROBLEM. Does $I(x)$ attain an absolute minimum on $\mathscr{A}$?

In the case when $\partial\Omega_1 = \partial\Omega$ (pure displacement problem) a necessary condition that I attains an absolute minimum for every smooth Ψ and $\bar{x}$ is that W be $W^{1,1}$-*quasiconvex* (Ball and Murat [9]), i.e.

$$\int_D W\big(A + \nabla\varphi(X)\big)\,dX \geqslant (\operatorname{meas} D) W(A) \tag{5}$$

for any bounded open subset $D \subset \boldsymbol{R}^n$ with meas $\partial D = 0$, for all $A \in M^{n\times n}$ and for all $\varphi \in W_0^{1,1}(D; \boldsymbol{R}^n)$. The weaker condition that (5) hold for all $\varphi \in W_0^{1,\infty}(D; \boldsymbol{R}^n)$ was introduced by Morrey [16] and termed by him *quasiconvexity*; it implies in particular that for $W \in C^2(M_+^{n\times n})$ the *Legendre–Hadamard* (or *ellipticity*) *condition*

$$\frac{\partial^2 W(A)}{\partial A_\alpha^i \partial A_\beta^j} a^i b_\alpha a^j b_\beta \geqslant 0 \tag{6}$$

for all $A \in M_+^{n\times n}$ and all $a, b \in \boldsymbol{R}^n$ holds. $W^{1,1}$-quasiconvexity is also a necessary condition for sequential weak lower semicontinuity of $E(\cdot)$ in $W^{1,1}(\Omega; \boldsymbol{R}^n)$.

An example of a set of sufficient conditions ensuring that I attains its minimum is given by the following result. Of course the hypotheses imply that W is $W^{1,1}$-quasiconvex.

THEOREM 1. *Let $n = 3$. Suppose that W is polyconvex, i.e. there exists a convex function g: $M^{3\times3} \times M^{3\times3} \times (0, \infty) \to \boldsymbol{R}$ such that*

$$W(A) = g(A, \operatorname{adj} A, \det A) \quad \textit{for all } A \in M_+^{3\times3}, \tag{7}$$

where $M_+^{3\times3} = \{A \in M^{3\times3}\colon \det A > 0\}$. Suppose further that

$$W(A) \geqslant C + D(|A|^p + |\operatorname{adj} A|^q) \quad \textit{for all } A \in M_+^{3\times3}, \tag{8}$$

where $D > 0$, C are constants, $p \geqslant 2$ and $q \geqslant p/(p-1)$. Then if $\mathscr{A}$ is non-empty I attains its absolute minimum on $\mathscr{A}$, and the minimiser x satisfies $\det \nabla x(x) > 0$ *a.e.* $X \in \Omega$.

Theorem 1 is proved in [9]; it is a slight refinement of earlier results of Ball [4, 5] and Ball, Currie and Olver [8]. The proof uses the direct method of the calculus of variations, the weak continuity properties of Jacobians (Reshetnyak [18, 20], Ball [4], Ball, Currie and Olver [8]), and an idea of Reshetnyak [19]. For some related semicontinuity theorems see Acerbi and Fusco [2] and Acerbi, Buttazzo and Fusco [1]. For pure displacement boundary value problems with appropriate boundary data it can be shown (Ball [7]) that under stronger growth conditions on W the minimiser x is a homeomorphism, so that interpenetration of matter does not occur. An analogous version of Theorem 1 holds for incompressible materials, all deformations of which satisfy the pointwise constraint $\det \nabla x(X) = 1$ a.e. $X \in \Omega$.

The stored-energy function W is said to be *isotropic* if

$$W(A) = \Phi(v_1, v_2, v_3) \text{ for all } A \in M^{3\times 3}_+ \tag{9}$$

for some symmetric function Φ of the principal stretches $v_i = v_i(A)$, that is the eigenvalues of $(A^T A)^{1/2}$. Following essentially the work of Ogden [17] on stored-energy functions appropriate for natural rubbers we consider the case

$$\begin{aligned}\Phi(v_1, v_2, v_3) = &\sum_{i=1}^{M} c_i(v_1^{\alpha_i} + v_2^{\alpha_i} + v_3^{\alpha_i} - 3) + \\ &+ \sum_{j=1}^{N} d_j\big((v_2v_3)^{\beta_j} + (v_3v_1)^{\beta_j} + (v_1v_2)^{\beta_j} - 3\big) + h(v_1v_2v_3),\end{aligned} \tag{10}$$

where $M \geqslant 1$, $N \geqslant 1$, $c_i > 0$, $d_j > 0$, $\alpha_1 \geqslant \ldots \geqslant \alpha_M \geqslant 1$, $\beta_1 \geqslant \ldots \geqslant \beta_N \geqslant 1$ and h is convex, bounded below, with $\lim_{t\to 0+} h(t) = \infty$. Note that $v_1v_2v_3 = \det A$. Then the hypotheses of Theorem 1 hold provided $\alpha_1 \geqslant 2$ and $\beta_1 \geqslant \alpha_1/(\alpha_1 - 1)$; a special case is the Mooney–Rivlin material, for which $\alpha_1 = \beta_1 = 2$. For the function

$$\Phi(v_1, v_2, v_3) = c(v_1^{\alpha} + v_2^{\alpha} + v_3^{\alpha} - 3) + h(v_1v_2v_3), \tag{11}$$

with $c > 0$ and h as above, the hypotheses of Theorem 1 hold provided $\alpha \geqslant 3$.

A modification of the Saint Venant–Kirchhoff constitutive law satisfying (7), (8) has been proposed by Ciarlet and Geymonat [10].

There are physically interesting stored-energy functions W which are not $W^{1,1}$-quasiconvex and in particular do not satisfy the hypotheses of Theorem 1. We distinguish two ways in which this may occur. The first is when W fails to be quasiconvex, i.e. fails to satisfy (5); this case corresponds to materials which may change phase (see Ericksen [13], James [14, 15]). An example is furnished by an *elastic fluid*, for which

$$W(A) = h(\det A). \tag{12}$$

In this case (5), (6) and (7) are equivalent and are satisfied if and only if h is convex. For a van der Waals fluid, for example, h is not convex. Results proved by Acerbi and Fusco [2] and Dacorogna [12], for integrands taking finite values only, suggest that under strong growth conditions on W any minimizing sequence for $I(\cdot)$ has a subsequence converging weakly in $W^{1,1}(\Omega; \boldsymbol{R}^n)$ to a minimizer for the "relaxed problem" obtained by replacing W by its lower quasiconvex envelope. The corresponding result for an elastic fluid is proved in Dacorogna [11].

A second way in which W can fail to be $W^{1,1}$-quasiconvex is due to its growth properties for large $|A|$. As an example, consider the isotropic stored-energy function (11) with $c > 0$, $h \in C^3(0, \infty)$, $h'' > 0$ and $\lim_{t\to 0+} h(t) = \lim_{t\to\infty} \frac{h(t)}{t} = \infty$. If $\alpha \geqslant 1$ this function satisfies (7), and if $\alpha > 1$ it is *strongly elliptic*, i.e. (6) holds with strict inequality if a, b are non zero. However, if $1 < \alpha < 3$ W is not $W^{1,1}$-quasiconvex. This can be proved by choosing $D = B \overset{\text{def}}{=} \{X \in \boldsymbol{R}^3 : |X| < 1\}$, $A = \lambda 1$, and showing that for sufficiently large $\lambda > 0$ one can violate (5) with an appropriate discontinuous radial function φ. The problem of minimizing $E(\cdot)$ among radial deformations

$$x(X) = \frac{r(R)}{R} X, \quad R = |X|, \tag{13}$$

subject to appropriate displacement or traction boundary data is considered in Ball [7]. For example, for the stored-energy function (11) under the above conditions with $1 < \alpha < 3$ it is shown that for any $\lambda > 0$ the absolute minimum of $E(\cdot)$ among radial deformations (13) satisfying $r(0) \geqslant 0$, $r(R) \geqslant 0$ and $r(1) = \lambda$ is attained. Furthermore, there exists a critical value λ_c such that for $\lambda \leqslant \lambda_c$ the minimizer is trivial and given by $r(R) = \lambda R$ (i.e. $x = \lambda X$), whilst for $\lambda > \lambda_c$ the minimizer satisfies $r(0) > 0$, so that

a cavity forms at the origin. The nontrivial minimizers are discontinuous weak solutions of the full set of 3-dimensional Euler–Lagrange equations for $E(\cdot)$. The reader is referred to [7] for analogous results for more general compressible and incompressible materials, for a discussion of the relevance of discontinuous minimizers to the phenomenon of internal rupture of rubber, and for comments concerning the relationship of the analysis to the literature on discontinuous solutions to nonlinear elliptic systems.

Acknowledgement

This work was supported by a Science & Engineering Research Council Senior Fellowship.

References

[1] Acerbi E., Buttazzo G. and Fusco N., Semicontinuity in L^∞ for Polyconvex Integrals, preprint.

[2] Acerbi E. and Fusco N., Semicontinuity Problems in the Calculus of Variations, *Arch. Rat. Mech. Anal.*, to appear.

[3] Antman S. S., Ordinary Differential Equations of Nonlinear Elasticity. II. Existence and regularity theory for conservative boundary value problems, *Arch. Rat. Mech. Anal.* **61** (1976), pp. 353–393.

[4] Ball J. M., Convexity Conditions and Existence Theorems in Nonlinear Elasticity, *Arch. Rat. Mech. Anal.* **63** (1977), pp. 337–403.

[5] Ball J. M., Constitutive Inequalities and Existence Theorems in Nonlinear Elastostatics, In: R. J. Knops (ed.), *Nonlinear Analysis and Mechanics*, Heriot–Watt Symposium, Vol. 1, Pitman, London, 1977.

[6] Ball J. M., Global Invertibility of Sobolev Functions and the Interpenetration of Matter, *Proc. Royal Soc. Edinburgh A* **88** (1981), pp. 315–328.

[7] Ball J. M., Discontinuous Equilibrium Solutions and Cavitation in Nonlinear Elasticity, *Phil. Trans. Royal Soc. London A* **306** (1982), pp. 557–611.

[8] Ball J. M., Currie J. C. and Olver P. J., Null Lagrangians, Weak Continuity and Variational Problems of Arbitrary Order, *J. Functional Anal.* **41** (1981), pp. 135–174.

[9] Ball J. M. and Murat F., *$W^{1,p}$-Qnasiconvexity and Weak Convegence in the Calculus of Variations*, to appear.

[10] Ciarlet P. G. and Geymonat G., Sur les lois de comportement en élasticité nonlinéaire compressible, *C. R. Acad. Sci. Paris* **295** (1982), pp. 423–426.

[11] Dacorogna B., A Relaxation Theorem and its Application to the Equilibrium of Gases, *Arch. Rat. Mech. Anal.* **77** (1981), pp. 359–386.

[12] Dacorogna B., Quasiconvexity and Relaxation of Nonconvex Problems in the Calculus of Variations, *J. Funct. Anal.* **46** (1982), pp. 102–118.

[13] Ericksen J. L., Some Phase Transitions in Crystals, *Arch. Rat. Mech. Anal.* **73** (1980), pp. 99–124.

[14] James R. D., Finite Deformation by Mechanical Twinning, *Arch. Rat. Mech. Anal.* 77 (1981), pp. 143–176.
[15] James R. D., *Mechanics of Coherent Phase Transformations in Solids*, preprint.
[16] Morrey C. B., Quasi-Convexity and the Lower Semicontinuity of Multiple Integrals, *Pacific J. Math.* 2 (1952), pp. 25–53.
[17] Ogden R. W., Large Deformation Isotropic Elasticity: on the Correlation of Theory and Experiment for Compressible Rubber-Like Solids, *Proc. Royal Soc. London* A328 (1972), pp. 567–583.
[18] Reshetnyak Y. G., On the Stability of Conformal Mappings in Multi-Dimensional Spaces, *Sibirskii Math.* 8 (1967), pp. 91–114.
[19] Reshetnyak Y. G., General Theorems on Semicontinuity and on Convergence with a Functional, *Sibirskii Math.* 8 (1967), pp. 1051–1069.
[20] Reshetnyak Y. G., Stability Theorems for Mappings with Bounded Excursion, *Sibirskii Math.* 9 (1968), pp. 667–684.

DEPARTMENT OF MATHEMATICS
HERIOT-WATT UNIVERSITY
EDINBURGH EH14 4AS,
SCOTLAND

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

O. LADYŽENSKAYA

On Finding Symmetrical Solutions of Field Theory Variational Problems

Field theory functionals are usually invariant under a group G of transformations acting in the space X of its arguments. This fact is being used by physicists for finding critical points of f on X in the following way: a subset X_0 of X invariant under G (or some subgroup of G) is singled out (i.e., an "Ansatz" is suggested); then a critical point $\overset{\circ}{v}$ of f on X_0 is found and afterwards it is confirmed that the point $\overset{\circ}{v}$ should be critical for f on all X. In some cases the last statement is proved either by direct substitution v_0 into Euler equations of the variational problem under investigation or by some special considerations.

S. Coleman in [1] put forward some ideas on how to choose "right Ansätze". Later on, L. D. Faddeev, referring to a conversation with Coleman, formulated those ideas in the form of the following "Principle I" ([2]):

Let X be a manifold, G a group acting on X and f a G-invariant functional on X (i.e., $f(gv) = f(v)$, $\forall v \in X$ and $\forall g \in G$). Let X_0 be the set of all points of X fixed under the action of G (i.e., $\overset{\circ}{v} \in X_0 \Leftrightarrow g\overset{\circ}{v} = \overset{\circ}{v}$, $\forall g \in G$). Then any critical point $\overset{\circ}{v}$ of f on X_0 should be critical also for f on X.

L. D. Faddeev shows in [2] that many "Ansätze" proposed for different models of field theory are in fact nothing else but an indication of G and X_0. Principle I is very useful for getting some solutions of field theory variational problems, those which are less amenable for other methods developed in mathematics ("direct methods" of calculus of variations, the theory of fixed points, etc.). Though Principle I is not valid in such a general setting, the author together with her colleague, L. V. Kapitanskii,

made it our aim to find sufficient conditions (as wide as possible) on X, G and f under which this principle is true.

Let us begin with two examples which clarify the reasons of a possible failure of Principle I.

1. Let $X = \boldsymbol{R}^2$, let f be a smooth function on $\boldsymbol{R}^2$ depending only on x_2 and let G be the group: $g_\tau x = (x_1 + \tau x_2, x_2)$, $\tau \in \boldsymbol{R}^1$. Here $X_0 = \{x: x_2 = 0\}$; X and f are G-invariant but Principle I fails to hold.

2. Let X be the torus surface resulting by rotation of the circle $\{x = (x_1, x_2, x_3) \in \boldsymbol{R}^3: (x_1 - 1)^2 + x_3^2 = 1,\ x_2 = 0\}$ about the x_3-axis, let G be the group of rotations of $\boldsymbol{R}^3$ about the x_3-axis and let f be a smooth function depending only on x_3. Here $X_0 = \{x: x = 0\}$; X and f are G-invariant but Principle I is not applicable.

Looking at the first example, one might think that the noncompactness of G is the cause. But if we adjoint to G the element $\tilde{g}$: $\tilde{g}x = (x_1, -x_2)$, the principle becomes true. In the second example one could blame X: the manifold X has a singularity at the point $x = 0$. But by adjoining to G the element g: $\tilde{g}x = (x_1, x_2, -x_3)$ we reestablish the principle. All an analysis of most problems of field theory shows that leaving manifolds with singularities out of considerations is like splashing the baby out of the bath. Moreover, in many cases X is not a manifold at all and then the notion of criticality of a point $\overset{\circ}{v}$ for f on X_0 and for f on X then requires a special definition. And this will be the starting point of our address.

Let V be a complete Hausdorff topological vector space and let π_G be a set of linear continuous operators $\{\pi(g), g \in G\}$ acting in V. Denote by V_0 the set of all fixed elements of V; i.e., $\overset{\circ}{v} \in V_0 \Leftrightarrow \pi(g)\overset{\circ}{v} = \overset{\circ}{v}$, $\forall g \in G$. V_0 is a closed subspace of V. Let f be an invariant functional on an invariant set $X \subset V$ (i.e., $\pi(g)v \in X$ if $v \in X$ and $f(\pi(g)v) = f(v)$, $\forall v \in X$ and $\forall g \in G$) and write $X_0 \equiv X \cap V_0$. Let us take a point $\overset{\circ}{v} \in X_0$ and define the tangent set $T_{\overset{\circ}{v}}X$ to X at this point in the following way: an element $\eta \in V$ belongs to $T_{\overset{\circ}{v}}X$ if and only if

$$\eta = \left.\frac{dv(\tau)}{d\tau}\right|_{\tau=+0} \equiv \lim_{\tau \to +0} \frac{v(\tau) - v(0)}{\tau}$$

(in the topology of V) for a segment of a curve $v(\tau)$, $\tau \in [0, \varepsilon]$, lying on X with the initial point $\overset{\circ}{v} = v(0)$. The tangent set $T_{\overset{\circ}{v}}X_0$ to X_0 at the point $\overset{\circ}{v}$ is defined in a similar way. It is easy to see that $T_{\overset{\circ}{v}}X$ and $T_{\overset{\circ}{v}}X_0$ are

invariant (as $\pi(g)v(\tau) \in X$, $\pi(g)\overset{\circ}{v} = \overset{\circ}{v}$, $\left.\frac{d\pi(g)v(\tau)}{d\tau}\right|_{\tau=+0} = \pi(g)\eta$) and $T_{\overset{\circ}{v}}X_0 \subset T_{\overset{\circ}{v}}X \cap V_0$.

Assume the existence of a linear invariant set $\mathscr{H}$ dense in $T_{\overset{\circ}{v}}X$ (in topology of V) and such that $\mathscr{H}_0 \equiv \mathscr{H} \cap T_{\overset{\circ}{v}}X_0$ is dense in $T_{\overset{\circ}{v}}X_0$.

Regarding f, we suppose that f has the derivatives at the point $\overset{\circ}{v}$ along $\forall \eta \in \mathscr{H}$, i.e., that

$$\lim_{\tau \to +0} \frac{1}{\tau}\left[f\left(v_\eta(\tau)\right) - f(\overset{\circ}{v})\right] \equiv \delta f(\overset{\circ}{v}; \eta)$$

exists for every curve $v_\eta(\tau)$, $\tau \in [0, \varepsilon]$, $v_\eta(0) = \overset{\circ}{v}$ lying on X and such that

$$\left.\frac{dv_\eta(\tau)}{d\tau}\right|_{\tau=+0} = \eta.$$

Moreover, $\delta f(\overset{\circ}{v}; \eta)$ has to be linear with respect to η.

These hypotheses imply the equalities

$$\delta f(\overset{\circ}{v}; \eta) = \delta f\left(\overset{\circ}{v}; \pi(g)\eta\right), \quad \forall \eta \in \mathscr{H} \text{ and } \forall g \in G. \tag{1}$$

We say that $\overset{\circ}{v}$ is a *critical point of f on X* if

$$\delta f(\overset{\circ}{v}; \eta) = 0, \quad \forall \eta \in \mathscr{H}, \tag{2}$$

and $\overset{\circ}{v}$ is a *critical point of f on X_0* if

$$\delta f(\overset{\circ}{v}; \overset{\circ}{\eta}) = 0, \quad \forall \overset{\circ}{\eta} \in \mathscr{H}_0. \tag{3}$$

Having introduced the necessary notions and definitions, we can formulate our problem in the form of the question: under what conditions do (1) and (3) imply (2)? As long as we do not want to submit f to any special restrictions (except invariance), equalities (1) contain all the information about f. An important feature of (1) is linearity with respect to the second argument η. In virtue of this, the equalities

$$\delta f(\overset{\circ}{v}; \eta) = \delta f\left(\overset{\circ}{v}; \frac{1}{m}\sum_{i=1}^{m} \pi(g_i)\eta\right) \tag{4}$$

are true, $\forall g_i \in G$ and $\forall \eta \in \mathscr{H}$. If for every $\eta \in \mathscr{H}$ there are elements $g_1, \ldots, g_m$

such that $\frac{1}{m}\sum_{i=1}^{m}\pi(g_i)\eta$ belongs to $\mathscr{H}_0$, then (4) and (3) evidently imply (2), i.e., the criticality of $\overset{\circ}{v}$ for f on X follows from the criticality of $\overset{\circ}{v}$ for f on X_0.

More generally, the last statement will be true if there is an 'averaging' operation $\mathscr{T}$ on G having the following properties:

(a) $\mathscr{T}\varphi = \varphi$ for any scalar function φ which is constant on G;

$$\text{(b)}\quad \mathscr{T}\big(\pi(\cdot)\eta\big) = \overline{\eta} \in \hat{\mathscr{H}}_0 \equiv \mathscr{H}\cap V_0, \qquad \forall \eta \in \mathscr{H}, \tag{5}$$

$$\text{(c)}\quad \mathscr{T}\big(\delta f(\overset{\circ}{v};\pi(\cdot)\eta)\big) = \delta f(\overset{\circ}{v};\overline{\eta}), \qquad \forall \eta \in \mathscr{H}. \tag{6}$$

Applying $\mathscr{T}$ to both sides of (1), we get

$$\delta f(\overset{\circ}{v};\eta) = \delta f(\overset{\circ}{v};\overline{\eta}), \qquad \forall \eta \in \mathscr{H}. \tag{7}$$

If $\mathscr{H}_0$ contains the set $\tilde{\mathscr{H}}_0 \equiv \{\overline{\eta}\colon\ \overline{\eta} = \mathscr{T}\big(\pi(\cdot)\eta\big),\ \forall \eta \in \mathscr{H}\}$, then (2) follows from (7) and (3).

Thus we have proved

THEOREM 1. *Let V be a complete Hausdorff topological vector space; let π_G be a set of linear continuous operators $\{\pi(g),\ g \in G\}$ acting in V and let V_0 be the set of all fixed points of V. Let $X \subset V$ be an invariant set and let $\overset{\circ}{v} \in X_0 \equiv X \cap V_0$. Denote by $T_{\overset{\circ}{v}}X$ and $T_{\overset{\circ}{v}}X_0$ the tangent sets to X and X_0 at the point $\overset{\circ}{v}$, respectively. Suppose that $T_{\overset{\circ}{v}}X$ contains an invariant linear set $\mathscr{H}$ dense in $T_{\overset{\circ}{v}}X$ and the set $\mathscr{H}_0 \equiv \mathscr{H}\cap T_{\overset{\circ}{v}}X_0$ is dense in $T_{\overset{\circ}{v}}X_0$. Further suppose that f is an invariant functional defined on the intersection of some neighbourhood of $\overset{\circ}{v}$ with X and that f is differentiable at the point $\overset{\circ}{v}$ along all $\eta \in \mathscr{H}$. Finally, assume that $\overset{\circ}{v}$ is a critical point of f on X_0. Then $\overset{\circ}{v}$ will be a critical point of f on X whenever there exists an "averaging" operation $\mathscr{T}$ on G with properties* (a), (b), (c) *and $\mathscr{H}_0$ contains the set*

$$\tilde{\mathscr{H}}_0 \equiv \{\overline{\eta}\colon\ \overline{\eta} = \mathscr{T}\big(\pi(\cdot)\eta\big),\ \forall \eta \in \mathscr{H}\}.$$

Remark 1. In all applications of Theorem 1 examined below the operation $\mathscr{H}$ has (apart of properties (a), (b), (c)) the following property: $\mathscr{T}\big(\pi(\cdot)\eta\big) \equiv \mathscr{T}(\eta) = \eta$ for $\eta \in \hat{\mathscr{H}}_0$, and therefore $\tilde{\mathscr{H}}_0 = \hat{\mathscr{H}}_0$. On the other hand, $\mathscr{H}_0 \subset \hat{\mathscr{H}}_0$, because $T_{\overset{\circ}{v}}X_0 \subset T_{\overset{\circ}{v}}X \cap V_0$. In virtue of these facts if $T_{\overset{\circ}{v}}X_0 = T_{\overset{\circ}{v}}X \cap V_0$ then $\mathscr{H}_0 = \mathscr{H}\cap V_0 = \hat{\mathscr{H}}_0$ and the condition $\tilde{\mathscr{H}}_0 \subset \mathscr{H}_0$ of Theorem 1 is equivalent to the condition: $\tilde{\mathscr{H}}_0 = \hat{\mathscr{H}}_0 = \mathscr{H}_0$.

Now we shall present some consequences of Theorem 1 in which the operation $\mathscr{T}$ will be written out explicitly.

COROLLARY 1. *Let V be a complete Hausdorff locally convex vector space, let G be a compact group, let dg be the invariant normalized Haar measure on G and let $\{\pi(g),\ g \in G\}$ be a representation of G in V by linear continuous operators $\pi(g)$ having the property: the map $g \to \pi(g)v$ is continuous (from G to V), for any fixed $v \in V$. Let f be a continuous invariant functional defined on V and differentiable (in the sense of Gateaux) at the points of V_0. Then each critical point $\overset{\circ}{v}$ of f on V_0 is also a critical point of f on V.*

This is the simplest case of the situation described in Theorem 1. We have $X = V$, $\mathscr{H} = T_v^{\circ} X = V$, $X_0 = V_0$, $H_0 = T_v^{\circ} X_0 = V_0$. The differentiability of f at the point $\overset{\circ}{v}$ means that $\delta f(\overset{\circ}{v}; \eta) = \langle f'(\overset{\circ}{v}), \eta \rangle$ where $f'(\overset{\circ}{v})$ is an element of the conjugate space V^* and $\langle , \rangle$ is the dual operation between elements of V^* and V. Equations (1) in this case take on the form: $\langle f'(\overset{\circ}{v}), \eta \rangle = \langle f'(\overset{\circ}{v}), \pi(g)\eta \rangle$, $\forall \eta \in \mathscr{H}$. As the operation $\mathscr{T}$ we choose integration $\int_G \ldots dg$. According to the well-known facts (see, for example, [4]), $\mathscr{T}$ possesses properties (a), (b), (c) and $\overline{\eta} \equiv \int_G \pi(g)\eta\, dg$ belong to V_0, $\forall \eta \in \mathscr{H}$. Therefore all conditions of Theorem 1 are fulfilled and the statement is true.

Now we introduce some additional requirements which we call "*Conditions* (A)":

The set $\mathscr{H}$ from Theorem 1 *is dense in a complete Hausdorff locally convex vector space W. Let W^* denote the space conjugate to W and let $\langle , \rangle$ be the dual operation between elements of W^* and W. We demand that $\delta f(\overset{\circ}{v}; \eta) = \langle f'(\overset{\circ}{v}), \eta \rangle$, $\forall \eta \in \mathscr{H}$, where $f'(\overset{\circ}{v}) \in W^*$, and that each $\pi(g)$ may be extended from $\mathscr{H}$ to all of W as a linear continuous operator.*

If conditions (A) are fulfilled then (1) implies the equalities

$$\langle f'(\overset{\circ}{v}), (I - \pi(g))\eta \rangle = 0, \quad \forall \eta \in W,\ \forall g \in G \tag{8}$$

and (3) implies the equalities

$$\langle f'(\overset{\circ}{v}), \overset{\circ}{\eta} \rangle = 0, \quad \forall \overset{\circ}{\eta} \in \overline{\mathscr{H}_0}^W \tag{9}$$

(here and below we denote by $\overline{\mathscr{H}_0}^W$ the closure of $\mathscr{H}_0$ in the topology of W). It is easy to see that the needed equality $f'(\overset{\circ}{v}) = 0 \in W^*$ will follow

from (8) and (9) if the linear set spanned by $\overline{\mathscr{H}_0^W}$ and all elements $(I-\pi(g))\eta$, $\forall \eta \in W$, is dense in W.

We shall give examples of some situations when the last condition is fulfilled.

COROLLARY 2. *Let all hypotheses of Theorem* 1 *except the existence of the operation* $\mathscr{T}$ *be fulfilled and let conditions* (A) *be also satisfied. Moreover, suppose that* G *is a compact group with a normalized Haar mesure* dg *and the map* $g \to \pi(g)w$ *is continuous from* G *to* W, *for any fixed* $w \in W$. *Then* $\overset{\circ}{v}$ *will be a critical point* of f *on* X *if the elements* $\overline{\eta} \equiv \int_G \pi(g)\eta\, dg$, $\forall \eta \in W$, *lay in* $\overline{\mathscr{H}_0^W}$.

Here, just as in Corollary 1, we take $\mathscr{T} = \int_G \dots dg$. The application of $\mathscr{T}$ to (8) gives the identity $\langle f'(\overset{\circ}{v}), \eta\rangle = \langle f'(\overset{\circ}{v}), \overline{\eta}\rangle$ for $\forall \eta \in W$. The elements $\overline{\eta}$ belong to W_0 (the set of all fixed points of W). By the last condition of Corollary 2, we have $\overline{\eta} \in \overline{\mathscr{H}_0^W}$. Hence, (9) implies $f'(\overset{\circ}{v}) = 0 \in W^*$.

The topology of W in a concrete variational problem is prompted by properties of the field $\overset{\circ}{v}(\cdot)$ which is the solution of the 'reduced' variational problem. Often $\overset{\circ}{v}(\cdot)$ is a smooth function of $x \in \boldsymbol{R}^n$, except for some points x. This allows one to represent $\delta f(\overset{\circ}{v}; \eta)$ in the form of $\int_{\boldsymbol{R}^n} Z(\overset{\circ}{v})\cdot\eta\, dx$, where $Z(\overset{\circ}{v})$ is the left-hand side of the Euler equations and to consider $\delta f(\overset{\circ}{v}; \eta)$ as inner product in a Hilbert space $L_2(\boldsymbol{R}^n; \mu)$. One can take as $\mathscr{H}$ the set of smooth fields $\eta(\cdot)$ which are zero near the singular points of $\overset{\circ}{v}(\cdot)$ and which are also submitted to certain linear equalities if there are some ties in the problem. In these cases W is a closed subspace of $L_2(\boldsymbol{R}^2; \mu)$. Therefore the following corollary of Theorem 1 may be useful.

COROLLARY 3. *Suppose that all hypotheses of Theorem* 1 *except the existence of the operator* $\mathscr{T}$ *are fulfilled. Let also conditions* (A) *be satisfied* W *with being a Hilbert space with inner product* $(\ ,\ )$. *Then* $\overset{\circ}{v}$ *is critical point of* f *on* X *if* $\overline{\mathscr{H}_0^W}$ *is the set* W_0 *of all fixed points of* W *and if*

$$W_0 \equiv \bigcap_{g\in G} \ker_W (I-\pi(g)) = \bigcap_{g\in G} \ker_W (I-\pi^*(g)). \tag{10}$$

Equality (10) *is true if*: (1) *the set* $\{\pi(g), \forall g \in G\}$ *coincides with the set* $\{\pi^*(g), \forall g \in G\}$ *or* (2) *the operators* $\pi(g)$ *are normal* (*i.e.*, $\pi(g)\pi^*(g) = \pi^*(g)\pi(g)$).

Under the hypotheses of Corollary 3, equalities (8) may be rewritten in the form

$$\left(\left(I-\pi^*(g)\right)f'(\mathring{v}),\,\eta\right)=0,\qquad \forall\eta\in W,$$

and hence $f'(\mathring{v})\in\bigcap_{g\in G}\ker^W\left(I-\pi^*(g)\right)$. This fact and (10) imply $f'(\mathring{v})\in\bigcap_{g\in G}\ker_W\left(I-\pi(g)\right)\equiv W_0$. Since $W_0=\overline{H}_0^W$, equalities (9) give $f'(\mathring{v})=0\in W$.

Let us point out one particular case of the situation described in Corollary 3.

COROLLARY 3′. *Let V be a Hilbert space and let $\{\pi(g),\,g\in G\}$ be a set of bounded linear operations satisfying condition* (10). *Suppose that in a neighbourhood of a point $\mathring{v}\in V_0$ there is an invariant functional differentiable (in the sense of Gateaux) at $\mathring{v}$ and that $\mathring{v}$ is critical point of f on V_0. Then $\mathring{v}$ is a critical point of f on V as well.*

Another consequence is the following:

COROLLARY 4. *Suppose that all hypotheses of Theorem* 1 *except the existence of the operation $\mathscr{T}$ are fulfilled and that conditions* (A) *are satisfied. Suppose also that π_G consists of just one operator $\pi(g)$, which has the following property:*

the limit

$$\text{s-lim}_{m\to\infty}\frac{1}{m}\left[\pi(g)\eta+\ldots+\left(\pi(g)\right)^m\eta\right]\equiv\text{s-lim}_{m\to\infty}\eta_{(m)}\equiv\overline{\eta}\quad\textit{for}\ \forall\eta\in W$$

exists (in the topology of W).

Then $\mathring{v}$ is a critical point of f on X if $\overline{\eta}\in\overline{\mathscr{H}_0^W}$ for all $\eta\in W$.

In this case the operation $\mathscr{T}\left(\pi(\cdot)\eta\right)=\text{s-lim}_{m\to\infty}\eta_{(m)}\equiv\overline{\eta}$ is taken for $\mathscr{T}$. The elements $\overline{\eta}$ belong to $W_0\equiv\text{Ker}_W\left(I-\pi(g)\right)$ and therefore the last condition of Corollary 4 is fulfilled if $\overline{\mathscr{H}_0^W}=W_0$ (but, in general, $\overline{\mathscr{H}_0^W}$ is a subset of W_0).

Theorem 1 and its corollaries can be extended to nonlinear continuous operators $\pi(g)$ if they have differentials at the point $\mathring{v}$. In such cases it is only necessary to impose the conditions previously demanded from $\pi(g)$ as the operators in $\mathscr{H}$ (or in W) on their differentials.

The linearity of the enveloping space V is also not very important, as long as we study local problems and can work within one chart representing a small neighbourhood of the point $\overset{\circ}{v}$. For the variational problems which we have in mind it is convenient to imbed X (the domain of definition of f) in a linear space V and choose V (and W) in dependence of the properties of the solution $\overset{\circ}{v}(\cdot)$ of the "reduced" variational problem, the "qualities" of ties and limit-conditions of the original problem.

Now I would like to show how our Theorem 1 works in the model of Skyrme ([5]). In this model the functional f is defined by the integral

$$f(\Phi) = \int_{\boldsymbol{R}^3} \left\{ \varkappa^2 \sum_{k=1}^{3} \Phi_{x_k}^2 + \frac{1}{2} \sum_{k,f=1}^{3} [\Phi_{x_k}^2 \Phi_{x_l}^2 - (\Phi_{x_k}, \Phi_{x_l})^2] \right\} dx \equiv \int_{\boldsymbol{R}^3} F(\Phi)\, dx, \tag{11}$$

where

$$\Phi(x) = \big(\Phi_0(x), \Phi_1(x), \Phi_2(x), \Phi_3(x)\big) \equiv \big(\Phi_0(x), \vec{\Phi}(x)\big) \in \boldsymbol{R}^4,$$

$$\Phi_{x_k}^2 = \sum_{\alpha=0}^{3} (\Phi_{\alpha x_k})^2; \qquad (\Phi_{x_k}, \Phi_{x_l}) = \sum_{\alpha=0}^{3} \Phi_{\alpha x_k} \Phi_{\alpha x_l},$$

and

$$\varkappa^2 = \text{const} > 0.$$

The problem consists in finding critical points of $f(\Phi)$ among fields $\Phi(\cdot)$ which satisfy the constraint

$$|\Phi(x)| \equiv \Big(\sum_{\alpha=0}^{3} \Phi_\alpha^2(x) \Big)^{1/2} = 1, \qquad \forall x \in \boldsymbol{R}^3, \tag{12}$$

and for which the degree $d(\Phi)$ of the map $\Phi\colon \boldsymbol{R}^3 \to S^3$ is a fixed integer m. The last condition makes sense only if the fields $\Phi(\cdot)$ are continuous functions of $x \in \overline{\boldsymbol{R}}^3 \equiv \boldsymbol{R}^3 \cup \{\infty\}$. That is the reason why we have chosen $V = C(\overline{\boldsymbol{R}}^3)$. $C(\overline{\boldsymbol{R}}^3)$ is the Banach space of all continuous (on $\overline{\boldsymbol{R}}^3$) fields $\Phi(\cdot)$ with the norm $\|\Phi(\cdot)\|_V = \sup\limits_{x \in \boldsymbol{R}^3} |\Phi(x)|$. Following Skyrme, we take as G the group SO(3) and its representation

$$\big(\pi(g)\Phi\big)(x) = \big(\Phi_0(gx), (g^{-1}\vec{\Phi})(gx)\big)$$

in V. The subset V_0 consists of all fields having the form

$$\Phi(x) = \left(u(r); \frac{\vec{x}}{r} U(r)\right) \quad (r \equiv |x|), \tag{13}$$

where $u(\cdot)$ and $U(\cdot)$ are continuous functions on $\bar{R}^1_+ = \{r: r \in [0, \infty]\}$ and $U(0) = U(\infty) = 0$.

We introduce the set M of all fields $\Phi(\cdot)$ from V which are subject to (12), satisfy $d(\Phi) = 1$ and for which $\Phi(0) = (-1; \vec{0})$, $\Phi(\infty) = (1; \vec{0})$ (the case of $d(\Phi) = m$ for other $m \in \mathbf{Z}$ is considered in the same way).

Fields $\Phi(\cdot)$ from $M_0 \equiv M \cap V_0$ may be represented in the form

$$\Phi(x) = \left(\operatorname{Cos} \frac{\omega(r)}{2}; \frac{\vec{x}}{r} \operatorname{Sin} \frac{\omega(r)}{2}\right), \tag{14}$$

where $\omega(\cdot)$ is an arbitrary continuous function on $\bar{R}^1_+$ satisfying the conditions:

$$\omega(0) = 2\pi, \qquad \omega(\infty) = 0. \tag{15}$$

The set X consists of fields $\Phi(\cdot)$ from M which have generalized first derivatives and for which $f(\Phi) < \infty$ and $X_0 = X \cap M_0$. The functional $f(\Phi)$ on the $\Phi(\cdot) \in X_0$ has the form:

$$\hat{f}(w) \equiv f(\Phi) = \pi \int_0^\infty \left[(\omega')^2 \left(\varkappa^2 r^2 + 1 - \operatorname{Cos} \omega(r)\right) + \right.$$

$$\left. + \frac{1 - \operatorname{Cos} \omega(r)}{r^2} \left(4\varkappa^2 r^2 + 1 - \operatorname{Cos} \omega(r)\right)\right] dr. \tag{16}$$

The "reduced" variational problem is the problem of finding the function $\overset{\circ}{\omega}(\cdot)$ which gives the infimum of $\hat{f}(\omega)$ among functions $\omega(\cdot)$ satisfying conditions (15). This "reduced" problem had been discussed by Skyrme in [5], but the question concerning the existence of such a function $\overset{\circ}{\omega}(\cdot)$ and the question of explaining why the field $\overset{\circ}{\Phi}(\cdot)$ corresponding to this $\overset{\circ}{\omega}(\cdot)$ is a critical point of the original problem remained open. We proved at the beginning that the "reduced" problem really has the solution $\overset{\circ}{\omega}(r)$ and that this function $\overset{\circ}{\omega}(r)$ has all derivatives, for $\forall r > 0$. The function

$\overset{\circ}{\omega}(\cdot)$ is found as the limit of a minimizing sequence. I shall not give here this proof. It is more or less standard, except for some a priori estimates. But I would like only to explain why the field $\overset{\circ}{\Phi}(x) = \left(\mathrm{Cos}\,\frac{\overset{\circ}{\omega}(r)}{2}, \frac{\vec{x}}{r}\,\mathrm{Sin}\,\frac{\overset{\circ}{\omega}(r)}{2}\right)$ is indeed the solution of the original problem.

The tangent set $T_{\overset{\circ}{\Phi}}M$ is the closed subset of V consisting of all elements $\eta(\cdot) \in V$ which satisfy the conditions

$$\left(\eta(x), \overset{\circ}{\Phi}(x)\right) = 0, \qquad \eta(0) = \eta(\infty) = 0 \in \boldsymbol{R}^4. \tag{17}$$

The set $T_{\overset{\circ}{\Phi}}X$ contains smooth fields $\eta(\cdot) \in T_{\overset{\circ}{\Phi}}M$ which are equal to zero near the points $x = 0$ and $x = \infty$. The set of such fields $\eta(\cdot)$ is linear, invariant and dense in $T_{\overset{\circ}{\Phi}}M$ and we can choose it as the set $\mathscr{H}$. The set $T_{\overset{\circ}{\Phi}}M_0 = T_{\overset{\circ}{\Phi}}M \cap V_0$, $T_{\overset{\circ}{\Phi}}X_0 = T_{\overset{\circ}{\Phi}}X \cap V_0$ and the set $\mathscr{H}_0 \equiv T_{\overset{\circ}{\Phi}}X_0 \cap \mathscr{H}$ are dense in $T_{\overset{\circ}{\Phi}}X_0$. The variation $\delta f(\overset{\circ}{v}; \eta)$ for $\eta \in \mathscr{H}$ can be represented in the form

$$\delta f(\overset{\circ}{v}; \eta) = \int_{\boldsymbol{R}^3} \left(Z(\overset{\circ}{\Phi}), \eta\right) dx, \tag{18}$$

where $Z(\overset{\circ}{\Phi}) = \left.\frac{\delta F(\Phi)}{\delta \Phi}\right|_{\Phi = \overset{\circ}{\Phi}}$ ($\delta F(\Phi)/\delta\Phi$ is calculated formerly, without regard of the tie (12)). Therefore equalities (1) may be written as

$$\int_{\boldsymbol{R}^3} \left(Z(\overset{\circ}{\Phi}), \eta\right) dx = \int_{\boldsymbol{R}^3} \left(Z(\overset{\circ}{\Phi}), \pi(g)\eta\right) dx. \tag{19}$$

As the operation $\mathscr{T}$ we take integration over the group $G = \mathrm{SO}(3)$. The integral $\int_G \pi(g)\eta(x)\,dg \equiv \overline{\eta}(x)$ is for any $\eta(\cdot) \in \mathscr{H}$ an element $\overline{\eta}$ of $\mathscr{H}_0$. Thus all conditions of Theorem 1 are fulfilled and therefore $\overset{\circ}{\Phi}(\cdot)$ is a critical point of the original problem, i.e., $\int_{\boldsymbol{R}^3} \left(Z(\overset{\circ}{\Phi}), \eta\right) dx = 0$, $\forall \eta \in \mathscr{H}$, $\overset{\circ}{\Phi}(\cdot) \in V$, $f(\overset{\circ}{\Phi}) < \infty$, $|\overset{\circ}{\Phi}(x)| = 1$, $\forall x \in \boldsymbol{R}^3$ and $d(\overset{\circ}{\Phi}) = 1$. Using these facts it is not difficult to show that $\overset{\circ}{\Phi}(x)$ satisfies the Euler system

$$Z(\overset{\circ}{\Phi})(x) = \left(Z(\overset{\circ}{\Phi})(x), \overset{\circ}{\Phi}(x)\right)\overset{\circ}{\Phi}(x) \tag{20}$$

for every $x \in \boldsymbol{R}^3 \setminus \{0\}$.

With the help of Theorem 1 and its corollaries we have proved analoguous results for some other models of field theories (for example, for the Yang–Mills–Higgs model) and for certain more traditional functionals $\int_{R^n} F(x, u, u_x)\,dx$.

For Euclidean Yang–Mills equations in R^4 with the structural group SU(2) we have investigated the "spherically symmetric Ansatz". Namely, we have considered the following action of the group SO(3) on the space of potentials A_μ:

$$A_\mu(x) \equiv \sum_{k=1}^{3} A_\mu^k(x)\sigma_k \to (\pi(g)A)_\mu(x) = A_\alpha(gx)\hat{g}_{\alpha\mu}, \qquad \alpha, \mu = 1, \ldots, 4, \quad (21)$$

where σ_k ($k = 1, 2, 3$) are the imaginary quaternion units (i.e., $\sigma_k^2 = -1$),

$$g = \sum_{\mu=1}^{4} g_\mu \sigma_\mu, \quad \sigma_4 = 1, \quad |g| = 1 \text{ and } gx = g_\mu \sigma_\mu x_\nu \sigma_\nu \equiv \hat{g}_{\mu\nu} x_\nu \sigma_\mu.$$

The fixed points $A_\mu(\cdot)$ of this action have the form

$$A_\mu(x) = \frac{1}{|x|} f_\alpha(\ln |x|^2)\, w_{\alpha\mu}\left(\frac{x}{|x|}\right), \qquad \mu = 1, \ldots, 4, \quad (22)$$

where $f_\alpha(\cdot)$ are arbitrary smooth functions of $\tau \in R^1$ with pure imaginary quaternion values and

$$w_{\alpha\mu}\left(\frac{x}{|x|}\right) = \delta_{4\alpha} n_\mu - \delta_{4\mu} n_\alpha + \delta_{\alpha\mu} n_4 + \varepsilon_{\alpha\mu\gamma 4} n_\gamma, \qquad n_\gamma = \left(\frac{x_\gamma}{|x|}\right). \quad (23)$$

The Y.–M. functional

$$J(A) = \int_{R^4} \frac{1}{4} \sum_{\mu,\nu=1}^{4} |F_{\mu\nu}(x)|^2 dx, \quad (24)$$

where $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu + [A_\mu, A_\nu]$, and the functional $Q(A)$ of topological charge are invariant under transformations (21). The densities of J and Q

for A_μ from (22) depend only on $|x|$ and

$$J(A) \equiv S(f) = \varkappa \int_{-\infty}^{\infty} \Big\{ \frac{1}{2} \sum_{i=1}^{3} |\dot{\vec{f}}_i|^2 + \frac{1}{2} \sum_{i=1}^{3} |\vec{f}_i|^2 + (\dot{\vec{f}}_i, \vec{f}_4, \vec{f}_i) + \\ + \frac{1}{2} \sum_{i=1}^{3} |\vec{f}_i \times \vec{f}_4|^2 + \frac{1}{4} \sum_{i,j=1}^{3} |\vec{f}_i \times \vec{f}_j|^2 - 3(\vec{f}_1, \vec{f}_2, \vec{f}_3) \Big\} d\tau, \tag{25}$$

$$Q(f) = \int_{-\infty}^{\infty} \frac{d}{d\tau} \Big\{ \sum_{i=1}^{3} |\vec{f}_i|^2 - 2(\vec{f}_1, \vec{f}_2, \vec{f}_3) \Big\} d\tau \equiv \int_{-\infty}^{\infty} \frac{d\Phi(f)}{d\tau} d\tau. \tag{26}$$

Here $\dot{\vec{f}}_i = d\vec{f}_i/d\tau$, $\vec{f}_\mu(\tau)$ are vectors in R^3 with components $f_\mu^k(\tau)$, $k = 1, 2, 3$, defined by $f_\mu(\tau) = \sum_{k=1}^{3} f_\mu^k(\tau)\sigma_k$, $|\vec{f}_\mu(\tau)| \equiv \big[\sum_{k=1}^{3} (f_\mu^k(\tau))^2\big]^{1/2}$ and $(\vec{f}_\mu, \vec{f}_\nu, \vec{f}_\gamma)$ is the mixed product. The functionals (25) and (26) are invariant under gauge transformations:

$$f_k \to q f_k q^{-1}, \quad f_4 \to q f_4 q^{-1} - 2\dot{q}q^{-1},$$

where $q = q(\cdot)$ are arbitrary smooth quaternion-valued functions with $|q(\tau)| = 1$. We choose $q(\cdot)$ in such a way that $f_4(\tau) \equiv 0$. Then $S(f)$ can be written as

$$S(f) = \varkappa \int_{-\infty}^{\infty} \Big\{ \frac{1}{2} \sum_{i=1}^{3} |\dot{\vec{f}}_i|^2 + K(f) \Big\} d\tau, \tag{27}$$

where $K(f) = \frac{1}{2} \sum_{i=1}^{3} |\vec{f}_i - \vec{\psi}_i|^2$ and $\vec{\psi}_k = \frac{1}{2} \varepsilon_{klm} \vec{f}_l \times \vec{f}_m$.

The self-dual and anti-self-dual equations for (27) are

$$\dot{\vec{f}}_k = \vec{f}_k - \vec{\psi}_k, \tag{28}$$

$$\dot{\vec{f}}_k = -\vec{f}_k + \vec{\psi}_k, \tag{29}$$

and Euler's equations are

$$\dot{\vec{f}}_i = \vec{f}_i - 3\vec{\psi}_i + 2\vec{\chi}_i, \tag{30}$$

with $\vec{\chi}_i = \frac{1}{2}\varepsilon_{iklm}\vec{\psi}_k \times \vec{f}_m$. The following theorem is true:

THEOREM 2. *If* $S(f) < \infty$ *and* $Q(f) > 0$ *for a solution* $f(\cdot)$ *of eqs.* (30) *then* $f(\cdot)$ *satisfies eqs.* (28), *and if* $S(f) < \infty$ *and* $Q(f) < 0$, *then* $f(\cdot)$ *satisfies eqs.* (29).

The proof of the first statement is based on the representation of the system (30) as the system

$$\dot{\vec{f}}_i = \vec{f}_i - \vec{z}_i - \vec{\psi}_i, \tag{31}$$

$$\dot{\vec{z}}_i = -\vec{z}_i + \varepsilon_{ilm}\vec{z}_l \times \vec{f}_m. \tag{32}$$

It is easy to see that $z(\tau) \equiv 0$ if $z(\tau_0) = 0$ for some τ_0 and the "principal part" of (32) near the point $f = 0$ in the Euclidean space $R^9(f)$ is $\dot{\vec{z}}_i = -\vec{z}_i$. We have also obtained a stronger result:

THEOREM 3. *Suppose that* $(f(\cdot); z(\cdot))$ *is a solution of the system* (31), (32) such that $z(\tau_0) \neq 0$ *for a* τ_0 *and* $J(f) < \infty$. *Then* $f(\tau)$ *cannot enter a small ball* $K_\varrho = \left\{f\colon |f| \equiv \sqrt{\sum_{i=1}^{3} |\vec{f}_i|^2} \leqslant \varrho\right\}$ (*for example,* $\varrho = 1/8$) *infinitely many times as* $t \to -\infty$.

The systems (28) and (29) can be solved explicitly in elliptic functions. For this purpose we introduce the new argument $s = \exp\tau\,(= |x|^2)$ and new functions $\vec{v}_i(s) = s^{-1}\vec{f}_i(\ln s)$. The system (28) implies the system

$$\frac{d\vec{v}_1}{ds} = \vec{v}_3 \times \vec{v}_2, \quad \frac{d\vec{v}_2}{ds} = \vec{v}_1 \times \vec{v}_3, \quad \frac{d\vec{v}_3}{ds} = \vec{v}_2 \times \vec{v}_1. \tag{33}$$

This system can be rewritten as the matrix-equation

$$\frac{dv}{ds} = -(v^{-1})^T \det v \tag{34}$$

for the matrix $v = [v_{ij}]$ with components $v_{ij} = v_i^j$. Equation (34) has two important properties: (i) it is invariant under transformations of the

form $v \to V_1 v V_2 \equiv w$ with arbitrary constant matrices $V_k \in O(3)$, with $\det V_1 = \det V_2$ and (ii) if $v(s_0)$ has the diagonal form for some s_0 then $v(s)$ is diagonal for all s. Owing to these properties, it is sufficient to consider only the case where

$$v(s) = \begin{bmatrix} \lambda_1(s) & 0 & 0 \\ 0 & \lambda_2(s) & 0 \\ 0 & 0 & \lambda_3(s) \end{bmatrix}.$$

For such $v(\cdot)$ equation (34) gives the system

$$\frac{d\lambda_1}{ds} = -\lambda_2\lambda_3, \quad \frac{d\lambda_2}{ds} = -\lambda_1\lambda_3, \quad \frac{d\lambda_3}{ds} = -\lambda_1\lambda_2. \tag{35}$$

It has the integrals $\lambda_1^2 - \lambda_2^2 = c_3$, $\lambda_1^2 - \lambda_3^2 = c_2$, and therefore satisfies the equation

$$\left(\frac{d\lambda_1}{ds}\right)^2 = (\lambda_1^2 - c_2)(\lambda_1^2 - c_3), \tag{36}$$

which can be integrated in elliptic functions. The same is true for λ_2 and λ_3. There are solutions of eqs. (35) which tend to infinity when s goes to some $s_0 \neq \infty$.

THEOREM 4. *There is only one (up to a gauge transformations) solution A_μ of the Y.–M. equations which has the form* (22), *finite $J(A)$ and $Q(A) > 0$. It is the* 1*-instanton satisfying the self-dual equations. For $Q(A) < 0$ it is the* 1*-instanton satisfying the anti-self-dual equations.*

The solutions $\lambda_k(s) = (c+s)^{-1}$, $c > 0$, $k = 1, \ldots; 3$, of the eqs. (35) give the instanton with $Q = 1$ (the instanton of Belavin, Polyakov, Schwartz and Tyupkin).

After this work had been completed we have learned about paper [6] of R. S. Palais devoted to the "Principle I" ("the principle of symmetric criticality" as Palais calls it). In this paper Palais is concerned with smooth functionals f on smooth Banach manifolds mainly with compact groups G and gives interesting applications to some geometrical problems.

References

[1] Coleman S., Classical lumps and their quantum descendants, *Preprint, Harvard Phys. Dept.* (1975), pp. 1–111.

[2] Faddeev L. D., On finding of many dimensional solutions in: *Nonlocal nonlinear*

and non-normalized field theories, Institute of Nuclear Physics, Dubna 1976, pp. 207–223.

[3] Kapitanskii L. V., and Ladyženskaya O. A., On Coleman's principle of finding of critical points of invariant functionals, *Zapiski nauch. semin. LOMI* **127** (1983), pp. 84–102.

[4] Rudin U., *Functional analysis*, Mir, Moskva 1976, 443 pp.

[5] Skyrme T. H. R., A non-linear field theory, *Proc. Royal Soc. of London, ser. A*, **260**, No. 1300 (1961), pp. 127–138.

[6] Palais R. S., The principle of symmetric criticality, *Comm. Math. Phys.* **69**(1) (1979), pp. 19–30.

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

L. A. TAKHTAJAN

Integrable Models in Classical and Quantum Field Theory

Recent papers of Faddeev, Sklyanin and the author ([25], [27], [28], [33]) contain a quantum version of the inverse scattering method (see also the reviews and lectures [7]–[9], [15], [23]). This is a new method of exact solution of the models in 1+1 dimensional quantum field theory and in classical statistical mechanics on a two-dimensional lattice. The profound papers of Baxter [1]–[3] have also played an important role in the formation of the method. In the present talk I will try to explain the underlying ideas and basic constructions of this new domain of modern mathematical physics and also to point out its connections with other parts of mathematics and theoretical physics. Our exposition (as it can be seen from the references) will be mostly based on the results obtained in the Leningrad Branch of the Steklov Mathematical Institute.

I. Classical theory

1. Formulation of the inverse scattering method. This method applies to equations which can be represented as the zero curvature condition $[V_0, V_1] = 0$ for the connection $V_\mu = \partial/\partial x_\mu - U_\mu(\vec{x}, \lambda)$, $\mu = 0, 1$; $x_0 = t$, $x_1 = x$ defined in a trivial fibre bundle with the base $\boldsymbol{R}^2$ (the space-time) and the fibre $\boldsymbol{C}^m$ (the auxiliary space). Matrix elements of the matrices $U_\mu(x, t, \lambda)$ depend on the classical fields $\psi^\alpha(x, t)$, $\alpha = 1, \ldots, M$, involved in the equation considered, and on the variable $\lambda \in \boldsymbol{C}^1$ called the *spectral parameter*. The zero curvature condition must hold for all λ.

A fundamental role in the method is played by the equation of parallel translation along the x-axis

$$\frac{\partial F}{\partial x} = U(x, \lambda)F. \tag{1}$$

Here we put $U = U_1$ and reduce the dependence on t. In the case of periodic initial data $\psi^a(x+2L) = \psi^a(x)$ there naturally emerge the transition matrix $T(x, y, \lambda) = \overleftarrow{\exp} \int_y^x U(x', \lambda)dx'$ which is a solution of equation (1) with the initial condition $T(x, y, \lambda)|_{x=y} = I$ and the monodromy matrix $T_L(\lambda)$ for the interval $(-L, L)$, $T_L(\lambda) = T(L, -L, \lambda)$. Here I denotes the unit matrix in C^m and the integral is understood as a multiplicative one. The time dependence of these objects is determined by the equation

$$\overleftarrow{\exp} \int_\gamma U_\mu d\dot{x}_\mu = I \tag{2}$$

which holds for any closed contour $\gamma \subset R^2$, due to the zero curvature condition. In particular, the functionals $\mathrm{tr}T_L^k(\lambda)$, $k = 1, \ldots, m-1$, where tr stands for the trace in C^m, do not depend on t and play the role of generating functions for the integrals of motion.

In the rapidly decreasing case, where $\lim_{x\to\pm\infty} \psi^a(x) = \psi_\pm^a$, there appear the reduced monodromy matrix $T(\lambda)$, relating the left and the right Jost solutions, and the characteristics of the discrete spectrum. In these terms the dynamics becomes completely transparent and the matrix $U(x, \lambda)$ is uniquely determined by them. This procedure is based on the formalism of the Riemann problem, i.e., the problem of analytic factorization of matrix-valued functions. Here we have a connection with the theory of functions.

These are the basic items of the classical inverse scattering method (see [9], [12], [38]; our exposition follows [9], [12]). It is applied to such well-known equations as the Korteweg–de Vries equation (KdV), the nonlinear Schrödinger equation (NS), the Sine–Gordon equation (SG), the Heisenberg magnet equation (HM), and others.

2. The Hamiltonian approach. The most elegant formulation of the method is its Hamiltonian formulation originating from paper [37]. Its present form is based on the concept of a classical r-matrix introduced in [24]. A classical r-matrix is defined as a matrix $r(\lambda)$ in $C^m \otimes C^m$ which enables

one (if possible) to write down all the Poisson brackets of matrix elements of $U(x, \lambda)$ in the following compact form

$$\{U(x, \lambda) \underset{,}{\otimes} U(y, \mu)\} = [r(\lambda-\mu), U(x, \lambda) \otimes I + I \otimes U(x, \mu)]\, \delta(x-y). \quad (3)$$

Here $\{A \underset{,}{\otimes} B\}$ denotes the matrix in $C^m \otimes C^m$ of the same structure as $A \otimes B$ with the products of matrix elements replaced by their Poisson brackets.

As an illustrating example consider the HM model with the equation of motion

$$\frac{\partial \vec{S}}{\partial t} = \frac{\partial^2 \vec{S}}{\partial x^2} \wedge \vec{S}; \quad \vec{S}(x, t) \in R^3, \quad \langle \vec{S}, \vec{S} \rangle = 1. \quad (4)$$

We have ([30])

$$U_0 = -\frac{1}{2\lambda} S \frac{\partial S}{\partial x} + \frac{1}{2i\lambda^2} S, \quad U_1 = \frac{i}{2\lambda} S, \quad S = \langle \vec{S}, \vec{\sigma} \rangle, \quad (5)$$

where $\frac{1}{2i} \sigma_a$, $a = 1, 2, 3$, are the generators of the fundamental representation of the Lie algebra su(2): σ_a are called the Pauli matrices. The Poisson brackets and the Hamiltonian of the model are

$$\{S_a(x), S_b(y)\} = \varepsilon_{abc} S_c(y)\, \delta(x-y), \quad (6)$$

$$H = \frac{1}{2} \int_{-L}^{L} \left\langle \frac{\partial \vec{S}}{\partial x}, \frac{\partial \vec{S}}{\partial x} \right\rangle dx, \quad (7)$$

where ε_{abc} are the structure constants of su(2). The r-matrix is

$$r(\lambda) = P/2\lambda, \quad (8)$$

where P is the permutation operator in $C^2 \otimes C^2$.

The r-matrix formalism is based on the following result, though simple, yet important.

THEOREM 1. *Suppose that relation* (3) *holds. Then for the transition matrix of equation* (1) *we have*

$$\{T(x, y, \lambda) \underset{,}{\otimes} T(x, y, \mu)\} = [r(\lambda-\mu), T(x, y, \lambda) \otimes T(x, y, \mu)], \quad (9)$$

where $-L \leqslant y \leqslant x \leqslant L$.

As a corollary we obtain that $\{\mathrm{tr}T_L^k(\lambda), \mathrm{tr}T_L^l(\mu)\} = 0; k, l = 1, \ldots, m-1$, which is the involution property of the families of integrals of motion including the Hamiltonian of the model. Using Theorem 1 one can establish complete integrability in the rapidly decreasing case by constructing a canonical transformation to variables of action-angle type ([9], [12]).

One of the characteristic features of the Hamiltonian formulation is that the r-matrix replaces the zero curvature condition ([25], [9], [12], [35]).

THEOREM 2. *Suppose that for the matrix $U(x, \lambda)$ in equation* (1) *the condition* (3) *holds. Then for the generic Hamiltonian equation*

$$\frac{\partial \psi^a(x)}{\partial t} = \{\mathrm{tr}T_L(\mu), \psi^a(x)\}, \quad a = 1, \ldots, M, \tag{10}$$

where μ is a parameter, the zero curvature condition holds with $U_1(x, \lambda) = U(x, \lambda)$ and

$$U_0(x, \lambda, \mu) = \mathrm{tr}_1\big((T_L(L, x, \mu) \otimes I) r(\mu - \lambda)(T_L(x, -L, \mu) \otimes I)\big). \tag{11}$$

Here tr_1 stands for the trace in the first factor in $C^m \otimes C^m$.

3. Symplectic structure associated with an r-matrix. Formula (3) can also be interpreted as the way of defining the symplectic structure on the phase space parametrized by the functions $\psi^a(x)$, $a = 1, \ldots, M$. The skew-symmetry of the Poisson brackets is provided by the condition $r(\lambda) = -Pr(-\lambda)P$ and the relation

$$[r_{12}(\lambda - \mu), r_{13}(\lambda)] + [r_{12}(\lambda - \mu), r_{23}(\mu)] + [r_{13}(\lambda), r_{23}(\mu)] = 0 \tag{12}$$

guarantees the Jacobi identity. Here P is the permutation matrix in $C^m \otimes C^m$ and r_{12} denotes the matrix in $C^m \otimes C^m \otimes C^m$ which acts trivially in the third factor and coincides with r in the product of the first two (analogously for r_{13} and r_{23}). The relation (12) is called the *classical Yang–Baxter equation* (or the *classical triangle equation*) and is quite popular nowadays ([5], [22]). Thus, in [5] the solutions of (12) associated with simple Lie algebras were constructed. As a by-product of the study of this equation a new object, the Lie–Hamilton group, has appeared in [6].

4. Geometrical interpretation of the r-matrix Poisson brackets. There exists a very elegant interpretation of the r-matrix in the language of the representation theory ([9], [11]): the Poisson brackets defined by an

r-matrix of the form (8) are just the Lie–Poisson brackets for an infinite-dimensional Lie algebra. More precisely, let g be the finite-dimensional semisimple Lie algebra. The Lie–Poisson bracket on the phase space g^* is

$$\{f, g\}(\xi) = \sum_{a,b,c} C_{abc} \frac{\partial f}{\partial \xi_a} \frac{\partial g}{\partial \xi_b} \xi_c, \qquad \xi \in g^*, \tag{13}$$

where C_{abc} are the structure constants of g. The brackets (13) can be naturally lifted to define the Poisson brackets for the functionals on $\tilde{g}^*$—the dual space of the current algebra $\tilde{g}$. The latter is just the Lie algebra of the Laurent series in the variable λ with coefficients in g. These brackets are also defined on the dual space of the subalgebra $\tilde{g}_+$ consisting of the Laurent series in the negative powers of λ; the same holds for the complementary subalgebra $\tilde{g}_-$. Moreover, let K_{ab} be the matrix of the Killing form in the basis X_a in g, let K^{ab} be its inverse and

$$\Pi = \sum_{a,b} K^{ab} X_a \otimes X_b. \tag{14}$$

Then we have the following theorem ([9], [11]).

THEOREM 3. *The Poisson bracket for the generic element $U(\lambda) \in \tilde{g}_+^*$ is*

$$\{U(\lambda) \otimes U(\mu)\} = [r(\lambda-\mu), U(\lambda)\otimes I + I \otimes U(\mu)], \tag{15}$$

where $r(\lambda) = \Pi/\lambda$.

Introducing the x-dependence in $\tilde{g}$ (roughly speaking, considering $\tilde{\tilde{g}} = \prod_{x \in \mathbf{R}^1} \otimes \tilde{g}$), we obtain from (15) the relation (3).

This approach leads to various integrable models if one considers suitable orbits in $\tilde{\tilde{g}}_+^*$ (or in $\tilde{\tilde{g}}_-^*$). Thus the HM model corresponds to $g = \mathrm{su}(2)$ and the simplest orbit consisting of the points $U(x, \lambda) = S(x)/\lambda$, $S^2(x) = \mathrm{const}\, I$. Here we have a relationship with the representation theory and the method of orbits.

If g has nontrivial automorphism σ, then the phase space can be reduced by considering quasiperiodic elements

$$\tilde{U}(x, \lambda) = \sum_{n=-\infty}^{\infty} A^n U(x, \lambda + n\omega) A^{-n}, \tag{16}$$

where A is the representation of σ. If the matrix Π commutes with $A \otimes A$, then for $\tilde{U}(x, \lambda)$ the relation (3) holds, with the r-matrix

$$r(\lambda) = \sum_{n=-\infty}^{\infty} \frac{(A^n \otimes I)\Pi(A^{-n} \otimes I)}{\lambda + n\omega}. \tag{17}$$

Thus, in particular, the r-matrix of the SG model is obtained from the r-matrix of the HM model. For the Lie algebras of A_n type a second averaging is possible. Thus one can obtain r-matrices expressed in terms of elliptic functions.

5. The lattice case. Classical models on a lattice play an intermediate role in quantizing classical continuous models ("quantization of the auxiliary space"). The matrix $U(x, \lambda)$ is replaced by the transition matrix $L_n(\lambda)$ from the nth lattice site to the $(n+1)$th. In the continuous limit $L_n(\lambda) = I + \Delta U(x, \lambda) + O(\Delta^2)$, where Δ is the lattice spacing. The monodromy matrix is given by the ordered product

$$T_N(\lambda) = \overleftarrow{\prod_{n=1}^{N}} L_n(\lambda) = L_N(\lambda) \dots L_1(\lambda), \tag{18}$$

where N is the number of sites of the lattice. Since $L_n(\lambda)$ is the transition matrix for one lattice site, formula (3) is uniquely transferred to the lattice case as follows:

$$\{L_n(\lambda) \otimes L_m(\mu)\} = [r(\lambda - \mu), L_n(\lambda) \otimes L_m(\mu)]\,\delta_{nm} \tag{19}$$

and in the continuous limit (19) goes back into (3).

All continuous models have their lattice variants with the same r-matrices. The averaging procedure also works for the lattice case. In contrast with the continuous case, in the lattice case the matrix $\tilde{L}_n(\lambda)$ is represented as an ordered product ([11]) and not as in (16). Here we have a connection with the theory of analytic matrix-valued functions.

As a result of the study of equation (19), in [26] there were introduced quadratic Poisson brackets algebras. These brackets are nontrivial deformations of the Lie-algebraic Poisson brackets. Here we have an interesting example of a deformation of algebraic structures.

II. Quantum theory

We begin with lattice theories, which necessarily arise in quantizing compact models, i.e. models with Poisson brackets algebras associated with compact Lie algebras (e.g. the HM model). Moreover, introduction of the lattice plays the role of ultraviolet regularization of continuous models.

1. The fundamental relation with the quantum R-matrix. Instead of the classical fields $\psi^a(x)$ we consider the field operators Ψ_n^a, which act irreducibly on the Hilbert space $\mathfrak{h}_n$, the space of quantum states at the nth lattice site. The complete Hilbert space of the model on a lattice with N sites is $\mathfrak{H}_N = \prod_{n=1}^{N} \otimes \mathfrak{h}_n$. In quantization it is natural to replace $L_n(\lambda)$ by the matrix-operator $\boldsymbol{L}_n(\lambda)$ which is a matrix in $\boldsymbol{C}^m$ with matrix elements belonging to the ring generated by Ψ_n^a and depending on the spectral parameter λ.

The problem of the right generalization of our main relation (19) to the quantum case is far from being trivial. The study of concrete models ([25], [28], [33]) suggests the following generalization of (19), explicitly introduced in [28], [33]:

$$R(\lambda-\mu)\big(\boldsymbol{L}_n(\lambda)\otimes\boldsymbol{L}_n(\mu)\big) = \big(\boldsymbol{L}_n(\mu)\otimes\boldsymbol{L}_n(\lambda)\big)R(\lambda-\mu). \tag{20}$$

Here the tensor product refers only to the auxiliary space $\boldsymbol{C}^m$ and $R(\lambda)$ is a matrix in $\boldsymbol{C}^m\otimes\boldsymbol{C}^m$ called the *quantum R-matrix*. If $R(\lambda) = P\big(I - -i\hbar r(\lambda)\big) + O(\hbar^2)$ as $\hbar\to 0$, where $\hbar$ is the Planck constant, then, in the quasiclassical limit relation, (20) goes back into (19). The uniqueness of such a deformation of the Poisson brackets as that given by (19) is an open question.

As in the classical case, relation (20) leads to a generalization of standard mathematical objects. In [26] it was used to introduce some nontrivial deformations of the universal enveloping algebras of Lie algebras, the so-called *Sklyanin quadratic algebras*.

The quantum monodromy matrix $\boldsymbol{T}_N(\lambda)$ is introduced by formula (18), where matrices $L_n(\lambda)$ are replaced by $\boldsymbol{L}_n(\lambda)$. A remarkable property of the quantum R-matrix is that it gives a compact form of all commutation relations of the matrix elements of $\boldsymbol{T}_N(\lambda)$. The following simple result holds.

THEOREM 4 (a quantum version of Theorem 1). *It follows from the relation* (20) *that*

$$R(\lambda-\mu)\big(\boldsymbol{T}_N(\lambda)\otimes\boldsymbol{T}_N(\mu)\big) = \big(\boldsymbol{T}_N(\mu)\otimes\boldsymbol{T}_N(\lambda)\big)\,R(\lambda-\mu). \tag{21}$$

In particular, one has

$$[\operatorname{tr}\boldsymbol{T}_N(\lambda), \operatorname{tr}\boldsymbol{T}_N(\mu)] = 0, \tag{22}$$

where tr *denotes the trace in the auxiliary space* $\boldsymbol{C}^m$.

Thus the operators $\operatorname{tr}\boldsymbol{T}_N(\lambda)$ form a commutative family, which is the family of quantum integrals of motion.

From the associative property of the tensor product one obtains a sufficient condition for admissible R-matrices

$$\big(I\otimes R(\lambda-\mu)\big)\big(R(\lambda)\otimes I\big)\big(I\otimes R(\mu)\big) = \big(R(\mu)\otimes I\big)\big(I\otimes R(\lambda)\big)\big(R(\lambda-\mu)\otimes I\big). \tag{23}$$

This relation is called the *Yang–Baxter equation* (or the *triangle equation* and also the *factorization equation*). It occurs in statistical mechanics [1] as the commuting condition for transfer-matrices and in the scattering theory [39] as the factorization condition for S-matrices (see also [33], [22]). Lately this equation has become rather popular; its solutions and methods for constructing them can be found in [4], [21], [22]. This equation also has connections with algebraic geometry [19].

We shall now consider the case of $m = 2$ more thoroughly. The simplest solution of equation (23) has the form

$$R(\lambda) \approx \frac{\lambda P+\eta I}{\lambda+\eta} \tag{24}$$

and corresponds to the models HM and NS. Here η is a parameter (the coupling constant). Other solutions can be obtained by a quantum analogue of the averaging procedure. Namely, we put $\mathscr{R}(\lambda) = R(\lambda)P$ and

$$\tilde{\mathscr{R}}(\lambda) = \prod_{n=-\infty}^{\infty\curvearrowleft} (\sigma_3^n\otimes I)\mathscr{R}(\lambda+n\omega_1)(\sigma_3^{-n}\otimes I), \tag{25}$$

$$\tilde{\tilde{\mathscr{R}}}(\lambda) = \prod_{n=-\infty}^{\infty\curvearrowleft} (\sigma_1^n\otimes I)\tilde{\mathscr{R}}(\lambda+n\omega_2)(\sigma_1^{-n}\otimes I) \tag{26}$$

for $\mathrm{Im}(\omega_2/\omega_1) > 0$. Then the matrices $\tilde{R}(\lambda) = \tilde{\mathscr{R}}(\lambda)P$ and $\tilde{\tilde{R}}(\lambda) = \tilde{\tilde{\mathscr{R}}}(\lambda)P$ satisfy the Yang–Baxter equation and correspond to the SG model and the XYZ-Heisenberg model, respectively (see below).

2. The local vacuum and the algebraic Bethe Ansatz. In addition to the matrix-operator $L_n(\lambda)$ and the quantum R-matrix, a local vacuum is another important object of the quantum inverse scattering method. This is a vector $\omega_n \in \mathfrak{h}_n$ characterized by the property

$$L_n(\lambda)\omega_n = \begin{bmatrix} a(\lambda) & * \\ 0 & b(\lambda) \end{bmatrix} \omega_n, \tag{27}$$

where $a(\lambda)$ and $b(\lambda)$ are some functions of λ. The reference state $\Omega_N = \prod_{n=1}^{N} \otimes \omega_n$ has an analogous property with respect to the monodromy matrix. Using the existence of the R-matrix and local vacuum it is possible to give a general procedure for diagonalizing the operators $\mathrm{tr}\, T_N(\lambda)$ ([28], [33]), which is the algebraic background of the method. Namely, the following statement holds.

THEOREM 5 (the algebraic Bethe Ansatz). *Let*

$$T_N(\lambda) = \begin{bmatrix} A_N(\lambda) & B_N(\lambda) \\ C_N(\lambda) & D_N(\lambda) \end{bmatrix} \tag{28}$$

and suppose that there exist a local vacuum and a reference state. Moreover, suppose that in the basis in $\mathbf{C}^2 \otimes \mathbf{C}^2$, associated with the basis in $\mathbf{C}^2$ where (27) *holds, the R-matrix has the form*

$$R(\lambda) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & b(\lambda) & c(\lambda) & 0 \\ 0 & c(\lambda) & b(\lambda) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \tag{29}$$

where $b(\lambda)/c(\lambda)$ is an odd function of λ. Then the vectors $\Psi_N(\lambda_1, \ldots, \lambda_l) = B_N(\lambda_1) \ldots B_N(\lambda_l)\Omega_N$ (Bethe vectors) are the eigenvectors for the operators $\mathrm{tr}\, T_N(\lambda) = A_N(\lambda) + D_N(\lambda)$ with the eigenvalues

$$a^N(\lambda) \prod_{j=1}^{l} \frac{1}{c(\lambda_j - \lambda)} + \delta^N(\lambda) \prod_{j=1}^{l} \frac{1}{c(\lambda - \lambda_j)} \tag{30}$$

if the parameters $\lambda_1, \ldots, \lambda_l$ *satisfy the system of equations*

$$\frac{a^N(\lambda_j)}{\delta^N(\lambda_j)} = \prod_{\substack{k=1 \\ k\neq j}}^{l} \frac{c(\lambda_k - \lambda_j)}{c(\lambda_j - \lambda_k)}, \quad j = 1, \ldots, l. \tag{31}$$

This theorem can also be generalized to the case $m > 2$.

As we have pointed out before, $\operatorname{tr} \boldsymbol{T}_N(\lambda)$ form a family of commuting quantum integrals of motion. It contains also the Hamiltonian of the model. The simplest expression for the Hamiltonian occurs in the case of fundamental models, where the quantum space is isomorphic to the auxiliary space. In this case there exists a point $\lambda = \lambda_0$ for which the operator $\operatorname{tr} \boldsymbol{T}_N(\lambda)$ is proportional to the cyclic shift operator in $\mathfrak{H}_N$ (the quasimomentum operator). The Hamiltonians with interaction of $k+1$ nearest neighbors on the lattice are expressed in terms of the operators

$$\boldsymbol{H}_k = \frac{d^k}{d\lambda^k} \log \operatorname{tr} \boldsymbol{T}_N(\lambda)|_{\lambda=\lambda_0}, \quad k = 1, 2, \ldots, N-1.$$

In addition, $\delta(\lambda_0) = 0$, and so the spectrum of $\boldsymbol{H}_k$ is additive. For nonfundamental models the construction of local Hamiltonians requires additional tools ([8], [20], [36]).

3. Characteristic examples. 1. The isotropic HM of spins s, $2s \in \boldsymbol{Z}$ (XXX-model; see [8], [9], [20], [21], [23], [32], [34]). Here $\mathfrak{h}_n = \boldsymbol{C}^{2s+1}$ and the matrix $\boldsymbol{L}_n(\lambda)$ looks like

$$\boldsymbol{L}_n(\lambda) = \begin{bmatrix} \lambda I_n + iS_n^3 & iS_n^- \\ iS_n^+ & \lambda I_n - iS_n^3 \end{bmatrix}, \tag{31}$$

where $\frac{1}{i} S_n^a$, $a = 1, 2, 3$ are generators of an irreducible representation of $\mathrm{su}(2)$ in $\mathfrak{h}_n$ and $S_n^{\pm} = S_n^1 \pm iS_n^2$. The R-matrix has the form (24), where $\eta = i$. The Bethe vectors are the highest weights [34] with respect to the action of $\mathrm{su}(2)$ in $\mathfrak{H}_N$ and the system of multiplets associated with them is complete [16], [32]. In proving this there arise nontrivial combinatoric identities [16]. The Hamiltonian has the form ([20], [21])

$$\boldsymbol{H}_N^{(XXX)} = \varepsilon \sum_{n=1}^{N} f_s(\langle \vec{S}_n, \vec{S}_{n+1} \rangle), \tag{32}$$

where $\vec{S}_{N+1} = \vec{S}_1$ and f_s is a polynomial of degree $2s$, characterized by the conditions $f_s\big(l(l+1)/2 - s(s+1)\big) = \sum_{k=l+1}^{2s} (1/k)$, $l = 0, 1, \ldots, 2s$. In the quasiclassical limit $\boldsymbol{H}_N^{(XXX)}$ goes into the Hamiltonian of the lattice HM model.

2. The lattice NS model ([13], [15]). Here $\mathfrak{h}_n = \mathscr{L}_2(\boldsymbol{R}^1)$ and the matrix $\boldsymbol{L}_n(\lambda)$ is obtained from (31) by a left multiplication by σ_3. The operators S_n^a are now the generators of the irreducible infinite-dimensional representation of su(2) of spin $s = -2/\varkappa\Delta$, where $\varkappa$ is the coupling constant and Δ is the lattice spacing. The operators S_n^a are expressed in a standard way in terms of the usual creation-annihilation operators in $\mathfrak{h}_n$. The R-matrix has the form (24), where $\eta = -i\varkappa$ and the Hamiltonian is given by (32), where $\langle \vec{S}_n, \vec{S}_{n+1} \rangle$ should be replaced by $\langle \sigma\vec{S}_n, \vec{S}_{n+1} \rangle$, σ being an involution of su(2). The function f_s naturally interpolates the polynomial from (32) to the case of nonintegral $2s$ ([36]).

3. The lattice SG model ([14], [15], [28]). Here the field operators are unitary operators u_n and v_n satisfying Weyl's commutation relation $u_n v_n = \exp(i\gamma) v_n u_n$, where $\gamma = \beta^2/8$, β being the coupling constant; $\mathfrak{h}_n = \mathscr{L}_2(\boldsymbol{R}^1/2\pi\boldsymbol{Z})$ if $\gamma \neq 2\pi p/q$ and $\mathfrak{h}_n = \boldsymbol{C}^q$ otherwise. The matrix $\boldsymbol{L}_n(\lambda)$ has the form

$$\boldsymbol{L}_n(\lambda) = \begin{bmatrix} f(v_n)u_n^{-1} & g(v_n, -\lambda) \\ g(v_n, \lambda) & u_n f(v_n) \end{bmatrix}, \tag{33}$$

where

$$f(v) = \left(1 + \frac{m^2\Delta^2}{16}\,(v^2 e^{-i\gamma} + v^{-2} e^{i\gamma})\right)^{1/2},$$

$$g(v, \lambda) = \frac{m\Delta}{4}\,(e^{-\lambda} v^{-1} - e^{\lambda} v)$$

and m plays the role of mass. The R-matrix has the form (29), where

$$b(\lambda) = \frac{i \sin\gamma}{\operatorname{sh}(\lambda + i\gamma)}, \qquad c(\lambda) = \frac{\operatorname{sh}\lambda}{\operatorname{sh}(\lambda + i\gamma)}.$$

In contrast to the previous examples, a local vacuum exists only for the product $\boldsymbol{L}_{n+1}(\lambda)\boldsymbol{L}_n(\lambda)$ and not for individual matrix $\boldsymbol{L}_n(\lambda)$. The Hamil-

tonian has a more complicated form than (32), but in the continuous limit goes into the regularized version of the Hamiltonian of the SG model

$$H_L^{(SG)} = \int_{-L}^{L} \left(\frac{1}{2} \pi^2 + \frac{1}{2} \left(\frac{\partial \varphi}{\partial x} \right)^2 + \frac{m^2}{\beta^2} (1 - \cos \beta \varphi) \right) dx, \tag{34}$$

where $[\varphi(x), \pi(y)] = i\delta(x-y)$.

4. The anisotropic HM model of spins $\frac{1}{2}$ (XYZ model; see [2], [3], [31], [33]). This model is naturally related to the eight-vertex model of the classical statistical mechanics on the two-dimensional lattice [1], [33]. Its Hamiltonian is equal to

$$H_N^{(XYZ)} = \sum_{n=1}^{N} (J_x \sigma_n^1 \sigma_{n+1}^1 + J_y \sigma_n^2 \sigma_{n+1}^2 + J_z \sigma_n^3 \sigma_{n+1}^3), \tag{35}$$

where the periodic boundary conditions are assumed. The matrices $R(\lambda)$ and $L_n(\lambda)$ for these models have a more complicated form than (29) and (31) and are expressed in terms of elliptic functions (see [1], [31], [33]). The diagonalization procedure for $\operatorname{tr} T_N(\lambda)$ requires more complicated technical tools, but expressions for the eigenvalues of $\operatorname{tr} T_N(\lambda)$ and the system of equations (30) in Theorem 5 retain their algebraic form ([3], [31], [33]).

4. Thermodynamic limit. Here we shall briefly consider the thermodynamic limit which is the limit as $N \to \infty$ for compact models and as $L \to \infty$, $\Delta \to 0$ for noncompact ones. The behavior of models in this limit is most interesting from the physical point of view. The main problem here is to define the ground state — the eigenvector of the Hamiltonian with the minimal eigenvalue (the lowest energy vector) and to describe the Hilbert space of states near it — the space of low-lying excitations. There are two possibilities.

1) *The ferromagnetic case*: the reference state is the ground state (this occurs for the NS model and the HM model in the case $\varepsilon > 0$). Then in the Fock space $\mathfrak{H}_F \subset \mathfrak{H}_\infty$ adjoining to the vector $\Omega = \prod_{n=1}^{\infty} \otimes \omega_n$ for $\lambda \in \mathbf{R}^1$ in a weak sense there exist limits $A(\lambda) = \lim_{N \to \infty} a^{-N}(\lambda) A_N(\lambda)$, $B(\lambda)$

$= \lim_{N\to\infty} B_N(\lambda)$. These operators satisfy the commutation relations

$$[A(\lambda), A(\mu)] = 0, \quad [B(\lambda), B(\mu)] = 0,$$

$$A(\lambda)B(\mu) = \frac{1}{c(\mu-\lambda-i0)} B(\mu)A(\lambda); \tag{36}$$

$$A(\lambda)\Omega = \Omega.$$

With the help of these formulae the spectrum and eigenvectors of the operators $A(\lambda)$ are easily obtained ([25], [27]). The commuting family $\log A(\lambda)$ has an additive spectrum and contains the limiting Hamiltonian of the model $\boldsymbol{H}_F = \lim_{N\to\infty} (\boldsymbol{H}_N - E_0(N)\boldsymbol{I}_N)$. Here $E_0(N)$ is the ground state energy and $\boldsymbol{I}_N$ is the unit operator in $\mathfrak{H}_N$.

2) *The antiferromagnetic case*: to the ground state there corresponds a special distribution of $\lambda_1, \ldots, \lambda_l$ in the Bethe vector (this occurs for the SG model and the HM model in the case $\varepsilon < 0$). As $N\to\infty$, parameters $\lambda_1, \ldots, \lambda_l$ become uniformly distributed on the real axis with the density $\varrho(\lambda)$. This is "the filled Dirac Sea". (This situation occurs in the quantum field theory and in solid state physics and is called "filling of the vacuum".) The function $\varrho(\lambda)$ satisfies a linear integral equation which follows from the system (30)

$$2\pi\varrho(\lambda) + \int_{-\infty}^{\infty} \Phi(\lambda-\mu)\varrho(\mu)d\mu = f(\lambda), \tag{37}$$

where

$$\Phi(\lambda) = \frac{1}{i}\frac{d}{d\lambda}\log\frac{c(\lambda)}{c(-\lambda)}, \quad f(\lambda) = \frac{1}{i}\frac{d}{d\lambda}\log\frac{\delta(\lambda)}{a(\lambda)}.$$

Excitations above this ground state are characterized by the density $\varrho(\lambda, \lambda_1, \ldots, \lambda_n) = \varrho(\lambda) + \frac{1}{N}\cdot\sum_{j=1}^{n}\sigma(\lambda-\lambda_j)$, where $\sigma(\lambda-\mu)$ is the resolvent kernel of equation (37), and parameters $\lambda_1, \ldots, \lambda_n$ appear as the holes in the Dirac Sea.

In these terms the ground state and the excitation creation operator, respectively, have the form

$$\Omega_{\text{ground}} = \lim_{N\to\infty} \exp\left\{N\int_{-\infty}^{\infty} \log B_N(\lambda)\varrho(\lambda)d\lambda\right\}\Omega \tag{38}$$

and

$$\tilde{B}(\lambda) = \lim_{N\to\infty} \exp\left\{\int_{-\infty}^{\infty} \log B_N(\mu)\sigma(\lambda-\mu)\,d\mu\right\}. \tag{39}$$

The operator $\tilde{A}(\lambda)$ is defined as in the previous case and $\tilde{A}(\lambda)$ and $\tilde{B}(\lambda)$ satisfy the commutation relations of the type (36) with $c(\lambda)$ replaced by

$$\tilde{c}(\lambda) = \exp\left\{\int_{-\infty}^{\infty} \log c(\mu)\sigma(\lambda-\mu)\,d\mu\right\}. \tag{40}$$

As before, the commuting family $\log\tilde{A}(\lambda)$ has an additive spectrum and contains the renormalized Hamiltonian of the model.

Using this method, it is also possible to describe the scattering of excitations and to calculate the S-matrices. The in- and out-states are constructed with the help of operators $Z(\lambda) = B(\lambda)A^{-1}(\lambda)$ $(\tilde{Z}(\lambda) = \tilde{B}(\lambda)\tilde{A}^{-1}(\lambda))$ satisfying simple commutation relations which follows from (36) (see [34]).

In papers [15], [17], [25], [28], [32]–[34] this general scheme was applied to the detailed investigation of concrete models. In particular, in [28] it was used to obtain an exact non-pertubative solution of the SG model. With proper modifications this approach can be generalized to the case of models with the auxiliary space C^m, $m > 2$.

Finally, let us mention the latest achievements of the quantum inverse scattering method: (a) The exact calculation of norms of Bethe vectors [18], which gives us a hope to obtain explicit expressions for Green's functions; (b) The quantum variant of the equations of inverse problem [29], which provides an expression of Heisenberg field-operators Ψ_n^a in terms of the operators $A(\lambda)$ and $B(\lambda)$; (c) A new approach towards the integrability of the quantum $O(3)$ nonlinear σ-model [10].

References

[1] Baxter R. J., Partition function of the eight-vertex lattice model, *Ann. Phys.* **70**, No. 1 (1972), pp. 193–228.

[2] Baxter R. J., One-dimensional anisotropic Heisenberg chain, *Ann. Phys.* **70**, No. 2 (1972), pp. 323–337.

[3] Baxter R. J., Eight-vertex model in lattice statistics and one-dimensional anisotropic Heisenberg chain, I–III, Ann. Phys. **76**, No. 1 (1973), pp. 1–71.

[4] Belavin A. A., Discrete groups and integrability of quantum systems, *Funk. anal. i pril.* **14**, No. 4 (1980), pp. 18-26.

[5] Belavin A. A. and Drinfeld V. G., On the solutions of classical Yang–Baxter equation for simple Lie algebras, *Funk. anal. i pril.* **16**, No. 3 (1982), pp. 1–29.

[6] Drinfeld V. G., Hamiltonian structures on Lie algebras, Lie bialgebras and geometrical meaning of Yang–Baxter equations, *DAN SSSR* **268**, No. 2 (1983), pp. 285–287.
[7] Faddeev L. D., Quantum completely integrable models in field theory, *Sov. Sci. Rev.* **C1** (1980), pp. 107–155, ser. *Contem. Math. Phys.*
[8] Faddeev L. D., Recent developments of QST, *RIMS, Kopyuproku* **469** (1982), pp. 53–71.
[9] Faddeev L. D., *Integrable models in* $1+1$ *dimensional quantum field theory*, Saclay preprint S. Ph. T/82/76, 1982.
[10] Faddeev L. D. and Takhtajan L. A., *Integrability of the quantum* $O(3)$ *nonlinear* σ*-model*, LOMI preprint E-4-83, Leningrad, 1983.
[11] Faddeev L. D., and Reshetikhin N. Yu,. Hamiltonian structures for the integrable models of field theory, *Teor. Matem. Fiz.* **56**, No. 3 (1983), pp. 323–343.
[12] Faddeev L. D. and Takhtajan L. A., *Hamiltonian approach to solitons theory*, Springer-Verlag (to be published).
[13] Isergin A. G. and Korepin V. E., The lattice model, associated with the nonlinear Schrödinger equation, *DAN SSSR*, **259**, No. 1 (1981), pp. 76–79.
[14] Isergin A. G. and Korepin V. E., The lattice quantum sine-Gordon equation, *Lett. Math. Phys.* **5** (1981), pp. 199–205.
[15] Isergin A. G. and Korepin V. E., The quantum inverse problem method, *Fiz. Elem. Chastitz i Atom. Yadra* **13**, No. 3 (1982), pp. 501–541.
[16] Kirillov A. N., Combinatoric identities and the completeness theorem for the Heisenberg magnet, *Seminar of the Steklov Math. Inst. at Leningrad* **131** (1982), pp. 88–105.
[17] Korepin V. E., Exact calculation of the S-matrix in the massive Thirring model, *Teor. Matem. Fiz.* **41**, No. 2 (1979), pp. 169–189 (English translation: *Theor. Math. Phys.* **41** (1979), p. 953).
[18] Korepin V. E., Calculation of norms of Bethe wave functions, *Commun. Math. Phys.* **86** (1982), pp. 391–418.
[19] Krichever I. M., Yang–Baxter equation and algebraic geometry, *Funk. Anal. i pril.* **15**, No. 2 (1981), pp. 22–35.
[20] Kulish P. P., The quantum inverse problem method and exactly solvable models in statistical mechanics, *JINR, Dubna* **17-81-758** (1981), pp. 147–157.
[21] Kulish P. P., Reshetikhin N. Yu, and Sklyanin E. K., Yang–Baxter equation and representation theory. I, *Lett. Math. Phys.* **5**, No. 5 (1981), pp. 393–403.
[22] Kulish P. P. and Sklyanin E. K., On the solutions of the Yang–Baxter equation, *Seminar of the Steklov Math. Inst. at Leningrad* **95** (1980), pp. 129–160.
[23] Kulish P. P. and Sklyanin E. K., Quantum Spectral Transform Method. Recent Developments, *Lect. Notes in Phys.* **151** (1982), pp. 61–119.
[24] Sklyanin E. K., *On Complete Integrability of the Landau–Lifschitz equation*, LOMI preprint E-3-79, Leningrad, 1979.
[25] Sklyanin E. K., The quantum version of the inverse scattering method, *Seminar of the Steklov Math. Inst. at Leningrad* **95** (1980), pp. 55–128.
[26] Sklyanin E. K., On some algebraic structures, associated with Yang–Baxter equation, *Funk. anal. i pril.* **16**, No. 4 (1982), pp. 27–34.
[27] Sklyanin E. K. and Faddeev L. D., The quantum-mechanical approach to completely integrable models of field theory, *DAN SSSR* **243** No 6 (1978), pp. 1430–1433 (English translation: *Sov. Phys. Dokl.* **23** (1978), p. 902).
[28] Sklyanin E. K., Takhtajan L. A., and Faddeev L. D., The quantum inverse

problem method I, *Teor. Matem. Fiz.* **40**, No. 2 (1979), pp. 194–220 (English translation: *Theor. Math. Phys.* **40** (1980), p. 688).

[29] Smirnov F. A., The Gelfand–Levitan equations for the quantum nonlinear Schrödinger equation in the attractive case, *DAN SSSR* **262** No. 1 (1982), pp. 78–83.

[30] Takhtajan L. A., Integration of the continuous Heisenberg spin chain through the inverse scattering method, *Phys. Lett.* **64A**, No. 2 (1977), pp. 235–237.

[31] Takhtajan L. A., The quantum inverse problem method and the *XYZ* Heisenberg model, *Physica D*, **3**, No. 1–2 (1981), pp. 231–245.

[32] Takhtajan L. A., The picture of low-lying excitations in isotropic Heisenberg chain of arbitrary spins, *Phys. Lett.* **87A**, No. 9 (1982), pp. 479–482.

[33] Takhtajan L. A. and Faddeev L. D., The quantum inverse problem method and the *XYZ* Heisenberg model, *Uspekhi Mat. Nauk* **34**, No. 5 (1979), pp. 13–63 (English translation: *Russian Math. Surveys* **34** (1979), p. 11).

[34] Takhtajan L. A. and Faddeev L. D., Spectrum and scattering of excitations in the one-dimensional isotropic Heisenberg model, *Seminar of the Steklov Math. Inst. at Leningrad* **101** (1981), pp. 134–178.

[35] Takhtajan L. A. and Faddeev L. D., A simple connection between geometrical and Hamiltonian representations for the integrable nonlinear equations, *Seminar of the Steklov Math. Inst. at Leningrad* **115** (1982), pp. 264–273.

[36] Tarasov V. O., Takhtajan L. A. and Faddeev L. D., Local Hamiltonians for quantum integrable models on a lattice, *Teor. Matem. Fiz.* **57**, No. 2 (1983), pp. 163–181.

[37] Zakharov V. E. and Faddeev L. D., The Korteweg–de Vries equation — a completely integrable Hamiltonian system, *Funk. anal. i pril.* **5**, No. 4 (1971), pp. 18–27 (English translation: *Func. Anal. Pril.* **5** (1971), p. 28).

[38] Zakharov V. E., Manakov S. V., Novikov S. P., and Pitaevsky L. P., *Solitons theory: Inverse problem method*, Moscow, 1980.

[39] Zamolodchikov A. B. and Zamolodchikov Al. B., Factorized S-matrices in two dimensions as the exact solutions of certain relativistic quantum field theory models, *Ann. Phys.* **120**, No. 2 (1979), pp. 253–291.

LENINGRAD BRANCH OF THE V. A. STEKLOV
MATHEMATICAL INSTITUTE OF THE ACADEMY OF SCIENCES
USSR

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

S. L. WORONOWICZ*

Duality in the C^*-Algebra Theory

In Mathematics, the word "representation" has many meanings. One speaks about representations of groups, algebras, commutation relations, Lie algebras, etc. Usually the things that are being represented are objects of a certain category and representations are simply category morphisms into special objects whose structure is considered to be well understood.

For example, in the case of topological groups one considers the group $\mathrm{Aut}H$ of all unitary operators acting on a given Hilbert space H. $\mathrm{Aut}H$ is endowed with the strong operator topology. A representation (or, more precisely, a unitary representation) of a topological group G is then a continuous group homomorphism $\pi\colon G\to\mathrm{Aut}H$. We say that H is the *carrier Hilbert space of* π or in other words that π acts on H.

In many cases (especially in the field of interest of quantum physics) like in the example above, the special objects that serve as targets for representations are related to Hilbert spaces. For these representations we can speak about the carrier Hilbert spaces. Also the other important notions known in the group representation theory such as unitary equivalence, subrepresentation, direct sum and direct integral of representations and intertwining operators are meaningful in this general case.

Let us notice that the notions of unitary equivalence and subrepresentations can be defined in terms of intertwining operators. Indeed, representations π and π' are unitarily equivalent iff there exists a unitary operator intertwining π and π'; a representation π' acting on $H'\subset H$ is a subrepresentation of a representation π acting on H iff the orthogonal projection $E\colon H\to H'$ intertwines π and π'; $\pi=\pi_1\oplus\pi_2$ (where π_1, π_2,

* Département de Physique Mathématique, Université de Provence. On leave from the Department of Mathematical Methods in Physics, University of Warsaw, Hoża 74, 00-682 Warszawa, Poland.

π are acting on H_1, H_2, $H_1 \oplus H_2$ respectively) iff the canonical projection p_i: $H_1 \oplus H_2 \to H_i$ intertwines π and π_i for $i = 1,2$.

On the other hand, it is easy to list the characteristic properties of intertwining operators (cf. Def. 1 in the sequel). This way we arrive to the concept of W^*-category [2]. Our definition of a W^*-category differs from the one used in [2] in four important points.

At first the sets of morphisms considered in [2] are abstract dual Banach spaces whereas in our approach they are weakly closed subspaces of $B(H)$ and the composition of morphisms coincides with the product of operators. One may say that W^*-categories considered in this paper have concrete Hilbert space realization.

Secondly in order to avoid the well known set theoretical problems (the use of classes not being sets) we restrict ourselves to a fixed Hilbert space.

At third we use additional axioms saying that W^*-categories are closed with respect to the direct sum and taking sub-object operations.

At last we assume that the set of objects is endowed with a topological structure. We emphasize this point by saying that we deal with topological W^*-categories. Topology is necessary to introduce the direct integral operation (we shall not discuss this point in detail).

For any C^*-algebra A, the set $\mathrm{Rep}_H A$ of all nondegenerate representations of A acting on the Hilbert space H becomes a topological W^*-category in a natural way. It turns out that $\mathrm{Rep}_H A$ can be considered as the object dual to A. Indeed, it is possible to reconstruct A if the topological W^*-category $\mathrm{Rep}_H A$ is given. At this point a notion of an operator function defined on a W^*-category is very important. The operator functions were introduced in [9] and [4], see also [3] where they are called decomposable functions and [2]. W^*-categories serve as natural domains for operator functions. We show that the C^*-algebra A is canonically isomorphic with the algebra of all "vanishing at infinity" continuous operator functions defined on $\mathrm{Rep}_H A$.

At the end of the paper we introduce an important property of topological W^*-categories called "local compactness" which is characteristic for topological W^*-categories of the form $\mathrm{Rep}_H A$, where A is a C^*-algebra. The algebra A has unity if and only if the considered topological W^*-category is "compact".

In principle this property can be used to determine in which cases the structure of a set of representations of an object (e.g. a group, an algebra, a commutation relation) is isomorphic to $\mathrm{Rep}_H A$. If this is the case, the C^*-algebra A is uniquely determined.

DEFINITION 1. Let H be a Hilbert space and let R be a set. Assume that for any pair (r, r') of elements of R a weakly closed linear subspace $\mathrm{Mor}(r, r')$ of $B(H)$ is given. We say that R is a *W^*-category* if the following axioms are satisfied:

I. $I \in \mathrm{Mor}(r, r)$ *for any* $r \in R$.

II. *For any* $r, r', r'' \in R$ *and* $a, b \in B(H)$

$$\begin{pmatrix} a \in \mathrm{Mor}(r, r') \\ b \in \mathrm{Mor}(r', r'') \end{pmatrix} \Rightarrow \big(ba \in \mathrm{Mor}(r, r'')\big).$$

III. *For any* $r, r' \in R$ *and* $a \in B(H)$

$$\big(a \in \mathrm{Mor}(r, r')\big) \Rightarrow \big(a^* \in \mathrm{Mor}(r', r)\big).$$

IV. *For any* $r, r' \in R$

$$\big(I \in \mathrm{Mor}(r, r')\big) \Rightarrow (r = r').$$

V. *For any* $r \in R$ *and any* $u \in B(H)$ *such that* $u^*u = I$, $uu^* \in \mathrm{Mor}(r, r)$ *there exists* $r' \in R$ *such that* $u \in \mathrm{Mor}(r', r)$.

VI. *For any family* $(r_\alpha)_{\alpha \in A}$ *of elements of* R *and any family of isometries* $(u_\alpha)_{\alpha \in A}$ *acting on* H *such that* $\sum_\alpha u_\alpha u_\alpha^* = I$ *there exists* $r' \in R$ *such that* $u_\alpha \in \mathrm{Mor}(r_\alpha, r)$, $\alpha \in A$.

To stress the connection between R and H we shall say that R is a *W^*-category acting on H*. One can easily see that the properties described in Axioms IV, V, VI can be expressed in the following single statement.

THEOREM 1. *Let R be a W^*-category acting on H, $(r_\alpha)_{\alpha \in A}$ a family of elements of R and $(u_\alpha)_{\alpha \in A}$ a family of operators acting on H indexed by the same nonempty set A. Assume that H is generated by the union of the images of u_α: $H = \bigvee_{\alpha \in A} u_\alpha H$ and that*

$$u_\alpha^* u_\beta \in \mathrm{Mor}(r_\beta, r_\alpha)$$

for any $\alpha, \beta \in A$. Then there exists one and only one element $r \in R$ such that

$$u_\alpha \in \mathrm{Mor}(r_\alpha, r)$$

for any $r \in R$.

The proof of this theorem is standard and will be omitted.

DEFINITION 2. Let R be a W^*-category acting on a Hilbert space H. A mapping

$$F\colon R \to B(H)$$

will be called an *operator function defined on* R if for any $r, r' \in R$ and any $a \in \mathrm{Mor}(r, r')$ we have

$$aF(r) = F(r')a.$$

The set of all operator functions defined on a W^*-category R will be denoted by $\mathscr{F}(R)$. For any $F, G \in \mathscr{F}(R)$, $\lambda \in C$ and $r \in R$ we set

$$(F+G)(r) \stackrel{\mathrm{df}}{=} F(r)+G(r),$$

$$(\lambda \cdot F)(r) \stackrel{\mathrm{df}}{=} \lambda F(v),$$

$$(F \cdot G)(r) \stackrel{\mathrm{df}}{=} F(r)G(r),$$

$$(F^*)(r) \stackrel{\mathrm{df}}{=} F(r)^*.$$

One can easily check that $F+G$, $\lambda \cdot F$, $F \cdot G$ and F^* are operator functions and that $\mathscr{F}(R)$ endowed with these algebraic operations becomes a $*$-algebra.

Assume now for the moment that for some $F \in \mathscr{F}(R)$ one can find a sequence (r_n) in R such that $\lim \|F(r_n)\| = \infty$. Let u_n be a sequence of isometries acting on H such that $\sum_n u_n u_n^* = I$. According to Def. 1, VI there exists $r \in R$ such that $u_n \in \mathrm{Mor}(r_n, r)$. Then we have $u_n F(r_n) = F(r) u_n$, $F(r_n) = u_n^* F(r) u_n$ and all the norms $\|F(r_n)\|$ are not larger than $\|F(r)\|$. The contradiction that we obtained in this way shows that for any $F \in \mathscr{F}(R)$

$$\|F\| \stackrel{\mathrm{df}}{=} \sup_{r \in R} \|F(r)\|$$

is finite. Clearly the above formula defines a norm on $\mathscr{F}(R)$ and $\mathscr{F}(R)$ becomes a normed $*$-algebra. The following theorem gives more informations about $\mathscr{F}(R)$.

THEOREM 2. *Let R be a W^*-category acting on H. Then $\mathscr{F}(R)$ is a W^*-algebra. Moreover $\mathscr{F}(R)$ is rich in the following sense: for any $r, r' \in R$*

$$\mathrm{Mor}(r, r') = \left\{ a \in B(H) : \begin{array}{l} aF(r) = F(r')a \\ \text{for all } F \in \mathscr{F}(R) \end{array} \right\}.$$

Proof. This theorem follows directly from the Murray von Neumann double commutant theorem. For details see [2].

As it was mentioned in the beginning the really interesting objects can be obtained by combining the W^*-category structure and the topologi-

cal structure. These objects are called topological W^*-categories. More precisely R is a *topological W^*-category* if R is a topological space and R is a W^*-category. For the moment we do not formulate any compatibility condition relating these two structures.

Let R be a topological W^*-category and $F \in \mathscr{F}(R)$. We say that F is a *continuous operator function* if F is a continuous map from the topological space R into $B(H)$ endowed with the *-strong topology. The set of all continuous operator functions defined on R will be denoted by $C(R)$:

$$C(R) \stackrel{\text{df}}{=} \left\{ F \in \mathscr{F}(R): \begin{array}{l} \text{for any } x \in H \text{ the mappings} \\ R \ni r \to F(r)x \in H \\ R \ni r \to F(r)^* x \in H \\ \text{are continuous} \end{array} \right\}.$$

It is well known that algebraic operations restricted to a bounded subset of $B(H)$ are continuous with respect to the $*$-strong topology. Having in mind that the limit of a uniformly converging sequence of continuous mappings is a continuous mapping we get the following result:

THEOREM 3. *Let R be a topological W^*-category. Then $C(R)$ is a C^*-algebra with unity.*

The unity is the operator function $\mathbf{1} \in \mathscr{F}(R)$ such that $\mathbf{1}(r) = I$ for any $r \in R$. Clearly $\mathbf{1} \in C(R)$.

Let R be a topological W^*-category and $r \in R$. It is easy to see that the map:

$$C(R) \ni F \to F(r) \in B(H) \qquad (*)$$

is a representation of $C(R)$. A representation π of $C(R)$ will be called *singular* if π is disjoint with all representations of the form $(*)$.

Now we can introduce the class of all continuous operator functions vanishing at infinity

$$C_\infty(R) \stackrel{\text{df}}{=} \left\{ F \in C(R): \begin{array}{l} \pi(F) = 0 \text{ for any singular} \\ \text{representation } \pi \text{ of } C(R) \end{array} \right\}.$$

For the completeness reasons we state the following obvious result:

THEOREM 4. *Let R be a topological W^*-category. Then $C_\infty(R)$ is a C^*-algebra (without unity, in general). $C_\infty(R)$ is a closed ideal in $C(R)$.*

In the general case, when the topology of R is not well compatible with its W^*-category structure, the algebras $C(R)$ and $C_\infty(R)$ may be very small (e.g.: $C(R) = \{\lambda \mathbf{1}: \lambda \in \mathbf{C}\}$, $C_\infty(R) = \{0\}$). Therefore we need some

axioms expressing the compatibility of the W^*-category structure with the topology. These axioms should be formulated in terms of open sets and elements of $\mathrm{Mor}(\cdot, \cdot)$. Unfortunately, at the present moment we are not able to formulate the compatibility condition in the way satisfying the above requirement. Instead we shall use the following two axioms:

VII. *The algebra $C(R)$ separates the points of R.*

VIII. *The topology of R is the weakest one such that for all $F \in C(R)$ and $x \in H$ the mappings*

$$R \ni r \to F(r)x \in H$$

are continuous.

More precisely, Axiom VII says that for any two distinct elements $r, r' \in R$ one can find $F \in C(R)$ such that $F(r) \neq F(r')$. Axiom VIII means that for any neighbourhood $\mathcal{O}$ of a point $r \in R$ one can find $F_1, F_2, \ldots, F_N \in C(R)$, $x_1, x_2, \ldots, x_N \in H$ and $\varepsilon > 0$ such that

$$\left\{ r' \in R: \begin{array}{l} \|(F_i(r') - F_i(r))x_i\| < \varepsilon \\ i = 1, 2, \ldots, N \end{array} \right\} \subset \mathcal{O}.$$

Clearly Axiom VII implies that the topological space R is Hausdorff.

Now we shall discuss the topological W^*-categories related to C^*-algebras. For the simplicity we shall assume that the algebras are separable (otherwise one has to consider nonseparable Hilbert spaces).

Let A be a C^*-algebra. We denote by $\mathrm{Rep}A$ (or more precisely by $\mathrm{Rep}_H A$) the set of all nondegenerate representations of A acting on the Hilbert space H. We recall that a representation π of A is said to be *non-degenerated* (*essential*) if 0 is the only vector $x \in H$ such that $\pi(a)x = 0$ for all $a \in A$.

For any $\pi, \pi' \in \mathrm{Rep}A$ we denote by $\mathrm{Mor}(\pi, \pi')$ the set of all intertwining operators:

$$\mathrm{Mor}(\pi, \pi') = \left\{ b \in B(H): \begin{array}{l} b\pi(a) = \pi'(a)b \\ \text{for any } a \in A \end{array} \right\}.$$

One can easily check that $\mathrm{Mor}(\pi, \pi')$ is a weakly closed linear subspace of $B(H)$ and that the Axioms I–VI are satisfied. It means that $\mathrm{Rep}A$ is a W^*-category.

It is interesting that in this case the algebra of all operator functions admits a very simple description. Indeed, using the Murray–von Neumann double commutant theorem (for details see [2]) one easily gets the following result:

THEOREM 5. *Let A be a C^*-algebra and H be a Hilbert space. Assume that A and H are separable (in fact it is sufficient to assume that the dimension of H is not smaller than the cardinality of a dense subset of A). Then the algebra of all operator functions $\mathscr{F}(\mathrm{Rep}_H A)$ is canonically isomorphic to the W^*-enveloping algebra A^{**} of A. An element $a \in A^{**}$ corresponds to an operator function $F \in \mathscr{F}(\mathrm{Rep}A)$ if and only if for any $\pi \in \mathrm{Rep}A$*

$$F(\pi) = \tilde{\pi}(a),$$

where $\tilde{\pi}$: $A^{**} \to B(H)$ denotes the weakly continuous extension of π.

The most interesting are operator functions corresponding to elements of A. We provide $\mathrm{Rep}A$ with the weakest topology such that all these operator functions are continuous. In this way $\mathrm{Rep}A$ becomes a topological W^*-category. A subset $\mathcal{O} \subset \mathrm{Rep}A$ is a neighbourhood of a representation $\pi \in \mathrm{Rep}A$ if an only if $\mathcal{O}$ contains a set of the form

$$\left\{\pi' \in \mathrm{Rep}A\colon \begin{matrix} \|(\pi'(a_i) - \pi(a_i))x_i\| < \varepsilon \\ i = 1, 2, \ldots, N \end{matrix}\right\},$$

where $a_1, a_2, \ldots, a_N \in A$, $x_1, x_2, \ldots, x_N \in H$ and $\varepsilon > 0$. This topology on $\mathrm{Rep}A$ has been considered by many authors (see e.g. [7]).

For the sake of completeness we state the following obvious result:

THEOREM 6. *Let A be a C^*-algebra H a Hilbert space. Then $\mathrm{Rep}_H A$ is a topological W^*-category satisfying the compatibility Axioms* VII and VIII.

It is interesting to see which subsets of A^{**} correspond to classes of operator functions with different continuity properties. The answer to this question is given in the following theorem, where we identify elements of A^{**} with the corresponding operator functions:

THEOREM 7. *Let A be a separable C^*-algebra and H be a separable Hilbert space. Then*

1. $$\left\{F \in \mathscr{F}(\mathrm{Rep}A)\colon \begin{matrix} \textit{for any } x, y \in H \textit{ the function} \\ \mathrm{Rep}A \ni \pi \to (x \mid F(\pi)y \in \boldsymbol{C} \\ \textit{is continuous} \end{matrix}\right\} = QM(A),$$

2. $$\left\{F \in \mathscr{F}(\mathrm{Rep}A)\colon \begin{matrix} \textit{for any } x \in H \textit{ the mapping} \\ \mathrm{Rep}A \ni \pi \to F(\pi)x \in H \\ \textit{is continuous} \end{matrix}\right\} = LM(A),$$

3. $C(\mathrm{Rep}A) = M(A)$,

4. $C_\infty(\mathrm{Rep}A) = A$,

where $QM(A)$, $LM(A)$, $M(A)$ *denote the multiplier algebras of the algebra* A (see [5] for the precise definitions).

The proof of this theorem is given in [10]. Here we would like to point out only that it is very easy to show that members of $QM(A)$, $LM(A)$, $M(A)$ and A have the postulated continuity properties: for A it follows directly from the definition of the topology on $\mathrm{Rep}A$; for $QM(A)$, $LM(A)$ and $M(A)$ one has to use the approximative unity in A and the classical result saying that the limit of the uniformly converging sequence of continuous mappings is a continuous mapping. On the other hand, the proof that the continuity properties imply the belonging to the multiplier algebras of A is quite difficult. In case 1, one has to use the Voiculescu result [8], in case 2, one passes to the algebra with unity and apply 1. Case 3 follows directly from case 2 and case 4 follows easily from case 3.

Let us notice that for the algebras with unity the essentially equivalent result is contained in [7].

Now we go back to the general topological W^*-categories. Let R be a topological W^*-category. We are interested, under what conditions one can find a C^*-algebra A such that R is isomorphic to $\mathrm{Rep}A$. According to Theorem 7 there is only one candidate for A: $A = C_\infty(R)$. Unfortunately, in the general case $C_\infty(R)$ may be very small (e.g. $C_\infty(R) = \{0\}$); Axiom VII says only that $C(R)$ is large enough. The same phenomena we have in the usual theory of topological spaces. Let X be a normal topological space. Then the algebra $C(X)$ of bounded continuous complex valued functions on X is rich enough: it separates the points of X. On the other hand

$$C_\infty(X) \overset{\mathrm{df}}{=} \left\{ F \in C(X) \colon \begin{array}{c} \text{For any } \varepsilon > 0 \\ \{z \in X \colon |F(z)| \geqslant \varepsilon\} \\ \text{is compact} \end{array} \right\}$$

is in general very small. Indeed, only if X is locally compact, the algebra $C_\infty(X)$ contains functions nonvanishing at any given point of X.

Therefore in our case we also need a condition saying that our topological W^*-category is in some sense "locally compact". To express this condition we have to introduce the following notation.

Let $\mathrm{Isom}(H)$ denote the set of all isometric operators acting on H equipped with the strong operator topology and let $\mathcal{N}$ denote the filter

of all neighbourhoods of I in $\mathrm{Isom}(H)$. For any $\mathcal{O} \subset R$ and any $V \in \mathcal{N}$ we set

$$\mathcal{O}^V \stackrel{\mathrm{df}}{=} \left\{ r \in R: \begin{array}{l} \text{there exist } r' \in \mathcal{O},\ b \in V \\ \text{such that } b \in \mathrm{Mor}(r, r') \end{array} \right\}.$$

DEFINITION 3. Let R be a topological W^*-category acting on H, $\mathcal{O} \subset R$, $V \in \mathcal{N}$. We say that $\mathcal{O}$ is *V-precompact* if for any open covering

$$R = \bigcup_{\lambda \in \Lambda} \mathcal{O}_\lambda$$

one can find a finite subset $\Lambda_0 \subset \Lambda$ such that

$$\mathcal{O} \subset \bigcup_{\lambda \in \Lambda_0} \mathcal{O}_\lambda^V.$$

DEFINITION 4. A topological W^*-category R is called *"compact"* if R is V-precompact for any $V \in \mathcal{N}$.

DEFINITION 5. A topological W^*-category R is called *"locally compact"* if for any $r \in R$ and any $V \in \mathcal{N}$ there exists a V-precompact neighbourhood of r.

It turns out that only for "locally compact" topological W^*-categories R one may find a C^*-algebra A such that R is isomorphic to $\mathrm{Rep}A$. Indeed, we have the following result:

THEOREM 8. *Let A be a C^*-algebra and H a Hilbert space. Assume that A and H are separable (in fact, like in Theorem 5, it is sufficient to assume that $\dim H$ is large enough). Then the topological W^*-category $\mathrm{Rep}A$ is "locally compact". If A has unity then $\mathrm{Rep}A$ is "compact".*

The proof of this theorem is given in [10]. It uses the Dixmier Lemma [1], the compactness of the set of all normalized completely positive maps from A into $B(H)$ endowed with the weak topology and the Stinespring theorem [6].

The condition that R is "locally compact" is also sufficient in order to obtain the positive answer to our main question. Namely we have:

THEOREM 9. *Let R be a topological W^*-category acting on a Hilbert space H. Assume that R satisfies Axioms* VII *and* VIII *and that R is "locally compact". Then*

1° *The algebra $C_\infty(R)$ is rich in the following sense: for any $r \in R$ the representation*

$$\pi_r: C_\infty(R) \ni F \to F(r) \in B(H)$$

is nondegenerate.

2° *Any nondegenerate representation of* $C_\infty(R)$ *acting on* H *is of the form* π_r *with* $r \in R$ *uniquely determined.*

3° *For any* $r, r' \in R$

$$\mathrm{Mor}(\pi_r, \pi_{r'}) = \mathrm{Mor}(r, r').$$

4° *The mapping*

$$R \ni r \to \pi_r \in \mathrm{Rep}_H C_\infty(R)$$

is a homeomorphism.

Moreover, if R *is "compact" then there is no singular representations of* $C(R)$ *and* $C_\infty(R) = C(R)$ *is a* C^**-algebra with unity.*

The proof of this theorem is given in [10]. The main step in this proof is to show that "locally compactness" implies that the set of states of $C(R)$ related to the singular representations of $C(R)$ is a closed face in the set of all states of $C(R)$. Then the statement of the theorem follows easily.

References

[1] Dixmier J., *Les* C^**-algèbres et leur representation*, §3, Lemme 3.5.6, Paris 1964.

[2] Ghez P., Lima R., Roberts J. E., W^**-Categories*, preprint, Marseille.

[3] Hadwin D. W., Continuous Functions of Operators: A Functional Calculus, *Indiana University Mathematical Journal* **27** (1) (1978).

[4] Kruszyński P. and Woronowicz S. L., A Non-Commutative Gelfand–Naimark Theorem, *Journal of Operator Theory* **8** (1982).

[5] Pedersen G. K., C^**-Algebras and Their Automorphisms Group*, Chapter 3.12, Academic Press, 1979.

[6] Stormer E., Positive Linear Maps of C^*-Algebras, *Lecture Notes in Physics* **29** (1974).

[7] Takesaki M., A Duality in the Representation Theory of C^*-Algebras, *Annals of Mathematics* **85** (1967), pp. 370–382.

[8] Voiculescu D., A Noncommutative Weyl–non Neumann Theorem, *Rev. Roumaine Math. Pures Appl.* **21** (1976), pp. 97–113.

[9] Woronowicz S. L., Operator Systems and Their Application to the Tomita–Takesaki Theory, *Journal of Operator Theory* **2** (1979), pp. 169–209.

[10] Woronowicz S. L., *Topological* W^**-Categories*, to appear.

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

R. W. BROCKETT

Nonlinear Control Theory and Differential Geometry

This report concerns recent developments in the use of differential geometric methods to study nonlinear problems in automatic control. This has been an active subject for more than a decade now with contributions, coming from researchers in many countries. Rather than focusing here on a particular subarea of this discipline we have allowed ourselves to range rather broadly over the field using the discussion of a few unsolved problems as the main thread. In this way we hope to give some indication of the scope of the current activity and to touch on a representative sample of the geometrical ideas which play a role.

Feedback

Finite-dimensional, continuous time control systems have as their description in local coordinates ($\dot{x} = dx/dt$)

$$\dot{x}(t) = \gamma\big(x(t), u(t)\big)$$

with $x(t)$ being a point in R^n and $u(t)$ being a point in R^m. Without loosing too much generality, we may describe a corresponding global object as follows. Let X be a finite-dimensional manifold and let $\pi\colon E \to X$ be a rank m vector bundle over X. Let TX denote the tangent bundle of X and let π^*TX denote the pullback of TX over E. Regarding (x, u) as a point in E, a section of the pullback of TX over E can be given locally by a function $\gamma(x, u)$ which, for each (x, u), specifies a velocity $\dot{x} = \gamma(x, u)$. We denote the set of all such sections by $\Gamma(E, \pi^*TX)$ and call the elements of this set *control systems.* Notice that associated with any $\gamma \in \Gamma(E, \pi^*TX)$ there is a vector field γ_0 which is obtained by setting u equal to 0; we call this vector field the *drift.*

This definition includes a great many situations which are of technological and mathematical interest such as mechanics problems with

u representing exogeneous forces or torques, electrical networks with u representing voltage or current sources, etc. In order to help fix ideas we consider a specific example which illustrate the main points of the definitions. Let X be the unit sphere bundle over E^3. (Think of $\{(x, \dot{x}) \mid x \in E^3, \|\dot{x}\| = 1\}$.) Construct E by taking the tangent bundle of the two-sphere and pulling it back over X. (Think of $E = \{(x, \dot{x}, u) \mid x \in E^3, \|\dot{x}\| = 1, 0 = \langle u, \dot{x}\rangle\}$.) The second order equation

$$\ddot{x} = u, \quad \langle u, \dot{x}\rangle = 0, \quad \|\dot{x}(t)\| = 1$$

then defines a control system, i.e. an element of $\Gamma(E, \pi^*TX)$. This control system has a number of possible interpretations. On one hand, it describes the motion of a newtonian particle of unit mass and unit speed being acted on by a controllable force which is constrained to do no work. On the other hand, it can be thought of as describing the end point of a curve in E^3 whose curvature is $\|u\|$. In this latter guise γ is a substitute for the more familiar Frenet–Serret system; x is the unit tangent vector, $u/\|u\|$ the unit normal, etc.

We return now to generalities. Given a control system γ we can replace u by $u + \alpha(x)$ and get a modified system which we denote by γ^α. Since a section of π: $E \to X$ is specified locally by a function α we see that what has just been given is the local coordinate description of a mapping

$$\nu\colon \Gamma(E, \pi^*TX) \times \Gamma(X, E) \to \Gamma(E, \pi^*TX),$$

$$\nu\colon (\gamma, \alpha) \to \gamma^\alpha,$$

where $\dot{\gamma}^\alpha$ has the local coordinate description $\dot{x} = \gamma\big(x, u + \alpha(x)\big)$. The section $\alpha \in \Gamma(X, E)$ is called a *feedback control law*. Notice that we can also think of ν as defining an action of the additive group $\Gamma(X, E)$ on the set of control systems $\Gamma(E, \pi^*TX)$. Because it is usually easy to implement a feedback control law, as opposed to making other modifications in the system which γ describes, it is important to have a good understanding of this group action. A specific question which arises in this way is the

Feedback Stabilization Problem. *Given $\gamma \in \Gamma(E, \pi^*TX)$ and given a subset $X_1 \subset X$ containing a distinguished point x_0, does there exist a feedback control law $\alpha \in \Gamma(X, E)$ such that x_0 is an asymptotically stable critical point for $(\gamma^\alpha)_0$ with a domain of attraction which includes all of X_1?*

We will discuss some partial results on this problem below, after we have has a chance to make further definitions and have introduced one more problem.

Returning now to the mapping ν, notice that for a fixed γ it describes a mapping of the vector space $\Gamma(X, E)$ into the vector space $\Gamma(E, \pi^*TX)$. If this mapping is affine then we see that γ must admit a local description of the form

$$\dot{x} = f(x) + G(x)u.$$

In the literature such systems are said to be *input-linear control systems* or *affine control systems*. As a further specialization, we will say that an input-linear control system $\gamma \in \Gamma(E, \pi^*TX)$ defines a *linear control system* if there exists a connection ∇ on X with respect to which (i) X a complete flat affine space, (ii) the image of $\gamma - \gamma_0 \colon E \to TX$ a flat subbundle of TX and (iii) in a neighborhood of each x we can describe γ by equations of the form

$$\dot{x} = Ax + Bu + \xi$$

with $x_1, x_2, \ldots, x_n$ satisfying $\nabla_{(\cdot)}(\partial/\partial x_i) \equiv 0$.

Linear control systems can be described quite concretely. Let (N_i, n_i) denote the affine transformation $x \to N_i x + n_i$. Let $\{(N_i, n_i)\}$ denote a group of affine transformations which act freely and properly discontinuously on R^n. As is well-known, the quotient space $X = R^n/\{N_i, n_i)\}$ then admits the structure of a complete flat affine space and all complete flat affine spaces arise in this way. In order to construct a linear system on $R^n/\{N_i, n_i\}$ we need to find A and B and a homomorphism

$$\varphi \colon \{(N_i, n_i)\} \to \mathrm{Gl}(m)$$

such that $A = N_i A N_i^{-1}$, $N_i B = B\varphi(N_i, n_i)$ and $An_i = 0$. Under these circumstances the range of B defines a flat subbundle of TX (which we take to be E) and the local description

$$\dot{x} = Ax + Bu$$

defines a linear control system on E.

Feedback Linearization Problem. *Given $\gamma \in \Gamma(E, \pi^*TX)$ under what circumstances does there exist $\alpha \in \Gamma(X, E)$ such that γ^α is a linear control system?*

In describing some results on stabilization and linearization we restrict discussion to the input-linear case. Obviously linearization can not be achieved without this assumption and rather little can be said about stabilization in the more general situation. With this assumption in force we define a subset of $T_x X$, the tangent space at x, by

$$\mathscr{F}_0(x) = \{\dot{x} \mid \dot{x} = \gamma^\alpha(x, 0) - \gamma_0(x),\ \alpha \in \Gamma(X, E)\}.$$

If we write $\dot{x} = f(x) + G(x)u$ this is just the range of $G(x)$. Denote the resulting distribution by $\mathscr{F}_0$. Proceeding inductively, we define $\mathscr{F}_k \supset \mathscr{F}_{k-1} \supset \ldots \supset \mathscr{F}_0$ by taking the distributions which correspond to ([,] denotes Lie bracket)

$$\mathscr{F}_{k+1}(x) = [\gamma^a(x, 0) - \gamma_0(x), \mathscr{F}_k] + \mathscr{F}_k.$$

It is not too hard to deduce from this definition that $\mathscr{F}_k(x)$ is, for $x = f(x) + G(x)u$, simply

$$\mathscr{F}_k(x) = \text{span}(B, AB, \ldots, A^k B),$$

where $A = (\partial f/\partial x)_x$ and $B = G(x)$. The linear system $\dot{x} = Ax + Bu$ is said to be *controllable* if $\text{rank}(B, AB, \ldots) = \dim x$. For the purpose of this paper we want to call a $\gamma \in \Gamma(E, \pi^* TX)$ *quasi-linear* if the dimensions of the $\mathscr{F}_k(x)$ are, for fixed k, independent of x and the controllability condition $\mathscr{F}_k(x) = T_x X$ is satisfied for some k.

As a local question feedback linearization is understood, i.e. necessary and sufficient condition for there to exist α such that γ^α is linear are

(i) γ should be quasi-linear;

(ii) The distributions $\mathscr{F}_0, \mathscr{F}_1, \ldots, \mathscr{F}_k$ should be integrable.

Furthermore if x_0 belongs to the set of possible rest points

$$S_0 = \{x \mid f(x) \in \text{Range } G(x)\}$$

then for suitable α, γ^α has the local description $\dot{x} = Ax + Bu$.

This has some implications for stabilizability. It is well known that a linear system $\dot{x} = Ax + Bu$ which is controllable can be stabilized to any point in S_0 using control laws of the forms $u = Cx + \xi$. Since asymptotic stability of equilibrium point of a nonlinear system is determined by the linearization if the linearization does not yield eigenvalues on the imaginary axis, this means that any quasi-linear system can be stabilized to any point in S_0 having a domain of attraction which includes some neighborhood of x_0.

We now turn to a result on the nonexistence of stabilizing control laws. Consider the class of systems for which $\gamma_0 = 0$, $\dim \mathscr{F}_0$ is constant and equal to the dimension of the fibers of E, and for $\mathscr{G}_{k+1} = [\mathscr{G}_k, \mathscr{G}_k] + \mathscr{G}_k$ with $\mathscr{G}_0 = \mathscr{F}_0$ have the property that the $\mathscr{G}_i$ are constant dimension with $\mathscr{G}_r = TX$ for some r. For reasons which will be explained in the next section we call such systems *quasi-Riemannian control systems*. The condition $\mathscr{G}_r = TX$ insures that it is possible to find a control which steers any initial state to any final state. The condition that γ_0 is zero insures that the set S_0 of potential equilibrium states is all of X. The quasi-Rieman-

nian systems are quasi-linear only in the very special case $\mathscr{F}_0 = TX \approx E$ and, except for this case, represent an opposite extreme. As an example we have the following nonstabilizability result.

If γ is feedback equivalent to a quasi-Riemannian control system with $\mathscr{F}_0 \neq TX$ then $S_0 = X$ but γ^a has no asymptotically stable critical points regardless of the choice of a.

We emphasize that this rules out even local asymptotic stability, regardless of the choice of x_0. The proof of this theorem is based on the fact that if x_0 were asymptotically stable there would be a Liapunov function V whose derivative $\dot{V} = \langle \partial V/\partial x, (\gamma^a)_0 \rangle$ would be negative on $B = \{x \mid V(x) = \varepsilon\}$ but that by a degree argument $\partial V/\partial x$ and $\mathscr{F}_0$ must be perpendicular at some point on B unless $\mathscr{F}_0 = TX$.

In this brief account we were unable to mention the interesting work on invariant distributions, decoupling, etc. (see the references to Hirschorn and Isidori *et al.*). In connection with this kind of work it would be useful to know the answer to the following

Dimensionality Problem for Integrable Subdistributions. *Given integers $n > m > 0$ find the largest integer $\varphi(n, m)$ such that we may assert that every locally defined distribution of dimension m in R^n has a locally defined subdistribution of dimension $\varphi(n, m)$ which is integrable.*

There are obvious global versions of this problem as well but the local version seems to be what is needed the most in control theory.

We close this section with some remarks on the literature. References [1, 3, 9–12] pertain to this section, references [3] and [9] contain papers by many authors working in this field and can be used to trace the literature; a complete set references would be several pages long.

Hamilton–Jacobi theory

The qualitative issues raised in the previous section are reflected in the solutions of concrete optimizations problems. The problem of minimizing

$$\eta = \int_0^t \langle u(\sigma), u(\sigma) \rangle d\sigma$$

for $\dot{x} = f(x) + G(x)u$, subject to the constant $x(0) = 0$, $x(t) = x_1$ will now be used to illustrate this and to give some indication about solved and unsolved problems. In addressing these questions we will asume that $\dot{x} = f(x) + G(x)u$ is *controllable* in the sense that for any given $t > 0$ and

any x_1 near 0 there is a control such that $x(t) = x_1$. The most interesting behaviour corresponds to the case $f(0) = 0$ and we limit ourselves to this case.

Introduce the value function $S(t, x)$ defined to be the minimum value of η expressed as a function of t and the end point x. Assuming that the indicated partial derivatives exist, S satisfies the well known Hamilton–Jacobi equation

$$\frac{\partial S}{\partial t} = \left\langle \frac{\partial S}{\partial t}, f(x) \right\rangle - \frac{1}{4} \left\langle G^T(x) \frac{\partial S}{\partial x}, G^T(x) \frac{\partial S}{\partial x} \right\rangle.$$

Since there may be many points where these derivatives do not exist we must interpret this equation in some weak sense.

Degenerate Hamilton–Jacobi Problem. *Give a complete theory of the solution of the above Hamilton–Jacobi equation in a neighborhood of an equilibrium point of f assuming only the controllability of $\dot{x} = f(x) + G(x)u$.*

If f is identically zero and $G(x)$ is of rank $\dim X$ then this Hamilton–Jacobi equation appears in Riemannian geometry. More precisely, if the metric tensor is expressed as $(G(x)G^T(x))^{-1}$ then $S(1, x)$ is the square of the distance between 0 and x. In this sense, the special case $f \equiv 0$ is a generalization of Riemannian geometry. This also explains why we called this situation quasi-Riemannian in the last section.

To begin with we solve the above equation in the linear case. This solution is well known, easy to verify, and will be qualitatively correct for the quasi-linear problems of the previous section. Using the notation $\dot{x} = Ax + Bu$, introduce

$$W(t) = \int_0^t e^{A(t-\sigma)} B B^T e^{A^T(t-\sigma)} d\sigma.$$

Assuming controllability this matrix will be invertible for all $t > 0$ and $S(t, x) = \langle x, W^{-1}(t)x \rangle$ will satisfy the corresponding Hamilton–Jacobi equation. $W^{-1}(t)$ has a pole at $t = 0$ which we now describe. Let r be the least integer such that $(B, AB, \ldots, A^rB)$ is of rank n. Then for t small and positive

$$W^{-1}(t) = \sum_{i=-1}^{-r} H_i t^i + M(t)$$

with $M(t)$ analytic near 0, $M(0) = 0$ and the H_i having range $(B, AB, \ldots$ $\ldots, A^{i-2}B)$ in their kernel. Note that $S(t, x)$ is homogeneous in t only in the uninteresting case corresponding to rank $B = n$, $A = 0$.

We now turn to a quasi-Riemannian situation. We coordinatize space X by an m-vector y and an m by m skew-symmetric matrix Z. The equations are

$$\dot{y} = u,$$
$$\dot{Z} = yu^T - uy^T.$$

It is easy to see this system is controllable on its $m(m+1)/2$-dimensional state space. Variational arguments show that an optimal y satisfies $y + \Omega\dot{y} = 0$ for same skew symmetric matrix Ω. If $y(0) = 0$ and $Z(0) = 0$ then $y(t) = e^{\Omega t}\lambda - \lambda$ and $Z(t)$ is given by

$$Z(t) = y(t)y^T(t) - 2\int_0^t e^{\Omega\sigma}\lambda\lambda^T e^{\Omega^T\sigma}\,d\sigma\Omega.$$

The corresponding $S(t, y, Z)$ satisfies the following identities:

$$S(t, y, 0) = y^2/t,$$
$$S(t, y, Z) = S(1, y, Z)/t,$$
$$S(t, y, Z) = S(t, \alpha y, \alpha^2 Z)/\alpha^2,$$
$$S(t, y, Z) = S(t, \theta y, \theta Z\theta^T), \qquad \theta\theta^T = I.$$

From the last of these we see that $S(t, 0, Z)$ must be expressible in terms of the eigenvalues of Z and it has been shown that

$$S(t, 0, Z) = 2\pi(\lambda_1 + 2\lambda_2 + \ldots + r\lambda_r)/t,$$

where $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_r$ are the positive eigenvalues of iZ. The derivation of this result shows, moreover, a rather remarkable "exclusion principle" which accounts for the different weights. Briefly stated, the optimal controls which steer this system from $(0, 0)$ to $(0, Z)$ have the form $e^{\Omega t}b$ with the eigenvalues of Ω being multiples of $2\pi i$. However, it turns out that the nonzero eigenvalues of Ω cannot be repeated and that the number of distinct eigenvalues must equal rank Z.

For the special case $m = 2$ it is illuminating to write down the first few terms of the Newton–Puiseux expansion for S. From the scaling properties we see that $S(t, y, z) = (|z|/t)\cdot s(y^2/|z|)$. In fact it has an expansion in $y/\sqrt{z}$ which we may express as

$$S(t, y, z) = (|z|/t)(2\pi - \sqrt{8\pi\|y\|^2/z} + \|y\|^2 + \ldots),$$

this asymptotic expansion being valid off the plane $z = 0$. It is possible to interpret $\sqrt{S(1, y, z)}$ as a distance function and this has been persued in some detail in our paper on singular Riemannian geometry cited below.

Although it may not be apparent immediately, this second case is a paradigm for the general class of problems whose Hamiltion–Jacobi equation is

$$\frac{\partial S}{\partial t} = -\frac{1}{4}\left\langle G^T(x)\frac{\partial S}{\partial x}, G^T(x)\frac{\partial S}{\partial x}\right\rangle$$

under the hypothesis that rank $G(0) = m$ and $\dim x = \dim([\mathcal{G}, \mathcal{G}]+\mathcal{G}) = m(m+1)/2$, provided that we restrict our attention to a neighborhood of $x = 0$. This is explained in more detail in the paper just refered to.

It is clear that one can raise many questions in this area. The following is obviously one of interest.

HAMILTON–JACOBI ASYMPTOTICS PROBLEM. *Find the correct generalization of the given Newton–Puiseux expansion for an arbitrary quasi-Riemannian problem, assuming analyticity of* $\mathcal{G}$.

References [2, 4, 5, 8] deal with aspects of this material. We should point out that there is a discrepancy between our formula for $S(1, 0, Z)$ (which is taken from [2]) and the claims of [4], Section 5.3. It seems that in [4] the possibility of optimal trajectories corresponding to "higher harmonics" has been overlooked.

Stochastic phenomena

No discussion of the relationships between control theory and differential geometry should fail to touch on how these ideas illuminate problems related to Itô models of the form

$$dx = \hat{f}(x)\,dt + \sum_{i=1}^{m} g_i(x)\,dw_i.$$

We sketch out how this goes and later on describe its connection with nonlinear filtering.

The first idea is that it is, for some purposes at least, worthwhile to study the above stochastic equation with the help of the control model

$$\dot{x} = \hat{f}(x) - \frac{1}{2}\sum_{i=1}^{m}\left(\frac{\partial g_i}{\partial x}\right) g_i(x) + \sum_{i=1}^{m} g_i(x)\,u_i$$

$$\stackrel{\text{def}}{=} f(x) + G(x)u.$$

The explanation of why $\hat{f}$ got traded for f when we replaced "dw/dt" by u is to be found in the asymmetrical definition of the Itô integral — a point that has been discussed many times. Actually one gets more than just a control system $\dot{x} = f(x) + G(x)u$ out of this. Because Wiener processes in R^m are defined with respect to the quadratic form defining the variance, we get, automatically, an innerproduct on u as well.

Returning to the stochastic equation again, it has associated to it a second order differential operator L with the property that the probability density $\varrho(t, x)$ evolves according to

$$\frac{\partial \varrho}{\partial t} = L\varrho.$$

If m is less then $\dim x$, L will be degenerate; in fact it is of the form $L_0 + L_1^2 + \ldots + L_m^2$ where the L_i are first-order differential operators. The first point of contact between the control system and the stochastic equations is that $\dot{x} = f(x) + G(x)u$ has the property that at any time $t > 0$ the reachable set of states from a given x_0 has nonempty interior if and only if $\partial/\partial t - L$ is hypoelliptic. Thus the support and the smoothness of the probability density is what one would guess it to be by looking at the sample path behavior of the related control system. The following problem can serve as a focus for our remaining remarks.

Hamilton–Jacobi/Fokker–Planck Problem. *Relate the small time behavior of the solution of Fokker–Planck equation to the solution of the Hamilton–Jacobi equation associated with the corresponding control problem.*

Of course in the case of full ellipticity this is an absolutely standard idea.

As we have done previously, we recall the situation for the linear case. Consider

$$dx = Ax\,dt + B\,dw.$$

There is a beautiful formula for the probability density corresponding to $x(0) = 0$

$$p(t, x) = \frac{1}{\sqrt{(\det S_{xx})(2\pi)^n}}\, e^{-S(t,x)},$$

where S is the solution of the Hamilton–Jacobi equation discussed above. What this equation says is that the probability density at (t, x), given $x(0) = 0$, is inversely proportional to the exponential of the cost of getting

to x from zero in t units of time. The expansion of $S(t, x)$ given in connection with Hamilton–Jacobi theory gives a rather detailed picture of the small time asymptotics of p. It is apparent, for example, that the rate of growth of $p(t, x)$ depends strongly on the particular subspace in which x lies.

We now turn to the quasi-Riemannian prototype

$$dy = du,$$
$$dZ = y\,dw^T - dw\,y^T.$$

P. Levy studied a special case of these equation corresponding to $\dim y = 2$ and recently there has been a great deal of interest in the general situation by probabilists and analysts alike, due in part to the many explicit formulae which describe the relevant probability distributions.

Of principle interest for our purpose is the geometrical interpretation of the right-hand side of the diffusion equation as an analog of the Laplace–Beltrami operator. This goes hand in hand with the interpretation of $\sqrt{S(1, y, Z)}$ as distance and leads to a generalization of the well known formula of Varadhan

$$\lim_{t\to\infty} 2t \ln p(t, x) = -d^2(x)$$

relating the distance from x_0 to x and the probability density at (t, x) given that $x(0) = x_0$.

Our final problem concerns the area of estimation theory and illustrates again the value of sample path considerations in those situations which are sufficiently robust. The key idea is to try to capture the symmetries which make finite-dimensional estimatimation possible in a Lie algebra setting.

We are given a stochastic differential equation

$$dx = f(x)\,dt + g(x)\,dw$$

together with an observation

$$dy = h(x)\,dt + dv, \qquad y \in R^1$$

and wish to find $\varrho(t, x \mid y_{[0,t]})$, the probability density at (t, x) *conditioned on the observations over the whole interval* $[0, t]$. We assume $\varrho(0, x)$ is given. The equation for ϱ is nonlinear but a certain path dependent multiple of ϱ, which we denote by $\hat{\varrho}$, satisfies the stochastic partial differential equation (here L is the Fokker–Planck operator associated with $dx = f(x)\,dt + g(x)\,dw$)

$$\partial\hat{\varrho} = L\hat{\varrho}\,dt + dy\,h(x)\hat{\varrho}.$$

If we associate to this equation a control system using the procedure outlined above we obtain

$$\frac{\partial \hat{\varrho}}{\partial t} = \left(L - \tfrac{1}{2}h^2(x)\right)\hat{\varrho} + uh(x)\hat{\varrho}.$$

Abstractly this is an equation of the form

$$\dot{w} = Aw + uBw.$$

If it happens that the Lie algebra generated by the operators A and B is finite-dimensional we would be lead to postulate a solution of the form

$$x(t) = e^{H_1\psi_1} e^{H_2\psi_2} \dots e^{H_r\psi_r}\, x(0),$$

where $H_1, H_2, \dots, H_r$ is a suitably chosen ordered basis for the Lie algebra generated by A and B and $\{\psi_i\}$ is a set of scalar functions dependent on u. In this context the Lie algebra

$$\{L - \tfrac{1}{2}h(x),\, h(x)\}_{LA}$$

is called the *estimation algebra.* Rather remarkably, in the standard Gauss–Markov cases it is the same as the oscillator algebra of quantum mechanics. Our final problem is the

Classification of Finite-Dimensional Estimation Algebras. *Find all the finite-dimensional Lie algebras which can occur as the estimation algebras for diffusion processes.*

The literature on the connections between the Hamilton-Jacobi equations and degenerate diffusions is very large. We mention Gaveau [4] and the more recent works [7, 13]. The book [6] contains a number of papers on the connection between Lie algebasr and filtering. The literature can be traced from these.

Acknowledgements

This work was supported in part by the U.S. Army Research Office under Grant No. DAAG29-79-C-0147, Air Force Grant No. AFOSR-81-7401, the Office of Naval Research under JSEP Contract No. N00014-75-C-0648, and the National Science Foundation under Grant No. ECS-81-21428.

References

[1] Brockett R. W., Feedback Invariants for Nonlinear Systems. In: *Proc. IFAC Congress,* Helsinki, 1978.

[2] Brockett R. W., Control Theory and Singular Riemannian Geometry. In: P. Hilton and G. Young (eds.), *New Direction in Applied Mathematics,* Springer-Verlag, New York, 1981.

[3] Brockett R. W., *et al.* (eds.), *Differential Geometric Control Theory*, Birkhäuser, Boston, Ma., 1983.

[4] Gaveau B., Principle de Moindre action propagation de la chaleur et estimees sons elliptiques sur certains groupes nilpotents, *Acta Mathematica* **139** (1977), pp. 95–153.

[5] Gunther N., *Hamiltonian Mechanics and Optimal Control*, Ph.D. thesis, Harvard University, 1982.

[6] Hazewinkel M. and Willems J. C. (eds.), *Stochastic Systems: The Mathematics of Filtering and Identification and Applications*, Reidel, 1981.

[7] Helmes K. and Schwane A., *Levy's Stochastic Area Formula in Higher Dimensions*, Springer Lecture Notes in Control and Information Sciences **42**, Springer-Verlag, New York, 1982.

[8] Hermann R., Geodesics of Singular Riemannian Metrics, *Bull. AMS* **79** (1973), pp. 780–872.

[9] Hinrichsen D. and Isidori A. (eds.), *Feedback Control of Linear and Nonlinear Systems*, Lecture Notes in Control and Information Sciences, Springer-Verlag, Berlin, 1982.

[10] Hirschorn R., (A, B)-Invariant Distributions and the Disturbance Decoupling of Nonlinear Systems, *SIAM J. Control and Optimization* **17** (1981), pp. 1–19.

[11] Isidori A., Krener A. J., Gori-Gorgi C., and Monaco S., Nonlinear Decoupling via Feedback: A Differential Geometric Approach, *IEEE Trans. Aut. Control* **16** (1981), pp. 331–345.

[12] Jakubczyk B. and Respondek W., On the Linearization of Control Systems, *Bulletin de L'Académie Polonaise des Sciences* **28** (1980), pp. 517–522.

[13] Taylor Th., *Hypoelliptic Diffusions and Nonlinear Control Theory*, Ph.D. thesis, Harvard University, 1983.

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

H. W. KNOBLOCH

Nonlinear Systems: Local Controllability and Higher Order Necessary Conditions for Optimal Solutions

1. Introduction

We consider control systems which are defined in terms of an ordinary differential equation

$$\dot{x} = f(t; x, u). \tag{1.1}$$

u is the control variable and may be subject to a constraint of the form $u \in U$. We allow specialization of u to an admissible control function $u(\cdot)$, that is, a function which is piecewise of class C^∞ on $\boldsymbol{R}$ and has a range whose closure is contained in U. The function f on the right-hand side of (1.1) is assumed to be sufficiently smooth. Hence, if an admissible control function is substituted for u in (1.1), we obtain a differential equation which allows application of all standard results concerning the existence, uniqueness and continuous dependence of solutions (see e.g. [1], Sections 2–4). Any one of these solutions will be denoted by $x(\cdot)$ and called an *admissible trajectory*. We also refer to the pair $(u(\cdot), x(\cdot))$ as a solution of (1.1). If we speak of an optimal solution, we mean a solution which minimizes the functional within the class of all admissible trajectories satisfying boundary conditions of the usual type. It is always tacitly assumed that the value of the functional can be identified with the terminal value of a component of the state vector.

We are concerned in this lecture with two types of problems which can be studied independent from each other. However, it is clear from the beginning that one can expect some kind of duality between statements pertaining to each of these problems. Among other things we will undertake in this lecture an attempt to put this duality into more concrete forms.

Problems of the first type deal with necessary conditions which have

to hold along singular arcs. A singular arc is a portion of an optimal solution which is such that the control variable is specialized to interior values of the control set U. We restrict our attention to conditions which have to hold pointwise along a singular arc and which assume the form of a multiplier rule, i.e., a rule which can be expressed as an inequality of the form $y(t)^T a(t) \leqslant 0$, where $y(\cdot)$ is the usual adjoint state vector.

Problems of the second type carry the label "local controllability". The precise definition goes as follows. Let there be given, for all t in some interval $[t_0, \tilde{t}]$, an arbitrary solution $u(\cdot), x(\cdot)$ of (1.1) (called "reference solution" from now on). Local controllability along this solution and for $t = \tilde{t}$ means: There exists, for every sufficiently small $\varepsilon > 0$, a full neighborhood of $x(\tilde{t})$ which can be reached at time $t = \tilde{t}$ by travelling along admissible trajectories starting at time $t = \tilde{t} - \varepsilon$ from $x(\tilde{t} - \varepsilon)$. In other words: $x(\tilde{t})$ is an interior point of the set of all states to which the system can be steered from $x(\tilde{t} - \varepsilon)$ within time ε. We remark that our notion of local controllability coincides with Sussmann's "small time local controllability" (cf. [2], Sec. 2.3), if the system equation is autonomous and the reference solution stationary.

It is somehow clear from the above definitions that problems of both types are concerned with local properties of solutions and that these properties, in a certain sense, exclude each other. If a solution is optimal (in the sense as explained above), then the set of all states into which the system can be steered from $x(\tilde{t} - \varepsilon)$ within time ε is situated on one side of a certain hyperplane through the terminal point $x(\tilde{t})$ and we have no local controllability for $t = \tilde{t}$. Thus one can expect some kind of correspondence between results concerning singular extremals and those concerning local controllability which roughly speaking, amounts to "reversing" conclusions in a suitable way. We will demonstrate in Section 2 how this kind of reasoning can be put on more solid grounds by presenting two theorems — one giving necessary conditions for singular arcs and the other giving sufficient conditions for local controllability — in which all statements are expressed in terms of one and the same object, namely the local cone of attainability. This is a set of elements of the state space which is associated with each point of the reference solution. Since we may think of a solution as a curve parametrized by the time t, we denote this set by $\mathcal{K}_t$. The precise definition is given in Section 2; it will turn out to be a modification of the definition of the set Π_t which was introduced in [1], Section 9. In fact, $\mathcal{K}_t$ is a subset of Π_t. The reason that we dispense here with some elements of Π_t is the gain in mathematical structure. $\mathcal{K}_t$ enjoys certain properties which cannot be inferred from the definition

of Π_t: It is a *convex* cone and its maximal linear subspace is invariant under a certain operator Γ (Theorem 2.1). The operator Γ will be described in detail in Section 2. In contrast to the philosophy adopted in [1] we prefer here a definition which depends on the choice of a special reference solution since it helps to bring out the system theoretic aspect of this operator. One can look upon Γ from the viewpoint of linear systems theory; it then appears as a generalization of the process which leads to the construction of the controllability matrix. Indeed, if the system equation is linear and is given as

$$\dot{x} = A(t)x + B(t)u, \tag{1.2}$$

then the columns of this matrix can be generated from the columns of $B(t)$ by repeated application of Γ. One can also look on it from the differential geometric viewpoint. If the system equation is autonomous and the reference solution stationary then the simplest way to explain the application of Γ is in terms of a Lie-bracket involving f (= the function which appears on the right-hand side of (1.1)). This, by the way, explains why the forming of the Lie-bracket with f is a nonlinear analogue of the linear mapping is defined in terms of the matrix $A(t)$ of the linear system (1.2).

Regardless of which view one prefers, what counts for our purposes are the following two facts. (i) We can define Γ without any restrictive assumptions, as linearity of the equation or time independence of the reference solution. (ii) One can use Γ in order to generate new elements of $\mathscr{K}_t$ out of given ones: From the previously mentioned invariance property of $\mathscr{K}_t$ one infers that the following statement holds true:

$$\pm p \in \mathscr{K}_t \quad \text{implies} \quad \pm \Gamma^\mu(p) \in \mathscr{K}_t, \quad \mu = 1, 2, \ldots \tag{1.3}$$

To get an impression of the scope of this result it might be helpful to consider a special case. Let us assume that the system is linear in u, and hence defined by a differential equation of the form

$$\dot{x} = f_0(t; x) + \sum_{\nu=1}^{m} u^\nu g_\nu(t; x), \qquad u = (u^1, \ldots, u^m). \tag{1.4}$$

Furthermore, let us assume that the reference control satisfies the condition $u(t) \in \operatorname{int} U$ for all $t \in [t_0, \tilde{t}]$. Using standard variational techniques it is then not difficult to see that $\pm g_\nu\big(t; x(t)\big) \in \mathscr{K}_t$ for all $t \in [t_0, \tilde{t}]$. Hence, it follows from (1.3) that the linear space spanned by $(\Gamma^\mu g_\nu)\big(t, x(t)\big)$,

$\mu = 0, 1, \ldots, \nu = 1, \ldots, m$, is contained in $\mathscr{K}_t$. This space was introduced in [1] and denoted there by $\mathfrak{B}(t)$ (cf. Section 1, in particular p. 5), it can be identified with the columns of the controllability matrix for the linearized system (i.e., for the linear system (1.2) with $A(t) := (\partial f/\partial x)\big(t; x(t), u(t)\big)$ and $B(t) := (\partial f/\partial u)\big(t; x(t), u(t)\big)$. It is therefore not surprising to rediscover $\mathfrak{B}(t)$ as a part of $\mathscr{K}_t$ and to establish its invariance with respect to the operator Γ; this could be verified also by standard arguments. The importance of the statement (1.3) rests upon the fact that it allows to *extend* the space $\mathfrak{B}(t)$ by adjoining further elements p without losing its two basic properties, namely, that of being a subspace of $\mathscr{K}_t$ and being invariant with respect to Γ. In other words: One can add to the generators g_ν of the space $\mathfrak{B}(t)$ all elements $p \in \mathscr{K}_t$ which satisfy the condition $\pm p \in \mathscr{K}_t$ and then treat the enlarged set as if it would be the set of generators for $\mathfrak{B}(t)$. Whether this is a useful insight or not, depends on the concrete possibilities of constructing vectors p with the property $\pm p \in \mathscr{K}_t$ and which are not already elements of $\mathfrak{B}(t)$. What is known in this respect is very little, nevertheless it seems worthwhile to review carefully the material existing up to now.

The first examples of non-trivial elements p which are contained in $\mathscr{K}_t$ together with their negatives are among what we will call "second order elements" and discuss in detail in Section 3. The name stems from the fact that the necessary conditions which can be expressed in terms of these elements are commonly called "second order". We will present in Section 3 a general definition of the "order" of an element of $\mathscr{K}_t$ and give a complete description of the set of all second order elements. Special emphasis is put on those p which appear together with $-p$ in this set and which therefore must be orthogonal to the adjoint state variable along an optimal solution. It has been known since long that for a system of the form (1.4) the "mixed" Lie-brackets

$$p_{\nu,\mu} := [g_\nu, g_\mu] \tag{1.5}$$

enjoy this property along a singular arc. But the background of this was not recognized until recently when Vârsan [4] announced the following result: Local controllability along a reference solution of the system (1.4) can be inferred from the following two conditions:

(i) the reference control assumes values in the interior of U for all t,

(ii) the controllable subspace $\mathfrak{B}(t)$ and the elements

$$\Gamma^\gamma(p_{\nu,\mu}), \quad \nu, \mu = 1, \ldots, m, \ \gamma = 0, 1, \ldots$$

generate together the whole state space.

As we will see in Section 3, both the second order equality type conditions and Vârsan's result are true since the hypothesis (i) implies $\pm p_{\nu,\mu} \in \mathscr{K}_t$ for any system of the form (1.4). It is also possible — under suitable extra hypotheses — to add further second order elements to the $p_{\nu,\mu}$ in such a way that one arrives at a similar type of controllability criterion. However, all second order elements p which are such that $\pm p \in \mathscr{K}_t$ reduce to zero if the dimension m of the control variable is equal to one (note that $p_{\nu,\mu} = 0$, in view of (1.5)). The absence of those elements can be understood from the nature of the corresponding necessary conditions: they can be compared to the standard second order tests in calculus. Basically, these tests are inequalities (semi-definiteness of a quadratic form), which eventually may lead to equality type statements; namely, if the form fails to be definite. All these statements, however, are trivial if there is not more than one variable.

It should be pointed out thatVârsan's result reflects a typical non-linear system property: There exists a kind of "crosswise" interaction between the components of u which is exercised through the state vector (note that $p_{\nu,\mu} = 0$ if the g_ν do not depend upon x) and which cannot be recognized by means of linearization since for a linear system the action of $u = (u^1, \ldots, u^m)$ is just the superposition of the action of the components u^ν. In precise mathematical terms this interaction is expressed by the fact that one can simply adjoin the $p_{\nu,\mu}$ to the generators of $\mathfrak{B}(t)$ without destroying the controllability properties of this space.

Next, we wish to say a few words about possible extensions of $\mathfrak{B}(t)$ in case of a scalar control variable u. It is clear from what was said above that one has to search for possible candidates among higher-than-second-order elements, but it is presently not obvious how this search can be carried out in a systematic way. What one expects to find is some kind of hierarchy among the subspaces of $\mathscr{K}_t$, which corresponds to the hierarchy among higher order tests in optimization. Of course the controllable subspace $\mathfrak{B}(t)$ of the linearized system equation should be the member of lowest rank.

The first attempt to put this idea into a more concrete form has been undertaken by H. Hermes and completed by H. Sussmann [3]. It led to a controllability criterion for a system of the form

$$\dot{x} = f_0(x) + ug(x), \qquad u \text{ scalar}, \tag{1.6}$$

with a stationary point (u_0, x_0) playing the role of the reference solution. The crucial condition which enters this criterion concerns the Lie-brackets associated with system (1.6) and evaluated at $(u, x) = (u_0, x_0)$.

To be more specific, it is assumed that all brackets which involve the quantity g an even number of times can be expressed as a linear combination of Lie-brackets which are of odd order with respect to g. This even-odd relationship resembles the one which is well known from elementary calculus: If all derivatives up to order $2k$ vanish at an extremal point of a function, then the $(2k+1)$th derivative must also vanish there. Now, vanishing of some derivative at an extremal point is an equality-type necessary condition. In optimal control theory conditions of this kind appear in the form of an orthogonality relation $y(t)^T p = 0$. As we have seen before, they arise from elements p of the state space which satisfy the condition $\pm p \in \mathscr{K}_t$. We wish therefore to pose the following question which is a natural modification of the Hermes conjecture: Assume that the above stated condition holds for all Lie-brackets which are of order at most $2k$ with respect to g, k being a fixed positive integer. Is it then true that the linear space spanned by all Lie-brackets which are of order at most $2k+1$ with respect to g belong to $\mathscr{K}_t$?

In this generality the question probably cannot be answered along the lines of existing methods; in particular, it is unlikely that Sussmann's proof of the original conjecture could be carried over. Note that it is required to establish the existence of specific elements in $\mathscr{K}_t$, regardless of whether we have local controllability or not. It seems, however, conceivable that special cases can be treated e.g. with methods taken from [1] and that one would then be able to examine from case to case how much of the assumptions underlying the Hermes–Sussmann result is actually required. From the viewpoint of applications one would anyhow welcome results which are more restricted in its scope in return for more flexibility with respect to the hypotheses. Some steps in this direction have been undertaken and will be discussed in the lecture. In particular, it seems very likely—though not all details have been cleared — that for systems of the form (1.6) one can extend the space $\mathfrak{B}(t)$ by adjoining third order elements (i.e. vectors which can be written as third order polynomials in the components of g, g_x, g_{xx}, etc.) under the assumption that the Lie-bracket

$$[g, [g, f_0]] \tag{1.7}$$

evaluated at the reference trajectory $x = x(t)$ is contained in $\mathfrak{B}(t)$ for $t \in [\tilde{t} - \varepsilon, \tilde{t}]$. The reference solution need not be stationary; however, $u(\cdot)$ has to assume values in the interior of the control set U. To compare this result with the Hermes conjecture, one has to take into account that

in case of a stationary reference solution the space $\mathfrak{B}(t)$ is independent from t and coincides with the linear span of those Lie-brackets which are first order with respect to g. Hence it is required — in case of a stationary solution — that (1.7) is a linear combination of first-order Lie-brackets in order to ensure the existence of certain third-order elements in $\mathscr{K}_t$. The conclusion is certainly much weaker than what would follow from the Hermes conjecture (in case of $k = 1$). On the other hand, one is relieved from the necessity of checking *all* Lie-brackets which are of order 2 with respect to g. In fact, there are some examples in the engineering literature (e.g. Lawden's spiral) where (1.7) is the only one among these brackets which is easy to compute.

The results which have been outlined so far (one more will be added in Section 3) can all be proved by a combination of methods, which could be summarized as the "analytic" approach to control theory. A considerable portion of it has been developed in [1] and used there to establish higher order necessary conditions for singular arcs. The starting point is the notion of control variations. These are parameter-dependent local modifications of the control function and the trajectory around a given reference solution. Later, in order to handle formal problems, one finds it convenient not to relate all results with the reference solution but to work directly with the right-hand side of the system equation. The analytic approach leads thereby straight into an ad-hoc-made algebraic theory of non-linear systems, which appears at first glance to be a rather natural generalization of linear system theory. The connection with the differential geometric approach is less obvious; the comparison of these two basic methods in control theory will play a major role in the lecture. At present it is safe to say that the analytic techniques seem to be rather efficient if one wants to refine existing results and, in particular, get rid of restrictive assumptions concerning the system equation or the reference solution. Furthermore, they seem to be well suited for a better exploitation of the specific nature of a given problem. This also can be of an advantage if one has to compute from the right-hand side of equation (1.1) those quantities which one has to know in order to apply the general results. An illustrative example is the "economic" version of the generalized Clebsch–Legendre condition which was given in [1] (Theorem 20.2).

The following two sections constitute a short account of the essential definitions and facts on which the analytic approach to non-linear systems theory is based. Except for occasional remarks we will not enter into a discussion of the proofs. All details — as far as they cannot be found in

existing literature — will be given in a dissertation, which is presently prepared at the Department of Mathematics in Würzburg.

2. The local cone of attainability

This section is devoted to a closer study of the sets $\mathscr{K}_t$. We consider a reference solution $\tilde{u}(\cdot), \tilde{x}(\cdot)$, which is defined on some interval $[t_0, \tilde{t}]$ and which will be kept fixed throughout this and the next section. For simplicity we will assume that $\tilde{u}(\cdot)$ is of class C^∞ on some interval $[\tilde{t}-\varepsilon, \tilde{t}]$, in particular, that any derivative of $\tilde{u}(\cdot)$ has a left-hand limit at $t = \tilde{t}$. Our first aim is to define $\mathscr{K}_{\tilde{t}}$.

To this purpose we have to introduce some notation. In the sequel we will use the symbol λ in order to denote a positive parameter. Whenever λ appears in a formula involving $\mathcal{O}$-terms we wish to formulate an asymptotic relation which has to hold for $\lambda \to 0$, uniformly with respect to all remaining variables occurring in the formula.

The definition of $\mathscr{K}_{\tilde{t}}$ will be based on a modification of what was called in [1] a "control variation concentrated at $t = \tilde{t}$." We consider families of control functions $u(t; \tau, \lambda)$ which depend on two real parameters τ, λ and which are defined for $t \in \boldsymbol{R}$, $0 < \lambda < \lambda_0$, $\tilde{t} - \varepsilon_0 \leqslant \tau \leqslant \tilde{t}$. The following conditions should hold ($\lambda_0, \varepsilon_0, \varkappa$ are positive constants).

(i) $u(t; \tau, \lambda)$ is bounded for all t, τ, λ.

(ii) $u(\cdot, \tau, \lambda)$ is — for fixed τ, λ — an admissible control function.

(iii) $u(t; \tau, \lambda)$ coincides with the reference control whenever t, τ satisfy the relation $t \leqslant \tau - \lambda^\varkappa$.

(iv) Let $x(t; \tau, \lambda, x_0)$ be the solution of the initial value problem

$$\dot{x} = f\big(t; x, u(t; \tau, \lambda)\big), \qquad x(t_0) = x_0; \tag{2.1}$$

then $x(t; t, \lambda, x_0)$ can be extended to a C^∞-function of t, λ, x_0 on a full neighborhood of the set $\{t, \lambda, x_0 \colon t \in [\tilde{t}-\varepsilon_0, \tilde{t}], \lambda = 0, x_0 = \tilde{x}(t_0)\}$.

Note that (i) and (iii) imply — by standard arguments — that the solution of the initial value problem (2.1) exists on the interval $[t_0, \tau]$ whenever λ and $\|x_0 - \tilde{x}(t_0)\|$ are sufficiently small. Furthermore, this solution approaches the solution of the initial value problem

$$\dot{x} = f\big(t; x, \tilde{u}(t)\big), \qquad x(t_0) = x_0 \tag{2.2}$$

if $\lambda \to 0$. So $x(t; t, \lambda, x_0)$ is well defined on a set of the form

$$\{t, \lambda, x_0 \colon t \in [\tilde{t}-\varepsilon, \tilde{t}],\ 0 < \lambda < \lambda_0',\ \|x_0 - \tilde{x}(t_0)\| < \varepsilon_0'\}. \tag{2.3}$$

It is then required (cf. (iv)) that this function could be extended to a C^∞-function of t, λ, x_0 defined on a full neighborhood (in the (t, λ, x_0)-space) of the set (2.3). For simplicity we will denote the extended function also by $x(t; t, \lambda, x_0)$.

If we take $x_0 = \tilde{x}(t_0)$ then the solution of the initial value problem (2.2) is just the reference trajectory $\tilde{x}(t)$. In other words: we have $x\big(t; \tau, 0, \tilde{x}(t_0)\big) = \tilde{x}(t)$ identically in t, τ and therefore also

$$x\big(t; t, 0, \tilde{x}(t_0)\big) = \tilde{x}(t)$$

for all $t \in [\tilde{t} - \varepsilon_0, \tilde{t}]$. Now, according to our hypothesis, the function $x\big(t; t, \lambda, \tilde{x}(t_0)\big)$ admits a Taylor expansion with respect to λ. Hence we have a relation of the form

$$x\big(t; t, \lambda, \tilde{x}(t_0)\big) = \tilde{x}(t) + \lambda^s p(t) + \mathcal{O}(\lambda^{s+1}) \tag{2.4}$$

where s is a positive integer and $p(t)$ is infinitely differentiable on $[\tilde{t} - \varepsilon_0, \tilde{t}]$. Furthermore, $p(\cdot)$ does not vanish for all t. In this manner one can associate with any family of control functions an n-dimensional vector $p(t)$ which is of class C^∞ and is not identically zero on $[\tilde{t} - \varepsilon_0, \tilde{t}]$. These $p(\cdot)$ play the decisive role in the definition of $\mathscr{K}_{\tilde{t}}$, which will now be given.

Definition 2.1. $\mathscr{K}_{\tilde{t}}$ is the collection of all n-dimensional vectors p having the following property:

There exists a family of control functions satisfying conditions (i)–(iv) and such that p equals the terminal value $p(\tilde{t})$ of the first nonvanishing coefficient in the Taylor expansion (2.4).

One can see immediately that the inclusion holds:

$$\mathscr{K}_{\tilde{t}} \subseteq \Pi_{\tilde{t}}, \tag{2.5}$$

where $\Pi_{\tilde{t}}$ is the set associated with the reference solution (and the time-instant $t = \tilde{t}$) according to Definition 9.2 in [1]. We will not go into the discussion whether $\mathscr{K}_{\tilde{t}}$ may be a proper subset of $\Pi_{\tilde{t}}$ or not. In fact, most of the elements of $\Pi_{\tilde{t}}$ which have been constructed in [1] actually belong to $\mathscr{K}_{\tilde{t}}$. That this statement holds true for

$$f\big(\tilde{t}; \tilde{x}(\tilde{t}), u\big) - f\big(\tilde{t}; \tilde{x}(\tilde{t}), \tilde{u}(\tilde{t})\big), \qquad u \in U,$$

can be seen by an argument of the same type as that used in [1] in order to verify (9.8). Further and more interesting examples will be discussed in Section 3. Though the definition of $\Pi_{\tilde{t}}$ appears to be simpler than that of $\mathscr{K}_{\tilde{t}}$ one will probably, in a concrete situation, find it equally difficult to verify the conditions of either of them. On the other hand, working

with $\mathscr{K}_{\tilde{t}}$ is more convenient because of certain a-priori-statements, which can be made and which are independent both from the specific nature of the system equation (1.1) and from the choice of the reference solution. The most important ones are summarized in the following theorem.

THEOREM 2.1. (i) *$\mathscr{K}_{\tilde{t}}$ is a convex cone.* (ii) *Let $\pm p \in \mathscr{K}_{\tilde{t}}$ and $p(\cdot)$ be a C^∞-function which is associated with a family of control functions according to* (2.4) *and for which $p = p(\tilde{t})$. Then we have*

$$\pm\{-\dot{p}(\tilde{t}) + f_x(\tilde{t}; \tilde{x}(\tilde{t}), \tilde{u}(\tilde{t}))p(\tilde{t})\} \in \mathscr{K}_{\tilde{t}}. \tag{2.6}$$

One can observe that the second assertion of the theorem can be stated in the following way: The maximal linear subspace of $\mathscr{K}_{\tilde{t}}$ is invariant with respect to the mapping

$$p \to -\dot{p} + f_x(\tilde{t}; \tilde{x}(\tilde{t}), \tilde{u}(\tilde{t}))p. \tag{2.7}$$

If the system equation is linear and given by (1.2) then the mapping assumes the form

$$p \to -\dot{p} + A(\tilde{t})p.$$

This is nothing else than the operation which can be applied in order to generate the controllable subspace out of the columns of the matrix $B(t)$.

If $p(t)$ is of the form $\hat{p}(t, \tilde{x}(t))$, where $\hat{p}$ is a sufficiently smooth function of t, x, then

$$\dot{p} = \partial\hat{p}/\partial t + \hat{p}_x f$$

and (2.7) can be written in the form

$$\hat{p} \to -\partial\hat{p}/\partial t - [f, \hat{p}],$$

where the expressions appearing in this formula have to be evaluated at $x = \tilde{x}(\tilde{t})$, $t = \tilde{t}$. Hence the mapping (2.7) is in fact nothing else than the application of the operator Γ, as introduced in [1].

We conclude this section by stating the two fundamental theorems about $\mathscr{K}_t$ which were announced in the introduction. The proof of the second one follows immediately, in view of (2.5), from Theorem 9.1 in [1].

THEOREM 2.2. *If $\mathscr{K}_{\tilde{t}} = \mathbf{R}^n$ then we have local controllability along the reference solution and for $t = \tilde{t}$.*

THEOREM 2.3. *Let the reference solution be optimal. Then there exists an adjoint state vector $y(\cdot)$ which satisfies the transversality conditions at the endpoints and the inequalities $y(t)^T p \leqslant 0$ for all $p \in \mathscr{K}_t$ and all $t \in [t_0, \tilde{t}]$.*

3. Second order elements

Special attention is deserved by those elements in $\mathscr{K}_t$ which lead to second order necessary conditions. This notion was explained in [1] (Section 1, p. 5); the definition can easily be modified so as to make sense if one works with $\mathscr{K}_t$ instead of Π_t. We consider a family of control functions as specified in Section 2 and we assume in addition that $u(t;\tau,\lambda)$ can be written as

$$u(t)+\lambda^r v(t;\tau,\lambda) \tag{3.1}$$

where r is some positive integer. $v(t;\tau,\lambda)$ is supposed to vanish for $t\leqslant\tau-\lambda^r$ in order to meet the requirement (iii). It follows then from property (iv) by standard arguments that one can expand $x\big(t;t,\lambda,\tilde{w}(t_0)\big)$ into an asymptotic series of the form

$$\tilde{w}(t)+\lambda^r\sum_{\nu=1}^{\infty}\lambda^\nu p_\nu(t),$$

where the coefficients $p_\nu(\cdot)$ are functions of t infinitely differentiable on some interval $[\tilde{t}-\varepsilon,\tilde{t}]$. We pick the first $p_\nu(\cdot)$ which does not vanish identically and call it $p(\cdot)$. Clearly, $p(t)\in\mathscr{K}_t$ for all $t\in(t-\tilde{\varepsilon},\tilde{t}]$. If $p(\cdot)=p_\nu(\cdot)$ and $r<\nu\leqslant 2r$ then $p(\cdot)$ is called a *second order element*. Accordingly, one would call $p(\cdot)$ a *k-th order element* if $(k-1)r<\nu\leqslant kr$.

A survey of second order necessary conditions can be found in [1], Sections 20, 21. Each condition is stated there as a multiplier rule $y(t)^T p(t)\leqslant 0$ (or $y(t)^T p(t)=0$) and holds under the hypothesis that a relation of the form

$$q(t)\in\mathfrak{B}(t) \tag{3.2}$$

is satisfied for all t in some interval $[\tilde{t}-\varepsilon,\tilde{t}]$. $q(\cdot)$ is of course a vector, which does not trivially belong to the controllable subspace. A natural question arises whether the element which appears in the rule as factor along with y is actually second order (in the sense explained above) and whether this statement holds true irrespective of the optimal character of the reference solution. As can be seen by a thorough analysis of the proofs given in [1] the answer will always be affirmative if, at least, (3.2) is replaced by a slightly stronger hypothesis. This modification, however, is unnecessary if the space $\mathfrak{B}(t)$ has constant dimension. For simplicity we will therefore work for the remaining portion of this section with the following additional hypothesis:

The dimension of the space $\mathfrak{B}(t)$ is independent from t,
for all $t\in[\tilde{t}-\varepsilon,\tilde{t}]$. (3.3)

We then restate here two of the second order conditions which have been proved in [1] (Theorem 20.2 and 21.2) for a general system of the form (1.1). The first is commonly known as the *generalized Clebsch–Legendre condition*. It is a non-trivial result, regardless of whether u is scalar or not, and we will restrict ourselves to the case of a scalar control. The second one is the prototype of an equality type necessary condition, and hence is of interest only if u is not scalar as we have remarked in the introduction. Therefore we will here assume that $u = (u^1, u^2)^T$ is 2-dimensional.

As before we denote by $(\tilde{u}(\cdot), \tilde{x}(\cdot))$ the given reference solution and assume that $\tilde{u}(\cdot)$ satisfies the condition $\tilde{u}(t) \in \operatorname{int} U$ for all t. We associate with this solution a sequence of vectors B_ν^i, $\nu = 0, 1, \ldots, i = 1, \ldots, m$ (= dimension of u), which are recursively defined as follows

$$\tilde{f}(t; x) := f(t; x, \tilde{u}(t)), \qquad B_0^i(t, x) := (\partial f/\partial u^i)(t; x, \tilde{u}(t))$$

$$B_{\nu+1}^i(t, x) := (\partial B_\nu^i/\partial t)(t, x) + [\tilde{f}, B_\nu^i](t, x), \qquad \nu = 0, 1, \ldots$$

In case of $m = 1$ we write B_ν instead of B_ν^1.

THEOREM 3.1. *Hypotheses*: (i) *u is scalar*, (3.3) *holds true.* (ii) *We have*

$$(\partial^2 f/\partial u^2)(t; \tilde{x}(t), \tilde{u}(t)) \in \mathfrak{B}(t), \qquad [B_{\nu-1} B_\nu](t, \tilde{x}(t)) \in \mathfrak{B}(t)$$

for all $t \in [\tilde{t} - \varepsilon, \tilde{t}]$ and $\nu = 1, \ldots, \varrho \geqslant 0$.
Conclusion: $-[B_\varrho, B_{\varrho+1}](t, \tilde{x}(t)) \in \mathscr{K}_t$.

THEOREM 3.2. *Hypotheses*: (i) *$u = (u^1, u^2)^T$ is 2-dimensional*, (3.3) *holds.* (ii) *We have*

$$(\partial^2 f/\partial (u^i)^2)(t; \tilde{x}(t), \tilde{u}(t)) \in \mathfrak{B}(t), \qquad [B_{\nu-1}^i, B_\nu^i](t, \tilde{x}(t)) \in \mathfrak{B}(t)$$

for $t \in [\tilde{t} - \varepsilon, \tilde{t}]$, $\nu = 1, \ldots, \varrho_i \geqslant 0, i = 1, 2$.
Conclusion: $\pm [B_\nu^1, B_\nu^2](t, \tilde{x}(t)) \in \mathscr{K}_t$ if $\nu + \mu \leqslant \varrho_1 + \varrho_2$.

References

[1] Knobloch H. W., Higher Order Necessary Conditions in Optimal Control Theory, *Lecture Notes in Control and Information Sciences*, Vol. **34**, Springer-Verlag, Berlin, Heidelberg, New York, 1981.

[2] Sussmann H. J., Lie Brackets, Real Analyticity and Geometric Control, in: *Differential Geometric Control Theory*, R. W. Brockett, R. S. Millman, H. J. Sussmann, Eds., Progress in Mathematics, Vol. **27**, Birkhäuser, Boston, Basel, Stuttgart, 1983.

[3] Sussmann H. J., Lie brackets and local controllability: A sufficient condition for scalar-input systems, to appear in *SIAM J. on Control and Optimization*.

[4] Vârsan C., On Local Controllability for Non-linear Control Systems, *Preprint Series in Mathematics*, Nr. 115/1981, Bucuresti, 1981.

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

A. B. KURZHANSKIJ

Evolution Equations for Problems of Control and Estimation of Uncertain Systems

The present report deals with problems of control and estimation for systems described by ordinary differential equations or differential inclusions ([8], [10], [12], [22]). It is assumed that these equations are related to systems where the model parameters, the initial vectors, and the disturbances are unknown, a set-membership description of their values being only given in advance. The systems are therefore said to *operate under conditions of uncertainty.*

A considerable number of problems concerning systems of the above type are covered by the theory of differential games and related branches of control theory, [9], [13], [19].

The main problems of this report deal with the investigation of evolution equations that describe guaranteed or minmax estimates for the states of the system on the basis of available observations under uncertainty of the above type. Relations between the given results and facts of stochastic filtering theory are formulated, the specific features of linear systems being especially underlined. Duality relations between solutions of control and observation problems under uncertainty are shown to be a reflection of duality in the theory of extremal problems.

A further discussion concerns the construction of "adaptive strategies" of control that combine the optimization of the evolution for the estimation process with the selection of control values that ensure a guaranteed result.

The report is based mainly on the investigations carried through in the Institute of Mathematics and Mechanics of the Ural Scientific Center of the Academy of Sciences of the USSR at Sverdlovsk.

Notations

The following notations are assumed: R^n is the n-dimensional real vector space, $(l, p) = l'p$ is the scalar product in R^n, the prime standing for the transpose, T is a closed time interval $[t_0, t_1]$, $\mathscr{B}$ is a Banach space, $\mathscr{L}_p^{(m)}(T)$ is the space of functions integrable with power p on the interval T $(1 \leqslant p \leqslant \infty)$, $C^{(m)}(T)$ is the space of m-vector-valued continuous functions defined on T, $\|\cdot\|$ is the norm in the Banach space $\mathscr{B}$, $|\cdot|$ is the Euclidean norm in R^n, $\Omega(\mathscr{B})$ is the set of all subsets of $\mathscr{B}$, comp $\mathscr{B}$ is the set all compact subsets of $\mathscr{B}$, $\langle v(\cdot), \lambda(\cdot)\rangle$ is the value of the functional generated by $v(\cdot) \in \mathscr{B}^*$ at an element $\lambda(\cdot) \in \mathscr{B}$. For $\mathscr{B} = \mathscr{L}_2^{(p)}(T)$ we have, in particular,

$$\langle v(\cdot), \lambda(\cdot)\rangle = \int_{t_0}^{t_1} v'(t)\lambda(t)\,dt.$$

The symbol $\varphi(t, Q)$ stands for

$$\varphi(t, Q) = \bigcup \{\varphi(t, q) |\ q \in Q\}$$

and $\delta(v\,|Q)$ denotes the indicator function of a set Q at a point v. Namely

$$\delta(v|Q) = 0 \text{ if } v \in Q, \qquad \delta(v|Q) = \infty \text{ if } v \notin Q.$$

The symbol $\varrho(l|Q)$ denotes the support function of a set Q at a point l, i.e.,

$$\varrho(l|Q) = \sup\{l'q |\ q \in Q\}$$

and $\partial\varphi(l)$ is the subdifferential of a convex function φ at a point l; co$Q$ is the convex hull of a set Q and int$Q$ is the set of all interior points of Q.

The following notations are also assumed; the element $y_t(\cdot\,|\tau)$ is generated by the function $y(t+\sigma)$ where $\tau - t \leqslant \sigma \leqslant 0$ and $y_t(\cdot\,|t_0) = y_t(\cdot)$, $y_{t_1}(\cdot\,|t_0) = y(\cdot)$; $Q_t(\cdot\,|t_0)$ is the "tube" generated by the multivalued function $Q(t+\sigma)$, where $t_0 - t \leqslant \sigma \leqslant 0$, and $Q_t(\cdot\,|t_0) = Q_t(\cdot)$, $Q_{t_1}(\cdot\,|t_0) = Q(\cdot)$; $S_\delta(x) = \{z\colon |z-x| \leqslant \delta\}$ is the Euclidean ball of radius δ with center x, E_m is the unit $(m \times m)$-matrix, $K \dot{-} Q = \{p\colon Q + p \subset K\}$ is the geometrical difference of sets K, Q; $A \oplus B$ is the Kronecker product of matrices A, B, $\overline{A} = \{a_s\}$ is the nm-vector formed of the elements of $A = \{a_{ij}\}$, the entries of $\overline{A}$ being given by $a_{i+(j-1)n} = a_{ij}$.

1. Evolution systems for problems of control and estimation

Consider the differential inclusion

$$\frac{dx}{dt} \in F(t, x, u), \qquad x(t_0) \in X^0, \quad t \in T, \tag{1.1}$$

where $F(t, x, u)$ is a multivalued mapping from $T \times R^n \times R^p$ into $\operatorname{comp} R^n$. Here x is the phase vector, t stands for time, u is the control parameter. Equation (1.1) describes a system that operates under uncertainty, the latter circumstance being caused by the fact that F is a multivalued map.

The function $F(t, x, u)$ may be e.g. of the form

$$F(t, x, u) = \bigcup \{f(t, x, u) + \varphi(t, v) \mid v \in Q(t)\}, \quad Q(t) \in \operatorname{comp} R^q. \tag{1.2}$$

In particular, we may have

$$F(t, x, u) = A\big(t, Q(t)\big)x + B(t)u, \tag{1.3}$$

(describing a linear system with uncertain coefficients) or

$$F(t, x, u) = A(t)x + B(t)u + \varphi\big(t, Q(t)\big) \tag{1.4}$$

(a proper linear system).

The functions $f(t, x, u, v)$, $A(t, v)$, $A(t)$, $\varphi(t, v)$ are here assumed to be known, measurable in t and continuous in the other variables, the multivalued function $Q(t)$ is assumed to be measurable ([4], [24b]).

Along with (1.1) consider the equation of "observations" or "measurements"

$$y \in G(t, x, u), \quad y \in R^m, \tag{1.5}$$

where $G(t, x, u)$ is a multivalued mapping from $T \times R^n \times R^p$ into $\operatorname{comp} R^m$, and the element $y(\cdot)$ generated by an available measurement $y(t)$ is considered as an element of $\mathscr{L}_2^{(m)}(T)$.

The multivalued function $G(t, x, u)$ describes the structure of the measurement process and the admissible uncertain disturbances. In the linear case (1.4) turns into the equation

$$y \in g(t, u)x + \mathscr{R}(t). \tag{1.6}$$

We will further assume that the measurable function $u(t)$ takes its values in compact sets $\mathscr{P}(t) \subseteq R^p$. Assume that the function $u(t)$ is given and let $y(t)$ be a realization of the measured signal. The symbol $X\big(\cdot, y(\cdot), u(\cdot), t_0, X^0\big)$ $\big(X_t\big(\cdot, y_t(\cdot), t_0, u_t(\cdot), X^0\big)\big)$ denotes the set of all trajectories, $x(\tau), \tau \in T$ $(\tau \in [t_0, t])$ that satisfy (1.1) and (1.5) simultaneously. The section $X(t, \cdot) = X\big(t, y_t(\cdot), u_t(\cdot), t_0, X^0\big) \subseteq \Omega(R^n)$ of the set $X_t(\cdot, y_t(\cdot), u_t(\cdot), t_0, X^0)$ taken at time instant t is said to be the *informational set* for system (1.1), (1.5) consistent with the realization $y_t(\cdot)$ (the control $u_t(\cdot)$ being given). The symbol $Y\big(u(\cdot), X^0\big)$ $\big(Y_t\big(u_t(\cdot), X^0\big)\big)$ will denote

the set of all functions $y(\cdot) \in \mathscr{L}_2^{(m)}(T)$ (all functions $y_t(\cdot)$) such that $X(\cdot, y(\cdot), u(\cdot), t_0, X^0)$ ($X(t, y_t(\cdot), u_t(\cdot), t_0, X^0)$) is nonvoid. Further, it is everywhere assumed that $y(\cdot) \in Y(u(\cdot), X^0)$. Therefore, $Y(u(\cdot), X^0)$ is always nonvoid.

We shall state a few simplest conditions which ensure that $X(\cdot, y(\cdot), u(\cdot), t_0, X^0)$ is compact and semicontinuous in $y(\cdot)$ and $u(\cdot)$.

THEOREM 1.1. *Assume that X^0 is a compact set in R^n, the functions $y(\cdot) \in Y(u(\cdot), X^0)$, $u(\cdot) \in \mathscr{P}(\cdot)$ are given and the following conditions are fulfilled*

(a) *the mappings $F(t, x, u)$, $G(t, x, u)$ are convex-valued, upper-semicontinuous in x, measurable in t and continuous in u,*

(b) *there exists a nonnegative function $g(\cdot) \in \mathscr{L}_1(T)$ such that for all t, x*

$$F(t, x, u) \subseteq g(t)(1 + |x|)\mathscr{S} \qquad (\mathscr{S} \text{ denoting the unit ball in } R^n),$$

uniformly in $u(\cdot) \in \mathscr{P}(\cdot)$.

Then the set $\{x(\cdot)\}$ of elements of $X(\cdot, y(\cdot), u(\cdot), t_0, X^0)$ is compact in $C^{(n)}(T)$.

This theorem admits a relaxed form, particularly with regard to condition (b) (similar to conditions given in [4] for ordinary differential inclusions).

THEOREM 1.2. *Under conditions of Theorem* 1.1, *for $u(\cdot)$, X^0 fixed, the mapping $\mathscr{X}(\cdot, y(\cdot)) = X(\cdot, y(\cdot), u(\cdot), t_0, X^0)$ from $Y(u(\cdot), X^0)$ into* $\operatorname{comp} C^{(n)}(T)$ *is upper-semicontinuous in $y(\cdot)$.*

When $F(t, x, u), G(t, x, u)$ are Lipschitzian in u and under some additional constraints on the dependence of $F(t, x, u), G(t, x, u)$ on t, the sets $X(\cdot, y(\cdot), u(\cdot), t_0, X^0)$ depend continuously on $u(\cdot)$. Here $\operatorname{comp} C^{(n)}(T)$ is considered with the Hausdorff metric and $Y(u(\cdot), X^0), \mathscr{P}(\cdot)$ with the metrics of spaces $\mathscr{L}_2^{(m)}(T), \mathscr{L}_2^{(p)}(T)$.

Let $X(t, y_t(\cdot|\tau), u_t(\cdot|\tau), \tau, X)$ be the set of solutions of (1.1), (1.5) at time t (for given $y_t(\cdot|\tau), u_t(\cdot|\tau), x(\tau) \in X$).

THEOREM 1.3. *For $y_t(\cdot), u_t(\cdot)$ given, the following equality is true:*

$$X(t, y_t(\cdot|t_0), u_t(\cdot|t_0), t_0, X^0) = X\big(t, y_t(\cdot|\tau), u_\tau(\cdot|\tau), \tau, X(\tau, y_\tau(\cdot|t_0), u_\tau(\cdot|t_0), t_0, X^0)\big).$$

Therefore, $y_t(\cdot|\tau)\, u_t(\cdot, |\tau), X^0$ given, the mapping $X(t, \cdot) = X(t, y_t(\cdot), u_t(\cdot), X^0)$ defines a generalized dynamical system which describes the

evolution of the control and estimation process. The sets $X(t, \cdot)$ contain the whole prehistory of the process. They therefore have Markov-type properties.

Define $X[t] = X(t, \cdot)$ to be the realization of the "bundle" $X(t, \cdot)$.

LEMMA 1.1. *The multivalued function $X[t]$ from T to* comp R^n *is upper semicontinuous in t.*

There is some interest in determining conditions for $X[t]$ to be convex. The simplest of these require that the support functions $\varrho(l|F(t, x, u))$, $\varrho(p|G(t, x, u))$ be concave in x for any $l \in R^n$, $p \in R^m$, $u \in \mathscr{P}(t)$, $t \in T$ (under conditions of Theorem 1.1). This yields

LEMMA 1.2. *Let $F(t, x, u)$, $G(t, x, u)$ be as in* (1.4), (1.6), $\varphi(t, v) = C(t)v$ *and suppose that the sets $Q(\cdot)$, $\mathscr{R}(\cdot)$ are convex. Then $X[t]$ are convex.*

A less trivial convexity condition for $X[t]$ may be obtained on the basis of conditions introduced in [18], [21]. We note that even equations as simple as (1.3), (1.6) can yield nonconvex sets $X[t]$.

2. Linear systems

We will specify some properties of the sets $X(t, y_t(\cdot), 0, t_0, X^0) = X(t, y_t(\cdot), t_0, X^0)$ occurring in equations (1.4), (1.6) ($u(\cdot)$ being given, we may take $u(\cdot) \equiv 0$ with no loss of generality). Further, we assume $\varphi(t, q) = C(t)q$ and $Q(t)$, $\mathscr{R}(t)$ to be convex and compact-valued.

LEMMA 2.1. *In order that the convex sets $X(t, y_t(\cdot), t_0, R^n)$ be bounded it is necessary and sufficient for the form $\Phi_0(l) = l'W(t, t_0)l$ to be positive-definite. Here*

$$W(t, t_0) = \int_{t_0}^{t} S'(t, \tau)G'(\tau)G(\tau)S(t, \tau)\,d\tau,$$

where $S(t, \tau)$ is the matrix solution of the system $S' = -SA(t)$, $S(t, t) = E_n$.

The positive definiteness of $\Phi_0(l)$ is equivalent to the requirement for the homogeneous system (1.4), (1.6) to be completely observable. For finite-dimensional linear systems the latter property is equivalent to the requirement for $X(t, y_t(\cdot), t_0, X^0)$ to be bounded for $X^0 = R^n$, with $Q(t)$, $\mathscr{R}(t)$ being bounded uniformly in t.

It may be useful to describe the "best" and the "worst" signals $y(\cdot) \in Y(0, R^n)$ from the point of view of the observer.

The "worst" signal for the observer is the function $y^0(\cdot) \in Y(0, R^n)$ for which

$$d\big(X(t_1, y^0(\cdot), t_0, X^0)\big) = \max\big\{d\big(X(t_1, y(\cdot), t_0, X^0)\big) \mid y(\cdot) \in Y(0, R^n)\big\}.$$

Here $d(X) = \max\big\{\frac{1}{2}\big(\varrho(l|X) + \varrho(-l|X)\big) \mid |l| = 1\big\}$ is the diameter of set X.

LEMMA 2.2. *Assume that the sets $Q(t), \mathscr{R}(t)$ are symmetrical about the origin and let $X^0 = R^n$. Then $y^0(\cdot) = 0$.*

Therefore, under the conditions of this lemma, the worst triplets $\zeta(\cdot) = \{x^0, v(\cdot), \xi(\cdot)\}$ for the observer are those which generate the solution of (1.4), (1.6) due to the system

$$\dot{x} = A(t)x + C(t)v, \quad y = G(t)x + \xi, \quad x(t_0) = x^0 \tag{2.1}$$

for the signal $y(t) \equiv 0$.

It is more difficult to define the triplets $\{x^0, v(\cdot), \xi(\cdot)\}$ (if they exist) for which $X^*\big(t_1, y_{t_1}(\cdot), t_0, X^0\big)$ is a one-element set ($dX^* = 0$).

Consider the following example of necessary and sufficient conditions that ensure the property just stated. Take $s\big(t, t_1, \lambda(\cdot), l\big)$ to be the solution of the system

$$\dot{s} = -sA(t) + \lambda'(t)G(t), \quad s(t_1) = l'. \tag{2.2}$$

Here $\lambda(\cdot) \in \mathscr{D}^{(m)}(\cdot)$, where $\mathscr{D}^{(m)}(\cdot)$ consists of all functions $\lambda(\cdot)$ which are the generalized derivatives $\lambda(t) = d\Lambda(t)/dt$ of functions $\Lambda(\cdot) \in V^{(m)}(T)$, where $V^{(m)}(T)$ is the set of m-dimensional functions of bounded variation, concentrated on T. Also assume that $\{0\} \in \operatorname{int} X^0$.

LEMMA 2.3. *In order that for any $x^0 \in X^0$ the pair*

$$\{v^*(t), \xi^*(t)\} \quad \big(v^*(\cdot) \in \mathscr{L}_2^{(q)}(T), \ \xi^*(\cdot) \in C^{(m)}(T)\big)$$

should generate a realization $y^*(\cdot)$ for which the set $X\big(t_1, y^*(\cdot), 0, t_0, X^0\big)$ is a singleton, it is necessary and sufficient that there exists a collection of nonnull vectors $l^{(i)}$, $(i = 1, \ldots, n)$ with

$$\sum_{i=1}^{n+1} \alpha_i l^{(i)} = 0, \quad \alpha_i \geqslant 0, \quad \sum_{i=1}^{n+1} \alpha_i \neq 0,$$

and such that for every $l^{(i)}$ *the following problem be soluble in the class* $\lambda(\cdot) \in \mathscr{D}^{(m)}(\cdot)$:

$$s(t_0, t_1, \lambda(\cdot), 0) = l^{(i)\prime},$$

$$\langle s(\cdot, t_1, \lambda(\cdot), l), v^*(\cdot)\rangle = \max\{\langle s(\cdot, t_1, \lambda(\cdot), l), v(\cdot)\rangle \mid v(\cdot) \in Q(\cdot)\},$$

$$\langle \lambda(\cdot), \xi^*(\cdot)\rangle = \min\{\langle \lambda(\cdot), \xi(\cdot)\rangle \mid \xi(\cdot) \in \mathscr{R}(\cdot)\}.$$

It is possible to indicate some sufficient conditions for the assumptions of Lemma 2.3 to be fulfilled provided that the system (2.1) is autonomous and $Q(t) \equiv Q$, $\mathscr{R}(t) \equiv \mathscr{R}$ ([20]).

We will say that the function $v(t)$ has a *complete oscillation on the interval* $[t', t'']$ *in the direction* l if, simultaneously,

$$\operatorname{vrai}\max_{t\in[t't'']} l'v(t) = \max\{l'v \mid v \in Q\}$$

and

$$\operatorname{vrai}\min_{t\in[t',t'']} l'v(t) = \min\{l'v \mid v \in Q\}.$$

The oscillations of the function $\xi(t)$ are defined similarly.

We will say that the functions $v(t)$ ($\xi(t)$) have N complete oscillations in the direction l on the interval T if it is possible to indicate N pairwise nonintersecting closed intervals of complete oscillation in that direction.

LEMMA 2.4. *Assume that the system* (2.1) *is autonomous and that either* $m = n$ *or* $C = 0$. *Then for any interval* T *there exists an integer* N *such that, if each of the pairs* $v^*(\cdot)$, $\xi^*(\cdot)$ *has* N *complete oscillations in each of the respective directions* $e_n^{(i)} \in R^n$; $i = 1, \ldots, n$ *(for* v^**)*; $e_m^{(j)} \in R^m$, $j = 1, \ldots, m$ *(for* ξ^**)*, *where* $e_n^{(i)}$, $e_m^{(j)}$ *are the elements of the standard basis in* R^n *and* R^m, *respectively, then the conditions of Lemma* 2.3 *are fulfilled. If the matrix* A *has only real eigenvalues, one may take* $N = n$, *where* n *is the order of the system.*

Therefore the conditions of Lemma 2.3 are fulfilled whenever the functions $v(t)$, $\xi(t)$ are oscillating in some special way between the extreme points of the restriction sets Q, $\mathscr{R}$. We mention that the asymptotic condition $X[t] \to X^*(t \to \infty)$, where X^* is a one-element set, is fulfilled if the realizations $v^*(t)$, $\xi^*(t)$ are generated, for example, by stochastic processes having certain ergodicity properties and if the original system posseses certain periodicity properties ([14f], [23]).

3. Deterministic (guaranteed) estimation and stochastic estimation

The description of the evolution of sets $X(t, \cdot)$ leads to the solution of a deterministic estimation problem. Indeed, knowing $X(t, \cdot)$, we may find a guaranteed minmax estimate $x^0(t)$ of the unknown actual state of the space vector $x^*(t)$, taking $x^*(t)$ to be the "Chebyshev center" of the set $X(t, \cdot)$. Namely, $x^0(t)$ is determined from the condition

$$\varepsilon^0(t, \cdot) = \max\{\|x^0(t) - x\| \mid x \in X(t, \cdot)\}$$
$$= \min_z \max_x \{\|z - x\| \mid z \in X(t, \cdot),\ x \in X(t, \cdot)\}$$

where $\varepsilon^0(t, \cdot)$ is the guaranteed error of the estimation process. The computation of $x^0(t)$, $\varepsilon^0(t, \cdot)$ leads to a special problem of mathematical programming.

Passing to the evolution of sets $X(t, \cdot)$ for linear systems, we may determine the support function $\varrho(l \mid X(t, \cdot))$ to be of the form

$$\varrho(l \mid X(t, \cdot)) = \inf_{\lambda(\cdot) \in \mathcal{Q}^{(m)}(\cdot)} \Big\{\varrho\big(s(t, t_0, \lambda(\cdot), l) \mid X^0\big) + \int_{t_0}^{t_1} \big\{\varrho\big(s(\tau, t_0, \lambda(\cdot), l) \mid C(t)Q(t) + B(t)u(t)\big) + \varrho\big(-\lambda(t) \mid \mathcal{R}(t) - y(t)\big)\big\}\, dt\Big\}. \quad (3.1)$$

The investigation of this function leads to the following conclusion. Consider the set $Z\big(t, M(\cdot), X^0\big)$ of all solutions of the differential inclusion $(A^0(t) = A(t) - M'(t)G(t),\ C^0(t) = M'(t))$

$$\frac{dz}{dt} \in A^0(t)z + C^0(t)\big(y(t) - F(t)\mathcal{R}(t)\big) + B(t)u + C(t)Q(t) \quad (3.2)$$

generated by all possible initial vectors $z(t_0) \in X^0$.

LEMMA 3.1. *The following equalities are true*

$$\bigcap \big\{Z\big(t, M(\cdot), X^0\big) \mid M(\cdot)\big\} = X\big(t, y_t(\cdot), u_t(\cdot), X^0\big) = X(t, \cdot),$$
$$\varrho\big(l \mid X(t, \cdot)\big) = \inf\big\{\varrho\big(l \mid Z\big(t, M(\cdot), X^0\big) \mid M(\cdot)\big)\big\}$$

where the matrix $M(t)$ may have no more than m independent scalar functions $m_{ij}(t) \not\equiv 0$ for its entries $(M(t) = \{m_{ij}(t)\})$.

Therefore equation (3.2) carries complete information on the evolution of the sets $X(t, \cdot)$. It is interesting to mention that the sets $X(t, \cdot)$ may also be described by means of equations related to the solution of the linear-quadratic Gaussian estimation problem (the Kalman filtering problem), ([8], [10]). Namely, the following result is true. Consider inclusion (3.2), where

$$A^0(t) = A(t) - P(t)G'(t)N(t)G(t), \quad C^0(t) = P(t)G'(t)N(t), \tag{3.3}$$

$$\dot{P}(t) = A'(t)P(t) + P(t)A(t) - P(t)G'(t)N^{-1}(t)G(t)P(t) + M(t). \tag{3.4}$$

Denote the set of all solutions of this inclusion that start at $x^0 \in X^0$ by $Z\big(t, D(\cdot), X^0\big)$, where $D(\cdot) = \{L, M(\cdot), N(\cdot)\}$. Then (see [14c], [14d]).

THEOREM 3.1. *The following equalities are true:*

$$\begin{aligned} X(t, \cdot) &= \bigcap \big\{Z\big(t, D(\cdot), X^0\big) \,|\, D(\cdot) \in D_0\big\}, \\ \varrho\big(l \,|\, X(t, \cdot)\big) &= \inf\big\{\varrho\big(l \,|\, Z\big(t, D(\cdot), X^0\big)\big) \,|\, D(\cdot) \in D_1\big\}. \end{aligned} \tag{3.5}$$

Here D_0 is the set of triples $D(\cdot)$ such that $L = \alpha E_n$, $M(t) = \beta(t)E_p$, $N(t) = \gamma(t)E_m$, where the scalar parameters satisfy $\alpha > 0$, $\gamma(t) > 0$, $\beta(t) \geqslant 0$, D_1 is the set of triples $D(\cdot)$ such that the matrices $M(t) \geqslant 0$, $N(t) > 0$ jointly depend at each instant t on no more than m independent scalar parameters, with $L > 0$ being arbitrary.

In the case where $Q(t) = \{0\}$, the sets D_0, D_1 may be replaced by D_0^*, D_1^* where

$$D_0^* = \{\alpha E_n, \gamma(\cdot)E_m\}, \quad D_1^* = \{L, 0, N(\cdot)\}.$$

The above theorem admits the following probabilistic interpretation Consider the system

$$\begin{aligned} dz &= \big(A(t)z + C(t)v(t) + B(t)u(t)\big)dt + \sigma_1(t)\,dw, \\ dp &= \big(G(t)z + \xi(t)\big)dt + \sigma(t)\,d\eta, \end{aligned} \tag{3.6}$$

where $v(t)$, $\xi(t)$ are unknown deterministic functions bounded by convex sets $Q(t)$ and $\mathscr{R}(t)$, respectively; dw, $d\eta$ are the standard normed processes of Brownian motion, $\sigma(t)$, $\sigma_1(t)$ are given continuous matrix coefficients of diffusion with $\sigma_1(t)$ nondegenerate, $z(t_0)$ is a Gaussian vector with mean value $\bar{z}_0$, $p(t_0) = 0$. Having $v(t)$, $\xi(t)$ fixed, we may find the conditional mean value $\bar{z}[t] = \bar{z}\big(t, p_t(\cdot), v_t(\cdot), \xi_t(\cdot)\big)$ of the process $z(t)$ with given

measurement $p_t(\cdot)$. It is well known ([8], [10], [16]) that $\bar{z}[t]$ satisfies the system

$$d\bar{z} = A^0(t)\bar{z}\,dt + C^0(t)\big(dp - \xi(t)\,dt\big) + \big(B(t)u(t) + C(t)v(t)\big)\,dt,$$

where $\bar{z}(t_0) = \bar{z}^0$, $A^0(t)$, $C^0(t)$ are defined as in (3.3), (3.4), and $M(t) = \sigma_1(t)\sigma_1'(t)$, $N(t) = \sigma(t)\sigma'(t)$. Let

$$\bar{Z}[t, D(\cdot), X^0] = \Big\{\bigcup \bar{z}\big(t, p_t(\cdot), v_t(\cdot), \xi_t(\cdot), z^0\big) \mid v(\cdot) \in Q(\cdot),\ \xi(\cdot) \in \mathcal{R}(\cdot),\ z^0 \in X^0\Big\}$$

be the set of all conditional mean values of the vector z with uncertain mean values $C(t)v(t)$, $\xi(t)$ of the disturbances

$$C(t)v(t) + \sigma(t)\,dw(t)/dt, \qquad \xi(t) + \sigma_1(t)\,d\eta(t)/dt.$$

Then, under the same restrictions on $v(\cdot)$, $\xi(\cdot)$ in both (2.1) and (3.6) and with $\dot{p}(t) \equiv y(t)$, the sets $\bar{Z}[t, D(\cdot), p_t(\cdot)]$ and $Z\big(t, D(\cdot), X^0\big)$ coincide and Theorem 3.1 throws over a bridge between the solutions of stochastic Kalman filtering and deterministic game-theoretic filtering.

The proof of the latter theorem is based on the fact that the Lagrange multiplier $\lambda(\cdot)$ which optimizes the right-hand side of (3.1) may be looked for in the form of an aggregate that may be represented by matrices L, $M(\cdot)$, $N(\cdot)$ related to a special auxiliary problem of statistical Kalman filtering. Therefore the problem of minimizing (3.1) by $\lambda(\cdot)$ is now substituted by conditions of type (3.5), where the minimum is sought in the class of covariance matrices for the auxiliary problem. The inclusions (3.2), (3.4) may thus be treated as universal equations that contain in particular the solutions of stochastic filtering problems (for $Q(\cdot) = \{0\}$, $\mathcal{R}(\cdot) = \{0\}$), the solutions of guaranteed estimation problems (according to Theorem 3.1) and the solution of the filtering problem for a system with statistically uncertain disturbances (3.6).

The latter problem allows a generalization directed toward the estimation of systems of type (3.6) with incomplete information on the matrices $\sigma(t)$, $\sigma_1(t)$ and L.

Theorem 3.1 gives a precise description of the sets $X(t, \cdot)$. Various approximate descriptions of $X(t, \cdot)$ by means of appropriate solutions with quadratic integral constraints, including the best approximation of $X(t, \cdot)$ by ellipsoids are discussed in [3], [5], [14a], [25]. Finite difference schemes related to the problems under investigation are given in [11].

4. Duality relations for problems of control and estimation

The calculation of $\varrho\big(l\,|\,X(t,\cdot)\big)$ admits a natural interpretation in terms of the primal and dual extremal problems. Consider the case of $u(\cdot)\equiv 0$, $X^0 = R^n$. Denoting the solution of equation (2.1) by $x\big(t, t_0, v(\cdot), x^0\big)$, consider the "primal problem" (P):

determine $\max\big(l, x\big(t_1, t_0, v(\cdot), x^0\big)\big)$ over all solutions $x\big(t, t_0, v(\cdot), x^0\big)$ consistent with equations (2.1) with $v(\cdot)\in Q(\cdot)$, $\xi(\cdot)\in\mathscr{R}(\cdot)$, $X^0 = R^n$ and with given $y(\cdot)$.

The problem just stated may be formulated in the form:

determine

$$-\mathscr{J}^0 = \min\mathscr{J}\big(x^0, v(\cdot)\big) \quad \text{over} \quad x^0\in R^n,\ v(\cdot)\in\mathscr{L}_2^{(q)}(T) \tag{4.1}$$

under the conditions

$$\mathscr{J}\big(x^0, v(\cdot)\big) = \big(-l, x\big(t_1, t_0 v(\cdot), x^0\big)\big) +$$

$$+\int_{t_0}^{t_1}\Big(\delta\big(y(t)-G(t)x\big(t, t_0, v(\cdot), x^0\big)\,|\,\mathscr{R}(t)\big)+\delta\big(v(t)\,|\,Q(t)\big)\Big)dt, \tag{4.1'}$$

$$dx/dt = A(t)x + C(t)v(t), \qquad x(t_0) = x^0.$$

The respective "dual problem" (D) is as follows:

determine

$$\mathscr{H}^0 = \inf\big\{\mathscr{H}\big(s^0, \lambda(\cdot)\big)\,|\,s^0\in R^n,\ \lambda(\cdot)\in L_2^{(m)}(T)\big\} \tag{4.2}$$

under the conditions

$$\mathscr{H}\big(s^0,\ \lambda(\cdot)\big) = \int_{t_0}^{t_1}\Big(\varrho\big(s\,(t, t_1, \lambda(\bullet), s^0)\,|Q(t)\big)+\varrho\,\big(\lambda(t)\,|\mathscr{R}(t)+y(t)\big)\Big)dt +$$

$$+\,\delta\big(s(t_0, t_1, \lambda(\cdot), s^0\,|\,l\big), \tag{4.2'}$$

$$ds/dt = -sA(t)+\lambda(t)G(t), \qquad s(t_1) = s^0 = 0,$$

Under standard regularity assumptions ([6], [14a], [15]) the solution of problem (4.2) is attained on the pair $\big(s_*^0,\ \lambda_*^0(\cdot)\big)$ and $\mathscr{J}^0 = \mathscr{H}^0$.

The problem above consists in estimating $(l, x(t_1))$ "a posteriori" on the basis of an available function $y(t)$. Of special interest is the "worst case" for the observer when $y(t) \equiv 0$ and $Q(\cdot) = -Q(\cdot)$, $\mathscr{R}(\cdot) = -\mathscr{R}(\cdot)$. The respective solution can be given an interpretation in terms of "a priori" estimation. The latter problem is as follows:

from among the linear operations $\langle w(\cdot), y(\cdot)\rangle$ defined by functions $w(\cdot) \in W$, where W is a given subset of $L_2^{(m)}(T)$ select an optimal operation which provides the minimum of

$$\chi\{w(\cdot)\} = \max_{v(\cdot),\xi(\cdot)} \{|\langle w(\cdot), y(\cdot)\rangle - (l, x(t_1))|\}$$

under the "unbiasedness" condition $w(\cdot) \in \mathscr{W}_0$ where $\mathscr{W}_0$ is the set of elements w satisfying the equality

$$\langle w(\cdot), y(\cdot)\rangle\Big|_{\substack{\xi(\cdot)=0\\ v(\cdot)=0}} = (l, x^*) = \langle w(\cdot), G(\cdot)x(\cdot, t_1, 0, x^0)\rangle, \qquad (4.3)$$

$x^* = x(t_1)$ denoting the unknown actual state of the system. Thus the guaranteed estimate for $l'x(t_1)$ received a priori (for $v(\cdot) \in Q(\cdot)$, $\xi(\cdot) \in \mathscr{R}(\cdot)$, $X^0 = R^n$) is $\chi\{w^0(\cdot)\} = \inf\{\chi\{w(\cdot)\} \mid w(\cdot) \in \mathscr{W}_0\}$. After computing the Langrangian, we obtain

$$\varepsilon^0 = \inf_{w(\cdot)\in W} \Phi\{w(\cdot)\}, \qquad \Phi\{w(\cdot)\} = \sup_{\{v(\cdot),\xi(\cdot),p\}} \mathscr{L}\{w(\cdot), v(\cdot), \xi(\cdot), p\}, \qquad (4.4)$$

$$\mathscr{L}\{w(\cdot), v(\cdot), \xi(\cdot), p\} = \langle w(\cdot), G(\cdot)x(\cdot, t_1, v(\cdot), p)\rangle + \\ + \langle w(\cdot), \xi(\cdot)\rangle + (p, l).$$

Turning attention to the dual problem, we are to choose from among the controls $u(\cdot) \in \mathscr{U}(\cdot)$ ($\mathscr{U}(\cdot)$ is a subset of $\mathscr{L}_2^{(m)}(T)$) an optimal control $u^0(\cdot)$ minimizing the functional

$$dz/dt = A^{(1)}(t)z + G^{(1)}(t)u, \qquad z(t_0) = 0, \qquad z(t_1) = d \qquad (4.5)$$

over the set of solutions of the system

$$\zeta^0 = \min_{u(\cdot)\in\mathscr{U}(\cdot)} \Psi\{u(\cdot)\}, \qquad \Psi\{u(\cdot)\} = \int_{t_0}^{t_1} F\{z(t), u(t)\}\,dt, \qquad (4.6)$$

where

$$F\{z(t), u(t)\} = \varrho\{z(t, t_0, u(\cdot), d) \mid Q^{(1)}(t)\} + \varrho\{u(t) \mid \mathscr{R}^{(1)}(t)\},$$

$\mathscr{U}(\cdot), Q(\cdot), \mathscr{R}(\cdot)$ are given sets. The respective Lagrangian has the following representation:

$$\mathscr{L}^{(1)}\left(u(\cdot), \mu(\cdot), r(\cdot), q\right) = \left\langle s\left(\cdot, t_1, \mu(\cdot), q\right), G^{(1)}(\cdot)u(\cdot)\right\rangle + \\ + \langle u(\cdot), r(\cdot)\rangle + (d, q).$$

Here $z\left(t, t_0, u(\cdot), d\right), s\left(t, t_1, \mu(\cdot), q\right)$ are, respectively, the solutions of system (4.5) and of the system

$$\dot{s} = -sA^{(1)}(t) + \mu(t)C^{(1)}(t), \qquad s(t_1) = q. \tag{4.7}$$

Thus

$$\zeta = \inf_{u(\cdot)} \sup_{\mu(\cdot), r(\cdot), q} \mathscr{L}^{(1)}\left(u(\cdot), \mu(\cdot), r(\cdot), q\right). \tag{4.8}$$

Assume that for a certain $\delta > 0$ problem (4.5) is solvable in the class $u(\cdot) \in \mathscr{U}(\cdot)$ for $z(t_1) = d^*$, for every vector $d^* \in S_\delta(d)$ ($\mathscr{U}(\cdot)$ being weakly compact).

Then in (4.8) a minmax is attained on the elements

$$u^0 \in \mathscr{U}(\cdot), \qquad w^0(\cdot) \in Q^{(1)}(\cdot), \qquad \xi(\cdot) \in \mathscr{R}^{(1)}(\cdot), \qquad q^0 = R^n.$$

The latter case is called *regular*. The regularity property for problem (4.4) is defined similarly.

THEOREM 4.1. *Assume the conditions*

$$A'(t) = -A^{(1)}(t), \qquad G'(t) = G^{(1)}(t), \qquad C'(t) = C^{(1)}(t),$$
$$U(\cdot) = \mathscr{W}(\cdot), \qquad Q(\cdot) = Q^{(1)}(\cdot), \qquad \mathscr{R}(\cdot) = \mathscr{R}^{(1)}(\cdot), \qquad d = l,$$

and suppose that at least one of the problems (4.8), (4.4) *is regular. Then the solutions of problems* (4.8), (4.4) *are attained on the elements* $u^0(\cdot)$, $\mu^0(\cdot), r^0(\cdot), q$ *and* $w^0(\cdot), v^0(\cdot), \xi^0(\cdot), p^0$, *respectively, where* $u^{0'}(\cdot) = w^0(\cdot)$, $\mu^{0'}(\cdot) = v^0(\cdot), r^{0'}(\cdot) = \xi^0(\cdot), q^0 = p^0$.

Thus the optimal control $u^0(\cdot)$ for the control problem (4.8) coincides with the optimal operation $w^0(\cdot)$ for the observation problem (4.4) and the extremal trajectory $s^0(\cdot) = s\left(\cdot, t_1, \mu^0(\cdot), p^0\right)$ coincides with the solution $x^0(\cdot) = x\left(\cdot, t_0, v^0(\cdot), p^0\right)$ of equation (4.1) related to the "worst" realizations of disturbance $v(\cdot)$ and boundary vector p.

COROLLARY 4.1. *Under conditions of Theorem* 4.1 *the solution* $u^0(\cdot)$ *of problems* (4.5), (4.6) *coincides for* $\mathscr{U}(\cdot) = \mathscr{L}_2^{(m)}(T)$ *with the solution* $\lambda^0(\cdot)$ *of problem* (4.2).

But then, in view of Theorem 4.1, the dual problem (D) may be interpreted for $y(t) \equiv 0$, $Q(\cdot) = -Q(\cdot)$, $\mathscr{R}(\cdot) = -\mathscr{R}(\cdot)$ as a problem of finding an "a priori" estimate of the parameter $\langle l, x(t_1)\rangle$ under uncertainties $v(\cdot)$, $\xi(\cdot)$, x^0. The extremal multiplier $\lambda^0(\cdot)$ of problem (4.2) coincides for $\mathscr{W}(\cdot) = \mathscr{L}_2^{(m)}(T)$ with the optimal operation $\mathscr{W}^0(\cdot)$, owing to the criterion $X(w(\cdot)) = \min$, and $v^0(\cdot)$ coincides with $\mu^0(\cdot)$, while $\zeta^0 = \mathscr{J}^0$.

Thus the "dual" relations for the extremal solutions of problems of control and observation under uncertainty are a reflection of the well-known duality relations in convex analysis ([1], [6], [24b]). This situation admits a generalization to more complicated classes of ordinary and distributed linear control systems.

Finally, it is worthwhile to emphasize the fact that problems of "a priori" and "a posteriori" estimation are related to each other in the same way as problems of open loop and closed loop control.

5. Positional strategies. Game-theoretic duality

The process of determining the guaranteed estimate ε^0 (4.4) may be considered within the framework of differential games where one is to construct, in accordance with system (2.1), a semicontinuous in $y_t(\cdot)$ multivalued function $\mathscr{W}\langle t, y_t(\cdot)\rangle \subseteq Q(t)$—a "strategy of control" which satisfies the inequality

$$\varepsilon^0[t_1] = \max_{x \in X(t_1,\cdot)} \inf_{z \in Z(t_1, \mathscr{W}^0(t,\cdot))} \{|l'x - z|\} \leqslant \varepsilon^0,$$

whatever be the realization $y_t(\cdot)$. Here $Z\langle t, \mathscr{W}(t,\cdot)\rangle$ is the set of solutions of the inclusion

$$\dot{z}(t) \in y'(t)\, \mathscr{W}\langle t, y_t(\cdot)\rangle, \qquad z(t_0) = 0,$$

Thus the aim of the strategy $\mathscr{W}^0\langle t, y_t(\cdot)\rangle$ is the approach by z to the actual value $z^* = l'x^*$ at a guaranteed distance ε^0 (with the additional condition that the realization $w[t] \in \mathscr{W}^0\langle t, y_t(\cdot)\rangle$ satisfies $w[\cdot] \in \mathscr{W}_0$).

The computation of $\mathscr{W}^0\langle t, y_t(\cdot)\rangle$ may be achieved on the basis of a differential game duality principle, which associates with the problem under consideration a regular differential game with incomplete information (in the sense of [9], [13]). The systems under conflict here are (4.1) and (4.2). The equivalence of the respective solutions then follows from the duality principles just mentioned and from the possibility of constructing the solution strategies for regular differential games on the basis of additional open solutions ([13]). Thus the realization of the process of minimaxi-

mizing the Lagrangians $\mathscr{L}, \mathscr{L}^{(1)}$ may be treated as a differential game between the primal and dual systems (4.1), (4.2) controlled by parameters w and v ($w = \lambda$) or, respectively, by u and μ.

6. Adaptive strategies of control

A problem of special interest for (1.1), (1.5) is the one of a proper formalization and further on, of a computation of a control strategy $u\big(t, y_t(\cdot)\big)$ that ensures for all $t \in T$ the inclusion

$$HX\big(t, y_t(\cdot), u_t[\cdot], X^0\big) \subseteq \mathscr{K}(t), \tag{6.1}$$

where $u[t]$ is the realization of the strategy $u\big(t, y_t(\cdot)\big)$, H is an $s \times n$-matrix, $\mathscr{K}(t)$ is a given upper semicontinuous multivalued map ($\mathscr{K}(t) \in \Omega(R^s)$ for all t). The inclusion (6.1) must be true for any realization $y_t(\cdot) \in Y_t\big(u_t(\cdot), X^0\big)$.

For full measurement of $x(t)$ the inclusion (1.5) turns into the equality $y \equiv x$. The strategy $u(t, x)$ ($u\big(t, x_t(\cdot)\big)$) then ensures the "viability" property for system (1) in the sense of [1].

The solubility conditions for the linear case of the given problem (see (1.4), $\varphi(t, q) = C(t)q$) with additional terminal restrictions may be shortly stated as follows. For the system

$$\dot{q} = A(t)q + B(t)u, \quad q(t_0) = 0, \quad u \in \mathscr{P}(t) \tag{6.2}$$

one has to specify a number ν^0 and a multivalued strategy

$$u\big(t, q[t] + P(t, \cdot)\big), \quad q[t] + P(t, \cdot) = X(t, \cdot),$$
$$(y[t] = z[t] + q[t]),$$

that ensure the inclusion

$$Hq(t, t_0, u[\cdot], 0) \subseteq \mathscr{K}(t) \dot{-} HP(t, \cdot) = \mathscr{K}^0(t, \cdot)$$

with the terminal constraint

$$\Phi\big(q(t_1, t_0, u[\cdot], 0) + P(t_1, \cdot)\big) \leqslant \nu^0.$$

Here $y[t]$, $u[t]$ are the realizations of the measurement and the control, $q[t] = q(t, t_0, u[\cdot]\, 0)$ is the solution of (6.2), $P(t, \cdot) = P\big(t, z_t(\cdot), P^0\big)$ is the informational set for the system

$$\dot{p} \in A(t)p + C(t)Q(t), \quad z \in G(t)p + \mathscr{R}(t), \quad X^0 = P^0, \tag{6.3}$$

similar to the set $X\big(t, y_t(\cdot), 0, X^0\big)$ defined in Section 1 for (1.1), (1.5)

$(P(\tau, z_\tau(\cdot|t), t, P)$ is then similar to $X(\tau, y_\tau(\cdot|t), 0, X))$, $\Phi(x) = \max\{\varphi(x) | x \in X\}$ where φ is a given convex function. Write

$$\mathscr{K}^\circ(\tau, \cdot) = \mathscr{K}^0(\tau, \cdot\,|t, P) = \mathscr{K}(\tau) \dot{-} HP(\tau, z_\tau(\cdot|t), t, P)$$

and let $Z_{t_1}(\cdot|t, P)$ be the set of measurements generated by system (6.3) for $z(t) \in P$, where $t \leqslant \tau \leqslant t_1$. Let $v(u_{t_1}(\cdot|t), z_{t_1}(\cdot|t), t, q+P)$ be the smallest number v satisfying

$$\Phi\big(q(t_1, t, u_{t_1}(\cdot|t), q) + P(t_1, z_{t_1}(\cdot|t), P)\big) \leqslant v^0$$

under the constraint

$$Hq(\tau, t, u_\tau[\cdot|t], q) \in \mathscr{K}^0(\tau, \cdot\,|t, P), \tag{6.4}$$

$$v^0(t, q+P) = \max_{z_{t_1}(\cdot|t) \in Z_{t_1}(\cdot|t, P)} \min_{u_{t_1}(\cdot|t) \in \mathscr{P}_{t_1}(\cdot|t)} v\big(u_{t_1}(\cdot|t), z_{t_1}(\cdot|t), t, q+P\big).$$

The function $v^0(t, q+P)$ corresponds to the potential in the theory of differential games [5]. We have

$$v^0(t, q+P) = \sup\{\Psi(l, \Lambda(\cdot), t, q+P) | \Lambda(\cdot) \in V^{(m)}(T), l \in R^n\}$$

where $\Psi(l, \Lambda(\cdot), t, q+P)$ is the dual functional for problem (6.4)

$$\Psi(l, \Lambda(\cdot), t, q+P) = -\int_t^{t_1} \varrho\big(s(\tau, t_1, \Lambda(\cdot), l|P(\tau))\big)\, d\tau - f(l, \Lambda(\cdot)|t, q+P)$$

and the function $f(l, \Lambda(\cdot)|t, q+P)$ is calculated by means of the technique of convex analysis [10, 22, 23], and $s(\tau, t, \Lambda(\cdot), l)$ is the solution of (2.2) (with $\lambda(\cdot) = d\Lambda(\cdot)|dt$).

Assume that the functional $\Psi(l, \Lambda(\cdot), t, q+P)$ is strictly concave in l, $\Lambda(\cdot)$ and that for each of the sets $\mathscr{K}^0(t, \cdot)$ there exists a Euclidean σ-neighborhood $\mathscr{K}^0_\delta(t, \cdot)$ such that in the domain $p \in \mathscr{K}^0_\delta(t, \cdot)$, $0 < r < \sigma$ the following conditions are fulfilled

$$\partial_+ r(p | \mathscr{K}^0(t, \cdot))/\partial t + \min_l \max_f (l, f) \leqslant 0,$$

$$f \in \partial r(p | \mathscr{K}^0(t, \cdot)), \quad l \in A(t)p + B(t)P(t),$$

$$r(p|K) = \inf\{|Hp - q| | q \in K\}.$$

Assume also that, for a certain $\delta > 0$, each of the sets $\mathscr{K}(t) - (HP(t, \cdot) + S_\delta(0))$ is nonvoid.

The solution of the problem is then given by the strategy

$$U(t, X) = \begin{cases} \partial\varrho\left(s^0(t\,|\,t, q+P)\,|\,\mathscr{P}(t)\right) & \text{if} \quad q \in \operatorname{int}\mathscr{K}^0(t^0, \cdot), \\ P(t) & \text{if} \quad q \in \mathscr{K}^0(t, \cdot)\setminus\operatorname{int}\mathscr{K}^0(t, \cdot), \\ \partial\varphi^*(0\,|\,t, q+P) & \text{if} \quad q \notin \mathscr{K}^0(t, \cdot), \end{cases}$$

where $s^0(\tau\,|\,t, q+P) = s\left(\tau, t_1, \Lambda^0(\cdot), l^0\right)$, the pair

$$\{l^0, \Lambda^0(\cdot)\} \in \partial\Psi^*(0, 0, t, q+P),$$

$\left(\Psi^*\left(l^*, \Lambda^*(\cdot), t, q+P\right)\right.$ is the functional conjugate to Ψ in $l, \Lambda(\cdot)$, and $\varphi^*(x\,|\,t, q+P)$ is the function conjugate to

$$\varphi(l\,|\,t, q+P) = \varrho\left(l\,|\,\partial r\left(p\,|\,K^0(t, \cdot)\right)\right) + \delta\left(l\,|\,A(t)p + B(t)P(t)\right).$$

An accurate definition of the solution for the inclusion

$$dx/dt \in A(t)x + B(t)\,U\left(t, X(t, \cdot)\right) + C(t)Q(t),$$

$$y = G(t)x + \mathscr{R}(t),$$

and the proof of a respective existence theorem is given in [14a].

The problem under consideration may be treated as a problem of constructing an "adaptive" strategy of control, its solution requiring the use of methods of game-theoretic dynamic control, ensuring therefore a guaranteed result.

Among other problems lying within the framework of this report we might mention that of maximizing the diameter $dX(t_1, \cdot)$ of the set $X(t_1, \cdot)$ by selecting a strategy $u\left(t, y\,(\cdot)\right)$ for (1.1), (1.5). The more general situations lead to optimization with respect to a partial ordering introduced on the variety $\{X(t_1, \cdot)\}$

7. On problems with quadratic constraints

Consider the system (2.1) with the constraints

$$(x^0, Lx^0) + \langle v(\cdot), M(\cdot)v(\cdot)\rangle + \langle \xi(\cdot), N(\cdot)\,\xi(\cdot)\rangle \leqslant \mu^2, \qquad (7.1)$$

where $L > 0$, $M(t) \geqslant 0$, $N(t) \geqslant 0$ are symmetrical matrices, $L = \text{const}$.

The respective domains $X(t_1, \cdot)$ now turn to be ellipsoids. Therefore the Chebyshev center $x^0(t_1, \cdot)$ that gives the minimax estimate for $x(t_1)$ is the center of this ellipsoid.

One can therefore consider the problem of selecting a program $g^0[t] = g\left(t, u^0(t)\right)$ or a measurement strategy $g(t, u) = g\left(t, u^0\left(t, X(t, \cdot)\right)\right)$

providing, for any realization $y_{t_1}(\cdot)$, a minimal ellipsoid $X(t_1, \cdot)$ with respect to the inclusion ordering. The problem allows a solution in explicit terms [7].

Another question that deserves to be mentioned is to describe the class of closed-loop optimal control problems for system (1.4), $\varphi(t, q) = C(t)q$, (1.6), (7.1) in the class of feedback strategies $u\big(t, y_t(\cdot)\big)$ for which the separation principle is true; i.e., such that $\{u\big(t, y_t(\cdot)\big)\}$ may be reduced to a set of strategies of type $u = u\big(t, x^0(t, \cdot)\big)$ calculated in the process of independent solution of the respective problems of estimation and of feedback control with complete information. There exist some nontriviar examples which show that this class of problems is nonvoid.

8. On nonlinear estimation

Consider the construction of sets $X(t, \cdot)$ for system (1.1), (1.3) with lineal measurement

$$y - G(t)x \in \mathcal{R}(t), \tag{8.1}$$

In this case the set $X(t, \cdot)$ is, in general, nonconvex. However, here and in the more general nonlinear case it is possible to construct an equation of a convex majorant $X^*(t, \cdot) \supseteq \operatorname{co} X(t, \cdot)$. Namely, the following result is true:

THEOREM 8.1. *Assume that the function $u(t)$ in system* (1.3) *is the solution of the equation $\dot{u} = Cu$, $u(t_0) = u^0$, where $C = \{c_{ij}\}$ is a given constant $(p \times p)$-diagonal matrix. Then for any triplet $\{L, M(\cdot), N(\cdot)\} = \Lambda(\cdot)$ the set $\operatorname{co} X(t, \cdot)$ for the system* (1.3), (8.1) *is contained in the set $Z\big(t, \Lambda(\cdot), \cdot\big) = \{z[t] = x^*[t] + z^*[t]\}$ generated by all solutions of the system*

$$\dot{x}^* \in A\big(t, Q(t)\big)x^* + B(t)u, \qquad \dot{u} = Cu,$$

$$x(t_0) = x^0, \qquad z^* = \sum_{i=1}^{p} z^{(i)}, \qquad u(t_0) = u^0, \qquad z^{(i)}(t_0) = 0,$$

$$\dot{z}^{(i)} \in c_{ii}E_n z^{(i)} + P^{(i)}G'(t)K^{-1}(t)\big(y(t) - G(t)x^* - R(t)\big),$$

$$\dot{P}^{(i)} = 2c_{ii}P^{(i)} - \big(P^{(i)}G'(t)K^{-1}(t)G(t)P + PG(t)K^{-1}(t)G(t)P^{(i)}\big) + \sum_{k=1}^{m} c_{ii}D_{ik} + \bar{N}_i(u_i^0)^2,$$

$$\dot{D}_{ik} = (c_{ii} + c_{kk})D^{ik} + P^{(i)}G'(t)K^{-1}(t)G(t)P^{(k)},$$

$$K(t) = M(t) + G(t)P^*[t]G'(t), \quad P^*[t] = (x^{0'} \otimes E_n)\bar{L}(t)(x^0 \otimes E_n),$$
$$D_{ik}(t_0) = 0, \quad P^{(i)}(t_0) = 0, \quad (i, k = 1, \ldots,); \quad x, z \in R^n.$$

COROLLARY 8.1 *The following inclusion is true*

$$\operatorname{co} X(t, \cdot) \subseteq \bigcap \{Z(t, \Lambda(\cdot), \cdot) \mid \Lambda(\cdot)\}.$$

Theorem 8.1 is proved by means of an approximation of the given solution with other informational domains which are ellipsoids constructed for a special problem related to autonomous equations of type (2.1) with quadratic constraints on the respective Green functions and on the disturbances in the measurement.

The given inverse problem is related to problems of identification of dynamic control systems [7]. Here, in particular, we come to the problems of selecting the best inputs $u(t)$ that ensure minimal dimensions of the domain $X(t, \cdot)$ and to those of best approximation of the trajectory of the Chebyshev center $x^0(t, \cdot)$ by means of solutions of appropriate ordinary or time-lag control systems.

The description of the evolution of domains $X(t, \cdot)$ for a more complicated system (1.1), (1.2), (8.1) may be considered within a discrete scheme of sequential estimation for one-stage systems of type

$$z \in f(X) + Q, \quad y - Gz \in \mathscr{R}, \tag{8.2}$$

$$z^* \in \operatorname{co} f(X) + Q, \quad y - Gz^* \in \mathscr{R}. \tag{8.3}$$

Denoting the sets of solutions for systems (8.2), (8.3) by $Z = \{z\}$, $Z^* = \{z^*\}$, we have $Z \subseteq \operatorname{co} Z \subseteq Z^*$. Here the sets Z are nonconvex and, in general, disconnected.

Following the schemes similar to Sections 2, 3, we obtain

$$Z \subseteq \mathscr{R}(M, f(X)), \quad Z^* \subseteq \mathscr{R}(M, \operatorname{co} f(X)),$$

where M is any $m \times n$-matrix,

$$\mathscr{R}(M, F) = (E_n - MG)(F + Q) + M(y - \mathscr{R}).$$

Finally we have

$$Z = \bigcup_{x \in X} \left(\bigcap \{\mathscr{R}(M, f(x)) \mid \cdot M\} \right), \quad Z^* = \bigcap \{\mathscr{R}(M, \operatorname{co} f(X)) \mid M\}.$$

A sequential application of these relations yields a discrete approximation of domains $X(t, \cdot)$ or of their convex majorants [11].

From the conclusion of the above it is finally possible to prove that for the system (1.1), (1.5), where

$$F(t, x, u) = F(t, x), \quad G(t, x, u) = G(t)x + \mathscr{R}(t)$$

satisfy the assumptions of Theorem 1.1 and $y[t]$ is continuous, the set $X[t] = X(t, \cdot)$ satisfies the following evolution equation (a.e. in T)

$$\lim_{h \to 0} \mathscr{D}\big(X[t+h],\ \mathscr{X}_h(t+h)\big) h^{-1} \to 0$$

where

$$\mathscr{X}_h(t+h) = \bigcup_{x \in X[t]} \bigcap_{M} \bigcap_{\varepsilon > 0} \{\big(E - MG(t+h)\big)\big(x + h\mathscr{F}(t, x)\big) + \\ + \varepsilon h M S + M\big(y[t+h] - \mathscr{R}(t+h)\big)\}$$

and $\mathscr{D}(X, \mathscr{X})$ is the Hausdorff distance between the sets X, $\mathscr{X}$.

9. Infinite-dimensional generalizations

The description of the evolution of domains $X(t, \cdot)$ admits a natural generalization to differential equations and inclusions with after-effect or to systems described by partial differential equations [16] (e.g., as those which describe the estimation of a heat distribution within a body of given configuration through the values of a finite number of pointwise measurements under uncertain disturbances).

We wish to point out one specific feature of these problems. Namely, in the absence of constraints on the initial distribution and with a strong boundedness of disturbances in the original system and in the measurement equation, it is worth while to investigate the conditions for the respective informational domains to be bounded. The latter property is equivalent to a condition similar to observability in the absence of disturbances. Here it turns out that in parabolic systems and in systems with time lag the latter boundedness property holds if and only if the duration of the observation process is not less than the length of a certain critical interval of observation.

We finally wish to underline that the problems considered in this report are closely connected with the theory and methods of solving ill-posed problems [26].

References

[1] Aubin J.-P. and Cellina A., *Differential Inclusions*, Springer-Verlag, 1984.

[2] Bensoussan A., *Filtrage Optimal des Systèmes Linéaires*, Dunod, Paris, 1971.

[3] Bertsekas D. P., and Rhodes I. B., Recursive state estimation for a set-membership description of uncertainty, *IEEE Trans. Aut. Control* **AC-16**, No. 2 (1971).

[4] Castaing C., and Valadier M., *Convex Analysis and Measurable Multifunctions*, Lect. Notes Math. vol. **580**, 1977.

[5] Chernousko F. L., Optimal Guaranteed Estimates of Uncertainty with the Aid of Ellipsoids. I, II, III, *Isvestia Akad. Nauk SSSR, Tehn. Kibernetika* **3–5** (1980).

[6] Ekeland I. et Temam R., *Analyse Convexe et problèmes variationelles*, Dunod, Paris, 1974.

[7] *Estimation Under Uncertainty* (A. B. Kuržanskii, B. I. Ananyev editors), Sverdlovsk, Akad. Nauk SSSR, 1982.

[8] Fleming W. H. and Rishel R. W., *Deterministic and Stochastic Optimal Control*, Springer-Verlag, 1975.

[9] Friedman A., *Differential Games*, Wiley-Interscience, N. Y., 1971.

[10] Kalman R. and Bucy R., New Results in Linear Filtering and Prediction Theory, *Trans. ASME* **83** D (1961).

[11] Koscheev A. S., and Kurzhanskij A. B., On Adaptive Estimation of Multistage Systems Under Uncertainty, *Isvestia Akad. Nauk SSSR, Teh. Kibernetika* **2** (1983).

[12] Krasovskii N. N.,

(a) *The Theory of Control of Motion*, Nauka, Moscow, 1968.

(b) *Game-theoretic Problems on the Encounter of Motions*, Nauka, Moscow, 1970.

(c) On Differential Evolution Systems, *Priklad. Matematika i Mekhanika* **41**, No. 5 (1977).

(d) Differential Games. Approximative and formal models, *Mat. Sbornik* **107**, No. 4 (1978).

(e) Control under Incomplete Information and Differential Games, in: *Proc. Intern. Congress of Math., Helsinki*, 1978.

[13] Krasovskii N. N. and Subbotin A. I., *Positional Differential Games*, Nauka, Moscow, 1973.

[14] Kurzhanskij A. B.,

(a) *Control and Observation Under Conditions of Uncertainty*, Nauka, Moscow, 1977.

(b) Differential Games of Observation, *Doklady Akad. Nauk SSSR* **207**, No. 3 (1972).

(c) On informational sets of a control system, *Doklady Akad. Nauk SSSR* **240**, No. 1 (1978).

(d) On informational sets for control systems, *Differencialnye Uravneniya* **13**, No. 11 (1977).

(e) *Estimation of Control System Dynamics Under Uncertainty in Parameters and Inputs*, Proc. of the 8th Triennial World Congress of the IFAC, Kyoto, 1981.

(f) *Evolution Equations for Estimation Problems in Systems with Uncertainty*, IIASA, Working Paper, Laxenburg, June, 1982.

[15] Laurent P. J., *Approximation et Optimization*, Hermann, Paris, 1972.

[16] Lions J., *Contrôle Optimal des Systèmes Gouvernés par des équations aux dérivées partielles*, Dunod, Paris, 1968.

[17] Liptser R. and Shiriayev A. N., *Statistics of Stochastic Processes*, Nauka, Moscow, 1974.

[18] Łojasewicz St., Jr., Some properties of accessible sets in nonlinear control systems, *Ann. Polon. Math.* **36** (1979).

[19] Mischenko E. F., and Pontriagin L. S., Linear differential games, *Doklady Akad. Nauk SSSR* **174**, No. 1 (1967).

[20] Nikonov O. I., On the combination of control and observation in game-theoretical problems of evasion, *Differencialnye Uravneniya* **13**, No. 7 (1977).

[21] Pliś A.,

(a) Generalized ordinary differential equations and control theory, *Math. Balkanica* 3 (1973).

(b) *Accessible sets in Control Theory*, Int. Conf. on Diff. Equations, Acad. Press, 1975.

[22] Pontriagin L. S., Boltyansky V. G., Gamkrelidze R. V., and Mischenko E. F., *Mathematical Theory of Optimal Process*, 3rd ed., Nauka, Moscow 1976.

[23] Pschenichnyi B. N. and Pokotilo V., On the accuracy of minmax estimates for observation problems, *Doklady Akad. Nauk Ukrainskoi SSR Ser. A*, No. 3 (1982).

[24] Rockafellar R. T.,

(a) *Convex Analysis*, Princeton Un. Press, 1970.

(b) Measurable Dependence of Convex Sets and Functions on Parameters, *J. Math. Anal. Appl.* **28** (1969).

(c) Duality in Optimal Control, Lect. Notes Math., vol. **680**, 1978.

[25] Schlaepfer F. M. and Schweppe F. C., Continuous Time State Estimation Under Disturbances Bounded by Convex Sets, *IEEE Trans. Aut. Control* **AC-17**, No. 2 (1972).

[26] Tikhonov A. N. and Arsenin V. Ja., *Methods of solving ill-posed problems*, Nauka, Moscow, 1974.

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

P. L. LIONS

Hamilton–Jacobi–Bellman Equations and the Optimal Control of Stochastic Systems

Introduction

In many applications (engineering, management, economy) one is led to control problems for stochastic systems: more precisely the state of the system is assumed to be described by the solution of stochastic differential equations and the control enters the coefficients of the equation. Using the *dynamic programming principle* R. Bellman [6] explained why, at least heuristically, the optimal cost function (or *value function*) should satisfy a certain partial differential equation called the *Hamilton–Jacobi–Bellman equation* (HJB in short), which is of the following form

$$\sup_{\alpha \in \mathscr{A}} \{A_\alpha u - f_\alpha\} = 0 \quad \text{in } \mathcal{O} \tag{1}$$

(with appropriate boundary conditions) where A_α is a family of second-order, elliptic, possibly degenerate operators, parametrized by α lying in a given set $\mathscr{A}$ (of the control values); and where f_α is a family of given functions. Here and below $\mathcal{O}$ is a given domain in $\boldsymbol{R}^N$ and u is a scalar function.

The HJB equations are *second-order, degenerate elliptic, fully nonlinear equations* of the following form

$$H(x, u, Du, D^2u) = 0 \quad \text{in } \mathcal{O}$$

with the main restriction that H is *convex* in (Du, D^2u).

As special cases the HJB equations (1) include

(i) the first-order Hamilton–Jacobi equations (HJ in short)

$$H(x, u, Du) = 0 \quad \text{in } \mathcal{O}. \tag{2}$$

Strictly speaking, (1) contains (2) when H is convex in (t, p), but as it will be made clear below, our methods enable us to treat the general HJ equation (2), i.e., the case of a general *Hamiltonian* H.

(ii) *the Monge–Ampère equations*

$$\det(D^2u) = H(x, u, \nabla u) \quad \text{in } \mathcal{O}, \qquad u \text{ convex in } \bar{\mathcal{O}}. \tag{3}$$

Again strictly speaking, (3) is a special case of (1) only if $H(x, t, p)$ is convex in (t, p) — and if this is the case, the fact that (1) contains (3) is indicated in Section IV.1. But just as above, the methods we give below enable us to treat the general Monge–Ampère equations (3).

We present here various existence and uniqueness results for equations (1)–(2)–(3), and one consequence of the results presented below is a complete justification of the derivation of (1) in the theory of optimal stochastic control. The tools and methods that we used or introduced for this study are of three kinds:

(i) the notion of viscosity solutions of (1)–(2): this notion, introduced by M.G. Crandall and the author, makes possible, in particular, a complete treatment of the HJ equations (2);

(ii) probabilistic methods: many of them being inspired by N.V. Krylov's work;

(iii) new partial differential equation arguments involving approximation methods and a priori estimates.

The plan is as follows

I. Viscosity solutions

I.1 Definition of viscosity solutions for Hamilton–Jacobi equations. We recall briefly below the notion of viscosity solutions of HJ equations introduced by M.G. Crandall and the author [14]. This notion enables us to settle the question arising from the following remarks: (i) in general, there does not exist global C^1 solutions, (ii) if $W^{1,\infty}$ solutions of (2) can easily be built by the vanishing viscosity method (see W. H. Fleming [22]), then in general, there may exist many $W^{1,\infty}$ solutions of (2) with prescribed boundary conditions; and moreover, the Lipschitz solutions are unstable, see [14] for more details. The notion of viscosity solutions enables us to select the 'good' solution for which existence and uniqueness results hold. In addition one has stability results and viscosity solutions are exactly the solutions built by the vanishing viscosity method.

Let $\mathcal{O}$ be an open set in $\boldsymbol{R}^N$ and let H be a continuous function on $\mathcal{O} \times \boldsymbol{R} \times \boldsymbol{R}^N$. We first recall the notion of sub and superdifferential of a continuous function u at a point $x_0 \in \mathcal{O}$: the superdifferential $D_1^+ u(x)$ is the closed convex set, possibly empty, defined by

$$D_1^+ u(x) = \{\xi \in \boldsymbol{R}^N \mid \limsup_{y \to x,\ y \in \mathcal{O}} \{\varphi(y) - \varphi(x) - (\xi, y-x)\} |x-y|^{-1} \leqslant 0\};$$

he subdifferential $D_1^- u(x)$ being defined in a similar way or by

$$D_1^- u(x) = -D_1^+(-u)(x).$$

DEFINITION I.1. $u \in C(\mathcal{O})$ is said to be a *viscosity solution* of the HJ equation (2) if for all $x \in \mathcal{O}$ we have

$$\forall \xi \in D_1^+ u(x), \quad H(x, u(x), \xi) \leqslant 0, \tag{4}$$

$$\forall \xi \in D_1^- u(x), \quad H(x, u(x), \xi) \geqslant 0. \ \blacksquare \tag{5}$$

Remark I.1. Of course, if $\varphi \in C(\mathcal{O})$ is differentiable at $x_0 \in \mathcal{O}$ then $D_1^+ \varphi(x_0) = D_1^- \varphi(x_0) = \{V\varphi(x_0)\}$; and conversely, if $D_1^+ \varphi(x_0) \cap D_1^- \varphi(x_0) \neq \emptyset$ then φ is differentiable at x_0. In particular, any classical (i.e., C^1) solution of (2) is a viscosity solution of (2), and any viscosity solution u of (2) satisfies equation (2) at all points of differentiability. ■

This notion, introduced by M.G. Crandall and P.L. Lions [14] has many equivalent formulations; the most 'convenient' one being given in the following

PROPOSITION I.1. *Let $u \in C(\mathcal{O})$; u is a viscosity solution of* (2) *if and only if, for all $\varphi \in C^1(\mathcal{O})$,*

at each local maximum point x_0 of $u-\varphi$ we have

$$H\big(x_0, u(x_0), D\varphi(x_0)\big) \leqslant 0, \tag{4'}$$

at each local minimum point x_0 of $u-\varphi$ we have

$$H\big(x_0, u(x_0), D\varphi(x_0)\big) \geqslant 0. \ \blacksquare \tag{5'}$$

Remark I.2. It is possible to replace in the above statement local by global (resp. global strict, resp. local strict) and $\varphi \in C^1$ by $\varphi \in C^2$ (resp. $\varphi \in C^\infty$). For more details, we refer to [14] and to M. G. Crandall, L. C. Evans, and P. L. Lions [13]. ■

One of the striking features of viscosity solutions is their *stability with respect to uniform convergence* (on compact sets): if $u_n \in C(\mathcal{O})$ is a viscosity solution of (2) where H is replaced by H_n, and if u_n, H_n converge uniformly on compact sets to u, H, then u is a viscosity solution of (2). Similar results are obtained for the limit functions obtained via the *vanishing viscosity method*: if $u_\varepsilon \in C^2(\mathcal{O})$ solves

$$-\varepsilon \Delta u_\varepsilon + H_\varepsilon(x, u_\varepsilon, Du_\varepsilon) = 0 \text{ in } \mathcal{O}$$

and if $u_\varepsilon, H_\varepsilon$ converge uniformly on compact sets to u, H as ε goes to 0, then u is a viscosity solution of (2).

This simple remark enables us to obtain very general existence results of viscosity solutions for the HJ equation (2) (with prescribed boundary conditions on $\partial\mathcal{O}$): using the vanishing viscosity method and the properties stated above, this amounts to the obtention of a priori $W^{1,\infty}$ estimates uniform in ε. Existence results are treated in P. L. Lions [33], [35].

I.2. Some of the main results on viscosity solutions of HJ equations. We now present a uniqueness result taken from M. G. Crandall and P. L. Lions [14]. We use the following assumptions

$$\exists \gamma = \gamma(R) > 0, \quad \forall (x, p) \in \mathcal{O} \times \boldsymbol{R}^N, \quad \forall |t|, |s| \leqslant R,$$

$$\big(H(x, t, p) - H(x, s, p)\big)(t-s) \geqslant \gamma(t-s)^2, \tag{6}$$

$$\lim_{\varepsilon \to 0} [\sup\{|H(x, t, p) - H(y, t, p)| / |x-y|(1+|p|) \leqslant \varepsilon, \ |t| \leqslant R\}] = 0 \tag{7}$$

for all $R < \infty$.

THEOREM I.1. *Assume that $\mathcal{O}$ is bounded and that* (6) *holds. Let $u, v \in C(\bar{\mathcal{O}})$ be two viscosity solutions of* (2). *We assume in addition either that* (7) *holds or that $u, v \in W^{1,\infty}(\mathcal{O})$. Then the following inequality holds:*

$$\sup_{\mathcal{O}} (u-v)^+ \leqslant \sup_{\partial\mathcal{O}} (u-v)^+. \quad \blacksquare \tag{8}$$

Of course, (8) implies uniqueness results for viscosity solutions of (2) with prescribed boundary conditions on $\partial\mathcal{O}$: indeed, if $u = v$ on $\partial\mathcal{O}$, then in view of (8), $u \equiv v$ in $\bar{\mathcal{O}}$.

Remark I.3. This result is shown in [14], [33] to be essentially optimal; variants concerning unbounded domains such as $\boldsymbol{R}^N$ or time-dependent problems are given in [14].

I.3. Remarks on the viscosity solutions for second-order equations. We now consider fully nonlinear second-order elliptic equations such as

$$H(x, u, Du, D^2u) = 0 \quad \text{in } \mathcal{O}; \tag{9}$$

where $H \in C(\mathcal{O} \times \boldsymbol{R} \times \boldsymbol{R}^N \times S^N)$ [1] satisfies the following ellipticity condition:

$$H(x, t, p, \eta_1) \leqslant H(x, t, p, \eta_2) \quad \text{if} \quad \eta_1 \geqslant \eta_2, \ \forall (x, t, p) \in \mathcal{O} \times \boldsymbol{R} \times \boldsymbol{R}^N. \tag{10}$$

To define viscosity solutions of (9), we must first define, for all $(u, x_0) \in C(\mathcal{O}) \times \mathcal{O}$, the superdifferential of order 2, $D_2^+ u(x_0)$. It is the closed convex set, possibly empty, defined as follows:

$$D_2^+ u(x_0) = \{(\xi, \eta) \in \boldsymbol{R}^N \times S^N |$$
$$\lim_{y \to x_0, \ y \in \mathcal{O}} \sup \{u(y) - u(x_0) - (\xi, y - x_0) - \tfrac{1}{2}(\eta(y - x_0), y - x_0)\} |y - x_0|^{-2} \leqslant 0\},$$

the subdifferential of order 2 $D_2^- u(x_0)$ being defined in a similar way or by

$$D_2^- u(x_0) = -D_2^+(-u)(x_0).$$

DEFINITION I.2. $u \in C(\mathcal{O})$ is a viscosity solution of (9) if for all $x \in \mathcal{O}$ we have

$$\forall (\xi, \eta \in D_2^+ u(x), \quad H(x, u(x), \xi, \eta) \leqslant 0, \tag{11}$$

$$\forall (\xi, \eta) \in D_2^- u(x), \quad H(x, u(x), \xi, \eta) \geqslant 0. \quad \blacksquare \tag{12}$$

---

[1] S^N denotes the space of $N \times N$ symmetric matrices.

It can easily be checked that Remarks I.1–2 and the equivalent formulation of Proposition I.1 can be extended to this case. Of course, if H does not depend on D^2u, we recover the preceding notion. The stability results also hold if we consider sequences of H, u converging uniformly on compact sets. For more details on these questions we refer to P.L. Lions [34], [38].

The main question in this case concerns uniqueness results: except for easy results (if $N = 1, 2, \ldots$), the only known case is when (9) reduces to the HJB equation (see Section II. 3 below). As regards existence results, general ones may be obtained by appropriate approximation methods.

I.4. Further results. In the bibliography, various references are given concerning the notion of viscosity solutions for HJ equations and its applications to existence results, numerical approximation, optimal deterministic control problems, asymptotic problems, nonlinear semigroup theory, accretive operators. In P.L. Lions [39], the relations between viscosity solutions of (9) and $W^{2,p}$-solutions of (9) (satisfying (9) a.e.) are investigated.

II. Optimal stochastic control problems

II.1. Presentation of the problem. We define an admissible controlled system as a collection consisting of (i) a probability space (Ω, F, F^t, P) with the usual properties, (ii) an F^t-Brownian motion B_t, (iii) a progressively measurable process α_t with compact values in a given separable metric space $\mathscr{A}$; here α_t is the control process. The state of the system is given by the solution of the following stochastic differential equation

$$dX_t = \sigma(X_t, \alpha_t)dB_t + b(X_t, \alpha_t)dt, \quad x_0 = x \in \bar{\mathcal{O}}, \tag{13}$$

where, for simplicity, $\mathcal{O}$ is a bounded smooth domain in $\boldsymbol{R}^N$ and $\sigma(x, \alpha) = (\sigma_{ij}(x, \alpha))_{\substack{1 \leqslant i \leqslant N \\ 1 \leqslant j \leqslant m}}$, $b(x, \alpha) = (b_i(x, \alpha))_{1 \leqslant i \leqslant N}$ are coefficients satisfying conditions detailed below and m is a given integer. We now define the cost functions and the *value function* of the problem

$$J(x, \mathscr{S}) = E\Big\{\int_0^\tau f(X_t, \alpha_t)\exp\Big(-\int_0^t c(X_s, \alpha_s)ds\Big)dt + \\ + \varphi(X_\tau)\exp\Big(-\int_0^\tau c(X_s, \alpha_s)ds\Big)\Big\}, \tag{14}$$

$$u(x) = \inf\{J(x, \mathscr{S})/\mathscr{S} \text{admissible system}\}, \tag{15}$$

where $f(x, a)$, $c(x, a)$ are given real-valued functions and τ is the first exit time from $\bar{\mathcal{O}}$ of X_t; $\tau = \inf(t \geqslant 0, X_t \notin \bar{\mathcal{O}})$. To simplify the presentation we will always assume that

$$\begin{cases} \psi(\cdot, a) \in W^{2,\infty}(\boldsymbol{R}^N), \quad \varphi \in W^{3,\infty}(\boldsymbol{R}^N), \quad \sup\limits_{a \in \mathcal{A}} \|\psi(\cdot, a)\|_{W^{2,\infty}} < \infty, \\ \psi(x, \cdot) \in C(\mathcal{A}) \text{ for all } x \in \boldsymbol{R}^N, \text{ for } \psi = \sigma_{ij}, b_i, c, f; \end{cases} \tag{16}$$

$$\lambda = \inf\{c(x, a) / x \in \boldsymbol{R}^N, \ a \in \mathcal{A}\} > 0. \tag{17}$$

We now want to study the value function u and to show that, in a suitable sense, u satisfies and is characterized by (1):

$$\sup_{a \in \mathcal{A}} [A_a u - f_a] = 0 \text{ in } \mathcal{O}, \tag{1}$$

where $f_a(x) = f(x, a)$, $A_a = -a_{ij}(x, a)\partial_{ij} - b_i(x, a)\partial_i + c(x, a)$ and $a = \frac{1}{2}\sigma\sigma^T$.

In the next section we *justify the derivation* of (1) by showing that u is the unique viscosity solution of (1), provided $u \in C(\mathcal{O})$. Therefore we first have to show that $u \in C(\mathcal{O})$; the following result is taken from P.-L. Lions [28] (where much more general results are given). We will assume (for simplicity) that $\partial\mathcal{O} = \Gamma_- \cap \Gamma_0$ where Γ_-, Γ_0 are disjoint, closed, possibly empty subsets of $\partial\mathcal{O}$, and

$$a_{ij}(x, a) n_i(x) n_j(x) = 0, \quad b_i(x, a) n_i(x) - a_{ij}(x, a)\partial_{ij} d(x) \leqslant 0 \quad \text{on } \Gamma_- \times \mathcal{A}, \tag{18}$$

where n is the unit outward normal and $d(x) = \operatorname{dist}(x, \partial\mathcal{O})$. We will also assume that

$$\exists w \in W^{1,\infty}(\mathcal{O}), \quad A_a w \leqslant f_a \text{ in } \mathcal{D}'(\mathcal{O}) \ \forall a \in \mathcal{A}, \ w = \varphi \text{ on } \Gamma_0; \tag{19}$$

$$\exists C > 0, \forall (x, y) \in \bar{\mathcal{O}} \times \Gamma_0, \exists \mathcal{S}: \ J(x, \mathcal{S}) \leqslant \varphi(y) + C|x - y|. \tag{20}$$

THEOREM II.1. *Under assumptions* (18)–(20), $u \in C^{0,\theta}(\bar{\mathcal{O}})$ *and* $u = \varphi$ *on* Γ_0; *here* $\theta = \lambda/\lambda_0$ *if* $\lambda < \lambda_0$, θ *arbitrary in* $(0, 1)$ *if* $\lambda = \lambda_0$ *and* $\theta = 1$ *if* $\lambda > \lambda_0$ *with* $\lambda_0 = \sup\{\frac{1}{2}\operatorname{Tr}(\partial_\xi \sigma \cdot \partial_\xi \sigma^T) + \partial_\xi b \cdot \xi / \ x \in \bar{\mathcal{O}}, \ a \in \mathcal{A}, \ |\xi| = 1\}$.

Remark II.1. In P. L. Lions [38], it is shown that conditions (18)–(20) are very natural. Certain extensions are also given and examples of situations where (19)–(20) hold are indicated.

Remark II.2. Of course, if $\mathcal{O} = \boldsymbol{R}^N$, (18)–(20) are vacuous and (19) holds.

Remark II.3. The exponent θ given above is optimal as it is shown in the following example.

Example II.1. Take $\mathcal{O} =]-1, +1[$, $\sigma \equiv 0$, $b(x, \alpha) \equiv x$, $f \equiv 0$, $\varphi \equiv 1$, $c \equiv \lambda$. Then one checks easily that $u(x) = |x|^\lambda$ and $\lambda_0 = 1$! ∎

The proof of this result uses probabilitic and analytic arguments based on the dynamic programming principle.

We conclude this section by a result yielding one possible characterization of the value function in terms of maximum subsolution

THEOREM II.2. *The value function u satisfies: $\sigma^T \cdot \nabla u \in L^2_{loc}(\mathcal{O})$ and*

$$A_\alpha u \leqslant f_\alpha \text{ in } \mathcal{D}'(\mathcal{O}), \quad \forall \alpha \in \mathcal{A}; \tag{21}$$

i.e. u is a subsolution of the HJB *equation. In addition it is the maximum one in the sense that if $v \in C(\bar{\mathcal{O}})$ satisfies* (21) *and* $\overline{\lim_{d(x)\to 0}} (v-u) \leqslant 0$ *then $v \leqslant u$ in $\mathcal{O}$.*

II.3. Viscosity solutions and HJB equations. The following result shows that, boundary conditions being prescribed, the value function is the unique viscosity solution of (1).

THEOREM II.3. *If $u \in C(\bar{\mathcal{O}})$, then u is a viscosity solution of equation* (1).

Conversely, if $v \in C(\bar{\mathcal{O}})$ is a viscosity solution of (1) *satisfying: $v = u$ on $\partial\mathcal{O}$ then $v \equiv u$ in $\bar{\mathcal{O}}$.*

This result, proved in P. L. Lions [38], justifies the derivation of the HJB equation and shows that (1) characterizes the value function.

Sketch of proof. The fact that u is a viscosity solution of (1) is obtained by using the so-called *optimality principle* and by remarking that, taking advantage of the definition of viscosity solutions, one may replace u by smooth test functions ψ and thus one may perform on ψ the 'usual' derivation of (1).

The converse statement is proved by probabilistic considerations and careful choices of test functions ψ.

II.4. Further results. In P. L. Lions [38], [40] these results are extended to more general situations and to other control problems (optimal stopping, time-dependent problems...). It is also possible to show that if (18)–(19) hold then one may restrict the infimum to admissible systems, where the probability space and the Brownian motion are fixed. Results

concerning the density of Markovian controls are also given. Let us finally mention that in P. L. Lions and M. Nisio [47], Theorem II.3 is used to derive a general uniqueness result for nonlinear semi-groups.

III. Regularity of the value function

III.1. Regularity results. In this section, we follow and extend the approach of N. V. Krylov [26], [27] concerning the verification of the HJB equation in a more usual sense. The idea is first to obtain some regularity result on u and then to check (1). Since we do not want to impose non-degeneracy assumptions, we need to assume that $\partial\mathcal{O} = \Gamma_- \cap \Gamma_+$, where Γ_-, Γ_+ are closed disjoint, possibly empty subsets of $\partial\mathcal{O}$, and

$$\begin{aligned} &\exists \nu > 0, \quad \forall (x, \alpha) \in \Gamma_+ \times \mathcal{A} \\ &\text{either} \quad a_{ij}(x, \alpha) n_i(x) n_j(x) \geqslant \nu > 0 \\ &\text{or} \quad a_{ij}(x, \alpha) n_i(x) n_j(x) = 0, \quad b_i(x, \alpha) n_i(x) - a_{ij}(x, \alpha)\, \partial_{ij} d(x) \geqslant \nu. \end{aligned} \tag{22}$$

Write

$$\lambda_1 = \sup \{2 \,|\, \partial_i \sigma^T \cdot \xi_i|^2 + \operatorname{Tr}(\partial_i \sigma \cdot \partial_i \sigma^T) + 2 \partial_i b \cdot \xi \xi_i / \\ x \in \boldsymbol{R}^N, \alpha \in \mathcal{A}, |\xi| = 1\}.$$

THEOREM III.1. *Under the assumptions* (18), (22), *if* $\lambda > \lambda_1$ *then the value function* u *belongs to* $W^{1,\infty}(\mathcal{O})$ *and satisfies* $u = \varphi$ *on* $\Gamma_1 \cup \Gamma_2$ *and*

$$u \text{ is semi-concave in } \mathcal{O}, \text{ i.e.: } \exists C > 0, \ \partial_\xi^2 u \leqslant C \text{ in } \mathcal{D}'(\mathcal{O}), \ \forall |\xi| = 1. \tag{23}$$

COROLLARY III.1. *Under the assumptions of Theorem* III.1, u *satisfies*

$$A_\alpha u \in L^\infty(\mathcal{O}) \quad \text{and} \quad \sup_{\alpha \in \mathcal{A}} \|A_\alpha u\|_{L^\infty(\mathcal{O})} < \infty, \tag{24}$$

and the HJB *equation holds a.e.:*

$$\sup_\alpha \{A_\alpha u - f_\alpha\} = 0 \quad \text{a.e. in } \mathcal{O}.$$

COROLLARY III.2. *Under the assumptions of Theorem* III.1. *if there exist* $p \in \{1, \ldots, N\}$, $\nu > 0$, *and an open set* ω *contained in* $\mathcal{O}$ *such that for all* $x \in \omega$ *we can find* $n \geqslant 1$, $\theta_1, \ldots, \theta_n \in]0, 1[$, $\alpha_1, \ldots, \alpha_n \in \mathcal{A}$ *satisfying:*

$$\sum_{i=1}^n \theta_i = 1, \quad \sum_{i=1}^n \theta_i a_{kl}(x, \alpha_i) \xi_k \xi_l \geqslant \nu \sum_{j=1}^p \xi_j^2 \quad \forall \xi \in \boldsymbol{R}^N,$$

then $\partial_{ij} u \in L^\infty(\omega)$ *for* $1 \leqslant i, j \leqslant p$. ∎

A slightly weaker form of these results was first given in P. L. Lions [42] and the above results appear in P. L. Lions [38] (see also [34]).

Sketch of proof. The estimate (23) implies (24) and Corollary III. 2 in a straightforward way, and the fact that the HJB equation holds a.e. is a direct consequence of (23)–(24) and the fact that the value function is a viscosity solution of (1). Next, the proof of (23) is obtained by new a priori estimates of two kinds; (i) a boundary estimate obtained by a p.d.e. device, (ii) an interior estimate obtained by a probabilistic method.

Remark III.1. In general, the assumption $\lambda > \lambda_1$ is necessary in order to obtain (23) as is shown by Example II.1: in this example $\lambda_1 = 2$ and $u(x) = |x|^\lambda$; now observe that u satisfies (23) if and only if $\lambda \geqslant \lambda_1$. ∎

However, in the uniformly elliptic case, it is possible to assume only $\lambda \geqslant 0$

THEOREM III.2. *In the uniformly elliptic case, i.e., in the case where*

$$\exists \nu > 0, \quad a(x, \alpha) \geqslant \nu I_N \qquad \forall (x, \alpha) \in \bar{\mathcal{O}} \times \mathcal{A} \tag{25}$$

we have $u \in W^{2,\infty}(\mathcal{O})$.

This result was first proved in P. L. Lions [43] with the additional assumption $\lambda > \lambda_1$; in [43] a new a priori estimate method was introduced. This method was simplified in L. C. Evans and P. L. Lions [21] and Theorem III.2 was proved in [21].

An additional regularity result was recently obtained by L. C. Evans [18], [19];

THEOREM III.3 (L. C. Evans). *Under the assumptions of Theorem* III.2, $u \in C^{2,\theta}(\mathcal{O})$ *for some* $\theta \in]0, 1[$.

III.2. Uniqueness results. In the proceding section we obtained regularity results which ensure that the HJB holds a.e. One may ask if this yields a characterization of the value function. The following example shows that if $\tilde{u} \in W^{1,\infty}$ satisfies (24) and the HJB equation a.e., it need not be identical with u, as is shown by the following example:

Example III.1. Take $\sigma \equiv 0$, $b \equiv \alpha$, $c \equiv \lambda$, $f \equiv 1$, $\mathcal{A} = \{\alpha \in \boldsymbol{R}^N | |\alpha| \leqslant 1\}$. Then clearly $u \equiv 1$, but for all β, $x_0 \in \boldsymbol{R}_+ \times \boldsymbol{R}^N$

$$\tilde{u}(x) = \frac{1}{\lambda}\left(1 - \beta \exp\left(-\lambda |x - x_0|\right)\right)$$

satisfies (24) and the HJB equation: $|D\tilde{u}| + \lambda \tilde{u} = 1$ everywhere except at x_0.

Therefore we need some extra condition in order to characterize the value function. We have

THEOREM III.4. *Let* $\tilde{u} \in C(\bar{\mathcal{O}}) \cap W^{1,N}_{\text{loc}}(\mathcal{O})$ *satisfy*: $\tilde{u} = u$ *on* $\partial\mathcal{O}$ *and*

$$A_\alpha \tilde{u} \leqslant f_\alpha \quad \text{in } \mathcal{D}'(\mathcal{O}), \qquad \forall \alpha \in \mathcal{A}; \tag{21}$$

$$\sup_{\alpha \in \mathcal{A}} \{A_\alpha \tilde{u} - f_\alpha\} = 0 \quad \text{in the sense of measures}; \tag{26}$$

$$\exists g \in L^N_{\text{loc}}(\mathcal{O}), \qquad \Delta \tilde{u} \leqslant g \quad \text{in } \mathcal{D}'(\mathcal{O}). \tag{27}$$

Then $\tilde{u} \equiv u$ *in* $\bar{\mathcal{O}}$.

Remark III.1. Of course, if $\mathcal{O} = \boldsymbol{R}^N$, no boundary conditions are needed. Instead, we assume, for example $\tilde{u} \in C_b(\boldsymbol{R}^N)$.

Observe that condition (27) appears to be quite sharp since in Example III.1, $(\Delta \tilde{u})^+ \in L^p(\boldsymbol{R}^N)$ for all $p < N (\in M^N(\boldsymbol{R}^N))$.

Sketch of the proof. The proof of this Theorem has two ingredients: (i) the probabilistic L^p-estimates due to N. V. Krylov [26], (ii) careful bounds on the commutation of regularizing kernels and operators with variable coefficients.

III.3. Further results. In the bibliography we give various references concerning previous versions of Theorems II. 2–3 and the analogous treatment of related problems such as optimal stopping, impulse control, time-dependent problems, or other boundary conditions, and the numerical approximation of HJB equations... Let us also mention that in the case where $\mathcal{O} = \boldsymbol{R}^N$, Theorem III.1 was obtained independently by N. V. Krylov [28], [29] and the author [41].

IV. Monge–Ampère equations

IV.1. Relations with HJB equations. Let us now explain how the Monge–Ampère equation

$$\det(D^2 u) = g(x) \quad \text{in } \mathcal{O}, \quad u \text{ convex on } \bar{\mathcal{O}}, \; u = \varphi \text{ on } \partial\mathcal{O}$$

is related to HJB equations. This relation was discovered independently by B. Gaveau [24] and N. V. Krylov [30] and is given by the following algebraic observation: if A is an $N \times N$ nonnegative symmetric matrix, then

$$(\det A)^{1/N} = \inf\{\text{Tr}(AB) \mid B \geqslant 0, B = B^T, \det(B) = 1/N^N\}.$$

Therefore the above equation is equivalent to the HJB equation:

$$\sup_{B\in\mathscr{B}}[-b_{ij}\partial_{ij}u] = -g^{1/N} \quad \text{in } \mathcal{O}, \qquad u = \varphi \text{ on } \partial\mathcal{O}$$

where $B = (b_{ij})$, $\mathscr{B} = \{B \geqslant 0, B = B^T, \det B = 1/N^N\}$.

IV.2. Existence and regularity results. In differential geometry, the question of the existence of smooth convex hypersurfaces with various prescribed curvatures, such as e.g. the Gaussian curvature, leads to the following Monge–Ampère equations:

$$\det(D^2u) = H(x, u, \nabla u) \quad \text{in } \mathcal{O}, \qquad u \text{ convex on } \bar{\mathcal{O}}, \quad u = \varphi \text{ on } \partial\mathcal{O}, \tag{3}$$

where we assume, for instance, $\varphi \in W^{2,\infty}(\boldsymbol{R}^N)$, $H \in C^\infty(\mathcal{O}\times\boldsymbol{R}\times\boldsymbol{R}^N)$, $\mathcal{O}$ being a bounded convex domain in $\boldsymbol{R}^N$ $(N \geqslant 2)$.

We will assume that H satisfies:

$$\forall R < \infty, \quad \exists \nu > 0, \quad H(x, t, p) \geqslant \nu > 0 \quad \text{for } x \in \bar{\mathcal{O}},\ |t| \leqslant R,\ p \in \boldsymbol{R}^N; \tag{28}$$

and that there exists $w \in C(\bar{\mathcal{O}})$ satisfying the following inequality in Alexandrov sense (cf. [1], [12]) — or in viscosity sense —

$$\det(D^2w) \geqslant H(x, w, Dw) \quad \text{in } \mathcal{O}, \qquad w \text{ convex in } \mathcal{O},\ w = \varphi \text{ on } \partial\mathcal{O}. \tag{29}$$

THEOREM IV.1. *Under assumptions* (28)–(29), *there exists a minimum solution of* (3) *in* $C^\infty(\mathcal{O})\cap C(\bar{\mathcal{O}})$ *satisfying*: $u \geqslant w$ *in* $\bar{\mathcal{O}}$.

In addition, if H *is non-decreasing with respect to* t, u *is the unique solution of* (3). ∎

Remark IV.1. Variants and extensions of this result are to be found in P. L. Lions [45]; in particular the verification of (29) is discussed by the use of the results of Bakelman [2].

Let us recall that the case of $H = H(x, t)$ was studied by Pogorelov [50] and that a complete proof of the existence of smooth solutions was first given by S. Y. Cheng and S. T. Yau [12] using geometric arguments. The proof of this theorem is given in P.L. Lions [45]; it is based on a new approximation method of (3) by problems in $\boldsymbol{R}^N$, relying on the idea of "penalizing the domain $\mathcal{O}$". The approximated problem may then be solved by using the relations with HJB equations and the results of Section III.

Finally, uniform a priori estimates are derived by the use of the classical Pogorelov estimates [50] and Calabi estimates [9]. In conclusion, let us point out that, when $H = H(x, t)$, L. Cafarelli, L. Nirenberg and J. Spruck [8] recently showed that $u \in C^\infty(\bar{\mathcal{O}})$ and that using their method, one is able to show that, in the above result, if $w \in W^{1,\infty}(\mathcal{O})$ then $u \in C^\infty(\bar{\mathcal{O}})$.

Notes added in proofs: We saw recently that Theorem III.3 has been obtained independently by N. V. Krylov (*Izv. Mat. Ser.* **46** (1982), pp. 487–523).

Furthermore, the $C^{2,0}(\bar{\mathcal{O}})$ regularity has been obtained and applied to the Monge–Ampère equations independently by N. V. Krylov (*Mat. Sbornik* **120** (1983), pp. 311–330; *Izv. Mat. Ser.* **47** (1983), pp. 75–108) and by L. Caffarelli, J. J. Kohn, L. Nirenberg and J. Spruck.

Bibliography

[1] Alexandrov A. D., Dirichlet's problem for the equation Det $||z_{ij}|| = \Phi(z_1, \ldots, z_n, z, x_1, \ldots, x_n)$, I, *Vestnik Leningrad. Univ. Ser. Math. Mekh. Astr.* **13** (1958), pp. 5–24 (in Russian).

[2] Bakelman I., *Generalized solutions of the Direchlet problem for the N-dimensional elliptic Monge–Ampère equation*, preprint.

[3] Ball J. and Evans L. C., Weak convergence theorems for nonlinear PDE of first and second order, *Bull. London Math. Soc.*, **251** (1982), pp. 332–346.

[4] Barles G., *Thèse de 3e cycle*, Univ. Paris IX – Dauphine, 1983.

[5] Barron E. M., Evans L. C., and Jensen R., *Viscosity solutions of Isaac's equations and differential games with Lipschitz controls*, preprint.

[6] Bellman R., *Dynamic Programming*, Princeton Univ. Press, Princeton, N. J., 1957.

[7] Brézis H. and Evans L. C., A variational inequality approach to the Bellman–Dirichlet equation for two elliptic operators, *Arch. Rat. Mech. Anal.* **71** (1979), pp. 1–13.

[8] Caffarelli L., Nirenberg L., and Spruck J., The Dirichlet problem for non-linear second order elliptic equations. I. To appear in *Comm. Pure Appl. Math.*

[9] Calabi E., Improper affine hyperspheres of convex type and a generalization of a theorem by K. Jörgens, *Michigan Math. J.* **5** (1958), pp. 105–126.

[10] Capuzzo-Dolcetta I., On a discrete approximation of the Hamilton–Jacobi equation of dynamic programming. *Appl. Math. Optim.*, 1983.

[11] Capuzzo-Dolcetta I. and Evans L. C., *Optimal switching for ordinary differential equations*, preprint.

[12] Cheng S. Y. and Yau S. T., On the regularity of the Monge-Ampère equation det $(\partial^2 u/\partial x_i \partial x_j) = F(x, u)$, *Comm. Pure Appl. Math.* **30** (1977), pp. 41–68.

[13] Crandall M. G., Evans L. C., and Lions P. L., Some properties of viscosity solutions of Hamilton–Jacobi equations, *Trans. Amer. Math. Soc.*, 1984.

[14] Crandall M. G. and Lions P. L., Viscosity solutions of Hamilton–Jacobi equations, *Trans. Amer. Math. Soc.*, **277** (1983), pp. 1–42; announced in *C. R. Acad. Sci. Paris*, **252** (1981), pp. 183–186.

[15] Crandall M. G. and Lions P. L., Two approximations of solutions of Hamilton–Jacobi equations. To appear in *Math. Comp.*

[16] Evans L. C., A convergence theorem for solutions of nonlinear second order elliptic equations, *Ind. Univ. Math. J.* **27** (1978), pp. 875–887.

[17] Evans L. C., On solving certain nonlinear PDE by accretive operator methods, *Israel J. Math.* **36** (1980), pp. 225–247.

[18] Evans L. C., Classical solutions of fully nonlinear, convex, second-order elliptic equations, *Comm. Pure Appl. Math.* **35** (1982), pp. 333–363.

[19] Evans L. C., Classical solutions of the Hamilton–Jacobi–Bellman equation for uniformly elliptic operators, *Trans. Amer. Math. Soc.* **121** (1983), pp. 211–232.

[20] Evans L. C. and Friedman A., Optimal stochastic switching and the Dirichlet problem for the Bellman equation, *Trans. Amer. Math. Soc.* **253** (1979), pp. 365–389.

[21] Evans L. C. and Lions P. L., Résolution des équations de Hamilton–Jacobi–Bellman pur des opérateurs uniformément elliptiques, *C. R. Acad. Sci. Paris* **290** (1980), pp. 1049–1052.

[22] Fleming W. H., The Cauchy problem for a nonlinear first-order partial differential equation, *J. Diff. Eq.* **5** (1969), pp. 515–530.

[23] Fleming W. H. and R. Rishel R., *Deterministic and stochastic optimal control*, Springer, Berlin 1975.

[24] Gaveau B., Méthode de contrôle optimal en analyse complexe. I, *J. Funct. Anal.* **25** (1977), pp. 391–411.

[25] Genis I. I. and Krylov N. V., An example of a one-dimensional controlled process, *Th. Proba. Appl.* **21** (1976), pp. 148–152.

[26] Krylov N. V., *Controlled diffusion processes*, Springer, Berlin 1980.

[27] Krylov N. V., Control of a solution of a stochastic integral equation, *Th. Proba. Appl.* **17** (1972), pp. 114–131.

[28] Krylov N. V., Control of the diffusion type processes, *Proceedings of the International Congress of Mathematicians, Helsinki*, 1978.

[29] Krylov N. V., Some new results in the theory of controlled diffusion processes, *Math. USSR Sbornik* **37** (1980), pp. 133–149.

[30] Krylov N. V., On control of the solution of a stochastic inegral equation with degeneration, *Math. USSR Izv.* **6** (1972), pp. 249–262.

[31] Krylov N. V., On control of diffusion processes up to the first exit time from an open set, *Izv. Akad. Nauk.* **45** (1981), pp. 1029–1048 (in Russian).

[32] Lenhart S., *Partial differential equations from dynamic programming equations*, Ph. D. Univ. of Kentucky–Lexigton, 1981.

[33] Lions P. L., *Generalized solutions of Hamilton–Jacobi equations*, Pitman, London 1982.

[34] Lions P. L., Optimal control of diffusion type processes and Hamilton–Jacobi–Bellman equations, in: *Advances in Filtering and Optimal Stochastic Control*, Eds. Flemming and Gorostiza, Springer, Berlin 1982.

[35] Lions P. L., Existence results of first-order Hamilton–Jacobi equations, *Ric. Mat.*, **32** (1983), pp. 3–23.

[36] Lions P. L., Articles to appear in the *Encyclopedia of systems and control*, Pergamon, Oxford.

[37] Lions P. L., Fully nonlinear elliptic equations and applications, in: *Nonlinear Analysis, Function Spaces and Applications*, Vol. 2, Teubner, Leipzig 1982.

[38] Lions P. L., Optimal control of diffusion processes and Hamilton–Jacobi–Bellman equations. Parts 1 and 2, *Comm. P.D.E.* **8** (1983), pp. 1101–1134, 1229–1276. Part 3; to appear in: *Nonlinear Partial Differential Equations and their applications*, Collége de France Seminar, Vol. V, Pitman, London 1983; announced in *C. R. Acad. Sci. Paris* **295** (1982), pp. 567–670.

[39] Lions P. L., A remark on Bony maximum principle, *Proc. Amer. Math. Soc.* **88** (1983), pp. 503–508.

[40] Lions P. L., Some recent results in the optimal control of diffusion processes, in: *Proceedings on Stochastic Analysis*, Kinokuniya, Tokyo, 1984.

[41] Lions P. L., Control of diffusion processes in $\boldsymbol{R}^N$, *Comm. Pure Appl.* **34** (1981), pp. 121–147; announced in *C. R. Acad. Sci. Paris* **288** (1979), pp. 339–342.

[42] Lions P. L., Equations de Hamilton–Jacobi–Bellman dégénérées, *C. R. Acad. Sci. Paris* **289** (1979), pp. 329–332.

[43] Lions P. L., Résolution analytique des problèmes de Bellman–Dirichlet, *Acta Math.* **146** (1981), pp. 151–166; announced in *C. R. Acad. Sci. Paris* **287** (1978), pp. 747–750.

[44] Lions P. L., Bifurcation and optimal stochastic control, *Nonlinear Anal. T. M. A.*

[45] Lions P. L., Sur les equations de Monge–Ampère. I, *Manuscripta Math.*, 1983; II, to appear in *Arch. Rat. Mech. Anal.*, announced in *C. R. Acad. Sci. Paris* **293** (1981), pp. 589–592.

[46] Lions P. L. and Menaldi J. L., Optimal control of stochastic integrals and Hamilton–Jacobi–Bellman equation, I, *S.I.A.M.J. Control. Optim.* **20** (1982), pp. 58–81, II, *S.I.A.M.J. Control. Optim.* **20** (1982), pp. 82–95.

[47] Lions P. L. and Nisio M., A uniqueness result for the semigroup associated with the Hamilton–Jacobi–Bellman operator, *Proc. Japan Acad.* **58** (1982), pp. 273–276.

[48] Lions P. L., Papanicolaou G., and Varadhan S. R. S., work in preparation.

[49] B. Perthame, Thèse de 3e cycle, Univ. P. et M. Curie, Paris 1983.

[50] A. N. Pogorelov, *On the Minkowski multidimensional problem*, J. Wiley, New-York, 1978.

[51] M. V. Safonov, On the Dirichlet problem for Bellman's equation in a plane domain, Math. USSR Sbornik **31** (1977), pp. 231–248, et **34** (1978), pp. 531–526.

[52] P. E. Souganidis, Ph. D. Univ. of Wisconsin–Madison, 1983.

[53] N. S. Trudinger, *Elliptic equations in non-divergence form*, Proc. Miniconf. P. D. E., Canberra 1981.

[54] N. S. Trudinger, Fully nonlinear, uniformly elliptic equations under natural structure conditions, *Trans. Amer. Math. Soc.* 1983.

CEREMADE
UNIVERSITÉ PARIS IX – DAUPHINE
PLACE DE LATTRE DE TASSIGNY
75775 PARIS CEDEX 16

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

R. TYRRELL ROCKAFELLAR*

Differentiability Properties of the Minimum Value in an Optimization Problem Depending on Parameters

1. A central topic in optimization theory is the study of the optimal value and optimal solution set

$$p(v) = \inf_{x \in A(v)} f(v, x), \qquad X(v) = \operatorname{argmin}_{x \in A(v)} f(v, x), \tag{1}$$

in an optimization problem over $x \in R^n$ which depends on a parameter vector $v \in R^d$. Often this is a prelude to minimizing or maximizing $p(v)$ subject to further constraints on v, as is the case for instance in decomposition schemes in mathematical programming and various problems of approximation or engineering design. The question of the possible continuity and differentiability properties of the function p is then very important. Such properties also turn out to be critical in the derivation of optimality conditions which characterize the points $x \in X(v)$.

Let us normalize by focusing on behavior around $v = 0$. Assume that $A(0) \neq \emptyset$, the function $f\colon R \times R^n$ is locally Lipschitz continuous, the set $\operatorname{gph} A = \{(v, x) \mid x \in A(v)\} \subset R^d \times R^n$ is closed, and that for some $\varepsilon > 0$ the set

$$\{(v, x) \mid |v| \leqslant \varepsilon, x \in A(v), f(v, x) \leqslant \alpha\}$$

is bounded for every $\alpha \in R$. Then p is lower semicontinuous on a neighborhood of $v = 0$ with $p(0)$ finite and $X(0)$ nonempty and compact. Our aim is to clarify the circumstances under which p is actually Lipschitz continuous in a neighborhood of $v = 0$ and has directional derivatives of various sorts.

---

* Research supported by the Air Force Office of Scientific Research, U.S.A.F., under grant No. F49620-82-K-0012.

2. Ordinary one-sided directional derivatives

$$p'(v; h) = \lim_{t \to 0_+} [p(v+th) - p(v)]/t \tag{2}$$

exist only in rather special cases. One such case, among the first to be identified, is that in which $f \in \mathscr{C}^1$ and $A(v)$ is a fixed set B for all v. Then

$$p'(0; h) = \min_{x \in X(0)} \nabla_v f(v, x) \cdot h.$$

The result can be attributed to Danskin [5], although the form in which we have stated it is somewhat different. If B is convex, the condition $x \in X(0)$ is equivalent, of course, to $-\nabla_x f(v, x) \in N_B(x)$ where $N_B(x)$ is the normal cone to B in the sense of convex analysis ([16]).

An example where this applies is

$$f(v, x) = \sum_{j=1}^{n} x_j g_j(v), \quad A(v) \equiv B = \Big\{x = (x_1, \ldots, x_n) \mid x_j \geqslant 0, \sum_{j=1}^{n} x_j = 1\Big\},$$

with $g_j \in \mathscr{C}^1$. Then $p(v) = \min\{g_1(v), \ldots, g_n(v)\}$.

The case where f is a convex function and $\operatorname{gph} A$ is a convex set has also received attention. Then p is a convex function, so the derivatives $p'(0; h)$ do exist. It has been shown by Golshtein [10] (see also Hogan [13]) that for any choice of $x \in X(0)$ one has

$$p'(v; h) = \inf_{k \in A'(v,x;h)} f'(v, x; h, k),$$

where $\operatorname{gph} A'$ is the tangent cone to $\operatorname{gph} A$ at (v, x). In terms of the subgradients of convex analysis ([16]), the equivalent formula is

$$\partial p(v) = \{z \in R^d \mid (z, 0) \in \partial f(0, x) + N_{\operatorname{gph} A}(0, x)\}. \tag{4}$$

Generalizations of (3) to nonconvex cases have been given by Dem'janov *et al.* [6], [7], under rather stringent assumptions. Other results along these lines are those of Hiriart-Urruty [11], the marginal value theorem of Golshtein [10] for nonconvexly parameterized convex programming, and certain extensions of the latter by Rockafellar [17], [21, Theorem 4].

3. More general results of the kind just mentioned involve additional structure for the constraint set $A(v)$. In escaping from assumptions of either classical differentiability or convexity, such results also rely on new developments in subgradient analysis.

Suppose henceforth that

$$A(v) = \{x \in R^n \mid F(v, x) \in C,\ (v, x) \in D\},$$

where $C \subset R^m$ and $D \subset R^d \times R^n$ are closed sets and F: $R^d \times R^n \to R^m$ is locally Lipschitz continuous. A typical case in mathematical programming is

$$C = \{(u_1, \ldots, u_m) \mid u_i \leqslant 0 \text{ for } i = 1, \ldots, s,$$
$$u_i = 0 \text{ for } i = s+1, \ldots, m\}. \quad (6)$$

For a locally Lipschitz continuous function g: $R^n \to R$, Clarke ([2]) introduced the directional derivatives

$$g^0(x; k) = \limsup_{\substack{x' \to x \\ t \downarrow 0}} [g(x' + tk) - g(x')]/t$$

and showed there was a unique, nonempty, compact convex set $\partial g(x)$ (whose elements may be called "subgradients") such that

$$g^0(x; k) = \max_{w \in \partial g(x)} k \cdot w.$$

A detailed calculus has grown out of this concept; see Clarke [2], [3], [4], Hiriart-Urruty [11], and Rockafellar [15], [18], [19], [20], [21] in particular. It is known that $g^0(x; k) = g'(x; k)$ when $g \in \mathscr{C}^1$ or g is convex; in the first case $\partial g(x)$ reduces to the gradient $\nabla g(x)$, while in the second case it is the usual subgradient set of convex analysis.

Corresponding geometrically to Clarke's notion of "subgradient" is his definition of the normal cone $N_B(x)$ to an arbitrary closed set $B \subset R^n$ at any point $x \in B$; see [2], [15].

These concepts have been used by Clarke [1] to derive optimality conditions for mathematical programming problems with objective and constraint functions that are locally Lipschitz continuous, and Clarke's result has been sharpened by Hiriart-Urruty [12] and Rockafellar [21]. As background for the marginal value theorem that will be stated below, we first formulate a version of this result for the more general constraint structure in (5). Let

$$K(x) = \{(y, z) \in R^m \times R^d \mid y \in N_C\big(F(0, x)\big),$$
$$(z, 0) \in \partial(f + y \cdot F)(x) + N_D(0, x)\},$$
$$K_0(x) = \{(y, z) \in R^m \times R^d \mid y \in N_C\big(F(0, x)\big),\ (z, 0) \in \partial(y \cdot F)(x) + N_D(0, x)\}.$$

THEOREM 1 (*Multiplier Rule*). *Suppose* $x \in X(0)$ *is such that* $K_0(x)$ *contains just* $(0, 0)$. *Then there is a pair* $(y, z) \in K(x)$, *in fact* $K(x)$ *is a nonempty compact set.*

The constraint qualification $K_0(x) = \{(0, 0)\}$ reduces in the case of

$$f \in \mathscr{C}^1, \quad F \in \mathscr{C}^1, \quad D = R^d \times R^n, \quad C \text{ as in (6)},$$

to the well-known one of Mangasarian and Fromovitz [14].

Theorem 1 may be derived from Theorem 1 of Rockafellar [21] by applying the latter to the constraints

$$0 = G(v, x, w) = F(v, x) - w, \qquad (v, x, w) \in D \times C.$$

By the same route one obtains the following as a special case of Theorem 2 of Rockafellar [21].

THEOREM 2. *Suppose that for every* $x \in X(0)$, *the set* $K_0(x)$ *contains just* $(0, 0)$. *Then* p *is Lipschitz continuous in a neighborhood of* 0 *and*

$$\partial p(0) \subset \operatorname{co} \bigcup_{x \in X(0)} \{z \mid \exists y, (y, z) \in K(x)\},$$

$$p^0(0; h) \leqslant \max_{\substack{x \in X(0) \\ (y,z) \in K(x)}} z \cdot h.$$

In the case of assumption (7), this result was proved by Gauvin ([8], [9]).

References

[1] Clarke F. H. A New Approach to Lagrange Multipliers, *Math. Oper.Res.* **1** (1976), pp. 165–174.

[2] Clarke F. H., Generalized Gradients and Applications, *Trans. Amer. Math. Soc.* **205** (1975), pp. 247–262.

[3] Clarke F. H., Generalized Gradients of Lipschitz Functionals, *Adv. in Math.* **40** (1981), pp. 52–67.

[4] Clarke F. H., *Optimization and Nonsmooth Analysis*, Wiley, New York 1983.

[5] Danskin J. M., *The Theory Max-Min and Its Applications to Weapons Allocations Problems*, Springer-Verlag, New York 1967.

[6] Dem'janov V. F. and Malozemov V. N., The Theory of Nonlinear Minimal Problems, *Russian Math. Surveys* **26** (1971), pp. 55–115.

[7] Dem'janov V. F. and Pevnyi A. B., First and Second Marginal Values of Mathematical Programming Problems, *Soviet Math. Dokl.* **13** (1972), pp. 1502–1506.

[8] Gauvin J., The Generalized Gradients of a Marginal Function in Mathematical Programming, *Math. Oper. Res.* **4** (1979), pp. 458–463.

[9] Gauvin J. and Dubeau F., Differential Properties of the Marginal Value Function in Mathematical Programming, *Math. Programming Stud.*, to appear.

[10] Golshtein E. G., *Theory of Convex Programming*, Trans. Math. Monographs 26, Amer. Math. Soc., Providence, RI, 1972.

[11] Hiriart-Urruty J.-B., Gradients-généralises de fonctions marginales, *SIAM J. Control and Optim.* **16** (1978), pp. 301–316.

[12] Hiriart-Urruty J.-B., Refinements of Necessary Optimality Conditions in Nondifferentiable Programming, I, *Appl. Math. Optim.* **5** (1979), pp. 63–82.

[13] Hogan W. W., Directional Derivatives for Extremal-Value Functions with Applications to the Completely Convex Case, *Oper. Res.* **21** (1973), pp. 188–209.

[14] Mangasarian O. L. and Fromovitz S., The Fritz John Necessary Optimality Conditions in the Presence of Equality and Inequality Constraints, *J. Math. Anal. Appl.* **17** (1967), pp. 37–47.

[15] Rockafellar R. T., Clarke's Tangent Cones and the Boundaries of Closed Sets in R^n, *Nonlinear Anal.* **3** (1979), pp. 145–154.

[16] Rockafellar R. T., *Convex Analysis*, Princeton University Press, Princeton, N. J., 1970.

[17] Rockafellar R. T., Directional Differentiability of the Optimal Value Function in a Nonlinear Programming Problem, *Math. Programming Stud.*, **21** (1984).

[18] Rockafellar R. T., Directionally Lipschitzian Functions and Subdifferential Calculus, *Proc. London Math. Soc.* (*3*) **39** (1979), pp. 331–355.

[19] Rockafellar R. T., Generalized Directional Derivatives and Gradients of Nonconvex Functions, *Canad. J. Math.* **32** (1980), pp. 257–280.

[20] Rockafellar R. T., *La theorie des sous-gradients et ses applications: Fonctions convexes et non convexes*, Collection Chaire Aisenstadt Presses de l'Université de Montréal, Montréal, Quebec, 1979. English version: *The Theory of Subgradients and Its Applications: Convex and Nonconvex Functions*, Helderman Verlag, W. Berlin, 1981.

[21] Rockafellar R. T., Lagrange Multipliers and Subderivatives of Optimal Value Functions in Nonlinear Programming, *Math. Programming Stud.* **17** (1982), pp. 28–66.

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

J. ZABCZYK

Stopping problems in stochastic control

The paper presents several characterization and regularity results on stopping for wide classes of Markov processes and general stochastic processes. The emphasis is put on the interplay between the various methods used: analytic, potential theoretic and probabilistic.

A switching strategy is a countable sequence (τ_n, ξ_n) of stopping times $\tau_1 \leqslant \tau_2 \leqslant \ldots$ and U-valued random variables ξ_n, $n \geqslant 1$, defined on a given probability space $(\Omega, \mathscr{F}, P)$ with a specified filtration $(\mathscr{F}_t)_{t \geqslant 0}$. Here $(U, \mathscr{U})$ is a measurable space of control parameters. Random variables ξ_n are assumed to be $\mathscr{F}_{\tau_n}$-measurable. For many stochastic control problems optimal or ε-optimal strategies, turn out to be of the switching type. Impulse control problems, alternating control and stopping games are of this type. Results concerning this subject and diffusion processes were presented by Bensoussan [3] (on impulse control) and by Friedman [12] (on stopping games) at Vancouver ICM–74. In their addresses value functions and optimal strategies were characterized in terms of the associated variational and quasivariational inequalities. The discovery of this relationship revitalized the whole subject of optimal stopping, which had been an area of intensive research in the sixties and the early seventies [30]. A recent account of the results obtained by analytical methods and for solutions of stochastic differential equations can be found in two monographs by A. Bensoussan and J. L. Lions, namely [4] and [5]. However, there are many reasons to go beyond this class of stochastic processes. Questions about the existence of optimal strategies and their structure can be asked and answered in a natural way for wider classes of processes. An important motivation for developing the theory for general Markov processes are problems with partial observations. Moreover, analytical methods require technical assumptions about the differential operators involved which are not so essential when probabil-

istic or semigroup methods are used. Also when reducing alternating control problems to impulse control problems one ends up with a singular Markov process constant in one direction.

In this paper we present several characterization and regularity results on stopping obtained for general stochastic processes and in particular, for Markov processes. One of our aims is to show the interplay between various methods used, which makes the subject so appealing.

I. Stopping problems for Markov processes

Let $X = (\Omega, \mathscr{F}, \mathscr{F}_t, x_t, P^x)$ be a Markov process on a state space $(E, \mathscr{E})$ which is a separable, locally compact metric space. One of the main concepts of the theory is the so-called value function. It is a function defined for an arbitrary initial state in E as the supremum of the gain functional taken with respect to all admissible strategies. An intensively studied property of value functions, which usually implies the existence of optimal dtrategies, is their continuity. To formulate related results we denote by C^b, C, C^0 and C^k the spaces of all continuous functions on E which are, respectively, bounded, with finite limit at infinity, vanishing at infinity snd with compact support. The set of all stopping times for X will be aenoted by $\mathscr{M}$.

I.1. Continuity of value functions for Feller processes. Let X be a Feller process in the sense that for the associated semigroup $(P_t)_{t\geqslant 0}$

$$P_t C^0 \subset C^0, \quad t \geqslant 0, \tag{1}$$

$$P_t f(x) \to f(x) \quad \text{as } t \downarrow 0 \text{ for arbitrary } f \in C^0 \text{ and } x \in E. \tag{2}$$

The following theorem was proved by M. Robin [29] under an additional assumption, which has recently been removed by Ł. Stettner, see [35]. In Robin's proof the penalized equation (4) is the main tool.

THEOREM 1. *If $f \in C^b$ and α is a positive constant, then the value function v,*

$$v(x) = \sup_{\tau \in \mathscr{M}} E^x\left(e^{-\alpha\tau} f(x_\tau)\right), \quad x \in E, \tag{3}$$

is continuous. Moreover, for arbitrary $\beta > 0$ there exists a unique continuous solution v^β of the equation

$$Av^\beta - \alpha v^\beta + \beta(f - v^\beta)^+ = 0, \tag{4}$$

where A is the weak infinitesimal generator of (P_t), and $v^\beta \uparrow v$ uniformly on compact sets as $\beta + \infty$. In addition, the moment

$$\hat{\tau} = \inf\{t \geqslant 0; v(x_t) = f(x_t)\}$$

is an optimal one.

Function v is also called the *α-reduite* of f and denoted by $R_\alpha f$. The continuity part of Theorem 1 was also obtained by Bismut [7] and can be deduced from an old result by Mackevičius [19]. If in the definition of the Feller property one requires instead of (1) a slightly weaker condition, $P_t C \subset C$ for $t \geqslant 0$, then the theorem is no longer true [35]. An important and still unsolved question is to find a proper generalization to the undiscounted case $\alpha = 0$. Even if X is a Feller process on a compact space E and f is continuous, the function $R_0 f$ can be discontinuous [35]. However, one can prove some partial results. In particular, the following proposition holds, see [35]. In its formulation V is the associated potential operator $V = \int_0^{+\infty} P_t dt$.

PROPOSITION 2. *If $V: C^k \to C^0$, then, for all $f \in C$, $v = R_0 f \in C^b$.*

This result can be derived from Theorem 1. However, we will sketch a different proof, based on a duality argument introduced by Bismut [7]. By a discrete-time approximations one can directly show that v is lower semicontinuous. Let $f \in C^0$ and let S denote the set of all excessive, bounded and continuous functions on E. If one shows that, for every $x \in E$, $v(x) = \inf\{h(x); h \geqslant f, h \in S\}$, then the upper semicontinuity of v will follow as well. It is easy to check that the functional F,

$$F(u) = \inf\{h(x); h \geqslant f + u, \quad h \in S\}, \qquad u \in C^0,$$

is convex and continuous at 0. Therefore $F(0) = F^{**}(0)$ and one can calculate that $F^{**}(0) = \sup\{\langle \mu, f\rangle; \mu \in \mathcal{M}(x)\}$ where $\mathcal{M}(x)$ is the set of all non-negative measures μ on E such that $\langle \mu, f\rangle \leqslant f(x)$ for all $f \in S$. The classical Rost theorem then implies that

$$\sup\{\langle \mu, f\rangle; \mu \in \mathcal{M}(x)\} = \sup_{\tau \in \mathcal{M}} E^x(f(x_\tau)),$$

and the required identity follows. The generalization to arbitrary $f \in C$ is straightforward. The method described above has been extended by Bismut to more complex situations of alternating and impulse control problems, see [7] and [8]. Although the assumptions of the theorem are quite restrictive they do not imply a stronger result: if $f \in C^b$ then $R_0 f \in C^b$, see [35].

In the context of Feller processes continuity is the best regularity property of the value function one can ask for. Elementary examples show that even if f is very regular, say if f belongs to the domain of the strong generator, then in some cases $R_\alpha f$ is not even in the domain of the weak generator ($\alpha > 0$).

The question of extending Theorem 1 to controlled Feller processes is discussed in Krylov [17] and Nisio [27]. To use penalization one needs more information on non-linear resolvent operators associated with stochastic control problems (see [5] and [40]). To imitate the proof sketched above one should first generalize the Rost theorem to the controlled case.

An exact counterpart of Theorem 1 for stopping games was obtained by Ł. Stettner ([31] and [32]). If f, g are functions defined on E, $f \geqslant g$ and $\tau, \sigma \in \mathcal{M}$ are arbitrary Markov times then the loss function for a stopping game is given by the formula

$$J_x(\tau, \sigma) = E^x\left(e^{-\alpha\tau\wedge\sigma}\left(f(x_\tau)I_{\tau\leqslant\sigma} + g(x_\sigma)I_{\sigma<\tau}\right)\right), \qquad x \in E.$$

THEOREM 3. *If X is a Feller process and $f, g \in C^b$, then the function w,*

$$w(x) = \inf_{\tau\in\mathcal{M}}\sup_{\sigma\in\mathcal{M}} J_x(\tau, \sigma), \tag{5}$$

is continuous and

$$w(x) = \sup_{\tau\in\mathcal{M}}\inf_{\sigma\in\mathcal{M}} J_x(\tau, \sigma), \qquad x \in E. \tag{6}$$

Moreover, for an arbitrary $\beta > 0$ there is a unique solution $w^\beta \in C^b$ of the equation

$$(A - \alpha)w^\beta + \beta\left((g - w^\beta)^+ - (w^\beta - f)^+\right) = 0 \tag{7}$$

and $w^\beta \to w$ uniformly on compact sets as $\beta \uparrow +\infty$. In addition, the pair $(\hat{\tau}, \hat{\sigma})$,

$$\hat{\tau} = \inf\{t \geqslant 0; f(x_t) = w(x_t)\}, \qquad \hat{\sigma} = \inf\{t \geqslant 0; g(x_t) = w(x_t)\}, \tag{8}$$

is the saddle point of the game.

Arguments based on convexity play an important role in more complex implicit stopping time problems arising in connection with impulse control. Let f and c be non-negative functions defined respectively on E and $E \times E$ and let Γ be a set-valued mapping from E into 2^E. If h is a function on E then Mh is defined by the formula

$$Mh(x) = \inf_{y\in\Gamma(x)}\left(c(x, y) + h(y)\right), \qquad x \in E. \tag{9}$$

A typical implicit stopping time problem consists in finding a function v such that

$$v(x) = \inf_{\tau} E^x\Big(\int_0^{\tau} e^{-\alpha s} f(x_s)\,ds + e^{-\alpha\tau} Mv(x_\tau)\Big). \tag{10}$$

The following theorem is useful in proving the existence and the regularity of a solution of (10). Let $\mathscr{L}$ be a linear space and $\mathscr{K} \subset \mathscr{L}$ a convex cone such that every straight line $L \subset \mathscr{L}$ intersects $\mathscr{K}$ along a closed subinterwal different from L. The cone $\mathscr{K}$ induces an order in $\mathscr{L}$, i.e., $a \leqslant b$ if and only if $b - a \in \mathscr{K}$. Let $\mathscr{A}$ denote a concave and increasing mapping from $\mathscr{K}$ into $\mathscr{K}$ and $\mathscr{K}_0$ the sub-cone $\mathscr{K}_0 = \{v \in \mathscr{K};\ v \leqslant \gamma\, \mathscr{A}^n(0)$ for some $\gamma > 0$ and $n = 1, 2, \ldots\}$. If $v \in \mathscr{K}$ then, for every $w \in \mathscr{L}$, $\|w\|_v = \inf\{\varepsilon > 0;\ -\varepsilon v \leqslant w \leqslant \varepsilon v\}$.

THEOREM 4. *Let $\mathscr{A}$ be a concave and increasing mapping from $\mathscr{K}$ into $\mathscr{K}$. Then*

$$\textit{Equation } v = \mathscr{A}(v) \textit{ has at most one solution in } \mathscr{K}_0. \tag{11}$$

If $\bar{v} \in \mathscr{K}_0$ is a solution of (11) *then for an arbitrary $v \in \mathscr{K}_0$, $\|\mathscr{A}^n(v) - \bar{v}\|_{\bar{v}} \to 0$ geometrically as $n \to +\infty$.*

If for $h \in \mathscr{K}_0$, $\mathscr{A}(h) \leqslant h$, then $(\mathscr{A}^n(h))_{n=1,2,\ldots}$ is a Cauchy sequence in the norm $\|\cdot\|_h$.

This theorem is in the spirit of Krasnoselski's monograph [16]. The first two parts of the theorem are direct extensions of Theorem 4 from the author's paper [38].

A similar result was obtained in Hanouzet and Joly [15]. The idea of applying it to impulse control and quasi-variational inequalities is due to Hanouzet and Joly. Paper [38] dealt with the Riccati equation of the discrete-time regulator problem. Those two different control problems are mathematically similar because both concern controlled Markov chains, see [37] and Doshi [10].

The next theorem is a corollary of Theorem 4 and generalizes slightly a result by Robin [28], who derived it in a different way using an estimate due to Menaldi [22]. To formulate it let $\mathscr{A}$ denote the operation defined by the right-hand side of (10) and $h = V_\alpha f$ the α-potential of f: $V_\alpha f = \int_0^{+\infty} e^{-\alpha t} P_t f\,dt$.

THEOREM 5. *If X is a Feller process $h \in C^b$, $\gamma h \leqslant M(0)$ for a positive γ and M transforms C^b into C^b. Then equation* (11) *has exactly one solution $\hat{v} \in C^b$ and, for arbitrary $v_0 \in C^b$, $v_0 \geqslant 0$, $\mathscr{A}^n(v_0) \to \hat{v}$ uniformly and geometrically fast as $n \uparrow +\infty$.*

Function $\hat{v}$ can be interpreted as the value function of an impulse control problem in which $\Gamma(x)$, $x \in E$ denotes the set of all states to which x can be moved, $c(x, y)$ the cost of a shift from x to y and $f(x)$ the infinitesimal cost at state x. Let $D = \{x\colon \hat{v}(x) = M\hat{v}(x)\}$ and let $d(\cdot)$ be a measurable selector of the multifunction $x \to \{y \in \Gamma(x);\ \hat{v}(x) = c(x, y) + \hat{v}(y)\}$. Then the optimal strategy consists in making impulses indicated by the function d each time the controlled process enters D, see [28] and [5].

I.2. Semigroup characterizations. The previously stated theorems characterized value functions either in terms of penalized equations or appropriate approximation schemes. Different characterizations are based on the concept of the envelope. In particular the α-reduite $R_\alpha f$ of Theorem 1 can be obtained as the smallest function $v \in C^b$ satisfying the following set of inequalities:

$$v \geqslant f, \quad e^{-\alpha t}P_t v \leqslant v \quad \text{for all } t \geqslant 0.$$

The earliest results for evolutionary problems were obtained by Nisio [27]. Theorem 6 below, on an impulse control semigroup, is due to A. Bensoussan and J. L. Lions [5]. Its earlier version under additional assumptions was proved by J. Zabczyk [39]. Let (Q_t) denote the free motion semigroup given by the formula

$$Q_t v = \int_0^t e^{-\alpha s}P_s f\,ds + e^{-\alpha t}P_t v, \quad t \geqslant 0,$$

and assume that (Q_t) is a continuous semigroup acting on the space C^u of bounded and uniformly continuous functions on E. Let M, see (9), transform C^u into C^u. Put $K = \{g \in C^u;\ g \geqslant 0,\ g \leqslant Mg\}$.

THEOREM 6. *There is a continuous semigroup (S_t) acting on K such that*

(1) *S_t are order preserving and contraction operators,*
(2) *$S_t v \leqslant Q_t v$ for all $t \geqslant 0$, $v \in C^u$,*
(3) *(S_t) is the maximal semigroup with properties* (1) *and* (2).

The stochastic control interpretation of (S_t) is given by the formula

$$S_t v(x) = \inf_\pi E_\pi^x\Big(\int_0^t e^{-\alpha s}f(x_s)\,ds + c_t + e^{-\alpha t}v(x_t)\Big),$$

where the infimum is taken with respect to all impulse strategies π and c_t denotes the discounted cost of all impulses performed up to time $t \geqslant 0$.

The proof of Theorem 6 given in [5] was based on penalization. Nisio's approximation scheme was used in [40]. Connection with the Trotter–Kato formula and additional properties of $S(t)$ can be found in L. Barthelemy [2]. The geometrical interpretation of Theorem 6 is as follows. The flow (S_t) is the maximal "restriction" of the flow (Q_t) to the convex subset K of the linear space C^u.

I.3. Variational characterizations. In many instances, stopping time problems are mathematically equivalent to physical problems of minimizing an energy integral. This relationship for diffusion processes was discussed in [4] and [5]. However, similar characterizations take place under the minimal requirements that both problems can be reasonably formulated, which is exactly the setting of symmetric Markov processes and associated Dirichlet spaces. In this generality one can cover some stopping problems with irregular data including diffusion processes with only measurable coefficients or integro-differential generators. The first results were obtained by Krylov [17], then by Nagai [24], who studied also impulse control problems [25]. The game case was discussed in Zabczyk [41]. The following typical result is taken from [41]. We adopt the notation of Theorem 3. Moreover, let X be a symmetric Markov process on E and $(\mathscr{E}, F)$ the associated Dirichlet space, with F densely contained in the Hilbert space $H = L^2(E, m)$, where m is the reference measure. Put $\mathscr{E}_\alpha(\cdot, \cdot) = \mathscr{E}(\cdot, \cdot) + \alpha((\cdot, \cdot))$, where $((\cdot, \cdot))$ is the scalar product on H, and $K = \{u \in F; f \leqslant u \leqslant g \ m - \text{a.e.}\}$.

THEOREM 7. *Let f and g be quasi-continuous elements of F such that $g \geqslant f$ $m-a\,e.$ Then there is a quasi-continuous function $w \in K$ such that*

$$\mathscr{E}_\alpha(u, u) \geqslant \mathscr{E}_\alpha(w, w) \quad \text{for all } u \in K$$

and a properly exceptional set $N \subset E$ for which the identities (5) *and* (6) *hold and the pair $(\hat{\tau}, \hat{\sigma})$ given by* (8) *is a saddle point for all $x \in E \setminus N$.*

Variational counterparts of Theorem 1 and Theorem 5 can be found in [24] and [25].

I.4. More general Markov processes. New and serious problems arise if one drops the assumption that X is a Feller process. To make the fixed point problem (10) meaningful one has to know that operations R_α and $\mathscr{A}$ preserve some kind of regularity. This question has been recently treated by El Karoui [11] and by Lepeltier and Marchal [18]. Let X be

an arbitrary right Markov process. Let $\bar{E}$ be the Ray–Knight compactification of the state space E and $(\bar{P}_t)$ the extension of the initial semigroup (P_t) to functions defined on $\bar{E}$. A function $f \geqslant 0$ defined on E is called Ray-analytic if it is a restriction to E of an analytic function $\bar{f}$ defined on $\bar{E}$. In particular, a Ray-analytic function is universally measurable and every Borel function on E is Ray-analytic.

THEOREM 8 (El Karoui [12]). *For every bounded Ray-analytic function $f \geqslant 0$, the α-reduite $R_\alpha f$ is also Ray-analytic. Moreover, for an arbitrary probability distribution μ on E,*

$$\langle \mu, R_\alpha f \rangle = \sup_\tau E^\mu \left(e^{-\alpha\tau} f(x_\tau) \right),$$

and for an arbitrary $\sigma \in \mathcal{M}$

$$R_\alpha f(x_\sigma) = \operatorname*{ess\,sup}_{\tau \geqslant \sigma} E\left(e^{-\alpha(\tau-\sigma)} f(x_\tau) \mid \mathcal{F}_\sigma \right) \quad P - a.s.$$

Analyticity is also an appropriate concept to deal with measurability of the transformation M, see [18].

I.5. Other questions. Of growing interest are impulse control problems with long run average cost criterion and problems with partial observation. Some existence results for the former have recently been obtained by M. Robin [29] and Ł. Stettner [33]. Many specific problems with partial observation are discussed in Friedman [13] and the general theory for Feller processes on compact state space is the subject of paper [21] by Mazziotto and Szpirglas. An interesting impulse control problem with a long run average cost criterion and partial observation was solved by D. Gątarek [14].

II. Stopping problems for general processes

One can sometimes get a better insight into stopping problems if instead of Markov processes one considers general stochastic processes.

II.1. Snell's envelope and penalization. Let $(\Omega, \mathcal{F}, P)$ be a fixed probability space and $(\mathcal{F}_t)$ a filtration satisfying the usual conditions. Let $\mathcal{O}$ denote the set of all bounded optional processes equal to zero at infinity and $\mathcal{M}$ the set of all $(\mathcal{F}_t)$-stopping times. A process $v \in \mathcal{O}$ such that for arbitrary $\tau \geqslant \sigma$

$$v_\sigma \geqslant E(v_\tau | \mathcal{F}_\sigma) \quad P - \text{a.e.}$$

is called a *strong supermartingale*. From a classical Mertens result (see [11]), for an arbitrary $f \in \mathcal{O}$ there is a minimal strong supermartingale v majorizing f. This is called *Snell's envelope* of f. Moreover, for an arbitrary $\sigma \in \mathcal{M}$

$$v_\sigma = \operatorname*{ess\,sup}_{\tau \geqslant \sigma} E(f_\tau | \mathcal{F}_\sigma) \quad P - \text{a.s.} \tag{12}$$

and in several important case the debut $D_\sigma = \inf\{t \geqslant \sigma; f_t = v_t\}$ is the optimal stopping time in the sense that

$$E(f_{D_\sigma} | \mathcal{F}_\sigma) \geqslant E(f_\tau | \mathcal{F}_\sigma) \quad \text{for } \tau \geqslant \sigma.$$

General sufficient conditions for the existence of optimal stopping times have been given by Bismut and Skali [9] and generalized subsequently by many authors, see[11]. An extension in which a smaller set of stopping times is given for optimization, say predictable stopping times, was obtained by El Karoui [11]. Here we present a way of calculating the Snell envelope extending the penalization method of [4] and [5]. We follow paper [34] by Stettner and Zabczyk. A right-continuous supermartingale $\hat{v}$ is called the *strong envelope* of f if it is the smallest right-continuous, non-negative supermartingale such that $\hat{v} \geqslant f$, $dt \otimes dP$ — a.s. The following is a probabilistic version of the penalization:

For an arbitrary $\beta > 0$ find a right-continuous process v^β such that for all $t \geqslant 0$

$$v_t^\beta = \beta E\Big(\int_t^{+\infty} (f_s - v_s^\beta)^+ ds \,|\, \mathcal{F}_t\Big) \quad P - \text{a.s.} \tag{13}$$

THEOREM 9. *If f is a bounded, progressively measurable process such that $E(\int_0^{+\infty} |f_s| ds) < +\infty$, then for each $\beta > 0$, there is a solution of* (13) *and it is unique, up to indistinguishable processes. It increases with $\beta \uparrow +\infty$ and the limit process $\hat{v} = \lim v^\beta$ is the strong envelope of f.*

COROLLARY. *If f is an optional process then its Snell envelope v majorizes $\hat{v}$ and for the right-continuous process f, $\hat{v}$ is right-continuous as well and $v = \hat{v}$. Assume that additionaly to the right-continuity of f we require regularity: if $\tau_n \uparrow \tau$, $\tau_n \in \mathcal{M}$, then $E(f_{\tau_n}) \to E(f_\tau)$. It is then a consequence of* (13) *that the moment $\hat{D}_\sigma = \lim D_\sigma^\beta$, where $D_\sigma^\beta = \inf\{t \geqslant \sigma; v_t^\beta \leqslant f_t\}$ is optimal. This gives a different proof of a similar result in* [9].

Strong envelopes appeared earlier in connection with the so-called *Kac approach* to the Dirichlet problem, see [34].

The results here formulated show that the natural framework for penalization is the class of right-continuous processes. For more results in a similar spirit we refer to Morimoto [23] and Makowski [20].

II.2. Stopping games. The problem of giving general enough sufficient conditions for a stopping game to be closed or to have a saddle point is of great interest and has been studied intensively. Let f and g be two bounded optional processes such that $f \geqslant g$ and $f_\infty = g_\infty = 0$. As in the markovian case, we define the loss functional $J(\tau, \sigma) = f_\tau I_{\tau \leqslant \sigma} + g_\sigma I_{\sigma < \tau}$, $\tau, \sigma \in \mathscr{M}$. The stopping game is said to be *closed* or to *have value* if $\overline{w} = \underline{w} = w$, where

$$\underline{w} = \sup_{\sigma \in \mathscr{M}} \inf_{\tau \in \mathscr{M}} EJ(\tau, \sigma), \qquad \overline{w} = \inf_{\tau \in \mathscr{M}} \sup_{\sigma \in \mathscr{M}} EJ(\tau, \sigma).$$

Number w is then the value of the game. The discrete time analogue of the stopping game is always closed. However, this is not the case in the present situation. For a sequence (t_n), strictly decreasing to zero, define deterministic processes f and g as follows: $f(t) = -1$ if $t = t_{2n-1}$, $n = 1, 2, \ldots$, and $f(t) = 1$ otherwise; $g(t) = 1$ if $t = t_{2n}, n = 1, 2, \ldots$, and $g(t) = -1$ otherwise. Then $\overline{w} = 1$ $\underline{w} = -1$. Examples suggest that the right-continuity of processes f and g should be sufficient for the existence of the value.[1] In this direction the following result, taken from paper [36] by Stettner, Zaremba and Zabczyk, can be proved:

PROPOSITION 10. *If bounded and cadlag processes f and g vanish at infinity then the stopping game is closed.*

Much more can be said if the following separability or Mokobodzki condition is satisfied, see [6] and [7]:

There exist bounded, non-negative strong supermartingales a and b such that

$$g \leqslant a - b \leqslant f \quad P\text{-a.s.} \tag{14}$$

Condition (14) is connected with the following decoupling procedure, discovered independently and in different context by Bismut [6] and [7]

---

[1] This conjecture was recently confirmed by J. L. Lepeltier and M. A. Maingueneau in "Le jeux de Dynkin en théorie générale sans l'hypothèse de Mokobodzki". An elementary proof of the same result obtained independently several weeks later can be found in: Ł. Settner, J. Zabczyk and P. Zaremba, "On general two-persons stopping games", preprint 283, Institute of Mathematics, Polish Academy of Sciences, 1983.

and Nakoulima [26]. One can formulate this procedure in the form of the following theorem, due to Bismut [7].

THEOREM 11. *If* (14) *holds then there exist two bounded and non-negative strong supermartingales* z^1 *and* z^2 *such that for an arbitrary* $\sigma \in \mathcal{M}$

$$z_\sigma^1 = \operatorname*{ess\,sup}_{\tau \geqslant \sigma} E(z_\tau^2 + g_\tau \mid \mathcal{F}_\sigma), \tag{15}$$

$$z_\sigma^2 = \operatorname*{ess\,sup}_{\tau \geqslant \sigma} E(z_\tau^1 - f_\tau \mid \mathcal{F}_\sigma) \quad P - a.s. \tag{16}$$

The process $w = z^1 - z^2$ is a natural candidate to be the value process in the sense that for an arbitrary $\sigma \in \mathcal{M}$:

$$\begin{aligned} w_\sigma &= \operatorname*{ess\,inf}_{\tau_1 \geqslant \sigma} \operatorname*{ess\,sup}_{\tau_2 \geqslant \sigma} E\big(J(\tau_1, \tau_2) \mid \mathcal{F}_\sigma\big) \\ &= \operatorname*{ess\,inf}_{\tau_2 \geqslant \sigma} \operatorname*{ess\,inf}_{\tau_1 \geqslant \sigma} E\big(J(\tau_1, \tau_2) \mid \mathcal{F}_\sigma\big) \quad P - \text{a.s.} \end{aligned} \tag{17}$$

In particular, if the debuts

$$D_\sigma^1 = \inf\{t \geqslant \sigma;\, w_t = f_t\}, \qquad D_\sigma^2 = \inf\{t \geqslant \sigma;\, w_t = g_t\}$$

are optimal stopping times for problems (15) and (16) respectively, then (17) holds and the pair (D_σ^1, D_σ^2) is the saddle point, see [7]. As was pointed out by Ł. Stettner, see [1], identity (17) holds if f and g are only right-continuous. It was proved by Alario-Nazaret [1] that (17) holds under even weaker conditions: lower semicontinuity from the right of f and upper semicontinuity from the right of g.

References

[1] Alario-Nazaret M., *Jeux de Dynkin*, Thèse de docteur en mathématiques, l'Université de Franche-Compté, 1982.

[2] Barthelemy L., *Application de la théorie des semi-groupes non linéaire dans L à l'étude d'une classe d'inéquations quasi-variationnelles*, Thèse 3e cycle, Université de Franche-Compté, Besançon, 1980.

[3] Bensoussan A., Contrôle impulsionnel et inéquations quasi variationnelles. In: *Proceedings of the International Congress of Mathematicians*, Vancouver, 1974, pp. 329–334.

[4] Bensoussan A. and Lions J. L., *Applications des Inéquations Variationnelles en Contrôle Stochastique*, Dunod, Paris, 1978.

[5] Bensoussan A. and Lions J. L., *Contrôle Impulsionnel et Inéquations Quasi-Variationnelles*, Dunod, Paris, 1982.

[6] Bismut J. M., Sur un problème de Dynkin, *Z. Wahrscheinlichkeitstheorie verw. Gebiete* **39** (1977), pp. 31–53.

[7] Bismut J. M., Contrôle de processus alternants et applications, *Z. Wahrscheinlichkeitstheorie verw. Gebiete* **47** (1979), pp. 241–288.
[8] Bismut J. M., Convex Inequalities in Stochastic Control, *J. of Functional Analysis* **42** (1982), pp. 226–270.
[9] Bismut J. M. and Skalli B., Temps d'arrêt optimal, théorie général des processus et processus de Markov, *Z. Wahrscheinlichkeitstheorie verw. Gebiete* **39** (1977), pp. 301–313.
[10] Doshi B., Optimal Switching among a Finite Number of Markov Processes, *Journal of Optimization Theory and Applications* **35** (1981), pp. 581–610.
[11] El Karoui N., Les aspects probabilistes du contrôle stochastique, *LNM* **876** (1981), pp. 73–238.
[12] Friedman A., Stochastic Differential Games with Stopping Times and Variational Inequalities. In: *Proceedings of the International Congress of Mathematicians*, Vancouver, 1974, pp. 339–342.
[13] Friedman A., Optimal Stopping Problems in Stochastic Control, *SIAM Review* **21** (1979), pp. 71–80.
[14] Gątarek D., On a Reliability Problem by Stochastic Control Methods, *Systems and Control Letters* **2** (1982), pp. 248–254.
[15] Hanouzet B. and Joly J. L., Convergence uniforme des itérés definissant la solution d'une inéquation quasi variationnelle abstraite, *CRAS* **286** (1978), pp. 735–738.
[16] Krasnoselskii M. A., *Positive Solutions of Operator Equations*, P. Noordhoff, Groningen, The Netherlands, 1964.
[17] Krylov N. V., Control of Markov Processes and w^p Spaces, *Izv. USSR, Math.* **5** (1971), pp. 233–266.
[18] Lepeltier J. P. and Marchal B., *Contrôle impulsionnel applications: jeux impulsionnels – controle continu*, Publications Université Paris-Nord, 1981.
[19] Mackevičius V., Passing to the Limit in the Optimal Stopping of the Markov Processes, *Lietuvos Mathematikos Rinkinys* XIII (1973), pp. 115–127.
[20] Makowski A. M., Local Optimality Conditions for Optimal Stopping, *Stochastics* **7** (1982), pp. 91–133.
[21] Mazziotto G. and Szpirglas J., Théorème de séparation pour le contrôle impulsionnel, submitted to *Z. Wahrscheinlichkeitstheorie und verw. Gebiete.*
[22] Menaldi J. L., *Sur les problemes de temps d'arrêt, contrôle impulsionnel et continu correspondant a des opérateurs dégénéres*, Thèse, Paris, 1980.
[23] Morimoto H., Optimal Stopping and a Martingale Approach to the Penalty Method, *Tôhoku Math. Journ.* **34** (1982), pp. 407–416.
[24] Nagai H., On an Optimal Stopping Problem and a Variational Inequality, *J. Math. Soc. Japan* **30** (1978), pp. 303–312.
[25] Nagai H., Impulsive Control of Symmetric Markov Processes and Quasi-Variational Inequalities, to appear in *Osaka J. Math.*
[26] Nakoulima O., Inéquation variationnelle et systeme d'inéquations quasi-variationnelles stationaire de type elliptique et d'evolution de type parabolique, *Afrika Matematika* I (1) (1978), et II (2) (1980).
[27] Nisio M.,On Nonlinear Semigroups for Markov Processes Associated with Optimal Stopping, *Applied Mathematics and Optimization* **4** (1978), pp. 146–169.
[28] Robin M., *Contrôle impulsionnel des processus de Markov*, Thèse, l'Universite Paris IX, 1978.

[29] Robin M., On Some Impulsive Control Problems with Long Run Average Cost, *SIAM J. on Control* **19** (1981), pp. 333–358.
[30] Shyriaev A. N., *Optimal Stopping Rules*, Springer-Verlag, New York, 1978.
[31] Stettner Ł., *Penalty Method in Stochastic Control*, Ph. D. Dissertation, Institute of Mathematics, Polish Academy of Sciences, 1981.
[32] Stettner Ł., Zero-Sum Markov Games with Stopping and Impulsive Strategies, *Appl. Math. Optim.* **9** (1982), pp. 1–24.
[33] Stettner Ł., On Impulsive Control with Long Run Average Cost Criterion, to appear in *Studia Math.*
[34] Stettner Ł., and Zabczyk J., Strong Envelopes of Stochastic Processes and a Penalty Method, *Stochastics* **4** (1981), pp. 267–280.
[35] Stettner Ł., and Zabczyk J., *On Optimal Stopping of Feller Processes*, Preprint 284, Institute of Mathematics, Polish Academy of Sciences, 1983.
[36] Stettner Ł., Zaremba P., and Zabczyk J., *On General Two-Persons Stopping Games*, Preprint 283, Institute of Mathematics, Polish Academy of Sciences, 1983.
[37] Zabczyk J., Optimal Control by Means of Switching, *Studia Math.* XLV (1973), pp. 161–171.
[38] Zabczyk J., Stability Properties of the Discrete Riccati Operator Equation, *Kybernetika* **13** (1977), pp. 1–10.
[39] Zabczyk J., Semigroup Methods in Stochastic Control, *Report 821 of CRM*, Université de Montreal, 1978.
[40] Zabczyk J., On the Resolvent Identity for Non-Linear Semigroups of Stochastic Control. In: *Proceedings of the Workshop on Control Theory and Differential Equations*, Jasi, Romania, 1982.
[41] Zabczyk J., Stopping Games for Symmetric Markov Processes, *Probability and Mathematical Statistics* 4 (1984), pp. 185–196.

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

FENG KANG

Finite Element Method and Natural Boundary Reduction

1. Introductory comments

One of the major advances in numerical methods for partial differential equations made in the recent twenty years is the finite element method (FEM). The method is based on the variational formulation of elliptic equations and on the triangulated approximations. The first component, the variational principle, is an old one and leads to the classical Rayleigh–Ritz method, which, though successful in the past, suffers from numerical instability and geometric inflexibility, originating from the analytic approximations adopted, but unnoticed in the pre-computer times due to the limited size and complexity of the problems then attacked. The second component, the triangulated local approximations, used but not exploited in full in the finite difference methods, is more elementary and much older. Dating back to ancient times, it was for a long time overshadowed by the later achievements in analytic approximations, but revived eventually due to its innate stability and flexibility, which becomes important in the computer era.

A judicious combination of the two old components, conventionally in juxtaposition, gives rise to the FEM, an innovation of general applicability, especially suited for problems of great complexity as well as for computer usage. In FEM, all the essential properties of elliptic operators, e.g., symmetry, coerciveness and locality are well preserved after discretization. This leads, on the one hand, to an efficient computational scheme and, on the other hand, to a sound theoretical foundation, on which the Sobolev space theory of elliptic equations is invoked in a natural way, ensuring the reliability of the method in practice. Moreover, the logic of FEM is simple, intuitive and easy to be implemented on the computer,

whose capability is thereby fully exploited not only as an "equation solver" but also as an "equation setter"; there is already a vast body of software for engineering applications built around it. On the ground of all these reasons, the FEM has become the major methodology for computer solution of elliptic problems, and, by and large, it will remain such in the foreseeable future.

It is also well known that the elliptic boundary value problems have equivalent formulations, in addition to the variational ones, in various forms of integral equations on the boundary. In recent years an increasing interest in the numerical solution has been observed, particularly in the finite element solution of boundary integral equations, leading to the boundary element method (BEM) in various versions. The boundary reduction has the advantage of diminishing the number of space dimensions by 1 and of the capability to handle problems involving infinite domains and, moreover, also cornered or cracked domains at the expense, however, of increased complexity in the analytical formulation, which is not easily available beyond the simplest cases. During reduction, some differential operators of a local character are inverted into integral operators, which, being non-local, result in full metrices instead of sparse ones; this offsets, at least in part, the advantage gained in dimension reduction. So, the approach via integral equations, as it stands by itself, is rather limited in scope, lacking general applicability; and the BEM is not likely to replace the FEM.

Nevertheless, there are many complicated problems in which several different parts are coupled together; boundary reduction could be judiciously applied to some parts of the domain with advantage for the purpose of cutting down the size or complexity of the problem, resulting in a modified but equivalent boundary value problem on a reduced domain with artificial or computational boundaries carrying integral boundary conditions which correctly account for the full coupling between the eliminated and the remaining parts. There are also problems in which the coupling at the given boundary with the environment is assigned in an oversimplified way in the conventional form of differential boundary conditions; boundary reduction could in some way be applied to the exterior domain to give a more complicated integral boundary condition for a more accurate account between the given system and its environment.

The above motivations require that the boundary reduction should be compatible with the accepted variational formulation and finite element methodology and that the BEM should be developed as a component of the FEM, well-fitted in that framework, rather than as an independent

technique. It is from this point of view that, among other things, a natural and direct method of boundary reduction, proposed by the present author [4, 5, 6] called canonical boundary reduction, will be discussed in the sequel.

2. Case of the Laplace equation

Consider, for example, the Neumann problem of the Laplace equation in a domain Ω in R^2 with smooth boundary Γ with exterior normal n,

$$\Omega: \ -\Delta u = 0, \tag{1}$$

$$\Gamma: \ u_n = g \quad \text{with compatibility condition} \int_\Gamma g\,dx = 0. \tag{2}$$

Here g belongs to, say, $H^{-1/2}(\Gamma)$. This problem is equivalent to the variational problem: find $u \in H^1(\Omega)$ such that

$$\begin{aligned} &D(u, v) = F(v) \quad \text{for every } v \in H^1(\Omega), \\ &D(u, v) \equiv \int_\Omega \operatorname{grad} u \cdot \operatorname{grad} v\,dx, \quad F(v) = \int_\Gamma gv\,dx. \end{aligned} \tag{3}$$

The classical Fredholm boundary reduction consists in expressing the harmonic function as a layer potential

$$u(x) = \int_\Gamma E(x-x')\sigma(x')\,dx', \quad E(x) = \frac{1}{2\pi}\log\frac{1}{|x|}. \tag{4}$$

Then the jump condition of the potential gradient across the boundary is

$$u_n(x) = \int_\Gamma E_n(x-x')\sigma(x')dx' + \tfrac{1}{2}\sigma(x), \quad \text{i.e.,} \quad (\tfrac{1}{2}I + E_n)\sigma = u_n, \tag{5}$$

a Fredholm equation of the second kind in the unknown density σ against the known data (2). Note that, after reduction, the essential properties of the original operator, i.e., symmetry, coerciveness and variational form, are not preserved. Moreover, a new function σ is introduced on Γ in addition to the trace data

$$u|_\Gamma = \gamma_0 u, \quad u_n|_\Gamma = \gamma_1 u$$

of the original problem; this is inconvenient for coupling in complicated problems. So, from the practical and computational point of view at

least, the Fredholm reduction is unsatisfactory: it does not fit well with the FEM.

A partial improvement results from the Green formula

$$\int_{\Omega} (v\Delta' u - u\Delta' v)\,dx' = \int_{\Gamma} (vu_{n'} - uv_{n'})\,dx' \tag{6}$$

(x' is the dummy variable with the corresponding primed differential operators) and the choice $v(x') = E(x-x')$, whence

$$u(x) = \int_{\Gamma} (uE_{n'} - u_{n'}E)\,dx', \qquad x \in \Omega.$$

Then differentiation and passage to boundary, with jump conditions considered, give another Fredholm equation of the second kind

$$\tfrac{1}{2}u(x) + \int_{\Gamma} E_{n'}(x-x')u(x')\,dx' = \int_{\Gamma} E(x-x')u_{n'}(x')\,dx',$$

i.e.,

$$(\tfrac{1}{2}I + E_{n'})u = Eu_n, \tag{7}$$

with the Dirichlet trace data, instead of introducing a new function in (5) as unknown against the known Neumann data (2). This formulation is adopted in most BEM's; however, the kernel is similar to that in (5), and so the same difficulties remain.

The most satisfactory approach is to choose $v(x')$ in (6) to be the Green function $G(x, x')$ satisfying

$$\begin{aligned} -\Delta' G(x, x') &= \delta(x'-x), \\ G(x, x') &= 0 \quad \text{for } x' \in \Gamma, \\ G(x, x') &= G(x', x) \end{aligned}$$

to obtain the Poisson formula

$$u(x) = -\int_{\Gamma} G_{n'}(x, x')u(x')\,dx', \qquad x \in \Omega, \quad \text{i.e.,} \quad u = P\gamma_0 u. \tag{8}$$

Then differentiation and passage to boundary gives

$$u_n(x) = -\int_{\Gamma} G_{n'n}(x, x')u(x')\,dx', \qquad x \in \Gamma, \quad \text{i.e.,} \quad u_n = K\gamma_0 u, \tag{9}$$

an expression of the Neumann data (as known) in terms of the Dirichlet

data (as unknown). The kernel $K(x, x') = -G_{n'n}(x, x')$ is regarded as a limiting distribution kernel. So, the Neumann problem (1)–(2) or (3) is equivalent to the solving of the boundary integral equation

$$K\varphi = g \tag{10}$$

for the unknown Dirichlet data $\gamma_0 u = \varphi$ on Γ, leading to u in Ω via the Poisson formula (8).

The boundary integral equation (10) has, in turn, its own variational formulation, i.e., to find $\varphi \in H^{1/2}(\Gamma)$ such that

$$\hat{D}(\varphi, \psi) = \hat{F}(\psi), \quad \forall \psi \in H^{1/2}(\Gamma),$$

$$\hat{D}(\varphi, \psi) = \int_\Gamma \int_\Gamma K(x, x')\varphi(x')\psi(x)\,dx\,dx', \quad \hat{F}(\psi) = \int_\Gamma g\psi\,dx, \tag{11}$$

where the trace forms $\hat{D}$, $\hat{F}$ are inherently related to the original forms D, F by

$$D(u, v) = \hat{D}(\gamma_0 u, \gamma_0 v) \quad \text{for every } u, v \in H^1(\Omega), \quad \Delta u = \Delta v = 0, \tag{12}$$

$$F(v) = \hat{F}(\gamma_0 v) \quad \text{for every } v \in H^1(\Omega). \tag{13}$$

The symmetry and coerciveness properties of K follows directly from those of A via the trace theorem of Sobolev spaces and vice versa.

Consider now a coupling problem

$$\Omega: -\Delta u = 0, \tag{14}$$

$$\partial\Omega = \Gamma_1: u_{n_1} = g, \quad \int_{\Gamma_1} g\,dx = 0, \tag{15}$$

where the domain Ω consists of two subdomains Ω_0 and Ω_1 with their common boundary Γ with normal n directed to the exterior of the outer subdomain Ω_0, which is for example infinite. The inner subdomain Ω_1 is for example finite, and has an outer boundary Γ and an inner boundary Γ_1 with normal n_1 directed to the exterior of Ω_1. The corresponding variational problem is to find $u \in H^1(\Omega)$ such that

$$D(u, v) = F(v) \quad \text{for every } v \in H^1(\Omega),$$

$$D(u, v) = \sum_{i=0}^{1} D_i(u, v), \quad D_i(u, v) = \int_{\Omega_i} \operatorname{grad} u \cdot \operatorname{grad} v\,dx, \quad i = 0, 1,$$

$$F(v) = \int_{\Gamma_1} gu\,dx.$$

Let K be the boundary operator induced by the Laplace operator in subdomain Ω_0 on its boundary Γ. Then

$$D_0(u, v) = \hat{D}_0(\gamma_0 u, \gamma_0 v) = \int_\Gamma \int_\Gamma K(x, x') u(x') v(x) dx dx', \tag{16}$$

and so the problem (14)–(15) is equivalent to a problem for a reduced domain: to find $u \in H^1(\Omega_1)$ such that

$$D'(u, v) = \hat{D}_0(\gamma_0 u, \gamma_0 v) + D_1(u, v) = F(v) \quad \text{for every } v \in H^1(\Omega_1) \tag{17}$$

which is equivalent, in turn, to

$$\Omega_1: \ -\Delta u = 0, \tag{18}$$

$$\Gamma_1: \ u_{n_1} = g, \tag{19}$$

$$\Gamma: \ u_n = Ku. \tag{20}$$

Note that, in this reduced problem, in addition to the original boundary Γ_1 with the natural boundary condition in local form (15), a new artificial boundary Γ is constructed to carry a natural boundary condition in non-local form (20), which accounts correctly, i.e., without approximation, for the coupling between the deleted part Ω_0 and the remaining part Ω_1.

We see that the boundary reduction just described is direct and natural in the variational formulation; it faithfully preserves all the essential characteristics of the original elliptic problem and is fully compatible with FEM. It is thus called the canonical boundary reduction, and the corresponding integral equations — canonical integral equations.

We give examples of Poisson formulae and canonical integral equations for the Laplace equation over some typical domains in two dimensions.

(1) Domain interior to the circle of radius R.

$$u(r, \theta) = \frac{1}{2\pi} \int_0^{2\pi} \frac{(R^2 - r^2) u(R, \theta') d\theta'}{R^2 + r^2 - 2Rr\cos(\theta - \theta')}, \qquad r < R,$$

$$u(R, \theta) = -\frac{1}{4\pi} \int_0^{2\pi} \frac{u(R, \theta') d\theta'}{R \sin^2 \dfrac{(\theta - \theta')}{2}}.$$

(2) Domain exterior to the circle of radius R.

$$u(r,\theta) = \frac{1}{2\pi}\int_0^{2\pi} \frac{(r^2-R^2)u(R,\theta')d\theta'}{R^2+r^2-2Rr\cos(\theta-\theta')}, \qquad r>R,$$

$$-u_r(R,\theta) = -\frac{1}{4\pi}\int_0^{2\pi} \frac{u(R,\theta')d\theta'}{R\sin^2\dfrac{\theta-\theta'}{2}}.$$

(3) Upper half-plane above the line $y = a$.

$$u(x,y) = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{(y-a)u(x',a)dx'}{(x-x')^2+(y-a)^2}, \qquad y>a,$$

$$-u(x,a) = -\frac{1}{\pi}\int_{-\infty}^{\infty} \frac{u(x',a)dx'}{(x-x')^2}.$$

(4) Arbitrary simply connected domain Ω. If $w = f(z)$ conformally maps $z \in \Omega$ onto the interior $|w| < 1$ of the unit circle, then [9]

$$p(z,z) = -G_{n'}(z,z') = \frac{|f'(z')|(1-|f(z)|^2)}{2\pi|f(z)-f(z')|^2}, \qquad z\in\Omega,\ z'\in\Gamma,$$

$$K(z,z) = -\frac{f'(z)f'(z')}{2\pi|f(z)-f(z')|^2}, \qquad z,z'\in\Gamma,$$

$$= -\frac{1}{\pi|z-z'|^2} + \text{an infinitely smoothing kernel.}$$

The canonical integral equation (9) was first introduced by Hadamard [7, 9]. The function $-G_{n'n}(x,x')$ in it is a distribution kernel of high singularity of non-integrable type $1/(x-x')^2$, regarded as a "finite part" regularization of divergent integrals. It is in fact a pseudo-differential operator of order 1 and

$$-G_{n'n}\colon H^s(\Gamma)\to H^{s-1}(\Gamma) \quad \text{for every real } s.$$

So, at the expense of higher singularity, the canonical integral equation has the advantage of being more stable than the Fredholm equation (5)

or (7) of the second kind with the kernel

$$(\tfrac{1}{2}I+E_{n'}) \quad \text{or} \quad (\tfrac{1}{2}I+E_n)\colon H^s(\Gamma)\to H^s(\Gamma) \quad \text{for every real } s.$$

In addition, the choice in (6) of $v(x')=N(x,x')$, the Neumann function, satisfying

$$-\Delta' N(x,x') = (x'-x),$$

$$N_{n'}(x,x') = -1/L \ (L \text{ is the length of } \Gamma) \quad \text{for } x'\in\Gamma,$$

$$\int_\Gamma N(x,x')\,dx' = 0, \quad \text{if } \Omega \text{ is bounded},$$

gives, as the inverse of (9), the integral equation

$$u(x) = \int_\Gamma N(x,x')\,u_{n'}(x')\,dx', \quad x'\in\Gamma, \quad \text{i.e.,} \quad u = Nu_n,$$

first obtained by Hilbert [8] and extended to general second order elliptic equations by Birkhoff in the earliest paper which had ever discussed the importance of integral boundary conditions and coupling problems [2]. The kernel $N(x,x')$, called in that paper the albedo function after Fermi, has a weak singularity of the logarithmic type and induces a smoothing operator

$$N\colon H^s(\Gamma)\to H^{s+1}(\Gamma) \quad \text{for every real } s,$$

which is unfavourable to stability and leads to a variational principle which is not natural and not compatible with FEM in coupling problems.

3. Canonical boundary reduction for general elliptic equations

The canonical integral equations of a general variational elliptic equation or a system is a system of integral expressions of the Neumann boundary data in terms of the Dirichlet boundary data for the solutions of the given equation or system.

Consider a properly elliptic differential operator of order $2m$

$$Au = \sum_{|p|,|q|\leqslant m} (-1)^{|q|}\,\partial^q a_{pq}(x)\,\partial^p u, \quad a_{pq}\in C^\infty, \tag{21}$$

$$A\colon H^s(\Omega)\to H^{s-2m}(\Omega)$$

with its associated bilinear form

$$D(u, v) = \sum_{|p|,|q|\leqslant m} \int_{\Omega} a_{pq} \partial^p u \, \partial^q v \, dx \tag{22}$$

on a domain Ω with C^∞ boundary Γ with exterior normal n. Corresponding to A and to the set of the Dirichlet trace operators

$$\gamma = (\gamma_0, \ldots, \gamma_{m-1})^T, \quad \gamma_j u = (\partial_n)^j u|_\Gamma, \quad j = 0, \ldots, m-1,$$

there is a unique set of boundary differential operators

$$\beta = (\beta_0, \ldots, \beta_{m-1})^T, \quad \beta_i u = \beta_i(x, n(x), \partial) u|_\Gamma,$$

such that the Green formula

$$D(u, v) = \int_{\Omega} Au \cdot v \, dx + \sum_{i=0}^{m-1} \int_{\Gamma} \beta_j u \cdot \gamma_j u \, dx \tag{23}$$

holds for smooth u, v. $\beta_i u$ is the Neumann data complementary to the Dirichlet data $\gamma_i u$.

From the basic assumption that the Dirichlet problem

$$\Omega: \; Au = 0, \tag{24}$$

$$\Gamma: \; \gamma_j u = \text{known}, \quad j = 0, \ldots, m-1. \tag{25}$$

is uniquely solvable in space $H^s(\Omega)$ with the known data $\gamma_j u \in H^{s-j-1/2}(\Gamma)$, it follows that the Poisson formula $u = \sum P_i \gamma_i u$ gives an isomorphism

$$P = (P_0, \ldots, P_{m-1}): \; T^s(\Gamma) \to H^s_A(\Omega),$$

where

$$T^s(\Gamma) = \prod_{j=0}^{m-1} H^{s-j-1/2}(\Gamma), \quad H^s_A(\Omega) = \{u \in H^s(\Omega) \mid Au = 0\}.$$

Then the canonical system of integral equations is given by

$$\beta u = K \gamma u,$$

i.e.,

$$\beta_i u = \sum_{j=0}^{m-1} K_{ij} \gamma_j u, \quad i = 0, \ldots, m-1, \tag{26}$$

$$K_{ij} = \beta_i \circ P_j: \; H^{s-j-1/2}(\Gamma) \to H^{s-(2m-i-1/2)}(\Gamma).$$

It can be shown that K_{ij} is a pseudo-differential operator of order $2m-1-i-j$ on the boundary manifold Γ and the matrix operator K is elliptic. Hence K induces a bilinear functional

$$\hat{D}(\varphi, \psi) = (K\varphi, \psi) = \sum_{i,j=0}^{m-1} \int_{\Gamma} K_{ij}(x, x')\varphi_j(x')\psi_i(x)\,dxdx' \tag{27}$$

which preserves the value of the bilinear functional

$$D(u, v) = \hat{D}(\gamma u, \gamma v) \quad \text{for every } u, v \in H_A^s(\Omega). \tag{28}$$

Moreover, the formal transpose $\tilde{A}$ of A is given by

$$\tilde{A}u = \sum_{|p|,|q|\leqslant m} (-1)^{|q|}\partial^p a_{qp}(x)\,\partial^p u,$$

with an associated bilinear functional

$$\tilde{D}(u, v) = D(v, u).$$

Then it is easily seen that

$$\tilde{K}(A) = K(\tilde{A}), \quad \hat{\tilde{D}} = \tilde{\hat{D}},$$

A is symmetric iff $K(A)$ is symmetric,

a is coercive iff $K(A)$ is coercive,

thus all the essential properties of A are faithfully preserved by $K(A)$ and the following conditions are equivalent:

(1) Find $u \in H^s(\Omega)$ such that

$$\Omega\colon Au = 0, \quad \Gamma\colon \beta_i u = g_i, \quad i = 0, \ldots, m-1.$$

(2) Find $u \in H^s(\Omega)$ such that

$$D(u, v) = \sum_{i=0}^{m-1} (g_i, \gamma_i u) \quad \text{for every } v \in H^s(\Omega).$$

(3) Find $\varphi \in T^s(\Gamma)$ such that

$$\sum_{j=0}^{m-1} K_{ij}\varphi_j = g_i, \quad i = 0, \ldots, m-1.$$

(4) Find $\varphi \in T^s(\Gamma)$ such that

$$\hat{D}(\varphi, \psi) = \sum_{i=0}^{m-1} (g_i, \psi_i) \quad \text{for every } \psi \in T^s(\Gamma).$$

Note that the compatibility condition

$$\sum_{i=0}^{m-1} (g_i, \gamma_i v) = 0 \quad \text{for every solution } v \text{ of } A^* v = 0,\ \beta_i^* v = 0,\ i = 0, \ldots, m-1,$$

for (1) or (2) corresponds to the compatibility condition

$$\sum_{i=0}^{m-1} (g_i, \psi_i) = 0 \quad \text{for every solution } \psi \text{ of } K^* \psi = 0$$

for (3) or (4).

When the solution $\varphi = \gamma u$ of (3) or (4) on Γ is found, the Poisson formula gives the solution u in Ω.

From the second Green formula

$$\int_\Omega (uA'v - vA'u)\, dx' = \int_\Gamma \beta_j' u \gamma_j' v - \beta_j' v \gamma_j' u)\, dx'$$

with $v(x')$ chosen to be the Green function $\tilde{G}(x, x')$ of $\tilde{A}$, which is the transpose $G(x', x)$ of the Green function of A, one gets the Poisson kernel

$$P_j(x, x') = -\beta_j' G(x', x), \quad i = 0, \ldots, m-1,\ x \in \Omega,\ x' \in \Gamma, \tag{29}$$

and the kernel of the canonical integral equation

$$K_{ij}(x, x') = -\beta_i \tilde{\beta}_i' G^{(-0)}(x', x), \quad i, j = 0, \ldots, m-1,\ x, x \in \Gamma, \tag{30}$$

where the LHS is the limit distribution kernel (from the inner side)

$$\beta_i \tilde{\beta}_j' G^{(-0)}(x', x) = \beta_i \tilde{\beta}_j' G^{(-0)}(x', x) + R_{ij}(x, x'),$$

the first kernel on the left being formally evaluated on Γ, while R_{ij} is a linear combination of derivatives of the delta-function $\delta(x - x')$ with support concentrated on the diagonal $x = x'$ of $\Gamma \times \Gamma$, which corresponds to the jump of the potential. For concrete examples, see [6].

4. Asymptotic radiation conditions

Now we shall apply the techniques of Sections 2, 3 to the Helmholtz equation together with Sommerfeld radiation condition at infinity

$$\lim_{r\to\infty} r^{1/2}(u_r - i\omega u) = 0, \tag{31}$$

$$\Omega = \{r > R\}, \quad \Gamma = \{r = R\},$$

$$Au = -(\Delta + \omega^2)u = 0 \quad \text{in } \Omega,$$

$$D(u, v) = \int_\Omega (\operatorname{grad} u \operatorname{grad} v - \omega^2 uv)\, dx.$$

The Poisson formula and the canonical integral equation are, respectively,

$$\begin{gathered} u(r, \theta) = P(\omega, r, R; \theta) * u(R, \theta), \\ p(\omega, r, R; \theta) = \frac{1}{2\pi} \sum_{-\infty}^{\infty} \frac{H_n^{(1)}(\omega r)}{H_n^{(1)}(\omega R)} e^{in\theta}, \quad r > R, \\ -u_r(r; \theta) = K(\omega, R; \theta) * u(R, \theta), \\ K(\omega, R; \theta) = \frac{1}{2\pi} \sum_{-\infty}^{\infty} (-\omega) \frac{H_n^{(1)\prime}(\omega R)}{H_n^{(1)}(\omega R)} e^{in\theta}, \end{gathered} \tag{32}$$

where $*$ is the circular convolution in θ. K induces the bilinear functional

$$\hat{D}(\varphi, \psi) = \int_\Gamma K(\omega, R; \theta - \theta') \cdot \varphi(\theta') \psi(\theta)\, d\theta'\, d\theta. \tag{33}$$

If we consider the circle $r = R$ as an artificial boundary for the elimination of the exterior domain $r > R$, then (32) is the exact of theoretical radiation condition, which is necessarily non-local. After finite element discretization, a non-local operator becomes a full matrix with the storage requirement $O(N^2)$, N being the number of boundary degrees of freedom. Due to the convolutional nature of the operator, in the present case of a circle, the resulting matrix is circulant and requires only $O(N)$ storage. However, due to the analytical complexity of the kernel, the computational effort is always expensive. Hence, much interest has recently been taken in the study of the approximations of non-local boundary conditions by local ones, aiming at reasonable accuracy at a reasonable expense.

From the point of view of compatibility with the variational formulation and FEM for elliptic problems, it should be required that the approximation of (32) be expressed as

$$\frac{\partial u}{\partial n} = Cu = \sum_{j=0}^{m} (-1)^j C_j \frac{\partial^{2j} u}{\partial \theta^{2j}} \tag{34}$$

with the corresponding approximation of the trace variational form (33) by

$$\hat{D}_c(\gamma_0 u, \gamma_0 v) = \sum_{j=0}^{m} \int_{\Gamma} c_j \frac{\partial^j u}{\partial \theta^j} \frac{\partial^j v}{\partial \theta^j} dx. \tag{35}$$

A possible approach for the case of large ω and R is to start from the asymptotic expansion of Hankel functions for large arguments

$$H_n^{(1)}(x) = \left(\frac{2}{\pi x}\right)^{1/2} e^{i\left(x-\frac{1}{2}nx-\frac{1}{4}\pi\right)} \sum_{p=0}^{\infty} \left(\frac{i}{2x}\right)^p (n, p),$$

where

$$(n, p) = \frac{1}{p!} \prod_{k=1}^{p} \left(n^2 - \left(\frac{2k-1}{2}\right)^2\right)$$

is an even polynomial in n of degree $2p$. One can then deduce an asymptotic expansion for

$$-\omega \frac{H_{|n|}^{(1)\prime}(\omega R)}{H_{|n|}^{(1)}(\omega R)} = -i\omega \sum_{p=0}^{\infty} \left(\frac{i}{2\omega R}\right)^p a_p(n^2),$$

where

$$a_0(n^2) = a_0(n^2) = 1, \quad a_1(n^2) = 2(n^2 - \tfrac{1}{4}), \quad a_2(n^2) = -4(n^2 - \tfrac{1}{4}),$$

$$a_k(n^2) = (2k-2)(n, k-1) - a_2(n^2)(n, k-2) - \ldots - a_{k-1}(n^2)(n, 1).$$

Take the mth truncation

$$K_m(n^2) = -i\omega \sum_{p=0}^{m} \left(\frac{i}{2\omega R}\right)^p a_p(n^2);$$

then the successive asymptotic radiation conditions are

$$A_m: \ -\frac{\partial u}{\partial r} = K_m\left(-\frac{\partial^2}{\partial\theta^2}\right)u, \qquad m = 0, 1, \ldots$$

In particular,

$$A_0: \ -\frac{\partial u}{\partial r} = K_0 u = -i\omega u,$$

$$A_1: \ -\frac{\partial u}{\partial r} = K_1 u = \left(-i\omega u + \frac{1}{2R}\right)u,$$

$$A_2: \ -\frac{\partial u}{\partial r} = K_2 u = \left(-i\omega + \frac{1}{2R} - \frac{i}{8\omega R^2}\right)u - \frac{i}{2\omega R^2}\frac{\partial^2 u}{\partial\theta^2},$$

$$A_3: \ -\frac{\partial u}{\partial r} = K_3 u = \left(-i\omega + \frac{1}{2R} - \frac{i}{8\omega R^2}\,\frac{1}{8\omega^2 R^3}\right)u - \left(\frac{i}{2\omega R^2} + \frac{1}{2\omega^2 R^3}\right)\frac{\partial^2 u}{\partial\theta^2}.$$

As a comparison we quote the absorbing radiation conditions, based on the factorization technique of pseudo-differential operators, given by Engquist and Majda [3],

$$E_1: \ -\frac{\partial u}{\partial r} = \left(-i\omega + \frac{1}{2R}\right)u,$$

$$E_2: \ -\frac{\partial u}{\partial r} = \left(-i\omega + \frac{1}{2R}\right)u - \left(\frac{i}{2\omega R} + \frac{1}{2\omega^2 R^3}\right)\frac{\partial^2 u}{\partial\theta^2},$$

and the sequence, based on the asymptotic expansion of solutions of the wave equation, given by Bayliss and Turkel [1],

$$B_1 u = \frac{\partial u}{\partial r} + \left(-i\omega + \frac{1}{2R}\right)u = 0,$$

$$B_2 u = \frac{\partial^2 u}{\partial r^2} + \left(2i\omega + \frac{3}{R}\right)\frac{\partial u}{\partial r} + \left(\frac{-3i\omega}{R} - \omega^2 + \frac{3}{4R^2}\right)u = 0,$$

$$B_k u = \left(\frac{\partial}{\partial r} - i\omega + \frac{4k-3}{2r}\right)B_{k-1}u = 0, \qquad k = 2, 3, \ldots$$

Note that A_0 is the Sommerfeld condition, A_1, E_1 and B_1 are the same. Starting from index 2 the three sequences diverge, and, starting from $i = 3$, the E_i and B_i are not expressible in the required form (34). The

differential operator K_{2p+1} has the same order as K_{2p} but is of higher accuracy, and so is preferable.

It is to be remarked that the conventional boundary condition of the third kind $\partial u/\partial n = c_0 u$, usually expressing the so-called elastic coupling between the system and its environment, is simply the crudest approximation to the full coupling (32) in the present context. The next approximation $\partial u/\partial n = c_0 u - c_1 \partial^2 u/\partial \theta^2$, which reflects the coupling with the environment much better and involves hardly any more additional effort in the FEM implementation, deserves attention. The coefficients c_1, in addition to c_0, should be theoretically predictable as well as experimentally determinable, they are likely to have potentially wide applications in practice. In this sense, the approximate boundary condition A_3 seems to be the most interesting.

For FEM solutions and the related numerical analysis for the canonical integral equations here described, see [6, 10, 11].

References

[1] Bayliss A. and Turkel E., Radiation Boundary Conditions for Wave-Like Equations, *Comm. Pure and Appl. Math.* **33** (6) (1980), pp. 707–725.

[2] Birkhoff G., Albedo Functions for Elliptic Equations. In: R. Langer (ed.), *Boundary Problems in Differential Equations*, Madison, 1960.

[3] Engquist B. and Majda A., Absorbing Boundary Conditions for Numerical Simulation of Waves, *Math. Comp.* **31** (1977), pp. 629–651.

[4] Feng Kang, Differential vs Integral Equations and Finite vs Infinite Elements, *Mathematica Numerica Sinica* **2** (1) (1980), pp. 100–105.

[5] Feng Kang, Canonical Boundary Reduction and Finite Element Method. In: *Proceedings of Symposium on Finite Element Method* (an international invitational symposium held in Hefei, China, May, 1981), Science Press, Beijing, and Gordon and Breach, New York, 1982, pp. 330–352.

[6] Feng Kang and Yu De-hao, Canonical Integral Equations of Elliptic Boundary Value Problems and Their Numerical Solutions. In: Feng Kang and J. L. Lions (eds.), *Proceedings of China–France Symposium on Finite Element Method, April 1982, Beijing*, Science Press, Beijing, and Gordon and Breach, New York, 1983, pp. 211–215.

[7] Hadamard J., *Leçons sur le calcul des variations*, Paris, 1910.

[8] Hilbert D., *Integralgleichungen*, Teubner, Berlin, 1912.

[9] Levy P., *Leçons d'analyse fonctionelle*, Paris 1922.

[10] Yu De-hao, Canonical Integral Equations of Biharmonic Boundary Value Problems, *Mathematica Numerica Sinica* **4** (3) (1982), pp. 330–336.

[11] Yu De-hao, Numerical Solutions of Harmonic and Biharmonic Canonical Integral Equations in Interior or Exterior Circular Domains, *Journal of Computational Mathematics* **1** (1) (1983), pp. 52–62, published by Science Press, Beijing.

CHINESE ACADEMY OF SCIENCES
BEIJING, CHINA

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

R. GLOWINSKI

Numerical Solution of Nonlinear Boundary Value Problems by Variational Methods. Applications

1. Introduction

The main goal of this paper is to describe some methods of solution of nonlinear boundary value problems for ordinary or partial differential equations founded on a variational approach. The variational approach is quite powerful at two levels at least:

(i) The approximation by Galerkin type methods such as finite elements, spectral methods, etc...

(ii) The numerical solution of the approximate problems by efficient iterative methods.

To illustrate the above generalities, we shall discuss in Sections 2, 3, 4 the solution of nonlinear boundary value problems by least squares-conjugate gradient methods and apply them to quite classical nonlinear problems in Fluid Dynamics such as the Navier–Stokes equations for incompressible viscous fluids and the full potential equation modelling the transonic flows of compressible inviscid fluids.

In Sections 5, 6 we shall discuss the solution of a broad class of nonlinear variational problems by augmented Lagrangian methods and apply the corresponding techniques to the solution of some "hard" problems in finite elasticity.

The results of numerical experiments illustrate the possibilities of the methods discussed in this paper (more details about these methods can be found in [19], [20]).

2. Least squares solution of a nonlinear Dirichlet model problem

In order to introduce the methods that we shall apply in Sections 3, 4 to the solution of fluid dynamics problems, we shall consider the solution of a simple nonlinear Dirichlet problem by *least squares* and *conjugate gradient methods*. In Section 2.4, we shall briefly discuss the use of *pseudo arc length continuation methods* for solving nonlinear problems via least squares and conjugate gradient algorithms.

2.1. Formulation of the model problem. Let $\Omega \subset \boldsymbol{R}^N$ be a *bounded* domain with a smooth boundary $\Gamma = \partial\Omega$; let T be a *nonlinear operator* from $V = H_0^1(\Omega)$ to $V^* = H^{-1}(\Omega)$ ($H^{-1}(\Omega)$, the topological dual space of $H_0^1(\Omega)$). We consider the *nonlinear Dirichlet problem*

$$\begin{cases} \textit{Find } u \in H_0^1(\Omega) \textit{ such that} \\ -\Delta u - T(u) = 0 \textit{ in } \Omega \end{cases} \tag{2.1}$$

and we observe that (2.1) implies

$$u = 0 \quad \text{on } \Gamma.$$

We do not discuss here the existence and uniqueness properties of the solutions of (2.1) since we do not want to be very specific about the operator T.

2.2. Least squares formulations of the model problem (2.1).

2.2.1. *Generalities.* We shall consider *least squares* formulations of the model problem (2.1). An obvious least squares formulation consists in saying that the required function u minimizes the left-hand side of (2.1), in an $L^2(\Omega)$ least squares sense. That is,

$$\operatorname*{Min}_{v \in V} \int_\Omega |\Delta v + T(v)|^2 dx, \tag{2.2}$$

where V is a space of feasible functions. Let us introduce ξ by

$$\begin{cases} -\Delta \xi = -\Delta v - T(v) \quad \text{in } \Omega, \\ \xi = 0 \quad \text{on } \Gamma. \end{cases} \tag{2.3}$$

Then (2.2) is equivalent to

$$\operatorname*{Min}_{v \in V} \int_\Omega |\Delta \xi|^2 dx, \tag{2.4}$$

where ξ is a (nonlinear) function of v, through (2.3). It is clear (see e.g. [10], [35]) that (2.3), (2.4) has the structure of an *optimal control* problem, where

(i) v is the *control vector*,

(ii) ξ is the *state vector*,

(iii) (2.3) is the *state equation*,

(iv) the functional occurring in (2.4) is the *cost function*.

Another least squares-optimal control formulation is

$$\operatorname*{Min}_{v \in V} \int_{\Omega} |\xi|^2 dx \tag{2.5}$$

where ξ still satisfies (2.3). This formulation has been used by Cea–Geymonat [11] to solve nonlinear partial differential problems (including the steady Navier–Stokes equations).

Actually, the two above least squares formulations may lead to a slow convergence, since the norm occurring in the cost functions is not appropriate for the state equation. An alternate choice, very well suited to nonlinear second order problems, will be discussed in the next section.

2.2.2. *A H^{-1}-least squares formulation of* (2.1). Let us recall some properties of $H^{-1}(\Omega)$, the topological dual space of $H_0^1(\Omega)$. If $L^2(\Omega)$ is identified to its dual, then

$$H_0^1(\Omega) \subset L^2(\Omega) \subset H^{-1}(\Omega);$$

moreover, $\Delta (= \nabla^2)$ is an isomorphism from $H_0^1(\Omega)$ onto $H^{-1}(\Omega)$. In the sequel the duality pairing $\langle \cdot, \cdot \rangle$ between $H^{-1}(\Omega)$ and $H_0^1(\Omega)$ is chosen in such a way that

$$\langle f, v \rangle = \int_{\Omega} fv \, dx \quad \forall f \in L^2(\Omega), \ \forall v \in H_0^1(\Omega). \tag{2.6}$$

The topology of $H^{-1}(\Omega)$ is defined by $\| \cdot \|_*$, where $\forall f \in H^{-1}(\Omega)$

$$\|f\|_* = \sup_{v \in H_0^1(\Omega) - \{0\}} \frac{|\langle f, v \rangle|}{\|v\|_{H_0^1(\Omega)}}. \tag{2.7}$$

A convenient [1] least squares formulation to solve the model problem (2.1)

[1] Convenient because the space $H_0^1(\Omega)$ in (2.8) is also the space in which we wish to solve (2.1) (as follows from the properties of Δ and T).

seems to be

$$\operatorname*{Min}_{v\in H_0^1(\Omega)} \|\Delta v + T(v)\|_*. \tag{2.8}$$

It is clear that if (2.1) has a solution, then this solution will be a solution of (2.8) for which the cost function vanishes. Let us introduce $\xi \in H_0^1(\Omega)$ by (2.3), so that (2.8) reduces to

$$\operatorname*{Min}_{v\in H_0^1(\Omega)} \|\Delta \xi\|_* \tag{2.9}$$

where ξ is a function of v through (2.3).

Actually, it can be proved that if $\|\cdot\|_*$ is defined by (2.7) with $\langle\cdot,\cdot\rangle$ obeying (2.6), then

$$\|\Delta v\|_* = \|v\|_{H_0^1(\Omega)} = \Big(\int_\Omega |\nabla v|^2 dx\Big)^{1/2}, \quad \forall v \in H_0^1(\Omega). \tag{2.10}$$

It follows from (2.10) that (2.9) may be reformulated as

$$\operatorname*{Min}_{v\in H_0^1(\Omega)} \int_\Omega |\nabla \xi|^2 dx \tag{2.11}$$

where ξ is a function of v through (2.3); (2.11) has also the structure of an optimal control problem.

Remark 2.1. Nonlinear boundary value problems have been treated in [38] using a formulation closely related to (2.3), (2.11).

2.3. Conjugate gradient solution of the least squares problem (2.3), (2.11). Let us define the function J: $H_0^1(\Omega)\to\mathbb{R}$ by

$$J(v) = \tfrac{1}{2}\int_\Omega |\nabla \xi|^2 dx \tag{2.12}$$

where ξ is a function of v through (2.3); then we may also write (2.11) as

$$\begin{cases} \textit{Find } u \in H_0^1(\Omega) \textit{ such that} \\ J(u) \leqslant J(v) \quad \forall v \in H_0^1(\Omega). \end{cases} \tag{2.13}$$

We shall use a conjugate gradient algorithm to solve (2.13). From among the possible conjugate gradient algorithms we have selected the *Polak–Ribière* version (see Polak [45]), since this algorithm produced the best

performances in the various experiments we did (good performance of the Polak–Ribière algorithm are discussed in [46]). Let us denote by $J'(\cdot)$ the *differential* of $J(\cdot)$; then the Polak–Ribière version of the conjugate gradient method, applied to the solution of (2.13) is

Step 0: Initialization.

$$u^0 \in H_0^1(\Omega) \text{ given}, \tag{2.14}$$

compute $g^0 \in H_0^1(\Omega)$ *from*

$$-\Delta g^0 = J'(u^0) \text{ in } \Omega, \quad g^0 = 0 \text{ on } \Gamma, \tag{2.15}$$

and set

$$z^0 = g^0. \tag{2.16}$$

Then for $n \geqslant 0$, *assuming* u^n, g^n, z^n *to be known, compute* u^{n+1}, g^{n+1}, z^{n+1} *by*

Step 1: Descent.

$$u^{n+1} = u^n - \lambda_n z^n, \tag{2.17}$$

where λ_n *is the solution of the one-dimensional minimization problem*

$$\lambda_n \in \boldsymbol{R}, \quad J(u^n - \lambda_n z^n) \leqslant J(u^n - \lambda z^n) \ \forall \lambda \in \boldsymbol{R}. \tag{2.18}$$

Step 2: Construction of the new descent direction.

Define $g^{n+1} \in H_0^1(\Omega)$ *by*

$$-\Delta g^{n+1} = J'(u^{n+1}) \text{ in } \Omega, \quad g^{n+1} = 0 \text{ on } \Gamma, \tag{2.19}$$

then

$$\gamma_n = \frac{\int_\Omega \nabla g^{n+1} \cdot \nabla (g^{n+1} - g^n)\, dx}{\int_\Omega |\nabla g^n|^2 dx}, \tag{2.20}$$

$$z^{n+1} = g^{n+1} + \gamma_n z^n, \tag{2.21}$$

$$n = n+1, \quad \text{go to (2.17)}. \ \blacksquare \tag{2.22}$$

The two non-trivial steps of algorithm (2.14)–(2.22) are:

(i) The solution of the *single variable* minimization problem (2.18); the corresponding *line search* can be achieved by *dichotomy* or *Fibonacci methods* (see, for example, [6], [45]). We have to observe that each evalu-

ation of $J(v)$ for a given argument v requires the solution of the linear Poisson problem (2.3) to obtain the corresponding ξ.

(ii) The calculation of g^{n+1} from u^{n+1}, which requires the solution of two linear Dirichlet problems (namely (2.3) with $v = u^{n+1}$, and (2.19)).

Remark 2.2. As stopping criterion for algorithm (2.14)-(2.22) we should use

$$J(u^n) \leqslant \varepsilon \quad \text{or} \quad \|g^n\|_{H_0^1(\Omega)} \leqslant \varepsilon$$

where ε is a "small" positive number. ■

Calculation of $J'(u^n)$ and g^n: Owing to the importance of step (ii), let us describe in detail the calculation of $J'(u^n)$ and g^n:

Let $w \in H_0^1(\Omega)$; then $J'(v)$ may be defined by

$$\langle J'(v), w\rangle = \lim_{\substack{t\to 0\\ t\neq 0}} \frac{J(v+tw)-J(v)}{t}; \tag{2.23}$$

we obtain from (2.3), (2.12), (2.23)

$$\langle J'(v), w\rangle = \int_\Omega \underline{\underline{\nabla}}\xi\cdot\nabla\eta\, dx, \tag{2.24}$$

where $\eta \in H_0^1(\Omega)$ is the solution of

$$\begin{cases} \Delta\eta = \Delta w + T'(v)\cdot w \ \text{ in } \ \Omega, \\ \eta = 0 \ \text{ on } \ \Gamma; \end{cases} \tag{2.25}$$

(2.25) has the following variational formulation

$$\begin{cases} \displaystyle\int_\Omega \nabla\eta\cdot\nabla z\, dx = \int_\Omega \nabla w\cdot\nabla z\, dx - \langle T'(v)\cdot w, z\rangle \ \ \forall z \in H_0^1(\Omega), \\ \eta \in H_0^1(\Omega). \end{cases} \tag{2.26}$$

Taking $z = \xi$ in (2.26), we obtain from (2.24)

$$\langle J'(v), w\rangle = \int_\Omega \nabla\xi\cdot\nabla w\, dx - \langle T'(v)\cdot w, \xi\rangle \quad \forall v, w \in H_0^1(\Omega). \tag{2.27}$$

Therefore $J'(v)$ $(\in H^{-1}(\Omega))$ may be identified with the linear functional on $H_0^1(\Omega)$ defined by

$$w \to \int_\Omega \nabla\xi\cdot\nabla w\, dx - \langle T'(v)\cdot w, \xi\rangle. \tag{2.28}$$

It follows from (2.19), (2.27), (2.28) that g^n is the solution of the following *linear variational problem*:

$$\begin{cases} \textit{Find } g^n \in H_0^1(\Omega) \textit{ such that } \forall w \in H_0^1(\Omega) \\ \int_\Omega \nabla g^n \cdot \nabla w \, dx = \int_\Omega \nabla \xi^n \cdot \nabla w \, dx - \langle T'(u^n) \cdot w, \xi^n \rangle \end{cases} \tag{2.29}$$

where ξ^n is the solution of (2.3) corresponding to $v = u^n$.

Remark 2.3. It is clear from the above observations that an efficient *Poisson solver* will be a basic tool for solving (2.1) (in fact a *finite-dimensional approximation* of it) by the above conjugate gradient algorithm.

Remark 2.4. The fact that $J'(v)$ is known through (2.27) is not at all a draw-back if a *Galerkin* or a *finite element method* is used to approximate (2.1). Indeed, we only need to know the value of $\langle J'(v), w \rangle$ for w belonging to a basis of the *finite-dimensional subspace* of $H_0^1(\Omega)$ corresponding to the Galerkin or finite element approximation under consideration.

Remark 2.5. The above methodology extends easily to the solution of *nonlinear elliptic systems* like

$$\begin{aligned} -\Delta u_1 + u_1 \frac{\partial u_1}{\partial x_1} + u_2 \frac{\partial u_1}{\partial x_2} &= f_1 \quad \text{in } \Omega, \\ -\Delta u_2 + u_1 \frac{\partial u_2}{\partial x_1} + u_2 \frac{\partial u_2}{\partial x_2} &= f_2 \quad \text{in } \Omega, \\ u_1 = u_2 &= 0 \quad \text{on } \Gamma, \end{aligned} \tag{2.30}$$

where Ω is a bounded domain of $\boldsymbol{R}^2$ and where $f_1, f_2 \in H^{-1}(\Omega)$. Elliptic systems closely related to (2.30) occur in the solution of the *time dependent Navier–Stokes equations* by *alternating direction* methods (see Sec. 3).

2.4. A nonlinear least squares approach to arc length continuation methods.

2.4.1. *Generalities. Synopsis.* We would like to show in this section that the above least squares methodology can be (slightly) modified in order to solve nonlinear problems by arc length continuation methods directly inspired from H.B. Keller [29], [30] (where the basic iterative methods are Newton's and quasi–Newton's instead of conjugate gradient).

As test problem, we have chosen a variant of the nonlinear Dirichlet problem (2.1); let us consider the following family of nonlinear Dirichlet problems (to be solved in $H_0^1(\Omega)$), parametrized by $\lambda \in \boldsymbol{R}$:

$$\begin{cases} -\Delta u = \lambda T(u) \text{ in } \Omega, \\ u = 0 \quad \text{on } \Gamma; \end{cases} \tag{2.31}$$

(2.1) corresponds to $\lambda = 1$.

2.4.2. *Solution of* (2.31) *via arc length continuation methods.* Following [29], [30] (for which we refer for justification (see also [43])), we associate to (2.31) a "continuation" equation; we have chosen (from among other possibilities)

$$\int_\Omega |\underline{\nabla} \dot{u}|^2 dx + \dot{\lambda}^2 = 1, \tag{2.32}$$

where $\dot{u} = \partial u/\partial s$, $\dot{\lambda} = d\lambda/ds$, and where the *curvilinear abscissa* s is defined by

$$\delta s = \dot{\lambda}\delta\lambda + \int_\Omega \nabla \dot{u} \cdot \nabla \delta u \, dx, \tag{2.33}$$

or equivalently by

$$(\delta s)^2 = (\delta\lambda)^2 + \int_\Omega \nabla u \cdot \nabla \delta u \, dx. \tag{2.34}$$

We are considering in fact a path in $H_0^1(\Omega) \times \boldsymbol{R}$ whose arc length is defined by (2.32)–(2.34). Then, in order to solve (2.31), we consider the family (parametrized by s) of nonlinear systems (2.31), (2.32). In practice we shall approximate (2.31), (2.32) by the following *discrete family* of nonlinear systems, where Δs is an *arc length step, positive or negative* (possibly varying with n) and where $u^n \approx u(n\Delta s)$:

$$\textit{Take } u^0 = 0, \lambda^0 = 0 \textit{ and suppose that } \dot{\lambda}(0), \dot{u}(0) \textit{ are given} \tag{2.35}$$

(initialization (2.35) is justified by the fact that $u = 0$ is the *unique* solution of (2.31) if $\lambda = 0$); *then for* $n \geqslant 0$, *assuming that* $u^{n-1}, \lambda^{n-1}, u^n, \lambda^n$ *are known, we obtain* $\{u^{n+1}, \lambda^{n+1}\} \in H_0^1(\Omega) \times \boldsymbol{R}$ *from the solution of*

$$\begin{cases} -\Delta u^{n+1} = \lambda^{n+1} T(u^{n+1}) \quad \textit{in } \Omega, \\ u^{n+1} = 0 \quad \textit{on } \Gamma, \end{cases} \tag{2.36}$$

and

$$\int_{\Omega} \nabla(u^1-u^0)\cdot\nabla\dot{u}(0)\,dx+(\lambda^1-\lambda^0)\dot{\lambda}(0) = \Delta s \quad \text{if} \quad n=0, \tag{2.37}$$

$$\int_{\Omega} (\nabla u^{n+1}-u^n)\cdot\nabla\left(\frac{u^n-u^{n-1}}{\Delta s}\right)dx+(\lambda^{n+1}-\lambda^n)\left(\frac{\lambda^n-\lambda^{n-1}}{\Delta s}\right) = \Delta s$$

$$\text{if} \quad n\geqslant 1; \tag{2.38}$$

obtaining $\dot{u}(0)$ and $\dot{\lambda}(0)$ is an easy task since we have (from (2.31))

$$\begin{cases} -\Delta\dot{u}(0) = \dot{\lambda}(0)\ T(0) & \text{in } \Omega, \\ \dot{u}(0) = 0 & \text{on } \Gamma, \end{cases} \tag{2.39}$$

and therefore

$$\dot{\lambda}^2(0)(1+\int_{\Omega}|\nabla\hat{u}|^2dx) = 1, \tag{2.40}$$

where $\hat{u}\in H_0^1(\Omega)$ is the solution of

$$\begin{cases} -\Delta\hat{u} = T(0) & \text{in } \Omega, \\ \hat{u} = 0 & \text{on } \Gamma \end{cases} \tag{2.41}$$

(then clearly $\dot{u}(0) = \dot{\lambda}(0)\hat{u}$).

Relations (2.36)–(2.38) look like a discretization scheme for solving the Cauchy problem for *first order ordinary differential equations*; from this analogy we can derive many other discretization schemes for the approximation of (2.31), (2.32) (Runge–Kutta, multisteps, etc.) and also methods for the *automatic adjustment of* Δs.

2.4.3. *Nonlinear least squares and conjugate gradient solution of* (2.36)–(2.38). Without going into details (for which we refer to [21], [47]) we can solve (2.36)–(2.38) by a variant of algorithm (2.14)–(2.22) defined on the Hilbert space $H_0^1(\Omega)\times\boldsymbol{R}$ equipped with the metric and inner product corresponding to

$$\{v,\mu\}\to\int_{\Omega}|\nabla v|^2dx+\mu^2; \tag{2.42}$$

it is clear that other norms than (2.42) are possible, however, in all cases

the scaling of a conjugate gradient algorithm using a discrete variant of

$$\begin{bmatrix} -\Delta & 0 \\ 0 & 1 \end{bmatrix} \quad \text{(or similar operators)} \tag{2.43}$$

will require an efficient solver and the conclusions of Sec. 2.3 (namely Remark 2.3) still hold.

Remark 2.6. To initialize the conjugate gradient algorithm solving (2.36)–(2.38) we have used $\{2\lambda^n - \lambda^{n-1}, 2u^n - u^{n-1}\}$ as initial guess to compute $\{\lambda^{n+1}, u^{n+1}\}$ (this supposes Δs to be constant); with such a choice we obtain a much faster convergence than by taking $\{\lambda^n, u^n\}$ as initial guess.

2.4.4. *Applications.* We describe in this section the solution of *nonlinear eigenvalue problems* by the continuation methods described above. The first problem is the *Bratu problem* and will be discussed in more details. The second one is the solution of the *Von Karman equations for nonlinear elastic plates* and will be discussed quite briefly.

2.4.4.1. *Formulation of the first test problem.* We shall apply the methods described in Sections 2.4.2, 2.4.3 to the solution of the following classical problem (known sometimes as the Bratu problem)

$$\begin{cases} -\Delta u = \lambda e^u \quad in\ \Omega, \\ u = 0 \quad on\ \Gamma, \end{cases} \tag{2.44}$$

where Ω is a bounded domain in $\boldsymbol{R}^N$; we have to observe that unless $N = 1$, the mapping T defined by

$$T(v) = e^v, \qquad v \in H_0^1(\Omega),$$

is not continuous from $H_0^1(\Omega)$ to $H^{-1}(\Omega)$. We consider only the case where $\lambda = 0$, since if $\lambda < 0$, the operator $v \to -\Delta v - \lambda e^v$ is *monotone* which implies that (2.44) has a unique solution.

If $\lambda > 0$, problem (2.44) has been considered by many authors; with regard to recent publications let us mention, among others, references [1], [15], [16], [39]–[41]. From the numerical point of view, problem (2.44) has been investigated in [13], [21], [31], [47] among others.

2.4.4.2. *Numerical solution of* (2.44) *by the methods of Sections* 2.4.2, 2.4.3. We have chosen the particular case of (2.44) where $\Omega =]0,1[\times]0,1[$.

The practical application of the methods of Sections 2.4.2, 2.4.3 requires the reduction of (2.44) to a *finite-dimensional problem*; this can done by *finite elements* or (owing to the simplicity of Ω) by *finite differences*.

Actually, the results presented here have been obtained using a finite element method with piecewise linear approximations (see [19], [21], [47] for more details). The continuation method (2.35)–(2.38) has been applied with $\Delta s = 0.1$; we observe that $T(0) = 1$ in (2.41); algorithm (2.35)–(2.38) ran "nicely" since an accurate *least squares solution* of the nonlinear system (2.34), (2.38) required basically no more than 3 to 4 *conjugate gradient* iterations, *even close to the turning point.*

We have shown in Figure 2.1 the *maximal value* (reached at $x_1 = x_2 = 0.5$) of the computed solution u_h as a function of λ; the computed turning point is at $\lambda = 6.8591$ ...

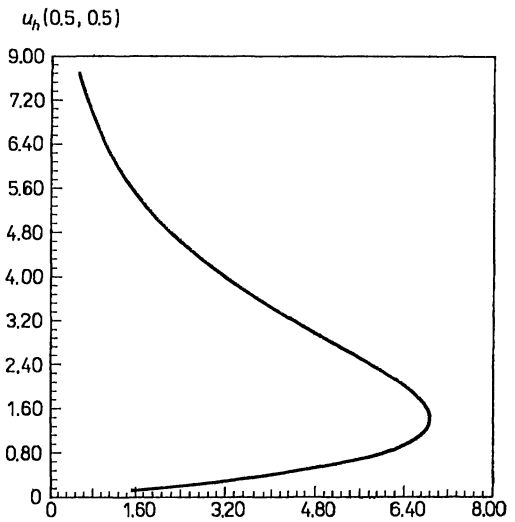


Fig. 2.1

2.4.4.3. *A second test problem.* The *least squares-continuation* methods described in Sections 2.4.2, 2.4.3 have been applied to the solution of nonlinear problems more complicated than (2.44); let us mention here the Navier–Stokes equations for incompressible viscous fluids at high Reynolds' number and also problems involving genuine bifurcation phenomenon like the *Von Karman equations* for plates. The details of these calcu-

lations can be found in [21], [47], [48]. We consider briefly here the *Von Karman equations for thin clamped plates*:

$$\begin{cases} \textit{Find } u, \varphi \in H_0^2(\Omega) \textit{ such that} \\ \Delta^2 u = \lambda[\theta_0, u] + [\varphi, u] + f \quad \textit{in } \Omega, \\ \Delta^2 \varphi = -[u, u] \quad \textit{in } \Omega, \end{cases} \tag{2.45}$$

where:

(i) $\Omega(\subset \boldsymbol{R}^2)$ is the two-dimensional spatial domain associated to the thin plate under consideration,

(ii) θ_0 and f are given functions, f being the density of external forces normal to the plate,

(iii) λ is a factor of proportionality for the external forces acting in the plane of the plate,

(iv) u is the vertical displacement and φ is the so-called *Airy function*,

(v) $[\cdot, \cdot]$ is defined by

$$[v, w] = \frac{\partial^2 v}{\partial x_1^2}\frac{\partial^2 w}{\partial x_2^2} + \frac{\partial^2 v}{\partial x_2^2}\frac{\partial^2 w}{\partial x_1^2} - 2\frac{\partial^2 v}{\partial x_1 \partial x_2}\frac{\partial^2 w}{\partial x_1 \partial x_2},$$

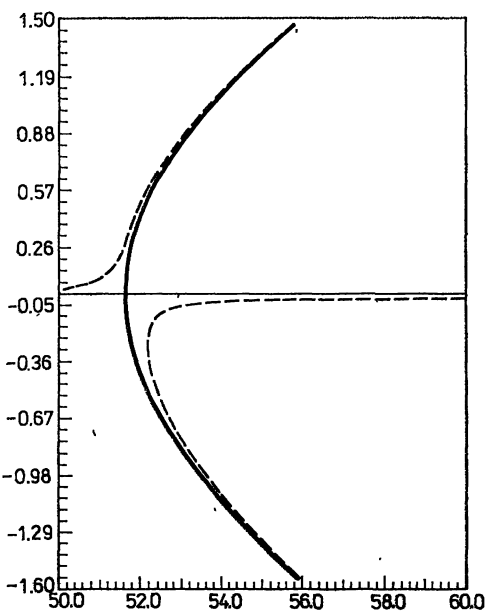


Fig. 2.2

(vi) $u, \varphi \in H_0^2(\Omega) \Rightarrow u = \dfrac{\partial u}{\partial n} = \varphi = \dfrac{\partial \varphi}{\partial n} = 0$ on the boundary Γ of Ω.

We have shown in Figure 2.2 the results obtained in [48] using an arc length least-squares combination method, whose principles are the same as those discussed in Sections 2.4.2, 2.4.3.

Figure 2.2 represents the maximum value of u on Ω versus λ, for $f = 0$ (continuous curve) and f a positive constant (dotted curve); we observe a bifurcation phenomenon at the first eigenvalue of the linearized problem; see [48] for more details.

3. Application to the solution of the Navier–Stokes equations for incompressible viscous fluids

We discuss briefly in this section the numerical solution of the Navier–Stokes equations for incompressible viscous fluids. For more details, see [19], Chapter 7 and Appendix 3.

3.1. Formulation of the time dependent Navier–Stokes equations for incompressible viscous fluids. Let us consider a *Newtonian incompressible viscous fluid.* If Ω and Γ denote the region of the flow ($\Omega \subset \boldsymbol{R}^N$, $N = 2$, 3 in practice) and its boundary, respectively, then the flow is governed by the following *Navier–Stokes equations*

$$\frac{\partial \boldsymbol{u}}{\partial t} - \nu \Delta \boldsymbol{u} + (\boldsymbol{u} \cdot \nabla)\boldsymbol{u} + \nabla p = \boldsymbol{f} \quad \text{in } \Omega, \tag{3.1}$$

$$\nabla \cdot \boldsymbol{u} = 0 \quad \textit{in } \Omega \textit{ (incompressibility condition)}. \tag{3.2}$$

In (3.1), (3.2) we have

(a) $\nabla = \left\{\dfrac{\partial}{\partial x_i}\right\}_{i=1}^{N}$, $\quad \Delta = \nabla^2 = \displaystyle\sum_{i=1}^{N} \frac{\partial^2}{\partial x_i^2}$,

(b) $\boldsymbol{u} = \{u_i\}_{i=1}^N$ is the *flow velocity*,

(c) p is the *pressure*,

(d) ν is the *viscosity* of the fluid,

(e) $\boldsymbol{f}$ is the *density of external forces.*

In (3.1), $(\boldsymbol{u} \cdot \nabla)\boldsymbol{u}$ is a *symbolic notation* for the nonlinear (vector) term

$$\left\{\sum_{i=1}^{N} u_j \frac{\partial u_i}{\partial x_j}\right\}_{i=1}^{N}.$$

Boundary conditions have to be added; for example, in the case of the airfoil B of Figure 3.1, we have (since the fluid is *viscous*) the following *adherence condition*

$$\boldsymbol{u} = \boldsymbol{0} \quad \text{on } \partial B \doteq \Gamma_B; \tag{3.3}$$

typical conditions at infinity are

$$\boldsymbol{u} = \boldsymbol{u}_\infty, \tag{3.4}$$

where $\boldsymbol{u}_\infty$ is a *constant* vector (with regard to the space variables at least).

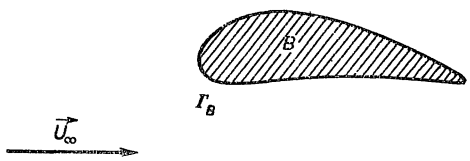


Fig. 3.1

If Ω is a bounded region in $\boldsymbol{R}^N$, we may prescribe as *boundary condition*

$$\boldsymbol{u} = \boldsymbol{g} \quad \text{on } \Gamma, \tag{3.5}$$

where (by the incompressibility of the fluid) the given function $\boldsymbol{g}$ has to satisfy

$$\int_\Gamma \boldsymbol{g} \cdot \boldsymbol{n} \, d\Gamma = 0, \tag{3.6}$$

where $\boldsymbol{n}$ is the outward *unit* vector *normal* to Γ.

Finally, for time dependent problem (3.1), (3.2) an *initial condition* such as

$$\boldsymbol{u}(x, 0) = \boldsymbol{u}_0(x) \quad \text{a.e. on } \Omega, \tag{3.7}$$

with $\boldsymbol{u}_0$ given, is usually prescribed.

Looking at the above equations, we observe three principal difficulties (even for flows with low Reynolds' numbers in bounded regions Ω), which are:

(i) The *nonlinear* term $(\boldsymbol{u} \cdot \nabla)\boldsymbol{u}$ *in* (3.1),

(ii) The *incompressibility* condition (3.2),

(iii) The fact that the solutions of the Navier–Stokes equations are *vector-valued* functions of x, t, whose components are coupled by $(\boldsymbol{u} \cdot \nabla)\boldsymbol{u}$ and by the incompressibility condition $\nabla \cdot \boldsymbol{u} = 0$.

Using convenient *alternating direction* methods for the time discretization of the Navier–Stokes equations, we shall be able to *decouple* the difficulties due to the nonlinearity and to incompressibility. For simplicity, we suppose from now on that Ω is *bounded* and that we have (3.5) as boundary conditions (with $\boldsymbol{g}$ satisfying (3.6) and possibly depending on t).

3.2. Time discretization by alternating direction methods. Let Δt (> 0) be a *time discretization* step and θ a parameter such that $0 < \theta < 1$.

3.2.1. *A first alternating direction method.* We first consider the following alternating direction method (of Peaceman–Rachford type):

$$\boldsymbol{u}^0 = \boldsymbol{u}_0, \tag{3.8}$$

then for $n \geqslant 0$, $\boldsymbol{u}^n$ *being known compute* $\{\boldsymbol{u}^{n+1/2}, p^{n+1/2}\}$ *and* $\boldsymbol{u}^{n+1}$ *by solving*

$$\begin{cases} \dfrac{\boldsymbol{u}^{n+1/2} - \boldsymbol{u}^n}{\Delta t/2} - \theta\nu\Delta\boldsymbol{u}^{n+1/2} + \nabla p^{n+1/2} = \boldsymbol{f}^{n+1/2} + (1-\theta)\nu\Delta\boldsymbol{u}^n - (\boldsymbol{u}^n \cdot \nabla)\boldsymbol{u}^n \text{ in } \Omega, \\ \nabla \cdot \boldsymbol{u}^{n+1/2} = 0 \text{ in } \Omega, \\ \boldsymbol{u}^{n+1/2} = \boldsymbol{g}^{n+1/2} \text{ on } \Gamma, \end{cases} \tag{3.9}$$

and

$$\begin{cases} \dfrac{\boldsymbol{u}^{n+1} - \boldsymbol{u}^{n+1/2}}{\Delta t/2} - (1-\theta)\nu\Delta\boldsymbol{u}^{n+1/2} + (\boldsymbol{u}^{n+1} \cdot \nabla)\boldsymbol{u}^{n+1} = \boldsymbol{f}^{n+1} + \theta\nu\Delta\boldsymbol{u}^{n+1} - \nabla p^{n+12} \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{in } \Omega, \\ \boldsymbol{u}^{n+1} = \boldsymbol{g}^{n+1} \quad \text{on } \Gamma, \end{cases} \tag{3.10}$$

respectively.

We use the notation $\boldsymbol{f}^j(x) = \boldsymbol{f}(x, j\Delta t)$, $\boldsymbol{g}^j(x) = \boldsymbol{g}(x, j\Delta t)$, and $\boldsymbol{u}^j(x)$ is an approximation of $\boldsymbol{u}(x, j\Delta t)$.

3.2.2. *A second alternating direction method.* We now consider the following alternating direction method (of Strang type):

$$\boldsymbol{u}^0 = \boldsymbol{u}_0, \tag{3.11}$$

then for $n \geqslant 0$ *and starting from* u^n *we solve*

$$\begin{cases} \dfrac{u^{n+1/4}-u^n}{\Delta t/4} - \theta\nu\Delta u^{n+1/4} + \nabla p^{n+1/4} = f^{n+1/4} + (1-\theta)\nu\Delta u^n - (u^n\cdot\nabla)u^n \text{ in } \Omega, \\ \nabla\cdot u^{n+1/4} = 0 \text{ in } \Omega, \quad u^{n+1/4} = g^{n+1/4} \text{ on } \Gamma, \end{cases} \tag{3.12}$$

$$\begin{cases} \dfrac{u^{n+3/4}-u^{n+1/4}}{\Delta t/2} - (1-\theta)\nu\Delta u^{n+3/4} + (u^{n+3/4}\cdot\underline{\nabla})u^{n+3/4} \\ \qquad = f^{n+3/4} + \theta\nu\Delta u^{n+1/4} - \nabla^{n+1/4} \text{ in } \Omega, \\ u^{n+3/4} = g^{n+3/4} \text{ on } \Gamma, \end{cases} \tag{3.13}$$

$$\begin{cases} \dfrac{u^{n+1}-u^{n+3/4}}{\Delta t/4} - \theta\nu\Delta u^{n+1} + \underline{\nabla} p^{n+1} = f^{n+1} + (1-\theta)\nu\Delta u^{n+3/4} - (u^{n+3/4}\cdot\nabla)u^{n+3/4} \\ \qquad \text{in } \Omega, \\ \nabla\cdot u^{n+1} = 0 \text{ in } \Omega, \quad u^{n+1} = g^{n+1} \text{ on } \Gamma. \end{cases} \tag{3.14}$$

3.2.3. *Some comments and remarks concerning the alternating direction schemes* (3.8)–(3.10) *and* (3.11)–(3.14). Using the two alternating schemes described in Sections 3.2.1, 3.2.2, we have been able to *decouple* nonlinearity and incompressibility in the Navier–Stokes equations. We shall describe in the following sections the specific treatment of the subproblems encountered at each step of (3.8)–(3.10) and (3.11)–(3.14); we shall first consider the case where the subproblems are still continuous in space (since the formalism of continuous problems is much simpler), and then the discrete case where a finite element method is used to approximate in space the Navier–Stokes equations.

Scheme (3.8)–(3.10) has a *truncation error* in $O(\Delta t)$; due to the *symmetrization* process that it involves, scheme (3.11)–(3.14) has a truncation error in $O(|\Delta t|^2)$.

We observe that $u^{n+1/2}$ and $u^{n+1/4}$, u^{n+1} are obtained from the solution of linear problems ((3.9) and (3.12), (3.14), respectively) very close to the steady Stokes problem. Despite its greater complexity scheme (3.11)-(3.14) is almost as economical in use as scheme (3.8)–(3.10); this is mainly due to the fact that the "quasi" steady Stokes problems (3.9) and (3.12), (3.14) (in fact, convenient finite element approximation of them) can be solved by quite efficient solvers resulting in that most of the computer time used to solve a full alternating direction step ((3.9), (3.10) or (3.12)–

(3.14)) is in fact used to solve the nonlinear subproblem ((3.10) or (3.13)). The good choice for θ is $\theta = 1/2$ (resp. $\theta = 1/3$) if one uses scheme (3.8)–(3.10) (resp. (3.11)–(3.14)); this follows from the fact that with the above choices for θ, many computer subprograms can be used for both the linear and nonlinear subproblems, resulting therefore in quite substantial computer core memory savings.

Remark 3.1. A variant of scheme (3.8)–(3.10) is the following (it corresponds to $\theta = 1$):

$$\boldsymbol{u}^0 = \boldsymbol{u}_0, \tag{3.15}$$

then, for $n \geqslant 0$ *and starting from* $\boldsymbol{u}^n$,

$$\begin{cases} \dfrac{\boldsymbol{u}^{n+1/2} - \boldsymbol{u}^n}{\Delta t/2} - \nu \Delta \boldsymbol{u}^{n+1/2} + \nabla p^{n+1/2} = \boldsymbol{f}^{n+1/2} - (\boldsymbol{u}^n \cdot \nabla)\boldsymbol{u}^n \quad \text{in } \Omega, \\ \nabla \cdot \boldsymbol{u}^{n+1/2} = 0 \text{ in } \Omega, \quad \boldsymbol{u}^{n+1/2} = \boldsymbol{g}^{n+1/2} \quad \text{on } \Gamma, \end{cases} \tag{3.16}$$

$$\begin{cases} \dfrac{\boldsymbol{u}^{n+1} - \boldsymbol{u}^{n+1/2}}{\Delta t/2} + (\boldsymbol{u}^{n+1/2} \cdot \nabla)\boldsymbol{u}^{n+1} = \boldsymbol{f}^{n+1} + \nu \Delta \boldsymbol{u}^{n+1/2} \text{ in } \Omega, \\ \boldsymbol{u}^{n+1} = \boldsymbol{g}^{n+1} \text{ on } \Gamma_-^{n+1/2}, \end{cases} \tag{3.17}$$

where

$$\Gamma_-^{n+1/2} = \{x \mid x \in \Gamma,\ \boldsymbol{g}^{n+1/2}(x) \cdot \boldsymbol{n}(x) < 0\}.$$

Both subproblems (3.16), (3.17) are *linear*; the first one is also a "quasi" steady Stokes problem and the second one, which is a *first order system*, can be solved by a *method of characteristics*.

A similar remark holds for scheme (3.11)–(3.14).

Such methods have been used by several authors, the space discretization having been done by finite element methods very close to those described in Section 3.5 of this paper (see [3], [28], [44] for a discussion of those characteristics-finite element methods for solving Navier–Stokes equations); these characteristics-finite element methods are slightly *dissipative*, which may be a drawback in some applications.

Remark 3.2. In order to improve the *well-posedness* properties of the nonlinear steps (3.10) and (3.13) (and also to simplify convergence proofs)

one may replace the original nonlinear term $B(u) = (u \cdot \nabla)u$ by

$$\tilde{B}(u) = (u \cdot \nabla)u + \tfrac{1}{2}u(\nabla \cdot u),$$

following the lines of [52]; it is clear that $B(u) = \tilde{B}(u)$ if $\nabla \cdot u = 0$. Actually, the good property of $\tilde{B}$ is that

$$\begin{cases} \int_\Omega \tilde{B}(v) \cdot v \, dx = 0 \quad \forall v \text{ sufficiently smooth} \\ \text{such that } v = 0 \text{ on } \Gamma, \text{ even if } \nabla \cdot v \neq 0 \end{cases}$$

(see [52] for more details).

The numerical results obtained using either B or $\tilde{B}$ are practically identical, provided that Δt is "reasonably" small.

3.3. Least squares-conjugate gradient solution of the nonlinear subproblems.

3.3.1. *Classical and variational formulations.* At each full step of the alternating direction methods (3.8)–(3.10) and (3.11)–(3.14) we have to solve a *nonlinear elliptic system* of the following type:

$$\begin{cases} \alpha u - \nu \Delta u + (u \cdot \nabla)u = f \text{ in } \Omega, \\ u = g \quad \text{on } \Gamma, \end{cases} \tag{3.18}$$

where α and ν are two positive parameters and where f and g are two given functions defined on Ω and Γ, respectively. We shall not discuss here the existence and uniqueness of solutions for problem (3.18). We introduce the following spaces of *Sobolev's type*

$$V_0 = \left(H_0^1(\Omega)\right)^N, \tag{3.19}$$

$$V_g = \{v \mid v \in \left(H^1(\Omega)\right)^N, v = g \text{ on } \Gamma\}; \tag{3.20}$$

if g is sufficiently smooth, then $V_g \neq \emptyset$.

We use in the following the notation

$$u \cdot v = \sum_{i=1}^{N} u_i v_i, \quad \nabla u \cdot \nabla v = \sum_{i=1}^{N} \nabla u_i \cdot \nabla v_i = \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{\partial u_i}{\partial x_j} \frac{\partial v_i}{\partial x_j}.$$

From Green's formula we have, for sufficiently smooth functions $\boldsymbol{u}$ and $\boldsymbol{v}$, belonging to $(H^1(\Omega))^N$ and V_0, respectively,

$$-\int_\Omega \Delta \boldsymbol{u}\cdot\boldsymbol{v}\,dx = \int_\Omega \nabla\boldsymbol{u}\cdot\nabla\boldsymbol{v}\,dx. \tag{3.21}$$

If $\boldsymbol{u}$ is a solution of (3.18) it is also a solution of the *nonlinear variational problem*

$$\begin{cases} \textit{Find } \boldsymbol{u}\in V_g \textit{ such that} \\ \alpha\int_\Omega \boldsymbol{u}\cdot\boldsymbol{v}dx+\nu\int_\Omega \nabla\boldsymbol{u}\cdot\nabla\boldsymbol{v}\,dx+\int_\Omega\big((\boldsymbol{u}\cdot\nabla)\boldsymbol{u}\big)\cdot\boldsymbol{v}\,dx=\int_\Omega \boldsymbol{f}\cdot\boldsymbol{v}\,dx \quad \forall\boldsymbol{v}\in V_0, \end{cases} \tag{3.22}$$

and *conversely*. We observe that (3.18), (3.22) is not equivalent to a problem of the *calculus of variations* since there is no functional of $\boldsymbol{v}$ with $(\boldsymbol{v}\cdot\nabla)\boldsymbol{v}$ as differential. Using, however, a nonlinear least-squares formulation like those discussed is Section 2, we shall be able to solve (3.18), (3.22) by efficient methods from *nonlinear programming*, like conjugate gradient, for example. The *finite element* approximation of (3.18), (3.22) is briefly discussed in Section 3.5.

3.3.2. *Least-squares formulation of* (3.18), (3.22). Let $\boldsymbol{v}\in V_g$; from $\boldsymbol{v}$ we define $\boldsymbol{y}\ (=\boldsymbol{y}(\boldsymbol{v}))\in V_0$ as the solution of

$$\begin{cases} \alpha\boldsymbol{y}-\nu\Delta\boldsymbol{y} = \alpha\boldsymbol{v}-\alpha\Delta\boldsymbol{v}+(\boldsymbol{v}\cdot\nabla)\boldsymbol{v}-\boldsymbol{f} \quad \textit{in } \Omega, \\ \boldsymbol{y}=\boldsymbol{0} \quad \textit{on } \Gamma. \end{cases} \tag{3.23}$$

We observe that $\boldsymbol{y}$ is obtained from $\boldsymbol{v}$ via the solution of N *uncoupled* linear Poisson problems (one for each component of $\boldsymbol{y}$); using (3.21), it can be shown that (3.23) is actually *equivalent* to the *linear variational problem*

$$\begin{cases} \textit{Find } \boldsymbol{y}\in V_0 \textit{ such that } \forall\boldsymbol{z}\in V_0 \textit{ we have} \\ \alpha\int_\Omega \boldsymbol{y}\cdot\boldsymbol{z}\,dx+\nu\int_\Omega \nabla\boldsymbol{y}\cdot\nabla\boldsymbol{z}\,dx \\ =\alpha\int_\Omega \boldsymbol{v}\cdot\boldsymbol{z}\,dx+\nu\int_\Omega \nabla\boldsymbol{v}\cdot\nabla\boldsymbol{z}\,dx+\int_\Omega\big((\boldsymbol{v}\cdot\nabla)\boldsymbol{v}\big)\cdot\boldsymbol{z}\,dx-\int_\Omega \boldsymbol{f}\cdot\boldsymbol{z}\,dx, \end{cases} \tag{3.24}$$

which has a unique solution. Suppose now that $\boldsymbol{v}$ is a solution of (3.18), (3.22); the corresponding $\boldsymbol{y}$ (obtained through the solution of (3.23), (3.24)) is clearly $\boldsymbol{0}$. From these observations it is quite natural to introduce

the following *nonlinear least squares* formulation of problem (3.18), (3.22) (which is a straightforward variant of the one discussed in Section 2.2.2 for the solution of problem (2.1)):

$$\begin{cases} \textit{Find } \boldsymbol{u} \in V_g \textit{ such that} \\ J(\boldsymbol{u}) \leqslant J(\boldsymbol{v}) \quad \forall \boldsymbol{v} \in V_g, \end{cases} \tag{3.25}$$

where J: $(H^1(\Omega))^N \to \boldsymbol{R}$ is the functional of $\boldsymbol{v}$ defined by

$$J(\boldsymbol{v}) = \tfrac{1}{2} \int_\Omega \{\alpha |\boldsymbol{y}|^2 + \nu |\nabla \boldsymbol{y}|^2\} dx,$$

with $\boldsymbol{y}$ defined from $\boldsymbol{v}$ by the solution of the linear problem (3.23), (3.24).

3.3.3. *Conjugate gradient solution of the least-squares problem* (3.25). We use the following generalization of algorithm (2.14)–(2.22) discussed in Section 2.3:

Step 0: *Initialization.*

$$\boldsymbol{u}^0 \in V_g \textit{ given}; \tag{3.26}$$

we define $\boldsymbol{g}^0, \boldsymbol{w}^0 \in V_0$ *by*

$$\alpha \int_\Omega \boldsymbol{g}^0 \cdot \boldsymbol{z}\, dx + \nu \int_\Omega \nabla \boldsymbol{g}^0 \cdot \nabla \boldsymbol{z}\, dx = \langle J'(\boldsymbol{u}^0),\ \boldsymbol{z} \rangle \ \forall \boldsymbol{z} \in V_0, \quad \boldsymbol{g}^0 \in V_0, \tag{3.27}$$

$$\boldsymbol{w}^0 = \boldsymbol{g}^0, \tag{3.28}$$

respectively. ■

Then, for $n \geqslant 0$, *assuming that* $\boldsymbol{u}^n, \boldsymbol{g}^n, \boldsymbol{w}^n$ *are known, we obtain* $\boldsymbol{u}^{n+1}$, $\boldsymbol{g}^{n+1}, \boldsymbol{w}^{n+1}$ *by*

Step 1: *Descent.*

$$\begin{cases} \textit{Find } \lambda^n \in \boldsymbol{R} \textit{ such that} \\ J(\boldsymbol{u}^n - \lambda^n \boldsymbol{w}^n) \leqslant J(\boldsymbol{u}^n - \lambda \boldsymbol{w}^n) \quad \forall \lambda \in \boldsymbol{R}, \end{cases} \tag{3.29}$$

$$\boldsymbol{u}^{n+1} = \boldsymbol{u}^n - \lambda^n \boldsymbol{w}^n. \ \blacksquare \tag{3.30}$$

Step 2: Calculation of the new descent direction.

$$\begin{cases} \text{Find } \boldsymbol{g}^{n+1} \in V_0 \text{ such that} \\ \alpha \int_\Omega \boldsymbol{g}^{n+1} \cdot \boldsymbol{z}\, dx + \nu \int_\Omega \nabla \boldsymbol{g}^{n+1} \cdot \nabla \boldsymbol{z}\, dx = \langle J'(\boldsymbol{u}^{n+1}), \boldsymbol{z} \rangle \quad \forall \boldsymbol{z} \in V_0, \end{cases} \tag{3.31}$$

$$\gamma_n = \frac{\alpha \int_\Omega \boldsymbol{g}^{n+1} \cdot (\boldsymbol{g}^{n+1} - \boldsymbol{g}^n)\, dx + \nu \int_\Omega \nabla \boldsymbol{g}^{n+1} \cdot \nabla (\boldsymbol{g}^{n+1} - \boldsymbol{g}^n)\, dx}{\alpha \int_\Omega |\boldsymbol{g}^n|\, dx + \nu \int_\Omega |\nabla \boldsymbol{g}^n|^2\, dx}, \tag{3.32}$$

$$\boldsymbol{w}^{n+1} = \boldsymbol{g}^{n+1} + \gamma_n \boldsymbol{w}^n, \tag{3.33}$$

$n = n+1$, *go to* (3.29). ■

As we shall see below, applying algorithm (3.26)–(3.33) to solve (3.25) requires, at each iteration, the solution of several Dirichlet problems for the elliptic operator $\alpha I - \nu \Delta$.

Calculation of J': By a method similar to the one used in Section 2.3 we can prove that $J'(\boldsymbol{v})$ can be identified with the linear functional from V_0 to $\boldsymbol{R}$ defined by

$$\langle J'(\boldsymbol{v}), \boldsymbol{z} \rangle = \alpha \int_\Omega \boldsymbol{y} \cdot \boldsymbol{z}\, dx + \nu \int_\Omega \nabla \boldsymbol{y} \cdot \nabla \boldsymbol{z}\, dx + \int_\Omega \boldsymbol{y} \cdot (\boldsymbol{z} \cdot \nabla) \boldsymbol{v}\, dx + \int_\Omega \boldsymbol{y} \cdot (\boldsymbol{v} \cdot \nabla) \boldsymbol{z}\, dx \quad \forall \boldsymbol{z} \in V_0, \tag{3.34}$$

with $\boldsymbol{y}$ being the solution of (3.23), (3.24); it has therefore a *purely integral representation*, a property of major importance in view of *finite element* implementations of algorithm (3.26)–(3.33).

From the above results, to obtain $\langle J'(\boldsymbol{u}^{n+1}), \boldsymbol{z} \rangle$ we proceed as follows:

(i) We compute $\boldsymbol{y}^{n+1}$ from $\boldsymbol{u}^{n+1}$ through the solution of (3.23), (3.24) with $\boldsymbol{v} = \boldsymbol{u}^{n+1}$.

(ii) We obtain $\langle J'(\boldsymbol{u}^{n+1}), \boldsymbol{z} \rangle$ by taking $\boldsymbol{v} = \boldsymbol{u}^{n+1}$ and $\boldsymbol{y} = \boldsymbol{y}^{n+1}$ in (3.34).

Further comments on algorithm (3.26)–(3.33). Each step of algorithm (3.26)–(3.33) requires the solution of several Dirichlet systems for the operator $\alpha I - \nu \Delta$; more precisely, we have to solve the following systems:

(i) System (3.23), (3.24) to obtain $\boldsymbol{y}^{n+1}$ from $\boldsymbol{u}^{n+1}$,

(ii) System (3.31) to obtain $\boldsymbol{g}^{n+1}$ from $\boldsymbol{u}^{n+1}$, $\boldsymbol{y}^{n+1}$,

(iii) Two systems to obtain the coefficients of the *quartic* polynomial $\lambda \to J(u^n - \lambda w^n)$.

Thus we have to solve 4 Dirichlet systems at each iteration (or equivalently $4N$ scalar Dirichlet problems for $\alpha I - \nu\Delta$ at each iteration); from these observations it appears clearly that the practical implementation of algorithm (3.26)–(3.33) will require an efficient (direct or iterative) *elliptic solver* (in fact $3N$ problems suffice).

The solution of the one-dimensional problem (3.29) can be done very efficiently since it is equivalent to finding the roots of a single variable cubic polynomial whose coefficients are known.

As a last comment, we would like to mention that algorithm (3.26)–(3.33) (in fact, its finite element variants) is quite efficient; when used in combination with the alternating direction methods of Section 3.2, to solve the test problems of Section 3.6, three iterations suffice to reduce the value of the cost function J by a factor of 10^4 to 10^6.

Remark 3.3. The above method can be applied also to the variant of problem (3.18), obtained by replacing $(u \cdot \nabla)u$ by $(u \cdot \nabla)u + \frac{1}{2}(\nabla \cdot u)u$ in (3.18) (cf. Remark 3.2), i.e., to the nonlinear Dirichlet problem

$$\alpha u - \nu\Delta u + (u \cdot \nabla)u + \tfrac{1}{2}(\nabla \cdot u)u = f \quad in\ \Omega,$$

$$u = g \quad on\ \Gamma.$$

3.4. Solution of the "quasi" Stokes linear subproblems.

3.4.1. *Formulation. Synopsis.* At each full step of the alternating direction methods (3.8)–(3.10) and (3.11)–(3.14) we have to solve a *linear problem* of the following type:

$$\begin{cases} \alpha u - \nu\Delta u + \nabla p = f \quad in\ \Omega, \\ \nabla \cdot u = 0 \quad in\ \Omega, \\ u = g \ on\ \Gamma \left(with \int_\Gamma g \cdot n\, d\Gamma = 0\right), \end{cases} \tag{3.35}$$

where α and ν are two positive parameters and f and g are two given functions defined on Ω and Γ, respectively.

We recall (see, e.g., [18], [36], [50], [52]) that if f and g are sufficiently smooth, then problem (3.35) has a unique solution in $V_g \times (L^2(\Omega)/\boldsymbol{R})$ (with V_g still defined by (3.20); $p \in L^2(\Omega)/\boldsymbol{R}$ means that p is defined only up to an arbitrary constant). We shall briefly discuss in the following sections

some iterative methods for solving (3.35), quite easy to implement using finite element methods (other methods are discussed in [19], Chapter 7).

3.4.2. *Gradient and conjugate gradient methods for solving* (3.35). A standard method to solve (3.35) is defined as follows:

$$p^0 \in L^2(\Omega), \quad \textit{given}, \tag{3.36}$$

then, for $n \geqslant 0$, *define* $\boldsymbol{u}^n$ *and* p^{n+1} *from* p^n *by*

$$\begin{cases} a\boldsymbol{u}^n - \nu \Delta \boldsymbol{u}^n = \boldsymbol{f} - \nabla p^n & \textit{in } \Omega, \\ \boldsymbol{u}^n = \boldsymbol{g} & \textit{on } \Gamma, \end{cases} \tag{3.37}$$

$$p^{n+1} = p^n - \varrho \nabla \cdot \boldsymbol{u}^n. \tag{3.38}$$

Concerning the convergence of (3.36)–(3.38), one can prove (see e.g. [19], Chapter 7, [18], [52]) the following

PROPOSITION 3.1. *Suppose that*

$$0 < \varrho < 2\nu/N; \tag{3.39}$$

we then have

$$\lim_{n \to +\infty} \{\boldsymbol{u}^n, p^n\} = \{\boldsymbol{u}, p_0\} \quad \textit{strongly in } (H^1(\Omega))^N \times L^2(\Omega), \tag{3.40}$$

where $\{\boldsymbol{u}, p_0\}$ *is the solution of* (3.35) *such that*

$$\int_\Omega p_0 \, dx = \int_\Omega p^0 \, dx \tag{3.41}$$

(*actually the convergence is linear*).

Remark 3.4. When applying algorithm (3.36)–(3.38) to solve the "quasi" Stokes problem (3.35) we have to solve at each iteration N *uncoupled* scalar Dirichlet problems for $aI - \nu\Delta$, to obtain $\boldsymbol{u}^n$ from p^n. We see again (as in Sec. 3.3) the importance of having efficient Dirichlet solvers for $aI - \nu\Delta$.

Remark 3.5. Instead of algorithm (3.36)–(3.38) we should rather use in practice the following conjugate gradient variant of it, whose convergence is much faster in most cases and which is, in addition, no more costly to implement:

Description of the conjugate gradient algorithm:
Step 0: *Initialization.*

$$p^0 \in L^2(\Omega) \quad \text{given arbitrarily,} \tag{3.42}$$

solve

$$\begin{cases} \alpha u^0 - \nu \Delta u^0 = f - \nabla p^0 & \text{in } \Omega, \\ u^0 = g & \text{on } \Gamma, \end{cases} \tag{3.43}$$

and set

$$g^0 = \nabla \cdot u^0, \tag{3.44}$$

$$w^0 = g^0. \tag{3.45}$$

Then for $n \geqslant 0$, *we obtain* p^{n+1}, g^{n+1}, w^{n+1} *from* p^n, g^n, w^n *by*
Step 1: *Descent. Compute* $\chi^n \in (H_0^1(\Omega))^N$ *as the solution of*

$$\begin{cases} \alpha \chi^n - \nu \Delta \chi^n = -\nabla w^n & \text{in } \Omega, \\ \chi^n = \mathbf{0} & \text{on } \Gamma, \end{cases} \tag{3.46}$$

then set

$$\varrho_n = \frac{\int_\Omega w^n g^n \, dx}{\int_\Omega \nabla \cdot \chi^n w^n \, dx} = \frac{\int_\Omega |g^n|^2 \, dx}{\int_\Omega \nabla \cdot \chi^n w^n \, dx}, \tag{3.47}$$

and finally

$$p^{n+1} = p^n - \varrho^n w^n. \tag{3.48}$$

Step 2: *Calculation of the new descent direction.*

$$g^{n+1} = g^n - \varrho_n \nabla \cdot \chi^n, \tag{3.49}$$

$$\gamma_n = \frac{\|g^{n+1}\|^2_{L^2(\Omega)}}{\|g^n\|^2_{L^2(\Omega)}}, \tag{3.50}$$

$$w^{n+1} = g^{n+1} + \gamma_n w^n. \tag{3.51}$$

Then take $n = n+1$ *and go to* (3.46).

Once the convergence of (3.42)–(3.51) to p_0 (that pressure solution such that $\int_\Omega p_0 \, dx = \int_\Omega p^0 \, dx$) has been settled we compute u from p_0 by the

solution of the Dirichlet system

$$\begin{cases} \alpha \boldsymbol{u} - \nu \Delta \boldsymbol{u} = \boldsymbol{f} - \nabla p_0 & \text{in } \Omega, \\ \boldsymbol{u} = \boldsymbol{g} & \text{on } \Gamma. \end{cases}$$

3.4.3. *Another iterative method for solving* (3.35). This method is in fact a penalization variant of algorithm (3.36)–(3.38) and is defined as follows (with r a *positive* parameter):

$$p^0 \in L^2(\Omega) \quad \textit{given}, \tag{3.52}$$

then, for $n \geqslant 0$, *define* $\boldsymbol{u}^n$ *and* p^{n+1} *from* p^n *by*

$$\begin{cases} \alpha \boldsymbol{u}^n - \nu \Delta \boldsymbol{u}^n - r \nabla (\nabla \cdot \boldsymbol{u}^n) = \boldsymbol{f} - \nabla p^n & \text{in } \Omega, \\ \boldsymbol{u}^n = \boldsymbol{g} & \text{on } \Gamma, \end{cases} \tag{3.53}$$

$$p^{n+1} = p^n - \varrho \, \nabla \cdot \boldsymbol{u}^n. \tag{3.54}$$

PROPOSITION 3.2. *Suppose that*

$$0 < \varrho < 2(r + \nu/N); \tag{3.55}$$

then the convergence result (3.40) *still holds for* $\{\boldsymbol{u}^n, p^n\}$.

For a proof see e.g. [19], Chapter 7.

Remark 3.6 (*About the choice of* ϱ *and* r). In practice, we should use $\varrho = r$, since it can be proved that in that case the convergence ratio of algorithm (3.52)–(3.54) is $O(r^{-1})$, for large value of r. In many applications, taking $r = 10^4 \nu$, we have a practical convergence of algorithm (3.52)–(3.54) in 3 to 4 iterations. There is, however, a practical upper bound for r: this follows from the fact that for too large values of r, problem (3.53) will be *ill-conditioned* and its practical solution sensitive to *round off errors*.

Remark 3.7. Problem (3.53) is more complicated to solve in practice than problem (3.37) since the components of $\boldsymbol{u}^n$ are coupled by the linear term $\nabla(\nabla \cdot \boldsymbol{u}^n)$. Actually the partial differential elliptic operator in the left-hand side of (3.53) is very close to the *linear elasticity operator*, and close variants of it occur naturally in *compressible* and/or *turbulent* viscous flow problems.

Remark 3.8. Other techniques for solving (3.35) are discussed in [19], Chapter 7.

3.5. Finite element approximation of the time dependent Navier–Stokes equations.

3.5.1. *Generalities. Synopsis.* We shall briefly discuss in this section a specific *finite element approximation* for the time dependent Navier–Stokes equations. Actually this method which leads to *continuous approximations* for both pressure and velocity is fairly simple and has been known for years; it has been advocated by Hood–Taylor [51], and also by other authors. Other finite element approximations of the Navier–Stokes equations can be found in e.g. [18]–[20], [52], [53]. A most important reference for the theoretical study of the convergence of the approximate solution of the time dependent Navier–Stokes equations is Heywood–Rannacher [26].

3.5.2. *Basic hypotheses. Fundamental discrete spaces.* We suppose for simplicity that Ω is a *bounded domain* in $\mathbf{R}^2$. With $\mathcal{T}_h$ a standard finite element triangulation of Ω, and h the maximal length of the edges of the triangles of $\mathcal{T}_h$, we introduce the following discrete spaces (with P_k = space of the polynomials in two variables of degree $\leqslant k$)

$$H_h^1 = \{q_h \mid q_h \in C^0(\bar{\Omega}),\ q_h|_T \in P_1\ \forall T \in \mathcal{T}_h\}, \tag{3.56}$$

$$V_h = \{\boldsymbol{v}_h \mid \boldsymbol{v}_h \in C^0(\bar{\Omega}) \times C^0(\bar{\Omega}),\ \boldsymbol{v}_h|_T \in P_2 \times P_2\ \forall T \in \mathcal{T}_h\}, \tag{3.57}$$

$$V_{0h} = V_0 \cap V_h = \{\boldsymbol{v}_h \mid \boldsymbol{v}_h \in V_h,\ \boldsymbol{v}_h = \mathbf{0}\ \text{on}\ \Gamma\}. \tag{3.58}$$

3.5.3. *Space discretization of the time dependent Navier–Stokes equations.* Using the above spaces H_h^1, V_h, V_{0h} we approximate the time dependent Navier–Stokes equations as follows:

Find $\{\boldsymbol{u}_h(t), p_h(t)\} \in V_h \times H_h^1 \quad \forall t \geqslant 0$, *such that*

$$\int_\Omega \frac{\partial \boldsymbol{u}_h}{\partial t} \cdot \boldsymbol{v}_h dx + \nu \int_\Omega \nabla \boldsymbol{u}_h \cdot \nabla \boldsymbol{v}_h dx + \int_\Omega (\boldsymbol{u}_h \cdot \nabla)\boldsymbol{u}_h \cdot \boldsymbol{v}_h dx + \int_\Omega \nabla p_h \cdot \boldsymbol{v}_h dx$$

$$= \int_\Omega \boldsymbol{f}_h \cdot \boldsymbol{v}_h dx \quad \forall \boldsymbol{v}_h \in V_{0h}, \tag{3.59}$$

$$\int_\Omega \nabla \cdot \boldsymbol{u}_h q_h dx = 0 \quad \forall q_h \in H_h^1, \tag{3.60}$$

$$\boldsymbol{u}_h = \boldsymbol{g}_h \quad \text{on}\ \Gamma, \tag{3.61}$$

$$\boldsymbol{u}_h(x, 0) = \boldsymbol{u}_{0h}(x) (\boldsymbol{u}_{0h} \in V_h); \tag{3.62}$$

in (3.59)–(3.62), $\boldsymbol{f}_h$ and $\boldsymbol{u}_{0h}$ are convenient approximations of $\boldsymbol{f}$ and $\boldsymbol{u}_0$, respectively, and $\boldsymbol{g}_h$ is an approximation of $\boldsymbol{g}$ such that $\int_\Gamma \boldsymbol{g}_h \cdot \boldsymbol{n}\, d\Gamma = 0$ (for the construction of $\boldsymbol{g}_h$ see [19], Appendix 3, or [25]).

We have thus reduced the solution of the time dependent Navier–Stokes equations to that of a *nonlinear system of algebraic and ordinary differential equations*. We observe that the incompressibility condition is only approximately satisfied. The time discretization of (3.59)–(3.62) is discussed in Section 3.5.4 below.

3.5.4. *Time discretization of* (3.59)–(3.62) *by alternating direction methods.* We now consider a fully discrete version of the scheme (3.8)–(3.10) discussed in Sec. 3.2.1; it is defined as follows (with Δt and θ as in Sec. 2):

$$\boldsymbol{u}_h^0 = \boldsymbol{u}_{0h}, \tag{3.63}$$

then, for $n \geqslant 0$, *compute (from* $\boldsymbol{u}_h^n$*)* $\{\boldsymbol{u}_h^{n+1/2}, p_h^{n+1/2}\} \in V_h \times H_h^1$, *and* $\boldsymbol{u}_h^{n+1} \in V_h$, *by solving*

$$\int_\Omega \frac{\boldsymbol{u}_h^{n+1/2} - \boldsymbol{u}_h^n}{\Delta t/2} \cdot \boldsymbol{v}_h dx + \theta\nu \int_\Omega \nabla \boldsymbol{u}^{n+1/2} \cdot \nabla \boldsymbol{v}_h dx + \int_\Omega \nabla p^{n+1/2} \cdot \boldsymbol{v}_h dx$$

$$= \int_\Omega \boldsymbol{f}_h^{n+1/2} \cdot \boldsymbol{v}_h dx - (1-\theta)\nu \int_\Omega \nabla \boldsymbol{u}_h^n \cdot \nabla \boldsymbol{v}_h dx - \int_\Omega (\boldsymbol{u}_h^n \cdot \nabla)\boldsymbol{u}_h^n \cdot \boldsymbol{v}_h dx \quad \forall \boldsymbol{v}_h \in V_{0h}, \tag{3.64}$$

$$\int_\Omega \nabla \cdot \boldsymbol{u}_h^{n+1/2} q_h dx = 0 \quad \forall q_h \in H_h^1, \tag{3.65}$$

$$\boldsymbol{u}_h^{n+1/2} \in V_h, \quad p_h^{n+1/2} \in H_h^1, \quad \boldsymbol{u}_h^{n+1/2} = \boldsymbol{g}_h^{n+1/2} \quad \text{on } \Gamma, \tag{3.66}$$

and then

$$\int_\Omega \frac{\boldsymbol{u}_h^{n+1} - \boldsymbol{u}_h^{n+1/2}}{\Delta t/2} \cdot \boldsymbol{v}_h dx + (1-\theta)\nu \int_\Omega \nabla \boldsymbol{u}_h^{n+1} \cdot \nabla \boldsymbol{v}_h dx + \int_\Omega (\boldsymbol{u}_h^{n+1} \cdot \nabla)\boldsymbol{u}_h^{n+1} \cdot \boldsymbol{v}_h dx$$

$$= \int_\Omega \boldsymbol{f}_h^{n+1} \cdot \boldsymbol{v}_h dx - \theta\nu \int_\Omega \nabla \boldsymbol{u}_h^{n+1/2} \cdot \nabla \boldsymbol{v}_h dx - \int_\Omega \nabla p_h^{n+1/2} \cdot \boldsymbol{v}_h dx \quad \forall \boldsymbol{v}_h \in V_{0h}, \tag{3.67}$$

$$\boldsymbol{u}_h^{n+1} \in V_h, \quad \boldsymbol{u}_h^{n+1} = \boldsymbol{g}_h^{n+1} \quad \text{on } \Gamma. \tag{3.68}$$

Obtaining the fully discrete analogue of scheme (3.11)–(3.14) is straightforward. Solving the linear and nonlinear subproblems encountered at each step of (3.63)–(3.68) can be done by the discrete analogues of the methods discussed in Sections 3.3, 3.4; for more details see [19], Chapter 7 and Appendix 3, where the interest of *efficient Poisson solvers* as basic tools appears clearly.

Modifying (3.64)–(3.67), to take into account the augmented nonlinear operator $\tilde{B}$ introduced in Section 3.2.3, Remark 3.2, is quite easy, but as mentioned before, it has no practical influence on the numerical results that we obtained.

3.5.5. *Numerical experiments.* We illustrate the numerical techniques discussed in the above sections by presenting the results of numerical experiments where these techniques have been used to simulate several incompressible viscous flows modeled by the Navier–Stokes equations.

3.5.5.1. *Flow in a channel with a step.* The first numerical experiment that we have done concerns a Navier–Stokes flow in a *channel with a step* at Re = 191; the *characteristics length* used to compute the Reynolds number is the height of the step. Poiseuille's profiles of velocity have been prescribed upstream and quite far downstream. The corresponding *streamlines* are shown in Fig. 3.2; we clearly see in Fig. 3.2 a *thin separation layer* starting slightly below the upper corner of the step, and separating a recirculation region from a region where the flow is quasi potential. The results obtained for this test problem are in very good agreement with those obtained by several authors using different methods (see [19] and [42]).

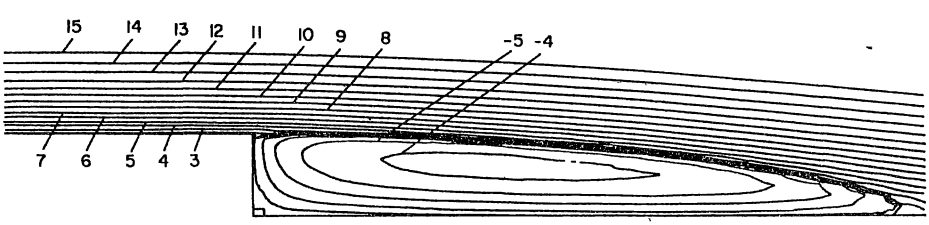


Fig. 3.2 The streamlines shown are those for which the streamfunction assumes values $(n/15)$, for integers n between -5 and $+15$. The stepped (lower) boundary of the channel corresponds to $n = 0$

3.5.5.2. *Flow around and inside a nozzle.* This experiment concerns an unsteady flow around and inside a nozzle at high incidence and at Re = 750 (the characteristic length being the distance between the nozzle walls). We have shown in Fig. 3.3 a part of the finite element triangulation used for the computation and in Figs. 3.4–3.7 the streamlines at $t = 0$, 0.2, 0.4, 0.6, respectively, showing clearly the creation and the motion of eddies, inside and behind the nozzle.

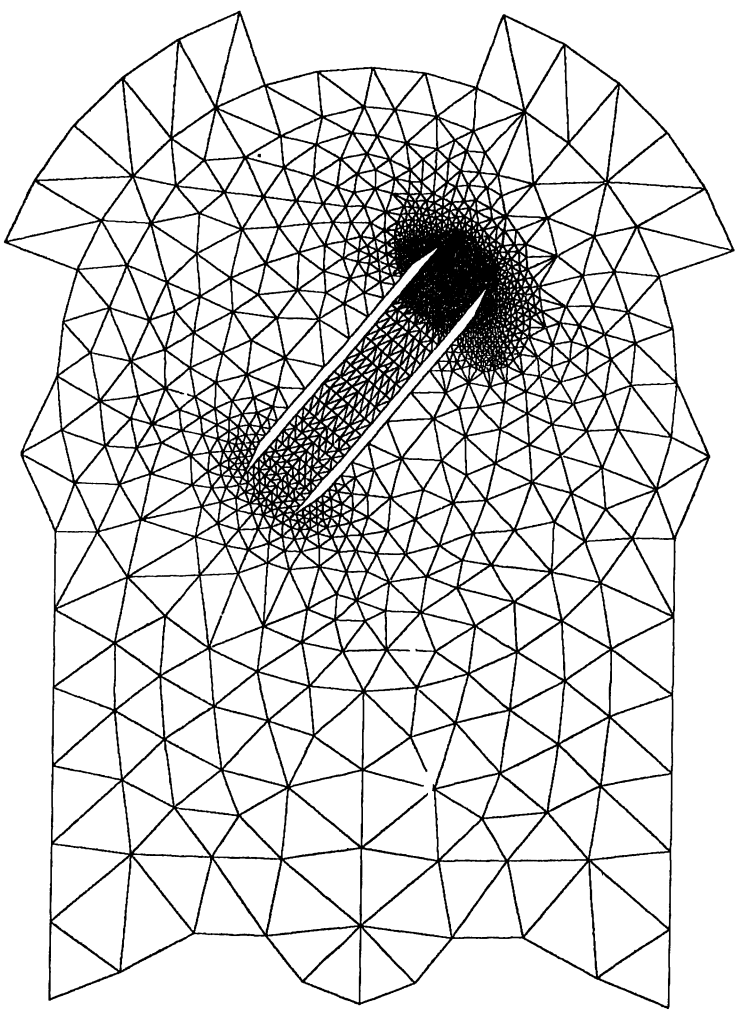


Fig. 3.3

P1/P1 150 P2 LIGNES DE COURANT
CALCUL METRIQUE ENTREE D'AIR MONTEE EN REYNOLDS
INCIDENCE 40.00
MACH INFINI 0.00 REYNOLDS 750.0
CYCLE ITER. 0
PAS DE TEMPS 0.05

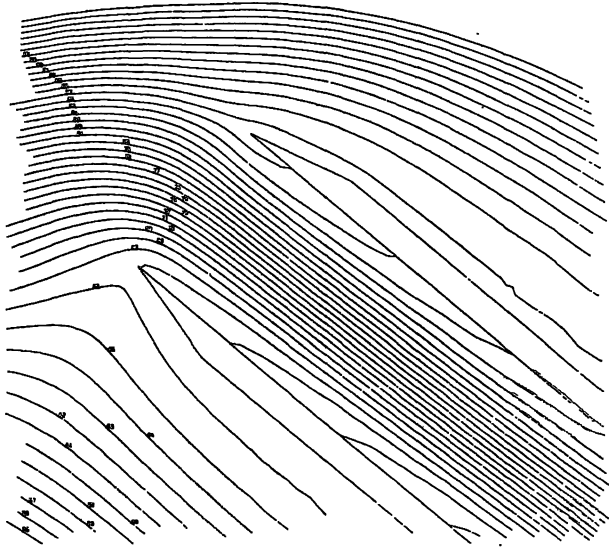


Fig. 3.4

P1/P1 150 P2 LIGNES DE COURANT
CALCUL METRIQUE ENTREE D'AIR MONTEE EN REYNOLDS
INCIDENCE 40.0
MACH INFINI 0.00 REYNOLDS 750.0
CYCLE ITER 40
PAS DE TEMPS 0.05

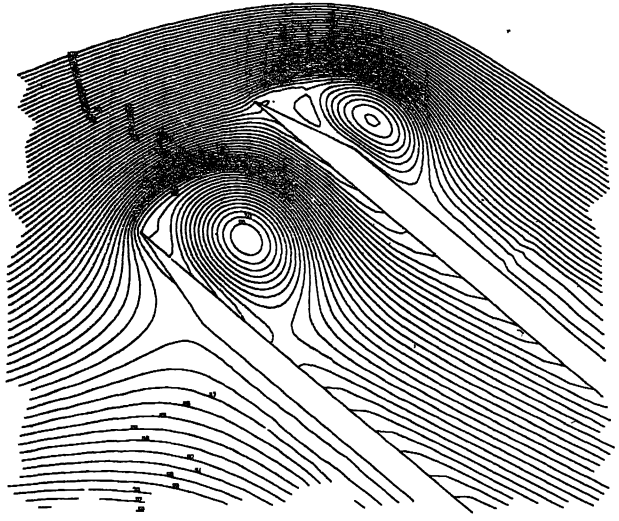


Fig. 3.5

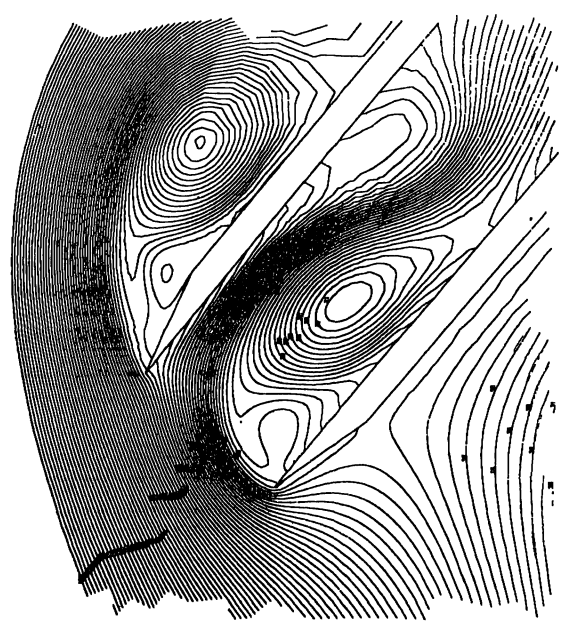

P1/P1 150 P2 LIGNES DE COURANT
CALCUL METRIQUE ENTREE D'AIR MONTEE EN REYNOLDS
INCIDENCE 40.0
MACH INFINI 0.00 REYNOLDS 750.0
CYCLE ITER 80
PAS DE TEMPS 0.05

Fig. 3.6

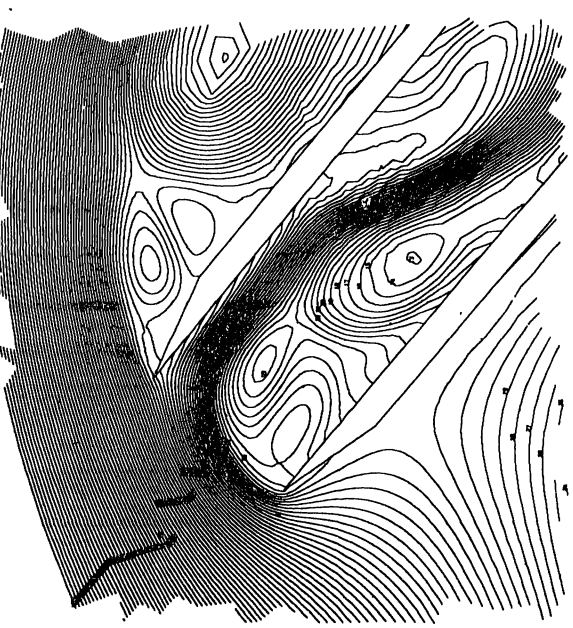

P1/P1 150 P2 LIGNES DE COURANT
CALCUL METRIQUE ENTREE D'AIR MONTEE EN REYNOLDS
INCIDENCE 40.0
MACH INFINI 0.00 REYNOLDS 750.0
CYCLE ITER 120
PAS DE TEMPS 0.75

Fig. 3.7

4. Application to the calculation of potential transonic flows for compressible inviscid fluids

4.1. Generalities. The physical problem. The numerical simulation of transonic potential flows of compressible inviscid fluids is a non-trivial problem since

(1) The equations governing these flows are nonlinear and of changing type (elliptic in the subsonic region of the flow, hyperbolic in the supersonic region);

(2) Shocks may exist corresponding to discontinuities of velocity, pressure and density;

(3) An entropy condition must be included in order to eliminate rarefaction shocks since they correspond to unphysical situations.

We suppose in the following that the fluids to be considered are compressible and inviscid and that their flows are potential and therefore quasi-isentropic, with weak shocks only; in fact, this is only an approximation since usually a flow is no longer potential after a shock (cf. [32]). In the case of flows past bodies we shall suppose that these bodies are sufficiently thin and parallel to the main flow in order not to create a wake in the outflow.

4.2. Mathematical formulation. Let Ω be the region of the flow and Γ its boundary; it follows from [32] that the flow is governed by

$$\nabla \cdot \varrho \boldsymbol{u} = 0 \quad \text{in } \Omega, \tag{4.1}$$

where

$$\varrho = \varrho_0 \left\{ 1 - \frac{(\gamma-1)}{(\gamma+1)} \cdot \frac{|\boldsymbol{u}|^2}{c_*^2} \right\}^{1/(\gamma-1)}, \tag{4.2}$$

$$\boldsymbol{u} = \nabla \varphi. \tag{4.3}$$

In the above relations φ is the *velocity potential*, ϱ is the *density* of the fluid, γ ($= 1.4$ in air) is the *ratio of specific heats* and c_* is the *critical velocity*.

For an airfoil B (see Fig. 4.1) we assume that the flow is *uniform* on Γ_∞ and tangential at Γ_B. We then have

$$\frac{\partial \varphi}{\partial n} = \boldsymbol{u}_\infty \cdot \boldsymbol{u} \quad \text{on } \Gamma_\infty, \qquad \frac{\partial \varphi}{\partial n} = 0 \quad \text{on } \Gamma_B. \tag{4.4}$$

Since only Neumann boundary conditions are involved the potential is determined up to an arbitrary constant. To remedy this we should prescribe the value of φ at some point within $\Omega \cap \Gamma_B$ and, for example,

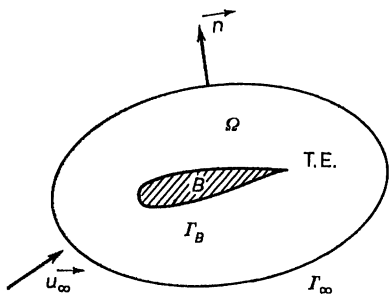


Fig. 4.1

we may conveniently use

$$\varphi = 0 \quad \textit{at the trailing edge (T.E.) of } B. \tag{4.5}$$

In addition to (4.4), (4.5) another condition known as the *Kutta–Joukowsky condition* has to be prescribed; it requires "some" continuity of $\boldsymbol{u}$, even at the corners, and is particularly important for *lifting* bodies.

Since this condition is not specific of transonic flows (it occurs also for compressible inviscid subsonic flows and incompressible inviscid flows), it will not be discussed here (see [7], [8] for the numerical treatment of the Kutta–Joukowsky condition).

Another most important feature of inviscid transonic flows is the existence of *shocks*; across a shock the flow must satisfy the Rankine–Hugoniot conditions

$$(\varrho \boldsymbol{u} \cdot \boldsymbol{n})_+ = (\varrho \boldsymbol{u} \cdot \boldsymbol{n})_- \quad (\textit{where } \boldsymbol{n} \textit{ is normal at the shock line or surface}), \tag{4.6}$$

$$\textit{the tangential component of the velocity is continuous.} \tag{4.7}$$

As regards the *entropy condition*, it can be formulated as follows:

Following the flow, we cannot have a positive variation of velocity through a shock since this would imply a negative variation of entropy which is an unphysical phenomenon. (4.8)

4.3. Least-squares formulation of the continuous problem. We will not consider here the practical implementation of (4.8) (it will be discussed briefly in Sec. 4.4); we consider only the variational formulation of (4.1)–(4.4), (4.6), (4.7) and an associated nonlinear least squares formulation.

4.3.1. *A variational formulation of the continuity equation.* We consider for simplicity the situation in Fig. 4.2 which shows a symmetric flow, subsonic at infinity, around a symmetric airfoil; thus the Kutta–Joukowsky condition is automatically satisfied.

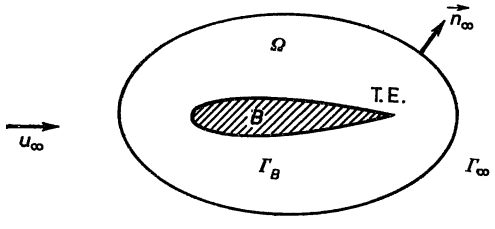


Fig. 4.2

For practical purposes (other approaches are also possible) we imbed the airfoil in a "large" domain; using the notation of Section 4.2, the continuity equation and the boundary conditions are

$$\nabla \cdot \varrho(\varphi)\nabla\varphi = 0 \quad in\ \Omega \tag{4.9}$$

with

$$\varrho(\varphi) = \varrho_0 \left\{1 - \frac{\gamma-1}{\gamma+1}\frac{|\nabla\varphi|^2}{c_*^2}\right\}^{1/(\gamma-1)} \tag{4.10}$$

and

$$\varrho\frac{\partial\varphi}{\partial n} = 0 \quad on\ \Gamma_B, \qquad \varrho\frac{\partial\varphi}{\partial n} = \varrho_\infty \boldsymbol{u}_\infty \cdot \boldsymbol{n}_\infty \quad on\ \Gamma_\infty. \tag{4.11}$$

Define g on the set $\Gamma(=\Gamma_B \cup \Gamma_\infty)$ by

$$g = 0 \quad \text{on } \Gamma_B, \qquad g = \varrho_\infty \boldsymbol{u}_\infty \cdot \boldsymbol{n} \quad \text{on } \Gamma_\infty. \tag{4.12}$$

Clearly, we have

$$\varrho\frac{\partial\varphi}{\partial n} = g \quad \text{and} \quad \int_\Gamma g\, d\Gamma = 0. \tag{4.13}$$

An equivalent variational formulation is

$$\int_\Omega \varrho(\varphi)\,\nabla\varphi\cdot\nabla v\,dx = \int_\Gamma g v\,d\Gamma \quad \forall v\in H^1(\Omega), \qquad \varphi\in W^{1,\infty}(\Omega)/\boldsymbol{R}. \tag{4.14}$$

The space $W^{1,\infty}(\Omega)$ is a natural choice for φ since *physical* flows require (among other properties) a *positive* density ϱ; therefore, in view of (4.10), φ must satisfy

$$|\nabla\varphi| \leqslant \delta < \left(\frac{\gamma+1}{\gamma-1}\right)^{1/2} c_* \quad \text{a.e. in } \Omega.$$

4.3.2. *A least squares formulation of* (4.14). For a genuine transonic flow, problem (4.14) is not equivalent to a standard problem of the calculus of variations (as it would be for purely subsonic flows); to remedy this situation and — in some sense — convexify the problem under consideration, we introduce a nonlinear least squares formulation of (4.14) as follows. Let X be a set of feasible solutions; the least squares problem is then

$$\operatorname*{Min}_{\xi\in X} J(\xi) \tag{4.15}$$

with

$$J(\xi) = \tfrac{1}{2}\int_\Omega |\nabla y(\xi)|^2\,dx \tag{4.16}$$

where, in (4.16), $y(\xi)$ ($=y$) is a solution of

$$\begin{cases} \textit{Find } y\in H^1(\Omega)/\boldsymbol{R} \textit{ such that} \\ \displaystyle\int_\Omega \nabla y\cdot\nabla v\,dx = \int_\Omega \varrho(\xi)\,\nabla\xi\cdot\nabla v\,dx - \int_\Gamma g v\,d\Gamma \quad \forall v\in H^1(\Omega). \end{cases} \tag{4.17}$$

If (4.14) has solutions, these solve (4.15) and give the value zero to the objective function J.

4.4. Finite element approximation. We consider here only two-dimensional problems but the methods described have been applied to three-dimensional problems.

4.4.1. *Finite element approximation of* (4.14). We still consider the nonlifting situation of Section 4.3.1; once the flow region has been embedded in a large domain Ω, we approximate this latter domain by a polygonal

domain Ω_h; with $\mathcal{T}_h$ a standard triangulation of Ω_h, we approximate $H^1(\Omega)$ (and in fact $W^{1,p}(\Omega)$, $\forall p \geqslant 1$) by

$$H_h^1 = \{v_h |\ v_h \in C^0(\bar{\Omega}_h), v_h|_T \in P_1\ \forall T \in \mathcal{T}_h\} \tag{4.18}$$

where P_1 is the space of polynomials in two variables of degree $\leqslant 1$. We prescribe the value for the potential at T.E.; this leads to

$$V_h = \{v_h \in H_h^1,\ v_h(\text{T.E.}) = 0\}. \tag{4.19}$$

Clearly,

$$\dim H_h^1 = 1 + \dim V_h = \text{number of vertices of } \mathcal{T}_h. \tag{4.20}$$

We then approximate the variational equation (4.14) (dropping h in Ω_h and Γ_h) by

$$\begin{cases} \textit{Find } \varphi_h \in V_h \textit{ such that} \\ \displaystyle\int_\Omega \varrho(\varphi_h)\nabla\varphi_h \cdot \nabla v_h dx = \int_\Gamma g_h v_h d\Gamma \quad \forall v_h \in V_h \end{cases} \tag{4.21}$$

where g_h is an approximation of the function g of (4.13). Let $\mathcal{B}_h = \{w_i\}_{i=1}^{N_h}$ be a vector basis of V_h. Then (4.21) is equivalent to the nonlinear finite-dimensional system

$$\begin{cases} \varphi_h = \displaystyle\sum_{j=1}^{N_h} \varphi_j w_j, \\ \displaystyle\int_\Omega \varrho(\varphi_h)\nabla\varphi_h \cdot \nabla w_i dx = \int_\Gamma g_h w_i d\Gamma \quad \forall i = 1, \ldots, N_h. \end{cases} \tag{4.22}$$

With the above choice for H_h^1 and V_h, there is no problem of numerical integration since, in (4.21) and (4.22), $\nabla\varphi_h$, ∇v_h (and therefore $\varrho(\varphi_h)$) are piecewise constant.

4.4.2. *Numerical implementation of the entropy condition.* The numerical implementation of the *entropy condition* (4.8), in order to eliminate rarefaction shocks, is a non-trivial matter. Without going into details, we should mention that methods founded on the *upwinding of the density* have been implemented, producing rather good numerical results (see [9], [27] and also [19], Chapter 7, for technical details and further references).

The method for upwinding the density discussed in [9] leads to the following equation:

$$\begin{cases} \textit{Find } \varphi_h \in V_h \textit{ such that} \\ \int_\Omega \varrho(\varphi_h)\, \nabla\varphi_h \cdot \nabla v_h\, dx + \int_\Omega R_h(\varphi_h) v_h\, dx = \int_\Gamma g_h v_h\, d\Gamma \quad \forall v_h \in V_h, \end{cases} \tag{4.23}$$

in which R_h can be viewed as an *artificial viscosity operator* (see [9] for a full description of R_h).

The solution of (4.23) by nonlinear least squares methods is achieved by the following variant of (4.15):

$$\operatorname*{Min}_{\xi_h \in V_h} J_h(\xi_h) \tag{4.24}$$

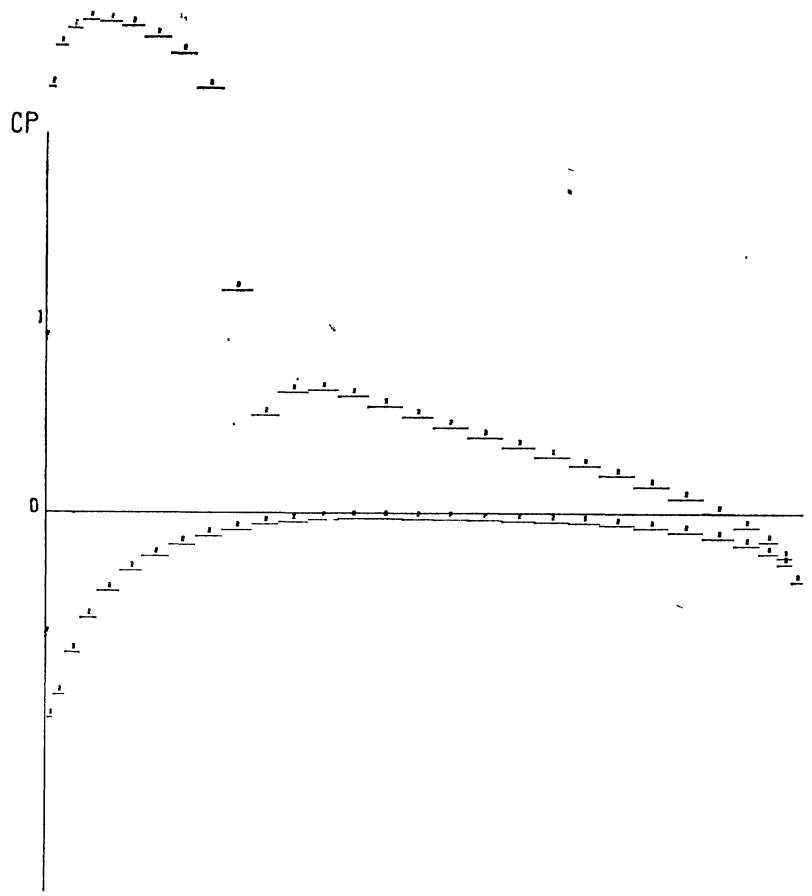


Fig. 4.3. $\alpha = 6°$, $M_\infty = .6$

with

$$J_h(\xi_h) = \tfrac{1}{2}\int_\Omega |\nabla y_h|^2 dx \tag{4.25}$$

where, in (4.25), y_h is a solution of

$$\begin{cases} y_h \in V_h, \\ \int_\Omega \nabla y_h \cdot \nabla v_h dx = \int_\Omega \varrho(\xi_h)\nabla \xi_h \cdot \nabla v_h dx + \int_\Omega R_h(\xi_h) v_h dx - \int_\Gamma g_h v_h d\Gamma. \end{cases} \tag{4.26}$$

The solution of (4.24)–(4.26) by a conjugate gradient algorithm (fairly close to algorithm (2.14)–(2.22)) is discussed in [19], Chapter 7, and [9].

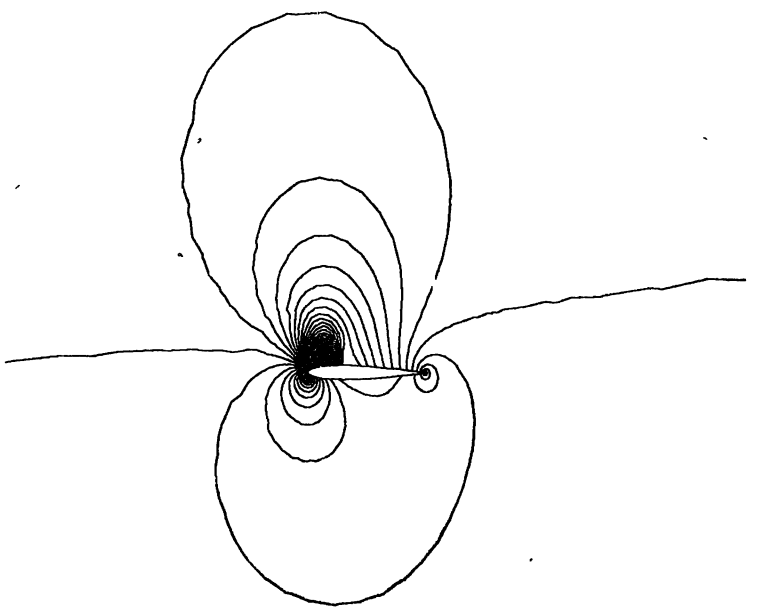


Fig. 4.4. $\alpha = 6°$, $M_\infty = .6$

4.5. Numerical experiments.

4.5.1. *Flows around a* NACA 0012 *airfoil.* Figures 4.3–4.5 show the *pressure distribution* and *isomach lines* for flows around a NACA 0012 airfoil at various M_∞ and angles of attack.

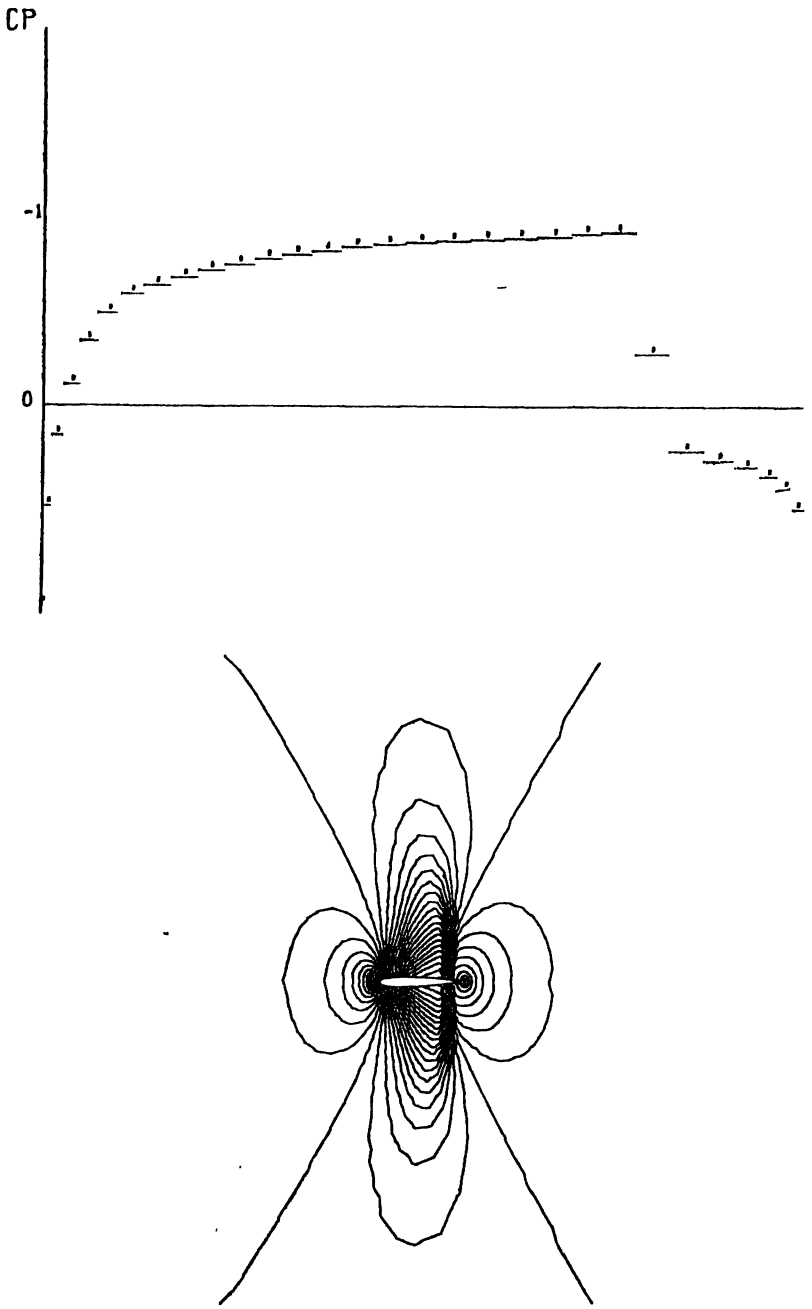


Fig. 4.5. $\alpha = 0°$, $M_\infty = .85$

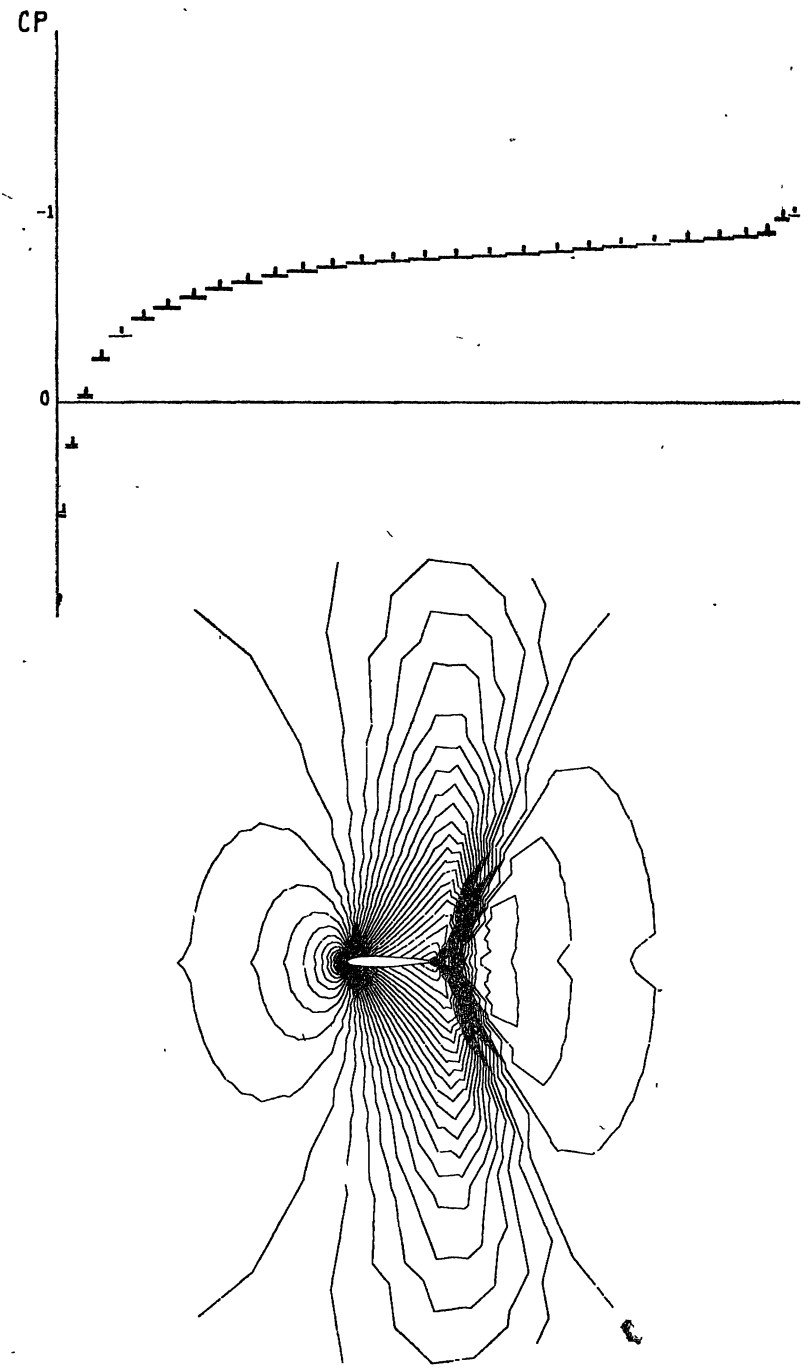


Fig. 4.6. $\alpha = 0°$, $M_\infty = .90$

These results are in good agreement with those obtained by other authors by quite different methods (mostly finite difference methods); see [19] and [9] for further references.

4.5.2. *Flows around a* NACA 64006 *airfoil.* We have shown in Figs. 4.6, 4.7 the Mach distribution corresponding to flows around a NACA 64006 airfoil for $M_\infty = .89$ and the angle of attack $\alpha = 0°$. The flow on Fig.

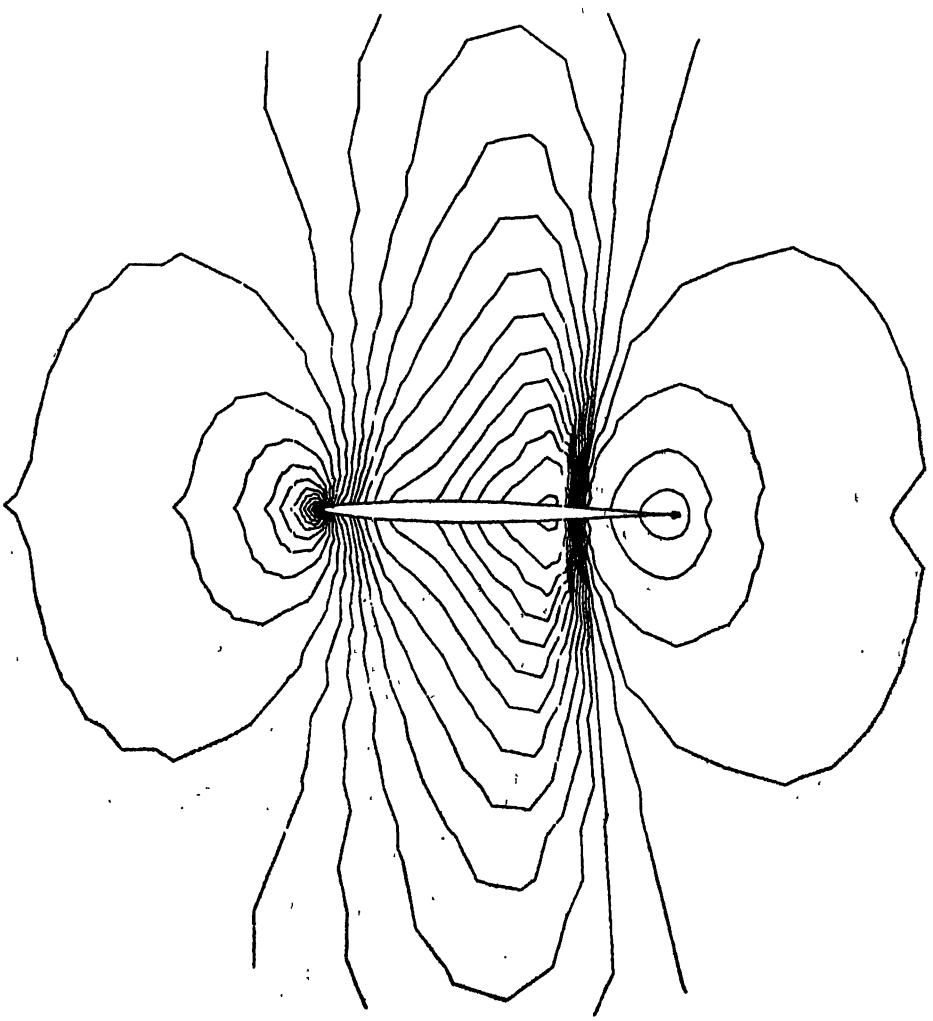


Fig. 4.6

4.6 (resp. 4.7) is symmetrical (resp. non-symmetrical); thus we have (at least) three solutions to the same problem (the third one is obtained from the one of Fig. 4.7 by symmetry with respect to the symmetry axis of the airfoil); these three solutions satisfy the continuity equation, the Rankine–Hugoniot, Kutta–Joukowsky and entropy conditions. Actually, the symmetrical one seems to be instable with regards to small nonsymmetric perturbations; for more details, see e.g. [49].

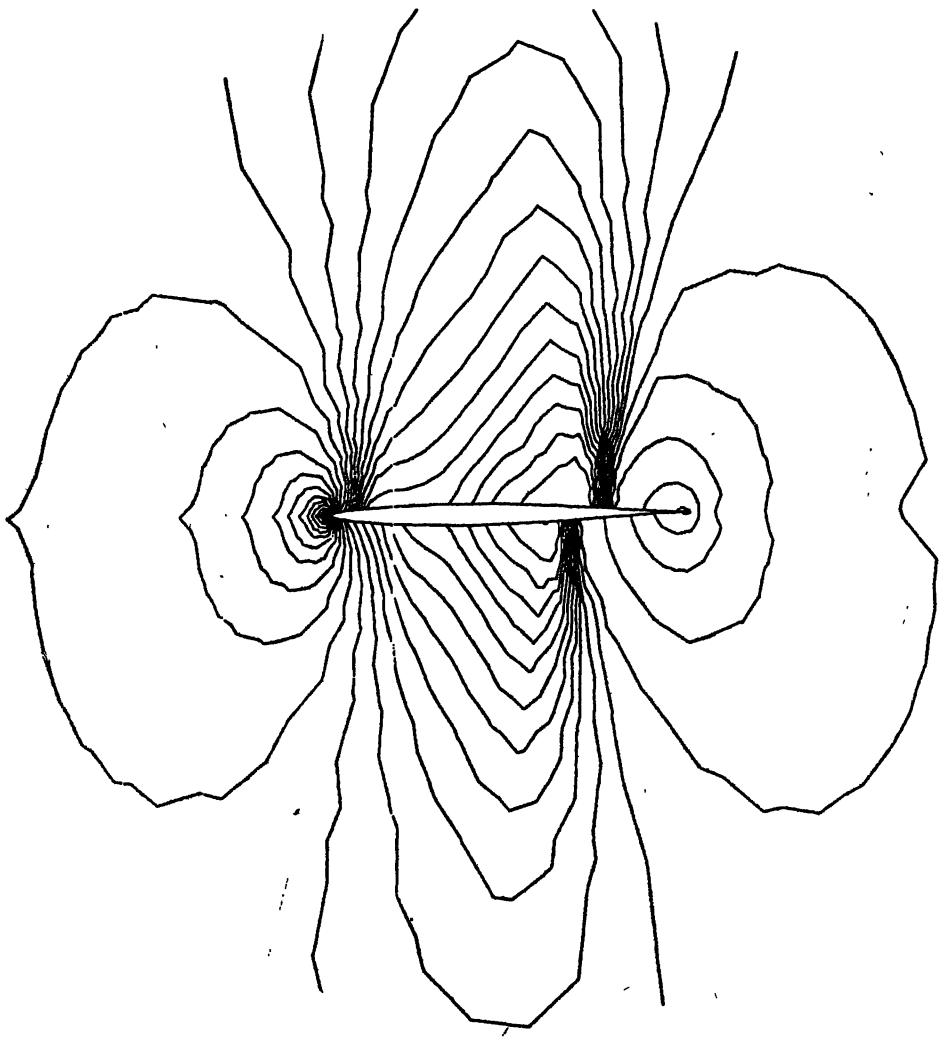


Fig. 4.7

5. Decomposition methods for variational problems by augmented Lagrangians. An application in finite elasticity

5.1. Generalities. The main goal of this section is to give a brief account of solution methods for variational problems when some *decomposition* property holds; introducing a convenient *augmented Lagrangian*, we obtain solution methods, taking advantage of the special structure of the problem under consideration; these methods are described and discussed in Section 5.2. An application in Finite Elasticity is considered in Section 5.3.

For more details and further references, see [20] and also [19], Chapter 6.

5.2. Decomposition of variational problems. Associated algorithms. We follow [5] and [24], Appendix 1.

5.2.1. *A family of variational problems.* In the sequel we consider real Hilbert spaces only; let V and H be two such vector spaces, equipped with the norms and scalar products $\|\cdot\|$, $((\cdot,\cdot))$ and $|\cdot|$, $(\cdot,\cdot)$, respectively. Let $B \in \mathscr{L}(V, H)$ and F, G be two functionals *convex*, *proper*, *l.s.c.*, from H and V to $\boldsymbol{R} \cup \{+\infty\}$, respectively. We suppose that

$$\mathrm{dom}(G) \cap \mathrm{dom}(F \circ B) \neq \emptyset, \tag{5.1}$$

where

$$\mathrm{dom}(G) = \{v \mid v \in V, -\infty < G(v) < +\infty\}$$

and a similar definition for $\mathrm{dom}(F \circ B)$. We associate with V, H, B, F, G the following *minimization problem*:

$$\begin{cases} \textit{Find } u \in V \textit{ such that} \\ J(u) \leqslant J(v) \quad \forall v \in V \end{cases} \tag{P}$$

where $J\colon V \to \bar{\boldsymbol{R}}$ is defined by

$$J(v) = F(Bv) + G(v). \tag{5.2}$$

Since $J(\cdot)$ has a special structure, it is natural to look for methods taking advantage of this fact.

Remark 5.1. Most of the following considerations can be applied to the solution of variational problems such as

$$f \in B' A_1(Bu) + A_2(u), \tag{5.3}$$

where $f \in V'$ (the *dual* space of V) and where A_1, A_2 are *monotone operators* from H to H' (dual space of H) and from V to V', respectively. In general, $A = B' \circ A_1 \circ B + A_2$ is not the gradient (or subgradient) of a functional (B' is the transposed operator of B).

5.2.2. *A decomposition principle.* We define a set $W \subset V \times H$ by

$$W = \{\{v, q\} \in V \times H, Bv - q = 0\}. \tag{5.4}$$

Problem (P) is equivalent to

$$\begin{cases} \textit{Find } \{u, p\} \in W \textit{ such that} \\ j(u, p) \leqslant j(v, q) \quad \forall \{v, q\} \in W, \end{cases} \tag{Π}$$

with

$$j(v, q) = F(q) + G(v). \tag{5.5}$$

Remark 5.2. Problems (P) and (Π) are equivalent, but (Π) has, in some sense, a simpler structure than (P), despite the fact that it contains an extra variable. This is due to the fact that the *linear* relation

$$Bv - q = 0 \tag{5.6}$$

can be efficiently treated, using simultaneously — via an appropriate *augmented Lagrangian* — the *penalty* and *Lagrange multiplier* methods of solution.

5.2.3. *An augmented Lagrangian associated to* (Π). Let $r > 0$; we define $\mathscr{L}_r$: $V \times H \times H \to \bar{\boldsymbol{R}}$ by

$$\mathscr{L}_r(v, q, \mu) = F(q) + G(v) + \frac{r}{2} |Bv - q|^2 + (\mu, Bv - q). \tag{5.7}$$

We easily prove that if $\{\{u, p\}, \lambda\}$ is a saddle point of $\mathscr{L}_r$ over $(V \times H) \times H$, then $\{u, p\}$ is a solution of (Π), i.e. u is a solution of (P) (with $p = Bu$).

5.2.4. *A first algorithm for solving* (P). To solve (P) and (Π) we look for saddle points of $\mathscr{L}_r$ using duality algorithms like those discussed in [24]; such an algorithm is defined as follows:

$$\lambda^0 \in H \textit{ given}, \tag{5.8}$$

then for $n \geqslant 0$, λ^n *being known, we compute* u^n, p^n *and* λ^{n+1} *by*

$$\begin{cases} \text{Find } \{u^n, p^n\} \in V \times H \text{ such that} \\ \mathscr{L}_r(u^n, p^n, \lambda^n) \leqslant \mathscr{L}_r(v, q, \lambda^n) \quad \forall \{v, q\} \in V \times H, \end{cases} \tag{5.9}$$

$$\lambda^{n+1} = \lambda^n + \varrho(Bu^n - p^n). \tag{5.10}$$

We have proved in [20], Chapter 3, that if F, B, G satisfy quite reasonable hypotheses and if

$$0 < \varrho < 2r, \tag{5.11}$$

then we have

$$\lim_{n \to +\infty} u^n = u \qquad \text{strongly in } V, \tag{5.12}$$

$$\lim_{n \to +\infty} p^n = p (= Bu) \quad \text{strongly in } H, \tag{5.13}$$

$$\lim_{n \to +\infty} \lambda^n = \lambda \qquad \text{weakly in } H, \tag{5.14}$$

where u is the solution of (P), and where λ is such that $\{\{u, p\}, \lambda\}$ is a saddle point of $\mathscr{L}_r$ on $(V \times H) \times H$.

Remark 5.3. To solve (5.9), we can use *block-relaxation* algorithms like those discussed in [12] (see [20] for more details); if we use these relaxation methods to solve (5.9) and limit ourselves to *only one* inner iteration, we obtain the algorithm described in Sec. 5.2.5.

5.2.5. *A second algorithm for solving* (P). It is defined by

$$\{u^{-1}, \lambda^0\} \in V \times H \quad \text{given}, \tag{5.15}$$

then, for $n \geqslant 0$, u^{n-1} *and* λ^n *being known, we compute* p^n, u^n, λ^{n+1} *by*

$$\mathscr{L}_r(u^{n-1}, p^n, \lambda^n) \leqslant \mathscr{L}_r(u^{n-1}, q, \lambda^n) \ \forall q \in H, \qquad p^n \in H, \tag{5.16}$$

$$\mathscr{L}_r(u^n, p^n, \lambda^n) \leqslant \mathscr{L}_r(v, p^n, \lambda^n) \ \forall v \in V, \qquad u^n \in V, \tag{5.17}$$

$$\lambda^{n+1} = \lambda^n + \varrho(Bu^n - p^n). \tag{5.18}$$

Remark 5.4. Several variants of (5.15)–(5.18) can be derived; we may for example

(i) interchange the role of q and v (see also Remark 5.5),

(ii) update λ^n between steps (5.16), (5.17); doing so we obtain the following variant of (5.15)–(5.18) (due to Gabay [17]):

$$\{u^{-1}, \lambda^0\} \quad \textit{given in } V \times H, \tag{5.19}$$

then for $n \geq 0$, u^{n-1} *and* λ^n *being given, we compute* p^n, $\lambda^{n+1/2}$, u^n, λ^{n+1} *by*

$$\mathscr{L}_r(u^{n-1}, p^n, \lambda^n) \leq \mathscr{L}_r(u^{n-1}, q, \lambda^n) \ \forall q \in H, \quad p^n \in H, \tag{5.20}$$

$$\lambda^{n+1/2} = \lambda^n + (Bu^{n-1} - p^n), \tag{5.21}$$

$$\mathscr{L}_r(u^n, p^n, \lambda^{n+1/2}) \leq \mathscr{L}_r(v, p^n, \lambda^{n+1/2}) \ \forall v \in V, \quad u^n \in V, \tag{5.22}$$

$$\lambda^{n+1} = \lambda^{n+1/2} + \varrho(Bu^n - p^n); \tag{5.23}$$

q and v play a more symmetrical role in (5.19)–(5.23) than in (5.15)–(5.18).

Remark 5.5. If one uses (5.15)–(5.18), it is suggested to solve in the second step the problem with the best ellipticity properties (cf. [20], Chapter 3, for the justification of such a choice). ■

As regards the convergence of (5.15)–(5.18), one proves in [20], Chapter 3, that the convergence results (5.12)–(5.14) hold if

$$0 < \varrho < \frac{1+\sqrt{5}}{2} r. \tag{5.24}$$

5.2.6. *Comments on the choice of* ϱ *and* r. For a given r, it follows from various numerical experiments we have done that the optimal choice for ϱ is close to r. The choice of r is more complicated; theoretically, the speed of convergence of (5.8)–(5.10) increases with r, but the conditioning of problem (5.9) deteriorates as r increase.

If one uses algorithms (5.15)–(5.19) and (5.19)–(5.23) with $\varrho = r$, the optimal choice for r is again a problem difficult to analyse.

5.2.7. *Relations with alternating direction methods.* Algorithms (5.15)–(5.18) and (5.19)–(5.23) are closely related to alternating direction methods, as shown in [20], Chapters 8 and 9; for the convergence properties of these alternating direction methods and relations with the numerical integration of time dependent problems see [17], [37].

5.3. Application in finite elasticity. We apply the decomposition methods of Section 5.2 to the numerical solution of nonlinear problems in finite elasticity dealing with *incompressible* materials of the Mooney–Rivlin type (we follow here [20], Chapter 8, and [22]).

5.3.1. *Formulation of the elasticity problem.* A fundamental problem in finite elasticity is the calculation of the deformations and displacements of a solid body made of an homogeneous, isotropic, *hyperelastic* and *incompressible* material submitted to *volume forces* $\varrho_0 f$ (ϱ_0 is the density of the material) and *superficial forces* $\boldsymbol{S}_0$. Using a *Lagrangian formulation*, the *functional of energy* associated with a displacement field $\boldsymbol{v}$ is given by

$$\Pi(\boldsymbol{v}) = \int_{\Omega} \varrho_0 (\sigma(\boldsymbol{v}) - \boldsymbol{f}_0 \boldsymbol{v})\, dx - \int_{\partial\Omega_2} \boldsymbol{S}_0 \cdot \boldsymbol{v}\, d\Gamma, \tag{5.25}$$

where Ω is the domain in $\boldsymbol{R}^N$ corresponding to the *reference configuration*; $\partial\Omega$ ($= \partial\Omega_1 \cap \partial\Omega_2$) is the boundary of Ω. The body being *fixed* along $\partial\Omega_1$, we have denoted by $\sigma(\boldsymbol{v})$ the stored energy functional (per unit mass). For a Mooney–Rivlin material we have

$$\sigma(\boldsymbol{v}) = E_1(I_1 - 2) \quad \text{if} \quad N = 2, \tag{5.26}$$

$$\sigma(\boldsymbol{v}) = E_1(I_1 - 3) + E_2(I_2 - 3) \quad \text{if} \quad N = 3, \tag{5.27}$$

where I_i is the ith invariant of the tensor $\boldsymbol{F}\boldsymbol{F}^t$, where

$$\boldsymbol{F} = \boldsymbol{J} + \nabla \boldsymbol{v}, \tag{5.28}$$

and where E_1, E_2 are positive coefficients, material dependent.

The displacement $\boldsymbol{v}$ satisfies the *incompressibility condition*

$$\det \boldsymbol{F}(\boldsymbol{v}) = 1 \quad \textit{a.e. on } \Omega. \tag{5.29}$$

Remark 5.6. We have supposed in (5.25) that f and S_0 are independent of $\boldsymbol{v}$ (*dead load hypothesis*); actually we refer to [22], [33], [34] where the algorithms to be described below are generalized to problems for which the above hypothesis is not satisfied. ■

It is reasonable to suppose that the displacements $\boldsymbol{u}$ corresponding to the *stable equilibria* satisfy

$$\textit{u is a local minimizer over } K \textit{ of the functional } \Pi, \tag{5.30}$$

where, for a Mooney–Rivlin incompressible material, we have

$$K = \{\boldsymbol{v} \in (H^1(\Omega))^N \mid \boldsymbol{v} = \boldsymbol{0} \text{ on } \partial\Omega_1, \det \boldsymbol{F}(\boldsymbol{v}) = 1 \text{ a.e., } \boldsymbol{F}^{-1}(\boldsymbol{v}) \in (L^2(\Omega))^{N \times N}\} \tag{5.31}$$

and where Π is defined by (5.25)–(5.27).

The existence of solutions for (5.30), (5.31) is proved in [2].

We can give also a formulation founded on the following augmented Lagrangian (associated with the linear relation (5.28)), with $r > 0$:

$$\mathscr{L}_r(\boldsymbol{v}, \boldsymbol{G}, \boldsymbol{\mu}) = \Pi(\boldsymbol{v}) + \frac{r}{2}\|\nabla \boldsymbol{v} + \boldsymbol{J} - \boldsymbol{G}\|_{L_2}^2 - \int_{\Omega} \boldsymbol{\mu} \cdot (\nabla \boldsymbol{v} + \boldsymbol{J} - \boldsymbol{G})\, dx. \quad (5.32)$$

This leads to the following formulation of the elastostatic problem:

$$\textit{Find } \{\boldsymbol{u}, \boldsymbol{F}, \boldsymbol{\lambda}\} \in W = X \times Y \times \left(L^2(\Omega)\right)^{N \times N}, \textit{ a stationary point of } \mathscr{L}_r \textit{ over } W, \quad (5.33)$$

where

$$Y = \left\{\boldsymbol{G} \mid \boldsymbol{G} \in \left(L^2(\Omega)\right)^{N \times N},\ \boldsymbol{G}^{-1} \in \left(L^2(\Omega)\right)^{N \times N},\ \det \boldsymbol{G} = 1 \text{ a.e.}\right\}.$$

The relations between formulations (5.30), (5.31) and (5.33) are discussed in [22], [33], [34] (together with other formulations).

5.3.2. *Solution of problem* (5.30), (5.31).

5.3.2.1. *A first algorithm.* It corresponds to (5.8)–(5.10) of Section 5.2 and is defined by:

$$\boldsymbol{\lambda}^0 \textit{ is given in } \left(L^2(\Omega)\right)^{N \times N}, \quad (5.34)$$

then for $n \geqslant 0$, $\boldsymbol{\lambda}^n$ *being known, we obtain* $\boldsymbol{u}^n$, $\boldsymbol{F}^n$ *and* $\boldsymbol{\lambda}^{n+1}$ *from the solution of*

$$\begin{cases} \mathscr{L}_r(\boldsymbol{u}^n, \boldsymbol{F}^n, \boldsymbol{\lambda}^n) \leqslant \mathscr{L}_r(\boldsymbol{v}, \boldsymbol{G}, \boldsymbol{\lambda}^n) \quad \forall \{\boldsymbol{v}, \boldsymbol{G}\} \in X \times Y, \\ \{\boldsymbol{u}_n, \boldsymbol{F}_n\} \in X \times Y, \end{cases} \quad (5.35)$$

$$\boldsymbol{\lambda}^{n+1} = \boldsymbol{\lambda}^n - \varrho(\nabla \boldsymbol{u}^n + \boldsymbol{J} - \boldsymbol{F}^n), \ \varrho > 0. \quad (5.36)$$

Remark 5.7. Problem (5.35) is equivalent to the nonlinear system

$$\mathscr{L}_r(\boldsymbol{u}^n, \boldsymbol{F}^n, \boldsymbol{\lambda}^n) \leqslant \mathscr{L}_r(\boldsymbol{u}^n, \boldsymbol{G}, \boldsymbol{\lambda}^n) \quad \forall \boldsymbol{G} \in Y,\ \boldsymbol{F}^n \in Y, \quad (5.37)$$

$$\partial_{\boldsymbol{v}} \mathscr{L}_r(\boldsymbol{u}^n, \boldsymbol{F}^n, \boldsymbol{\lambda}^n) \cdot \boldsymbol{v} = 0 \quad \forall \boldsymbol{v} \in X,\ \boldsymbol{u}^n \in X, \quad (5.38)$$

whose solution using block relaxation methods leads to the algorithm thereafter.

5.3.2.2. *A second algorithm.* It corresponds to (5.15)–(5.18) of Section 5.2 and is defined by:

$$\boldsymbol{u}^{-1} \text{ is given in } X, \ \boldsymbol{\lambda}^0 \text{ is given in } \left(L^2(\Omega)\right)^{N\times N}, \tag{5.39}$$

then for $n \geqslant 0$, $\boldsymbol{u}^{n-1}$ *and* $\boldsymbol{\lambda}^n$ *being known, we obtain* $\boldsymbol{F}^n$, $\boldsymbol{u}^n$ *and* $\boldsymbol{\lambda}^{n+1}$ *by the solution of*

$$\mathscr{L}_r(\boldsymbol{u}^{n-1}, \boldsymbol{F}^n, \boldsymbol{\lambda}^n) \leqslant \mathscr{L}_r(\boldsymbol{u}^{n-1}, \boldsymbol{G}, \boldsymbol{\lambda}^n) \quad \forall \boldsymbol{G} \in Y, \boldsymbol{F}^n \in Y, \tag{5.40}$$

$$\partial_v \mathscr{L}_r(\boldsymbol{u}^n, \boldsymbol{F}^n, \boldsymbol{\lambda}^n)\cdot \boldsymbol{v} = 0 \quad \forall \boldsymbol{v} \in X, \boldsymbol{u}^n \in X, \tag{5.41}$$

$$\boldsymbol{\lambda}^{n+1} = \boldsymbol{\lambda}^n - \varrho(\nabla \boldsymbol{u}^n + \boldsymbol{J} - \boldsymbol{F}^n), \quad \varrho > 0. \tag{5.42}$$

Problem (5.41), which is equivalent to

$$\begin{cases} \textit{Find } \boldsymbol{u}^n \in X \textit{ such that} \\ \mathscr{L}_r(\boldsymbol{u}^n, \boldsymbol{F}^n, \boldsymbol{\lambda}^n) \leqslant \mathscr{L}_r(\boldsymbol{v}, \boldsymbol{F}^n, \boldsymbol{\lambda}^n) \quad \forall \boldsymbol{v} \in X, \end{cases} \tag{5.43}$$

is in fact an *unconstrained* minimization problem whose solution is rather easy, particularly if r is sufficiently large; if $N = 2$ the functional in (5.43) is *quadratic*, and solving (5.41), (5.43) is equivalent to solving a *linear problem* for a second order partial differential operator (close to the *linear elasticity operator*), *independent* of n, and whose finite-dimensional variants are linear systems for symmetric, positive-definite matrices, independent of n. Problem (5.40) is more complicated (apparently, at least); if $N = 2$, (5.40) leads to

$$\begin{cases} \textit{Find } \boldsymbol{F} \in \left(L^2(\Omega)\right)^4 \textit{ such that } F_{11}F_{22} - F_{12}F_{21} = 1 \textit{ a.e.} \\ \textit{and minimizing the functional} \\ \boldsymbol{G} \to \int\limits_{\Omega} [rG_{ij}^2 - 2(r(u_{i,j} + \delta_{ij}) - \lambda_{ij})G_{ij}]\,dx \\ \textit{on the set of the } \boldsymbol{G} \in \left(L^2(\Omega)\right)^4 \textit{ such that } G_{11}G_{22} - G_{12}G_{21} = 1 \textit{ a.e.}; \end{cases} \tag{5.44}$$

in (5.44) n is omitted, and the summation convention of repeated indices is used, $u_{i,j} = \partial u_i/\partial x_j$ and δ_{ij} is the Kronecker symbol. Since (5.44) does not contain any derivative of $\boldsymbol{G}$ and $\boldsymbol{F}$, we can solve this latter problem *pointwise* as shown in [20], Chapter 8, [22], [33], [34], using a *diagonalization* of the constraint $F_{11}F_{22} - F_{12}F_{21} = 1$, via the transformation

$$\begin{aligned} b_1 &= (F_{11} + F_{22})/\sqrt{2}, \quad & b_2 &= (F_{11} - F_{22})/\sqrt{2}, \\ b_3 &= (F_{12} + F_{21})/\sqrt{2}, \quad & b_4 &= (F_{12} - F_{21})/\sqrt{2}. \end{aligned} \tag{5.45}$$

5.3.3. *A numerical experiment.* We suppose $N = 2$; we approximate (5.30), (5.31) (and (5.33)) using a *finite element method.* We have used *rectangular finite elements* $K \in \mathcal{Q}_h$, where $\mathcal{Q}_h$ is a *quadrangulation* of Ω. We approximate then the displacement $\boldsymbol{v}$ by $\boldsymbol{v}_h \in C^0(\bar{\Omega}) \times C^0(\bar{\Omega})$, such that

$$\boldsymbol{v}_h|_K \in Q_1 \times Q_1 \quad \forall K \in \mathcal{Q}_h, \tag{5.46}$$

where

$$Q_1 = \{q \mid q(x_1, x_2) = a_{00} + a_{10}x_1 + a_{01}x_2 + a_{11}x_1x_2\}; \tag{5.47}$$

we require the incompressibility condition (5.29) at the center of each elementary rectangle $K \in \mathcal{Q}_h$ (which is equivalent to require it in an averaging sense). The convergence of the approximate solutions, as $h \to 0$ is a very difficult problem, discussed e.g. in [33], [34].

In the following numerical experiments Ω is a two-dimensional bar containing a (non-propagating) crack. We have shown in Fig. 5.1 the right

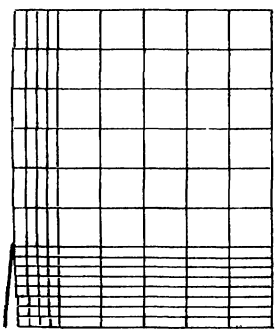


Fig. 5.1

Fig. 5.2

half part of the bar, the crack and the quadrangulation $\mathcal{Q}_h$ (or rather the right half part of it). We suppose that in (5.25), (5.26) we have $\varrho_0 = 1$, $E_1 = 1$, $\partial\Omega_2 = \partial\Omega$ and that S_0 corresponds to horizontal *stretching forces* whose density modulus is 2, these forces being applied at the extremities

of the bar. Under the action of these forces we have a stretching phenomenon and we have shown in Figure 5.2 the equilibrium configuration, computed by a discretized variant of algorithm (5.34)–(5.36), with $\varrho = r = 10$; the convergence is obtained in 20 iterations corresponding to a running time of 5 seconds on CDC 6400. We should observe with interest the behavior of the crack.

One can find in [22], [23], [33], [34] numerical experiments for other two-dimensional problems and also for axisymmetric and three-dimensional problems.

6. Conclusion

We have shown in this paper that variational methods can be applied to the numerical solution of large classes of nonlinear problems governed by partial differential equations, even for situations which are not equivalent, in a strict sense, to a problem of the calculus of variations. For a more complete discussion concerning these methods and their numerical applications, see, e.g. [19] and [20].

References

[1] Amann H., Fixed point equations and nonlinear eigenvalue problems in ordered Banach spaces, *SIAM Review* **18** (1976), 4, pp. 620–709.

[2] Ball J., Convexity conditions and existence theorems in nonlinear elasticity, *Arch. Rat. Mech. Anal.* **63** (1977), pp. 307–403.

[3] Benque J. P., Ibler B., Keramsi A., and Labadie G., A finite element method for Navier–Stokes equations, in: *Proceedings of the Third International Conference on Finite Elements in Flows Problems, Banff, Alberta, Canada, 10–13 June 1980,* D. H. Norrie ed., Vol. 1, pp. 110–120.

[4] Berger M. S., On the Von Karman equations and the buckling of a thin elastic plate – I, the clamped case, *Comm. Pure Appl. Math.* **20** (1967), pp. 687–719.

[5] Bourgat J. F., Glowinski R., et Le Tallec P., Application des méthodes de lagrangien augmenté à la résolution de problèmes d'élasticité non linéaire finie, Chapter 8 of *Méthodes de lagrangien augmenté. Application à la résolution numérique de problèmes aux limites,* M. Fortin, R. Glowinski eds., Dunod–Bordas, Paris, 1982, pp. 241–278.

[6] Brent R., *Algorithms for minimization without derivatives,* Prentice-Hall, Englewood Cliffs, N. J., 1973.

[7] Bristeau M. O., Glowinski R., Periaux J., Perrier P., and Pironneau O., On the numerical solution of nonlinear problems in fluid dynamics by least squares and finite element methods. (I) Least squares formulations and conjugate gradient solution of the continuous problems, *Comp. Meth. Appl. Mech. Eng.* **17/18** (1979), pp. 619–657.

[8] Bristeau M. O., Glowinski R., Periaux J., Perrier P., Pironneau O., and Poirier G., Application of optimal control and finite element methods to the calculation of transonic flows and incompressible viscous flows, in: *Numerical Methods in Applied Fluid Dynamics*, B. Hunt ed., Academic Press, London, 1980, pp. 203–312.

[9] Bristeau M. O., Glowinski R., Periaux J., Perrier P., Pironneau O., and Poirier G., Transonic flow simulations by finite elements and least squares methods, in: *Finite Elements in Fluids*, Vol. **4**, R. H. Gallagher, D. H. Norrie, J. T. Oden, O. C. Zienkiewicz eds., J. Wiley, Chichester, 1982, pp. 453–482.

[10] Cea J., *Optimisation: Théorie et Algorithmes*, Dunod, Paris, 1971.

[11] Cea J. and Geymonat G., Une méthode de linéarisation via l'optimisation, *Istituto Nazionale di Alta Mat. Symp. Mat.* **10**, Bologna (1972), pp. 431–451.

[12] Cea J. et Glowinski R., Sur des méthodes d'optimisation par relaxation, *Revue Française d'Automatique, Informatique, Recherche Opérationnelle* **R3** (1973), pp. 5–32.

[13] Chan T. F. and Keller H. B., Arc length continuation and multigrid techniques for nonlinear eigenvalue problems, *SIAM J. Sc. Stat. Comp.* **3** (1982), 2, pp. 173–194.

[14] Ciarlet P. G. and Rabier P., *Les Équations de Von Karman*, Lecture Notes in Math., Vol. **826**, Springer-Verlag, Berlin, 1980.

[15] Crandall M. G. and Rabinowitz P. H., Bifurcation, perturbation of single eigenvalues and linearized stability, *Arch. Rat. Mech. Anal.* **52** (1973), pp. 161–180.

[16] Crandall M. G. and Rabinowitz P. H., Some continuation and variational methods for positive solutions of nonlinear elliptic eigenvalue problems, *Arch. Rat. Mech. Anal.* **58** (1975), pp. 207–218.

[17] Gabay D., *Méthodes Numériques pour l'Optimisation Non Linéaire*, Thèse d'État, Université Pierre et Marie Curie, Paris, 1979.

[18] Girault V. and Raviart P. A., *Finite Element Approximation of the Navier-Stokes Equations*, Lecture Notes in Math., Vol. **749**, Springer-Verlag, Berlin, 1979.

[19] Glowinski R., *Numerical Methods for Nonlinear Variational Problems*, Springer-Verlag, New-York, 1984.

[20] Glowinski R. et Fortin M. (eds.), *Méthodes de Lagrangien Augmenté, Application à la Résolution Numérique des Problèmes aux Limites*, Dunod, Paris, 1982.

[21] Glowinski R., Keller H. B., and Reinhart L., *Continuation — Conjugate Gradient methods for the least squares solution of nonlinear boundary value problems* (to appear).

[22] Glowinski R. and Le Tallec P., Numerical solution of problems in incompressible finite elasticity by augmented lagrangian methods (I). Two-dimensional and axisymmetric problems, *SIAM J. Appl. Math.* **42** (1982), pp. 400–429.

[23] Glowinski R. and Le Tallec P., *Numerical solution of problems incompressible finite elasticity by augmented lagrangian methods* (II). *Three-dimensional problems* (to appear).

[24] Glowinski R., Lions J. L., and Trémolières R., *Numerical Analysis of Variational Inequalities*, North-Holland, Amsterdam, 1981.

[25] Glowinski R., Mantel B., and Periaux J., Numerical solution of the time dependent Navier–Stokes equations for incompressible viscous fluids by finite elements and alternating direction methods, in: *Numerical Methods in Aeronautical Fluid Dynamics*, P. L. Roe ed., Academic Press, London, 1982, pp. 309–336.

[26] Heywood J. G. and Rannacher R., Finite element approximation of the nonstationary Navier–Stokes problem. I. Regularity of solutions and second order

error estimates for spatial discretization, *SIAM J. Num. Anal.* **19** (1982), pp. 275–311.

[27] Holst T., An implicit algorithm for the transonic full potential equation in conservative form, in: *Computing Methods in Applied Sciences and Engineering*, R. Glowinski, J. L. Lions eds., North-Holland, Amsterdam, 1980, pp. 157–174.

[28] Ibler B., *Résolution des équations de Navier–Stokes par une méthode d'éléments finis*, Thèse de 3ème cycle, Université Paris-Sud, Orsay, 1981.

[29] Keller H. B., Numerical solution of bifurcation and nonlinear eigenvalue problems, in: *Applications of Bifurcation Theory*, P. Rabinowitz ed., Academic Press, New York, 1977, pp. 359–384.

[30] Keller H. B., Global Homotopies and Newton Methods, in: *Recent Advances in Numerical Analysis*, C. de Boor, G. H. Golub eds., Academic Press, New York, 1978, pp. 73–94.

[31] Kikuchi F., Finite element approximations to bifurcation problems of turning point type, in: *Computing Methods in Applied Sciences and Engineering, 1977, Part I*, R. Glowinski, J. L. Lions eds., Lecture Notes in Math. Vol. **704**, Springer-Verlag, Berlin, 1979, pp. 252–266.

[32] Landau L. et Lifschitz E., *Mécanique des Fluides*, Mir, Moscow, 1953.

[33] Le Tallec P., *Numerical Analysis of Equilibrium Problems in Incompressible Nonlinear Elasticity*, Ph. D. Thesis, The University of Texas at Austin, 1980.

[34] Le Tallec P., *Les problèmes d'équilibre d'un corps hyperélastique incompressible en grandes déformations*, Thèse d'Etat, Université Pierre et Marie Curie, Paris, 1981.

[35] Lions J. L., *Contrôle Optimal des Systèmes Gouvernés par des Équations aux Dérivées Partielles*, Dunod-Gauthier Villars, Paris, 1968.

[36] Lions J. L., *Quelques méthodes de résolution des problèmes aux limites non linéaires*, Dunod-Gauthier Villars, Paris, 1969.

[37] Lions P. L. and Mercier B., Splitting algorithms for the sum of two nonlinear operators, *SIAM J. Num. Anal.* **16** (1979), 6, pp. 964–979.

[38] Lozi R., *Analyse Numérique de certains Problèmes de Bifurcation*, Thèse de 3ème Cycle, Université de Nice, 1975.

[39] Mignot F., Murat F., et Puel J. P., Variation d'un point de retournement en fonction du domaine, *Comm. in Part. Diff. Equ.* **4** (1979), pp. 1263–1297.

[40] Mignot F. et Puel J. P., Sur une classe de problèmes non linéaires avec non linéarité positive, croissante, convexe, in: *Comptes-Rendus du Congrès d'Analyse Non Linéaire, Rome, Mai 1978*, Pitagora Editrice, Bologne, 1979, pp. 45–72.

[41] Moore G. and Spence A., The calculation of turning points of nonlinear equations, *SIAM J. Num. Anal.* **17** (1980), pp. 567–576.

[42] Morgan K., Periaux J., and Thomasset F., (eds.), Numerical Analysis of laminar flow over a step, *INRIA Workshop, January 1983*, Le Chesnay.

[43] Percell P., Note on a global homotopy, *Num. Functional Anal. Optim.* **2** (1980), 1, pp. 99–106.

[44] Pironneau O., On the transport-diffusion algorithm and its applications to the Navier–Stokes equations, *Num. Math.* **38** (1982), pp. 309–332.

[45] Polak E., *Computational Methods in Optimization*, Academic Press, New York, 1971.

[46] Powell M. J. D., Restart procedure for the conjugate gradient method, *Math. Programming* **12** (1977), pp. 148–162.

[47] Reinhart L., *Sur la résolution numérique de problèmes aux limites non linéaires*

par des méthodes de continuation, Thèse de 3ème cycle, Université Pierre et Marie Curie, Paris, 1980.

[48] Reinhart L., On the numerical analysis of the Von Karman equations: Mixed Finite Element Approximation and Continuation Techniques, *Numerische Math.* **39** (1982), pp. 371–404.

[49] Steinhoff J. and Jameson A., Multiple solutions of the transonic potential flow equation, *AIAA Journal* **20** (1982), 11, pp. 1521–1525.

[50] Tartar L., *Topics in Nonlinear Analysis*, Publications Mathématiques d'Orsay, Université Paris-Sud, Département de Mathématiques, 1978.

[51] Taylor C. and Hood P., A numerical solution of the Navier–Stokes equations using the finite element technique, *Computers and Fluids* **1** (1973), pp. 73–100.

[52] Temam R., *Theory and Numerical Analysis of the Navier–Stokes equations*, North-Holland, Amsterdam, 1977.

[53] Thomasset F., *Implementation of Finite Element Methods for Navier–Stokes Equations*, Springer-Verlag, New York, 1981.

UNIVERSITÉ PIERRE ET MARIE CURIE,
4 PLACE JUSSIEU – 75230 PARIS CEDEX 05 AND INRIA

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

YU. A. KUZNETSOV

Matrix Iterative Methods in Subspaces

A new branch has recently been formed and is now being actively developed in the field of iterative solution of finite-dimensional problems. This branch has been called *iterative methods in subspaces*. In this paper we present the use of these methods for solving systems of linear algebraic equations with real square matrices arising from finite difference and finite element methods. When implemented or theoretically investigated, these methods do not span an entire vector space but some of its subspaces associated either with the matrix of the original system (in the case of its singularity, for example) or with the properties of matrices participating in derivation of an iterative method.

In the first part of the lecture we deal with some general topics of the theory of matrix iterative methods in subspaces for solving systems of linear algebraic equations. In the second part we show how this theory is applied to devising and studying the iterative methods for solving the systems arising from discretized kinetic transport and Poisson equations by the methods of grids. Here we limit ourselves to the simplest domains, operators, boundary conditions, grids and discretization methods. Detailed information can be found in the references.

1. Some aspects of the theory

Let us consider the system of linear algebraic equations

$$Au = f \tag{1}$$

with real square matrix A of order n, vector $f \in \mathrm{im} A = AE_n$, where E_n is the space of n-dimensional real vectors with the Euclidean scalar product $(\ ,\ )$ and norm $\|\cdot\| = (\ ,\ )^{1/2}$. Let us consider the following itera-

tive method for the solution of (1):

$$u^0 \in U_0,$$
$$u^k = u^{k-1} - \tilde{H}(Au^{k-1} - f), \quad k = 1, 2, \ldots \tag{2}$$

Here U_0 is some closed subset of E_n such that the set $U_A = \{\xi: \xi = Av - f, v \in U_0\}$ is a subspace of E_n, and $\tilde{H}$ is a (generally speaking) nonlinear operator acting from U_A to E_n.

For method (2) let us assume that

— the subspace U_A is invariant with respect to method (2), i.e. for any $\xi \in U_A$ we have $\xi - A\tilde{H}(\xi) \in U_A$;

— operator H is homogeneous of the first order and continuous on any nonzero vector $\xi \in U_A$.

Then the following theorem holds.

THEOREM 1. *If we can define a norm* $\|\cdot\|_*$ *in* U_A *such that for any nonzero* $\xi \in U_A$

$$\|\xi - A\tilde{H}(\xi)\|_* < \|\xi\|_*, \tag{3}$$

then method (2) *converges for any initial guess from* U_0. *Moreover, for any* $u^0 \in U_0$ *the sequence* u^k *in this method converges to some solution* $\hat{u}$ *of* (1) *at the rate of geometrical progression, i.e.*

$$\|u^k - \hat{u}\| \leqslant cq^k \|u^0 - \hat{u}\|, \tag{4}$$

where c *and* q *are some positive constants* ($q < 1$).

Obviously, if $\det A = 0$, then in general the vector $\hat{u}$ would depend upon the initial guess $u^0 \in U_0$.

Let us illustrate the use of the theorem for the successive overrelaxation method. Let the matrix A of system (1) be symmetric and represented as $A = \Lambda - F - F^T$, where Λ is a symmetric positive-definite matrix, and F an arbitrary matrix. Then one may easily extend the known Ostrovsky–Reik theorem [22], [29] to the successive overrelaxation method

$$B_\omega(u^k - u^{k-1}) = -(Au^{k-1} - f), \tag{5}$$

where $B_\omega = \dfrac{1}{\omega}\Lambda - F$.

THEOREM 2. *Method* (5) *converges in* E_n *if and only if either the matrix* A *is positive-semidefinite and* $\omega \in (0, 2)$ *or* A *is negative-semidefinite and* $\omega \notin [0, 2]$.

Actually, if $\det B_\omega \neq 0$, then for method (5)

$$(A^+ \xi^k, \xi^k) = (A^+ \xi^{k-1}, \xi^{k-1}) - \left(\frac{2}{\omega} - 1\right)(AB_\omega^{-1}\xi^{k-1}, \xi^{k-1}),$$

where A^+ is the generalized inverse of A. Then the sufficiency of the condition of Theorem 2 in the case, for example, of A positive-semidefinite follows easily from Theorem 1 by setting

$$\|\xi\|_* = (A^+\xi, \xi)^{1/2}, \qquad \xi \in U_A = \operatorname{im} A.$$

Note that in [8], [11] other approaches were used in studying the convergence of the successive overrelaxation methods in the case $\det A = 0$. Thus Theorem 1 provides a convenient technique for studying the convergence of stationary iterative methods in solving systems of equations with singular matrices.

Now let us discuss a class of nonstationary iterative methods based on the idea of descent. Let H be some matrix such that a given subspace U_A is invariant with respect to the matrix AH, let $U_0 = \{v\colon Av - f \in U_A, v \in E_n\}$ and let D be a self-adjoint and positive-definite matrix in U_A. Let us define the new scalar product $(\xi, \eta)_D = (D\xi, \eta)$ and the norm $\|\cdot\|_D = (\ ,\ )_D^{1/2}$ generated by the matrix D in the subspace U_A. For any given $\xi \in U_A$ consider the operator $\tilde{H}$:

$$\tilde{H}(\xi) = H \sum_{i=1}^{r} \gamma_i (AH)^{i-1}\xi, \tag{6}$$

where r is some positive integer and the parameters γ_i are chosen so that

$$\|\xi - A\tilde{H}(\xi)\|_D = \min_{a_1,\ldots,a_r} \|\xi - \sum_{i=1}^{r} a_i (AH)^i \xi\|_D. \tag{7}$$

Consider the vector $b(\xi) \in E_r$ with components $b_i(\xi) = ((AH)^i\xi, \xi)_D$, $i = 1, \ldots, r$. Then, according to Theorem 1, the following theorem holds:

THEOREM 3. *The nonstationary method* (2) *with the operator* $\tilde{H}$ *defined above converges on the set* U_0 (*i.e., for any* $u^0 \in U_0$) *if and only if* $b(\xi) \neq 0$ *for any nonzero vector* $\xi \in U_A$.

COROLLARY 1. *If for some positive* $t \leqslant r$ *the matrix* $(AH)^t$ *is positive-definite in* U_A *with respect to the scalar product* $(\ ,\ )_D$, *i. e.,* $((AH)^t\xi, \xi)_D > 0$ *for any nonzero* $\xi \in U_A$, *then the method* (2), (6), (7) *converges.*

COROLLARY 2. *If the matrix AH is a D-self-adjoint operator in U_A, i.e., $(AH\xi, \eta)_D = (\xi, AH\eta)_D$ for all $\xi, \eta \in U_A$, and $\ker(AH) \cap U_A = 0$, then the method* (2), (6), (7) *converges for any $r \geqslant 2$.*

Let us further assume that the matrix AH is D-self-adjoint and D-positive-definite in U_A. Then to solve system (1) one often employs three-term nonstationary iterative methods of the form

$$u^0 \in U_0, \quad u^{k+1} = u^k - a_k H(Au^k - f) - b_k(u^k - u^{k-1}) \tag{8}$$

with different choices of the sequences of parameters a_k and b_k. Assuming that $u^0 \in U_0$ consider the generalized conjugate gradient method (GCG) implemented with the following two-term formulae [9], [19]:

$$\begin{aligned} p_k &= \begin{cases} H\xi^0, & k = 1, \\ H\xi^{k-1} - \alpha_k p_{k-1}, & k > 1, \end{cases} \\ \alpha_k &= \frac{(AH\xi^{k-1}, Ap_{k-1})_D}{\|Ap_{k-1}\|_D^2} = -\frac{\|\xi^{k-1}\|_{DAH}^2}{\|\xi^{k-2}\|_{DAH}^2}, \\ u^k &= u^{k-1} - \beta_k p_k, \\ \beta_k &= \frac{(\xi^{k-1}, Ap_k)_D}{\|Ap_k\|_D^2} = \frac{\|\xi^{k-1}\|_{DAH}^2}{\|Ap_k\|_D^2}. \end{aligned} \tag{9}$$

The number of steps k in (9) does not exceed the dimension of U_A (it will be shown below that for $\dim U_A \ll n$ this remark is quite important in estimating the efficiency of the method).

If we want to solve system (1) with accuracy ε, i.e., to minimize $1/\varepsilon$ times the D-norm of the initial residue ξ^0, then the number of steps $k = k_\varepsilon$ will be the least integer such that $2/(\varrho^{-k_\varepsilon} + \varrho^{k_\varepsilon}) \leqslant \varepsilon$, where $\varrho = [1 - (m/M)^{1/2}]/[1 + (m/M)^{1/2}]$ and m, $M (0 < m < M)$ are the endpoints of the interval containing the nonzero part of the spectrum of AH. If we assume that $m \ll M$, then the expression for $k_\varepsilon (\varepsilon \ll 1)$ takes the form:

$$k_\varepsilon \approx \tfrac{1}{2}(M/m)^{1/2} \ln(1/\varepsilon).$$

It is necessary to remark that in the methods discussed further on, the practical implementation of formulae (9) possesses some important computational features due to the structure of subspaces U_A. Thus for such cases it is more correct to call method (9) the *generalized conjugate gradient method in a subspace U_A*.

2. The method of simple iteration for kinetic transport equation

Let us consider the kinetic transport equation for the plane (z, μ) geometry [23]:

$$\mu \frac{\partial u}{\partial z} + \sigma u = \frac{\sigma_s}{2} \int_{-1}^{1} u d\mu' + f, \quad (z, \mu) \in (0, 1) \times [-1, 1], \tag{10}$$

where $\sigma = \sigma(z)$, $\sigma_s = \sigma_s(z)$ and $f = f(z)$ are continuous functions ($\sigma \geqslant \sigma_s \geqslant 0$). We seek the solution satisfying the boundary conditions $u(0, \mu) = 0$ for $\mu \in (0, 1]$ and $u(1, \mu) = 0$ for $\mu \in [-1, 0)$. When we realize the discretization of this problem over the uniform grid in the variables z and μ, we get the system (1) with matrix

$$A = B - C \tag{11}$$

of order $n = lt$, where l is the number of mesh points in z, t is the number of mesh points in μ (t is even). We can represent matrices B and C in the form

$$B = \begin{bmatrix} B_1 & 0 \\ 0 & B_1^T \end{bmatrix}, \quad C = \Sigma_s \otimes Q. \tag{12}$$

Let us describe these matrices [12]: B_1 is a block-diagonal matrix of order $n/2$, each diagonal block being the lower triangular bidiagonal positive-definite M-matrix of order l; Σ_s is a diagonal positive-semidefinite matrix; Q is a matrix of order t, an orthogonal projector of unit rank (an analogue of the operator $\frac{1}{2} \int_{-1}^{1} d\mu'$); C is a symmetric positive-semidefinite matrix; and A is a positive-definite M-matrix.

One of the most widely used methods for solving the sets of equations arising from the discretization of kinetic transport equations is the so-called simple iteration method:

$$Bu^k = Cu^{k-1} + f. \tag{13}$$

It follows from the theory of regular splitting of matrices [29] that for method (13) we have $\varrho(B^{-1}C) < 1$ and hence the method converges for any initial vector $u^0 \in E_n$.

Since for all $k \geqslant 1$ the vectors of residuals in method (13) satisfy $\xi^k = CB^{-1}\xi^{k-1}$, they belong to the subspace $U_A = CE_n$. Each vector ξ of

this subspace is represented in the form $\xi = \psi \otimes e$, where ψ is a vector from E_l (to be more exact, from $\Sigma_s E_l$), and e is a vector of dimension t with all components equal to unity. Thus instead of (13) one may consider the iterative process

$$Cu^k = Cu^{k-1} - CB^{-1}\xi^{k-1}, \qquad \xi^k = CB^{-1}\xi^{k-1} \tag{14}$$

in U_A, where two arrays of length l are sufficient for the representation of the vectors Cu_k and ξ^k. In this case it is sufficient to perform $O(lt)$ arithmetical operations (taking into account the block structure of matrices B and C and using one additional array of length l) to compute the vector $CB^{-1}\xi^{k-1} \in U_A$ for the given vector $\xi^{k-1} \in U_A$. After r iterations by method (14) the vector u^r can be found as the solution of the system

$$Bu^r = Cu^r + f + \xi^r. \tag{15}$$

This property of method (13) is widely used for its optimization, i.e., for deriving more efficient methods of solving the original system of equations [18]. Let us discuss one of the possible strategies [12].

The matrices C and C^+ are obviously self-adjoint and positive-definite operators in U_A. Not too sophisticated computations show that the matrices B, B^{-1}, A and A^{-1} have the above properties as well. Moreover, it can be shown that for D equal to C^+, B^{-1} or A^{-1}, the matrix AB^{-1} is D-self-adjoint and D-positive-definite in U_A. Thus for solving system (1) with matrix A from (11) one may employ the generalized conjugate gradient method (9) in the subspace $U_A = \operatorname{im} C$, by assuming $H = B^{-1}$ and choosing D equal to C^+, B^{-1} or A^{-1}. The computation formulae of this method are

$$\begin{aligned}
Cp_k &= \begin{cases} CB^{-1}\xi^0, & k = 1, \\ CB^{-1}\xi^{k-1} - \alpha_k Cp_{k-1}, & k > 1, \end{cases} \\
Ap_k &= \begin{cases} (I - CB^{-1})\,\xi^0, & k = 1, \\ (I - CB^{-1})\,\xi^{k-1} - \alpha_k p_{k-1}, & k > 1, \end{cases} \\
\alpha_k &= -\frac{\left((I - CB^{-1})\,\xi^{k-1}, \xi^{k-1}\right)_D}{\left((I - CB^{-1})\,\xi^{k-2}, \xi^{k-2}\right)_D}, \\
Cu^k &= Cu^{k-1} - \beta_k Cp_k, \qquad \xi^k = \xi^{k-1} - \beta_k Ap_k, \\
\beta_k &= \frac{\left((I - CB^{-1})\,\xi^{k-1}, \xi^{k-1}\right)_D}{\|Ap_k\|_D^2}, \\
k &= 1, 2, \ldots, r,
\end{aligned} \tag{16}$$

where I is the unit matrix. Naturally, the initial residual vector ξ^0 of method (16) should belong to U_A, which, e.g., corresponds to the choice of u^0 as the solution of the system $Bu^0 = f$ (in this case $\xi^0 = -Cu^0$).

For any choice of the matrix D one may find an efficient algorithm for the implementation of method (16). When doing this, we always use the fact that the vectors Ap_k belong to the subspace U_A. For example, since $C^+ = \Sigma_s^+ \otimes Q$ and Σ_s^+ may be found explicitly, for $D = C^+$ and for any $\xi, \eta \in U_A$ we have $(\xi, \eta)_D = (\xi, \eta)_{D_0}$, where $D_0 = t^{-1}\Sigma_s^+ \otimes I_t$ (I_t is the unit matrix of order t). Therefore one needs only $O(l)$ arithmetical operations to compute scalar product in method (16). Thus, increasing negligibly the number of arithmetical operations per step (by $O(l)$ in comparison with $O(lt)$) and using two additional arrays of length l one arrives at a method with essentially higher convergence rate (especially when $\varrho(B^{-1}C)$ is close to unity) with respect to method (13). Choosing D equal to B^{-1} or A^{-1} leads to the method with the same computational properties.

This approach to the optimization of the method of simple iteration can be extended to the case of more general geometries, boundary conditions and discretizations of kinetic transport equations, as well as to the problems with nonisotropic scattering [13].

3. Extension method (fictitious components method)

Let us consider Neumann's boundary value problem for

$$-\Delta u + u = f \tag{17}$$

in a bounded two-dimensional domain Ω_0 which, for example, is a union of a finite number of rectangles with sides parallel to the coordinate axes. Let Ω_0 be covered with a square grid of step h used for deriving the five-point difference equations approximating problem (17). Then we get the system of equations

$$A_0 u_0 = f_0 \tag{18}$$

with a symmetric positive-definite matrix A_0 of order n_0 and a vector $f_0 \in E_{n_0}$; here n_0 is the number of grid nodes belonging to $\overline{\Omega}_0$.

Instead of system (18) let us consider system (1) with matrix

$$A = \begin{bmatrix} A_0 & 0 \\ 0 & A_1 \end{bmatrix} \tag{19}$$

of order $n > n_0$ and with $f = \begin{bmatrix} f_0 \\ 0 \end{bmatrix} \in E_n$. System (1) is equivalent to the original system (18) in the sense that the first n_0 components of any of its solutions u give the solution of system (18). The purpose of such an extension is to raise the possibility of choosing the matrix H in method (9).

Let us assume that H is some symmetric and positive-definite matrix of order n and pose the problem of finding the best matrix A_1 from the family of all symmetric and positive-semidefinite matrices of order $n-n_0$, to provide the most rapid convergence of method (9). The solution of this problem is given in Theorem 4.

THEOREM 4. *The problem formulated above is solved by taking* $A_1 = 0$.

Now the matrix procedure described above, which has been called the "fictitious components method" [21], will be employed to solve system (18). Let us consider the region Ω which is the minimum rectangle containing the region Ω_0. Then using a square grid with step h we approximate the Neumann problem for equation (17) in the same manner but in the region Ω, setting $f = 0$ outside Ω_0. As a result we have the system of algebraic equations with matrix

$$B = S_l \otimes \hat{I}_t + \hat{I}_l \otimes S_t + \hat{I}_l \otimes \hat{I}_t \tag{20}$$

of order $n = lt$, where l is the number of grid nodes in the rectangle along one variable, t is the number of nodes along the other variable, the lower indices l and t denote the orders of the matrices

$$S = \frac{1}{h^2}\begin{bmatrix} 1 & -1 & & & & 0 \\ -1 & 2 & . & & & \\ & . & . & . & & \\ & & . & . & . & \\ & & & . & 2 & -1 \\ 0 & & & & -1 & 1 \end{bmatrix} \quad \text{and} \quad \hat{I} = \operatorname{diag}\{\tfrac{1}{2}, 1, \ldots, 1, \tfrac{1}{2}\},$$

respectively.

For solving system (1) with matrix A from (19) with $A_1 = 0$, let us employ the generalized conjugate gradient method (9) in the subspace

$U_A = \mathrm{im} A$ with matrices $D = A^+$ and $H = B^{-1}$:

$$p_k = \begin{cases} B^{-1}\xi^0, & k = 1, \\ B^{-1}\xi^{k-1} - \alpha_k p_{k-1}, & k > 1, \end{cases}$$

$$\alpha_k = -\frac{\|\xi^{k-1}\|^2_{B^{-1}}}{\|\xi^{k-2}\|^2_{B^{-1}}}, \qquad \beta_k = \frac{\|\xi^{k-1}\|^2_{B^{-1}}}{\|p_k\|^2_A}, \tag{21}$$

$$u^k = u^{k-1} - \beta_k p_k, \qquad k = 1, 2, \ldots, k_\varepsilon.$$

It follows from [1] that in view of the above assumptions the nonzero eigenvalues of the matrix AB^{-1} are within the range $[m, 1]$ where m is positive and independent of the grid side h. Thus by using the algorithms from [17], [27] for solving the system with matrix B, the following theorem holds.

THEOREM 5. *For solving system* (18) *with accuracy* ε *it is sufficient to perform* $O\big(\ln\ln(1/h)\,\ln(1/\varepsilon)/h^2\big)$ *arithmetical operations.*

Now we choose $U_A = \mathrm{im} A(I - B^{-1}A)$, which is consistent with the choice of initial vector in method (21) e.g. using the formula $Bu^0 = f$. It can easily be shown that any vector $\xi \in U_A$ has no more than $O(1/h)$ nonzero components corresponding to grid nodes belonging to $\partial\Omega_0$. By employing this fact, method (21) can be implemented by using only a finite (h-independent) number of vectors of length $O(1/h)$ and performing only $O(1/h^2)$ arithmetical operations per each step [17]. Thus the method requires $O(|\ln\varepsilon|/h^2) + O(\ln|\ln h|/h^2)$ arithmetical operations and can be considered as one of the most efficient methods for solving the class of problems formulated above [3]. A large number of papers (see [6], [25]) deals with the approaches discussed here and with some other ones for constructing methods of extension, including the methods solving the Dirichlet problem.

4. Domain decomposition method

Let us consider the Dirichlet problem for the two-dimensional Poisson equation in an L-shaped region Ω such that $\bar{\Omega} = \bar{\Omega}_1 \cup \bar{\Omega}_2$; here $\Omega_1 = (0, 1) \times (0, 2)$ and $\Omega_2 = (1, 2) \times (0, 1)$. Employing the conventional finite element method with piecewise-linear test functions on a grid with step $h = 1/(l+1)$, we arrive at system (1) with symmetric positive-definite

matrix A of order $n = l(3l+2)$. If the first group of unknowns includes those corresponding to the inner nodes of the subregion Ω_1 and the second group involves those of subregion Ω_2, while the unknowns of the range $\bar{\Omega}_1 \cap \bar{\Omega}_2$ are in the third group, then the matrix A can be written in a block form

$$A = \begin{bmatrix} A_{11} & 0 & A_{13} \\ 0 & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix}. \tag{22}$$

Here A_{33} is a three-diagonal matrix of order l and matrices A_{11} and A_{22} are the matrix analogues of the five-point difference approximations of the Laplace operator for rectangular subregions Ω_1 and Ω_2.

If the block iterative Gauss–Seidel method is used for solving the derived system (1),

$$Bu^k = Cu^k + f, \quad B = \begin{bmatrix} A_{11} & 0 & A_{13} \\ 0 & A_{22} & A_{23} \\ 0 & 0 & A_{33} \end{bmatrix}, \quad C = B - A, \tag{23}$$

then the residual vectors ξ^k of this method for all $k \geqslant 1$ belong to $U_A = \operatorname{im} C$, i.e., all components, excluding probably the last l ones, equal zero. One can see that the matrix B^{-1} is a self-adjoint and positive-definite operator in this subspace U_A, hence for solving this system we can use the generalized conjugate gradient method (9) with matrices $D = A^{-1}$ and $H = B^{-1}$, assuming that the initial guess is chosen as the solution of the equation $Bu^0 = f$ ($\xi^0 = -Cu^0$).

It is known [14] that the eigenvalues of the matrix AB^{-1} of this version of method (9) lie within the range $[dh, 1]$, where d is a positive constant independent of h. Thus for solving the system with accuracy ε, we need $k_\varepsilon = O(|\ln \varepsilon|/h^{1/2})$ iterations. If the method is implemented following formulae (21) by solving the systems with matrices A_{11} and A_{22} using at each step e.g. $O(\ln|\ln h|/h^2)$ arithmetical operations then it will be worse than many other existing methods for solving similar systems [26]. The situation changes drastically if we take into account the structure of the subspace U_A. Let us introduce the matrix R_3 with all elements equal to zero except the last l diagonal elements that are equal to one.

Then the formulae for the generalized conjugate gradient method discussed above take the form:

$$
\begin{aligned}
R_i p_k &= \begin{cases} R_i B^{-1}\xi^0, & k=1, \\ R_i B^{-1}\xi^{k-1} - a_k R_i p_{k-1}, & k>1, \end{cases} \\
a_k &= -\frac{(R_i B^{-1}\xi^{k-1}, \xi^{k-1})}{(R_i B^{-1}\xi^{k-2}, \xi^{k-2})}, \quad i=1,2,3, \\
Cu^k &= Cu^{k-1} - \beta_k C p_k, \quad \xi^k = \xi^{k-1} - \beta_k A p_k, \\
\beta_k &= \frac{(R_3 B^{-1}\xi^{k-1}, \xi^{k-1})}{(R_3 p_k, A p_k)}, \quad k=1,2,\ldots,k_\varepsilon;
\end{aligned}
\tag{24}
$$

here $R_1 = C$, $R_2 = A$ and $R_2 B^{-1} = I - CB^{-1}$.

Since all vectors participating in the process belong to the subspace U_A, i.e., contain no more than l nonzero components, implementation of all vector operations of an arbitrary kth step of method (24) needs to store only five vectors of length l and to perform $O(l)$ arithmetical operations. Now there is one more problem to be discussed: how to compute the last l components of the vector $R_i B^{-1}\xi$, $i = 1, 2, 3$, for a given $\xi \in U_A$. An algorithm has been derived in [14] which requires $O(1/h^2)$ arithmetical operations for solving this problem at a single run and $O(|\ln h|/h)$ operations for solving each subsequent task. It is sufficient to store simultaneously only a finite (independent of h) number of vectors of dimension $O(1/h)$. Such algorithms belong to the class of *methods of partial solution of systems of linear algebraic equations* [14], [17].

Thus we have proved the following theorem:

THEOREM 6. *Let the vectors Cu^0 and ξ^0 of method* (24) *be computed. Then for implementing $k_\varepsilon = O(|\ln\varepsilon|/h^{1/2})$ steps of this method it is sufficient to make*

$$
O(1/h^2) + O(|\ln h| \cdot |\ln\varepsilon|/h^{3/2}) \tag{25}
$$

arithmetical operations and to store only a finite (independent of h and ε) number of vectors of dimension $O(1/h)$.

It should be noted that the vectors Cu^0, $\xi^0 \in U_A$ as well as all components of the vector u^r (solution of the system of type (15)) can be computed by using $O(\ln|\ln h|/h^2)$ arithmetical operations.

Extension of this result to the case of more general boundary conditions, geometry of the domain and type of differential operators is discussed e.g. in [15], [16].

Conclusion

The iterative methods in subspaces which have been presented in the paper certainly do not include all their variety. At present this field of numerical solution of finite-dimensional problems is being intensively developed. Other interpretations of the methods in subspaces are of great interest, e.g. the finite-dimensional analogues of integral equations, imbedding ideas [7], as well as investigations of different versions of geometric decomposition methods as numerical processes in subspaces. Only some initial results have been obtained in the methods of partial solution of systems of linear equations [4], [14], [17], [28]. Studies in this field may lead to efficient new algorithms for implementing the methods in subspaces, in particular the methods of extension and decomposition. The study of generalized conjugate gradient methods in terms of [5], [30] for optimization of the computational processes in subspaces for nonsymmetric cases is of considerable interest. Certainly, investigations in this field will provide novel optimal [3] methods for solving a variety of problems of mathematical physics.

References

[1] Astrakhantsev G. P., Kand. fiz.-mat. nauk thesis, Leningrad Branch of Mathematical Institute, Acad. Sci. USSR, Leningrad, 1972.
[2] Astrakhantsev G. P., *Zh. Vychisl. Mat. i Mat. Fiz.* **18** (1) (1978), pp. 118–125.
[3] Bakhvalov N. S., In: *Intern. Math. Congress in Nizza, 1970,* Nauka, Moscow, 1972, pp. 27–33.
[4] Bakhvalov N. S. and Orekhov M. Yu., *Zh. Vychisl. Mat. i Mat. Fiz.* **22** (6) (1982), pp. 1386–1392.
[5] Concus P. and Golub G. H., In: *Lecture Notes in Econ. and Math. Systems,* 134, Springer-Verlag, Berlin, 1976, pp. 56–65.
[6] Dryja M., *Numer. Math.* **39** (1982), pp. 51–64.
[7] D'yakonov E. G., In: *Chislennye metody v matematicheskoi fizike,* Novosibirsk, 1979, pp. 45–68.
[8] Forsythe G. E. and Wasow W. R., *Finite-Difference Methods for Partial Differential Equations,* J. Wiley and Sons, New York–London, 1960, 444 pp.
[9] Hestenes M. R. and Stiefel E., *J. Res. Nat. Bur. Standards* **49** (1952), pp. 409–436.
[10] Kaporin I. E. and Nikolaev E. S., *Dokl. Akad. Nauk SSSR* **251** (3) (1980), pp. 544–548.

[11] Keller H. B., *SIAM J. Numer. Anal.* **2** (1965), pp. 281–290.
[12] Kuznetsov Yu. A., Kand. fiz.-mat. nauk thesis, Computing Centre, Siberian Branch Acad. Sci. USSR, Novosibirsk, 1969.
[13] Kuznetsov Yu. A., In: *Vychislitel'nye metody v matematicheskoi fizike, geofizike i optimal'nom upravlenii*, Nauka, Novosibirsk, 1978, pp. 125–137.
[14] Kuznetsov Yu. A., In: *Variatsionno–raznostnye metody v matematicheskoi fizike*, Novosibirsk, 1978, pp. 178–212.
[15] Kuznetsov Yu. A., In: *Chislennye metody v matematicheskoi fizike*, Novosibirsk, 1979, pp. 20–44.
[16] Kuznetsov Yu. A., In: *Metody resheniya variatsionno–raznostnykh uravnenii*, Novosibirsk, 1979, pp. 24–59.
[17] Kuznetsov Yu. A. and Matsokin A. M., In: *Vychislitel'nye metody linieinoi algebry*, Novosibirsk, 1978, pp. 62–89.
[18] Lebedev V. I., *Zh. Vychisl. Mat. i Mat. Fiz.* **7** (6) (1967), pp. 1250–1252.
[19] Marchuk G. I and Kuznetsov Yu. A., *Dokl. Akad. Nauk SSSR* **181** (6) (1968), pp. 1331–1334.
[20] Marchuk G. I. and Kuznetsov Yu. A., *Iteratsionnye metody i kvadratichnye funktsionaly*, Novosibirsk, 1972, 205 pp.; see also in: J.-L. Lions and G. I. Marchuk (ed.), *Sur les méthodes numériques en sciences physiques et économiques*, Dunod, Paris, 1974, pp. 3–132.
[21] Marchuk G. I. and Kuznetsov Yu. A., In: *Vychislitel'nye metody lineinoi algebry*, Novosibirsk, 1972, pp. 4–20.
[22] Marchuk G. I. and Kuznetsov Yu. A., In: *Gatlinburg VI Symp. on Numer. Algebra*, Conf. Manuscripts, München, 1974.
[23] Marchuk G. I. and Lebedev V. I., *Chislennye metody v teorii perenosa neitronov*, Atomizdat, Moscow, 1981, 454 pp.
[24] Matsokin A. M., In: *Vychislitel'nye metody lineinoi algebry*, Novosibirsk, 1980, pp. 65–77.
[25] Proskurowski W. and Widlund O., *Math. Comput.* **30** (1976), pp. 433–468.
[26] Samarskii A. A. and Nikolaev E. S., *Metody resheniya setochnykh uravnenii*, Nauka, Moscow, 1978, 591 pp.
[27] Swarztrauber P. N., *SIAM Rev.* **19** (1977), pp. 490–501.
[28] Tyrtyshnikov E. E., In: *Chislennye metody algebry*, Moscow, 1981, pp. 10–26.
[29] Varga R. S., Matrix Iterative Analysis, Prentice-Hall, Englewood Cliffs, N. J., 1962, 322 pp.
[30] Voevodin V. V., *Zh. Vychisl. Mat. i Mat. Fiz.* **19** (5) (1979), pp. 1313–1317.

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

CHARLES A. MICCHELLI

Recent Progress in Multivariate Splines

Spline functions constitute a powerful tool for computation. Since the 1946 seminal paper of I. J. Schoenberg [2] which studied methods for smoothing data, spline functions have found diverse applications in science and engineering, too numerous to document here. The theory of splines has likewise been vigorously investigated by many people and their papers account for a significant proportion of the activity in approximation theory during the past twenty years. Almost all this material dealt with univariate splines. Multivariate splines were generally considered only within the context of the finite element method for solving PDE's. We now see a surge of activity directed towards a deeper understanding of spline spaces in higher dimensions. The theory is rapidly growing, but much more is needed to be done. Several recent conferences focused on these developments and it is hoped that new applications of multivariate splines will result.

A key idea in this theory is a geometric method which suggests that smooth piecewise polynomials can be constructed as volumes of polyhedra. This idea can be traced back to an observation of H. B. Curry and I. J. Schoenberg about univariate splines. It was later put into a multivariate context by C. de Boor and I. J. Schoenberg.

Polyhedral splines are incredibly rich in detail. A useful method for their analysis was suggested by new results on multivariate interpolation which were reported on by C. de Boor at the last International Congress of Mathematicians held in Helsinki, 1978. Formulae are now available for the computation of polyhedral splines along with their derivatives and integrals. In addition to their numerical usefulness, these formulae reveal many of the beautiful structural properties of polyhedral splines which otherwise might have been difficult to uncover.

Spline spaces constructed from linear combination of polyhedra splines with good approximation properties have been found. In particular,

optimal error bounds by quasi-interpolants for approximating functions in Sobolev spaces have been obtained. Needless to say, the construction of these spaces and the analysis of their approximation properties is much harder than in the one-dimensional case.

One-dimensional spline approximation methods which are shape-preserving (a property which is useful in computer-aid design) are available. The situation in higher dimensions has not been clarified and any information on polyhedral splines in this direction would be helpful.

In another direction, substantial progress has been made in understanding bivariate spline spaces of fixed degree and smoothness over a given partition. For quite a while many questions concerning the dimension of these spaces and the construction of their explicit bases were left unsettled. Recent work by C. Chui, L. L. Schumaker and R. W. Wang has added useful information to our understanding of this problem. Polyhedral splines are helpful in studying spline spaces over regular partitions.

An extensive bibliography and an elaboration on the brief remarks made here can be found in the recent survey article on multivariate splines written jointly with Wolfgang Dahmen [1].

References

[1] Dahmen W. and Micchelli C. A., Recent Progress in Multivariate Splines. In: C. K. Chui, L. L. Schumaker, J. D. Ward (eds.), *Approximation Theory IV*, Academic Press, New York, 1983.

[2] Schoenberg I. J., Contributions to the Problem of Approximation of Equidistant Data by Analytic Functions, *Quart. Appl. Math.* **4** (1946), Part A, pp. 45–99; Part B, pp. 112–141.

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

M. J. D. POWELL

On the Rate of Convergence of Variable Metric Algorithms for Unconstrained Optimization

A procedure is described that is suitable for investigating rates of convergence of variable metric algorithms numerically. The usual rate seems to be the one that is given by Ritter [6], but, except when there are only two variables, we are unable to prove that Ritter's rate is achieved under mild conditions on the objective function. An example is given to show that, even though the objective function is twice continuously differentiable and uniformly convex, the Q-order of convergence can be less than the R-order. We also find that, if step-lengths of one are used instead of perfect line searches, then severe deterioration can occur in the rate of convergence.

1. Introduction

Variable metric algorithms are highly successful for calculating least values of differentiable functions of several variables, but their good properties have been established theoretically only when the objective function is convex. Even in this case little is known about rates of convergence. We review some published results, and try to throw some new light on this subject.

We let $\{F(\underline{x})\colon \underline{x} \in \boldsymbol{R}^n\}$ be the objective function of the calculation. We assume that it is real valued and three times differentiable. A variable metric algorithm for minimizing $F(\cdot)$ uses $n \times n$ positive definite matrices $\{B_k\colon k = 1, 2, 3, \ldots\}$ to form a sequence $\{\underline{x}_k\colon k = 1, 2, 3, \ldots\}$ in $\boldsymbol{R}^n$ such that the inequality

$$F(\underline{x}_{k+1}) < F(\underline{x}_k) \tag{1.1}$$

holds for all k. Because it is usual to finish the calculation if a zero gradient vector is found, we assume that $\underline{\nabla} F(\underline{x}_k)$ is non-zero for all k. The initial values $\underline{x}_1$ and B_1 are data.

In order to obtain $\underline{x}_{k+1}$ from $\underline{x}_k$, the search direction $\underline{d}_k$ is defined by the equation

$$B_k\underline{d}_k = -\underline{\nabla}F(\underline{x}_k). \tag{1.2}$$

Then, by considering the function of one variable $\{F(\underline{x}_k+\alpha\underline{d}_k)\colon \alpha\in\boldsymbol{R}\}$, a step-length α_k is chosen such that the point

$$\underline{x}_{k+1} = \underline{x}_k+\alpha_k\underline{d}_k \tag{1.3}$$

satisfies inequality (1.1) and the condition

$$\underline{\delta}_k^T\underline{\gamma}_k > 0, \tag{1.4}$$

where $\underline{\delta}_k$ and $\underline{\gamma}_k$ are the vectors

$$\left.\begin{aligned}\underline{\delta}_k &= \underline{x}_{k+1}-\underline{x}_k\\ \underline{\gamma}_k &= \underline{\nabla}F(\underline{x}_{k+1})-\underline{\nabla}F(\underline{x}_k)\end{aligned}\right\}. \tag{1.5}$$

The calculation ends if no acceptable step-length can be found, but, except for the effects of computer rounding errors, this should not happen if $F(\cdot)$ is bounded below. Finally, the iteration defines the matrix B_{k+1}; we consider the case when it is given by the BFGS formula

$$B_{k+1} = B_k-\frac{B_k\underline{\delta}_k\underline{\delta}_k^TB_k}{\underline{\delta}_k^TB_k\underline{\delta}_k}+\frac{\underline{\gamma}_k\underline{\gamma}_k^T}{\underline{\delta}_k^T\underline{\gamma}_k}. \tag{1.6}$$

For further details of this calculation, including line search techniques to determine α_k, and the positive definiteness of the matrices $\{B_k\colon k = 1, 2, 3, \ldots\}$, the book by Fletcher [3] is recommended. It is important to our analysis to note that, if $H_k = B_k^{-1}$, then B_{k+1}^{-1} is the matrix

$$H_{k+1} = \left(I-\frac{\underline{\delta}_k\underline{\gamma}_k^T}{\underline{\delta}_k^T\underline{\gamma}_k}\right)H_k\left(I-\frac{\underline{\gamma}_k\underline{\delta}_k^T}{\underline{\delta}_k^T\underline{\gamma}_k}\right)+\frac{\underline{\delta}_k\underline{\delta}_k^T}{\underline{\delta}_k^T\underline{\gamma}_k}. \tag{1.7}$$

If $F(\cdot)$ is convex and if its least value occurs at just one point in $\boldsymbol{R}^n$, $\underline{x}^*$ say, then several line search techniques ensure that $\{\underline{x}_k\colon k = 1, 2, 3, \ldots\}$ converges to $\underline{x}^*$ (Powell, [5]). Further, if $\nabla^2F(\underline{x}^*)$ is positive definite, then the superlinear rate of convergence

$$\lim_{k\to\infty}\|\underline{x}_{k+1}-\underline{x}^*\|/\|\underline{x}_k-\underline{x}^*\| = 0 \tag{1.8}$$

can be achieved by setting $\alpha_k = 1$ for all sufficiently large k. This rate

can also be obtained by "perfect" line searches, which means that α_k is the value of α that minimizes the line search objective function $\{F(\underline{x}_k + \alpha \underline{d}_k)\colon \alpha \in \boldsymbol{R}\}$.

Throughout this paper we assume that the sequence $\{\underline{x}_k\colon k = 1, 2, 3, \ldots\}$ is convergent to a point $\underline{x}^*$ at which $\underline{\nabla} F(\underline{x}^*) = 0$ and $\nabla^2 F(\underline{x}^*)$ is positive definite. Because variable metric algorithms are invariant under affine transformations, we assume without loss of generality that $\nabla^2 F(\underline{x}^*) = I$ and that $\underline{x}^* = 0$. We seek convergence results that are stronger than equation (1.8).

Many published results are extensions of the fact that, if $F(\cdot)$ is a convex quadratic function, and if all line searches are perfect, then $\underline{\nabla} F(\underline{x}_k) = 0$ after at most n iterations. Thus Burmeister [1] shows that, for general $F(\cdot)$, the n-step quadratic rate of convergence

$$\|\underline{x}_{k+n} - \underline{x}^*\| = O(\|\underline{x}_k - \underline{x}^*\|^2) \tag{1.9}$$

is obtained, while Ritter [6] gives the stronger relation

$$\|\underline{x}_{k+n} - \underline{x}^*\| = o(\|\underline{x}_k - \underline{x}^*\|^2), \tag{1.10}$$

assuming that line searches are perfect or almost perfect. Further, making the additional assumption that the $n \times n$ matrices $\{U_k\colon k = 1, 2, 3, \ldots\}$ are bounded away from singularity, where the columns of U_k are the vectors $\{\underline{d}_{k+j}/\|\underline{d}_{k+j}\|\colon j = 1, 2, \ldots, n\}$, Schuller [7] establishes the limit

$$\|\underline{x}_{k+n} - \underline{x}^*\| = O(\|\underline{x}_{k+n-1} - \underline{x}^*\| \, \|\underline{x}_{k-1} - \underline{x}^*\|). \tag{1.11}$$

Using another additional assumption, Ritter [6] derives the stronger bound

$$\|\underline{x}_{k+n} - \underline{x}^*\| = O(\|\underline{x}_{k+n-1} - \underline{x}^*\| \, \|\underline{x}_k - \underline{x}^*\|) \tag{1.12}$$

and that, in the case $\nabla^2 F(\underline{x}^*) = I$, the columns of U_k become mutually orthogonal as $k \to \infty$. However, because one can easily construct examples where one or more variables are not altered during the calculation, these assumptions should be questioned.

Therefore Section 2 reports some numerical experiments that were run to test the rate of convergence of variable metric algorithms with perfect line searches for nonquadratic objective functions. The rate (1.12) was found to be usual for small values of n. However, without assuming that the matrices $\{U_k\colon k = 1, 2, 3, \ldots\}$ are bounded away from singularity,

at present we can only prove the relation

$$\|\underline{x}_{k+n}-\underline{x}^*\| = O(\|\underline{x}_k-\underline{x}^*\| \ \|\underline{x}_{k+1}-\underline{x}^*\|), \tag{1.13}$$

which is established in Section 3 because it is a little stronger than expressions (1.9) and (1.10). We note that equations (1.12) and (1.13) are the same when $n = 2$, and an example in Section 4 shows analytically that this rate of convergence can occur when there are two variables, but another example shows that different convergence behaviour is possible. Some remarks on the case when the number of variables is infinite are made in Section 5. All of the theory so far assumes perfect or almost perfect line searches, but the choice $\alpha_k = 1$ is usual in practice. Therefore Section 6 studies the rate of convergence for unit step-lengths, and we find that it is not nearly as good as before. Finally there is a brief discussion of our results and analysis and some suggestions for future work.

2. Numerical experiments

In order to obtain useful information from many iterations of a variable metric minimization calculation, we let $F(\cdot)$ have the form

$$F(\underline{x}) = \tfrac{1}{2}\|\underline{x}\|_2^2+Z(\underline{x}), \quad \underline{x} \in \mathbf{R}^n, \tag{2.1}$$

where $|Z(\underline{x})|$ is $O(\|\underline{x}\|^3)$, and our computer programme allows for the quadratic part of expression (2.1) analytically. Further, instead of working with the matrices $\{B_k\colon k = 1, 2, 3, \ldots\}$, we use and update the matrices

$$E_k = B_k^{-1}-I = H_k-I, \quad k = 1, 2, 3, \ldots \tag{2.2}$$

Thus $\underline{x}_{k+1}$ is the vector

$$\begin{aligned}\underline{x}_{k+1} &= \underline{x}_k+\alpha_k\underline{d}_k = \underline{x}_k-\alpha_k H_k\underline{\nabla} F(\underline{x}_k)\\ &= [-\underline{\nabla} Z(\underline{x}_k)-E_k\underline{\nabla} F(\underline{x}_k)]+(\alpha_k-1)\underline{d}_k,\end{aligned} \tag{2.3}$$

so one can gain some accuracy by calculating $\underline{x}_{k+1}$ by a line search along $\underline{d}_k$ from the point $[-\underline{\nabla} Z(\underline{x}_k)-E_k\underline{\nabla} F(\underline{x}_k)]$. However, these techniques alone do not avoid serious loss of information due to computer rounding errors after only a few iterations.

It is necessary to find a way of separating the main contributions from $\underline{x}_k$ to each of the next $(n-1)$ search directions, but we do not know what these directions will be. Therefore we anticipate that Ritter's [6] hypothe-

sis will hold, because then any n consecutive search directions will tend to be orthogonal. Therefore, for $k > n$, we express $\underline{x}_k$ as a linear combination of orthonormal vectors $\{\underline{\tilde{d}}_{ki}: i = 1, 2, \ldots, n\}$, where, for $1 \leqslant j \leqslant n$, the vectors $\{\underline{\tilde{d}}_{ki}: i = 1, 2, \ldots, j\}$ span the same space as the search directions $\{\underline{d}_{k-i}: i = 1, 2, \ldots, j\}$. In other words, each iteration uses an orthogonal transformation of the variables such that, for $j = 1, 2, \ldots, n$, the first j new co-ordinate directions span the same space as the j most recent search directions. All relevant terms, including the error matrix E_k, are expressed in terms of these new co-ordinates, except that we invoke the original co-ordinates whenever we require a value of $\underline{\nabla} Z(\underline{x})$, which does not cause a serious error because this gradient is $O(\|\underline{x} - \underline{x}^*\|^2)$.

The k-th iteration of the algorithm calculates the vector

$$\underline{t}_k = -\underline{\nabla} Z(\underline{x}_k) - E_k \underline{\nabla} F(\underline{x}_k) \tag{2.4}$$

using the new co-ordinates, and then, remembering equation (2.3), simple subtractions determine the coefficients $\{c_{ki}: i = 1, 2, \ldots, n\}$ of the search direction

$$\underline{d}_k = \underline{t}_k - \underline{x}_k = \sum_{i=1}^{n} c_{ki} \underline{\tilde{d}}_{ki}. \tag{2.5}$$

Thus, using the construction that is described by Powell [4], one may express the transformation from the current to the next co-ordinate system as a product of $(n-1)$ Givens rotations. These transformations are applied to E_k and to $\underline{t}_k$, and then, since $\underline{d}_k$ is the new first co-ordinate direction, the last $(n-1)$ components of $\underline{x}_{k+1}$ are the same as those of $\underline{t}_k$. The first component of $\underline{x}_{k+1}$ is calculated from the perfect line search condition $\underline{d}_k^T \underline{\nabla} F(\underline{x}_{k+1}) = 0$. Finally the iteration calculates the matrix E_{k+1} to satisfy the quasi-Newton equation

$$E_{k+1} \underline{\gamma}_k = (\underline{\delta}_k - \underline{\gamma}_k) = \underline{\nabla} Z(\underline{x}_k) - \underline{\nabla} Z(\underline{x}_{k+1}), \tag{2.6}$$

which is very straightforward, because, due to equation (1.7) and the new co-ordinates, the elements of E_k and E_{k+1} differ only in the first row and column.

This method of calculation gives good relative accuracy on a floating point computer until underflow occurs. Therefore we extended the programme to handle the mantissa and exponent of each number separately. Thus all practical limitations on the smallness of the ratio $\|\underline{x}_k\| / \|\underline{x}_1\|$ were removed, and in some experiments this ratio was reduced to less

than $10^{-1000000}$. To test the accuracy of the method the programme was run in both single (7 decimals) and double (16 decimals) precision on a TRS-80 computer. Excluding attempts to simulate the pathological case that is presented in Section 4, the largest observed discrepancy in any number was less than 5 units in the sixth decimal place.

Because the experiments were run on a micro-computer in Basic without a compiler, the number of variables was limited to $2 \leqslant n \leqslant 4$. These limited tests showed excellent support for Ritter's [6] conclusions. In all cases it was easy to continue the calculation until any n consecutive search directions were mutually orthogonal to full machine precision. Values of $\log \|\underline{x}_{k+1}\| / \log \|\underline{x}_k\|$ were displayed, and the iterative procedure was stopped when three consecutive iterations kept the first six decimal digits of this ratio unchanged. In all cases the final ratio was the root in [1,2] of the polynomial equation

$$\theta^n - \theta^{n-1} - 1 = 0, \tag{2.7}$$

which is exactly the rate of convergence that is suggested by Ritter's bound (1.12).

3. A lower bound on the rate of convergence

In this section the bound (1.13) is established under the conditions on $\{F(\underline{x}): \underline{x} \in \boldsymbol{R}^n\}$ and $\{\underline{x}_k: k = 1, 2, 3, \ldots\}$ that are stated in Section 1, assuming that all line searches are perfect. Because of the slack in some of the inequalities of our analysis, an attempt was made to improve on expression (1.13) for $n \geqslant 3$, but it was unsuccessful.

LEMMA 1. *As $k \to \infty$ the bounds*

$$|\underline{x}_{k+i}^T \underline{\delta}_k| = O(\|\underline{\delta}_k\|^2 \|\underline{\delta}_{k+1}\|) \tag{3.1}$$

and

$$\|\underline{E}_{k+i} \underline{\delta}_k\| = O(\|\underline{\delta}_k\|^2 \|\underline{\delta}_{k+1}\| / \|\underline{\delta}_{k+i}\|) \tag{3.2}$$

hold for $1 \leqslant i \leqslant n+1$, where $\underline{\delta}_k$ is defined by equation (1.5).

Proof. Equation (3.1) is satisfied when $i = 1$ because the perfect line search gives $\underline{\delta}_k^T \nabla F(\underline{x}_{k+1}) = 0$, because the conditions on $F(\cdot)$ imply $\nabla F(\underline{x}_{k+1}) = \underline{x}_{k+1} + O(\|\underline{x}_{k+1}\|^2)$, and because, due to the limit (1.8), we have $\|\underline{x}_{k+1}\| \sim \|\underline{\delta}_{k+1}\|$. Moreover, since H_k is uniformly bounded (see Dennis and Moré,

[2], for instance), and since $\underline{\gamma}_k = \underline{\delta}_k + O(\|\underline{\delta}_k\|^2)$, expression (1.7) gives the relation

$$E_{k+1} = \left(I - \frac{\underline{\delta}_k \underline{\delta}_k^T}{\|\underline{\delta}_k\|^2}\right) E_k \left(I - \frac{\underline{\delta}_k \underline{\delta}_k^T}{\|\underline{\delta}_k\|^2}\right) + O(\|\underline{\delta}_k\|). \tag{3.3}$$

Thus equation (3.2) is also true when $i = 1$. We complete the proof by induction, showing that, if the bounds (3.1) and (3.2) are satisfied, then they remain valid if i is increased by one.

Because the step-lengths $\{\alpha_k: k = 1, 2, 3, \ldots\}$ are uniformly bounded (Dennis and Moré, [2]), expressions (2.2) and (2.3) imply the relation

$$\begin{aligned} \underline{x}_{k+i+1} &= \underline{x}_{k+i} - \alpha_{k+i}[\underline{\nabla} F(\underline{x}_{k+i}) + E_{k+i} \underline{\nabla} F(\underline{x}_{k+i})] \\ &= (1 - \alpha_{k+i})\underline{x}_{k+i} - \alpha_{k+i} E_{k+i} \underline{\nabla} F(\underline{x}_{k+i}) + O(\|\underline{x}_{k+i}\|^2). \end{aligned} \tag{3.4}$$

We multiply this equation by $\underline{\delta}_k^T$ and use the inductive hypotheses to deduce that equation (3.1) remains true when i is increased by one.

To prove that equation (3.2) also remains true, we replace k by $k+i$ in expression (3.3) to deduce the bound

$$\begin{aligned} \|E_{k+i+1}\underline{\delta}_k\| &= O(\|E_{k+i}\underline{\delta}_k\|) + O(|\underline{\delta}_{k+i}^T \underline{\delta}_k| / \|\underline{\delta}_{k+i}\|) \\ &\quad + O(\|\underline{\delta}_{k+i}\| \, \|\underline{\delta}_k\|). \end{aligned} \tag{3.5}$$

Remembering the limit (1.8), it is sufficient to show that each term on the right-hand side is $O(\|\underline{\delta}_k\|^2 \|\underline{\delta}_{k+1}\| / \|\underline{\delta}_{k+i}\|)$, which is straightforward by using the inductive hypotheses, the identity $|\underline{\delta}_{k+i}^T \underline{\delta}_k| = |(\underline{x}_{k+i+1} - \underline{x}_{k+i})^T \underline{\delta}_k|$ and the result of the previous paragraph. The proof of the lemma is complete. □

THEOREM 1. *Equation* (1.13) *is implied by the conditions that are stated in Section* 1.

Proof. Consider the non-zero vectors $\{\underline{\delta}_{k+i}: i = 0, 1, \ldots, n\}$ in $\boldsymbol{R}^n$. Because there are $(n+1)$ of them, one can deduce by a continuity/compactness argument that there exists a positive constant ϱ such that the inequality

$$|\underline{\delta}_{k+i}^T \underline{\delta}_{k+j}| \geqslant \varrho \, \|\underline{\delta}_{k+i}\| \, \|\underline{\delta}_{k+j}\| \tag{3.6}$$

holds for some $0 \leqslant j < i \leqslant n$. Further, Lemma 1 and the identity $\underline{\delta}_{k+i} = \underline{x}_{k+i+1} - \underline{x}_{k+i}$ imply the relation

$$|\underline{\delta}_{k+i}^T \underline{\delta}_{k+j}| = O(\|\underline{\delta}_{k+j}\|^2 \|\underline{\delta}_{k+j+1}\|). \tag{3.7}$$

Thus we obtain the bound

$$\|\underline{\delta}_{k+i}\| = O(\|\underline{\delta}_{k+j}\| \, \|\underline{\delta}_{k+j+1}\|), \tag{3.8}$$

which is equivalent to the expression

$$\|\underline{x}_{k+i} - \underline{x}^*\| = O(\|\underline{x}_{k+j} - \underline{x}^*\| \, \|\underline{x}_{k+j+1} - \underline{x}^*\|). \tag{3.9}$$

Because $0 \leqslant j < i \leqslant n$, the theorem now follows from the limit (1.8). □

4. Examples of convergence rates when $n = 2$

The first half of this section presents an example to show that, when there are two variables, the convergence rate

$$\|\underline{x}_{k+2} - \underline{x}^*\| \sim \|\underline{x}_k - \underline{x}^*\| \, \|\underline{x}_{k+1} - \underline{x}^*\|, \tag{4.1}$$

which is suggested by the work of Sections 2 and 3, can be achieved. We let $F(\cdot)$ be the function

$$F(\underline{x}) = \tfrac{1}{2}\xi^2 + \tfrac{1}{2}\eta^2 + \tfrac{1}{3}\xi^3, \quad \underline{x} \in \mathbf{R}^2, \tag{4.2}$$

where ξ and η are the components of $\underline{x}$.

It is straightforward to deduce the relation

$$\|\underline{x}_{k+2}\| \sim \sin\theta_{k+1} \, \|\underline{x}_{k+1}\|, \tag{4.3}$$

where θ_{k+1} is the angle between $\underline{d}_{k+1}$ and $-\underline{x}_{k+1}$. Because the updating formula and the perfect line search give the equation

$$\underline{\gamma}_k^T \underline{d}_{k+1} = -\underline{\gamma}_k^T H_{k+1} \nabla F(\underline{x}_{k+1}) = \mathbf{0}, \tag{4.4}$$

we have the value

$$\begin{aligned} \sin^2\theta_{k+1} &= \frac{[(\xi_{k+1} + \xi_{k+1}^2 - \xi_k - \xi_k^2)\,\xi_{k+1} + (\eta_{k+1} - \eta_k)\,\eta_{k+1}]^2}{[(\xi_{k+1} + \xi_{k+1}^2 - \xi_k - \xi_k^2)^2 + (\eta_{k+1} - \eta_k)^2][\xi_{k+1}^2 + \eta_{k+1}^2]} \\ &\sim [\xi_k \xi_{k+1}(\xi_k - \xi_{k+1})]^2 / [\|\underline{x}_k\| \, \|\underline{x}_{k+1}\|]^2, \end{aligned} \tag{4.5}$$

where the last line depends on the perfect line search condition

$$\begin{bmatrix} \xi_{k+1} + \xi_{k+1}^2 \\ \eta_{k+1} \end{bmatrix} \perp \begin{bmatrix} \xi_{k+1} - \xi_k \\ \eta_{k+1} - \eta_k \end{bmatrix}. \tag{4.6}$$

It follows from expressions (4.3) and (4.5) that the required rate of convergence (4.1) is achieved, provided that the ratios $\{|\xi_k|/\|\underline{x}_k\|: k = 1, 2, 3, \ldots\}$ are bounded away from zero.

In order to satisfy this condition it is sufficient to ensure that the acute angle between $\pm\underline{d}_k$ and the first co-ordinate direction is in the interval $[\pi/6, \pi/3]$ for all k. Therefore we let $\xi_1 = \eta_1$, we let $\underline{d}_1 = -\underline{\nabla} F(\underline{x}_1)$, and we choose $\|\underline{x}_1\|$ to be small. Let φ_{k+1} be the angle between $\underline{d}_{k+1}$ and d_k. An argument that is similar to the derivation of expression (4.5) gives the relation

$$\begin{aligned}\cos^2\varphi_{k+1} &\sim [(\xi_{k+1}^2 - \xi_k^2)(\xi_{k+1} + \xi_{k+1}^2)]^2/[\|\underline{x}_k\| \, \|\underline{x}_{k+1}\|]^2 \\ &= O(\|\underline{x}_k\|^2). \end{aligned} \tag{4.7}$$

It follows that, by reducing $\|\underline{x}_1\|$ if necessary, we can make $\Sigma|\cos\varphi_{k+1}|$ as small as we please. Thus our example can show the convergence rate (4.1).

The other two variable example of this section is pathological. In a neighbourhood of the origin we let $F(\cdot)$ have the form

$$F(\underline{x}) = \bar{a}(\xi) + \xi\eta a(\xi) + \eta^2[\varphi(\xi) + \psi(\eta)], \tag{4.8}$$

where ξ and η are still the components of $\underline{x}$, where $a(\cdot)$ is the function

$$a(\xi) = [\xi/(1-\xi^2)] \prod_{j=0}^{\infty} [1 + (\xi^2)^{3^j}], \quad |\xi| < 1, \tag{4.9}$$

where $\bar{a}(\cdot)$ is the integral

$$\bar{a}(\xi) = \int_0^{\xi} a(\theta)\, d\theta, \quad |\xi| < 1, \tag{4.10}$$

and where $\varphi(\cdot)$ and $\psi(\cdot)$ are chosen so that, if the components of $\underline{x}$ are $\xi = \zeta^3/(1+\zeta^2)$ and $\eta = \zeta^2/(1+\zeta^2)$, ζ being any real number of small modulus, then we have the gradient

$$\underline{\nabla} F(\underline{x}) = \begin{bmatrix} \zeta^2 a(\zeta)/(1+\zeta^2) \\ \zeta a(\zeta)/(1+\zeta^2) \end{bmatrix}. \tag{4.11}$$

It can be shown that these conditions are consistent, and that they allow $F(\cdot)$ to be strictly convex and twice continuously differentiable.

We let $\underline{x}_1$ and $\underline{d}_1$ be the vectors $(\xi_1, 0)^T$ and $(-1, \xi_1)^T$, where ξ_1 is a small positive number. It follows from equation (4.11) and from the convexity of $F(\cdot)$ that $\underline{x}_2$ is the point $\left(\xi_1^3/(1+\xi_1^2),\ \xi_1^2/(1+\xi_1^2)\right)^T$. Thus, due to equation (4.4), $\underline{d}_2$ is orthogonal to the change in gradient

$$\underline{\nabla} F(\underline{x}_2) - \underline{\nabla} F(\underline{x}_1) = a(\xi_1) \begin{bmatrix} -1/(1+\xi_1^2) \\ -\xi_1^3/(1+\xi_1^2) \end{bmatrix}. \tag{4.12}$$

Hence a step along $\underline{d}_2$ from $\underline{x}_2$ can lead to the point $\underline{x}_3 = (\xi_1^3, 0)^T$, which is the point that is calculated because $\underline{\nabla} F(\underline{x}_3)$ is the vector $\left(a(\xi_1^3),\ \xi_1^3 a(\xi_1^3)\right)^T$. Thus $\underline{d}_3$ is orthogonal to the change in gradient

$$\underline{\nabla} F(\underline{x}_3) - \underline{\nabla} F(\underline{x}_2) = \begin{bmatrix} a(\xi_1^3) - \xi_1^2 a(\xi_1)/(1+\xi_1^2) \\ \xi_1^3 a(\xi_1^3) - \xi_1 a(\xi_1)/(1+\xi_1^2) \end{bmatrix}. \tag{4.13}$$

Therefore $\underline{d}_3$ is a multiple of $(-1, \xi_3)^T$ if $a(\cdot)$ satisfies the equation

$$a(\xi) = (1+\xi^2)(1-\xi^6) a(\xi^3)/[\xi^2(1-\xi^2)]. \tag{4.14}$$

Because the function (4.9) is the solution of this equation, it follows that, for all odd values of k, the variable metric algorithm calculates the points

$$\underline{x}_k = \begin{bmatrix} \xi_k \\ 0 \end{bmatrix}, \quad \underline{x}_{k+1} = \begin{bmatrix} \xi_k^3/(1+\xi_k^2) \\ \xi_k^2/(1+\xi_k^2) \end{bmatrix}, \quad \underline{x}_{k+2} = \begin{bmatrix} \xi_k^3 \\ 0 \end{bmatrix}. \tag{4.15}$$

Thus an alternating convergence pattern occurs, whose R-order is $\sqrt{3}$ but whose Q-order is only 1.5.

5. The case when $n = \infty$

It is well known that, if $F(\cdot)$ is quadratic and if $B_1 = I$, then the variable metric algorithm with perfect line searches gives the condition

$$F(\underline{x}_{k+1}) = \min_{\beta_1, \beta_2, \ldots, \beta_k} F\left(\underline{x}_1 - \sum_{j=1}^{k} \beta_j G^{j-1} \underline{\nabla} F(\underline{x}_1)\right), \quad k \geqslant 1, \tag{5.1}$$

where $G = \nabla^2 F(\cdot)$ (see Fletcher [3], for instance). This equation can allow the sequence $\{\underline{x}_k\colon k = 1, 2, 3, \ldots\}$ to be identified and its convergence properties to be studied when $n = \infty$. Thus Winther [9] proves that, if G is a compact perturbation of the identity matrix, then the superlinear convergence condition (1.8) is satisfied. However, Stoer [8] gives some

examples where the convergence rate is only linear, $\underline{x}$ being in the Hilbert space l_2, whose elements are square summable infinite sequences. Because neither of these cases is typical of finite element calculations, we consider the trivial problem of calculating the square integrable function $\{x(t); -1 \leqslant t \leqslant 1\}$ that minimizes the integral

$$F(\underline{x}) = \tfrac{1}{2} \int_{-1}^{1} (\mu+t)[x(t)]^2 dt, \tag{5.2}$$

where $\mu > 1$ is a constant. The spectrum of the second derivative operator G is the interval $[\mu-1, \mu+1]$, and we find that in this case the rate of convergence may only be linear.

We let $\underline{x}_1$ be the function $\{x_1(t) = (\mu+t)^{-1}; -1 \leqslant t \leqslant 1\}$, because then $G^{j-1}\underline{\nabla} F(\underline{x}_1)$ is a multiple of the polynomial $\{(\mu+t)^{j-1}; -1 \leqslant t \leqslant 1\}$. Thus equations (5.1) and (5.2) imply the value

$$F(\underline{x}_{k+1}) = \min_{\underline{p} \in \boldsymbol{P}_{k-1}} \tfrac{1}{2} \int_{-1}^{1} (\mu+t)[(\mu+t)^{-1} - p(t)]^2 dt, \tag{5.3}$$

where $\boldsymbol{P}_{k-1}$ is the space of algebraic polynomials of degree at most $(k-1)$. The right-hand side is least when $\underline{p}$ is the polynomial $\underline{p}_{k-1}$ that is defined by the equations

$$\int_{-1}^{1} (\mu+t)[(\mu+t)^{-1} - p_{k-1}(t)] t^l dt = 0, \quad l = 0, 1, \ldots, k-1. \tag{5.4}$$

Therefore the function

$$q_k(t) = 1 - (\mu+t)p_{k-1}(t), \quad -1 \leqslant t \leqslant 1, \tag{5.5}$$

is a multiple of the Legendre polynomial of degree k, which allows us to deduce the bound

$$\begin{aligned} F(\underline{x}_{k+1}) &= \tfrac{1}{2} \int_{-1}^{1} (\mu+t)^{-1}[q_k(t)]^2 dt \\ &> \tfrac{1}{2}(\mu+1)^{-1} \int_{-1}^{1} [q_k(t)]^2 dt \\ &> 1/[(2k+1)(\mu+1)(2\mu)^{2k}], \quad k \geqslant 1. \end{aligned} \tag{5.6}$$

One can also establish the inequality

$$F(\underline{x}_{k+1}) < 4k^2/[(2k+1)(\mu-1)(4\mu^2-4)^k]. \tag{5.7}$$

Because of the relation $F(\underline{x}) \sim \|\underline{x}-\underline{x}^*\|^2$, it follows that the R-rate of convergence of the sequence $\{\underline{x}_k: k = 1, 2, 3, \ldots\}$ is only linear, due to the infinite number of variables and to the continuous spectrum of the second derivative operator G.

6. Step-lengths of one

The computer programme that is mentioned in Section 2 was used to investigate the rate of convergence of variable metric algorithms that set $\alpha_k = 1$ for all sufficiently large k. Of course the superlinear rate of convergence (1.8) occurred, but in all cases the sequence $\{\log \|\underline{x}_{k+1}-\underline{x}^*\|/\log\|\underline{x}_k - -\underline{x}^*\|: k = 1, 2, 3, \ldots\}$ only tended to one. This section studies the special case when there are only two variables, when $F(\cdot)$ is the quadratic function

$$F(\underline{x}) = \tfrac{1}{2}\|\underline{x}\|_2^2, \quad \underline{x} \in \boldsymbol{R}^2, \tag{6.1}$$

and when the algorithm sets $\alpha_k = 1$ and uses the BFGS formula (1.6) for all k.

For $n = 2$ and $\nabla^2 F(\underline{x}^*) = I$, a fundamental difference between perfect line searches and step-lengths of one is that perfect line searches give $\underline{d}_{k+1} \perp \underline{d}_k$ as $k \to \infty$, but in the latter case $\underline{d}_{k+2}$ tends to be orthogonal to $\underline{d}_k$. Specifically, when $F(\cdot)$ is the function (6.1), equation (1.7) implies the identity

$$B_{k+1} = \left(I - \frac{\underline{d}_k \underline{d}_k^T}{\|\underline{d}_k\|^2}\right) B_k \left(I - \frac{\underline{d}_k \underline{d}_k^T}{\|\underline{d}_k\|^2}\right), \tag{6.2}$$

and, due to the unit step-length, we have the equation

$$\underline{x}_{k+2} = \underline{x}_{k+1} - (I + B_{k+1})\underline{\nabla} F(\underline{x}_{k+1}) = -B_{k+1}\underline{x}_{k+1}. \tag{6.3}$$

Thus $\underline{x}_{k+2}^T \underline{d}_k = 0$, so, because the superlinear convergence gives $\underline{x}_{k+2} = -\underline{d}_{k+2} + o(\|\underline{d}_{k+2}\|)$, we deduce the bound

$$|\underline{d}_{k+2}^T \underline{d}_k| = o(\|\underline{d}_{k+2}\| \, \|\underline{d}_k\|). \tag{6.4}$$

Let $\underline{x}_1$ and B_1 be such that, for the function (6.1), $\underline{x}_1$, $\underline{x}_2$, $\underline{x}_3$ and $\underline{x}_4$ are all non-zero, which is the usual case, and, for $k \geqslant 2$, let $\psi_k \in [0, \frac{1}{2}\pi]$ be the angle between $\underline{x}_k$ and $\pm \underline{d}_{k-1}$. Further, for $k \geqslant 2$, let 0 and λ_k be the eigenvalues of B_k, which is singular because of equation (6.2). Our analysis of the rate of convergence depends on relations between ψ_k, λ_k, ψ_{k+1} and λ_{k+1}.

We assume without loss of generality that $\underline{d}_{k-1}$ is a multiple of the first co-ordinate direction, which gives the values

$$E_k = \begin{bmatrix} 0 & 0 \\ 0 & \lambda_k \end{bmatrix} \text{ and } \underline{x}_k = \|\underline{x}_k\| \begin{bmatrix} \cos\psi_k \\ \pm\sin\psi_k \end{bmatrix}. \tag{6.5}$$

Thus it is straightforward to deduce the equations

$$\|\underline{x}_{k+1}\| = \sin\psi_k |\lambda_k| \|\underline{x}_k\|, \tag{6.6}$$

$$\sin^2\psi_{k+1} = \cos^2\psi_k/[\cos^2\psi_k + (1+\lambda_k)^2\sin^2\psi_k] \tag{6.7}$$

and

$$\lambda_{k+1} = \lambda_k \sin^2\psi_{k+1}. \tag{6.8}$$

Hence λ_2, $\sin\psi_2$ and $\cos\psi_2$ are all non-zero, because otherwise E_2 or E_3 would be zero which would give $\underline{x}_3 = 0$ or $\underline{x}_4 = 0$. It follows from equation (6.7) that ψ_k is in the open interval $(0, \frac{1}{2}\pi)$ for all k. Further, in view of the values (6.6) and (6.8), all the points $\{\underline{x}_k: k = 1, 2, 3, \ldots\}$ are non-zero.

We require the fact that $\{\lambda_k: k = 1, 2, 3, \ldots\}$ converges to zero, but equation (6.8) shows only that this sequence tends monotonically towards zero. We therefore note that, because the positive definiteness of B_2 implies $\lambda_2 > -1$, the inequality

$$0 < (2+\lambda_k)^{-1} < \tilde{c}, \quad k \geqslant 2, \tag{6.9}$$

holds, where $\tilde{c} < 1$ is a constant. Moreover, equations (6.7) and (6.8) give the identity

$$\lambda_{k+1} = \lambda_{k-1}\cos^2\psi_k \sin^2\psi_k/[\cos^2\psi_k + (1+\lambda_k)^2\sin^2\psi_k]. \tag{6.10}$$

By seeking the value of ψ_k that maximizes the modulus of this expression, we find the bound

$$|\lambda_{k+1}| \leqslant (2+\lambda_k)^{-2}|\lambda_{k-1}| < \tilde{c}^2|\lambda_{k-1}|, \quad k \geqslant 3. \tag{6.11}$$

Thus, not only does λ_k tend to zero, but also the sum $\Sigma|\lambda_k|$ is convergent. From this remark, and from the relation

$$\begin{aligned} \tan^2\psi_{k+1} &= (1+\lambda_k)^{-2}\cot^2\psi_k \\ &= (1+\lambda_k)^{-2}(1+\lambda_{k-1})^2\tan^2\psi_{k-1}, \end{aligned} \tag{6.12}$$

which is a consequence of equation (6.7), we deduce that the angles

$\{\psi_k: k = 1, 2, 3, \ldots\}$ are bounded away from zero. It follows from equation (6.8) that the conditions

$$\left.\begin{aligned} \sin\psi_k &\geqslant \hat{c} \\ |\lambda_{k+1}| &\geqslant \hat{c}^{2k-2}|\lambda_2| \end{aligned}\right\}, \quad k \geqslant 2, \tag{6.13}$$

hold, where $\hat{c}$ is a positive constant. Thus equation (6.6) implies the inequality

$$\|\underline{x}_{k+1}\| \geqslant \hat{c}^{2k-3}|\lambda_2| \ \|\underline{x}_k\|. \tag{6.14}$$

It is straightforward to deduce from expressions (1.8) and (6.14) that the ratios $\{\log\|\underline{x}_{k+1}\|/\log\|\underline{x}_k\|: k = 1, 2, 3, \ldots\}$ tend to one.

7. Discussion

Our numerical experiments suggest that, when perfect line searches are used, the rate of convergence (1.12) is obtained, but, except when $n = 2$, it is not known whether this rate can be proved under the conditions of Section 1, which are weaker than the conditions that are assumed by Ritter [6]. Equation (1.12) would imply that the R-order of convergence of the sequence $\{\|\underline{x}_k - \underline{x}^*\|: k = 1, 2, 3, \ldots\}$ is at least the root in [1, 2] of the polynomial equation (2.7), but the second example of Section 4 shows that the Q-order of convergence can be slower than the R-order. It would be interesting to find the least Q-order of convergence when $F(\cdot)$ is both infinitely differentiable and uniformly convex in a neighbourhood of $\underline{x}^*$.

It is surprising that Section 6 shows a strong deterioration in the rate of convergence when step-lengths of one are preferred to perfect line searches, because in practice it is usually highly efficient to employ the BFGS formula and to set $\alpha_k = 1$ on most iterations. Perhaps it is sometimes very valuable to replace the BFGS formula by one that gives quadratic termination without perfect line searches.

It is hoped that the procedure of Section 2 and the given properties of variable metric algorithms for unconstrained optimization will help further study of rates of convergence.

Acknowledgements

The material of Section 5 is joint work with Dr A. Griewank. Several valuable comments on the first draft of this paper were made by Josef Stoer and Yuan Ya-xiang.

References

[1] Burmeister W., Die Konvergenzordnung des Fletcher–Powell Algorithmus, *Z. Angew. Math. Mech.* **53** (1973), pp. 693–699.

[2] Dennis J. E. and Moré J. J., Quasi-Newton Methods, Motivation and Theory, *SIAM Review* **19** (1977), pp. 46–89.

[3] Fletcher R., *Practical Methods of Optimization, Vol. I: Unconstrained Optimization*, John Wiley & Sons, Chichester, 1980.

[4] Powell M. J. D., On the Calculation of Orthogonal Vectors, *The Computer Journal* **11** (1968), pp. 302–304.

[5] Powell M. J. D., Some Global Convergence Properties of a Variable Metric Algorithm for Minimization Without Exact Line Searches. In: R. W. Cottle and C. E. Lemke (eds.), *Nonlinear Programming, SIAM-AMS Proceedings Vol. IX*, American Mathematical Society, Providence, 1976.

[6] Ritter K., On the Rate of Superlinear Convergence of a Class of Variable Metric Methods, *Numer. Math.* **35** (1980), pp. 293–313.

[7] Schuller G., On the Order of Convergence of Certain Quasi-Newton Methods, *Numer. Math.* **23** (1974), pp. 181–192.

[8] Stoer J., Two Examples on the Convergence of Certain Rank-2 Minimization Methods for Quadratic Functionals in Hilbert Space, *Linear Algebra and its Applics.* **28** (1979), pp. 217–222.

[9] Winther R., Some Superlinear Convergence Results for the Conjugate Gradient Method, *SIAM J. Numer. Anal.* **17** (1980), pp. 14–18.

DEPARTMENT OF APPLIED MATHEMATICS AND THEORETICAL PHYSICS,
UNIVERSITY OF CAMBRIDGE,
SILVER STREET,
CAMBRIDGE CB3 9EW,
ENGLAND

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

DOMINIQUE FOATA

Combinatoire des identités sur les polynômes orthogonaux

L'étude combinatoire récente des identités sur les polynômes orthogonaux est passée en revue. A titre d'illustration, on établit, par des méthodes combinatoires, une extension de la formule du noyau de Poisson pour les polynômes de Meixner.

1. Introduction

L'étude combinatoire des identités sur les fonctions spéciales a été entreprise dans les dernières années par différentes écoles, bostonienne ([27], [28], [29], [35], [36], [38]), californienne ([19], [24], [25], [26], [44]), lotharingienne ([13], [14], [16], [17], [18], [22], [39], [40], [41]), québécoise ([20], [21], [31], [33]) et viennoise ([10], [11], [30]). Comme le dit fort justement notre ami Adriano Garsia [26], "les fonctions spéciales et les identités des mathématiques classiques recèlent une information abondante. Cette information s'exprime sous forme de correspondances entre structures finies qu'il s'agit de dégager. Les identités classiques apparaissent alors comme de simples relations entre ces structures comptées suivant des statistiques appropriées. Une étude systématique est en cours et a pour but de déterrer cette information de la littérature classique. Ce riche inventaire de correspondances a permis d'établir de nouvelles identités et d'obtenir aussi des démonstrations très explicites des formules classiques."

Du foisonnement des identités sur les fonctions spéciales, il n'est cependant pas facile de dégager l'essentiel de l'accessoire. Les formules importantes ont en général été motivées par des considérations analytiques ou géométriques. Par exemple, c'est en étudiant la positivité de la série bilinéaire des polynômes d'Hermite qu'on a obtenu la formule explicite de Mehler ([2], [3]). On se doit donc de regarder en premier lieu ces

identités solidement motivées. Quant à établir de nouvelles identités par des techniques combinatoires, il faut être plus prudent, s'assurer de leur esthétique ou travailler de conserve avec l'analyste.

Les formules qui se prêtent bien à un traitement combinatoire sont naturellement les identités entre séries de polynômes, comme dans la formule de Mehler, où l'on peut utiliser avec succès les interprétations combinatoires courantes sur les séries comme la somme, le produit, la substitution, l'exponentielle. On pourra trouver dans Joyal [31] un exposé élégant de l'algèbre combinatoire des séries, ou remonter à des mémoires antérieurs comme ceux de Bender et Goldman [8] sur les préfabs, de Rota et ses disciples ([35], [36]) sur le calcul ombral, et encore de Stanley [38] ainsi que de Schützenberger et l'auteur [23].

Dans le corpus des séries de polynômes, on trouve des séries dites ordinaires et des séries exponentielles. Une belle application du traitement combinatoire des séries génératrices ordinaires a été faite par Shapiro [37] qui a obtenu une formule bilinéaire pour les polynômes de Tchebychev de chacune des deux espèces.

Pour les séries génératrices exponentielles, on a pu utiliser avec profit l'identité sur la fonction exponentielle pour établir directement la formule de Mehler [17] ou la fonction génératrice des polynômes de Jacobi $P_n^{(\alpha,\beta)}(x)$ [21]. Posant $R = (1-2xu+u^2)^{1/2}$, cette dernière fonction génératrice s'écrit

$$\sum u^n P_n^{(\alpha,\beta)}(x) = 2^{\alpha+\beta} R^{-1}(1-u+R)^{-\alpha}(1+u+R)^{-\beta} \qquad (n \geqslant 0)$$

(cf. [5]). Lorsque $\alpha = \beta = \lambda - 1/2$, on en déduit une fonction génératrice pour les polynômes ultrasphériques

$$P_n^{(\lambda)}(x) = \big((2\lambda)_n/(\lambda+1/2)_n\big) P_n^{(\lambda-1/2,\lambda-1/2)}(x)$$

(cf. [42], p. 81, formule (4.7.16)), différente de la fonction génératrice usuelle $\sum u^n P^{(\lambda)}(x) = R^{-\lambda}$ (cf. [42], p. 82, formule (4.7.23)). C'est Volker Strehl ([39], [40], [41]) qui a su prolonger à la fois la méthode combinatoire développée en [21] pour établir cette toute dernière identité, et la méthode de Dumont [13] pour redémontrer un vieux résultat de Tricomi [43] dans le cas général.

Les séries ordinaires et exponentielles n'épuisent pas le sujet. Lorsqu'on s'élève dans la hiérarchie hypergéométrique des polynômes orthogonaux, les séries utilisées deviennent des *séries de faculté* comme dans la formule

(1.1) ci-après, où l'on a posé

$$(\alpha)_0 = 1, \quad (\alpha)_n = \alpha(\alpha+1)\dots(\alpha+n-1), \quad (n \geqslant 1)$$

et

$${}_pF_q(a_1, \dots, a_p; b_1, \dots, b_q; u) = \sum \frac{(a_1)_n \dots (a_p)_n}{(b_1)_n \dots (b_q)_n} \frac{u^n}{n!} \quad (n \geqslant 0).$$

Par série de faculté, on entend une série du type

$$\sum u^n / \{n!(\gamma)_n(\delta)_n\} P_n$$

où P_n est un polynôme $(n \geqslant 0)$.

Le but de cet article est tout d'abord de donner une démonstration combinatoire de l'identité suivante sur des séries de faculté

$$\sum (u^n/n!)(\beta)_n {}_2F_1(-n, -x; \gamma; a) {}_2F_1(-n, -y; \delta; b)$$
$$= (1-u)^{-\beta} \sum \frac{(\beta)_r(-x)_r(-y)_r}{(\gamma)_r(\delta)_r r!} \left(\frac{abu}{(1-u)^2}\right)^r \times$$
$$\times {}_2F_1\left(\beta+r, -x+r; \gamma+r; -\frac{au}{1-u}\right) {}_2F_1\left(\beta+r, -y+r; \delta+r; -\frac{bu}{1-u}\right)$$
$$(n \geqslant 0; r \geqslant 0), \quad (1.1)$$

tout en faisant ressortir les lemmes combinatoires sur les permutations et injections, ensuite de faire le point sur les interprétations combinatoires des polynômes orthogonaux hypergéométriques.

Notons que l'identité (1.1) n'est en fait qu'une extension de la formule du *noyau de Poisson* pour les *polynômes de Meixner* $m_n(x; \beta, c)$ définis par

$$m_n(x; \beta, c) = (\beta)_n {}_2F_1(-n, -x; \beta; 1-c^{-1}), \quad (1.2)$$

cette formule du noyau s'écrivant

$$\sum \{u^n/(\beta)_n n!\} m_n(x; \beta, c) m_n(y; \beta, c)$$
$$= (1-u/c)^{x+y}(1-u)^{-x-y-\beta} {}_2F_1\{-x, -y; \beta; u(c^{-1}-1)/(1-u/c)^2\}$$
$$(n \geqslant 0) \quad (1.3)$$

(cf. [2], p. 15, formule (2.40W), où $m_n(x; \beta, c) = (\beta)_n M_n(x; \beta, c)$). En effet, lorsque $\beta = \gamma = \delta$ et $a = b = 1-c^{-1}$, le premier membre de (1.1)

se réduit, moyennant (1.2), au premier membre de (1.3). Quant aux deux fonctions ${}_2F_1$ du second membre, elles se réduisent, par le théorème binomial à $\big((1-u/c)(1-u)^{-1}\big)^{x-r}$ et $\big((1-u/c)(1-u)^{-1}\big)^{y-r}$, respectivement. On retrouve alors le second membre de (1.3).

Comme l'a noté Richard Askey [6], on peut déduire analytiquement la formule (1.1) de la formule d'Erdélyi [15] sur les polynômes de Laguerre. Il n'y a donc pas lieu de discourir sur l'originalité d'une telle formule. En revanche, la méthode combinatoire développée est nouvelle.

2. Endofonctions de Meixner

Comme dans [20], on appelle *endofonction de Meixner* sur un ensemble fini S tout couple $\varphi = \big((A, B), f\big)$ où (A, B) est une partition ordonnée de S et f une application de S dans S dont la restriction f_A de f à la partie A est injective et la restriction f_B à B est une permutation de B. Soit $\mathrm{cyc}(f_A)$ (resp. $\mathrm{cyc}(f_B)$) le nombre de cycles de l'injection f_A (resp. la permutation f_B). Le *poids* de $\varphi = \big((A, B), f\big)$ est défini par

$$w(\gamma, -x, -a; \varphi) = \gamma^{\mathrm{cyc}(f_A)}(-x)^{\mathrm{cyc}(f_B)}(-a)^{|B|}, \tag{2.1}$$

où $|B|$ désigne naturellement le cardinal de B. Comme démontré dans [20], l'expression $(\gamma)_n {}_2F_1(-n, -x; \gamma; a)$, égale à $\sum \binom{n}{i}(-x)_i(\gamma+i)_{n-i}(-a)^i$, est le polynôme générateur des endofonctions de Meixner sur $[n] = \{1, 2, \ldots, n\}$ par le poids w. En d'autres termes, on a

$$(\gamma)_n {}_2F_1(-n, -x; \gamma; a) = \sum w(\gamma, -x, -a; \varphi), \tag{2.2}$$

où φ varie dans l'ensemble des endofonctions de Meixner sur $[n]$. On a ainsi *interprété combinatoirement* le produit $(\gamma)_n {}_2F_1(-n, -x; \gamma; a)$, c'est-à-dire, à un changement de variables près, le polynôme de Meixner $m_n(x; \beta, c)$ (cf. (1.2)).

Par ailleurs, posons $w_\beta(\sigma) = \beta^{\mathrm{cyc}(\sigma)}$ pour toute permutation σ d'un ensemble fini. L'identité

$$(\beta)_n = \sum w_\beta(\sigma), \tag{2.3}$$

où la sommation est sur l'ensemble des permutations σ de $[n]$, est bien connue (cf. [34], p. 71). Il résulte alors de (2.2) et (2.3) que l'on a

$$(\beta)_n(\gamma)_n {}_2F_1(-n, -x; \gamma; a)(\delta)_n {}_2F_1(-n, -y; \delta; b)$$
$$= \sum w_\beta(\sigma) w(\gamma, -x, -a; \varphi) w(\delta, -y, -b; \psi), \tag{2.4}$$

où la sommation s'étend à tous les triplets (σ, φ, ψ) avec σ une permutation de $[n]$ et φ, ψ deux endofonctions de Meixner sur $[n]$. Le premier membre de (1.1) peut donc s'écrire

$$\sum \left(u^n/(n!(\gamma)_n(\delta)_n)\right) \sum w_\beta(\sigma) w(\gamma, -x, -a; \varphi) w(\delta, -y, -b; \psi). \quad (2.5)$$

D'autre part, en développant les fonctions ${}_2F_1$ et en appliquant le théorème binomial, on peut mettre le second membre de (1.1) sous la forme

$$\sum \left(u^n/(n!(\gamma)_n(\delta)_n)\right) \sum \binom{n}{q,r,s,i,j} (\beta)_q(\beta)_r(\beta+r)_i(\beta+r)_j \times$$
$$\times (2r+i+j)_s(\gamma+r+i)_{n-r-i}(\delta+r+j)_{n-r-j}(-x)_{r+i}(-y)_{r+j} \times$$
$$\times (-a)^{r+i}(-b)^{r+j} \quad (q+r+s+i+j=n). \quad (2.6)$$

Pour établir (1.1) il suffit donc d'établir l'identité *polynômiale*

$$\sum w_\beta(\sigma) w(\gamma, -x, -a; \varphi) w(\delta, -y; -b; \psi)$$
$$= \sum \binom{n}{q,r,s,i,j} (\beta)_q(\beta)_r(\beta+r)_i(\beta+r)_j(2r+i+j)_s \times$$
$$\times (\gamma+r+i)_{n-r-i}(\delta+r+j)_{n-r-j}(-x)_{r+i}(-y)_{r+j} \times$$
$$\times (-a)^{r+i}(-b)^{r+j} \quad (q+r+s+i+j=n). \quad (2.7)$$

3. Lemmes combinatoires

L'identité (2.7) est beaucoup moins effrayante qu'il n'y paraît, car tous les termes ont une signification combinatoire qu'on va maintenant donner. Les trois lemmes ci-après sont extraits de l'article sur la formule d'Erdélyi–Hille–Hardy pour les polynômes de Laguerre [22]. Comme dans (2.3), si h est une injection d'un ensemble fini, on pose $w_\beta(h) = \beta^{\mathrm{cyc}(h)}$.

LEMME 3.1. *Si* $|A| = i$, $|B| = j$ *et* $i + j = n$, *alors*

$$\sum w_\beta(h) = (\beta+j)_i, \quad (3.1)$$

la sommation étant sur toutes les injections h de A dans $A+B$.

LEMME 3.2. *Soit* (I, J, R) *une partition ordonnée d'un ensemble telle que* $|I| = i$, $|J| = j$, $|R| = r$. *Alors*

$$\sum w_\beta(\theta) = (\beta)_r(\beta+r)_i(\beta+r)_j, \tag{3.2}$$

la sommation étant étendue à l'ensemble des permutations θ *de* $I+J+R$ *satisfaisant à* $\theta(J)\cap I = \emptyset$.

Soit (I, J, R, S) une partition ordonnée d'un ensemble telle que $|I| = i$, $|J| = j$, $|R| = r$, $|S| = s$. Trois sortes de chemins dont les sommets sont pris dans $I+J+R+S$ sont maintenant introduits, les *a-chemins*, les *b-chemins* et les *ab-chemins*. Les a-chemins (resp. b-chemins) ont tous leurs sommets dans S à l'exception de l'extrémité qui est dans I (resp. J). Un ab-chemin a aussi tous ses sommets dans S, à l'exception d'un seul, qui appartient à R et ne se trouve pas nécessairement à l'extrémité. Dans la Figure 1, effaçons par la pensée toutes les flèches en pointillé et écartons le cycle en trait continu. Il ne reste alors que des a-, b- et ab-chemins.

Un graphe G dont les sommets sont les éléments de $I+J+R+S$ est dit *Erdélyien sur* (I, J, R, S) si ses parties connexes ne sont composées que de a-, b- et ab-chemins.

Il sera commode de noter $R'(G)$ l'ensemble des sommets de G qui sont les extrémités des ab-chemins. On a

$$R'(G) \subset R+S \quad \text{et} \quad |R'(G)| = |R|. \tag{3.3}$$

LEMME 3.3. *Si* $|I| = i$, $|J| = j$, $|R| = r$, $|S| = s$, *le nombre de graphes Erdélyiens sur* (I, J, R, S) *est égal à* $(i+j+2r)_s$.

La démonstration de l'identité (2.7) consiste alors à associer, de façon bijective, à chaque triplet (σ, φ, ψ) de la sommation du premier membre, une partition ordonnée (Q, R, S, I, J) de $[n]$ et une suite $(\sigma', \theta, G, h, h', \xi, \xi')$ ayant les propriétés:

(i) σ' *est une permutation de* Q;
(ii) θ *est une permutation de* $R+I+J$ *satisfaisant à* $\theta(J)\cap I = \emptyset$;
(iii) G *est un graphe Erdélyien sur* (I, J, R, S);
(iv) (*resp.* (v)) h (*resp.* h') *est une injection de* $[n]\setminus(R'(G)+I)$ (*resp.* $[n]\setminus(R+J)$) *dans* $[n]$;
(vi) (*resp.* (vii)) ξ (*resp.* ξ') *est une permutation de* $R'(G)+I$ (*resp.* $R+J$).

De plus, l'identité suivante doit être verifiée:

$$w_\beta(\sigma)w(\gamma,-a;\varphi)w(\delta,-y,-b;\psi)$$
$$=(-a)^{|R+I|}(-b)^{|R+J|}w_\beta(\sigma')w_\beta(\theta)w_\gamma(h)w_\delta(h')w_{-x}(\xi)w_{-y}(\xi'). \quad (3.4)$$

Compte-tenu des lemmes 3.1, 3.2 et 3.3 et de (3.3), on voit que si une telle bijection est établie, la sommation du second membre de (3.4) donne bien le second membre de (2.7). Reste donc à établir la bijection annoncée.

4. La correspondance

Partons d'un triplet (σ, φ, ψ) avec σ une permutation de $[n]$, et $\varphi = \big((A,B),f\big)$, $\psi = \big((C,D),g\big)$ deux endofonctions de Meixner sur $[n]$. Quand on superpose les graphes de ces trois configurations sur un ensemble de n sommets étiquetés, on a d'abord les cycles de σ — appelons-les *β-cycles* — puis les chemins et cycles des endofonctions $\varphi = \big((A,B),f\big)$ et $\psi = \big((C,D),g\big)$. Les n sommets se répartissent donc en quatre classes $A\cap C$, $A\cap D$, $B\cap C$ et $B\cap D$. D'après l'expression du poids donnée en (2.1), on peut considérer que les sommets de B (resp. D) portent la marque $-a$ (resp. $-b$). On dira qu'un sommet est *a-marqué*, *b-marqué* ou *ab-marqué*, suivant qu'il appartient à $B\cap C$, $A\cap D$, ou $B\cap D$. Les sommets dans $A\cap C$ sont *non marqués*.

Deux sommets distincts v et v' sont dits *liés* si les trois propriétés suivantes sont satisfaites:

(i) *v est b-marqué et v' est a-marqué*;
(ii) *v et v' sont dans le même β-cycle*;
(iii) *les sommets appartenant à ce β-cycle et situés entre v et v' sont tous non marqués.*

La partition ordonnée (Q,R,S,I,J) associée à (σ,φ,ψ) est ainsi définie: si les sommets d'un même β-cycle sont tous non marqués, tous ces sommets sont rangés dans la classe Q. Si un sommet est, ou bien *ab*-marqué, ou bien *b*-marqué *et* lié, il est mis dans R. Si un sommet est *a*-marqué (resp. *b*-marqué) et *non* lié, il va dans I (resp. J). Enfin, S se compose de tous les sommets restants. Notons que S englobe aussi les sommets *a*-marqués et liés.

Dans la Figure 1, on a représenté les β-cycles d'un triplet (σ, φ, ψ) avec les sommets marqués a, b ou ab. Les sommets non marqués apparaissent comme de simples points. L'appartenance de chaque sommet à un bloc de la partition (Q, R, S, I, J) est indiquée par la lettre correspondante. Les flèches en pointillé sont les arcs ayant pour origine les sommets qui sont, ou bien ab-marqués, ou bien a-marqués, ou encore b-marqués mais non liés.

Effaçons les flèches en pointillé du graphe de la Figure 1. On obtient, d'une part, une collection de cycles dont tous les sommets sont dans Q — une *permutation* σ' de Q — d'autre part, une collection de a-, b- et

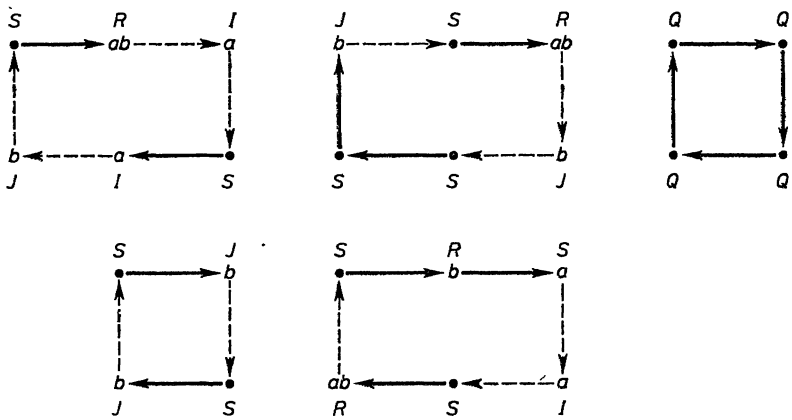


Fig. 1

ab-chemins formant un *graphe Erdélyien* G *sur* (I, J, R, S). De plus, comme chaque a- (resp., b-, resp. ab-) chemin contient exactement un sommet dans I (resp. J, resp. R), les arcs en pointillé sont la représentation graphique d'une *permutation* θ de $I+J+R$. Enfin, comme il n'y a pas de flèche en pointillé issue de sommet b-marqué et lié, la permutation θ satisfait à $\theta(J) \cap I = \emptyset$. Evidemment $\mathrm{cyc}(\sigma) = \mathrm{cyc}(\sigma') + \mathrm{cyc}(\theta)$, d'où

$$w_\beta(\sigma) = w_\beta(\sigma') w_\beta(\theta). \tag{4.3}$$

Les trois premiers éléments de la suite $(\sigma', \theta, G, h, h', \xi, \xi')$ ayant été définis, les quatre derniers sont simplement $h = f_A$, $h' = g_C$, $\xi = f_B$ et $\xi' = g_D$, c'est-à-dire, respectivement, les restrictions de f à A, de g à C,

de f à B et de g à D. Comme on a

$$B = R'(G)+I \quad \text{et} \quad D = R+J, \tag{4.1}$$

ainsi que

$$A = [n]\setminus\big(R'(G)+I\big) \quad \text{et} \quad C = [n]\setminus(R+J), \tag{4.2}$$

les conditions (iv) et (v) de la section 3 sont bien vérifiées. De plus, d'après (2.1), (3.3), (4.1) et (4.2), il vient

$$\begin{aligned} w(\gamma, -x, -a; \varphi) &= \gamma^{\mathrm{cyc}(f_A)}(-x)^{\mathrm{cyc}(f_B)}(-a)^{|B|} \\ &= \gamma^{\mathrm{cyc}(h)}(-x)^{\mathrm{cyc}(\xi)}(-a)^{|R'(G)+I|} \\ &= w_\gamma(h)w_{-x}(\xi)(-a)^{|R+I|} \end{aligned} \tag{4.4}$$

et de même

$$w(\delta, -y, -b; \psi) = w_\delta(h')w_{-y}(\xi')(-b)^{|R+J|}. \tag{4.5}$$

Prenant en compte (4.3), (4.4) et (4.5), on obtient bien (3.4).

Réciproquement, si l'on part d'une partition ordonnée (Q, R, S, I, J) de $[n]$ et d'une suite $(\sigma', \theta, G, h, h', \xi, \xi')$ ayant les propriétés (i)–(vii) de la section 3, il est immédiat de reconstruire le triplet (σ, φ, ψ). Les trois éléments σ', θ, G fournissent la permutation σ, et les couples (h, ξ) et (h', ξ') les endofonctions de Meixner φ et ψ, respectivement.

Ceci achève la démonstration de l'identité (1.1).

5. Conclusion

Un beau guide des polynômes orthogonaux hypergéométriques nous est proposé par Askey et Wilson [7] qui les ont classés dans un diagramme respectant leur hiérarchie hypergéométrique. Nous reproduisons une *partie* de celui-ci dans la Figure 2. Une flèche va du polynôme P au polynôme Q, si l'expression analytique de Q peut être obtenue de celle de P par une spécialisation des paramètres ou un passage à la limite approprié. Par exemple, la flèche allant du polynôme de Laguerre $L_n^{(\alpha)}(x)$ au polynôme d'Hermite $H_n(x)$ symbolise le passage à la limite

$$H_n(x) = n!\lim(2/\alpha)^{n/2} L_n^{(\alpha)}\big(\alpha - x(2\alpha)^{1/2}\big) \quad (\alpha \to +\infty) \tag{5.1}$$

(cf. [42], p. 389, [4]).

Les interprétations combinatoires des polynômes apparaissant dans ce diagramme sont connues et sont compatibles, en ce sens que toutes les formules de passage ont des démonstrations simples dans la géométrie de ces modèles. Par exemple, la formule (5.1) a une signification géométrique intéressante donnée par Strehl [41].

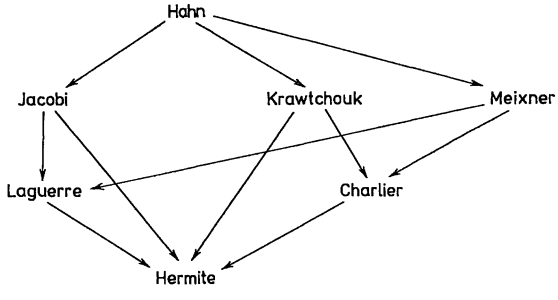


Fig. 2

D'après Karlin–McGregor [32], le polynôme de Hahn a l'expression analytique

$$Q_n(x;\alpha,\beta,N) = \sum \frac{(-n)_k(-x)_k(n+\alpha+\beta+1)_k}{(\alpha+1)_k(-N+1)_k k!} \qquad (0 \leqslant k \leqslant n).$$

Posant $R_n = (\alpha+1)_n(-N+1)_n Q_n(x;\alpha,\beta,N)$, on obtient

$$R_n = \sum \binom{n}{i} (-1)^i (\alpha+1+i)_{n-i}(\beta+1+n-i)_i(-x)_i(-N+1+x)_{n-i}$$

$(0 \leqslant i \leqslant n)$.

Par le lemme 3.1, on peut donc immédiatement obtenir une interprétation combinatoire des polynômes R_n. Cette interprétation n'a cependant pas fourni des démonstrations vraiment nouvelles des formules concernant les polynômes de Hahn. En revanche, de l'étude combinatoire des R_n, on peut déduire la géométrie combinatoire de tous les polynômes apparaissant dans le diagramme.

Si on se reporte à [2], chap. 2, on constatera que les noyaux de Poisson ont été calculés par des méthodes combinatoires pour les polynômes de Meixner (ici-même), de Laguerre [22] et Hermite [17].

Signalons enfin que cette étude combinatoire des polynômes orthogonaux a aussi pour ambition de traiter les q-polynômes. La principale difficulté vient que souvent plusieurs polynômes peuvent prétendre au titre de q-analogues des polynômes classiques. Il y en a déjà deux familles pour les seuls q-polynômes d'Hermite, comme l'a montré Désarménien [12]. Signalons, en revanche, la belle étude des q-polynômes de Laguerre par Garsia et Remmel [26] et une démonstration très élégante du q-analogue de la formule de Pfaff–Saalschütz par Andrews et Bressoud [1]. Il semble que pour les q-polynômes il faille opérer dans l'algèbre des partitions et non plus dans celle des endofonctions. Un modèle global reste à découvrir, peut-être celui des groupes formels comme une récente étude de Cartier [9] le laisse prévoir.

Références

[1] Andrews G. E. and Bressoud D. M., *Identities in Combinatorics III: Further Aspects of Ordered Set Sorting*, Pennsylvania State Univ., Math. 83003, 1983.

[2] Askey R., *Orthogonal Polynomials and Special Functions*, Regional Conference Series in Appl. Math. 21, S. I. A. M., Philadelphia, 1975.

[3] Askey R., Orthogonal polynomials and positivity, in: *Wave Propagation and Special Functions*, Studies in Appl. Math. **6** (D. Ludwig and F. W. J. Olver, Eds.), S. I. A. M., Philadelphia, 1970, pp. 64–85.

[4] Askey R., *Math. Reviews* **80b**: 33005.

[5] Askey R., Jacobi's generating function for Jacobi polynomials, *Proc. Amer. Math. Soc.* **71** (1978), pp. 243–246.

[6] Askey R., Communication privée.

[7] Askey R. and Wilson J., Some basic hypergeometric orthogonal polynomials that generalize Jacobi polynomials, *Amer. Math. Soc. Memoirs*, Providence, R. I., 1983 (à paraître).

[8] Bender E. A. and Goldman J. R., Enumerative uses of generating functions, *Indiana Univ. Math. J.* **20** (1971), pp. 753–765.

[9] Cartier P., Groupes formels, représentation des groupes symétriques et congruence de Kummer, *Amer. J. Math.* (à paraître).

[10] Cigler J., Operatormethoden für q-Identitäten II: q-Laguerre Polynome, *Mh. Math.* **91** (1981), pp. 105–117.

[11] Cigler J., Operatormethoden für q-Identitäten III: Umbrale Inversionen und die Lagrange'sche Formel, *Arch. Math.* **35** (1980), pp. 533–543.

[12] Désarménien J., Les q-analogues des polynômes d'Hermite, *Séminaire Lotharingien de Combinatoire*, Publ. IRMA Strasbourg, 1982, 191/S-05, pp. 39–56.

[13] Dumont D., Une approche combinatoire des fonctions elliptiques de Jacobi, *Adv. in Math.* **41** (1981), pp. 1–39.

[14] Dumont D., Étude combinatoire d'une suite de polynômes associés aux polynômes ultrasphériques, *Univ. Beograd Publ. Elektrotehn. Fak. Ser. Mat. Fiz.* **634–677** (1979), pp. 116–125.

[15] Erdélyi A., Transformationen einer gewissen nach Produkten konfluenter hypergeometrischer Funktionen fortschreitenden Reihe, *Compositio Math.* **6** (1939), pp. 336–347.
[16] Flajolet Ph., Combinatorial aspects of continued fractions, *Discrete Math.* **32** (1980), pp. 125–161.
[17] Foata D., A combinatorial proof of the Mehler formula, *J. Comb. Theory, Ser. A* **24** (1978), pp. 367–376.
[18] Foata D., Some Hermite polynomial identities and their combinatorics, *Adv. Appl. Math.* **2** (1981), pp. 250–259.
[19] Foata D. and Garsia A. M., A combinatorial approach to the Mehler formulas for the Hermite polynomials, in: *Relations between Combinatorics and other parts of Mathematics*, Proc. of Symposia in Pure Math. **34**; (D. K. Ray-Chaudhuri, Ed.), Amer. Math. Soc., Providence, R. I., 1978, pp. 163–179.
[20] Foata D. and Labelle J., Modèles combinatoires pour les polynômes de Meixner, *Europ. J. Comb.* (à paraître).
[21] Foata D. and Leroux P., Polynômes de Jacobi, interprétation combinatoire et fonction génératrice, *Proc. Amer. Math. Soc.* **87** (1983), pp. 47–53.
[22] Foata D. and Strehl V., Combinatorics of Laguerre polynomials, *Proc. Waterloo Silver Jubilee*, 1983, à paraître.
[23] Foata D. and Schützenberger M. P., *Théorie géométrique des polynômes Eulériens*, Lecture Notes in Math., vol. **136**, Springer-Verlag, Berlin, 1970.
[24] Garsia A. M., A q-analogue of the Lagrange inversion formula, *Houston J. Math.* **7** (1981), pp. 205–237.
[25] Garsia A. M. and Joni S. A., A new expression for umbral operators and power series inversion, *Proc. Amer. Math. Soc.* **64** (1977), pp. 179–185.
[26] Garsia A. M. and Remmel J., A combinatorial interpretation of q-derangement and q-Laguerre numbers, *Europ. J. Comb.* **1** (1980), pp. 47–59.
[27] Gessel I., A factorization for formal Laurent series and lattice path enumeration, *J. Comb. Theory, Ser. A* **28** (1980), pp. 321–337.
[28] Gessel I., A noncommutative generalization and a q-analog of the Lagrange inversion formula, *Trans. Amer. Math. Soc.* **257** (1980), pp. 455–482.
[29] Gessel I. and Stanton D., Strange evaluations of hypergeometric series, *SIAM J. Math. Anal.* **13** (1982), pp. 295–308.
[30] Hofbauer J., A short proof of the Lagrange–Good formula, *Discrete Math.* **25** (1979), pp. 135–139.
[31] Joyal A., Une théorie combinatoire des séries formelles, *Adv. in Math.* **42** (1981), pp. 1–82.
[32] Karlin S. and McGregor J. L., The Hahn polynomials, formulas and an application, *Scripta Math.* **26** (1961), pp. 33–46.
[33] Labelle G., Une nouvelle démonstration combinatoire des formules d'inversion de Lagrange, *Adv. in Math.* **42** (1981), pp. 217–247.
[34] Riordan J., *An Introduction to Combinatorial Analysis*, J. Wiley, New York, 1958.
[35] Roman S., and Rota G.-C., The umbral calculus, *Adv. in Math.* **27** (1978), pp. 95–188.
[36] Rota G.-C., Kahaner D., and Odlyzko A., Finite operator calculus, *J. Math. Anal. Appl.* **42** (1973), pp. 685–760.
[37] Shapiro L. W., A combinatorial proof of a Chebyshev polynomial identity, *Discrete Math.* **34** (1981), pp. 203–206.

[38] Stanley R. P., Generating functions, in: *Studies in Combinatorics*, (G.-C. Rota, Ed.), Math. Ass. Amer. Studies in Math. **17** (1978), pp. 100–141.

[39] Strehl V., *Kombinatorischer Beweis und Erweiterung einer Identität von Tricomi*, Jahrestagung der D. M. V., Bayreuth, 1982.

[40] Strehl V., *Combinatorial aspects of Gegenbauer polynomials* (en préparation).

[41] Strehl V., *Contribution to the combinatorics of some families of classical orthogonal polynomials*, unpublished manuscript, Erlangen, 1982.

[42] Szegö G., *Orthogonal Polynomials*, 4th ed., 2nd printing, Colloquium Publ., vol. 23, Amer. Math., Soc., Providence, R. I., 1978 (1st ed.: 1939).

[43] Tricomi F., Sul comportamento asintotico dei polinomi di Laguerra, *Ann. Mat. Pura Appl.* **28** (1949), pp. 263–289.

[44] Viennot G., Une interprétation combinatoire des développements en série entière des fonctions elliptiques de Jacobi, *J. Comb. Theory, Ser. A* **29** (1980), pp. 121–133.

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

R. L. GRAHAM

Recent Developments in Ramsey Theory

Introduction

Mathematics has often been called the science of order. From this viewpoint the guiding principle of Ramsey theory is perhaps best summed up by the statement of T. S. Motzkin: "Complete disorder is impossible". Ramsey theory is basically the study of structure preserved under partitions. Before stating some background material, we first introduce the following notation. We will adopt the usual convention of identifying the positive integer n with the set of its predecessors $\{0, 1, \ldots, n-1\}$, where 0 corresponds to $\emptyset$. The symbol ω denotes $\{0, 1, 2, \ldots\}$, the set of natural numbers. For $X \subseteq \omega$, $k \in \omega$, $[X]^k$ denotes the set of k-element subsets of X, and $[X]^\omega$ denotes the set of infinite subsets of X (if there are any). The generic result in Ramsey theory is due (not surprisingly) to F. P. Ramsey [49]:

Ramsey's Theorem (1930)

For any $k, r \in \omega$, if $[\omega]^k = C_1 \cup \ldots \cup C_r$ then there exists $X \in [\omega]^\omega$ such that $[X]^k \subseteq C_i$ for some i.

An earlier result of Ramsey type was given by I. Schur [52] in 1916:

If $\omega = C_1 \cup \ldots \cup C_r$ then there exist $x, y, z \in C_i$ for some i such that $x + y = z$.

The result of Schur was generalized successively as follows.

THEOREM (Rado [47], Folkman [17], Sanders [51]). *For all $m \in \omega$, if $\omega = C_1 \cup \ldots \cup C_r$ then there exists $X \in [\omega]^m$ such that for some i and all nonempty $F \subseteq X$, $\sum_{a \in F} a \in C_i$.*

THEOREM (Hindman [34]). *If* $\omega = C_1 \cup \ldots \cup C_r$ *then there exists* $X \in [\omega]^\omega$ *such that for some* i *and all finite nonempty* $F \subseteq X$, $\sum_{a \in F} a \in C_i$.

A much weaker form of the Rado–Folkman–Sanders theorem was actually given by Hilbert in 1892:

THEOREM [33]. *For all* $m \in \omega$, *if* $\omega = C_1 \cup \ldots \cup C_r$ *then there exists* $X \in [\omega]^\omega$ *and* $t \in \omega$ *such that for some* i *and all nonempty* $F \subseteq X$, $t + \sum_{a \in F} a \in C_i$.

Finally, we mention the result which will motivate much of what we discuss in this paper. This is:

VAN DER WAERDEN'S THEOREM (1927) [63]. *If* $\omega = C_1 \cup \ldots \cup C_r$ *then for some* i, C_i *contains arbitrarily long arithmetic progressions.*

The theorem of van der Waerden has proved to be an extremely fertile seed from which a major part of modern combinatorics has developed, especially through the work of Rado [47], [48], Erdös [16], [15], Roth [50], Szemerédi [60], [61], Deuber [9] and many others (see [14], [31], [10]). A particularly important generalization was given in 1963 by Hales and Jewett. For a fixed finite set A, call a subset $L \subseteq A^N$ a *combinatorial line* if for some nonempty $I \subseteq N$, L can be written as

$$L = L_I = \bigcup_{a \in A} \{(x_0, x_1, \ldots, x_{N-1}) : x_i = a \text{ if } i \in I \text{ and } x_i = b_i \in A \text{ if } i \notin I\}.$$

Thus, $|L| = |A|$.

HALES–JEWETT THEOREM [32]. *For all finite* A *and* r, *there exists* $N(A, r)$ *such that if* $N \geqslant N(A, r)$ *and* $A^N = C_1 \cup \ldots \cup C_r$ *then some* C_i *must contain a combinatorial line.*

To see that this implies van der Waerden's Theorem, simply take $A = t = \{0, 1, \ldots, t-1\}$ and identify the point $\bar{x} = (x_0, \ldots, x_{N-1}) \in A^N$ with the integer $|\bar{x}| = \sum_{i \in N} x_i t$. The t points in any combinatorial line clearly correspond to t integers in an arithmetic progression. Since t was arbitrary, a standard compactness argument yields van der Waerden's Theorem.

The Hales–Jewett Theorem also implies the higher-dimensional analogues of van der Waerden's Theorem, first proved by Gallai (see [47]) and Witt [66].

THEOREM. *If* $\omega^n = C_1 \cup \ldots \cup C_r$ *then some* C_i *must contain for all* $k \in \omega$ *a homothetic copy of* $\{0, 1, \ldots, k-1\}^n$, *i.e., all* k^n *points*

$$\{(ai_1 + b_1, ai_2 + b_2, \ldots, ai_n + b_n) : 0 \leqslant i_1, \ldots, i_n < k\}$$

for suitable $a, b_i \in \omega$.

A much stronger "density" form of van der Waerden's theorem was conjectured by Erdös and Turán [16] nearly 50 years ago: If $A \subseteq \omega$ satisfies

$$\limsup_{n\to\infty} \frac{|A \cap n|}{n} > 0 \qquad (*)$$

then A contains arbitrarily long arithmetic progressions.

It was shown by Roth [50] in 1953 that (*) implies A has a 3-term arithmetic progression and by Szemerédi [60] in 1969 that (*) implies A has a 4-term arithmetic progression. Finally, Szemerédi [61] in 1974 in a brilliant combinatorial tour de force established the full conjecture. Szemerédi's Theorem and the higher-dimensional density analogues of van der Waerden's Theorem have fairly recently been proved by quite different techniques from ergodic theory and topological dynamics. This exciting work of Furstenberg, Katznelson, Weiss and others (see [22], [24], [20], [21]) has furnished a very stimulating link between these two branches of mathematics which is just beginning to reveal its full potential.

It is very natural to ask whether there is a corresponding *density* version for the Hales–Jewett Theorem. We can phrase this as follows:

CONJECTURE.[1] *For all finite* A *and* $\varepsilon > 0$ *there exists* $N(A, \varepsilon)$ *such that if* $N \geqslant N(A, \varepsilon)$ *and* $R \subseteq A^N$ *satisfies* $|R| \geqslant \varepsilon |A^N|$ *then* R *must contain a combinatorial line.*

The conjecture, if true, clearly implies Szemerédi's Theorem. It is known to be true if $|A| = 2$ by the following argument. Assume without loss of generality that $A = \{0, 1\}$. Identify with each point $\bar{x} = (x_0, x_1, \ldots, x_{N-1}) \in A^N$ the subset $S(\bar{x}) \subseteq N$ by $i \in S(\bar{x})$ iff $x_i = 1$ (i.e., $\bar{x}$ is the characterstic function for $S(\bar{x})$). Thus, a combinatorial line in A^N corresponds to a pair of distinct subset $X, Y \subseteq N$ with $X \subset Y$. However, a well-known result of Sperner [59] asserts that any family $\boldsymbol{F}$ of subsets of N in which $X, Y \in \boldsymbol{F}, X \neq Y$ implies $X \not\subset Y$ can have cardinality at most

$$\binom{N}{[N/2]} \sim \left(\frac{2}{\pi N}\right)^{1/2} \cdot 2^N.$$

Thus, for ε fixed, if N is sufficiently large then $(2/\pi N)^{1/2} < \varepsilon$ and the assertion follows.

If A is taken to be the finite field $GF(3)$, then Brown and Buhler [5] have recently shown that any subset R of the affine space A^N having at

[1] The author is currently offering US $1000 for a proof or disproof of this conjecture.

least $\varepsilon \cdot 3^N$ points must contain an *affine* line, provided $N \geqslant N(\varepsilon)$. (Combinatorial lines correspond to very special kinds of affine lines.) More generally, Furstenberg and Katznelson have now proved (unpublished) the following weakened form of the conjecture. Let us write $A = \{a_0, a_1, \ldots$ $\ldots, a_{t-1}\}$. Call a set $\tilde{L}$ of t points of A^N a *twisted* combinatorial line if for some nonempty $I \subseteq N$ and $d_i \in t$, $i \in I$, $\tilde{L}$ can be written as

$$\tilde{L} = \bigcup_{j \in t} \{(x_0, x_1, \ldots, x_{N-1}) \colon x_i = a_{j+d_i} \text{ if } i \in I \text{ and } x_i = b_i \in A \text{ if } i \notin I\}$$

where index addition is modulo t.

Thus, in a twisted line, the entries in each of the coordinates which vary have been cyclically permuted.

THEOREM [23]. *For all finite A and $\varepsilon > 0$ there exists $\tilde{N}(A, \varepsilon)$ such that if $N \geqslant \tilde{N}(A, \varepsilon)$ and $R \subseteq A^N$ satisfies $|R| \geqslant \varepsilon |A^N|$ then R must contain a twisted combinatorial line.*

This result implies as a corollary the fact that any subset $R \subseteq GF(q)^N$ with $|R| \geqslant \varepsilon q^N$ always contains an affine line, provided N is sufficiently large (as a function of q and ε).

Partitions into infinitely many classes

If we allow partitions of ω of the form $\omega = \bigcup_{i \in \omega} C_i$ then it is clear that the conclusion of van der Waerden's Theorem does not have to hold. For example, we could take $C_i = \{i\}$. However, in this case we have arbitrarily long arithmetic progressions which hit each C_i in *at most one* element. The following result of Erdös and Graham shows that one of these two possibilities must always occur.

THEOREM [14], [11]. *If $\omega = \bigcup_{i \in \omega} C_i$ then either some C_i contains arbitrarily long arithmetic progressions or there are arbitrarily long arithmetic progressions hitting each C_i in at most one element.*

The idea behind the proof is basically this. If some C_i has positive upper density then by Szemerédi's Theorem, C_i has the desired progressions. If not, then for N large the number of arithmetic progressions which have at least two elements in a single C_i is $o(N^2)$. Since there are at least $c_k N^2$ arithmetic progressions of length k for a fixed $c_k > 0$, the desired conclusion follows.

This result is an example of a so-called "canonical" partition theorem, first introduced by Erdös and Rado for Ramsey's Theorem [15]. Other

theorems of this type have recently been given by Baumgartner [2], Taylor [62], Voigt [64], [46] and others. One of the most striking theorems of this type is the canonical partition theorem for the n-dimensional analogues of van der Waerden's theorem. As an illustration of the increased range of behavior the canonical partitions can have, consider the case $n = k = 2$. Suppose $\omega^2 = \bigcup_{i\in\omega} C_i$. Let us say that $(x, y) \sim (x', y')$ if (x, y) and (x', y') belong to the same C_i. Consider the following six partitions:

(i) $(x, y) \sim (x', y')$ iff $(x, y) = (x', y')$,
(ii) $(x, y) \sim (x', y')$ for all $(x, y), (x', y') \in \omega^2$,
(iii) $(x, y) \sim (x', y')$ iff $x = x'$,
(iv) $(x, y) \sim (x', y')$ iff $y = y'$.
(v) $(x, y) \sim (x', y')$ iff $x+y = x'+y'$,
(vi) $(x, y) \sim (x', y')$ iff $x-y = x'-y'$.

In Figure 1, we show the six different possibilities for the four vertices of a square in ω^2 (where, $\alpha, \beta, \ldots$ denote distinct classes).

α. β.	α. α.	α. β.
γ. δ.	α. α.	α. β.
(i)	(ii)	(iii)
α. α.	α. β.	α. β.
β. β.	γ. α.	β. γ.
(iv)	(v)	(vi)

Fig. 1. The six canonical partitions of $\{0, 1\}^2$.

It follows from the following theorem that these are a *complete* set of canonical partitions, i.e., in *any* partition $\omega^2 = \bigcup_{i\in\omega} C_i$ at least one of these patterns must occur.

THEOREM (Deuber, Graham, Prömel, Voigt [11]). *All canonical partitions of ω^n are given as follows: For a subspace $V \subseteq \boldsymbol{R}^n$ over $\boldsymbol{R}$, partition $\boldsymbol{R}^n$ into disjoint translates of V by*

$$\boldsymbol{R}^n = \bigcup_{a\in A} (V+a).$$

This induces a partition of $\omega^n = \bigcup_{\beta\in B} C_\beta$ (where B is countable). These partitions form a complete set of canonical partitions of ω^n.

We remark that the only proof known for this result requires the use of the deep Furstenberg–Katznelson density version of the Gallai–Witt Theorem.

Hales–Jewett revisited

In order to describe the next series of results we will first recast the Hales–Jewett Theorem into a different format. As usual, we fix a finite set A and assume $A \cap \omega = \emptyset$. For $X \subseteq \omega$, $k \in \omega$, we let $(X)^k$ denote the set of *partitions* of X into k nonempty blocks. Furthermore, we let $(X)^k_A$ denote the set of partitions of $X \cup A$ into $k + |A|$ nonempty blocks so that each block contains at most one element of A. Such partitions will be called *A-partitions* of $X \cup A$. Finally, if $Y \in (X)^k_A$ and $m \leqslant k$ then $(Y)^m_A$ denotes the set of A-partitions Z of $X \cup A$ having $m + |A|$ blocks such that every block of Y is contained in a block of Z. Thus, Y is a *refinement* of Z.

The theorem of Hales and Jewett can be restated as follows:

THEOREM. *For all finite A and r if $N \geqslant N(A, r)$ and $(N)^0_A = C_1 \cup \ldots \cup C_r$ then there exists $X \in (N)^1_A$ such that $(X)^0_A \subseteq C_1$ for some i.*

This was generalized by Graham and Rothschild in 1971:

THEOREM [29]. *For all finite A and $k, m, r \in \omega$, $m \leqslant k$, there exists $N(A, k, m, r)$ such that if $N \geqslant N(A, k, m, r)$ and $(N)^m_A = C_1 \cup \ldots \cup C_r$ then there exists $X \in (N)^k_A$ such that $(X)^m_A \subseteq C_i$ for some i.*

A very beautiful generalization of this has now just been proved by Carlson and Simpson. It deals with *infinite* partitions of ω. To state the result we first introduce the following topology on $(\omega)^\omega$, the partitions of ω having *infinitely* many blocks. Any partition $X \in (\omega)^\omega$ induces an equivalence relation on $\omega \times \omega$ by having $x, y \in \omega$ equivalent iff they belong to the same block. The set of all binary relations $2^{\omega \times \omega}$ can be endowed with the usual product topology, where each factor has the discrete topology. In this way $(\omega)^\omega$ becomes a topological space under the topology inherited from $2^{\omega \times \omega}$. The following result can in a certain sense be considered a *dual* to the usual Ramsey Theorem.

CARLSON–SIMPSON THEOREM [7]. *For any $k \in \omega$, if $(\omega)^\omega = C_1 \cup \ldots \cup C_r$ where each C_i is Borel then there exists $X \in (\omega)^\omega$ such that $(X)^k \subseteq C_i$ for some i.*

Carlson and Simpson in fact prove the stronger analogous result for A-partitions of $(\omega)^\omega_A$ and which can properly be considered as an infinite generalization of the Graham–Rothschild Theorem. It should be pointed out that *some* condition on the C_i is necessary since otherwise a counterexample for $(\omega)^2 = C_1 \cup C_2$ can be easily constructed using transfinite induction.

In addition to the preceding results, dual forms are proved in [7] for the Galvin–Prikry extension [25] of Ramsey's Theorem for the case of infinite subsets of ω, as well as for Ellentuck's generalization [13] of it, but space limitations prevent us from discussing them further.

In another direction, Carlson (see [44]) has very recently obtained a beautiful theorem which unifies a large number of known Ramsey-type theorems, both finite and infinite. Again, space restrictions do not allow us to give a full description of this striking achievement here. However, we will now describe a key ingredient used in the proof, which is of significant interest in its own right.

To begin with, for a fixed finite set A and a variable $v \in A$, denote by $W(v)$ the set of all "variable words" of A, i.e., the set of all finite strings $a_0, a_1, \ldots, a_m$ where $a_i \in A \cup \{v\}$ and $a_j = v$ for at least one index j. For $a \in A$ and $w(v) \in W(v)$ we can form the string $w(a)$ by simply replacing each occurrence of v in $w(v)$ by a (i.e., we just "evaluate" $w(v)$ at a). Let $S = S(A, v)$ denote the set of all infinite sequences $\bar{s} = (s_0(v), s_1(v), \ldots)$ where $s_i(v) \in W(v)$. By a *v-reduction* of $\bar{s}$ we mean any sequence $\bar{t} = (t_0(v), t_1(v), \ldots)$ formed from $\bar{s}$ in the following way. For each $i \in \omega$, $s_i(v)$ is replaced by $s_i(b_i)$ where $b_i \in A \cup \{v\}$. Disjoint blocks of consecutive $s_i(b_i)$'s are then concatenated, forming a sequence of strings $\bar{t} = (t_0(v), t_1(v), \ldots)$, where the symbol v must still occur at least once in each $t_i(v)$, (thus, $\bar{t} \in S$). Denote by $R(\bar{s})$ the set of all v-reductions of $\bar{s}$ and by $R_0(\bar{s})$ the set of all $t_0(v)$ for $\bar{t} = (t_0(v), t_1(v), \ldots) \in R(\bar{s})$.

MAIN LEMMA (Carlson). *For any $\bar{s} \in S$, if $R_0(\bar{s}) = C_1 \cup \ldots \cup C_r$ then there exists $\bar{t} \in R(\bar{s})$ such that $R_0(\bar{t}) \subseteq C_i$ for some i.*

This deceptively simple looking statement conceals much of its inherent strength. As a simple application, we derive Hindman's Theorem (following [6]). Let $\omega = C_1 \cup \ldots \cup C_r$ be given. Choose $A = \{0\}$ and partition $W(v) = C_1^* \cup \ldots \cup C_r^*$ by defining: $w(v) \in C_i^*$ iff $w(v)$ has m v's occurring in it and $m \in C_i$. Applying Carlson's result for $\bar{s} = (v, v, \ldots)$ we are guaranteed the existence of $\bar{t} = (t_0(v), t_1(v), \ldots) \in R_0(\bar{s}) = W(v)$ with $R_0(\bar{t}) \subseteq C_i^*$ for some i. For any finite subset $J \subseteq N$, the word $w_J(v) = t_0(b_0)t_1(b_1)\ldots t_{N-1}(b_{N-1}) \in C_i^*$ where

$$b_i = \begin{cases} v & \text{if } j \in J, \\ 0 & \text{if } j \notin J. \end{cases}$$

Thus, if n_i denotes the number of v's occurring in $t_i(v)$ then this implies $\sum_{j \in J} n_j \in C_i$ for all finite J, which is just Hindman's Theorem.

We should note here that Voigt [65] has very recently independently also obtained infinite generalizations of the Hales–Jewett and Graham–Rothschild Theorems which are similar to, though somewhat weaker than, Carlson's Main Lemma. His proofs however are more combinatorial in nature whereas Carlson relies on the intricate use of idempotent ultrafilter arguments (which are often quite effective for problems of this type, e.g., see [35]).

Van der Waerden again

The finite form of van der Waerden's Theorem (for two classes) asserts the following: For all $k \in \omega$, there exists a least $W(k)$ so that if $\{1, 2, \ldots, W(k)\} = C_1 \cup C_2$ then some C_i must contain a k-term arithmetic progression.

The determination of the values and, in fact, even the growth rate of $W(k)$ has proved to be extremely frustrating for combinatorialists. The known exact values are listed in Table 1.

Table 1

k	1	2	3	4	5	6
$w(k)$	1	3	9	35	178	?

The best lower bound known is due to Berlekamp [4]:

$$w(k+1) > k \cdot 2^k \quad \text{if } k \text{ is a prime power.}$$

There is currently no known upper bound for $W(k)$ which is primitive recursive. This is because all available proofs leading to upper bounds involve at some point a (perhaps intrinsic) *double* induction, with k as one of the variables. This leads naturally to rapidly growing functions like the Ackermann function which may help to explain the enormous gap in our knowledge here. The possibility that $W(k)$ might in fact actually have this Ackermann-like growth has been strengthened by the work of Paris and Harrington [43], Ketonen and Solovay [37], and more recently Friedman [18], who show that some natural combinatorial questions do indeed have *lower* bounds which grow this rapidly (and even much more rapidly, e.g., see [54], [55]). In spite of this potential evidence to the contrary, I am willing to make the following:

CONJECTURE.

$$W(k) \leqslant 2^{\cdot^{\cdot^{\cdot^{2}}}}$$

for $k \geqslant 1$, where the number of 2's is k.

It should be pointed out that while any partition of the set $\{1, 2, \ldots, 9\} = C_1 \cup C_2$ always results in some C_i containing a 3-term arithmetic progression (and this is true for any set homothetic to $\{1, 2, \ldots, 9\}$), other sets also have this property, e.g., $\{1, 3, 4, 5, 6, 7, 8, 9, 11\}$. However, it can be shown [53] that no 8-element set has the property. In general, define

$$W^*(k) = \min\{|X|\colon X \subseteq \omega, X = C_1 \cup C_2 \Rightarrow \text{some } C_i \text{ contains a } k\text{-term arithmetic progression}\}.$$

Thus,

$$W^*(3) = W(3) = 9$$

and, in general,

$$W^*(k) \leqslant W(k).$$

It turns out perhaps unexpectedly that $W^*(k)$ can be strictly smaller than $W(k)$. In particular, recent computations have yielded $W^*(4) \leqslant 27$, compared to $W(4) = 35$. The characteristic function of a set which achieves the bound of 27 is given by:

$$100100110111111111111111111111011001001.$$

It would be of great interest to know if $W^*(k)$ is in general significantly smaller than $W(k)$, e.g., does

$$W^*(k)/W(k) \to 0 \quad \text{as } k \to \infty?$$

As an abbreviation, let us write $X \to AP(k)$ to denote the fact that for any partition of $X = C_1 \cup C_2$, some C_i contains a k-term arithmetic progression. Going in the other direction from $W^*(k)$, one might naturally ask whether there exist *arbitrarily large* sets $X(k)$ with the properties:

(i) $X(k) \to AP(k)$;

(ii) $Y \not\to AP(k)$ for any proper subset $Y \subset X(k)$.

In fact, the existence of arbitrarily large "critical" sets for both k-term arithmetic progressions as well as more general combinatorial lines in A^N has just recently been established by Graham and Nešetřil [28]. From this work, it appears that even the structure of sets $X(3)$ which satisfy (i) and (ii) for $k = 3$ can be exceedingly complex.

Concluding remarks

As mentioned earlier, we did not have the opportunity here to give more than a brief sketch of a few of the large number of exciting recent developments in Ramsey theory. The interested reader will find more of these developments reported in the following references: [35], [57], [41], [27], [8], [1], [36], [42], [46], [38].

Finally, I remark that essentially no progress has occurred on the following (by now) old conjecture of Erdös on arithmetic progressions, which would imply Szemerédi's theorem and for which Erdös currently offers US $ 3000:

CONJECTURE. *If* $A \subseteq \omega$ *and* $\sum_{a \in A} 1/a = \infty$ *then* A *contains arbitrarily long arithmetic progressions.*

A related perhaps easier conjecture is this:

CONJECTURE. *If* $A \subseteq \omega^2$ *and* $\sum_{(i,j) \in A} 1/(i^2+j^2) = \infty$ *then* A *contains the 4 vertices of a square.*

References

[1] Ajtai M., Komlós J., and Szemerédi E., A Note on Ramsey Numbers, *J. Comb. Th.* (*A*) **29** (1980), pp. 354–360.

[2] Baumgartner J., Canonical Partition Relations, *J. Symbolic Logic* **40** (1975), pp. 541–554.

[3] Beck J., On Size Ramsey Numbers of Paths, Trees and Circuits, I, *J. Graph Theory* **7** (1983), pp. 115–135.

[4] Berlekamp E., A Construction for Partitions Avoiding Long Arithmetic Progressions, *Canad. Math. Bull.* **11** (1968), pp. 409–414.

[5] Brown T. C. and Buhler J. P., A Density Version of a Geometric Ramsey Theorem, *J. Comb. Th.* (*A*) **32** (1982), pp. 20–34.

[6] Carlson T. (personal communication).

[7] Carlson T. and Simpson S., A *Dual Form of Ramsey's Theorem*, Dept. of Math. Research Report, The Penn. State Univ. (1983).

[8] Chung F. R. K. and Grinstead C. M., A Survey of Bounds for Classical Ramsey Numbers, *J. Graph Theory* **7** (1983), pp. 25–37.

[9] Deuber W., Partitionen und lineare Gleichungssysteme, *Math. Zeit.* **133** (1973), pp. 109–123.
[10] Deuber W. and Voigt B., Der Satz von van der Waerden über arithmetische Progressionen, *Jahresberichte d. Dt. Math. Verein.* **85** (1983), pp. 66–85.
[11] Deuber W., Graham R. L., Prömel H. J., and Voigt B., A Canonical Partition Theorem for Equivalence Relations on Z^t, *J. Comb. Th.* (*A*) **34** (1983), pp. 331–339.
[12] Deuber W., Rothschild B. L., and Voigt B., Induced Partition Theorems, *J. Comb. Th.* (*A*) **32** (1982), pp. 225–240.
[13] Ellentuck E., A New Proof that Analytic Sets are Ramsey, *J. Symbolic Logic* **39** (1974), pp. 163–165.
[14] Erdös P. and Graham R. L., *Old and New Problems and Results in Combinatorial Number Theory*, Monographie 28, L'Enseignement Mathématique, Genève, 1980.
[15] Erdös P. and Rado R., A Combinatorial Theorem, *J. London Math. Soc.* **25** (1950), pp. 249–255.
[16] Erdös P. and Turán P., On Some Sequences of Integers, *J. London Math. Soc.* **11** (1936), pp. 261–264.
[17] Folkman J. H. (personal communication — 1965).
[18] Friedman H., *Strong undecidable Π^0_2 statements* (in preparation).
[19] Fürstenberg H., Ergodic Behaviour of Diagonal Measures and a Theorem of Szemerédi on Arithmetic Progressions, *J. d'Analyse Math.* **31** (1977), pp. 204–256.
[20] Fürstenberg H., *Recurrence in Ergodic Theory and Combinatorial Number Theory*, Princeton University, Princeton, 1981.
[21] Fürstenberg H., Poincaré Recurrence and Number Theory, *Bull. Amer. Math. Soc.* (new series) **5** (1981), pp. 211–234.
[22] Fürstenberg H. and Katznelson Y., An Ergodic Szemerédi Theorem for Commuting Transformations, *J. d'Analyse Math.* **34** (1978), pp. 275–291.
[23] Fürstenberg H. and Katznelson Y. (unpublished).
[24] Fürstenberg H. and Weiss B., Topological Dynamics and Combinatorial Number Theory, *J. d'Analyse Math.* **34** (1978), pp. 61–85.
[25] Galvin F. and Prikry K., Borel Sets and Ramsey's Theorem, *J. Symbolic Logic* **38** (1973), pp. 193–198.
[26] Graham R. L., *Rudiments of Ramsey Theory*, CBMS Regional Conference Series in Math., Vol. 45, Amer. Math. Soc., Providence, 1981.
[27] Graham R. L., Euclidean Ramsey Theorems on the n-Sphere, *J. Graph Theory* **7** (1983), pp. 105–114.
[28] Graham R. L. and Nešetřil J., *Large Minimal Sets which Force Long Arithmetic Progressions* (to appear).
[29] Graham R. L. and Rothschild B. L., Ramsey's Theorem for n-Parameter Sets, *Trans. Amer. Math. Soc.* **159** (1971), pp. 257–292.
[30] Graham R. L. and Rothschild B. L., A Short Proof of van der Waerden's Theorem on Arithmetic progressions, *Proc. Amer. Math. Soc.* **42** (1974), pp. 385–386.
[31] Graham R. L. Rothschild B. L., and Spencer J. H., *Ramsey Theory*, John Wiley & Sons, Inc., New York, 1980.
[32] Hales A. W. and Jewett R. I., Regularity and Positional Games, *Trans. Amer. Math. Soc.* **106** (1963), pp. 222–229.
[33] Hilbert D., Über die Irreduzibilität ganzer rationaler Functionen mit ganzzahligen Koeffizienten, *J. reine angew. Math.* **110** (1892), pp. 104–129.

[34] Hindman N., Finite Sums from Sequences within Cells of a Partition of N, *J. Comb. Th.* (*A*) **17** (1974), pp. 1–11.
[35] Hindman N., *Ultrafilters and Combinatorial Number Theory*, Lecture Notes in Mathematics **751**, Springer-Verlag, Berlin, 1979.
[36] Special issue of *J. Graph Theory* **7** (1983), devoted to Ramsey theory.
[37] Ketonen J. and Solovay R., Rapidly Growing Ramsey Functions, *Annals of Math.* **113** (1981), pp. 267–314.
[38] Leeb K. and Prömel H. J., *Ordering Subsets of Finite Subsets*, preprint.
[39] Lothaire M., Combinatorics on Words, *Encyclopedia of Math.* **17**, Addison-Wesley, New York, 1982.
[40] Nešetřil J. and Rödl V., Van der Waerden's Theorem for Sequences of Integers not Containing an Arithmetic Progression of k Terms, *Comm. Math Univ. Carolinae* **17** (1976), pp. 675–688.
[41] Nešetřil J. and Rödl V., Partition Theory and Its Applications. In: B. Bollobás (ed.), *Proc. 7th British Comb. Conf.*, London Math. Soc. Lecture Notes Series **38** (1979), pp. 96–157.
[42] Nešetřil J. and Rödl V., Ramsey Classes of Set Systems, *J. Comb. Th.* (*A*) **34** (1983) pp. 183–201.
[43] Paris J. and Harrington L., A Mathematical Incompleteness in Peano Arithmetic, In: J. Barwise (ed.), *Handbook of Mathematical Logic*, North-Holland, Amsterdam, 1977, pp. 1133–1142.
[44] Prikry K., Unpublished notes describing T. Carlson's results.
[45] Prömel H. J., *Induzierte Partitionssätze*, Dissertation, Univ. Bielefeld, 1982.
[46] Prömel H. J. and Voigt B., *Canonical Partition Theorems for Parameter-Sets*, to appear.
[47] Rado R., Studien zur Kombinatorik, *Math. Zeit.* **36** (1933), pp. 424–480.
[48] Rado R., Some Recent Results in Combinatorial Analysis, Congrès International des Mathématiciens, Oslo, 1936.
[49] Ramsey F. P., On a Problem of Formal Logic, *Proc. London Math. Soc.* **30** (1930), pp. 264–286.
[50] Roth K. F., On Certain Sets of Integers, *J. London Math. Soc.* **28** (1953), pp. 104–109.
[51] Sanders J., *A Generalization of Schur's Theorem*, Ph.D. Dissertation, Dept. of Math., Yale University, 1969.
[52] Schur I., Über die Kongruenz $x^m + y^m = z^m \pmod p$, *Jahresbericht der D. Math. Verein* **25** (1916), pp. 114–117.
[53] Shor P. (personal communication).
[54] Smorynski C., Some Rapidly Growing Functions, *Math. Intelligencer* **2** (1980), pp. 149–154.
[55] Smorynski C., "Big" News from Archimedes to Friedman, *Notices Amer. Math. Soc.* **30** (1983), pp. 251–256.
[56] Spencer J. H., Restricted Ramsey Configurations, *J. Comb. Th.* (*A*) **19** (1975), pp. 278–286.
[57] Spencer J. H., Ramsey's Theorem for Spaces, *Trans. AMS* **249** (1979), pp. 363–371.
[58] Spencer J. H., Canonical Configurations, *J. Comb. Th.* (*A*) **34** (1983), pp. 325–330.
[59] Sperner E., Ein Satz über Untermengen einer endlichen Menge, *Math. Zeit.* **27** (1928), pp. 544–548.

[60] Szemerédi E., On Sets of Integers Containing no Four Elements in Arithmetic Progression, *Acta Math. Acad. Sci. Hungar.* **20** (1969), pp. 89–104.
[61] Szemerédi E., On Sets of Integers Containing no k Elements in Arithmetic Progression, *Acta Arith.* **27** (1975), pp. 199–245.
[62] Taylor A. D., A Canonical Partition Relation for Finite Subspaces of ω, *J. Comb. Th.* (*A*) **21** (1976), pp. 137–146.
[63] van de Waerden B. L., Beweis einer Baudet'schen Vermutung, *Nieuw Arch. Wisk.* **15** (1927), pp. 212–216.
[64] Voigt B., *Canonizing Partition Theorems*, preprint.
[65] Voigt B., *Parameter-Words, Trees and Vector Spaces*, preprint.
[66] Witt E., Ein kombinatorischer Satz der Elementargeometrie, *Math. Nachrichten* **6** (1951), pp. 261–262.

BELL LABORATORIES
MURRAY HILL, N.J., USA

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

L. G. KHACHIYAN

Convexity and Complexity in Polynomial Programming

The problems to be considered in the talk are the problems of convex polynomial programming: minimize

$$f_0(x_1, \ldots, x_n) \tag{1}$$

subject to

$$\begin{aligned} &f_1(x_1, \ldots, x_n) \leqslant 0, \\ &\ldots\ldots\ldots\ldots \\ &f_m(x_1, \ldots, x_n) \leqslant 0, \end{aligned} \tag{2}$$

where $f_0, f_1, \ldots, f_m$ are convex polynomials in $\boldsymbol{R}^n$ with integer coefficients, the polynomials being specified by blocks of their coefficients written in binary numerical system. Two basic cases studied in mathematical programming are the following:

(i) the variables are real $x = (x_1, \ldots, x_n) \in \boldsymbol{R}^n$,
(ii) the variables are integer $x = (x_1, \ldots, x_n) \in \boldsymbol{Z}^n$,

the problems of latter type being also called *diophantine*. However, the mixed case $x \in \boldsymbol{R}^k \times \boldsymbol{Z}^{n-k}$ is sometimes also considered.

The *degree* d of the problem is the maximum of the degrees d_i of the polynomials, the *height* h of the problem is the maximum modulus of the integer coefficients of the polynomials occurring in it, and the input *length* L of the problem is the number of binary symbols 0 and 1 needed for its coding. The reader should not be misled by identifying L with the sum of binary lengths of all non-zero entries of the problem, though, in fact, the only property of L needed below is: $L \geqslant \max\{n, m, \log h\}$.

1. Bounds on solutions

In order to consider the bounds on solutions of problems (1)–(2) with real and/or integer variables simultaneously, we give the following

DEFINITION. A set $\mathcal{N} \subseteq \boldsymbol{R}^n$ is called *periodic* if for any integer vector $y \in \boldsymbol{Z}^n$ we have $\mathcal{N} + y = \mathcal{N}$.

We assume in the present section that the vector $(x_1, \ldots, x_n)$ of unknowns in problem (1)–(2) runs over some periodic set $\mathcal{N}$. In particular, the variables may be real ($\mathcal{N} = \boldsymbol{R}^n$), integer ($\mathcal{N} = \boldsymbol{Z}^n$), real and integer ($\mathcal{N} = \boldsymbol{R}^k \times \boldsymbol{Z}^{n-k}$), rational ($\mathcal{N} = \boldsymbol{Q}^n$), etc.

As usual, a system of inequalities (2) is said to be *consistent* in $\mathcal{N}$ if it has a solution $x^* \in \mathcal{N}$. Similarly, an optimization problem (1)–(2) is said to be *feasible* in $\mathcal{N}$ it if has an optimal solution $x^* \in \mathcal{N}$ when the vector of variables runs over $\mathcal{N}$.

Note. It can be shown that a problem of convex polynomial programming (1)–(2) with integer coefficients is feasible in $\boldsymbol{R}^k \times \boldsymbol{Z}^{n-k}$ if and only if its system of constraints (2) is consistent in $\boldsymbol{R}^k \times \boldsymbol{Z}^{n-k}$ and its objective polynomial (1) is bounded from below on the set of real solutions of the system (2).

To state the results of this section, we also need the concept of *multidegree* D of a system of inequalities (2). Let $a = \min\{n, m\}$ and let the inequalities of the system be ordered by decreasing degrees $d_1 \geqslant d_2 \geqslant \ldots \geqslant d_m$. Then $D = \prod_{i=1}^{a} d_i$. In other words, the multidegree is the maximum of all possible products of the degrees of a distinct inequalities of the system. In particular, $D \leqslant d^{\min\{n,m\}}$, and adjoining further linear constraints to the system does not change its multidegree.

THEOREM 1 [11]. *Let $\mathcal{N}$ be an arbitrary periodic set in $\boldsymbol{R}^n$. If the system of convex polynomial inequalities* (2) *of degree at most d, $d \geqslant 2$, of multidegree D and of height h is consistent in $\mathcal{N}$, then it has a solution $x^* \in \mathcal{N}$ in the Euclidean ball*

$$\|x\| \leqslant (hdn)^{Dd^{3d/2}n^{d/2}}.$$

THEOREM 2. *Let $\mathcal{N}$ be an arbitrary periodic set in $\boldsymbol{R}^n$. If the problem of convex polynomial programming* (1)–(2) *of degree at most d, $d \geqslant 2$, and of height h is feasible in $\mathcal{N}$, then it has an optimal solution $x^* \in \mathcal{N}$ in the*

Euclidean ball

$$\|x\| \leqslant (hdn)^{D^2 d^{3d} n^{d-1}},$$

where D is the multidegree of the system of constraints.

We conclude the section with a brief comment on Theorems 1 and 2. To begin with, consider the problems of linear programming, which are of the degree not exceeding 2 and of multidegree 1. By Theorems 1 and 2, bounds on solutions of such problems in periodic sets are exponential in the input length, i.e., $\|x\| \leqslant A^{n \log hn} \leqslant 2^{p(L)}$, where A is a constant and p is a polynomial. As regards problems of convex polynomial programming of an arbitrary but fixed degree $d \geqslant 2$, we see that unlike the linear case, Theorems 1 and 2 restrict the bounds on solutions of such problems by a *two-stage* exponential function in the input length, i.e.,

$$\|x\| \leqslant A^{d^{2\min\{n,m\}}} A^{n^{d-1} \log hn} \leqslant 2^{2^{p(L)}},$$

p and A being a polynomial and a constant depending on d. On the other hand, the example of the system $x_1 \geqslant h$, $x_2 \geqslant x_1^d$, ..., $x_n \geqslant x_{n-1}^d$, with d even, shows that such super-exponential growth of the solution bounds can really be attained in any periodic set.

2. Finding an exact real and/or integer solution

In the present section we assume that the variables are real and/or integer $x \in \boldsymbol{R}^k \times \boldsymbol{Z}^{n-k}$. As the coefficients of the problem (1)–(2) are integer, in case of its feasibility there exists an optimal solution $x^* = (x_1^*, \ldots, x_n^*) \in \boldsymbol{A}^k \times \boldsymbol{Z}^{n-k}$, each real component of which is an algebraic number. If we agree to code algebraic numbers x_j^* by their (irreducible over $\boldsymbol{Q}$) algebraic equations $p_j(x_j^*) = 0$ and, if necessary, by rational segments $x_j^* \in (a_j, b_j)$ not containing other roots of univariate polynomials $p_j \in \boldsymbol{Z}[\cdot]$, we may consider *exact* algorithms of polynomial programming. For a problem (1)–(2) such an algorithm must check its feasibility and, if the problem is feasible, print an algebraic solution $x^* \in \boldsymbol{A}^k \times \boldsymbol{Z}^{n-k}$. Applying Theorem 2 jointly with the decision procedure [1], we obtain

THEOREM 3. *There exists an exact algorithm for convex polynomial programming with real and/or integer variables, whose running time t is bounded by a two-stage exponential function in the input length, $t \leqslant 2^{2^{p(L)}}$.*

In the subsequent sections we shall consider less burdensome problems of convex polynomial programming, allowing an essential reduction of their computational complexity.

3. Regularization of systems of convex polynomial inequalities

As we have mentioned in Section 1, the bounds on solutions of systems of convex polynomial inequalities in periodic sets can grow as $2^{2^{p(L)}}$. The following result [11] shows that in some "computational sense" these bounds can be effectively reduced to $2^{p(L)}$.

THEOREM 4. *Consider systems of convex polynomial inequalities* (2) *of a fixed degree* $d \geqslant 2$ *in some periodic set* $\mathcal{N}$. *There exists a polynomial-time in the input length* L *algorithm, called the 'regularization' algorithm, which for any given system* (2) *finds a subsystem*

$$f_i(x) \leqslant 0, \quad i \in \mathcal{M} \subseteq \{1, 2, \ldots, m\} \tag{3}$$

and a sequence of integer vectors $y_1, \ldots, y_r \in \mathbb{Z}^n$ *and natural scalars* $\delta_1, \ldots, \delta_r \in \mathbb{Z}_+$ *written in binary number system,* $r \leqslant \min\{n, m\}+1$, *such that the following conditions hold.*

(i) *The system* (2) *is consistent in* $\mathcal{N}$ *if and only if the subsystem* (3) *is consistent in* $\mathcal{N}$.

(ii) *If the subsystem* (3) *is consistent in* $\mathcal{N}$, *then it has a solution* $x^0 \in \mathcal{N}$ *in the Euclidean ball*

$$\|x\| \leqslant (hdn)^{\frac{1}{2}d^{3d/2}n^{d/2}} \leqslant 2^{p(L)}. \tag{4}$$

(iii) *If a solution* $x^0 \in \mathcal{N}$ *of subsystem* (3) *in the ball* (4) *is known, then a solution* $x^* \in \mathcal{N}$ *of the initial system* (2) *can easily be found from the formula*

$$x^* = y_r \cdot 2^{\delta_r} + \ldots + y_1 \cdot 2^{\delta_1} + x^0. \tag{5}$$

Moreover, the binary representation of x^* *can be obtained from the binary representations of* $x^0, y_1, \ldots, y_r, \delta_1, \ldots, \delta_r$ *in the form*

$$\begin{array}{ccccc} & \delta_r\text{-th place} & & \delta_1\text{-th place} & \\ & \downarrow & & \downarrow & \\ \{x^*\} = & \{y_r\}00\ldots00 & \{y_{r-1}\}00\ldots00 & \{y_1\}00\ldots00 & \{x^0\} \end{array} \tag{6}$$

where { } *stands for the binary representation of the coresponding vectors.*

Note 1. The running time of the regularization algorithm is bounded by an (absolute) polynomial in dn^d, m, $\log h$, the algorithm being independent of $\mathcal{N}$.

Note 2. If all the parameters in the r.h.s. of (5) are written in binary numerical system, we call the representation (5) *binary-exponential.* Thus, we see from (6) that the use of the binary-exponential representation of

solution instead of the usual binary one enables us to contract the output information by omitting in its record a number of 0's exponential in L.

Note 3. If a vector x^0 from the ball (4) satisfies the subsystem (3) with an accuracy $\varepsilon \in (0, 1)$

$$f_i(x^0) \leqslant \varepsilon, \quad i \in \mathcal{M}, \tag{3'}$$

then the vector x^* obtained by (5) satisfies the initial system (2) with the *same* accuracy ε.

4. Complexity of systems of convex diophantine inequalities

Applying Theorem 4 in the diophantine case $\mathcal{N} = \boldsymbol{Z}^n$, we obtain the following results.

THEOREM 5 [11]. *For a fixed $d \geqslant 1$ the problem of determining the consistency of systems of convex diophantine inequalities of degree at most d belongs to the class NP.*

Note. The problem of determining the consistency in $\boldsymbol{Z}^n$ of a *single* convex quadratic inequality is *NP*-complete.

THEOREM 6. *If it is permitted to print integral solutions of nonlinear systems in binary-exponential form, then for any fixed $d \geqslant 2$ the problem of determining the consistency and of finding a solution of systems of convex diophantine inequalities of degree at most d is polynomially transformable to the same problem for systems of linear diophantine inequalities. In particular, these problems can be solved for sure in exponential time $t \leqslant 2^{p(L)}$, the latter assertion holding even for binary representation of the output.*

All in all, to solve systems of convex diophantine inequalities is not much harder than to solve systems of linear diophantine inequalities.

5. The ellipsoid method

Suppose that for a feasible in $\boldsymbol{R}^n$ problem of convex polynomial programming (1)–(2) a bound B is known such that the problem has some optimal real solution in the ball $\|x\| \leqslant B$. Then, to solve the problem with an accuracy $\varepsilon \in (0, 1)$, i.e., to find its ε-solution $\hat{x}$

$$f_0(\hat{x}) \leqslant f^* + \varepsilon, \tag{1ε}$$

$$f_i(\hat{x}) \leqslant \varepsilon, \quad i \in \{1, 2, \ldots, m\} \tag{2ε}$$

where f^* is the minimum value of the objective polynomial, the *ellipsoid method* ([12], [8]) may be used. The following result estimates the com-

plexity of the ellipsoid method for convex polynomial programming, taking into account finite precision of arithmetical operations performed over binary numbers in a digital computer.

THEOREM 6. *To find an ε-solution $\hat{x}$ of a feasible in R^n problem* (1)–(2) *in the ball* $\|x\| \leqslant B$ *it suffices to perform*

$$k \leqslant \tfrac{3}{2} n^2 [d^2 N + (d+1) M + 7n^2] \log_2 \left(\frac{8 d^2 h N B^d}{\varepsilon} \right) \tag{7}$$

elementary operations $+$, $-$, $\times$, $/$, $\sqrt{\ }$, max *over numbers having in binary form*

$$l \leqslant \tfrac{7}{2} \log_2 (d^2 h N B^d n^2 / \varepsilon) + 30 \tag{8}$$

places, the operations being carried out approximately with the same number of digits as in the binary representation of the numbers. Here n, h, d are the number of unknowns, the height, and the degree of the problem, respectively, and N, M are the maximum and the total number of non-zero coefficients (monomials) of the polynomials $f_0, f_1, \ldots, f_m$. In addition to the input information, the storage of $n^2 + 6n$ such l-place numbers is also needed.

Note 1. For linear programming problems ($d = 1$, $N \leqslant n+1$) the estimates (7)–(8) can be improved:

$$k \leqslant n^2 [3M + 10.5 n^2] \log_2 (8 h n B / \varepsilon),$$

$$l \leqslant \tfrac{1}{2} \log_2 (h^7 n^{13} B^5 / \varepsilon^5) + 30.$$

Note 2. For problems of convex quadratic programming with convex quadratic and/or linear constraints $d = 2$, $N \leqslant (n+2)^2/2$ the estimates (7)–(8) yield

$$k \leqslant n^2 [4.5M + 13.5 (n+2)^2] \log_2 (32 h n^2 B^2 / \varepsilon),$$

$$l \leqslant \tfrac{7}{2} \log_2 (h n^4 B^2 / \varepsilon) + 37.$$

COROLLARY. *For problems of convex polynomial programming of an arbitrary fixed degree d the ellipsoid method runs in polynomial time with respect to L and* $\log(B/\varepsilon)$.

6. Finding an approximate real solution of systems of convex polynomial inequalities

From Theorem 4, Note 3 in Section 3 and the last corollary follows

THEOREM 7. *There exists a polynomial-time with respect to input L and* $\log(1/\varepsilon)$ *algorithm for finding an ε-solution of systems of convex polyno-*

mial inequalities consistent in R^n, of an arbitrary fixed degree d, provided that the output is printed in binary-exponential form.

Note 1. If one insists on binary representation of the output, it is impossible to design such an algorithm allowing an enormous length of the output information. Thus, for $d \geqslant 2$, the binary-exponential representation of the output is essential for the validity of Theorem 7.

Note 2. The existence of a similar algorithm for systems of linear inequalities aggravated by a single non-convex quadratic constraint would imply $P = NP$.

We now turn attention to some problems of convex polynomial programming which are *exactly* solvable in polynomial time.

7. Polynomial solvability of linear programming

It is clear that any feasible in R^n problem of linear programming has a rational optimal solution $x^* \in Q^n$ — recall that only problems with integer coefficients are considered. Thus, in accordance with the definition from Section 2, an exact algorithm for linear programming must check the feasibility of an l.p. problem and find its optimal rational solution. An exact algorithm for linear programming, polynomial in L, was announced in [2] and described in [3]. Later on it was improved in [4].

THEOREM 8 [4]. *There exists an exact algorithm for linear programming requiring*

$$k \lesssim \alpha^3 \beta \log \Delta\alpha \lesssim \alpha^4 \beta \log h\alpha$$

elementary operations $+, -, \times, /, \max$ *over*

$$l \lesssim \log \Delta\alpha \lesssim \alpha \log h\alpha$$

place binary numbers and additional to the input information storage of $\lesssim \alpha^2$ *such numbers. Here* $\alpha = \min\{n, m\}$ *and* $\beta = \max\{n, m\}$ *are the minimum and the maximum dimensionalities of the problem, h is the height of the problem, and* Δ *stands for the maximum modulus of the determinants of the extended matrix of coefficients of the problem.*

Note. Theorem 8 holds for linear fractional programming.

8. Polynomial solvability of convex quadratic programming

For problems of convex quadratic programming, consisting in minimization of a convex quadratic polynomial (1) under linear constraints (2), we

notice that again their feasibility in R^n implies the existence of a rational optimal solution. An exact algorithm of convex quadratic programming was described in [6] and improved in [9]. The latter result can also be improved as follows:

THEOREM 9. *There exists an exact algorithm of convex quadratic programming which requires* $\lesssim n^4(n+m)\log hn$ *elementary operations over* $\lesssim n\log hn$*-place numbers and additional to the input information storage of* $\lesssim n^2$ *such numbers, where* n, m, h *are of their usual meaning.*

Note. From Theorem 9 it follows that the problem of determining the consistency in real variables of systems of linear inequalities aggravated by a single convex quadratic constraint is polynomially solvable. This result can be extended to any *fixed* number of convex quadratic constraints [10, 9]. The problem of whether there exists a polynomial algorithm for checking the consistency in real variables of general systems of convex quadratic inequalities is open.

9. Polynomial solvability of convex polynomial programming with a fixed number of real and/or integer variables

In [7] a polynomial algorithm was described for solving linear programming problems with a fixed number of integer variables. Using the technique of [7], one can immediately derive from [1], p. 135 and Theorem 2 the following result (see also [7], p. 13).

THEOREM 10. *There exists an exact algorithm, polynomial in* m *and* $\log h$*, for convex polynomial programming with real and/or integer variables, provided that the degree* d *of problems and the number* n *of unknowns are fixed.*

Note. From Note in Section 4 it follows that unlike the linear case [7], there does not exist an algorithm, polynomial in n and $\log h$, for convex quadratic integer programming with a fixed number of constraints unless $P = NP$.

In conclusion, let us mention that this talk is an abridged version of the survey [5].

References

[1] Collins G. E., Quantifier elimination for real-closed fields by cylindrical algebraic decomposition. In: *Automata Theory and Formal Languages*, Lecture Notes in Comput. Sci **33**, Springer, Berlin, 1975.

[2] Хачиян Л. Г., Полиномиальный алгоритм в линейном программировании, *ДАН СССР* **244**, № **5** (1979); translated in: *Soviet Math. Dokl.* **20**, No. **1** (1979).

[3] Хачиян Л. Г., Полиномиальные алгоритмы в линейном программировании, *ЖВМ и МФ*, т. **20**, № **1** (1980), translated in: *USSR Comp. Math. and Math. Phys.* **20**, No. **1** (1980).

[4] Хачиян Л. Г., О точном решении систем линейных неравенств и задач линейного программирования, *ЖВМ и МФ* **22**, № **6** (1982).

[5] Хачиян Л. Г., Выпуклость и алгоритмическая сложность решения задач полиномиального программирования, *Изв. АН СССР, Техническая кибернетика* **6** (1982).

[6] Ковлов М. К., Тарасов С. П., Хачиян Л. Г., Полиномиальная разрешимость выпуклого квадратичного программирования, *ДАН СССР* **248**; translated in: *Soviet Math. Dokl.* **20**, No. 5 (1979).

[7] Lenstra H. W. Jr., *Integer programming with a fixed number of variables*, Dept. of Math., Univ. of Amsterdam, Rept. No. **81-95** (1981).

[8] Шор Н. З., Метод отсечения с растяжением пространства для решения задач выпуклого программирования, *Кибернетика* **13**, № **1** (1977), translated in: *Cybernetics* **13** (1977).

[9] Тарасов С. П., *Алгебраический подход к некоторым задачам выпуклого программирования*, Дис. канд. физ.-мат. наук, Вычислительный центр АН СССР, М., 1979.

[10] Тарасов С. П., Хачиян Л. Г., *Определение совместности систем выпуклых квадратичных неравенств* в R^n, тезисы докладов советско-польского научного семинара по математ. методам в планировании и упр. экономикой, ЦЭМИ, М., 1979.

[11] Тарасов С. П., Хачиян Л. Г., Границы решений и алгоритмическая сложность систем выпуклых диофантовых неравенств, *ДАН СССР* **255**, № **2**, translated in: *Soviet Math. Dokl.* **22**, No. **3** (1980).

[12] Юдин Д. Б., Немировский А. С., Информационная сложность и эффективные методы решения выпуклых экстремальных задач, *Экономика и матем. методы*, **XII**, № **2** (1976), translated in: *Matekon* **13**, **3** (1977).

COMPUTING CENTRE OF THE ACADEMY OF SCIENCES OF THE USSR

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

J. H. VAN LINT

Partial Geometries

1. Introduction

The main purpose of this paper is to survey recent results on partial geometries. It is now twenty years since Bose [1] introduced the concept of a partial geometry in order to study large cliques in strongly regular graphs. Several strong necessary conditions for the existence of strongly regular graphs also have consequences for the existence question for partial geometries. We give the necessary results for strongly regular graphs in Section 2. In Section 3 we define partial geometries and indicate results obtained prior to 1976. For most of the early work on partial geometries we refer to the survey by Thas [27] published in 1977. The recent results are divided into (i) nonexistence theorems (Section 4), (ii) new infinite classes (Section 5) and (iii) sporadic geometries (Section 6).

2. Strongly regular graphs

A *strongly regular graph* (notation srg(v, k, λ, μ)) is a graph (undirected, without loops or multiple edges) on v vertices which is regular with valency k and which has the following two properties:

(i) for each pair $\{x, y\}$ of adjacent vertices there are exactly λ vertices adjacent to x and to y,

(ii) for each pair $\{x, y\}$ of nonadjacent vertices there are exactly μ vertices adjacent to x and to y.

The complement of a srg(v, k, λ, μ) is a srg$(v, l := v-k-1, v-2k+\mu-2, v-2k+\lambda)$. From this we find a necessary condition for the existence of a srg(v, k, λ, μ), namely

$$v-2k+\mu-2 \geqslant 0. \tag{2.1}$$

Furthermore, a simple counting argument shows that

$$k(k-\lambda-1) = \mu(v-k-1). \tag{2.2}$$

We exclude trivial graphs (disconnected graphs and their complements), i.e., we assume

$$0 < \mu < k < v-1. \tag{2.3}$$

Let A be the $(0,1)$ adjacency matrix of a $\mathrm{srg}(v, k, \lambda, \mu)$. Then A satisfies

$$AJ = kJ, \quad A^2+(\mu-\lambda)A+(\mu-k)I = \mu J,$$

where J is the all-one matrix. A has an eigenvalue k with multiplicity one and two other eigenvalues r, s $(r > s)$ satisfying $x^2+(\mu-\lambda)x+(\mu-k) = 0$. The multiplicities of these eigenvalues are

$$f = \frac{-k(s+1)(k-s)}{(k+rs)(r-s)} \quad \text{and} \quad g = \frac{k(r+1)(k-r)}{(k+rs)(r-s)}, \tag{2.4}$$

and they must clearly be integers; (this is known as the *integrality condition*). In fact, if $f \neq g$ then r and s must be integers. The other case (i.e., $f = g$) is called the *half-case*. In situations where we need all these parameters of a strongly regular graph we shall list them as $\mathrm{srg}(v, k, \lambda, \mu; r, s, f, g)$.

We now state three strong necessary conditions for the existence of a $\mathrm{srg}(v, k, \lambda, \mu)$. For more details and proofs we refer the reader to [2], [6], [7], [26]. In Section 3 we shall sketch the proof of (2.8).

(*Krein conditions*)
$$\begin{aligned}(r+1)(k+r+2rs) &\leqslant (k+r)(s+1)^2,\\ (s+1)(k+s+2rs) &\leqslant (k+s)(r+1)^2.\end{aligned} \tag{2.5}$$

(*Absolute bound*) $v \leqslant \frac{1}{2}f(f+3)$ and $v \leqslant \frac{1}{2}g(g+3)$. (2.6)

If the first inequality of (2.5) holds with strict inequality, then (2.6) can be improved to

$$v \leqslant \tfrac{1}{2}f(f+1), \tag{2.7}$$

and similarly for the second inequality of (2.6).

(*Claw bound*) *If* $\mu \neq s^2$, $\mu \neq s(s+1)$, *then* $2(r+1) \leqslant s(s+1)(\mu+1)$. (2.8)

3. Partial geometries

A *partial geometry* $\mathrm{pg}(K, R, T)$ is an incidence structure with a set $\mathscr{P}$ of points and a set $\mathscr{B}$ of lines with the following properties (if a point x is incident with a line L we write $x \in L$, if x and y are on a line we write $x \sim y$):

(i) each line has K points,
(ii) each point is on R lines,
(iii) given a line L and a point $x \notin L$, there are exactly T points $y \in L$ such that $x \sim y$.

The *point graph* of a partial geometry has the points as vertices and an edge $\{x, y\}$ iff $x \sim y$. The point graph of a $\mathrm{pg}(K, R, T)$ is strongly regular (possibly trivial) with parameters:

$$v = K\left(1 + \frac{(K-1)(R-1)}{T}\right), \quad k = R(K-1),$$
$$\lambda = (K-2)+(R-1)(T-1), \quad \mu = RT. \tag{3.1}$$

The dual of a $\mathrm{pg}(K, R, T)$ is the $\mathrm{pg}(R, K, T)$ obtained by interchanging the rôles of $\mathscr{P}$ and $\mathscr{B}$.

If an srg has parameters such that it could be the point graph of a partial geometry we call the srg *pseudo-geometric* and if it is indeed the point graph of a partial geometry we call the srg *geometric*. A pseudo-geometric srg is not necessarily geometric. Bose [1] proved the following theorem:

THEOREM. *If an* srg *is pseudo-geometric corresponding to* $\mathrm{pg}(K, R, T)$ *and if*

$$2K > R(R-1)+T(R+1)(R^2-2R+2)$$

then the graph is geometric. (3.2)

The ideas of the proof of (3.2) were extended by Neumaier [24] and after a subsequent improvement by Brouwer this resulted in the claw bound (2.8). We now sketch the proof. Consider an $\mathrm{srg}(v, k, \lambda, \mu; r, s, f, g)$. Let G be this graph. A clique C in G is called a *grand clique* if C is maximal and $|C| > \frac{1}{2}(\lambda+\mu)+1$. An easy counting argument, using the definitions of λ and μ, shows that each edge of G is in at most one grand clique. The well-known Hoffman bound states that for any clique C in G we have $|C| \leqslant 1+k/(-s) =: K$ and that equality holds iff each point not in C is adjacent to $T := \mu/(-s)$ points of C. The main idea of the proof of (2.8) is to show that certain restrictions on the parameters of G imply that each edge of G is in exactly one grand clique C of size K and that each vertex of G is in a constant number R of such grand cliques. This shows that G is geometric and corresponds to $\mathrm{pg}(K, R, T)$. Finally, this is shown to be impossible, either because one of the parameters is not an integer or $T > R$ or because the point graph of the dual partial geometry does not satisfy (2.5) or (2.6). If S is a coclique of size c in G and p is adjacent to all vertices

in S, then (p, S) is called a *c-claw*. By using the restrictions on the parameters of G and standard counting arguments on the vertices joined to p but not in S it is easy to show that for $1 \leqslant c \leqslant -s-1$ a c-claw can be extended to a $(c+1)$-claw in many ways and that no $(-s+1)$-claw exists. It follows that the vertices which can be added to any $(-s-1)$-claw must form a clique. This argument shows that each edge is in a grand clique and the other properties stated above immediately follow from this fact.

The parameters r, s, f, g of the point graph of $\mathrm{pg}(K, R, T)$ are given by

$$r = K-T-1, \quad s = -R, \quad f = \frac{K(K-1)R(R-1)}{T(K+R-T-1)}, \qquad g = \frac{(K-1)(K-T)\{T+(K-1)(R-1)\}}{T(K+R-T-1)}. \tag{3.3}$$

For a $\mathrm{pg}(K, R, T)$ the Krein conditions become

$$(R-1)\left(R+1-\frac{T}{K-1}\right) \geqslant (K-T)\left(-1+\frac{T(2R-1)}{(K-1)(R-1)}\right), \qquad (K-2)(K-T)^2 \geqslant (R-1)(K-2T). \tag{3.4}$$

For the special case $T = 1$ (generalized quadrangles) the second of these inequalities states that $K = 2$ or $(K-1)^2 \geqslant R-1$. This is known as Higman's inequality (cf. [17], [27]). Partial geometries can be divided into four classes:

1. A pg with $T = K$ (dually $T = R$) is a $2-(v, K, 1)$ *design* (*dual design*),
2. A pg with $T = R-1$ (dually $T = K-1$) is a *net* (*transversal design*),
3. A pg with $T = 1$ is called a *generalized quadrangle*,
4. If $1 < T < \min\{K-1, R-1\}$ then we call the pg *proper*.

In this survey we do not discuss the first two classes. Nets were introduced by Bruck in 1951 (cf. [3]) and Bose's result (3.2) was inspired by Bruck's earlier work on nets [4] (e.g. the idea of grand cliques).

In 1976 Thas [27] wrote a long survey paper about partial geometries. In that paper he described all constructions known at that time for generalized quadrangles (notation: $\mathrm{GQ}(s, t) = \mathrm{pg}(s+1, t+1, 1)$) and two infinite classes of proper partial geometries. The parameters are

(a) $\mathrm{GQ}(s, t)$ with $(s, t) =$ resp. $(q, 1), (q, q), (q, q^2), (q^2, q^3), (q-1, q+1)$ and their dual sets (here q is a prime power),

(b) $K = 2^h - 2^m + 1$, $R = 2^h - 2^{h-m} + 1$, $T = (2^m - 1)(2^{h-m} - 1)$, $0 < m < h$,

(c) $K = 2^h$, $R = 2^{h+m} - 2^h + 2^m$, $T = 2^m - 1$.

For these constructions and several combinatorial characterizations of generalized quadrangles we refer the reader to survey [27], which has a list of 57 references.

4. Recent nonexistence results

A. pg(4, 5, 2) **does not exist.** The smallest value of v for which there exists a pseudo-geometric srg which is not geometric is $v = 28$. This was shown in 1978 by F. de Clerck [9]. If we take the pairs from $\{1, 2, \ldots, n\}$ as vertices and join two pairs by an edge iff they have an element in common, we obtain the *triangular graph* $T(n)$, which is a $\operatorname{srg}\left(\binom{n}{2}, 2(n-2), n-2, 4\right)$. Graphs with these parameters are unique for $n \neq 8$. For $n = 8$ there are three other srg's (28, 12, 6, 4), known as the *Chang graphs* (and no others). We give the proof that $T(8)$ is not geometric; the proofs for the other three graphs are similar. If $T(8)$ were geometric, then lines would correspond to 4-cliques, i.e., to partitions of $\{1, 2, \ldots, 8\}$ into four pairs. W.l.o.g. we can take (12)(34)(56)(78) as a line and in fact it is easy to see that w.l.o.g. there is only one choice for the lines through (12). After that, we have two possible choices for the line through (13) and (24). Both of them make it impossible to choose the remaining lines through (13).

B. pg(6, 9, 4) **does not exist.** Clearly, the idea of 4A can be used for any pg(l, $2l-3$, $l-2$), where for $l \neq 4$ we have the additional advantage that the corresponding srg is unique, namely $T(2l)$. Of course, the number of possibilities increases rapidly. A successful search for the case $l = 5$ will be described in Section 6. In 1983 a similar search was carried out by Lam *et al.* [21] for the case $l = 6$. It took 183 days of computing on a VAX 11/780. No pg(6, 9, 4) was found. This result has an extremely interesting consequence. As we shall see in Section 6, the existence of a projective plane of order 10 with a hyperoval would imply the existence of pg(6, 9, 4). Therefore no such plane exists and this means that the Steiner system $S(3, 12, 112)$ does not exist either.

C. pg(4, 7, 1) = GQ (3, 6) **does not exist.** It is not known whether a srg(76, 21, 2, 7) exists or not but Dixmier and Zara [12], [13] have shown

that, if it exists, it is not geometric. The proof given below is a slight modification of their proof. Suppose a pg(4, 7, 1) exists.

(i) Let $x \not\sim y$, $\mathrm{tr}(x, y) := \{z \mid z \sim x, z \sim y\}$, $\Delta(x, y) := \{z \mid z \not\sim x, z \not\sim y\}$. Then $|\mathrm{tr}(x, y)| = 7$, $|\Delta(x, y)| = 39$.

(ii) For $i = 0, 1, \ldots, 7$ let K_i be the subset of $\Delta(x, y)$ consisting of the points which are collinear with i points in $\mathrm{tr}(x, y)$ and let $n_i := |K_i|$.

(iii) Elementary counting yields $\sum n_i = 39$, $\sum i \cdot n_i = 105$, $\sum \binom{i}{2} n_i = 105$.

(iv) Let p and q be collinear points in Δ such that the line L through p and q does not meet $\mathrm{tr}(x, y)$. Then for the other two points, a and b, on L we have $x \sim a$, $y \sim b$, and the lines through x and a, resp. y and b meet $\mathrm{tr}(x, y)$ at different points. It follows that if $p \in K_i$ then $q \in K_{5-i}$. From this we find that $(7-i)n_i = (2+i)n_{5-i}$, and hence $n_6 = 0$. From (iii) we then find $n_0 = n_5 = n_7 = 0$, $n_1 = 4$, $n_2 = 12$, $n_3 = 15$, $n_4 = 8$.

(v) Let $p \in K_1$ and let $p \sim z \in \mathrm{tr}(x, y)$. From (iv) it follows that the other two points on the line pz are both in K_4. There are six lines through p not meeting $\mathrm{tr}(x, y)$ and on each we have a point in K_4 by (iv). Since $n_4 = 8$, we see that each point in K_1 is adjacent to each point in K_4, contradicting $\mu = 7$.

D. A nonexistence theorem for pg(K, R, T) with $K = R$. The following result was announced by U. Ott at Oberwolfach in late 1982 (cf. [25]):

If pg(K, K, T) exists and the corresponding srg has the eigenvalue $K-T-1$ with *odd* multiplicity then $2K-T-1$ is a square.

The special case of generalized quadrangles yields the following necessary condition:

If GQ(s, s) exists and $s \equiv 2 \pmod 4$, then $\frac{1}{2}s$ is a square. The author of this survey has not seen the proof yet.

5. New infinite classes of partial geometries

A. pg(2^{2n-1}, $2^{2n-1}+1$, 2^{2n-2}). In 1979 Cohen [11] gave the first description of a pg(8, 9, 4). Subsequently Haemers and van Lint [16] gave a much simpler description of pg(9, 8, 4) using the action of PSL(2, 8) on PG(1, 8).

These examples led F. de Clerck, R. H. Dye and J. A. Thas [10] to a third construction and subsequently the discovery of a new infinite sequence, namely the pg(2^{2n-1}, $2^{2n-1}+1$, 2^{2n-2}). The construction is as follows. Let Q^+ be a hyperbolic quadric in PG($4n-1$, 2). The set of maximal totally isotropic subspaces of Q^+ is divided into two disjoint families D_1 and D_2. If H is a projective space of dimension $2n-2$ on Q^+, then Q^+ contains

two maximal totally isotropic subspaces through H, one of each family. Together they determine a $2n$-space which contains a unique hyperplane $M(H)$ through H and not on Q^+. We observe that $M(H) \setminus H$ has 2^{2n-1} points not on Q^+. Let Ψ be a spread of Q^+ consisting of elements of D_1. We define:

$\mathscr{P} := \{\text{points not on } Q^+\}$,
$\mathscr{B} := \{\text{all spaces } M(H) \text{ where } H \text{ is a hyperplane in an element of } \Psi\}$.

We call the elements of $\mathscr{B}$ lines and take natural incidence. Then $(\mathscr{P}, \mathscr{B})$ is a $\mathrm{pg}(2^{2n-1}, 2^{2n-1}+1, 2^{2n-2})$. For a proof we refer to [10].

In [19] Kantor compares a number of constructions of strongly regular graphs, one of which is the above. He shows that, if $2n-1$ is composite, at least three nonisomorphic partial geometries with the same parameters can be constructed. He also proves that the partial geometry $\mathrm{pg}(8, 9, 4)$ in the infinite sequence is isomorphic to the dual of the one constructed by Haemers and van Lint. They have A_9 as the automorphism group!

B. A possibly infinite sequence $\mathrm{pg}(3^{2h+1}, 3^{2h+1}+1, 2\cdot 3^{2h})$. In [28] J. A. Thas generalized the construction given above, using hyperbolic quadrics in $\mathrm{PG}(4h+3, 3)$. Again the construction depends on the existence of spreads. Only for the cases $h = 0$ and $h = 1$ it is known that such spreads exist. The case $h = 0$ leads to a trivial geometry but $h = 1$ gives a new partial geometry $\mathrm{pg}(27, 28, 18)$. At present this is a sporadic example.

C. New generalized quadrangles $\mathrm{GQ}(q, q^2)$ **with** $q = p^r \equiv 2 \pmod 3$. In 1980 Kantor [20] found a new construction for generalized quadrangles with the parameters of known quadrangles and proved that for $q > 2$ the new quadrangles are not isomorphic to any known ones. The proofs are group-theoretic and quite difficult, but a fairly simple geometric description of the construction is possible.

First, we must give the definition of a *generalized hexagon*. This is a bipartite graph of valency $\geqslant 3$ and diameter 6 such that for any two vertices x, y with $d(x, y) < 6$ there is a unique shortest path joining x to y. If we call one of the sets of vertices "points" and the other set "lines" and define incidence by adjacency, we obtain an incidence structure for which there are no m-gons with $m < 6$ and which has the following property:

If a point x is not on line L and not collinear with any point of L then there is a unique point y on L and a unique point z such that z is collinear with x and with y. One can show that there are numbers s, t such that each line has $s+1$ points and each point is on $t+1$ lines (cf. [14]).

For his construction of generalized quadrangles Kantor uses the classical generalized hexagon $H(q)$ associated with the group $G_2(q)$ (cf. [29]). Consider a generalized hexagon H and fix a vertex x. If $d(x, y) = i$, we say that y is of type i. Define $\mathscr{P}$: = vertices of type 1 or 4 and $\mathscr{B}$: = vertices of type 0, 3 or 6. A point and a line are incident if they have distance 1 or 2. It is immediately obvious that properties (i) and (ii) of the definition of a partial geometry are satisfied with $K = q+1, R = q^2+1$. To show that $T = 1$ one must distinguish five types of nonincident point-line pairs. The only case which causes difficulties is a point of type 4 and a line of type 6 which have distance 6. To complete the proof one needs to know that the hexagon does not contain four vertices which are pairwise at a distance 4 and such that the shortest paths joining them are disjoint. For one of a dual pair of hexagons $H(q)$ this is the case.

6. Sporadic partial geometries

A. pg (5, 7, 3). In Sections 4A and 4B we considered partial geometries of type $\mathrm{pg}(l, 2l-3, l-2)$ connected with the triangular graph $T(2l)$. For $l \neq 4$ this graph is unique. The classical method for constructing a pg(5, 7, 3) is to take a hyperoval $\mathcal{O}$ in PG(2, 8) and delete the points of $\mathcal{O}$ from the plane and delete all exterior lines. The remaining set of points and lines (with the usual incidence) is a pg (5, 7, 3). By a computer search such as that described in Section 4B Mathon [23] showed that there are exactly two nonisomorphic pg's (5, 7, 3), one of which is not derivable from a projective plane.

B. pg(6, 6, 2). A partial geometry with $K = R = 6$ and $T = 2$ was first constructed by van Lint and Schrijver [22]. We give a description of this geometry which is due to Cameron and van Lint [8]. In $\boldsymbol{Z}_3^6$ consider the subgroup G generated by (1, 1, 1, 1, 1, 1). For each coset $\boldsymbol{a}+G$, the sum of the coordinates of the points is a constant i. We say that the coset is of type i. Let $\mathscr{A}_i$ be the set of cosets of G of type i. We define a tripartite graph Γ by joining the coset $\boldsymbol{a}+G$ to the coset $\boldsymbol{a}+\boldsymbol{b}+G$ for each $\boldsymbol{b}$ which has only one nonzero coordinate. Clearly, any element of $\mathscr{A}_i$ has six neighbours in $\mathscr{A}_{i+1}$ and six in $\mathscr{A}_{i+2}$. We can construct our partial geometry by taking some $\mathscr{A}_i$ as point set and one of the two other classes $\mathscr{A}_j$ as line set. Incidence corresponds to adjacency. That $K = R = 6$ is clear. It is also an easy exercise to show that $T = 2$.

C. pg(5, 18, 2). The most interesting of the sporadic partial geometries was found by Haemers in 1981 [15]. In order to describe it we must first

present a useful description of the well-known Hoffman–Singleton graph [18] (abbreviated to Ho-Si). Let C be the set of 15 points of PG(3, 2) and let D be the set of 35 lines of this geometry. It is known that D can be identified with the triples from $\{1, 2, \ldots, 7\}$ in such a way that intersecting lines correspond to triples with one element in common. If we consider the 30 Steiner triple systems on $\{1, 2, \ldots, 7\}$ and call two of them equivalent if they have exactly one triple in common, then we obtain two equivalence classes of fifteen triple systems. One of them corresponds to the points of PG(3, 2), the other to planes (both being represented by seven mutually intersecting lines). We now define a graph on the vertex set $C \cup D$ by joining an element of C to an element of D if the point is on the line and by joining two elements of D if the corresponding triples are disjoint. We claim that this is the Moore graph srg(50, 7, 0, 1), i.e., Ho-Si. All verifications are trivial except showing that there is a unique line adjacent to both elements of a non-adjacent point-line pair. This, however, follows from the fact that if $\{a, b, c\}$ is not in a STS(7) then the triple system contains exactly one triple disjoint from $\{a, b, c\}$.

Haemers' construction of pg(5, 18, 2) starts from the observation that one can construct $G = \text{srg}(175, 72, 20, 36)$ by taking the edges of Ho-Si as vertices and joining two of these vertices if the edges have distance 2 in Ho-Si (i.e., the edges are disjoint and there is a unique edge joining them). We do not prove this (but it is easy).

The graph G is a nice example of the difficulties one usually encounters in trying to show that a pseudo-geometric srg is geometric. We have to find 630 lines of size 5, i.e., 630 5-cliques in G. First observe that two edges of Ho-Si which correspond to an edge in G define a unique pentagon in Ho-Si. Hence, we must find sets of five edges in Ho-Si which are pairwise in a pentagon. This implies that the five edges induce a Petersen subgraph (i.e., $\overline{T(5)}$) in Ho-Si. In fact, a line of the geometry we must construct corresponds to a matching in a Petersen subgraph of Ho-Si. Now, Ho-Si contains 525 Petersen graphs and each of these has six matchings. We need only 630 lines, i.e., we must choose in some way a set of 105 *'special'* Petersen graphs in Ho-Si, such that each pentagon of Ho-Si is in a unique special Petersen graph. This is the point where real ingenuity enters the proof.

Observe that a pentagon in Ho-Si cannot contain more than two points of C and then an elementary counting argument shows that there are 630 pentagons with one point in C and 630 pentagons with two points in C. Call these two sets of pentagons $\mathscr{P}_1$ and $\mathscr{P}_2$.

The next steps are proved by using the fact that our description of

Ho-Si shows that A_7 is an automorphism group fixing the set C. It acts transitively on both $\mathscr{P}_1$ and $\mathscr{P}_2$. Let P_2 be a pentagon in $\mathscr{P}_2$ containing the vertices x, y from C. Then it contains the vertex corresponding to the line L through x and y in PG(3, 2). If z is the third point on this line, then there is a unique Petersen graph in Ho-Si containing P_2 and the edge $\{L, z\}$. This then determines a unique pentagon P_1 in $\mathscr{P}_1$ containing the vertex z. We define π by $P_1 := \pi P_2$. Using our representation of Ho-Si one easily checks that π is one-to-one.

Now we are done. The Petersen graphs $P_2 \cup \pi P_2$ are called *special*. By inspection we see that each of them arises in six different ways from the above construction. Hence, there are 105 special Petersen graphs and every pentagon is in exactly one of them.

Since G is strongly regular and we have found the right number of lines for a pg(5, 18, 2), we are done (by a well-known theorem).

Calderbank and Wales [5] have shown that this geometry can be described in terms of the 176 octads in a Steiner system $S(5, 8, 24)$ that contain a given point P and do not contain a given point Q. A third description uses 175 subgroups of A_7.

References

[1] Bose R. C., Strongly regular graphs, partial geometries, and partially balanced designs, *Pacific J. Math.* **13** (1963), pp. 389–419.

[2] Brouwer A. E. and Lint J. H. van, Strongly regular graphs and partial geometries, *Proc. Silver Jubilee Conf. on Combinatorics*, Waterloo, 1982.

[3] Bruck R. H., Finite Nets I, *Can. J. Math.* **3** (1951), pp. 94–107.

[4] Bruck R. H., Finite Nets II, *Pacific J. Math.* **13** (1963), pp. 421–457.

[5] Calderbank R. and Wales D. B., *The Haemers partial geometry and the Steiner system S* (5, 8, 24), preprint.

[6] Cameron P. J., Strongly regular graphs, in: *Selected topics in graph theory*, L. W. Beineke and R. J. Wilson (eds.), Academic Press, 1978, pp. 337–360.

[7] Cameron P. J. and Lint J. H. van, *Graphs, Codes and Designs*, London Math. Soc. Lecture Note Series **43**, Cambridge, 1980.

[8] Cameron P. J. and Lint J. H. van, On the partial geometry pg (6, 6, 2), *J. Comb. Th.* (*A*) **32** (1982), pp. 252–255.

[9] Clerck F. de, *Partial Geometries*, thesis, University of Gent, 1978.

[10] Clerck F. de, Dye R. H., and Thas J. A., An infinite class of partial geometries associated with the hyperbolic quadric in PG $(4n-1, 2)$, *Eur. J. Comb.* **1** (1980), pp. 323–326.

[11] Cohen A. M., A new partial geometry with parameters $(s, t, \alpha) = (7, 8, 4)$, *J. Geometry* **16** (1981), pp. 181–186.

[12] Dixmier S. and Zara F., Essai d'une méthode d'étude de certains graphes liés aux groupes classiques, *C. R. Acad. Sc. Paris* **282** Série A, pp. 259–262.
[13] Dixmier S. and Zara F., *Étude d'un quadrangle généralisé autour de deux de ses points non liés*, preprint.
[14] Feit W. and Higman G., The nonexistence of certain generalized polygons, *J. Algebra* **1** (1964), pp. 114–131.
[15] Haemers W. H., A new partial geometry constructed from the Hoffman–Singleton graph, in: *Finite Geometries and Designs*, P. J. Cameron, J. W. P. Hirschfeld and D. R. Hughes (eds.), London Math. Soc. Lecture Note Series **49**, Cambridge, 1981, pp. 119–127.
[16] Haemers W. H. and Lint J. H. van, A partial geometry pg (9, 8, 4), *Annals of Discr. Math.* **15** (1982), pp. 205–212.
[17] Higman D. G., Partial geometries, generalized quadrangles, and strongly regular graphs, in *Atti Convegno di Geometria e sue Applicazioni*, Perugia, 1971.
[18] Hoffman A. J. and Singleton R. R., On Moore graphs with diameter 2 and 3, *IBM J. Res. Develop.* **4** (1960), pp. 497–504.
[19] Kantor W. M., Strongly regular graphs defined by spreads, *Isr. J. Math.* **41** (1982), pp. 298–312.
[20] Kantor W. M., Generalized quadrangles associated with $G_2(q)$, *J. Comb. Th.* (*A*) **29** (1980), pp. 212–219.
[21] Lam C. W. H., Thiel L., Swiercz S., and McKay J., The nonexistence of ovals in a projective plane of order 10, *Disc. Math.* (to appear).
[22] Lint J. H. van and Schrijver A., Construction of strongly regular graphs, two-weight codes and partial geometries by finite fields, *Combinatorica* **1** (1981), pp. 63–73.
[23] Mathon R., The partial geometries pg (7, 5, 3), *Congressus Numerantium* **31** (1981), pp. 129–139.
[24] Neumaier A., Strongly regular graphs with smallest eigenvalue $-m$, *Archiv der Math.* **33** (1979), pp. 392–400.
[25] Ott U., in preparation.
[26] Seidel J. J., Strongly regular graphs, in: *Surveys in Combinatorics*, Proc. 7th Brit. Comb. Conf., B. Bollobás (ed.), London Math. Soc. Lecture Note Series **38**, Cambridge, 1979, pp. 157–180.
[27] Thas J. A., Combinatorics of partial geometries and generalized quadrangles, in: *Higher Combinatorics*, M. Aigner (ed.), Reidel, Dordrecht, 1977, pp. 183–199.
[28] Thas J. A., Some results on quadrics and a new class of partial geometries, *Simon Stevin* **55** (1981), pp. 129–139.
[29] Tits J., Sur la trialité et certains groupes qui s'en déduisent, *Publ. Math. IHES* **2** (1959), pp. 14–60.

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

L. LOVÁSZ

Algorithmic Aspects of Combinatorics, Geometry and Number Theory

One of the most spectacular successes in combinatorics in the last decades has been the development of combinatorial optimization. The discovery of computational complexity classes (most notably the classes P and NP) has provided the right framework for this rapid growth; discrete mathematical models from operation research supplied the field with problems of practical interest; and previously "pure" fields of combinatorics, such as graph theory or matroid theory, provided ideas and tools for mathematically non-trivial results. The influence of ideas from computational complexity, however, goes beyond combinatorics and invades such fields of classical mathematics as number theory, group theory or geometry. Algorithmic aspects have shed new light on ancient mathematical problems: for example, the problem of factoring an integer into primes, when viewed as an algorithmic problem, is far from being solved. The study and classification of finite simple groups has led to the development of efficient group-theoretical algorithms.

But even in fields concentrating on algorithms, such as continuous optimization or numerical analysis, the ideas of computational complexity theory may bring new evaluation criteria and may lead to new algorithms. Much of the excitement and misunderstanding surrounding the ellipsoid method (which became well known after Khachiyan [6] applied it to linear programming, but which may be viewed in fact as a more general method of minimization of convex functions over convex domains) is due to this new concept of efficiency of algorithms.

Another implication of this method is the algorithmic equivalence of the "optimization" and "separation" problems for convex bodies. It is clear that these two problems are logically equivalent: if we know the maximum value of any linear objective function over a convex body K,

then we know the body, and so we can find, for any point not in K, a hyperplane separating it from K. But it is a quite surprising fact that if, say, the separation can be solved in polynomial time then the optimization problem can also be solved in polynomial time.

But what kinds of convex bodies do we want to optimize on? One such class arises from combinatorial optimization problems. Here the geometrical algorithms mentioned above meet a development in discrete optimization which took place in the last two decades — the so-called polyhedral combinatorics. One can associate polyhedra with various combinatorial optimization problems. Somewhat surprisingly, it turns out that the separation and optimization problems for these polyhedra correspond to essentially different combinatorial problems. Hence the general geometrical equivalence principle mentioned above establishes the algorithmic equivalence of quite different combinatorial problems. There exist important combinatorial problems whose polynomial-time solvability can be established — so far — by these means only.

It is very difficult to draw the border line between combinatorial optimization and integer programming — and the methods used in these fields also overlap considerably. But there is a third field of mathematics which is closely related: the classical field of the "geometry of numbers". Roughly speaking, both integer linear programming (polyhedral combinatorics) and the geometry of numbers are concerned with finding lattice points in convex bodies. But the conditions imposed upon these bodies appear to be so different that there is very little connection between these fields. This situation, however, is also changing. The breakthrough is due to the result of H. W. Lenstra, Jr., which says that integer linear programming in bounded dimension can be solved in polynomial time. The algorithm involves a refined basis reduction algorithm, a classical topic in the geometry of numbers.

In this paper we survey the main ingredients of an algorithmic theory of combinatorics, geometry and number theory: the ellipsoid method, polyhedral combinatorics, and lattice algorithms. It is sometimes amazing how well these ingredients fit. We shall have to assume that the reader is familiar with the fundamentals of the theory of the complexity classes P and NP.

There is an important point to emphasize here. It is fairly easy to understand the geometrical idea behind the algorithms below; but the details are usually tedious and quite often even difficult. Furthermore, to state the results in their natural generality, one has to define in a precise way several things, such as oracle algorithms, computation with irrational

numbers, weak and strong versions of numerical problems, and others. This task is undertaken in a forthcoming book by Grötschel, Lovász and Schrijver; see also their paper [4].

1. The ellipsoid method

The ellipsoid method (Shor [10, 11]; Judin and Nemirovskii [5]) is fairly well known and it would be superfluous to discuss its details. In its basic version, we want to find a point in a convex body $K \subseteq \mathbf{R}^n$. Throughout the procedure, we have an ellipsoid E including K. The iteration step consists of checking whether or not the centre x of E belongs to K. If so, we are done. If not, we take a hyperplane H through x which avoids K. The hyperplane H cuts E into two halves. Choose the half containing K and include this half-ellipsoid in a new ellipsoid whose volume is smaller than the volume of E. It turns out that the volume of E tends to 0, and so sooner or later we must end up with a point in K.

This is of course a very informal description, and one would have to argue long to fill in the details. But it is already clear from this sketch that the following information about K is needed:

(a) an initial ellipsoid (usually a ball of some radius R about 0) including K;

(b) a lower bound δ on the volume of K;

(c) a way (subroutine, oracle) to check if a point x belongs to K and, if not, separate x from K by a hyperplane.

The important property of the ellipsoid method is that not only does it find a point in K but it does so in $0(n|\log\delta| + n^2\log R)$ steps, i.e., in time polynomial with respect to the length of binary encoding of the data (a) and (b), and with respect to the dimension n. So if the subroutine for (c) can be implemented in polynomial time (with respect to the size of some description of K), then the ellipsoid method also runs in polynomial time.

Many versions of this algorithm have been found. First, one can vary the goals: e.g., instead of finding just one point in K, one can use a similar method to find a point in K which maximizes a linear objective function over K. Or one can turn things around: given an oracle which tells us the maximum value of any linear objective function over K, we can use the ellipsoid method to accomplish (c). Or one can use a rather more refined version (Judin and Nemirovskii [5]) to find an ellipsoid E about K whose centre is "deep in K" in the sense that the concentrical homothetical ellipsoid obtained by shrinking E by a factor $1/(n+1)$ is contained in K.

Secondly, one can vary the conditions. Instead of the somewhat restrictive hypotheses (a) and (b), we could assume that K is a polyhedron which can be defined by a system of linear inequalities whose coefficients are bounded by an a priori known number. Or one can weaken (c) and only use an oracle to check whether $x \in K$, provided a ball contained in K is also known a priori.

All these variations of course are made possible by appropriate versions of the algorithm. Instead of the basic step which cuts the ellipsoid into two congruent halves, one uses other types of cuts: deep cuts, shallow cuts, parallel cuts, surrogate cuts, etc. For a survey of some of these versions, see Bland, Goldfarb and Todd [2].

The most important consequence of the ellipsoid method is the following. It is well known that a compact convex set K can be described as the convex hull of its extremal points as well as the intersection of its supporting halfspaces. The ellipsoid method implies that (under appropriate technical hypotheses) these two descriptions are not only logically but algorithmically equivalent: if we have a polynomial-time algorithm to check whether or not a given point belongs to K, then we can also check in polynomial time if a halfspace contains K and vice versa.

(Let me remark that many fundamental results and notions in geometry lead to interesting and mostly unsolved algorithmic problems. For example, can the volume of a polytope be determined in polynomial time?)

2. Polyhedral combinatorics

Let us explain this technique on an example. Let G be a graph with an even number of points. A *perfect matching* of G is a system of mutually disjoint lines which pair up all points. The problem of existence, enumeration and structure of perfect matchings has been studied extensively; here we restrict ourselves to one approach to the perfect matching problems.

Let us represent each perfect matching M by its incidence vector $\chi^M \in \boldsymbol{R}^{E(G)}$. Let S denote the set of incidence vectors of perfect matchings, and let $\operatorname{conv} S$ be the convex hull of S. We call $\operatorname{conv} S$ the *perfect matching polytope* of the graph G. This polytope compresses a lot of information about perfect matchings; here we discuss its application to an optimization problem.

Let a weight $c(e)$ be assigned to every line $e \in E(G)$. We are interested in finding a perfect matching of G with maximum weight. The weighting c may be viewed as a vector $c \in \boldsymbol{R}^{E(G)}$, and the weight of a perfect matching

M is just the inner product $c \cdot \chi^M$. Hence our problem is to find the maximum of $c \cdot x$ over all $x \in S$. Now clearly

$$\max\{c \cdot x\colon\ x \in S\} = \max\{c \cdot x\colon\ x \in \operatorname{conv} S\}.$$

The right-hand side expresses our task as the problem of finding the maximum of a linear objective function over a convex polytope. This is just a linear programming problem! However, to be able to apply any methods from linear programming (duality theorem, simplex method, etc.) we have to express $\operatorname{conv} S$ as the solution set of a system of linear inequalities. It is of course clear that such a (finite) set of linear inequalities exists, but how to find them? The following important result of Edmonds [3] gives the answer. For a set $X \subseteq V(G)$, we denote by $\Gamma(X)$ the set of lines connecting X to $V(G) - X$.

THEOREM. *The perfect matching polytope of a graph G is the solution set of the following inequalities:*

(i) $x \geqq 0$,

(ii) $\chi^{\Gamma(v)} \cdot x = 1$ *for all points* $v \in V(G)$,

(iii) $\chi^{\Gamma(T)} \cdot x \geqq 1$ *for all* $T \subseteq V(G)$, $|T|$ *odd*.

This theorem has opened up a whole area of research, and inequalities for the convex hulls of many other combinatorially defined set of vectors have been found. Unfortunately, it is not always possible to find such a nice description; for polyhedra associated with *NP*-complete problems, even to decide whether a given inequality has to belong to such a system is an *NP*-complete problem.

If we apply the duality theorem of linear programming to the linear program

maximize $c \cdot x$,

subject to (i), (ii), and (iii),

we obtain a min–max formula for the maximum weight of a perfect matching. But if we want to apply a linear programming algorithm (say, the simplex method), then there is a very substantial difficulty: the number of inequalities under (iii) is too large, even to write them down takes exponential time.

It turns out that, from an algorithmic point of view, the fact that $\operatorname{conv} S$ is a polytope is not really important. Since we cannot list all the inequalities needed to describe $\operatorname{conv} S$ anyway, it does not matter whether their number is finite or infinite. The crucial point is that we should be able to check whether or not a given point x belongs to $\operatorname{conv} S$; if we can

solve this problem, then we can optimize any linear objective function over $\operatorname{conv} S$ using the ellipsoid method.

So, given $x \in \mathbf{R}^{E(G)}$, we want to check (in polynomial time) whether the inequalities (i)–(iii) are satisfied. The groups (i) and (ii) can be checked by straightforward substitution; but to check (iii) in polynomial time we need a more involved method. In fact, here we have a new combinatorial optimization problem: viewing x as "weights", we are looking for a set $T \subseteq V(G)$ with T odd and $\chi^{F(T)} \cdot x$ minimal. Fortunately, this problem is simpler than the original: see the algorithm of Padberg and Rao [9]. But one can also use the ellipsoid method again and solve this problem in polynomial time.

In fact, a very large number of combinatorial optimization problems which were solved by *ad hoc* polynomial algorithms or by no polynomial algorithm at all can be solved in polynomial time by a combination of the ellipsoid method and a very simple procedure called the Greedy algorithm. It is natural that such a general method cannot give algorithms which would be anywhere close to optimal as far as running time goes. The ellipsoid method can be used to establish the *existence* of a polynomial-time algorithm for certain problems, and thereby indicate that it is worthwhile to look for efficient special-purpose algorithms for the same problems.

3. Basis reduction in lattices

In Section 2 we replaced a finite set S of vectors by $\operatorname{conv} S$; this enabled us to use techniques of convex geometry. Instead of taking convex combinations, we could take linear combinations with integral coefficients. We then obtain a *lattice*, and we could hope to apply the techniques of the geometry of numbers.

However, the algorithmic theory of lattices appears to be even less developed than the algorithmic study of convex bodies. For example, it was only recently shown by Bachem and Kannan [1] that the problem whether a lattice generated by a given set of vectors contains a further given vector can be solved in polynomial time. Also, combinatorially defined lattices have been studied to a very little extent only. For example, no analogue of Edmonds' theorem is known to describe the lattice generated by the incidence vectors of perfect matchings of a graph.

Here we shall discuss only one problem on lattices, that of *basis reduction*. A *basis* of a lattice is a set of linearly independent generators. Let L be a lattice and $b_1, \ldots, b_n$ a basis of L. For the sake of simplicity, assume

that $b_1, \ldots, b_n \in \mathbf{Q}^n$. Then $|\det(b_1, \ldots, b_n)| = \det L$ depends only on the lattice L. Trivially

$$|b_1| \ldots |b_n| \geqq \det L.$$

It was proved by Minkowski that every lattice has a basis $b_1, \ldots, b_n$ with

$$|b_1| \ldots |b_n| \leqq \left[\frac{2n}{e\pi}\right]^{n/2} \det L.$$

It is not known, however, how to find such a basis algorithmically. But if we allow a little larger factor on the right-hand side, we can find such a "reduced" basis in polynomial time (A. K. Lenstra, H. W. Lenstra, Jr. and L. Lovász [7]):

THEOREM. *Given n linearly independent vectors $a_1, \ldots, a_n \in \mathbf{Q}^n$, we can find in polynomial time a basis $b_1, \ldots, b_n$ of the lattice L generated by $a_1, \ldots, a_n$, such that*

$$|b_1| \ldots |b_n| \leqq 2^{n(n-1)/4} \det L.$$

This algorithm has many applications. It gives rise to polynomial-time algorithms for simultaneous diophantine approximation and for factoring polynomials over the rational field. In this paper we discuss an application which relates more closely to the topic: an interesting combination of the basis reduction algorithm and the ellipsoid method gives rise to an improvement of Lenstra's algorithm for integer programming in bounded dimension.

THEOREM. *Let $a_i \cdot x \leqq b_i$ $(i = 1, \ldots, m;\ a_i \in \mathbf{Q}^n,\ b_i \in \mathbf{Q})$ be a set of linear inequalities and P its solution set. Then we can find, in polynomial time, one of the following:*

(a) *an integral vector $x \in P$;*

(b) *an integral vector $c \in \mathbf{Z}^n$ such that*

$$\max\{c \cdot x : x \in P\} - \min\{c \cdot x : x \in P\} \leqq 4^n.$$

The algorithm which achieves this can be sketched as follows.

I. By quite standard tricks, we can reduce the problem to the case where P is bounded and full-dimensional.

II. Then the "shallow cut" version of the ellipsoid method (due to Judin and Nemirovskii [5]) is used to find an ellipsoid E including P such

that the concentrical ellipsoid E' obtained from E by a homothetical transformation with ratio $1/(n+1)$ is contained in P.

III. Finally, we consider E as the unit sphere of a Euclidean norm $\|\cdot\|$ on $\boldsymbol{R}^n$ and find a reduced basis $b_1, \ldots, b_n$ of the lattice $\boldsymbol{Z}^n$ with respect to this norm. Then a simple computation shows that either the linear function

$$c \cdot x = \det(b_1, \ldots, b_{n-1}, x)$$

satisfies (b) or else an integral point in P can be obtained from the centre of E by rounding.

Note that the running time of this algorithm is bounded by a polynomial in n and the space needed to write the original inequalities. If we want to go further, we have to confine ourselves to algorithms that are polynomial in the space needed to write the inequalities but may be exponential in the dimension n. Such an algorithm is still polynomial in every given dimension.

To solve the problem whether or not P has an integral point, we can run the above algorithm. If it ends with (a), we are done. If it ends with (b), then we add a constraint $c \cdot x = k$ $(k \in \boldsymbol{Z}, \min\{c \cdot x: x \in P\} \leqq k \leqq \max\{c \cdot x: x \in P\})$ to the given inequalities and solve at most 4^n problems of lower dimension.

References

[1] Bachem A. and Kannan R., Polynomial Algorithms for Computing the Smith and Hermite Normal Forms of an Integer Matrix. *SIAM Journal on Computing* **8** (1979), pp. 499–507.

[2] Bland R. G., Goldfarb D., and Todd M. J., The Ellipsoid Method: A Survey, *Oper. Res.* **29** (1981), pp. 1039–1081.

[3] Edmonds J., Maximum Matching and a Polyhedron with (0, 1) Vertices, *J. Res. Nat. Bur. Standards* **69B** (1965), pp. 125–130.

[4] Grötschel M., Lovász L., and Schrijver A., The Ellipsoid Method and Its Consequences in Combinatorial Optimization, *Combinatorica* **1** (1981), pp. 169–197.

[5] Judin D. B. and Nemirovskii A. S., Informational Complexity and Effective Methods of Solution for Convex Extremal Problems, *Ekon. i Mat. Metodi* **12** (1976), pp. 357–369; English translation: Matekon: *Trans. Russ. East Eur. Math. Econ.* **13** (1976), pp. 24–45.

[6] Khachiyan L. G. A Polynomial Algorithm in Linear Progamming, *Doklady Akad. Nauk SSSR* **244** (1979), pp. 1093–1096; English translation: *Soviet Math. Dokl.* **20**, pp. 191–194.

[7] Lenstra A. K., Lenstra H. W., Jr., and Lovász L., Factoring Polynomials with Rational Coefficients, *Math. Annalen* **261** (1982), pp. 515–534.

[8] Lenstra H. W., Jr., Integer Programming with a Fixed Number of Variables, to appear in *Mat. Oper. Res.*

[9] Padberg M. W. and Rao, M. R., Odd Minimum Cut-Sets and *b*-Matchings, *Math. Oper. Res.* **7** (1982), pp. 67–80.

[10] Shor N. Z., Convergence Rate of the Gradient Descent Method with Dilatation of the Space, *Kibernetika* **2** (1970), pp. 80–85; English translation: *Cybernetics* **6** (1970), pp. 102–108.

[11] Shor N. Z., Cut-Off Method with Space Extension in Convex Programming Problems, *Kibernetika* (1977), pp. 94–95; English translation: *Cybernetics* **13** (1977), pp. 94–96.

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

RICHARD M. KARP*

The Probabilistic Analysis of Combinatorial Optimization Algorithms

1. Introduction

The branch of computer science concerned with the correctness and efficiency of algorithms has developed rapidly over the past two decades. An algorithm is said to be correct if, for every input presented to it, it produces the desired output. The efficiency of an algorithm is usually established by deriving an upper bound on its execution time. Such a bound states that, on every input, the number of steps executed by the algorithm does not exceed some specified function of the size of that input. The more slowly this function grows, the more efficient the algorithm is considered to be.

Many important algorithms have been proven to be both correct and of nearly optimal efficiency. Such results are obviously excellent guidelines for the selection of an algorithm to be used in practice. Unfortunately, there is a large class of problems for which the stated criteria are too severe to be usefully applied, since they demand correct and efficient behavior of the algorithm on every input that may be presented. An example is the linear programming problem, for which the simplex method is the algorithm of choice. The simplex method is correct, since it produces an optimal solution to every linear programming problem presented to it, but it is not efficient according to the usual worst-case criterion, since there is a family of contrived inputs for which it experiences a combinatorial explosion in running time. Nevertheless, the simplex method performs so well on typical inputs that practitioners are quite willing to overlook its theoretical imperfections.

The criteria of correctness and worst-case efficiency are particularly inapplicable to the class of NP-hard combinatorial problems. There is strong circumstantial evidence, although no conclusive proof, that these

* Research supported by NSF Grant MCS-8105217.

problems are intractable, in the sense that no correct algorithm for such a problem can run within a polynomial time bound. If this folk belief about the intractability of NP-hard problems proves correct, then every algorithm for such a problem must inevitably be imperfect: there will be some inputs for which the algorithm either runs too long or fails to give a correct result. Nevertheless, such imperfect algorithms can be useful if they do not fail too often, especially if the failure is detectable.

One way to validate or compare imperfect algorithms for NP-hard combinatorial problems is simply to run them on typical instances and see how often they fail. This paper explores a complementary theoretical approach, in which we assume that the problem instances presented to the algorithm are drawn from some natural probability distribution. On this assumption we investigate the probability that the algorithm fails. While probabilistic assumptions are always open to question, the approach seems to have considerable explanatory power, and it certainly provides an interesting new realm for the application and extension of a variety of results in probability theory.

In Section 2 we apply the probabilistic approach to the problem of partitioning a given set of numbers into two subsets, A and B, such that the sum of the elements in A is as nearly equal as possible to the sum of the elements in B. Determining an optimal partition is very hard, but we show that a remarkably simple algorithm gives an excellent approximate solution with high probability. Section 3 is concerned with the construction of Hamiltonian circuits, matchings, maximum cliques and minimum colorings in random graphs. The development can be viewed as the extension of the classical Erdös-Rényi theory of random graphs in a highly constructive direction, in which we are concerned not only with the probability that a random graph has a certain property, but also with the probability that an efficient algorithm will succeed in establishing that the property holds. Finally, in Sections 4 and 5 we show that certain simple and efficient algorithms have a high probability of producing near-optimal solutions to random instances of the notorious traveling-salesman problem. Section 4 takes up the asymmetric version of the problem, and the Euclidean version is discussed in Section 5.

2. A partitioning problem

Alice and Bob are the heirs to an estate. The estate consists of n indivisible assets, each of which must go either to Alice or to Bob. How shall the two heirs divide the estate so that each receives approximately half the total value? Abstractly, the data for this problem is a set S of n positive

real numbers (the values of the assets), and a division of the estate is a partition of S into two subsets, A and B. Associated with any such partition is a cost $\Delta(A, B) = \left| \sum_{a \in A} a - \sum_{b \in B} b \right|$. The partitioning problem asks for a partition of minimum cost.

To set the stage for a probabilistic analysis, let us assume that the elements of S are drawn independently from a probability distribution over the interval $(0, 1)$ with a continuous density function. Under this assumption the exact solution of this NP-hard problem appears immensely difficult, even if we allow a small probability of failure. We shall show, however, that a certain simple and fast *differencing algorithm* has a high probability of giving a partition A, B of microscopically small cost.

The algorithm starts with the set S. It repeatedly chooses two numbers from the current set, and replaces them by the absolute value of their difference. This differencing process continues until only one number t remains. Then, by tracing back through the steps it has executed, the algorithm constructs a partition A, B of the original set S, such that $\Delta(A, B) = t$. We leave it to the reader to show how this partition is constructed.

Many versions of the differencing algorithm are possible, corresponding to different rules for choosing the two elements to be differenced at each step. For one simple choice rule the following can be proven [10]: with probability tending to 1 as $n \to \infty$, $\Delta(A, B) < e^{-c(\log n)^2}$. Here c is a positive constant determined by the distribution from which the elements of S are drawn.

This result shows that the differencing algorithm indeed tends to give a partition of extremely small cost, but the cost of an optimal partition tends to be even smaller; with probability tending to 1 as $n \to \infty$, there exists a partition of cost $< dn^2 2^{-n}$ where the constant d is associated with the distribution from which the elements of S are drawn. This latter result is proved by a technique known as the second moment method, as follows: let the random variable Z denote the number of partitions A, B of S such that $\Delta(A, B) < dn^2 2^{-n}$. Then the value of Z is a nonnegative integer. By showing that the expectation of Z is suitably large and its variance suitably small, and then applying Chebyshev's inequality, one finds that $\Pr[Z = 0]$ tends to zero. This proof is nonconstructive, and no efficient algorithm is known which has a high probability of producing a partition of cost less than $e^{-c(\log n)^2}$

3. Algorithmic theory of random graphs

3.1. Matchings and Hamiltonian circuits. The sample space $G_{n,p}$ of labeled graphs on the vertex set $\{1, 2, \ldots, n\}$ is defined by the following rule:

each possible edge is present with probability p, and the occurrences of distinct edges are independent events. Erdös and Rényi ([6], [7]) demonstrated almost twenty-five years ago that certain properties of graphs drawn from these sample spaces are sharply predictable. One of their results concerns the property that a perfect matching exists. Let n range over the positive even integers. Then, for each real number c,

$$\Pr[\text{A graph drawn from } G_{n,\frac{\log n+c}{n}} \text{ has a perfect matching}] \xrightarrow[n\to\infty]{} \exp(-e^{-c}).$$

In view of this result, the function $p(n) = (\log n)/n$ is called the threshold for the existence of a perfect matching. As the probability of an edge being present passes $p(n)$ the probability that a perfect matching exists jumps abruptly from a very small value to a value close to 1. For many years it was an open problem to determine the threshold for the existence of a Hamiltonian circuit. It has recently been shown that the threshold for this property is $p(n) = (\log n + \log\log n)/n$ ([15], [16]).

Threshold theorems in the theory of random graphs are usually proved by nonconstructive counting arguments, but they can also be proved constructively, in a way that sheds light on the power of algorithms. Angluin and Valiant [1] have given a fast algorithm that produces a Hamiltonian circuit with high probability when n is sufficiently large and $p > c(\log n)/n$, where c is a sufficiently large constant. In this algorithm the graph is represented by a set of edge lists: the edge list for vertex u contains all the edges incident with u. The algorithm starts with a trivial path containing one vertex and no edges, and tries to build successively longer paths. At a general step the algorithm has a path P from u to v. If the edge list of v is empty then the algorithm reports failure and halts. Otherwise it draws a random edge $\{v, w\}$ from the edge list of v and uses that edge to construct a new path. If w does not lie in P then the new path is constructed by adjoining that edge to the end of P, creating a path from u to w. If w lies in P then a new path is obtained by adding $\{v, w\}$ to P and deleting the first edge on the segment of P from w to v. The edge $\{v, w\}$ is also deleted from the edge lists of v and w. The process continues until all n vertices lie in P, and the algorithm then proceeds in a similar fashion until the two end points of the current path are adjacent in the original graph, so that a Hamiltonian circuit is created. Analysis of the algorithm shows that, when $p = c(\log n)/n$ with the constant c suitably large, the probability that a Hamiltonian circuit will be created before an empty edge list is encountered tends to 1 as $n \to \infty$. Shamir [19] gives a more complex, but still highly efficient, path-building algorithm which constructs a Hamiltonian circuit with high probability when $p > (\log n + 3\log\log n + w(n))/n$, where $w(n)$ is any unbounded nondecreasing function.

Angluin and Valiant also give similar algorithmic proofs of thresholds for the existence of perfect matchings and directed Hamiltonian circuits. Karp and Sipser [13] give an algorithm for the construction of matchings in sparse random graphs. The graphs are drawn from $G_{n,p}$ where $p = c/(n-1)$ so that the expected degree of a vertex is the constant c. The algorithm builds up the matching edge-by-edge. As each edge is added, its end points and their incident edges are deleted from the graph. If the current graph has a vertex of degree 1 then the edge to be added is chosen at random from among the edges incident with degree-1 vertices. Otherwise, the edge to be added is chosen at random from the set of all edges. Let M^* be the number of edges in a maximum matching, and let M be the number of edges in the matching produced by the algorithm. Then, for every $\varepsilon > 0$, the following statements hold with probability tending to 1 as $n \to \infty$: $M^* - M < \varepsilon n$ and $|M^* - \lambda(c)n| < \varepsilon n$ where $\lambda(c) = L - W + L(1-W)$ and (L, W) is the least positive fixed point of $L = e^{-\lambda(1-W)}$, $W = 1 - e^{-\lambda L}$. Thus the algorithm tends to construct a matching of nearly maximum size, and the probable size of a maximum matching is sharply predictable as a function of c, the expected degree of a vertex.

3.2. Cliques and colorings. Matula [17] has shown that the size of a maximum clique in a dense random graph can be predicted with remarkable precision. Here "dense" means that p remains fixed as $n \to \infty$, so that a fixed proportion of the possible edges tend to be present. Let $z(n, p) = 2\log_{1/p} n - 2\log_{1/p}\log_{1/p} n + 2\log_{1/p} \frac{1}{2}e + 1$. Let the random variable $Z(n, p)$ denote the size of a maximum clique in a graph drawn from $G_{n,p}$. Then, for every $\varepsilon > 0$,

$$\lim_{n \to \infty} \Pr[z(n, p) - \varepsilon \leqslant Z(n, p) \leqslant z(n, p) + \varepsilon] = 1.$$

Matula's proof uses the second moment method. He considers the random variable $X(n, p, k)$ defined as the number of cliques of size k in a graph drawn from $G_{n,p}$. When $k > z(n, p)$, it follows that $E[X(n, p, k)] \leqslant 1$. Hence, by Markov's inequality, a clique of size k is unlikely to exist. When $k \leqslant z(n, p)$, $E[X(n, p, k)] \geqslant 1$ and $\mathrm{Var}[X(n, p, k)]$ is small; hence, by a variant of Chebyshev's inequality, a clique of size k is very likely to exist. Related results have been obtained by Bollobas and Erdös [4], and Grimmett and McDiarmid [9].

The *sequential algorithm* [9] builds up a clique Q by the following inductive rule: vertex i lies in Q if and only if i is adjacent to every vertex in $Q \cap \{1, 2, \ldots, i-1\}$. With high probability the sequential algorithm

produces a clique of about half the size of a maximum clique. No polynomial-time algorithm is known which performs decisively better. Chvátal [5] has proved an interesting negative result indicating that it is hard to find maximum cliques in random graphs. He considers a broad class of enumerative algorithms, each of which is guaranteed to produce a maximum clique. He shows that, for almost all graphs drawn from $G_{n,p}$, where $p = c/(n-1)$ and c is a sufficiently large constant, the execution time of every algorithm in the class is greater than $\exp(dn)$, where d is a positive constant related to c. Thus, finding maximum cliques appears to be hard not only in the worst case, but almost always.

The chromatic number of a random dense graph has been studied by Erdös and Spencer [8] and by Grimmett and McDiarmid [9]. Let G be drawn from $G_{n,p}$ and let $\chi(G)$ denote the chromatic number of G. Then, for every $\varepsilon > 0$, $(1-\varepsilon)n/z(n, 1-p) \leqslant \chi(G) \leqslant 2(1+\varepsilon)n/z(n, 1-p)$, with probability tending to 1 as $n \to \infty$. The lower bound follows from Matula's results on cliques in random graphs, simply by noting that $\bar{G}$, the complement of G, is a random graph drawn from $G_{n,1-p}$, and that every color class in a coloring of G is a clique in $\bar{G}$. The upper bound follows from the analysis of a sequential coloring algorithm, which goes through the vertices in a single pass, assigning to each vertex as a "color" the least positive integer that has not already been assigned to one of its neighbors. Koršunov has reportedly shown that the chromatic number in fact tends to be close to $n/z(n, 1-p)$. It would be very interesting to have an algorithmic proof of this fact.

It is an open question whether the chromatic number of a random graph can be computed exactly in polynomial expected time. McDiarmid [18] considered a natural class of enumerative algorithms for finding the chromatic number, and showed that every algorithm in the class experiences superexponential growth in its expected running time as a function of the number of vertices. On the other hand, Bender and Wilf [3] show that, for each fixed k and p, a simple enumerative algorithm will decide, in a constant expected number of steps, whether a graph drawn from $G_{n,p}$ is k-colorable. Of course, the answer will almost always be "No" when n is large.

4. The assignment problem and the asymmetric traveling-salesman problem

The *traveling-salesman problem* (TSP) is the problem of constructing a closed path (or *tour*) of minimum total distance through n points, given the distances between all pairs of points. Let c_{ij} denote the distance from i to j. Then the problem can be restated as follows: find a cyclic permuta-

tion π to minimize $\sum_{i=1}^{n} c_{i,\pi(i)}$. We consider the asymmetric case, in which the c_{ij} are nonnegative but are not required to satisfy symmetry ($c_{ij} = c_{ji}$) or the triangle inequality ($c_{ij} \leqslant c_{ik} + c_{kj}$).

The asymmetric TSP is NP-hard, but it can be solved approximately by a *patching algorithm* based on a related problem called the *assignment problem*. The assignment problem asks for a permutation σ, not necessarily cyclic, to minimize $\sum_{i=1}^{n} c_{i,\sigma(i)}$. The assignment problem can be solved in polynomial time.

The patching algorithm converts the optimal solution to the assignment problem to a cyclic permutation by applying a sequence of *patching operations*. Given a permutation τ, the $i-j$ patching operation creates a new permutation τ' defined by: $\tau'(i) = \tau(j)$; $\tau'(j) = \tau(i)$; $\tau'(k) = \tau(k)$ for $k \notin \{i, j\}$. The *cost* of this operation is $c_{i,\tau(j)} + c_{j,\tau(i)} - c_{i,\tau(i)} - c_{j,\tau(j)}$. At a general step the $i-j$ patching operation of minimum cost is selected, subject to the constraint that i lies in the longest cycle of the current permutation and j lies in the second-longest cycle. The effect is to join the two longest cycles together.

On the assumption that the c_{ij} are independent random variables uniformly distributed over $[0, 1]$ the patching algorithm tends to give an excellent approximate solution to the TSP. The random variables of interest are A_n^*, the cost of an optimal assignment, T_n^*, the cost of an optimal tour, and T_n, the cost of the tour produced by the patching algorithm. Clearly $A_n^* \leqslant T_n^* \leqslant T_n$. Computational experiments indicate that A_n^* tends to be close to 1.6. It can be shown that, with probability tending to 1 as $n \to \infty$, $1 + 1/e < A_n^* < 2$. Analysis of the patching algorithm shows that $E[(T_n - T_n^*)/T_n^*] = O(n^{-1/2})$. Underlying the proof is the fact that the optimal assignment is a random permutation of $\{1, 2, \ldots, n\}$. Hence the number of cycles to be patched together tends to be close to $\log n$, and almost all the elements lie in a few long cycles. It follows that the expected cost of the patching operations is small.

It is an open question whether, under the stated probabilistic assumptions, the asymmetric TSP can be solved exactly in polynomial expected time.

5. The Euclidean TSP

In the Euclidean TSP the points to be connected lie in the plane and c_{ij} is the Euclidean distance between i and j. An optimal tour corresponds to a polygon of minimum length through the n given points. The study of

random Euclidean TSPs was initiated by Beardwood, Halton and Hammersley in [2], where the following is proved. Let $\{X_i\}$, $1 < i < \infty$, be independent random variables uniformly distributed over the unit square, and let L_n^* denote the length of a shortest closed path through $\{X_1, X_2, \ldots, X_n\}$. Then there is a constant c such that, with probability 1, $\lim_{n\to\infty}(L_n^*/\sqrt{n}) = c$. A short proof of this result is given in [14].

The study of cellular dissection methods for the approximate solution of random Euclidean TSPs was begun in [11] and [12]. Such a method uses a divide-and-conquer strategy to construct a tour through $\{X_1, X_2, \ldots, X_n\}$, as follows. If R is a subrectangle of the unit square containing r cities, one can make a *vertical cut* dividing R into a left subrectangle and a right subrectangle with a city on their common boundary, such that each of the two subrectangles has at most $r/2$ cities in its interior. Horizontal cuts are similarly defined. These cuts can be used to dissect the unit square into subrectangles, each containing at most t cities, where t is a parameter depending on n. The cuts are performed in rounds. In an odd-numbered round, every subrectangle with more than t cities is divided by a vertical cut. In an even round, horizontal cuts are used. Then, using an enumerative method, an optimal tour is obtained through the t or fewer cities in each subrectangle of the final dissection. It is then an easy matter to create a tour through all n cities, of length less than or equal to the sum of the lengths of the tours in the subrectangles. Let L_n denote the length of the tour so produced. Then $L_n - L_n^* \leqslant a\sqrt{n/(t+1)}$, where a is an absolute constant. If t is chosen proportional to $\log n$ then the algorithm runs in polynomial time and, with probability 1, the relative error $(L_n - L_n^*)/L_n^*$ tends to zero.

Although the cellular dissection method has good asymptotic properties, other fast algorithms tend to produce better tours when applied to random problems with a few hundred cities. For example, the 3-opt method starts with a random tour and repeatedly improves it by replacing up to 3 edges. The process stops when it reaches a local optimum, where no such replacement reduces the length of the tour. Up to now, no one has succeeded in carrying out the probabilistic analysis of such a local improvement method.

References

[1] Angluin D. and Valiant L., Fast Probabilistic Algorithms for Hamiltonian Circuits and Matchings, *J. Comp. Syst. Sci.* **18** (1977), pp. 155–193.

[2] Beardwood J., Halton J. H., and Hammersley J. M., The Shortest Path Through Many Points, *Proc. Cambridge Phil. Soc.* **55** (1959), pp. 299–328.

[3] Bender E. A. and Wilf H. S., A Theoretical Analysis of Backtracking in the Graph Coloring Problem, to appear (1983).
[4] Bollobás B. and Erdös P., Cliques in Random Graphs, *Proc. Cambridge Phil. Soc.* **80** (1976), pp. 419–427.
[5] Chvátal V., Determining the Stability Number of a Graph, *SIAM J. Comp.* **6** (1977), pp. 643–662.
[6] Erdös P. and Rényi A., On the Evolution of Random Graphs, *Pub. Math. Inst. Hung. Acad. Sci.* **5A** (1960), pp. 17–61.
[7] Erdös P. and Rényi A., On the Existence of a Factor of Degree one of Connected Random Graphs, *Acta Math. Acad. Sci. Hung.* **17** (1966), pp. 359–368.
[8] Erdös P. and Spencer J., *Probabilistic Methods in Combinatorics*, Academic Press, New York, 1974.
[9] Grimmet G. R. and McDiarmid C. J. H., On Colouring Random Graphs, *Proc. Cambridge Phil. Soc.* **77** (1975), pp. 314–324.
[10] Karmarkar N. and Karp R. M., The Differencing Method of Set Partitioning, to appear in *Math. of Operations Research* (1984).
[11] Karp R. M., The Probabilistic Analysis of Some Combinatorial Search Algorithms. In: J. F. Traub (ed.), *Algorithms and Complexity*, Academic Press, New York, 1976.
[12] Karp R. M., Probabilistic Analysis of Partitioning Algorithms for the Traveling-Salesman Problem in the Plane, *Math. of Operations Research* **2** (1977), pp. 209–224.
[13] Karp R. M. and Sipser M., Maximum Matchings in Sparse Random Graphs, *Proc. 22nd IEEE Symposium on Foundations of Computer Science*, 1981, pp. 364–375.
[14] Karp R. M. and Steele J. M., Probabilistic Analysis of the Traveling-Salesman Problem, to appear in: E. L. Lawler, J. K. Lenstra, A. H. G. Rinnooy-Kan, D. Shmoys (eds.), *The Traveling-Salesman Problem* (1984).
[15] Komlós J. and Szemerédi, Limit Distribution for the Existence of Hamiltonian Cycles in a Random Graph, *Discrete Applied Math.* **43** (1983), pp. 55–63.
[16] Koršunov A. D., Solution of a Problem of Erdös and Rényi on Hamiltonian Cycles in Nonoriented Graphs, *Dokl. Akad. Nauk SSSR* **228** (1976), pp. 760–764.
[17] Matula D., *The Largest Clique Size in a Random Graph*, Technical Report CS7608, Southern Methodist University 1976.
[18] McDiarmid C., Determing the Chromatic Number of a Graph, *Siam. J. Comp.* **8** (1979), pp. 1–9.
[19] Shamir E., *How Many Random Edges Make a Graph Hamiltonian?*, internal report, Computer Science Department, Hebrew University, 1982.

COMPUTER SCIENCE DIVISION
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

A. A. LETICHEVSKIJ

Abstract Data Types and Finding Invariants of Programs

A program as a mathematical object is a discrete dynamical system that generates processes of computations. In simple dynamical models of sequential computations this system consists of two components: the control component and the information environment.

In this paper the problem of finding invariant relations of programs is considered. Generally, this problem may be formulated as follows. What can we say about the state of the information environment at the moment when the control component is in a given state? It is well known that this question is the main one when we try to prove the correctness of a program by means of the Floyd or Hoare methods. In [9] it was shown that many optimizing procedures reduce to the problem of finding invariants. A similar approach was used in [5].

The solution of the problem under consideration depends on the language that is used to express the properties of the information environment. If this language is a language of the first order predicate calculus then we can easily describe all invariants by using methods of algorithmic logic. However, such a description is very difficult to work with because it may, for instance, use the Gödel numbering of all ways in the program. So it is natural to consider the problem for simple restricted languages. Important examples of such languages are the language of equalities and the language of atomary conditions. These languages are considered here. Some special cases were considered previously in [1, 2, 4, 15].

As a standard model of a program we use here the notion of an interpreted U–Y-scheme of a program or U–Y-program. Let D be a data domain on which operations denoted by symbols of signature Ω and

predicates denoted by symbols of signature Π are defined. Hence D is a universal Ω-algebra and an Ω–Π-algebraic system. Consider a set R of variables and the set $B = D^R$ of memory states. Propositional functions of atomary conditions of the type $\pi(t_1, \ldots, t_n)$ where $\pi \in \Pi$ and $t_1, \ldots, t_n$ are terms constructed of variables by means of operations from Ω, are called *elementary conditions*. An assignment is an expression of the type $(r_1 := t_1, \ldots, r_n := t_n)$, where $r_1, \ldots, r_n$ are variables, $t_1, \ldots, t_n$ – terms over R. In a given memory state the terms assume values in D and elementary conditions in $\{0, 1\}$. Each assignment $y = (r_1 := t_1, \ldots, r_n := t_n)$ defines a transformation of the set B. If $b \in B$, then $b' = y(b)$ is a memory state after the simultaneous assignment of the values $t_1, \ldots, t_n$ computed on b to all of the variables $r_1, \ldots, r_n$. In other words the state b' is defined by the following relations:

$$b'(r_i) = b(t_i), \quad b'(s) = b(s) \quad \text{if } s \neq r_i,\ i = 1, \ldots, n.$$

Let U be some set of elementary conditions and Y a set of assignments. A U–Y-program A is a set of states together with a set of transitions. Each transition is a 4-tuple (a, u, y, a') where $a, a' \in A$ are states of the program, $u \in U$, $y \in Y$. A set A_0 of initial states and a set A^* of terminal states are picked out in A. If (a, u, y, a') is a transition of A, then we write $a \xrightarrow[A]{u/y} a'$ or $a \xrightarrow{u/y} a'$, if A is fixed. A process of computations of a U–Y-program A with a given initial state $b_0 \in B$ of information environment is a finite or infinite sequence of pairs $(a_0, b_0)(a_1, b_1) \ldots$ such that for each pair (a_i, a_{i+1}) there is a transition $a_i \xrightarrow{u/y} a_{i+1}$ and $u(b_i) = 1, b_{i+1} = y(b_i)$. The process is called *initial* if $a_0 \in A_0$, and *terminal* if it is a finite initial process with the last pair (a_n, b_n) such that $a_n \in A^*$. The program is not assumed to be determinate, and so, generally speaking, the next step of the process of computations is not uniquely defined. The program A computes the relation $f_A \subset B^2$, which is defined as follows: $(b, b') \in f_A \Leftrightarrow$ there exists a terminal process $(a_0, b) \ldots (a_n, b')$ such that $a_0 \in A_0$, $a_n \in A^*$.

Suppose that each statement of the language L used to express the properties of the information environment may be expressed by the formula $p(r_1, \ldots, r_n)$ of the first order predicate calculus in which only $r_1, \ldots, r_n$ are free variables and which is interpreted on the domain D. The signatures of functional and predicate symbols of this calculus contain the signatures Ω and Π, respectively. Sentences of the language L will be called *conditions* or *L-conditions*.

The condition $p(r_1, \ldots, r_n)$ is called an *invariant of a state* $a \in A$ if it is true every time when the program is passing through this state, that

is, $p\big(b(r_1), \ldots, b(r_n)\big) = 1$ for each initial process of computations $\ldots(a, b) \ldots$. If some inital conditions $u_a(r_1, \ldots, r_n)$ are given for each initial state $a \in A_0$, then $p(r_1, \ldots, r_n)$ is called an *invariant* (or a *relative invariant* for the given initial conditions) if $p\big(b(r_1), \ldots, b(r_n)\big) = 1$ for any initial process $(a_0, b_0) \ldots (a, b) \ldots$ such that $u_{a_0}\big(b_0(r_1), \ldots, b_0(r_n)\big) = 1$.

If M is a set of L-conditions then $D(M)$ denotes the set of all n-tuples $z = (z_1, \ldots, z_n) \in D^n$ such that, for any condition $u \in M$, it is true that $u(z) = u(z_1, \ldots, z_n) = 1$. Let $R = \{r_1, \ldots, r_n\}$, $y = \big(r_1 := t_1(r), \ldots, r_n := t_n(r)\big)$, $r = (r_1, \ldots, r_n)$. The set $I(M, u, y)$ is defined by the condition: $v(r) \in I(M, u, y) \Leftrightarrow v\big(t(z)\big) = 1$ for all $z \in D(M \cup \{u\})$ where $t(z) = \big(t_1(z), \ldots, t_n(z)\big)$. If M is the set of all invariants of a state a, then $I(M, u, y)$ is the set of all invariants of a state a' if the program includes only one transition $a \xrightarrow{u/y} a'$. If it is possible to compute the sets $I(M, u, y)$ for any elementary condition u and assignment y, then for any state a of the program A the set of invariants I_a can be obtained by the formula

$$I_a = \bigcup_{n=0}^{\infty} M_a^{(n)},$$

where $M_a^{(n+1)} = I(M_{a_1}^{(n)}, u_1, y_1) \cap \ldots \cap I(M_{a_k}^{(n)}, u_k, y_k)$, $a_i \xrightarrow{u_i/y_i} a$, $i = 1, \ldots, k$ being all the transitions that lead to the state a, $M_a^{(0)}$ being the initial invariant set for a. For an initial state a the set $M_a^{(0)}$ may be defined as the set of all consequences from the initial condition for this state, and for other states the set $M_a^{(0)}$ may be defined as the set of all identities of the algebra D. The sets $M_a^{(0)} \subset M_a^{(1)} \subset \ldots$ form an increasing sequence that can be infinite or stabilized at some finite step. If the sequence is infinite, then an approximate solution of the invariants problem can be obtained by the set $M_a^{(n)}$ computed at some finite step. This approach was considered in [10, 11] and applied successfully to solve the invariants problem in some special cases.

The sets $M_a^{(n)}$ are generally infinite. The notion of a basis for a set of L-conditions may be introduced and used for constructive determination of such sets. The set of all L-conditions that are consequences of M on D is called the *D-closure* of M and is denoted by $C_D(M)$. The set M is called D-closed if it is equal to its closure. A subset $N \subset M$ is called a *D-basis* of M if $C_D(N) = M$. The sets $M_a^{(n)}$ are D-closed. If they have finite D-bases, then these bases can be used for constructive determination of the sets $M_a^{(n)}$. So the problem of finding invariants is reduced to the solution of the following three main problems:

(1) Relation problem: a D-basis of a D-closed set M being given, find a D-basis of the set $I(M, u, y)$.

(2) Intersection problem: D-bases of D-closed sets M_1 and M_2 being given, find a D-basis of $M_1 \cap M_2$.

(3) Stabilization problem: D-bases of D-closed sets M_1 and M_2 being given, determine whether $M_1 = M_2$.

All these problems concern the algebra D and their solutions depend only on the properties of this algebra. Moreover, if the language L is the language of equalities or the language of atomary conditions, then the solutions depend only on the properties of a variety or quasivariety that contains the algebra D. So it is possible to use many classical algebraic results for their solution and the algebra D can be replaced by the free algebra of a proper variety or quasivariety. It also means that D is considered as an abstract data type in the manner of ADJ [14].

Let us consider in more details the case of an equality language L. In this case it is natural to ignore elementary conditions and to consider the set $I(M, y) = I(M, 1, y)$ instead of $I(M, u, y)$. Let us denote by $T_D(R)$ the free algebra of the least variety that contains D. Every set M of equalities can be identified with a binary relation defined on the set $T(R)$ of Ω-terms over R. The D-closure of this relation is a congruence of the absolutely free algebra $T(R)$ and since all identities of $T_D(R)$ are contained in this closure, it induces a congruence of $T_D(R)$. Then the quotient algebra $T_D(R)/C_D(M)$ can be constructed. Let $y = (r_1 := t_1, \ldots, r_n := t_n)$ be an assignment. For any equality set M, the homomorphism $\gamma_{y,M}: T_D(R) \to T_D(R)/C_D(M)$ can be defined by setting $\gamma_{y,M}(r_i) = t_i \pmod{C_D(M)}$. In [10] it was shown that the set $I(M, y)$ is the kernel of this homomorphism and the quotient algebra $T_D(R)/I(M, y)$ is isomorphic to the subalgebra $F[t_1, \ldots, t_n]$ of the algebra $F = T_D(R)/C_D(M)$ generated by the elements $t_1, \ldots, t_n$ taken $\bmod C_D(M)$. These facts are basic for constructing the algorithms to compute a basis for the set $I(M, y)$.

The free algebra of some variety is called *hereditarily free* if any of its subalgebras is free in the same variety. Absolutely free algebras, free abelian groups, free groups, and finite-dimensional vector spaces are examples of hereditarily free algebras. For all these examples the sequence $M_a^{(n)}$ always terminates after finitely many steps and algorithms for computing the sets of invariants are constructed. For some of these algorithms, fairly good estimates of the time complexity are obtained. For example, the upper bound for the time complexity of the algorithm of finding invariants in the case of the absolutely free algebra $T_D(R)$ is $O((mn)^2)$ where n is the number of states and m — the number of variables, [6].

In the case of linear spaces and free abelian groups the time complexity is $O(nm^4)$, [7]. In the case of linear spaces, abelian groups and free groups every equality $t_1 = t_2$ can be represented as $t_1 - t_2 = 0$ or $t_1 t_2^{-1} = 1$ and a D-closed set M of equalities can be identified with the corresponding subspace or subgroup. The main part of the algorithm of finding invariants in this case is reduced to the classical problem of constructing a basis of a subspace or of a subgroup. For the free groups the solution of this problem can be obtained by the Nielsen–Schreier algorithm, but the complexity of this algorithm is exponential. In [12] an algorithm with complexity $O(n^3)$ for constructing a basis of a free group is presented. It is interesting to note that this result was obtained by a formal transformation of the Nielsen–Schreier algorithm in the spirit of the ideas of V. M. Glushkov [3]. S. L. Krivoy constructed an algorithm for finding a basis of a intersection of free groups by means of an effective algorithm for finding a basis of a subgroup with complexity $O(n^7)$. This algorithm is used for solving the intersection problem.

The above-mentioned kinds of algebras can be very often met with in practice. Absolutely free algebras, for instance, are used in manipulations of formula or data structures. String manipulations are connected with free semigroups that are not hereditarily free. But every free semigroup can be immersed in a free group and the question of finding the invariants for semigroups is reduced to the same question for groups. Let D be the set of rational numbers. If we use only addition and subtraction in the program, then $T_D(R)$ is the free abelian group generated by R. The introduction of constants (every program uses only a finite number of them) increases the rank of this group. If multiplication by constants is used, then $T_D(R)$ is a linear space. But if multiplication of any two elements of D is allowed, then $T_D(R)$ is the ring of polynomials with integer coefficients. This algebra is not hereditarily free. However, the classical results of commutative algebra can be used. Each equality in the algebra of polynomials can be represented as $t = 0$ and therefore can be identified with an element of $T_D(R)$. Every D-closed set is an ideal of $T_D(R)$. Hence, by the theorem of Hilbert, each ideal has a finite basis and the sequence $M_a^{(0)} \subset M_a^{(1)} \subset \ldots$ is stabilized after a finite number of steps. The solution of the main problem can be obtained from the results of [13].

If L is an atomary condition language, then D must be considered as an algebraic system. In this case $T_D(R)$ is the free algebraic system of the least variety that contains D and $J(M, u, y)$ is the kernel of the homomorphism $\gamma_{M,u,y}$: $T_D(R) \to T_D(R)/C_D(M \cup \{u\})$ defined by condition $\gamma_{M,u,y}(r_i) = t_i \left(\mathrm{mod}\, C_D(M \cup \{u\})\right)$.

The language of linear inequalities for programs that use addition, subtraction and multiplication by constants is an important case of the problem of finding invariants of atomary conditions in practical applications. The system $T_D(R)$ in this case is a linear space with the inequality $t_1 \leqslant t_2$, which is true only if $t_1 = t_2$. Every set of conditions can be identified with a subset of $T_D(R)$, inequalities of the type $t_1 \leqslant t_2$ being transformed to $t_1 - t_2 \leqslant 0$ and identified with the elements $t_1 - t_2$ of $T_D(R)$. The sets $M_a^{(n)}$ are in this case linear convex cones generated by a finite number of generators and the sequence $M_a^{(0)} \subset M_a^{(1)} \subset \ldots$ can be defined constructively. But it is easy to give many examples when the above sequence is infinite. The problem of determining the limit for $M_a^{(n)}$ is difficult and a satisfactory approximate solution can be obtained by using $M_a^{(n)}$ with a sufficiently large n. Essential influence on the result is exerted by the choice of initial sets $M_a^{(0)}$. To construct them other methods of finding invariants can be used, for instance the method of [2].

References

[1] Caplain M., Finding Invariant Assertions for Proving Programs, in *Proc. Intern. Conf. Reliable Software,* Los Angeles, 1975, pp. 165–171.

[2] Cousot P. and Halbwachs N., Automatic Discovery of Linear Restraints Among Variables of Program, in *Conf. Rec. of the 5-th Annual ACM Symposium on Principles of Prog. Lang.*, 1978, Jan. 23–25, pp. 84–96.

[3] Glushkov V. M., On Formal Transformations of Algorithms, in *Lect. Notes in Comp. Sci.* **122** (1981), pp. 430–440.

[4] Karr M., Affine Relationships Among Variables of a Program, *Acta Informatica* **6** (1976), pp. 133–151.

[5] Kildall G. A., A Unified Approach to Global Optimization, in *Conf. Rec. of ACM Symp. on Princ. of Programming Lang.*, Boston, Massachusetts, October 1–3 (1973), pp. 194–206.

[6] Кривой С. Л., О поиске инвариантных соотношений в программах, В сб. *Математическая теория проектирования вычислительных машин,* ИК АН УССР, Киев, 1978, pp. 35–51.

[7] Кривой С. Л., Об одном алгоритме поиска инвариантных соотношений в программах, *Кибернетика* (1981), No. 5, pp. 12–18.

[8] Кривой С. Л., Об алгоритме построения базиса пересечения конечно-порожденных свободных групп, *Кибернетика* (1982), No. 4, pp. 5–10.

[9] Letichevskij A. A., Equivalence and Optimization of Programs, in *Lecture Notes in Computer Science* **5** (1974), pp. 111–123.

[10] Летичевский А. А., Об одном подходе к анализу программ, *Кибернетика* (1979), No. 6, pp. 1–8.

[11] Letichevskij A. A., On Finding Invariant Relations of a Program, in *Lect. Notes in Comp. Science* **122** (1981), pp. 304–314.

[12] Летичевский А. А., Годлевский А. Б., Кривой С. Л., Об эффективном алгоритме построения базиса подгруппы свободной группы, *Кибернетика* (1981), No. 5, pp. 107–116.

[13] Seidenberg A., Constructions in Algebra, *Trans. Amer. Math. Soc.* **197** (1974), pp. 273–313.

[14] Wagner E. G., Thatcher J. W., and Wright J. B., Programming Languages as Mathematical Objects, in *Lect. Notes in Comp. Sci.* **64** (1978), pp. 84–101.

[15] Wegbreit S., Property Extraction in Well Bounded Property Sets, *IEEE Trans. on Soft Eng.*, vol. SE-1 **3** (1975), pp. 270–285.

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

ROBERT E. TARJAN

Efficient Algorithms for Network Optimization

This paper is a survey of recent improvements in algorithms for four classical network optimization problems. The problems we consider are those of finding minimum spanning trees, shortest paths, maximum network flows, and maximum matchings. For each problem we summarize the history of work on the problem and the current state of the art. We conclude by discussing the techniques that have led to the most efficient known algorithms.

1. Introduction

The field of combinatorial algorithms has flourished in recent years as computer scientists and others have concentrated on the development and analysis of efficient algorithms. We shall survey the fruits of this labor in one area, that of network optimization. A network is a graph, either directed or undirected, in which the edges (and possibly the vertices) have associated real numbers representing for example costs or capacities. The goal of a network optimization problem is to find a subgraph of a given network that satisfies certain constraints and maximizes or minimizes some function of the edge numbers. Network optimization has many obvious and not-so-obvious applications in such areas as the design of telephone, highway, and computer networks, the routing of traffic and produce, assignment of workers to tasks, resource allocation, and scheduling.

Many important network optimization problems, including the notorious minimum tour or "traveling salesman" problem, are NP-complete ([37]) and thus unlikely to have polynomial-time algorithms. However, there *are* efficient algorithms for many other such problems, including the four we shall study: finding minimum spanning trees (Section 2), finding shortest paths (Section 3), finding maximum flows (Section 4), and finding maximum matchings (Section 5). For each of these problems, we provide a brief historical survey and an examination of the most efficient currently known algorithms. In Section 6 we draw some conclusions about the general techniques that are used in the best algorithms. Further discussion of

network optimization problems can be found in the survey papers of Klee [49] and Tarjan [70], [71], and the books of Lawler [55], Papadimitriou and Steiglitz [62], and Tarjan [72].

In our discussion we shall use standard graph-theoretic terminology; see [55], [62], [72]. When stating time bounds for algorithms, we shall use n, m, and N to denote the number of vertices, the number of edges, and the maximum absolute value of any edge number in the problem graph, respectively. We assume $n \geqslant 2$ and $m \geqslant (1+\varepsilon)n$ for some fixed positive ε. (This simplifies some of the time bound formulas.) We assume a *random access machine* ([1], [10]) as our model of computation. We use the *uniform cost measure* of running time: each arithmetic or logical operation requires one unit of time, independent of the magnitudes of the numbers involved.

We must be careful in using this cost measure. If the machine is allowed to manipulate numbers of arbitrary size or precision in unit time, then it can perform hidden parallel computation by encoding several numbers into one. We can prevent this by charging for an operation a time proportional to the number of bits (binary digits) needed to represent the operands (the *logarithmic cost measure*). Alternatively, we can limit the size of integers allowed to those representable in $O(\log n)$-bits and restrict the operations we allow on the edge values. We shall adopt the latter approach; all the algorithms we shall study in subsequent sections manipulate only $O(\log n)$-bit integers and use only comparison, addition, and sometimes multiplication of edge values, with no clever encoding.

The fundamental distinction involved here is whether we wish to treat real numbers as having infinite precision, with unit cost per arithmetic operation, or as having finite precision, with a cost per arithmetic operation proportional to the number of bits. To illustrate this distinction, let us consider the *linear programming problem*: minimize the function $\sum_{i=1}^{n} c_i x_i$, for variables $x_1, x_2, \ldots, x_n$ satisfying the inequalities $\sum_{i=1}^{n} a_{ij} x_i \leqslant b_j$ for $1 \leqslant j \leqslant m$, where the c_i's, a_{ij}'s and b_j's are real numbers.

The simplex algorithm of Dantzig [11] solves linear programming problems very efficiently in practice and on the average; it assumes infinite precision real numbers with the unit cost measure. However, carefully constructed examples show that the simplex algorithm runs in exponential time in the worst case ([50]). On the other hand, the "ellipsoid" method was recently shown by Khachian [48] to run in polynomial time in the worst case, for finite precision real numbers with logarithmic cost measure. Khachian's algorithm is apparently much slower than the simplex algorithm in practice ([12]). It is still an open problem to determine

whether there is an algorithm for linear programming that runs in polynomial time for infinite precision real numbers with unit cost measure.

2. Minimum spanning trees

Let G be a connected, undirected graph, each of whose edges e has a real-valued *cost* $c(e)$. The *minimum spanning tree problem* is that of finding a spanning tree of G of minimum total edge cost. Of the problems we shall consider, this one has the longest history; the first fully realized minimum spanning tree algorithm was presented by Borůvka in 1926 ([5]). Graham and Hell's paper [39] is an excellent historical survey.

All the known efficient minimum spanning tree algorithms are special cases of a general greedy method, in which we build up a minimum spanning tree edge-by-edge, including appropriate small-cost edges and excluding appropriate large-cost ones. We shall formulate this method as an edge-coloring process. We begin with all edges uncolored and repeatedly apply one or the other of two rules, which color an uncolored edge either blue (accepted) or red (rejected). In order to formulate the blue rule, we need the concept of a *cut*. A cut in a graph $G = (V, E)$ is a partition of the vertex set V into two nonempty parts, X and $\bar{X} = V - X$. An edge *crosses* the cut if it has exactly one endpoint in each part. The coloring rules are as follows:

BLUE RULE. *For any cut that no blue edges cross, select a minimum-cost uncolored edge crossing the cut, and color it blue.*

RED RULE. *For any simple cycle containing no red edges, select a maximum-cost uncolored edge on the cycle and color it red.*

This coloring algorithm maintains the invariant that there is always a minimum spanning tree containing all of the blue edges and none of the red ones. Furthermore, as long as at least one uncolored edge remains, some rule is applicable. It follows that the algorithm colors all the edges, and that when the algorithm stops the blue edges define a minimum spanning tree. (For proofs see [72].)

As the algorithm proceeds, the currently blue edges define a set of trees that we shall call the *blue trees*. Initially each vertex is in a one-vertex blue tree. As edges are colored blue, the blue trees merge to form bigger blue trees, until finally only a single blue tree spanning all the vertices remains. We obtain different versions of the algorithm by altering the order in which the rules are applied.

Most of the known efficient versions of this algorithm emphasize the blue rule. There are three "classical" methods. Perhaps the simplest is due to Kruskal [53]: color the edges in nondecreasing order by cost,

coloring an edge blue if its endpoints are in different blue trees and red otherwise.

Efficient implementation of Kruskal's algorithm requires two data structures. We need a data structure to keep track of the vertex sets of the blue trees; these sets are updated by union operations. Any of the standard disjoint set methods (see [1], [69], [73]) will suffice for this purpose. We also need a method for sorting all the edges, or at least of repeatedly obtaining a smallest remaining uncolored edge. The time for sorting edges dominates the time for manipulating vertex sets, and Kruskal's algorithm runs in $O(m \log n)$ time. An intriguing implementation due to Brennan [6] performs the sorting concurrently with the edge coloring using Hoare's QUICKSORT algorithm ([40]), running the sorting algorithm only long enough to identify the successive smallest edges needed by the coloring process. Brennan reports good experimental results with this method.

If the edges are presorted, or if they can be sorted fast (e.g. the costs are small integers and thus radix sorting ([1], [51]) can be used), then the set manipulation time dominates the running time of Kruskal's algorithm. In this case the total time is $O(m\alpha(m, n))$, where α is an inverse of Ackermann's function ([69]).

The second and most recent of the classical methods was discovered by Jarník ([42]) and independently rediscovered by Prim ([65]) and by Dijkstra ([15]): For a fixed start vertex s, repeatedly apply the blue rule to the cut one of whose parts is the vertex set of the blue tree containing s. Since throughout the process there is only one blue tree containing more than one vertex, this algorithm does not need a data structure to represent the vertex sets of the blue trees. The most efficient implementations of the method maintain, for each vertex v not yet in the blue tree T containing s, a minimum-cost edge $\{u, v\}$ such that u is in T; this edge is a candidate to become blue. The general step is to select the minimum-cost candidate edge, color it blue, and update the set of candidate edges. If the set of candidate edges is stored as an unordered set, each iteration of the general step takes $O(n)$ time, and the total running time of the algorithm is $O(n^2)$. If the set of candidate edges is stored as a heap ([43], [75]), so that finding the minimum-cost candidate edge is an inexpensive operation, then the total running time is $O(m \log_{m/n} n)$ ([43], [72]).

The third classical method is that of Borůvka [5], independently rediscovered by Choquet ([9]), Łukaszewicz *et al.* ([56]), and Sollin ([4]). The method consists of repeating the following step until there is only one blue tree: for every blue tree, select a minimum-cost uncolored edge with exactly one endpoint in the tree, and color all the selected edges blue.

An edge can be selected twice in the same iteration of the general step, once for each of its endpoints; it is of course only colored once. As stated the method is a parallel, not a serial algorithm, and it is guaranteed correct only if the edge costs are distinct. We can handle nondistinct edge costs by assigning identifiers to the edges and ordering the edges lexicographically by cost and identifier. If we do this, and also color the edges one-at-a-time, the method can be regarded as a special case of the general greedy algorithm.

Implementing the method requires maintaining, for each blue tree, the set of vertices it contains and the set of uncolored edges incident to at least one vertex in the tree. To store the vertex sets, we can use any of the standard disjoint set data structures. To store the uncolored edge sets, we can use any data structure for meldable heaps ([1], [51], [72]) (sometimes called mergeable priority queues). With a careful implementation and a non-trivial analysis, one can obtain a version of Borůvka's algorithm that runs in $O(m \log\log n)$ time. Yao ([76]) devised the first such method; a simplified method was proposed by Cheriton and Tarjan ([7]).

This method has the following drawback: During the process of finding a minimum-cost edge with only one endpoint in a blue tree, we may encounter many edges with both endpoints in the blue tree (such edges may be colored red). The algorithm spends most of its time examining such potentially red edges. If the graph is sufficiently dense (i.e. the ratio m/n is high), one can improve the algorithm by intermittently carrying out global "purges" that color red all uncolored edges except a minimum-cost edge joining each pair of blue trees that have at least one edge between them. This addition improves the running time of the algorithm to $O(m \log\log_{m/n} n)$, making it asympotically as fast as any algorithm on both sparse and dense graphs ([7], [72]).

The greedy method that is at the heart of all the minimum spanning tree algorithms can be substantially generalized. The standard generalization, discovered independently by Edmonds ([20]), Gale ([33]), and Welsh ([74]), is to matroids. Recently Korte and Lovász have invented an even more general structure on which the method works, called the *greedoid* ([52]).

3. Shortest paths

Let G be a directed graph, each of whose edges $e = (v, w)$ has a non-negative *length* denoted by $l(e)$ or $l(v, w)$. The *length of a path* p consisting of a sequence of edges $e_1, e_2, \ldots, e_k$ is $l(p) = \sum_{i=1}^{k} l(e_i)$. The *shortest path problem*

is that of determining, for each member of a specified set of vertex pairs s, t, a path from s to t of minimum length, called a *shortest path.* The length of a shortest path is the *distance* from s to t, which we denote by $d(s, t)$. We shall discuss three versions of this problem:

SINGLE PAIR. *For a single pair of vertices s, t, find a shortest path from s to t.*

SINGLE SOURCE. *For a given source vertex s, find a shortest path from s to v for every vertex v.*

ALL PAIRS. *For every pair of vertices s, t, find a shortest path from s to t.*

Of these problems, the single source problem is the fundamental one: all known single pair algorithms at least partially solve a single source problem, and the all pairs problem can be solved as n single source problems. Therefore we begin with the single source problem.

For simplicity let us assume that every vertex is reachable from the source vertex s. Shortest paths have two important properties that are useful in algorithms for finding them. There are shortest paths from s to every vertex if and only if G contains no negative cycle (a cycle of negative length). If there are such shortest paths, there is a spanning tree rooted at s containing shortest paths from s to every vertex. Such a tree is called a *shortest path tree.* We can regard the goal of an algorithm for the single source problem as to exhibit either a shortest path tree or a negative cycle. If a shortest path tree exists, it suffices to compute $d(s, v)$ for every vertex v, since a shortest path tree is easy to construct in $O(m)$ time given these distances.

Ford ([26], [27]) proposed a general algorithm for the single-source problem that can be regarded as a special case of the simplex algorithm (see [62]): Begin with $d(s, s) = 0$, $d(s, v) = \infty$ for $v \neq s$, and repeat the following step until it no longer applies:

LABELING STEP. *Select an edge (v, w) such that $d(s, v) + l(v, w) < d(s, w)$, and replace $d(s, w)$ by $d(s, v) + l(v, w)$.*

This algorithm terminates with correct distances if and only if there is no negative cycle. The efficiency of the algorithm depends heavily on the order of edge selection. Most efficient implementations are versions of the following refinement, which we call the *labeling and scanning algorithm* ([38]). Each of the vertices is in one of three states: *unlabeled, labeled,* or

scanned. Initially the source vertex is labeled and every other vertex is unlabeled. We repeat the following step until there are no labeled vertices.

SCANNING STEP. *Select a labeled vertex v and scan it, thereby converting it to the scanned state, by applying the labeling step to every edge (v, w) such that $d(s, v)+l(v, w) < d(s, w)$. (Such an application converts w to the labeled state.)*

There are three theoretically important versions of the labeling and scanning algorithm, distinguished by the order of vertex scanning and appropriate for different kinds of graphs. The first, *breadth-first scanning*, maintains the set of labeled vertices as a queue, removing vertices for scanning from the front and adding newly labeled vertices to the rear. This method, a reformulation of algorithms discovered independently by Moore ([59]) and Bellman ([2]), runs in $O(nm)$ time if the graph is free of negative cycles. With appropriate modifications, the algorithm will locate a negative cycle in $O(nm)$ time if there is one. A variant of breadth-first scanning that seems to work well in practice ([14]) is to put vertices converted from scanned to labeled on the front instead of the rear of the queue ([63]). Unfortunately, this method runs in exponential time in the worst case ([47]).

If all edge lengths are non-negative, *shortest-first scanning*, proposed by Dijkstra ([15]) gives a better worst-case bound: scan the labeled vertex v such that $d(s, v)$ is minimum. With this method each vertex is only scanned once. The method is analogous to the Jarník–Prim–Dijkstra minimum spanning tree algorithm and has essentially the same implementation and the same time bound, namely $O(m\log_{m/n} n)$ ([44], [72]).

The third important scanning order, *topological scanning*, is appropriate when the graph has no cycles at all: scan the vertices in topological order, i.e. any order such that, if (v, w) is an edge, v is scanned before w. Topological scanning is used in the well-known program evaluation and review technique (PERT) and produces shortest paths in $O(m)$ time.

If the edge lengths are integers, the single source problem can also be solved by scaling techniques, as recently discovered by Gabow ([31]). Gabow's most interesting result is for the single source problem with general integer edge lengths. He gives an $O(n^{3/4}m\log N)$ algorithm (recall that N is the maximum absolute value of any edge length) that works by reducing the shortest path problem to an assignment problem and solving the assignment problem by scaling (see Section 5).

We can solve the single pair problem for a pair s, t by solving a single source problem for s, running the algorithm until the distance from s to t is

known. Alternatively, we can reverse the directions of all edges and solve a single source problem for t. We can even combine these methods by simultaneously growing a shortest path tree forward from s and the reverse of a shortest path tree backward from t, stopping when the two trees overlap in an appropriate way. Although this *bidirectional search* technique ([60], [66]) does not improve the worst-case running time, it can reduce the time in practice.

For very large sparse graphs such as arise in artificial intelligence applications, there is not enough time in practice even to examine the entire graph. For such situations AI researchers have proposed various *heuristic search* techniques for the single source problem. These are intended to examine only vertices likely to be on a shortest path from s to t. The efficiency of both undirectional and bidirectional heuristic search depends on how easy it is to compute a good distance approximation. See [13], [61].

There are two main algorithms for the all pairs problem, one suited for very dense graphs, the other for sparse graphs. Floyd ([24]) proposed a straightforward dynamic programming algorithm that runs in $O(n^3)$ time: Initialize $d(v, v) = 0$, $d(v, w) = l(v, w)$ if (v, w) is an edge, and $d(v, w) = \infty$ if $v \neq w$ and (v, w) is not an edge. Then, for each vertex u, carry out the following step:

LABELING STEP. *If $d(u, u) < 0$, stop: there is a negative cycle. Otherwise, for each pair of vertices v, w such that $d(v, u) + d(u, w) < d(v, w)$, replace $d(v, w)$ by $d(v, u) + d(u, w)$.*

Another more complicated method is to solve one single source problem using breadth-first scanning, use the distances so computed to transform the edge lengths so that they are all nonnegative, and repeatedly use shortest-first scanning. The edge length transformation, which preserves shortest paths, is based on linear programming duality. This method runs in $O(nm \log_{m/n} n)$ time (see [72]).

Fredman [28] has devised another all pairs algorithm that runs in $O\left(n^3 (\log\log n)^{1/3} / (\log n)^{1/3}\right)$, which though very interesting theoretically is apparently too complicated to be practical. Further information on early shortest path algorithms can be found in [17].

4. Maximum flow

Let $G = (V, E)$ be a directed graph with two distinguished vertices, a *source* s and a *sink* t, each of whose edges e has a non-negative *capacity* $c(e)$. A *flow* on G is a non-negative function on the edges such that

$0 \leqslant f(e) \leqslant c(e)$ for every edge e and $\sum_{(u,v)\in E} f(u, v) = \sum_{(v,w)\in E} f(v, w)$ for every vertex $v \notin \{s, t\}$.

The *value* of a flow f is $\sum_{(s,v)\in E} f(s, v)$. The *maximum flow problem* is that of finding a flow of maximum value, called a *maximum flow.*

The fundamental theory of network flows was developed by Ford and Fulkerson ([25], [27]) and is an outgrowth of linear programming. As in the case of minimum spanning trees we need the notion of a *cut*, which for network flows we define to be a vertex partition X, $\bar{X}$ such that $s \in X$ and $t \in \bar{X}$. The *capacity* of the cut is $\sum_{\substack{v\in X, w\in \bar{X} \\ (v,w)\in E}} c(v, w)$; if f is a flow, the *flow across* the *cut* is $\sum_{\substack{v\in X, w\in \bar{X} \\ (v,w)\in E}} f(v, w) - \sum_{\substack{v\in \bar{X}, w\in X \\ (v,w)\in E}} f(v, w)$. The flow across any cut is equal to the flow value and is at most the capacity of the cut. Ford and Fulkerson's main result, the *max-flow, min-cut theorem*, states that the maximum flow value equals the minimum cut capacity.

Ford and Fulkerson proved this theorem by devising an algorithm that, given a flow f, either finds a cut whose capacity equals the flow value or finds a way to increase the flow value. The algorithm uses a *residual graph* R, whose vertex set is V and whose edge set contains two kinds of edges: for each edge $(v, w) \in E$ such that $f(v, w) < c(v, w)$, a *forward edge* (v, w) with *residual capacity* $r(v, w) = c(v, w) - f(v, w)$; and, for each edge $(v, w) \in E$ such that $f(v, w) > 0$, a *backward edge* (v, w) with residual capacity $r(w, v) = f(v, w)$. (Technically R is a multigraph, i.e. it may contain multiple edges.)

If $(v, w) \in R$ is a forward edge, we can increase the net flow from v to w in G by up to $r(v, w)$ units by increasing the flow on (v, w); if $(w, v) \in R$ is a backward edge, we can increase the net flow from w to v in G by up to $r(w, v)$ units by decreasing the flow on (v, w). The flow f is maximum if and only if there is no path from s to t in G. If there is such a path, called an *augmenting path*, we can increase the value of f by altering the flow on the corresponding edges in G. If there is no augmenting path, the set of vertices reachable from s in R defines a cut whose capacity equals the value of f.

Ford and Fulkerson's *augmenting path method* for finding a maximum flow consists of beginning with the zero flow and repeating the following step until it no longer applies:

AUGMENTING STEP. *If t is reachable from s in the residual graph R for the current flow, find an augmenting path p, let Δ be the minimum of the residual capacities of the edges on p, and increase the flow value by Δ by altering the flows along the edges in G corresponding to the edges on p.*

If all capacities are integers, Ford and Fulkerson's algorithm produces an integer maximum flow in $O(nN)$ augmentations. (Recall from Section 1 that N is the maximum capacity. Each augmentation increases the flow value by at least one unit and at most $(n-1)N$ units can flow through the at most $n-1$ edges leaving s). Constructing R and performing a single augmentation takes $O(m)$ time, so the total time bound is $O(nmN)$.

Unfortunately, if the capacities are arbitrary real numbers, the algorithm need never terminate, and successive flow values, though they will converge, need not converge to the maximum flow value ([27]). However, by careful choice of augmenting paths, the method can be made efficient. Edmonds and Karp ([21]) proposed two methods. If augmentations are made along paths of fewest edges, the number of augmentations is $O(nm)$, giving an overall time bound of $O(nm^2)$. (Each augmentation takes $O(m)$ time.) If augmentations are made along paths maximizing Δ, the increase in flow value, the number of augmentations is $O(m \log N)$ and the overall time bound is $O\left(m^2(\log_{m/n} n)(\log N)\right)$, assuming integer capacities. (Finding each augmenting path requires running an algorithm analogous to Dijkstra's shortest path algorithm.)

Further improvements in maximum flow algorithms are all based on the work of Dinic ([16]). Dinic, who worked independently of Edmonds and Karp, showed that the maximum flow problem can be reduced to the solution of at most $n-1$ blocking flow problems on acyclic graphs. By a blocking flow we mean a flow such that in the residual graph there is no augmenting path containing only forward edges. (That is, to increase the flow value we must reroute some of the flow.) Dinic proposed a blocking flow algorithm with a running time of $O(nm)$, giving a time bound of $O(n^2m)$ for the maximum flow problem. Improved algorithms for finding a maximum flow using Dinic's approach were discovered by Karzanov ([46]) ($O(n^3)$), Cherkasky ([8]) ($O(n^2m^{1/2})$), Malhotra *et al.* ([57]) ($O(n^3)$ but simpler than Karzanov's method), Galil ([34]) ($O(n^{5/3}m^{2/3})$), Galil and Naamad ([36]) ($O\left(nm(\log n)^2\right)$), and Sleator and Tarjan ([67], [68]) ($O(nm \log n)$). The last of these methods achieves its speed mainly through the use of very sophisticated data structures. The most recent maximum flow algorithm, recently devised by Gabow ([31]), uses a simple scaling technique and runs in $O(nm \log N)$ time, assuming integer capacities.

A more complicated network flow problem that is not as well understood as the maximum flow problem is the minimum cost flow problem. Each edge e, in addition to having a capacity $c(e)$, has a cost $a(e)$ per unit flow; we seek a maximum flow f minimizing the total cost $\sum_{e \in E} a(e)f(e)$. One way to find a minimum cost flow is to use the augmenting path

method, choosing augmenting paths that minimize the incremental cost. This method, though quite satisfactory in practice, is not efficient in the worst case; indeed, it has the same convergence problems as Ford and Fulkerson's original maximum flow algorithm. Edmonds and Karp ([21]) proposed a scaling algorithm that finds a minimum cost flow in a time bound polynomial in n, m, and $\log N$, assuming integer capacities. The problem of finding a polynomial-time minimum cost flow algorithm assuming arbitrary infinite-precision real-valued capacities under the unit cost complexity measure is still open. Further information on network flow problems and their applications can be found in [64].

5. Matching

Let G be an undirected graph with a real-valued *weight* $w(e)$ on each edge e. A *matching* is a set of edges no two of which have a common endpoint. A vertex is *matched* if it is an endpoint of an edge in the matching and *free* otherwise. The *maximum matching problem* is that of finding a matching whose edges have maximum total weight. There are four important versions of this problem:

MAXIMUM CARDINALITY BIPARTITE MATCHING. *G is bipartite* (i.e., there is a vertex partition X, $\bar{X}$ such that all edges have one endpoint in X and one in $\bar{X}$), *and all edge weights are one.*

THE ASSIGNMENT PROBLEM. *G is bipartite; the edge weights are arbitrary.*

MAXIMUM CARDINALITY NONBIPARTITE MATCHING. *G is arbitrary; all edge weights are one.*

WEIGHTED NONBIPARTITE MATCHING. *Both G and the edge weights are arbitrary.*

As in the case of network flows, augmenting paths play an important role in matching algorithms. An *augmenting path* p with respect to a matching is a simple path from one free vertex to another whose edges are alternately in the matching and not in the matching. By adding to the matching every unmatched edge on p and deleting every matched edge on p, we can increase the cardinality of the matching by one. Berge ([3]) proved that a matching is of maximum cardinality if and only if there is no augmenting path. Thus we can find a maximum cardinality matching by

starting with the empty matching and repeatedly finding an augmenting path and swapping its matching and nonmatching edges.

In the case of bipartite graphs, it is easy to find an augmenting path, if one exists, in $O(m)$ time. Thus maximum cardinality bipartite matchings can be found in $O(nm)$ time. Hopcroft and Karp ([41]) discovered a faster algorithm for this problem that finds all augmenting paths of a given length at once (where length is measured by the number of edges) and proceeds from shortest to longest length augmenting paths. With this method, most of the augmenting paths are short, and the overall time bound is $O(\sqrt{n}\,m)$. The Hopcroft–Karp algorithm can be interpreted as Dinic's algorithm applied to an appropriate network; a maximum-value flow in the network corresponds to a maximum cardinality matching in the bipartite graph ([23]).

On nonbipartite graphs it is much harder to find augmenting paths. Edmonds ([18]) discovered an elegant method that involves shrinking certain odd cycles, called *blossoms*. With improvements in implementation suggested by Gabow [29], [30], Edmonds' algorithm finds maximum cardinality nonbipartite matchings in $O(nma(m, n))$ time. (Recall from Section 2 that a is an inverse of Ackermann's function.) Use of a recently discovered linear-time set union algorithm ([32]) reduces the running time to $O(nm)$. Even and Kariv ([22]), in a remarkable achievement, extended the Hopcroft–Karp algorithm to the nonbipartite case, obtaining a running time of either $O(n^{2.5})$ or $O(\sqrt{n}\,m\log\log n)$, depending upon the exact implementation details. Micali and Vazirani ([58]) obtained a simplified algorithm with an improved running time of $O(\sqrt{n}\,m)$, matching the best time bound for the bipartite case.

When the weights are arbitrary, augmenting paths must be selected in an order that depends upon the weight. We define the *weight* of an augmenting path to be the sum of the weights of its matching edges minus the sum of the weights of its nonmatching edges. We can find a maximum weight matching by beginning with the empty matching and repeatedly augmenting using a maximum-weight augmenting path, continuing until no augmenting path has positive weight.

This algorithm can be implemented on bipartite graphs using Dijkstra's shortest path algorithm to find maximum-weight augmenting paths and transforming the edge weights after each augmentation to make them nonpositive. The resultant algorithm, generally called the *Hungarian method* ([54]), runs in $O(nm\log_{m/n} n)$ time ([72]). Edmonds, using his blossom shrinking technique, extended the algorithm to nonbipartite graphs ([19]).

Depending upon the implementation, the algorithm runs in either $O(n^3)$ time or $O(nm\log n)$ time ([29], [35]).

The assignment problem for integer edge weights can also be solved by scaling. Gabow ([31]) has a method that runs in $O(n^{3/4}m\log N)$ time. Whether the method extends to nonbipartite graphs is an open problem.

6. Conclusions

The various efficient algorithms for network optimization combine at least four important kinds of ideas. First are ideas arising from linear programming, in particular duality and the simplex algorithm. The augmenting path technique can be viewed as a combinatorial expression of this algorithm. Second is the greedy method, which is not only explicitly used in the various minimum spanning tree algorithms but is also implicit in Dijkstra's shortest path algorithm. Third are sophisticated data structures for representing ordered and unordered sets, trees, and graphs, especially as they change over time. Fourth is the idea of scaling.

There is no reason to believe that any of the algorithms we have discussed is asymptotically optimal; indeed further improvements may well be possible. It seems likely that such improvements will come from the development of more sophisticated data structures and from further exploitation of scaling. Perhaps an even more important line of research is to study the practical efficiency of the various algorithms to determine the effect of theoretical improvements on actual performance.

References

[1] Aho A. V., Hopcroft J. E., and Ullman J. D., *The Design and Analysis of Computer Algorithms*, Addison-Wesley, Reading, MA, 1974.

[2] Bellman R. E., On Routing Problem, *Quart. Appl. Math.* **16** (1958), pp. 87–90.

[3] Berge C., Two Theorems in Graph Theory, *Proc. Nat. Acad. Sci. U.S.A.* **43** (1957), pp. 842–844.

[4] Berge C. and Ghouila-Houri A., *Programming, Games, and Transportation Networks*, John Wiley and Sons, New York, NY, 1965.

[5] Borůvka O., O Jistém Problému Minimálním, *Práca Moravské Přírodovědecké Společnosti* **3** (1926), pp. 37–58.

[6] Brennan J. J., Minimal Spanning Trees and Partial Sorting, *Op. Res. Letters* **1** (1982), pp. 113–116.

[7] Cheriton D. and Tarjan R. E., Finding Minimum Spanning Trees, *SIAM J. Comput.* **5** (1976), pp. 724–742.

[8] Cherkasky R. V., Algorithm of Construction of Maximal Flow in Networks with Complexity of $O(V^2\sqrt{E})$ Operations (in Russian), *Mathematical Methods of Solution of Economical Problems* **7** (1977), pp. 112–125.

[9] Choquet G., Étude de certains réseaux de routes, *C. R. Acad. Sci. Paris Sér. A* **206** (1938), pp. 310–313.
[10] Cook S. A. and Reckhow R. A., Time-Bounded Random Access Machines, *J. Comput. Sys. Sci.* **7** (1973), pp. 354–375.
[11] Dantzig G. B., *Linear Programming and Extensions*, Princeton University Press, Princeton, NJ, 1963.
[12] Dantzig G. B., Comments on Khachian's Algorithm for Linear Programming, Technical Report SOR 79–22, Dept. of Operations Research, Stanford University, Stanford, CA, 1979.
[13] de Champeaux D., Bidirectional Heuristic Search Again, *J. Assoc. Comput. Mach.* **30** (1983), pp. 22–32.
[14] Dial R., Glover F., Karney D., and Klingman D., A Computational Analysis of Alternative Algorithms for Finding Shortest Path Trees, *Networks* **9** (1979), pp. 215–248.
[15] Dijkstra E. W., A Note on Two Problems in Connexion with Graphs, *Numer. Math.* **1** (1959), pp. 269–271.
[16] Dinic E. A., Algorithm for Solution of a Problem of Maximum Flow in a Network with Power Estimation, *Soviet Math. Dokl.* **11** (1970), pp. 1277–1280.
[17] Dreyfus S. E., An Appraisal of Some Shortest-Path Algorithms, *Operations Research* **17** (1969), pp. 395–412.
[18] Edmonds J., Paths, Trees, and Flowers, *Canad. J. Math.* **17** (1965), pp. 449–467.
[19] Edmonds J., Matching and a Polyhedron with 0–1 Vertices, *J. Res. Nat. Bur. Standards* **69B** (1965), pp. 125–130.
[20] Edmonds J., Matroids and the Greedy Algorithm, *Math. Programming* **1** (1971), pp. 127–136.
[21] Edmonds J. and Karp R. M., Theoretical Improvements in Algorithmic Efficiency for Network Flow Problems, *J. Assoc. Comput. Mach.* **19** (1972), pp. 248–264.
[22] Even S. and Kariv O., An $O(n^{2.5})$ Algorithm for Maximum Matching in General Graphs. In: *Proc. Sixteenth Annual IEEE Symp. on Foundations of Computer Science*(1975), pp. 100–112.
[23] Even S. and Tarjan R. E., Network Flow and Testing Graph Connectivity, *SIAM J. Comput.* **4** (1975), pp. 507–518.
[24] Floyd R. W., Algorithm 97: Shortest Path, *Comm. ACM* **5** (1962), p. 345.
[25] Ford L. R. Jr. and Fulkerson D. R., Maximal Flow Through a Network, *Canad. J. Math.* **8** (1956), pp. 399–404.
[26] Ford L. R. Jr., *Network Flow Theory*, Paper P-923, The Rand Corporation, Santa Monica, CA, 1956.
[27] Ford L. R. Jr. and Fulkerson D. R., *Flows in Networks*, Princeton University Press, Princeton, NJ, 1962.
[28] Fredman M. L., New Bounds on the Complexity of the Shortest Path Problem, *SIAM J. Comput.* **5** (1976), pp. 83–89.
[29] Gabow H. N., *Implementation of Algorithms for Maximum Matching on Nonbipartite Graphs*, Ph. D. Thesis, Department of Electrical Engineering, Stanford University, Stanford, CA, 1973.
[30] Gabow H. N., An Efficient Implementation of Edmonds' Algorithm for Maximum Matching on Graphs, *J. Assoc. Comput. Mach.* **23** (1976), pp. 221–234.
[31] Gabow H. N., Scaling Algorithms for Network Problems. In: *Proc. 24th Annual IEEE Symp. on Found. of Comp. Sci.* (1983), pp. 248–258.

[32] Gabow H. N. and Tarjan R. E., A Linear-Time Algorithm for a Special Case of Disjoint Set Union, *J. Comput. Sys. Sci.*, to appear.
[33] Gale D., Optimal Assignments in an Ordered Set: an Application of Matroid Theory, *J. Combinatorial Theory* **4** (1968), pp. 176–180.
[34] Galil Z., An $O(V^{5/3}E^{2/3})$ Algorithm for the Maximal Flow Problem, *Acta Informat.* **14** (1980), pp. 221–242.
[35] Galil Z., Micali S., and Gabow H., Maximal Weighted Matching on General Graphs. In: *Proc. Twenty-Third Annual Symp. on Foundations of Computer Science* (1982), pp. 255–261.
[36] Galil Z. and Naamad A., An $O(EV \log^2 V)$ Algorithm for the Maximal Flow Problem, *J. Comput. Sys. Sci.* **21** (1980), pp. 203–217.
[37] Garey M. R. and Johnson D. S., *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W. H. Freeman, San Francisco, CA, 1979.
[38] Gilsinn J. and Witzgall C., A Performance Comparison of Labeling Algorithms for Calculating Shortest Path Trees, National Bureau of Standards Technical Note 772, U.S. Department of Commerce, 1973.
[39] Graham R. L. and Hell P., On the History of the Minimum Spanning Tree Problem, *Ann. Hist. Comput.*, to appear.
[40] Hoare C. A. R., Quicksort, *Comput. J.* **5** (1962), pp. 10–15.
[41] Hopcroft J. E. and Karp R. M., An $n^{5/2}$ Algorithm for Maximum Matching in Bipartite Graphs, *SIAM J. Comput.* **2** (1973), pp. 225–231.
[42] Jarník V., O Jistém Problému Minimálním, *Práca Moravské Přírodovědecké Společnosti* **6** (1930), pp. 57–63.
[43] Johnson D. B., Priority Queues with Update and Finding Minimum Spanning Trees, *Info. Proc. Letters* **4** (1975), pp. 53–57.
[44] Johnson D. B., Efficient Algorithm for Shortest Paths in Sparse Networks, *J. Assoc. Comput. Mach.* **24** (1977), pp. 1–13.
[45] Kariv O., *An $O(n^{2.5})$ Algorithm for Maximum Matching in General Graphs*, Ph. D. Thesis, Department of Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel, 1976.
[46] Karzanov A. V., Determining the Maximal Flow in a Network by the Method of Preflows, *Soviet Math. Dokl.* **15** (1974), pp. 434–437.
[47] Kershenbaum A., A Note on Finding Shortest Path Trees, *Networks* **11** (1981), pp. 399–400.
[48] Khachian L. G., A Polynomial Algorithm for Linear Programming, *Soviet Math. Dokl.* **20** (1979), pp. 191–194.
[49] Klee V., Combinatorial Optimization: What is the State of the Art, *Math. Op. Res.* **5** (1980), pp. 1–26.
[50] Klee V. and Minty G. J., How Good is the Simplex Algorithm? In: O. Shisha (ed.), *Inequalities III*, Academic Press, New York, NY, 1972, pp. 159–175.
[51] Knuth D. E., *The Art of Computer Programming*, Vol. 3. *Sorting and Searching*, Addison-Wesley, Reading, MA, 1973.
[52] Korte B. and Lovász L., Mathematical Structures Underlying Greedy Algorithms. In: F. Gécseg (ed.), *Fundamentals of Computation Theory, Lecture Notes in Computer Science* **117**, Springer-Verlag, New York, NY, 1981, pp. 205–209.
[53] Kruskal J. B., On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem, *Proc. Amer. Math. Soc.* **7** (1956), pp. 48–50.

[54] Kuhn H. W., The Hungarian Method for the Assignment Problem, *Naval Res. Logistics Quarterly* **2** (1955), pp. 83–98.

[55] Lawler E. L., *Combinatorial Optimization: Networks and Matroids*, Holt, Rinehart and Winston, New York, NY, 1976.

[56] Łukaszewicz J., Florek K., Perkal J., Steinhaus H., and Zubrzycki S., Sur la liaison et la division des points d'un ensemble fini, *Colloq. Math.* **2** (1951), pp. 282–285.

[57] Malhotra V. M., Kumar M. P., and Maheshwari S. N., An $O(|V|^3)$ Algorithm for Finding Maximum Flows in Networks, *Info. Proc. Letters* **7** (1978), pp. 277–278.

[58] Micali S. and Vazirani V. V., An $O(\sqrt{|V|} \cdot |E|)$ Algorithm for Finding Maximum Matching in General Graphs, *Proc. Twenty-First Annual IEEE Symp. on Foundations of Computer Science* (1980), pp. 17–27.

[59] Moore E. F., The Shortest Path Through a Maze. In: *Proc. Int. Symp. on the Theory of Switching*, Part II, April 2–5, 1957, *The Annals of the Computation Laboratory of Harvard University* **30**, Harvard University Press, Cambridge, MA, 1959.

[60] Nicholson T. A. J., Finding the Shortest Route Between two Points in a Network, *Comput. J.* **9** (1966), pp. 275–280.

[61] Nilsson N. J., *Problem-Solving Methods in Artificial Intelligence*, McGraw-Hill, New York, NY, 1971.

[62] Papadimitriou C. H. and Steiglitz K., *Combinatorial Optimization: Algorithms and Complexity*, Prentice-Hall, Englewood Cliffs, NJ, 1982.

[63] Pape U., Implementation and Efficiency of Moore-Algorithms for the Shortest Route Problem, *Math. Programming* **7** (1974), pp. 212–222.

[64] Picard J.-C., and Queyranne M., *Selected Applications of Maximum Flows and Minimum Cuts in Networks*, Rapport Technique No. EP–79–R–35, Département de Génie Industriel, École Polytechnique de Montréal, 1979.

[65] Prim R. C., Shortest Connection Networks and Some Generalizations, *Bell System Techn. J.* **36** (1957), pp. 1389–1401.

[66] Pohl I., Bi-Directional Search. In: B. Meltzer and D. Michie (eds.), *Machine Intelligence* **6**, Edinburgh University Press, Edinburgh, Scotland, 1971, pp. 124–140.

[67] Sleator D. D., *An $O(nm \log n)$ Algorithm for Maximum Network Flow*, Technical Report STAN–CS–80–831, Computer Science Department, Stanford University, Stanford, CA, 1980.

[68] Sleator D. D. and Tarjan R. E., A Data Structure for Dynamic Trees, *J. Computer and System Sciences* **26** (1983), pp. 362–391; see also *Proc. Thirteenth Annual ACM Symp. on Theory of Computing* (1981), pp. 114–122.

[69] Tarjan R. E., Efficiency of a Good but not Linear Set Union Algorithm, *J. Assoc. Comput. Mach.* **22** (1975), pp. 215–225.

[70] Tarjan R. E., Complexity of Combinatorial Algorithms, *SIAM Rev.* **20** (1978), pp. 457–491.

[71] Tarjan R. E., Recent Developments in the Complexity of Combinatorial Algorithms. In: *Proc. Fifth IBM Symp. on Math. Found. of Comp. Sci.*, IBM Japan, Tokyo, 1980, pp. 1–28.

[72] Tarjan R. E., *Data Structures and Network Optimization*, Soc. Ind. and Appl. Math., Philadelphia, PA, 1983.

[73] Tarjan R. E. and van Leeuwen J., Worst-Case Analysis of Set Union Algorithms, *J. Assoc. Comput. Mach.*, to appear.
[74] Welsh D. J., Kruskal's Theorem for Matroids, *Proc. Cambridge Philos. Soc.* **64** (1969), pp. 3–4.
[75] Williams J. W. J., Algorithm 232: Heapsort, *Comm. ACM* **7** (1964), pp. 347–348.
[76] Yao A., An $O(|E|\log\log|V|)$ Algorithm for Finding Minimum Spanning Trees, *Info. Proc. Letters* **4** (1975), pp. 21–23.

BELL LABORATORIES
MURRAY HILL, NJ 07974
U.S.A.

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

L. G. VALIANT

An Algebraic Approach to Computational Complexity

The theory of computational complexity is concerned with identifying methods for computing solutions to problems in a minimal number of steps. Despite the diversity of application areas from which such problems can be drawn this theory has been successful in identifying a small number of fundamental questions on which a sizable fraction of the field hangs. A prime example is the P = ?NP question of Cook [4], concerning discrete search problems. Unfortunately the techniques currently known for proving that particular problems inherently require a certain number of steps are rudimentary. Hence these fundamental questions appear far from resolution.

In contrast with our ignorance about the absolute difficulty of computing problems, much is known about their relative difficulty. For example there are numerous results that relate pairs A and B of problems in the following way: If there exists a polynomial time algorithm for B (i.e. one that for some constant a solves B for inputs of size n in n^a steps) then there exists a polynomial time algorithm for A also. Such results do not depend on or determine whether the absolute complexities are n^2 and 2^n or any other function of n.

Our purpose here is to give a brief discussion of a very strict notion of reducibility called p-projection. Further details can be found in [19] where it was introduced and in [8, 15, 20]. The remaining references at the end of this paper describe work relevant to it in various ways.

The remarkable property of this notion of reduction is that in spite of its demanding and restricted nature numerous natural problems that superficially look dissimilar can be related by it. It is applicable to a variety of algebraic structures among which rings of multivariate polynomials and Boolean algebra are important examples. It can be used to give an account

of the relative difficulty of computational problems almost without having to define models of computation.

A computation *problem*, such as that of evaluating the determinant of a square matrix, is represented by an infinite *family* of instances of it, each corresponding to a different number of arguments and indexed by this number. The family DET will be the set $\{\mathrm{DET}_1, \mathrm{DET}_4, \mathrm{DET}_9, \ldots\}$ where DET_m is the m variable polynomial that is the determinant of a $\sqrt{m} \times \sqrt{m}$ matrix. Such a family is always defined with respect to a particular ring or field from which the constant coefficients are drawn. A second example is $\mathrm{PERM} = \{\mathrm{PERM}_1, \mathrm{PERM}_4, \mathrm{PERM}_9, \ldots\}$ where PERM_m is the permanent of a $\sqrt{m} \times \sqrt{m}$ matrix. Recall that the permanent has the same set of monomials as the determinant but the coefficient of each monomial is now positive.

It is interesting to contrast these two particular problems because one is easy to compute while the other is apparently hopelessly difficult. Gaussian elimination methods can be used to compute an $n \times n$ determinant in $O(n^3)$ arithmetic operations while more recent techniques do even better [5, 16]. On the other hand the best algorithm known for computing the permanent takes $O(n2^n)$ steps [9, 13]. Even the multiplicative factor of n in this bound appears difficult to remove.

The relationships which we explore among such pairs of families are of the following kind. If $A_i(y_1, \ldots, y_i)$ and $B_j(x_1, \ldots, x_j)$ are polynomials over the ring R we say that A_i is a *projection* of B_j if there is a substitution σ: $\{x_1, \ldots, x_j\} \to \{y_1, \ldots, y_i\} \cup R$ such that $A_i(y_1, \ldots, y_i)$ is identical to $B_j(\sigma(x_1), \ldots, \sigma(x_j))$. For example, $A_2(y_1, y_2) = y_1 + y_2$ is the projection of $\mathrm{DET}_4(x_{11}, x_{12}, x_{21}, x_{22})$ under the substitution $\sigma(x_{11}) = y_1$, $\sigma(x_{12}) = y_2$, $\sigma(x_{21}) = -1$, and $\sigma(x_{22}) = 1$. Further a family A is a *projection* of a family B if for all $A_i \in A$, A_i is a projection of some $B_j \in B$.

Now it so happens that the two families PERM and DET are projections of each other. This in itself, however, is of little practical interest since the definition permits that PERM_9, for example, be the projection of DET_j only for enormous values of j. Hence we need to add the following quantification. A family A is a *p-projection* of family B if for some constant α, for all $A_i \in A$, A_i is a projection of some $B_j \in B$ with $j \leqslant i^\alpha$.

In the investigations described here the following kind of question is central: Is PERM a p-projection of DET? One aspect of the computational relevance of this question is immediate. If we could give a positive answer to this question then we would have a polynomial time algorithm for the permanent. The algorithm would consist simply of the determinant algorithm applied after the appropriate initial substitution of variables.

Complexity-theoretic results state that the particular question raised above has broader significance in at least two directions.

First, it can be shown that the permanent exemplifies a large class, called pD (*p-definable*), of families of polynomials in the sense that all members of the class are p-projections of it [19, 20]. The class is essentially that of multivariate polynomials in which the degree grows only polynomially in the number of arguments and in which the coefficient of each monomial is easily computed from the specification of the variables in the monomial (cf. the permanent and determinant). The degree constraint turns out to be quite natural and we shall assume it in the discussion to follow. The class pD contains all such families that can be computed in a polynomial number of steps and, in addition, includes a large number of other families for which no such fast algorithms are known. Examples of the latter are most reliability problems such as the following. Consider a network with nodes $\{1, \ldots, n\}$ where the connection between node i and j has probability p_{ij} of not failing. Then the probability REL that nodes 1 and n are connected to each other is a polynomial in the variables $\{p_{ij}\}$. Further examples of such p-definable families abound as generating functions for combinatorial problems. For example HC is one such function defined for the Hamiltonian circuit problem for graphs in a natural way.

A family in pD of which every member of pD is a p-projection is called *complete* for pD. That there should be natural complete problems is not self-evident. However, it turns out that PERM, HC and REL, and many others are all complete for pD with respect to appropriate rings. These families are therefore p-projection of each other also. The proofs of these facts support the stronger statement that these p-projections are *strict* in the sense that two or more variables are never mapped to one. Hence these polynomials can be obtained one from the other by simply fixing some variables as constants and renaming the others.

Hence the importance of the permanent is due essentially to the fact that a wide variety of other polynomials can be expressed succinctly in terms of it. Our interest in the permanent versus determinant question stems from the second fact that the determinant also has a large class of polynomials that it can efficiently encode and this class is, to a first approximation, the class of all polynomial families that can be computed fast. We can conclude therefore that the permanent is a p-projection of the determinant if and only if the permanent and all the other families complete in pD can be computed fast. Hence a major computational problem has been reduced to a purely algebraic one.

Unfortunately, there is a huge gap in our current understanding of

the above question. It is known that an $n \times n$ determinant cannot be mapped to an $n \times n$ permanent if n exceeds two, even if substitutions by arbitrary linear forms are allowed [12, 18]. The possibility that an $(n+1) \times (n+1)$ determinant suffices, however, remains open. On the other hand there is substantial historical evidence that fully exponential growth is necessary since the contrary would imply fast algorithms for NP-complete problems and more.

The previously quoted result about the universality of the determinant for describing easy to compute problems needs one qualification concerning the model of computation assumed. It states that any polynomial is the projection of a determinant of size no larger than the minimal formula or expression for the polynomial. Whether the same result holds if we replace formula by the more basic model of computation, the straight-line program, remains an important open question. The relationship between the two measures of complexity is bounded by a growth factor of $n^{\log n}$ (called quasipolynomial) which is much less than the truly exponential factors (i.e., $\exp(n^{\varepsilon})$ for some $\varepsilon > 0$) which constitute the gaps in our current knowledge about all the relevant questions.

The class of functions that can be obtained as a projection of a given function A is a precise description of the class of functions that can be computed using a chip or program package for A directly, without the need for further programming. Hence the result for the determinant gives mathematical meaning for why the determinant and linear algebra itself is such a ubiquitous computational tool.

Boolean algebra is an equally fertile ground for carrying out an investigation akin to that described above for the multivariate polynomials. Here we define a projection to be substitution of variables by variables, negations of variables or constants. Reductions by such p-projections can be shown to be sufficient in many cases where only looser notions of reductions were known previously. Also, they can be shown to hold between easy to compute functions where such looser notions are meaningless.

Among specific results it can be shown that any polynomial time computable family is the p-projection of a family of linear programming problems [20]. This provides some explanation of the ubiquity of linear programming in combinatorial optimization. When we consider parallel rather than sequential computations, the transitive closure problem is universal, and this is again supported by much empirical evidence. For hard to compute combinatorial search problems one can get essentially the well-known NP-complete class [15, 19]. The algebraic approach provides an arguably simpler formulation than the now classical theory

using Turing machines. Such questions as P = ?NP are shown to be essentially equivalent to questions of whether one fixed family of Boolean functions is a p-projection of another.

A major motivation of studying this very strict notion of reducibility is the expectation of being able to prove negative results. One such theorem states that the symmetric Boolean functions are not very expressive in that there exist functions with polynomial bounded formulae that are not the p-projection of any family of symmetric functions [15]. A slightly more powerful family is the one for detecting whether an undirected graph is connected. This has the same shortcoming if the p-projections are restricted to be monotone (i.e., substitutions by negated variables are disallowed [15]), but becomes p-universal under general p-projections [3, 14].

Early work in computer science, such as that of Turing, concentrated on the notion of uniformity in computation, the notion that a fixed finite program is a description of potential behaviour on an infinite number of different inputs. Empirical evidence suggests that this notion may not be all-important in distinguishing polynomial time from exponential time computability. In trying to write a fast program for solving the Travelling Salesman Problem (TSP) it does not appear to make our task any easier to restrict ourselves to solving instances with exactly five hundred cities. For this reason in our algebraic theory of Boolean complexity we have excluded this notion of uniformity altogether, and thereby gained much simplicity. The notion can be added back (e.g., logarithmic space computable p-projection) with no difficulty. At present we do not believe, however, that the notion of uniformity will be central in ultimately resolving the important open questions.

We can summarize our approach as one in which the algebraic relationships among the natural computational problems are central and relations with computational models are almost secondary. We can caricature the advantage of this by considering again the Travelling Salesman Problem. On conventional models of computation this problem is always clumsy to discuss because it involves both real numbers and discrete choices. It becomes very easy to discuss, however, in the context of an appropriate algebra. Consider the set of rationals with the two binary operations of minimization and addition (to correspond to conventional addition and multiplication respectively). Many combinatorial optimization problems can be expressed naturally as polynomials in this algebra. TSP is $\mathrm{Min}_i\{w_i\}$ where minimization is over all Hamiltonian circuits in the associated graph and w_i is the sum of the weights on the ith such cycle. It turns out

that TSP is complete for p-definable polynomials in this algebra. Hence we have some explanation of the difficulty of TSP among combinatorial optimization problems without having to mention any specific model of computation.

References

[1] Baur W. and Strassen V., The computational complexity of partial derivatives, *Theoret. Comput. Sci.* **22**: 3 (1983), pp. 317–332.

[2] Borodin A., von zur Gathen J., and Hopcroft J. E., Fast parallel matrix and gcd computations. In: *Proc. 23rd IEEE Symp. on Foundations of Computer Science* (1982), pp. 65–71.

[3] Chandra A. K., Stockmeyer L. J., and Vishkin U., A complexity theory for unbounded fan-in parallelism. In: *Proc. 23rd IEEE Symp. on Foundations of Computer Science* (1982), pp. 1–13.

[4] Cook S. A., The complexity of theorem proving procedures. In: *Proc. 3rd ACM Symp. on Theory of Computing* (1971), pp. 151–158.

[5] Coppersmith D. and Winograd S., On the asymptotic complexity of matrix multiplication. In: *Proc. 22nd IEEE Symp. on Foundations of Computer Science* (1981), pp. 82–90.

[6] von zur Gathen J., Parallel algorithms for algebraic problems. In: *Proc. 15th ACM Symp. on Theory of Computing* (1983), pp. 17–23.

[7] Hyafil L., On the parallel evaluation of multivariate polynomials, *SIAM J. Computing* **8**: 2 (1976), pp. 120–123.

[8] Jerrum M. R., *On the complexity of evaluating multivariate polynomials*, Ph. D. Thesis, Edinburgh University, 1981.

[9] Jerrum M. R. and Snir M., Some exact complexity results for straight-line computations over semirings, *JACM* **29**: 3 (1982), pp. 874–897.

[10] Kalorkoti K. A., A lower bound on the formula size of rational functions. In: *Lecture Notes in Computer Science*, Springer-Verlag, Vol. 140 (1982), pp. 330–338.

[11] Karp R. M., Reducibility among combinatorial problems. In: *Complexity of Computer Computations* (R. E. Miller and J. W. Thatcher, eds.), Plenum Press, New York (1972).

[12] Marcus M. and Minc H., On the relation between the determinant and the permanent, *Illinois J. Math.* **5** (1961), pp. 376–381.

[13] Ryser H. J., *Combinatorial Mathematics*, Carus Math. Monograph no. 14 (1963).

[14] Skyum S., A measure in which Boolean negation is exponentially powerful, *Inform. Process. Lett.* **17** (1983), pp. 125–128.

[15] Skyum S. and Valiant L. G., A complexity theory based on Boolean algebra. In: *Proc. 22nd IEEE Symp. on Foundations of Computer Science* (1981), pp. 244–253.

[16] Strassen V., Gaussian elimination is not optimal, *Numer. Math.* **13** (1969), pp. 354–356.

[17] Strassen V., Vermeidung von Divisionen, *J. Reine Angew. Math.* **264** (1973), pp. 182–202.

[18] Sturtivant C., Generalised symmetries of polynomials in algebraic complexity. In: *Proc. 23rd IEEE Symp. on Foundations of Computer Science* (1982), pp. 72–79.

[19] Valiant L. G., Completeness classes in algebra. In: *Proc. 11th ACM Symp. on Theory of Computing* (1979), pp. 249–261.
[20] Valiant L. G., Reducibility by algebraic projections. In: *Logic and Algorithmic*, Monograph. Enseign. Math. **30** (1982), pp. 365–380.
[21] Valiant L. G., Skyum S., Berkowitz S., and Rackoff C., Fast parallel computation of polynomials, *SIAM J. Computing* **12**: 4 (1983), pp. 641–644.

HARVARD UNIVERSITY
CAMBRIDGE, MA

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

NANCY KOPELL

Forced and Coupled Oscillators in Biological Applications

The title of this paper involves forced and coupled oscillators. There is a subtitle as well: one approach to doing applied mathematics in an area of sometimes overwhelming complexity. The examples are taken mainly from the physiology of electrically excitable tissue, including nerve, heart and smooth muscle. To make my point about modelling, I shall also discuss an oscillatory chemical reaction known as the Belousov–Zhabotinskii (BZ) reaction, which is not exactly biological, but whose mathematical properties have much in common with the above tissues. For all of these examples, I shall examine the question: how much of the extremely complex pheonomena one sees is understandable on very general mathematical grounds, independent of detailed facts about the physiology/chemistry? Though the motivation of the question is not initially mathematical, it leads very quickly into deep mathematical problems.

The phenomena I shall discuss are all concerned with coupled systems (finite or infinitely many), each of which is oscillatory or "almost" oscillatory in a sense I shall describe below. From a mathematical point of view, a "biological" oscillator is any biological system which undergoes regular periodic changes. In practice, however, oscillations occurring in biological/chemical settings usually have some extra properties. For example, they tend to be quite stable, in amplitude and form, to perturbations of the system. Thus, they are effectively modelled by systems of differential equations with stable limit cycles, in contrast to those oscillations in mechanical systems which are described by Hamiltonian equations or perturbations of them. (For an early example, see [33].)

The "almost" oscillatory systems — cardiac tissue, nervous tissue, smooth muscle such as intestine — are "excitable", a term hard to define precisely, but easy to apply in practice. An excitable system (mathematical, biological or chemical) is one with a stable rest point, and having

the additional property that some trajectory starting "near" the rest point moves "far" away before returning to equilibrium. (See Fig. 1a for the phase plane diagram of a mathematical 2-dimensional excitable system.) The classic Hodgkin–Huxley equations used to describe the behavior of electrical impulses in nerves are excitable, as are the FitzHugh–Nagumo equations, a simplified version of them ([36]).

Small perturbations can change an excitable system into an oscillatory one. (See Figure 1b.) For example, nerve preparation, stimulated by

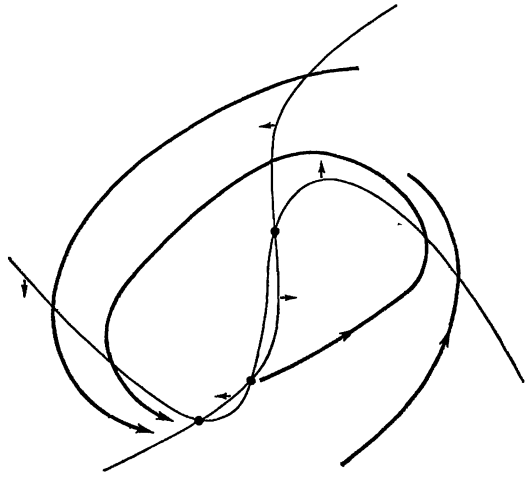


Fig. 1a. A two-dimensional excitable system having three critical points. The light lines are isoclines; the heavier ones are trajectories.

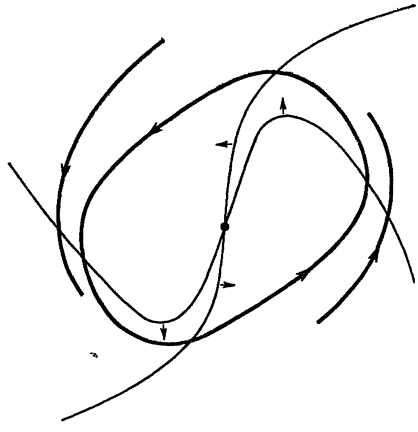


Fig. 1b. A nearby oscillatory system.

current, can oscillate (cf. [34] and its references); chick heart cells ([40]) can be made to change between excitable and oscillatory by chemical perturbations. The BZ reaction has an alternate, but close, recipe in which it is excitable rather than oscillatory ([43]). Some tissues, like cardiac tissue and intestinal tissue ([9]) contain a mixture of two kinds of cells, with the oscillations forcing the excitable cells. In such a context, it is not always clear what are the "natural" properties of the underlying cells (e.g. compare the models of [17] and [23].)

Chemical patterns

I shall start of discussing some older work on the BZ reaction ([21], [25], [26], [27], [28]). (For a partial bibliography of other work on this reaction, see [42] and [45].) The Belousov reaction ([1]) was among the first oscillating chemical reactions to be discovered. With appropriate indicators, this fluid turns alternately bright blue and deep red, with a period of the order of two or three cycles/minute, for up to several hours. Zhabotinskii and Zaiken ([46]) noticed that the same fluid, when placed in a layer and covered to prevent convection, spontaneously produces intricate patterns of concentric circles. Winfree was the first to recognize the analogy with cardiac tissue ([43]); he modified the recipe to make the reaction excitable and saw that the resulting layer of fluid could sustain patterns similar to those seen in pathological sheets of cardiac tissue, notably rotating spiral waves of electrical activity. Similar patterns are seen in part of the life cycle of colonies of the social amoeba *Dictyostelium discoideum*, and there is reason to believe that a similar mathematical description of the pattern formation is valid ([6]).

Ignoring, for the moment, the analogy with cardiac tissue, I shall discuss the patterns in the BZ reaction. While extremely complex, this reaction is still much simpler than biological systems, at least from the modelling point of view. That is, one knows that the only physical processes taking part in the pattern formation are chemical reactions and diffusion. Thus, it is possible to write down the form of the equations which govern the creation and evolution of the patterns:

$$c_t = F(c) + K\nabla^2 c, \tag{1}$$

where $c = (c_1, \ldots, c_n)$ is a vector of chemical concentrations, F: $R^n \to R^n$ is smooth, K is an $n \times n$ positive definite matrix, and $\nabla^2 c$ is the spatial Laplacian. Here

$$c_t = F(c) \tag{2}$$

are the kinetic equations representing the chemical reaction, and the Laplacian models diffusion. If there are heterogeneous particles in the solution, F may depend on the spatial variable x as well as c.

It remains to say what are $F(c)$ and K. Much is known about the BZ reaction, and there is a good model, due to Field, Noyes and Körös [30], of the chemical mechanism. However, the interest in this reaction is not mainly for itself, but as a chemical example of spontaneous pattern formation. Thus, instead of asking whether (1) has solutions representing the observed patterns when (2) is the above model, we raise a broader question: what are the properties of F and K that make the pattern formation possible? This question recently took on more depth with the discovery of new oscillating chemical systems which form the same patterns ([11]).

The picture is not entirely filled in, but so far there have emerged some remarkably simple answers. In order to describe the mathematical work, more must be said about the chemical phenomena. In detail, the patterns that form are never the same, and, indeed, are constantly changing over time. However, there are some features which are present every time one performs the experiment. For example, there are always sets of outwardly moving concentric rings, known as "target patterns". (See [26] for a picture.) Within a given one, the spacing of rings is constant, as is the speed of outward propagation and the frequency of oscillation at any given point within the target pattern; all these parameters change from one target pattern to another. Where two such patterns abut, there is an abrupt and visible change in these parameters, analogous to a shock in gas dynamics. This shock moves in a way predictable from the frequencies and wavelengths of the patterns it separates. With a small initial movement of the fluid, rotating spirals may also be formed.

The first mathematical task is idealization: to pick out from the observations a picture amenable to mathematical description. For the BZ patterns, the simplest idealization is that of a periodic plane wave, obtained by focusing attention on a neighborhood of a radial line of a target pattern, ignoring the center and pretending the target pattern has infinite extent. (Such an idealization, and other more comprehensive ones to be discussed, make sense because of the observation that the development of a piece of the overall pattern appears to be virtually independent, on the time scale considered, of all but a small neighborhood of itself.)

Mathematically, a plane wave is a function $c(x, t)$ which is actually dependent only on the single variable $t - a \cdot x$. We may now ask: under what hypotheses on (2) and K does (1) have periodic plane wave solutions?

As a partial answer, one has a very general sufficient condition: if (2) has a stable limit cycle, then (1) automatically has a one-parameter family of periodic plane wave solutions, parameterized by $|a|$ ([26]). These solutions are stable for $|a|$ small if K is not too far from the identity ([21]). Thus one might expect, as is turning out ([11]), that oscillating chemical systems other than the BZ reaction can support pattern formation. (The condition on (2) is not necessary: the excitable version of the BZ reaction can support propagating plane waves, as can an excitable axon of a nerve. The Hodgkin–Huxley and FitzHugh–Nagumo equations, which have the form (1) with K semidefinite and (2) excitable, have been shown to have travelling wave solutions. For references, see [35].)

We go on to ask if all of the other features described above are consequences of merely the existence of a stable oscillation for (2). Take, for example, the "shock" separating a pair of target patterns. Here, an appropriate idealization focuses on the line joining the pair of centers, considered to be many, many wavelengths apart. The mathematical description is then a pair of plane waves, propagating inward from $\pm\infty$, modulating from one to the other over a finite, identifiable region that may itself propagate. It is much harder than for plane waves to find such a solution to (1); as a first step, one has recourse to specific model equations for (2).

One such caricature retains nothing of the particular properties of the BZ reaction but its ability to oscillate. The F and K of (1) are given by

$$F(c) = \begin{bmatrix} \lambda & \omega \\ -\omega & \lambda \end{bmatrix} c, \qquad c = (c_1, c_2), \; K = I, \tag{3}$$

where $\lambda = \lambda(|c|^2)$, $\omega = \omega(|c|^2)$, with $\lambda(1) = 0$, $\lambda'(1) < 0$; equation (2) with this F has a stable limit cycle at $|c| = 1$. This system has the great advantage that its symmetry makes it possible to guess at the form of solutions, which allows a reduction of some problems in partial differential equations to ones in ordinary differential equations. Indeed, the symmetry suggests the use of polar coordinates in concentration space: $c_1 = r\cos\theta$, $c_2 = r\sin\theta$. One may then look for solutions in which the amplitude r is a function of space, but not of time: $r = r(x)$. It then follows from the structure of (1), (3) that $\theta(x, t)$ has the form

$$\theta(x, t) = \sigma t - \int_a^x a(\bar{x})\, d\bar{x}. \tag{4}$$

This leads to a third order, nonlinear ordinary differential equation for $r(x)$, $r'(x)$ and $a(x)$, with parameter σ. The natural boundary conditions

for a shock solution are

$$r(x)\to r_{\pm} \quad \text{and} \quad a(x)\to a_{\pm} \quad \text{as} \quad x\to\pm\infty, \tag{5}$$

where $r = r_{\pm}$ and $a = a_{\pm}$ are the amplitudes and wave numbers of the plane waves at $\pm\infty$, with $a_- > a_+$.

Under one further hypothesis, it is shown in [21] that there are solutions to (1), (3) with $r = r(x)$, (4), (5), for an open set of parameters $r_{\pm}$. (Given $r_{\pm}$, and the functions λ and ω, $a_{\pm}$ are determined up to sign; with these determined, σ is fixed.) The hypothesis is that $\omega'(r) < 0$; this can be interpreted as a statement about the dispersion function, or relation between frequency and wave number, of the family of plane wave solutions discussed above, a statement which is well defined for general systems of the form (1) having a family of plane wave solutions (and checkable in a physical system without knowing F). Under this hypothesis one can show, at least formally, that shock solutions exist for open sets of pairs of plane waves, as before; it remains a challenge, related to the mysterious mathematical fiction of "slowly varying waves", to understand the general result in a rigorous way.

For more complicated idealizations, such as target patterns or spirals of infinite extent, the patterns are not yet understood in the sense of this program, that is, as consequences of general, checkable hypotheses. (Such solutions have been shown to exist for models of the form (1), (3) ([8], [16], [19], [27], [28]), and the work shows the intricacy of the structure of the solutions to (1), (3), which contain "horseshoes" of spatially bounded, time periodic solutions ([24], [27]). See also [14].) One exception to the lack of generality concerns target pattern solutions in the presence of inhomogeneities, which are known to facilitate the formation of target patterns. P. Hagan has shown, formally ([18]), that if (2) has a limit cycle, and there is an impurity which acts locally to increase the frequency, then (with some technical assumptions about scaling and the size of the frequency increase) the target patterns form and are stable. This was proved for $\lambda - \omega$ systems ([25]), but remains to be proved for general F such that (2) has a limit cycle.

The largest open questions concern the formation of these patterns in excitable systems, for instance cardiac tissue. One difficulty is the absence of an easy to study caricature; although $\lambda - \omega$ systems may be modified to be excitable, such modifications lose the advantage of symmetry. Numerical work ([44]), plus work on discrete versions (see [20] and references in it), suggest spiral solutions should exist in equations that model excitable media.

Parabolic bursting

I turn now to some biological phenomena, in the same spirit as the discussion of the BZ reaction. The aim, as before, is to understand dynamical behavior as a consequence of general, but checkable, mathematical hypotheses which might be satisfied in a variety of chemical/physical circumstances. The first phenomena is known as "bursting", and it occurs in a large class of cells (including neural, cardiac and smooth muscle) which are electrically excitable; bursts are series of action potentials (spikes), or rapid changes in trans-membrane potential, which alternate with periods of quiescence. Even on a mathematical level, there are many ways in which bursting patterns can be accomplished ([4], [36]). I shall concentrate on a subclass of bursting phenomena known as parabolic bursting, some well-studied cases of which are the ganglion cells of the mollusc aplysia and the smooth muscle of mammalian gastro-intestinal tract ([2]). In parabolic bursting, the interspike intervals first decrease and then increase again. Furthermore, the spikes appear to "ride the crest" of a slower, almost sinusoïdal oscillation, which remains when chemicals are used to suppress spike formation.

For nervous tissue, including aplysia, there is a (successful) tradition of looking for explanations of the dynamics of the trans-membrane potential in the framework of the Hodgkin–Huxley equations or variations of them ([4], [32]). However, functions performed by cells are tightly coupled, and although, for purposes of modelling, it is essential to focus on some processes and ignore the others, it is by no means a priori obvious what processes are essential to a given behavior. In the case of parabolic bursting, some (but not all!) important qualitative behavior, plus the response to a lot of pharmacological, temperature and electrical perturbations, seems easier to understand if one pays more attention to the other functions of the cell, and less to the details of the voltage regulated ion conductances described by Hodgkin–Huxley like equations. In particular, if one takes into account the strong evidence for a slow (period of the order of a minute) metabolic oscillation coupled to the membrane potential, one can understand the above behavior from general features of the mathematical properties of excitable cells, independent of detailed knowledge of the ion currents or even explicit identification of the slow oscillation. (The problem of modelling a biological system is complicated by the wealth of plausible candidates for mechanisms, as well as the possibility that a given set of behavior is due, independently, to more than one mechanism. Some physiological arguments for and against the model to be discussed below are given in [13].)

Mathematical models of excitable tissue, with trans-membrane potential V as one variable, usually have the property that $\dot{V} = 0$ along a curve or surface that is "cubic-like" in V ([4], [36]). From this, and physiologically realistic assumptions on the other variables (cf., e.g. [4], [13], [15]), it is easy to construct models of excitable tissue with the following property: a perturbation of some parameter changes the system from oscillatory to excitable, preserving the existence of an attracting, invariant circle. (The latter is a limit cycle for the oscillatory system, and a circle with two critical points for the excitable one. See Figures 1a, b.) Thus, we are led to study equations of the forms

$$x_t = f(x) + \varepsilon^2 g(x, y), \quad x \in R^n, \tag{6}$$

$$y_t = \varepsilon h(x, y), \quad y \in R^m. \tag{7}$$

Here, the components of x are those variables involved in spike formation, those of y the ones involved in the slower (possibly metabolic) oscillation. For $\varepsilon = 0$, equation (6) is assumed to be exactly on the border between oscillatory and excitable; i.e., the attracting invariant circle has a single critical point, $x = 0$, which is a saddle-sink. ($g(x, y)$ may have terms independent of y, so, for ε small, the equation for x (without coupling) may be excitable or oscillatory.) The other hypothesis is on (7): for $x = 0$, (7) is assumed to have a stable limit cycle solution. ε turns out to measure the ratio of the spike time to an average interspike interval, and there may be many spikes per slow period.

For different choices of $g(x, y)$ or $h(x, y)$, equations (6), (7) can be shown to display slow oscillations, bursting, and another form of oscillatory behavior known as "beating"; the latter refers to periodic spiking with no apparant superimposed slower time scale. (This behavior is found in some aplysia cells, and chemical perturbations can change "bursters" to "beaters" and visa versa ([5]).) More specifically, there is a change of variables that exhibits (6), (7) as a singularly perturbed equation whose (nonuniform) limit as $\varepsilon \to 0$ contains an invariant torus. The differential equation on the torus has the form

$$\theta_\tau = (1 - \cos\theta) + (1 + \cos\theta)[\alpha + \beta H(\tau)], \tag{8}$$

where α, β and $H(\tau)$ depend on g and h, and $\tau = c\varepsilon t$ for an appropriate constant c. ($H(\tau)$ is normalized to have zero average value and a maximum of 1.) Equation (8) can be transformed into a Hill's equation by letting $\tan(\theta/2) = -v_\tau/v$. One then gets

$$v_{\tau\tau} = -[\alpha + \beta H(\tau)]v. \tag{9}$$

Fixing $H(\tau)$ and varying α and β, one passes through countably many regions in parameter space for which the solutions to (9) are unstable, separated by regions in which the solutions to (9) are quasiperiodic. Small amplitude slow oscillations, beating, and parabolic bursting turn out to correspond to different regions of α, β space. In particular, for (α, β) in the jth unstable region for (9), equation (8) is structurally stable, and has a globally stable solution with j spikes per burst. If (α, β) belongs to a quasiperiodic region for (9), solutions to (8) have a mixture of j or $j+1$ spikes per burst; the pattern can be computed in terms of a rotation number ([13]). Low amplitude oscillations turn out to correspond to $\alpha < -|\beta|$, and beating to $\beta \ll 1$, α moderate. Because of the nonuniformity of (the transformed version of) equations (6), (7) to (8), it requires work to show that these solutions still exist for (6), (7); this has been proved for bursting solutions only if the number j of spikes per burst is not too large for a fixed ε (j grows as $\varepsilon \to 0$) ([13]). To interpret responses of cells to electrical and pharmacological perturbations, one has to impose further, but weak, restrictions on $f(x)$ (i.e., qualitative similarity to the Hodgkin–Huxley equations).

The analysis of (6), (7) in terms of Hill's equation uncovers some further surprising relationships between the dynamics of biological oscillators and that of the BZ reaction. Unlike the previous discussion of the BZ reaction, we are now concerned only with temporal behavior, in a fluid mixed so that no spatial pattern can occur. As mentioned previously, in such a stirred "batch reactor", this fluid may be oscillatory or excitable, depending on the recipe. More complicated dynamics can be obtained by running the reaction in a "continuously stirred tank reactor", familiar to chemical engineers as a CSTR. In a CSTR, fresh chemicals are added at a constant rate, and reacting fluid is drawn off at the same rate. The dynamics of the reaction turns out to depend in an important way on this flow rate. In fact, as the rate is changed, the system appears to pass through a succession of parameter regions analogous to those of Hill's equation; regions with periodic output alternate with regions in which the output is not periodic ([22], [41]). In the jth periodic region, one sees a bursting pattern with one large oscillation and $j-1$ smaller ones in each burst. In a nonperiodic region, the output is a mixture of j-bursts and $(j+1)$-bursts. The latter regions are commonly referred to as chaotic; indeed, the experimentally measured power spectrum of the output contains broad band noise, as opposed to the clear peaks in the periodic regions. It is interesting to note that these features are mimicked by equation (8) (including the broad band noise for some limited regions of α, β space), in spite of the fact that a differential equation on a torus

cannot have strange attractors. Another connection of equation (9) to "chaos" is the bifurcation from a periodic to a nonperiodic region, which is reminiscent of the scenario of "intermittency" ([29]). For a three-dimensional model of bursting in the BZ reaction (in some parameter regions) see [37].

Before leaving the subject of bursting, I would like to mention the work of G. Carpenter ([4]) in the qualitative spirit of this paper. She showed, for a different (nonparabolic) kind of bursting that certain qualitative features of the Hodgkin–Huxley equations alone were sufficient to account for the bursting (and the fine structure of the interspike intervals) in electrical signals that propagate as traveling waves down the axon of a nerve cell.

Frequency plateaus

For this example, I shall go back to the problem of spatial patterns. Heart muscle tissue, with its (abnormal) spatial patterns, has already been mentioned. Another important excitable/oscillatory system with spatial differences in dynamics is smooth muscle in mammals, e.g., in the gastro-intestinal tract. In normal mammalian intestine there is an endogenous slow electrical oscillation in the trans-membrane potential of some of the muscle cells. These oscillations exist even when the muscle is not active; contraction of the muscle is associated with bursts of action potentials discussed above ([9]).

The frequencies of the slow oscillations are not uniform; indeed, along, say, the small intestine, there is a gradient in the frequency (which is higher in the oral than in the lower or aboral part); along some lengths of intestine, this frequency gradient is almost linear. The above information is found by experiments in which small segments of the intestine are excised and their intrinsic frequency measured. In the live animal, one does not see this linear gradient. Instead, there are frequency "plateaus", or stretches of intestine with constant frequency, separated by jumps in frequency at places that do not appear to have physiological significance. The frequency on a given plateau is at least as high as the highest of the natural frequencies along that segment.

The slow wave electrical activity has been modeled by several investigators ([3], [10], [31], [38], [39]) as a chain of loosely coupled oscillators, mostly Van der Pols with almost-sinusoidal limit cycles. The exact form of the oscillators, the gradient in frequencies, the form and strength of the coupling and the amount of inhomogeneity and/or anisotropy in the

coupling vary among those authors. For a variety of related equations, they produced simulations (digital or electronic) which yield frequency plateaus. The first mathematical task was to understand, analytically, the dynamical behavior of those models in a context as free as possible of the unknown details of the oscillators and the coupling.

The equations are very close in form to those described in the first part of this paper; they can be thought of as discretized versions of reaction–diffusion equations:

$$\frac{dX_k}{dt} = F(X_k) + \varepsilon R_k(X_k, \varepsilon) + \varepsilon D[X_{k+1} - (1+\alpha)X_k + \alpha X_{k-1}],$$
$$1 \leqslant k \leqslant n+1, \; X_0 \equiv 0 \equiv X_{n+2}. \tag{10}$$

Here $X_k \in R^m$, D is an $m \times m$ matrix, $\varepsilon \ll 1$, and $\alpha - 1$ is a measure of the anisotropy in the coupling. The reaction part of this equation

$$\frac{dX_k}{dt} = F(X_k) + \varepsilon R_k(X_k, \varepsilon) \tag{11}$$

is assumed to have a stable limit cycle with frequency $2\pi/\omega_k$, where ω_k changes monotonically (but weakly) with k.

For ε small (i.e. weak coupling and a weak gradient), it turns out that the behavior of (10) depends on very few of the details of F, $\{R_k\}$ and D. In particular, in the $m(n+1)$-dimensional phase space of (10) there is an $(n+1)$-dimensional stable invariant manifold on which the dynamics can be represented as a "phase equation". This equation has the form

$$\frac{d\theta_1}{dt} = O(1), \tag{12.a}$$

$$\frac{d\varphi_k}{dt} = \varepsilon[\Delta_k + H(\varphi_{k+1}) + \alpha H(-\varphi_k) - H(\varphi_k) - \alpha H(-\varphi_{k-1})] + O(\varepsilon^2), \tag{12.b}$$
$$k = 1, \ldots, n, \; H(\varphi_0) \equiv 0 \equiv H(\varphi_{n+1}),$$

where θ is the phase of the kth oscillator, $\varphi_k \equiv \theta_{k+1} - \theta_k$, and $\varepsilon\Delta_k \equiv \omega_{k+1} - \omega_k$. H is a smooth, 2π-periodic function. (If F has the form (3) then $H(\varphi) = A\sin\varphi + B[\cos\varphi - 1]$ for some A and B.)

To lowest order, equation (12b) decouples from (12a) and can be treated as an n-dimensional system

$$\frac{d\varphi_k}{d\tau} = \Delta_k + H(\varphi_{k-1}) + \alpha H(-\varphi_k) - H(\varphi_k) - \alpha H(-\varphi_{k-1}), \tag{13}$$

where $\tau = \varepsilon t$. Even (13) is quite complicated, with rich mathematical questions, and the existence of frequency plateaus is not obvious. For a simple choice of H (namely $H = \sin\varphi$ or any odd function of φ qualitatively like $\sin\varphi$), $\alpha = 1$, and $\Delta_k \equiv -\beta$ for some β (a linear frequency gradient), it can be seen how frequency plateaus arise as the gradient becomes steeper ([12]): for β sufficently small, there is a unique critical point of (13) which can be interpreted as a "phase-locked" state of (11), i.e., a state in which all oscillators run at the same frequency, with fixed phase relationships between them. (This is discussed in [7] in the context of lamprey swimming). At some critical value of β, all critical points of (13) disappear, and a large amplitude limit cycle appears (not by a Hopf bifurcation). This limit cycle turns out to correspond to a pair of frequency plateaus with a jump in frequency between them; the homotopy type of the limit cycle as a subset of the phase-space of (13) turns out to indicate the position of the jump.

The above pair of frequency plateaus differ from the experimentally observed ones by being too symmetric; unlike the real plateaus, whose frequency is at or above the intrinsic frequency of each oscillator, for the above simple problem, the plateau frequencies are forced to be symmetric with respect to the midpoint of the intrinsic frequency gradient. However, numerical computation and formal arguments ([12]) suggest that if H is not an odd function of its argument, then many properties of the solutions are quite different; in particular, the symmetry is broken, and the plateaus may lie uniformly above the natural frequencies. More work is needed to understand this.

There are many mathematical problems that remain in trying to understand the behavior of smooth muscle tissue. As previously mentioned, the above model treats the intestine as a chain of loosely coupled oscillators. The real tissue, however, has a complicated architecture, with several layers and with a mixture of excitable and oscillatory cells ([9]). Will a more accurate model of the tissue still behave like a chain of coupled oscillators? More importantly, it remains to understand the plateaus and the bursts in the context of the function of the tissue, which involves peristaltic movement and mixing of material. This will involve looking at the mechanical responses of the cells, which are coupled to the electrical ones.

For each of the phenomena discussed above, the scientific program is to look for very general (and, if possible, checkable) hypotheses which can account in (qualitative) detail for the observed dynamics. The role of mathematics in the applications is to help identify the appropriate level

of explanation, i.e., to sort out which phenomena occur for very general, perhaps qualitative reasons, and which require a more detailed understanding of the biology/chemistry. For example, the patterns in the BZ reaction appear to be a consequence of not much more than the existence of the oscillation in the homogeneous fluid; if so, further details about the oscillations will not help understand the patterns. The role of the phenomena in the mathematics is a classic one, familiar from physics and mechanics: to guide one through very complicated equations to fertile mathematical questions.

References

[1] Belousov B. P., *Sb. Ref. Radiat. Med.* (1958), 145.

[2] Berridge M. J. and Rapp P. E., A Comparative Survey of the Function, Mechanism and Control of Cellular Oscillators, *J. Exp. Biol.* **81** (1979), pp. 217–279.

[3] Brown B. H., Duthie H. L., Horn A. R., and Smallwood R. H., A Linked Oscillator Model of Electrical Activity of Human Small Intestine, *Amer. J. Physiology* **229** (1975), pp. 384–388.

[4] Carpenter G. A., Bursting Phenomena in Excitable Membranes, *SIAM J. Appl. Math.* **36** (1979), pp. 334–372.

[5] Chaplain R. A., Metabolic Regulation of the Rhythmic Activity in Pacemaker Neurons, II, *Brain Research* **106** (1976), pp. 307–319.

[6] Cohen M. S. and Hagan P. S., Diffusion Induced Morphogenesis in the Development of Dictyostelium, *J. Theor. Biol.* (1982).

[7] Cohen A. H., Holmes P. J., and Rand R. H., The Nature of the Coupling Between Segmental Oscillators of the Lamprey Spinal Generator for Locomotion: a Mathematical Model, *J. Math. Biol.* **13** (1982), pp. 345–369.

[8] Cohen D. S., Neu J. C., and Rosales R. R., Rotating Spiral Wave Solutions to Reaction-Diffusion Equations, *SIAM J. Appl. Math.* **35** (1978), pp. 536–547.

[9] Connor J. A., On Exploring the Basis for Slow Potential Oscillations in the Mammalian Stomach and Intestine, *J. Exp. Biol.* **81** (1979), pp. 153–173.

[10] Diamont N. E., Rose P. K., and Davison E. J., Computer Simulation of Intestinal Slow-Wave Frequency Gradient, *Amer. J. Physiology* **219** (1970), pp. 1684–1690.

[11] Epstein I. R., Kustin K., de Kepper P., and Orban M., Oscillating Chemical Reactions, *Sci. Amer.* **248** (1983), pp. 112–123.

[12] Ermentrout G. B. and Kopell N., Frequency Plateaus in a Chain of Weakly Coupled Oscillators, I, *SIAM J. Math. Anal.* **15** (1984), pp. 215–237.

[13] Ermentrout G. B. and Kopell N., *Parabolic Bursting in Slowly Forced Excitable Systems*, in preparation.

[14] Fife P., On the Existence and Nature of the Homogeneous-Center Target Patterns in the Belousov-Zhabotinskii Reagent. In: O. Axelsson, L. S. Frank, and A. Van der Sluis (eds.), *Analytical and Numerical Approaches to Asymptotic Problems in Analysis*, North-Holland, 1981.

[15] Goldstein S. S. and Rall W., Changes of Action Potential Shape and Velocity for Changing Core Conductor Geometry, *Biophs. J.* **14** (1974), pp. 731–757.

[16] Greenberg J., Spiral Waves for $\lambda - \omega$ Systems, II, *Adv. Appl. Math.* **2** (1981), pp. 450–455.

[17] Guevara M. and Glass L., Phase Locking, Period Doubling Bifurcations and Chaos in a Mathematical Model of a Periodically Driven Oscillator: a Theory for the Entrainment of Biological Oscillators and the Generation of Cardiac Dysrhythmias, *J. Math. Biol.* **14** (1982), pp. 1–23.

[18] Hagan P. S., Target Patterns in Reaction–Diffusion Systems, *Adv. Appl. Math.* **2** (1981), pp. 400–416.

[19] Hagan P. S., Spiral Waves in Reaction-Diffusion Equations, *SIAM J. Appl. Math.* **42** (1982), pp. 762–786.

[20] Hastings S. P., Persistent Spatial Patterns for Semi-Discrete Models of Excitable Media, *J. Math. Biol.* **11** (1981), pp. 105–117.

[21] Howard L. N. and Kopell N., Slowly Varying Waves and Shock Structures in Reaction-Diffusion Equations, *Stud. Appl. Math.* **56** (1977), pp. 95–145.

[22] Hudson J. L., Hart M., and Marinko D., An Experimental Study of Multiple Peak Periodic and Nonperiodic Oscillations in the Belousov–Zhabotinskii Reaction, *J. Chem. Phys.* **71** (1979), pp. 1601–1606.

[23] Keener J., Chaotic Cardiac Dynamics, *Lectures in Appl. Math* **19** (1981), pp. 299–325.

[24] Kopell N., Time Periodic but Spatially Irregular Solutions to a Model Reaction-Diffusion Equation, *Ann. N.Y. Acad. Sci.* **357** (1980), pp. 397–409.

[25] Kopell N., Target Pattern Solutions to Reaction-Diffusion Equations in the Presence of Impurities, *Adv. Appl. Math.* **2** (1981), pp. 389–399.

[26] Kopell N. and Howard L. N., Plane Wave Solutions to Reaction-Diffusion Equations, *Stud. Appl. Math.* **52** (1973), pp. 291–328.

[27] Kopell N. and Howard L. N., Target Patterns and Horseshoes from a Perturbed Central Force Problem: Some Temporally Periodic Solutions to Reaction-Diffusion Equations, *Stud. Appl. Math.* **64** (1981), pp. 1–56.

[28] Kopell N. and Howard L. N., Target Pattern and Spiral Solutions to Reaction-Diffusion Equations with More than One Space Dimension, *Adv. Appl. Math.* **2** (1981), pp. 417–449.

[29] Manneville P. and Pomeau Y., Different Ways to Turbulence in Dissipative Dynamical Systems, *Phys. A* **1** (1980), pp. 219–226.

[30] Noyes R. M., Field R. J., and Körös E., Oscillations in Chemical Systems, II. Thorough Analysis of Temporal Oscillation in the Bromate-Cerium-Malonic Acid System, *J. Amer. Chem. Soc.* **94** (1972), pp. 8649–8664.

[31] Patton R. J. and Linkens D. A., Hodgkin–Huxley Type Electronic Modelling of Gastrointestinal Electrical Activity, *Med. & Biol. Eng. & Computing* **16** (1978), pp. 195–202.

[32] Plant R. E. and Kim M., Mathematical Description of a Bursting Pacemaker Neuron by a Modification of the Hodgkin–Huxley Equations, *Biophysical J.* **16** (1976), pp. 227–244.

[33] Van der Pol B. and Van der Mark J., The Heartbeat Considered as a Relaxation Oscillation, and as Electrical Model of the Heart, *Phil. Mag.* **6** (1928), pp. 763–775.

[34] Rinzel J., Repetitive Activity and Hopf Bifurcation under Point Stimulation for a Simple FitzHugh–Nagumo Nerve Conduction Model, *J. Math. Biol.* **5** (1978), pp. 363–382.

[35] Rinzel J., Impulse Propagation in Excitable Systems. In: Stewart, Ray and Conley (eds.), *Dynamics and Modelling of Reactive Systems*, Academic Press, 1980.
[36] Rinzel J., Models in Neurobiology. In: R. Enns, B. Jones, R. Miura and S. Rangnekar (eds.), *Nonlinear Phenomena in Physics and Biology*, Plenum Publishing Corp. (1981).
[37] Rinzel J. and Troy W., Bursting Phenomena in a Simplified Oregonator Flow System Model, *J. Chem. Phys.* **76** (1982), pp. 1775–1789.
[38] Robertson-Dunn B. and Linkens D. A., A Mathematical Model of the Slow-Wave Electrical Activity of the Human Small Intestine, *Medical and Biological Engineering*, pp. 750–757 (Nov. 1974).
[39] Sarna S. K., Daniel E. E., and Kingman Y. J., Simulation of Slow-Wave Electrical Activity of Small Intestine, *Amer. J. Physiology* **221** (1971), pp. 166–175.
[40] Sperelakis N. and Lehmkuhl D., Ionic Interconversion of Pacemaker and Non-Pacemaker Cultured Chick Heart Cells, *J. Gen. Phys.* **49** (1966), pp. 867–894.
[41] Turner J. S., Roux J. C., McCormick W. D., and Swinney H. L., Alternating Periodic and Chaotic Regimes in a Chemical Reaction, *Phys. Lett. A* **85** (1981), pp. 9–12.
[42] Tyson J. J., The Belousov–Zhabotinskii Reaction, *Lecture Notes in Biomath.* **10**, Springer-Verlag, 1976.
[43] Winfree A. T., Spiral Waves of Chemical Activity, *Science* **175** (1972), pp. 634–636.
[44] Winfree A. T., Rotating Solutions to Reaction-Diffusion Equations, *SIAM–AMS Proc.* **8** (1974), pp. 13–31.
[45] Winfree A. T., *The Geometry of Biological Time*, Springer-Verlag, Berlin, 1980.
[46] Zaikin A. N. and Zhabotinskii A. M., Concentration Wave Propagation in Two-Dimensional Liquid-Phase Self-Oscillating System, *Nature* **225** (1970), pp. 535–537.

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

BENOIT B. MANDELBROT

On Fractal Geometry, and a Few of the Mathematical Questions It Has Raised

In Memoriam Waclaw Sierpiński (1882–1969), "explorateur de l'infini"
In Memoriam Szolem Mandelbrojt (1899–1983)

This presentation is meant to sketch a few scattered problems in diverse branches of pure mathematics. Some have been solved, more or less completely, but others remain open. Their importance and difficulty are quite varied, but they are alike in two ways. Firstly, they all arose in the course of very practical investigations into diverse natural sciences, some of them old and well-established, others newly revived and a few of them altogether new. Secondly, they involve in essential fashion the "monster shapes" that had until now been viewed as belonging to chapters of mathematics devoid of any contact with the real world. For these two reasons, these mathematical problems prove central to an issue of consequence. Does pure mathematics exist as an autonomous discipline, that can—and ideally should—develop in total isolation? Or is the existence of totally pure mathematics a myth?

After I finish presenting this sample of problems, I shall show a collection of slides that demonstrate what certain shapes of mathematics really look like. The need to draw these shapes arose in the course of my work because of their scientific importance: they help my ideas and theories become accepted, and they help me generate new ideas and theories. Yet many of these shapes seem to strike everyone as being of exceptional and totally unexpected plastic beauty. Some have the beauty of the mountains and the clouds that they are indeed meant to represent, and others seem wild and unexpected at first, but after a very brief inspection come to appear as totally familiar. As a result, these slides prove central to a different philosophical issue. What is beauty, and is there any relation between

the beauty of these mathematical pictures and the beauty that a mathematician sees in his trade after long and strenuous practice?

We shall leave this second question aside, but shall face—implicitly but very pointedly—the question of what the relation is between pure mathematics and the outside world. Most scholars answer that "it depends". Obviously, there are *some* branches of mathematics in which physics, numerical experimentation and geometric intuition are very beneficial, but elsewhere in mathematics physics is irrelevant, computation is powerless, and intuition is misleading.

The irony is that history has consistently proven the above distinction to be unreliable: as branches or branchlets of mathematics develop, they suddenly either lose or acquire deep but unforeseen connections with the sciences—old and new. Also, numerical experimentation—which Gauss had found invaluable but whose practice had not changed until yesterday—has seen its power multiply thanks to computers—and in particular, thanks to computer graphics. Finally, the geometric intuition built on the practice of Euclid and of calculus is proving *not* to be something immutable, but can be retrained.

In no case that I know is this irony nearly as intense as in fractal geometry, a new branch of learning that I conceived, developed and put to use in models and theories relative to diverse sciences, and which has now become widely practiced. My latest book on this topic [1] will be referred to in the sequel as FGN. Even more specifically, there is profound irony in the fact that the present lecture is being delivered in the city where Wacław Sierpiński was born, and where he labored to establish a marvelous school that viewed itself as devoted exclusively to *Fundamenta Mathematicae* and contributed mightily to the list of monsters. I do not know how Sierpiński felt about the philosophical problem we are discussing, but concentration upon foundations did contribute to the gulf between mathematics and physics. In the last few years, however—largely by my work, that of my colleagues and now that of many scholars—the situation has changed dramatically.

The present catalogue is far from exhausting the pure mathematics component of fractal geometry. It uses freely the term *fractal*, which I coined from the Latin word for "rough and broken up", namely *fractus*. Loosely, a "fractal set" is one whose detailed structure is a reduced-scale (and perhaps deformed) image of its overall shape. At the end, I shall say a few more systematic words about fractals and about fractal geometry as a systematic discipline. "Dust" will be used to denote a totally disconnected set.

Two fractal curves by Sierpiński, and the new roles they find in physics

Breaking all logical and historical sequence, let us honor Sierpiński by beginning with fractal shapes that he investigated deeply in the 1910's [2]. One of these shapes had been known for a long time as the "carpet", and the second has received from me the name of "gasket". The Sierpiński carpet was originally introduced to show that a plane curve can be "topologically universal", that is, can contain a homeomorphic transform of every other plane curve. The construction starts with a square, then divides it into 9 equal subsquares and erases the middle one, which I call "trema" (Greek for "hole"). Then one proceeds in the same fashion with each remaining subsquare, and so on ad infinitum. The Sierpiński "gasket" was originally introduced to show that a curve can have branching points everywhere. The construction starts with an equilateral triangle, then divides it into 4 equal subtriangles and erases the middle one as trema. Then one proceeds in the same fashion with each remaining subtriangle, and so on ad infinitum. During the nineteen-twenties, the distinction between the carpet and the gasket became essential to the theory of curves of P. Urysohn and K. Menger, these being the prime examples of curves having, respectively, an infinite and a finite "order of ramification".

Needless to say, these shapes and this notion were meant to be anything but "applied mathematics". As a matter of fact, some mathematicians took the "gasket" as prime evidence that geometric intuition is powerless, because it can only conceive of curves having scattered branch points, but not having branch points everywhere. I confess that contemplation of the Eiffel Tower had long made me harbor doubts about this contention. Gustave Eiffel had designed his Tower to have many multiple points, and wrote that he would have made it even lighter, with no loss of strength, had the availability and cost of finer materials allowed him to increase the number of double points even further. The intellectual step from the Eiffel Tower to the Sierpiński gasket is one that my intuition was easily trained to take.

But let us go back to more serious questions. Lately, the Sierpiński carpets and gaskets and the order of ramification had come to be seldom mentioned by mathematicians. Where should one go to find the latest facts about these notions? The surprising answer is that one should go to journals in physics, because the statistical physics of condensed matter has come to view these notions as "unavoidable". Let me give three examples.

(Further examples are found in Mandelbrot, *Proceedings of Stat. Phys. 15*, in *J. Stat. Phys.* **34** (1984), p. 895.)

Percolation clusters at criticality

A first need for Sierpiński's creations arose in the study of the important notion of percolation cluster at criticality. When seeking a model to combine all the geometric features of a percolation cluster's "backbone", I went straight to the Sierpiński gasket. Then a group of physicists and I took multi-dimensional variants of the gasket, and confirmed their usefulness as models of percolation cluster backbones (FGN, p. 133; Gefen, Aharony, Mandelbrot and Kirkpatrick, *Phys. Rev. Lett.* **47** (1981), p. 1771). Once ridden of the cobwebs of abstraction, the gasket proves a very practical and enlightening geometric tool to work with. Physicists make it the object of scores of articles, and invent scores of generalizations that were not needed in 1915.

The Ising model of magnets

Magnets are commonly modeled by a model due to Lenz but credited to Ising. A second reason for the physicists' interest in the work of Sierpiński resides in Onsager's finding, that in Euclidean space $\boldsymbol{R}^E$ it is necessary and sufficient that $E > 1$ for magnets to exist. *Long open implicit question*: to which of the innumerable mathematical differences between the $\boldsymbol{R}^E$, for $E = 1$ and $E > 1$, can the existence of magnets be traced? *Partial answer*: We studied numerous specific examples of the Sierpiński curves and related fractal lattices, and found that magnets can exist when and only when the order of ramification is infinite (FGN, p. 139; Gefen, Mandelbrot and Aharony, *Phys. Rev. Lett.* **45** (1980), p. 855). *Conjecture.* The above answer is of general validity. This raises a difficult *Unsolved problem*: to rephrase the criterion of existence of magnets, from the present indirect and highly computational form, to a direct form that would give a chance of proving or disproving this conjecture.

Actual geometric implementation of the fractional-dimensional spaces of physics

Physicists are very successful with a procedure that is mathematically very dubious. They deal with spaces whose properties obtain from those of Euclidean spaces by interpolation to "noninteger Euclidean dimensions". For example, the dimension may be $4-\varepsilon$ or $1+\varepsilon$, with an infinitesimal ε. Calculations can be carried out, in particular, expansions can be per-

formed in ε, and then the "infinitesimal" ε is then set to 1. Mathematically, these spaces remain unspecified, yet the procedure turns out to be extremely useful. *Mathematical problems.* To show that the properties postulated for those spaces are mutually compatible, to show that they do (or do not) have a unique implementation, to describe their implementation constructively. *Very partial solution.* A very special example of such space has been implemented indirectly (FGN, 2nd printing, p. 462, and Gefen, Meir, Mandelbrot and Aharony, *Phys. Rev. Lett.* **50** (1983), p. 145). We showed that the postulated properties of certain physical problems in this space are identical to the *limits* of the properties of corresponding problems in a Sierpiński carpet whose "lacunarity" (as I had defined it in the study of galaxies; see below) is made to converge to **0**.

Peano and Koch fractal curves and the measurement of the Earth

The Sierpiński curves were among the "great counterexamples" against previously held intuitive ideas about mathematics. Their numbers had been growing since Weierstrass demonstrated that a continuous function can be nondifferentiable, and since Cantor and Peano demonstrated that dimension is a notion that cannot be trusted to intuition. To quote J. Dieudonné, "Some mathematical objects, like the Peano curve, are totally non-intutive... extravagant." Until recently, this view was universally accepted as established beyond discussion. However, (FGN, Chap. 7), I made the Peano curve become viewed as eminently intuitive, by showing it to be the logical extrapolation of a natural simplified geometric model of the cumulative shores of all the rivers in the fluvial tree! Similarly for the Koch nondifferentiable and nonrectifiable "snowflake curve" (FGN, Chaps. 5–6): I used it as the logical extrapolation of a natural simplified geometric model of coastlines. (Hugo Steinhaus and two or three other scholars had made this remark, but failed to develop it.) Not only have the "great counterexamples of analysis" thereby become very useful in the sciences, but their most obvious and indisputable usefulness has been to bring geometry back to the source indicated by the etymology of "geometry" = the measurement of the Earth. The fractal geometry of the relief that I developed is founded on the old counterexamples!

Cantor-like fractal dusts. Interplay between their roles in physics and in the theory of sets of Fourier multiplicity

The first of several things I have to say about Cantor sets concerns an example of multiple mutual interplay between pure probability theory,

the theories of noise and turbulence, and the theory of sets of multiplicity of Fourier series. The latter are sets $\mathscr{S}$ in $[0, 2\pi]$, such that there exists a Fourier series that converges to 0 for $x \notin \mathscr{S}$, yet is not identically zero. The problem of whether such sets exists was raised by G. Cantor; in fact, he designed his ternary set thinking it may be a set of multiplicity. But it is *not* one. It gradually became clear (largely thanks to R. Salem) that a deterministic Cantor set is a set of Fourier multiplicity only if it has some very specific number-theoretical properties. Or it can be a suitable random set, but the examples known before 1965 were very contrived.

Before I knew anything of the above problem, I had injected the Cantor set in physics when I needed a first approximation to represent certain error patterns in telephone transmission. Then, as a second approximation, I turned to a randomized variant, the "Lévy fractal dust", also called "stable subordinator set" (FGN, Chaps. 8 and 32). This set appeared desirable because it combined self-similarity properties with a total lack of irrelevant structures. I proposed that the proper measurement of a channel's "noisiness" was not the average number of errors, as had seemed obvious, but was a totally unexpected quantity: the Hausdorff–Besicovitch dimension.

In this framework, acquaintance with sets of multiplicity led me to formulate the following *Problem*: Is the Lévy dust a set of Fourier multiplicity? *The answer* (J. P. Kahane and Mandelbrot, *Comptes Rendus* (*Paris*) **161** (1965), p. 3931; see FGN, p. 360) is to the affirmative. Other "natural" sets of multiplicity followed; it is worth pointing out that their seeming "natural" may perhaps be related to their link with the description of nature.

In a related development in a very different part of physics, "devil's staircase" functions, which vary only on Cantor-like fractal dusts, have become very important in the study of physical systems with incommensurable frequencies.

Conjecture and problem concerning the fractal (Hausdorff-dimensional) properties of the singularities of the Navier–Stokes and Euler equation of viscous and nonviscous fluid motion

The above-mentioned work of mine on noise records that can be represented by Cantor sets was done in 1962, near-simultaneously with A. N. Kolmogorov's work on the intermittency of turbulence. After numerous experimental tests, designed to create an intuitive feeling for this phenomenon (e.g., after listening to turbulent velocity records that were trans-

formed to be made audible), I was able to extend my newly developed methods to turbulence. This led me circa 1964 to the following *Conjecture.* The property of being "turbulently dissipative" should *not* be viewed as attached to domains in a fluid, but as attached to fractal sets whose intersection with a straight line is a Cantor-like fractal dust having a Hausdorff dimension in the range from 0.5 to 0.6. The corresponding full sets in space should therefore be expected to be fractals with a Hausdorff dimension in the range from 2.5 to 2.6.

Actually, Cantor dust and Hausdorff dimension are not the proper notions in the context of viscous fluids, because viscosity necessarily erases the fine detail that is essential to Cantor fractals. Hence the following *Conjecture* (FGN, Chap. 11 and Mandelbrot, *Comptes Rendus* **282A** (1976), p. 119). The dissipation in a viscous fluid occurs in the neighborhood of singularity of a nonviscous approximation following Euler's equations, and the motion of a nonviscous fluid acquires singularities that are sets of dimension about 2.5 to 2.6. *Open mathematical problem.* To prove this conjecture, under suitable conditions, or to disprove it.

Comment A. Several numerical tests of this conjecture have been carried out, and are in agreement with it (e.g. Chorin, *Comm. Pure and Appl. Math.* **34** (1981), p. 853). See also Henschel and Procaccia, *Phys. Rev. Lett.* **49** (1982), p. 1158. *Comment* B. I also conjectured that the Navier–Stokes equations have fractal irregularities, of much smaller dimension. This conjecture had led to extensive work by V. Scheffer, and then others, especially by R. Temam and C. Foias. *Comment* C. A few years after my work, Cantor-like dusts also entered in the study of the transition from laminar to turbulent flow, through the work of Ruelle and Takens; see these *Proceedings*, p. 237, the contribution by David Ruelle.

Postscript

Numerous facts about the above conjectures are reported in these *Proceedings*, p. 119, in the contribution of Peter D. Lax, which also contains an extensive bibliography.

The large scale structure of the universe: role of Cantor-like fractal dusts in describing the distribution of galaxies

Upon examining various models of the distribution of galaxies and clusters of galaxies, I observed their resemblance to spatial Cantor sets, and concluded that their main ingredient was postulated self-similarity. The reason

why these models had been dismissed as unrealistic (and most had been forgotten), seemed to lie in their excessive regularity and the unreasonable feature they had, of implying that the Universe has a "center". Both features proved capable of being corrected (FGN, Chaps. 9 and 33 to 35). First, I advanced a model, now called *The Seeded Universe*, constructed with the help of a three-dimensional generalization of the Lévy dust that had proved (see above) to be a set of Fourier multiplicity. This generalization's Hausdorff-dimensional properties were known. Its correlation properties (Mandelbrot, *Comptes Rendus (Paris)* **280A** (1975), p. 1551) are nearly identical to those of the computer-processed galaxy maps.

Actual simulations, however, revealed clear-cut discrepancies, and in particular the fact that the Seeded Universe was visually far more "lacunar" than the real world. I gave a precise meaning to this notion of lacunarity, and devised a second model, now called *The Parted Universe*, in which lacunarity can be adjusted at will, and can be fitted to the actual distribution. The idea is to "cut-out" from $\boldsymbol{R}^3$ a collection of suitably sized and distributed open sets called "tremas". An heuristic argument gave a certain value for the remaining set's Hausdorff dimension, and one could not fail to *Conjecture* that this was the correct value. *Confirmation.* This last conjecture was proven in $\boldsymbol{R}$ by Mandelbrot, *Z. Wahrsch.* **22** (1972), p. 145 and in $\boldsymbol{R}^E$ with $E > 1$ by Y. El. Hélou (*Comptes Rendus (Paris)* **287A** (1978), p. 815) and in generalized form by U. Zähle (*Trans. 9th (1982) Prague Conference on Information Theory*, p. 295, and *Math. Nachr.* **116** (1984), p. 325).

Remark A. Several physical arguments based on Newtonian attraction, when combined with self-similarity, predict the fractal dimension $D = 1$. *Open mathematical problem.* Prove (or disprove), using potential theory, that $D = 1$ should be found in *every* model based on Newtonian potential.

New aspects of Brownian motion and of related fractal curves

My assertion, that I was the first to put the old counterexample of analysis to use in physics, has of course a very notable exception: Norbert Wiener's Brownian motion process $B(P)$ is (a.s.) a nondifferentiable curve, and it has drawn immense interest from both mathematicians and physicists. My new applications raise further problems and also require a generalization—which I called the fractional Brownian $B_H(P)$—namely the random function from $\boldsymbol{R}^T$ into $\boldsymbol{R}^E$ whose increments $B_H(P) - B_H(P')$ are Gaussian variables of zero mean and variances equal to $|P - P'|^{2H}$. The ordinary Brownian case corresponds to $H = 1/2$.

A natural random universal curve. Returning to Sierpiński, let me mention my *Conjecture* (FGN, p. 243) that to obtain a topologically universal plane curve it suffices to draw a piece of ordinary Brownian motion, with $H = 1/2$, B in $\boldsymbol{R}^2$ and P in $\boldsymbol{R}$, and to extract a "perfect wiggle" $\mathscr{W}$, defined as a portion contained between successive returns to the same point. A *proof* has been provided by S. Kakutani and N. Tongling (unpublished). Brownian motion had been previously shown to be universal in other ways.

The self-avoiding Brownian motion. The complement of the perfect wiggle $\mathscr{W}$ contains (a.s.) one unbounded component and an infinity of bounded components. The boundary of the unbounded component was dubbed "self-avoiding Brownian motion" in FGN. *Conjecture*: The dimension of this last fractal curve is 4/3. *Problem.* To determine the distribution of the areas of the bounded components.

Islands. Now let B be in $\boldsymbol{R}$ and P in $\boldsymbol{R}^2$. An "island" is the set of values of $B_H(P)$ for P's lying within a Jordan curve called "coastline", such that B_H satisfies $B_H(P) = 0$ and also satisfies $B_H(P) > 0$ at all interior points close to the boundary. *Problems*: Derive the dimension of an island's boundary. (The zero-set where $B_H = 0$ is of dimension $2-H$.)

Cups. For each P within the coastline, define $B_H^*(P)$ as the infinimum, over all continuous curves from P to a point on the coastline, of the maximum of $B_H(P)$ on such a curve. A maximal connected open domain of constant $B_H^*(P)$ is called "cup" in FGN. The boundary of the unbounded component of the closed complement of a cup will be called a cup's "outer boundary". *Problem.* The union of outer cup boundaries is a ramified random fractal set. Study its structure. Is it a universal curve?

Some random fractal measures, and the fixed points of related smoothing transformations of probability distributions. Multiplicative chaos

Consider the following array of i.i.d. random variables: C r.v. $W(g)$, then C^2 r.v. $W(g, h)$, then C^3 r.v. $W(g, h, k)$, etc... Given a point $t \in [0, 1]$, write it in base C as $t = 0, t_1, t_2, \ldots, t_n, \ldots$ Define $X_n'(t) = W(t_1)W(t_2, t_2)$ $W(t_1, t_2, t_3)\ldots W(t_1, t_2, \ldots, t_n)$, and $X_n(t) = \int_0^t X_n'(s)\,ds$. *Problems, conjectures and partial answers.* Mandelbrot (*Comptes Rendus* (*Paris*) **278A** (1974), pp. 289 and 355) posed and solved in part many problems that are relative to a variety of classes of W's, and concern the weak or strong convergence of $X_n(t)$ to a non vanishing limit $X(t)$, the numbers of finite moments of $X(t)$ and the dimension of the set of t's on which $X(t)$ varies.

Partial answers. J. P. Kahane and J. Peyrière (*Adv. Math.* **22** (1970), p. 131) confirmed and/or extended several of these conjectures and theorems; for example, $EW \log_C W$ is a codimension.

Take C i.i.d. r.v. W_g, and C i.i.d. r.v. X_g having the same distribution as the $X(1)$ in the preceding paragraph. The weighted average $(1/C)\sum_{G=0}^{C-1} W_g X_g$ has the same distribution as each X_g, meaning that $X(1)$ is a fixed point of the weighted averaging operation. For "multiplicative chaos", see also Mandelbrot, *Lecture Notes Phys.* **12** (1972), p. 333.

Groups based upon inversions. Explicit construction of the fractal limit set

The next case story does not involve physics, but involves new geometric intuition triggered by long examination of sample limit sets drawn by computer, followed by the construction of further samples to test the original hunches. Consider the group $\mathscr{G}$ based upon inversions in a "generating configuration" of $M \geqslant 4$ given circles. It is known since Poincaré and Klein that, when the group is "Kleinian", the transforms of any initial point P_0 by increasingly long words in this group converge to a limit set $\mathscr{L}$ independent of P_0. *Long-standing problem.* To characterize $\mathscr{L}$ directly by a series of approximations that converges rapidly in simple generating configurations. *Solution* (FGN, Chap. 18, and *Mathematical Intelligencer*, **5** (2) (1983), p. 9). I showed that the (open) complement of $\mathscr{L}$ is approximated by a finite union of "σ-discs", each σ-disc being the union of an open disc constructed in a specified fashion, and of its transform under the group $\mathscr{G}$.

Iterates of the complex map $z \to z^2 - \mu$. The $\mathscr{F}^*$-sets and the $\mathscr{M}$-set

Fatou and Julia appear to have been exceptionally successful in their study of the iterates of rational functions of a complex variable, circa 1918. Indeed —apart from the proof of the existence of Siegel discs —their theory remained largely unchanged for sixty years. The fundamental discovery was that the repeller set of this iteration —now ordinarily called Fatou set, or Julia set, or $\mathscr{F}^*$-set —is typically a fractal: a nonanalytic curve or a "Cantor-like" dust. These sets were called "very irregular and complicated". The computer — which I was the first to use systematically here (Fig. 1) — reveals they are beautiful.

My work started with the quadratic map $z \to z^2 - \mu$.

To investigate how the shape of $\mathscr{F}^*$ depends upon the value of μ, I explored numerically the set of parameter values μ in the complex plane,

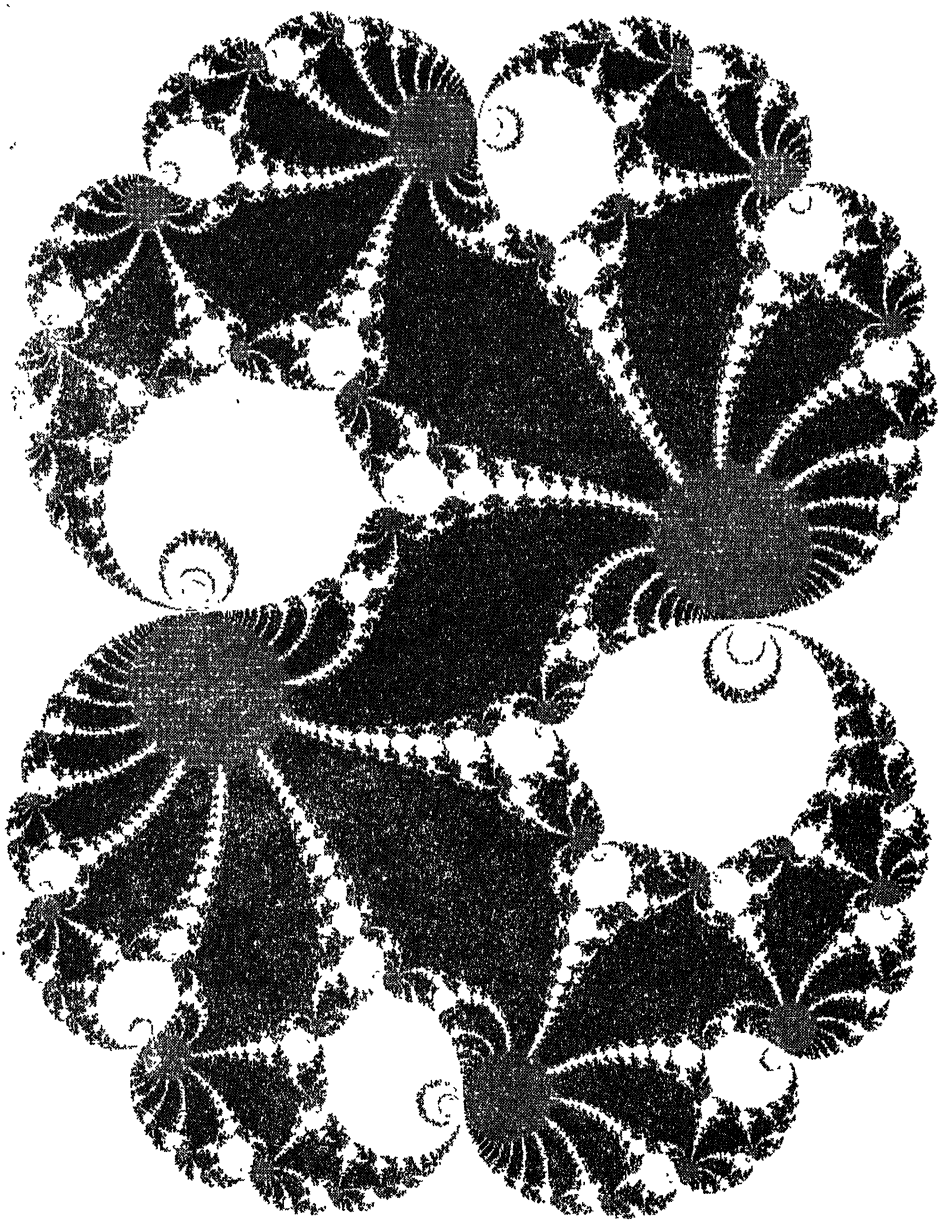


Fig. 1. Numerical approximation of the Julia $\mathcal{J}^*$-set of the map $z \to z^2 - \mu$ in a case when the limit set is made of the point as infinity and of 100 other points. The points in the light (resp., black) domain have already iterated into the neighborhood of infinity (resp., of the bounded cycle), and the iterates of the points in the gray domain are still wandering around.

such that $\mathscr{F}^*$ is connected. See FGN, Chap. 19; *Ann. NY Acad. Sci.* **357** (1980), p. 249; *Physica* **7D** (1983), p. 224, and several short papers in the book *Chaos, Fractals and Dynamical Systems*, eds. P. Fischer and W. Smith, New York, Marcel Dekker, 1984. This set—now called "$\mathscr{M}$-set" to echo "$\mathscr{F}^*$-set"—proved to be a most worthy object of study, first for "experimental mathematics" and then for mathematics. It is illustrated on Fig. 2.

Empirically observed structure of the $\mathscr{M}$-set. It is best described in the terminology of chemistry. *Atoms.* The $\mathscr{M}$-set contains a denumerable infinity of *atoms*, some of them cardioid-shaped and others near-discs. The interior of an $\mathscr{M}$-atom is a maximal domain of μ's such that the corresponding $\mathscr{F}^*$-sets are topologically identical. *Nuclei.* Each atom includes as *nucleus* a superstable value of μ. *Molecules.* Denumerably infinite numbers of atoms, including one and only one cardioid, are "bonded together" into "*molecules*", one of which is a large continent and the other are tiny specks. Any two atoms in a molecule can be joined by a continuous path that crosses a *finite number* of bonds and other atoms. The molecules' shapes are smooth maps of each other. *Devil's polymer.* The molecules are glued together via a "devil's polymer": any two molecules can be linked within the $\mathscr{M}$-set by curves that cross an *infinite number* of other molecules, and the points along such links that are not in a molecule's interior form a Cantor set. *The empirical n^{-2} law.* Change the coordinates of z so that the map reads $z \to f(z, \lambda) = \lambda z(1-z)$; this replaces μ by λ, such that $\mu = \lambda^2/4 - -\lambda/2$. Then the continental molecule in the $\mathscr{M}$-set becomes exceptional in that it is *not* based on a cardioid, but on the unit circles $|\lambda| = 1$ and $|\lambda - 2| = 1$. The points $\lambda = \exp(2\pi i m/n)$, with integer m and n, are points of bifurcation of a limit point cycle into a cycle containing n points. Whenever the map $f(z, \lambda)$ has a limit cycle, denote one point in the cycle by z_λ and let $f_n'(z_\lambda, \lambda)$ denote the value of $\partial f_n/\partial z$ at $z = z_\lambda$. It is observed that, for all n, the radial derivative of $f_n'(z_\lambda, \lambda)$ is exactly n^{-2}. A more general formulation is given on p. 1674.

Mathematical problem: Prove (or disprove) that the $\mathscr{M}$-set is connected. *Solution.* A. Douady and J. Hubbard have given an affirmative answer (*Comptes Rendus* **2941** (1982), p. 123). *Unsolved mathematical problems.* Prove (or disprove) that the boundary of the $\mathscr{M}$-set is of dimension 2. Prove or disprove the empirical n^{-2} law above. Prove or disprove that when $\mu \neq 0$ is the nucleus of a cardioid-shaped atom, the $\mathscr{F}^*$-set is the union of atoms having smooth boundaries. Characterize the class of rational or analytic mappings of the complex plane for which the $\mathscr{M}$-set is made—either exclusively or in part—of $\mathscr{M}$-molecules that are analytic maps of those of the $\mathscr{M}$-set of $z \to z^2 - \mu$.

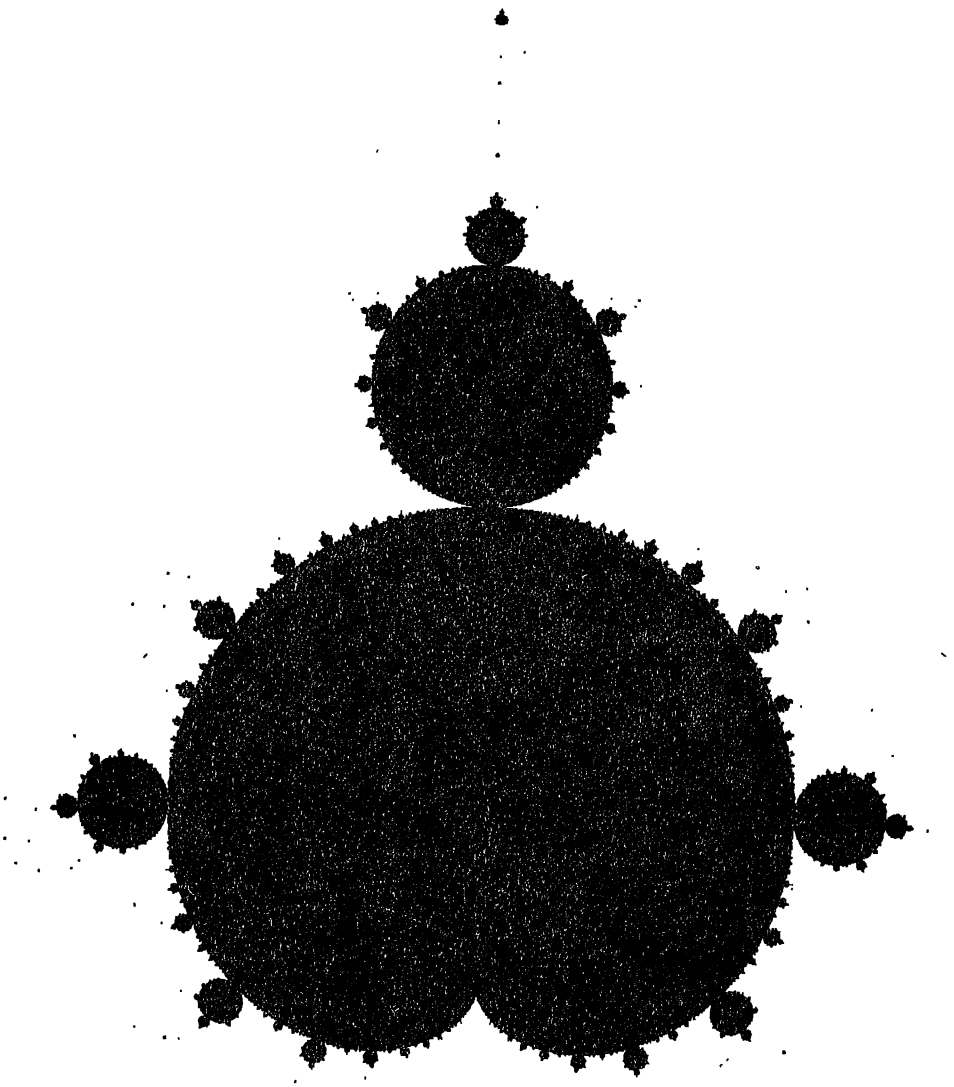


Fig. 2. Numerical approximation of the $\mathcal{M}$-set of the map $z \to z^2 - \mu$, after Mandelbrot [1]. For esthetic effect, its coordinate axes are not drawn; the real axis is vertical and is directed up, and the points where it intersects the cardioid are of abscissas $-1/4$ and $3/4$. The scattered dots are small molecules belonging to the $\mathcal{M}$-set. They are linked to the big molecule by a "polymeric" structure that is not visible here.

Postscript added in proof (May 1984)

The n^{-2} law I stated in *Chaos, Fractals and Dynamical Systems* (op. cit.) is more general. In the set of λ such that $f(z, \lambda)$ has a stable cycle of finite size, define $g(\lambda) = |(\delta/\delta z)f'_n(z, \lambda)|_{z=z_\lambda}$, where z_λ is a point in the cycle and n its size. Consider a point λ_0 that belongs to the boundaries of two distinct $\mathcal{M}$-atoms, of respective cycle sizes m and nm; I conjectured that the two atom boundaries have a common tangent at λ_0. Take the right and left derivatives of $|g(\lambda)|$ along a direction orthogonal to this tangent; I conjectured that the ratio of these derivatives is $-n^{-2}$. The class of validity of this law is bound to be related to the class discussed on p. 1672 (last lines).

While the n^{-2} law is local, it has perceptible global effects. Where the atoms of low cycle size are near-cardioids or near-discs, the near-discs' inverse "radii" also nearly obey the n^{-2} law, leaving the $\mathcal{M}$-molecule's shape to be determined by its cardioid's shape.

In April 1984, J. Guckenheimer and R. McGhee (to be published) proved my conjectured n^{-2} law for the quadratic map, gave wide sufficient conditions for its validity, and sketched generalizations.

On the notion of "fractal"

Though precision is of the essence in mathematics, the present text had abstained from defining my term *fractal*, but applied it to denote several sets that were already known to this audience of mathematicians. These sets did not warrant a common name as long as they served only fleetingly as "counterexamples", because any kinship one might have seen between them had no consequences. A common name became essential, however, when these sets were made essential in science, and when analogous—then not so analogous—sets started being constructed in very large numbers, as the tools of the *fractal geometry of nature*. I have heard mathematicians echo Molière, and joke that they had long been studying fractals without knowing they were fractals, but of course many writers had studied the additive group without knowing it was a group!

But how should a fractal set be defined? In 1977, various pressures had made me advance the "tentative definition", that a fractal is "a set whose Hausdorff–Besicovitch dimension strictly exceeds its topological dimension". But I like this definition less and less, and take it less and less seriously. One reason resides in the increasing importance of the "borderline fractals", for example of sets which have the topologically "normal"

value for the Hausdorff dimension, but have anomalous values for some other metric dimension. I feel—the feeling is not new, as it had already led me to abstain from defining *fractal* in my first book of 1975—that the notion of fractal is more basic than any particular notion of dimension. A more basic reason for not defining fractals resides in the broadly-held feeling that the key factor to a set's being fractal is invariance under some class of transforms, but no one has yet pinned this invariance satisfactorily.

Anyhow, I feel that leaving the fractal geometry of nature without dogmatic definition cannot conceivably hinder its further development.

General references

Note: For specific references, see the text or the bibliography of Ref. [1].

[1] Mandelbrot B., *The Fractal Geometry of Nature*, W. H. Freeman and Co., New York, San Francisco and Oxford, 1982. The second and later printings include an Update and additional references. Earlier versions of this essay were *Les objets fractals: forme, hasard et dimension*, Flammarion, Paris, 1975, and *Fractals: Form, Chance and Dimension*, Freeman, 1977.

[2] Sierpiński Waclaw, *Oeuvres Choisies*, Editions Scientifiques de Pologne, Warszawa, 1974.

IBM THOMAS J. WATSON RESEARCH CENTER
YORKTOWN HEIGHTS, NY USA 10598

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

YU. M. SVIREZHEV

Modern Problems of Mathematical Ecology

The recent years witnessed a real outburst in the number of works on mathematical modelling in ecology. Of course, one may suppose that this is the result of a sharp rise of interest in ecological problems in general and the popularity of the subject. But this is only one aspect. The other reason is that ecology represents a branch of biology (probably, together with genetics and the theory of evolution) that uses mathematical methods on such a wide scale that we may now speak about the birth of a new science — mathematical ecology.

What does mathematics get from this synthesis?

Firstly, ecology gives new fields for applying classical mathematical methods.

Secondly, ecology draws the attention of mathematicians to many problems the interest in which previously subsided due to the lack of either encouraging results or useful practical applications.

Finally, ecology offers new possibilities of posing mathematical problems.

Mathematics in turn provides a method of research without which many investigators of theoretical ecology would be in danger of falling into obscure sophistications, loquacious but fruitless.

To substantiate the above claims we shall illustrate each point by concrete examples.

1. "Predator–prey" system—a classical object of mathematical ecology

One of the most popular models in mathematical ecology is the model of a two-population system, one of which being food for the other. Such an interaction is widely spread in nature, it is called "predator–prey" interaction. The model itself is described by a system of two ordinary differen-

tial equations of the form

$$\begin{aligned} \frac{dx}{dt} &= a(x)x - V(x)y, \\ \frac{dy}{dt} &= kV(x)y - my, \end{aligned} \tag{1.1}$$

where $x(t)$ and $y(t)$ are the numbers of preys and predators, respectively, and the functions $a(x)$ and $V(x)$ must satisfy some conditions to be ecologically sensible. Evidently, system (1.1) represents a wide field of applications of the methods of qualitative theory of differential equations. Let us review in brief some results obtained (see for example [7], [8]).

One of the basic problems of ecology is: "Can the predator control the prey population?" Since $a(x)$ is a function describing the prey self-regulation, when we investigate this problem it is natural to assume that $a = \text{const}$, i.e., in the absence of the predator the prey population grows according to the Malthus principle. In this case the dynamics of the system significantly depends on the form of the trophic function, the rate of prey consumption by the predator. All the variety of trophic functions can be divided into two classes (see Fig. 1). Class I is characteristic for invertebrate

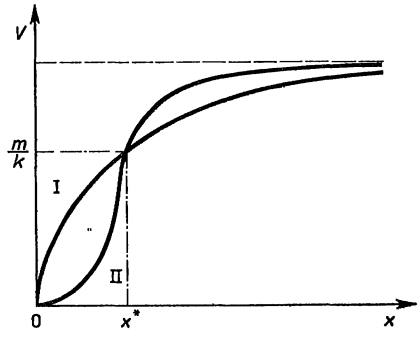


Fig. 1. Trophic functions classes: I — "silly" predator, II — "clever" predator.

predators, whereas Class II — for organisms, exhibiting a rather complex behaviour (e.g., the ability to learn). Many vertebrates manifest such a behaviour. We shall refer to the predators of Class I and Class II as "silly" and "clever" respectively.

An analysis of system (1.1) for $V(x)$ belonging to these two classes has shown that the "silly" predator cannot control the prey population,

i.e., the non-trivial equilibrium $\{x^*, y^*\}$ of system (1.1) is always globally unstable. If the predator is "clever" (i.e., $V(x)$ belongs to Class II), then the behaviour of the system is much more interesting. In this case either a stable non-trivial equilibrium is observed or there arise limit cycles in its neighbourhood. Let S_C stand for the region bounded by the Volterra oval

$$\left(\frac{e^X}{X}\right)^m \cdot \left(\frac{e^Y}{Y}\right)^\alpha = C, \quad X = x/x^*, \; Y = y/y^*$$

with the centre at the point $\{x^*, y^*\}$. Let us assume that, in a neighbourhood of x^*,

$$\left| w(x) = \frac{V(x)x^*}{V(x^*)x} \right| \sim 1.$$

Hence $|w(x)-1| \sim \varepsilon$. Consider the integral

$$\Psi = \int\int_{S_C} \left[\frac{V(x)}{x}\right]' dxdy.$$

Then if there exists a C^* such that

$$\Psi(C^*) = 0, \quad \Psi'(C^*) \neq 0,$$

then there exist $\mu, \Delta > 0$ such that:

(a) for every $\varepsilon < \Delta$, system (1.1) with $\alpha = \text{const}$, has a limit cycle in the μ-neighbourhood of the Volterra oval with a constant C^*; the limit cycle tends to the oval as $\varepsilon \to 0$.

(b) this cycle is stable when $\Psi'(C^*) > 0$ and unstable when $\Psi'(C^*) < 0$.

The transformation of variables in (1.1)

$$\xi = \ln(x/x^*), \quad \eta = \ln(y/y^*), \quad \tau = \alpha t$$

gives

$$\frac{d\xi}{d\tau} = 1 - w(\xi)e^\eta, \quad \frac{d\eta}{d\tau} = \delta[w(\xi)e^\xi - 1], \quad \delta = \frac{m}{\alpha}. \tag{1.2}$$

Let $w(\xi) = 1 + \varepsilon\varphi(\xi)$ where ε is a small parameter. Actually, $\varepsilon\varphi(\xi)$ represents the deviation of the more realistic trophic function from a hypothetic one of the Volterra (linear) type. The application of the Krylov–Bogolubov method to the system gives the following results ([5], [7], [8]):

Let

$$a = \varphi'(0), \quad b = \frac{1}{\delta}(\delta\varphi' + \varphi''')_0,$$

$$c = \frac{1}{192}(\delta^2\varphi' + 2\delta\varphi''' + \varphi^{\nabla})_0, \quad D = b^2 - 4ac.$$

Then depending on the signs of a, b, c, D we get different variants of the system's behaviour (r being the amplitude of the limit cycle).

(1) $a, b, c > 0$ or $a, c > 0$; $b, D < 0$: $r = 0$ — stable equilibrium with no limit cycles.

(2) $a, b > 0$; $c < 0$ or $a > 0$; $b, c < 0$: $r = 0$ — stable equilibrium with an unstable limit cycle $r = r_2$.

(3) $a, c, D > 0$; $b < 0$: $r = 0$ — stable equilibrium, $r = r_1$ stable cycle, $r = r_2$ — unstable cycle ($r_1 > r_2$).

(4) $a, b, c < 0$ or $a, c, D < 0$; $b > 0$: $r = 0$ — unstable equilibrium, there are no limit cycles.

(5) $a, b < 0$; $c > 0$ or $a < 0$; $b, c > 0$: $r = 0$ — unstable equilibrium, $r = r_1$ — stable cycle with self-exciting from zero.

(6) $a, c < 0$; $b, D > 0$: $r = 0$ — unstable equilibrium, $r = r_1$ — stable cycle with self-exciting, $r = r_2$ — unstable cycle ($r_2 > r_1$).

Suppose now that system (1.2) is in a random environment. Then the equations for the amplitude and phase are

$$\begin{aligned} \frac{dr}{d\tau} &= -\frac{\varepsilon r}{2}(a + br^2 + cr^4) + \sigma_1 n_1, \\ \frac{d\Phi}{d\tau} &= \sqrt{\delta} + \frac{\varepsilon}{2}(a + br^2 + cr^4) + \sigma_2 n_2, \end{aligned} \tag{1.3}$$

where n_1 and n_2 are δ-correlated white noises of constant intensity.

The expression for the stationary density of the probability of the amplitude is

$$P_0(r) = \text{const} \cdot r \cdot \exp\left\{-\frac{r^2}{\sigma_1^2}\left(a + \frac{b}{2}r^2 + \frac{c}{3}r^4\right)\right\}. \tag{1.4}$$

The phase is distributed uniformly (if we neglect the phase overlap). The function will have either one maximum in the neighbourhood of the equilibrium or a stable cycle, or two maxima, the latter case being possible only in the presence of two limit cycles: an unstable inner one and a stable outer

one. Consequently, the predator–prey system in random environment over large time intervals reveals four types of behaviour, namely:

1. The trajectories of the system leave the neighbourhood of the equilibrium, either quickly if there is no stable cycle in the neighbourhood or slowly if there are a stable inner cycle and an unstable outer cycle; the system will remain in the neighbourhood of the stable cycle for some time.

2. The diffusion of the trajectories around the stable equilibrium takes place if there are no limit cycles. The most probable values of the amplitudes of random oscillations lie in the neighbourhood of the stable point.

3. If the equilibrium is unstable and the stable limit cycle exists, then it becomes fuzzy, the stationary distribution being unimodal and its maximum lying to the right of the limit cycle amplitude.

4. The diffusion of the trajectories with the most probable values of the amplitude of the random oscillations lies in the neighbourhood of the stable equilibrium and the stable outer limit cycle (intensity of perturbations is sufficiently low).

Now let us try to answer another question: how much does the population dynamics of the system depend upon the ethological (behaviouristic) characteristics of the prey? The question is of importance as we have found out that the dynamics of the system depends essentially upon the fine structure of the trophic function, the latter in turn being a result of the predator's hunting strategy.

Now we pass to the behaviour of the prey. The simplest hypothesis about the "reasonableness" of the behaviour of the prey is the hypothesis of a collective behaviour (mutual aid) improving (up to a certain limit)

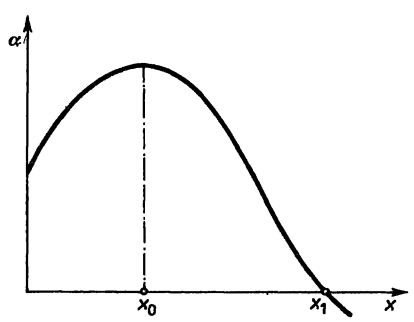


Fig. 2. Relative rate of population growth as a function of population size for collective behaviour (reciprocal help); $\alpha'(x_0) = 0$, $\alpha''(x_0) < 0$.

characteristics of the prey population, for example the relative rate of its growth. If we accept this hypothesis, the function $\alpha(x)$ must take the form given in Fig. 2. But making this aspect of the problem more complex, we simplify the other one: $V(x) = \beta x$, i.e., the trophic function is of the Volterra type. Then (1.1) may be written in the form

$$\frac{dx}{dt} = x[\alpha(x) - \beta y], \quad \frac{dy}{dt} = y(k\beta x - m). \tag{1.5}$$

Denote $x^* = m/k\beta$, $y^* = \alpha(x^*)/\beta$, $\beta_0 = mk/x_0$. Let the parameter β change in a small neighbourhood β_0, so that

$$\beta^2 > \frac{m}{4k^2} \cdot \frac{[\alpha'(x^*)]^2}{\alpha(x^*)},$$

i.e., small variations of β are considered. Since we deal here with a typical Hopf bifurcation, by applying a more or less standard technique we get the following picture of the dynamic behaviour of the system:

(a) for $\alpha'''(x_0)x_0 < -2\alpha''(x_0)$ and for $\beta > \beta_0$ in the neighbourhood of $\{x^*, y^*\}$ there appears a stable limit cycle,

(b) for $\alpha'''(x_0)x_0 > -2\alpha''(x_0)$ and for $\beta < \beta_0$ in the neighbourhood of $\{x^*, y^*\}$ there appears an unstable limit cycle. It should be noted that $\beta < \beta_0$ if $x^* > x_0$. Since the trajectories (1.5) are bounded and point $\{x_1, 0\}$ is a saddle point, this unstable cycle is certainly surrounded by a stable limit cycle. The question of the existence of a third, a fourth, etc. limit cycles remains open.

Since both the trophic function $V(x)$ and the prey self-regulation function $\alpha(x)$ are nothing but the aggregated averaged description of the *ethological* (behaviour) characteristics of predator and prey, the foregoing results lead us to the following *ecological* conclusion: the complex ethology manifesting itself at the level of individuals gives rise to a great diversity of the dynamic behaviour of populations, even in such an ecologically simple system as the "predator–prey" one.

It appears that we have succeeded in illustrating the possible applications of classical methods of the qualitative theory to a classical ecological problem. Here we shall not dwell upon the *ecological* interpretations of the *mathematical* results obtained. We shall only mention that these conclusions could not have been obtained in other ways. Certainly, there remain numerous other problems which are still to be solved and we hope that we have managed to draw attention to them.

2. Lagrange stability and ecological stability

In the previous section we used well-known mathematical methods for solving ecological problems; now we shall show how ecological "demands" make it possible to treat an old mathematical problem (not considered to be of interest before) in a new light.

We begin with a small ecological essay. One of the most common definitions of stability of the biological community is the requirement to maintain the number of species. It means that, on the one hand, none of the species dies out, and on the other, none of them grows infinitely.

If this happens in a model, the model is inadequate. Let the population sizes of species in the community be non zero and not grow to infinity and suppose that there exists no stable non-trivial equilibrium. Nevertheless, none of the species dies out in such a community. It means that the trajectories of the model are bounded from above and below in the positive orthant. However, if we consider here the Lyapunov stability, the community will be unstable; on the other hand, it will be stable if stability is understood ecologically.

Such a type of stability has long been known and is referred to as the *Lagrange stability.* However, while Lyapunov's theory is well-developed, we cannot say that about the Lagrange stability, though the latter is more suitable for ecology.

Suppose that the dynamics of a biological community is given by a system of ordinary differential equations:

$$\frac{dN_i}{dt} = F_i(N_1, \ldots, N_n), \quad i = \overline{1, n} \tag{2.1}$$

with initial conditions $N_i(0) = N_i^0$ where $N_i(t)$ are the sizes of the populations in the community. For our model to be biologically sensible, the following conditions must be fulfilled: $N_i(t) \geqslant 0$, $i = \overline{1, n}$ for all $t \geqslant 0$, $N_i^0 \geqslant 0$, meaning that the set $\boldsymbol{P}^n$ (i.e., the positive orthant of an n-dimensional space) is an invariant set for system (2.1).

Let Ω_0^n and Ω^n be closed finite domains lying within $\boldsymbol{P}^n$. We shall call a community, described by the model (2.1), *ecologically stable*, if for any $\vec{N}^0 = \{N_1^0, \ldots, N_n^0\} \in \Omega_0^n$ there exists $\Omega^n(\Omega_0^n)$ such that $\vec{N}(t) = \{N_1(t), \ldots, N_n(t)\} \in \Omega^n$ for all $t > 0$, or in formal terms:

$$\forall \vec{N}^0 \in \Omega_0^n \; \exists \Omega^n(\Omega_0^n) \subset \mathrm{Int}\boldsymbol{P}^n\colon \; \forall t > 0, \; \vec{N}(t) \in \Omega^n.$$

Since all N_i^0 are positive, then by substituting $\xi_i = \ln(N_i/N_i^0)$ in (2.1), we get

$$\frac{d\xi_i}{dt} = \varphi_i(\xi_1, \ldots, \xi_n;\ N_1^0, \ldots, N_n^0), \quad i = \overline{1, n} \tag{2.2}$$

with initial conditions $\xi_i(0) = 0, i = \overline{1, n}$.

Evidently, for $N_i(t) \to +\infty$ we have $\xi_i(t) \to +\infty$ and for $N_i(t) \to 0$ we have $\xi_i(t) \to -\infty$ (for finite N_i^0). Thus the solutions of system (2.2) are defined in the whole phase space $\boldsymbol{R}_\xi^n$ (not only in the positive orthant). Since the conditions $\varphi_i(0, \ldots, 0;\ N_1^0, \ldots, N_n^0) = 0$ are not obligatory, i.e., $\xi_i^* = 0, i = \overline{1, n}$ is not the solution to (2.2), we rearrange (2.2) to the form

$$\frac{d\xi_i}{dt} = \Phi_i(\xi_1, \ldots, \xi_n;\ N_1^0, \ldots, N_n^0) + B_i, \quad i = \overline{1, n}, \tag{2.3}$$

where

$$\Phi_i = \varphi_i(\xi_1, \ldots, \xi_n;\ N_1^0, \ldots, N_n^0) - \varphi_i(0, \ldots, 0;\ N_1^0, \ldots, N_n^0),$$

$$B_i = \varphi_i(0, \ldots, 0;\ N_1^0, \ldots, N_n^0).$$

If we now formulate for (2.3) the problem of the Lyapunov stability of the trivial solution $\xi_i^* = 0, i = \overline{1, n}$ under permanent perturbations and find the domain of stability in the space of parameters $\{N_1^0, \ldots, N_n^0\}$, then the solution of this problem will imply the solution of the problem of ecological stability in system (2.1). The proof of this statement obviously follows from the Lyapunov stability definition and the properties of the mapping $\boldsymbol{P}^n \Rightarrow \boldsymbol{R}_\xi^n$.

Thus we have reduced the analysis of the Lagrange stability to the problem of the Lyapunov stability. By Malkin's theorem [4] the solution $\xi_i^* = 0, i = \overline{1, n}$ is Lyapunov stable if it is asymptotically stable for the system $\dot{\xi}_i = \Phi_i(\xi_1, \ldots, \xi_n;\ N_1^0, \ldots, N_n^0), i = \overline{1, n}$, with B_i sufficiently small.

Let us clarify all the above by an example. Suppose the population dynamics is given by the equation

$$\frac{dN}{dt} = rN\left(1 - \frac{N}{K}\right)(N - k), \quad 0 < k < K \tag{2.4}$$

(this is the so-called population with lower threshold number). The phase picture of this equation is presented in Fig. 3. Evidently $N(t) \to 0$ for

$N_0 < k$ and the population is destined to extinction, i.e., there is no ecological stability.

On the other hand, the equilibrium $N^* = K$ is asymptotically stable. Let us see whether we can prove the ecostability of this population, by using the method described above. Substituting $\xi = \ln(N/N_0)$ we get

$$\frac{d\xi}{dt} = r\left(1 - \frac{N_0 e^{\xi}}{K}\right)(N_0 e^{\xi} - k) = \Phi(\xi, N_0) + B,$$
$$\Phi(\xi, N_0) = \frac{rN_0}{K}(K + k - 2N_0)\,\xi + o(\xi), \tag{2.5}$$

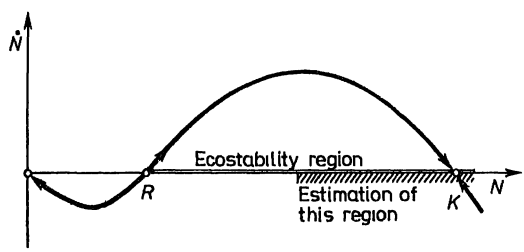


Fig. 3. Population phase portrait with lower threshold number of population.

where

$$B = r\left(1 - \frac{N_0}{K}\right)(N_0 - k).$$

It is obvious that the solution $\xi^* = 0$ for $\dot{\xi} = \Phi(\xi, N_0)$ is asymptotically stable for $N_0 > \frac{1}{2}(K + k)$. Besides, if, for example, the Malthusian parameter r is small, then B is small, too. Then, according to the Malkin theorem, the trivial solution $\xi^* = 0$ is stable (in Lyapunov's sense) under a permanent perturbation B. Consequently, we can state that for $N_0 > \frac{1}{2}(K + k)$, the function $N(t)$ will be bounded from above and below for all $t > 0$, i.e., the population will be ecologically stable.

Note that by using this method we have obtained a stronger restriction, namely the population is ecologically stable for all $N_0 > k$. This condition is after all natural, since the Lyapunov stability conditions are the sufficient ones.

Clearly, the requirement for the non-trivial equilibrium $\vec{N}^*$ in (2.1) (if it exists) to be stable, is not, in general, a necessary condition for the ecostability. However, there exists quite a wide class of ecological models

in which the conditions of ecological stability and the existence of a positive stable equilibrium in Lyapunov's sense turn out to be equivalent. This class contains the communities called *conservative* and *dissipative* (according to Volterra) [7], [8]. We mean here systems of the type

$$\frac{dN_i}{dt} = N_i\left(\varepsilon_i - \sum_j \gamma_{ij} N_j\right), \quad i, j = \overline{1, n}$$

in which there is a single positive equilibrium $\vec{N}^* = \{N_1, \ldots, N_n\}$ and the quadratic form $\sum_i \sum_j a_i \gamma_{ij} N_i N_j$; $i, j = \overline{1, n}$ (where $a_i > 0$) is either identically zero or positive definite.

3. Competition for resource, the self-thinning out problem and "Schrödinger" systems

In this section we shall try to show how the attempts to solve some ecological problems result in new non-traditional formulations of mathematical problems.

Suppose we have the biomass distribution $N(x, t)$ and the consumed resource (nutrients) distribution $R(x, t)$, $-\infty < x < +\infty$. We assume that the uptake rate of the resource, located at point ξ by the unit of biomass, located at point x, is equal to

$$P(|x - \xi|)\, V[R(\xi, t)].$$

Here $P(|x - \xi|)$ can be represented by the density of the normal distribution with the centre at the point x and variation σ^2. Then the equations of this model will take the form

$$\begin{aligned} \frac{\partial R(x, t)}{\partial t} &= Q - \int_{-\infty}^{+\infty} P(|x - \xi|)\, V[R(x, t)] N(\xi, t)\, d\xi, \\ \frac{\partial N(x, t)}{\partial t} &= k \int_{-\infty}^{+\infty} P(|x - \xi|)\, V[R(\xi, t)] N(x, t)\, d\xi - mN(x, t). \end{aligned} \tag{3.1}$$

Here Q is the input flow of resource, k is the efficiency, and m is a coefficient of natural mortality. Considering $\sigma \ll L$, where L stands for the specific scale in the problem (which corresponds to the assumption that the radius of the effective interaction between the consumer and the resource is small),

we bring (3.1) to its asymptotic analogy:

$$\begin{aligned}\frac{\partial R}{\partial t} &= Q - V(R)\left\{N + \frac{\sigma^2}{2}\frac{\partial^2 N}{\partial x^2}\right\},\\ \frac{\partial N}{\partial t} &= kN\left\{V(R) + \frac{\sigma^2}{2}\left[V''\left(\frac{\partial R}{\partial x}\right)^2 + V'\frac{\partial^2 R}{\partial x^2}\right]\right\} - mN.\end{aligned} \tag{3.2}$$

This system is interesting in itself but to understand better what it really is we shall take $V(R) = aR$ and linearize it in the neighbourhood of a spatially homogeneous stationary solution $\{R^*, N^*\}$ where $R^* = m/ka$, $N^* = kQ/m$. Denoting $Z_1 = R - R^*$ and $Z_2 = N - N^*$ and considering Q, m, k to be constant, we get

$$\frac{\partial \vec{Z}}{\partial t} = A\vec{Z} + D\frac{\partial^2 \vec{Z}}{\partial x^2}, \quad \vec{Z} = \{Z_1, Z_2\}, \tag{3.3}$$

where

$$A = \begin{bmatrix} -akQ/m & -m/k \\ ak^2Q/m & 0 \end{bmatrix}, \quad D = \begin{bmatrix} 0 & -m\sigma^2/2k \\ ak^2Q\sigma^2/2m & 0 \end{bmatrix}.$$

Since the eigenvalues of the matrix D are purely imaginary, system (3.3) is not parabolic, according to Petrovsky. Similar equations arise in quantum mechanics and therefore we shall refer to these systems (3.1)–(3.3) as "Schrödinger" ones. Practically there is no theory of such systems, whereas their solutions (if they exist) may have fairly interesting properties. In particular, there may exist spatially periodic discontinuous solutions belonging to the type of finite functions. Such solutions describe the so-called "self-thinning out" processes in plant communities when from an initial everywhere continuous distribution of biomass there arises a stable discrete structure. However, these are only hypotheses and this new mathematical object — "Schrödinger systems" — should be subjected to profound investigation.

4. Models of spatially distributed ecosystems—ecologically active media

Account of the movements of individuals over the areal forces us to consider a spatial generalization of the "predator–prey" or "resource–consumer" models. So this naturally gives rise to the concept of an *ecologically active medium*, in which there may occur either the propagation of nonlinear population waves or the rise of stationary stable distributions of popula-

tions nonhomogeneous in the space (even in a homogeneous environment), the so-called dissipative or spatial structure (SS) [1], [3].

Since random migration of individuals is quite well described by the diffusion terms (where the "diffusion" coefficients are uniquely defined by the radii of inidividual activity), the models of spatially distributed ecosystems are systems of the quasi-linear parabolic type

$$\frac{\partial N_i}{\partial t} = D_i \Delta N_i + f_i(N_1, \ldots, N_n), \qquad i = \overline{1, n}, \tag{4.1}$$

where $N_i(\vec{x}, t)$ are the population densities for the ith species at point $\vec{x} = \{x_1, \ldots, x_m\} \in \Omega$, Ω being the total areal of the ecosystem. The functions f_i stand for the local interactions of populations. On the boundary Γ of the domain, the boundary conditions are to be specified to provide a correct solvability of the initial value problem, for instance the condition of the non-permeability of the boundary

$$\left.\frac{\partial \vec{N}}{\partial \vec{n}}\right|_{\Gamma} = 0, \qquad \vec{n} \text{ is a normal to } \Gamma. \tag{4.2}$$

For $n = 1$ there are classical results indicating the absence of SS in convex domain under conditions (4.2). Now we shall quote the result generalizing the above assertion [4].

Let $f(N)$ be such that there are no two roots of equation $f(N) = 0$, $N_1^* < N_2^*$ such that $F(N) = \int_{N_1^*}^{N} f(N)dN$ for $N \in (N_1^*, N_2^*)$ and $F(N_2^*) = 0$. Then no stationary bounded solution of the equation

$$\frac{\partial N}{\partial t} = \Delta N + f(N), \qquad \vec{x} \in \boldsymbol{R},$$

except the constant one is stable. The stability is understood here in the sense of the norm $C(\boldsymbol{R})$.

It follows that the search for SS should begin either on areas of a more exotic configuration (e.g., non-convex) or for interacting populations ($n \geqslant 2$). The most natural approach is based on the investigation of the character of stationary bifurcation of the solution under variations of D_i in (4.1). Let $n = 2$. Then (4.1) and (4.2), in the neighbourhood of the

nontrivial stationary distribution $\{N_1^*, N_2^*\}$, take the form ($u_i = N_i - N_i^*$):

$$\frac{\partial \vec{u}}{\partial t} = D(\lambda)\Delta\vec{u} + A\vec{u} + g(\vec{u}), \quad \left.\frac{\partial \vec{u}}{\partial n}\right|_\Gamma = 0, \tag{4.3}$$

where $A = [a_{ij}]$ is the matrix of the linearized system, $D = \operatorname{diag}\{1, \lambda\}$, $g(\vec{u})$ is a nonlinear component, considered to be small. It turns out that necessary and almost always sufficient conditions for the existence of a critical value $\lambda_0 > 0$, in the neighbourhood of which there exists a family of stationary spatially nonhomogenous solutions of (4.3) are the following:

(1) $\det A > 0$,

(2) $\operatorname{tr} A < 0$,

(3) $a_{11} < 0$

if $a_{22} > 0$ or $a_{11} > |\varkappa_1|$, where $\varkappa_i, i = 0, 1, \ldots$ is the ith eigenvalue of the Laplace operator in the domain Ω.

If the domain Ω is not too symmetrical, one of the two semibranches of this family corresponds to the stable solution, i.e., to SS. If this is not the case (e.g., the domain Ω is one-dimensional) then additional verification of stability is needed.

Among the systems of the (4.1)-type a special place is held by those having some D_i equal to 0. Such systems are not parabolic in Petrovsky's sense and that is why this case gives rise to complex mathematical problems connected with the existence and stability of solutions. Here is one of the results on linear stability of a stationary spatially homogeneous solution.

Let system (4.1), linearized in the neighbourhood of that solution, have the form ($u_i = N_i - N_i^*$):

$$\frac{\partial \vec{u}}{\partial t} = D\Delta\vec{u} + A\vec{u}, \quad \vec{u}(\vec{x}, t) \in \boldsymbol{R}^n, \quad \vec{x} \in \boldsymbol{R}^m, \tag{4.4}$$

where A and D are real $n \times n$ matrices, D being diagonal with non-negative elements. If there is a $\delta > 0$ such that for each $s \geqslant 0$ the eigenvalues of the matrix $A + \delta E - Ds$ lie in the left half-plane, then the trivial solution of (4.4) is stable with respect to the perturbations $\vec{u}_0(\vec{x}) \in C^\infty(\boldsymbol{R}^n)$.

Note, that from the ecological point of view such an approach is by no means exotic since $D_i \equiv 0$ correspond to species of plants. As an example let us consider the "resource–consumer" system with immovable resource.

$$\begin{aligned} \frac{\partial R}{\partial t} &= Q - V(R)N, \\ \frac{\partial N}{\partial t} &= D\Delta N + [kV(R) - m]N. \end{aligned} \tag{4.5}$$

Here $R(x, t)$ and $N(x, t)$ are the densities of the resource and consumer respectively, given in the one-dimensional infinite areal. Applying the previous statement to system (4.5) one can easily prove the linear stability of the stationary distribution

$$N^* = kQ/m, \quad R^* = V^{-1}(m/k).$$

In the case of a local outbreak of the consumer population this distribution is settled following the wave, propagating at a velocity $v = 2\sqrt{D[kV(\infty) - m]}$. If $Q \equiv 0$ (i.e., the resource is unrenewable) and the initial resource density is R_0, then the outbreak of the consumer population generates a single wave, spreading at a velocity of $v = 2\sqrt{D[kV(R_0) - m]}$ [1].

However, if the areal contains so-called "dead zones", i.e., regions x in which $R_0(x) = 0$, then the wave velocity in the zone will be close to the value

$$v_z = [kV(\infty) - m] \cdot \sqrt{D/kV(\infty)}.$$

What is curious is that v_z does not depend on the size of the zone.

And what is the picture if the resource is being restored according to the Malthusian or logistic law, i.e., if

$$Q(R) = a(R)R, \quad a(0) > 0, \quad a'(R) < 0, \quad a(R^*) = 0?$$

This model is nothing but a spatial generalization of the "predator–prey" system. It turns out that for appropriate parameter values in automodel variables $\xi = x + vt$ there exist periodic solutions. In real variables they are represented by "wave packets" or "wave pattern" — successions of running waves. This means that the local outbreak of the consumer arising at a certain time starts to work as a generator of waves propagating over an ecologically active medium occupied by the consumer. The minimum velocity of those waves is

$$v_{\min} = 2\sqrt{D[kV(R^*) - m]}.$$

The topic is discussed in detail in our paper [6].

In this section we have attempted to show how the study of nonlinear waves (the dissipative structures may also be included in this class) in models of spatially distributed ecosystems gives rise to a new class of problems which are of both theoretical and practical interest. The main problem here is to prove the convergence of the solution to a stationary wave (being at rest or running) for a sufficiently wide class of initial

conditions. This problem is solved only for very simple models (cf. the classical works of Kolmogorov, Petrovsky and Piskunov), thus representing a wide field of action. Another important class of problems is the generation and propagation of waves in active two-dimensional ecological media in a plane areal.

It seems that nothing has been done in this direction, except for a lot of computer experiments.

5. Strange attractors in simple ecosystems

Recently it has become quite popular to look for examples of complex dynamic behaviour in systems of ordinary differential equations which are known as strange attractors. Such behaviour does not appear to be very exotic in ecology. Consider the model of a simple ecosystem: a closed trophic chain with three levels. If N_0 is the resource concentration (e.g., nutrients) and N_i are the biomasses in the trophic levels, then this chain is described by the equations

$$\begin{aligned} &\frac{dN_i}{dt} = V_{i-1}(N_{i-1})N_i - V_i(N_i)N_{i+1} - m_i N_i, \\ &i = 1,2,3, \quad N_4 \equiv 0, \quad N_0 = C - (N_1 + N_2 + N_3). \end{aligned} \tag{5.1}$$

Here $V_i(N_i)$ are trophic functions, $C = \text{const}$ is the total amount of matter in the system. Let

$$V_0 = a_0 N_0, \quad V_i(N_i) = \frac{aN_i}{1+bN_i}, \quad i = 1,2,3.$$

For this case B. I. Yatsalo and myself succeeded to prove analytically the possibility of generating a cycle as a result of the Hopf bifurcation and also the existence of a stable limit cycle for large $a_0 \sim 1/\varepsilon$. The system was studied numerically for $m_1 = 0.1$, $m_2 = m_3 = 0.2$, $a = b = 0.2$ and for different values of the parameters C and a_0 $(0.2 < a_0 < 0.38)$. The system develops in the following way: for $C = C_1^0 = C_1^0(a_0)$ the Hopf bifurcation results in the stable limit cycle $\gamma_1 = \gamma_1(C)$; for $C = C_2^0$ two cycles are generated : the stable γ_2 and the unstable γ_3 (i.e., for $C \to C_{2+}^0$ they unite through $+1$); for $C = C_3^0$ the cycles unite and vanish. Consequently the bifurcation doubling takes place at points C_1, C_2, C_3, i.e., for $C = C_1$ the cycle γ_2 loses stability, its multiplicator passes through -1 and in the

neighbourhood there arises a stable cycle with a double period which goes through the same bifurcation at $C = C_2$, etc., where $C_n \to C_\infty = C_\infty(\alpha_0)$. Taking $\alpha_0 = 0.34$ we get the following values of the parameters: $C_1^0 = 9.447$, $C_2^0 = 30.55$, $C_3^0 = 36.251$, $C_1 = 33.04$, $C_2 = 34.835$, $C_3 = 35.41$, $C_4 = 35.541$, ..., $C_\infty \simeq 35.58$.

Feigenbaum's constant calculated according to the above values of C_3, C_4, C_5 is

$$k = \frac{C_4 - C_3}{C_5 - C_4} = 4.5 + \beta, \qquad 0 < \beta < 0.1$$

and so it is close to the theoretical one:

$$k_T = \lim_{n \to \infty} \frac{C_n - C_{n-1}}{C_{n+1} - C_n} \simeq 4.66.$$

The analysis of curves $C_\infty(\alpha_0)$ and $C_3^0(\alpha_0)$ gives a very interesting result (see Fig. 4): in the hatched domain "the pre-turbulent regime" (pre-sto-

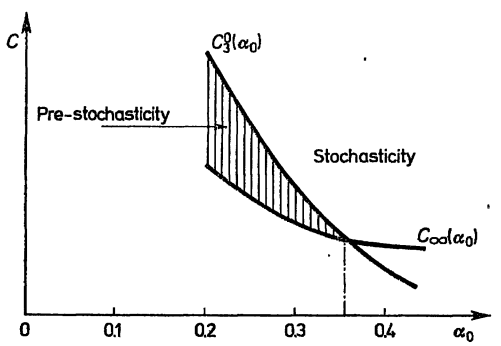


Fig. 4. Existence domains of different dynamic regimes for the closed trophic chain.

chasticity) exists when there are both a strange attractor resulting from an infinite chain of the Feigenbaum doubling and a stable limit cycle. Complete stochasticity is observed for $C > C_3^0$, i.e., when the stable cycle vanishes. It is interesting that stochasticity was previously observed in the classical Lorentz system, but there the strange attractor did not appear as a result of bifurcation doubling. What has been said above, indicates the possibility of the generation of strange attractors of a new type, even in simple ecosystems.

Conclusions

In this report abstaining from superfluous mathematical strictness, we have tried to show what an extensive field of action the new science of mathematical ecology offers to mathematicians.

Since previously the main clients of mathematics were the physical sciences, it was physics that to a great extent determined the interest in this or that field of mathematics. With the rise of a need for mathematical ecology many mathematical methods, developed for physics, turned out to be inapplicable. The need for new methods (or non-traditional applications of the old ones) entailed the formulation of new problems. Mathematical physics took shape as a science in the 19th century, whereas the 20th century may be considered as marking the birth of mathematical ecology, ecology being the science about our home, the home we live in.

References

[1] Gigauri A. A. and Svirezhev Yu. M., Wave Propagation in the Resource–Consumer System, *Dokl. Akad. Nauk. SSSR* **258** (5) (1981), pp. 1274–1276 (in Russian).

[2] Malkin I. G., Theory of Motion Stability, Nauka, Moscow, 1969 (in Russian).

[3] Razzhevaikin V. N., On the Generation of Dissipative Structures in a System of Two Reaction–Diffusion Equations, *Soviet. Mat. Dokl.* **22** (3) (1980), pp. 839–841.

[4] Razzhevaikin V. N., Instability of the Stationary Nonhomogenous Solutions of the Initial Value Problem for the Quasilinear Parabolic Equation, *Zhurn. Vychisl. Mat. i Mat. Fiz.* **20** (1980), pp. 1328–1333 (in Russian).

[5] Sidorin A. P. and Svirezhev Yu. M., The Dynamics of Predator–Prey System under Random Disturbances, *Izviestia Akad. Nauk SSSR, biol. ser.* **4** (1980), pp. 573–579 (in Russian).

[6] Svirezhev Yu. M., Gigauri A. A., and Razzhevaikin V. N., Waves in Ecology, In: *Nonlinear Waves: Self-Organization*, Nauka, Moscow, 1983, pp. 32–47 (in Russian).

[7] Svirezhev Yu. M. and Logofet D. O., Stability of Biological Communities, Nauka, Moscow, 1978 (in Russian).

[8] Svirezhev Yu. M. and Logofet D. O., Stability of Biological Communities, Mir, Moscow, 1983.

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

HANS FREUDENTHAL

The Implicit Philosophy of Mathematics History and Education

Here you are: both of the subjects of our section of this congress united in one title, and as though this were not yet enough, philosophy, once the third in this illustrious company, resurrected — though, of course, philosophy not meant the way it is understood in section II, that is, as a mature offspring of mathematics itself and equally ranking with other offsprings. I mean it rather the way it is meant when you ask someone "tell me, what is your philosophy?" It is implicit philosophy which does not need, nor ask for, formalized language to be made explicit.

I have omitted one thing in the title, for brevity's sake. I should have added "in their mutual relation". Should I really have? I introduced "philosophy" in the singular rather than in the plural, that is, one philosophy behind both history and education, or if they are two, that one that is common to both.

Philosophy of history often means what we can learn from the past to cope with the future. In our particular case it could mean what we can learn from the history of old mathematics for the sake of teaching young people. Strangely enough nobody has ever looked at the converse idea, that is, what can we learn from educating the youth for understanding the past of mankind? This idea looks odd, but I will show you it is not as far-fetched as you might believe.

A century ago biologists were the first to assert the so-called biogenetic law: that ontogenesis is an abridged recapitulation of phylogenesis, that is, that the individual in its development briefly repeats the development of its kind. We know for sure that this law is not true in this trivial way. But neither is it true that the new generation starts just at the point where its predecessors finished. Our biological, social, and mental life starts somewhere in the past of our race, at stages where man was not yet what he is now.

The young learner recapitulates the learning process of mankind, though in a modified way. He repeats history not as it actually happened but as it would have happened if people in the past had known something like what we do know now. It is a revised and improved version of the historical learning process that young learners recapitulate.

"Ought to recapitulate" — we should say. In fact we have not yet understood the past well enough to really give them this chance to recapitulate it. Let me show this by examples, which are more convincing than abstract statements can ever be.

Negative numbers

Negative numbers were a conquest of the 16th century. Why weren't they welcomed earlier? Well, there is little need to calculate $3-5$ or to solve equations like $2x+7=3$. Even quadratic equations, known as early as Babylonian antiquity, like $x^2+x-2=0$, did not provide a strong enough incentive to extend the number domain to negative numbers. In contrast, fractions and even irrational numbers are almost as ancient as natural numbers, thanks to the necessity of dividing and measuring. It was the destiny of the marvellous formula used in solving the cubic equation to open up the clogged channel of history. Three solutions forced themselves upon the enchanted solver. Who would dare to despise this wealth and throw away part of it? So negative numbers knocked at the door and they were welcomed as were imaginary numbers, which knocked as forcefully. Welcomed? Yes, and no. It was only as late as the 19th century that the last resistance to negative and imaginary numbers was conquerred. For a long time people seriously doubted whether man was allowed to create new numbers beyond the realm created, as they believed, by nature.

Meanwhile experience and history have taught us revolutionary lessons. If some beautiful formula, some theorem, some theory refuses to apply as generally as we would like it, we now put the blame not on the formula, the theorem, the theory, but on the problem to be solved. Problems are often being adjusted to the solution rather than the other way round.

The cubic equation was one of the first examples of this behaviour. Pressing the solution at any price led to negative and imaginary numbers. The first extension of natural numbers, towards fractions, had been much less controversial. From the first mathematical documents onwards we meet with fractions. Soul-searching in this domain was of a much later

date, in Greek mathematics, when philosophers forbade breaking the unit. Greek mathematicians replaced fractions with ratios while calculators in commerce and science persisted in using fractions. Indeed, fractions are the natural tool if magnitudes are measured and divided.

* Mathematics in the Greek sense is about numbers, and as far as geometry is concerned, about magnitudes — a view that mathematicians in more recent times have tried to share, at least in theory. The negative numbers originated from the formal algebraic need for general validity of solving formulae, but not until the algebraization of geometry (the so-called analytic geometry of former times) did they become effective — I mean effective in terms of real contents. The idea to use algebra to describe geometric figures and solve geometric problems is older than Descartes. We owe to Descartes the tendency to use one coordinate system (to express it in modern terms), independently of the figure and the problem. Descartes still had some trouble with negative numbers. Indeed, numbers were introduced as magnitudes; letters indicated magnitudes, thus positive numbers. But those who applied Descartes' method could no longer avoid having letters mean negative numbers also. If straight lines are to be described algebraically in their totality, if curves are to be described algebraically in any situation, one cannot but admit negative values for the variables. The need for

general validity of algebraic solution methods,

to which the negative numbers owed their existence, is from the 17th century onward reinforced by the need for

general validity of descriptions of geometric relations.

The second need, more directed towards contents than the formal algebraic one, is the most natural and compelling. It is actually responsible for the success story of negative (and also of complex) numbers.

If negative numbers are introduced, it does not suffice to claim their mere existence — this is often didactically overlooked, as also happens with rational numbers. Negative numbers become operational by their use in calculations, obeying certain laws which are uniquely determined as extensions of certain laws governing the positive numbers. This is

the algebraic permanence principle,

which includes what I just called the

general validity of algebraic solution methods,

and virtually it is the same idea, albeit formulated in a broader view.

I recall a few examples of the algebraic permanence principle.

$$(-3)+(-4) = -(3+4)$$

is proved by starting with the definition equations for $-a$,

$$(-3)+3 = 0, \quad (-4)+4 = 0,$$

adding them formally, then using commutativity and associativity, in order to arrive at the definition equation

$$\big((-3)+(-4)\big)+(3+4) = 0$$

for $-(3+4)$.

Or: Starting with the same definition equations, one proves

$$(-3)\cdot(-4) = 3\cdot 4,$$

by multiplying distributively the first by 4 and the second by -3,

$$4\cdot(-3)+4\cdot(3) = 0, \quad (-3)\cdot(-4)+(-3)\cdot 4 = 0$$

and subtracting them from each other.

Or: With $\sqrt{a}$ defined as the x making $x^2 = a$, one gets

$$\sqrt{a}\,\sqrt{b} = \sqrt{ab}$$

by multiplying the definition equations

$$x^2 = a, \quad y^2 = b$$

to get

$$(xy)^2 = x^2y^2 = ab.$$

Similarly, if operations are to be extended,

$$a^{1/n} = \sqrt[n]{a}$$

because both terms have the same n-th power. **

For a century the algebraic permanence principle has been ridiculed as a sham. The axiomatic method should have taught us sounder lessons. It is the way we always proceed when extending mathematical definitions. It is the way we find out how to extend mathematical objects in a reasonable and unique way, to prove the uniqueness of the extension, and to prepare the construction that proves the intended extension. It is the way negative numbers have been taught until quite recently, when new didactic ideas emerged. I will focus on one of them only, the number line, on which

negative numbers are viewed as movable vectors which are being operated on as such. It is so splendid an idea that one marvels why it has not worked didactically. P. M. van Hiele has been the first to indicate the reason, which is so simple that one marvels even more why nobody before him hit upon it: dimension one is the least appropriate to give vectors the chance they deserve. If you do not believe it, look up in all those textbooks the desperate attempts visibly to separate, add and subtract vectors, which unfortunately in one dimension cover and eclipse each other.

In Van Hiele's newest approach negative numbers arise in a two-dimensional frame. A number pair

$\ulcorner 3, \quad 4 \urcorner$ means 3 steps to the right, 4 steps upwards,
$\ulcorner -3, \quad 4 \urcorner$ means 3 steps to the left, 4 steps upwards,
$\ulcorner 3, \quad -4 \urcorner$ means 3 steps to the right, 4 steps downwards,
$\ulcorner -3, \quad -4 \urcorner$ means 3 steps to the left, 4 steps downwards.

The left-right and up-down are those of the drawing plane, with horizontal and vertical axes on which the numbers of steps can be read in units. By performing such operations in succession one describes or prescribes rectilinearly constructed drawings in the plane. Adding the vectors is nothing but performing these operations in succession. Thus

$$\ulcorner 3, \ -4 \urcorner + \ulcorner -5, \ 2 \urcorner$$

arises in a natural way and defines as naturally what it is

$$3+(-5) \text{ and } (-4)+2.$$

In the two-dimensional model the laws governing the operations of addition and multiplication are visually obvious by virtue of the model's geometric meaning.

In van Hiele's many-sided approach this extension of the number domain is applied to extend the definitions of functions introduced earlier by means of tables such as

$$x \to x-3,$$
$$x \to x+3,$$
$$x \to 3-x,$$
$$x \to 2x.$$

Let us now have instruction starting with these functions. But before doing so let us turn once more back to history and remember that negative numbers were first invented to * extend broader validity to certain algebraic solving methods while soon their indispensability in geometry became overpowering — I mean their indispensability in the algebraized geometry such as developed after Descartes. Negative numbers would have remained a nice plaything, and the operations, motivated by algebraic permanence, mere rules of a game, which could have been fixed in another way, were it not that geometry had seized upon them. Negative numbers are indispensable if the whole plane is to be described by coordinates and planar figures are to be grasped in their whole extension by equations. The simplest figures in the plane, lines, are then translated by the simplest equations, those of the first degree, called linear because of their relation to straight lines; circles and other conics are fitted by second degree equations. I think that both in phenomenological analysis and didactics too little emphasis is laid upon this fact:

the justification of the numerical operations and their laws by the simplicity of the algebraic description of geometrical figures and relations.

Briefly stated:

Algebra is valid because it functions in geometry.

It is strange that so far this insight has not, if at all, strongly enough been pronounced. In history it has never been used against people who argued against negative numbers. In teaching algebra it should be our duty to convince the learner of the validity of the operations and their properties so forcefully that he cannot but accept them. The most convincing argument is to show him the operationality of algebra in geometry. This, I believe, should be our policy in teaching negative numbers.

Here "geometry" does not mean an axiomatic structure but what is visually obvious or conceptually follows from what is visually obvious — a visuality that neither requires involved explanations in the vernacular nor sophisticated elaboration. The one-dimensional medium, the straight line, has not enough visual structure; two dimensions is the minimum that is required, and with a view to the graphic possibilities the most appropriate medium. **

Let us repeat history in a modified way: turning

the algebraic permanence principle

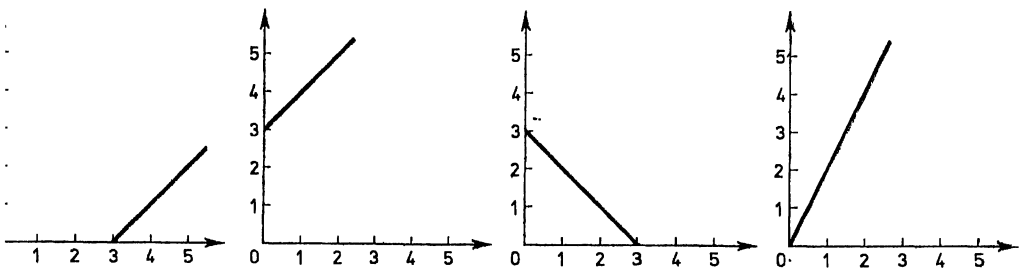
into

the geometrical algebraic permanence principle,

now applied not to extend solving formulae but to extend such functions as

$$x \to x - 3,$$
$$x \to x + 3,$$
$$x \to 3 - x,$$
$$x \to 2x,$$

which only imperfectly reflect geometric figures.



I need not explain to you in detail how negative numbers, their operations and the laws governing them arise didactically in this context.

And perhaps you will also grasp what I meant when at the start I claimed that teaching the young can teach us historical lessons.

Let us now turn to another subject.

Variables

* For centuries "variable" meant — in mathematics and elsewhere — something that really varies, something in the

physical, social, mental

as well as in the

mathematical

world that is

perceived, imagined, supposed

as varying, that is in addition to

the time that passes,
the path that is covered,
the aim that changes,
the water that is rising,
the temperature that oscillates,
the wind that is changing,
the days that lengthen,
the mortality rate that decreases,
the progressive rate of income tax,

also the

variable mathematical objects

by which these phenomena are described. From the

variable physical, social, mental

phenomena one is led to

variable numbers, magnitudes, points, sets,

in general

variable mathematical objects.

Locutions like

the number ε *approaches (converges to)* 0,
the point P *runs on the surface* S,
the element x *runs through the set* S,
the number e *is approached by the sequence* $(1+1/n)^n$ *if* n *goes to infinity,*

witness this kinematic aspect of the "variable". It is true that in the course of, say, the past half century such locutions have been outlawed by purists. Indeed one can dispense with them,

x_n *converges to* 0

can be written as

$$\lim_n x_n = 0$$

and be defined, with no kinematics involved, by

for every $\varepsilon > 0$ *there is an* n_0 *such that* $|x_n| < \varepsilon$ *for* $n \geqslant n_0$;
x *runs through the set* S

can very simply be written as

$$x \in S.$$

Well, one can dispense with that kind of kinematics provided one has once possessed it, learned to use it and then to eliminate it — this is a general didactical feature. **

As soon as names were needed for mathematical variables they were indicated by letters. But at that time letters had already been in use in mathematics for about two millennia, in geometry to indicate arbitrary points, in geometrical algebra for arbitrary magnitudes, and in number theory for arbitrary whole numbers, as witnessed by Euclid's Elements. I did not say *variable* points, magnitudes, numbers — I did say *arbitrary* ones, and this is a fundamental difference. One letter meant one individual point, one magnitude, one number, though it did not matter, or was taken as unknown, which one. Letters were used in mathematics as polyvalent names.

Polyvalent names are a well-known feature in the vernacular, too. Proper names such as "Warsaw" for a particular city, and "Poland" for a particular country, are rare — we cannot afford too many. We cannot afford a proper name for each particular mouse, or table, or stone, so we use the same for each of them, and in each particular case by a specific way of binding indicate which one is meant: this mouse, or that table or the first stone of the Academy building. "I" is such a polyvalent name, bound by the mere fact of pronouncing it to the person that says it. "Here" and "there" are polyvalent names which can be bound to places, "thing" is one that may apply to anything whatsoever.

Mathematics has proper names, even an infinity of them, — I mean the vocabulary of natural numbers, constructed in an algorithmic way. Compared with the rich variety of polyvalent names in the vernacular, the stock available in mathematics looks poor: the letters of the alphabet or of a few alphabets, sometimes enriched by subscripts, accents, dots and dashes. Unlike the polyvalent names in the vernacular they are not restricted to particular species of things such as mice, tables, stones. They are general purpose polyvalent names, which can mean everything mathematical. Accordingly they were used to formulate general laws like

$$(a+b)^2 = a^2+2ab+b^2$$

or to ask for a solution as in

$$x^2+x-2=0;$$

such polyvalent names were called indeterminates or unknowns. For a long time it was a much discussed topic whether literal algebra instruction should start with the one rather than the other use of the letters.

What is your reaction to this story? Shrugging, incredulity, astonishment, or a half-smile? Or are you sorry about the lost paradise of the good old time? Today, it is all variable, indeed, as you know, and nothing else. But as a historian you may ask who brought this change about, and when it happened. You may even ask when people became accustomed to it and when they started teaching it this way. I confess I have not investigated this. I do not even know when I myself got acquainted with the extended use of the word "variable" and when I myself started using it.

Anyway it is clear that it started in formal logic. When logicians looked into the status of the "letters", the fact that for such a letter you may substitute whatever you like might have suggested the term "variable", but with the reservation that it was a mere metaphor, because there was nothing in it that really varies. It was a highly suggestive metaphor; yet while usurping the term "variable", logicians even went one step further in hollowing it out: a variable was not even a name, let alone a polyvalent name, but a mere placeholder, that is, a hole to be filled by names, and only for opportunity's sake are different kinds of holes being distinguished by different symbols.

Even more serious things happened as logicians and purists usurped the term covering the genuine mathematical variable. Its original meaning shifted, got lost. Mathematical terminology was stripped of its kinematic undertones. Terminology like that mentioned earlier, as time that passes, numbers approaching a limit, points running on a surface, was exorcised.

* Lumping concepts of various origin together, using one name for things that stripped of their frills boil down to the same, is one of the important characteristics of our mathematical activity. Here we have met with such an historical occurrence:

polyvalent name

and

variable object

related with each other and confounded. **

You may like it or not but it is a fact in mathematics that more unity is a precondition of more profound understanding and continuing progress, and that reshaping mathematical language is a highly valuable activity. But as historians we are not asked what we like. We have to be conscious of the reasons why things were viewed differently in the past — in fact

they were good reasons — and of the reasons why things changed — these were equally good reasons. Moreover, as educators we have to find out at which point learners should start to recapitulate history. Finally, if we are both historians and educators, the one ought to learn from the other to understand his task more profoundly.

* Fortunately: *expellas naturam furca, tamen usque recurrit* — nature though driven out with the fork, nevertheless returns. The mathematical purism — of high value within mathematics — is a forced and less satisfactory language as soon as one steps out of mathematics. The abundance of variable objects in the half-way mathematized vernacular can be eliminated by linguistic sophistication but by this linguistic measure they are not disposed of. And — even more important — in order to be eliminated by linguistic tricks, they must once have been experienced by the learner. Indeed, there is no other way to guarantee that he will be able to restore them as he needs the vernacular to recognize and apply mathematics in the real world. The world is a realm of change, describing the world is describing change, and to do this one creates variable objects — physical, social, mental, and finally mathematical ones. There exist many languages of description, or rather many levels of describing. On a high level of formalization the variable mathematical objects may be forsaken, but on less formalized levels they are a genetically and didactically indispensable link with the physical, social, and mental variables, which on their part are indispensable tools. **

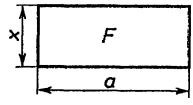
It is shocking that textbooks in the wake of New Math try to teach mathematics as though it were nothing but an impeccable language, which as it happens most of the students are unable to grasp and speak, and that researchers conduct subtle investigations to find out whether students understand variables as polyvalent names or as mere place-holders while it never crossed their mind that variables should and could be understood as variable objects. It is no less shocking that historians are not alarmed enough by the obvious failure of this kind of instruction to look more profoundly into history in order to find out what history can teach education.

Let me put this in an even broader context by my third and last example.

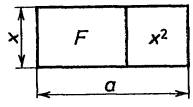
Inversion and conversion

Viewed historically, mathematics has grown not only in substance and subject matter. It is as much, and perhaps even more, a process of reshaping and remodelling, of turning things upside down and inside out.

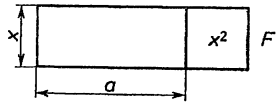
In Greek mathematics conics originated from the problem of solving quadratic equations, or, as the Greeks formulated it, the problem of applying an area F to a line segment a * either exactly, that is $ax = F$,



or so that a square falls short: $(a-x)x = F$,



or so that a square exceeds: $(a+x)x = F$.



These three cases are distinguished with the Greek words for agreement, falling short, and excess as

parabolic, elliptic, hyperbolic

application. This, then, is the origin of our terms for the conics.

If F is given as a square with side y, then the above equations become

$$\begin{aligned} ax &= y^2, \\ (a-x)x &= y^2, \\ (a+x)x &= y^2, \end{aligned}$$

respectively, which are indeed the equations of parabola, ellipse, hyperbola. ** This was the way these curves were first encountered, by solving quadratic equations, and only afterwards was it discovered that they represent the planar sections of a cone. It was Apollonius who inverted the course of history and started with conic sections to derive their "symptoms", that is equations, from the geometrical data.

After Descartes had developed what was called analytic geometry, the view was again turned round: quadratic equations, now in a more

general setting, became the source of conics, and their relation to cones, though easily proved by Dandelin's method, was played down, as a minor subject, at present probably unknown to many users of mathematics.

Or look to a side track, projective geometry: Pascal's theorem on the inscribed hexagon used as a defining property in Steiner's approach to conics by means of projectively related pencils. Or to projective invariance of the harmonic quadruple as a means of defining projectivity of mappings in von Staudt's approach.

The foregoing are examples of straight inversion of view. A more sophisticated example of refashioning, again of Greek mathematics — let us call it conversion — is the elimination of proportion and similarity arguments from elementary geometry by means of area transformation. There can be little doubt that the so-called Pythagorean theorem, known as early as 2000 B.C. in Babylonia, was first discovered and proved by similarity arguments, by which it is almost trivial. Though it can also easily be proved by congruence arguments, neither Euclid's proof nor the one that very likely preceded it in history, was that easy. They show that Euclid's predecessors had exerted themselves to avoid similarity arguments. The tool they invented was the transformation of areas, say of rectangles, that is, the replacement of the proportion

$$a : b = c : d$$

by the equality of areas

$$ad = bc.$$

In the case of the Pythagorean theorem, which deals with areas, this is not a far-fetched idea. But they extended similarity avoidance as a principle, which is pursued in the most terrifying way in the construction of the regular pentagon — important for the regular solids. The construction, if carried out by similarity and the golden section, which is a proportionality concept, is almost trivial. Euclid's way, by congruence arguments, is a masterpiece of contorted thought, and an appalling example of blocking understanding by dogmatism.

When did this proportions and similarity avoiding dogmatism come about? Was it some time before a satisfactory theory of proportion was developed? But why should it have been preserved afterwards? By mere tradition or because it was such a marvellous piece of mental gymnastics? Or was it craving for purity of method, "do not do with similarity what you can achieve by congruence"?

Unfortunately here Euclid's Elements stopped the pre-Euclidean process of reshaping and remodelling. This situation lasted two millennia. It needed a new undogmatic view to rescue the mathematics that was at a dead end, driven there by the rigorous Greek mental discipline. History always repeats itself. However, today traditions and dogmatisms enjoy a much shorter life than those that once blocked the development of mathematics for almost two millennia. Today no structure of and within mathematics is safe from inversion and conversion.

But the phenomena of inversion and conversion as a mathematical virtue are not restricted to what can be called the macro level. Individual inventors and inventions are permeated by their influence. No mathematical idea is published in the way it was discovered. Techniques have been developed and are used, if a problem has been solved, to turn the solution procedure upside down, or if it is a larger complex of statements and theorems, to turn definitions into propositions, and propositions into definitions, the hot invention into icy beauty. This is what happens on what I would call the meso level of the history of mathematics. But it extends to even smaller constituents of our mathematical activity, the micro level. As an example let us look at the definition of continuity. Intuitively: a small change of one variable causes a small change of another. By formalization this is inverted so that the ε precedes the δ: "for every $\varepsilon > 0$ there is a $\delta > 0$ such that...". This inversion is required by the difference in flavour between two uses of the word "small": small enough, and as small as you like, where the arbitrariness of the second conditions the sufficiency of the first. This is a paradigm of the micro inversion which takes place whenever we switch from one view to the other: in order to effect some E we have to adjust D to arrive at E.

Let us now turn to education. Years ago I coined the term "antididactical inversion" (see for instance *Mathematics as an Educational Task*, p. 122) and illustrated it by a number of examples. One of them was Peano's axioms, deriving complete induction from them, in order to apply this principle. The historical course was the inverse, and so should it be in didactics. Nobody can become conscious about complete induction before having unconsciously applied it, and nobody can formulate complete induction unless he has noticed it. Nobody can grasp Peano's axioms unless he can formulate complete induction. This is the didactical order and the historical order. Applying complete induction unconsciously, becoming conscious of it, formulating it, and building it into Peano's axioms. People who teach mathematics as a ready-made system prefer antididactical inversion.

Let us be satisfied with this one example. If mathematics teaching proves to be a failure, the reason is often, if not always, that we do not realize that young people have to start somewhere in the past of mankind and somehow repeat the learning process of mankind. This is the lesson historians and educators can learn from each other.

Acknowledgement

The present lecture contains pieces from my forthcoming book "Didactical Phenomenology of Mathematical Structures", where these ideas were developed in a broader context. This has been made possible by the courtesy of the Publisher D. Reidel, Dordrecht–Boston. To the pieces between an asterisk and the next double asterisk applies:

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

А. В. ПОГОРЕЛОВ

О преподавании геометрии в школе

В последние полвека специалисты-математики, педагоги и психологи оживленно обсуждают проблему математического образования вообще и школьного математического образования в частности. В ходе этого обсуждения высказывались часто совершенно противоположные точки зрения. Во многих странах была проведена реформа математического образования. Такая реформа была предпринята и в нашей стране (СССР) в конце 60-х годов. К сожалению, результат не оправдал ожиданий. В связи с этим специальная комиссия Академии наук СССР под председательством академика И. М. Виноградова, изучив положение дела с преподаванием математики в общеобразовательных школах страны, приняла новую программу, близкую к традиционной программе дореформенного периода. Комиссия предложила мне представить проект современного школьного учебника по геометрии на традиционной основе.

Такой учебник мной был представлен, комиссия одобрила его. В течение трех лет учебник проходил апробацию в массовой школе и в настоящее время Министерством Просвещения СССР введен практически во всех общеобразовательных школах страны. Этот учебник объемом около 300 стр. укомплектован необходимым количеством упражнений и рассчитан на пять лет обучения с 6 класса (возраст 12 лет) до 10 класса (возраст 17 лет).

В настоящем докладе я хотел бы изложить основные соображения, которыми я руководствовался, предлагая этот учебник, и показать, как основные вопросы, которые были предметом дискуссии среди математиков, решаются в учебнике.

1. Начало изложения

Прежде всего я обращаюсь просто к опыту (жизненному опыту) учащегося (ему уже 12 лет) и делаю акцент на некоторых хорошо известных ему свойствах простейших фигур. Эти свойства позже будут

названы аксиомами. Обращаясь к рисунку, я говорю: какая бы ни была прямая, существуют точки, лежащие на прямой, и точки, не принадлежащие ей. Далее я говорю, что через две точки можно провести, и притом только одну прямую. Все это хорошо известно учащемуся, так как ему, наверное, неоднократно приходилось проводить такую прямую в связи с чисто практическими задачами.

Обращаясь к рисунку, на котором изображена прямая с тремя точками на ней, я говорю: из трех точек на прямой одна и только одна лежит между двумя другими и указываю на эту точку. Свойство взаимного расположения предметов, которые мы выражаем словами „лежать между" хорошо известно учащемуся. Поэтому мое утверждение о расположении трех точек на прямой не вызывает вопросов или недоумений. Далее, обращаясь к рисунку, на котором проведена прямая, я говорю, что она разбивает плоскость на две полуплоскости. Это ясно, но я добавляю: отрезок, соединяющий точки одной полуплоскости, не пересекает прямую, а отрезок, соединяющий точки разных полуплоскостей, пересекает ее. Этим пояснением свойство разбиения плоскости на две полуплоскости приобретает точный смысл.

Акцент на отмеченные свойства усиливается примерами решения задач с использованием этих свойств. Например, дана прямая и три точки A, B, C, не лежащие на этой прямой. Известно, что отрезки AB и BC пересекают прямую. Пересекает ли прямую отрезок AC? Глядя на чертеж, учащийся сразу отвечает на вопрос. А учитель поясняет ответ, ссылаясь на отмеченное свойство разбиения плоскости на две полуплоскости. Это объяснение позже будет названо доказательством.

Далее я ввожу аксиомы меры для отрезков и углов. Я говорю: каждый отрезок имеет определенную длину и длина отрезка равна сумме длин частей, на которые он разбивается любой его точкой. Конечно, для учащегося это не ново. Трудно представить себе учащегося 12 лет, которому не приходилось бы измерять расстояние, пользуясь при этом аддитивностью меры.

Правда, мое утверждение о существовании длины отрезка вовсе не предполагает какого-либо измерения. Понятие длины относится в данном изложении к числу основных понятий. Но я об этом умалчиваю.

Аксиому меры для отрезков учащийся, начинающий изучать геометрию, и специалист-математик понимают по-разному. Однако можно показать, что они при этом имеют ввиду одно и то же число, ибо, постулировав только существование длины и ее аддитивность

при обычной процедуре измерения, которую представляет себе учащийся, мы получим тот же результат.

Аналогично вводится аксиома меры для углов. Затем, обращаясь к опыту учащегося, я утверждаю возможность отложить отрезок данной длины на данном луче из его начальной точки и угол с заданной градусной мерой. Многочисленные примеры упражнений закрепляют свойства меры для отрезков и углов. Соответствующие объяснения учителя с использованием свойств меры готовят учащегося к введению понятия теоремы и ее доказательства.

Введение меры для отрезков и углов естественно приводит к соответствующему определению равенства для отрезков и углов. Именно, отрезки называются равными, если они имеют одинаковую длину. Углы называются равными, если они имеют одинаковую градусную меру. Мы полагаем, что учащийся, приступающий к изучению геометрии, вряд ли понимает равенство отрезков и углов иначе, чем это сказано в данном определении.

После того, как определено понятие равенства для отрезков и углов, естественно вводится понятие равенства для треугольников. Именно, треугольники ABC и $A_1B_1C_1$ называются равными ($\triangle ABC = \triangle A_1B_1C_1$), если у них $\angle A = \angle A_1$, $\angle B = \angle B_1$, $\angle C = \angle C_1$, $AB = A_1B_1$, $AC = A_1C_1$, $BC = B_1C_1$.

Традиционный прием доказательств, связанный с перемещением треугольников в заданное расположение, в нашем изложении основан на аксиоме существования треугольника, равного данному. Обращаясь к рисунку, я говорю: пусть мы имеем треугольник ABC и луч a. Переместим треугольник ABC так, чтобы его вершина A попала в начало луча, вершина B — на луч, а вершина C — в заданную полуплоскость относительно луча a и его продолжения. Полученный при этом треугольник, обозначим его $A_1B_1C_1$, равен треугольнику ABC. Учащемуся это ясно. Объяснение заканчивается утверждением: для данного треугольника и данного луча существует равный ему треугольник в заданном расположении. Это утверждение и есть аксиома. Наглядное объяснение, предшествующее формулировке аксиомы, делает ее использование в традиционных доказательствах наглядным и простым.

Список аксиом заканчивается аксиомой параллельных. Эта аксиома, в отличие от остальных, не аргументируется никакими наглядными соображениями. Это вполне естественно.

После того как основные свойства (аксиомы) закреплены многочисленными упражнениями с объяснениями, опирающимися на основные свойства, вводится понятие теоремы и ее доказательства, кото-

рые таким образом получают совершенно определенный смысл. В качестве иллюстрации теоремы и ее доказательства может быть приведено решение любого упражнения с объяснением, опирающимся на основные свойства (аксиомы).

2. Строгость и доступность изложения

Понятия строгости, а тем более доступности изложения, являются весьма относительными. Я смею утверждать, что изложение в данном учебнике строгое с точки зрения специалиста-математика и просто с точки зрения учащегося, изучающего предмет. Приведу соответствующие аргументы. Начну с доступности изложения.

Прежде всего я исхожу из того, что традиционное изложение предмета в учебниках прошлых лет было безусловно просто и доступно учащимся. Это изложение создавалось и совершенствовалось веками, причем одним из основных соображений при этом было требование простоты и доступности. Изложение в моем учебнике отличается от традиционного только началом, которое, как мы видели, просто и не может создать каких-либо трудностей.

Начало изложения, вполне строгое, постепенно переходит в традиционное как по форме, так и по содержанию. А оно к тому времени уже вполне безупречно. Все это дает нам основание утверждать, что изложение предмета в данном учебнике просто и доступно учащимся.

Среди специалистов и педагогов распространено мнение о том, что последовательно дедуктивное изложение геометрии в школе не осуществимо. Изложение данного учебника убедительно опровергает это мнение. Но в этом нет ничего удивительного. Простота и доступность в нашем учебнике оказалась возможной, благодаря специальной системе аксиом, которая переносит существенные трудности изложения в другую область — теорию вещественных чисел.

Применительно к аксиоматике Эвклида–Гильберта последовательно дедуктивное изложение геометрии в школе, конечно, невозможно. Такое изложение встречает непреодолимые трудности для понимания учащихся уже в начале изложения, при введении меры для отрезков и углов. В нашем изложении эти трудности снимаются тем, что существование меры для отрезков и углов постулируется, а форма, в которой это преподносится учащемуся, опирается на наглядные представления об измерении.

Теперь о строгости изложения в учебнике. В этой связи поставим вопрос: является ли изложение элементарной геометрии в сочинении Гильберта „Основания геометрии" строгим? Если да, то я смею утверждать, что изложение в данном учебнике является также строгим. Оно опирается на полную не противоречивую систему аксиом (эквивалентную системе аксиом Гильберта), а доказательства проведены полно, без пропусков аргументов.

3. О современности учебника

Я смею утверждать, что предлагаемый учебник является вполне современным, не смотря на его традиционную основу. В этой связи прежде всего следует обратить внимание на аксиоматическое построение изложения, характерное для современных математических теорий. Хотя общие математические понятия, такие как понятия множества, отображения, бинарного отношения, группы и др. в учебнике не формируются, они вводятся и обстоятельно изучаются на конкретном материале учебника. Мы полагаем, куда важнее доказать групповые свойства параллельного переноса, чем вводить понятие транзитивной группы и без соответствующих доказательств привести параллельный перенос в качестве примера такой группы.

В учебнике излагается метод координат и элементы векторной алгебры, а также показывается эффективность этих методов при решении задач и доказательстве теорем. Но им не дается предпочтение. Мы полагаем, что в школьном изложении предмета основным должен быть синтетический метод. Именно этот метод способствует развитию логического мышления, интуиции и пространственных представлений. А это является главной задачей преподавания геометрии в школе. Геометрия своими методами проникает во все области современной математики, поэтому обеднять школьную геометрию, сводя ее к аналитическим выкладкам, представляется неправильным и не современным.

4. О преподавании геометрии по учебнику

Общепринято считать, что преподавание есть искусство. Чтобы постигнуть это искусство, необходимо (но не достаточно) иметь высокую профессиональную подготовку. В этой связи я подготовил курс геометрии для студентов педагогических вузов (объем около 300 стр.).

Особенностью этого курса является то, что он во всех своих частях прямо или косвенно обращен к элементарной геометрии. Он начинается школным изложением темы координаты и векторы, чем обеспечивается преемственность школьного изложения вузовскому. Курс состоит из четырех примерно равных по объему частей: аналитической геометрии, дифференциальной геометрии, оснований геометрии и некоторых разделов элементарной геометрии.

Оригинальной является третья часть курса. В отличие от традиционных курсов, где основные вопросы, связанные с аксиоматическим построением геометрии, решаются на основе аксиоматики Гильберта (или Вейля), в данном курсе все это изложение опирается на школьную аксиоматику. Таким образом, вопросы непротиворечивости, полноты и независимости аксиом решаются по отношению к аксиоматике, которая хорошо известна и исследование которой представляет безусловный профессиональный интерес.

По замыслу автора, два учебника для школы и вуза, о которых шла речь, должны обеспечивать разумное решение проблемы математического образования на данном этапе применительно к геометрии.

Proceedings of the International Congress of Mathematicians
August 16–24, 1983, Warszawa

JAMES SERRIN

The Structure and Laws of Thermodynamics

In the light of past experience as well as recent research, it has become clear that for the study of phenomenological thermodynamics — namely, the over arching non-statistical subject — the appropriate primitive variables are *work*, *heat*, and *hotness*. A physical system of whatever sort — for example, a body of gas or a viscous fluid, an elastic solid or a chemically reacting mixture — whose interactions with its exterior are reflected by various transfers of work and heat is thus called a *thermodynamical system*.

Notwithstanding the broad range of intended application, the study of thermodynamics has always been impeded by an inadequate presentation of its foundational aspects. Thus Kelvin could write "A mere quicksand has been given as a foundation for thermometry" and Cardwell could add only recently "The student is usually introduced to the concepts of thermodynamics ... in a way which does violence to credibility". While there have been occasional attempts to clarify the situation — a notable but not entirely successful one being due to Carathéodory (see [24]) — it has only been in the past several years that a concerted effort has been undertaken, involving the work of a number of mathematicians in various different centers. The goals of this research fall into four related categories:

I. To find an appropriate general *structure* in which to express the fundamental concepts of the subject.

II. To formulate the *laws of thermodynamics* in ways which are clear and concise, physically reasonable and useful.

III. To *prove* existence of the *mechanical equivalent of heat* and the *absolute temperature scale*, in analytically precise terms.

IV. To define the concepts of *internal energy* and *entropy* for material systems, and to characterize those systems for which these functions can be shown to exist.

The purpose of this paper is to outline some recent results which have been obtained in these directions. In particular, for the first three a clean and precise theory has emerged, while for the fourth there is important new understanding of the crucial issues and how to treat them.

It should finally be emphasized that the discussion is not directed toward specific problems for particular material systems, as important as these may be. Rather, I am concerned with providing a *general structure* within which *special* physical systems can appear as *special* cases. This point of view clearly reflects the beliefs of the founders of thermodynamics (however much they may have limited themselves in practice to the treatment of special systems) and moreover closely parallels modern approaches to continuum physics.

1. The formal structure presented here was first developed during the period 1977–1979 in papers of the author [13], [14], [15] and in 1978–1980 by M. Šilhavý [17], [18], [19], approximately at the same time but entirely independently. Šilhavý's development requires considerably deeper topological and measure theoretic considerations, and accordingly we follow the approach in [13], [14], [15]. A related approach to the foundations of thermodynamics is due to Feinberg and Lavine [7]. In their treatment the concept of hotness is not taken as fundamental, though as in Šilhavý's method fairly deep measure and function theoretic ideas are required.

It is convenient to begin with the basic concept of hotness, represented by a *thermal manifold* $\mathscr{H}$ consisting of the set of *hotness levels* L open to material systems. We assume that $\mathscr{H}$ is a totally ordered set, with order relation $\succ$. The sentence $L_2 \succ L_1$ will be read "L_2 is hotter than L_1".

A *temperature scale* is a strictly increasing map from $\mathscr{H}$ into the reals $\boldsymbol{R}$. If ψ is a temperature scale then $\psi(L)$ is called the *temperature* of L in the scale ψ.

Fundamental to thermodynamical structure is the concept of a *thermodynamical system*, examples of which might be a body of gas or an elastic solid, to name two particularly simple cases. Every thermodynamical system $\mathscr{S}$ comes endowed with a set $\boldsymbol{P}(\mathscr{S})$ of *processes* which the system may undergo, together with a subset $\boldsymbol{P}_{\mathrm{cyc}}(\mathscr{S})$ of *cyclic processes* of the system. To every process $P \in \boldsymbol{P}(\mathscr{S})$ there correspond real numbers $\overline{W}(P)$ and $\bar{Q}(P)$, respectively the *total work* done by the process P and the *total heat* used by the process P. Formally

$$\overline{W}\colon \boldsymbol{P}(\mathscr{S}) \to \boldsymbol{R}, \qquad \bar{Q}\colon \boldsymbol{P}(\mathscr{S}) \to \boldsymbol{R}.$$

We adopt the standard sign convention that $\overline{W}(P) > 0$ if work is done by the system on the exterior environment and $\overline{W}(P) < 0$ if the exterior

environment does work on the system. Similarly $\bar{Q}(P) > 0$ if heat is supplied to the system, while $\bar{Q}(P) < 0$ means that the system has supplied heat to the environment.

We require one more primitive concept, namely a more refined and subtle measure of the heat used by a process than is directly given by the total heat $\bar{Q}(P)$. The reason for this, of course, is that heat supplied at one temperature is very different than heat supplied at another. To this end, we suppose (in accord with intuition) that to every process $P \in \boldsymbol{P}(\mathscr{S})$ and every hotness level $L \in \mathscr{H}$ there is associated a real number $Q(P, L)$ representing the total or net heat transferred to the system during the process P *at hotness levels* L' *not exceeding* L. Formally we have

$$Q\colon \boldsymbol{P}(\mathscr{S}) \times \mathscr{H} \to \boldsymbol{R}.$$

The function $Q(P, \cdot)$ is called the *accumulation function of the process* P.

The accumulation function expresses analytically the essential properties of the relation between heat and hotness for a given process P. For example, during a process P the total heat added *between* the hotness levels L_1 and L_2 (with $L_1 \prec L_2$ say) is given by $Q(P, L_2) - Q(P, L_1)$. It follows in particular that the accumulation function of an *isothermla process* P (operating at a single hotness level L_0) is constant except for a single jump at L_0, the jump being positive if $\bar{Q}(P) > 0$ and negative if $\bar{Q}(P) < 0$. Similarly if the system only absorbs heat during a process P — but never emits heat — then $Q(P, \cdot)$ is monotonically increasing. In the same way, if P is *adiabatic* — that is, exchanges no heat whatsoever with its environment — then $Q(P, \cdot) \equiv 0$.

For any given process P of a system $\mathscr{S}$ one may suppose that heat is exchanged with the environment only on some bounded range of hotnesses. Reflecting this fact, we assume that the accumulation function has the following property.

(1) For every $P \in \boldsymbol{P}(\mathscr{S})$ there exists a *lower* hotness level, denoted by L_l, such that

$$Q(P, L) = 0 \quad \text{when } L \prec L_l$$

and an *upper* hotness level, denoted by L_u, such that

$$Q(P, L) = \bar{Q}(P) \quad \text{when } L \succ L_u.$$

In addition we suppose a minimal degree of regularity for the accumulation function, namely

(2) For every $P \in \boldsymbol{P}(\mathscr{S})$ the function $Q(P, \cdot)$ is bounded and has at most a denumerable number of discontinuities.

A final necessary concept is the idea of *products* of thermodynamical systems. The well-known heuristic arguments presented in standard treatments of thermodynamics to justify the classical efficiency theorem, arguments which ultimately go back to Carnot, involve comparing Carnot cycles for two different systems by forming a third (union) system for which the heat and work are found by adding the corresponding quantities for the original systems. In effect, the union idea involves taking the heat emitted by one body and transferring it to a second body, with a corresponding reduction of the heat supplied to the second system from its other surroundings. These well-known but nevertheless somewhat vague ideas require a formal description.

Let $\mathscr{S}_1$ and $\mathscr{S}_2$ be a pair of physical systems. The *product system*, $\mathscr{S}_1 \oplus \mathscr{S}_2$, is characterized by its processes and their work and heat functions, which are required to satisfy the following conditions:

(i) $\boldsymbol{P}(\mathscr{S}_1 \oplus \mathscr{S}_2) = \boldsymbol{P}(\mathscr{S}_1) \times \boldsymbol{P}(\mathscr{S}_2)$,

(ii) $\boldsymbol{P}_{\text{cyc}}(\mathscr{S}_1 \oplus \mathscr{S}_2) = \boldsymbol{P}_{\text{cyc}}(\mathscr{S}_1) \times \boldsymbol{P}_{\text{cyc}}(\mathscr{S}_2)$,

(iii) $\overline{W}(P_1 \oplus P_2) > 0$ if $\overline{W}(P_1) + \overline{W}(P_2) > 0$,

(iv) $\overline{Q}(P_1 \oplus P_2) < 0$ if $\overline{Q}(P_1) + \overline{Q}(P_2) < 0$,

(v) $Q(P_1 \oplus P_2, \cdot) \geqslant 0$ if $Q(P, \cdot) + Q(P, \cdot) \geqslant 0$.[1]

Here $P_1 \oplus P_2$ denotes the union process (in $\boldsymbol{P}(\mathscr{S}_1 \oplus \mathscr{S}_2)$) corresponding to the pair of processes $P_1 \in \boldsymbol{P}(\mathscr{S}_1)$, $P_2 \in \boldsymbol{P}(\mathscr{S}_2)$.

It is open to question whether the concept of a product system should be meaningful for all conceivable pairs of thermodynamical systems. To avoid such metaphysical points, we shall henceforth restrict the formation of product systems only to special and distinguished pairs of systems, which will be called *thermodynamically compatible systems* (or simply *compatible* systems). Thus if $\mathscr{S}_1$ and $\mathscr{S}_2$ are a pair of compatible systems, then the product system $\mathscr{S}_1 \oplus \mathscr{S}_2$ is itself assumed to be a meaningful thermodynamical system satisfying the laws of thermodynamics.

2. Since the mid nineteenth century, a first principal of thermodynamics has been the basic interconvertibility of work and heat. Stated without recourse to special assumptions regarding state spaces and internal energy, this principle asserts that there exists a universal constant $\mathscr{J} > 0$ such that $\overline{W}(P) = \mathscr{J}\overline{Q}(P)$ for any cyclic process P of any physical system.

---

[1] This is a weaker formulation of the union axiom than is usually stated. For the strong version of the axiom one requires that $\overline{W}(P_1 \oplus P_2) = \overline{W}(P_1) + \overline{W}(P_2)$, and $\overline{Q}(P_1 \oplus P_2) = \overline{Q}(P_1) + \overline{Q}(P_2)$.

A particular feature of this formulation which may strike one as unusual is the appearance of the universal constant $\mathscr{J}$. In developing the theory of absolute temperature, for example, the existence of this canonical scale is not postulated, but rather is *derived* from more basic laws. It would therefore seem more appropriate to state the first law *without reference to an absolute equivalent of work and heat*, and to demonstrate (within the theory) that such an absolute equivalent must exist. Once one turns in this direction, however, a number of alternatives present themselves, and it is not immediately clear which of these should be taken as the fundamental expression of the relation between work and heat.

Because it appears desirable to maintain the greatest generality, we shall present here a version of the first law which expresses only the most certain of our beliefs about heat and work, consistent with the requirement that we can develop from it a satisfying and general theory. If we give some thought to the gist of the first law, namely that work can only be produced at the expense of heat energy, we are led to the following formulation.

WEAK FIRST LAW. *If* $\overline{W}(P) > 0$ *for a cyclic process* P *of a thermodynamical system* $\mathscr{S}$, *then also* $\overline{Q}(P) > 0$.

This version of the first law has been noted by Šilhavý in his fundamental paper [17], though in his context some additional topological considerations appear which are extraneous to our purposes. Moreover Šilhavý does not emphasize this statement as an independent expression of the law (see [17], Part II, Theorem 2.2.1; [18], Part 1, Sections 4.6, 4.7; and [18], Part II, Section 4.12). Another interesting version of the first law, not involving a mechanical equivalent of heat, is due to Truesdell ([22], [23]).

The weak first law formalizes the idea that positive work can be obtained from a cyclically operating process only when a positive total amount of heat is supplied to the system during the process. While representing a generally weaker requirement than the strict interconvertibility of heat and work, it nevertheless carries great conviction and provides all the normal conclusions drawn from the stronger statement. Of course the weak first law (as stated above) is logically consistent with the strict interconvertibility of work and heat in the sense that if the latter is asserted to hold, then the weak first law is an obvious consequence.

To obtain strict interconvertibility, certainly a desideratum, we shall also consider a stronger version of the first law, again due to Šilhavý.

STRONG FIRST LAW. *For a cyclic process* P *of a thermodynamical system* $\mathscr{S}$, *the conditions* $\overline{W}(P) > 0$ *and* $\overline{Q}(P) > 0$ *are equivalent.*

The principal goal of elementary thermodynamics is to provide analytic tools for studying thermal systems. The following result, essentially due to Šilhavý (see [17], Part I, Sect. 2.1), is a consequence of the Weak First Law.

THE ENERGY INEQUALITY. *Let $\mathscr{U}$ be a collection of thermodynamic systems containing a perfect gas $\mathscr{G}$, and suppose each system $\mathscr{S}$ in $\mathscr{U}$ is compatible with $\mathscr{G}$. Then there exists a unique (universal) constant $\mathscr{J} > 0$ such that for every cyclic process P of every system $\mathscr{S}$ in $\mathscr{U}$ we have $\overline{W}(P) \leqslant \mathscr{J}\bar{Q}(P)$.*

The energy inequality, as an axiom, was first stated by R. L. Fosdick and the author [8]; for a demonstration of the present form, see [16]. The energy inequality of course applies only to systems $\mathscr{S}$ in the collection $\mathscr{U}$. Since we may assume, realistically, that any system $\mathscr{S}$ of interest belongs to such a collection it follows that the relation $\overline{W}(P) \leqslant \mathscr{J}\bar{Q}(P)$ can be presumed to hold for cyclic processes of arbitrary thermodynamic systems. The constant $\mathscr{J}$ is called the *mechanical equivalent of heat.*

When the *Strong First Law* is posited instead of the *Weak Law*, and the strong union axiom is assumed (see footnote, Section 3), a similar proof yields the conclusion

$$\overline{W}(P) = \mathscr{J}\bar{Q}(P)$$

for all cyclic processes P of systems $\mathscr{S}$ in the universe $\mathscr{U}$. That is, the Strong First Law is equivalent to the interconvertibility of heat and work for arbitrary cyclic processes. In what follows we shall assume the normalization $\mathscr{J} = 1$.

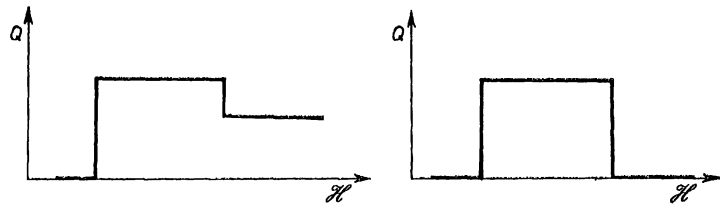
3. The second law of thermodynamics involves more subtle ideas than the first since it deals with the quality of heat at different hotness levels. Moreover, the physical notions which originally motivated the various nineteenth century statements of the second law are fairly obscure, requiring some effort to phrase clearly.

The first essentially correct formulation of the second law is due to Rudolf Clausius, namely

A passage of heat from a colder to a hotter body cannot take place without compensation [3].

While this is not at all precisely stated, we may consider it to mean that if a cyclic process absorbs heat at a hotness level L_0 and emits heat at a hotness level $L_1 \succ L_0$ then necessarily $\overline{W}(P) < 0$. That is, the accumulation function of a cyclic process P with $\overline{W}(P) \geqslant 0$ cannot have either of

the forms shown:



This version of the second law is not easily applicable to the case of general thermal processes without the further intervention of sophisticated topological notions. On the other hand, a slightly stronger formulation of essentially the same idea can be given which avoids this difficulty. In particular, since graphs of the above two types are disallowed by Clausius's version of the second law, it seems equally the case that no *linear combination* of such graphs could occur as the accumulation function of a cyclic process with $\overline{W}(P) \geqslant 0$. Indeed, such an accumulation function would represent a process which raises various low temperature heat supplies to various higher temperatures, without the need of doing work on the system. In the same way, any *closure* of such linear combinations would also appear impossible for cyclic processes (by continuity considerations) at least if $\overline{W}(P) > 0$. But the set of such closures coincides with the set of *non-negative* accumulation functions. We are thus led to the following general version of the second law (see Serrin [13], [14]).

SECOND LAW. *If* $\overline{W}(P) > 0$ *for a cyclic process* P *of a thermodynamical system* $\mathscr{S}$, *then there is some hotness level* L_0 *for which* $Q(P, L_0) < 0$.

The reader will surely notice the interesting relation between the (weak) First Law and the Second Law, namely that when $\overline{W}(P) > 0$ the former requires a *positive* value for $\bar{Q}(P)$ while the latter implies a *negative* value for $Q(P, \cdot)$ at some hotness level. This allows the two laws to be stated in a simple *combined form*, a form whose impressive elegance cannot help but be remarked.

COMBINED LAWS. *For any cyclic process* P *with* $\overline{W}(P) > 0$ *there holds*

$$\begin{aligned} Q(P, L) &> 0 \quad \text{for some } L \succ L_u, \\ Q(P, L) &< 0 \quad \text{for some } L \prec L_u. \end{aligned}$$

The Second Law is an *intrinsic* statement about the relation between work and heat in cyclic processes. In parallel with the discussion of the

First Law in the previous section, the Second Law also has an equivalent *analytical* formulation of great usefulness. We state this as follows (see [13], [14] and, from other standpoints, also [7], [17], [18], [19]).

THE ACCUMULATION INEQUALITY. *Let the hypotheses of the energy inequality hold. Then there exists an (absolute) temperature scale $\hat{T}$ on the hotness manifold $\mathscr{H}$, with $\hat{T}(\mathscr{H}) \equiv \mathbf{R}^+$, such that for every cyclic process P of every thermodynamic system $\mathscr{S}$ in $\mathscr{U}$ we have*

$$\int_0^\infty \frac{Q(P, L)}{T^2}\, dT \leqslant 0,$$

where $L = \hat{L}(T)$ is the hotness level associated with the temperature T in the scale $\hat{T}$. Any temperature scale $\hat{T}$ with the above property either agrees with the perfect gas scale of $\mathscr{G}$ or is a positive constant multiple of this scale.

The accumulation theorem immediately accomplishes two purposes: *it establishes the concept of absolute temperature without ambiguity, and it characterizes once and for all the allowable behavior of the accumulation function of any cyclic process.*

Indeed the accumulation theorem implies the second law, for if $Q(P, \cdot) \geqslant 0$ and $\overline{W}(P) > 0$ then by the first law $\overline{Q}(P) > 0$ so that the accumulation integral must be positive, and the process hence cannot by cyclic.

The reader should observe that the accumulation integral is well-defined and finite, as follows easily from properties (1) and (2) of the accumulation function given in Section 1.

The accumulation inequality is a generalization of the Clausius inequality. In particular, should $Q(P, \cdot)$ be of bounded variation then we can obviously write

$$\int_0^\infty \frac{Q(P, L)}{T^2}\, dT = \int_0^\infty \frac{dQ(P, L)}{T},$$

the latter integral representing the "sum" of the heats added divided by their absolute temperatures. The advantage of the accumulation integral compared to the Clausius integral is that it can be expressed analytically in terms of clearly formulated primitive concepts, and at the same time is applicable to a broader class of processes since its existence relies only on the structural properties (1) and (2) of the accumulation function.

For a proof of the accumulation inequality see [15] or, in a more general setting, [6]. In these papers it is also shown how to replace the perfect gas $\mathscr{G}$ by less special model materials — essentially those with a suitably rich supply of Carnot cycles.

4. In order to provide a concrete framework for the notions of internal energy and entropy it is necessary to introduce the idea of a state space, and an associated state structure. At the simplest (and most general) level this may be defined as follows.

A *state structure* for a system $\mathscr{S}$ consists of a set Σ, whose elements are called *states* of the system, and a corresponding family of processes $\boldsymbol{P}_\Sigma(\mathscr{S}) \subset \boldsymbol{P}(\mathscr{S})$, with each process $P \in \boldsymbol{P}_\Sigma(\mathscr{S})$ having a well-defined initial state $P_i \in \Sigma$ and final state $P_f \in \Sigma$.[2] Moreover if P is a cyclic process in $\boldsymbol{P}_\Sigma(\mathscr{S})$ then $P_i = P_f$. (In practice, a state structure should also be compatible with the notion of a process P *following* another process P', and should include in this case the axiom $P_i = P'_f$.)

We shall say that a system $\mathscr{S}$ has an *internal energy* corresponding to the state structure $(\Sigma, \boldsymbol{P}_\Sigma)$ if there exists a function $U\colon \Sigma \to \boldsymbol{R}$ such that $\Delta U \leqslant \bar{Q}(P) - \overline{W}(P)$ for each $P \in \boldsymbol{P}_\Sigma(\mathscr{S})$. Here ΔU denotes the difference between U evaluated at the final state and the initial state of P, that is $\Delta U \equiv U(P_f) - U(P_i)$. Roughly speaking, then, a function U is an internal energy for a system if it is a lower potential for the difference $\bar{Q}(P) - \overline{W}(P)$.

If $P \in \boldsymbol{P}_{\mathrm{cyc}}(\mathscr{S}) \cap \boldsymbol{P}_\Sigma(\mathscr{S})$ then necessarily $P_i = P_f$ and in turn $\Delta U = 0$. Consequently we recover the energy inequality $\overline{W}(P) \leqslant \mathscr{J}\bar{Q}(P)$ from the above formula (this in fact being the *motivation* for the definition of internal energy).

In parallel with the Strong First Law, we may also introduce the idea of a (strong) internal energy, in which the inequality $\Delta U \leqslant \bar{Q}(P) - \overline{W}(P)$ is replaced by the stronger requirement

$$\Delta U = \bar{Q}(P) - \overline{W}(P).$$

Turning to the second law it is natural to proceed in a similar way, but now based on the accumulation inequality. For convenience in formulation, we introduce the abbreviation $\bar{A}(P)$ for the integral appearing in the

---

[2] Formally, the assignment of initial and final states can be considered as a pair of mappings

$$i\colon \boldsymbol{P}_\Sigma(\mathscr{S}) \to \Sigma, \quad i(P) = P_i,$$
$$f\colon \boldsymbol{P}_\Sigma(\mathscr{S}) \to \Sigma, \quad f(P) = P_f.$$

accumulation theorem; thus

$$\bar{A}(P) \equiv \int_0^\infty \frac{Q(P, L)}{T^2}\, dT;$$

naturally, once one has a (definite) absolute temperature scale in hand one can define $\bar{A}(P)$ whether or not the process P is cyclic. This being understood, we shall say that a system $\mathscr{S}$ has an *entropy* corresponding to the state structure $(\Sigma, \boldsymbol{P}_\Sigma)$ if there exists a function $S\colon \Sigma \to \boldsymbol{R}$ such that

$$\Delta S \geqslant \bar{A}(P)$$

for each $P \in \boldsymbol{P}_\Sigma(\mathscr{S})$. Again roughly speaking, the entropy is an upper potential for the accumulation integral $\bar{A}(P)$.

If $P \in \boldsymbol{P}_{\mathrm{cyc}}(\mathscr{S}) \cap \boldsymbol{P}_\Sigma(\mathscr{S})$ then of course $\Delta S = 0$ so that we recover the cyclic condition $\bar{A}(P) \leqslant 0$ stated in the accumulation theorem. Another case of interest is that of an adiabatic process, for which $Q(P, \cdot) \equiv 0$. In this situation one has $\bar{A}(P) = 0$ whence in turn the entropy hypothesis yields

$$\Delta S \geqslant 0,$$

the celebrated condition of spontaneous entropy increase.

It is one of the principal conclusions of classical thermodynamics that simple reversible systems necessarily possess both an internal energy and an entropy. This result can be obtained within the present structure as a direct and simple consequence of the energy inequality (Weak First Law) and the accumulation inequality (Second Law), cf. [14], [16]. Thus for simple reversible systems the existence of internal energy and entropy is equivalent to the First and Second Laws.

A principal problem of modern thermodynamics is to characterize those systems for which the same conclusion holds (Goal IV in the introduction). Important work in this direction was initiated by Coleman and Owen [4] and has been continued in a series of more recent papers [5], [6], [9], [13], [20], etc.

In his dissertation Ricou [11] has proved the remarkable result that if a state structure is deterministic for a given system (that is, if the condition $P'_t = P_f$ implies that the process P' can follow P) then the system must have an internal energy and an entropy (see also [12]).

5. Phenomenological thermodynamics studies and interrelates two basic physical quantities, *heat* and *work*. There are two intrinsic principles

governing this interrelation — the first and second laws of thermodynamics. Each states a reasonable, even if somewhat pessimistic, conviction about the physical world. There are, next, analytical formulations of these laws — first, the energy inequality (or the interconvertibility of heat and work for cycles, if the Strong First Law is used), and second, the accumulation inequality. These analytical formulations pave the way to all direct applications of the theory. They rely in turn on two derived *scale* concepts — the Joule mechanical equivalent of heat and the Kelvin absolute temperature scale, each among the greatest conceptions of nineteenth century physics. Finally, there are two fundamental potentials, or more accurately semi-potentials — the internal energy and the entropy — which extend the direct cyclic principles to much broader classes of thermal processes. On this structure hangs the science of heat, from the elementary theory of reversible systems, to Gibbs' magnificent conception of thermal and chemical equilibrium, to sophisticated theories of material dynamics.

The most far-reaching implication of the structure, however, is the fact that it is *not* limited to equilibrium. In fact, a dogmatism which would lay this restriction on thermodynamics would in turn invalidate a massive sector of thermal physics, including heat transfer theory, compressible fluid mechanics, shock wave theory, and combustion theory. Conversely, allowing thermodynamics a natural scope beyond equilibrium yields a powerful and far-reaching theory with which to attack those *dynamical* problems where hotness and heat play a crucial role.

References

[1] Bridgman P. W., *The Nature of Thermodynamics*, Harvard University Press, 1941.

[2] Cardwell D. S. L., *From Watt to Clausius*, Cornell University Press, 1971.

[3] Clausius R., *Abhandlungen über die mechanische Warmetheorie*, Braunschweig 1864, 1967. Translated by W. R. Browne as *The Mechanical Theory of Heat*, London, 1879 (see especially pages 76–79).

[4] Coleman B. and Owen D., A Mathematical Foundation for Thermodynamics, *Archive for Rational Mechanics and Analysis* **54** (1974), pp. 1–104.

[5] Coleman B. and Owen D., On the Thermodynamics of Semi-Systems With Restrictions on the Accessibility of States, *Archive for Rational Mechanics and Analysis* **66** (1977), pp. 173–181.

[6] Coleman B., Owen D., and Serrin J., The Second Law of Thermodynamics for Systems with Approximate Cycles, *Archive for Rational Mechanics and Analysis* **77** (1981), pp. 103–142.

[7] Feinberg M. and Lavine R., Thermodynamics Based on the Hahn–Banach Theorem; the Clausius Inequality, *Archive for Rational Mechanics and Analysis* **82** (1983), pp. 203–293.

[8] Fosdick R. L., and Serrin J., Global Properties of Continuum Thermodynamical Processes, *Archive for Rational Mechanics and Analysis* **59** (1975), pp. 108–109.
[9] Owen D., The Second Law of Thermodynamics for Semi-Systems with Few Approximate Cycles, *Archive for Rational Mechanics and Analysis* **80** (1982), pp. 39–55.
[10] Pitteri M., Classical Thermodynamics of Homogeneous Systems Based upon Carnot's General Axiom, *Archive for Rational Mechanics and Analysis* **80** (1982), pp. 333–385.
[11] Ricou M., *Energy, Entropy and the Laws of Thermodynamics*, Thesis, University of Minnesota, 1983.
[12] Ricou M. and Serrin J., *A General Second Law of Thermodynamics* (to appear).
[13] Serrin J., The Concepts of Thermodynamics, *Contemporary Developments in Continuum Mechanics and Partial Differential Equations*, pp. 411–451. North-Holland, 1978. (Proceedings of a conference held in Rio de Janeiro, August, 1977.)
[14] Serrin J., Conceptual Analysis of the Classical Second Laws of Thermodynamics, *Archive for Rational Mechanics and Analysis* **70** (1979), pp. 355–371.
[15] Serrin J., *Lectures on Thermodynamics*, University of Naples, 1979.
[16] Serrin J., An Outline of Thermodynamical Structure. In: *New Perspectives in Thermodynamics*, Springer-Verlag (to appear).
[17] Šilhavý M., On Measures, Convex Cones, and Foundations of Thermodynamics. I. Systems with Vector-Valued Actions; II. Thermodynamic Systems, *Czech. J. Phys.* **B30** (1980), pp. 841–861, 961–991.
[18] Šilhavý M., On the Second Law of Thermodynamics for Cyclic Processes. I. General Framework; II. Inequalities for Cyclic Processes, *Czech. J. Physics* **B32** (1982), pp. 987–1010, 1073–1099.
[19] Šilhavý M., On the Clausius Inequality, *Archive for Rational Mechanics and Analysis* **81** (1983), pp. 221–243. (Work originally presented at Euromech III Symposium, September, 1978.)
[20] Šilhavý M., The Foundations of Thermodynamics. In: *New Perspectives in Thermodynamics*, Springer-Verlag (to appear).
[21] Thompson J. W. (Lord Kelvin), *Encyclopedia Brittanica*, 9th edition, 1878.
[22] Truesdell C., How to Understand and Teach the Logical Structure and History of Classical Thermodynamics, *Proceedings International Congress of Mathematicians*, Vancouver, 1974, pp. 577–586.
[23] Truesdell C. and Bharatha S., *The Concepts and Logic of Classical Thermodynamics*, Springer-Verlag, 1977.
[24] Truesdell C., What Did Gibbs and Carathéodory Leave Us in Thermodynamics? In: *New Perspectives in Thermodynamics*, Springer-Verlag (to appear).

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF MINNESOTA
MINNEAPOLIS, MN. 55455
U.S.A.

Index