JOURNAL DE MATHÉMATIQUES PURES ET APPLIQUÉES

Neuvième série publiée par H. VILLAT.

Cette revue internationale parait trimestriellement. Publie des travaux de mathématiciens de tous pays, en français et en anglais, des mémoires originaux sur toutes les branches des mathématiques.

Le JOURNAL DE MATHEMATIQUES PURES ET APPLIQUEES qui se trouve être le plus ancien des journaux scientifiques du monde est indispensable aux savants qui dominent et créent la mathématique moderne.

Dans ses pages ont paru d'innombrables découvertes sur des sujets essentiels.

C'est la revue des chercheurs qui doivent demeurer au courant de la science qui s'élabore. 1 Tome par an.

Tarif d'abonnement 1971 :

— France : 165 F — Etranger : 185 F

ANNALES SCIENTIFIQUES DE L'ÉCOLE NORMALE SUPÉRIEURE

Revue internationale, parait trimestriellement, publie les travaux les plus récents des mathématiciens de tous les pays.

Tarif d'abonnement 1971 :

— France : 150 F — Etranger : 200 F

GAUTHIER-VILLARS, Editeur, PARIS

GAUTHIER-VILLARS, Editeur, PARIS

TARIF 1971

PERIODIQUES SCIENTIFIQUES

ABONNEMENTS ET ANNEES ANTERIEURES

Tous nos abonnements sont payables d'avance et partent de janvier

Titre et périodicité des revues	Tarif France	Etranger	
		F.F.	
Annales d'Histochimie (4 N^{os})	90	110	
Annales de l'Ecole Normale Supérieure (4 N^{os})...........	150	200	
Annales de Physique Biologique et Médicale (4 N^{os})	80	96	
Annales de l'Institut Henri Poincaré (2 séries)	174	195	
Série A (Physique théorique) (2 tomes de 4 N^{os})	116	130	
Série B (Calcul des Probab. et Statistiq.) (1 tome de 4 N^{os}).	58	65	
Bulletin de la Société Mathématique de France			
(4 N^{os} + 4 suppléments)	150	150	
Bulletin des Sciences Mathématiques (4 N^{os})	120	145	
Comptes rendus de l'Académie des Sciences : Hebdomadaire de 2 tomes par an			
1er Fascicule. Séries AB : Sciences Math. et Phys.	470	765	
2^{e} Fascicule. Série C : Sciences Chimiques	405	675	
3^{e} Fascicule. Série D : Sciences Naturelles	450	755	
Les trois Fascicules	980	1 655	
Journal de Mathématiques pures et appliquées (4 N^{os})	165	185	
Journal de Mécanique (4 N^{os})	110	130	
Mathématiques et Sciences Humaines (4 N^{os})	35	45	
Œcologia Plantarum (4 N^{os})	80	96	
Physiologie Végétale (4 N^{os})	95	110	
Revue d'Ecologie et de Biologie du Sol (4 N^{os})	80	96	
Revue de Chimie Minérale (6 N^{os})	175	200	

BULLETIN D'ABONNEMENT 1971

à retourner aux Editions GAUTHIER-VILLARS — 55, quai des Grands-Augustins, Paris 6^{e}

Veuillez m'inscrire pour un abonnement d'un an à ______________________

Je vous adresse le montant de cette commande soit

France ______________________ Etranger ______________

Par chèque : □ bancaire ou □ postal (C C P Paris 29 323) — Par mandat*

** Rayer la mention inutile*

NOM ______________________ Adresse ______________

ATTENTION !

Cet abonnement peut être pris en charge par le Laboratoire, l'Organisme ou la Société dont vous dépendez. Ceux-ci peuvent imputer la dépense à leur budget "Documentation scientifique". Dans ce cas, c'est bien volontiers que nous ferons parvenir la facture en plusieurs exemplaires.

ACTES
DU
CONGRÈS INTERNATIONAL DES MATHÉMATICIENS

1970

ACTES
DU
CONGRÈS INTERNATIONAL DES MATHÉMATICIENS

1970

publiés sous la direction du
Comité d'Organisation du Congrès

3

Mathématiques appliquées (E)
Histoire et Enseignement (F)

GAUTHIER-VILLARS ÉDITEUR
55, quai des Grands-Augustins, Paris 6^{e}
1971

E

MATHÉMATIQUES APPLIQUÉES

(Tome 3 : pages 1 à 330)

Actes, Congrès intern. Math., 1970. Tome 3, p. 3 à 8.

E1 - ASPECTS MATHÉMATIQUES DE LA THÉORIE QUANTIQUE DES CHAMPS

QUANTUM FIELD THEORY MODELS

by James GLIMM

Quantum field theory uses singular mathematical operations, and new mathematical developments will be needed to formulate it rigorously. In order to make progress in understanding the correct mathematical basis for renormalization in quantum field theory, we have restricted ourselves to the world of two (or occasionally three) dimensional space time. Our main result is an existence theorem. In two space time dimensions, quantum fields exist. The fields satisfy a nonlinear equation, for example

$$(1) \qquad \varphi_{tt}(x\,,t) - \varphi_{xx}(x\,,t) + m^2\varphi(x\,,t) + \varphi^{2n-1}(x\,,t) = 0,$$

and in this equation, it is the nonlinear term, φ^{2n-1}, which produces the interaction between the particles and the nontriviality of the theory. The field itself, $\varphi(x\,,t)$, is a bilinear form, densely defined on a Hilbert space $\mathfrak{H}$. A classical field is a real valued function which is a solution of the same equation, and the distinction between the quantum field and the classical field is that the quantum field satisfies the commutation relations

$$(2) \qquad [\varphi(x\,,t), \varphi_t(y\,,t)] = i\delta(x\text{-}y).$$

After averaging over x and t,

$$(3) \qquad \varphi(f) = \int \varphi(x\,,t) f(x\,,t)\,dxdt, \qquad f \in C_0^\infty$$

is a self adjoint operator, and so the field φ is also called an operator valued distribution. From this point of view, the distinction between classical and quantum fields is that in the case of quantum fields, the operators are not commutative.

Our existence theorem applies to two nonlinear interactions : the φ^{2n} interaction as in (1) and the Yukawa interaction. In four dimensions these same interactions are typical of those used to describe the strong interactions of mesons, neutrons and protons. I mention in passing that in (1) we may replace the nonlinear term by

$$(4) \qquad P'(\varphi(x\,,t)),$$

(1) Supported in part by the U. S. Air Force office of Scientific Research Contract AF-49(638)-1 719.

where P is a nonnegative polynomial. The powers, $\varphi^{2n-1}(x,t)$, of the bilinear form φ do not exist, and what occurs in the equation (1) is the renormalized power, or Wick power. These Wick powers are explicitly defined in many standard sources and cause no complications. In the two dimensional Yukawa theory more serious renormalizations occur, as we shall see, and in higher dimensions the renormalizations become progressively more difficult and are the main obstacle to a satisfactory mathematical theory of quantum fields.

THEOREM 1. – *For the two dimensional $P(\varphi)$ and Yukawa interactions, the quantum fields exist as densely defined bilinear forms and as operator valued distributions. The fields satisfy the correct nonlinear field equations and the canonical commutation relations. For the $P(\varphi)_2$ interaction, all of the Haag-Kastler axioms and many of the Wightman axioms are verified.*

A number of people have contributed to the proof of this theorem (see [7]) and I mention in particular my collaborator, A. Jaffe, and also E. Nelson, J. Cannon and L. Rosen.

The proof of this theorem is based on the Hamiltonian method. We introduce a Hamiltonian or energy operator H, and write an explicit formula for the time evolution in the fields,

$$\varphi(x,t) = e^{itH}\varphi(x,0)e^{-itH} \tag{5}$$

It is a simple calculation, using the canonical commutation relations (2) and the definition of H to check that (5) is a solution of the nonlinear field equation. Thus the main mathematical problem, as far as the existence of the fields is concerned, is to show that H is a self adjoint operator. (As a technical digression, I mention that the H to which we are referring contains a space cutoff. Because of the finite propagation speed, i.e. the hyperbolic character of the field equation (1), space cutoff or locally correct $H's$ can be used in (5) to define fields which satisfy (1), as required).

The operator H, of course, is specified by the physics of the problem. For the $P(\varphi)_2$ interaction, H is given in the form

$$H = A + B \tag{6}$$

where A and B are noncommuting self adjoint operators, and it is a well defined mathematical problem to show that H is essentially self adjoint on the domain

$$\mathcal{D}(A) \cap \mathcal{D}(B) \tag{7}$$

THEOREM 2. – [2, 10, 11] *For the $P(\varphi)_2$ model, $H = A + B$ is essentially self adjoint on the domain* (7).

As one might guess, the proof uses a mixture of estimates and functional analysis. The functional analysis centers around the Trotter product formula, or the Feynmann-Kac formula, or the Hilbert resolvent identity. The most fundamental of the estimates is the positivity of the energy,

$$-\text{const.} \leqslant A + B, \tag{8}$$

due to Nelson [9]. To explain this inequality, I will mention that

$$A = H_0 = \text{free energy}$$

is given explicitly and is obviously positive, while

$$B = H_1 = \text{interaction energy}$$

is an integral over x of the interaction energy density

$$H_1(x) = P(\varphi(x, 0)).$$

Since P is a positive polynomial, one expects $H_1(x)$ to be positive. However at this point the Wick ordering in the definition of $P(\varphi)$ interferes. The main point in the proof of (8) is to show that the Wick ordering does not interfere strongly, and that $H_1(x)$ is "nearly positive" in some sense. Thus for the $P(\varphi)_2$ interaction, we might characterize H in (6) by saying that it is a nearly positive but large perturbation of a positive self adjoint operator.

I will now turn to the Yukawa$_2$ interaction. In this case H is not an operator, but a formal expression of the form

$$H = A + B + C \tag{9}$$

where $A = H_0$ is the free energy operator as before and $B = H_1$ is the interaction energy. Now B is not an operator but rather a densely defined bilinear form, and C is the sum of two infinite terms,

$$C = c_e D_e + c_m D_m$$

where

$$c_e = \left(\int_{-\infty}^{\infty} (p^2 + 1)^{-1/2} dp\right)^2 = \infty$$

$$c_m = \int_{-\infty}^{\infty} (p^2 + 1)^{-1/2} dp = \infty,$$

D_e is a (finite) multiple of the identity operator and $D_m = D_m^*$ is a specific self adjoint operator. C is one of the famous infinite counterterms of quantum field theory. From a mathematical point of view, its role is to cancel other infinites which enter the theory because of singularities in B. (Recall, for example, that B is a bilinear form and not an operator). From the point of view of physics, the role of C is to shift the spectrum of H, so that H will be positive and so that the elementary particles described by H will have specified masses. (In technical terms, C is the sum of a vacuum energy renormalization $c_e D_e$ and a mass renormalization $c_m D_m$).

Of course the sum $A + B + C$ has no direct mathematical meaning, and in order to reason with complete mathematical rigor, we perform the following three steps.

Step 1. – Introduce approximations, so that all infinities become finite and all operators become self adjoint. We consider

$$H_j = A + B_j + \left(\int_{-j}^{j} (p^2 + 1)^{-1/2} dp\right)^2 D_e + \int_{-j}^{j} (p^2 + 1)^{-1/2} D_m$$
$$= A + B_j + C_j.$$

Here the subscript j on the operator B refers to a cutoff in the permitted range of integration for an integral defining B, in analogy with the definition of C_j. By an application of the Kato perturbation theory and some simple estimates, one shows that

$$H_j = H_j^*$$

STEP 2. – Perform cancellations between large finite quantities and estimate the remainders. We perform this step in the resolvent.

THEOREM 3. – *Let* $R_j(\zeta) = (H_j - \zeta)^{-1}$. *Then*

$$\lim_{j,k \to \infty} \|R_j(\zeta) - R_k(\zeta)\| = 0$$

Again the positivity of the energy is a basic tool,

$$0 \leqslant H_j + \text{const.},$$

where the constant must be independent of j. It is worth mentioning that the cancellation of the large finite quantities makes essential use of ideas coming from conventional physics and in particular from formal perturbation theory. This should not be surprising, because formal perturbation theory tells us not only that the infinities should occur, but also how they should be cancelled. We may also perform the cancellations on a dense domain for H, by constructing a dense set of vectors θ such that

$$(10) \qquad \theta = \lim \theta_j,\ \theta_j \in \mathcal{D}(H_j)$$

and such that the limit

$$(11) \qquad \lim H_j \theta_j$$

exists.

STEP 3. – Remove the approximations and obtain H as a limit. It is an easy matter to combine operator theory with the estimates of step 2 and thereby prove

THEOREM 4. – [5] *The limit*

$$R(\zeta) = \lim_{j \to \infty} R_j(\zeta)$$

is the resolvent of a self adjoint operator H, and in (11) we have

$$\lim_{j \to \infty} H_j \theta_j = H\theta$$

These three steps are essentially the same as the steps used to cancel other infinities in mathematics, for example in the expressions

$$df/dx = \lim_{\Delta x \to 0} \Delta f / \Delta x \qquad \text{and} \qquad P\int_{-1}^{1} x^{-1} dx = 0.$$

Beyond the basic question of existence, we want to obtain detailed properties of our quantum fields. For the space cutoff theory we have extensive knowledge of the spectrum of H.

THEOREM 5. – [3, 8, 10, 11] *Let H be a space cutoff Hamiltonian for the $P(\varphi)_2$ theory. Then H is a positive operator and H has finite dimensional spectrum in any spectral interval $[0, m - \epsilon]$, $\epsilon > 0$. The continuous spectrum for H has suport equal to the interval $[m, \infty]$. Zero is a simple eigenvector for H.*

The transition from the lower semibounded Hamiltonian in (8) to the positive Hamiltonian in Theorem 5 is obtained by adding a finite constant to H (a finite vacuum energy renormalization). The eigenvector Ω for the eigenvalue zero is the vacuum (of the space cutoff theory). In order to remove the space cutoff in the vacuum and the Hamiltonian, difficult estimates were required to show that the constant in (8) has a dependence on the space cutoff volume which is at most linear. This and other estimates are combined with the GNS construction from the theory of C^*-algebras to obtain a Hamiltonian H which does not depend on any space cutoff [4].

THEOREM 6. – [4] *In the $P(\varphi)_2$ theory, let H be the Hamiltonian, with the space cutoff removed. Then H is a positive operator and zero is an eigenvalue for H.*

We see that in two space time dimensions, quantum fields with nontrivial interactions have a rigorous mathematical existence. These fields satisfy many of the properties required by physics. The method of construction is consistent with conventional ideas in physics and the remaining axioms are all valid on the level of formal perturbation theory. At this point the major open problem for these models is to obtain further properties of the solutions, such as the existence of scattering states and a discrete mass spectrum. Possibly the main conclusion of this work is that the infinite cancellations and renormalizations in quantum field theory, first discovered in formal perturbation theory, are seen to have an existence in the rigorous solutions of our models.

REFERENCES

[1] CANNON J. and JAFFE A. — Lorentz covariance of the $((\varphi^4)_2$ quantum field theory, *Commun. Math. Phys.*, 17, 1970, p. 261-321.

[2] GLIMM J. and JAFFE A. — A (φ^4) quantum field theory without cutoffs, I. *Phys. Rev.* 176, 1968, p. 1945-1951.

[3] GLIMM J. and JAFFE A. — The $(\varphi^4)_2$ quantum field theory without cutoffs: II. The field operators and the approximate vacuum, *Ann. Math.* 91, 1970, p. 362-401.

[4] GLIMM J. and JAFFE A. — The $(\varphi^4)_2$ quantum field theory without cutoffs. III, The physical vacuum, *Acta Math.*, 125, 1970, p. 203-267.

[5] GLIMM J. and JAFFE A. — Self-adjointness of the Yukawa$_2$ Hamiltonian, *Ann. of Phys.*, 60, 1970, p. 321-383.

[6] GLIMM J. and JAFFE A. — The Yukawa$_2$ quantum field theory without cutoffs, *Jour. Funct. Analysis*, to appear.

[7] Glimm J. änd Jaffe A. — Quantum Field Theory Models, in :· *Les Houches Summer School Lectures,* 1970, ed. by C. De Witt and R. Stora, Gordon and Breach, New York.

[8] HØegh-Krohn R. — *On the spectrum of the space cutoff :* P(φ) : *Hamiltonian in two space-time dimensions,* to appear.

[9] Nelson E. — *A quartic interaction in two dimensions, in Mathematical theory of elementary particles,* ed. by R. Goodman and I. Segal, M.I.T. Press, Cambridge, 1966.

[10] Rosen L. — A φ^{2n} field theory without cutoffs, *Commun. Math. Phys.* 16, 1970, p. 157-183.

[11] Rosen L. — The $(\varphi^{2n})_2$ quantum field theory : higher order estimates, *Comm. Pure Appl. Math.,* to appear.

[12] Rosen L. — *The* $(\varphi^{2n})_2$ *quantum field theory : Lorentz covariance,* to appear.

[13] Segal I. — Notes toward the construction of nonlinear relativistic quantum fields I : The Hamiltonian in two space-time dimensions as the generator of a C*-automorphism group, *P.N.A.S.* 57, 1967, p. 1178-1183.

Courant Institute of Mathematical Sciences
New-York University
251 Mercer Street,
New-York, N.Y. 10012 (USA)

Actes, Congrès intern. Math., 1970. Tome 3, p. 9 à 14.

FEYNMAN INTEGRALS

by Tullio REGGE

The appearance in 1946 of the celebrated papers of Feynman on quantum electrodynamics (QED) marked the beginning of a long stream of papers attempting to elucidate all the implications of this new formalism. Feynman was extremely successful in that he succeeded in bringing together a radical simplification in the computational techniques of field theory and a startling pictorial and physical interpretation of those techniques. Today we know that the Feynman-Dyson expansion for field theory is probably inadequate for anything but QED. It is not amenable to a rigorous mathematical treatment and there are reasons to believe that it represents at best an asymptotic expansion.

This argument has cooled off somewhat the initial enthusiasm for a detailed investigation of Feynman relativistic amplitudes (FRA). It is often argued that, even if we knew in every detail the analytic properties of the FRA, this information would be useless unless a substantial progress could be made on the question of convergence. I think that this attitude is not warranted. If we adopt the original point of view of Heisenberg of the autonomous role of the S-matrix, I do not think that FRA are less relevant than in a conventional field theory. This because there are reasons to believe that the local analytic behaviour of S-matrix elements is closely related to the behaviour of suitable FRA. How this can be achieved systematically we do not know but certainly examples abound in the literature, some of them of great intrinsic value, as discussed at length by Chew and collaborators.

In this talk I cannot claim that I am anywhere closer to this final goal of a purely analytical definition of the S-matrix than anybody else. Rather, I wish to treat FRA as a testing ground for some conjectures which could be crucial later in achieving the final goal. Also I would like to point out which properties of the FRA I consider worth looking for.

Finally, I shall discuss the role of the Hilbert-Riemann problem in connection with the construction of an S-matrix.

I assume that everybody here already knows the definition of FRA. The class of integrals of physical interest is rather narrow for the purpose of mathematical investigations and it is better to widen it in two directions. FRA in physics use the form

$$A(\mathcal{D}, k, z) = \lim_{\epsilon \to 0} \int dp^{4\nu} \prod_i \frac{1}{p_i^2 - z_i - i\epsilon}$$

Here $\frac{1}{p_i^2 - z_i - i\epsilon}$ is the propagator associated with the i-th internal line, $z_i = m_i^2$

the associated (square) mass. ν is the number of loops in the Feynman diagram $\mathcal{D}$. Clearly p_l is a 4-vector in Minkowsky space. The k,z dependence in A are the set of external 4-momenta and the set of internal masses respectively.

I shall generalize the above definition in two ways :

(A) Following Speer I replace $\frac{1}{p_l^2 - z_l - i\epsilon}$ by $\frac{1}{(p_l^2 - z_l - i\epsilon)^{\lambda_l}}$ where λ_l is a complex parameter. In this way by giving λ_l a sufficiently large real part the integral for A will converge. Speer then teaches us how to renormalize a divergent integral for this generalized FRA.

(B) We put no restriction to the dimension d of the Minkowski space. It is to be noted that in many nonrelativistic applications also $d = 3$ is relevant. It is even possible to have d complex through a suitable definition of the FRA. The generalization A) removes some unpleasant degeneracy occurring in the physical case. This case can be easily retrieved by a limiting process. It is also much easier to check the role of each individual internal line in the FRA through the corresponding z_l and λ_l. So far I think that it is technically advisable to consider as complex variables all the variables appearing in the FRA but the λ's treated as parameters.

Landau first realized that FRA are singular on algebraic varieties (Landau varieties) which can be constructed out of very simple rules with an interesting geometrical and physical interpretation (Coleman and Norton). The original Landau singularities do not exhaust the singular set, there are second kind singularities whose physical interpretation is controversial. I think (following Chew) that they will play, in an S-matrix theory, a completely different role from the standard Landau singularity. Anyway a given FRA is defined on a noncompact algebraic variety X-W where X,W is a pair of a algebraic variety. X is spanned by k and z, W is the singular set (including Landau's set). A can be proved to have W as a branching locus. Strictly speaking, A is then a function on $\widetilde{X\text{-}W}$ the universal covering space and not on X-W. A is a member of the so-called Nilsson class of multivalued functions on X-W.

More precisely :

1) A is polynomially bounded in the distance from the singular set. W is not an essential singularity.

2) Given an open ball U, not intersecting W, there are k independent functions $A_1 \ldots A_k$, k depending on $\mathcal{D}$ and on λ, such that every analytic continuation A_g of A is of the form

$$A_g = \sum_{i=1}^{k} c_{gi} A_i$$

where c_{gi} are constants.

It is well known that A_g depends on the loop used to carry the continuation and only on the homotopy class of this loop in the fundamental group $\pi_1(X\text{-}W, U)$. Therefore we label A_g the result of the continuation of A along $g \in \pi_1(X\text{-}W, U)$. Clearly then $A_{i,g}$ is the corresponding continuation of A_i along g.

We have then for $1 \leqslant j \leqslant k$, $h \in \pi_1$:

$$A_{j,g} = \sum_{i=1}^{k} c_{i,j}(g) A_i$$

and $c_{ij}(g)$ identifies a $K \times K$ matrix $\mathscr{L}(g)$ in Hom (V_k). The map $\mathscr{L}(\pi_1) \xrightarrow{\mathscr{L}}$ Hom (V_k) is then a group representation. $\mathscr{L}(\pi_1)$ is called the monodromy group. We call R the ring generated by π_1, B the matrix ring generated by $\mathscr{L}(\pi_1)$. Clearly the map $\mathscr{L}$ can be extended to R and $\mathscr{L}(R) = B$.

It is possible to construct π_1, R, B, $\mathscr{L}(\pi_1)$ explicitly for classes of FRA. The interested reader is advised to look into the references for technical details.

Some points are however worth discussing here. Among these is the role of the so-called local conditions. A set of proposals for a careful definition of these conditions is contained in (1). Here I can only quote a few examples just to get the feeling : These examples are relevant to Landau singularities only, second kind singularities exhibit a much more complicated behaviour.

Example 1 — Let a FRA be given in terms of local coordinates $z_1 \ldots z_n$. Let $z_1 = 0$ be the local equation of a Landau singularity. Then near $z_1 = 0$ we have the decomposition (see ref. 30)

$$A = z_1^{\mu} R^1 + R$$

where μ depends linearly on λ_l. The net effect of an analytic continuation g around $z_1 = 0$ is to change A into A_g

$$A_g = e^{2i\pi\mu} z_1^{\mu} R^1 + R$$

The corresponding $\mathscr{L}(g)$ then satisfies the equation, setting $\mathscr{L}(g) - 1 = a$:

$$a^2 = Aa$$

where
$$A = e^{2i\pi\lambda} - 1.$$

Therefore $\mathscr{L}((g-1)^2 - A(g-1)) = 0$. This we refer to as a local condition.

Example 2 — Let now $z_1 = 0$, $z_2 = 0$ be two Landau varieties crossing transversally. Let as in Ex. 1 g go around z_1 once and h around z_2. Then the theory of π_1 developed by Van Kampen (see refs. 2 and 3) tells us that $gh = hg$ but there are also Cutkowsky-Steinman relations to the effect that

$$\mathscr{L}((g-1)(h-1)) = \mathscr{L}((h-1)(g-1)) = 0$$

which we also name local conditions. Other local conditions hold for tacnodes and cuspidal singularities.

The point here is to consider the two-sided ideal $J \subset R$ generated by the elements $(g-1)(h-1)$, $(h-1)(g-1)$, $(g-1)^2 - A(g-1)$, etc. For all the FRA examined so far in refs. (2, 3) we have

(I) $B \xrightarrow{i} R/J$ where i is an isomorphism.

In refs. (1, 2, 3) this is a standard conjecture for all FRA. Its significance and practical value are remarkable. It means that π_1 is all the global information we need in order to glue together the local analytic properties of A near singularities.

Other systematic properties.

I may venture to list here other conjectures which emerged as a result of the examples discussed so far.

(II) $\mathcal{R}(\pi_1)$ is pseudounitary for real $\lambda_l \neq$ integer. Westwater has proposed a general scheme for proving this conjecture.

(III) B is a complete matrix ring.

Consider then the Landau variety $z_l = 0$ where z_l is an internal mass of $\mathcal{D}$ as in Example 1. Consider $R^1|_{z_l=0}$ and $R|_{z_l=0}$. It can be proved that R^1 is essentially the FRA for the diagram $\mathcal{D}'$ obtained from $\mathcal{D}$ by removal of the line i. Similarly, R is the FRA where the line i contains a massless particle. Moreover $\pi_1(\mathcal{D}')$ is a subgroup of $\pi_1(\mathcal{D})$ and $aB(\mathcal{D})a$ is isomorphic to $B(\mathcal{D})$. Finally $(a - A)B(\mathcal{D})(a - A)$ is isomorphic to $B(\mathcal{D})$ (massless case).

Another obvious relation holds for quotient diagrams. Let A be a subdiagram of $\mathcal{D}$. $\mathcal{D}/A$ is obtained from $\mathcal{D}$ by identifying all the points of A. By the Landau rules $\mathcal{D}$ has all the singularities of $\mathcal{D}/A$ and possibly more. Therefore $\pi_1(\mathcal{D}/A)$ is a factor group of $\pi_1(\mathcal{D})$.

All these relations, although they have not been exploited systematically, point out to some regularity and naturality in the dependence of π_1 and related groups and rings on the topology of $\mathcal{D}$. They make it very plausible the possibility of recursive construction of these sets for any FRA.

So far we have focussed our attention on π_1, B, $\mathcal{R}(\pi_1)$, R without really having a motivation for doing so. It has been known for a long time that the search of a function with given X, W, π_1, $\mathcal{R}(\pi_1)$, the celebrated Hilbert-Riemann problem, has always a solution. In fact there is a whole class of solutions of the form

$$(HR) \qquad A(z) = \sum_{\alpha=1}^{k} R_a(z) \, . \, A^a(z)$$

where R_a are arbitrary rational functions on X, $A^a(z)$ are k independent solutions of the Hilbert-Riemann problem.

This is not surprising. FRA with spin can be expressed in terms of a finite number of invariant FRA all of which share the same singular set W and have the same π_1, $\mathcal{R}(\pi_1)$, X. The class HR therefore accomodates all these amplitudes.

In order to have a unique solution, further conditions have to be imposed.

A natural condition is that A should not have singularities (even polar ones) outside the set W. This excludes rational functions $R_a(z)$ singular outside W. But more important is the specification of the parameter μ in Ex. 1, and not just of $e^{2i\mu\pi}$ as required in $\mathcal{R}(\pi_1)$. Similar specifications have to be given for all singularities.

In this case the final solution is

$$A = \sum_{\alpha=1}^{k} R_a A^a(z)$$

where now R_a are constants and $A^a(z)$ are k independent solutions. It is therefore enough to specify the value of A at k preselected sites to determine it. The only obstacle is that there is no singular guarantee that such a solution exists. There are partial results by Lappo-Danilievsky to the effect that solutions to this restricted problem can be constructed for sufficiently small exponents μ. This matter really needs to be investigated further with much more powerful methods than the traditional ones, involving only one complex variable, used by Plemelj, Garnier, Schlesinger, Lappo Danilevsky, and others.

Final comments.

If the restricted problem is solvable, then π_1, $\mathcal{R}(\pi_1)$ and local exponents μ determine the FRA up to k constants. If this result holds true or, to be more modest, can be formulated for infinitely many variables and infinite k, as is the case with the S-matrix, then there is a chance of having a purely analytic definition for the S-matrix and of implementing the Heisenberg-Chew program. There is no need to stress the difficulty of such a task.

Finally, it is only a matter of justice and of fairness to acknowledge here the crucial contributions of the pioneering work of the French group of Pham, Fotiadi, Lascoux, Froissart, who have done the first job of using systematically the powerful tool of algebraic topology to the investigation of FRA.

BIBLIOGRAPHY

[1] Ponzano G., Regge T., Speer E.R. and Westwater M.J. — *Commun. Math. Phys.*, 15, 1969, p. 83-132.

[2] Regge T. — The Fundamental Group of Poincaré and the Analytic Properties of Feynman Relativistic Amplitudes, *Nobel Symposium 8; Elementary Particle Theory,* ed. Nils Svartholm, New York, Interscience, 1969.

[3] Ponzano G. and Regge T. — *The Monodromy Group of One Loop Relativistic Feynman Integrals,* Published in a volume on the occasion of the 60th birthday of the Academician N.N. Bogoliubov.

[4] Ponzano G., Regge T., Speer E.R. and Westwater M.J. — The Monodromy Rings of One-Loop Feynman Amplitudes, to be published in *Commun. Math. Phys.*

[5] Regge T., Speer E.R. and Westwater M.J. — The Monodromy Rings of the Necklace Graphs, to appear.

[6] Cutkosky R.E. — *J. Math. Phys.* 1, 1960, p. 429-433.

[7] Polkinghorne J.C. and Screaton G.R. — *Nuovo Cimento* 15, 1960, p. 925.

[8] Landau L.D. — *Nucl. Phys.* 13, 1959, p. 181-192.

[9] Berge C. — *The Theory of Graphs,* London, Methuen & Co., Ltd, 1962.

[10] Speer E.R. — *Generalized Feynman Amplitudes,* Princeton University Press, 1969.

[11] Regge T. — Algebraic Topology Methods in the Theory of Feynman Relativistic Amplitudes, *Batelle Rencontres 1967 Lectures in Mathematics and Physics,* ed. C.M. DeWitt, J.A. Wheeler. New York, W.A. Benjamin, 1968.

[12] Ahlfors L.V. — *Complex Analysis* (lst edition), New York, McGraw-Hill, 1953.

[13] Ince E.L. — *Ordinary Differential Equations,* New York, Dover, 1956.

[14] Hepp K. — *Commun. Math. Phys.,* 2, 1966, p. 301.

[15] Gel'fand I.M. and Shilov G.E. — *Generalized Functions,* Vol. I., New York, Academic Press, 1964.

[16] Fotiadi D., Froissart M., Lascoux J. and Pham F. — *Topology* 4, 1965, p. 159-191.

[17] Kinoshita T. — *J. Math. Phys.,* 3, 1962, p. 650-677.

[18] Van der Waerden B.L. — *Modern Algebra,* Vol. II. New York, Frederic Ungar Publishing Co., 1948.

[19] Risk C. — *J. Math. Phys.* 9, 1968, p. 2168-2172.

[20] Sard A. — *Bull. Amer. Math. Soc.* 48, 1942, p. 883-890.

[21] Eden R.J., Landshoff P.V., Olive D.I. and Polkinghorne J.C. — *The Analytic S-Matrix,* Cambridge University Press, 1966.

[22] Westwater M.J. — *Helv. Phys. Acta* 40, 1967, p. 389-400.

[23] Polkinghorne J.C. — *Analytic Properties in Perturbation Theory,* Lectures in Theoretical Physics, Brandeis Summer Institute, 1961, Vol. 1. New York, W.A. Benjamin, 1961.

[24] Okun L.B. and Rudik A.P. — *Nucl. Phys.* 14, 1960, p. 261-288.

[25] Hironaka H. — *Ann. Math.* 79, 1964, p. 109-326.

[26] Orlik P. and Wagreich P. — *Isolated Singularities of Algebraic Surfaces with C* Action,* Institute for Advanced Study preprint, 1970.

[27] Pham F. — *Introduction à l'Etude Topologique des Singularités de Landau,* Paris, Gauthier-Villars, 1967.

[28] Boyling J.B. — *Nuovo Cimento* 53, 1968, p. 351-375.

[29] Regge T. — *Talk delivered at the Accademia dei Lincei during the Mendeleyev meeting on Symmetry and Periodicity in Elementary Particles,* September, 1969.

[30] Speer E.R. and Westwater M.J. — *Generic Feynman Amplitudes,* (to appear in *Comm. Math. Phys.*).

Institute for Advanced Study
School of Natural Sciences
Princeton,
New Jersey 08 540 (USA)

Actes, Congrès intern. Math., 1970. Tome 3, p. 15 à 19.

ÉTATS D'ÉQUILIBRE DES SYSTÈMES INFINIS EN MÉCANIQUE STATISTIQUE

par David RUELLE

Au cours de ces dernières années, la mécanique statistique a été l'objet d'un sérieux effort de compréhension mathématique. Cet effort a conduit à la découverte de structures simples et remarquables que je voudrais décrire rapidement. Je me limiterai à un sujet : les états d'équilibre des systèmes infinis, et j'envisagerai un type de système particulièrement simple : les gaz sur un réseau[1].

1. Etats d'un gaz sur un réseau.

On utilise l'espace $\mathbf{Z}^\nu$ des ν-tuples d'entiers ($\nu > 0$) pour décrire un réseau (cristallin). Chaque $x \in \mathbf{Z}^\nu$ peut être occupé par zéro (0) ou une (1) particule. Dans ce modèle discrétisé d'un gaz, les *configurations* sont les fonctions X : $\mathbf{Z}^\nu \rightarrow \{0,1\}$, c'est-à-dire les éléments de

$$\mathfrak{X} = \{0,1\}^{\mathbf{Z}^\nu}$$

On confondra dans ce qui suit une configuration X et le sous-ensemble de $\mathbf{Z}^\nu$ dont X est fonction caractéristique, c'est-à-dire que l'on identifiera $\mathfrak{X}$ à $\mathfrak{P}(\mathbf{Z}^\nu)$.

On prend sur $\{0, 1\}$ la topologie discrète donc $\mathfrak{X}$, comme produit, est compact. Les *états* du système sont les mesures de probabilité sur $\mathfrak{X}$ et forment un ensemble convexe compact E (pour la topologie vague).

2. Interactions. Energie.

On appellera *interactions* les fonctions réelles Φ *définies sur les parties finies de* $\mathbf{Z}^\nu$ et telles que

$$\Phi(\emptyset) = 0 \tag{11}$$

$$\|\Phi\|_x = \sum_{X \ni x} |\Phi(X)| < \infty \qquad \text{pour tout } x \tag{12}$$

Pour X fini $\subset \mathbf{Z}^\nu$ on définit une *énergie* U par

$$U(X) = U_\Phi(X) = \sum_{Y \subset X} \Phi(Y)$$

3. Etats de Gibbs. Limite thermodynamique.

L'idée fondamentale de la mécanique statistique classique de l'équilibre est d'associer à une fonction énergie U des mesures de probabilité ρ^Λ pour les systèmes "finis" Λ, puis d'étudier la limite $\Lambda \rightarrow \infty$.

[1] Parmi les auteurs des résultats décrits ci-dessous, citons entre autres Dobrushin, Lanford et Robinson. Voir [6] ch. 6 et 7, [1], [2] et [5].

Soit Λ fini $\subset \mathbf{Z}^\nu$. On appelle *état de Gibbs* la mesure de probabilité sur $\{0,1\}^\Lambda$ (identifié à $\mathfrak{P}(\Lambda)$) définie par

$$\rho^\Lambda(X) = Z^{-1}\, e^{-U_\Phi(X)} \quad \text{pour } X \subset \Lambda$$

$$Z = \sum_{X:X\subset\Lambda} e^{-U_\Phi(X)}$$

Si $\Delta \subset \Lambda$, on notera ρ^Λ_Δ la projection de ρ^Λ sur $\mathfrak{P}(\Delta)$:

$$\rho^\Lambda_\Delta(\{X\}) = \sum_{Y\subset\Lambda\setminus\Delta} \rho^\Lambda(\{X\cup Y\}) \qquad \text{pour } X\subset\Delta$$

3.1. *Soit* (Λ_n) *une suite de sous-ensembles finis de* $\mathbf{Z}^\nu$ *tendant vers* $\infty(\Delta \subset \Lambda_n$ pour tout Δ fini et n assez grand). *On peut extraire de* (Λ_n) *une suite partielle* (Λ'_n) *telle que pour tout* Δ *fini, la limite suivante existe*

$$\lim_{n\to\infty} \rho^{\Lambda'_n}_\Delta = \sigma_\Delta$$

Il existe alors une mesure de probabilité (unique) σ sur $\mathfrak{X} = \mathfrak{P}(\mathbf{Z}^\nu)$ *dont la projection sur* $\mathfrak{P}(\Delta)$ *soit* σ_Δ *pour tout* Δ *fini.*

On appelle $\sigma \in E$ une *limite thermodynamique d'états de Gibbs.* Nous voulons interpréter ces limites comme états d'équilibre d'un système infini, mais il est avantageux d'introduire une définition plus générale.

4. Etats d'équilibre.

Pour $X, Y \subset \mathbf{Z}^\nu$, X fini, nous définissons

$$W(X,Y) = \sum_{U:U\subset X\cup Y} \Phi(U)$$

où la somme est étendue aux U finis tels que $U\cap X \neq \emptyset$ et $U \cap Y \neq \emptyset$. Il suit de (I 2) que $W(X,.)$ est une fonction continue sur $\mathfrak{X}$. Si X, Y sont finis et disjoints on a

$$U(X\cup Y) = U(X) + U(Y) + W(X,Y)$$

Pour Δ fini $\subset \mathbf{Z}^\nu$, $X\subset\Delta$ et $Y \subset \mathbf{Z}^\nu\setminus\Delta$, soit

$$f_{\Delta,Y}(X) = Z_Y^{-1}\, e^{-U(X)-W(X,Y)}$$

avec

$$Z_Y = \sum_{X\subset\Delta} e^{-U(X)-W(X,Y)}$$

Remarquons que l'état de Gibbs est donné par $\rho^\Lambda(\{X\}) = f_{\Lambda,\emptyset}(X)$.

Suivant Dobrushin [1] nous dirons que $\sigma\in E$ est un *état d'équilibre* si pour tout Δ il existe une mesure de probabilité $\tilde\sigma_\Delta$ sur $\mathfrak{P}(\mathbf{Z}^\nu\setminus\Delta)$ telle que

$$(*) \qquad \tilde\sigma_\Delta(\{X\}) = \int f_{\Delta,Y}(X)\,\sigma_\Delta(dY)$$

Cela revient à dire que $f_{\Delta,Y}(X)$ peut s'interpréter comme probabilité conditionnelle de trouver la configuration X dans Δ si Y est réalisée dans $\mathbf{Z}^\nu\setminus\Delta$.

4.1. *Toute limite thermodynamique d'états de Gibbs est un état d'équilibre.*

4.2. *L'ensemble K_Φ des états d'équilibre est convexe compact, c'est un simplexe de Choquet.*

Un état d'équilibre peut donc être représenté de manière unique comme résultante d'états d'équilibre extrémaux.

Soit $\mathcal{A} = \mathcal{C}(\mathcal{X})$ la C^*-algèbre des fonctions complexes continues sur $\mathcal{X} = \mathcal{P}(\mathbf{Z}^\nu)$. Pour Λ fini $\subset \mathbf{Z}^\nu$ on définit $\mathcal{A}_\Lambda$ comme la sous-algèbre des fonctions $X \to A(X)$ qui ne dépendent que de $X \cap \Lambda$ ($\mathcal{A}_\Lambda$ est isomorphe à $\mathcal{C}(\mathcal{P}(\Lambda))$. On a alors la caractérisation suivante :

4.3. *Un état d'équilibre σ est extrémal si et seulement si la condition (de Cluster) suivante est satisfaite.*

(C) *Pour tout $A \in \mathcal{A}$ il existe Δ fini $\subset \mathbf{Z}^\nu$ tel que*

$$(B \in \mathcal{A}_\Lambda \quad \text{et} \quad \Lambda \cap \Delta = \emptyset) \Rightarrow \|\sigma(AB) - \sigma(A)\,\sigma(B)\| < \|B\|$$

4.4. *Remarque.* La construction de Gel'fand-Naïmark-Segal donne une représentation π de $\mathcal{A}$ dans les opérateurs bornés sur $L^2(\sigma)$. Soit

$$\mathcal{B} = \cap_\Delta \; [U_{\Lambda:\Lambda\cap\Delta=\varphi} \, \pi(\mathcal{A}_\Lambda)]^-$$

où $^-$ désigne la fermeture faible. La décomposition d'un état d'équilibre en états d'équilibre extrémaux correspond à la diagonalisation de l'*algèbre à l'infini* $\mathcal{B}$. L'état σ est extrémal si et seulement si $\mathcal{B}$ est triviale.

5. Unicité ou non-unicité des états d'équilibre.

On peut dans certains cas démontrer l'unicité de l'état d'équilibre (on y arrive en transformant les équations d'équilibre (*) en une équation intégrale linéaire non homogène à noyau borné dans un espace de Banach adéquat). Dans d'autres cas on sait qu'il y a plusieurs états d'équilibre distincts (voir [2]).

6. Rôle de l'invariance par translation.

Nous n'avons jusqu'ici fait aucun usage de la structure additive de $\mathbf{Z}^\nu$. On pourrait donc reformuler ce qui précède (et ce serait plus naturel) en remplaçant $\mathbf{Z}^\nu$ par un ensemble dénombrable "abstrait" N. Nous allons maintenant considérer des interactions invariantes par translations, c'est-à-dire satisfaisant

$$(13) \qquad \Phi(X + x) = \Phi(X) \quad \text{si } X \text{ est fini} \subset \mathbf{Z}^\nu \quad \text{et} \quad x \in \mathbf{Z}^\nu$$

Ces interactions forment un espace de Banach $\mathcal{B}$ pour la norme

$$\|\Phi\| = \sum_{X \ni 0} |\Phi(X)|$$

L'introduction de l'invariance par translation conduit à des développements complètement nouveaux de la théorie, permettant en particulier de caractériser les états d'équilibre invariants par un principe variationnel.

7. Entropie.

Les translations de $\mathbf{Z}^\nu$ définissent des homéomorphismes $\tau_x : X \to X + x$ de $\mathcal{X}$. Soit I l'ensemble des états invariants par ces homéomorphismes. L'ensemble

$I \subset E$ est convexe compact, et c'est un simplexe de Choquet. Si $\sigma \in I$, $(\mathfrak{X}, \sigma, \tau)$ est un système dynamique abstrait.

Soit σ_Λ une mesure de probabilité sur $\mathfrak{X}(\Lambda)$; on lui associe une entropie

$$S_\Lambda = - \sum_{X \subset \Lambda} \sigma_\Lambda(\{X\}) \log \sigma_\Lambda(\{X\})$$

Soit aussi $N(\Lambda) = \text{card } \Lambda$.

7.1. *Si σ_Λ désigne la projection sur $\mathfrak{X}(\Lambda)$ d'un état $\sigma \in I$, alors la limite suivante existe*

$$s(\sigma) = \lim N(\Lambda)^{-1} S_\Lambda$$

quand Λ tend vers l'infini dans un sens convenable[1]. *La fonction $s(.)$ est affine semi-continue supérieurement sur I.*

En fait pour $\nu = 1$, $s(\sigma)$ n'est autre que l'invariant de Kolmogorov-Sinaï du système dynamique $(\mathfrak{X}, \sigma, \tau)$.

8. Pression.

8.1. *Soit $\Phi \in \mathcal{B}$, alors la limite suivante existe*

$$P(\Phi) = \lim N(\Lambda)^{-1} \log Z = \lim N(\Lambda)^{-1} \log \sum_{X : X \subset \Lambda} e^{-U_\Phi(X)}$$

quand Λ tend vers l'infini dans un sens convenable[2]. *La fonction $P(.)$ est continue* (en fait $|P(\Phi) - P(\Psi)| \leqslant \|\Phi - \Psi\|$) *et convexe*[3].

La quantité P est (essentiellement) la pression thermodynamique.

9. Principe variationnel.

A toute interaction $\Phi \in \mathcal{B}$ nous associons $A_\Phi \in \mathcal{C}(\mathfrak{X})$; A_Φ est une "variable aléatoire" ou "observable" donnant la contribution d'un point (0) du réseau à l'énergie U_Φ. Nous posons

$$A_\Phi(X) = \sum_{Y : 0 \in Y \subset X} \frac{\Phi(Y)}{N(Y)}$$

9.1. *Si $\Phi \in \mathcal{B}$ on a l'identité*

$$(**) \qquad P(\Phi) = \max_{\sigma \in I} [s(\sigma) - \sigma(A_\Phi)]$$

[1] Par exemple Λ est un parallélipipède $\{x \in \mathbf{Z}^\nu : 0 \leqslant x^i < a^i \text{ pour } i = 1, \ldots, \nu\}$ et $a^1, \ldots, a^\nu \to \infty$

[2] Voir précédente note en bas de page.

[3] On montre que P est strictement convexe : $P\left(\frac{1}{2}\Phi + \frac{1}{2}\Psi\right) = \frac{1}{2}P(\Phi) + \frac{1}{2}P(\Psi)$ seulement si $\Phi = \Psi$.

9.2. *Le maximum de (**) est atteint précisément pour les états invariants σ qui sont des états d'équilibre.*

Si K_Φ est l'ensemble des états d'équilibre, le maximum de (**) est donc atteint sur $I \cap K_\Phi$.

9.3. *L'ensemble $I \cap K_\Phi$ est non-vide, il est convexe compact, c'est un simplexe de Choquet, c'est une face du simplexe I.*

Un état d'équilibre invariant σ a donc une décomposition unique en états d'équilibre invariants extrémaux. Comme $I \cap K_\Phi$ est une face de I, cette décomposition est la même que la décomposition de σ en états invariants extrémaux (décomposition ergodique). On l'interprète physiquement comme décomposition de σ en *phases thermodynamiques pures.*

9.4. *Il existe un "grand" sous-ensemble $D \subset \mathcal{B}$ (D est résiduel) tel que $I \cap K_\Phi$ est réduit à un point si $\Phi \in D$.*

Donc "en général" il n'existe qu'une phase *thermodynamique pure associée* à une interaction $\Phi \in \mathcal{B}$. (C'est une forme faible de la *règle des phases de Gibbs*). Quand une phase thermodynamique pure a une décomposition non-triviale en états d'équilibre extrémaux (voir 4.2) on dit que l'on se trouve en présence d'une *symétrie brisée* (la symétrie dont il s'agit est l'invariance par translations du réseau). Dobrushin [2] a donné des exemples explicites de symétries brisées.

BIBLIOGRAPHIE

[1] Dobrushin R.L. — Gibbsian probability field for lattice systems with pair interactions, *Funkts. Analiz i ego Pril,* 2, p. 31-43.

[2] Dobrushin R.L. — The question of uniqueness of a gibbsian probability field and problems of phase transitions. *Funkts. Analiz i ego Pril.* 2, 1968, p. 44-57.

[3] Haag R., Hugenholtz N.M. and Winnink M. — On the equilibrium states in quantum statistical mechanics. *Commun. Math. Phys.* 5, 1967, p. 215-236.

[4] Lanford O.E. — *The KMS states of a quantum spin system in Systèmes à un nombre infini de degrés de liberté,* C.N.R.S., Paris, 1970.

[5] Lanford O.E., Ruelle D. — Observables at infinity and states with short range correlations in statistical mechanics. *Commun. Math. Phys.* 13, 1969, p. 194-215.

[6] Ruelle D. — *Statistical mechanics. Rigorous results,* Benjamin, New York, 1969.

[7] Takesaki M. — *Tomita's theory of modular Hilbert algebras and its applications.* Springer, Berlin, 1970.

I.H.E.S
Route de Chartres,
91. Bures-Sur-Yvette (France)

Actes, Congrès intern. Math., 1970, Tome 3, p. 21 à 26.

ANALYTIC FUNCTIONS OF SEVERAL COMPLEX VARIABLES AND AXIOMATIC QUANTUM FIELD THEORY

by V. S. VLADIMIROV

At the International Congress of Mathematicians in 1958 (Edinburgh) N.N. Bogolyubov made our common report "On some Mathematical Problems of Quantum Field Theory" [1], in which particularly a problem of a rigorous prove of dispersion relations in quantum field theory was discussed and in connection with this it was pointed out the importance of the problem of analytic extension of generalized functions. The following decade confirms these trends : now the theory of functions of several complex variables found many applications in quantum field theory. On the other hand the quantum field theory found oneself as a source of many nontrivial problems in the theory of analytic functions.

1. In axiomatic quantum field theory the physical quantities arise as boundary values of some classes of analytic functions of several complex variables holomorphic in some primitive domains defined by axioms.

But in the complex space C^N of dimension $N \geqslant 2$ an aribitrary domain is not in general a domain of holomorphy. Therefore a nontrivial problem of construction an envelope of holomorphy $\mathcal{H}(\mathcal{D})$ of a given domaine $\mathcal{D}$ appears. Next step would consist in finding a corresponding integral representation for functions holomorphic in $\mathcal{H}(\mathcal{D})$. This representation ought to express values of a function in $\mathcal{H}(\mathcal{D})$ by means of its values on an "essential" part of the boundary of $\mathcal{H}(\mathcal{D})$. By this we may naturally hope that for classes of functions under consideration the values over the "essential" part of the boundary can be determined from experiments. Passing to the integral representation obtained from the corresponding boundary values we get the so-called (many dimensional) dispersion relations between quantities observed in experiments. Realization of this programme in the frame of some system of axioms would give firstly a possibility to verify experimentaly the consistency of the axioms considered and secondly would lead to an analytical approach which is capable to predict results of experiments.

2. Now we shall turn our attention to the axiomatic quantum field theory in Wightman's approach [2]. I shall not discuss here this wellknown system of axioms. It has been exposed in detail in the books by Streater and Wightman [3], by Jost [4] and by Bogolyubov, Logunov and Todorov [5].

The analysis of the axioms in this approach is reduced to investigations of the vacuum expectation values of the product of field operators $A_j(x_j)$ (Wightman's function)

$$W^J_{n+1}(x_0, \dots, x_n) = \langle \Psi_0, A_{j_0}(x_{j_0}) \dots A_{j_n}(x_{j_n}) \Psi_0 \rangle, \; n = 1, 2, \dots$$

where J is the permutation $(0, 1, \dots, n) \to (j_0, j_1, \dots, j_n)$. It turns out that all functions W^J_{n+1} are the boundary values in sense of $\mathcal{S}'$ of a single function $w_n(\zeta_1 - \zeta_0, \dots, \zeta_n - \zeta_{n-1})$ holomorphic with respect to variables

$$z_k = \zeta_k - \zeta_{k-1}, \; k = 1, 2, \dots, n$$

in the domain T_n — the union of the permutted extended tubes τ'^J_n over all permutations J ; the extended tube τ'_n is the union of the images $\Lambda\, \tau^+_n$ of the tube domain $\tau^+_n = \mathbf{R}^{4n} + i V^{\times n}_+$ by all complex (proper) Lorentz transformations $\Lambda \in L_+(C)$; here V_+ is the future light cone. Hence the function w_n has some definite properties of growth.

Note that this list of linear properties of functions w_n may be extended by some (nonlinear) conditions so that the theory can be reconstructed completely (up to the unitary equivalence).

It is interesting that the domain τ'_n, in contrary to τ^+_n, contains real points so-called Jost points J_n.

3. Thus we have a sequence of the primitive domains T_n, $n = 1, 2, \dots$ (Evidently the domain T_1 is $\mathbf{C}^4$ with the "cut" $z^2 = \rho$, $\rho \geqslant 0$).

For $n \geqslant 2$ the domain T_n is not a domain of holomorphy and therefore a nontrivial problem arises to construct its envelope of holomorphy $\mathcal{H}(T_n)$. $\mathcal{H}(T_2)$ has been constructed by Källén and Wightman [6] and the corresponding integral representation by Källén and Tall [7] (for functions with sufficiently regular boundary values).

For $n \geqslant 3$ $\mathcal{H}(T_n)$ is not yet constructed because of increasing complexity of the problem. It is not clear if $\mathcal{H}(T_n)$ is a schlicht domain although the domain T_n is schlicht (Tomozava [8]). A very general theorem was proved by Ruelle [9] : each completely space-like point, i.e. a real point $z_k = x_k - x_{k-1}$, $k = 1, 2, \dots, n$ for which $(x_k - x_j)^2 < 0$, $k \neq j$ is contained in $\mathcal{H}(T_n)$. For $n \geqslant 3$ this set is larger than the Jost points set J_n.

4. The retarded functions have similar primitive domain of analyticity Δ_n in momentum space of variables $k = (k_0, \dots, k_n)$. But the presence of nonzero threshold masses in spectral conditions and the appearance of some linear relations between retarded functions (Steinmann's indentities) lead to a further extension of the primitive domain Δ_n (Steinmann [10], Ruelle [11], Araki [12]). Note : if all the thershold masses are positive then the domain Δ_n is star-shaped, $\mathcal{H}(\Delta_n)$ as well ; hence $\mathcal{H}(\Delta_n)$ is schlicht.

A great number of papers have been devoted to the important case of the four-point function ($n = 3$) in the frame of the linear theory. In these papers the dispersion relations with respect to s for scattering amplitudes

$$T(s, t), s = (k_0 + k_1)^2, t = (k_0 + k_2)^2$$

for various two-point scattering processes were proved, the analyticity properties T with respect to both variables s and t were established, the analyticity properties of partial wave amplitudes in s as well, the "crossing"-property was proved and so on. Especially it should be noted important results obtained by Bros, Epstein and Glaser [13-14] and Hepp [15]. In particular some estimates from below of the envelope of holomorphy $\mathcal{H}(\Delta_n)$ obtained by these authors allow to overcome a difficulty in restriction of generalized functions on the mass surface.

5. The listed results have been obtained by very complicate and laborious ways; in such a short report it is impossible to represent them with some completeness. Therefore I shall restrict myself to the short exposition of the following two results i) the so-called "C-convex hull" theorem and ii) some generalization of the integral representation of the Jost-Lehmann-Dyson type([1]). Besides we shall hint at some applications of these results in quantum field theory.

Using the "edge of the wedge" theorem, firstly discovered and proved by Bogolyubov ([16], 1956) for functions of retarded and advanced types, we get a wellknown primitive domain of analyticity of the Dyson type

$$\mathcal{D} = (\mathbf{R}^N + iC) \cup (\mathbf{R}^N - iC) \cup \widetilde{\mathcal{O}}$$

where C is a (convex) proper cone in $\mathbf{R}^N$ with vertex at the origin and $\widetilde{\mathcal{O}}$ is some neighborhood in $\mathbf{C}^N$ of the real open set $\mathcal{O}$. It is nessary to construct the envelope of holomorphy $\mathcal{H}(\mathcal{D})$. For arbitrary cone C and open set $\mathcal{O}$ it has not yet been constructed. (For future light cone $C = V_+$ it was made by Bros, Messia and Stora [17] provided that $\mathcal{O}$ is bounded by two space-like surfaces). Nevertheless it is possible to indicate some real points which are contained in $\mathcal{H}(\mathcal{D})$ and not contained in $\mathcal{D}$, namely :

The inclusion $\text{ch}_c(\mathcal{O}) \subset \mathcal{H}(\mathcal{D})$ is valid. (It is the content of the so-called "C-convex hull" theorem, Vladimirov [18-19], Borchers [20]). Here $\text{ch}_c(\mathcal{O})$ is C-convex hull of $\mathcal{O}$ which is the smallest open set which contains $\mathcal{O}$ and has the property : if a C-like curve is contained in $\text{ch}_c(\mathcal{O})$ then each homotopic C-like curve is also contained in $\text{ch}_c(\mathcal{O})$. (A smooth curve is called C-like if at each point its tangent lies in the cone C).

From the "C-convex hull" theorem follows a corollary : if a tempered distribution vanishes in $\mathcal{O}$ and the support of its Fourier transform is contained in $(-C^* \cup C^*) + K$, where K is a compactum and C^* is the conjugate cone to C, then it vanishes also in $\text{ch}_c(\mathcal{O})$. Thus we have here a peculiar property of quasi-analyticity for the class of distributions under consideration. This fact was firstly observed by Dyson (details see in [21]).

[1] Another interesting generalization of the Jost-Lehman-Dyson representation using $\mathcal{L}_2$-estimates by Hömander [32] for $\bar{\partial}$-operator was given by Seneor [31].

6. We shall now give two applications of the last result in axiomatic quantum field theory.

(i) Wigthman [22] put a question : if a commutator vanishes for $(x - y)^2 < - l^2$ would it vanishe for $(x - y)^2 < 0$? The affirmative answer to this question follows immediately from the corollary of 5 because of equality

$$\mathrm{ch}_{V_+} (\xi^2 < - l^2) = (\xi^2 < 0)$$

without using the $L_+^\uparrow$-invariance axiom (Petrina [23]). Another proof using $L_+^\uparrow$-invariance was given by Wightman [24]. He obtained an even more strong result : if a commutator. vanishes for x and y varying in some space-like separated set then it vanishes for $(x - y)^2 < 0$.

(ii) Let $\mathcal{A}(\varphi)$ be the "smeared" field operators provided that their supports φ are contained in a given open set $\mathcal{O} \subset \mathbf{R}^4$. Denote by $\mathcal{R}' (\mathcal{O})$ the von Neumann algebra of bounded operators weakly commuting with all operators $\mathcal{A}(\varphi)$. From 5, we get immediately Borcher's theorem [20] (see also Wightman [25]) :

$$\mathcal{R}' (\mathrm{ch}_{V_+}(\mathcal{O})) = \mathcal{R}' (\mathcal{O})$$

without refering to the $L_+^\uparrow$-invariance of the theory. Moreover in that last equation the $\mathrm{ch}_{V+} (\mathcal{O})$ can be replaced by a larger hull : the real section of $\mathcal{H}(\mathcal{D})$ (Araki [26]).

These results indicate a close connection between Lorentz invariance, spectrality and locality of the theory. In this aspect there are also investigations by Bogolyubov and Vladimirov [27, 35], Streater [28] and Bros, Epstein and Glaser [29].

7. Now we shall present some generalization of the wellknown Jost-Lehman-Dyson representation.

Let a function $f(z)$ be holomorphic in the domain $\mathcal{D}$ (see 5) where $\mathcal{O} = - C$ and behaves like $O(|z|^a \; |\mathrm{Im}\, z|^{-\beta})$ as $|z| \to \infty$ or $|\mathrm{Im}\, z| \to 0$.

Then $f(z)$ can be represented in the form (Vladimirov and Jarinov [30])

$$(1) \qquad f(z) = \frac{i^N}{2\pi} l^m (z) \int_{\mathrm{pr}\, C^*} \left(\Delta_\sigma^{(N-1)} (\lambda), \frac{1}{(z, \sigma) - \lambda}\right) d\sigma$$

where $l(z)$ is an admissible polynomial for the domain $\mathcal{D}$; $m \geqslant 0$ an integer depending only on f ; $\Delta_\sigma (\lambda)$ is a continuous function of σ on pr C^* with values in $\mathcal{D}'_{\mathcal{L}_2,\lambda}$ with respect to the variable λ ; it vanishes for $\lambda < 0$. For a given admissible polynomial $l^m (z)$ the function $\Delta_\sigma(\lambda)$ is unique and it is determined by the Radon transform of the boundary values of $f(z) l^{-m} (z)$. From representation (1) it follows that all functions $f(z)$ under consideration are holomorphic in $\mathbf{C}^N$ with "cuts" $(z, \sigma) = \lambda, \lambda \geqslant 0$, $\sigma \in$ pr C^*. The representation (1) admits a generalization to the Dyson type domains $\mathcal{D}$ in which $\mathcal{O}$ is a union of some translations of cones C and $- C$ [30].

For $N = 2$, $C = [\mathrm{Im}\, s > 0, \mathrm{Im}\, t > 0]$ the formula (1) gives a new integral representation of the Nakanishy type for the scattering amplitudes $T(s, t)$ in perturbation theory, corresponding to the plane convergent Feynmann diagramms.

8. Here we almost did not touch the analytic properties of the Feynmann amplitudes. A progress in this direction achieved last years came from using new powerfull methods of homological algebra. These investigations are principally connected with the names of Pham, Lascoux, Froissart, Fotiadi, Federbush, Regge, . . . It is presented in books by Hwa and Teplitz [33] and by Pham [34] where further references can be found.

ЛИТЕРАТУРА

[1] Bogolyubov N.N. and Vladimirov V.S. — *Proc. I.C.M.*, 1958, p. 19-32, (1959).

[2] Wightman A.S. — *Phys. Rev.*, 101, 1956, p. 860-866.

[3] Streater R.F. and Wightman A.S. — *P.C.T., spin and statistics, and all that*, Benjamin, 1964.

[4] Jost R. — *The general theory of quantized fields*, Providence, 1965.

[5] Боголюбов Н. Н., Логунов А. А. и Тодоров И. Т. — Основы аксиоматического подхода в квантовой теории поля, « Наука », 1969.

[6] Källén G. and Wightman A.S. — *Mat. Fys. Skr. Dan. Vid. Selsk.*, I, N 6, 1958.

[7] Källén G. and Tall J. — *Helv. Phys. Acta*, 33, 1961, p. 753-772.

[8] Tomozava Y. — *J. Math. Phys.*, 4, 1963, p. 1240-1252.

[9] Ruelle D. — *Helv. Phys. Acta*, 32, 1959, p. 135-137.

[10] Steinmann O. — *Helv. Phys. Acta*, 33, 1960, p. 257-298; p. 347-362.

[11] Ruelle D. — *Nuovo Cimento*, 19, 1961, p. 356-376.

[12] Araki H. — *J. Math. Phys.*, 2, 1961, p. 163-177.

[13] Bros J., Epstein H., Glaser V. — *Nuovo Cimento*, 31, 1964, p. 1265-1302.

[14] Bros J., Epstein H., Glaser V. — *Commun. Math. Phys.*, I, 1965, p. 240-264.

[15] Hepp K. — *Helv. Phys. Acta*, 37, 1964, p. 639-658.

[16] Боголюбов Н. Н., Медведев Б. В. и Поливанов М. К. — Вопросы теории дисперсионных соотношений, Физматгиз, 1958.

[17] Bros J., Messia A., Stora R. — *J. Math. Phys.*, 2, 1961, p. 639-651.

[18] Владимиров В. С. — ДАН СССР, 134, 1960, стр. 251-254.

[19] Владимиров В. С. — Труды Матем. Ин-та им. В. А. Стеклова, 60, 1961, стр. 101-144.

[20] Borchers H.J. — *Nuovo Cimento*, 19, 1961, p, 787-793.

[21] Владимиров В. С. — Методы теории функций многих комплексных переменных, « Наука », 1964.

[22] Wightman A.S. — *Les problèmes mathématiques de la théory quantique des champs*, Paris, 1959, p. 1-38.

[23] Петрина Д. Я. — УМЖ, 13, № 4, 1961, стр. 109-111.

[24] Wightman A.S. — *J. Indian. Math. Soc.*, 24, 1960-1961, p. 625-677.

[25] Wightman A.S. — *Ann. Inst. Henri Poincaré*, Sect. A., I, 1964, p. 403-420.

[26] Araki H. — *Helv. Phys. Acta*, 36, 1963, p. 132-139.

[27] Боголюбов Н. Н. и Владимиров В. С. — НДВШ, физ.-мат науки, № 3, 1958, стр. 26-35.

[28] Streater R.F. — *J. Math. Phys.*, 3, 1962, p. 256-261.

[29] Bros J., Epstein H. and Glaser V. — *Commun. Math. Phys.*, 6, 1967, p. 77-100.

[30] Владимиров В. С. и Жаринов В. В. — О представлении типа Иоста-Лемана-Дайсона, ТМФ, 3, 1970, стр. 305-319.

[31] Seneor R. — *Commun. Math. Phys.*, II, 1969, p. 233-256.

[32] HÖRMANDER L. — *Acta Math.*, 113, 1965, p. 89-152.
[33] HWA C. and TEPLITZ V.L. — *Homology and Feynman integrals*, Benjamin, 1966.
[34] PHAM F. — *Introduction à l'étude topologique des singularités de Landau*, Paris, 1967.
[34] Боголюбов Н.Н. и Владимиров В.С. — Представления и-точечных функций, препринт Р 2-5662, ОИЯИ, 1971.

Steklov Mathematical Institute
Vavilova street 42,
Moscow V 333
U.R.S.S

Actes, Congrès intern. Math., 1970. Tome 3, p. 27 à 32.

E2 - THÉORIE DE LA RELATIVITÉ

PROBLÈMES MATHÉMATIQUES EN RELATIVITÉ

par Yvonne CHOQUET-BRUHAT

Introduction

La Relativité Générale est une très belle théorie, dont les axiomes sont simples. Cependant, en plus des problèmes d'interprétation physique et de vérification expérimentale (qui sont particulièrement à l'ordre du jour ces derniers temps) la Relativité Générale pose des problèmes mathématiques, dont certains sont résolus et beaucoup d'autres non. Ces problèmes sont, comme dans toute théorie, de deux types, construction de modèles et propriétés générales. C'est de ce deuxième groupe de problèmes que je parlerai ici, dans le cadre de la Relativité Générale classique.

1. Définitions

Un espace-temps, de la relativité générale, est une variété différentiable de dimension 4 de classe C^p, $p \geqslant 1$, munie d'une métrique riemanienne hyperbolique, de signature (+ ——). Deux espaces-temps isométriques sont considérés comme identiques. Etant donné une structure $C^p (p \geqslant 1)$ sur une variété V, il existe toujours une structure C^∞ sur V qui induit cette structure C^p : on ne restreint pas la généralité des espaces temps en ne considérant que des variétés C^∞.

V étant C^∞ on peut y considérer des tenseurs distributions quelconques. La généralité des métriques sera cependant limitée par la condition d'hyperbolicité si on désire la conserver. Si cette condition impose la posivité sur les vecteurs d'un champ (C^∞) de cônes convexes, on ne trouvera pour ces métriques que des mesures.

2. Equations d'Einstein

Un espace-temps (V_4, g) est dit einsteinien si le tenseur de Ricci de sa métrique g est égal à un tenseur donné sur V_4 c'est-à-dire si

$$(2\text{-}1) \qquad S_{\alpha\beta} \equiv R_{\alpha\beta} - \frac{1}{2} g_{\alpha\beta} R = T_{\alpha\beta} \quad , \quad R = R^{\alpha}_{\alpha}$$

A cause des identités de Bianchi, T, dit tenseur d'impulsion énergie soit satisfaire aux "conditions de conservation"

$$(2\text{-}2) \qquad \nabla_\alpha \, T^{\alpha\beta} = 0$$

Sauf dans des cas de symétrie de l'espace-temps il faudra pour que la définition de $R_{\alpha\beta}$ (non linéaire en g) ne dépende pas du mode de calcul que les coefficients de connexion soient des fonctions. Sous cette hypothèse il peut n'y avoir ni existence, ni unicité, des arcs géodésiques ; la distribution **T** est d'ordre $\leqslant 1$.

3. Théorèmes locaux

Dans le vide, le tenseur **T** est nul. Dans le cas général **T** doit représenter au mieux les propriétés des sources. Les choix de $T_{\alpha\beta}$ ont été inspirés par l'analogie avec la théorie classique, les lois de la dynamique de la relativité restreinte et le principe de covariance : bien des modèles de milieux continus, dont les équations macroscopiques sont assez bien connues en mécanique classique, n'ont pas encore de représentation satisfaisante en Relativité Générale. Un modèle ayant de bonnes propriétés est le "fluide parfait chargé sans induction" ($\tau_{\alpha\beta}$ tenseur de Maxwell)

$$(3\text{-}1) \qquad T_{\alpha\beta} = (p + p)\, u_\alpha\, u_\beta - p g_{\alpha\beta} + \tau_{\alpha\beta} \quad , \quad u^\alpha\, u_\alpha = 1$$

Aux équations 2-1, 2-2 sont à joindre les équations de Maxwell

$$(3\text{-}2) \qquad \oint \nabla_\alpha\, F_{\beta\gamma} = 0$$

et soit

$$(3\text{-}3.a) \qquad \nabla_\lambda\, F^{\lambda\mu} = J^\mu \equiv \lambda\, u^\mu + \sigma\, F^{\lambda\mu}\, u_\lambda \qquad \text{(conductivité } \sigma \text{ finie)}$$

soit

$$(3\text{-}3.b) \qquad F^{\lambda\mu}\, u_\mu = 0 \qquad \text{(conductivité infinie)}$$

On montre (cf Y. Choquet-Bruhat, A. Lichnerowicz, O. Friedrichs) que en coordonnées bien choisies (harmoniques) le système 3-2, 3-3.a (ou 3-3.b), $T_{\alpha\beta}$ donné par 3-1, est hyperbolique strict ou non strict au sens de Leray-Ohya selon que la conductivité est nulle, finie[1] ou infinie, pour lequel le problème de Cauchy local a une solution dans un espace de Sobolev ou un espace de Gevrey, dont le domaine de dépendance est déterminé par le cône défini par la métrique g (pour une équation d'état $\varphi = \varphi(p)$ convenable*), donc exhibe le caractère causal que l'on attend d'un système relativiste.

De nombreuses études ont été faites sur les modèles dissipatifs (Lichnerowicz, Pham Mau Quan, Pichon, Choquet-Bruhat, Mahjoub...). Un modèle récemment introduit, directement issu de la mécanique statistique, est :

$$(3\text{-}4) \qquad T_{\alpha\beta}\,(x) = \int_{P_x} f(x, p)\, p^\alpha\, p^\beta\, \overline{\omega}$$

où $f(x, p)$ est une fonction de distribution à une particule définie sur un espace de phase qui est un fibré de base V_4 et de fibre P_x, sous ensemble de l'espace

[1] Une autre variable thermodynamique que ρ et p peut être introduite, ainsi qu'une équation supplémentaire (cf Taub. Pichon Lichnerowicz), sans changer la nature du système dans le cas adiabatique (Lichnerowicz).

tangent à V_4 défini par exemple par une des relations : (outre $p^0 > 0$ expriment l'orientation temporelle en coordonnées adaptées)

$$g_{\alpha\beta}\, p^\alpha\, p^\beta = m_i^2$$

$i = 1, \ldots, N$ (particules de masses $m_i \geqslant 0$ données) ou

$$m^2 \leqslant g_{\alpha\beta}\, p^\alpha\, p^\beta \leqslant M^2$$

(particules de masses comprises entre $m > 0$ et $M < \infty$, cas des amas stellaires)

$\bar{\omega}$ est la forme de Leray sur P_x.

La fonction f est astreinte à vérifier l'équation de Liouville

$$p^\alpha \frac{\partial f}{\partial x^\alpha} - \Gamma^\alpha_{\lambda\mu}\, p^\lambda\, p^\mu \frac{\partial f}{\partial p^\alpha} = 0 \tag{3-5}$$

dont la vérification entraîne les conditions de conservation (3-4).

On peut démontrer un théorème d'existence et d'unicité locales, en coordonnées harmoniques de la solution du problème de Cauchy pour le système couplé d'Einstein-Liouville, dans des espaces de Sobolev H_μ, satisfaisant au principe de causalité. Ce résultat s'étend au cas électromagnétique. On peut aussi sans doute mettre au 2ème membre de (3-5) un opérateur de collision, avec une section efficace convenable, comme le font prévoir les travaux de Pichon sur l'équation de Boltzman linéarisée en relativité restreinte (existence globale et comportement asymptotique), Bitcheler et Bancel (étude locale de l'équation non linéaire).

4. Problème de Cauchy intrinsèque et global

Un système einsteinien, en coordonnées quelconques, est un système mal posé au sens de Cauchy. Cependant le problème véritable, de nature géométrique, a une solution possédant des propriétés satisfaisante, si les théorèmes locaux précédents sont vérifiés, et si les données initiales satisfont certaines équations (contraintes). Je me bornerai à parler des équations du vide. L'extension aux modèles avec sources est aisée quand les théorèmes locaux d'existence et d'unicité (§ 3) sont connus.

On aboutit, après interprétation géométrique, aux définitions suivantes :

Definition. – Une "donnée initiale" $\mathfrak{I}$ pour les équations d'Einstein est une variété Σ de dimension 3, une métrique riemanienne $\bar{\mathrm{g}}$ (définie < 0) sur Σ, et un tenseur du second ordre $\mathbf{P}$ sur Σ.

Une solution de ce problème de Cauchy, est un espace temps $(V_4\ , \mathrm{g}) = M$, à tenseur de Ricci nul, tel qu'il existe un difféomorphisme Λ de Σ sur une sous variété S et M, tel que la métrique et la 2ème forme fondamentale induites par M sur S soient identiques à l'image par Λ de $\bar{\mathrm{g}}$ et $\mathbf{P}$.

Il est bien connu qu'une condition nécessaire pour que ce problème de Cauchy ait une solution est que la donnée initiale $\mathfrak{I}$ vérifie sur Σ les équations, appelées contraintes :

$$\bar{R} + P_{ij}\, P^{ij} - (P)^2 = 0 \tag{4-1}$$

$$(4\text{-}2)\cdot \qquad \overline{\nabla}_j \, (P^{ij} - \delta^j_i \, P) = 0 \qquad P = P^i_i$$

($\overline{\nabla}$ dérivation covariante dans la métrique $\overline{\mathrm{g}}$)

On montre, en utilisant les théorèmes d'existence et d'unicité en coordonnées harmoniques que les équations (4-1), (4-2) sont aussi suffisantes pour assurer l'existence d'une solution, si les données Σ, $\overline{\mathrm{g}}$, **P** sont assez régulières. Les meilleurs résultats (quant à la régularité minimum) sont obtenus en utilisant la théorie de Leray-Dionne, on a :

THEOREME 1 — *A toute donnée initiale* $\mathfrak{J}$, *vérifiant les contraintes, et telle que*

$$\Sigma \in C^{s+\nu} \quad , \quad \overline{\mathrm{g}} \in \widetilde{H}_{4+\nu} \quad , \quad \mathrm{P} \in \widetilde{H}_{3+\nu} \quad , \quad \nu \geqslant 0$$

correspond une solution des équations d'Einstein, $\mathrm{g} \in \widetilde{H}_{4+\nu}$

($\widetilde{H}_m$ est l'espace de Sobolev des fonctions localement de carré sommable ainsi que leurs dérivées d'ordre $\leqslant m$ au sens des distributions).

A ce théorème d'existence correspond un théorème local d'unicité géométrique :

THEOREME II — *Deux solutions des équations d'Einstein, pour une donnée initiale* $\mathfrak{J}$, *sont extensions d'une même solution, si* $\nu \geqslant 1$.

M' est une extension de M si M est isométrique à un sous-ensemble de M', dans une isométrie ψ *qui "conserve la variété initiale"* ($\Lambda^{-1} \, \psi \, \Lambda'$ *est l'identité sur* Σ).

Aucun théorème global d'existence n'est connu jusqu'à présent (global signifiant ici que l'espace temps solution est complet en un sens ou un autre). Un certain nombre de théorèmes de non existence globale sont connus (le premier, du à Lichnerowicz traitant le cas stationnaire, des plus récents dus essentiellement à Penrose et Hawking, faisant intervenir des hypothèses diverses sur les données initiales ou les solutions cherchées. Le problème de l'existence de telles solutions non minkoskiennes, avec des propriétés globales de régularité et de comportement asymptotique convenables est encore un problème ouvert.

Par contre un théorème global d'unicité a été démontré récemment (Y. Choquet-Bruhat and R. Geroch (1969)), faisant suite à une version plus faible (Y. Choquet-Bruhat (1968)), dans la classe des espaces-temps globalement hyperboliques.

Globale hyperbolicité. (Cette notion a été introduite par Leray pour l'étude globale des équations aux dérivées partielles hyperboliques linéaires). (V_4, g) est dit globalement hyperbolique si l'ensemble des chemins temporels ou isotropes joignant deux points x et y est compact (dans la topologie de l'espace des chemins).

Une propriété fondamentale des variétés globalement hyperbolique est que leur topologie est équivalente à la topologie de l'ordre définie par la causalité.

Une surface de Cauchy, dans un espace-temps, est une sous variété de dimension 3 telle que tout chemin temporel ou isotrope la rencontre une fois et une seule. Geroch a montré qu'un espace-temps admet une surface de Cauchy si et seulement si il est globalement hyperbolique.

Une solution $M = (V_4 \, , \mathrm{g})$ est dite un developpement de la donnée initiale $\mathfrak{J} = (\Sigma, \overline{\mathrm{g}}, \mathrm{P})$ si M est globalement hyperbolique et admet S comme surface de

Cauchy. On munit l'ensemble des développements de $\mathcal{J}$ d'une structure qu'on montre être un ordre partiel par la notion d'extension, et on démontre :

THEOREME III. – *Toute donnée initiale $\mathcal{J}$ régulière et vérifiant les contraintes admet un développement maximal et un seul.*

5. Contraintes (4-1, 4-2)

Il n'est pas facile, à cause de la non linéarité de ces équations, d'en trouver des solutions globales $\overline{\mathbf{g}}$, **P** sur une variété V_3. Des solutions particulières ont été calculées de diverses manières, et, d'autre part, par divers choix des inconnues des formes variées ont été proposées pour la recherche des solutions générales. Je vais ici proposer un nouveau choix, qui a l'avantage d'être un système de quatre équations aux dérivées partielles à quatre inconnues, de type elliptique, sur une variété V_3 quelconque ; on trouve :

$$(5\text{-}1) \qquad -8\,\Delta^*\varphi + R^*\varphi + L\,\varphi^5 = -2T_0^0\,\varphi^5\,,\ L = P_{ij}\,P^{ij} - (P)^2$$

$$(5\text{-}2) \qquad -\Delta^*\lambda_i^* + 8\,\nabla_i^*\left(\lambda_k^*\,\frac{\partial_k^*\varphi}{\varphi}\right) - \frac{1}{\varphi}(6\,P_{ik}\,\partial_k^*\varphi - 2\,P_j^j\,\partial_i^*\varphi)$$

$$+\frac{1}{2}\nabla_j^*\,(A_{ij}^* - \delta_{ij}A^*) = T_i^0\,\varphi^2$$

où $\Delta^*\lambda_i = \nabla_j^*\nabla_j^*\lambda_i - R_{ij}^*\lambda_j$.

Les inconnues sont φ et λ_i^* ; $\mathbf{g}^*$ est une métrique définie < 0 quelconque avec, $\overline{\mathbf{g}} = \varphi^4\ \mathbf{g}^*$ (choix déjà donné par Lichnerowicz pour résoudre 4-1), A^* un tenseur symétrique arbitraire (lié à $\partial_0\ \mathbf{g}^*$). Le système (5-1), (5-2) n'est pas fortement elliptique à cause des termes en dérivées secondes de φ dans (5-2).

A. Simon-Vaillant a, dans sa thèse démontré l'existence sur R^3 d'une solution d'un système elliptique analogue[1] sous des hypothèses faiblement gravitationnelles. Sa méthode s'applique certainement à ce système. Il serait très intéressant d'obtenir des résultats d'existence, ou de non existence, pour des variétés plus générales que R^3, en particulier compactes.

6. Ondes gravitationnelles

Ce terme est utilisé pour désigner des solutions des équations d'Einstein qui ont des caractères communs avec les solutions d'autres équations (en particulier de Maxwell) que les physiciens appellent ondes : discontinuités ou oscillations qui se propagent en transportant de l'énergie.

On montre qu'il n'y a pas d'ondes de choc gravitationnelles intrinsèques et que les signaux gravitationnels à haute fréquence se propagent sans déformation. Une onde gravitationnelle à haute fréquence étant cherchée sous forme d'une perturbation oscillatoire d'une métrique donnée $\overset{0}{g}{}^{\alpha\beta}$:

[1] Ce système avait l'inconvénient, n'étant pas covariant, de ne pas pouvoir être utilisé sur une variété à plusieurs cartes.

$$(6\text{-}1) \qquad g^{\alpha\beta}(x, \omega\,\varphi) = \overset{0}{g}{}^{\alpha\beta}(x) + \frac{1}{\omega}\overset{1}{g}{}^{\alpha\beta}(x, \omega\,\varphi) + \frac{1}{\omega^2}\overset{2}{g}{}^{\alpha\beta}(x, \omega\,\varphi) + \dots$$

(6.1) correspond aux ondes asymptotiques des équations linéaires de Gårding-Kotaké-Leray : ω paramètre réel, grand, φ fonction que l'on trouve être nécessairement solution de l'équation caractéristique de la métrique donnée.

On trouve pour la partie significative[1] de cette perturbation, $\overset{0}{g}{}^{ij}$, si on choisit $x^0 = \varphi$, un système d'équations différentielles linéaires de propagation le long des rayons associés aux surfaces d'onde $\varphi = c^{te}$. De plus $\overset{0}{g}{}^{ij}$ doit satisfaire à 4 conditions algébriques linéaires dont la vérification sur une variété transverse aux rayons entraine la vérification dans le domaine balayé par ces rayons (ces ondes gravitationnelles ont donc 2 modes de polarisation).

On constate d'autre part que le fait pour(6-1)d'être solution (approchée d'ordre ω^{-1}) des équations d'Einstein impose la présence d'un second membre à ces équations, et une inégalité reliant cette source à une certaine forme quadratique de la perturbation qu'on peut raisonnablement interpréter comme énergie de radiation gravitationnelle : le couplage avec les équations de Maxwell voit s'ajouter à ce terme la classique énergie électromagnétique.

REFERENCES

On trouvera des bibliographies dans :

[1] Bruhat Y. — Cauchy problem, In *Gravitation, an introduction to current research,* (4. witten ed.), J. Wiley, 1962.

[2] Lichnerowicz A. — *Théories Relativistes de la gravitation et de l'électromagnétisme,* Masson, 1955.

[3] Lichnerowicz A. — *Relativistic hydrodynamics and magnetohydrodynamics,* Benjamin, 1967.

[4] Choquet-Bruhat Y. and Geroch R. — Global aspects of the Cauchy problem in General Relativity, *Comm. Maths. Phys.* 14, 1969, p. 328-335.

{5] Choquet-Bruhat Y. — *Solutions Radiatives approchées des équations d'Einstein, comm. Maths. Phys.* 12, 1969, p. 16-35.

Faculté des Sciences de Paris
Département de Mathématique
11, Quai Saint Bernard
Paris 5ème (France)

[1] C'est-à-dire invariante par les changements de coordonnées oscillatoires conservant la métrique de base.

Actes, Congrès intern. Math., 1970. Tome 3, p. 33

μ SPACE STATISTICAL MECHANICS IN GENERAL RELATIVITY THEORY

by Jürgen EHLERS

The material of that address is contained in the following two papers :

J. EHLERS, General-relativistic kinetic theory of Gases, published in the *Proceedings of the C.I.M.I.* (Centro Internazionale Matematico Estivo) summer session on relativistic fluid dynamics, held at Bressanone, June 1970 ; Edizioni Cremonese, Roma 1971.

J. EHLERS, Relativity and kinetic theory, published in *Rendiconti della Scuola Internazionale di Fisica "Enrico Fermi"*, corso XLVII, Roma 1969.

Max-Planck-Institut für
Physik und Astrophysik
8 Müncher 23
Föhringer Ring 6
Deutschland

Actes, Congrès intern. Math., 1970, Tome 3, p. 35 à 39.

SYMPLECTIC STRUCTURE AND QUANTIZATION OF THE EINSTEIN GRAVITATION THEORY

by L.D. FADDEEV

Quantum Field Theory once again attracts the attention of the mathematically minded theoretical physicists, endeavouring to prove its internal selfconsistency. The simplest nonlinear hyperbolic equation

$$\partial_0^2 u - \nabla^2 u + m^2 u + \lambda u^3 = 0$$

serves as a popular model for this goal. It is argued that this equation adequatly illustrates the more physical systems of differential equations, i.e. those of Electrodynamics. On the other hand one can not exclude the possibility that only after the inclusion of the gravitational field into the general system of quantum fields this theory will release itself from inherent infinities.

The problem of quantization of the gravitational field in its very beginning raises the questions having no analogue in the case of scalar equation above. This in great extent expository report treats just them. More exactly I shall describe the measure on the set of paths in the space of gravitational fields over which one must integrate the functional exp $\{i \times \text{action}\}$ in the Feynman formulation of quantum field theory. I shall rest on purely formal level and not dwell upon the mathematical difficulties of the exact definition of the quantum field dynamics which are by no means resolved till now.

1. Asymptotically flat gravitational field.

The description of the gravitational field includes such four dimensional pseudo Riemannian manifold V, metric g having Lorentz signature, that (V, g) is globally hyperbolic in sence of Bruhat-Leray [1]. Suppose that V is homeomorphic to the Euclidean space E_4 and g is asymptotically Galilean : on any three-dimensional space - like section

$$g = g_0 + O(r^{-1}) \quad ; \quad \partial g = O(r^{-2})$$

when three-dimensional distance r from some point goes to infinity g_0 being Minkowskian metric (cf. [2]). Let M denote the set of such metrics on fixed manofold V, M_0 being the subset of metrics with vanishing Ricci tensor.

Let diff$_a$ V be group of all diffeomorphisms V, preserving the described asymptotical structure, diff$_0$ V being the normal subgroup consisting of the diffeomorphisms which tend to identity at infinity. Then $P = \text{diff}_a V/\text{diff}_0 V$ is a group of isometries of the Minkowskian space. It is clear how to introduce the action of diff$_a$ V and diff$_0$ V on the sets of metrics M and M_0.

Classical asymptotically flat gravitational fields in vacuum are in one to one correspondence with the classes of metrics in the set $G = M_0/\mathrm{diff}_0\ V$. The factorisation reflects the important physical principle which has often misleading name of general covariance.

The Poincare group nontrivially acts on G which we can following Segal call the phase space of the gravitational theory. Thus the dynamics of this theory, which usually is called the general relativity, does not differ from that of special relativity (cf. Fock in [3]). It is the highly exotic construction of the phase space which distinguishes the theory of gravitation and leads to nontrivial questions when quantizing.

2. Feynman functional integral

The most universal though purely heuristic way to quantize the given classical system is the Feynman functional integral [4], [5]. I shall formulate its essentials in terms of mechanical system with finite number degrees of freedom. Let Γ be corresponding phase space, $\dim \Gamma = 2n$, $(p\,,q)$ – canonical coordinates, $\Omega = \Sigma\ dp\ \Lambda\ dp$ — symplectic form, $H(p\,,q)$ – energy function. Then

$$\psi\ (q\ (t_2), t_2) = \int \exp\left\{ i \int_{t_1}^{t_2} \Sigma\ p\ dq - H\ d\tau \right\} \psi\ (q\ (t_1), t_1) \prod_{\tau} (\Omega(\tau)/2\pi)^n$$

gives us the evolution of the wave function, describing the quantum mechanical state of the system. The integral over the paths $q(\tau)\,, p(\tau)\,, t_1 \leqslant \tau \leqslant t_2$ in R.H.S. can be defined by means of the limiting procedure, based on the finite dimensional approximation of the set of paths ; for instance one can use piece-wise linear $q(\tau)$ and piece-wise constant $p(\tau)$. In general the result depends on the approximation and so the Feynman integral by no means gives the exact meaning to the principle of correspondence. Nevertheless in the most important examples it leads to the unambigous results.

The Feynman measure $\exp\ \{i \times \text{action}\} \prod (\Omega/2\pi)^n$ is symplectically invariant up to the multiple depending on the ends of the paths. This dependence trivialises in the most important case when $t_1 \to -\infty$ and $t_2 \to \infty$. Thus the knowledge of the symplectic structure of given classical system allows to describe its quantum analogue by means of the Feynman integral.

The phase space for the nonlinear scalar equation written above can be simply parametrized in terms of the Cauchy data $\varphi(\vec{x}) = u(\vec{x}\,, t)\,, \pi(\vec{x}) = \partial_0\ u(\vec{x}\,, t)$. The corresponding Feynman measure assumes the form

$$\exp\left\{ i \int \left[\partial_0\ \psi\ \pi - \frac{1}{2}\pi^2 - \frac{1}{2}(\nabla\ \varphi)^2 - \frac{m^2}{2}\ \varphi^2 - \frac{\lambda}{4}\varphi^4 \right] dx \right\} \prod_x d\varphi(x)\ d\pi(x)$$

When integrating functionals depending only on $\varphi(x)$ one can trivially integrate over $\pi(x)$ and the measure transforms into the manifestly Lorentz invariant one

$$\exp\left\{ i \int \frac{1}{2}\ (\partial_0 \varphi)^2 - (\nabla \varphi)^2 - m^2\ \varphi^2 - \frac{\lambda}{2}\varphi^4\ \ dx \right\} \prod_x d\varphi(x)$$

Analogous expression in case of Quantum Electrodynamics was used by Feynman in the end of 40 –s to develope the manifestly covariant perturbation technics in terms of celebrated now diagrams.

The naive generalization of the last formula for the case of gravitation theory leads to the expression

$$\exp\{i \int R \, d\mu\} \prod_x dg(x)$$

where R is scalar curvature (the Lagrange function for the Einstein equation), $d\mu$ — Riemannian volume element of V, dg — volume element in the space of Lorentz definite matricies. It was Feynman himself who showed that this formula contradicts the general properties of quantum mechanics [6]. Thus to derive the true generalization one must return to its symplectic form. The knowledge of the symplectic structure of the gravitational field is nessesary for this.

3. Symplectic structure of the gravitational field

The main result elucidated by papers of Dirac [7], Arnowitt-Deser-Misner [8], Schwinger [9], De-Witt [10] and others can be formulated as follows (cf. [11]). The manifold G has symplectic structure and action of Poincare group on G is canonical. This result was obtained by means of the parametrization of the phase space G in terms of the Caushy data for the Einstein equations. The factorization mentioned above involves this procedure and makes the description of G highly implicite.

Let S be three dimensional manifold homeomorphic to Euclidean space E_3. Consider manifold C of pairs γ, h, where γ –Riemannian metric on S and h – symmetric tensor of the rank 2. Let at infinity

$$\gamma = \gamma_0 + O(r^{-1}) \quad ; \quad \partial\gamma = O(r^{-2}) \quad ; \quad h = O(r^{-2}),$$

where γ_0 is Euclidean metric. Manifold C has natural symplectic structure, γ playing the role of coordinates and h – that of momenta. In particular the Poisson brackets $\{\ ,\ \}$ are defined for arbitrary functionals of γ and h. Given γ, the operations h^2, Trh, δh can be defined for symmetric tensors h [12], and scalar product $<,>$ is defined for arbitrary tensor fields sufficiently decreasing at infinity.

Let us consider S now as a Cauchy surface for the Einstein equations, γ and h being the first and second quadratic forms induced on S by metric g. The Gauss-Kodazzi conditions

$$\pi \equiv d\,Trh - \delta h = 0 \quad ; \quad H \equiv \rho + Tr h^2 - (Trh)^2 = 0,$$

where ρ is scalar curvature of γ, are necessary for this. We shall consider these equations as the constraints on the canonical variables γ and h. It is remarkable that these constraints are in involution. To formulate this condition in detail let us construct the following functionals on C :

$$T(X) = <\pi, X> \quad ; \quad T(f) = <H, f>,$$

where X and f are arbitrary vector and scalar fields on S with compact support. Then

$$\{T(X), T(Y)\} = T([X,Y]) \quad ; \quad \{T(X), T(f)\} = T(Xf)$$
$$\{T(f), T(g)\} = T(gX_f - fX_g),$$

where X_f –vector field γ– dual to the one form df. The last conditions allow us to introduce certain symplectic manifold connected with the submanifold C_0 of the solutions of the equations

$$\pi = 0 \quad , \quad H = 0$$

Let us describe the corresponding procedure on the finite dimensional example. Let Γ, Ω be as above and φ_i, $i = 1, \ldots, m$, $m < h$ be functions (constraints) on Γ such that $\varphi_i = 0$ define the submanifold Γ_0 of codimension m. Let further $\{\varphi_i, \varphi_k\}$ vanish on Γ_0. Then Hamiltonian vector fields Z_φ generated by constraints are tangent to Γ_0 and span the maximal involutive distribution singular for the form Ω_0 induced on Γ_0 by Ω (cf. [13], [14]). The space Γ^* of maximal integral manifolds for this distribution has symplectic structure and plays the role of physical phase space, defined by constraints in involution.

Returning to the case of gravitation let C^* be manifold, obtained by such procedure from C and constraints $T(X)$ and $T(f)$. Analyzing the Einstein equations one can show that C^* parametrizes G and find the functionals on C^* which generate the canonical action of the Poincare group. It is worth to mention here the geometrical meaning of constraints : $T(X)$ generates the group $\text{diff}_0\, S$ while $T(f)$ corresponds to the change of the surface S.

4. Feynman measure for gravitational field

The formulation of Feynman integral in the case of system with constraints was developed in my paper [14]. Let χ_i, $i = 1, \ldots, m$ be arbitrary functions on Γ such that $\chi_i = 0$ define submanifold of Γ_0 of codimension m and matrix $\mathcal{D} = \|\{\chi_i, \varphi_k\}\|$ is non-degenerate. Then Feynman measure can be written as follows

$$\exp\{i \int \Sigma\, p\, dp - H\, d\tau\} \prod_\tau \det \mathcal{D}\ \delta(\varphi)\ \delta(\chi)\ \Omega^n\ (2\pi)^{m-n}$$

and when considered on the set of functions on Γ^* it does not depend on the choice of the supplementary conditions χ_i. Introducing Lagrange multipliers λ_i conjugate with the constraints we can rewrite the last expression in the form

$$\exp\{i \int \Sigma\, p\, dq - H\, d\tau - \Sigma\, \lambda_i \varphi^i\, d\tau\} \prod_\tau \det \mathcal{D}\ \delta(\chi)\ (\Omega/2\pi)^n\, d\lambda$$

In the case of gravitation Dirac showed a convenient choice of the supplementary conditions. Geometrically they are the minimality of S and harmonicity of the coordinates on it. These conditions as well as corresponding matrix $\mathcal{D}$ can be described by the systems of differential equations. There exists the natural way for the regularization of det $\mathcal{D}$, the latter depending only on γ.

The Lagrange multipliers can be interpreted as vector and scalar fields completing γ up to the four dimensional metric g. The momenta h enter into the action and constraints quadratically so that it is possible to integrate over them explicitely.

These considerations lead to the expression of the Feynman measure in the form $\exp \{i \int R\, d\mu\}\ \Phi\ (g) \prod dg$, where $\Phi(g)$ is functional not invariant to $\mathrm{diff}_0\, V$. Change of supplementary conditions would change $\Phi(g)$ but not the measure on the set of functionals on $B = M/\mathrm{diff}_0\, V$ which can be considered as the set of paths in G. Having one description of Feynman measure we can calculate it in different parametrizations of G. Especially convenient parametrization is given by the choice of harmonic coordinates on V. Explicit calculation in this case leads to the expression of Feynman measure first given by V. Popov and me in [15]. The rules for Feynman diagrams following from this expression coincide with those obtained independently by De-Witt [16] and Mandelstam [17].

REFERENCES

[1] Choquet-Bruhat Y. — *Battelle Rencontres,* 1967, p. 84.

[2] Lichnerowicz A. — *Theories Relativistes de la Gravitation,* Masson, Paris, 1955.

[3] Fock V.A. — *Theory of Space, Time and Gravitations,* Moscow, 1955.

[4] Feynman R. — *Rev. Mod. Phys.,* 20, 1948, p. 367.

[5] Feynman R. — *Phys. Rev.* 84, 1951, p. 108.

[6] Feynman R. — *Acta Phys. Polonica,* 24, 1963. p. 697.

[7] Dirac P.A.M. — *Proc. Roy. Soc.* A 246, 1958, p. 326-333.

[8] Arnowitt R., Deser S., Misner C. — *in Gravitation; An Introduction to Current Research,* (ed. L. Witten), Wiley, N.-Y., 1962.

[9] Schwinger J. — *Phys. Rev.* 130, 1963, p. 1253; 132, 1963, p. 1317.

[10] De-Witt B. — *Phys. Rev.* 160, 1967, p. 1113.

[11] Faddeev L.D. — *5-th Internat. Conf. on Gravitation and Relativity,* Tbilici, 1968.

[12] Lichnerowicz A. — *Publ. Math. I.H.E.S.,* No. 10, 1961.

[13] Herman R. — *Phys. Rev.* 177, 1969, p. 2449.

[14] Faddeev L.D. — *Theor. and Math. Phys.* I, 1969, p. 3.

[15] Popov V.N., Faddeev L.D. — *Preprint Inst. Theor. Phys. Kiev.* 1967.

[16] De-Witt B. — *Phys. Rev.* 162, 1967, p. 1195-1239.

[17] Mandelstam S. — *Phys. Rev.* 175, 1968, p. 1604.

Mathematical Institute
Fontanka 25,
Leningrad D 11 (URSS)

Actes, Congrès intern. Math., 1970. Tome 3, p. 41 à 45.

SOME RECENT WORK ON GLOBAL PROPERTIES OF SPACETIMES

by Robert GEROCH *

The fundamental object in general relativity is a *spacetime* : a connected, Hausdorff, 4-dimensional differentiable manifold M (without boundary) on which there is specified a smooth (C^∞) metric tensor field g of signature $(-, +, +, +)$. The points of M represent the events (idealized occurrences having no extension in space or time, e.g., the snapping of one's fingers) of the physical world, while the metric gives the result of measuring the spatial distance or temporal interval between any pair of nearby events. For example, the collection of all events directly experienced by an observer is a timelike curve[1]. The elapsed time between two events on the curve, as measured by this observer, is given by the length of the curve between the events.

Most of the early work on spacetimes was confined to local properties : the suitability of M, g as a model of space and time, the structure and consequences of Einstein's equation, etc. While it is still true that many of the outstanding problems in general relativity are local, the study of spacetimes in the large has become increasingly important in recent years [1, 2, 3]. Much of this work is still in the exploratory stage : which intuitive ideas about the global structure of a spacetime can be formulated precisely ? what is the physical interpretation of various conditions ? how are apparently different conditions related ? Of particular interest are conditions which can be justified as being "physically reasonable requirements", for they serve to select, from the class of all spacetimes, a subclass which presumably includes a very special spacetime – that which represents our own physical universe. We shall here briefly discuss a few of these conditions (involving causality, orientability, determinism, and singularities), their interpretations, and the theorems relating them.

One important example of an additional condition which can be imposed on spacetimes is the absence of causality violations. If a spacetime admits self-intersecting timelike curves, then an observer represented by such a curve would be in a position to influence his own past. Such possibilities are certainly not observed in our own local region of the universe. Since, furthermore, causality violations would be difficult to reconcile with our intuitive idea of what "time" is, we are led to consider as physically reasonable only those spacetimes in which

[1] By a (smooth) *curve* we understand a smooth mapping $\gamma : R \to M$ together with all mappings obtainable from this one by smooth reparameterizations. A vector is said to be *spacelike, null,* or *timelike* according as its norm is, respectively, positive, zero, or negative. A *timelike curve* is a smooth curve whose tangent vector is everywhere timelike.

there are no self-intersecting timelike curves. Define a *spacelike section* as a closed, connected, 3-dimensional submanfiold S (without boundary) of M whose normal is every-where timelike. Spacelike sections represent "all space at one instant of time". A consequence of requiring causality nonviolation is, for example, a severe restriction on the possibilities for the "topology of space" to change from one epoch to another :

THEOREM 1 (*4*). – *Let M, g be a spacetime, and Let M' be a compact, connected, 4-dimensional submanifold whose boundary is the union of disjoint spacelike sections S and S'. Then either S and S' are diffeomorphic, or else M has self-intersecting timelike curves.*

For some purposes one requires a causality condition stronger than the absence of self-intersecting timelike curves. In fact, there exists a condition of this type which is, in some sense, the strongest. With each metric g on M, associate the cross-section, C_g, of the bundle of second-rank covariant tensors on M. A spacetime M, g is said to be *stably causal* if there exists a neighborhood U of C_g such that no spacetime M, g' with $C_g{}' \subset U$ has self-intersecting timelike curves. (More generally, every property of spacetimes has a stable version. Since a spacetime represents a physical system, and since with any measurement there is associated a certain error, it is natural to introduce conditions which ignore "small variations" of the metric). An interesting consequence of stable causality is the existence of a covering of the spacetime by disjoint spacelike sections. That is to say, in such a spacetime one can introduce a (never unique) global notion of simultaneity :

THEOREM 2 (*5*). – *A time-orientable*[1] *spacetime M, g is stably causal if and only if it admits a smooth scalar field t whose gradient is everywhere timelike.* The connected components of the surfaces t = const. are spacelike sections.

The requirement that spacetimes have no self-intersecting timelike curves stems, at least in part, from the fact that causality violations have never been observed. We next consider an example –orientability– in which a global condition arises more directly from local experiments. In fact, there are three types of orientability, with respect to time, parity, and charge. At $p \in M$, (i) the space of oriented, one-dimensional, timelike subspaces of the tangent space at p consists of two connected components (physically, the "future" and "past" time directions), (ii) the space of oriented, spacelike, three-dimensional subspaces of the tangent space at p consists of two connected components (the two possible "spatial parities") and (iii) the charged particles in the spacetime at p can be divided into two classes, "positively" and "negatively" charged[2]. We define a bundle K over M whose fibre (eight points) over $p \in M$ consists of combinations of one choice from each of (i), (ii), and (iii), and whose group G is isomorphic with $Z_2 \times Z_2 \times Z_2$. This K induces a homomorphism, $\Psi : \pi_1(M) \to G$, from the first homotopy group of M. Let T, P, and C denote the elements of G which interchange, respectively, the choices (i), (ii), and (iii). Local physics (of elemen-

[1] A spacetime is said to be *time-orientable* if its bundle of timelike vectors is not connected.

[2] Whereas (i) and (ii) involve only the spacetime itself, (iii) depends on additional physical observations in the spacetime.

tary particles) appears to be invariant under the combination *PCT*, but not invariant under *P*, *C*, or *PC* separately. The requirement that a spacetime be consistent with these local observations places restrictions on Ψ : *P* noninvariance, for example, means that, given a choice of (i) and (iii), one can determine a preferred element of (ii) by means of an experiment. Thus, $P \notin \Psi[\pi_1(M)]$.

THEOREM 3 (*6*). – *A spacetime M, g consistent with PCT invariance and P, C, and PC noninvariance has the property that $\Psi[\pi_1(M)]$ is contained in the subgroup of G consisting of the identity and PCT. (In particular, M must be orientable).*

A third condition on spacetimes involves the possibility of determining the metric, and perhaps also other fields, from data given at an initial time. Let the spacelike section S represent an "initial time". We wish to formulate the idea that any fields on the spacetime which are subject to suitable hyperbolic equations (e.g., Einstein's equation for the metric, Maxwell's equations for an electromagnetic field, the hydrodynamic equations for a fluid, etc.) will be completely and uniquely determined throughout M by appropriate initial data on S. Since all signals reaching $p \in M$ (and therefore all signals capable of influencing the situation at p) must "travel in timelike or null directions", we might expect every influence on p to have once been registered on S provided S has the following property : for every timelike or null curve, $\gamma : R \rightarrow M$, without endpoint[1] and with $p \in \gamma[R]$, $\gamma[R] \cap S$ is a single point. If this is true for every $p \in M$, and if M, g is time-orientable, we say that S is a *Cauchy surface* for the spacetime. (There is a theorem to the effect that a field satisfying a suitable hyperbolic equation is determined on M from initial data on any Cauchy surface). That a spacetime have a Cauchy surface is a very strong condition. For example,

THEOREM 4 (*7,1*). – *Let M, g be a spacetime with a Cauchy surface S. Then there exists a 3-dimensional manifold V and a diffeomorphism $\Lambda : M \rightarrow V \times R$ such that, for each $a \in R$, $\Lambda^{-1}[V, a]$ is a Cauchy surface.*

In particular, such an M, g is stably causal. (In fact, the existence of a Cauchy surface is a stable property of spacetimes). Theorem 4 shows that a spacetime on which fields are predictible from initial data on S also has the property that the underlying manifold M is "predictible" from S. (For M is diffeomorphic with $S \times R$.) The study of global properties would be enormously simplified if, in some way, one could establish that our own universe has a Cauchy surface. In fact, the existence of a Cauchy surface is often assumed in the treatment of essentially local problems in order to avoid global questions. One of the important global problems in general relativity is to decide whether or not the assumption of the existence of a Cauchy surface can be justified on physical grounds.

We discuss one further property connected with Cauchy surfaces. A condition for the existence of a Green's function in the theory of hyperbolic equations is

[1] A point $p \in M$ is said to be an *endpoint* of the curve represented by $\gamma : R \rightarrow M$ if $\gamma(\lambda) \rightarrow p$ as $\lambda \rightarrow +\infty$ or as $\lambda \rightarrow -\infty$.

the following (*8*) : a spacetime is said to be *globally hyperbolic*[1] if it is stably causal and, for $p , q \in M$, the union of the images of all timelike curves with endpoints p , q has compact closure in M. It is perhaps not suprising that the existence of a Cauchy surface should be related to the condition for the existence of a Green's function.

THEOREM 5 (*7*). – *A time-orientable spacetime has a Cauchy surface if and only if it is globally hyperbolic.*

Finally, we give an example of a theorem from one of the most important classes of results in global general relativity – the singularity theorems. Many of the global conditions on spacetimes arose originally from the study of singularities. There is a tendency for the metric of a solution of Einstein's equation to become badly behaved. Unfortunately, since the manifold and metric of a spacetime must necessarily be smooth, we cannot express the notion "badly behaved" directly in terms of the behavior of g at points of M. A spacetime is said to be *causally incomplete* if it contains a timelike or null geodesic with just one endpoint and with finite affine length. (Intuitively, we may think of incompleteness as resulting from "actual singular points" having been "removed" from M). The Einstein equation is $R_{ab} = T_{ab} - \frac{1}{2} T^m_m g_{ab}$, where R_{ab} is the Ricci tensor and T_{ab} is a tensor field on M representing the stress and energy distribution of matter. We would not, of course, expect to reach any conclusions about incompleteness without some condition on the curvature. It turns out, fortunately, that the required curvature condition –that $R_{ab}\, \xi^a\, \xi^b$ be positive for every timelike or null vector ξ^a– is, when expressed in terms of T_{ab} via Einstein's equation, a reasonable condition to impose on the matter. That $\left(T_{ab} - \frac{1}{2} T^m_m g_{ab}\right) \xi^a\, \xi^b$ be positive for all timelike or null ξ^a at $p \in M$ means, physically, that "local energy density is positive and not dominated by stresses".

THEOREM 6 (*9*). – *Let M, g be a spacetime in which (i) there are no self-intersecting timelike curves, (ii) there is a compact spacelike section, and (iii) for every non zero timelike or null vector* ξ^a, $R_{ab}\, \xi^a\, \xi^b > 0$. *Then M, g is causally incomplete.*

Note that (iii) requires that there be at least some matter everywhere in M. Condition (ii) is the statement that the spacetime represent a "closed universe". (The singularity theorems in the non-closed case are considerably weaker).

The study of singular spacetimes is far from complete. One would like, for example, to describe singularities locally, to classify them, etc. A first step in such a program would be to obtain a prescription, given an incomplete spacetime M, g, for constructing a space $\bar{M}$ representing M with "additional singular points attached". A very promising approach to this problem has recently been introduced

[1] To simplify the discussion, we have replaced the standard definition of global hyperbolicity by an equivalent one.

by Schmidt (*10*). The 10-dimensional bundle of frames([1]), B, of M can be parallelized by ten vector fields linearly independent at each point. (Six are vertical fields generating infinitesimal rotations of the frames at each point ; the other four are horizontal fields representing parallel transport of the frame along each of its four vectors). The sum of the outer products of these fields defines a positive-definite metric field –and hence a uniform structure– on B. Let $\overline{B}$ denote the completion of B as a uniform space. The natural action of the Lorentz group on B (generating rotations in the fibres) preserves the uniform structure, and so can be extended (uniquely) to an action on $\overline{B}$. Let $\overline{M}$ be the topological space obtained as the quotient of $\overline{B}$ by this action. Each fibre of B is itself an orbit under this Lorentz action, and so we recover M as a dense subspace of $\overline{M}$. The additional points, $\overline{M} - M$, are to represent the "singular points" of the spacetime.

Much work remains to be done to show that this construction is an appropriate and useful one. It is perhaps not overly optimistic, however, to hope that it will lead eventually to a classification of singular points, and then possibly to theorems relating the various types of singular points to other properties of the spacetime.

REFERENCES

[1] Penrose R. — Article in *Battelle Recontres,* C. DeWitt, J.A. Wheeler, eds, (Benjamin, New York, 1968).
[2] Geroch R. — Article in *International School of Physics Enrico Fermi, Course XLVII,* (to be published by Academic Press).
[3] *Seminar on the Bearings of Topology upon General Relativity,* Bern, 1970, (proceedings to be published in J. Relativity and Gravitation).
[4] Geroch R. — *J. Math. Phys.,* 8, 1967, p. 782.
[5] Hawking S.W. — *Proc. Roy. Soc.,* A 308, 1969, p. 433.
[6] Geroch R. — *PhD thesis,* Dept. Physics, Princeton U., 1967.
[7] Geroch R. — *J. Math. Phys.,* 11, 1970, p. 437.
[8] Choquet-Bruhat Y. — Article in *Battelle Recontres,* C. DeWitt, J.A. Wheeler, eds, (Benjamin, New York, 1968).
[9] Hawking S.W., Penrose R. — *Proc. Roy. Soc.,* A 314, 1970, p. 529.
[10] Schmidt B. — *J. Relativity and Gravitation,* (to appear).

The Enrico Fermi Institute
University of Chicago
Chicago, Illinois
60637 U.S.A.

[1] A *frame* at $p \in M$ is an ordered collection of four vectors at p which are mutually orthogonal, and whose norms are -1, $+1$, $+1$, and $+1$, respectively.

Actes, Congrès intern. Math., 1970, Tome 3, p. 47 à 56.

MAGNÉTOHYDRODYNAMIQUE RELATIVISTE ET ONDES DE CHOC

par André LICHNEROWICZ

Introduction.

Au cours des dernières années, la théorie des fluides relativistes, et particulièrement la magnétohydrodynamique, se sont révélées importantes au double point de vue mathématique et physique ; le système différentiel de la magnétohydrodynamique donne l'exemple d'un système suggéré par la physique qui est hyperbolique, sans être strictement hyperbolique. D'autre part, en astrophysique théorique, interviennent, particulièrement dans le système solaire, mais aussi à plus grande échelle, des ondes de choc magnétohydrodynamiques.

Le but principal de cet exposé est l'esquisse de la théorie rigoureuse des ondes de choc en magnétohydrodynamique relativiste. La forme relativiste des conditions de compressibilité d'un fluide joue ici un rôle important. De ces conditions et d'une généralisation convenable de l'équation d'Hugoniot, on peut déduire une étude complète des vitesses des ondes et de la thermodynamique des chocs.

1. Fluide parfait thermodynamique.

a) Soit V_4 un espace-temps donné, muni d'une métrique lorentzienne de signature $+ - - -$. En coordonnées locales $ds^2 = g_{\alpha\beta}\, dx^\alpha\, dx^\beta$ $(\alpha, \beta = 0, 1, 2, 3)$. Dans un domaine de V_4, un fluide parfait est décrit par un tenseur d'énergie

$$T^{(f)}_{\alpha\beta} = (\rho + p)\, u_\alpha\, u_\beta - p\, g_{\alpha\beta}$$

où ρ est la densité propre d'énergie, p la pression, u_α le vecteur-vitesse unitaire orienté vers le futur ; ρ se compose d'une densité de matière et d'une énergie interne. Nous posons :

$$\rho = c^2\, r \left(1 + \frac{\epsilon}{c^2}\right) \qquad (r > 0)$$

où r est la densité de matière du fluide et ϵ son énergie interne spécifique. Considérons les scalaires :

$$\rho + p = c^2\, r \left(1 + \frac{\epsilon}{c^2} + \frac{p}{c^2\, r}\right) \quad , \quad i = \epsilon + \frac{p}{r} = \epsilon + p\, V \qquad \left(\text{où } V = \frac{1}{r}\right)$$

i est l'enthalpie spécifique. A la place de i, j'ai introduit systématiquement *l'indice du fluide* $f = 1 + i/c^2$. Le tenseur d'énergie peut s'écrire :

$$T^{(f)}_{\alpha\beta} = c^2\, r f\, u_\alpha\, u_\beta - p\, g_{\alpha\beta}$$

b) La *température propre* Θ du fluide et son *entropie spécifique* S peuvent être définies, comme en hydrodynamique classique par la relation différentielle

$$\Theta\, dS = d\epsilon + p\,dV = di - V\,dp = c^2 df - V\,dp \qquad (\Theta > 0)$$

Ainsi :

$$c^2\ df = V\,dp + \Theta\ ds \tag{1-1}$$

En relativité, la variable thermodynamique $\tau = fV$ ("volume dynamique") joue un rôle important et se substitue souvent au volume spécifique. Nous considérons $\tau = \tau(p\,,S)$ comme une fonction donnée de p et S qui constitue l'équation d'état du fluide.

Soit Σ une hypersurface régulière, d'équation locale $\varphi = 0$; nous posons $l_a = \partial_a \varphi$. La vitesse de Σ par rapport au fluide est donnée par :

$$\frac{(v^\Sigma)^2}{c^2} = \frac{(u^a\ l_a)^2}{(u^a\ l_a)^2\ - l^a\ l_a}$$

$v^\Sigma < c$ est équivalent à $l^a\ l_a < 0$ (Σ orientée dans le temps). Si v est la *vitesse sonique* du fluide, on a $v^2/c^2 = 1/\gamma$ avec :

$$c^2\ \tau'_p = -\ V^2\ (\gamma - 1)$$

Nous supposons que $\tau(p\,,S)$ vérifie les *conditions de compressibilité* suivantes (extension des conditions classiques dites de Hermann Weyl)

$$(H_1) \qquad \tau'_p < 0 \qquad \tau'_S > 0$$

et

$$(H_2) \qquad \tau''_{p^2} < 0$$

$\tau'_p < 0$ est équivalent à $\gamma > 1$ ou $v < c$

2. Le système de la magnétohydrodynamique relativiste

a) Un champ électromagnétique est défini par deux tenseurs antisymétriques, dont l'un H est le tenseur champ électrique-induction magnétique. Si $^*H$ est le tenseur dual, les vecteurs spatiaux :

$$e_\beta = u^a\ H_{a\beta} \qquad b_\beta = u^a\ (^*H)_{a\beta}$$

sont respectivement le champ électrique et l'induction magnétique par rapport à la direction temporelle u et $u^\beta\ e_\beta = u^\beta\ b_\beta = 0$. Si μ (constante donnée) est la perméabilité magnétique du fluide, $b_\beta = \mu\ h_\beta$ où h est le champ magnétique. Le courant électrique est, en première approximation, la somme de deux termes :

$$J^\beta = \lambda\ u^p\ + \sigma\, e^b$$

où λ est la densité propre de charge électrique et σ la *conductivité* du fluide.

La magnétohydrodynamique est ici l'étude des propriétés d'un fluide parfait de conductivité infinie, $\sigma = \infty$; J et par suite σe étant essentiellement finis, on a nécessairement $e = 0$. Le champ électromagnétique se réduit par rapport au fluide au champ magnétique.

En ajoutant à $T^{(f)}_{\alpha\beta}$ le tenseur d'énergie correspondant, on obtient pour tenseur d'énergie complet :

$$(2\text{-}1)\quad T_{\alpha\beta} = (c^2 r f + \mu |h|^2)\, u_\alpha\, u_\beta - q\, g_{\alpha\beta} - \mu\, h_\alpha\, h_\beta \qquad (\text{où } q = p + \frac{1}{2}\mu\, |h|^2)$$

$|h|^2 = - h^\rho\, h_\rho$ est strictement positif.

b) Le système différentiel de la magnétohydrodynamique est donné par les considérations suivantes ; nous supposons d'abord que la densité de matière (qui correspond àu nombre spécifique de particules) est conservative. Si ∇ est l'opérateur de dérivation covariante :

$$(2\text{-}2)\qquad \nabla_\alpha\, (r\, u^\alpha) = 0$$

Les équations de Maxwell donnent seulement ici :

$$(2\text{-}3)\qquad \nabla_\alpha\, (h^\alpha\, u^\beta - u^\alpha\, h^\beta) = 0$$

et les équations de la dynamique relativiste s'écrivent :

$$(2\text{-}4)\qquad \nabla_\alpha\, T^{\alpha\beta} = 0$$

Ce système entraîne l'équation de flot adiabatique $u^\alpha\, \partial_\alpha\, S = 0$

c) L'analyse de ce système montre que les variétés caractéristiques sont, outre les ondes tangentielles ($u^\alpha\, l_\alpha = 0$), les *ondes magnétosoniques*, solutions de :

$$(2\text{-}5)\quad P(l) \equiv c^2\, r f\, (\gamma - 1)\, (u^\alpha l_\alpha)^4 + (c^2\, r f + \mu\, |h|^2\, \gamma)\, (u^2\, l_\alpha)^2\, l^\beta\, l_\beta - \mu\, (h^\alpha\, l_\alpha)^2\, l^\beta\, l_\beta = 0$$

et les *ondes d'Alfven* solutions de :

$$(2\text{-}6)\qquad D\,(l) \equiv (c^2\, r f + \mu\, |h|^2)\, (u^\alpha\, l_\alpha)^2 - \mu\, (h^2\, l_\alpha)^2 = 0$$

Si $v < c$ (ou $\tau'_p < 0$), on montre que (2-5) et (2-6) définissent des vitesses v^{ML}, v^{MR}, v^A vérifiant les inégalités

$$v^{ML} \leqslant \substack{v \\ v^A} \leqslant v^{MR} < c$$

Si on pose $\beta = \sqrt{c^2\, r f + \mu\, |h|^2 / \mu}$, (2-6) montre que les ondes d'Alfven, sont engendrées par les trajectoires de deux champs de vecteurs :

$$A^\alpha = \beta\, u^\alpha + h^\alpha \qquad B^\alpha = \beta\, u^\alpha - h^\alpha$$

où A et B sont orientés vers le futur.

d) Les cônes d'onde de la magnétohydrodynamique se composent d'un cône du 4ème degré à deux nappes convexes, d'un système de deux hyperplans tangents au cône précédent correspondant aux ondes d'Alfven et de l'hyperplan orthogonal à u (cas général).

Le système de la magnétohydrodynamique est hyperbolique, sans être strictement hyperbolique au sens de Garding-Leray. A partir d'un théorème récent de Leray et Ohia, j'ai pu établir que le problème de Cauchy correspondant admet une solution unique et qu'il y a domaine d'influence.

Les discontinuités infinitésimales d'une solution se propagent le long des bicaractéristiques ou rayons. La direction des rayons d'Alfven est invariante par l'opérateur de discontinuité infinitésimale, alors qu'il n'en est pas de même pour celle des rayons magnétosoniques.

3. Equations générales de choc.

On suppose toujours les $g_{\alpha\beta}$ et leurs dérivées premières continus.

a) Une *onde de choc* est une hypersurface Σ de V_4 telle que u^α, h^α ou l'une des variables thermodynamiques est discontinue à la traversée de Σ. Je montrerai que, sous les conditions de compressibilité $\tau'_p < 0$, $\tau'_S > 0$, Σ est *nécessairement orientée dans le temps.* Si on décompose u^β et h^β selon une composante normale à Σ, on obtient :

$$\text{(3-1)} \quad u^\beta = v^\beta + \frac{u^\alpha\, l_\alpha}{l^\alpha\, l_\alpha}\, l^\beta \quad , \quad h^\beta = t^\beta + \frac{h^\alpha\, l_\alpha}{l^\alpha\, l_\alpha}\, l^\beta \quad (v^\beta\, l_\beta = t^\beta\, l_\beta = 0)$$

Soit Y un état du fluide et du champ en $x \in \Sigma$, défini par les valeurs de p, S, u^β, h^β ; Y_0 est l'état antérieur au choc, Y_1 l'état postérieur ; $[Q]$ est la discontinuité $Q_1 - Q_0$ d'une quantité à travers Σ.

Le système fondamental est satisfait au sens des distributions. On en déduit, par un raisonnement classique, les *équations de choc* :

$$\text{(3-2)} \qquad [r\, u^\alpha]\, l_\alpha = 0 \qquad [h^\alpha\, u^\beta - u^\alpha\, h^\beta]\, l_\alpha = 0 \qquad [T^{\alpha\beta}]\, l_\alpha = 0$$

Nous obtenons l'invariance du scalaire

$$a(Y) = r\, u^\alpha\, l_\alpha$$

l'invariance du vecteur tangent à Σ

$$V^\beta\, (Y) = (h^\alpha\, l_\alpha)\, u^\beta - \frac{a}{r}\, h^\beta$$

et celle du vecteur

$$W^\beta\, (Y) = \left(c^2\, \tau + \mu\, \frac{|h|^2}{r^2}\right)\, a\, r\, u^\beta - q\, l^\beta - \mu\, (h^\alpha\, l_\alpha)\, h^\beta$$

Si $a = 0$ (choc tangentiel), on a $u_0^\alpha\, l_\alpha = u_1^\alpha\, l_\alpha = 0$ et Σ a une vitesse nulle par rapport au fluide avant et après le choc. L'étude de ce cas est triviale et nous supposons $a \neq 0$.

b) Si on décompose W^β tangentiellement et normalement à Σ, on obtient un vecteur tangent et un scalaire e tous deux invariants dans le choc. Ces différents invariants conduisent aux résultats suivants : les deux variables thermodynamiques et les trois scalaires $|h|^2$, $u^\alpha\, l_\alpha$, $h^\alpha\, l_\alpha$ satisfont les cinq relations :

$$r_1\ u_1^2\ l_a = r_0\ u_0^a\ l_a = a \tag{3-3}$$

$$f_1\ u_1^2\ l_a = f_0\ h_0^2\ l_a = b \tag{3-4}$$

$$\frac{(h_1^a\ l_a)^2}{a^2} - \frac{|h_1|^2}{r_1^2} = \frac{(h_0^2\ l_a)^2}{a^2} - \frac{|h_0|^2}{r_0^2} = H \tag{3-5}$$

$$q_1 - \frac{c^2\,a^2}{l^a\ l_a}\,\tau_1 = q_0 - \frac{c^2\ a^2}{l^a\ l_a}\ \tau_0 = e \tag{3-6}$$

$$\chi_1\ \alpha_1^2 = \chi_0\ \alpha_0^2 = L \tag{3-7}$$

où

$$\alpha = c^2\ \tau - \mu H = D\ (l)/a^2 \qquad \chi = |h|^2 + \frac{a^2}{l^a\ l_a} H \tag{3-8}$$

Les quantités (3-8) ont les propriétés suivantes : $\alpha = 0$ exprime que Σ est onde d'Alfven pour l'état Y ; on a $(l^a\ l_a)\,\chi \leqslant 0$ et $\chi = 0$ si et seulement si l est dans le 2-plan $(u\ , h)$; $l^a\ l_a \geqslant 0$ implique $H \leqslant 0$ et par suite $\alpha > 0$.

Les composantes tangentielles de la vitesse et du champ magnétique vérifient

$$(h_1^a\ l_a)\ \nu_1^\beta - (u_1^a\ l_a)\ t_1^\beta = (h_0^a\ l_a)\ \nu_0^\beta - (u_0^a\ l_a)\ t_0^\beta \tag{3-9}$$

$$(c^2\ r_1 f_1 + \mu\ |h_1|^2)\ (u_1^a\ l_a)\ \nu_1^\beta - \mu\ (h_1^a\ l_a)\ t_1^\beta = (c^2\ r_0 f_0 + \mu |h_0|^2)\ (u_0^a\ l_a)\ \nu_0^\beta - \mu\ (h_0^a\ l_a)\ t_0^\beta \tag{3-10}$$

Le déterminant des premiers membres de (3-9), (3-10) est $D_1(l) = a^2\ \alpha_1$. Si $\alpha_1 \neq 0$, (3-9), (3-10) donnent ν_1^β, t_1^β, en fonction de quantités qui seront déduites des cinq équations scalaires.

c) Un *choc d'Alfven* est un choc tel que $\alpha_0 = \alpha_1 = 0$. Un tel choc peut-être de type A ou B, exactement comme une onde d'Alfven. On montre que si $\tau_p' < 0$, les variables thermodynamiques et les scalaires $|h|^2$, $u^a\ l_a$, $h^a\ l_a$ sont invariants au cours du choc. La direction du champ magnétique tangentiel après le choc est indéterminée, mais détermine celle de la vitesse tangentielle d'après la condition suivante : le vecteur A^a(resp. B^a) est invariant par un choc d'Alfven de type A (resp. B).

On établit aussi que *pour qu'une onde de choc* Σ *telle que* $\alpha_0\ \alpha_1 = 0$ *soit compatible avec les ondes d'Alfven, il faut et il suffit que* Σ *définisse un choc d'Alfven* $(\alpha_0 = \alpha_1 = 0)$. Les cas $\alpha_1 = 0\,, \chi_0 = 0\ (\alpha_0 \neq 0)$ et $\alpha_0 = 0\,, \chi_1 = 0\ (\alpha_1 \neq 0)$ sont interdits.

4. Fonction d'Hugoniot relativiste.

a) Un état initial Y_0 étant donné en $x \in \Sigma$, nous considérons dans la suite les états vérifiant les deux conditions :

$$H(Y) = H\ (Y_0) = H \qquad L\ (Y) = L\ (Y_0) = L \tag{4-1}$$

si bien que :

$$\chi = \frac{L}{(c^2\ \tau - \mu H)^2} \qquad \left(\tau \neq \frac{\mu H}{c^2}\right)$$

Nous substituons à q la variable :

$$\overline{q} = p + \frac{1}{2}\mu\chi = q + \frac{1}{2}\mu\,\frac{a^2}{l^a\ l_a}H$$

et à (3-6) la relation :

$$\overline{q}_1 - \overline{q}_0 = \frac{c^2\,\alpha^2}{l^a\ l_a}(\tau_1 - \tau_0) \tag{4-2}$$

Sous les conditions (4-1), un état thermodynamique $(\tau\ , p)$ du fluide définit (pour $\tau \neq \mu H/c^2$) un point Z du plan $(\tau\ , q)$. Entre $\tau\ , S$ et q, on a la relation :

$$q = p\ (\tau, S) + \frac{1}{2}\mu\,\frac{L}{(c^2\ \tau - \mu H)^2} \tag{4-3}$$

où $p = p\ (\tau\ , S)$ se déduit par inversion de $\tau = \tau\,(p\ , S)$

b) Introduisons la *fonction d'Hugoniot* $\mathcal{H}\,(Z_0, Z)$ de Z, pour un point initial donné Z_0 :

$$\mathcal{H}\ (Z_0,\ Z) = c^2\ (f^2 - f_0^2) - (\tau + \tau_0)\ (p - p_0) + (\tau - \tau_0)\frac{1}{2}\mu\left(\chi + \chi_0 - 2\,\frac{\chi_0\ \alpha_0}{\alpha}\right)$$

On a $\mathcal{H}\ (Z_0, Z_0) = 0$ et on peut montrer que l'on peut substituer à (3-4) l'équation d'Hugoniot $\mathcal{H}\,(Z_0, Z_0) = 0$. En différentiant $\mathcal{H}\,(Z_0, Z)$, on obtient d'après (1-1)

$$d\mathcal{H} = 2\,f\,\Theta\ dS + (\tau - \tau_0)\ dq - (q - q_0)\ d\tau \tag{4-4}$$

En différentiant $\mathcal{H}$ le long d'une droite Δ du plan $(\tau\ , \overline{q})$, il vient :

$$d\,\mathcal{H} = 2\,f\,\Theta\ dS \tag{4-5}$$

De plus, si Δ est la droite (Z_0, Z_1), on obtient par un calcul direct :

$$\tau'_S\ \alpha\,dS = \frac{P\ (l)}{a^2\ \ l^a\ l_a}d\tau = \frac{P\ (l)}{c^2\ \ a^4}d\overline{q} \tag{4-6}$$

5. Orientation dans le temps des ondes de choc.

On montre aisément que, sous les conditions (H_1) , $l^a\ l_a \geqslant 0$ implique $P(l) > 0$. De plus α est > 0 et d'après (4-6), dS/dq est > 0 le long de $(Z_0\ , Z_1)$. Mais comme :

$$\mathcal{H}\ (Z_0\ , Z_0) = \mathcal{H}\ (Z_0\ , Z_1) = 0$$

la fonction $\mathcal{H}\ (Z_0\ , Z)$ est stationnaire en un point au moins de $(Z_0\ ,Z_1)$ et il en est de même pour S d'après (4-5). Ainsi on a nécessairement $l^a\ l_a < 0$.

THEOREME. – *Si le fluide satisfait les hypothèses* (H_1), *toute onde de choc est nécessairement orientée dans le temps. Si* v_0^Σ *et* v_1^Σ *sont les vitesses de* Σ *par rapport au fluide avant et après le choc, on a* $v_0^\Sigma < c$, $v_1^\Sigma < c$.

6. Thermodynamique des chocs.

On sait que $\chi = k^2$ est positif et (3-7) peut s'écrire $k_1 \alpha_1 = k_0 \alpha_0$. L'équation d'Hugoniot devient :

$$\mathcal{H}(Z_0, Z_1) = c^2 (f_1^2 - f_0^2) - (\tau_0 + \tau_1)(p_1 - p_0) + (\tau_1 - \tau_0) \frac{1}{2} \mu (k_1 - k_0)^2 = 0 \tag{6-1}$$

D'après (1-1) :

$$c^2 f'_p = V > 0 \qquad c^2 f'_S = \Theta > 0$$

et en dérivant :

$$\frac{\partial}{\partial p} (c^2 f^2) = 2\tau \tag{6-2}$$

On suppose naturellement qu'en chaque point $x \in \Sigma$, on a $S_0 \leqslant S_1$. On déduit des conditions de compressibilité que $S_1 = S_0$ entraîne qu'on a en x un choc d'Alfven. De (6-1), (6-2) on déduit

THEOREME. —*Pour un choc effectif qui n'est pas choc d'Alfven, on a sous les hypothèses de compressibilité* (H_1), (H_2)

$$S_1 > S_0 \qquad p_1 > p_0 \qquad f_1 > f_0 \qquad \tau_1 < \tau_0 \tag{6-3}$$

Supposons $S_1 > S_0$; nous allons établir $p_1 > p_0$. Si p_1 était $\leqslant p_0$, on a par intégration de (6-2) pour $S = S_0$:

$$c^2 \{f^2(p_0, S_0) - f^2(p_1, S_0)\} = 2 \int_{p_1}^{p_0} \tau(p, S_0)\, dp \leqslant (p_0 - p_1) (\tau(p_0, S_0) + \tau(p_1, S_0))$$

d'après (H_2). On en déduit a fortiori :

$$c^2 (f_1^2 - f_0^2) - (\tau_0 + \tau_1)(p_1 - p_0) > 0$$

De (6-1) on déduit $\tau_1 < \tau_0$ ce qui est en contradiction avec $p_1 \leqslant p_0$, $S_1 > S_0$ et (H_1). Les autres démonstrations sont analogues.

Ainsi $\alpha_1 < \alpha_0$. On peut montrer qu'une onde de choc qui n'est pas choc d'Alfven n'est compatible avec les ondes d'Alfven que si $\alpha_0 \alpha_1 > 0$. On obtient ainsi deux types de chocs : les *chocs lents* pour lesquels $\alpha_1 < \alpha_0 < 0$, les *chocs rapides* pour lesquels $0 < \alpha_1 < \alpha_0$

7. Courbes isentropiques et vitesses des ondes de choc.

a) Un état initial Y_0 étant donné, nous considérons l'ensemble des états Y vérifiant :

$$H(Y) = H(Y_0) = H \qquad k\alpha = k_0\,\alpha_0 \qquad (\text{avec } \alpha\,\alpha_0 > 0)$$

Dans le plan $(\tau\,, q)$, la droite $\tau = \mu H/c^2$ est interdite. Dans ce plan, la courbe isentropique correspondant à la valeur S de l'entropie est défini par :

$$q = p\,(\tau\,, S) + \frac{1}{2}\,\mu\,\frac{k_0^2\,\alpha_0^2}{(c^2\,\tau - \mu H)^2} \tag{7-1}$$

En dérivant le long de $\mathfrak{S}$, on a :

$$\left(\frac{d^2\,q}{d\,\tau^2}\right)_{\mathfrak{S}} = -\frac{1}{\tau'^3_p}\,M \tag{7-2}$$

où

$$M = \tau''_{p^2} - 3\,\mu\,c^4\,\tau'^3_p\,\frac{k^2}{\alpha^2} > 0 \qquad (\text{sous } H_1\,, H_2)$$

et la courbe $\mathfrak{S}$ est convexe.

b) Soit Δ une droite $\overline{q} - \overline{q}_0 = m\,(\tau - \tau_0)$ issue de Z_0.

Lemme. – *Sous les hypothèses (H_1), (H_2) on a en tout point Z_s de Δ où S est stationnaire*

$$\left(\frac{d^2\,S}{d\,\tau^2}\right)_{\Delta} < 0$$

En dérivant deux fois (7-1) le long de Δ, on a en effet :

$$\left(\frac{d^2\,S}{d\,\tau^2}\right)_{\Delta} = \left(\frac{\tau'_p}{\tau'_s}\cdot\frac{d^2\,q}{d\tau^2}\right)_{\mathfrak{S}} = -\left(\frac{1}{\tau'^2_p\,\tau'_s}\,M\right)_{Z_s} \quad < 0$$

Considérons un choc $Z_0 \to Z_1$. Il résulte du lemme que le point Z_s de la droite $(Z_0\,, Z_1)$ où $S(\tau)$ est stationnaire est unique et correspond à un maximum pour S. De (4-6) on déduit :

$$\left(\frac{dS}{d\tau}\right)_0 \Rightarrow \alpha_0\,P(l)_0 > 0 \qquad \left(\frac{dS}{d\tau}\right) > 0 \Rightarrow \alpha_1\,P(l)_1 < 0$$

En interprétant les signes de α et $P(l)$, on obtient :

Theoreme — *Sous les hypothèses (H_1), (H_2), les vitesses v_0^{Σ} et v_1^{Σ}, d'une onde Σ vérifient les inégalités :*

1) *pour un choc rapide*

$$v_0^{ML} < v_0^{A} < v_0^{MR} < v_0^{\Sigma} \qquad v_1^{ML} < v_1^{A} < v_1^{\Sigma} < v_1^{Ml}$$

2) *pour un choc lent :*

$$v_0^{ML} < v_0^{\Sigma} < v_0^{A} < v_0^{MR} \qquad v_1^{\Sigma} < v_1^{ML} < v_1^{A} < v_1^{MR}$$

c) Considérons, dans le plan $(\tau, \overline{q})$, la courbe d'Hugoniot $\mathcal{H}$ définie par $\mathcal{H}(Z_0, Z) = 0$ et la courbe isentropique $\mathcal{S}_0$ correspondant à $S = S_0$. On établit

THEOREME. — *$\mathcal{H}$ et $\mathcal{S}_0$ admettent un contact du second ordre en Z_0. Pour un choc faible $(S_1 - S_0)$ est du troisième ordre par rapport à la puissance $(p_1 - p_0)$ du choc :*

$$S_1 - S_0 = \left(\frac{M}{12 f \Theta}\right)_{Z_0} (p_1 - p_0)^3 + \ldots$$

8. Existence de solutions pour les équations de choc.

En inversant $\tau = \tau(p, S)$, on obtient $S = S(p, \tau)$. *On suppose $S(p, \tau)$ telle que les hypothèses (H_1), (H_2) soient satisfaites pour $\tau \leqslant \tau_0$ et des valeurs arbitrairement grandes de p.* Je veux prouver que si $v_0^A < v_0^{MR} < v_0$ (resp. $v_0^{ML} < v_0 < v_1^A$), les équations de choc admettent une solution non triviale unique telle que $v_1^A < v_1^{\Sigma}$ (resp $v_1^{\Sigma} < v_1^{\Sigma} < v_1^A$)

a) Considérons la branche de $\mathcal{S}_0$ telle que $\tau \leqslant \tau_0$ et $\alpha\alpha_0 > 0$. Le long de $_0$

$$\left(\frac{d\mathcal{H}}{d\tau}\right)_{\bullet} = (\tau - \tau_0)\left(\frac{d\overline{q}}{d\tau}\right) - (\overline{q} - \overline{q}_0)$$

et d'après (7-2) :

$$\left(\frac{d^2\mathcal{H}}{d\tau^2}\right) = (\tau - \tau_0)\left(\frac{d^2\overline{q}}{d\tau^2}\right) < 0$$

Par suite $(d\mathcal{H}/d\tau)$ est > 0 et $\mathcal{H}(Z_0, Z) < 0$ sur la branche considérée.

b) Considérons une droite Δ de pente m issue de Z_0. On a :

$$\left(\frac{d\mathcal{H}}{d\tau}\right)_{\Delta} = 2 f \Theta \left(\frac{dS}{dt}\right)_{\Delta} \tag{8-1}$$

Soit m_0 la pente en Z_0 de $\mathcal{S}_0$. De l'hypothèse générale faite et de la convexité de $\mathcal{S}_0$, il résulte que *si $m < m_0$*, Δ rencontre la branche considérée de $\mathcal{S}_0$ en un point Z_A et un seul et $\mathcal{H}(Z_0, Z_A) < 0$. Mais, le long de Δ, $S(Z_0) = S(Z_A) = S_0$' et $S(\tau)$ est nécessairement stationnaire entre Z_0 et Z_A en un point Z_s (unique) qui est, d'après le lemme, un maximum pour S sur Δ. On a $(dS/d\tau)_{\Delta} < 0$ en Z_0 et d'après (8-1), $(d\mathcal{H}/d\tau)_{\Delta} < 0$ en Z_0. Il existe donc au moins un point Z_0. Il existe donc au moins un point Z_1, de Δ entre Z_0 et Z_A tel que $\mathcal{H}(Z_0, Z_1) = 0$.

Sur Δ, S croit de Z_0 à Z_s, passe par un maximum en Z_s puis décroit constamment. D'après (8-1), il en est de même pour $\mathcal{H}$ et le point Z_1 est unique *A tout nombre $m < m_0$ correspond un point unique Z_1 du contour d'Hugoniot de Z_0 (avec $\tau_1 < \tau_0$ et $\alpha_1\alpha_0 > 0$) tel que $q_1 - q_0/\tau_1 - \tau_0 = m$.* Pour

$m = c^2 a^2/l^\alpha l_\alpha < m_0, Z_1$ satisfait (4-2) et $\mathcal{H}(Z_0, Z_1) = 0$. De la connaissance de Z_1, on déduit immédiatement l'existence d'une solution non triviale unique telle que $\alpha_1 \alpha_0 > 0$ des équations de choc ; l'inégalité $c^2 a^2/l^\alpha l_\alpha < m_0$ est équivalente à $\alpha_0 P(l)_0 > 0$. Il vient :

THEOREME.i – *Si $S(p, \tau)$ vérifie (H_1), (H_2) pour $\tau \leqslant \tau_0$ et des valeurs arbitrairement grandes de p, pour chaque état Y vérifiant $\alpha_0 P(l)_0 > 0$, les équations générales de choc admettent une solution non triviale unique telle que $\alpha_0 \alpha_1 > 0$.*

BIBLIOGRAPHIE

[1] CHOQUET-BRUHAT Y. — *Astron. Acta* 6, 1960, p. 354-365.

[2] HOFFMANN F. and TELLER E. — *Phys. Rev.* 80, 1950, p. 692-702.

[3] ISRAEL W. — *Proc. Roy. Soc.* A 259, 1960, p. 129-143.

[4] LICHNEROWICZ A. — *Relativistic hydrodynamics and magneto hydrodynamics New-York,* Benjamin, 1967.

[5] LICHNEROWICZ A. — *Comm. Math. Phys.* 12, 1969, p. 145-174 et *C.R. Acad. Sc. Paris,* 268, 1969, p. 256-260.

[6] LUCQUIAND J.C. — *C.R. Acad. Sc. Paris,* 270, 1970, p. 85-88.

Collège de France
11, Place Marcelin-Berthelot,
Paris 5ème (France)

Actes, Congrès intern. Math., 1970. Tome 3, p. 57 à 63.

E3 - PROBLÈMES MATHÉMATIQUES DE LA MÉCANIQUE DU CONTINU

ANOTHER USEFUL SINGULAR PERTURBATION DEVICE

by G.F. CARRIER

1. Introduction.

Frequently, in the course of an engineering or scientific investigation, one finds it convenient to describe the solution of a boundary value problem in terms of a conventional expansion in a parameter ϵ. Sometimes, however, such a conventional expansion (e.g. a convergent series, an asymptotic series, a multiscaling procedure, or matched asymptotic expansions) either fails to be applicable or provides an unnecessarily cumbersome description. In such cases *ad hoc* methods occasionally salvage the situation. Here we present an exceptionally simple device which has been useful to the author and we indicate the manner in which that device can be formalized so that the procedure falls within the framework of established methods. It seems easiest to introduce the idea via an illustrative example which was constructed for that purpose.

Suppose that one is required to find $u(x, \epsilon)$ such that

$$(1) \qquad \frac{\partial}{\partial x} u(x, \epsilon) = \frac{1}{\sqrt{x + \epsilon(u + a)}} \qquad \text{in} \quad 0 < x < 1$$

with $0 < \epsilon \ll 1$, $0 < a = O(1)$, and with

$$(2) \qquad u(\mathrm{i}) = 2.$$

One could initiate a perturbation procedure by writing

$$u(x, \epsilon) = u_0(x) + \epsilon u_1(x) + \dots$$

and expanding the square root in a power series in ϵ. Unfortunately, there is no conventional expansion of the square root in powers of ϵ which is useful in all of $0 < x < 1$. However, it is easily argued that u will be positive on the interval and it can be anticipated that $u(0, \epsilon)$ will be of order $\epsilon^{1/2}$. Accordingly, we expect that there is a small positive number γ such that $[x + \epsilon(u + a)]^{1/2}$ can be accurately approximated by $\sqrt{x + \gamma}$. With that approximation,

$$u_x(x, \epsilon) \simeq \frac{1}{\sqrt{x + \gamma}}$$

and

$$u(x\,,\epsilon) \simeq 2\,[\sqrt{x+\gamma} - \sqrt{1+\gamma} + 1]\,.$$

Thus

$$u(0\,,\epsilon) \simeq 2\left[\sqrt{\gamma} - \frac{\gamma}{2} + O(\gamma^2)\right].$$

If no refinements are to be made, γ must be given by

$$\gamma = \epsilon\,(a + u\,(0\,,\epsilon))$$

so that

$$\gamma = \epsilon a + O\,(\epsilon^{3/2}a^{1/2})$$

or, more precisely,

$$\gamma = \epsilon\,(a + 2\,[\sqrt{\gamma} - \sqrt{1+\gamma} + 1]).$$

It is a little difficult to assess the accuracy of this approximate result. Furthermore, a mere successive approximation continuation of the process becomes very clumsy. Accordingly, it is worth while to formalize the procedures in such a way as to eliminate these disadvantages.

Thus, we define a new problem

$$w_x(x\,,\epsilon\,,\alpha) = \frac{1}{\sqrt{x + \alpha m + \epsilon\,(w - \beta)}} \qquad \text{in} \quad 0 < x < 1 \tag{3}$$

with

$$w\,(1\,,\epsilon\,,\alpha) = 2. \tag{4}$$

The parameters β and $m = \beta + a$ can be chosen later but the simplest and most obvious choise is $m = a$. Note that, in the degenerate case $\alpha = \epsilon$

$$w\,(x\,,\epsilon\,,\epsilon) \equiv u\,(x\,,\epsilon)$$

where u is defined by Eqs. (1) and (2).

We can now write

$$w\,(x\,,\epsilon\,,\alpha) = w_0(x\,,\alpha) + \epsilon\, w_1(x\,,\alpha) + \ldots$$

so that

$$w_{0\,,x} = \frac{1}{\sqrt{x + \alpha m}} \tag{5}$$

$$w_{1\,,x} = \frac{w_0 - \beta}{2\,(x + \alpha m)^{3/2}} \tag{6}$$

$$w_{2\,,x} = \ldots$$

It now follows that

$$(7) \qquad w_0(x, \alpha) = 2[\sqrt{x + \alpha m} - \sqrt{1 + \alpha m} + 1] = 2\sqrt{x + \alpha m} - k$$

$$(8) \qquad w_1(x, \alpha) = - \left[\ln \frac{x + \alpha m}{1 + \alpha m} - \frac{k + \beta}{\sqrt{x + \alpha m}} + \frac{k + \beta}{\sqrt{1 + \alpha m}}\right] \cdot$$

With the choice, $m = a$,

$$w_0(0, \alpha) = 2\sqrt{\alpha a} - k$$

and

$$w_1(0, \alpha) = - [\ln \alpha a + O\sqrt{\alpha}]$$

so that $w(0, \epsilon, \alpha) = 2\sqrt{\alpha a} - k - \epsilon \ln(\alpha a) + O(\epsilon\sqrt{\alpha})$
and, of course,

$$(9) \qquad w(0, \epsilon, \epsilon) = 2\sqrt{\epsilon a} - \epsilon a - \epsilon \ln \epsilon a + O(\epsilon^{3/2}).$$

It is easily verified that the inclusion of higher order terms cannot destroy the truth of Eq. (9).

There are several exercises which the reader may find interesting ; he can compare (1) the accuracy of w_0 as an approximation to $u(x, \epsilon)$ when m is so chosen that $w_1(0, \alpha) = 0$, and (2) the accuracy of w_0 when $m = a$; but he should also note the labor required to carry out each calculation. He might also find it interesting to compare the description of Eq. (7) and (8) with the exact solution

$$(10) \qquad (\phi - \epsilon)^2 + 2\epsilon^2 \ln(\phi + \epsilon) = x + \lambda$$

where

$$\phi = \sqrt{x + \epsilon(u + a)}$$

and

$$\lambda = (\sqrt{1 + \epsilon[u(1, \epsilon) + a]} - \epsilon)^2 .$$

He should verify easily $\left(\text{by trying } u' = x^n + \frac{1}{\sqrt{x + \epsilon(u + a)}} \text{ and } u(1) = 2 + n^{-1}\right)$ that the process can work even when a closed form solution is not available. The most fascinating exercise, however, is that which leads to the conclusion that the case $a = 0$ is too delicate for this procedure to provide a convenient result.

Perhaps the most useful and noteworthy aspect of the result is this : Each term of the series obtained, i.e. Eqs. (7), (8), . . . , differs from its predecessor by a factor, ϵ, in its coefficient ; all of the $\sqrt{\epsilon}$ and $\ln \epsilon$ variations are carried in the $w_n(x, \epsilon)$. This is not advantageous in sorting out asymptotic ordering, but it is *very* advantageous in the efficient calculation of, for example, $u(0, \epsilon)$.

We can now summarize the foregoing in a more abstract form. Suppose that

$$\mathscr{L}(u, \epsilon) = 0$$

is such that the series

$$u(x\,,\epsilon) = u_0(x) + \epsilon\, u_1(x) + \dots$$

fails to be useful. It then is advantageous to seek an operator $\mathcal{M}$ such that

$$\mathcal{M}(w\,,\epsilon\,,\alpha) = 0\,,$$

such that

$$w = \Sigma\, \epsilon^n\, w_n(x\,,\alpha)$$

is useful for a range of α including $\alpha = \epsilon$, and such that

$$\mathcal{M}(w\,,\epsilon\,,\epsilon) \equiv \mathcal{L}(w\,,\epsilon)\,.$$

2. A Rotating Fluid Problem.

There is another problem whose concise discussion may be informative. A typical, low Rossby number investigation in the dynamics of rotating fluids requires that we find ν and ψ such that

$$(1) \qquad \epsilon\,\Delta\,\Delta\,\psi = -\,\nu_z$$

$$(2) \qquad \epsilon\,\Delta\,\nu = \psi_z$$

in

$$-\,1 < z < 1 \quad \text{and} \quad -\,\infty < x < \infty$$

with

$$\psi\,(x\,,\pm\,1\,,\epsilon) = \psi_z(x\,,\pm\,1\,,\epsilon) = 0\,,$$

$$(3) \qquad \nu(x\,,1\,,\epsilon) = V_+(x\,,\epsilon)\quad,\quad \nu(x\,,-\,1\,.\,\epsilon) = V_-(x\,,\epsilon)$$

and with $0 < \epsilon \ll 1$; in these equations Δ denotes $\dfrac{\partial^2}{\partial x^2} + \dfrac{\partial^2}{\partial z^2}$.

It is well known that one can use boundary layer ideas (matched asymptotic expansions) to construct the solution of Eqs. (1), (2) and (3). An informal process provides the description

$$(4) \qquad \psi \simeq \epsilon^{1/2}\,[\Psi\,(\eta\,,z\,,\epsilon) + \psi^{(1)}\,(\eta\,,\xi_1\,,\epsilon) + \psi^{(2)}\,(\eta\,,\xi_2\,,\epsilon)]$$

$$(5) \qquad \nu \simeq [\,V(\eta\,,z\,,\epsilon) + \nu^{(1)}\,(\eta\,,\xi_1\,,\epsilon) + \nu^{(2)}\,(\eta\,,\xi_2\,,\epsilon)]$$

where

$$(6) \qquad \eta = \epsilon^{-1/3}\,x \quad,\quad \xi_1 = (z - 1)\,\epsilon^{-1/2} \quad \text{and} \quad \xi_2 = (z + 1)\,\epsilon^{-1/2}\,.$$

The differential equations which Ψ, $\psi^{(1)}$, ... must obey are

$$(7) \qquad \left(\frac{\partial^2}{\partial \eta^2} + \epsilon^{2/3}\,\frac{\partial^2}{\partial z^2}\right)\quad V = \epsilon^{1/6}\,\Psi_z\,,$$

$$(8) \qquad \left(\frac{\partial^2}{\partial \eta^2} + \epsilon^{2/3}\,\frac{\partial^2}{\partial z^2}\right)^2\quad \Psi = -\,\epsilon^{-1/6}\,V_z\ ;$$

and

$$(9) \qquad \left(\frac{\partial^2}{\partial \xi^2} + \epsilon^{1/3} \frac{\partial^2}{\partial \eta^2}\right) \nu^{(j)} = \psi^{(j)}_{\xi_j}$$

$$(10) \qquad \left(\frac{\partial^2}{\partial \xi_j^2} + \epsilon^{1/3} \frac{\partial^2}{\partial \eta^2}\right)^2 \psi^{(j)} = -\nu^j_{\xi_j}$$

for $j = 1$ and for $j = 2$.

The latter equations imply that

$$(11) \qquad \nu^{(1)} \simeq A_1 \, e^{\sqrt{i\xi_1}} + B_1 \, e^{\sqrt{-i\xi_1}}$$

$$(12) \qquad \psi^{(1)} \simeq \sqrt{iA_1} \; e^{\xi_1\sqrt{i}} + B_1 \sqrt{-i \, e^{\xi_1\sqrt{-i}}}$$

and there is a similar description for $\nu^{(2)}$, $\psi^{(2)}$. When these are substituted into Eq. (3) and when the A_j, B_j are eliminated, we obtain

$$(13) \qquad V(\eta, 1, \epsilon) \cdot\cdot \sqrt{2}\, \Psi(\eta, 1, \epsilon) + \sqrt{\epsilon}\, \Psi_z(\eta, 1, \epsilon) = V_+$$

$$(14) \qquad V(\eta, -1, \epsilon) + \sqrt{2}\, \Psi(\eta, -1, \epsilon) + \sqrt{\epsilon}\, \Psi_z(\eta, -1, \epsilon) = V_- \cdot$$

It is readily argued (correctly) that the Ψ_z terms are of higher order, mathematically, and, *more importantly,* that they do not represent a contribution of primary importance to the physical balance. Thus, we ignore those terms in Eqs. (13) and (14).

It happens to be true that, as Eq. (13) and (14) seem to imply, the major features of the structure of the flow configuration require contributions from both V and Ψ. The difficulty is that Eqs. (7) and (8) seem to imply that $\Psi = O(\epsilon^{-1/6} V)$ or that Ψ is unimportant in the balance. In an informal procedure it is easy to ignore the apparent inconsistency of Eqs. (7) and (8) with (13) and (14) (less the $\sqrt{\epsilon}$ terms) and to use

$$(15) \qquad V_{\eta\eta} = \epsilon^{1/6} \Psi_z$$

$$(16) \qquad \Psi_{\eta\eta\eta\eta} = -\epsilon^{-1/6} V_z$$

together with

$$(17) \qquad V(\eta, 1, \epsilon) - \sqrt{2}\, \Psi(\eta, 1, \epsilon) = V_+$$

$$(18) \qquad V(\eta, -1, \epsilon) + \sqrt{2}\, \Psi(\eta, -1, \epsilon) = V_-$$

to find V and Ψ (as has been done in several places including [1]).

The V and Ψ so obtained provide an excellent description of the phenomenon but V and Ψ each depend on ϵ in such a way that the continuation of the process to obtain perturbation series for Ψ, $\psi^{(1)}$, etc., in powers of ϵ would be difficult.

Alternatively, one can distinguish terms of order $\epsilon^{1/6}$ from terms of order unity even though ordinarily, $\epsilon^{1/6}$ will not be particularly small. When this is done the same flow field can be inferred but the series is more cumbersome and the primary structure is not entirely in the terms of order unity. This is implicit in the approach used in [2].

It is our intent, here, to indicate how we can keep the convenience of the informal analysis within a framework which permits error estimates and evidence of internal consistency.

We particularly want to do this in such a way that the power of ϵ which multiply successive terms of the expansions have no larger a ratio than $\epsilon^{1/3}$.

The choices which follow are dictated directly by the criteria ; (1) the first term in the expansion of Ψ and the first term in the expansion of V must be of the same order in ϵ in equations which replace (7) and (8) ; (2) no powers of ϵ other than $n/3$ should appear in any equations. These criteria are met by writing :

$$(19)\quad \psi = \epsilon^{1/3}\, \alpha^{1/6}\, \Psi(\eta\,, z\,, \epsilon\,, \alpha) + \epsilon^{1/2}\, [\phi^{(1)}(\eta\,.\, \xi_1\,, \epsilon\,, \alpha) + \phi^{(2)}(\eta\,, \xi_2\,, \epsilon\,, \alpha)],$$

$$(20)\quad V = W(\eta\,, z\,, \epsilon\,, \alpha) + w^{(1)}(\eta\,, \xi_1\,, \epsilon\,, \alpha) + w^{(2)}(\eta\,, \xi_2\,, \epsilon\,, \alpha)$$

where each of the functions on the right side of these equations will be expanded in conventional series in powers of ϵ.

Replacing $v^{(j)}$ by $w^{(j)}$ and $\psi^{(j)}$ by $\phi^{(j)}$, Eqs. (9) and (10) still apply, but Eqs. (7) and (8) are replaced by

$$(21)\qquad \left(\frac{\partial^2}{\partial\eta^2} + \epsilon^{2/3}\, \frac{\partial^2}{\partial z^2}\right) W = \alpha^{1/6}\, \Psi_z$$

$$(22)\qquad \left(\frac{\partial^2}{\partial\eta^2} + \epsilon^{2/3}\, \frac{\partial^2}{\partial z^2}\right)^2 \Psi = -\,\alpha^{-1/6}\, W_z\,.$$

Eqs. (3) now imply (with $V_+(x) \equiv W_+(\eta)\,, \ldots$)

$$(23)\qquad W(x\,, \pm\, 1\,, \epsilon\,, \alpha) + w^{(j)}(x\,, 0\,, \epsilon\,, \alpha) = W_\pm(\eta)$$

$$(24)\qquad \Psi(x\,, \pm\, 1\,, \epsilon\,, \alpha) + \psi^{(j)}(x\,, 0\,, \epsilon\,, \alpha) = 0$$

$$(25)\qquad \epsilon^{1/3}\, \alpha^{1/6}\, \Psi_z(x\,, \pm\, 1\,, \epsilon\,, \alpha) + \psi^{(j)}_{\xi_j}(x\,, 0\,, \epsilon\,, \alpha) = 0$$

where $j = 1$ when $z = 1$ and $j = 2$ when $z = -\,1$.

It is apparent that Eqs. (9), (10) and (21) to (25) define functions W, Ψ, $w^{(j)}$, $\phi^{(j)}$ which can be expanded in the conventional framework of the matched asymptotic expansion methods, and that, for example,

$$W = \sum_0^\infty \epsilon^{n/3}\, W_n(\eta\,, z\,, \alpha)\,.$$

Furthermore, when W is evaluated at $\alpha = \epsilon$, the zero^th order contributions are those implied by Eqs. (15) through (18). It is not as obvious that the range of validity of the expansion includes $\alpha = \epsilon$, but such is the case.

Another difficulty in connection with such expansions arises, of course, when

$W_{\pm}(\eta)$ has a scale in x as small as $\epsilon^{1/2}$ but that is a different kind of difficulty to which the foregoing is irrelevant.

Once again the abstract summary of the substance of this section is given by the last paragraph of Section 1.

This study was supported by the National Science Foundation under Contract N00014-67-A-0298-0002.

REFERENCES

[1] CARRIER G.F. — Phenomena in Rotating Fluids, *Proceedings, 11th Internat'l Congress of Applied Mechanics,* Munich, Germany, 1964.

[2] PEDLOSKY J. — Geophysical Fluid Dynamics, to be published in the *Proceedings of the American Mathematical Society's Summer,* 1970, Symposium on Mathematical Problems in Geophysics.

Harvard University
Division of Engineering and Applied Physics,
Pierce Hall,
Cambridge,
Massachusetts 02 138 (USA)

Actes, Congrès intern. Math., 1970. Tome 3, p. 65 à 69.

THE FLOW OF FLUIDS IN POROUS MEDIA

by Jim DOUGLAS Jr

1. Introduction.

This lecture will outline the developments over the past two years of Todd Dupont and the author on problems of flow of fluids in porous media arising from petroleum reservoir engineering. Two specific problems will be discussed and a third mentioned. The first is the single-phase flow of a non-ideal gas. This flow can be described by a nonlinear parabolic equation. The second is the two-phase flow of water and oil. If the fluids are assumed incompressible (this is the more difficult situation mathematically) and immiscible, a degenerate nonlinear parabolic system in two dependent variables results. The time derivative of only one appears explicitly. The third problem comes from a so-called "beta factor" description of three-phase (oil, gas, water) flow. Again a very strongly nonlinear parabolic system is generated.

Galerkin methods are used both to treat the theoretical questions of existence, uniqueness, and continuous dependence and the practical problem of approximating the solutions for the single-phase and two-phase problems. Approximate solution by Galerkin procedures is discussed for the three-phase problem, but the other issues remain untouched in this case.

2. Single-Phase Flow.

Let us consider the flow of gas in a porous medium [3, 4]. For simplicity, consider the domain to be horizontal and radially symmetric so that a single space variable can be used to describe the geometry. Then the flow can be described by the parabolic equation

$$\frac{\partial}{\partial t}(2\pi r\phi\rho) = \frac{\partial}{\partial r}\left(2\pi r\frac{k\rho}{\mu}\,\frac{\partial p}{\partial r}\right), \tag{2.1}$$

where

$$0 < r_{\min} \leqslant r < r_{\max} < \infty , \tag{2.2}$$

and $\phi = \phi_1(r)\,\phi_2(p)$ is the porosity of the medium, ρ is the density of the gas, $k = k(r)$ is the permeability of the medium, $\mu = \mu(p)$ is the gas viscosity, and p is the gas pressure. Assume that

$$\zeta = \zeta(p) = \frac{p}{\rho} \tag{2.3}$$

is the equation of state (assuming the temperature fixed). It is convenient to make the following change of variables :

$$q = q(p) = \int_0^p \frac{\tau \, d\tau}{\zeta(\tau)\,\mu(\tau)} \, . \tag{2.4}$$

Let

$$c(r) = 2\pi r \phi_1(r) \quad , \quad a(r) = 2\pi r k(r) \tag{2.5}$$

and

$$f(q) = \frac{p\,\phi_2(p)}{\zeta(p)} \quad , \quad f'(q) = \mu \left\{ \phi_2 \left(\frac{1}{p} - \frac{\zeta'}{\zeta} \right) + \phi_2' \right\} \, . \tag{2.6}$$

Then the differential equation reduces to the form

$$c(r) f'(q) \, \frac{\partial q}{\partial t} = \frac{\partial}{\partial r} \left(a(r) \, \frac{\partial q}{\partial r} \right) . \tag{2.7}$$

Initial values for p and, hence, for q need to be specified. Either p or the flow, $a \partial q / \partial r$, should be given for $0 < t \leqslant T$ at the boundaries.

Existence, uniqueness, and continuous dependence of the solution on the data can be treated (for arbitrary dimension) quite easily by exactly the same methods as will be discussed in the next section. Let us consider the problem of approximating the solution by a Galerkin method. Assume that flows are specified and that $\mathfrak{M} \subset H^1((r_{\min}, r_{\max}))$. If $\{v_1, \ldots, v_N\}$ is a basis for $\mathfrak{M}$ and

$$\tilde{q}(r, t) = \sum_{j=1}^{N} \eta_j(t) \, v_j(r) \tag{2.8}$$

is the approximate solution, the Galerkin method for determining $\eta(t)$ is given by the system

$$C(\eta(t)) \, \frac{d\eta(t)}{dt} + A\eta(t) = \gamma(t) \, , \tag{2.9}$$

where

$$C(\eta) = \left(\int_{r_{\min}}^{r_{\max}} c(r) f' \left(\sum_k \eta_k v_k \right) v_i v_j dr \right) ,$$

$$A = \left(\int_{r_{\min}}^{r_{\max}} a(r) \, \frac{dv_i}{dr} \, \frac{dv_j}{dr} \, dr \right) , \tag{2.10}$$

$$\gamma_i = a \, \frac{\partial q}{\partial r} \, v_i \, \Bigg|_{r_{\min}}^{r_{\max}} .$$

Note that A is independent of time. The specification of initial values translates into a specification of $\eta(0)$. It can be proved [3] that for any reasonable choice of a family of subspaces $\mathfrak{M}$ converging to H^1, the solution $\tilde{q}$ converges to q

under very reasonable assumptions. Moreover, the estimates hold with constants independent of $\mathfrak{M}$. Similar estimates are valid when (2.9) is solved by a Crank-Nicolson differencing in time or by various predictor-corrector versions of the Crank-Nicolson relation. The proofs are independent of the dimension of the domain.

A practical implementation of the Galerkin approximation has been made using an Hermite quintic basis modified to allow jump discontinuities in the coefficients $c(r)$ and $a(r)$ [4]. The technique has proved to be most satisfactory.

3. Two-Phase Flow.

Let water and oil flow simultaneously in a porous medium. Consider the fluids to be in capillary equilibrium and to be incompressible and immiscible. Then this flow can be described by the degenerate parabolic system [1, 5, 6]

$$(3.1)\qquad \begin{aligned} c(u)_t &= \nabla\cdot(a\nabla u) + \nabla\cdot(b\nabla v) + f, \\ 0 &= \nabla\cdot(b\nabla u) + \nabla\cdot(a\nabla v) + g \quad, \quad x\in\Omega \quad, \quad 0<t\leqslant T\,, \end{aligned}$$

where

$$c = c(x\,,u) \quad, \quad a = a(x\,,u) \quad, \quad b = b(x\,,u) \quad, \quad f = f(x\,,u\,,\nabla u\,,\nabla v),$$

and

$$g = g(x\,,u\,,\nabla u\,,\nabla v)\,.$$

The variable u represents the capillary pressure and v the average of the two phase pressures. Assume that the coefficients are smooth in u and v and bounded and measurable in x. Also, assume that

$$(3.2)\qquad a(x\,,u) \geqslant \alpha > 0 \quad, \quad |b(x\,,u)| \leqslant (1-\epsilon)\,a(x\,,u)\,.$$

While it is possible to assign several types of boundary conditions, the choice arising most naturally is the following :

$$(3.3)\qquad \left.\begin{aligned} a\frac{\partial u}{\partial n} + b\frac{\partial v}{\partial n} &= -\psi\,, \\ b\frac{\partial u}{\partial n} + a\frac{\partial v}{\partial n} &= \psi\,, \end{aligned}\right. \qquad x\in\partial\Omega^+\,,$$

and

$$(3.4)\qquad \frac{\partial u}{\partial n} = 0 \quad, \quad a\frac{\partial v}{\partial n} = \varphi\,, \qquad x\in\partial\Omega^-,$$

where $\partial\Omega^+$ is the portion of $\partial\Omega$ through which injection takes place and $\partial\Omega^-$ the production portion. Consistency with incompressibility requires that

$$(3.5)\qquad \int_{\partial\Omega^+} \psi\,d\sigma + \int_{\partial\Omega^-} \varphi\,d\sigma + \int_\Omega g\,dx = 0.$$

We also assume the normalization

$$\int_\Omega v\,dx = 0 . \tag{3.6}$$

Initially, u, but not v, must be specified.

The weak form of the above system is the following :

$$\begin{aligned} &< c(u)_t \,, z >_\Omega + < a(u) \nabla u \,, \nabla z >_\Omega + < b(u) \nabla v \,, \nabla z >_\Omega \\ &\qquad = - < \psi \,, z >_{\partial\Omega^+} + < \frac{b}{a}\,(u)\, \varphi \,, z >_{\partial\Omega^-} + < f \,, z >_\Omega \,, \\ &< b(u) \nabla u \,, \nabla z >_\Omega + < a(u) \nabla v \,, \nabla z > \\ &\qquad = < \psi \,, z >_{\partial\Omega^+} + < \varphi \,, z >_{\partial\Omega^-} + < g \,, z >_\Omega \,, \\ &\qquad\qquad z \in H^1(\Omega) \;,\; 0 < t \leqslant T \,, \end{aligned} \tag{3.7}$$

where

$$< \alpha \,, \beta >_X = \int_X \alpha\beta \, d\xi \;, \quad d\xi \text{ being Lebesgue measure.} \tag{3.8}$$

The Galerkin method can be used to demonstrate the existence of a solution of (3.7) such that $u \in L^2(0\,, T\,; H^1(\Omega)) \cap L^\infty(0\,, T\,; L^2(\Omega))$, $u_t \in L^2(0\,, T\,; H^1(\Omega)')$, and $v \in L^2(0\,, T\,; H^1(\Omega))$. A uniqueness result can be proved in the neighborhood of a smooth solution ; i.e., if $u \in C^1(\overline{\Omega} \times [0\,, T])$ and $(u\,, v)$ is a solution of (3.7) having $u(0) = u_0$, then for data in a neighborhood of $(u_0\,, \psi\,, \varphi)$ the weak solution depends continuously on the data.

The practical approximation of smooth solutions of (3.7) can also be accomplished by Galerkin methods. In the continuous time case, the error estimate [3] bounds

$$\|u - U\|^2_{L^\infty(0\,,T\,;L^2(\Omega))} + \|u - U\|^2_{L^2(0\,,T\,;H^1(\Omega))} + \|v - V\|^2_{L^2(0\,,T\,;H^1(\Omega))} \,, \tag{3.9}$$

$(U\,, V)$ being the approximate solution, in terms of the ability to approximate $(u\,, v)$ within the Galerkin subspace. Convergence proofs for the differenced-in-time versions of the Galerkin process have been obtained only under the simplifying assumption that $c(x\,, u) = c_1(x) u$. The Crank-Nicolson equation can be shown to be second order correct in Δt and the backward equation in which the coefficients and the boundary conditions are evaluated at the old time level is first order correct Predictor-corrector methods can also be employed. The nonlinear boundary conditions apparently need to be handled implicitly ; also, two passes through the corrector seem to be required to maintain second order accuracy in Δt. This second pass results from the degeneracy of the differential system, since it is not required for a non-degenerate system [2, 3].

Earlier (and not totally unblemished) versions of these results can be found in [1, 5].

4. Three-Phase Flow.

The ' beta factor" description of the flow of oil, gas, and water in a porous medium leads to a nonlinear parabolic system of three equations. Some of the coefficients depend discontinuously on the solution, and, in general, the system fails to fall into any of the commonly considered classes of parabolic systems. Dupont, H.H. Rachford, Jr., and the author have succeeded in devising an efficient Galerkin method for its approximate solution (at least as measured by the computer), but since we have no proofs the problem will not be treated further here.

REFERENCES

[1] Douglas J. Jr. and Dupont T. — The numerical solution of water-flooding problems in petroleum engineering, *Studies in Numerical Analysis 2,* S.I.A.M., Philadelphia, 1970, p. 53-63.

[2] Douglas J. Jr. and Dupont T. — Galerkin methods for parabolic equations, *S.I.A.M.J. Numer. Anal.,* 7, 1970, p. 575-626.

[3] Douglas J. Jr. and Dupont T. — *The flow of fluids in porous media,* to appear.

[4] Douglas J. Jr., Dupont T. and Henderson G.E. — *Simulation of gas well performance by variational methods,* SPE 2891 and to appear.

[5] Douglas J. Jr, Dupont T. and Rachford H.H. Jr. — The application of variational methods to waterflooding problems, *J. Canadian Petroleum Technology* 8, 1969, p. 79-85.

[6] Douglas J. Jr., Peaceman D.W. and Rachford H.H. Jr. — A method for calculating multi-dimensional immiscible displacement, *Trans. A.I.M.E.,* 216, 1957, p. 297-308.

University of Chicago
Dept. of Mathematics,
5 734 University Avenue
Chicago,
Illinois 60 637 (USA)

Actes, Congrès intern. Math., 1970. Tome 3, p. 71 à 77.

PROBLÈMES UNILATÉRAUX EN MÉCANIQUE DES MILIEUX CONTINUS

par G. DUVAUT

1. Introduction.

Nous qualifierons d'*unilatéral* tout problème dans lequel intervient un seuil séparant deux types de comportement ou dans lequel un paramètre doit rester dans un ensemble convexe.

Les premiers problèmes unilatéraux en mécanique ont été introduits, semble-t-il, par :

Hill [9] avec les matériaux de Von Mises
Prager [21] avec les matériaux à blocage
Signorini (puis Fichera [8]) avec le problème de Signorini
Moreau [19] avec le phénomène de cavitation.

Toutes ces situations mécaniques conduisent à des problèmes mathématiques qui s'expriment par des *inéquations variationnelles.* Ces dernières années d'énormes progrès ont été faits dans ce domaine grâce aux travaux de :

Fichera [8] Lions et Stampacchia [17] Lewy et Stampacchia [13]	pour les cas elliptiques
Lions [14] [15] Brézis et Lions [2]	pour les cas paraboliques et hyperboliques.

Devant ces progrès, il était nécessaire de rechercher toutes les situations mécaniques ou physiques conduisant à des inéquations variationnelles. Nous avons fait ce travail -qui a d'ailleurs conduit à des développements mathématiques nouveaux [5], [1]– et l'ensemble des résultats sera publié dans [7]. Ici nous nous proposons d'indiquer comment certaines situations mécaniques conduisent à des inéquations variationnelles. Nous examinerons avec quelques détails :

(1) les parois semi perméables

(2) les phénomènes d'asservissement en thermique

(3) les phénomènes de frottement en élasticité

et nous renvoyons à la bibliographie pour :

(1) les problèmes de plasticité [9] [10] [18] [3] [11] [12]

(2) les écoulements de fluides de Bingham [5] [7]

(3) lc phénomènc de claquage d'antenne en électromagnétisme [5] [7].

2. Les parois semi-perméables.

Considérons une enceinte thermique qui occupe une région ouverte $\Omega \subset R^3$, de frontière Γ régulière, de normale extérieure unitaire $\vec{n}$. La frontière est constituée, en totalité ou en partie, d'une membrane semi-perméable à la chaleur, par exemple qui laisse seulement *sortir* la chaleur. (Cette description en terme de chaleur peut aussi bien être donnée en termes d'écoulement de fluide visqueux dans un milieu poreux limité par une paroi semi-perméable).

La température (ou la pression) à l'intérieur de Ω satisfait à l'équation de diffusion

$$(1) \qquad \frac{\partial u}{\partial t} - \Delta u = f \quad \text{dans} \quad \Omega$$

ou f représente une densité volumique d'apport de chaleur.

Si de plus Γ est en contact, à l'extérieur de Ω, avec un milieu à la température 0, nous avons les conditions à la frontière :

$$(2) \qquad \left\{ \begin{array}{ll} u < 0 \ , & \dfrac{\partial u}{\partial n} = 0 \\[2ex] u \geqslant 0 \ , & -\dfrac{\partial u}{\partial n} = ku \end{array} \right.$$

ou k représente le coefficient de conduction de la paroi Γ. (On admet la loi de Fourier pour la diffusion de la chaleur, ou de Darcy s'il s'agit de fluide en molieux poreux). D'où

Probleme 1. – *Trouver* $u = u\,(x\,,t)$ *qui satisfait* (1), (2) *et* $u\,(x\,,0) = u_0(x)$.

Supposons maintenant que la paroi Γ soit infiniment conductrice pour la chaleur sortante ; le coefficient k est alors infiniment grand et les conditions (2) deviennent

$$(2\,\text{bis}) \qquad \left\{ \begin{array}{l} u < 0 \Rightarrow \dfrac{\partial u}{\partial n} = 0 \\[2ex] u = 0 \Rightarrow \dfrac{\partial u}{\partial n} \leqslant 0 \end{array} \right.$$

d'où le

Probleme 1bis. – *Trouver* $u\,(x\,,t)$ *qui satisfasse* (1), (2 bis) *et* $u\,(x\,,0) = u_0(x)$.

Introduisons les formes bilinéaires suivantes

$$(u\,,v) = \int_\Omega u(x)\,v(x)\,dx \quad , \quad a(u\,,v) = \int_\Omega \text{grad}\,u.\ \text{grad}\,v.\ dx$$

et la fonction scalaire $\phi_k(\xi)$ définie par

$$\phi_k(\xi) = 0 \ \text{ si } \ \xi \leqslant 0 \quad , \quad \phi_k(\xi) = k\xi \ \text{ si } \ \xi > 0 \quad , \quad (k > 0).$$

THEOREME 1. – *Le problème 1 (resp. 1bis) est "équivalent" à trouver u_k (resp. u) tel que* :

$$(3)\quad \begin{cases} \left(\dfrac{\partial u_k}{\partial t}, v\right) + a(u_k, v) + \displaystyle\int_\Gamma \phi_k(u)\, v\, d\Gamma = (f, v)\,, \ \forall v \in H^1(\Omega)\,, \forall t\,, \\ u_k(t) \in H^1(\Omega). \end{cases}$$

(resp.

$$(3\text{bis})\quad \begin{cases} \left(\dfrac{\partial u}{\partial t}, v - u\right) = a(u, v - u) \geqslant (f, v - u)\,, \ \forall v \in H^1(\Omega)\,, v|_\Gamma \leqslant 0\,, \forall t, \\ u(t) \in H^1(\Omega)\,, u|_\Gamma \leqslant 0). \end{cases}$$

Conséquences et résultats. – Le théorème 1 permet de reposer les problèmes 1 et 1 bis dans un cadre fonctionnel précis. On établit alors existence et unicité d'une solution u_k (resp. u) dans la classe fonctionnelle

$$L^2(0, T; H^1(\Omega)) \cap L^\infty(0, T; L^2(\Omega))\ ;$$

de plus $\partial u_k/\partial t$ (resp. $\partial u/\partial t$) appartient à $L^\infty(0, T; L^2(\Omega))$; ceci quel que soit l'instant $T > 0$.

Remarque 1. – La condition 2bis peut aussi s'écrire

$$(4)\qquad -\frac{\partial u}{\partial n} \in \phi(u)$$

où $\xi \to \phi(\xi)$, $\xi \in R$, $\phi(\xi) \subset R$ est une *application multivoque* dont le graphe est indiqué sur la figure ci-contre.

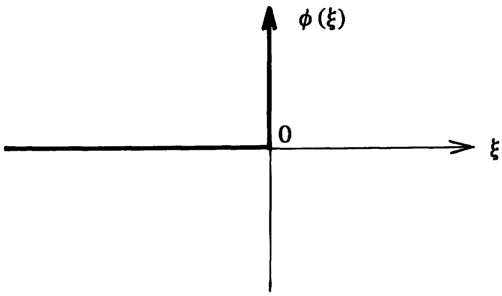


Figure 1

Généralisation. – La remarque 1, ainsi que la considération d'autres problèmes du type de 1 bis, tels des problèmes de climatisation (on veut par exemple maintenir la température à la frontière Γ entre certaines limites) conduit à poser le problème général suivant, PROBLEME 2. – *Trouver* $u(x\,,t)$ *qui satisfasse* (1) *et*

$$(5) \qquad -\frac{\partial u}{\partial n} \in \phi(u) \quad \text{sur} \quad \Gamma \;, \quad \text{p.p.t} \in \left]0\,,T\right[\,.$$

et (6)

$$u(x\,,0) = u_0(x)\,,$$

où (ξ, $\phi(\xi)$), $\xi \in R$ *est un graphe* maximal monotone *dans* R^2 (*c'est-à-dire un graphe monotone tel que toute parallèle à la* $2^{ème}$ *bissectrice le coupe en un point*).

L'existence et l'unicité d'une solution du problème 2 a été établie simultanément par Tartar et Brézis et résulte de la théorie des opérateurs maximaux monotones dans les espaces de Hilbert.

Les mêmes types de problèmes peuvent se poser en remplaçant les flux de chaleur à la frontière par des *injections volumiques* de chaleur ; on parvient ainsi au problème que l'on énonce sous sa forme générale,

PROBLEME 3. – *Trouver* $u = u(x\,,t)$, *avec* $u|_\Gamma = 0$, $u(x\,,0) = u_0(x)$, *solution de*

$$(7) \qquad \frac{\partial u}{\partial t} - \Delta u = f + g \quad \text{dans} \quad \Omega$$

où f *est donnée et*

$$(8) \qquad -g \in \phi(u)$$

où ϕ *est une application multivoque de graphe maximal monotone.*

Ce problème possède également une unique solution grâce aux théories déjà indiquées.

3. Asservissement.

Le contrôle effectué sur la température u dans les problèmes 2 et 3 peut s'effectuer sur $\partial u/\partial t$. On obtient alors les :

PROBLEME 4. – *Trouver* $u = u(x\,,t)$, *avec* $u(x\,,0) = u_0(x)$, *qui satisfasse* (1) *et*

$$-\frac{\partial u}{\partial n} \in \phi\left(\frac{\partial u}{\partial t}\right).$$

PROBLEME 5. – *Trouver* $u = u(x\,,t)$, *avec* $u(x\,,0) = u_0(x)$ *qui satisfasse* (7) *et*

$$-g \in \phi\left(\frac{\partial u}{\partial t}\right).$$

L'application de R dans R : $\xi \to \phi(\xi)$ est *multivoque de graphe maximal monotone.*

On démontre *existence et unicité pour les problèmes 4 et 5* par les méthodes indiquées précédemment.

4. Elasticité linéaire avec frottement (Cas statique).

Dans les problèmes statiques *classiques* d'élasticité linéaire les données aux limites sont des densités de forces de contraintes ou des déplacements. Nous introduisons ici le problème où, sur une partie Γ_f au moins de la frontière Γ de Ω, les déplacements tangentiels ont lieu avec frottement. La loi de frottement retenue est la loi de Coulomb avec le coefficient de frottement $f > 0$. sur la partie complémentaire $\Gamma - \Gamma_f = \Gamma_u$ de la frontière on supposera (par exemple) les déplacements donnés.

On arrive ainsi au :

PROBLEME 6. – *Trouver* $u(x) = u_i(x)$, $i = 1, 2, 3,$ *tel que*

$$\sigma_{ij,j} + f_i = 0 \quad \text{dans} \quad \Omega \,, \tag{8}$$

$$\sigma_{ij} = a_{ijkh} \; \epsilon_{kh}(u) \tag{9}$$

$$u_i = U_i \quad \text{sur} \quad \Gamma_u \tag{10}$$

$$\sigma_N = F_N \quad \text{sur} \quad \Gamma_f (F_N < 0), \tag{11}$$

$$\left.\begin{array}{l} |\sigma_T| < f|F_N| \Rightarrow u_T = 0 \\ |\sigma_T| = f|F_N| \Rightarrow \exists \lambda \geqslant 0 \;,\; u_T = -\lambda \sigma_T \end{array}\right\} \text{sur} \quad \Gamma_f \tag{12}$$

on posera dans la suite $g = f|F_N|$.

Les coefficients d'élasticité a_{ijkh} possèdent les propriétés habituelles

$$\left\{\begin{array}{l} a_{ijkh} = a_{jikh} = a_{khij} \\ a_{ijkh} \; \epsilon_{kh} \; \epsilon_{ij} \geqslant \alpha \epsilon_{ij} \epsilon_{ij} \;, \qquad \alpha > 0. \end{array}\right. \tag{13}$$

Les données $\{f_i\}, \{U_i\}$, F_N, représentent respectivement une densité volumique de forces, un vecteur déplacement, une densité surfacique de forces normales. On établit alors le

THEOREME 2. – *Toute solution assez régulière* $(u \in (H^1(\Omega))^3)$ *du problème 6 satisfait à*

$$\left\{\begin{array}{l} u \in U_{ad} \,, \\ a(u\,, v-u) + \int_{\Gamma_f} g(|V_T| - |u_T|)\, d\Gamma \geqslant (f, v-u) + (F_N, v_N - u_N)_{\Gamma_f} \end{array}\right. \tag{14}$$

$$\forall v \in U_{ad} = \{v \,|\, v \in (H^1(\Omega)^3) \quad , \quad v_i = U_i \quad \text{sur} \quad \Gamma_U\} . \tag{15}$$

(on a posé

$$a(u\,, v) = \int_\Omega a_{ijkh} \; \epsilon_{kh}(u)\, \epsilon_{ij}(u)\, dx \;,\; (f, v) = \int_\Omega f_i v_i dx \;,\; (F, v)_{\Gamma_f} = \int_{\Gamma_f} Fv \; d\Gamma).$$

La relation (14) (15) *est équivalente à : toute solution du problème 6 minimise dans* U_{ad} *la fonctionnelle*

$$(16) \qquad I(v) = \frac{1}{2}\, a(v\,,v) + \int_{\Gamma_f} g\,|V_T|\, d\Gamma - (f\,,v) - (F_N\,,V_N)_{\Gamma_f}.$$

Ces propriétés sont caractéristiques (au moins formellement).

Cette situation est analysée en détail dans [6] où l'on établit sous des hypothèses precises l'existence et l'unicité des champs de déformations et contraintes solutions. On y trouvera aussi un énoncé dual de celui exprimé ici.

Variantes.

(1) Au lieu de donner σ_N sur Γ_f on donne U_N, c'est-à-dire que, dans l'énoncé du problème 6, on remplace (11) par

$$(11\text{ bis}) \qquad u_N = U_N \quad \text{sur} \quad \Gamma_f$$

et (12) devient alors

$$(12\text{ bis} \qquad \begin{array}{l} |\sigma_T| < f\,|\sigma_N| \Rightarrow u_T = 0 \\ |\sigma_T| = f\,|\sigma_N| \Rightarrow \exists\,\lambda \geqslant 0 \quad , \quad u_T = -\lambda\,\sigma_T . \end{array}$$

On obtient ainsi le *problème 6 bis.*

On montre alors :

THEOREME 3. – *Toute solution régulière u du problème 6 bis satisfait à*

$$(17) \quad a(u\,,v-u) + \int_{\Gamma_f} f\,|\sigma_N(u)|\,(|v_T| - |u_T|)\, d\Gamma \geqslant (f\,,v-u) \quad , \quad \forall\, v \in U_{ad}$$

où

$$(18) \quad U_{ad} = \{v \mid v \in (H^1(\Omega))^3 \quad , \quad v_i = U_i \quad \text{sur} \quad \Gamma_U\,,\, v_N = U_N \quad \text{sur} \quad \Gamma_f\}.$$

Cette propriété est formellement caractéristique.

Ce problème 6 bis n'a pas reçu de solution satisfaisante. L'unicité en particulier n'est pas établie à notre connaissance.

(2) Dans les problèmes 6 et 6 bis le frottement agit sur les déplacements tangentiels. On peut également le faire agir sur le déplacement normal, et même considérer des coefficients de frottement différents pour les déplacements vers l'intérieur et vers l'extérieur. On établit alors l'existence et l'unicité et on peut obtenir la solution du problème de Signorini [4] [8] comme cas limite, l'un des coefficients de frottement tendant vers + ∞ et l'autre vers 0.

(3) *Problème de Signorini avec frottement.* Il s'énonce,

PROBLEME 7. – *Trouver* $u = \{u_i(x)\}$, *solution de (8), (9),*

$$(19) \qquad \sigma_{ij}\, n_j = F_i \qquad \text{sur } \Gamma - \Gamma_f,$$

$$(20) \qquad \left\{ \begin{array}{l} u_N \leqslant 0 \\ u_N < 0 \Rightarrow \sigma_{ij}\, n_j = 0 \\ u_N = 0 \Rightarrow \sigma_N \leqslant 0 \\ \text{et} \left\{ \begin{array}{l} |\sigma_T| < f\,|\sigma_N| \Rightarrow u_T = 0 \\ |\sigma_T| = f\,|\sigma_N| \Rightarrow \exists\,\lambda \geqslant 0 \quad , \quad u_T = -\lambda\,\sigma_T \end{array} \right. \end{array} \right\} \text{sur } \Gamma_f$$

Pour toute solution u éventuelle du problème 7 on a la propriété caractéristique.

THEOREME 4. – *Toute solution $u \in (H^1(\Omega))^3$ du problème 7 est caractérisée par*

$$(21)\ a(u\,, v - u) + \int_{\Gamma_f} f\,|\sigma_N(u)|\,(|v_T| - |u_T|)\,d\Gamma \geqslant (f\,, v - u) + (F\,, v - u)_{\Gamma - \Gamma_f} \quad \forall v \in (H^1(\Omega))^3\,.$$

A notre connaissance ce problème est ouvert (10 septembre 1970).

BIBLIOGRAPHIE

[1] BREZIS H. — *Inéquations variationnelles,* Paris, 1970.

[2] BREZIS H. et LIONS J.L. — *C.R. Acad. Sc. Paris* 264, 1967, p. 928-931.

[3] COUTRIS N. — *C.R. Acad. Sc. Paris,* T. 270, 25 mai 1970, p. 1377-1380.

[4] DUVAUT G. — *C.R. Acad. Sc. Paris,* mai 1969, Tome 268, p. 1044-1046.

[5] DUVAUT G. et LIONS J.L. — *C.R. Acad. Sc. Paris,* sept. 1969, déc. 1969, janvier 1970, juin 1970.

[6] DUVAUT G. et LIONS J.L. — Un problème d'élasticité avec frottement, Article pour *le journal de Mécanique* (à paraître).

[7] DUVAUT G. et LIONS J.L. — *Sur les inéquations en Mécanique et en Physique,* En préparation.

[8] FICHERA G. — *Mem. Accad. Naz. Lincei,* 1964, Ser. 8. Vol. 7, p. 91-140.

[9] HILL R. — *The mathematical theory of plasticity,* Clarendon Press, Oxford, 1950.

[10] KOITER W.I. — *Progress in Solid Mechanics,* Tome 1, 1960, p. 165-221. North Holland.

[11] LANCHON H. — *C.R. Acad. Sc. Paris,* oct. 1969, Tome 269, p. 791-794.

[12] LANCHON H et DUVAUT G. — *C.R. Acad. Sc. Paris.* Mars 1967, Tome 264, p. 520-523.

[13] LEWY H. et STAMPACCHIA G. — *Com. Pure Applied Math.* 22, 1969, p. 153-188.

[14] LIONS J.L. — *Contrôle optimal des systèmes gouvernés par des équations aux dérivées partielles,* Dunod, Gauthier-Villars, 1968.

[15] LIONS J.L. — *Quelques méthodes de résolution de problèmes aux limites non linéaires,* Dunod, Gauthier-Villars, 1969.

[16] LIONS J.L. — *Ce Volume.*

[17] LIONS J.L. et STAMPACCHIA G. — *Com. in pure and Applied Math.* XX, 1967, p. 493-519.

[18] MANDEL G. — Séminaire de Plasticité, *Publ. Sc. Tech. Ministère de l'Air,* N.T. 116, 1962.

[19] MOREAU J. — *Journal de Mécanique,* 5, No. 4, 1966.

[20] NAYROLLES J. — *Journal de Mécanique,* à paraître.

[21] PRAGER W. — Locking material, *Trans. of the soc. of rheology,* Vol. 1, 1957.

Centre Universitaire de Saint Denis
Département de Mathématique
93. Saint Denis (France)

Actes, Congrès intern. Math., 1970. Tome 3, p. 79 à 84.

UNILATERAL CONSTRAINTS IN ELASTICITY*

by Gaetano FICHERA

Let A be an elastic body, occupying in its natural configuration a bounded domain of the space, which we denote by the same letter A. It is convenient, in order to include in our study both the cases of 3-dimensional and r-dimensional elasticity, to consider A as a domain of the r-dimensional cartesian space X^r. We suppose that A is such that (i) $\partial A = \partial \overline{A}$, (ii) A satisfies a restricted cone-hypothesis, (iii) the boundary ∂A of A can be decomposed in a finite set of non overlapping $(r-1)$-cells each one very smooth, for instance, of class C^∞ ([1]). Let Σ be a subset of ∂A formed by a finite set of non overlapping C^∞ $(r-1)$-cells. We shall denote Σ as the "surface Σ". We assume that Σ is frictionless and that A is supported by Σ. Let us now suppose that the elastic body is submitted to a given system of body forces and to a given system of surface forces on $\Sigma^* = \partial A - \Sigma$ (if not empty). Assume that the system of the given forces is such that the body is in equilibrium.

The analytic equilibrium conditions, according to the classical theory of elasticity, were, for the first time, written down explicitly by A. Signorini [1]. They are the following

$$(1) \quad \sigma_{ik|k} = f_i \text{ in } A \ , \qquad (2) \quad \sigma_{ik}\nu_k = \varphi_i \quad \text{on} \quad \Sigma^* \ ,$$

and on Σ

$$\text{either } (3) \begin{cases} u_i \nu_i = 0 \ , \\ \sigma_{ik}\,\nu_i \nu_k \geqslant 0 \ , \\ \sigma_{ik}\,\nu_i \tau_k = 0 \ , \end{cases} \qquad \text{or } (4) \begin{cases} u_i \nu_i > 0 \ , \\ \sigma_{ik}\,\nu_i \nu_k = 0 \ , \\ \sigma_{ik}\,\nu_i \tau_k = 0 \ ; \end{cases}$$

f_i and φ_i are functions determined by the given system of body and surface forces ; σ_{ik} are the rectangular components of the stress tensor ; u is the displacement vector ; ν is the inner normal to ∂A ; τ is *any* tangent vector to Σ.

If the first set of conditions is satisfied in a point of Σ, it means that the body is still supported by Σ and, since Σ is frictionless, the reaction of the constraints is purely normal.

(*) This research has been sponsored in part by the Aerospace Research Laboratories through the European Office of Aerospace Research, OAR, United States Air Force, under Grant EOOAR-69-0066.

[1] By a $C^\infty(r-1)$-dimensional cell of X^r we mean the range of a r-vector valued function $x = x(t)$ defined in the $(r-1)$-dimensional closed domain T, $t_k \geqslant 0$ $(k = 1, \dots, r-1)$, $t_1 + \dots + t_{r-1} < 1$ of class C^∞ in T, univalent in T, and such that the jacobian matrix $\partial x/\partial t$ has the maximum rank in every point of T.

Whereas, if the second set of conditions is satisfied, the body is detached from the supporting surface in the considered point and there is no reaction at all.

Of course, Eqs. (1)-(4) must be completed by the stress-strain relation

$$(5) \qquad W_{\epsilon_{ik}} + \sigma_{ik} = 0 ,$$

where $W(x , \epsilon)$ is the elastic potential, i.e. a given function depending on x and on the strain tensor ϵ, whose rectangular components are denoted by ϵ_{ik}. The analytic properties of $W(x , \epsilon)$ are determined by the kind of elasticity we wish tc consider for the body A.

The newness of the boundary value problem (1)-(5) with respect to the classical ones of elasticity depends on the circumstance that we have here a *unilateral constraint*, i.e. the body A is not permitted to go below the supporting surface Σ. This makes so that on Σ the boundary conditions are expressed either by (3) or by (4). It is not known a priori whether in a point of Σ the unknown function must satisfy the first or the second set of conditions. For this reason Signorini suggested for such kind of boundary conditions the name of *ambiguous boundary conditions.*

Let us denote by $\mathcal{U}_\Sigma$ the class of r-vector valued functions v belonging to $H^1(A)$ and satisfying on Σ the boundary condition $v_i\, \nu_i \geqslant 0$ [in the sense of functions of $H^1(A)$] ; $\mathcal{U}_\Sigma$ is the class of the *admissible displacements* belonging to $H^1(A)$. Problem (1)-(5) admits the following *weak formulation.*

Find $u \in \mathcal{U}_\Sigma$ *such that*

$$(6) \int_A \frac{\partial}{\partial \epsilon_{ik}} W[x , \epsilon_{ik}(u)]\, [\epsilon_{ik}(v) - \epsilon_{ik}(u)]\, dx \geqslant \int_A (v_i - u_i) f_i\, dx \; . \\ + \int_{\Sigma^*} (v_i - u_i)\, \varphi_i\, d\sigma$$

for any $v \in \mathcal{U}_\Sigma$ [1].

It is not difficult to see that if u is a solution of the weak problem and satisfies proper smoothness conditions, then u is a solution of problem (1)-(5).

The existence and uniqueness theorem for (6) was given for the first time in 1963 (see [2], [3]). Moreover, since it was shown that, in general, problems like (1)-(5) cannot have a solution in the classical sense (see [3]), it was proved, in connection with the general case of linear elasticity for inhomogeneous anisotropic bodies, the equivalence of (6) with a suitable generalization of problem (1)-(5).

These researches were the starting point of a new chapter of investigations in the theory of partial differential equations, connected with problems which generalize problem (6) and are nowadays denoted as "variational inequalities".

At present, problems connected with unilateral constraints play a central role in analysis and applied mathematics. Let us quote, as a prove, the circumstance that in this Congress at least four lectures are dedicated to this subject.

[1] By $\epsilon_{ik}(u)$ we denote the strain tensor corresponding to the displacement u.

The aim of the present paper is to emphasize one of the most delicate point connected with the solution of problem (6), which seems to have been somewhat overlooked by people who have later worked in this area, i.e. the problem of compatibility conditions for problem (6).

Moreover we shall refer on some recent regularization results concerning the solution of the problem near Σ.

1. Existence and uniqueness theorem.

Let R be the linear variety of all the rigid displacements ρ (in the sense of classical elasticity) i.e. displacements of the form $\rho = a + Bx$ where a is a constant vector and B a skewsymmetric $r \times r$-matrix with constant entries. Let $R_\Sigma = R \cap \mathfrak{U}_\Sigma$. From (6) it readily follows that a *necessary condition for the existence of a solution of problem (6) is*

$$(7) \qquad \int_A \rho_i f_i \, dx + \int_{\Sigma^*} \rho_i \varphi_i \, d\sigma \leqslant 0 \qquad (\rho \in R_\Sigma).$$

This necessary condition was already known to Signorini [1]. However condition (7) is not sufficient to ensure existence and a more strict condition is needed. To this end let us denote by R_Σ^* the subset of R_Σ formed by all the *bilateral* rigid displacements of R_Σ, i.e. a vector ρ of R_Σ belongs to R_Σ^* if and only if $-\rho$ belongs to R_Σ. We say that (7) is satisfied in the strong sense if the equality sign holds if and only if $\rho \in R_\Sigma^*$.

From now on we suppose that the ϵ_{ik}'s are the linearized strain components, i.e.

$$\epsilon_{ik}(u) = 2^{-1} \, (u_{i|k} + u_{k|i}).$$

The following general theorem of analysis holds (see [3]).

(I) *Let* $W(x\,,\epsilon)$ *satisfy the following hypotheses* : (i) $W(x\,,\epsilon)$ *is continuous for every* $(x\,,\epsilon)$; (ii) *for every* x, $W(x\,,\epsilon)$ *is a convex function of* ϵ ; (iii) *there exists a constant* $\lambda_0 > 0$ *such that* $W(x\,,\epsilon) > \lambda_0\, \epsilon_{ik}\, \epsilon_{ik}$; (iv) *if* K_W *denotes the subspace of* $H'(A)$ *formed by the functions* $u(x)$ *such that* $W[x\,,\epsilon(u)] \in \mathcal{L}^1(A)$, *then* $\mathfrak{U}_\Sigma \subset K_W$. *Suppose* $f \in \mathcal{L}^2(A)$, $\varphi \in \mathcal{L}^2(\Sigma^*)$.

Let conditions (7) be satisfied in the strong sense.

Under the assumed hypotheses there exists the absolute minimum of the functional

$$I(u) = \int_A W[x\,,\epsilon(u)]\, dx - \int_A u_i f_i \, dx - \int_{\Sigma^*} u_i \varphi_i \, d\sigma$$

in the class $\mathfrak{U}_\Sigma$.

From this theorem, under the additional hypothesis that $W(x\,,\epsilon)$ has the partial derivatives $W_{\epsilon_{ik}}(x\,,\epsilon)$ which are continuous for every $(x\,,\epsilon)$, the existence theorem for problem (6) follows.

All the hypotheses assumed on $W(x, \epsilon)$ are satisfied in the particular case that

$$(8) \qquad W(x, \epsilon) = \frac{1}{2} a_{ih,jk}(x) \, \epsilon_{ih} \, \epsilon_{jk} ,$$

where the $a_{ih,jk}(x)$'s are C^∞ real valued functions of x defined in X^r, satisfying the symmetry conditions

$$a_{ih,jk}(x) \equiv a_{jk,ih}(x) \equiv a_{hi,jk}(x) \equiv a_{ih,kj}(x)$$

and such that the above quadratic form be positive definite in the variables $\epsilon_{ik} = \epsilon_{ki}$.

This is the important case of linear elasticity for an inhomogeneous anisotropic body. In particular, assuming

$$W(\epsilon) = \epsilon_{ih} \, \epsilon_{ih} + 2^{-1} (\nu - 1) \, \epsilon_{ii} \, \epsilon_{hh}$$

where ν is a constant greater than $r^{-1}(r-2)$ we have the classical case of a homogeneous, isotropic body.

From now on we suppose that $W(x, \epsilon)$ is given by (8). If we consider the symmetric quadratic form in $H^1(A)$

$$B(u, u) = \int_A W[x, \epsilon(u)] \, dx$$

and denote by $B(u, v)$ the corresponding bilinear polar form, inequalities (6) become

$$(6') \qquad 2B(u, u - v) \geqslant F_i(u_i - v_i) \quad ; \quad u \in \mathcal{U}_\Sigma \quad , \quad \forall \, v \in \mathcal{U}_\Sigma ,$$

where

$$F_i(p) = \int_A f_i \, p \, dx + \int_{\Sigma^*} \varphi_i \, p \, d\sigma .$$

It must be remarked that, since $B(u, u)$ is *not* coercive on the space $H^1(A)$, existence theorem for (6') does not follow from the projection theorem on a closed convex set of $H^1(A)$ [1].

Let R_F be the subspace of R formed by the vectors ρ such that $F_i(\rho_i) = 0$. Obviously $R^*_\Sigma \subset R_\Sigma$. The following uniqueness theorem holds [3] :

(II) *If u is a solution of (6'), every other solution is given by $u + \rho$, where ρ is an arbitrary displacement of R_F such that $u + \rho \in \mathcal{U}_\Sigma$.*

[1] Inequalities (6') are the first case of "variational inequalities" which have been considered in the literature (see (44) and (45) of [3]) and probably the only one for which the *necessary* and sufficient conditions for the existence have been given. Later sufficient conditions have been given for the existence theory connected with non symmetric bilinear forms $B(u, v)$, i.e. with problems which are not "variational". The present author disagrees with the use of the term "variational inequalities" in the non symmetric case.

2. Necessity of the strong condition.

As far as the condition (7) is concerned let us distinguish the following three cases :

(1) $F_i(\rho_i) = 0$ for every $\rho \in R$.

(2) Condition (7) is satisfied in the strong sense.

(3) Condition (7) is satisfied but *not* in the strong sense.

Let us suppose that $r = 3$, Σ be a domain of the plane $x_3 = 0$ and $\bar{A} - \Sigma$ be contained in the half-space $x_3 > 0$.

If hypothesis (1) is satisfied, then the system of the given forces f_i and φ_i is equilibrated. That means that a solution of problem (1)-(5) exists, when we replace conditions (3), (4) by the classical boundary condition $\sigma_{ik} \nu_k = 0$. Since the solution is determined up to an arbitrary rigid displacement ρ, we can dispose of ρ in such a way to have a solution u which satisfies in Σ the set of conditions (4). *Let us exclude that condition (1) be satisfied.* The following theorem holds [3] :

(III) *Condition (2) is necessary for the existence of a solution of problem (6').*

Condition (3) has a physical interpretation. Since the system of the given forces is not equilibrated, it is easy to see that the system is equivalent to a single force orthogonal to the plane $x_3 = 0$ directed down-wards and applied in a point of the *central-axis* of the system, i.e. the straight line $x_1 = x_1^0, x_2 = x_2^0$ with

$$x_1^0 = \frac{F_3(x_1) - F_1(x_3)}{F_3(1)} \qquad x_2^0 = \frac{F_3(x_2) - F_2(x_3)}{F_3(1)} .$$

Condition (3) is satisfied if and only if (x_1^0, x_2^0) is contained in the *convex hull* $K(\Sigma)$ of Σ, i.e. in the intersection of all the closed half-planes which contain Σ (see [3]). Moreover condition (2) is satisfied when and only when (x_1^0, x_2^0) is an *interior* point of $K(\Sigma)$.

Condition (3) is necessary and sufficient for the equilibrium of A conceived as a rigid body. However if A is elastic, the more strict condition (2) is necessary and sufficient for its equilibrium. This is a remarkable fact since, contrary to the situation which we have in classical problems of linear elasticity, the conditions which are equivalent to the equilibrium of the rigid body are not anymore equivalent to the equilibrium of the elastic body.

3. Regularity properties of the solution.

There is no problem in obtaining regularity properties for the solution of (6') in any point interior to A or in the neighborhood of a *regular* point of Σ^* having a positive distance from Σ. In this case the problem reduces to the case of classical boundary value problems for linear elliptic systems. As far as regularization in points of Σ is concerned, situation is much more difficult. In [3] it was shown that the reaction of the constraint Σ can be expressed by a measure function $t(B)$, on the σ-ring of the Borel subsets B of Σ, defined by the condition

$$\int_{\Sigma} v_i \, dt_i + \int_{A} \sigma_{ik}(u) \, \epsilon_{ik}(v) \, dx + F_i(v_i) = 0$$

where v is any function of $H^1(A) \cap C^0(\overline{A})$.

We are now in position to give more information concerning the smoothness of u in the neighborhood of a regular point x^0 of Σ having positive distance from Σ^*. Let I be a spherical domain of center x^0, having positive distance from Σ^* and such that $J = \overline{I} \cap \overline{A}$ is C^∞-homeomorphic to the closed semiball Σ^+, $t \geqslant 0$, $|y|^2 + t^2 \leqslant 1$, of the r-dimensional space $(y_1, \ldots, y_{r-1}, t)$. Moreover the homeomorphism which maps J onto Σ^+, maps the set $\overline{I} \cap \partial A$ onto the $(r-1)$-dimensional ball $t = 0$, $|y| \leqslant 1$. The following theorem holds :

(IV) *The solution u is uniformly Hölder continuous in J with any exponent* < 1, *for* $r = 2$, *and with exponent* $1/2$, *for* $r = 3$. *For* $r = 4$ *u belongs to any space* $\mathcal{L}^p(J)$ *with* $0 < p < \infty$. *For* $r > 4$, $u \in \mathcal{L}^{\frac{2r}{r-4}}(J)$. *We have* $u_{|i} \in \mathcal{L}^p(J)$ *with* $0 < p < \infty$ *for* $r = 2$ *and* $u_{|i} \in \mathcal{L}^{\frac{2r}{r-2}}(J)$ *for* $r > 2$.

This results are consequences of the following estimate.

(V) *If we denote by* $\| \ \|_\nu$ *the norm in the space* $H^\nu(J)$, *the following estimate holds*

$$\|u\|_2 \leqslant c \, \{ \|f\|_0 + \|u\|_0 \},$$

where the constant c depends only on J and on the coefficients of W [1].

From this result it follows in particular the proof of a conjecture due to Prof. G.I. Barenblatt, according to which the measure function $t(B)$ is absolutely continuous in the neighborhood of any regular point of Σ.

REFERENCES

[1] Signorini A. — Questioni di elasticità non linearizzata e semi-linearizzata, *Rend di Matem. e delle sue Appl.*, v. XVIII, 1959.

[2] Fichera G. — Sul problema elastostatico di Signorini con ambigue condizioni al contorno, *Rend. Acc. Naz. Lincei*, s. VIII, v. XXXIV, fasc. 2, febbraio 1963.

[3] Fichera G. — Problemi elastostatici con vincoli unilaterali : il problema di Signorini con ambigue condizioni al contorno, *Atti Acc. Naz. Lincei*, Memoria presentata il 10-IX-63, s. VIII, vol. VII, fasc. 5, 1964.

Istituto Matematico Università di Roma
00185 Roma (Italie)

[1] The proof of this theorem will be given in a forthcoming paper.

Actes, Congrès intern. Math., 1970, Tome 3, p. 85 à 94.

THE EXTERIOR PROBLEM FOR THE NAVIER-STOKES EQUATIONS

by Robert FINN

The work on which I shall report is motivated by the physical problem of determining the motion of a viscous incompressible fluid, in which a rigid body B is moving with prescribed velocity, or under the action of known forces. The body is assumed of finite size, and surrounded by fluid which extends to infinity, where it is at rest.

The motion is assumed to be governed by the Navier-Stokes equations, which after suitable normalization can be written [1]

$$(1) \qquad \begin{aligned} \Delta \mathbf{w} - \mathbf{w} \cdot \nabla \mathbf{w} - \nabla p &= \frac{\partial \mathbf{w}}{\partial t} \\ \nabla \cdot \mathbf{w} &= 0. \end{aligned}$$

Here $\mathbf{w} = \mathbf{w}(\mathrm{x}\,; t)$ is the fluid velocity vector, p the (scalar) pressure and t the time. The fluid is assumed to adhere to B at its boundary Σ, and to be at rest at infinity. Thus

$$(2) \qquad \begin{aligned} \mathbf{w}|_{\Sigma} &= \mathbf{w}^{\Sigma} = A(t) \cdot \mathrm{x} \\ \mathbf{w}|_{\infty} &= \mathbf{w}^{\infty} = 0 \end{aligned}$$

where A(t) in an antisymmetric matrix. No conditions are imposed on the pressure. Only the cases of direct physical interest

$$\left\{ \begin{array}{l} \mathrm{x} = (x_1\,, x_2) \\ \mathbf{w} = (w_1\,, w_2) \end{array} \right. , \text{ or } \left\{ \begin{array}{l} \mathrm{x} = (x_1\,, x_2\,, x_3) \\ \mathbf{w} = (w_1\,, w_2\,, w_3) \end{array} \right. , \quad (\text{dimension } n = 2 \text{ or } 3)$$

will be considered in this report. We restrict attention also to flows that are *stationary*, that is, flows for which the velocity field relative to Σ is time independent, and to flows that are close (not necessarily infinitesimally close) to given stationary flows. In a stationary flow, the time coordinate can be eliminated from (1) by a Galilean transformation, so that (1) becomes

$$(3) \qquad \begin{aligned} \nabla \mathbf{w} - \mathbf{w} \cdot \nabla \mathbf{w} - \nabla p &= 0 \\ \nabla \cdot \mathbf{w} &= 0 \end{aligned}$$

[1] It is notationally convenient in this paper to absorb the Reynolds number, which appears in most formulations of (1), into the prescribed data. A disadvantage of this procedure is that the results, as obtained, are not in non-dimensional form. Their conversion to such a form is of course a routine procedure.

and the limiting conditions (2) become

$$(4)\qquad \begin{aligned} \mathbf{w}|_{\Sigma} &= 0 \\ \mathbf{w}|_{\infty} &= \mathbf{w}^{\infty} = \text{const.} \end{aligned}$$

At some points it will be convenient to consider somewhat more general conditions

$$(5)\qquad \begin{aligned} \mathbf{w}|_{\Sigma} &= \mathbf{w}^{\Sigma}(\mathbf{x}) \ , \ \mathbf{x} \in \Sigma \ , \\ \mathbf{w}|_{\infty} &= \mathbf{w}^{\infty} = \text{const.} \end{aligned}$$

The first attempt to study an exterior problem for a viscous fluid seems due to Stokes [1], who considered the linearization that arises in the case of infinitesimally slow motion. The resulting equations

$$(6)\qquad \begin{aligned} \Delta \mathbf{w} - \nabla p &= 0 \\ \nabla \cdot \mathbf{w} &= 0 \end{aligned}$$

are appropriately named the Stokes equations, and have a mathematical interest in themselves. Stokes constructed the striking explicit solution of (6)

$$(7)\qquad \mathbf{w} = \mathbf{w}^{\infty} - \frac{3}{4a} \nabla \wedge \left(\mathbf{r}^{2} \nabla \wedge \frac{\mathbf{w}^{\infty}}{\mathbf{r}} \right) - \frac{a}{4} \nabla \wedge \nabla \wedge \frac{\mathbf{w}^{\infty}}{\mathbf{r}} \ , \quad \mathbf{r} = |\mathbf{x}| ,$$

which represents an (infinitesimal) velocity field satisfying (6), in the case Σ is a 2-sphere of radius a in $n = 3$ space. The solution (7) is not physically realistic, as the symmetry property $\mathbf{w}(\mathbf{x}) = \mathbf{w}(-\mathbf{x})$ is at variance with the physically expected "wake" region behind B. This anomaly appears again in the values calculated from (7) for the force on B, which agree with experiments only for data that are much smaller than would be expected on the basis of analogy with other results of linearization procedures ([1]).

In the case $n = 2$, Stokes found that it was impossible to satisfy the conditions required by a formal expansion, and he concluded, without formal proof but with an insight justified by later developments, that no solution exists. This is the "Stokes Paradox" of hydrodynamics. The phenomenon has been studied by a number of authors (see, e.g. [4, 5] and the references cited therein).

Perhaps the clearest way to interpret the anomalous behaviour is in terms of the forces acting on the obstacle [4]. These appear naturally in the formal representations of a solution of (6) in terms of the fundamental solution tensor E, e. Such a tensor was introduced by Lorentz [6], who found the explicit form

[1] Remarkably it was just this limiting case of arbitrarily small data that was essential for the Millikan oil-drop experiment to determine the charge on the electron. Proofs of the asymptotic validity of the force formula in this case are given in [2, 3].

$$(8)\qquad \begin{aligned} E_{ij}(\mathbf{x}) &= \begin{cases} \dfrac{1}{4\pi}\left(\delta_{ij}\log\dfrac{1}{|\mathbf{x}|}+\dfrac{x_i x_j}{|\mathbf{x}|^2}\right), & n=2\\[2ex] \dfrac{1}{8\pi}\left(\delta_{ij}\dfrac{1}{|\mathbf{x}|}+\dfrac{x_i x_j}{|\mathbf{x}|^3}\right), & n=3 \end{cases}\\[2ex] e_i(\mathbf{x}) &= \begin{cases} \dfrac{1}{2\pi}\dfrac{\partial}{\partial x_i}\log|\mathbf{x}|, & n=2\\[2ex] \dfrac{1}{4\pi}\dfrac{\partial}{\partial x_i}\dfrac{1}{|\mathbf{x}|}, & n=3 \end{cases} \end{aligned}$$

For fixed i, the vectors $E_{ij}(\mathbf{x})$, together with the "pressure" $e_i(\mathbf{x})$, define a solution of (6) in all space except for the point $\mathbf{x} = 0$, where the singularity permits a representation for a general solution analogous to that of classical potential theory. A study of the properties of this representation yields the asymptotic estimate [4, 7]

$$(9)\qquad \mathbf{w}(\mathbf{x}) = \mathbf{w}^\infty + \mathbf{E}\cdot\mathcal{F} + O(|\mathbf{x}|^{1-n})$$

for any solution $\mathbf{w}(\mathbf{x})$ of (6) known to satisfy $|\mathbf{w}(\mathbf{x})| = o(|\mathbf{x}|)$ as $\mathbf{x} \to \infty$. Here $\mathcal{F}$ is the force exerted on B in the fluid motion. If in addition it is known that $\mathbf{w}^\Sigma = 0$, we obtain

$$(10)\qquad \mathcal{F}\cdot\mathbf{w}^\infty = 2\int_{\mathcal{E}} (\operatorname{def}\mathbf{w})^2\, d\mathbf{x}$$

where $\mathcal{E}$ is the (exterior) domain of definition of $\mathbf{w}(\mathbf{x})$, and the deformation tensor $\operatorname{def}\mathbf{w}$ has components $(\operatorname{def}\mathbf{w})_{ij} = \dfrac{1}{2}\left(\dfrac{\partial w_i}{\partial x_j}+\dfrac{\partial w_j}{\partial x_i}\right)$.

From (10) we find that in any non-trivial flow for which

$$\mathbf{w}^\Sigma = 0 \qquad \text{and} \qquad |\mathbf{w}(\mathbf{x})| = o(|\mathbf{x}|)\,,$$

there is a non-zero resistance force in the direction $\mathbf{w}^\infty$. In particular, $\mathcal{F} \neq 0$. But if $n = 2$, then the principal diagonal components of E become logarithmically infinite with $|\mathbf{x}|$. From (9) follows : *if* $n = 2$ *there is no solution* $\mathbf{w}(\mathbf{x})$ *of* (6) *in* $\mathcal{E}$ *for which* $|\mathbf{w}(\mathbf{x})| = o(\log|\mathbf{x}|)$ *at infinity and* $\mathbf{w}(\mathbf{x}) = 0$ *on* Σ.

The "Paradox" can be shown to arise from a non-uniformity of the perturbation (linearization) procedure in an exterior domain. It is the essential reason for the difficulties that have arisen in attempts to solve the exterior problem for the Navier-Stokes equations, as solutions could not be obtained by perturbation procedures from the solutions of (6). This is the case even if $n = 3$, as the physically unrealistic behaviour of the solutions then finds its mathematical expression in the singular behaviour of functionals that appear formally in the perturbation. The paradox will be clarified from another point of view in § 5.

2. – The first general contribution to the theory of (3) in an exterior domain is due to Leray [8], and is based on his discovery of an a-priori bound for the

Dirichlet integral $D[\mathbf{w}] = \int |\nabla \mathbf{w}|^2 d\mathbf{x}$ for any solution of (3) in a bounded domain, depending only on the data $\mathbf{w}^{\Sigma}$. It is remarkable that if Σ is composed of an inner boundary Σ_0 and a circle (or sphere) Σ_R of (large) radius R, then the bound on $D[\mathbf{w}]$ is independent of R whenever the data on Σ_R are constant. This result permitted Leray to prove the existence of a suitable family of solutions $\mathbf{w}_R(\mathbf{x})$ in expanding annular domains, and to prove their equicontinuity in compact subdomains. Thus *a solution of* (3) *exists in* $\mathcal{E}$, *is locally smooth, and assumes prescribed data* (3) *on* Σ. Leray showed that *if* $n = 3$ *the solution assumes the data* $\mathbf{w}^{\infty}$ *in the sense* $\int_{\mathcal{E}} \frac{|\mathbf{w} - \mathbf{w}^{\infty}|^2}{|\mathbf{x} - \mathbf{y}|^2} d\mathbf{x} < C < \infty$ *uniformly for all* $\mathbf{y} \in \mathcal{E}$. In fact, *the result* $\mathbf{w}(\mathbf{x}) \to \mathbf{w}^{\infty}$ *holds* [9, 10 p. 138], and it is known that $\nabla \mathbf{w} \to 0$ at infinity. However, nothing further has been proved about the asymptotic behaviour of these solutions. It is not known whether they exhibit a "wake" region behind B, nor is it known whether they are unique, even for small data. If $n = 2$ and the physical data $\mathbf{w}^{\Sigma} = 0$ are imposed, the possibility that $\mathbf{w}(\mathbf{x}) \equiv 0$ in $\mathcal{E}$ has not been excluded.

3. – Oseen [11] sought to overcome the singularity in the perturbation by linearizing (3) about the solution $\mathbf{w} \equiv \mathbf{w}^{\infty}$. For the function $\mathbf{u}(\mathbf{x}) = \mathbf{w}(\mathbf{x}) - \mathbf{w}^{\infty}$ we find

$$\begin{aligned} \Delta \mathbf{u} - \mathbf{w}^{\infty} \cdot \nabla \mathbf{u} - \nabla p &= \mathbf{u} \cdot \nabla \mathbf{u} \\ \nabla \cdot \mathbf{u} &= 0 \end{aligned} \tag{11}$$

and in the limit for small $\mathbf{u}(\mathbf{x})$ one obtains the equations

$$\begin{aligned} \Delta \mathbf{u} - \mathbf{w}^{\infty} \cdot \nabla \mathbf{u} - \nabla p &= 0 \\ \nabla \cdot \mathbf{u} &= 0 \end{aligned} \tag{12}$$

which are to be solved under the conditions

$$\begin{aligned} &\mathbf{u} = \mathbf{u}^{\Sigma} \text{ on } \Sigma \\ &\mathbf{u} \to \mathbf{w}^{\infty} \text{ at infinity.} \end{aligned} \tag{13}$$

This procedure does not lead to significantly better agreement with experiment (4). It does, however, avoid the Stokes Paradox and yield the qualitatively anticipated asymptotic behaviour at infinity. Oseen gave explicitly a fundamental tensor $\mathbf{E}$, $\mathbf{e}$, in terms of invariant differential operations on a single scalar [11].

4. – The Oseen equations may be taken as starting point for a different approach to the exterior problem [3]. Consider first the three-dimensional case, $n = 3$. We first observe that for any prescribed data $\mathbf{w}^{\Sigma}$ (c.f. footnote (3)) there is a

---

(3) The condition $\oint_{\Sigma} \mathbf{w}^{\Sigma} \cdot \nu d\sigma = 0$ (ν = unit exterior normal, $d\sigma$ = surface element), imposed by Leray, may be unnecessary in an exterior domain (c.f. the discussion in [2]).

(4) See, e.g., Olmstead and Gautesen [12, 13], where it is shown that in any flow described by a solution of (12, 13) with $\mathbf{u}^{\Sigma} = u$ the force does not change in magnitude under reversal of the direction of $\mathbf{w}^{\infty}$.

(unique) solution of (12) in $\mathcal{E}$ which vanishes at infinity. Since the fundamental tensor E, e, also vanishes at infinity, this implies the existence of a Green's tensor $G_{ij}(\mathbf{x}, \mathbf{y})$, whose components, as functions of either variable for fixed values of the other, vanish on Σ and at infinity. Let $\mathbf{w}^0(\mathbf{x})$ be the solution of the linearized problem (12, 13), and set $\mathbf{u}^0(\mathbf{x}) = \mathbf{w}^0(\mathbf{x}) - \mathbf{w}^\infty$. We are led to the representation

$$(14) \qquad \mathbf{u}(\mathbf{x}) = \mathbf{u}^0(\mathbf{x}) - \int_{\mathcal{E}} \mathbf{u}(\mathbf{y}) \cdot \mathbf{u}(\mathbf{y}) \cdot \nabla G(\mathbf{x}, \mathbf{y}; \mathbf{w}^\infty)\, d\mathbf{y} \equiv \mathbf{T}\mathbf{u}$$

which holds for any solution $u(x)$ of (11) in $\mathcal{E}$ that has reasonable asymptotic properties at infinity. We consider (14) as an integral equation for the unknown $\mathbf{u}(\mathbf{x})$, in the class of all functions that are locally (say) Hölder continuous and satisfy $\lim_{x \to \infty} \sup |\mathbf{x}|\ |\mathbf{u}(\mathbf{x})| < \infty$. An existence theorem in this class for the non-linear system (11) follows from the crucial technical lemma

$$(15) \qquad |\mathbf{x}| \int |\mathbf{y}|^{-2} |\ \nabla G(\mathbf{x}, \mathbf{y}; \mathbf{w}^\infty)|\, d\mathbf{y} < H < \infty$$

uniformly in $\mathbf{x} \in \mathcal{E}$ and in $\mathbf{w}^\infty$, as $\mathbf{w}^\infty \to 0$. The lemma is proved in [3]. It implies that if a norm $\|\mathbf{u}\| = \sup_{\mathcal{E}} |\mathbf{x}|\ |\mathbf{u}(\mathbf{x})|$ is introduced, then if $\|\mathbf{u}^0\| < 1/4H$, the operator $\mathbf{T}\mathbf{u}$ will contract the sphere $\|\mathbf{u}\| < 1/2H$ and will carry this sphere into itself. Thus, we may conclude that *solutions exist whenever the data are small.* It is not difficult to prove that the solutions are locally smooth and satisfy the original system (11), and, using estimates due to Odqvist [14] for the Green's tensor of (6) in a bounded domain, one may show that *the data* $\mathbf{w}^\Sigma$ *are approached smoothly.* Further, the solution tends strictly to $\mathbf{w}^\infty$ at infinity and exhibits a paraboloidal "wake" region behind B in the direction $\mathbf{w}^\infty$. Precisely, *let* $r = |\mathbf{x}|$ *and let* ϕ *be the angle subtended at* $\mathbf{x}$ *by a half ray from the origin in the direction* $\mathbf{w}^\infty$. *There then holds* [15, 3]

$$(16) \qquad \begin{aligned} |\mathbf{w} - \mathbf{w}^\infty| &< Cr^{-2} && \text{if } |\phi| > \phi_0 > 0 \\ |\mathbf{w} - \mathbf{w}^\infty| &< Cr^{-(1+\sigma)} && \text{if } |\phi| < \phi_0\, r^{-(1-\sigma)/2},\ 0 \leqslant \sigma \leqslant 1. \end{aligned}$$

In the case of the physical data $\mathbf{w}^\Sigma \equiv 0$, (16) *is best possible, in the sense that* $|\mathbf{w} - \mathbf{w}^\infty| = o(r^{-1})$ *implies* $\mathbf{w} \equiv \mathbf{w}^\infty = 0$. *For sufficiently small data, the solution obtained in this way is unique among all solutions in the given class that assume the same data* (and not only locally, as is implied by the contraction property of $\mathbf{T}\mathbf{u}$). A complete description of the results, together with their proofs, appears in [3].

5. – The two dimensional case cannot be treated directly by the methods of the preceding section. In fact, if $n = 2$ an analogue of the basic lemma (15),

$$(17) \qquad |\mathbf{x}|^{\frac{1-\epsilon}{2}} \int_{\mathcal{E}} |\mathbf{y}|^{-1} |\nabla G(\mathbf{x}, \mathbf{y}; \mathbf{w}^\infty)|\, d\mathbf{y} < H(\mathbf{w}^\infty) < \infty \quad , \quad \text{any } \epsilon > 0\,,$$

still holds for any fixed $\mathbf{w}^\infty$, but it is unlikely that it holds uniformly as $\mathbf{w}^\infty \to 0$.

Thus, the preceding method yields the existence of a solution for data $\mathbf{w}^\Sigma$ in some functional neighbourhood of $\mathbf{w}^\infty$, but the size of the neighbourhood shrinks as $\mathbf{w}^\infty \to 0$. Without further information the existence of a solution for the physical data $\mathbf{w}^\Sigma = 0$ cannot be inferred from (17), even for small $|\mathbf{w}^\infty|$. In [5, 16], this problem is approached by studying separately two perturbations that can be distinguished in a natural way in the problem as originally posed. We begin by writing $\mathbf{u}(\mathbf{x}\,;\mathbf{w}^\infty) = \dfrac{\mathbf{w}(\mathbf{x}) - \mathbf{w}^\infty}{|\mathbf{w}^\infty|}$, for which (3) takes the form

$$\Delta\mathbf{u} - \mathbf{w}^\infty\cdot\nabla\mathbf{u} - \nabla p = \tau\mathbf{u}\cdot\nabla\mathbf{u}, \qquad \tau = |\mathbf{w}^\infty| \tag{18}$$
$$\nabla\cdot\mathbf{u} = 0\,.$$

We now relax the requirement $\tau = |\mathbf{w}^\infty|$ and consider independently the two perturbations $\tau \to 0$, $\mathbf{w}^\infty \to 0$, under fixed prescribed data for $\mathbf{u}(\mathbf{x}\,;\mathbf{w}^\infty)$. Suppose the first perturbation has already been made. To facilitate the second, we write $\mathbf{w}^\infty = \lambda\alpha$, α a fixed vector and $0 < \lambda \leqslant 1$, so that (18) takes the form

$$\Delta\mathbf{u} - \lambda\alpha\cdot\nabla\mathbf{u} - \nabla p = 0 \tag{19}$$
$$\nabla\cdot\mathbf{u} = 0$$

the linearity of which permits us to write the prescribed data in the form

$$\mathbf{u} \to \mathbf{u}^\Sigma \qquad \text{on } \Sigma \tag{20}$$
$$\mathbf{u} \to 0 \qquad \text{at infinity.}$$

Under reasonable smoothness requirements on Σ and on $\mathbf{u}^\Sigma$, we may now assert the following results relevant to the singular perturbation $\lambda \to 0$, when $n = z$.

(1) *Given* λ, $\mathbf{u}^\Sigma$, *if* $\lambda \neq 0$ *there is a unique solution* $\mathbf{u}(\mathbf{x}\,;\lambda)$ *of* (19) *in* $\mathcal{E}$ *satisfying the limiting conditions* (20).

(2) *As* $\lambda \to 0$ *the solutions* $\mathbf{u}(\mathbf{x}\,;\lambda)$ *remain uniformly bounded in Dirichlet norm, that is, there is a constant A such that*

$$\int_{\mathcal{E}} |\nabla\mathbf{u}(\mathbf{x}\,;\lambda)|^2\,d\mathbf{x} < A < \infty$$

uniformly in λ *on any interval* $0 < |\lambda| < \lambda_0$.

(3) *There exists* $\mathbf{u}^0(\mathbf{x}) = \lim\limits_{\lambda\to 0} \mathbf{u}(\mathbf{x}\,;\lambda)$ *uniformly on any bounded set of points* $\mathbf{x} \in \mathcal{E}$; $\mathbf{u}^0(\mathbf{x})$ *satisfies the Stokes equations* (6) *and* $\mathbf{u}^0(\mathbf{x}) \to \mathbf{u}^\Sigma$ *on* Σ.

(4) *The function* $\mathbf{u}^0(\mathbf{x})$ *is the unique solution of* (6), *such that* $\mathbf{u}^0(\mathbf{x}) \to \mathbf{u}^\Sigma$ *on* Σ *and* $\int_{\mathcal{E}} |\nabla\mathbf{u}|^2\,d\mathbf{x} < \infty$.

(5) *Let* $\mathfrak{F}(\lambda)$ *be the force exerted on B by the flow* $\mathbf{u}(\mathbf{x}\,;\lambda)$; *let* $\mathfrak{F}^0$ *be the force arising from the flow* $\mathbf{u}^0(\mathbf{x})$. *There holds* $\mathfrak{F}^0 = \lim\limits_{\lambda\to 0} \mathfrak{F}(\lambda)$.

(6) $\mathfrak{F}^0 = 0$.

(7) *There exists* $\mathbf{u}_0^\infty = \lim\limits_{x\to\infty} \mathbf{u}^0(\mathbf{x})$, *and* $|\mathbf{u}_0^\infty| \neq \infty$. *In general,* $\mathbf{u}_0^\infty \neq 0$.

(8) *There holds* $u_0^\infty = \frac{1}{4\pi} \lim_{\lambda\to 0} \mathfrak{F}(\lambda) \log \frac{1}{\lambda}$ ·

We note from (7) that the limiting condition at infinity is in general lost in the perturbation. Properties (1) - (7) can be regarded as a clarification of the Stokes Paradox from the point of view of singular perturbation theory. Property (8) has a consequence that is important for the non-linear theory :

There holds $|\mathfrak{F}(\lambda)| < \frac{C}{\log (1/\lambda)}$ *as* $\lambda \to 0$. $\mathfrak{F}(\lambda)$ *has the asymptotic direction* u_0^∞.

From properties (4) and (8) we find :

Suppose the physical data $u^\Sigma = - w^\infty$ *are imposed. Then* $\mathfrak{F}(\lambda)$ *is asymptotically independent of the shape or size of the obstacle B.*

For by (4) we find in this case $u^0(x) \equiv u^\Sigma = - w^\infty$, so that $u_0^\infty = - w^\infty$, regardless of the choice of *B*.

6. – A closer study of the perturbation $\lambda \to 0$ in (19) yields the following result : *let* $0 < \epsilon < 1/2$. *Set* $\xi = (\xi_1 , \xi_2)$,

$$h_i(\xi) = \begin{cases} \log \frac{2}{|\xi|} & \text{if} \quad 0 < |\xi| \leqslant 1 \\ |\xi|^{-1/2} & \text{if} \quad |\xi| > 1, \quad i = 1 \\ |\xi|^{-(1-\epsilon)/2} & \text{if} \quad |\xi| > 1, \quad i = 2, \end{cases}$$

Suppose the coordinates are chosen so that $w^\infty = (w^\infty , 0)$. *Then there is a constant C such that the solution* $u(x ; \lambda)$ *of* (19, 20) *satisfies*

$$|u_i(x ; \lambda)| < C\, h_i(\lambda x) \frac{1}{\log (1/\lambda)} \qquad \text{as} \quad \lambda \to 0 .$$

The results of this and of the preceding section are proved in [5].

7. – We now apply the preceding results to the two dimensional nonlinear problem [16]. Set $w^\infty = \lambda\alpha$, set $u(x ; \lambda) = \frac{1}{\lambda} [w(x ; \lambda) - \lambda\alpha]$. Then $u(x ; \lambda)$ is to satisfy

$$\text{(21)} \qquad \begin{aligned} \Delta u - \lambda\alpha \cdot \nabla p - \nabla p &= \lambda u \cdot \nabla u \\ \nabla \cdot u &= 0 \end{aligned}$$

and a condition of small limiting data for $w(x ; \lambda)$ becomes a condition of fixed (or bounded in a suitable norm) data for $u(x ; \lambda)$. The existence theorem of § 5 implies the existence of a Green's tensor $G(x , y ; \lambda)$, and we are led again to an integral equation

$$\text{(22)} \qquad \begin{aligned} \hat{u}(x ; \lambda) &= u(x ; \lambda) - \lambda \int_{\mathfrak{G}} u(y ; \lambda) \cdot u(y ; \lambda) \cdot \nabla G(x , y ; \lambda)\, dy \\ &\equiv Tu \end{aligned}$$

where $\hat{u}(x\,;\lambda)$ is the solution of (19) with the same data. The basic lemma of § 4 now becomes

$$(23) \qquad \lambda \int_{\mathcal{E}} h_j(\lambda y)\, h_k(\lambda y) \left| \frac{\partial G_{ij}(x\,, y\,;\lambda)}{\partial y_k} \right| dy < H h_i(\lambda x)\ , \quad i = 1,\, 2$$

for fixed $H < \infty$, all $x \in \mathcal{E}$, as $\lambda \to 0$. It is proved in [16]. Note that in (23) it is necessary to distinguish the two velocity components, which apparently behave differently at infinity.

The lemma implies that **Tu** contracts the function sphere

$$\mathcal{S}_H \ : \ |u_i(x\,;\lambda)| < \frac{1}{2H}\, h_i(x\,;\lambda)$$

and carries $\mathcal{S}_H$ into itself if

$$|\hat{u}_i(x\,;\lambda)| < \frac{1}{4H}\ h_i(x\,;\lambda).$$

But $|\hat{u}_i| < C\, h_i(x\,;\lambda)\ \dfrac{1}{\log \dfrac{1}{\lambda}}$ for a fixed constant C (§ 6). *Thus, for small* λ *the properties of contraction mappings imply the existence of a solution of the boundary problem for* (21).

8. – *It is immediate that the solution is unique in a functional neighbourhood.* In contradistinction to the three dimensional case (§ 4), uniqueness in the large has not yet been proved if $n = 2$. *If* $w^{\Sigma} = 0$, *the solutions exhibit a parabolic wake region in the direction* w^{∞} [17]. The following limiting properties are analogous to those shown for the Oseen equations in [5].

(1) *Suppose* $u^{\Sigma} \to u_0^{\Sigma}$ *in a suitable norm, as* $\lambda \to 0$. *Then if* $w(x\,;\lambda)$ *are the corresponding solutions of* (3, 5), *the functions* $u(x\,;\lambda) = \dfrac{w(x\,;\lambda) - \lambda\alpha}{\lambda}$ *converge uniformly in bounded subsets of* $\mathcal{E}$ *to a (unique) solution* $u^0(x)$ *of the Stokes equations* (6), *with boundary data* u_0^{Σ} *and finite Dirichlet integral.*

(2) *There exists* $u_0^{\infty} = \lim\limits_{x \to \infty} u^0(x)$, *and* $|u_0^{\infty}| < \infty$.

(3) *Let* $\mathcal{J}(\lambda)$ *be the force on* Σ *due to the flow* $w(x\,;\lambda)$. *Then*

$$u_0^{\infty} = \frac{1}{4\pi} \lim_{\lambda \to 0} \frac{1}{\lambda}\, \mathcal{J}(\lambda) \log \frac{1}{\lambda}$$

Again we find *that* $\mathcal{J}(\lambda)$ *has the asymptotic direction* u_0^{∞}, *and that if the physical data* $w(x\,;\lambda) \to 0$ *on* Σ *are imposed,* $\mathcal{J}(\lambda)$ *is asymptotically independent of the choice of* Σ.

9. – The following result is due to D. Clark [18]. *Let* $w(x)$ *be a three dimensional solution of* (3) *in* $\mathcal{E}$, *such that for some* $w^{\infty} \neq 0$ *and* $\epsilon > 0$, *there holds*

$|x|^{\frac{1}{2}+\epsilon}$ $|w(x) - w^{\infty}| < M < \infty$ *for all* $x \in \mathcal{E}$. *Then along any ray extending to infinity in a direction not coinciding with that of* w^{∞}, *there holds* $|\text{rot } w(x)| < C_1 e^{-C_2|x|}$ *for certain positive constants* C_1, C_2. The result applies in particular to the solutions described in § 5 to § 8. The significance is seen most clearly in the "physical" case $w^{\Sigma} = 0$, for which it can be shown (c.f. the discussions in [15, 3]) that the estimates $|\nabla w| = \begin{cases} O(|x|^{-2} & \text{outside the wake} \\ O(|x|^{-\frac{3}{2}} & \text{in the wake} \end{cases}$ cannot be improved. Thus, as $x \to \infty$, the flow is asymptotically (and exponentially) potential in directions outside the wake. In the direction w^{∞}, Clark's estimate no longer holds.

Clark also obtained an analogous result for the case $n = 2$, and, for $n = 3$, general asymptotic estimates and expansions for solutions known to decay at prescribed rates at infinity.

10. – The family of solutions studied by Clark (§ 9) was introduced in [15], where they were referred to as "physically reasonable" solutions for evident reasons. The stability of such solutions has been investigated by Heywood [19], who considered the initial value problem (equation (1)) that arises when a physically reasonable solution w(x) is initially disturbed. Heywood proved that *if* w(x) *is sufficiently small in the norm introduced in* § 4, *and if the initial disturbance is small then there is a unique time dependent solution* w(x ; *t*) *that converges to* w(x) *in the Dirichlet norm and in* L_2^{loc} as $t \to \infty$. The proof is obtained by combining the methods used by Ladyzhenskaia [10] for the case of a bounded domain, with the method used to prove the global uniqueness of the solutions described in § 4 [3].

Note added in proof : I have just had the pleasure of seeing a preliminary version of a new paper of Heywood, who considers a solution of (3, 5) with $n = 3$, for which (a) $|x|$ $|w(x) - w^{\infty}|$ is uniformly small in $\mathcal{E}$, and (b) $\mathcal{J} = 0$. Under these conditions, Heywood shows that the given flow can be achieved, in Dirichlet norm and in L_2^{loc} , as limit of a time dependent motion (in a reasonable class) starting from rest, and also that no other solution of (3, 5) with the same data can be achieved in this way. Thus, if, in this situation, a Leray solution (§ 2) is distinct from the solution of § 4, it cannot be achieved as limit of a time dependent motion in the given class.

The work described in this report was supported in part by NSF grant and AFOSR contract with Stanford University. Some of it was completed during the author's tenure at the Université de Paris as a Guggenheim Fellow.

REFERENCES

[1] Stokes G. G. — On the effect of the internal friction of fluids on the motion of pendulums, *Math. and Phys. Papers,* Vol. III, 1851, p. 1.

[2] Finn R. — On the steady-state solutions of the Navier-Stokes equations, III. *Acta Math.* 105, 1961, p. 197-244.

[3] Finn R. — On the exterior stationary problem for the Navier-Stokes equations, and associated perturbation problems, *Arch. Rational Mech. Anal.,* 19, 1965, p. 363-406.

[4] Chang I-Dee and Finn R. — On the solutions of a class of equations occurring in continuum mechanics, with application to the Stokes paradox, *Arch. Rational Mech. Anal.* 7, 1961, p. 388-401.

[5]Finn R. and Smith D.R. — On the linearized hydrodynamical equations in two dimensions, *Arch. Rational Mech. Anal.* 25, 1967, p. 1-25.

[6] Lorentz H.A. — *Abhandlungen über theoretische Physik,* Bd. 1, S., p. 23-42, Leipzig 1907.

[7] Finn R. — On the Stokes Paradox and related questions, In : *Nonlinear Problems.* Madison : The University of Wisconsin Press, 1963.

[8] Leray J. — Etude de diverses équations integrales non linéaires et de quelques problèmes que pose l'hydrodynamique, *J. Math. Pures Appl.,* 9, 1933, p. 1-82.

[9] Finn R. — On steady-state solutions of the Navier-Stokes partial differential equations, *Arch. Rational Mech. Anal.,* 3, 1959, p. 381-396.

[10] Ladyzhenskaia O.A. — *The mathematical theory of viscous incompressible flow,* New York : Gordon and Breach (Transl. from Russian), (Second edition, 1969).

[11] Oseen C.W. — *Neuere Methoden und Ergebnisse in der Hydrodynamik,* Leipzig: Akademische Verlagsgesellschaft m. b. H. 1927.

[12] Olmstead W.E. and Gautesen A.K. — A new paradox in viscous hydrodynamics, *Arch. Rational Mech. Anal.,* 29, 1968, p. 58-65.

[13] Olmstead W.E. — Force relationships and integral representations for the viscous hydrodynamical equations, *Arch. Rational Mech. Anal.* 31, 1968, p. 380-389.

[14] Odqvist F.K.G. — *Die Randwertaufgaben der Hydrodynamik zäher Flüssigkeiten,* Stockholm, P.A. Norstedt and Söner 1928, See also : *Math. Z.* 32, 1930, p. 329-275.

[15] Finn R. — Estimates at infinity for stationary solutions of the Navier-Stokes equations, *Bull. Math. de la Soc. Sci. Math. Phys. de la R.P.R.* 3 (51), 1959, p. 387-418.. See also : *Amer. Math. Soc. Proc. Symposia Pure Math.* 4, 1961, p. 143-148.

[16] Finn R. and Smith D.R. — On the stationary solutions of the Navier-Stokes equations in two dimensions, *Arch. Rational Mech. Anal.* 25, 1967, p. 26-39.

[17] Smith D.R. — Estimates at infinity for stationary solutions of the Navier-Stockes equations in two dimensions, *Arch. Rational Mech. Anal.* 20, 1965, p. 341-372.

[18] Clark D. — The vorticity at infinity for solutions of the stationary Navier-Stokes equations in exterior domains, *Indiana Math J.,* 20, 1971, p. 633-654.

[19] Heywood J.G. — On stationary Solutions of the Navier-Stokes Equations as Limits of Nonstationary Solutions, *Arch. Rational Mech. Anal.,* 37, 1970, p. 48-60.

Stanford University
Dept. of Mathematics,
Stanford,
California 94305 (U.S.A.)

Actes, Congrès intern. Math., 1970. Tome 3, p. 95 à 97.

NUMERICAL DESIGN OF SHOCKLESS TRANSONIC AIRFOILS

by P.R. GARABEDIAN and D.G. KORN

To improve the performance of fast subsonic aircraft it has become desirable to shape the wings so as to suppress shock waves when the flow past them becomes transonic. Modern theories of the quasi-linear equations of motion suggest that a weak solution, by which we mean a solution with shocks, should not only exist, but should also be unique, so that elimination of the shocks becomes practical at a specified speed. This can be achieved within the framework of inviscid fluid dynamics if boundary layer separation is prevented. Our aim here is to sketch a mathematical method for computing shock-free transonic flows that has led to effective procedures for the design of supercritical airfoils whose force-break Mach number is relatively high.

We solve the standard equation

$$(c^2 - u^2)\phi_{xx} - 2uv\phi_{xy} + (c^2 - v^2)\phi_{yy} = 0$$

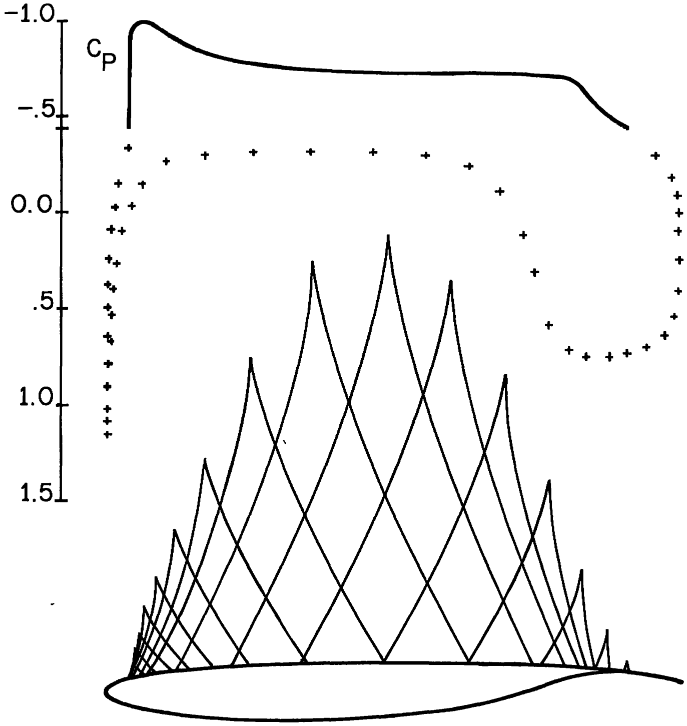
of two-dimensional gas dynamics numerically by means of a stable finite difference scheme drawn from our earlier treatment of the inverse detached shock problem. Complex characteristics are used to construct the analytic continuation of a desired solution in the four-dimensional complex extension of the classical hodograph plane. The method enables us to transform any analytic function of one complex variable into a transonic flow by solving characteristic initial value problems with data assigned along certain paths of integration in two initial complex planes. In this formulation of the equations of motion the sonic line becomes a singularity that must be circumvented by appropriate deformation of the paths of integration.

The analytic functions we use to define initial data are suggested by a knowledge of incompressible flow. Additional logarithmic terms and polynomials serve to control the airfoil shape and the pressure coefficient C_p on it in an appropriate fashion. These techniques turn out to be successful in the transonic as well as the subsonic range.

Since all the functions we consider must be calculated for complex values of their arguments, one difficulty that is encountered in computing them is to specify their branches correctly. Another more serious difficulty occurs in visualizing the geometry of the four-dimensional domain of two independent complex

The work presented in this paper is supported by the AEC Computing and Applied Mathematics Center, Courant Institute of Mathematical Sciences, New York University, under Contract AT(30-1)-1480 with the U.S. Atomic Energy Commission.

variables that underlies our method. To overcome such problems we have found it very helpful in practice to plot diagrams of the points in the initial complex planes through which pass characteristics intersecting the sonic line or the body in the real hodograph plane.



SHOCKLESS TRANSONIC AIRFOIL AT M= 8, C_L=.7, T/C= I

A major problem in operating our transonic flow program on the computer is to choose the variety of parameters appearing in the complex initial data so as to achieve a closed profile in the physical plane. Our procedure for doing this is at the present time more of an art than a science. By solving a suitable system of linear algebraic equations we do, however, impose a set of natural conditions on the gradient of the velocity potential ϕ at the points in the hodograph plane corresponding to the nose and the tail of the airfoil. In particular, at the image of the tail we require that ϕ have a multiple critical point consistent with the Kutta-Joukowski condition. Moreover, by locating the latter point near the sonic line we obtain a camber of the airfoil that generates high lift for a moderate adverse pressure gradient in much the same way as might be brought about by using a flap. Finally, all limiting lines are kept inside the body by narrowing it in a not always obvious fashion.

We are now able to calculate the $(x\,,y)$-coordinates of a typical transonic airfoil on the C.D.C. 6600 computer at N. Y. U. in as little as one minute of machine time. We have developed supercritical airfoils without shocks whose camber is quite similar to that of the famous Whitcomb upside-down wing. Experimental work will be done to show that these airfoils have good aerodynamic properties and can be used in the design of transonic aircraft. In the figure we display an example of one of our profiles that was calculated at the free-stream Mach number $M = .8$ and yielded a lift coefficient $C_L = .7$ while maintaining a thickness-chord ratio T/C = .1. We believe that this represents nearly the fattest two-dimensional wing that can be achieved in the range specified when we exclude shock waves in order to control drag rise. It is to be observed that the supersonic zone indicated by the Mach lines is unusually large, so that when shocks do appear at neighboring off-design conditions they should occur well toward the back of the wing.

New York University
Courant Institute of Mathematical Sciences
251 Mercet Street,
New York, N.Y. 10012 (USA)

Actes, Congrès intern. Math., 1970. Tome 3, p. 99 à 104.

SUR LA CONSTRUCTION DES RÉSEAUX DANS LES DOMAINES COMPLIQUÉS D'UNE FAÇON AUTOMATIQUE POUR LES ÉQUATIONS AUX DIFFÉRENCES FINIES

par S.K. GODOUNOV

Le problème de la construction des réseaux plans dans des domaines de forme compliquée, d'une façon automatique, se présente souvent dans différents problèmes de la mécanique des milieux continus.

Dans l'ouvrage [1] nous avons utilisé des réseaux mobiles de nature simple dans les coordonnées Euleriennes, pour résoudre divers problèmes de l'hydrodynamique (voir figure 1 de notre ouvrage [1] et figure 2 de l'ouvrage [2] où l'on faisait des calculs d'après le schéma de [1]). La figure 3 présente un réseau dans un domaine assez compliqué qu'il n'est plus possible de construire d'une manière élémentaire. Un tel réseau construit par l'ordinateur, selon une des méthodes décrites dans ce qui suit, a été utilisé pour la solution des équations elliptiques.

Le problème de la construction automatique des réseaux, dans les domaines à frontières curvilignes, n'est pas un problème nettement posé. Il se rattache plus au courant des idées de la cybernétique qu'à la théorie des méthodes numériques.

On ne connaît que deux publications concernant ce problème : celle de Winslow [3] 1966 et le nôtre [4] 1967. Un exemple de réseau calculé par Winslow est présenté sur la figure 4.

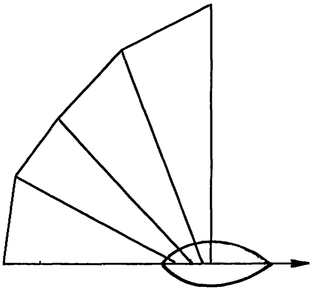


Figure 1

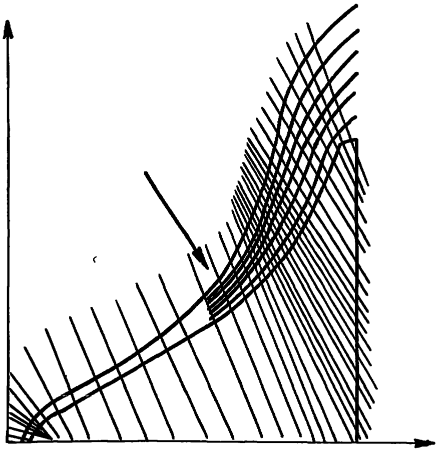


Figure 2

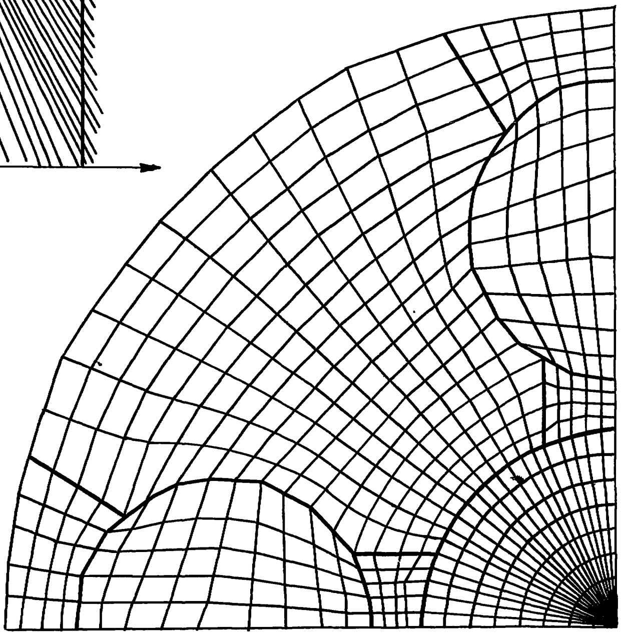


Figure 3

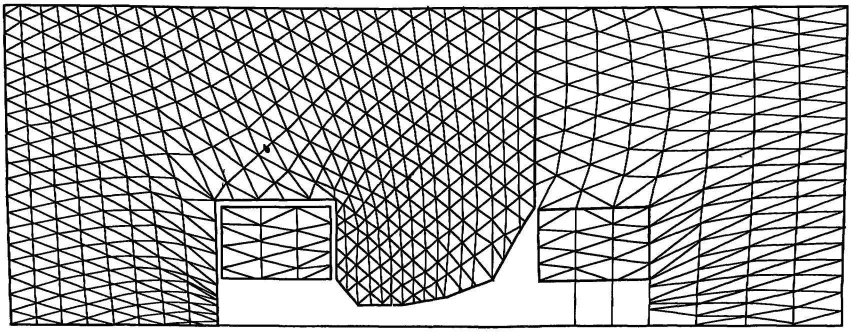


Figure 4

Notre première méthode a été fondée sur les méthodes numériques des applications conformes. Supposons qu'on doive construire un réseau de façon que les points frontières du réseau soient fixés sur la frontière du domaine. La figure 5 montre la solution de ce problème dans un quadrilatère curviligne réalisée par l'ordinateur selon la méthode suivante : on doit marquer un ensemble fini des points situés sur la frontière. Sur les arêtes opposées le nombre des points marqués doit être le même.

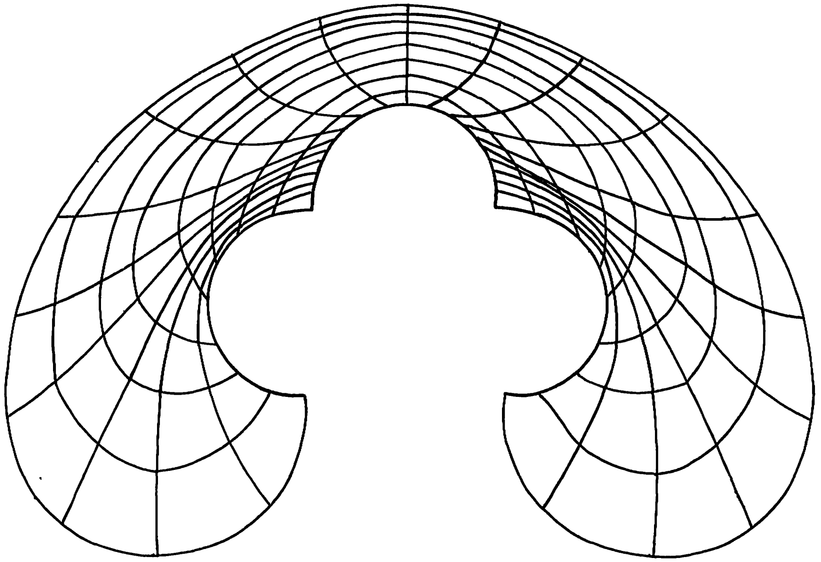


Figure 5

On trouve un rectangle qui admet une représentation conforme sur le domaine donné (l'image des sommets du rectangle étant les sommets du domaine). Alors des images des points marqués de la frontière du domaine sont certains points de la frontière du rectangle.

On n'a qu'à lier par des droites les points opposés du rectangle pour le recouvrir par un réseau rectiligne.

Le réseau désiré sur le domaine initial est alors l'image du réseau rectiligne.

D'autres exemples des réseaux construits après le schéma en question seront montrés au cours du rapport.

La méthode de Winslow pour résoudre ce problème (la construction des réseaux à points frontières marquès) est basée sur la construction d'une certaine application $u = u(x\ ,y)\ , v = v(x\ ,y)$ du carré unitaire, à réseau rectangulaire régulier, sur le domaine considéré. L'application est choisie en minimisant la fonctionnelle :

$$\int_0^1 \int_0^1 \frac{u_x^2 + u_y^2 + v_x^2 + v_y^2}{u_x v_y - u_y v_x} \quad dx\, dy$$

L'application doit réaliser une correspondance donnée à l'avance des points marqués de la frontière. Cette fonctionnelle est invariable par rapport aux transformations conformes du plan (u, v). Nous avons modifié la méthode de Winslow introduisant dans le carré (x, y) un réseau rectiligne oblique de même nature que dans l'ouvrage [4]. La figure 6 donne l'exemple du réseau calculé par cette méthode. Les figures 7, 8 représentent un domaine où la construction du réseau

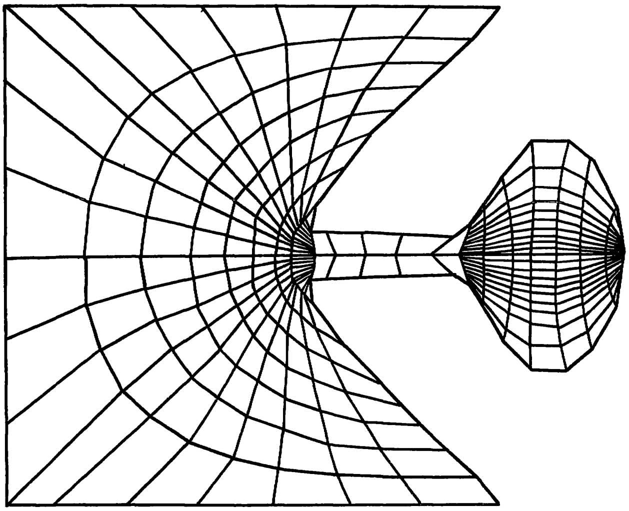


Figure 6

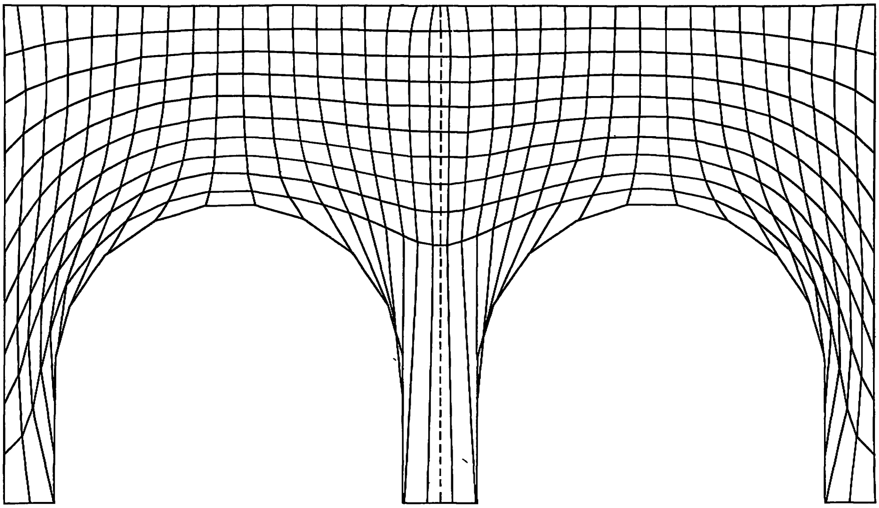


Figure 7

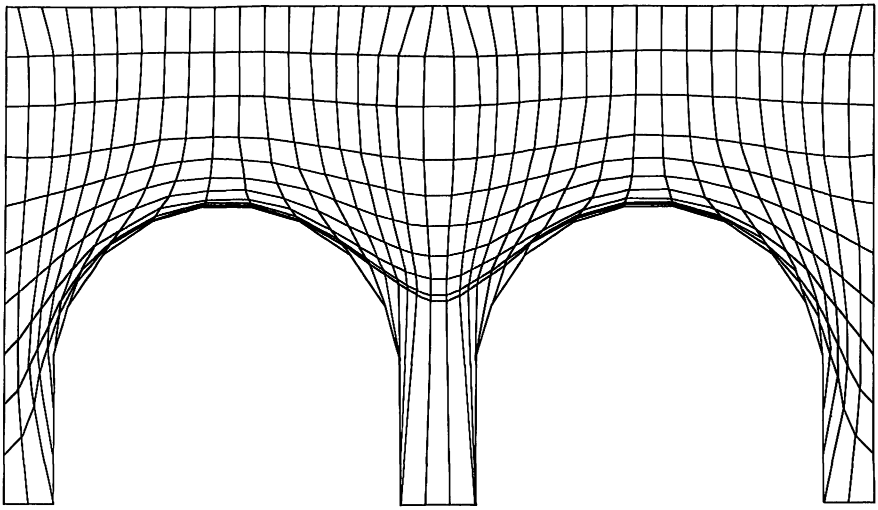


Figure 8

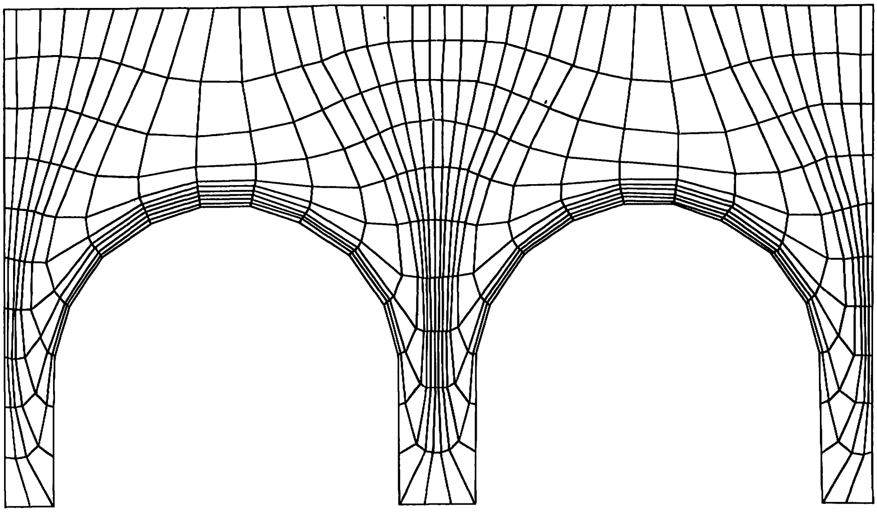


Figure 9

par la modification de la méthode Winslow n'a pas réussi. Un réseau satisfaisant pour le même domaine est représenté sur la figure 9. Il est obtenu à l'aide de l'application $u = u(x, y)$, $v = v(x, y)$ du réseau régulier du carré unitaire. Cette application minimise la fonctionnelle :

$$\int\int \left[\frac{1}{a} \; (u_x^2 + v_x^2 + \epsilon a_x^2) + a(u_y^2 + v_y^2 + \epsilon a_y^2) \right] dx\, dy$$

$a = a(x, y)$ étant un paramètre "de régularisation" déterminé au cours de la minimisation de la fonctionnelle ; $\epsilon > 0$ étant assez petit sert à rendre correct le problème qui ne l'est pas quand $\epsilon = 0$.

LITTERATURE

[1] Годунов С. К., Забродин А. В., Прокопов Г. П. — Разностная схема для двумерных нестационарных задач газовой динамики и расчёт обтекания с отошедшей ударной волной. Журнал вычислительной математики и математической Физики, 1961, т. 1, № 6.

[2] Taylor T.D., Masson B.S. — *Application of the unsteady numerical method of Godunov to computation of supersonic flows past bell shaped bodies* (Preprint of Northrop Corporate Laboratories, Hamthorne, California, U.S.A., 1969).

[3] Winslow A.M. — Numerical solution of the quasilinear Poisson equation in a nonuniform triangle mesh., *Journ. Comp. Phys.*, 1, 2, 1966, p. 149-172.

[4] Годунов С. К., Прокопов Г. П. — О расчётах конформных отображений и построении разностных сеток, Журнал Вычислительной Математики и Математической Физики, 1967, т. 7, № 5.

Computing Center of the Siberian Branch
of the USSR Academy of Sciences
Novosibirsk 90 (URSS)

Actes, Congrès intern. Math., 1970. Tome 3, p. 105 à 113.

MATHEMATICAL PROBLEMS ARISING IN PLASMA PHYSICS

by Harold GRAD

1. Introduction.

We present a selection of problems, mainly as mathematical conjectures, which are suggested by questions arising in plasma physics ; (for a physical survey of this field, see [1]). Those parts of the subject with an appreciable mathematical structure are primarily associated with magneto-fluid dynamics [2] or with guiding center theory [3]. Magneto-fluid dynamics is formulated as a system of differential equations. Guiding center theory only occasionally reduces to differential equations, but we shall concentrate on these cases.

Our plan will be to formulate a number of open questions and conjectures (and occasional theorems) which we believe are likely to develop into significant mathematical structures. We have found that problems which arise physically are rarely formulated, *ab initio*, as partial differential equations ; if as differential equations, they are rarely recognizable as elliptic, hyperbolic, or parabolic ; even when standard in type, the auxiliary data (boundary, initial, etc.) are frequently posed in non-standard forms. Some of these problems are accessible to relatively minor modifications of standard techniques, while others will certainly require development of new methods. Our emphasis will be on problems which are not too far removed from current mathematical practice.

One class of problems is formulated variationally. Unusual questions arise from non-positive variational functions associated with non-elliptic Euler equations ; also from non-local boundary conditions. In contrast to the conventional situation where weak solutions are introduced for mathematical convenience, we find cases where the equations (and the physical problem) insist that any solutions must be weak.

From the point of view of differential equations, unusual features arise from composite or mixed type and in non-simple topology. With regard to domain of dependence, an elliptic equation is global and a hyperbolic equation is local. These distinctions become blurred in a composite system, and even more so in a toroidal domain where the periodic coordinate is timelike.

2. Magnetic Diffusion.

The diffusion of a magnetic field through a solid conductor (skin effect) is governed by a vector diffusion equation,

$$\frac{\partial \mathbf{B}}{\partial t} + \operatorname{curl} \operatorname{curl} \mathbf{B} = 0\,, \qquad \operatorname{div} \mathbf{B} = 0. \tag{2.1}$$

In two dimensions **B** has a stream function which satisfies the simple diffusion equation,

$$\frac{\partial\psi}{\partial t} = \Delta\psi. \tag{2.2}$$

In a conducting *fluid,* the pressure balance between plasma and field requires that current contours, $\Delta\psi$ = const., coincide with flux contours, ψ = const., or

$$\Delta\psi = f(\psi\,, t). \tag{2.3}$$

This constraint is, in general, incompatible with (2.2). But in a fluid, (2.2) is replaced by

$$\frac{\partial\psi}{\partial t} + \mathbf{u}\cdot\nabla\psi = \Delta\psi \tag{2.4}$$

and we ask the question, whether a velocity field can be found which maintains the pressure balance (2.3) as the field diffuses.

For simplicity, assume that the flux contours are a set of simple closed curves and introduce the area average

$$<\phi>_\psi = \oint_\psi \phi\, ds/|\nabla\psi| \Big/ \oint_\psi ds/|\nabla\psi|\,. \tag{2.5}$$

We impose the condition that there is no net flow through each flux contour,

$$\oint_\psi u_n ds = 0 \quad \text{or} \quad <u\cdot\nabla\psi> = 0\,, \tag{2.6}$$

and formulate the plausible conjecture :

In a given plane domain D, specify ψ and f on ∂D (e.g. $\psi = 0, f = 0$), also $f(\psi\,, 0) = g(\psi)$ initially. Then $\psi(x\,, y\,, t), f(\psi\,, t)$ and $\mathbf{u}(x\,, y\,, t)$ satisfying (2.3), (2.4), (2.6) are uniquely determined.

Differentiating the constraint (2.3),

$$\Delta(\partial\psi/\partial t) = f_t + f_\psi\ (\partial\psi/\partial t), \tag{2.7}$$

and assuming that for given $\psi(x\,, y)$ the operator

$$L = \Delta - f_\psi \tag{2.8}$$

is invertible, $L^{-1} = G$, we have

$$\frac{\partial\psi}{\partial t} = G(f_t) \tag{2.9}$$

Taking mean values, $f = \widehat{G}(f_t)$, where $\widehat{G}$ is the restriction (in domain and range) of G to functions of a single argument ψ, and formally inverting, we have

$$f_t = \widehat{G}^{-1}(f). \tag{2.10}$$

We interpret (2.10) as a nonlinear diffusion equation for $f(\psi\,, t)$. A simple energy inequality shows that, with L negative, if a solution exists, it decays to zero (uniqueness is more difficult).

We can separate off an auxiliary linear problem, to determine the properties of the (self-adjoint) restriction $\widehat{G}$ of G in terms of *given* contours ψ over which to average. This type of problem also arises in several other situations given below.

If the contours ψ = const. are not simple, we can interpret (2.6) as being evaluated over the entire set ψ = const., in which case $f(\psi)$ is single-valued. Or we can require $< \mathbf{u} \cdot \nabla\psi > = 0$ on each connected component of ψ = const., in which case (2.10) determines distinct values of f on each component.

The qualitative behavior when $\widehat{G}$ is non-negative [whether (2.10) is unstable, or not well posed, etc.] is of great interest and is not transparent.

3. Force Free Fields.

We turn next to a group of unusual variational and related differential equation problems and look first at the simplest, the force free field [4].

The admissibility condition div $\mathbf{B} = 0$ for a vector field $\mathbf{B}(x)$ can always be parametrized (locally) as $\mathbf{B}$ = curl $\mathbf{A}$ where $\mathbf{A}$ can be further restricted by one of the two side conditions

$$(3.1)\qquad \begin{aligned} &\text{(a) div } \mathbf{A} = 0 \\ &\text{(b) } \mathbf{A} \cdot \mathbf{B} = 0\,, \quad \text{or} \quad \mathbf{A} = \frac{1}{2}\,(\alpha\nabla\beta - \beta\nabla\alpha) \end{aligned}$$

(in neither case is $\mathbf{A}$ unique). An interior variation of the Dirichlet integral, $\int \frac{1}{2} B^2 dV$, gives the Euler equation (after a standard integration by parts)

$$(3.2)\qquad \begin{aligned} &\text{(a) } \mathbf{J} \equiv \text{curl } \mathbf{B} = 0 \\ &\text{(b) } \mathbf{J} \times \mathbf{B} = 0\,, \quad \text{or} \quad \mathbf{J} = \sigma(\alpha\,,\beta)\mathbf{B} \end{aligned}$$

Case (a) is harmonic, (b) is a force free field.

Turning to the boundary variation and assuming that the normal component, B_n, is prescribed, we recover no information in case (a) and obtain the natural boundary condition $J_n = 0$ in case (b) which, together with the Euler equation implies $\mathbf{J} = 0$ in the domain. In other words, the different parametrizations (a) and (b) give a different split into interior and boundary variations, but eventually yield the same conclusion.

However, in a tubular domain (Fig. 1), we find that a force free field, $\mathbf{J} \times \mathbf{B} = 0$, renders Dirichlet's integral stationary if, in addition to the boundary condition on B_n (say $B_n = 0$ on S_0, $B_n > 0$ on S', $B_n < 0$ on S'') we fix the ends (on S' and S'') of each magnetic line, (compatible with the given B_n).

For example, taking $\mathbf{B} = \nabla\alpha \times \nabla\beta$, [case (b) in (3.1)], the complete boundary condition is that the magnetic line coordinates α and β are specified at both ends. Fixing only B_n allows α and β to be varied to any α', β' satisfying

$$\partial(\alpha\,,\beta)/\partial(\alpha'\,,\beta') = 1\,.$$

In the original formulation, in terms of **B**, giving the end connection is a global condition ; in terms of (α , β) it is a local boundary condition. We also remark that the *composite* Euler equation, $\mathbf{J} \times \mathbf{B} = 0$, (with real characteristics as well as an elliptic cone) is exempt form the standard requirement that a variational equation be elliptic because the variational function $\int \frac{1}{2} |\nabla\alpha \times \nabla\beta|^2 \, dV$ is not positive definite.

We not only conjecture the existence (subject to suitable regularity) of the force free field just described, but we predict that the proofs in the following cases should follow from standard analytic estimates :

(1) In the tubular domain, figure 1, $\mathbf{J} = \lambda\sigma(\alpha , \beta)\mathbf{B}$ should be uniquely solvable for given smooth $\sigma(\alpha , \beta)$ and sufficiently small λ [to pin down $\sigma(\alpha , \beta)$, α and β are fixed at one end]. One might iterate curl $\mathbf{B}_{\nu+1} = \lambda\sigma_\nu \mathbf{B}_\nu$, $\mathbf{B}_\nu \cdot \nabla\sigma_\nu = 0$. Giving $\sigma(\alpha , \beta)$ is equivalent to specifying J_n at one end in addition to the usual elliptic data on B_n.

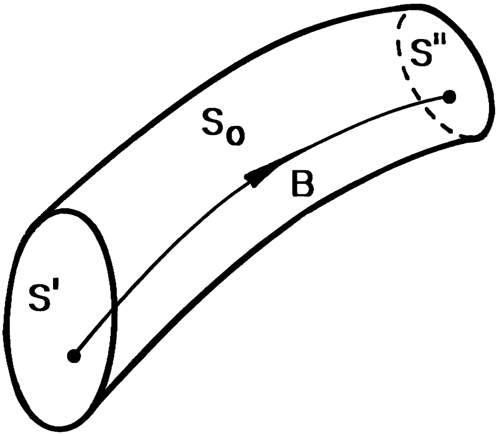


Figure 1 : Tubular domain

(2) For sufficiently small λ, there is a one-to-one correspondence between the specification of $\sigma(\alpha , \beta)$ and the specification of the end-mapping,

$$(\alpha' , \beta') \leftrightarrow (\alpha'' , \beta'') \quad \text{from} \quad S' \text{ to } S'' .$$

(3) $\int \frac{1}{2} |\nabla\alpha \times \nabla\beta|^2 dV$ is an absolute minimum for sufficiently small λ. One can see directly that the second variation is positive.

In two dimensions and axial symmetry, proofs of these and a number of other results are direct.

Many further conjectures have been made (in some special cases, proved) for the more general static equilibrium [4], [5], [6]

$$\mathbf{J} \times \mathbf{B} = \nabla p \qquad (3.3)$$
$$\operatorname{div} \mathbf{B} = 0$$

for which the variational function is $\int \left(\frac{1}{2}\mathbf{B}^2 - p\right) dV$ where $p = p(\alpha, \beta)$ is a given function. This type of problem requires somewhat more refined estimates than the force free field.

Still more generally, the convex variational function

$$\int f(B, \alpha, \beta)\, dV \qquad (3.4)$$
$$\partial f/\partial B > 0, \quad \partial^2 f/\partial B^2 > 0$$

yields the equations of static equilibrium for an anisotropic, guiding center plasma [3], [7],

$$\mathbf{J} \times \mathbf{B} = \operatorname{div} \mathfrak{P} \qquad (3.5)$$
$$P_{ij} = p_{\|} B_i B_j / B^2 + p_{\perp}(\delta_{ij} - B_i B_j / B^2)$$

on identifying

$$p_{\|} = \frac{1}{2} B^2 - f \qquad (3.6)$$
$$p_{\perp} = B(\partial f/\partial B) - \frac{1}{2} B^2 - f$$

Problems in which the magnetic line mapping, S' to S'', is given or J_n is specified at one end are formulated as for the force free field. The variational function $f(B, \alpha, \beta)$ is exactly what is determined by specifying the microscopic guiding center particle distribution on each magnetic line [3].

4. Toroidal problems.

An entirely new complication arises in a toroidal domain from the presence of real characteristics which do not intersect a spacelike manifold on which to assign initial values. In the purely elliptic problem, curl $\mathbf{B} = 0$, div $\mathbf{B} = 0$, a real characteristic is introduced when we ask whether there exist flux surfaces, $\mathbf{B} \cdot \nabla \psi = 0$. The scalar pressure equilibrium (3.3) can be written in the form

$$\operatorname{div} \mathbf{B} = 0 \qquad \mathbf{B} \cdot \nabla p = 0$$
$$\operatorname{curl} \mathbf{B} = \mathbf{J} \qquad \mathbf{B} \cdot \nabla \zeta = 1 \qquad (4.1)$$
$$\mathbf{J} = \nabla \zeta \times \nabla p$$

in which the elliptic cone and the doubly counted real characteristic are visible.

In the case of the harmonic vector, curl $\mathbf{B} = 0$, it has long been known that a complete set of magnetic surfaces does not exist in general in a toroidal domain with $B_n = 0$. For a smooth perturbation of an elementary geometry which does have a full set of surfaces, surfaces will exist on a set of positive measure [8], [9] separated by a dense set of gaps with different topology (islands) or ergodic behavior.

With curl $\mathbf{B} = \mathbf{J}$, one can always adjust $\mathbf{J}$ so as to provide surfaces. For example, given an arbitrary smooth set of toroidal surfaces, $p(x, y, z) =$ const., also two arbitrarily given sets of periods, say fluxes $\psi_1(p)$ and $\psi_2(p)$, or currents

$$\oint_{C_i} \mathbf{B} \cdot d\mathbf{x} = I_i(p) \,,$$

we can easily construct a field with div $\mathbf{B} = 0$ and $\mathbf{J} \times \mathbf{B}$ parallel to ∇p (this construction is unique). But the more stringent requirement, $\mathbf{J} \times \mathbf{B} = \nabla p$, eliminates any possibility of smooth solutions except in highly symmetric geometries [10]. For example, take a known harmonic field with islands ; in the same domain it is clearly impossible to find a family of solutions of $\mathbf{J} \times \mathbf{B} = \lambda \nabla p$ which converges to the harmonic field as $\lambda \to 0$ since a smooth function λp implies the existence of a complete set of magnetic surfaces for every λ. The same can be shown for an approach to a limiting force free field. It is clear, in particular, that studies of *analytic* solutions of $\mathbf{J} \times \mathbf{B} = \nabla p$ (e.g. as in [11]) are inappropriate except in symmetric geometries.

There is a large literature of formal physical calculations, e.g. of stability, which are based on the *assumption* of the existence of a smooth equilibrium. There are also formal parameter expansions which allow term by term calculations of nonexistent equilibria. An important question is how to interpret or otherwise salvage this literature.

Usually, when one is presented with nonexistence in a physical problem, one turns to more elaborate models. But, in this case, very convincing arguments can be given to the effect that existence will not be recouped by adding resistivity, finite Larmor radius, fluid flow, etc. and in fact the nonexistence of steady states is a correct *physical* prediction [10].

One possibility for reinterpreting formal stability calculations is to show that if a truncated pseudo-equilibrium expansion is taken as an initial state, the time dependent system (which is symmetric hyperbolic and well-posed) will have a solution which, in some sense, does not stray very far for some time. Some indications can be given of this possibility [12].

A more interesting mathematical resolution of this difficulty is to look for weak solutions of the static equilibrium system. It is easily seen that to exist in any sense, either p must be flat in the gaps (e.g. $\nabla p = 0$ in an interval surrounding each rational value of the rotation number), or else J must be unbounded in the same regions – in other words, either p or B is not continuously differentiable.

Basing our conjecture on formal variational analysis and qualitative properties of the differential system (e.g. that on a magnetic surface, the vector field B satisfies an elliptic equation) [4], we need to specify two profiles (two periods on each surface), e g.

$$\text{(4.2)}\qquad \begin{aligned}&\text{(a) } \psi_1(p)\ ,\ \psi_2(p) \quad \text{or} \quad p(\psi_1)\ ,\ \psi_2(\psi_1)\\ &\text{(b) } \psi_1(V)\ ,\ \psi_2(V)\\ &\text{(c) } I_1(V)\ ,\ I_2(V)\end{aligned}$$

where $\psi = \int \mathbf{B}\cdot d\mathbf{S}$ on an open 'cut' surface and $I = \oint \mathbf{B}\cdot d\mathbf{x}$ on a closed curve. The respective variational functions are

$$\text{(4.3)}\qquad \begin{aligned}&\text{(a) } \int\left[\frac{1}{2}B^2 - p(\psi)\right]dV\\ &\text{(b) } \int\frac{1}{2}B^2\,dV\\ &\text{(c) } \int\frac{1}{2}B^2dV - \int(I_1d\psi_1 + I_2d\psi_2)\end{aligned}$$

The first problem is to formulate the constraints (4.2) so that they are meaningful for an admissible vector field B which has gaps in its surfaces. Consider a class of vector fields in D each with a set Σ of positive measure of closed surfaces S of the given topology. To each surface S is assigned the volume V of its interior and the measure $\nu < V$ of the intersection of Σ with the interior of S. We can use $0 < V < 1$ or $0 < \nu < \nu_0$ as parameter spaces. The fluxes ψ_1, ψ_2 are continuous and (say) monotone on V ; p is continuous, with flat segments on V. On ν, p is continuous, but ψ is not. We can impose (for an admissible field) and conjecture (for an extremal) that the rotation number, $\alpha = d\psi_2/d\psi_1$, is continuous on ν. The current, I, is taken to be absolutely continuous on ν if, within the gaps, $\mathbf{B}$ is harmonic, curl $\mathbf{B} = 0$; or I is singular on ν (but continuous on V) if $\mathbf{B}$ is force free, $\mathbf{J} = \sigma\mathbf{B}$, in the gaps. These are different but equally valid specifications. Alternatively, we have $dp = 0$, $dI = 0$, $d\sigma = 0$ on the complement of Σ.

With these specifications, we can verify that

$$\int\frac{1}{2}B^2dV = \frac{1}{2}\int(I_1d\psi_1 + I_2d\psi_2)$$

exists as a Stieljes integral. The absolutely continuous and singular parts of the measures $d\psi_i$ on ν correspond to Σ and its complement respectively. Side conditions such as $\psi_i(V)$ or $I_i(V)$ would be considered to be given as a continuous function on $0 < V < 1$; a particular restriction to $0 < \nu < \nu_0$ (i.e. to the set of surfaces Σ) would depend on the specific admissible field $\mathbf{B}$. Also, instead of $p(\psi_1)$, we might specify a function $P(\psi_1) = \int \psi(\psi_1)d\psi_1$, continuous on $0 < V < 1$ but not on $0 < \nu < \nu_0$.

The weak form of div $\mathbf{B} = 0$ is of course $\int \mathbf{B}\cdot\nabla\phi dV = 0$, and for

$$\text{curl }(\mathbf{J}\times\mathbf{B}) = 0, \qquad \text{it is} \qquad \int\mathbf{B}\cdot(\mathbf{B}\cdot\nabla\mathbf{A})dV = 0$$

where div $\mathbf{A} = 0$ and ϕ and $\mathbf{A}$ have compact support.

The above description is, of course, only an indication of what may be necessary to give meaning to this unusual variational and differential system. One step that has been carried out is a proof of the uniqueness of the first variation from a known equilibrium [13]. Another is the case of a free boundary, in which the inverse problem can be solved for an analytic initial manifold provided that the initial rotation number is not too close to a rational [12]. From this result for the inverse problem, we conjecture that for the direct problem, instead of a continuum of free boundaries as in two dimensions (one for each specified area), there will be a set of allowable volumes of positive measure.

There is a tremendous assortment of steady flow problems in toroidal domains which present a combination of composite and mixed types (changes of type over the domain). Most such problems are not well posed (with a physical conclusion that only time dependent flows can be observed). In cases in which the degenerate real characteristic (magnetic line) opens up into a real cone, one would expect only trivial solutions since the domain of dependence is the entire domain rather than a 'gap'. But in the presence of so many possibilities, it seems likely that many more problems of mathematical interest will be found.

Acknowledgment.

This work was supported by the U.S. Atomic Energy Commission under contract AT(30 - 1) - 1480.

REFERENCES

[1] Grad H. — Frontiers of Physics Today : Plasmas, *Physics Today,* 22, 1969, p. 34.

[2] Chu C.K. and Grad H. — Magnetohydrodynamics, in *Research Frontiers in Fluid Dynamics,* ed. R.J. Seeger and G. Temple, John Wiley and Sons, 1964.

[3] Grad H. — The Guiding Center Plasma, *Proc. of the Symposia in Applied Mathematics,* New York, 1965, Vol. 18 (American Mathematical Society 1967), pp. 162-248.

[4] Grad H. and Rubin H. — Hydromagnetic Equilibria and Force-Free Fields, in *Second United Nations Conference on the Peaceful Uses of Atomic Energy,* Geneva, 1958, Vol. 31, pp. 190-197.

[5] Grad H. — *Proc. of the International Congress of Mathematicians,* Stockholm, 1962, pp. 560-583.

[6] Grad H. — *Phys. Fluids* 7, 1964, p. 1283.

[7] Grad H. — Magnetic Properties of a Contained Plasma, to appear in *Transactions of New York Academy of Sciences.*

[8] Arnold V.I. — *Usp. Math. Nauk.* 18, 1963, p. 13.

[9] Moser J. — *Nachr Akad. Wiss.* Gottingen, II. Math. Physik., Kl. I, 1962, p. 1.

[10] Grad H. — *Phys. Fluids* 10, 1967, p. 137.

[11] Arnold V.I. — *Comptes Rendus,* Paris, 261, (1965), p. 17.

[12] GRAD H. — *Problems in Magnetostatic Equilibrium,* New York University, Report MF-62, April 1970
[13] BINEAU M. — *Variational Equations of Hydromagnetic Equilibria,* to appear.

New York University
Courant Institute of Mathematical Sciences
251 Mercer Street,
New York, N.Y. 10012 (USA)

Actes, Congrès intern. Math., 1970. Tome 3, p. 115 à 122.

PROBLÈME AUX LIMITES INTERIEURES POUR L'EQUATION DE BOLTZMANN

par Jean-Pierre GUIRAUD

RESUME

Le problème aux limites intérieur pour l'équation de BOLTZMANN linéaire a été précédemment résolu. La solution obtenue appartient à un espace de HILBERT $L^2(\Omega, H)$ de fonctions de carré sommable relativement à $\underset{\sim}{x}$ (variable d'espace parcourant le domaine convexe Ω), à valeurs dans un espace de fonctions de carré sommable relativement à $\underset{\sim}{\xi}$ (variable décrivant l'espace des vitesses). Il s'agit d'une solution généralisée. Quelques propriétés de régularité de cette solution sont obtenues ici. Ces propriétés sont suffisantes pour traiter le problème aux limites intérieur, relatif à l'équation non linéaire, dans l'approximation des mouvements de faible amplitude.

SUMMARY

The interior boundary value problem for BOLTZMANN equation has been solved previously. This solution was obtained in a HILBERT space $L^2(\Omega, H)$ of square integrable functions relatively to the space variable $\underset{\sim}{x}$ (in the convex domain Ω), having their value in a space of square integrable functions of the velocity variable $\underset{\sim}{\xi}$. This is a generalized solution. Some regularity properties of this solution are given here. Theese properties are just sufficient for dealing with the weakly non linear full BOLTZMANN equation.

L'équation de Boltzmann intervient dans de nombreuses questions de physique mathématique et l'expression elle-même recouvre une réalité très diversifiée. C'est ainsi qu'il y a assez peu de points communs entre l'équation de Boltzmann qui tire son origine du problème du transport des neutrons et celle qui tire son origine de la théorie cinétique des gaz. C'est uniquement de la seconde qu'il sera question ici, soit, pour la fonction $F(t, \underset{\sim}{x}, \underset{\sim}{\xi})$

$$(1) \qquad \frac{\partial F}{\partial t} + \underset{\sim}{\xi} \cdot \underset{\sim}{\nabla} F = J(F, F) \quad ; \quad (t, \underset{\sim}{x}) \in R^1 \otimes \Omega \quad , \quad \underset{\sim}{\xi} \in \vec{E}^3 \ ;$$

où Ω désigne un domaine de E^3, espace usuel de la mécanique classique. La signification physique de l'équation (1), son origine, ainsi que les principales propriétés de l'opérateur de collision $J(F, F)$, se trouvent exposés, entre autres, dans un article de Grad [1958] et dans le livre de Cercignani [1969]. On ne s'intéresse ici qu'aux solutions qui sont de la forme

$$(2) \qquad F = F_M(1+f) \quad ; \quad F_M = n\,\omega = n\,(2\pi)^{-3/2} \; \exp\,(-|\underset{\sim}{\xi}|^2/2)\,,$$

étant convenu que n est une constante, que les unités ont été convenablement choisies et que f est petit en un certain sens. L'équation (1) devient alors

$$(3) \qquad \frac{\partial f}{\partial t} + \underset{\sim}{\xi}\cdot\underset{\sim}{\nabla} f + Lf = Q\,(f\,,f)\,,$$

où L et Q désignent respectivement un opérateur linéaire et un opérateur quadratique, agissant tous deux sur f, considéré comme fonction de $\underset{\sim}{\xi}$, et indépendamment de $\underset{\sim}{x}$. On désigne par $Q\,(f\,,g)$ la fonctionnelle bilinéaire associée à la fonctionnelle quadratique $Q\,(f\,,f)$.

Nous adopterons ici la forme de l'opérateur de collision qui correspond au modèle d'interaction entre molécules qui sont des sphères rigides. Grad [1963], puis Cercignani [1967 a] ont discuté, sous des conditions un peu plus générales, la nature de l'opérateur L qui s'écrit

$$(4) \qquad Lf = \nu\cdot f - Af\,,$$

où $\nu.$ désigne l'opérateur de multiplication scalaire par $\nu\,(|\underset{\sim}{\xi}|)$ et où A désigne un opérateur intégral. Tous deux agissent dans l'espace de Hilbert H associé au produit scalaire

$$(5) \qquad (f\,,g)_H = \int \omega f \bar{g}\, d\underset{\sim}{\xi} \quad , \quad (f,f) = \|f\|_H^2 \;.$$

L'opérateur A est borné dans H alors que l'opérateur $\nu.$ ne l'est pas en général puisque l'on a, dans le cas des sphères rigides,

$$(6) \qquad a(1 + |\underset{\sim}{\xi}|) \leqslant \nu \leqslant b\,(1 + |\underset{\sim}{\xi}|) \;;$$

en revanche, $\nu.$, A et L sont tous trois autoadjoints et L est non négatif. Le noyau de l'opérateur L est un sous-espace de dimension 5 et une base en est constitué par

$$(7) \qquad \psi_0 = 1 \;,\; \psi_1 = \xi_1 \;,\; \psi_2 = \xi_2 \;,\; \psi_3 = \xi_3 \;,\; \psi_4 = 6^{-1/2}(|\underset{\sim}{\xi}|^2 - 3) \;;$$

cette base est orthonormale. Si f est arbitraire dans H, l'on pose

$$(8) \qquad f = \sum_{a=0}^{4} \;(f\,,\psi_a)_H\; \psi_a \;+ Pf\,,$$

et l'on a l'inégalité suivante, capitale pour toute la théorie,

$$(9) \qquad (Lf\;,\;f)_H \geqslant \mu\|Pf\|_H^2 \quad , \quad \mu > 0\;.$$

Cette inégalité a été établie par Grad [1963] puis par Cercignani [1967 *a*], sous des hypothèses différentes, qui englobent en particulier le cas des sphères rigides.

Le problème aux limites a été traité par Cercignani [1967 b] et par Guiraud [1968 - 1970], mais uniquement pour l'équation de Boltzmann linéaire (obtenue en faisant $Q = 0$). Pao [1967] a traité l'équation complète, dans le cas faiblement non linéaire, pour la configuration spéciale du problème de Couette, mais sans prendre en compte les conditions aux limites réelles. Le problème aux valeurs initiales et aux limites a été traité par Guiraud [1968 b] comme application du théorème de Hille et Phillips et des résultats obtenus dans l'étude du problème aux limites. Scharf [1967 - 1969] d'une part, Fetz et Shen [1970] d'autre part, ont systématiquement appliqué la théorie des semi-groupes en traitant les conditions aux limites d'une manière abstraite. Le présent travail est destiné à présenter succinctement les résultats obtenus par Guiraud [1970], relativement au problème aux limites pour l'équation linéaire, et à annoncer des résultats plus récents (dont les démonstrations seront publiées ailleurs), relatifs au problème aux limites pour l'équation complète, dans le cas faiblement non linéaire.

Dans tout ce qui suit, le domaine Ω est convexe, borné, et son bord $\partial\Omega$ vérifie des conditions de Liapounoff ; la normale unitaire à $\partial\Omega$, pointant vers Ω, est désignée par $\underset{\sim}{n}$. Pour expliciter les conditions aux limites, il est commode de définir l'opérateur de trace $\mathcal{B}$ qui, à f, associe sa trace $\mathcal{B}f$ sur $\partial\Omega$, et d'utiliser les opérateurs $\mathcal{J}^{+}$ et $\mathcal{J}^{-}$ qui sont, en chaque point de $\partial\Omega$, les opérateurs qui, à $\mathcal{B}f$, associent ses restrictions $\mathcal{J}^{+}\mathcal{B}f$ et $\mathcal{J}^{-}\mathcal{B}f$ à $\underset{\sim}{\xi}\cdot\underset{\sim}{n} > 0$ et $\underset{\sim}{\xi}\cdot\underset{\sim}{n} < 0$, respectivement. Les conditions aux limites se traduisent alors par

$$(10) \qquad \mathcal{J}^{+}\mathcal{B}f = \mathcal{G}\mathcal{J}^{-}\mathcal{B}f + \Phi^{+},$$

Φ^{+} étant donné sur $\partial\Omega$. L'opérateur $\mathcal{G}$ est un opérateur linéaire assujetti à certaines conditions (Guiraud [1970]).

Le cadre fonctionnel, dans lequel se laisse développer naturellement la théorie de l'existence, est constitué, pour les fonctions f, par l'espace $L^{2}(\Omega\,,H)$, pour les traces $\mathcal{B}f$, par les espaces $L^{2}(\partial\Omega\,,H)$, $L^{2}(\partial\Omega\,,\widetilde{H}_{1})$ et $L^{2}(\partial\Omega\,,\widetilde{H}_{\nu})$, où $\widetilde{H}_{\rho}$ n'est autre que l'espace de Hilbert $L^{2}(\rho\, d\mu\,(\underset{\sim}{\xi}))$ avec $d\mu = |\underset{\sim}{\xi}\cdot\underset{\sim}{n}|\;\omega\, d\underset{\sim}{\xi}$, ρ désignant une fonction de $|\underset{\sim}{\xi}|$. Le produit scalaire dans $L^{2}(\Omega\,,H)$ est noté

$$((f\,,g)) = \int_{\Omega} (f\,,g)_{H}\, dx$$

et la norme correspondante est notée $|||f|||$. La norme dans l'espace $L^{2}(\partial\Omega\,,\widetilde{H}_{\rho})$ et le produit scalaire correspondant sont notés $[\mathcal{B}f\,,\mathcal{B}g]_{\rho}$ et $[\![\mathcal{B}f]\!]_{\rho}$.

Un premier problème aux limites

$$(11) \qquad \underset{\sim}{\xi}\cdot\underset{\sim}{\nabla}f + \nu f = \varphi \quad , \quad \mathcal{J}^{+}\mathcal{B}f = \Phi^{+}\,,$$

se résout trivialement par

$$(12) \qquad f = U\varphi + E\Phi^{+},$$

avec des opérateurs U et E dont la représentation est explicite. Pour traiter le problème aux limites complet

$$(13) \qquad \underset{\sim}{\xi}\cdot\underset{\sim}{\nabla}f + Lf = \varphi \quad , \quad \mathcal{J}^{+}\mathcal{B}f = \mathcal{G}\mathcal{J}^{-}\mathcal{B}f + \Phi^{+}\,,$$

il faut résoudre successivement deux problèmes d'inversion. Dans un premier temps l'on traite la première équation (11) avec la seconde condition (13) ; ce qui revient à chercher $f^{+} = \mathcal{J}^{+}\mathcal{B}f$, tel que l'on ait

$$(14) \qquad (1 - \mathcal{V})f^{+} = \mathcal{G}\mathcal{J}^{-}\mathcal{B}\,U\varphi + \Phi^{+}\;,\; \mathcal{V} = \mathcal{G}\mathcal{J}^{-}\mathcal{B}E\;,$$

moyennant quoi la solution du problème aux limites

$$(15) \qquad \underset{\sim}{\xi}\cdot\underset{\sim}{\nabla} f + \nu f = \varphi \quad , \quad \mathcal{J}^{+}\mathcal{B} f = \mathcal{G}\mathcal{J}^{-}\mathcal{B} f + \Phi^{+}$$

est donnée par

$$(16) \quad f = \mathcal{W}\varphi + \mathcal{E}\,\Phi^{+} = U\varphi + E(1-\mathcal{V})^{-1}\,\mathcal{G}\mathcal{J}^{-}\,\mathcal{B}U\varphi + E\,\Phi^{+} + E(1-\mathcal{V})^{-1}\,\Phi^{+}\,.$$

Dans un second temps, la recherche de la solution de (13) revient à inverser l'opérateur $(I - \mathcal{W}A)$, puisque (13) équivaut (en un sens généralisé) à

$$(17) \qquad (1-\mathcal{W}A)f = \mathcal{W}\varphi + \mathcal{E}\,\Phi^{+}.$$

Les deux théorèmes suivants ont été établis par Guiraud [1970].

THEOREME 1. — *Le domaine Ω est convexe, borné, et son bord $\partial\Omega$ vérifie des conditions de Liapounoff. La double inégalité* (6) *est satisfaite. L'opérateur $\mathcal{G}$ vérifie deux séries de conditions : d'une part, il satisfait aux relations*

$$(18) \qquad \begin{aligned} &\text{(i)} \ \ \mathcal{G}\,1^{-} = 1^{+}\,, \\ &\text{(ii)} \ \int \mathcal{G} f^{-}\,d\mu = \int f^{-}\,d\mu\,, \\ &\text{(iii)} \int d\mu\cdot\mathcal{G}\otimes\mathcal{G}\;|F^{-}|^{2} \geqslant \gamma \iint d\mu\otimes d\mu\;|F^{-}|^{2}\,, \end{aligned}$$

avec une constante γ positive (strictement), alors que d'autre part il est borné de $L^{2}(\partial\Omega\,,\widetilde{H}_{1})$ dans lui-même, dans $L^{2}(\partial\Omega\,,H)$ et dans $L^{2}(\partial\Omega\,,\widetilde{H}_{\nu})$. Dans ces conditions l'opérateur $(I - \mathcal{V})$ est inversible dans $L^{2}(\partial\Omega\,,\mathcal{J}^{+}\widetilde{H}_{\nu})$ et son inverse, noté T, est borné dans cet espace. En outre, les opérateurs $\mathcal{W}$ et $\mathcal{E}$, définis par (16) *vérifient les estimations.*

$$(19) \qquad \begin{aligned} &|||\nu\,\mathcal{W}\varphi||| \leqslant \text{Const}\;|||\varphi||| \quad , \quad |||\underset{\sim}{\xi}\cdot\underset{\sim}{\nabla}\,\mathcal{W}\varphi||| \leqslant \text{Const}\;|||\varphi|||\,, \\ &|||\nu\;\mathcal{E}\;\Phi^{+}||| \leqslant \text{Const}\;\llbracket\Phi^{+}\rrbracket_{\nu} \quad , \quad |||\underset{\sim}{\xi}\cdot\underset{\sim}{\nabla}\;\mathcal{E}\;\Phi^{+}||| \leqslant \text{Const}\;\llbracket\Phi^{+}\rrbracket_{\nu}\,, \\ &\llbracket\mathcal{B}\;\mathcal{W}\varphi\rrbracket_{\nu} \leqslant \text{Const}\;|||\varphi||| \quad , \quad \llbracket\mathcal{B}\;\mathcal{E}\;\Phi^{+}\rrbracket_{\nu} \leqslant \text{Const}\;\llbracket\Phi^{+}\rrbracket_{\nu}. \end{aligned}$$

La condition (18 iii) doit être interprêtée dans le sens suivant : F^{-} est une fonction de $\underset{\sim}{\eta}$ et $\underset{\sim}{\zeta}$, définie pour $\underset{\sim}{\eta}\cdot\underset{\sim}{n} < 0$ et $\underset{\sim}{\zeta}\cdot\underset{\sim}{n} < 0$; $\mathcal{G}\otimes\mathcal{G}$ désigne le produit tensoriel de $\mathcal{G}$ par lui-même et définit un opérateur agissant sur les fonctions de $\underset{\sim}{\eta}$ et $\underset{\sim}{\zeta}$ $(\underset{\sim}{n}\cdot\underset{\sim}{\eta} < 0,\ \underset{\sim}{n}\cdot\underset{\sim}{\zeta} < 0)$ pour donner des fonctions de $\underset{\sim}{\xi}$ $(\underset{\sim}{\xi}\cdot\underset{\sim}{n} > 0)$; enfin $d\mu\otimes d\mu = |\underset{\sim}{\eta}\cdot\underset{\sim}{n}|\;|\underset{\sim}{\zeta}\cdot\underset{\sim}{n}|\;\omega(|\underset{\sim}{\eta}|)\;\omega(|\underset{\sim}{\zeta}|)\;d\underset{\sim}{\eta}\;d\underset{\sim}{\zeta}$.

Le second théorème annoncé ci-dessus est relatif à la résolution de l'équation (17). On rencontre quelques difficultés dans cette résolution. D'abord, même lorsque A est compact comme opérateur agissant dans H, $\mathcal{W}A$ ne l'est pas dans $L^{2}(\Omega\,,H)$ et cela interdit d'appliquer l'alternative de Fredholm sous sa forme abstraite. Il faut rechercher une méthode indirecte. Soit $\mathcal{L} = \underset{\sim}{\xi}\cdot\underset{\sim}{\nabla} + L$ l'operateur qui intervient dans l'équation de Boltzmann et soit $\mathcal{L}^{*} = -\underset{\sim}{\xi}\cdot\underset{\sim}{\nabla} + L$ son adjoint formel ; soit, de même $\mathcal{G}^{*}$ l'adjoint de $\mathcal{G}$ et posons $\mathcal{L}^{*} = \mathcal{M}^{*} - A$, $\mathcal{L} = \mathcal{M} - A$, puis définissons l'espace V^{*} comme suit

$$V^* : \{ f \mid f \in L^2(\Omega , H) \ , \ \mathcal{M}^* f \in L^2(\Omega , H) \ , \mathcal{T}^- \mathcal{B} f = \mathcal{G}^* \mathcal{T}^+ \mathcal{B} f \} \tag{20}$$

on peut montrer (Guiraud [1968 - 1970]) que la condition nécessaire et suffisante pour que f soit solution de (17), c'est-à-dire pour que f soit solution généralisée de (15), avec φ donné dans $L^2(\Omega , H)$ et Φ^+ donné dans $L^2(\partial\Omega , \tilde{H}_\nu)$ est que la relation suivante ait lieu pour tout $h \in V^*$

$$((f \ , \ \mathcal{L}^* h)) = [\Phi^+ , \mathcal{T}^+ \mathcal{B} h]_1 + ((\varphi , h)) \tag{21}$$

Ce point étant acquis, la résolution de (17) s'effectue, grâce à (21), par la méthode des projections orthogonales dans $L^2(\Omega , H)$, pourvu que l'on dispose d'une estimation a priori dans l'espace V, analogue à V^* :

$$V : \{ f \mid f \in L^2(\Omega , H) \ , \ \mathcal{M} f \in L^2(\Omega , H) \ , \mathcal{T}^+ \mathcal{B} f = \mathcal{G} \mathcal{T}^- \mathcal{B} f \} . \tag{22}$$

Il est clair que les fonctions constantes font partie du noyau de l'opérateur $\mathcal{L}$; soit alors $\mathcal{X}$, le sous-espace de V formé des fonctions orthogonales aux constantes dans $L^2(\Omega , H)$, et soit Π l'opérateur de projection orthogonale sur $\mathcal{X}$, dans $L^2(\Omega , H)$, l'estimation a priori en question est la suivante :

$$|||\Pi f||| \leqslant \text{Const } |||\mathcal{L} f||| \quad , \quad f \in V . \tag{23}$$

Cette estimation, obtenue par Guiraud [1968 - 1970], peut être établie en utilisant (9), (18) et la version formelle du théorème H qui est due à Darrozes et Guiraud [1966]. Les hypothèses requises sont celles du théorème 1, complétées par (9) et une condition très peu restrictive, portant sur l'opérateur $\mathcal{G}^*$, qui est formulée ci-dessous.

HYPOTHESE H^* — On peut trouver un système de 4 fonctions $\Phi_\alpha(\underset{\sim}{x} , \underset{\sim}{\xi})$ $\alpha = 1, 2, 3, 4$, définies pour $\underset{\sim}{x} \in \partial\Omega$, continûment différentiables sur $\partial\Omega$, telles que l'on ait, en chaque point de $\partial\Omega$

$$(\psi_\alpha , \underset{\sim}{\xi} \cdot \underset{\sim}{n} \Phi_\beta)_H = \delta_{\alpha\beta} \quad , \quad \Phi_\beta^- = \mathcal{G}^* \Phi_\beta^+ \quad ; \ \alpha = 0, 1, 2, 3, 4 \ ; \beta = 1, 2, 3, 4. \tag{24}$$

On peut alors établir le second des deux théorèmes annoncés ci-dessus.

THEOREME 2. — *Les hypothèses sont celles du théorème 1, complétées par* (9) *et par l'hypothèse H^*. Le problème aux limites*

$$\underset{\sim}{\xi} \cdot \underset{\sim}{\nabla} f + L f = \varphi \quad , \quad \mathcal{T}^+ \mathcal{B} f = \mathcal{G} \mathcal{T}^- \mathcal{B} f + \Phi^+ \ , \tag{25}$$

avec φ donné dans $L^2(\Omega , H)$ et Φ^+ donné dans $L^2(\partial\Omega , \tilde{H}_\nu)$, n'est résoluble que si la condition de compatibilité suivante est satisfaite

$$((\varphi , 1)) + [\Phi^+ , 1]_1 = 0. \tag{26}$$

Lorsque cette condition est remplie, il existe une famille à un paramètre de solutions généralisées de (15), *qui sont des solutions de* (17), *et qui sont données par*

$$f = \mathcal{C} \varphi + (1 + \mathcal{C} A) \ \mathcal{E} \Phi^+ + \text{const.} \tag{27}$$

L'opérateur $\mathcal{C}$ est borné dans $L^2(\Omega , H)$ alors que les opérateurs $\mathcal{E}$ et $(1 + \mathcal{C} A) \mathcal{E}$ sont bornés de $L^2(\partial\Omega , \tilde{H}_\nu)$ dans $L^2(\Omega , H)$.

Reprenons maintenant le problème aux limites non linéaire

$$\underset{\sim}{\xi} \cdot \underset{\sim}{\nabla} f + Lf = Q(f,f) + \varphi \quad , \quad \mathcal{J}^{+} \mathcal{B} f = \mathcal{G} \mathcal{J}^{-} \mathcal{B} f + \Phi^{+} \, , \tag{28}$$

et opérons formellement, nous obtenons

$$f = \mathcal{C}\varphi + (1 + \mathcal{C}A)\ \mathcal{E}\ \Phi^{+} + \mathcal{C}Q(f,f) + \text{Const} \, , \tag{29}$$

de telle sorte que nous sommes amenés à rechercher les points fixes de l'application

$$f \to \mathcal{F}(f) = \mathcal{C}\varphi + (1 + \mathcal{C}A)\ \mathcal{E}\ \Phi^{+} + \mathcal{C}Q(f,f). \tag{30}$$

Soit $\mathcal{N}(f)$ et $\mathcal{N}(\Phi^{+})$, deux normes, convenablement choisies, telles que l'on ait

$$\begin{gathered} \mathcal{N}(\mathcal{C}\varphi) \leqslant \alpha\ \mathcal{N}(\varphi) \ ; \\ \mathcal{N}((1 + \mathcal{C}A)\ \mathcal{E}\,\Phi^{+}) \leqslant \beta\ \hat{\mathcal{N}}(\Phi^{+}) \, , \\ \mathcal{N}(\mathcal{C}Q(f,f)) \leqslant \gamma\,(\mathcal{N}(f))^{2} \ , \end{gathered} \tag{31}$$

et que, plus généralement, si $(f,g) \to Q(f,g)$ désigne la fonctionnelle bilinéaire associée à $f \to Q(f,f)$, l'on ait aussi

$$\mathcal{N}(\mathcal{C}Q(f,g)) \leqslant \gamma\ \mathcal{N}(f)\ \mathcal{N}(g) \ ; \tag{32}$$

l'application $f \to \mathcal{F}(f)$ applique alors la boule $\mathcal{N}(f) < x_0$ dans elle-même, pourvu que $x_0 < x^{*}$, où x^{*} désigne la plus petite racine de l'équation

$$x = \alpha\,\mathcal{N}(\varphi) + \beta\,\hat{\mathcal{N}}(\Phi^{+}) + \gamma x^{2}. \tag{33}$$

Par ailleurs, cette application est contractante dans la même boule pourvu que $2\gamma x_0 < 1$. L'existence d'un point fixe de (29) est alors assurée pourvu que

$$4\gamma(\mathcal{N}(\varphi) + \hat{\mathcal{N}}(\Phi^{+})) < 1 \ , \tag{34}$$

et ce point fixe est unique dans la boule $2\gamma\ \mathcal{N}(f) < 1$.

Tel est le schéma de principe de la méthode appliquée par Pao [1967] dans le cas du problème de Couette avec la condition aux limites artificielle $\mathcal{J}^{+}\ \mathcal{B}f = \Phi^{+}$, avec Φ^{+} donné. La prise en compte des conditions aux limites réelles et le passage à un domaine Ω tridimensionnel conduisent à des difficultés qui nécessitent une technique de démonstration assez lourde, et seuls les résultats de l'étude seront donnés ici.

Les normes $\mathcal{N}$ et $\hat{\mathcal{N}}$ qu'il convient d'utiliser ont été indiquées par Grad [1963], soit

$$\begin{aligned} \mathcal{N}(f) &= \sup_{\underset{\sim}{x} \in \Omega \, ; \, \underset{\sim}{\xi} \in \vec{E}^{3}} \omega^{1/2}(|\underset{\sim}{\xi}|)\,(1 + |\underset{\sim}{\xi}|^{2})^{r/2}\,|f(\underset{\sim}{x}, \underset{\sim}{\xi})| = <f>_{r} < +\infty \, , \\ \hat{\mathcal{N}}(\Phi) &= \sup_{\underset{\sim}{x} \in \partial\Omega \, , \, \underset{\sim}{\xi} \in \vec{E}^{3}} \omega^{1/2}(|\xi|)\,(1 + |\underset{\sim}{\xi}|^{2})^{r/2}\,|\Phi(\underset{\sim}{x}, \underset{\sim}{\xi})| = <\Phi>_{r} < +\infty. \end{aligned} \tag{35}$$

Le point capital est évidemment constitué par la démonstration de (31) et (32). En ce qui concerne (32) l'on dispose des inégalités suivantes dues à Grad

$$<Q(f,f)>_{r-1} \leqslant \text{const} <f>_{r} \quad , \quad <Af>_{r+1} \leqslant \text{const} <f>_{r} \ , \tag{36}$$

et, grâce à (6), il suffit d'établir la seconde estimation (31) ainsi que la suivante

$$< \nu \mathcal{C} \varphi >_r \leqslant \alpha_1 < \varphi >_r . \tag{37}$$

Ce sont les démonstrations de ces inégalités qui sont techniquement longues.

Nous donnons ci-dessous, outre l'énoncé du résultat final, quelques énoncés intermédiaires qui peuvent présenter un intérêt intrinsèque.

THEOREME 3. – *Les hypothèses concernant Ω et $\mathcal{G}$ sont celles du théorème* 1 ; *en outre $\mathcal{G}$ est donné par un noyau $\Gamma(\underset{\sim}{x}, \underset{\sim}{\xi}, \underset{\sim}{\eta})$ qui vérifie*

$$\Gamma \leqslant \Gamma_1(|\underset{\sim}{\xi}|)\ \Gamma_2(|\underset{\sim}{\eta}|) , \tag{38}$$

$$\left\{ \begin{aligned} & \int \omega^{-1/2}(|\eta|)\ (1 + |\eta|^2)^{-r/2}\ \Gamma_2(|\underset{\sim}{\eta}|)\ d\underset{\sim}{\eta} < + \infty , \\ & \sup_{\underset{\sim}{\xi} \in E^3}\ \omega^{1/2}(|\underset{\sim}{\xi}|)\ \nu(|\underset{\sim}{\xi}|)\ (1 + |\underset{\sim}{\xi}|^2)^{r/2}\ \Gamma_1(|\underset{\sim}{\xi}|) < + \infty ; \end{aligned} \right. \tag{39}$$

alors, pour $r > \dfrac{3}{2}$, si $T\Psi^+ = f^+$ est la solution de $(1 - \mathcal{V})f^+ = \Psi^+$ l'on a

$$< \nu T \Psi^+ >_r \leqslant \text{Const} < \Psi^+ >_r , \tag{40}$$

avec une constante qui dépend, en particulier, de r.

THEOREME 4. – *Les hypothèses sont celles du théorème* 1 ; *en outre, $\partial\Omega$ est pourvu de rayons de courbure continus non nuls et $\mathcal{G}$ est défini par un noyau Γ qui vérifie* (38) *avec*

$$\begin{aligned} & \int \omega^{-1}(|\underset{\sim}{\eta}|)\ (\Gamma_2(|\underset{\sim}{\eta}|))^2\ d\underset{\sim}{\eta} < + \infty , \\ & \int \omega(|\underset{\sim}{\xi}|)\ \ (\Gamma_1(|\underset{\sim}{\xi}|))^2\ \ d\underset{\sim}{\xi} < + \infty ; \end{aligned} \tag{41}$$

dans ces conditions, l'estimation suivante a lieu

$$\sup_{\underset{\sim}{x} \in \Omega}\ \|(\mathcal{W} A)^4 \varphi\|_H(\underset{\sim}{x}) \leqslant \text{const}\ |||\varphi||| . \tag{42}$$

THEOREME 5. – *Les hypothèses sont toutes celles qui figurent dans l'un ou l'autre des théorèmes 3 ou 4 ; alors l'estimation suivante a lieu*

$$< \nu \mathcal{W} \varphi >_r \leqslant \text{const}\ (|||\varphi||| + < \varphi >_r) . \tag{43}$$

THEOREME 6. – *Les hypothèses sont celles du théorème 5 ; alors, pour $r > \dfrac{3}{2}$, l'on a*

$$\begin{aligned} & < \mathcal{C} \varphi >_{r+1} \leqslant \alpha < \varphi >_r , \\ & < (1 + \mathcal{C} A)\ \mathcal{E}\ \Phi^+ >_r \leqslant \beta < \Phi^+ >_r , \\ & < \mathcal{C} Q(f, g) >_r \leqslant \gamma < f >_r < g >_r . \end{aligned} \tag{44}$$

THEOREME 7. – *Les hypothèses sont celles du théorème 5, alors, pour* $r > \frac{3}{2}$, φ *et* Φ^+ *étant donnés de telle sorte que*

$$4\gamma\,(\alpha < \varphi >_r + \beta < \Phi^+ >_r) < 1\ , \tag{45}$$

l'application

$$f \to \mathcal{C}\varphi + (1 + \mathcal{C}A)\,\mathcal{E}\,\Phi^+ + \mathcal{C}Q(f,f)\,, \tag{46}$$

admet un point fixe $f = \mathcal{R}(\varphi\,,\Phi^+)$ *unique dans la boule*

$$2\gamma < f >_r < 1\ . \tag{47}$$

Dans ces conditions, le problème aux limites non linéaire

$$\underset{\sim}{\xi}\cdot\underset{\sim}{\nabla} f + Lf = Q(f,f) + \varphi \quad , \quad \mathcal{J}^+\mathcal{B}f = \mathcal{G}\,\mathcal{J}^-\,\mathcal{B}f + \Phi^+ \tag{48}$$

admet une famille à un paramètre de solutions généralisées

$$f = \mathcal{R}(\varphi\ ,\ \Phi^+) + \text{const.} \tag{49}$$

Remerciements.

Le présent travail a été mené à bien dans le cadre d'un contrat de collaboration extérieure accordé par l'O.N.E.R.A., que l'auteur tient à remercier ici.

REFERENCES

[1] CERCIGNANI C. — *Physics of Fluids,* 10, 1967 (a), p. 2097-2104. *Journal of Mathematical Physics,* 8, 8, 1967 (b), p. 1653-56. *Mathematical methods in kinetic theory,* 1969 (b), Plenum press.

[2] FETZ J. et SHEN S.F. — *Communication au 7e symposium sur la dynamique des gaz raréfiés,* Pise, 29 juin-3 juillet 1970.

[3] GRAD H. — Volume XII de *Handbuch der Physik,* Springer, 1958, p. 205-294; dans *Rarefied Gas Dynamics,* Editeur Laurmann, Academic Press, 1963, p. 26-59; dans Applications of non linear partial differential equations in mathematical physics, *Proceedings of S.I.A.M.,* 1965, Vol. XVII., p. 154-183.

[4] DARROZES J.S. et GUIRAUD J.P. — *C.R. Ac. Sci. T.,* 262 A, 1966, p. 1368-71.

[5] GUIRAUD J.P. — Journal de Mécanique T. 7, n° 2, 1968 (a), p. 171-203; *O.N.E.R.A.* N.T. 1968 (b), p. 132; *Journal de Mécanique* T. 9, n° 3, 1970, à paraître.

[6] SCHARF C. — *Helvetica Physica Acta* 40, n° 7, 1967, p. 929-45; *Helvetica Physica Acta* 42, n° 1, 1969, p. 5-22.

[7] PAO Y. — *Journal of Mathematical Physics* 8, 9, 1967, p. 1893-98.

Université de Paris VI, UER 49
Quai St Bernard, Paris 05
ou
ONERA
29, Avenue de la Division Leclerc
92 - Chatillon-sous-Bagneux
France

Actes, Congrès intern. Math., 1970. Tome 3, p. 123 à 130.

ON THE MATHEMATICAL FOUNDATION OF SHELL THEORY

by W.T. KOITER

1. Introduction.

Since the birth of a first reasonably satisfactory, if admittedly approximate two-dimensional linear theory of thin shells at the hands of Love more than eighty years ago [1], numerous attempts have been made to strengthen the foundations of shell theory. An adequate review of the enormous literature is impossible here because of lack of space. We refer to [2] for a survey of recent progress.

It is the purpose of the present paper to provide as sound and concise a foundation of the theory as possible, by means of concrete estimates of the error of the stress distribution obtained from linear shell theory as an approximation to the solution of the actual three-dimensional problem. The present analysis is based on general estimates of the errors of approximate solutions in elasticity theory [3]. The argument follows essentially the lines developed in [4], but it includes now substantial improvements due to Danielson [5] ([1]).

2. Basic error estimates of the linear theory of elasticity.

Let $\underline{\tau}$ denote an arbitrary distribution of the symmetric stress tensor in a three-dimensional elastic body. The associated complementary elastic energy is a positive definite homogeneous quadratic functional $C_2[\underline{\tau}]$ of this stress field.

Let $\underline{\sigma}$ denote the unique actual solution for the stress distribution in the body for the linear boundary value problem of the theory of elasticity, and let $\underline{u}$ denote the associated displacement field. The existence of this solution is ensured under some weak conditions on the shape of the body and the boundary data, and the displacement field is also unique, except for a possible additional rigid-body displacement field in the absence of kinematic boundary conditions [e.g. 6].

Let $\underline{u}^0$ denote any kinematically admissible displacement field, i.e. a displacement field which is continuous and piecewise continuously differentiable, and which satisfies the kinematic boundary conditions. Let $\underline{\sigma}^0$ denote the stress field associated with $\underline{u}^0$ by the (linear) strain-displacement relations and Hooke's law.

Let $\underline{\sigma}^*$ denote any statically admissible stress distribution, i.e. a piecewise continuous and continuously differentiable distribution of the symmetric stress tensor which satisfies the equations of equilibrium in the interior of the body as well as the dynamic boundary conditions on its surface.

[1] The author is indebted to Professor Danielson for his permission to make use of his refined results before their independent publication.

For the actual solution $\underline{\underline{\sigma}}$ of the boundary value problem, and for any pair of a statically admissible stress distribution $\underline{\underline{\sigma}}^*$ and a stress distribution $\underline{\underline{\sigma}}^0$, associated by Hooke's law with a kinematically admissible displacement field u^0, Prager and Synge have established the basic equality [3]

$$C_2\left[\underline{\underline{\sigma}} - \frac{1}{2}(\underline{\underline{\sigma}}^0 + \underline{\underline{\sigma}}^*)\right] = \frac{1}{4} C_2\,[\underline{\underline{\sigma}}^0 - \underline{\underline{\sigma}}^*]. \tag{2.1}$$

This equation implies that $\frac{1}{2}(\underline{\underline{\sigma}}^0 + \underline{\underline{\sigma}}^*)$ may be regarded as an "approximation" to the actual solution, if the complementary energy associated with the stress difference $\underline{\underline{\sigma}}^0 - \underline{\underline{\sigma}}^*$ is "small". In this case we may also consider each of the stress fields $\underline{\underline{\sigma}}^0$ or $\underline{\underline{\sigma}}^*$ as an "approximation", in view of the inequalities

$$C_2\,[\underline{\underline{\sigma}} - \underline{\underline{\sigma}}^0] \leqslant C_2\,[\underline{\underline{\sigma}}^0 - \underline{\underline{\sigma}}^*]\,, \tag{2.2}$$

$$C_2\,[\underline{\underline{\sigma}} - \underline{\underline{\sigma}}^*] \leqslant C_2\,[\underline{\underline{\sigma}}^0 - \underline{\underline{\sigma}}^*]\,, \tag{2.3}$$

which follow easily from (2.1).

In other words, the root mean square error of the "approximate" solutions for the stress distribution $\frac{1}{2}(\underline{\underline{\sigma}}^0 + \underline{\underline{\sigma}}^*)$, $\underline{\underline{\sigma}}^0$ or $\underline{\underline{\sigma}}^*$ may be estimated in terms of the root mean square value of the stress difference $\underline{\underline{\sigma}}^0 - \underline{\underline{\sigma}}^*$. In the case of inequality (2.2) we have moreover

$$C_2\,[\underline{\underline{\sigma}} - \underline{\underline{\sigma}}^0] = P_2\,[\underline{u} - \underline{u}^0], \tag{2.4}$$

where the right-hand member represents the elastic energy as a functional of the displacement field $\underline{u} - \underline{u}^0$ which vanishes on the part of the surface where the displacement vector is specified. Rayleigh's principle may now be invoked to estimate the root mean square error of $\underline{u}^0$ as an "approximation" to the actual displacement field $\underline{u}$. This error estimate is again obtained in terms of the root mean square value of the stress difference $\underline{\underline{\sigma}}^0 - \underline{\underline{\sigma}}^*$.

3. Equations of linear shell theory.

Let $\underline{r}(x^\alpha)$, $\alpha = 1, 2$, denote the radius vector from a fixed origin in space to a generic point on the middle surface of the undeformed shell as a vector-valued function of the pair of Gaussian surface coordinates. The tangential base vectors are $\underline{a}_\alpha = \underline{r}_{,\alpha}$, where the comma preceding the subscript α denotes partial differentiation with respect to the coordinate x^α. The reciprocal base is defined by $\underline{a}_\alpha \cdot \underline{a}^\beta = \delta_\alpha^\beta$. The covariant and contravariant metric tensors of the middle surface are given by $a_{\alpha\beta} = \underline{a}_\alpha \cdot \underline{a}_\beta$ and $a^{\alpha\beta} = \underline{a}^\alpha \cdot \underline{a}^\beta$. These tensors are employed in lowering and raising indices of surface tensors. The determinant of the covariant metric tensor is denoted by a, the covariant alternating tensor by $\epsilon_{\alpha\beta}$. The normal to the middle surface is defined by $\underline{n} = \frac{1}{2}\,\epsilon^{\alpha\beta}\,\underline{a}_\alpha \times \underline{a}_\beta$. The second fundamental tensor is specified by $b_{\alpha\beta} = \underline{n} \cdot \underline{r}_{,\alpha\beta}$. Covariant surface differentiation with respect to a coordinate x^α is denoted by an additional subscript α preceded by a vertical stroke. All derivatives in the analysis are assumed to be continuous.

A point in shell space is identified by its distance z to the middle surface and by the surface coordinates of its projection on the middle surface. The shell faces $z = \pm \frac{1}{2} h$, where h is the constant shell thickness, are surfaces parallel to the middle surface. The coordinate z is orthonormal to the surface coordinates, and the components $g_{\alpha\beta}$ of the spatial metric tensor g_{ij} (with determinant g) reduce to $a_{\alpha\beta}$ at the middle surface. The edge of the shell is assumed to be a ruled surface formed by normals to the middle surface along an edge curve on this surface. Let $\underline{\nu}$ denote the unit vector in the tangent plane, normal to the edge curve and positive outwards. The positive sense on the edge curve is defined by the tangential unit vector $\underline{t} = \underline{n} \times \underline{\nu}$.

A deformation of the middle surface is described by the two-dimensional displacement field

$$(3.1) \qquad \underline{u}(x^{\kappa}) = u_{\alpha}\underline{a}^{\alpha} + w\underline{n} ,$$

where the surface vector u_{α} and invariant w are functions of the Gaussian surface coordinates. Apart from a rigid-body displacement, the deformation is specified completely by the differences between the first and second tensors $\bar{a}_{\alpha\beta}$, $\bar{b}_{\alpha\beta}$ in the deformed configuration and the similar undeformed tensors. Hence we employ as strain measures (1)

$$(3.2) \qquad \gamma_{\alpha\beta} = \frac{1}{2} (\bar{a}_{\alpha\beta} - a_{\alpha\beta}) , \quad \rho_{\alpha\beta} = \bar{b}_{\alpha\beta} - b_{\alpha\beta}.$$

The associated expressions in terms of the displacement components are [7]

$$(3.3) \qquad \gamma_{\alpha\beta} = \frac{1}{2} (u_{\alpha|\beta} + u_{\beta|\alpha}) - b_{\alpha\beta}w ,$$

$$(3.4) \qquad \rho_{\alpha\beta} = w_{|\alpha\beta} - b^{\kappa}_{\alpha}b_{\kappa\beta}w + b^{\kappa}_{\alpha}u_{\kappa|\beta} + b^{\kappa}_{\beta}u_{\kappa|\alpha} + b^{\kappa}_{\alpha|\beta}u_{\kappa}.$$

In shell theory we consider virtual deformations of the Kirchhoff-Love type, in which normals to the undeformed middle surface move to normals of the deformed middle surface without any change in length. They are *completely* specified by arbitrary variations of the first and second fundamental tensors, subject of course to the conditions of compatibility implied by (3.3) and (3.4). The internal virtual work per unit area of the middle surface in a Kirchhoff-Love type virtual deformation is therefore specified by an invariant expression

$$(3.5) \qquad n^{\alpha\beta}\,\delta\gamma_{\alpha\beta} + m^{\alpha\beta}\,\delta\rho_{\alpha\beta} ,$$

where $n^{\alpha\beta}$ and $m^{\alpha\beta}$ are *symmetric* tensors of stress resultants and stress couples respectively.

The external loads on the shell faces are reduced to statically equivalent loads acting in the middle surface, along the lines described in detail by Naghdi [8]. For the sake of brevity we omit body forces, and we assume that the reduction

(1) In our previous work [2, 4, 7] we have employed the notations $\bar{\rho}_{\alpha\beta}$, $\bar{n}^{\alpha\beta}$ and $\bar{m}^{\alpha\beta}$ for the same quantities written here without a bar.

of loads to the middle surface introduces no surface couples. The reduced loads per unit area of the middle surface are described by a surface vector p^α and a surface invariant p ; we assume that the distribution of normal loads over the two faces is such that p is of the same order of magnitude. The loads on the shell edge are likewise reduced to statically equivalent line loads along the edge curve on the middle surface, a force $\underline{N}$ and a couple $\underline{M}$, both per unit arc length, where the couple vector lies in the tangent plane. We write

$$\underline{N} = N_{(\nu)}\underline{\nu} + N_{(t)}\underline{t} + Q\underline{n} = N^\alpha \underline{a}_\alpha + Q\underline{n} \,, \tag{3.6}$$

$$\underline{M} = M_{(\nu)}\underline{\nu} + M_{(t)}\underline{t} = \epsilon_{\alpha\beta} M^\beta \underline{a}_\alpha \,. \tag{3.7}$$

From the principle of virtual work we now obtain by standard methods the equations of equilibrium [2, 4, 7, 8]

$$(n^{\beta\alpha} + b^\alpha_\kappa m^{\beta\kappa})|_\beta + b^\alpha_\kappa m^{\beta\kappa}|_\beta + p^\alpha = 0 \,, \tag{3.8}$$

$$-m^{\alpha\beta}|_{\alpha\beta} + c_{\alpha\beta}m^{\alpha\beta} + b_{\alpha\beta}n^{\alpha\beta} + p = 0 \,, \tag{3.9}$$

as well as the dynamic boundary conditions

$$(n^{\beta\alpha} + 2\, b^\alpha_\kappa m^{\beta\kappa})\nu_\beta = N^\alpha + b^\alpha_\kappa M^\kappa \,, \tag{3.10}$$

$$-m^{\alpha\beta}|_\alpha \nu_\beta - (m^{\alpha\beta}\nu_\alpha t_\beta)_{,s} = Q - M_{(\nu),s} \,, \tag{3.11}$$

$$m^{\alpha\beta}\nu_\alpha\nu_\beta = -M_{(t)} \,, \tag{3.12}$$

where the subscript s preceded by a comma denotes partial differentiation with respect to the arc length along the edge curve. We emphasize that equations (3.8) and (3.9) ensure the equilibrium of any shell element of finite thickness h, bounded by the shell faces and by the ruled surface formed by the normals to the middle surface through a closed infinitesimal curve on this surface.

Whereas the foregoing equations are all fully exact, the approximate nature of shell theory enters the picture at the stage where constitutive relations are introduced between the stress resultants $n^{\alpha\beta}$ and stress couples $m^{\alpha\beta}$ on the one hand, and the strain measures $\gamma_{\alpha\beta}$ and $\rho_{\alpha\beta}$ on the other hand. Leaving aside for the time being any question as to their accuracy, we shall employ the conventional uncoupled constitutive equations for an isotropic material

$$n^{\alpha\beta} = \frac{Eh}{1-\nu^2}\,[(1-\nu)\gamma^{\alpha\beta} + \nu a^{\alpha\beta}\gamma^\kappa_\kappa] \;\; , m^{\alpha\beta} = \frac{Eh^3}{12(1-\nu^2)} \; [(1-\nu)\rho^{\alpha\beta} + \nu a^{\alpha\beta}\rho^\kappa_\kappa], \tag{3.13}$$

where E denotes Young's modulus and ν is Poisson's ratio. The present linear theory of shells is complete in the sense that uniqueness of the solution of equations and boundary conditions (3.3), (3.4), (3.8) - (3.13) is ensured, apart from an additive rigid-body displacement. We shall assume in the sequel that the solution of the problem of shell theory has been obtained.

4. Construction of statically admissible stress distribution [4]

We introduce the tensor of pseudo-stresses $s^{ij}(z) = \sqrt{g/a}\ \sigma^{ij}(z)$, where $\sigma^{ij}(z)$ denotes the stress tensor in shell space. We construct a linear distribution over the shell thickness of pseudo-stresses parallel to the middle surface

$$s^{\alpha\beta}(z) = s_0^{\alpha\beta} + z s_1^{\alpha\beta},$$

which is statically equivalent to the stress resultants and stress couples of shell theory. We obtain the equations

$$(4.1)\ n^{\alpha\beta} = h s_0^{\alpha\beta} - \frac{1}{12} h^3 b^\alpha_\kappa b^\beta_\lambda s_0^{\kappa\lambda}\ ,\ m^{\alpha\beta} = \frac{1}{12} h^3 s_1^{\alpha\beta} + \frac{1}{24} h^3 (b^\alpha_\kappa s_0^{\kappa\beta} + b^\beta_\kappa s_0^{\kappa\alpha})\ ,$$

which determine $s_0^{\alpha\beta}$ and $s_1^{\alpha\beta}$ uniquely.

We write the equations of equilibrium for a volume element of the shell in terms of pseudo-stresses, in the form due to Naghdi [8]

$$(4.2)\qquad [(\delta^\alpha_\kappa - z b^\alpha_\kappa) s^{\kappa 3}(z)]_{,3} - b^\alpha_\kappa s^{\kappa 3}(z) + [(\delta^\alpha_\kappa - z b^\alpha_\kappa) s^{\kappa\beta}(z)]_{|\beta} = 0\ ,$$

$$(4.3)\qquad s^{33}(z)_{,3} + s^{\alpha 3}(z)_{|\alpha} + (\delta^\alpha_\beta - z b^\alpha_\beta) b_{\alpha\kappa} s^{\beta\kappa}(z) = 0.$$

Employing the solution of (4.1), equations (4.2) may be considered, at every point x^α of the middle surface, as a pair of ordinary differential equations of the first order for the pseudo-stresses $s^{\alpha 3}(z)$. Starting from the known values of these pseudo-stresses at the face $z = -\frac{1}{2}h$, we may integrate these equations. The overall equilibrium of a shell element expressed by (3.8) and (3.9) ensures that the solution thus obtained agrees on the shell face $z = \frac{1}{2}h$ with the tangential surface load specified on that face. Employing the solutions of equations (4.1) and (4.2), we may now solve (4.3) in the same way for $s^{33}(z)$.

The exact solution of equations (4.1) - (4.3) is required for a precise error estimate. The approximate solution for the stress tensor

$$(4.4)\ \sigma^{\alpha\beta}(z) = \frac{1}{h} n^{\alpha\beta} - \frac{12}{h^3} z m^{\alpha\beta}\ ,$$

$$(4.5)\ \sigma^{\alpha 3}(z) = -\frac{z}{h} n^{\alpha\beta}\Big|_\beta + \frac{3}{2h}\left(\frac{4z^2}{h^2} - 1\right) m^{\alpha\beta}\Big|_\beta\ ,$$

$$(4.6)\ \sigma^{33}(z) = \sigma^{33}(0) + \frac{z^2}{2h} n^{\alpha\beta}\Big|_{\alpha\beta} - \frac{z}{h} b_{\alpha\beta} n^{\alpha\beta} + \frac{6z^2}{h^3} b_{\alpha\beta} m^{\alpha\beta}$$
$$- \frac{3z}{2h}\left(\frac{4z^2}{3h^2} - 1\right) m^{\alpha\beta}\Big|_{\alpha\beta},$$

where the constant of integration $\sigma^{33}(0)$ depends on the distribution of normal loads over the shell faces, is adequate for an order of magnitude estimate.

5. Construction of admissible displacement field.

In the absence of kinematic boundary conditions, we may always construct a three-dimensional admissible displacement field

$$(5.1) \qquad \underline{u}(x^\kappa, z) = u_\alpha(z)\underline{a}^\alpha + w(z)\underline{n}$$

from the middle surface displacement field of shell theory (3.1), if we impose $u_\alpha(0) = u_\alpha$, $w(0) = w$. The usual Kirchhoff-Love type displacement field, defined by

$$(5.2) \qquad u_\alpha(z) = u_\alpha - (w_{,\alpha} + b^\kappa_\alpha u_\kappa)z \quad , \quad w(z) = w ,$$

is inadequate for our purposes. We modify (5.2) as follows

$$(5.3) \quad u_\alpha(z) = u_\alpha - (w_{,\alpha} + b^\kappa_\alpha u_\kappa)z - \frac{3(1 + \nu)}{Eh} m^\beta_\alpha\Big|_\beta z$$

$$- \frac{1}{2Eh}\left[2(1+\nu)n^\beta_\alpha\Big|_\beta - \nu n^\beta_\beta\Big|_\alpha\right] z^2 + \frac{2}{Eh^3}\left[2(1+\nu)m^\beta_\alpha\Big|_\beta - \nu m^\beta_\beta\Big|_\alpha\right] z^3 ,$$

$$(5.4) \qquad w(z) = w - \frac{\nu}{Eh} n^\beta_\beta z + \frac{6\nu}{Eh^3} m^\beta_\beta z^2 .$$

The additional terms in (5.4) are essential for a justification of shell theory [4], the additional terms in (5.3) were proposed by Danielson [5], and they lead to a substantial improvement of the error estimate.

Lack of space prevents us from writing down explicitly the components of the stress tensor $\underline{\underline{\sigma}}^0$, associated by Hooke's law with the admissible displacement field defined by (5.1), (5.3) and (5.4). Their evaluation, however, is a matter of straightforward algebra.

6. Error estimates of shell theory.

We consider the three-dimensional stress boundary value problem of the linear theory of elasticity for our shell under edge loads which are specified in accordance with the stress distribution of section 4. This stress distribution then satisfies all requirements on a statically admissible stress distribution $\underline{\underline{\sigma}}^*$ in the sense of section 2. Likewise the stress distribution $\underline{\underline{\sigma}}^0$, defined in section 5, is derived from a kinematically admissible displacement field. The solution of the equations of shell theory thus permits by section 2 a *precise* energy error estimate of the stress distributions $\underline{\underline{\sigma}}^*$, $\underline{\underline{\sigma}}^0$ or $\frac{1}{2}(\underline{\underline{\sigma}}^* + \underline{\underline{\sigma}}^0)$ as approximations to the actual stress distribution $\underline{\underline{\sigma}}$. The following qualitative argument shows that *this error is indeed always small in thin shells under sufficiently smooth loads.*

We introduce the concept of wave length L of the deformation pattern of shell theory by the order of magnitude relations

$$n^{\alpha\beta}\Big|_\kappa = 0\left(\frac{n}{L}\right) \quad , \quad m^{\alpha\beta}\Big|_\kappa = 0\left(\frac{m}{L}\right) ,$$

where n and m represent the maximum value of a component of the stress resultants and stress couples in a region of the middle surface of area $0(L^2)$ in the neighbourhood of the point under consideration. We emphasize that this wave length is a property of the solution of the equations of two-dimensional shell theory. Let R denote the minimum principal radius of curvature of the middle surface in the region under consideration. A thin shell is characterized by the property that $\epsilon = h^2/L^2 + h/R$ is a small number. For the difference between $\underline{\sigma}^0$ and $\underline{\sigma}^*$ we have now from sections 4 and 5 the estimate for *all* components

$$\underline{\sigma}^0 - \underline{\sigma}^* = 0(\epsilon\sigma_m), \tag{5.1}$$

where σ_m is the maximum absolute value of a component of $\underline{\sigma}^0$ or $\underline{\sigma}^*$ in the region under consideration. It follows that the root mean square error of all three approximate solutions $\frac{1}{2}(\underline{\sigma}^0 + \underline{\sigma}^*)$, $\underline{\sigma}^0$ or $\underline{\sigma}^*$ is of order $\epsilon\sigma_m$. In other words, *the relative error of the solution of shell theory is of order* $\epsilon = h^2/L^2 + h/R$. A further improvement of this estimate seems to be impossible, because the constitutive equations of shell theory (3.13) are known to involve errors of this order of magnitude [9].

7. Concluding remarks.

The scope of the present foundation of shell theory may be widened in several directions [4]. It may be generalized to the case of irregular edge tractions, provided that the difference between the prescribed edge tractions and the edge tractions in accordance with section 4 is not "excessive", it may be extended to the case of kinematic boundary conditions along part or whole of the shell edge, and it may be generalized to the case of a "slowly" variable shell thickness. A severe restriction remains, however, that the basic error estimates of section 2 have been established only in the linear theory.

REFERENCES

[1] Love A.E.H. — *The mathematical theory of elasticity,* 4th ed. Cambridge University Press 1927.

[2] Koiter W.T. — Foundations and basic equations of shell theory; a survey of recent progress. *Proc. Second I.U.T.A.M. Symp. on the Theory of Thin Shells,* September 1967, edited by F. Niordson, Springer-Verlag, Berlin, 1969, p. 93-105.

[3] Prager W. and Synge J.L. — Approximations in elasticity based on the concepts of function space. *Quart. Appl. Math.* 5, 1947, p. 241-269. Cf. also J.L. Synge, *The hypercircle in mathematical physics,* Cambridge University Press (1957).

[4] Koiter W.T. — On the foundations of the linear theory of thin elastic shells. *Proc. Kon. Ned. Ak. Wet.* B 73, 1970, p. 169-195.

[5] Danielson D.A. — Private communication, July 20, 1970.

[6] Mikhlin S.G. — *The problem of the minimum of a quadratic functional.* Holden-Day, San Francisco, 1965.

[7] Koiter W.T. — On the nonlinear theory of thin elastic shells. *Proc. Kon. Ned. Ak. Wet.* B 69, 1966, p. 1-54.

[8] Naghdi P.M. — Foundations of elastic shell theory. *Progr. in Solid Mech.* 4, edited by I. N. Sneddon and R. Hill. North-Holland Publ. Co., Amsterdam, 1963, p. 1-90.

[9] Koiter W.T. — A consistent first approximation in the general theory of thin elastic shells. *Proc. I.U.T.A.M. Symp. on the Theory of Thin Elastic Shells,* August 1959. North-Holland Publ. Co., Amsterdam, 1960, p. 12-32.

University of Technology
Delft (Pays-Bas)

Actes, Congrès intern. Math., 1970. Tome 3, p. 131 à 135.

ON THE STABILITY OF STEADY FLOWS INDEPENDENT OF A SPATIAL COORDINATE

by A.G. KULIKOVSKIY

In this paper some approximate methods of investigation of stability are considered. It is assumed that the systems under consideration are independent of time t and of one of spatial coordinates say x, and have enough large length in the direction of this coordinate.

At the beginning it is assumed for simplicity, that disturbances are described by a system of linear partial differential equations with two independent variables and with constant coefficients

$$A_{ij} \frac{\partial u_j}{\partial t} + B_{ij} \frac{\partial u_j}{\partial x} + C_{ij} u_j = 0 \tag{1}$$

As it is known, this system has a solution of the form exp $i(Kx - \omega t)$. For the existence of such solution it is necessary, that K and ω be connected by the relation

$$| - A_{ij}\, i\omega + B_{ij}\, iK + C_{ij} | = 0 \tag{2}$$

As to the system (1) it is supposed, that this system is correct in sense of Petrovsky [1]. This means, that for any real K all solutions ω of equation (2) have bounded imaginary part, i.e.

$$Im\, \omega < M \tag{3}$$

(M is a constant). In other words, if $Im\, \omega > M$, all the roots $K_j(\omega)$ of equation (2) have nonzero imaginary parts, thus forming two sets of the roots : the upper set, for which $Im\, K > 0$ and the lower one, for which $Im\, K < 0$.

Denote the number of K_j belonging to the upper set by S, and the number of K_l belonging to the lower set by $n - S$. (n is the total number of K_m). Further, we shall assume that all the roots have been enumerated in such a way, that

$$Im\, K_1(\omega) > Im\, K_2(\omega) > \cdots > Im\, K_n(\omega) \tag{4}$$

Such enumeration may be performed for all ω except at the points of the curves described by the equations

$$Im\, [K_i(\omega) - K_j(\omega)] = 0.$$

It is assumed that at the ends of some closed interval of the x-axis say $-L \leqslant x \leqslant L$ the solution of system (1) satisfy some boundary conditions. As to the boundary conditions, the following two assumptions are made. First, it is assumed that each boundary condition connects the unknown quantities and their derivatives at one of the ends of the interval

$$P_{ej}\left(\frac{\partial}{\partial t}, \frac{\partial}{\partial x}\right) u_j \Big|_{x=-L} = 0 \tag{5}$$
$$P_{mj}\left(\frac{\partial}{\partial t}, \frac{\partial}{\partial x}\right) u_j \Big|_{x=L} = 0$$

where P_{ij} are some polynomials of $\frac{\partial}{\partial t}, \frac{\partial}{\partial x}$. Secondly it is assumed that the mixed initial-boundary problem for system (1) with boundary conditions (5) is correct. For this it is necessary [2] that the number of the boundary conditions at $x = -L$ should be equal to S, and the number of the boundary conditions at $x = L$ should be equal to $n - S$. Thus in equalities (5) we have

$$l = 1, 2, \dots S \quad , \quad m = S + 1, \dots n \tag{6}$$

Now it may be shown, that in a general case the asymptotic behaviour of solutions of equations (1) satisfying the boundary conditions (5) with $t \to \infty$ is of the form

$$f(x)\, e^{i\omega_* t} \tag{7}$$

where ω_* is the eigenfrequency with the largest imaginary part and $f(x)$ is an eigenfunction which in the case under consideration is of the form

$$f(x) = \sum_{l=1}^{n} C_l\, e^{iK_l(\omega)x} \tag{8}$$

In the expressions (7) and (8) the index, indicating the number of unknown functions is omitted for the simplicity.

It is shown [3] that, when $L \to \infty$, all eigenfunctions and eigenfrequencies are divided into two parts.

One part consists of separate points in the complex plane ω, the distance between which does not tend to zero, when $L \to \infty$. The corresponding eigenfunctions are of the form

$$\sum_{l=1}^{S} C_l\, e^{iK_l(\omega)x} \tag{9}$$

or

$$\sum_{l=S+1}^{n} C_l\, e^{iK_l(\omega)} \tag{10}$$

The eigenfrequencies corresponding to the eigenfunctions (9) are determined by the properties of the system of equations (1) and the boundary conditions on the left end ($x = -L$) only. They are independent of the boundary conditions on the other end. The eigenfrequencies corresponding to the eigenfunctions (10) are determined by the system of equations and the boundary conditions on the right end ($x = L$) only. Such eigenfunctions and eigenfrequencies will be termed boundary the ones. If there is a boundary eigenfrequency ω with $Im\, \omega > 0$, we shall say, that boundary instability takes place.

The other part of the spectrum consists of the points, the distance between which is of order $1/L$. These points are at a distance of the order $1/L$ from the curve

$$(11) \qquad Im\,[K_s(\omega) - K_{s+1}(\omega)] = 0$$

The corresponding eigenfunction contains all terms in expression (8). But everywhere exept in some neighbourhoods of the ends of the interval the main terms in this expression are the terms with numbers S and $S + 1$. The other terms are small in comparison with these two terms. Such eigenfrequencies and eigenfunctions will be termed the global ones. If the equation (11) has a solution ω with $Im\ \omega > 0$, we shall say, that global instability occurs. Since the equation (11) is independent of boundary conditions, the global instability is determined only by the properties of the system of equations, describing the behavior of the small disturbances.

If the global instability takes place, then there exists [3] a branch $K_j(\omega)$, such that

$$(12) \qquad Im\,K_j(\omega) = 0$$

for some ω with

$$(13) \qquad Im\,\omega > 0.$$

The relations (12), (13) are usually used as a criterion of instability in unbounded system (see for example [4]).

It is shown beside, that if the unbounded system is absolutely unstable, then the bounded one is globally unstable. The condition of the absolute instability consists [5] in satisfying one of the equations

$$(14) \qquad K_i(\omega) = K_j(\omega)$$

with $Im\,\omega > 0$. In the above equation $K_i(\omega)$ and $K_j(\omega)$ are the roots of the dispersion equation (2), having different signs, when $Im\,\omega > M$.

It must be stressed, that the condition of global instability (11), (13) is different in general from the condition of instability of unbounded system (12), (13) and from the condition of absolute instability (13), (14).

The criterion of global instability, obtained for partial differential equations with two independent variables, may be applied to more complicate cases.

For some systems it may occur that to each ω corresponds an infinite set of K_j. As before, we shall assume, that there exists a constant M, such that for ω with $Im\ \omega > M$ the set of the K_j is divided into two sets, the upper one and the lower one with different signs of imaginary parts. Then, as above, we must take in the equation (11) for K_s that value of K from the upper set, which has the least imaginary part, and for K_{s+1} we must take that value of K from the lower set, which has the largest one.

For example, the stability of the parallel flow of a viscous incompressible fluid in a flat tube,the walls of which are two parallel planes of a large but finite length, is considered. It is shown, that this flow is globally unstable [6]. But in the case when the boundary conditions at the ends of the tube exclude the variation of the flow of the fluid through the tube, the flow is globally stable [7]. For the comparison we mention, that such flow in the unbounded tube is unstable (see [4]) but the instability is not absolute [8].

The second example is the stability of a collisionless uniform electron plasma contained between two parallel planes, the distance between which is sufficiently large. It is shown, that global instability takes place when the inequality

$$\nu\,[f'(\nu) - f'(-\nu)] > 0 \tag{15}$$

takes place in the region $|\nu| \gg \bar{\nu}$, where ν is the component of the velocity, which is orthogonal to the planes bounding the plasma, $\bar{\nu}$ is the square mean velocity of electrons, f is the distribution function, $f'(\nu) = \partial f/\partial \nu$.

We mention for comparison, that unbounded plasma is unstable [9], when the inequality

$$\nu\, f'(\nu) > 0$$

takes place in the region $|\nu| \gg \bar{\nu}$.

In conclusion it should be noted, that with the problems, considered here, is closely connected the problem of stability of physical systems weakly depending on x (e.g. depending on x/L when $L \to \infty$). Recently this problem has been investigated in detail in connection with applications to the theory of plasma. The main object of the investigation was the systems of differential equations of the second and forth order with respect to the spatial variable [10 - 14]. The speciality of the weakly nonuniform systems consists in the fact, that the waves of perturbances may be reflected not only from boundaries of the interval, but also from some interior points (the phenomenon of Stokes). This may lead to the formation of eigenfunctions of a special form. In the simplest case of an equation of the second order the eigenfrequencies are found from the equation given below (written with the same degree of accuracy as equation (11)) :

$$\int_{x_1}^{x_2} Im\,[K_1(\omega\,,x) - K_2(\omega\,,x)]\,dx = 0 \tag{16}$$

where K_1 and K_2 are the roots of the dispersion equation, in which x is concidered as a parameter, and the integration is taken in the complex plane x between the points in which $K_1(\omega\,,x) = K_2(\omega\,,x)$. This equation differs from the equation (11) only by an integration with respect to x. It is possible, that in the general case too the eigenfrequencies ω should be determined from some equation of a similar type. But, as far as the author knows, there is no sufficiently complete investigation of the case of differential equations of the n-th order.

REFERENCES

[1] Петровский И. Г. — О проблеме Коши для систем линейных уравнений с частными производными в области неаналитических функций. Бюлл. Моск. Университета, секция А, математика и механика, т. I, № 7, 1938.

[2] Соболев С. Л. — О смешанных задачах для уравнений в частных производных с двумя независимыми переменными, ДАН, т. 122, № 4, 1958.

[3] Куликовский А. Г. — Об устойчивости однородных состояний, ПММ, т. 30, № 1, 1966.

[4] Lin C.C. — *The Theory of Hydrodynamic stability*, 1955.

[5] Rolland P. — Instabilities and growing waves, *Phys. Rev.*, v. 140, No. 38 (1965).

[6] Куликовский А. Г. — Об устойчивости течений слабо сжимаемой жидкости в плоской трубе большой но конечной длины, ПММ, т. 32, № 1, 1968.

[7] Куликовский А. Г. — Об устойчивости течения Пуазейля и некоторых других плоскопараллельных течений в плоской трубе большой но конечной длины при больших числах Рейнольдса, ПММ, т. 30, № 5, 1966.

[8] Иорданский С. В., Куликовский А. Г. — Об абсолютной устойчивости некоторых плоскопараллельных течений при больших числах Рейнольдса, ЖЭТФ, 1965, т. 49, № 10.

[9] Ландау Л. Д. — О колебаниях электорнной плазмы, ЖЭТФ, т. 16, № 7, 1946.

[10] Силин В. П. — Колебания слабо неоднородной плазмы, ЖЭТФ, т. 44, № 4, 1963.

[11] Силин В. П., Рухадзе А. А. — Метод геометрической оптики в электродинамике неоднородной плазмы, Усп. Физ. Наук, т. 82, № 3, 1964.

[12] Днестровский Ю. Н., Костомаров Д. П. — Об асимптотике собствнных значений для несамосопряженных краевых задач, Ж. вычисл. матем. и мат. физики, т. 4, № 2, 1964.

[13] Заславский Г. М., Моисеев С. С., Сагдеев Р. З. — Асимптотические методы гидродинамической теории устойчивости, ПМТФ, № 5, 1964.

[14] Рухадзе А. А., Саводченко В. С., Тригер С. А. — Метод геометрической оптики для дифференциальных уравнений четвертого порядка в приложениях к низкочастотным колебаниям плазмы, ПМТФ, № 6, 1965.

Steklov Mathematical Institute
Vavilova street 42,
Moscow V 333 (URSS)

Actes, Congrès intern. Math., 1970. Tome 3, p. 137 à 142.

NON LOCAL CAUCHY PROBLEMS IN FLUID DYNAMICS

by L.V. OVSIANNIKOV

Summary.

The problems of unsteady incompressible fluid motions with free boundary leading to Cauchy problems for non-linear non local differential equations are considered. The special case of them is well known Cauchy-Poisson problem about waves on the water surface. The paper contains existence theorems of this class of problems in spaces of analytical functions. The results are founded on some new a priori estimates for elliptic type equations in the analytical case. It is convenient to solve free boundary problem in some cases in Lagrangian coordinates. As an example serves the problem about upflow of the bubble in the water of infinite depth.

Introduction.

Non local Cauchy problems arise in the theory of unsteady motion of incompressible liquid with not known in advance moving boundaries. The significant feaucher of these problems is that the domain of definition has not fixed boundaries and is, indeed, one of elements of the solution. To this class of problems belongs those investigated by Lichtenstein [2] about mixed, potential-vortex flows, and some problems about motions with strong tangential discontinuities which serve as the model for a special type of viscous flows.

This paper is devoted to non local problems arising in the theory of *potential* unsteady motion of inviscid incompressible liquid with *free boundaries.* The term "free boundary" designate that part of the moving liquid boundary which consist of the same liquid particles all the time of movement and carry prescribed values of hydrodynamical pressure.

The well known Cauchy-Poisson problem about waves spreading on the surface of ocean as the result of an initial disturbance belongs to this type. Although this problem was formulated mathematically more than 150 years ago, it remains unsolved till to now. The large amount of investigations concerning Cauchy-Poisson problem consider different kinds of approximate theories, the most significant among them are linear theory and shallow water theory. In the precise formulation there are few exact solutions and some investigations of the special cases only. The detailed exposition of these achievements was given by Prof. J.J. Stoker [8].

The situation here was that until the last time we had not even a general enough existence and uniqueness theorems "in small". Now it becames clear that, at any rate, in the class of analytical functions the free boundary problem is correct :

its solution exists, is unique and continuously depends on the initial data for sufficiently small time interval. The appropriate theory was elaborated during last years in the Institute of Hydrodynamics, Siberian Branch of Acad. Sci. USSR [3 - 7]. The theory developed was founded on the estimate technique which resembles to the elaborated by J. Leray and Y. Ohya [1].

Mathematical formulation.

At first we remember the formulation of the problem about potential liquid motion with free boundary in the simplest case. Let R^3 be three-dimensional Euclidean space of points $x = (x, y, z)$ and R^+-positive semiaxe of time t. On $R^3 \times \overline{R}^+$ the function $h = h(\vec{x}, t)$ is given (the potential of mass powers).

Problem A. – There are given the domain $\Omega \subset R^3$ and the harmonic in Ω function $\varphi_0(\vec{x})$. It needs to seek the domain Ω_t with boundary $\partial\Omega_t$ depending on $t \in R^+$, and harmonic in Ω_t function $\varphi = \varphi(\vec{x}, t)$ so that on $\partial\Omega_t$ for $t \in \overline{R}^+$

$$(1) \qquad \varphi_t + \frac{1}{2}\,|\nabla_x \varphi|^2 + h = p(t)$$

$$(2) \qquad \vec{n} \cdot \nabla_x \varphi = V_n$$

and on $t = 0$

$$(3) \qquad \Omega_0 = \Omega \quad , \quad \varphi(\vec{x}, 0) = \varphi_0(\vec{x}) .$$

Here ∇_x is the operator-gradient in R^3, $\vec{n}$-unit normal vector on $\partial\Omega_t$ and V_n-the speed of displacement (in R^3) of surface $\partial\Omega_t$ in the direction $\vec{n}$. The given function $p(t)$ can depend (in a previously known manner) on the domain Ω_t.

When in the motion take part given in $R^3 \times \overline{R}^+$ rigid surfaces (they belong to $\partial\Omega_t$), the only condition (2) on them have to be justified with the known value V_n.

Particularly, in Cauchy-Poisson problem about waves on the surface of finite depth water it is taken $h = a\,z$, $a = \text{const}$, $p(t) \equiv 0$ and the domain Ω is defined by inequalities $0 < z < g_0(x, y)$, $-\infty < x, y < +\infty$. Here the boundary $z = 0$ (bottom) is at rest and (2) has the form $\partial\varphi/\partial z = 0$. It is convenient here to seek free boundary in the form $z = g(x, y, t)$ and to consider the value of potential on this boundary putting $f(x, y, t) = \varphi(x, y, g(x, y, t), t)$. This definits the mapping $\vec{v} : (f, g) \to \nabla_x \varphi|_{z=g}$ by means of which the problem A is transformed into the following.

Problem A′. – To find functions $f = f(x, y, t)$ and $g = g(x, y, t)$ $(-\infty < x, y < +\infty, t \geqslant 0)$ satisfying equations

$$(4) \qquad f_t = -\frac{1}{2}\,|\vec{v}(f, g)|^2 - ag$$

$$(5) \qquad g_t = \vec{v}(f, g) \cdot \nabla_x (z - g)$$

$$(6) \qquad f(x, y, 0) = f_0(x, y) \quad , \quad g(x, y, 0) = g_0(x, y)$$

Here f_0, g_0 are the given functions and $f_0(x, y) = \varphi_0(x, y, g_0(x, y))$.

The problem A′ is the typical non local Cauchy problem in the theory of unsteady liquid motions with free boundary. It is possible to motivate the náturalness of its consideration in the class of analytical functions [4].

This way is connected with construction of Banach spaces of analytical functions (moreover these spaces have to be Banach rings) and estimation technique for nonlinear non local transformations of $\vec{\nu}(f, g)$ type.

Spaces B_ρ.

The following construction is used in considerations with Banach spaces of analytical functions. Let Ω be open set in $R^n = R^n(x)$. Let a norm N is defined on the set of functions $f : R^n \to R^s$, $f \in C_\infty(\Omega)$. Denote as (Ω, N) the closure of the set of those f which have $N(f) < \infty$. Let $D_j = \partial/\partial x_j$ $(j = 1, 2, \ldots, n)$. It is required that the norm N possesses of two properties :

(1) $f, g \in (\Omega, N) \Rightarrow N(fg) \leqslant N(f) \cdot N(g)$ (the Banach ring property) ;

(2) For any sequence $\{f_K\}_1^\infty \subset (\Omega, N)$ such that $\{D_j f_K\}_1^\infty \subset (\Omega, N)$ and by $K \to \infty : f_K \overset{N}{\to} f$, $D_j f_K \overset{N}{\to} g$ it follows that $g = D_j f$ $(j = 1, 2, \ldots, n)$ (closureness of the derivation operators).

Definition. – Let $\rho \geqslant 0$. We define the space

$$B_\rho = \{f \mid f \in (\Omega, N) \ , \ \|f\|_\rho < \infty\}$$

where

$$(7) \qquad \|f\|_\rho = \sum_{m=0}^{\infty} \frac{\rho^m}{m\,!} \max_{|\beta| = m} N(D^\beta f)$$

Lemma. B_ρ is Banach space.

Proof of the lemma uses the property (2) only. The main properties of the norm $\|\cdot\|_\rho$ are

$$(a) \qquad \|fg\|_\rho \leqslant \|f\|_\rho \cdot \|g\|_\rho$$

$$(b) \qquad \|D_j f\|_\rho \leqslant \frac{\partial}{\partial\rho} \|f\|_\rho \quad , \quad (j = 1, 2, \ldots, n)$$

The property (a) is essential for the purpose of estimation of superpositions of analytical mappings. This construction possess the great generality and may be useful in different questions (not only in hydrodynamical problems). It differs from the analogous construction given in [1] in the point that we use any norms N and not a formal series (7).

If there are several sets Ω and norms N in one consideration then symbols B_ρ and $\|f\|_\rho$ are changed into $B_\rho(\Omega, N)$ and $\|f ; \Omega, N\|_\rho$ correspondingly. We denote also $\|f\|'_\rho = \|f\|_\rho - \|f\|_0$.

A priori estimate in layer.

The first step in proving the existence theorem for problem A′ was the construction of a priori estimate for the special second order boundary value problem [3]. Let $\Omega \subset R^n$ be the layer $0 < x_n < 1$ and S_0 , S_1-planes $x_n = 0$, $x_n = 1$. In the definition (7) there are used norms $N_{K+\alpha}(K \geqslant 2,\ 0 < \alpha < 1)$ in spaces of $C_{K+\alpha}$ type and derivatives D^β with $\beta_n = 0$ only. There are considered differential operators

$$Pu = \sum_{|\beta| \leqslant 2} a_\beta(x)\, D^\beta u \quad , \quad Bu = D_n u + \sum_{j=1}^{n-1} b_j(x)\, D_j u$$

where P is uniformly elliptic in $\overline{\Omega}$. The a priori estimate has the form of assertion about existence such a constant E that

$$\begin{aligned}(8)\quad & \|u\,;\Omega\,,N_{K+\alpha}\|_\rho\,\{1 - E[\rho + \|P\,;\Omega\,,N_{K-2+\alpha}\|'_\rho + \|B\,;S_0\,,N_{K-1+\alpha}\|'_\rho]\} \\ & \leqslant E[\,\|Pu\,;\Omega\,,N_{K-2+\alpha}\|_\rho + \|Bu\,;S_0\,,N_{K-1+\alpha}\|_\rho \\ & + \|u\,;S_1\,,N_{K+\alpha}\|_\rho + \|u\,;\Omega\,,N_0\|_0].\end{aligned}$$

This is sufficient in order to prove the theorem [3] :

THEOREM 1. – *If the data (6) and the value $\rho_0 > 0$ are such that*

$$f_0\,,g_0 \in B_{\rho_0}(R^2\,,N_{2+\alpha})$$

then there exist values $\rho_1 > 0$ and $t_1 > 0$ such that the solution of the problem A' exists and is unique in $B_\rho(R^2\,,N_{2+\alpha})$ for every $\rho \leqslant \rho_1$ in the interval $0 \leqslant t \leqslant t_1$. This solution is holomorphic function of t in point $t = 0$.

As the consequence this gives the existence and uniqueness theorem of the solution of Cauchy-Poisson problem in analytical case.

Method of Lagrangian coordinates.

Lagrangian coordinates $\vec{\xi} = (\xi\,,\eta\,,\zeta)$, t are introduced by means of the equations

$$(9)\qquad \partial\vec{x}/\partial t = \nabla_x\,\varphi(\vec{x}\,,t) \quad , \quad \vec{x}|_{t=0} = \vec{\xi}\,.$$

The transformation $T_t : \vec{\xi} \to \vec{x}$ depends on time t and has to be determined with the solution of hydrodynamical problem together. In Lagrangian coordinates the domain of definition of the solution is fixed and coinside with the given Ω. The formulation of the problem A in Lagrangian coordinates will be as follows [5] :

PROBLEM L. – To find functions $\vec{x} = \vec{x}(\vec{\xi}\,,t)$, $\varphi = \varphi(\vec{x}\,,t)$ which satisfy equations in the domain Ω

$$(10)\qquad \vec{x}_t = M^{*-1}\nabla\varphi \quad , \quad \operatorname{div}\,(M^{-1}M^{*-1}\nabla\varphi) = 0$$

boundary condition on $\partial\Omega$

$$(11) \qquad \varphi_t = \frac{1}{2} |M^{*-1}\nabla\varphi|^2 - h + p(t)$$

and initial conditions in Ω by $t = 0$

$$(12) \qquad \vec{x} = \vec{\xi} \quad , \quad \varphi = \varphi_0(\vec{\xi}) \quad (\Delta\varphi_0 = 0)$$

Here $M = \partial\vec{x}/\partial\vec{\xi}$ is Jacobi matrix of the transformation T_t, M^{-1} and M^* designate the inverse and the transposed matrices correspondingly ; the operations ∇ and div are performed in variables $\vec{\xi}$.

This Lagrange representation was used in the problem on upflow of the bubble arizing by the underwater explosion [7]. Here Ω is the halfspace $\zeta < 0$ except for the ball Q $(0 \leqslant r \leqslant 1)$, $r^2 = \xi^2 + \eta^2 + (\zeta + \zeta_0)^2$, $\zeta_0 > 1$. The plane $S(\zeta = 0)$ and the sphere Γ $(r = 1)$ are free boundaries and $\partial\Omega = S \cup \Gamma$. The pressure $p(t) = 0$ on S and $p(t) = p_0[(3/4\pi)|T_tQ|]^{-\gamma}$ on Γ where $|T_tQ|$ is the volume of T_tQ in $R^3(x)$, $\gamma = \text{const.} > 1$, $p_0 = \text{const.} > 0$. In the field of gravity $h = az$. We put $\varphi_0 \equiv 0$ in (12). We designate this problem as L^0. The pecularity of the problem L^0 is the presence of two free boundaries and unboundedness of the domain Ω.

Using a fixed real σ, $3 < 4\sigma < 4$, the following norms are constructed $(i = 1, 2)$

$$N_i(u) = \sup_{\Omega} |(1 + |\vec{\xi}|^2)^{i\sigma} u(\vec{\xi})| + \sup_{\Omega\times\Omega} (|u(\vec{\xi}) - u(\vec{\xi}')| \cdot |\vec{\xi} - \vec{\xi}'|^{-\sigma}).$$

Appropriate B_ρ-spaces are defined by means of slightly changed norms of (7) type. Instead of operators $D_j = \partial/\partial x_j$ are used the special differential operators with analytical coefficients $\widetilde{D}_j$ $(j = 1, 2, 3, 4)$. They are constructed in a manner that $\widetilde{D}_1$ is the operator of normal derivation relatively to S and Γ and $\widetilde{D}_j$ $(j = 2, 3, 4)$ are tangential to S and Γ. The principal step is the proof of a priori estimate in B_ρ-space of the solution of following boundary problem in Ω

$$(13) \qquad \Delta u = \operatorname{div}\vec{w} \quad , \quad u|_{\partial\Omega} = f \quad , \quad u \to 0 \quad (|\vec{\xi}| \to \infty)$$

THEOREM 2. – *It exists analytical function* $\omega(\rho) \gg 0$ *such that for the solution of problem* (13) *the estimate is valid*

$$(14) \qquad \|\nabla u \,; \Omega\,, N_1\|_\rho \leqslant \omega(\rho) \left[\|w \,; \Omega\,, N_2\|_\rho + \frac{\partial}{\partial\rho}\|f \,; \partial\Omega\,, N_1\|_\rho\right]$$

We consider t-analytical solutions which have a form of power series of time t.

THEOREM 3. – *t-analytical solution of the problem L^0 exists and is unique. There exist real number $\rho_1 > 0$ and decreasing function $\theta(\rho)$, $\theta(\rho_1) = 0$, such that the solution $\nabla\varphi(\vec{\xi}, t)$, $\vec{x}(\vec{\xi}, t) - \vec{\xi}$ of the problem L^0 belongs to a space $B_\rho(\Omega\,, N_1)$, $\rho < \rho_1$ when $t < \theta(\rho)$.*

Coefficients of the t-oower series are obtained step by step. The convergence is proved by the method of majorants including the use of the estimate (14).

Conclusion.

There are very many problems about motions with free boundary in what it would be worth to apply B_ρ-spaces. Till now we have not existence theorems in the exact formulation for such a problems as unsteady interaction of jets, cumulation theory, waves on sloping beaches, floating bodies, junction and division of finite liquid masses and so on. For new applications the technique of B_ρ-spaces must be essentially improved. There is a number of questions waiting the answer. For example, is it possible to obtain estimates like (8) or (14) for domains with angular points? How to transform into equivalent Cauchy problems the "mixed" problems where free boundary intersects rigid wall and the point of intersection moves along the wall? The question of principle is about the behaviour of the solution "in large" in time. Perhaps it might be clarified by consideration of diversions from some exact solutions simply defined for unbounded time [6].

REFERENCES

[1] Leray J. et Ohya Y. — Equations et Systemes Non-linéaires, Hyperboliques Non-Stricts. *Math. Annalen* 170, 1967, p. 167-205.

[2] Lichtenstein L. — *Grundlagen der Hydromechanik,* Springer, Berlin, 1929.

[3] Nalimov V.I. — A priori Estimates of the Solutions of Elliptic Equations with Application to Cauchy-Poisson Problem. *Dokl. Akad. Nauk SSSR,* 189, 1969, p. 45-48.

[4] Ovsiannikov L.V. — Singular Operator in the Scale of Banach Spaces. *Dokl. Akad. Nauk S.S.S.R.,* 163, 1965, p. 819-822.

[5] Ovsiannikov L.V. — About Motion of a Finite Mass of Liquid. *Fluid Dynamics Trans.* 3, (PWN, Warszawa, 1967), p. 75-81.

[6] Ovsiannikov L.V. — On the Disturbances of an Unsteady Motion of Liquid with Free Boundary. *Fluid Dynamics Trans.* 4, (PWN, Warszawa, 1969), p. 105-113.

[7] Ovsiannikov L.V. — On a Bubble Upflow *Some Problems of Mathematics and Mechanics,* Akad. Nauk S.S.S.R., 1970.

[8] Stoker J.J. — *Water Waves: The Mathematical Theory iwth Applications* (Interscience Publishers, Inc., New York, 1957).

Institute of Hydrodynamics, Siberian Dept. of Acad. Sci. URSS.
Novosibirsk 90 (URSS)

Actes, Congrès intern. Math., 1970. Tome 3, p. 143 à 150.

SUR LE CALCUL DE LA SOLUTION DE QUELQUES PROBLÈMES DE LA THÉORIE DES ONDES

par Maurice ROSEAU

Nous nous proposons de discuter le problème de la diffraction d'ondes liquides de gravité, de grande longueur, en profondeur constante dans un bassin indéfini, par un dièdre d'arête normale au fond, compte tenu des effets de Coriolis dus à la rotation Ω, supposée constante et parallèle à l'arête, du système de référence lié au dièdre. C'est, on le voit, une généralisation du problème de Sommerfeld, à la solution de laquelle on peut appliquer une méthode de calcul introduite ailleurs [1, 2] à propos d'autres problèmes de théorie des ondes.

Avec $u\,(x\,,y)\,e^{i\mu t}$, $v\,(x\,,y)\,e^{i\mu t}$, composantes horizontales de la vitesse du fluide, $\varphi\,(x\,,y)\,e^{i\mu t}$ élévation du niveau de l'eau au dessus du plan horizontal $x\,,y$, $\mu = 2\pi/\lambda$, λ longueur d'onde, t désignant le temps, on sait que dans le cadre d'une théorie linéarisée, on peut écrire les équations du mouvement :

$$i\mu u + 2\Omega v + g\frac{\partial\varphi}{\partial x} = 0 \tag{1}$$

$$i\mu v - 2\Omega u + g\frac{\partial\varphi}{\partial y} = 0 \tag{2}$$

$$i\mu\varphi = -h\left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}\right) \tag{3}$$

h profondeur constante du bassin, g accélération de la pesanteur. On résout (1) et (2) en u, v, et portant dans (3) on a :

$$\Delta\varphi + \frac{\mu^2 - 4\Omega^2}{gh}\,\varphi = 0 \tag{4}$$

On suppose

$$k^2 = \frac{\mu^2 - 4\Omega^2}{gh} > 0 \quad , \quad k > 0 \tag{5}$$

hypothèse naturelle si l'on admet le possibilité qu'il existe des ondes de surface à l'infini. Les conditions aux limites qui expriment le glissement du fluide sur les parois du dièdre sont :

$$(i\mu\sin\beta - 2\Omega\cos\beta)\,\varphi_x - (2\Omega\sin\beta + i\mu\cos\beta)\,\varphi_y = 0 \ , \ y = x\,tg\,\beta\,, x < 0 \tag{6}$$

$$(i\mu\sin\beta + 2\Omega\cos\beta)\varphi_x - (2\Omega\sin\beta - i\mu\cos\beta)\,\varphi_y = 0 \ , y = -x\,tg\,\beta\,, x < 0 \tag{7}$$

avec β donné tel que $\pi/2 < \beta < \pi$.

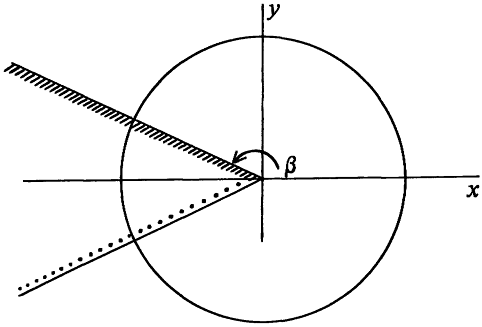


Figure 1

Le problème consiste à obtenir une représentation de $\varphi(x\,,y)$ vérifiant (4) dans le secteur $|\arg(x + iy)| \leqslant \beta$, les conditions aux limites (5) et (6), régulière au bord $x = y = 0$ et confondue à l'infini, au moins dans certaine direction avec une onde plane incidente donnée.

Le point de départ est la représentation :

$$(8)\quad \varphi = \int_C \exp\left\{ i\frac{k}{2}\left[x\left(\zeta + \frac{1}{\zeta}\right) + iy\left(\zeta - \frac{1}{\zeta}\right)\right]\right. \cdot g(\zeta)\,d\zeta$$

$$+\int_\Gamma \exp\left\{ i\frac{k}{2}\left[x\left(\zeta + \frac{1}{\zeta}\right) - iy\left(\zeta - \frac{1}{\zeta}\right)\right]\right\} \cdot h(\zeta)\,d\zeta$$

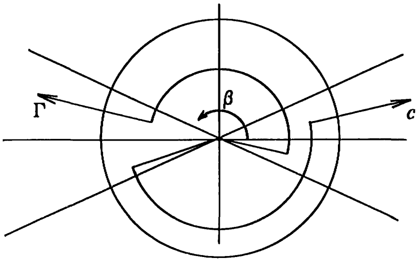
où $g(\zeta)\,,h(\zeta)$ sont des fonctions analytiques de ζ, C, Γ des contours joignant l'origine au point à l'infini, tracés sur le plan complexe ζ, $-\pi \leqslant \arg\zeta \leqslant \pi$, muni d'une coupure suivant la partie négative de l'axe réel.

Des considérations simples sur la convergence des intégrales qui interviennent dans (8) permettent de conclure que C et Γ doivent être construits de telle sorte que les facteurs exponentiels sous les signes $\int$ tendent vers 0, quand ζ décrit C ou Γ à l'approche des points 0 et ∞, pour tout $x + iy = \rho\, e^{i\theta}$ tel que $-\beta \leqslant \theta \leqslant \beta$, $\rho > 0$. Cet argument permet de mettre en évidence des secteurs d'ouverture $\pi - \beta$ dans lesquels doivent être situées les branches de C et Γ et l'on se rend compte ainsi qu'il sera nécessaire d'utiliser deux représentations distinctes selon que $0 \leqslant \theta \leqslant \beta$ ou $-\beta \leqslant \theta \leqslant 0$.

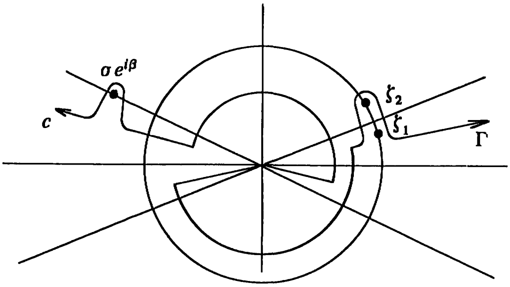
Pour réunir les branches des contours C ou Γ on utilise des arcs circulaires centrés à l'origine, de rayon < 1, tracés sur le feuillet $-\pi \leqslant \arg\zeta \leqslant \pi$; dans le cas $-\beta \leqslant \theta \leqslant 0$, C et Γ ont des enclaves respectivement autour de $\sigma\, e^{i\beta}$,

$$(9)\qquad \sigma = \left(\frac{\mu + 2\,\Omega}{\mu - 2\,\Omega}\right)^{1/2} ,$$

(on peut sans inconvénient supposer $\sigma > 1$), et $\zeta_1\,, \zeta_2$, ζ_1 donné tel que $|\zeta_1| = 1$, $0 \leqslant \arg\zeta_1 \leqslant \pi$, et ζ_2 tel que $|\zeta_2| = 1$, $\arg\zeta_1 + \arg\zeta_2 = 2\pi - 2\beta$, $\zeta_1 \neq \zeta_2$. (ζ_1 est le paramètre arbitraire lié, on le verra plus loin, à la définition de l'onde plane incidente).



$0 \leqslant \theta \leqslant \beta$ Figure 2



$-\beta \leqslant \upsilon \leqslant 0$ Figure 3

Les conditions aux limites.

La fonction φ définie par (8) satisfait (4) ; explicitant (6) on trouve :

$$\int_C \exp\left[i\frac{k}{2}\rho\left(\zeta e^{i\beta}+\frac{1}{\zeta e^{i\beta}}\right)\right]\cdot\left[(\mu-2\Omega)\,\zeta e^{i\beta}-(\mu+2\Omega)\frac{1}{\zeta e^{i\beta}}\right]\cdot g(\zeta)\,d\zeta$$

$$-\int_\Gamma \exp\left[i\frac{k}{2}\rho\left(\zeta e^{-i\beta}+\frac{1}{\zeta e^{-i\beta}}\right)\right]\cdot\left[(\mu+2\Omega)\zeta e^{-i\beta}-(\mu-2\Omega)\frac{1}{\zeta e^{-i\beta}}\right]\cdot h(\zeta)\,d\zeta=0$$

Dans ces intégrales on fait les changements de variable $u=\zeta e^{i\beta}$, $u=\zeta e^{-i\beta}$; les contours C et Γ de la figure 2 sont changés en C', Γ' respectivement par les rotations $[0,\beta]$ et $[0,-\beta]$ (on a, dans la figure 4, modifié les rayons des parties circulaires de C' et Γ' ce qui, on le verra plus loin, est licite tant qu'ils sont <1).

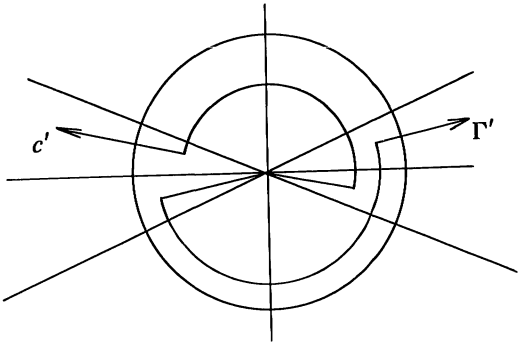


Figure 4

On peut récrire la condition aux limites :

$$(10)\quad \int_{C'} \exp\left[i\frac{k}{2}\rho\left(u+\frac{1}{u}\right)\right]\cdot\left[(\mu-2\Omega)u-(\mu+2\Omega)\frac{1}{u}\right]\cdot g(ue^{-i\beta})\,e^{-i\beta}\,du$$

$$-\int_{\Gamma'} \exp\left[i\frac{k}{2}\rho\left(u+\frac{1}{u}\right)\right]\cdot\left[(\mu+2\Omega)u-(\mu-2\Omega)\frac{1}{u}\right]\cdot h(ue^{i\beta})\,e^{i\beta}\,du=0$$

qui suggère que g et h satisfassent à :

$$\left[(\mu - 2\Omega)\, u - (\mu + 2\Omega)\,\frac{1}{u}\right]\; g(u\, e^{-i\beta})\, e^{-i\beta}$$

$$= \left[(\mu + 2\Omega)\, u - (\mu - 2\Omega)\,\frac{1}{u}\right]\; h(u\, e^{i\beta})\, e^{i\beta}$$

ou :

$$(11) \qquad g(\zeta) = \frac{\sigma^2\,\zeta^2\, e^{2i\beta} - 1}{\zeta^2 e^{2i\beta} - \sigma^2}\; h(\zeta e^{2i\beta})\, e^{2i\beta}$$

Par le théorème de Cauchy on voit que la condition (6) sera satisfaite s'il n'existe pas de singularité de $\left(\sigma^2 u - \frac{1}{u}\right) h(u\, e^{i\beta})$ sur le domaine connexe du feuillet $-\pi \leqslant \arg \zeta \leqslant \pi$ dont les frontières sont C' et Γ'.

La seconde condition aux limites, relative au bord $\theta = -\beta$, traitée de manière analogue conduit à :

$$(12)\; -\int_{C''} \exp\left[i\frac{k}{2}\rho\left(u + \frac{1}{u}\right)\right] \cdot \left[(\mu - 2\Omega)\, u - (\mu + 2\Omega)\frac{1}{u}\right] \cdot g(u\, e^{i\beta})\, e^{i\beta}\, du$$

$$+ \int_{\Gamma''} \exp\left[i\frac{k}{2}\rho\left(u + \frac{1}{u}\right)\right] \cdot \left[(\mu + 2\Omega)\, u - (\mu - 2\Omega)\frac{1}{u}\right] \cdot h(u\, e^{-i\beta})\, e^{-i\beta}\, du$$

C'' et Γ'' déduits de C et Γ (fig. 3) respectivement par les rotations $[0\,, -\beta]$ et $[0\,, +\beta]$.

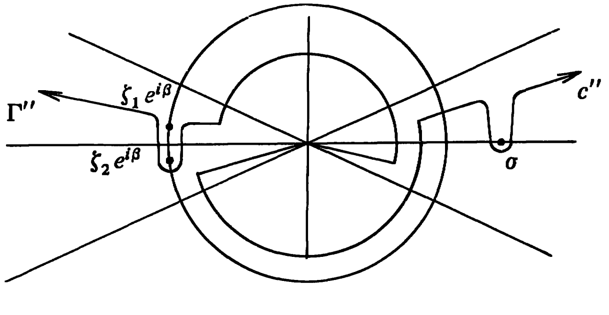


Figure 5

De (12) on est conduit à l'équation fonctionnelle :

$$(13) \qquad g(\zeta) = \frac{\sigma^2 \zeta^2 e^{-2i\beta} - 1}{\zeta^2\, e^{-2i\beta} - \sigma^2}\;\; h(\zeta\, e^{-2i\beta})\, e^{-2i\beta}$$

Avec $H(\zeta) = \zeta\, h(\zeta)$, $G(\zeta) = \zeta\, g(\zeta)$ l'on va considérer :

$$(14) \qquad H(\zeta\, e^{2i\beta}) = \frac{\zeta^2 e^{2i\beta} - \sigma^2}{\sigma^2 \zeta^2 e^{2i\beta} - 1} \cdot \frac{\sigma^2 \zeta^2 e^{-2i\beta} - 1}{\zeta^2 e^{-2i\beta} - \sigma^2}\; H(\zeta\, e^{-2i\beta})$$

$$(15) \qquad G(\zeta) = \frac{\sigma^2 \zeta^2 e^{-2i\beta} - 1}{\zeta^2 e^{-2i\beta} - \sigma^2} H(\zeta e^{-2i\beta}) = \frac{\sigma^2 \zeta^2 e^{2i\beta} - 1}{\zeta^2 e^{2i\beta} - \sigma^2} H(\zeta e^{2i\beta})$$

Solution des équations (14) (15).

On cherche une solution particulière de (14) en posant :

$$(16) \qquad H(z^{2\beta/\pi}) = \exp L(z)$$

$L(z)$ holomorphe dans $-\pi \leqslant \arg z \leqslant \pi$, satisfaisant à :

$$L(r\, e^{i\pi}) = m(r) + L(r\, e^{-i\pi}) \quad , \quad r > 0$$

avec :

$$m(r) = \log q(r)$$

$$(17) \qquad q(r) = \frac{r^{4\beta/\pi} e^{2i\beta} - \sigma^2}{\sigma^2 r^{4\beta/\pi} e^{2i\beta} - 1} \cdot \frac{\sigma^2 r^{4\beta/\pi} e^{-2i\beta} - 1}{r^{4\beta/\pi} e^{-2i\beta} - \sigma^2}$$

et $\arg \log q(r) = 0$, si $r = 0$ ou $r = \infty$.

On note que :

$$|q(r)| = 1 \quad , \quad q(r) = q\left(\frac{1}{r}\right)$$

et

$$m(r) = i \arg q(r) \quad , \quad m(r) = O\,(r^{-4\beta/\pi}) \quad \text{si} \quad r \to \infty$$

$$= O\,(r^{4\beta/\pi}) \quad \text{si} \quad r \to 0$$

On obtient :

$$L(z) = -\frac{1}{2\pi} \int_0^{+\infty} \frac{\arg q(r)\, dr}{r + z} \quad , \quad \text{dans} \quad -\pi \leqslant \arg z \leqslant \pi$$

d'où :

$$(18) \quad H(\zeta) = \exp \left\{ -\frac{1}{2\pi} \int_0^{+\infty} \frac{\arg q(r)\, dr}{r + \zeta^{\pi/2\beta}} \right\} \quad , \quad \text{dans} \quad -2\beta \leqslant \arg \zeta \leqslant 2\beta$$

On peut établir directement :

(a)
(19) $(H(\zeta) - 1)\, \zeta^{\pi/2\beta}$ est borné dans $-2\beta \leqslant \arg \zeta \leqslant 2\beta$

(b)
(20) $H(0) = \sigma^{2-\pi/2\beta}$, $(H(\zeta) - \sigma^{2-\pi/2\beta}) \cdot \zeta^{-\pi/2\beta}$ est borné dans
$-2\beta \leqslant \arg \zeta \leqslant 2\beta$

(c) $H(\zeta) = \overline{H(\bar{\zeta})}$

Il est clair que le prolongement analytique de $H(\zeta)$ à l'extérieur du secteur $-2\beta \leqslant \arg \zeta \leqslant 2\beta$ sera obtenu grâce à (14) ; il est évident, par (18), que $H(\zeta)$ n'a ni zéro, ni pôle dans $-2\beta \leqslant \arg \zeta \leqslant 2\beta$; grâce à (14) on voit que cette conclusion peut être étendue au secteur $-(\pi + \beta) < \arg \zeta < \pi + \beta$. On obtient $G(\zeta)$ par (15) et l'on voit que, dans le secteur $-\pi \leqslant \arg \zeta \leqslant \pi$, $G(\zeta)$ a pour seuls pôles : $\zeta = \sigma\, e^{-i\beta}$, $\zeta = \sigma\, e^{i\beta}$ pour seuls zéros : $\zeta = \dfrac{1}{\sigma}\, e^{-i\beta}$, $\zeta = \dfrac{1}{\sigma}\, e^{i\beta}$, tous simples.

Les équations (15) montrent que $H(\zeta)\cdot G(\zeta)$ est périodique par rapport à $\arg\zeta$, de période 2β ; compte tenu des singularités connues dans $|\arg\zeta| \leqslant \beta$, de (19) et (20) on obtient par le théorème de Liouville :

$$H(\zeta)\cdot G(\zeta)\cdot\frac{\zeta^{\pi/\beta}+\sigma^{\pi/\beta}}{\zeta^{\pi/\beta}+\left(\frac{1}{\sigma}\right)^{\pi/\beta}} = \sigma^2 \tag{21}$$

Multipliant les deux membres de (14) par $H(\zeta)$, on reconnait que

$$H(\zeta\, e^{i\beta})\cdot H(\zeta\, e^{-i\beta})\;\frac{\sigma^2\,\zeta^2-1}{\zeta^2-\sigma^2}$$

est périodique par rapport à $\arg\zeta$, de période 2β ; par examen des zéros et des pôles dans $-\beta \leqslant \arg\zeta \leqslant \beta$, utilisant (19), (20) et le théorème de Liouville on trouve :

$$H(\zeta\, e^{i\beta})\; H(\zeta\, e^{-i\beta})\;\frac{\sigma^2\zeta^2-1}{\zeta^2-\sigma^2}\;\cdot\;\frac{\zeta^{\pi/\beta}-\sigma^{\pi/\beta}}{(\sigma\zeta)^{\pi/\beta}-1} = \sigma^{2-\pi/\beta} \tag{22}$$

Enfin de (14) on observe que $H(\zeta)\cdot H\left(\frac{1}{\zeta}\right)$, $\left(\arg\frac{1}{\zeta}=-\arg\zeta\right)$, est périodique par rapport à $\arg\zeta$ de période 4β, d'où l'on déduit :

$$H(\zeta)H\left(\frac{1}{\zeta}\right)=\sigma^{2-\pi/\beta}\,. \tag{23}$$

Finalement on adoptera pour solution de (11) et (13) :

$$h(\zeta)=\frac{1}{\zeta}H(\zeta)\cdot\left[c+\frac{a\,\zeta_1{}^{\pi/2\beta}}{\zeta^{\pi/2\beta}-\zeta_1{}^{\pi/2\beta}}+\frac{b\,\zeta_2{}^{\pi/2\beta}}{\zeta^{\pi/2\beta}-\zeta_2{}^{\pi/2\beta}}\right] \tag{24}$$

$$g(\zeta)=\frac{1}{\zeta}H(\zeta)\cdot\left[c-\frac{a\,\zeta_1{}^{\pi/2\beta}}{\zeta^{\pi/2\beta}+\zeta_1{}^{\pi/2\beta}}-\frac{b\,\zeta_2{}^{\pi/2\beta}}{\zeta^{\pi/2\beta}+\zeta_2{}^{\pi/2\beta}}\right] \tag{25}$$

où $H(\zeta)$, $G(\zeta)$ sont définis via (18) et (15), a, b, c désignant des constantes qu'on précisera plus loin, ζ_1, ζ_2 les singularités introduites plus haut. On s'assure de façon immédiate que les deux représentations adoptées pour $\varphi(x\,,y)$ suivant que $-\beta \leqslant \theta \leqslant 0$ ou $0 \leqslant \theta \leqslant \beta$ sont prolongement analytique l'une de l'autre.

Retour sur les conditions aux limites.

La première condition aux limites est satisfaite, car l'on sait que $\left(\sigma^2 u-\frac{1}{u}\right)$ $h\,(u\, e^{i\beta})$ n'a pas de pôle sur le domaine connexe du feuillet $-\pi \leqslant \arg\zeta \leqslant \pi$ dont les frontières sont C' et Γ' (fig. 4).

La deuxième condition aux limites nous amène à considérer les singularités de $\left(\sigma^2 u-\frac{1}{u}\right)$ $h\,(u\, e^{-i\beta})$ dans le domaine connexe du feuillet $-\pi \leqslant \arg\zeta \leqslant \pi$

de frontière C'' et Γ'' (fig. 5). On y trouve les pôles simples $\zeta_1 e^{i\beta}$, $\zeta_2 e^{i\beta}$ et les résidus correspondants pour (12) sont :

$$\frac{2a\beta}{\pi}\exp\left[\frac{ik\rho}{2}\left(\zeta_1 e^{i\beta}+\frac{1}{\zeta_1 e^{i\beta}}\right)\right]\cdot\left(\sigma^2\zeta_1 e^{i\beta}-\frac{1}{\zeta_1 e^{i\beta}}\right)H(\zeta_1)$$

$$\frac{2b\beta}{\pi}\exp\left[\frac{ik\rho}{2}\left(\zeta_2 e^{i\beta}+\frac{1}{\zeta_2 e^{i\beta}}\right)\right]\cdot\left(\sigma^2\zeta_2 e^{i\beta}-\frac{1}{\zeta_2 e^{i\beta}}\right)H(\zeta_2)$$

On apprécie la raison du choix $\zeta_1 e^{i\beta}=\frac{1}{\zeta_2 e^{i\beta}}$, qui permet de rendre nulle la somme de ces résidus en imposant à a et b :

$$(26)\qquad a\left(\sigma^2\zeta_1 e^{i\beta}-\frac{1}{\zeta_1 e^{i\beta}}\right)H(\zeta_1)+b\left(\sigma^2\zeta_2 e^{i\beta}-\frac{1}{\zeta_2 e^{i\beta}}\right)H(\zeta_2)=0$$

Etude de la solution au voisinage de l'arête.

Compte tenu des identités (21), (23), des résultats (19) et (20) l'on obtient, pour que la solution $\varphi(x, y)$ demeure bornée au voisinage de $x = y = 0$, la condition :

$$(27)\qquad c=\frac{a+b}{1+\sigma^{\pi/\beta}}$$

Les dérivées partielles premières de $\varphi(x, y)$ sont alors de l'ordre de $\rho^{-(1-\pi/2\beta)}$ près de $x = y = 0$.

Etude du comportement de la solution à l'infini.

Elle fait jouer le rôle essentiel aux singularités ζ_1, ζ_2 de $h(\zeta)$ et $\sigma e^{i\beta}$ de $g(\zeta)$ $\left(\text{éventuellement } \zeta_1 e^{-2i\beta} \text{ si } \beta>\frac{2\pi}{3}\right)$; on modifie les contours C et Γ de manière qu'ils soient constitués de parties rectilignes aboutissant au cercle unité et d'arcs portés par ce même cercle, processus qui fait apparaitre éventuellement des résidus et l'on s'assure que les intégrales prises sur les parties rectilignes, celles prises sur les parties circulaires tendent vers 0 respectivement comme $\frac{1}{\rho}$, $\rho^{-1/2}$. On se contentera d'indiquer les conclusions relatives au cas $\pi-\beta<\arg\zeta_1<\beta$: avec $\theta_1=\arg\zeta_1$, $\theta_2=\arg\zeta_2$ on trouve pour $\rho\sim\infty$:

$$\begin{aligned}
&\theta_1<\theta\leqslant\beta && \varphi\sim 0\\
&\theta_2<\theta<\theta_1 && \varphi\sim-4i\beta a\,H(\zeta_1)\exp\left[ik(x\cos\theta_1+y\sin\theta_1)\right]\\
&-\beta<\theta<\theta_2 && \varphi\sim-4i\beta a\,H(\zeta_1)\exp\left[ik(x\cos\theta_1+y\sin\theta_1)\right]\\
& && \quad -4i\beta b\,H(\zeta_2)\exp\left[ik(x\cos\theta_2+y\sin\theta_2)\right]\\
&\theta=-\beta && \varphi\sim-4i\beta(a\,H(\zeta_1)+b\,H(\zeta_2))\exp\left[ik\rho\cos(\theta_1+\beta)\right]\\
& && \quad +p\exp\left[i\frac{k}{2}\rho\left(\sigma+\frac{1}{\sigma}\right)\right]
\end{aligned}$$

On observe ainsi l'apparition d'une onde de bord sur la face $\theta = -\beta$, qui correspond à la singularité $\zeta = \sigma e^{i\beta}$ de $g(\zeta)$; on a :

$$p = i\pi \frac{\sigma^4 - 1}{\sigma^2} \left(c - \frac{a \zeta_1^{\pi/2\beta}}{i \sigma^{\pi/2\beta} + \zeta_1^{\pi/2\beta}} - \frac{b \zeta_2^{\pi/2\beta}}{i \sigma^{\pi/2\beta} + \zeta_2^{\pi/2\beta}} \right) H(\sigma e^{-i\beta})$$

On peut normaliser la solution obtenue en prenant $a = \dfrac{i}{4\beta H(\zeta_1)}$ et calculant b et c par (26) et (27).

La solution construite correspond à la diffraction par le dièdre de l'onde incidente $\exp [ik(x \cos \theta_1 + y \sin \theta_1)]$.

Ajoutons enfin que la différence φ^* entre la solution φ et les termes qui la représentent asymptotiquement à l'infini satisfait à la condition de radiation :

$$\lim_{\rho \to \infty} \rho^{1/2} \left| \frac{\partial \varphi^*}{\partial \rho} - ik\varphi^* \right| = 0 .$$

REFERENCES

[1] Roseau M. — Diffusion dans un sol perméable d'ondes liquides entretenues par la marée. *Annales scientifiques de l'Ecole Normale Supérieure,* t. 77, 1960, p. 1-40.

[2] Roseau M. — Sur le problème de Sommerfeld. *Journal de Mécanique,* 1967.

Unïversité Paris VI
Dept. de Mécanique
11, Quai Saint Bernard,
Paris 5$^{\text{ème}}$ (France)

Actes, Congrès intern. Math., 1970. Tome 3, p. 151 à 156.

E4 - THÉORIE DU CONTROLE OPTIMAL

A CONSTRUCTIVE APPROACH TO THE MAXIMUM PRINCIPE

par A.V. BALAKRISHNAN [+]

In control theory, the Maximum Principle has for the most part been a theory of necessary conditions that a postulated optimal solution must satisfy. There is a weakness here inherent in proving properties of an empty set, and Young [10] for instance cites a paradox of Perron in illustration. In partial resolution, McShane [11] and Young [10] (among others) prove existence of "relaxed" controls (also known as "chattering" controls, based on the "generalized curves" of Young), and then deduce an appropriately modified version of the Maximum Principle. The present approach goes one step further and is totally constructive. We develop a computational scheme which at the same time yields the Maximum Principle for a constructed limiting solution.

To be specific (and to stay within the space limitation) we shall consider a particular class of problems, not the most general.

Minimize

$$\int_0^T g(t\,;x(t)\,;u(t))\,dt \tag{1}$$

where T is fixed and finite, subject to :

$$x(t) = f(t\,;x(t)\,;u(t)) \quad \text{a.e.} \tag{2}$$

$$x(0) = x_1\,, \; x(T) = x_2 \tag{3}$$

$$\phi(t\,;x(t)\,;u(t)) = 0 \quad \text{a.e.} \tag{4}$$

The controls are Lebesgue measurable, and subject to additional constraints C. We assume that there is at least one such control including (2) thru (4), with finite (1), so that the infimum, denoted $g(0)$, is less than plus infinity.

Our approach is to replace this by an approximating non-dynamic problem. For each $\epsilon > 0$, minimize :

$$\frac{1}{2\epsilon}\int_0^T \|\dot{x}(t) - f(t\,;x(t)\,;u(t))\|^2\,dt + \frac{1}{2\epsilon}\int_0^T \|\phi(t\,;x(t);u(t))\|^2\,dt + \int_0^T g(t\,;x(t)\,;u(t))\,dt \tag{5}$$

[1] Department of System Science, UCLA. Research supported in part under grant No. 68-1 408, AFOSR, Applied Mathematics Division.

over the class of controls $u(t)$, Lebesgue measurable and subject to C, and over the class of absolutely continuous functions $x(t)$ satisfying the end-conditions (3) and

$$(6)\quad \int_0^T (\|\dot{x}(t) - f(t\,;x(t)\,;u(t))\|^2 + \|\phi(t\,;x(t)\,;u(t))\|^2)\,dt \leqslant M < \infty$$

(the last condition can be eliminated in $\inf g(t\,;x\,;u) > -\infty$). Note that because $x(t)$ is not required to satisfy the dynamic equations, we can incorporate additional "phase-plane" constraints, such as inequality constraints for example. Let $u_n(t)$, $x_n(t)$ be a minimizing sequence for (5). Let

$$\delta(\epsilon) = \text{Lim inf}\frac{1}{2}\int_0^T (\|\dot{x}_n(t) - f(t\,;x_n(t)\,;u_n(t))\|^2 + \frac{1}{2}\|\phi(t\,;x_n(t);u_n(t))\|^2$$

$$G(\epsilon) = \text{Lim sup}\int_0^T g(t\,;x_n(t)\,;u_n(t))\,dt$$

and let $d(\epsilon)$ denote the supremum of $\delta(\epsilon)$ over all such minimizing sequences, and correspondingly $g(\epsilon)$ the infimum of $G(\epsilon)$. Let $h(\epsilon)$ denote the infimum of (5), so that

$$(7)\qquad h(\epsilon) = d(\epsilon)/\epsilon + g(\epsilon) \leqslant g(0)$$

We have then the basic estimate on how well the epsilon problem approximates the original problem. Assuming merely that $g(.)$, $\phi(.)$, $f(.)$ are say continuous, we have :

Theorem. – *Suppose $g(\epsilon_0)$ is finite for some ϵ_0. Then $g(\epsilon)$ is finite for every ϵ less than ϵ_0, and $g(\epsilon)$ is monotone non-decreasing as ϵ decreases, and similarly $d(\epsilon)$ is monotone non-increasing. Moreover $h(\epsilon)$ is monotone and is differentiable omitting a countable number of points with*

$$(8)\qquad h'(\epsilon) = -d(\epsilon)/\epsilon^2\ ,$$

and

$$g'(\epsilon) + d'(\epsilon) = 0 \text{ a.e.}$$

Also :

$$(9)\qquad d(\epsilon)/\epsilon \leqslant g(0+) - g(\epsilon) \to 0$$

To proceed further, we need to be more specific. Thus we assume that the functions $f(.), g(.), \phi(.)$ are C^1 in x. Further, since ours is a constructive approach with existence included, we specify the simplest blanket conditions :

$$u(t) \in U \text{ compact a.e.}$$

and

$$[x, f(t\,;x\,;u)] = 0[1 + \|x\|^2]$$

[Alternately we can assume (ad hoc) only the minimal needed properties that these conditions imply, as in McShane [11] for example. We forego this generality in the interest of simplicity, especially since it is not an intrinsic limitation of our approach]. We may then modify (6) as :

$$\int_0^T ||x(t)||^2 \, dt < m < \infty \tag{10}$$

Note that in this case $g(\epsilon)$ in (7) is always finite.

Our computational procedure for minimizing (5) is to pick any "admissible" state function $x(t)$ (i.e., absolutely continuous with given end-conditions, and satisfying (10)). Then we choose a control that minimizes (5). The important point here is the simple one that this can be done by minimizing the integrand in (5). And thereby hangs the Maximum Principle. Let $\mathcal{M}$ denote the class of regular probability measures on (the Lebesgue subsets of) U, and let $C(t\,;x)$ denote the compact convex set of points χ :

$$\chi = \int_U f(t\,;x\,;u)\,d\mu, \quad \int_U \phi(t\,;x\,;u)\,d\mu, \quad \int_U g(t\,;x\,;u)\,d\mu$$

as $d\mu$ ranges over $\mathcal{M}$. This is of course the closed convex hull of the set

$$\{f(t\,;x\,;u), \quad \phi(t\,;x\,;u), \quad g(t\,;x\,;u), \quad u \in U\}$$

It will be convenient to use the notation from now on :

$$\overline{f}(t\,;x\,;\mu) = \int_U f(t\,;x\,;u)\,d\mu$$

and similarly for the other functions. For each point χ in $C(t\,;x)$ let

$$r(\epsilon\,;t\,;y\,;\chi) = \frac{1}{2\epsilon}\,(||y - \overline{f}(t\,;x\,;\mu)||^2 + ||\overline{\phi}(t\,;x\,;\mu) + \overline{g}(t\,;x\,;\mu)$$

and let

$$\overline{m}(\epsilon\,;t\,;y\,;x) = \operatorname*{Inf}_{\chi \epsilon C(t\,;x)} r(\epsilon\,;t\,;y\,;\chi) \tag{11}$$

To obtain the Maximum Principle, let us note that the infimum in (11) is attained, and moreover letting

$$\overline{m}(\epsilon\,;t\,;y\,;x) = r(\epsilon\,;t\,;y\,;\chi_0)$$

we have, since $r(\ldots)$ is simply a quadratic functional on $C(t\,;x)$, that

$$\frac{d}{d\theta} r(\epsilon\,;t\,;y\,;\chi_0 + \theta(\chi - \chi_0))_{\theta=0} \geqslant 0$$

This differentiation yields :

$$[\Psi, \overline{f}(t\,;x\,;\mu_0)] + [\phi_0, \overline{\phi}(t\,;x\,;\mu_0)] - \overline{g}(t\,;x\,;\mu_0) = \\ \operatorname{Max}\, [\Psi, \overline{f}(t\,;x\,;\mu)] + [\phi_0, \overline{\phi}(t\,;x\,;\mu)] - \overline{g}(t\,;x\,;\mu) \tag{12}$$

where

$$\chi_0 = \overline{f}(t\,;x\,;\mu_0),\ \overline{\phi}(t\,;x\,;\mu_0),\ \overline{g}(t\,;x\,;\mu_0)\ ;\ \Psi = (y - \overline{f}(t\,;x\,;\mu_0))/\epsilon\ ;$$

$$\phi_0 = \overline{\phi}\,(t\,;x\,;\mu_0)/\epsilon$$

and is recognized as the prototype of the Maximum Principle. Next let us note that if $x_n(.)$, $u_n(.)$ is a minimizing sequence for (5), then we may take a subsequence of $x_n(t)$ to converge uniformly to $x_\epsilon(t)$ say, while there is a relaxed control $d\mu_\epsilon(t ; u)$ such that

$$\int_0^T dt \int_U k(t ; u)\, d\mu_\epsilon(t ; u) = \text{Lim} \int_0^T k(t ; u_n(t))\, dt$$

for every continuous $k(t ; u)$. Moreover

$$h(\epsilon) = \int_0^T \overline{m}(t ; x_\epsilon(t) ; \dot{x}_\epsilon(t))$$

$h(\epsilon)$ being also the infimum in the class of relaxed controls.

Next let

$$\Psi(\epsilon ; t) = (\dot{x}_\epsilon(t) - \overline{f}(t ; x_\epsilon(t) ; \mu_\epsilon(t)))/\epsilon$$

$$\phi_\epsilon(t) = \overline{\phi}(t ; x_\epsilon ; (t) ; \mu_\epsilon(t))/\epsilon$$

Then a first variation yields :

$$(13) \quad \dot{\Psi}(\epsilon ; t) + \overline{f_1}(t ; x_\epsilon(t) ; \mu_\epsilon(t))^* \, \Psi(\epsilon ; t) - \overline{\phi}_1(t ; \mu_\epsilon(t))^* \, \phi_\epsilon(t) - \overline{g}_1(t ; x_\epsilon(t) ; \mu_\epsilon(t)) = 0$$

where the subscripts denote gradient with respect to x. An obvious substitution in (13) yields the approximate Maximum Principle :

$$(14) \quad [\Psi(\epsilon ; t), \overline{f}(t ; x_\epsilon(t) ; \mu_\epsilon(t))] + [\phi_\epsilon(t), \overline{\phi}(t ; x_\epsilon(t) ; \mu_\epsilon(t))]$$
$$-\overline{g}(t;x_\epsilon(t);\mu_\epsilon(t)). = \underset{\mu}{\text{Max}}\{[\Psi(\epsilon;t), \overline{f}(t ; x(t) ; \mu)] + [\phi_\epsilon(t), \overline{\phi}(t ; x(t) ; \mu)] - \overline{g}(t ; x_\epsilon(t) ; \mu)\}$$

Finally let ϵ go to zero. Then taking a suitable subsequence, $x_\epsilon(t)$ converges uniformly to $x_0(t)$ say, and $d\mu_\epsilon(t ; u)$ in the weak-star sense to $d\mu_0(t ; u)$ say, such that

$$g(0 +) = \int_0^T \overline{g}(t ; x_0(t) ; \mu_0(t))\, dt$$

$$\dot{x}_0(t) = \overline{f}(t ; x_0(t) ; \mu_0(t)) \text{ a.e.}$$

$$\overline{\phi}(t ; x_0(t) ; \mu_0(t)) = 0 \text{ a.e.}$$

Also $g(0 +)$ is the infimum for the control problem in the class of relaxed controls. The question now is that of taking limits in (13) and (14). We note that we need only the limits in the weak sense in (L_2). We have no problems if for some sequence ϵ_n going to zero, and some t_0,

$$||\Psi(\epsilon_n ; t_0)||^2 + \int_0^T ||\phi_{\epsilon_n}(t)||^2 \, dt < \infty$$

Suppose then that

$$\liminf_{\epsilon \to 0} \|\Psi(\epsilon\,;0)\|^2 + \int_0^T \|\phi\epsilon(t)\|^2\, dt = +\infty$$

If

$$\liminf \int_0^T \|\phi_\epsilon(t)\|^2\, dt < \infty$$

then we divide thru in (14) by

$$k_n = \|\Psi(\epsilon_n\,;0)\|$$

and working with a subsequence, we obtain in the limit :

$$[\Psi(t), f(t\,; x_0(t)\,; \mu_0(t))] = \text{Max } [\Psi(t)\,; f(t\,; x_0(t)\,; \mu)]$$

$$\Psi(t) + f_1(t\,; x_0(t)\,; \mu_0(t))^* \,\Psi(t) = 0 \;;\; \Psi(0) \neq 0$$

Otherwise we divide by k_n where

$$k_n^2 = \|\Psi(\epsilon_n\,;0)\|^2 + \int_0^T \|\phi_{\epsilon_n}(t)\|^2\, dt$$

In this case there is the possibility that the [weak] limits :

$$\lim \Psi(\epsilon_n\,;0)/k_n = 0 = \lim \phi_{\epsilon_n}(t)/k_n$$

are both zero. To avoid this, additional conditions involving derivatives of $\phi(t\,;x\,;u)$ with respect to u can be imposed, for example those in [12]. The key consideration is whether $d(\epsilon)/\epsilon^2$ is finite as ϵ goes to zero and has a computational significance, as (8) shows. If the solution to the problem (5) is unique, then $h(\epsilon)$ is actually absolutely continuous in $\epsilon > 0$. The references listed may be consulted for more detail as well as application to other classes of problems.

REFERENCES

[1] BALAKRISHNAN A.V. — On a New Computing Technique in Optimal Control, *S.I.A.M. Journal on Control,* May 1968.

[2] BALAKRISHNAN A.V. — A Computational Approach to the Maximum Principle, to be published in the *Journal of Computer and System Sciences.* Also, available as Report No. 70-70, U.C.L.A.-Engr, 1970.

[3] HESTENES M.R. — Multiplier and Gradient Methods, *Journal of Optimization Theory and Applications,* November 1969.

[4] HUANG S.-C. — A Constructive Approach to the Maximum Principle for Differential-Difference Problems using Balakrishnan's ε-Technique, *Journal of Optimization Theory and Applications,* January 1970.

[5] JONES A.P. and MCCORMICK G.P. — A Generalization of the Method of Balakrishnan : Inequality Constraints and Initial Conditions, *S.I.A.M. Journal on Control,* May 1970.

[6] DE JULIO S. — Numerical Solution of Dynamical Optimization Problems, *S.I.A.M. Journal on Control,* May 1970.

[7] LIONS J.L. — *Controle Optimal de Systemes Gouvernes par des Equations aux Derivées Partielles,* Dunod, Paris, 1968.

[8] BALAKRISHNAN A.V. — On a New Computing Technique in System Identification, *Journal of Computer and System Sciences,* June 1968.

[9] SAKAWA Y. — An Application of the Balakrishnan Epsilon Technique to the Solution of Pursuit-Evasion Games », to be published in the *Journal of Computer & System Sciences.*

[10] YOUNG L.C. — *Calculus of Variations and Optimal Control Theory,* W.B. Saunders, 1969.

[11] MCSHANE E.J. — Relaxed Controls and Variational Problems, *S.I.A.M. Journal on Control,* August 1967.

[12] BALAKRISHNAN A.V. — The Epsilon Technique. A Constructive Approach to the Maximum Principle, in *Control Theory and the Calculus of Variations,* Academic Press, 1969.

[13] TAYLOR L.W. et al. — *Experience Using Balakrishnan's Epsilon Technique to Compute Flight Profiles,* A.I.A.A. Paper No. 69-75.

University of California
Dept. of System Science
Los Angeles,
California 90 024 (USA)

Actes, Congrès intern. Math., 1970. Tome 3, p. 157 à 161.

EXISTENCE THEOREMS FOR PROBLEMS OF OPTIMIZATION WITH DISTRIBUTED AND BOUNDARY CONTROLS

by Lamberto CESARI

1. Existence theorems with state equations in strong form.

We are interested here in control problems in a fixed domain $g \subset E_\nu$ for which the state variable x is an element of a Banach space S with norm $\|x\|$, for which state equations –in either strong or weak forms– and unilateral constraints are expressed in terms of general not necessarily linear operators on S, mapping S into vector-valued L-integrable functions on G and ∂G respectively, and of arbitrary measurable vector-valued control functions u on G and v on ∂G.

Precisely, we denote by Γ a closed subset of ∂G, by μ a suitable measure function on Γ, by T the set of all measurable vector functions $u(t) = (u^1, \dots, u^m)$, $t \in G$ (distributed controls), and by $\overset{\circ}{T}$ the set of all μ-measurable vector functions $v(t) = (v^1, \dots, v^{m'})$, $t \in \Gamma$ (boundary controls). We denote by $\mathcal{L}$, $\mathcal{J}$, $\mathcal{M}$, $\mathcal{K}$ given operators $\mathcal{L} : S \to (L_p(G))^r$, $\mathcal{J} : S \to (L_p(\Gamma))^{r'}$, $\mathcal{M} : S \to (L_p(G))^s$, $\mathcal{K} : S \to (L_p(\Gamma))^{s'}$, where $p \geqslant 1$, and r, r', s, s' are integers. The images under $\mathcal{L}, \mathcal{J}, \mathcal{M}, \mathcal{K}$ of elements $x \in S$ will be denoted by $z, \mathring{z}, y, \mathring{y}$, or $z(t) = (z^1, \dots, z^r) = (\mathcal{L}x)(t)$, $t \in G$; $\mathring{z}(t) = (\mathring{z}^1 \dots, \mathring{z}^{r'}) = (\mathcal{J}x)(t)$, $t \in \Gamma$; $y(t) = (y^1, \dots, y^s) = (\mathcal{M}x)(t)$, $t \in G$; $\mathring{y}(t) = (\mathring{y}^1, \dots, \mathring{y}^{s'}) = (\mathcal{K}x)(t)$, $t \in \Gamma$.

We consider the problem of finding elements $x \in S$, $u \in T$, $v \in \overset{\circ}{T}$, so as to minimize a cost functional of the form

$$(1)\quad I[x, u, v] = \int_G f_0(t, (\mathcal{M}x)(t), u(t))\, dt + \int_\Gamma g_0(t, (\mathcal{K}x)(t), v(t))\, d\mu,$$

subject to state equations (strong form)

$$(2)\qquad (\mathcal{L}x)(t) = f(t, (\mathcal{M}x)(t), u(t)) \qquad \text{a.e. in } G,$$

$$(3)\qquad (\mathcal{J}x)(t) = g(t, (\mathcal{M}x)(t), v(t)) \qquad \mu\text{-a.e. on } \Gamma,$$

and unilateral constraints on the values of $u, v, y, \mathring{y}$ of the forms

$$(4)\qquad u(t) \in U(t, (\mathcal{M}x)(t)) \subset E_m \qquad \text{a.e. in } G,$$

$$(5)\qquad v(t) \in V(t, (\mathcal{K}x)(t)) \subset E_{m'} \qquad \mu\text{-a.e. on } \Gamma,$$

$$(6)\qquad (\mathcal{M}x)(t) \in A(t) \subset E_s \qquad \text{a.e. in } G,$$

$$(7)\qquad (\mathcal{K}x)(t) \in B(t) \subset E_{s'}, \qquad \mu\text{-a.e. on } \Gamma,$$

Here we assume that for any $t \in cl\, G$ a given subset $A(t)$ of E_s is assigned, and we denote by A the set of all $(t, y) \in E_{\nu+s}$ with $t \in cl\, G$, $y \in A(t)$. We assume that for any $(t, y) \in A$ a given subset $U(t, y)$ of E_m is assigned (distributed control space), and we denote by M the set of all (t, y, u) with $(t, y) \in A$, $u \in U(t, y)$. Then, $f_0(t, y, u)$, $f(t, y, u) = (f_1, \ldots, f_r)$ are given functions on M. Analogously, we assume that for any $t \in \Gamma$ a given subset $B(t)$ of $E_{s'}$ is assigned, and we denote by B the set of all $(t, \overset{\circ}{y}) \in E_{\nu+s'}$, with $t \in \Gamma$, $\overset{\circ}{y} \in B(t)$. We assume that for any $(t, \overset{\circ}{y}) \in B$ a given subset $V(t, \overset{\circ}{y})$ of $E_{m'}$ is assigned (boundary control space), and we denote by $\overset{\circ}{M}$ the set of all $(t, \overset{\circ}{y}, \nu)$ with $(t, \overset{\circ}{y}) \in B$, $\nu_o \in V(t, \overset{\circ}{y})$. Then, $g_0(t, \overset{\circ}{y}, \nu)$, $g(t, \overset{\circ}{y}, \nu) = (g_1, \ldots, g_{r'})$ are given functions on $\overset{\circ}{M}$. We denote by $I_1[x, u]$ and $I_2[x, \nu]$ the two integrals in (1) on G and Γ respectively, whose sum is $I[x, u, \nu]$.

Usually, S is a Sobolev space in G, state equation (2) represents a system of r partial differential equations in G, and state equation (3) represents either boundary data, or a system of r' partial differential equations on Γ (or on ∂G). But the situation may be quite different, since the operators $\mathcal{L}$, $\mathcal{J}$, $\mathcal{M}$, $\mathcal{K}$ need not be differential operators. All we shall require is a set of axioms (P) in the present setting, as well as in the setting of n°. 2 with state equations in the weak form. In all cases the existence theorems follow in a natural way from uniformly proved closure theorems and lower closure theorems, extending the usual lower semicontinuity theorems for classical free problems of the calculus of variations.

(P_1) **Hypotheses on** $\mathcal{L}$, $\mathcal{J}$, $\mathcal{M}$, $\mathcal{K}$. If x, $x_k \in S$, $k = 1, 2, \ldots$, and $x_k \to x$ weakly in S as $k \to \infty$, then $\mathcal{L}x_k \to \mathcal{L}x$ weakly in $(L_p(G))^r$, $\mathcal{J}x_k \to \mathcal{J}x$ weakly in $(L_p(\Gamma))^{r'}$, $\mathcal{M}x_k \to \mathcal{M}x$ strongly in $(L_p(G))^s$, $\mathcal{K}x_k \to \mathcal{K}x$ strongly in $(L_p(\Gamma))^{s'}$.

(P_2) **Hypotheses on the set** Γ **and measure** μ. We assume that the closed set $\Gamma \subset \partial G$ is the union of finitely many sets $\Gamma_1, \ldots, \Gamma_N$, each Γ_j being the image of a $(\nu - 1)$-dimensional interval I under a transformation T_j of class K (Morrey), say $T_j : I \to \Gamma_j$. To simplify the exposition we assume here that μ is the hyperarea measure defined on Γ by the mappings T_j.

(P_3) **Hypotheses on** f_0, g_0, f, g. We assume that the sets A, B, M, $\overset{\circ}{M}$ are closed, that the functions f_0, $f = (f_1, \ldots, f_r)$ are continuous on M, and that the functions g_0, $g = (g_1, \ldots, g_{r'})$ are continuous on $\overset{\circ}{M}$. Also, we assume that there is some integrable function $\phi(t) \geqslant 0$, $t \in G$, such that $f_0(t, y, u) \geqslant -\phi(t)$ for all $(t, y, u) \in M$; and that there is some μ-integrable function $\psi(t) \geqslant 0$, $t \in \Gamma$, such that $g_0(t, \overset{\circ}{y}, \nu) \geqslant -\psi(t)$ for all $(t, \overset{\circ}{y}, \nu) \in \overset{\circ}{M}$.

These conditions, however, can be relaxed. For instance, f_0 and g_0 may be assumed to be only lower semicontinuous on M and $\overset{\circ}{M}$ respectively. Also, the bounds below can be relaxed as follows : For every point $(\bar{t}, \bar{y}) \in A$ there are a neighborhood $N_\delta(\bar{t}, \bar{y})$, real numbers r, $b = (b_1, \ldots, b_r)$, and an L-integrable function $\phi(t) \geqslant 0$, $t \in N_\delta(\bar{t}, \bar{y})$, such that $f_0(t, y, u) - r - \Sigma_j b_j f_j(t, y, u) \geqslant -\phi(t)$ for all $(t, y) \in N_\delta(\bar{t}, \bar{y})$ and $u \in U(t, y)$. An analogous relaxed requirement can be formulated for g_0.

Kuratowski's concept of upper semicontinuity of variable sets and modifications. We shall discuss these concepts on the sets $U(t, y)$, but they apply to the sets

$V(t, \overset{\circ}{y})$ as well, and to the sets $\tilde{\tilde{Q}}(t,y)$ and $\tilde{\tilde{R}}(t,y)$ we shall define below. For $(\bar{t},\bar{y}) \in A$ and $\delta > 0$ let $N_\delta(\bar{t},\bar{y})$ be the set of all $(t, y) \in A$ at a distance $\leqslant \delta$ from $(\bar{t},\bar{y})$, and let $U(\bar{t},\bar{y};\delta)$ be the union of all $U(t,y)$ with $(t,y) \in N_\delta(\bar{t},\bar{y})$. The sets $U(t,y)$ are said to satisfy property (U) at a point $(\bar{t},\bar{y}) \in A$ provided $U(\bar{t},\bar{y}) = \cap_\delta\, cl\, U(\bar{t},\bar{y};\delta)$. The sets $U(t,y)$ are said to satisfy property (U) in A if they have this property at every $(\bar{t},\bar{y}) \in A$. If A is closed, then the sets $U(t,y)$ satisfy property (U) in A if and only if M is closed. Property (U) is the original Kuratowski concept of upper semicontinuity.

The sets $U(t,y)$ are said to satisfy property (Q) at a point $(\bar{t},\bar{y}) \in A$ provided $U(t,y) = \cap_\delta\, cl\, co\, U(\bar{t},\bar{y};\delta)$. The sets $U(t,y)$ are said to satisfy property (Q) in A if they have this property at every $(\bar{t},\bar{y}) \in A$.

For every $(t,y) \in A$ we denote by $\tilde{\tilde{Q}}(t,y)$ the set of all points $(z^0, z) \in E_{r+1}$ with $z^0 \geqslant f_0(t,y,u)$, $z = f(t,y,u)$, $u \in U(t,y)$. For every $(t,\overset{\circ}{y}) \in B$ we denote by $\tilde{\tilde{R}}(t,\overset{\circ}{y})$ the set of all points $(z^0, z) \in E_{r'+1}$ with $z^0 \geqslant g_0(t,y,v)$, $z = g(t,\overset{\circ}{y},v)$, $v \in V(t,\overset{\circ}{y})$. We shall require below properties (U) or (Q) on the sets $\tilde{\tilde{Q}}(t,y)$ and $\tilde{\tilde{R}}(t,y)$. As proved by Cesari, property (Q) for the sets $Q(t,y)$ and $R(t,\overset{\circ}{y})$ is the natural extension of Tonelli's and McShane's concept of seminormality for free problems. Also, great many criteria for property (Q) have been recently proved. For instance, suitable growth conditions imply property (Q). A triple (x,u,v), $x \in S$, $u \in T$, $v \in \overset{\circ}{T}$, is now said to be admissible provided relations (2), (3), (4), (5), (6), (7) hold, $f_0(t, (\mathfrak{M}x)(t), u(t))$ is L-integrable in G, and $g_0(t, (\mathcal{K}x)(t), v(t))$ is μ-integrable on Γ.

We shall consider nonempty classes Ω of admissible triples (x,u,v). Any such class is said to be closed if the following holds : For any sequence (x_k, u_k, v_k) of elements of Ω, and elements $x \in S$, such that $x_k \to x$ weakly in S,

$$l_1 = \lim I_1[x_k, u_k] < +\infty \quad , \quad l_2 = \lim I_2[x_k, v_k] < +\infty$$

and there are elements $u \in T$, $v \in \overset{\circ}{T}$ such that (x,u,v) is admissible, $I_1[x,u] \leqslant l_1$, $I_2[x,v] \leqslant l_2$, then there is also one of these triples (x,u,v) which also belongs to Ω. This definition of a closed class Ω is justified by lower closure theorems, which precisely guarantee under conditions (P_1), (P_2), (P_3), that there are elements $u \in T$, $v \in \overset{\circ}{T}$ such that (x,u,v) is admissible and $I_1 \leqslant l_1$, $I_2 \leqslant l_2$. Given any class Ω of admissible triples (x,u,v), we denote by $\{x\}_\Omega$ the set of all $x \in S$ such that $(x,u,v) \in \Omega$ for some $u \in T$, $v \in \overset{\circ}{T}$.

EXISTENCE THEOREM 1. – *Under hypotheses (P_1), (P_2), (P_3), if the sets $\tilde{\tilde{Q}}(t,y)$ and $\tilde{\tilde{R}}(t,y)$ have property (Q) in A and B respectively, if Ω is a nonempty closed class of admissible triples (x,u,v) such that the set $\{x\}_\Omega$ is weakly sequentially relatively compact in S, then the cost functional (1) attains its infimum in Ω.*

Condition (Q) above can be relaxed. For instance, it is enough to know that there is a countable decomposition $G = \cup_j H_j$ into measurable disjoint sets H_j, such that H_0 has measure zero, and the sets $\tilde{\tilde{Q}}(t,y)$ have property (Q) on each set $A_j = A \cap (H_j \times E_s)$ with respect to A_j, $j = 1, 2, \ldots$ An analogous remark holds for the sets $\tilde{\tilde{R}}(t,\overset{\circ}{y})$.

If S is a Sobolev space $(W_p^l\,(G))^n$ with $p > 1$, $l \geqslant 1$, $n \geqslant 1$, and norm $\|x\|$, if G is of class K, and $I[x\,,u\,,v] \leqslant C$, $(x\,,u\,,v) \in \Omega$ implies $\|x\| \leqslant c$ for some constant c which may depend on C, then the requirement concerning the set $\{x\}_\Omega$ in Theorem 1 is certainly trivial. For $p = 1$ this is not the case, and often growth conditions must be required which are suitable generalizations of Tonelli's. Nagumo's, and McShane's analogous growth conditions for free problems. For instance, the following growth condition has been found to be relevant : (ϵ) Given $\epsilon > 0$, there is a function $\phi_\epsilon \geqslant 0$, $\phi_\epsilon \in L\,(G)$, such that

$$|f(t\,,y\,,u)| \leqslant \phi_\epsilon\,(t) + \epsilon f_0\,(t\,,y\,,u) \qquad \text{for all} \qquad (t\,,y\,,u) \in M.$$

An analogous growth condition can be expressed in terms of g and g_0.

There are situations where (P_1) holds in a stronger form. We denote by (P_1') the hypothesis analogous to (P_1) where $x_k \to x$ weakly in S implies $\mathcal{J} x_k \to \mathcal{J} x$ strongly in $(L_p\,(\Gamma))^{r'}$ (instead of weakly as in (P_1)).

EXISTENCE THEOREM 2. – *Under hypotheses (P_1'), (P_2), (P_3), if the sets $\tilde{\tilde{Q}}\,(t\,,y)$ have property (Q) in A and the sets $\tilde{\tilde{R}}\,(t\,,y)$ have property (U) in B, if Ω is a nonempty closed class of admissible triples $(x\,,u\,,v)$ such that the set $\{x\}_\Omega$ is weakly sequentially relatively compact, then the cost functional (1) attains its infimum in Ω.*

If some components of $\mathring{y}_k = \mathcal{J} x_k$ converge weakly and the remaining components converge strongly to the corresponding components of $y = \mathcal{J} x$, then an intermediate existence theorem can be proved, where a suitable intermediate property between (U) and (Q) is used (D.E. Cowles).

2. State equations in the weak form.

We consider now the case where the functional equations (2), (3) (state equations) are written in weak form, as it is customary in partial differential equations theory. To this purpose let W denote a suitable normed space of test functions $w = (w_1\,,w_2)$, where $w_1 : G \to (L_q(G))^r$ and $w_2 : \Gamma \to (L_q(\Gamma))^{r'}$ are vector-valued functions defined in G and Γ respectively, and $p^{-1} + q^{-1} = 1$ with the usual conventions. Let W^* be the dual space of W.

We shall use the same general notations as in n°. 1. Instead of the operators $\mathcal{L}$, $\mathcal{J}$, we shall consider only one operator $\mathcal{F} : \cdot S \to W^*$. Instead of the hypothesis (P_1) we shall now require : (P_1'') If $x, x_k \in S$, $k = 1,2,\ldots$, and $x_k \to x$ weakly in S, then $\mathcal{F} x_k \to \mathcal{F} x$ in the weak star topology on W^*, and $\mathcal{M} x_k \to \mathcal{M} x$ strongly in $(L_p(G))^s$, $\mathcal{K} x_k \to \mathcal{K} x$ strongly in $L_p\,(\Gamma))^{s'}$.

For $x \in S$ we denote by $f_{x,u,v}$ the element of W^* defined by

$$f_{x,u,v}\,w = \int_G f(t,\,(\mathcal{M}\,x)\,(t),\,u\,(t))\,w_1(t)\,dt + \int_\Gamma g(t,\,(\mathcal{K}\,x)\,(t),\,v(t))\,w_2\,(t)\,d\mu$$

for all $w = (w_1\,,w_2) \in W$, and where fw_1 and gw_2 are inner products. Instead of the state equations (2), (3), we shall now consider the unique state equation in the weak form $\mathcal{F} = f_{x,u,v}$, or equivalently

(8) $\qquad (\mathcal{F} x)\,w = f_{x,u,v}\,w \qquad \text{for all } w \in W, \qquad$ or

$$(9) \quad (\mathcal{F}x)(w_1, w_2) = \int_G f(t, (\mathcal{M}x)(t), u(t))\, w_1(t)\, dt + \int_\Gamma g(t, (\mathcal{H}x)(t), v(t))\, w_2(t)\, d\mu$$

for all $w = (w_1, w_2) \in W$. We are now interested in the problem of the minimum of a cost functional of the form (1) subject to state equation (8) and unilateral constraints (4), (5), (6), (7).

In the present situation suitable growth conditions must be required, as for instance : (P_4) If $p = 1$ we assume that f_0, f as well as g_0, g satisfy the (ϵ) condition above. If $p > 1$ we simply assume that there are functions $\phi \geqslant 0$, $\phi \in L(G)$, $\varphi \geqslant 0$, $\varphi \in L(\Gamma)$ and constants $a > 0$, $b > 0$ such that

$$|f(t, y, u)|^p \leqslant \phi(t) + af_0(t, y, u) \qquad \text{for all } (t, y, u) \in M\,;$$

$$|g(t, \mathring{y}, v)|^p \leqslant \varphi(t) + bg_0(t, \mathring{y}, v) \qquad \text{for all } (t, \mathring{y}, v) \in \mathring{M}\,;$$

A triple (x, u, v), $x \in S$, $u \in T$, $v \in \mathring{T}$ is now said to be admissible provided relations (4), (5), (6), (7), (8) hold, and $f_0(t, (\mathcal{M}x)(t), u(t))$ is L-integrable in G and $g_0(t, (\mathcal{H}x)(t), v(t))$ is μ-integrable in Γ. We shall then consider nonempty classes Ω of admissible triples (x, u, v), where the definition of closedness is analogous to the one in n°. 1.

EXISTENCE THEOREM 3. – *Under hypotheses (P''), (P_2), (P_3), (P_4), if the sets $\tilde{\tilde{Q}}(t, y)$ and $\tilde{\tilde{R}}(t, y)$ have property (Q) in A and B respectively, if Ω is a nonempty closed class of admissible triples such that the set $\{x\}_\Omega$ is weakly sequentially relatively compact, then the cost functional (1) attains its infimum in Ω.*

Existence theorems analogous to 1, 2, 3 have been also proved, where the operators $\mathcal{L}$, $\mathcal{T}$, $\mathcal{M}$, $\mathcal{H}$, $\mathcal{F}$ depend both on x and suitable components of the controls u and v.

For the results mentioned above, as well as for further ones concerning a number of different aspects of the wide subject, we refer to the work of the author and to the related work of T.S. Angell, R.F. Baum, D.E. Cowles, T. Nishiura, J.R. La Palm, D.A. Sanchez, M.B. Suryanarayana.

University of Michigan,
Dept. of Mathematics,
Ann Arbor, Michigan 48 104
U.S.A.

Actes, Congrès intern. Math., 1970. Tome 3, p. 163 à 167.

OPTIMAL STOCHASTIC CONTROL

by Wendell H. FLEMING

1. Introduction.

Stochastic optimization theories deal with mathematical models of random phenomena for which certain parameters are to be chosen in some "best" way. Frequently the random phenomena occur over time, in which case the optimization model deals with stochastic processes. In some problems the parameters to be optimized are constants ; in others these parameters may be taken as unknown functions of time, or more generally functions of certain data available at each instant of time.

Among such problems, let us consider optimal control models of the following general description. The state of a system (physical, economic, etc.) is at each instant of time t a vector $\xi(t)$ in some n-dimensional R^n. The states evolve according to differential equations (or difference equations if time is discrete), depending on certain control parameters u and also on certain random parameters. At each instant t, certain observations are allowed. The control vector $u(t)$ chosen at time t is to be (in a suitably defined sense) a function of the observations available up to t. The problem is to minimize some functional of the state and control processes on the time interval of operation of the system.

Our purpose is to indicate some results about such problems obtained during recent years, emphasizing the case of continuous time and state parameters. For a more complete background see the books [1], [12] and the survey articles [7], [21].

2. Linear-quadratic problems

One class of problems for which a rather explicit solution exists consists of those of linear regulator type. In a linear regulator problem, the system equations are linear in state and control, and the random parameters enter as an additive white noise term. Linear, white noise corrupted measurements of the states are made. The criterion to be minimized is quadratic and positive. {For a complete formulation and precise statements of results see the references cited above}. The most important result about linear regulators is that the optimal control has the form $u(t) = F(t)\,\hat{\xi}(t)$, where $\hat{\xi}(t)$ is the conditional expectation of $\xi(t)$ given the data observed up to time t. The matrices $F(t)$ satisfy (in the continuous time case) a matrix differential equation of Riccati type. The estimates $\hat{\xi}(t)$ for the states evolve according to the equations of the Kalman filter. Under controllability and observability assumptions the solution is valid on infinite time intervals.

These results have been extended to linear regulator problems for distributed parameter systems, whose states evolve according to linear parabolic partial differential equations. See [3], [14].

3. Dynamic programming methods.

Let us assume that the state $\xi(t)$ itself is observed at each instant t. We consider the following model. The states evolve according to stochastic differential equations

$$(3.1) \qquad d\xi = f(t, \xi(t), u(t))\, dt + \sigma(t, \xi(t))\, dw, \quad t \geqslant s,$$

with initial data $\xi(s) = x$. Here w is a brownian motion process. We require that $u(t) \in K$, where the "control set" K is given. Moreover, $u(t) = Y(t, \xi(t))$, where the function $Y(\cdot, \cdot)$ is a *control policy*. Let $(s, x) \in Q$, where $Q = (T_0, T) \times B$ is a cylinder and B is open with boundary ∂B a compact, smooth manifold. Let

$$J(s, x; Y) = E\left\{\int_s^\tau L(t, \xi(t), u(t))\, dt + \Phi(\tau, \xi(\tau))\right\},$$

where τ is the exit time from Q of $(t, \xi(t))$ and E denotes expected value. The problem is to minimize J, for given initial data (s, x).

Let us assume that : K is compact, convex ; f, σ, L, Φ are $C^{(1)}$ and bounded together with their first order partial derivatives (these assumptions can be weakened [7, 5.5]). Let a $= \frac{1}{2}\sigma\sigma^*$, where * is matrix transpose, and assume

(3.2) the characteristic values of $a(\cdot, \cdot)$ are bounded below by $c > 0$.

In order to insure the existence of an optimal control policy, we admit all bounded measurable Y with values in K. For any such Y the solution ξ of (3.1) together with the initial data is well defined, according to Stroock and Varadhan [20], since we assume (3.2). Let

$$(3.3) \qquad \varphi(s, x) = \min_Y J(s, x; Y).$$

Then φ satisfies the partial differential equation

$$(3.4) \quad \frac{\partial\varphi}{\partial s} + \sum_{i,j=1}^{m} a_{ij}(s, x)\frac{\partial^2\varphi}{\partial x_i\, \partial x_j} + \min_{y \in K}\left[L(s, x, y) + \frac{\partial\varphi}{\partial x} f(s, x, y)\right] = 0$$

with the boundary data

$$(3.5) \qquad \varphi(s, x) = \Phi(s, x), \ (s, x) \in \partial Q - \{T_0\} \times B$$

See [7]. Condition (3.2) means that (3.4) is uniformly parabolic ; and this implies continuity of the partial derivatives of φ appearing in (3.4). A policy Y is optimal if

(3.6) $L(s, x, y) + \frac{\partial\varphi}{\partial x} f(s, x, y) = \min$ on K for $y = Y(s, x)$, almost everywhere in Q.

In many applications of interest (3.2) does not hold. Instead, it may be assumed that both control and noise enter the same components of the system equations (roughly speaking). Rishel [18] has recently shown that φ is a generalized solution of the degenerate parabolic equation (3.4), in the sense of [6, p. 269]. The method of [18] uses a transformation formula of Girsanov [9]. The assumptions in [18] imply that the state process ξ has a transition density. In fact, if Y is a smooth control policy, then the backward operator

$$\partial/\partial s + \Sigma\, a_{ij}\, \partial^2/\partial x_i \partial x_j + \Sigma f_i(s\,, x\,, Y)\, \partial/\partial x_i$$

of the state process is hypoelliptic [11].

The boundary problem (3.4) – (3.5) can rarely be explicitly solved ; the linear regulator is the best known example for which this is possible. Various methods of approximate solution have been proposed and used with some success. One of these, due to Kushner [13], introduces an optimal control problem for Markov chains corresponding to a finite difference analogue of (3.4) – (3.5) {more precisely, [13] considers the autonomous case in which (3.4) is elliptic}. An effective iterative scheme for finding the optimum is given, corresponding to the Gauss-Seidel method for linear equations.

If the noise intensity $|\sigma|$ is small, then one can try to solve approximately the stochastic problem using the (closed-loop) solution of the corresponding deterministic problem (with $\sigma \equiv 0$). The latter problem is a standard one in control theory. For simplicity let $\sigma = (2\epsilon)^{1/2}$ (identity), and denote the solution of (3.4) – (3.5) by φ^ϵ. The second order term in (3.4) is now $\epsilon \Delta_x \varphi^\epsilon$. When $\epsilon = 0, \varphi^0(s,x)$ is the minimum of J in the deterministic problem. Information about approximate solutions can be obtained (because of (3.1), (3.6)) from the formulas

$$(3.7) \qquad \begin{aligned} \varphi^\epsilon &= \varphi^0 + \epsilon\theta + o(\epsilon), \\ \varphi^\epsilon_x &= \varphi^0_x + \epsilon\theta_x + o(\epsilon) \end{aligned}$$

as $\epsilon \to 0$. Here θ satisfies a linear first order equation obtained by formally taking $\partial/\partial\epsilon$ in (3.4) and setting $\epsilon = 0$. The characteristic curves of this equation are optimal trajectories for the deterministic control problem. Formulas (3.7) are not usually correct everywhere. In [8] they are obtained in regions where φ^0 is sufficiently smooth and where there is a well-behaved optimal control policy for the deterministic problem.

4. Optimal stopping problems.

In these problems the control parameters u do not enter the system equations (3.1). Given a function G, the controller wishes to choose a stopping time ζ for the Markov process ξ such that $EG[\zeta, \xi(\zeta)]$ is minimum. The dynamic programming approach leads to a free boundary problem for the backward equation of the ξ process. See Chernoff [4]. A combined optimal stopping and control problem is treated by Grigelionis and Shiryaev [10].

5. Problems with partial observations.

Let us now suppose that the data available to the controller at time t is $\eta(r)$ for $s \leqslant r \leqslant t$, where the "data process" η evolves according to stochastic differential equations

$$(5.1) \qquad d\eta = g(t, \xi(t), \eta(t))dt + \tilde{\sigma}(t)d\tilde{w}, \ s \leqslant t,$$

with $\eta(s) = 0$, where $\tilde{w}$ is a brownian motion independent of w and $\xi(s)$. The problem is now to minimize J among all controls which use (in some sense) only the data available. In a few cases (dynamical equations (3.1), (5.1) linear in ξ, η, u, $\sigma = \sigma(t)$, $\tau \equiv T$ fixed) a separation principle splits the problem into one of Kalman filtering and another of the type in Section 3 [21]. This applies in particular to the linear regulator (Section 2).

The general partially observable problem is quite difficult. Essentially, one should regard as the "state" at time t not $\xi(t)$ but rather the conditional distribution of $\xi(t)$ given the data. The discrete parameter case was treated by Dynkin [5] and others. For the continuous parameter case, a dynamic programming equation corresponding to (3.4) was formally derived by Mortensen [16]. Necessary and sufficient conditions of dynamic programming type were later obtained rigorously by Rishel [17]. For an approach using the theory of conditional Markov processes, see Stratonovich [19].

We have assumed that f, g, σ, $\tilde{\sigma}$ in (3.1), (5.1) are known. When this is not the case one has a system identification problem ; see Balakrishnan [2].

REFERENCES

[1] Astrom K. — *Introduction to Stochastic Control Theory,* New York, Academic Press, 1970.

[2] Balakrishnan A.V. — in *Stochastic Optimization and Control,* Wiley, 1968, p. 65-89.

[3] Bensoussan A. — *Rend. Mat. Rome* 2, 1969, p. 135-173.

[4] Chernoff H. — *Sankhya Ser.* A 30 1968, p. 221-252.

[5] Dynkin E.B. — *Theor. Probab. Appl.* 10, 1965, p. 1-14.

[6] Fleming W.H. — *J. Math. Anal. Appl.* 16, 1966, p. 254-279.

[7] Fleming W.H. — *S.I.A.M. Review,* 11, 1969, p. 470-509.

[8] Fleming W.H. — *S.I.A.M. J. Control,* submitted.

[9] Girsanov I.V. — *Theor. Probab. Appl.* 5, 1960, p. 285-301.

[10] Grigelionis B.I. and Shiryaev A.N. — *Problemy Peredachi Informacii* 4 1968, p. 60-72.

[11] Hormander L. — *Acta Math.* 119, 1968, p. 147-171.

[12] Kushner H. — *Introduction to Stochastic Control Theory,* Holt, Rinehart and Winston, New York, 1971.

[13] Kushner H. and Kleinman A. — *I.E.E.E. Trans. on Automatic Control,* AC-13, 1968, p. 344-353.

[14] Kushner H. — *S.I.A.M. J. Control* 6 1968, p. 596-614.

[15] Mandl P. — *Theor. Probab. Appl.,* 12, 1967, p. 68-76.

[16] Mortensen R.E. — *Internat. J. Control,* 4, 1966, p. 455-464.

[17] Rishel R. — *S.I.A.M. J. Control,* to appear.

[18] Rishel R. — *Weak solutions of the partial differential equation of dynamic programming,* Bell Labs, Tech. Memo. MM-70-4165-2.

[19] Stratonovich R.L. — *Conditional Markov Processes and their Application to the Theory of Optimal Control,* American Elsevier, 1968.

[20] Stroock D.W. and Varadhan S.R.S. — *Comm. Pure Appl. Math.,* 22, 1969, p. 345-400, p. 479-530.

[21] Wonham W. M. — *Random Differential Equations in Control Theory, Probabilistic Methods* in *Appl. Math.,* Vol. II, Academic Press, 1969.

Brown University
Dept. of Mathematics,
Providence,
Rhode Island 02912 (USA)

Actes, Congrès intern. Math., 1970, Tome 3, p. 169 à 176.

CONDITIONS NÉCESSAIRES DU PREMIER ORDRE DANS LES PROBLÈMES D'EXTREMUM

Par R. V. GAMKRELIDZE et G. L. KHARATISHVILI

1. Introduction.

Après la découverte du principe du maximum de Pontrjaguine dans la théorie du contrôle optimal, [1], [2], la nature logique des conditions nécessaires du premier ordre dans les problèmes d'extremum a été définitivement élucidée : il s'est avéré qu'en fin de compte elles se ramènent à des théorèmes de séparation d'ensembles convexes dans des espaces linéaires.

Il convient de remarquer qu'avant la découverte du principe du maximum, une telle relation a été utilisée à fond par McShane dans son article [3] lors de la démonstration de la règle des multiplicateurs dans le problème de Lagrange sans aucune hypothèse quant à la régularité du problème.

Parmi les nombreux travaux parus après la découverte du principe du maximum et consacrés spécialement à l'étude générale des conditions nécessaires de premier ordre et à la construction axiomatique de la théorie des problèmes extrémaux, qui est liée à cette étude, nous ne noterons que l'ouvrage de Neustadt [4], où est donnée une bibliographie détaillée.

Nous exposerons ici l'une des axiomatiques possibles des problèmes extrémaux et des conditions nécessaires de premier ordre, basée sur le principe du maximum et la notion de régime optimal glissant [5], [6], [7].

2. Cas fini.

Afin de motiver les définitions générales données ci-dessous, nous commencerons par le cas le plus simple : les extrema d'une fonction de n variables.

Sur l'ouvert $\mathcal{O} \subset E_z^n$ de l'espace de dimension n, E_z^n, de points $z = \begin{pmatrix} z^1 \\ \vdots \\ z^n \end{pmatrix}$, soient k fonctions scalaires différentiables $p^1(z), \ldots, p^k(z)$ et soit $\tilde{z} \in \mathcal{O}$ un point d'extremum conditionnel de la fonction $p^1(z)$ avec les conditions $p^2(z) = \ldots p^k(z) = 0$, c'est-à-dire qu'il existe un voisinage du point $\tilde{z}$, soit $V_{\tilde{z}} \subset \mathcal{O}$, tel que pour l'application

$$(1) \qquad P : z \to \begin{pmatrix} p^1(z) \\ \vdots \\ p^k(z) \end{pmatrix} = P(z) \in E_P^k, \qquad z \in \mathcal{O},$$

le point $P(\tilde{z})$ est soit le point "le plus haut" soit "le plus bas" d'intersection de l'axe "vertical" p^1 avec l'image $P(V_{\tilde{z}})$. Si nous désignons par $dP_{\tilde{z}} : E^n_{\delta z} \to E^k_{dp}$ la différentielle de l'application (1) au point $\tilde{z}$, alors la règle des multiplicateurs de Lagrange affirme l'existence d'un k-uplet non nul $\lambda = (\lambda_1, \ldots, \lambda_k)$ orthogonal au sous-espace $dP_{\tilde{z}}(E^n_{\delta z}) \subset E^k_{dp}$, c'est-à-dire tel que

$$(2) \qquad \lambda\, dP_{\tilde{z}}(\delta z) = \sum_{i=1}^{k} \lambda_i\, dp^i(\delta z) = 0 \qquad \forall \delta z \in E^n_{\delta z}.$$

Lors de l'obtention de la règle des multiplicateurs, le fait que l'image du point extrémal $P(\tilde{z})$ appartient à la frontière de l'ensemble $P(V_{\tilde{z}}) \subset E^k_p$ (n'en est pas un point intérieur) est déterminant. Pour cette raison, la notion de point extremum admet la généralisation naturelle suivante, avec laquelle la règle des multiplicateurs reste en vigueur.

En conservant le même sens que ci-dessus à $\mathcal{D}$, P, $dP_{\tilde{z}}$, supposons que M est un sous-ensemble arbitraire de $\mathcal{D}$. Par Φ_z, $z \in M$, nous désignons le filtre des voisinages du point z relativement à M, déterminé par la base $\{M \cap V_z, V_z$ voisinage arbitraire de $z\}$. Nous dirons que le point $\tilde{z} \in M$ est *point critique de l'application* $P : M \to E^k_p$ (restriction à M de l'application (1)), ou que le filtre $\Phi_{\tilde{z}}$ est *critique pour l'application* (1), ou enfin, nous dirons que l'application (1) est *critique pour le filtre* $\Phi_{\tilde{z}}$, s'il existe un élément $W \in \Phi_{\tilde{z}}$, $W \subset \mathcal{D}$, tel que le point $P(\tilde{z})$ appartient à la frontière de l'ensemble $P(W) \subset E^k_p$.

Si le filtre $\Phi_{\tilde{z}}$ est *convexe*, c'est-à-dire s'il existe une base du filtre constituée d'ensembles convexes, alors une condition nécessaire pour que l'application (1) soit critique pour le filtre $\Phi_{\tilde{z}}$ est *qu'il existe un ensemble convexe* $M \cap V_{\tilde{z}} = W \in \Phi_{\tilde{z}}$. *tel que l'ensemble* $dP_{\tilde{z}}(W - \tilde{z}) \subset E^k_{dp}$ *contienne le 0 de l'espace* E^k_{dp} *comme point frontière.*

De là découle directement la règle des multiplicateurs : si un vecteur

$$\lambda = (\lambda_1, \ldots, \lambda_k)$$

est orthogonal à un hyperplan d'appui, de dimension $k - 1$ *de l'ensemble convexe* $dP_{\tilde{z}}(W - \tilde{z})$ *en 0, et orienté de façon correspondante, alors on a*

$$(3) \qquad \lambda\, dP_z(\delta z) = \sum_{i=1}^{k} \lambda_i\, dp^i_z(\delta z) \leqslant 0 \qquad \forall \delta z \in k(w - \tilde{z}),$$

où $k(W - \tilde{z})$ *est le cône convexe de sommet 0, engendré par* $W - \tilde{z}$.

Ainsi, la règle des multiplicateurs est juste si M est un *ensemble localement convexe*, c'est-à-dire si, $\forall z \in M$, le filtre Φ_z est convexe. Comme il est facile de le voir, cela permet pour la détermination d'un point d'extremum conditionnel d'introduire des limitations non seulement sous forme d'égalités, mais également d'inégalités.

3. Cas général.

Une généralisation mineure et assez naturelle des définitions données au § 2 nous conduit à la formulation du problème d'extremum général incluant les problèmes de calcul des variations et d'optimum pour une fonction d'une variable indépendante.

E_z désignera maintenant un espace vectoriel réel arbitraire de points z (sans topologie), $\mathcal{O}$ *une partie finiment ouverte* de E_z, (l'intersection de $\mathcal{O}$ avec une quelconque variété linéaire de dimension finie L de E_z est ouverte pour la topologie de dimension finie de L). Soit l'application

$$P : \mathcal{O} \to E_p^k \tag{4}$$

dont la restriction est continue sur tout ensemble de la forme $\mathcal{O} \cap L$ (où L est de dimension finie !). La dimension finie de E_p^k ne permet de considérer que des problèmes d'extremum pour les fonctions d'une seule variable indépendante, c'est-à-dire des systèmes optimaux avec des paramètres concentrés ; les systèmes avec des paramètres distribués exigent l'étude d'applications dans des espaces vectoriels topologiques et ne seront pas envisagés ici, v. [8].

Nous dirons que *l'application* (4) *a une différentielle au point* $z \in \mathcal{O}$ s'il existe une application linéaire

$$dP_z : E_{\delta z} \to E_{dp}^k \tag{5}$$

telle que pour toute variété linéaire *finie* $L \subset E_z$ passant par z, la restriction de l'application (4) à $\mathcal{O} \cap L$ a une différentielle (au sens habituel) coïncidant avec la restriction de (5) à $L - z$.

Dans E_z nous introduirons diverses topologies *vectorielles métrisables* θ, qui en font des espaces vectoriels topologiques E_z^θ. Si Φ est un filtre arbitraire de E_z, nous dirons que *l'application* (4) *est déterminée sur le filtre* Φ (resp. *continue pour la topologie* θ) s'il existe une base du filtre dont tous les éléments soient contenus dans $\mathcal{O}$ (et telle que la restriction de l'application (4) à chacun d'eux soit continue pour θ). Enfin, nous appelerons enveloppe convexe $[\Phi]$ du filtre Φ le filtre déterminé par une base dont les éléments sont les enveloppes convexes d'une base arbitraire de Φ.

DEFINITIONS. – Le filtre Φ sur lequel est déterminée l'application (4) est dit *critique pour* (4) (ou (4) est dit *critique sur* Φ), si pour tout point z, appartenant à tous les éléments de Φ, il existe un ensemble $W \in \Phi, W \subset \mathcal{O}$ tel que $P(z)$ appartienne à la frontière de l'ensemble $P(w) \subset E_p^k$.

Si Φ_1 est inclus dans le filtre Φ_2 et si l'application (4) est critique sur Φ_1, elle est également critique sur Φ_2, c'est-à-dire que plus le filtre est fin, plus "faible" est la notion correspondante d'application critique.

Si nous supposons maintenant que l'application différentiable (4) est critique pour le filtre *convexe* Φ, alors, on peut appliquer la règle des multiplicateurs dont la formulation et la démonstration coïncident mot à mot avec ($\tilde{3}$) ; il est seulement nécessaire de supposer que $\tilde{z}$ appartient à tous les éléments de Φ. Ainsi ce cas est équivalent au cas fini et ne fait intervenir que la structure linéaire de l'espace de dimension infinie E_z sans avoir recours à une quelconque topologie θ dans E_z. Toutefois, quand on a réduit des problèmes de variations ou d'optimum au schéma indiqué ici, les filtres correspondants Φ ne sont pas toujours convexes, mais ils ont une structure beaucoup plus fine. Par conséquent, l'objectif de l'axiomatique proposée est de déterminer une classe de filtres Φ assez large pour comprendre tous les problèmes intéressants d'une part et, d'autre part, muni d'une structure assez riche pour que l'on puisse montrer une condition nécessaire d'extremum contenant les conditions nécessaires correspondantes de premier ordre, en particulier, le principe du maximum dans le contrôle optimal. Une telle classe est indiquée au § 4 et appelée par nous classe de *filtres quasi-convexes.*

On la définit à l'aide de l'introduction sur E_z d'une topologie métrisable arbitraire θ dont elle dépend (à la différence de la notion de filtre convexe, qui ne fait intervenir que la structure linéaire de l'espace E_z). En ayant la notion de filtre quasi-convexe, on peut formuler la condition nécessaire générale de point critique de la façon suivante :

Condition nécessaire de point critique. Soit l'application (4) déterminée sur l'enveloppe convexe $[\Phi]$ d'un filtre arbitraire Φ de E_z et critique sur Φ. S'il existe sur E_z une topologie vectorielle métrisable θ, pour laquelle Φ est quasi-convexe et pour laquelle l'application (4) est continue sur $[\Phi]$, alors, pour tout point $\widetilde{z}$, appartenant à tous les éléments du filtre Φ, en lequel l'application (4) a une différentielle (5), il existe un élément $W \in \Phi$, tel que le 0 de l'espace E^k_{dp} appartient à la frontière de l'ensemble $dP_{\widetilde{z}}\,([W] - \widetilde{z}) \subset E^k_{dp}$ (*).

Comme dans le cas fini, il découle de là la règle des multiplicateurs :

$$(6) \qquad \exists \lambda = (\lambda_1, \ldots, \lambda_k) \neq 0 \quad \lambda dP_{\widetilde{z}}(\delta z) \leqslant 0 \quad \forall \delta z \in k([w] - \widetilde{z})$$

Si $dP^*_{\widetilde{z}} : (E^k_{dp})^* \to (E_{\delta z})^*$ est l'application duale de (5), alors, à la règle des multiplicateurs (6) on peut donner la forme du "principe de maximum" :

$$dP^*_{\widetilde{z}}\,(\lambda)\, z \leqslant dP^*_z(\lambda)\widetilde{z} \qquad \forall z \in \widetilde{z} + K([W] - \widetilde{z})$$

Enfin, si la restriction de l'application (5) à un ensemble quelconque $M \supset [W]$ est continue dans la topologie θ, alors, au lieu du cône $K([W] - \widetilde{z})$ dans (6), on peut prendre le cône $K([\overline{W}]^\theta_M - z)$ où $[\overline{W}]^\theta_M$ est la fermeture de l'ensemble $[W]$ relativement à M dans la topologie θ.

4. Filtres quasi-convexes.

Definition. — Le filtre Φ dans E_z est dit *quasi-convexe dans la topologie métrisable* θ de l'espace E_z (θ-*quasi-convexe*), si pour tout $W \in \Phi$ et tout entier $s \geqslant 0$, il existe un élément $\widetilde{W} = \widetilde{W}(w, s) \in \Phi$ tel que quels que soient les $s + 1$ points $z_0, z_1, \ldots, z,$ de $\widetilde{W}$ et quel que soit $\epsilon > 0$, on peut trouver une application continue pour θ.

$$(7) \qquad \varphi : [z_0, z_1, \ldots, z_s) \to W,$$

satisfaisant à la condition $\varphi(z) \in z + B^\epsilon_z, \forall z \in [z_0, z_1, \ldots, z_s]$, où B^ϵ_z est là sphère de rayon ϵ et de centre z dans la métrique qui induit θ. Dans (7) sur l'enveloppe convexe $[z_0, z_1, \ldots, z_s]$ on considère la topologie de dimension finie habituelle, et sur W, la topologie induite par la topologie θ.

Il est évident que tout filtre convexe est quasi-convexe dans toute topologie. En outre si la topologie vectorielle θ_1 est plus faible que la topologie vectorielle θ_2, alors tout filtre quasi-convexe pour θ_2 est également quasi-convexe pour θ_1.

(*) $[W]$ est l'enveloppe convexe de l'ensemble W.

La réduction des problèmes d'optimum au schéma exposé est automatique à un point près : il reste à indiquer une topologie θ dans laquelle le filtre Φ est quasi-convexe et l'application (4) continue selon θ sur le filtre Φ. Plus θ est faible, plus grandes sont les chances que Φ soit un filtre quasi-convexe selon θ, mais plus faibles sont les chances que l'application (4) reste continue selon θ. La topologie nécessaire, si elle existe, se trouve quelque part "au milieu" et sa découverte effective constitue justement la solution du problème de variations. Les choix de topologies θ convenables dans les cas non-triviaux est suggéré par l'étude des régimes optimaux glissants, et la preuve de la quasi-convexité s'appuie sur un lemme qui donne une approximation du régime optimum glissant par le contrôle habituel, cf. [5], [6]. Enfin la preuve de la continuité selon θ de l'application (4) sur $[\Phi]$ repose sur les théorèmes généraux disant que les solutions des équations différentielles ordinaires dépendent de façon continue des données initiales et des seconds membres.

5. **Exemple.**

Soient G un ouvert de dimension n, $\mathcal{J}$ un intervalle de l'axe des temps, et $f(t\,,x)$ une fonction arbitraire à n dimensions sur $\mathcal{J} \times G$, continument différentiable en x, mesurable, de même que la matrice $f_x(t\,,x)$ en t, et telle que pour tout compact $K \subset G$ il existe une fonction $m(t)$ sommable sur $\mathcal{J}$ (dépendante de K et f) satisfaisant à la condition $|f(t\,,x)| + |f_x\,(t\,,x)| \leqslant m(t)\; \forall t \in \mathcal{J}\,,\ \forall x \in K$. Appelons E_f l'espace linéaire des classes de fonctions équivalentes f et considérons l'équation différentielle

$$\frac{dx}{dt} = f\,(t\,,x), \qquad f\,(t\,,x) \in E_f. \tag{8}$$

Soit $\mathcal{D}$ un ensemble de points $z = (t_1\,,t_2\,,x_1\,,f)\,,t_1\,,t_2 \in \mathcal{J},x_1 \in G\,,f \in E_f$, dans "l'espace des équations" $E_z = E^2_{\binom{t_1}{t_2}} \times E^n_{x_1} \times E_f$, tel que $\forall z \in \mathcal{D}$ l'équation correspondante (8) admet la solution $x(t)$ sur un segment d'extrémités $t_1\,,t_2$ avec la condition initiale $x(t_1) = x_1$. Ceci détermine l'application S de $\mathcal{D}$ dans l'"espace des valeurs aux bornes" $E^{2+2n}_{(t_1,t_2,x_1,x_2)^T}$ d'après la formule

$$(t_1\,,t_2\,,x_1\,,f) \mapsto (t_1\,,t_2\,,x_1\,,x_2)^T \quad \text{où} \quad x_2 = x(t_2)\,.$$

L'ensemble $\mathcal{D}$ est finiment ouvert et la restriction de S à $\mathcal{D} \cap L$ (L-variété linéaire de dimension finie) est continue sur $\mathcal{D} \cap L$ (la solution de l'équation (8) dépend continument des paramètres). Par conséquent, si l'application

$$Q : E^{2+2n}_{(t_1,t_2,x_1,x_2)^T} \to E^k_p$$

est différentiable, alors l'application

$$P = Q \circ S : \mathcal{D} \to E^k_p \tag{9}$$

a en chaque point $\widetilde{z} = (\widetilde{t}_1\,,\widetilde{t}_2\,,\widetilde{x}_1\,,\widetilde{f})$ pour lequel $\widetilde{f}(t\,,x)$ est continue aux points $(\widetilde{t}_1\,,\widetilde{x}_1)\,,(\widetilde{t}_2\,,\widetilde{x}_2)$ la différentielle

$$(10)\quad \begin{aligned} dP_{\tilde z}(\delta z) &= \frac{\partial Q}{\partial t_1}\,\delta t_1 + \frac{\partial Q}{\partial t_2}\,\delta t_2 + \frac{\partial Q}{\partial x_1}\,\delta x_1 + \frac{\partial Q}{\partial x_2}\,\delta x_2\,, \\ \delta x_2 &= \tilde f(\tilde t_2, \tilde x_2)\,\delta t_2 + \Gamma(\tilde t_2)\Big[\delta x_1 - \tilde f(\tilde t_1, \tilde x_1)\,\delta t_1 \\ &\qquad + \int_{t_1}^{t_2} \Gamma^{-1}(t)\,\delta f(t, \tilde x(t))\,dt\Big] \end{aligned}$$

où $\tilde x(t)$, $\tilde t_1 \leqslant t \leqslant \tilde t_2$, $\tilde x(\tilde t_1) = \tilde x_1$, est solution de l'équation $\dfrac{dx}{dt} = \tilde f$ et où la matrice $\Gamma(t)$ satisfait à l'équation

$$\frac{d\Gamma}{dt} = \tilde f_x\,(t, \tilde x(t))\,\Gamma \quad , \quad \Gamma(\tilde t_1) = I\,\cdot$$

Soit θ_f unẹ topologie vectorielle métrisable dans E_f et soit, dans E_f, $\Phi_{\tilde f}$ un filtre quasi-convexe selon θ_f, dont tous les éléments contiennent $\tilde f$. Alors le filtre $\Phi_{\tilde z}$ donné par les éléments de base $W_{\tilde z} = V_{\left(\begin{smallmatrix}\tilde t_1\\ \tilde t_2\end{smallmatrix}\right)} \times V_{\tilde x_1} \times W_{\tilde f}$ où $V_{\left(\begin{smallmatrix}\tilde t_1\\ \tilde t_2\end{smallmatrix}\right)}$, $V_{\tilde x_1}$ sont voisinages des points $\begin{pmatrix}\tilde t_1\\ \tilde t_2\end{pmatrix}$, $\tilde x_1$ et $W_{\tilde f} \in \Phi_{\tilde f}$, est quasi-convexe dans l'espace vectoriel topologique métrisable

$$E_z^\theta = E^2_{\left(\begin{smallmatrix}t_1\\ t_2\end{smallmatrix}\right)} \times E^n_{x_1} \times E_f^{\theta f} \quad \text{et} \quad \forall\, W \in \Phi_{\tilde z}\ \ \tilde z = (\tilde t_1, \tilde t_2, \tilde x_1, \tilde f) \in W.$$

Admettons que l'application (9) soit critique et continue selon θ sur le filtre $\Phi_{\tilde z}$. D'après la règle des multiplicateurs (6) et la formule (10), il découle que

$$\exists \lambda = (\lambda_1 \ldots \lambda_k) \neq 0 \quad \exists W_{\tilde f} \in \Phi_{\tilde f} : \forall \begin{pmatrix}\delta t_1\\ \delta t_2\end{pmatrix} \in E^2_{\left(\begin{smallmatrix}\delta t_1\\ \delta t_2\end{smallmatrix}\right)}$$

$$\forall \delta x_1 \in E^n_{\delta x_1} \quad \forall \delta f \in K([W_{\tilde f}] - \tilde f)$$

$$\lambda\left(\frac{\partial Q}{\partial t_1} - \frac{\partial Q}{\partial x_2}\,\Gamma(\tilde t_2)\,\tilde f_1\right)\delta t_1 + \lambda\left(\frac{\partial Q}{\partial t_2} + \frac{\partial Q}{\partial x_2}\,\tilde f_2\right)\,\delta t_2 +$$

$$\lambda\left(\frac{\partial Q}{\partial x_1} + \frac{\partial Q}{\partial x_2}\,\Gamma(\tilde t_2)\right)\delta x_1 + \lambda\,\frac{\partial Q}{\partial x_2}\,\Gamma(\tilde t_2)\int_{\tilde t_1}^{\tilde t_2} \Gamma^{-1}(t)\,\delta f(t, \tilde x(t))\,dt \leqslant 0$$

ou

$$(11)\left\{\begin{aligned} &\lambda\left(\frac{\partial Q}{\partial t_1} - \frac{\partial Q}{\partial x_2}\,\tilde\Gamma_2\tilde f_1\right) = \lambda\left(\frac{\partial Q}{\partial t_2} + \frac{\partial Q}{\partial x_2}\,\tilde f_2\right) = 0\ ,\ \lambda\left(\frac{\partial Q}{\partial x_1} + \frac{\partial Q}{\partial x_2}\,\tilde\Gamma_2\right) = 0 \\ &\lambda\ \frac{\partial Q}{\partial x_2}\,\tilde\Gamma_2 \int_{\tilde t_1}^{\tilde t_2} \Gamma^{-1}(t)\,\delta f(t, \tilde x(t))\,dt \leqslant 0\,. \end{aligned}\right.$$

A l'aide de la fonction $\tilde\psi(t) = \lambda\dfrac{\partial Q}{\partial x_2}\,\Gamma(\tilde t_2)\,\Gamma^{-1}(t)$ satisfaisant à l'équation

$$(12)\qquad \frac{d\tilde\psi}{dt} = -\,\tilde\psi\,\tilde f_x\,(t, \tilde x(t))$$

l'équation (11) peut être écrite sous la forme

$$
\begin{aligned}
&\lambda \frac{\partial Q}{\partial t_1} = \tilde{\psi}(t_1)\tilde{f}(\tilde{t}_1, \tilde{x}_1) \\
(13) \qquad &\lambda \frac{\partial Q}{\partial t_2} = -\tilde{\psi}(\tilde{t}_2)\tilde{f}(\tilde{t}_2, \tilde{x}_2) \\
&\lambda \frac{\partial Q}{\partial x_1} = -\tilde{\psi}(\tilde{t}_1),
\end{aligned}
$$

et l'inégalité (11) sous la forme

$$
(14) \quad \int_{\tilde{t}_1}^{\tilde{t}_2} \tilde{\psi}(t)\,\tilde{f}(t, \tilde{x}(t))\,dt \geqslant \int_{\tilde{t}_1}^{\tilde{t}_2} \tilde{\psi}(t)\,f(t, \tilde{x}(t))\,dt \quad \forall f \in \tilde{f} + K([W_{\tilde{f}}] - \tilde{f}).
$$

Introduisons la fonction de Hamilton du problème

$$
H(\psi, x, t) = \psi f(t, x) = \sum_{i=1}^{n} \psi_i f^i(t, x) ;
$$

en particulier $\tilde{H}(\psi, x, t) = \psi\tilde{f}(t, x)$. D'après (8) et (12), $\tilde{x}(t)$ et $\tilde{\psi}(t)$ satisfont au système de Hamilton sur le segment $\tilde{t}_1 \leqslant t \leqslant \tilde{t}_2$

$$
\begin{aligned}
\frac{d\tilde{x}}{dt} &= \frac{\partial}{\partial \psi}\tilde{H}(\tilde{\psi}(t), \tilde{x}(t), t) \\
\frac{d\tilde{\psi}}{dt} &= -\frac{\partial}{\partial x}\tilde{H}(\tilde{\psi}(t), \tilde{x}(t), t)
\end{aligned}
$$

L'inégalité (14) peut être écrite sous la forme "du principe du maximum sous sa forme intégrale"

$$
\int_{\tilde{t}_1}^{\tilde{t}_2} \tilde{H}(\tilde{\psi}(t), \tilde{x}(t), t)\,dt \geqslant \int_{\tilde{t}_1}^{\tilde{t}_2} H(\tilde{\psi}(t), \tilde{x}(t), t)\,dt
$$

$$
\forall H(\psi, x, t) = \psi f(t, x), \; f \in \tilde{f} + K([W_{\tilde{f}}] - \tilde{f})
$$

et l'équation (13) sous la forme de la condition suivante de transversalité : le $2 + 2n$–uplet

$$
(\tilde{H}(\tilde{\psi}(\tilde{t}_1), x(\tilde{t}_1), \tilde{t}_1) - \tilde{H}(\tilde{\psi}(\tilde{t}_2), x(\tilde{t}_2, \tilde{t}_2) - \tilde{\psi}(\tilde{t}_1) + \tilde{\psi}(\tilde{t}_2))
$$

est orthogonal à l'hyperplan de dimension $(2 + 2n - k)$, $Q(t_1, t_2, x_1, x_2) = \text{cte}$ au point $(t_1, t_2, x_1, x_2)^T$.

L'ensemble des conditions énumérées est trivial si $\psi(t) \equiv 0$. Afin d'exclure ce cas, nous devons exiger que la matrice, d'ordre $k \times (2 + 2n)$, $\left(\frac{\partial Q}{\partial t_1}, \frac{\partial Q}{\partial t_2}, \frac{\partial Q}{\partial x_1}, \frac{\partial Q}{\partial x_2}\right)$ soit de rang k au point $(\tilde{t}_1, \tilde{t}_2, \tilde{x}_1, \tilde{x}_2)$.

ЛИТЕРАТУРА

[1] Болтянский В. Г., Гамкрелидзе Р. В., Понтрягин Л. С. — К теории оптимальных процессов, ДАН СССР, 110, 1956, стр. 1-10.

[2] Понтрягин Л. С., Болтянский В. Г., Гамкрелидзе Р. В., Мищенко Е. Ф. — Математическая теория оптимальных процессов, М., Изд-во физ.-мат. лит-ры, 1961.

[3] McShane E. J. — On Multipliers for Lagrange Problems, *Am. J. Math.*, 61, 1939, c p. 809-819.

[4] Neustadt L. W. — An abstract variational theory with applications to a broad class of optimization problems I, General Teory, *J. SIAM Contr.*, 4, 1966, c p. 505-527; II. Applications, *Ibid.*, 5, 1967, c p. 90-137.

[5] Гамкрелидзе Р. В. — О скользящих оптимальных режимах, ДАН СССР, 134, № 6, 1962, стр. 106-128.

[6] Gamkrelidze R. V. — On some extremal problems in the theory of differential equations, *J. SIAM Contr.*, 1965, c p. 106-128.

[7] Гамкрелидзе Р. В., Харатишвили Г. Л. — Экстремальные задачи в линейных топологических пространствах. Известия АН СССР, Серия математическая, 33, 1969, стр. 781-839.

[8] Lions J. L. — *Contrôle optimal de Systèmes gouvernés par des équations aux dérivées partielles*, Dunod - Gauthier-Villars, Paris, 1968.

Institut Steklov
Moscou
U.R.S.S.

Actes, Congrès intern. Math., 1970. Tome 3, p. 177 à 181.

EXTREMAL STRATEGIES IN A DIFFERENTIAL GAME

by N.N. KRASOVSKII

Consider the following differential game. The phase vector x of the system is described by the equation

$$(1) \qquad \dot{x} = f(t, x, u, v)$$

where u and v are the control vectors governed correspondingly by the first and second players, and subjected to the constraints

$$(2) \qquad u \in \mathfrak{U} \quad , \quad v \in \mathfrak{V}$$

where $\mathfrak{U}$ and $\mathfrak{V}$ are compact sets ; the function f is continuous and satisfies the Lipschitz condition in x. Also given is a set $\mathfrak{M}$ in the space $\{t, x\}$. The aim of the first player is the rendez-vous with set $\mathfrak{M}$, which hence determines the moment θ for the conclusion of the game. Given is a functional

$$\gamma = \varphi(x[t],\ u[t],\ v[t]\ ;\ t_0 \leqslant t \leqslant \theta)$$

which is to be minimized by the first player and maximized by the second one. The initial position $\{t_0, x_0\}$ is fixed. The examples are : the game of pursuit $y[t] \to z[t]$ with pay-off

$$(3) \qquad \gamma = \theta - t_0$$

and with set $\qquad [x = \{y, z\} : (z - y) \in \mathfrak{M}\,]$;

the game of approaching to $\mathfrak{M}$ with pay-off

$$(4) \qquad \gamma = \int_{t_0}^{\theta} \psi(t, x[t],\ u[t],\ v[t])\, dt + \omega(\theta, x[\theta])$$

the game of approaching to $\mathfrak{M}$ with pay-off

$$(5) \qquad \gamma = \max_{t_0 \leqslant t \leqslant \theta} \omega(t, x[t])$$

Games of such kind were the object of research in a number of works (see e.g. [1-13]). The following report is essentially devoted to extremal strategies which are constructed here with the aid of the so-called target absorption sets. In the positional game with complete information considered in the sequel the strategies U and V are identified with systems of sets $\{\mu(du)\}_{\{t,x\}}$ and $\{\nu(dv)\}_{\{t,x\}}$ which are formed of regular measures μ and ν normalized on $\mathfrak{U}$ and $\mathfrak{V}$. The motion $x[t] = x[t, t_0, x_0\ ;\ U,\ V]$ is defined as the limit of Euler's broken curves $x_\Delta\,[t]$

$$\dot{x}_\Delta [t] = \int\int f(t, x_\Delta [t], u, v)\, \mu(du)_{\{\tau_i, x_\Delta[\tau_i]\}}\, \nu(d\nu)_{\{\tau_i, x_\Delta[\tau_i]\}}$$

$$(\tau_i \leqslant t < \tau_{i+1} ; \tau_0 = t_0, \max (\tau_{i+1} - \tau_i) = \Delta \to 0)$$

The approach-evasion game : given are the closed sets $\mathfrak{M}_c$ and $\mathfrak{N}_c$ in space $\{t , x\}$. The aim of the first player is to ensure an approach to $\mathfrak{M}_c$, i.e. to determine a strategy U_c, which guarantees a rendez-vous $\{\theta , x [\theta]\} \in \mathfrak{M}_c$ under the condition $\{t , x [t]\} \in \mathfrak{N}_c$ $(t_0 \leqslant t \leqslant \theta)$ for any motion $x [t] = x [t , t_0 , x_0 ; U_c , V]$; the aim of the second player is to ensure an evasion, i.e. to determine a strategy V_c which guarantees the departure of point $\{t , x [t]\}$ from $\mathfrak{N}_c$ prior to its rendez-vous with $\mathfrak{M}_c$, for any motion $x [t] = x [t , t_0 , x_0 , U , V_c]$.

We presuppose that the duration of the game has an upper bound given by a sufficiently large number T.

The following alternative is valid : for an arbitrary position $\{t_0 , x_0\}$ it is either that the game of approach is soluble or that the game of evasion is soluble. As a consequence one arrives on a number of assertions on saddle-point type situations for typical differential games. For example in game (3) assuming

$$\mathfrak{M}_c[\{t , x\} \vdots\ t_0 \leqslant t \leqslant c, z - y \in \mathfrak{R}] ; \mathfrak{N}_c [\{t , x\} : t_0 \leqslant t]$$

we determine the value $c_0 = \theta^0$ such that strategy U_c ensures a rendez-vous $(z [\theta] - y [\theta]) \in \mathfrak{R}$ with $\theta \leqslant \theta^0$ (or $c_0 = \infty$) ; however if $\theta^* < \theta^0$ then there exists a strategy V^* which ensures an evasion for $t_0 \leqslant t \leqslant \theta^*$. In game (5) assuming $\mathfrak{M}_c = \mathfrak{M}$ and $\mathfrak{N}_c [\{t , x\} : t_0 \leqslant t , \omega(t , x) \leqslant c]$ we determine the value c^0 such that the strategy U_c guarantees a rendez-vous $\{\theta , x [\theta]\} \in \mathfrak{M}$ for $\gamma \leqslant \gamma^0$ (under the condition that there exists a number $c < \infty$ for which an auxiliary evasion problem is unsoluble) ; if however $\gamma^* < \gamma^0$ then there exists a strategy V^* which guarantees the validity of the inequality $\gamma > \gamma^*$ (or eliminates the possibility of a rendez-vous). In game (4) choosing appropriate $\mathfrak{M}_c$ and $\mathfrak{N}_c$ we determine the values c_0 and c^0 so that : (1) for $c > c^0$ there exists a strategy U_c, which ensures a rendez-vous $\{\theta , x [\theta]\} \in \mathfrak{M}$ for $\gamma < c$ whereas for $t < \theta$ it prevents the exit of position $\{t , x [t]\}$ into the neighbourhood of $\mathfrak{M}$ for $\gamma > c$; (2) if $c < c_0$, then there exists a strategy V_c which eliminates the possibility of a rendez-vous for $\gamma < c$; if however $c > c_0$, then there exists a strategy U^c which directs the position $t , x [t]$ to a given arbitrary small neighbourhood of $\mathfrak{M}$ for $\gamma < c$ before the conclusion of the game. Hence the question of whether c_0 and c^0 are equal does arise.

The structure of the approach-evasion game as well as the nature of the strategies U_c and V_c are determined by the absorption sets. Suppose U^T is a strategy determined by the set $\{\mu (du)\}$ of all regular measures μ for any position $\{t , x\}$.

Denote with $V^{(j)}$ the strategy which is given for every $j = 1, 2, 3, 4$ by systems of sets $\{\nu(d\nu)\}_{\{t , x\}}$ of the following type : $(j = 1)$ $\{\nu\}_{\{t , x\}}$ does not depend on x , $\{\nu\}_{\{t , x\}} = \{\nu\}_t$; $(j = 2)$ every $\{\nu\}_{\{t , x\}}$ consists of a sole element $\nu_{\{t , x\}}$, $\nu_{\{t , x\}}$ are weakly continuous in t and x ; $(j = 3)$ $\{\nu\}_{\{t , x\}}$ are convex and weakly upper semicontinuous with respect to the inclusion property in x and in t from the right ; $(j = 4)$ $\{\nu\}_{\{t, x\}}$ are arbitrary sets of regular measures ν. With $Q^{(t^*)}$ we denote the intersection of set Q in space $\{t , x\}$ with the hyperplane $t = t^*$. The absorption set $\mathcal{W}_j^I(t_* , \theta \; ; \mathcal{M}_c^{(\theta)})$ $(\mathcal{W}_j^{II}(t_* , \theta \; ; \mathcal{M}_c))$ is the set of all points w such that for any selected strategy $V^{(j)}$ there exists a motion

$$x[t] = x[t , t_* , w ; U^T , V^{(j)}], \; (t_* \leqslant t \leqslant \theta),$$

which satisfies the conditions (for I) $x[\theta] \in \mathcal{M}_c^{(\theta)}$, $x[t] \in \mathcal{N}_c^{(t)}$ $(t_* \leqslant t \leqslant \theta)$; (for II) $x[\tau] \in \mathcal{M}_c^{(\tau)}$, $x[t] \in \mathcal{N}_c^{(t)}$ $(t_* \leqslant t \leqslant \tau \leqslant \theta)$.

The sets $\mathcal{W}_4^I$ are strongly stable i.e. for any

$$t_* < \theta, x_* \in \mathcal{W}_4^I\,(t_* , \theta \; ; \mathcal{M}_c^{(\theta)}), \;\; t^* \in (t_* , \theta]$$

and $V^{(1)}$ there exists a motion $x[t] = x[t , t_* , x_* \; ; U^T , V^{(1)}]$ such that $x[t^*] \in \mathcal{W}_4^I\,(t^* , \theta \; ; \mathcal{M}_c)$ and $x[t] \in \mathcal{N}_c^{(t)}$ $(t_* \leqslant t \leqslant t^*)$. The sets $\mathcal{W}_4^{II}$ are stable, that is for any $t_* < \theta , x_* \in \mathcal{W}_4^I\,(t_* , \theta , \mathcal{M}_c) , t^* \in (t_* , \theta]$ and $V^{(1)}$ there exists a motion $x[t] = x[t , t_* , x_* \; ; U^T , V^{(1)}]$ such that either $x[t^*] \in \mathcal{W}_4^{II}\,(t^* , \theta , \mathcal{M}_c)$ and $x[t] \in \mathcal{N}_c^{(t)}$ $(t_* \leqslant t \leqslant t^*)$ or $x[\tau] \in \mathcal{M}_c^{(\tau)}$ and $x[t] \in \mathcal{N}_c^{(t)}$ $(t_* \leqslant t \leqslant \tau \leqslant t^*)$ The strategy $U^{(e)}$ being extremal to $\mathcal{W}(t)$ is given by sets $\{\mu\}_{\{t , x\}}^{(e)}$ as determined by the relation

$$\min_{\nu} \int\!\!\int (x^0 - x)' f(t , x , u , \nu) \mu^{(e)}(du)\, \nu(d\nu)$$
$$= \max_{\mu} \min_{\nu} \int\!\!\int (x^0 - x)' f(t , x , u , \nu)\, \mu(du)\, \nu(d\nu)$$

where x^0 is the point in $\mathcal{W}(t)$ nearest to x, and where the prime denotes the transpose. If the sets $\mathcal{W}(t)$ are stable and $x_0 \in \mathcal{W}(t_0)$, then the strategy $U^{(e)}$ solves the problem of approach for an instant not later then θ. The sets $\mathcal{W}_4^i$ are constructed recurrently : $\mathcal{W}(\theta , \theta) = \mathcal{M}^{(\theta)}$ and for

$$i = I \quad \mathcal{W}(\theta - n\Delta , \theta) = \mathcal{W}_1^I(\theta - n\Delta , \theta - (n - 1)\Delta \; ; \mathcal{W}(\theta - (n - 1)\Delta \, , \, \theta)),$$

whereas for

$$i = II \quad \mathcal{W}(\theta - n\Delta , \theta) = \mathcal{W}_1^{II}\,(\theta - n\Delta , \theta - (n - 1)\Delta \; ;$$
$$[\theta - n\Delta , \theta - (n - 1)\Delta] \times \mathcal{W}(\theta - (n - 1)\Delta , \theta) \cup \mathcal{M}_c) \; ;$$

the next step is the limit transition for $\Delta = 1/2^k$, $k \to \infty$.

The situations when $\mathcal{W}_j^i$ coincide with $\mathcal{W}_{j^*}^{i^*}$ for $i < i^*$ and $j < j^*$ are of special interest for the sake of constructing $U^{(e)}$ effectively as well as for the classification of games. The sufficient conditions for the coincidence $\mathcal{W}_1^I = \mathcal{W}_j^I$ for $j = 1$, $j = 2$ or $j = 3$ may be deduced from appropriate fixed-point theorems [14]. Thus for the system

$$\dot{x} = A(t)x + B(t)u - C(t)\nu \tag{7}$$

with $\mathcal{M}_c$ convex it is sufficient for $\mathcal{W}_1^I = \mathcal{W}_4^I$ that with $x_* \notin \mathcal{W}_1^I\,(t_* , \theta)$ the auxiliary problem

$$\epsilon^0\,(t_* , x_* , \theta) = \max_{V^{(1)}} \min_{x[t]} \rho\,(\mathcal{M}_c^{(\theta)}, x[\theta , t_* , x_* , U^T , V^{(1)}]) \tag{8}$$

would have a solution $\{V_0^{(1)}\}$ in the form of a convex set. Here $\rho(\mathcal{M}_c^{(\theta)}, x)$ is the distance from x to $\mathcal{M}_c^{(\theta)}$. The set $\mathcal{W}_1^I(t, \theta)$ is described by the inequality.

$$(9) \qquad \max_{\|l\|=1} [\rho^{(2)}(t, \theta, l) - \rho^{(1)}(t, \theta, l) - \rho^{(\mathcal{M}}(l, \theta) - l'X(\theta, t)x] \leqslant 0$$

where

$$\rho^{(\mathcal{M})} = \max_{-p \in \mathcal{M}} [l'p]$$

$$\rho^{(1)} = \max_{u \in \mathcal{U}} \left[\int_t^\theta l'X(\theta, \tau)B(\tau)u(\tau)d\tau\right]$$

$$\rho^{(2)} = \max_{v \in \mathcal{V}} \left[\int_t^\theta l'X(\theta, \tau)C(\tau)v(\tau)d\tau\right]$$

$X(t, t_0)$ is the fundamental matrix for (7), and $\|l\|$ is the euclidean norm of l. For $x \notin \mathcal{W}_1^I(t, \theta)$ the left part (9) is $\epsilon^0(8)$. The condition for the convexity of $\{V_0^{(1)}\}$ is reduced to the unicity of the vector $l = l^\circ$ which maximizes (9) for the values of x and t which lead to an $\epsilon^\circ > 0$. In the stated regular case the strategy $U^{(e)}$ acquires the same sense as in the principle of extremal aiming [12] in game (3) where $y = x^{(1)}$ $(x^{(1)} = Ax^{(1)} + Bu)$, $z = x^{(2)}$ $(x^{(2)} = Ax^{(2)} + Cv)$ and $\mathcal{P} = -\mathcal{M}_c^{(\theta)}$. For (7) with $\mathcal{M}_c$ convex the sets $\mathcal{W}_1^I(t, \theta)$ coincide with $\mathcal{W}_j^I(t, \theta)$ $(j = 2, 3)$, i.e. the transition from programmed controls $V^{(1)}(v(t))$ to the continuous $V^{(2)}$ or to the regularly-discontinuous $V^{(3)}$ feedback controls $(v[t, x[t]])$, does not enlarge the possibility of an evasion from $\mathcal{M}_c$ at a fixed moment θ. The results are further propagated to the system $\dot{x} = A(t)x + f(t, u, v)$.

The conditions for the coincidence $\mathcal{W}_j^i = \mathcal{W}_{j*}^{i*}$ $(j \leqslant j^*)$ for nonlinear systems or for $i^* = II$ are less effective.

If for any x in the neighbourhood of stable sets $\mathcal{W}(t)$ the point x^0 from (6) is unique, then the extremal (with respect to $\mathcal{W}(t)$) strategy $U^{(e)} = U^{(3)}$ and the motion $x[t]$ are identified with the solutions of the corresponding contingent equation

$$(10) \qquad \dot{x}[t] \in \mathcal{F}_{U,V}(t, x[t])$$

where $\mathcal{F}_{U,V}$ are convex. It is then possible to approximate $U^{(e)}$ with continuous strategies. Example : the regular case (7) and (9). In the case when x^0 is nonunique ($\mathcal{W}(t)$ has a nonsmoothe concave boundary), $U^{(e)}$ may happen to be essentially discontinuous ($U^{(e)} = U^{(4)}$) and will not admit a continuous approximation. Then the motions $x[t]$ cannot be identified with the set of all the solutions of (10). The rendez-vous with $\mathcal{M}_c$ will then be achieved only by the constructive solutions of these equations, i.e. only by the limite curves for the broken lines $x_\Delta[t]$.

The stable sets $\mathcal{W}(t)$ were defined for the purpose of constructing rendez-vous strategies $U_c = U^{(e)}$. For constructing evasion strategies $V_c = V^{(e)}$ one should consider the v-stable sets $\mathcal{W}^*(t)$ which are determined here by an obvious inversion. It is rather frequent that sets $\mathcal{W}^*(t)$ are irregular even in simple problems. Example : problems (3) on the evasion of point $z[t]$ $(\dot{z}_1 = v_1, \dot{z}_2 = v_2 ; \|v\| \leqslant \nu)$ from point $y[t]$ $(\dot{y}_1 = y_3, \dot{y}_2 = y_4, \dot{y}_3 = u_1, \dot{y}_4 = u_2, \|u\| \leqslant \mu)$ with rendez-vous condition $(z - y) \in \mathfrak{P} [\{y , z\} : y_1 = z_1 , y_2 = z_2]$. Here it is always simple to construct an evasion strategy V_c. The strategy is not approachable by those of type $V^{(2)}$ or $V^{(3)}$ although it may be constructed as a strategy extremal to certain sets $\mathfrak{W}^* (t)$ $(t_0 \leqslant t \leqslant \theta)$. Hence it follows that there are sharpened hollows on the boundary for $\mathfrak{W}^*(t)$. Contrary to the previous discussion it may happen however that $V_c = V^{(3)}$ and that the evasion strategy V_c is approachable by continuous strategies $V^{(2)}$. Example : system (7) where

$$co\, [X (\sigma , t) B (t)\, \mathfrak{U}] = co\, [X(\sigma , t)\, C(t)\, \mathfrak{V}] + Q$$

with Q convex, $x_0 \notin \mathfrak{W}_1^I (t_0 , \theta , \mathfrak{M}^{(\theta)})$, $\epsilon^\circ (t_0 , x_0 , \sigma) > 0$ $(t_0 \leqslant t \leqslant \sigma \leqslant \theta)$. Here the evasion strategy is constructed effectively from condition min max $(d\lambda/dt)$ where $\lambda (t , x , \theta) = \int_t^\theta [\epsilon^0 (t , x , \sigma)]^{-1}\, d\sigma$.

REFERENCES

[1] Fleming W.H. — *J. Math. Anal. Appl.*, v. 3, 1961, p. 102.

[2] Nardzewski C.R. — Adv. in Game Theory. *Ann. of Math. Studies,* 1964, p. 113.

[3] Isaacs R. — *Differential Games,* John Wiley, New York, 1965.

[4] Pontriagin L.S. — *Dokl. Akad. Nauk S.S.S.R.*, v. 175, No. 4, 1967, p. 764.

[5] Pschenichnyi B.N. — *Dokl. Akad. Nauk S.S.S.R.*, v. 184, No. 2, 1969, p. 285.

[6] Roxin E. — *J. Optimization Theory Appl.*, v. 3, No. 3, 1969, p. 153.

[7] Varayia P., Lin J. — *S.I.A.M. J. Control,* v. 7, No. 1, 1969, p. 141.

[8] Petrov N.N. — *Dokl. Akad. Nauk S.S.S.R.*, v. 190, No. 6, 1970, p. 1289.

[9] Friedman A. — *J. Diff. Equations,* v. 7, No. 1, 1970, p. 92.

[10] Pontriagin L.S., Misčenko E.F. — *Dokl. Aakd. Nauk S.S.S.R.*, v. 189, No. 4, 1969, p. 721.

[11] Smolyakov E.R. — *Dokl. Akad. Nauk S.S.S.R.*, v. 191, No. 1, 1970, p. 39.

[12] Krasovskii N.N. — *Prikl. Mat. Meh.*, v. 32, No. 5, 1968, p. 793.

[13] Krasovskii N.N., Subbotin A.J. — *Dokl. Akad. Nauk S.S.S.R.*, v. 190, No. 3, 1970, p. 593.

[14] Krasovskii N.N. — *Prikl. Mat. Meh.*, v. 34, No. 2, 1970, p. 197.

Ural State University
Lenin street 51,
Sverdlovsk, K-83 (URSS)

Actes, Congrès intern. Math., 1970. Tome 3, p. 183 à 185.

OPTIMAL CONTROL : A THEORY OF NECESSARY CONDITIONS

by Lucien W. NEUSTADT

A broad class of optimal control problems are included as special cases of the following problem.

PROBLEM 1. – Let there be given a real Banach space $\mathfrak{X}$ and real linear topological vector space $\mathfrak{P}$ and $\mathfrak{Z}$, together with a subset Π of $\mathfrak{P}$, an open set A in $\mathfrak{X}$, and a convex body (i.e., convex set with non-empty interior) $Z \subset \mathfrak{Z}$. Further, let there be given a set $\mathcal{U}$ of continuously differentiable operators from A into $\mathfrak{X}$ each of whose elements has at most one fixed point in A, and continuous functions $\varphi : A \times \Pi \to R^m$, $\phi^0 : A \times \Pi \to R$, and $\phi_1 : A \times \Pi \to \mathfrak{Z}$. Then we wish to find a triple $(x_0 , T_0 , \pi_0) \in A \times \mathcal{U} \times \Pi$ such that $x_0 = T_0 x_0$, $\varphi(x_0 , \pi_0) = 0$, $\phi_1 (x_0 , \pi_0) \in Z$, and such that $\phi^0 (x_0 , \pi_0) \leqslant \phi^0 (x , \pi)$ for all $(x , T , \pi) \in A \times \mathcal{U} \times \Pi$ that satisfy the relations

$$x = Tx, \varphi(x , \pi) = 0, \quad \text{and} \quad \phi_1 (x , \pi) \in Z.$$

Our aim is to find necessary conditions which solutions of this problem must satisfy in the form of a generalized Lagrange multiplier rule. Obviously, this will be possible only if we make suitable hypotheses on the problem data. Of course, these hypotheses must be sufficiently weak so that our necessary conditions will be applicable to a sufficiently broad class of problems.

We now turn to these hypotheses.

We first imbed $\mathcal{U}$ into the linear vector space of all continuously differentiable operators from A into $\mathfrak{X}$. This space, which we shall denote by $\mathcal{C}$, may be considered to be a linear topological space if we endow it with the topology of pointwise convergence, defined by taking as a local base all sets of the form

$$\{T : T \in \mathcal{C}, ||Tx_j|| < \epsilon \text{ for } j = 1, \ldots , \nu\},$$

where $\{x_1 , \ldots , x_\nu\}$ is some finite subset of A and $\epsilon > 0$. Let us denote by $\mathcal{C}_1$ the set of all $T \in \mathcal{C}$ which have exactly one fixed point in A, and let φ_0 denote the map from $\mathcal{C}_1$ into A which assigns to each T its fixed point.

Let us make two definitions. A non-empty set B in a real linear vector space $\mathfrak{Y}$ will be said to be *finitely open in itself* if, for every $y_0 \in B$ and every finite

This research was supported in part by the United States Air Force Office of Scientific Research under grant AF-AFOSR-1 029-67, and in part by the National Science Foundation under grant GK-15 787.

subset $\{y_1, \ldots ; y_m\}$ of B, there is a number $\epsilon > 0$ such that $y_0 + \Sigma_1^m \lambda^i (y_i - y_0) \in B$ whenever $0 \leqslant \lambda^i < \epsilon$ for $i = 1, \ldots, m$. If $\mathcal{Y}$ and $\mathfrak{Z}$ are real linear topological spaces and $B \subset \mathcal{Y}$, then a function $\varphi : B \to \mathfrak{Z}$ will be said to be *dually differentiable* at a point $y_0 \in B$ if there is a continuous linear function $D\varphi : \mathcal{Y} \to \mathfrak{Z}$ such that

$$\lim_{\substack{\epsilon \to 0^+ \\ y \to 0}} \frac{\varphi(y_0 + \epsilon(\tilde{y} + y)) - \varphi(y_0)}{\epsilon} = D\varphi(\tilde{y}) \text{ for all } \tilde{y} \in \mathcal{Y}.$$

In this case, we shall refer to $D\varphi$ as a *dual differential* of φ at y_0.

We point out that if a non-empty set B in a real linear vector space is either convex, finitely open, or the intersection of two sets each of which is finitely open in itself, then B is finitely open in itself. Also, if B is an open set in a real Banach space $\mathcal{Y}$, and a function $\varphi : B \to \mathfrak{Z}$, where $\mathfrak{Z}$ is also a real Banach space, is Fréchet differentiable at a point $y_0 \in B$, then φ is dually differentiable at y_0.

We now make the following hypotheses on the data of Problem 1. In these hypotheses, (x_0, T_0, π_0) is to be considered a solution of the problem.

HYPOTHESIS 1. – The set Π is finitely open in itself.

HYPOTHESIS 2. – For each T in the convex hull, co $\mathcal{U}$, of $\mathcal{U}$ and each x in A, we have that $T \in \mathcal{C}_0$ and that $[E - DT(x; \cdot)]$ — where E denotes the identity operator on $\mathfrak{X}$ and $DT(x; \cdot)$ the Fréchet differential of T at x — is a (linear) homeomorphism of $\mathfrak{X}$ onto itself.

HYPOTHESIS 3. – For every finite subset $\{T_1, \ldots, T_\rho\}$ of $\mathcal{U}$, there exist a subset $\mathcal{U}_0$ of $\mathcal{U}$ and a number $\epsilon_0 \in (0, 1/\rho)$ with the following properties :

(A) The subset $\mathcal{R}$ of co $\mathcal{U}$ defined by

$$\mathcal{R} = \left\{ T : T = T_0 + \sum_1^\rho \lambda^i (T_i - T_0), 0 \leqslant \lambda^i \leqslant \epsilon_0 \text{ for } i = 1, \ldots, \rho \right\}$$

is contained in $\mathcal{C}_1$.

(B) The function φ_0 is continuous on $\mathcal{R} \cup (\mathcal{C}_1 \cap \mathcal{U}_0)$.

(C) For each neighborhood N of 0 in $\mathcal{C}$, there exists a continuous function

$$\gamma : \{\lambda : \lambda = (\lambda^1, \ldots, \lambda^\rho) \in R^\rho, 0 \leqslant \lambda^i \leqslant \epsilon_0 \text{ for } i = 1, \ldots, \rho\} \to N$$

such that

$$T_0 + \sum_1^\rho \lambda^i (T_i - T_0) + \gamma(\lambda^1, \ldots, \lambda^\rho) \in \mathcal{C}_1 \cap \mathcal{U}_0 \text{ for all } \lambda = (\lambda^1, \ldots, \lambda^\rho).$$

HYPOTHESIS 4. – The functions φ, ϕ^0, and ϕ_1 are dually differentiable at (x_0, π_0).

Hypothesis 4 can be considerably weakened with respect to ϕ^0, and also with respect to ϕ_1 in case Z is a closed, convex cone (with vertex at 0), as is true in most applications. However, space limitations prevent us from going into detail.

Our promised necessary conditions are then as follows :

THEOREM 1. – *Let* (x_0, T_0, π_0) *be a solution of Problem 1 such that Hypotheses 1, 2, and 4 hold. Further, suppose that either* $\mathcal{W}$ *is finitely open in itself or that Hypothesis 3 holds. Then there exist a vector* $\alpha \in R^m$, a number $\beta^0 \leqslant 0$, *and a linear continuous functional* l *on* $\mathfrak{Z}$, *not all zero, such that*

$$(1) \quad (\alpha \cdot D_1 \varphi + \beta^0 D_1 \phi^0 + l \circ D_1 \phi_1) \circ [E - DT_0]^{-1} (Tx_0)$$

$$\leqslant (\alpha \cdot D_1 \varphi + \beta^0 D_1 \phi^0 + l \circ D_1 \phi_1) \circ [E - DT_0]^{-1} (x_0) \text{ for all } T \in \mathcal{U},$$

$$(2) \quad (\alpha \cdot D_2 \varphi + \beta^0 D_2 \phi^0 + l \circ D_2 \phi_1)(\pi - \pi_0) \leqslant 0 \text{ for all } \pi \in \Pi,$$

$$(3) \quad l(z) \geqslant l \circ \phi_1 (x_0) \text{ for all } z \in \overline{Z},$$

where, in (1) and (2), $D_1 \varphi$ *and* $D_2 \varphi$ *denote the partial dual differentials (defined in an obvious manner) of* φ *with respect to its first and second arguments, respectively, at* (x_0, π_0), *and similarly for* $D_1 \phi^0$, *etc. ,* DT_0 *denotes the Fréchet differential of* T_0 *at* x_0, *and E the identity operator on* $\mathfrak{X}$.

Note that if Z is a (convex) cone (with vertex at 0), then (3) holds if and only if $l \circ \phi_1 (x_0) = 0$ and $l(z) \geqslant 0$ for all $z \in \overline{Z}$.

One of the principal applications of our results is to the special case where $\mathfrak{X} = \mathcal{C}^n(I)$, the space of continuous functions from a compact interval $I = [t_1, t_2]$ into R^n. In this case, typically, $\mathcal{U}$ consists of a class of integral operators such that the equation $x = Tx$ represents a functional differential equation of retarded type (which includes ordinary differential equations as a special case) or a Volterra integral equation. The constraint $\varphi(x) = 0$ then typically corresponds to constraints on the initial and final values of the "trajectory" x, and the constraint $\phi_1(x) \in Z$ corresponds to an inequality restriction on the phase coordinates, possibly of the form $g(x(t), t) \leqslant 0$ for all $t \in I$.

Under such conditions, Theorem 1 yields necessary conditions which generalize the Pontryagin maximum principle [5].

REFERENCES

[1] DUBOVITSKII A. Ya. and MILYUTIN A.A. — Extremum problems in the presence of constraints, *Zh. Vychisl. Mat. i Mat. Fiz.*, 5, 1965, p. 395-453.

[2] GAMKRELIDZE R.V. and KHARATISHVILI G.L. — Extremal problems in linear topological spaces, *Izv. Akad. Nauk S.S.S.R. Ser. Mat.*, 33, 1969, p. 781-839.

[3] HALKIN H. — *Nonlinear nonconvex programming in an infinite-dimensional space, Mathematical Theory of Control*, A.V. Balakrishnan and L.W. Neustadt, eds., Academic Press, New York, 1967, p. 10-25.

[4] NEUSTADT L.W. — A general theory of extremals, *J. Comput. System Sci.*, 3, 1969, p. 57-92.

[5] PONTRYAGIN L.S., BOLTYANSKII V.G., GAMKRELIDZE R.V. and MISHCHENKO E.F. — *The Mathematical Theory of Optimal Processes*, John Wiley, New York, 1962.

[6] WARGA J. — Control problems with functional restrictions, *S.I.A.M. J. Control*, 8, 1970, No. 3, p. 360-371.

University of Southern California
Dept. of Electrical Engineering
Los Angeles
California 90007 (USA)

Actes, Congrès intern. Math., 1970. Tome 3, p. 187 à 192.

CONVEXITY IN EXISTENCE THEORY OF OPTIMAL SOLUTION

by Czesław OLECH

In existence problem of optimal solutions to some optimal control problems one has to deal with sets of integrable functions consisting of all integrable selections of a set - valued mapping. More precisely, there is given a mapping Q

$$G \ni t \to Q(t) \subset R^N,$$

where G is a bounded domain of R^n and one considers the set $\Omega = \Omega_Q \subset L_1\ (G,\ R^N)$ of Lebesgue integrable functions from G into R^N defined by

$$\Omega = \{v \in L_1\ (G, R^N) \mid v(t) \in Q(t) \text{ a.e. in } G\}. \tag{1}$$

Roughly speaking, the so called minimizing sequence can be located in a set of the form (1), and we wish to know that such a sequence contains a convergent subsequence, and that the limit is again in Ω. This limit would be a candidate for optimal solution. Thus certain closedness and local compactness properties (minimizing sequences are in general bounded in L_1 norm) of sets defined by (1) are very much dezirable in existence theory. Of course the properties in question depend on topology we choose for Ω, and since we like to have some local compactness rather weak topologies should be considered.

Denote by clco $Q\ (t)$ the closed convex hool of $Q\ (t)$, by $\Omega_1 = \Omega_{\text{clco}\ Q}$, by $\overline{\Omega}$ the weak closure of Ω. Under rather mild regularity condition for the map Q one can prove the inclusion : $\Omega_1 \subset \overline{\Omega}$.

This inclusion shows that in order to have weak closedness property of Ω the values of Q should be closed and convex. In fact, if Q is such then Ω is convex and strongly closed, hence also weakly closed.

This is one place when convexity interveans in existence theory.

However bounded sequences in L_1 are not in general weakly precompact. To avoid this unpleasent situation we consider a weaker topology for Ω. The one which is convenient is the weak* topology of $C_0^*(G, R^N)$, where by $C_0(G, R^N)$ we denoted the space of uniformly continuous function from G into R^N with constant value on the boundary of G. Now, we imbed isometriedly Ω into the space of regular measures in G with values in R^N. Below the weak* topology will always mean the weak* topology of $C_0^*(G, R^N)$.

We ask again similar question : under which conditions of Q the set Ω is weak* closed. Here convexity and closedness of values of Q do not suffice.

For maps Q of a special form a necessary and sufficient condition is known (Lasota and Olech). Namely if

$$(2)\qquad Q(t) = \{q \mid < q, a > \leqslant h(t)\},$$

where $a \neq 0$ is fixed and $h : G \to R \cup \{+\infty\}$,

is Lebesgue measurable, then Ω_Q is weak* closed if and only if for almost all t, for which $h(t)$ is finite, h restricted to a neighborhood of t is integrable.

It would be convenient to give a name, say property P, to the assumption concerning function h in the latter statement.

Since we have $\Omega_{\cap Q_i} = \cap \Omega_{Q_i}$ where by $\cap Q_i$ we denoted the map $t \to \cap Q_i(t)$ thus a sufficient condition for weak* closedness of Ω is that Q can be represented by

$$(3)\qquad Q(t) = \bigcap_{i=1}^{\infty} \{q \mid < q, a_i > \leqslant h_i(t)\},$$

where h_i has Property P for each i and a_i are fixed.

The sufficient condition for Q to be representable in the form (3) is that Q is upper semi continuous in the Cesari [1] sense ; that is

$$(4)\qquad Q(t_0) = \bigcap_{\epsilon > 0} \text{clco} \bigcup_{|t - t_0| < \epsilon} Q(t) \qquad \text{for each } t_0 \in G.$$

Manifestly, if (4) holds then values of Q has to be closed and convex. Denote by $C_{Q(t)}$ the so called asymptotic cone of $Q(t)$; that is

$$(5)\qquad C_{Q(t)} = \{c \mid \lambda c + q \in Q(t) \quad \text{whenever } \lambda > 0 \quad \text{and } q \in Q(t)\}.$$

If Q satisfies (4) than so does $C_{Q(t)}$. By $D(t)$ we denote the polar cone $C^0_{Q(t)}$ of the asymptotic cone of $Q(t) \cdot D(t)$ is given by

$$(6)\qquad D(t) = \{d \mid < d, c > \leqslant 0 \qquad \text{for each } c \in C_{Q(t)}\}.$$

$D(t)$ is again a closed convex cone. Denote by $M(t)$ the smallest subspace of R^N containing $D(t), M(t) = D(t) - D(t)$, and by intrint $D(t)$ (the intrinsic interior) the interior of $D(t)$ in the topology of $M(t)$.

If $Q(t)$ satisfies (4) and $a \in$ intrint $D(t_0)$ then the support function

$$g(t, a) = \sup < q, a >, \quad \text{if } q \in Q(t),$$

is finite and upper semi continuous in t at t_0.

To obtain the representation (3) for $Q(t)$ from semicontinuity it is enough to choose $\{a_i\}$ so that $D(t) \cap \{a_i\}$ is a dence subset of $D(t)$ for each i, (this is possible, since because of (4) the set-valued function M takes on at the most denumerably many different values), and to define $h_i(t) = g(t, a_i)$ if $a_i \in$ intr. int $D(t)$ and $\neq \infty$ elsewhere.

So defined functions h_i satisfy property P since they are upper semi continuous, hence locally integrable, at each t at which they are finite.

One can check also that $D(t)$ is lower semi continuous if (4) holds.

Now we assume that $Q(t)$ is given by (3) with h_i satisfying property P for each i and we shall give a characterization of the weak* closure of Ω in the space $C^*_0(G, R^N)$.

First let us put

$$(7)\qquad \widetilde{D}(t) = \text{clco}\,\{d \mid d = \lambda a_t,\ \lambda > 0 \quad \text{and } h_t \text{ is locally integrable at } t\},$$

$$(8)\qquad \widetilde{C}(t) = \widetilde{D}^0(t) \quad (\text{the polar of } \widetilde{D}(t)).$$

Manifestly both $\widetilde{D}(t)$ and $\widetilde{C}(t)$ are closed convex cones and we note that $\widetilde{C}(t) = C_{Q(t)}$ a.e. in G. One can prove that $\widetilde{D}(t)$ is lower semi continuous and $C(t)$ is upper semi continuous in the Cesari sense.

Let measure m belongs to $C_0^*(G, R^N)$ and denote by m_a and m_s, the absolutly continuous part of m and singular part of m, respectively. Then m belongs to the weak* closure of Ω iff

$$(\text{i})\qquad \frac{dm_a}{dt}(t) \in Q(t) \text{ a.e. in } G$$

and

$$(\text{ii})\qquad m_s \text{ is } \widetilde{C}\text{-valued.}$$

The latter means that for any measure μ such that m_s is absolutly continuous with respect to μ, $\frac{dm_s}{d\mu}(t) \in \widetilde{C}(t)$ μ-a.e. in G.

In particular, if $\widetilde{C}(t) \subset C$ where C is a fixed convex and closed cone, then for any measurable $A \subset G, m_s(A) \in C$.

If $\widetilde{C}(t)$ is flat and contained in a fixed subspace of R^N then more can be said about compactness of bounded sequence of Ω. For this purpose assume that

$$(9)\qquad \widetilde{C}(t) \subset C,$$

where C is a fixed closed convex and proper cone and besides assume that $X = C - C$ is a proper subspace of R^N. Since C is proper thus the polar C^0 has non empty interior.

It is not dificult to see that in this case subsets of Ω of the form

$$(10)\qquad \Omega(M, d) = \{u \in \Omega \mid < d, \int_G u(t)\,dt > \geqslant M > -\infty\}$$

are bounded in L_1 norm if $d \in \text{int}\, C^0$.

Denote by Y the orthogonal complement of X. Then (iii) the orthogonal projection of $\Omega(M, d)$ into Y is weakly sequentialy compact in the weak topology of $L_1(G, Y)$.

In fact if $\{u_a\} \subset \Omega(M, d)$ then $\{u_a\}$ is bounded in L_1 norm and as such is weak* precompact. For simplicity, suppose u_a is weak* convergent to a measure m. Denoting by $u_{ax}(t)$ and $u_{ay}(t)$ the orthogonal projections of $u_a(t)$ into X and Y, respectively, and by m_x, m_y the analogous decomposition of the measure m, we see by (i) and (ii) that m_y is absolutly continous. Thus denoting $u_{0y}(t) = \frac{dm_y}{dt}(t)$, we conclude that the weak* limit of u_{ay} is again a function in $L_1(G, Y)$. Additionally we can prove using (3) and (9) that $\{u_{ay}\}$ is equiabsolutly integrable, thus the convergence of u_{ay} to u_{0y} is weak in L_1.

Statements (i), (ii), (iii) form a generalization of a lemma the author used in proving some existence theorenes of optimal solution, in [6] for one dimensional case and in [7] for multidimensional case. A similar result obtained more recently R.T. Rochafellar [9] as a consequence of some formulas derived by him for conjugates of certain convex integral functionals on Banach spaces of measurables or continuous vector-valued functions.

To ilustrate how the above statements can be usefull in getting existence theorem in optimal control problems, we shall consider a problem, Cesari [2, 3] refers to as multidimentional Lagrange problem.

The problem is to find minimum of a functional

$$I(v, z) = \int_G f(t, z, v)\, dt \tag{11}$$

in a class of pairs (v, z), $v : G \to R^s$, is measurable, $z : G \to R^k$ belongs to Sobolew space $H_1^1\ (G, R^k)$, satisfying : (a) system of partial differential equations of the form

$$\nabla z = g(t, z, v), \tag{12}$$

(b) constrains of the form $(t, z(t)) \in A \subset R^n \times R^k$ and $v(t) \in V(t, z(t)) \subset R^s$ a.e. in G and

(c) a boundary condition $z \in z_0 + H_{10}^1(G, R^k)$, where $z_0 \in H_1^1$ is fixed and H_{10}^1 is the subspace of H_1^1 of functions vanishing at the boundary of G. There z is referred to as a space variable, v is called control parameter or control function. A pair (v, z) satisfying (a), (b) and (c) is called an admissible pair. Thus the problem is to minimize (11) in the class of admissible pairs. This problem is a straight forward generalization of Pontryagin optimal control problem to both the multidemsional case ($t \in G \subset R^n$, $n > 1$), as well as to unbounded V.

The following existence theorem is proved in [7].

EXISTENCE THEOREM. — *Assume that $f(t, z, v)$, $g(t, z, v)$ are continuous in z, v for each t fixed, and measurable in t for each fixed z, v, that the set A in (b) is closed, that the set-valued function $V(t, z)$ is upper semi continuous in both variables (by that we mean that the graph $(t, z, v) \mid v \in V(t, z)$ is closed), that the set*

$$P(t, z) = \{(q, p) \mid p = g(t, z, v), f(t, z, v) \leqslant q,\ v \in V(t, z)\} \tag{13}$$

is convex, closed and upper semicontinuous in z for each t fixed, finally assume that for each $d \in R^{nk}$ there is an integrable function ψ_d from G into R such that

$$< d, g(t, z, v) > - f(t, z, v) \leqslant \psi_d(t) \text{ a.e. in } G \tag{14}$$

and each $(t, z) \in A$, $v \in V(t, z)$

Under these assumptions there exists an optimal solution provided the class of admissible pairs is not empty.

This theorem contains an existence theorem of Morrey (cf. [5], p. 24) for weak solution to the variational problems involving multiple integral. This is the case, when there is no constrain of type (b) and $\nu = \nabla z$. In this case $P(t, z)$ reduces to the epigraph of f and assumptions concerning P mean that f is convex with respect to ν. Assumption (14) corresponds to the so called growth condition known in the calculus of variation. Putting $d = 0$, in (14), we get a lower bound for (11). Thus there exist a minimizing sequence $(\nu_a, z_a,)$ for the problem.

The two main steps in a proof of this result goes as follows.

Consider the set-valued function $Q(t)$ given by

$$Q(t) = \bigcap_a \{(q,p) \mid -q + <d, p> \leqslant \psi_d(t) \text{ a.e. in } G\} \subset R \times R^{nk}.$$

For this mapping the cone $C(t) = \text{const} = \{(q, p) \mid p = 0, q \geqslant 0\}$ and (i), (ii), and (iii) hold. We can see also that

$$(q_a(t), p_a(t)) = (f(t, z_a(t), \nu_a(t)), \nabla z_a(t)) \in Q(t) \text{ a.e. in } G.$$

Thus by (iii)$p_a(t)$ contains weakly convergent subsequence and we may assume that p_a itself converges weakly to p_0. By (i) and (ii) $q_a$converges to a measure m whose singular part is nonnegative. Denoting by

$$q_0(t) = \frac{dm_a}{dt}(t)$$

we know that $(q_0(t), p_0(t)) \in Q(t)$ a.e. in G. Taking into acount the boundary condition (b) and some basic facts from the theory of Sobolev spaces we can conclude that $z_a \to z_0$ weakly in H_1^1 and that $\nabla z_0(t) = p_0(t)$. This means, in particular, that $z_a \to z_0$ strongly in L_1, hence we may assume that

(15) $$z_a(t) \to z_0(t) \text{ a.e. in } G.$$

The second step uses the fact, which is a consequence of regularity assumption on f and g and of (14), that the set-valued function $P(t, z)$ is upper semi-continuous in Cesari sence with respect to z. This together with the obvious inclusion

$$(q_a(t), p_a(t)) \in \text{clco} \bigcup_{\beta \geqslant \gamma} P(t, z_\beta(t)), \quad \text{if } \alpha \geqslant \gamma$$

and (15) gives that

(16) $$(q_0(t), \nabla z_0(t)) \in P(t, z_0(t)) \text{ a.e. in } G.$$

To complete the argument we need only to deduce from (16) that there is a measurable ν_0 such that

$$\nabla z_0(t) = g(t, z_0(t), \nu_0(t)), \ \nu_0(t) \in V(t, z_0(t)) \quad \text{and} \quad q_0(t) \geqslant f(t, z_0(t), y_0(t))$$

which requires a kind of implicit function theorem or a selection theorem very often referred to as a Filipov Lemma. Now it is a simple matter to check that (ν_0, z_0) is the optimal solution.

The way of proving existence theorems just skeched does not separates two properties : precompactness of a minimizing sequence and upper semicontinuity of functional in question but rather we obtained both simultanously.

From the characterization of the weak* closure of sets of the form (3) we described above one can also prove lower semicontinuity of some functional (or lower closure properties in the terminology of Cesari). Roughly speaking the difference in such results is that we already are assuming certain convergence for z_a, thus we need not the condition (14) to hold for each d. This is the case when the asymptotic cone of $Q(t)$ need not be constant as above but may change with t.

The convexity assumption in the existence theorems (convexity of sets $P(t, z)$) appears to be essential. However there is a class of one dimensional optimal control problems in which existence can be obtained without convexity assumption. For example, the classical two point variational problem of minimizing an integral functional

$$\int_a^b f(t, \dot{z}(t))\, dt,$$

which does not depend on function z but only on its derivative. In those problems it appears to be sufficient for deriving existence to know that the integral of set valued fonction Q is a closed set. For details we refer the reader to [8].

REFERENCES

[1] Cesari L. — Existence theorems for weak and usual optimal solutions in Lagrange problems with unilateral constrains, *I. Trans. Am. Math. Sic.* 124, 1966, p. 3⁹6-472.

[2] Cesari L. — Existence theorems for multidimensional Lagrange problems. *J. Optimization Theory Appl.* 1, 1967, p. 87-172.

[3] Cesari L. — Sobolev spaces and multidimensional Lagrange problems of optimization. *Ann. Scuola Norm. Sup. Pisa.* 22, 1960, p. 193-227.

[4] Lasota A. and Olech C. — On Cesari's semicontinuity condition for set valued mappings, *Bull. Acad. Polon. Sci., ser. sci. math. astr. phys.* 16, 1968, p. 711-716.

[5] Morrey C.B. Jr. — *Multiple Integrals in the Calculus of Variations.* Springer-Verlag, New York, 1966.

[6] Olech C. — Existence theorems for optimal problems with vector valued cost functions. *Trans. Am. Math. Soc.* 136, 1969, p. 157-180.

[7] Olech C. — Existence Theorems for Optimal Control Problems Involving Multiple Integrals. *J. Diff. Eq.* 6, 1969, p. 512-526.

[8] Olech C. — Integrals of set-valued functions and linear optimal control problems. *Proc. International Conference on Math. Control Theory in Brussels,* May 1969, p. 109-125.

[9] Rockafellar R.T. — *Integrals which are convex functionals,* II.

Instytut Matematyczny PAN
ul. Solskiego 30
Kraków
Pologne

Actes, Congrès intern. Math., 1970. Tome 3, p. 193 à 199.

E5 - COMBINATORIQUE - ALGÈBRE FINIE

RECENT DEVELOPMENTS IN ENUMERATION THEORY

by N.G. de BRUIJN

Dedicated to Marshall Hall, Jr. on the Occasion
of his Sixtieth Birthday

1. Introduction.

In this lecture we indicate a number of recent trends to extend Pólya's enumeration theorem, and we shall try to cover these by means of a small number of theorems. We first cite Pólya's theorem (with formulation and notation of [3]).

THEOREM 1 (Pólya). – *Let D and R be finite sets, let G be a group of permutations of D, let $\mathfrak{A}$ be a commutative algebra over the rationals, let w be a mapping of R into $\mathfrak{A}$. Two mappings f_1, f_2 of D into R are called equivalent if $g \in G$ exists such that $f_1 = f_2 g$. The equivalence classes are called mapping patterns. Every $f \in R^D$ is given as a weight $W(f) = \Pi_{d \in D}\, w(f(d))$. All f in a pattern have the same weight ; this is called the weight of the pattern. Then the sum of the weights of the patterns is*

$$P_G\Big(\sum_{r\in R} w(r),\ \sum_{r\in R} (w(r))^2,\ \sum_{r\in R} (w(r))^3, \ldots\Big).$$

Here $P_G(x_1, x_2, \ldots)$ is the cycle index of G :

$$P_G(x_1, x_2, \ldots) = |G|^{-1} \sum_{g\in G} x_1^{b_1(g)}\, x_2^{b_2(g)} \ldots,$$

where $b_i(g)$ is the number of cycles of length i in the permutation g.

The theorem has been extended in many directions. Admittedly, quite often such generalisations are consequences of the theorem itself. This is not to be wondered at, since this kind of enumeration theory is a matter of exposition and organisation of things which are in essence trivial. The usual mathematician's attitude is that generalisations of a theorem are deeper and more respectable than the theorem itself, but in a field like enumeration this does not always hold : many nice results are specializations of trivial statements.

Some tendencies to generalize Pólya's theorem are the following.

(i) The weights can be generalized, and in particular they can be chosen such that they bear information about how often an element of R is taken as a function value ([2]).

(ii) A more flexible point of view is obtained if we do not speak about a group G of permutations of a set, but about an abstract group G with a representation χ by means of permutations of that set.

(iii) We can ask for the number of patterns which are invariant with respect to a fixed permutation ([5]).

(iv) We can introduce a permutation group H of R and define equivalence by: $f_1 \sim f_2$ if $g \in G, h \in H$ exist with $f_2 = hf_1\, g^{-1}$ ([2], [3]).

(v) Instead of defining equivalence by means of two groups G and H as under (iv), we can consider two different representations of a single group, one acting on D and one on R (Theorem 4 of [2] ; [6]). (Example, patterns of mappings of the set of vertices of a cube into the set of edges of that cube). The case of two different permutation groups G and H can be seen as a special case : both G and H are representations of G x H.

(vi) We have the posibility of partitioning D and R, and to restrict the f's by requiring that each part of D is mapped into a corresponding part of R ([11] ; [10] ; [8], Theorem 2).

(vii) There are results where the cycle index is replaced by

$$|G|^{-1} \sum_{g \in G} \tau(g)\, x_1^{b_1(g)}\, x_2^{b_2(g)} \ldots,$$

where τ is a group character.

(viii) There is the possibility to drop the requirement that weights are constant over a pattern : we can just speak of the average of $W(f)$, averaged over the pattern.

The standard derivation for Pólya's theorem is in three stages.

1/ Develop Burnside's lemma for the sum of the weights of the orbits induced in a set by a permutational representation of a group.

2/ Determine the sum of the weights of those mappings of D into R which are invariant under a fixed permutation of D.

3/ Apply Burnside's lemma to the set R^D, with the representation $g \to \pi(g)$, defined by $\pi(g) f = fg^{-1}$.

We shall present generalized versions for each one of these stages. The first extension of Burnside's lemma (Theorems 1a, 1b) was implicit in [5], the second one (Theorems 2a, 2b) appeared in [7] as a generalisation of the Burnside-like lemma needed in a Pólya-like theorem by S.G. Williamson. Another special case of Theorem 2b is Theorem 1 of [4] ; in that case $P(x)$ is not an averaged character but a cycle index. The theorems we present for the second stage were implicit in [2]. As to the third stage, we adopt the point of view taken in [6].

We give no proofs in this paper. A detailed exposition will probably appear in the forthcoming book [9].

2. Theorems of the Burnside type.

THEOREM 1a. – *Let X be a finite set, let G be a finite group, let $\mathfrak{A}$ be a vector* *space over the rationals, let W be a mapping of X into $\mathfrak{A}$, let π be a representation of G by permutations of X, and let ρ be a fixed permutation of X. In X we have an equivalence relation : x_1 and x_2 are called equivalent if there is a $g \in G$ with $\pi(g)\, x_1 = x_2$. The equivalence classes are called $(G\, , \pi)$-patterns. Let C_x be the pattern to which x belongs. Then we have*

$$(2.1) \qquad \sum_{x \in V(\rho)} \frac{W(x)}{|C_x|} = \frac{1}{|G|} \sum_{g \in G} \; \sum_{x \in U(g,\rho)} W(x)$$

Here $V(\rho) = \{x \in X \mid \rho x \in C_x\}$, $U(g\, , \rho) = \{x \in X \mid \pi(g)x = \rho x\}$.

THEOREM 1b. – *If we have, in addition to the assumptions of the Theorem 1a,*

(i) *for each $g \in G$ there is a $g' \in G$ with $\rho\pi(g)\rho^{-1} = \pi(g')$,*

(ii) *for all $g \in G$, $x \in X$ we have $W(x) = W(\pi(g)x)$,*

then $V(\rho)$ is the union of a set of patterns (that is, the patterns invariant under ρ) and $W(x)$ is constant over each pattern. If the common value of the weights of the elements of a pattern is called the weight of the pattern, then the sum of the weights of the patterns invariant under ρ is expressed by the right-hand side of (2.1).

THEOREM 2a. – *In the situation of Theorem 1a we add the assumption that $\mathfrak{A}$ is a commutative algebra over the rationals and that τ is a mapping of G into $\mathfrak{A}$. For every $x \in X$ we denote by G_x the group $= \{x \in X \mid \pi(g)x = x\}$, and by $P(x)$ the average of $\tau(g)$ over G_x.* Then we have

$$(2.2) \qquad \sum_{x \in X} \frac{W(x)\, P(x)}{|C_x|} = \frac{1}{|G|} \sum_{g \in G} \tau(g) \sum_{x \in U(g)} W(x),$$

where $U(g) = \{x \in X \mid \pi(g)x = x\}$.

THEOREM 2b. – *If we moreover assume that*

$$W(x) = W(\pi(g)x) \quad \text{and} \quad \tau(hgh^{-1}) = \tau(h)$$

for all g, $h \in G$, $x \in X$, then $W(x)\, P(x)$ is constant over each pattern, whence the left-hand side of (2.2) can be seen as $\Sigma\, W(x)\, P(x)$ where a single x is taken from each pattern.

3. Theorems for summing weights of invariant mappings.

Let R and D be finite sets and let $\mathfrak{A}$ be a commutative algebra over the rationals. If $f \in R^D$, then the *degree* of f is the mapping δf of R into the sets $N^* = \{0, 1, 2, \ldots\}$, defined by

$$(\delta f)(r) = |\{d \in D \mid f(d) = r\}| \qquad (r \in R).$$

(cf J. Riguet, Appendix 5 in [1]).

We introduce a doubly indexed set of variables $x_{r,m}$ $(r \in R\, , m \in N^*)$. If q maps R into N^*, we attach to q the monomial

$$M(q) = \prod_{r \in R} x_{r,q(r)}.$$

We discuss the $M(q)$'s rather than the q's, since we can operate with the $M(q)$'s in the ring of all polynomials in the $x_{r,n}$. We assume that the $x_{r,n}$ are in $\mathfrak{A}$.

If S is a finite set, η a permutation of S, and if Ω attaches an element of $\mathfrak{A}$ to each subset of S, then we define

$$\Theta(S\,,\eta\,,\Omega) = \Omega(S_1)\ldots\Omega(S_k),$$

if $S_1, \ldots, S_k$ are the orbits of η in S.

We introduce the polynomials $d_m\,(\zeta_1\,,\zeta_2,\ldots)$ by the development

$$\exp(\zeta_1 x + \zeta_2 x^2 + \zeta_3 x^3 + \ldots) = \sum_{m=0}^{\infty} d_m\,(\zeta_1\,,\zeta_2,\ldots)\,x^m.$$

THEOREM 3. – *Let μ and σ be permutations of D and R, respectively. We define Ω_1 by*

$$\Omega_1(B) = \frac{\partial}{\partial z_{|B|}}$$

for every $B \subset D$, and Ω_2 by

$$\Omega_2\,(C) = \sum_{m=0}^{\infty} d_m\,(|C|\,z_{|C|},\,|C|\,z_{2|C|},\ldots)\prod_{r\in C} x_{r,m}$$

for every $C \subset R$. Then we have, if $\Sigma_f^{(\mu,\sigma)}$ denotes the sum over all $f \in R^D$ with $\sigma f \mu^{-1} = f$,

$$\Sigma_f^{(\mu,\sigma)}\,M(\delta f) = [\Theta(D\,,\mu\,,\Omega_1)\,\Theta(R\,,\sigma\,,\Omega_2)]_{z_1=z_2=\ldots=0}$$

with obvious formalism : the differentation is carried out in the ring of all polynomials in the z_i with coefficients in $\mathfrak{A}$. An alternative expression is

$$\Sigma_f^{(\mu,\sigma)}\,M(\delta f) = \Theta\,(D\,,\mu\,,\Omega_1)\,\Theta(R\,,\sigma\,,\Omega_3)\prod_{r\in R}\Big(\sum_{n=0}^{\infty}\frac{x_{r,n}\,y_r^n}{n!}\Big)$$

with

$$\Omega_3\,(C) = \exp\Big(|C|\sum_{m=1}^{\infty} z_{m|C|}\prod_{r\in C}\Big(\frac{\partial}{\partial y_r}\Big)^m\Big),$$

and the differentiations are carried out at the point where all z_i and all y_r are zero.

We get simpler expressions if we take a simpler weight, viz.

$$W_0(f) = \prod_{d\in D} w(f(d)),$$

where w is a mapping of R into $\mathfrak{A}$. This can be obtained from $M(\delta f)$ if we take $x_{r,m} = (w(r))^m$. We find

THEOREM 4. – *Defining Ω_4 by*

$$\Omega_4\,(C) = \exp\Big(|C|\sum_{m=1}^{\infty} z_{m|C|}\Big(\prod_{r\in C} w(r)\Big)^m\Big)\,,$$

we have

$$\Sigma_f^{(\mu,\sigma)} W(f) = [\Theta(D\,,\mu\,,\Omega_1)\,\Theta(R\,,\sigma\,,\Omega_4)]_{z_1=z_2=\ldots=0}.$$

Another expression for the same sum is

$$\Sigma_f^{(\mu,\sigma)} W(f) = \prod_B \lambda(|B|,\sigma),$$

where B runs through the set of orbits of μ, and

$$\lambda(k\,,\sigma) = \sum_{r\in R,\,\sigma^k r=r} w(r)\,w(\sigma r)\ldots w(\sigma^{m-1}r).$$

We briefly mention a case of summing weights which plays a role in the situation indicated in section 1 under (vi). Let D be partitioned into $D_1,\ldots,D_k$. For each i we take a weight function W_i, attaching a weight $W_i(f)$ to each mapping of D_i into R. Now if f maps D into R, and $f\,|\,D_i$ is the restriction of f to D_i, we define $W(f) = W_1(f\,|\,D_1)\ldots W_k(f\,|\,D_k)$. Assume that μ transforms each D_i into itself, and denote by μ_i the restriction of μ to D_i. Then we have

$$\sum_{f\in R^D}^{(\sigma,\mu)} W(f) = \prod_{i=1}^{k} \sum_{f\in R^{D_i}}^{(\mu_i,\sigma)} W_i(f).$$

The situation mentioned in section 1 under (vi) is not more general, since it can be obtained by suitable choice of the weight functions.

5. Theorems of the Pólya type.

We keep the notation of the previous section. Moreover, let G be a finite group, let χ and ζ be representations of G by permutations of R and D, respectively. And let μ and σ be fixed permutations of R and D, respectively, not necessarily of the form $\chi(g)$ or $\zeta(g)$.

These $\chi\,,\zeta$ induce a representation π of G by permutations of R^D:

$$\pi(g) : f \to \zeta(g)\,f(\chi(g))^{-1}\ ;$$

similarly, μ and σ induce a permutation ρ of R^D, viz.

$$\rho : f \to \sigma f \mu^{-1}.$$

We now apply Theorem 1a, with $X = R^D$, $W(f) = M(\delta(f))$. We have

$$\sum_{f\in U(g,\rho)} W(f) = \Sigma_f^{(\mu^{-1}\chi(g),\,\sigma^{-1}\zeta(g))} M(\delta f)$$

with the notation of Theorem 3. Applying that theorem, we obtain the following.

THEOREM 5a. – *In R^D we have equivalence defined by $f_1 \sim f_2$ if $g\in G$ exsists such that $f_1 = \zeta(g)\,f_2\,(\chi(g))^{-1}$. Equivalence classes are called mapping patterns. The mapping pattern to which f belongs in called C_f. If $V(\rho)$ is the set of all $f\in R^D$ for which $\sigma f\mu^{-1} \sim f$, we have*

$$\sum_{f\in V(\rho)} \frac{M(\delta f)}{|C_f|} = \frac{1}{|G|} \sum_{g\in G} [\Theta(D,\mu^{-1}\chi(g)\,,\Omega_1)\Theta(R,\sigma^{-1}\zeta(g),\Omega_2)]_{z_1=z_2=\ldots=0}.$$

THEOREM 5b. – *In addition to the conditions of Theorem 5a, we assume that for all $r \in R$, $m \in N^*$ we have $w(r, m) \in \mathfrak{A}$ such that*

$$w(r, m) = w(\zeta(g)\, r, m)$$

for all $r \in R$, $m \in N^*$, $g \in G$. And we assume that for every $g \in G$ there exists a $g' \in G$ such that simultaneously

$$\mu^{-1}\chi(g)\,\mu = \chi(g'), \quad \sigma^{-1}\zeta(g)\,\sigma = \chi(g').$$

If $f_1 \sim f_2$, then $\sigma f_1 \mu^{-1} \sim \sigma f_2 \mu^{-1}$, whence ρ maps patterns into patterns. A pattern mapped into itself is called *ρ-invariant*. The *weight* of a mapping pattern is taken to be

$$W_1(f) = \prod_{r \in R} w(r, (\delta f)(r))$$

(i.e. , the value obtained from $M(\delta f)$ upon the substitution $x_{r,n} = w(r, n)$) where f is an arbitrary function in that pattern (we have $f_1 \sim f_2 \Rightarrow W_1(f_1) = W_1(f_2)$). Under these conditions the sum of the weights of the ρ-invariant mapping patterns is

$$\frac{1}{|G|} \sum_{g \in G} [\theta(D, \mu^{-1}\chi(g), \Omega_1)\,\theta(R, \sigma^{-1}\zeta(g), \Omega_5)]_{z_1 = z_2 = \ldots = 0},$$

where Ω_5 is obtained from Ω_2 if we replace all $x_{r,m}$ by the corresponding $w(r, m)$.

Needless to say we can use the other expressions of theorems 3 and 4 in the same manner. Furthermore we can build a theorem of Pólya type if we use the superposition principle explained at the end of section 3. And we can get more Pólya type theorems if we apply the Burnside type theorem 2.

REFERENCES

[1] BERGE C. — *Théorie des graphes et ses applications,* Collection Universitaire de Mathematiques 2, Dunod, Paris, 1958. Translation by Alison Doig, *The Theory of Graphs and its Applications,* John Wiley and Sons, New York, 1962.

[2] DE BRUIJN N.G. — Generalization of Pólya's Fundamental Theorem in Enumerative Combinatorial Analysis, *Nederl. Akad. Wetensch. Proc. Ser.* A 62 = Indag. Math. 21, 1959, p. 59-69.

[3] DE BRUIJN N.G. — Pólya's Theory of Counting, ch. 5 in *Applied Combinatorial Mathematics,* ed. by E.F. Beckenbach, 1964, p. 144-184.

[4] DE BRUIJN N.G. — Enumerative combinatorial problems concerning structures, *Nieuw Archief Wiskunde* (3) 11, 1963, p. 142-161.

[5] DE BRUIJN N.G. — Color patterns that are invariant under a given permutation of the colors. *J. of Combinatorial Theory* 2, 1967, p. 418-421.

[6] DE BRUIJN N.G. — *Enumeration of mapping patterns.* (To appear).

[7] De Bruijn N.G. — Generalization of S. G. Williamson's generalization of Burnside's lemma and Pólya's theorem. Notitie nr. 56, 1968-1969. Internal Report, Department of Mathematics, Technological University, Eindhoven.

[8] De Bruijn N.G. and Klarner D.A. — Enumeration of generalized graphs. *Nederl. Akad. Wetensch. Proc. Ser. A* 72 = Indag. Math. 31, 1969, p. 1-9.

[9] De Bruijn N.G. and Klarner D.A. — *Pattern Enumeration.* (To appear).

[10] Polya G. — Kombinatorische Anzahlbestimmungen für Gruppen, Graphen und chemische Verbindungen, *Acta Math.* 68, 1937, p. 145-254.

[11] Robinson R.W. — Enumeration of Colored Graphs, *J. of Combinatorial Theory* 4, 1968, p. 181-190.

Technological University
Dept. of Mathematics,
Eindhoven (Pays-Bas)

Actes, Congrès intern. Math., 1970. Tome 3, p. 201 à 210.

ON THE APPLICATION OF COMBINATORIAL ANALYSIS TO MUMBER THEORY GEOMETRY AND ANALYSIS

by P. ERDÖS

In this lecture I will discuss the application of some well known and less well known theorems in combinatorial analysis to various other branches of mathematics. In other words I will not mention combinatorial types of reasoning the use of which is of course very wide-spread (e.g. the classical proof of Carleson on the almost everywhere convergence of Fourier series of functions in L_2 is full of combinatorial reasoning), but will restrict myself to cases where definite quotable theorems are used. My paper in no ways claim to give a complete survey of all the applications of combinatorial theorems and is certainly heavily biased towards my own work. Though combinatorics has been successfully applied to many branches of mathematics these can not be compared neither in importance nor in depth to the applications of analysis in number theory or algebra to topology, but I hope that time and the ingenuity of the younger generation will change this.

First we discuss some applications of Ramsey's theorem. The classical theorem of Ramsey states as follows : Let S be an infinite set. Split the k-tuples of S into r classes. Then there is an infinite subset S_1 of S all whose k-tuples are in the same class. The finite form of Ramsey's theorem states that to every k and $u_1, \ldots, u_r$ there is a smallest integer $R_k^{(r)}(u_1, \ldots, u_r)$ so that if we split the k-tuples of a set $|\mathfrak{S}| = R_k(u_1, \ldots, u_r)$ into r classes then for at least one i there is an $S_i \subset S$, $|S_i| \geqslant u_i$ all whose k-tuples are in the i-th class. The exact determination, or even good estimation of $R_k^{(r)}(u_1, \ldots, u_r)$ is a difficult problem which is very far from being solved and we do not discuss it here.

Ramsey's theorem was often rediscovered. Szekeres [1] rediscovered it in connection with the problem of Miss Klein. Miss Klein observed that if there are 5 points in the plane no three of them on a straight line then there are always 4 of them which determine a convex quadrilateral. She then asked : Is there a smallest $f(n)$ so that if there are $f(n)$ points in the plane no three on a line then there are always n of them which determine the vertices of a convex n-gon. Szekeres observed that

$$f(n) \leqslant R_4^{(2)}(5, n) \tag{1}$$

since an n-gon all whose quadrilaterals are convex is itself convex. Thus Ramseys theorem immediątely gives a positive answer to Miss Klein's question.

(1) gives a very poor upper bound for $f(n)$. Szekeres in fact conjectured $f(n) = 2^{n-2} + 1$. This is Miss Klein's result for $n = 4$ and for $n = 5$ it was

proved by Turán and E. Makai by methods of elementary geometry. $n > 5$ is not settled so far. Szekeres and I proved $f(n) \geqslant 2^{n-2} + 1$ [2] (there are some minor inaccuracies in our proof which were corrected by Kalbfleisch), and I [1] proved $f(n) \leqslant \binom{2n-4}{n-2}$.

Ramsey originally discovered his theorem for the purpose of some logical applications. Hajnal, Rado and I in our partition calculus [3] systematically studied the generalisations of Ramseys theorem to higher cardinal numbers, our results have applications to logic and model theory, also Hajnal and Juhász applied our results to set theoretic topology, but I do not discuss these transfinite applications here. Also Ramsey's theorem has many generalisations and extensions but I can not discuss them here. (Erdős-Rado London Journal 1950, Nash-Williams. . . Cambridge Phil. Soc.).

It is obvious that if there are $n + 2$ points in n dimensional space then not all the distances can be equal. Schoenberg and Seidel in fact determined the minimum of the ratio of the maximal distance divided by the minimal distance. Several years ago Coxeter asked me to determine or estimate the smallest integer $f(n)$ so that if there are $f(n)$ points in n dimensional space then they determine at least three different distances. It immediately follows from Ramseys theorem that

$$f(n) \leqslant R_2\,(n+2, n+2) \tag{2}$$

(2) in fact is a very poor estimate, probably $f(n) < c_1 n^{c_2}$ and perhaps $f(n) = \left(\frac{1}{2} + o\,(1)\right) n^2$. By fairly complicated arguments I can prove $f(n) < \exp\,(n^{1-\epsilon})$.

Let $A\,(k\,, n)$ be the smallest integer so that if there are given any $A\,(k\,, n)$ points in k-dimensional space one can always find n of them so that all their distances are distinct. It seems quite difficult to determine $A\,(k\,, n)$ even for $k = 1$, only crude upper bounds are known for the general case. $A\,(2\,, 3) = 7$ and Croft proved $A\,(3\,, 3) = 9$ [4].

Ramsey's theorem easily implies that to every $\epsilon > 0$ and n there is a $B\,(\epsilon\,, n)$, so that if there are $B\,(\epsilon\,, n)$ points in the plane then there are always n of them $x_1, \ldots, x_n$ which determine a convex polygon for which the angle $(\chi_i\,, \chi_j\,, \chi_r)$ is greater than $\pi - \epsilon$ (for every $1 \leqslant i < j < r \leqslant n$).

A well known theorem of Schur states that if we split the integers not exceeding $e\,n\,!$ into n classes then the equation $x + y = z$ is satisfied in at least one of the classes. V.T. Sós communicated to me the following simple proof of Schur's theorem : Consider the partition of pairs $(i\,, j)$ so that the pair $(i\,, j)$, $i < j$, belongs to the r-th class if $j - i$ belongs to the r-th class. By Ramsey's theorem at least one of the new classes contains a triangle $(i\,, j\,, l)$, but then

$$(j - i) + (l - j) = l - i,$$

or $x + y = z$ is solvable in the original class. It is known that

$$R_3^{(n)}\,(3, \ldots, 3) \leqslant [\,e\,n\,!\,]$$

which completes the proof of Schur's theorem.

The determination of the exact bound in Schur's theorem is a very difficult problem, probably $e\,n\,!$ can be replaced by c^n. The value of $f_2^{(r)}(3, \ldots, 3)$ is not known for $r > 3$. [5].

It seems that the following theorem of J. Sanders can not be proved so simply : To every r and n there is an $f_r(n)$ so that if we split the integers from 1 to $f_r(n)$ into r classes there are n distinct integers $a_1 < \ldots < a_n$ so that all the $2^n - 1$ sums

$$\sum_{i=1}^{n} \epsilon_i\, a_i, \qquad \epsilon_i = 0 \text{ or } 1, \text{ not all } \epsilon_i = 0$$

are in the same class. Rados results [10] imply the theorem of Sanders.

Graham and Rotschild [6] have a very general theorem from which this follows as a special case. They have the following very interesting problem : Split the integers into two (or more generally into r) classes. Is it true that there is an infinite sequences $a_1 < \ldots$ so that all the sums

$$(3) \qquad \sum_{i} \epsilon_i\, a_i \qquad \epsilon_i = 0 \text{ or } 1, \text{ not all } \epsilon_i = 0$$

are in the same class ? (in (3) of course only a finite number of ϵ's are 1). It is not even known if there is an infinite sequence $a_1 < \ldots$ for which $a_1 < a_2 < \ldots$ and $a_i + a_j$, $1 \leqslant i < j < \infty$ all belong to the same class.

On the other hand it immediately follows from Ramsey's theorem that for every l there is an infinite subsequence $a_1^{(l)} < \ldots$ so that all distinct sums taken l at a time belong to the same class. I do not know if there is an infinite subsequence $a_{i_1} < \ldots$ so that for every t, $(t = 1, 2, \ldots)$, all distinct sums taken t at a time belong to the same class, the class may depend on t.

It is easy to see that every infinite sequence of integers contains an infinite subsequence so that either no two members of the subsequence divide each other or each term of the subsequence divides the subsequent one. This follows immediately from Ramsey's theorem but perhaps is not a good example of its use since the direct proof is easier.

Császár [7] proved the following theorem which arose in his joint work with Czipszer : Let $g_1(x) \ldots, g_n(x)$ be n bounded real functions and $f(x)$ another real function. Assume that there are two real numbers $\epsilon > 0$, $\delta > 0$ so that whenever $f(x) - f(y) > \epsilon$ there is an i, $1 \leqslant i \leqslant n$ so that $g_i(x) - g_i(y) > \delta$. Then $f(x)$ is also bounded. Császár gave a direct proof of this theorem and V.T. Sós observed that it immediately follows from Ramsey's theorem.

A theorem of Van der Waerden states that if we split the integers into two classes at least one of them contains an arbitrarily long arithmetic progression. The finite form of Van der Waerden's theorem states that there is a smallest $f(n)$ so that if we split the integers from 1 to $f(n)$ into two classes at least one of them contains an arithmetic progression of n terms. No satisfactory upper bound is know for $f(n)$, the best lower bound is due to Berlekamp [8].

Van der Waerden's theorem also has many applications e.g. A Brauer [9] proved that if $p > p_0(k)$ is a sufficiently large prime then there are k consecutive quadratic residues and non-residues mod p. Rado [10] gives many interesting generalisations and applications to new number theoretic and combinatorial problems. The theorem of Graham and Rotschild [6] can be considered as a generalisation of Van der Waerdens theorem. Finally I would like to drow your attention to a beautiful conjecture of Rota [6] which seems very deep. Added in proof : Rotas conjecture has been proved by Graham, Leeb and Rotschild.

Turán and I [11] raised the following problem in combinatorial number theory : Denote by $r_k(n)$ the maximum value of l for which there exists a sequence of integers $a_1 < \ldots < a_l \leqslant n$ which do not contain an arithmetic progression of k terms. Determine or estimate $r_k(n)$. If we could prove that for every k there is an $n_0(k)$ so that for $n > n_0(k)$ $r_k(n) < n/2$, then Van der Warden's theorem would immediately follow. Unfortunately this has never been proved.

It is known that

$$n^{1-c_1/\sqrt{\log n}} < r_3(n) < c_2 n/\log\log n$$

The lower bound is due to Behrend [12] and the upper to Roth [13]. Szemerédi. [14] proved $r_4(n) = \sigma(n)$.

I would like to mention one more old conjecture of mine from combinatorial number theory : Let $g(n) = \pm 1$ be an arbitrary function. Then to every c there is a d and m so that

$$\left| \sum_{k=1}^{m} g(kd) \right| > c$$

Now we give some applications of combinatoral inequalities and extremal problems. Let $a_1 < \ldots < a_k \leqslant n$ be a sequence of integers no a divides any other. Then it is easy to see that max $k = [(n+1)/2]$. On the other hand if we assume that no a divides the product of two others then [15]

$$(4) \qquad \pi(n) + c_1 n^{2/3} < \max k < \pi(n) + c_2 n^{2/3}/(\log n)^2$$

The proof of both the upper and the lower bound in (4) uses combinatorial results. The lower bound uses Steiner triplets and the upper bound the trivial result that a graph of n vertices and n edges contains a circuit. Assume now that all the products $a_i a_j$ are distinct. Then [16]

$$(5) \qquad \pi(n) + c_3 n^{3/4}/(\log n)^{3/2} < \max k < \pi(n) + c_4 n^{3/4}/(\log n)^{3/2}$$

Here the lower bound uses the existence of finite geometries for $n = p^2 + p + 1$ and the upper bound uses the following result : Let $\mathfrak{G}$ be a graph of t_1 vertices which contains no rectangle, further assume that there are t_2 vertices so that every edge of our graph is incident to one of these vertices. Then the number $R(\mathfrak{G})$ of edges of $\mathfrak{G}$ is less than

$$t_1 + t_1\left[\frac{t_2}{t_1^{1/2}}\right] + t_2^2\left(1 + \left[\frac{t_2}{t_1^{1/2}}\right]\right)^{-1}$$

Thus, in particular, if $\mathcal{G}$ has n vertices and contains no rectangle then $R(\mathcal{G}) < cn^{3/2}$. W. Brown and Rényi, V.T. Sós and I proved that [17] if $\mathcal{G}$ contains no rectangle then

$$\max c(\mathcal{G}) = \left(\frac{1}{2} + \sigma(1)\right) n^{3/2}.$$

Assume now that the number of solutions of $a_i a_j = m$ is bounded. Then we have the following result : Let $a_1 < \dots < a_k \leqslant n$, $n > n_0(\epsilon , l)$. Assume

$$(6) \qquad k > (1 + \epsilon)\, n\, (\log\log n)^{l-1}/(l - 1)\ !\ \log n$$

Then for some m the number of solutions of $m = a_i a_j$ is $\geqslant 2^l$. (6) is best possible in the sense that it fails if $1 + \epsilon$ is replaced by $1 - \epsilon$ [18].

(6) implies that if $a_1 < \dots$ is an infinite sequence of integers so that every large integer can be written in the form $a_i a_j$ then the number of solutions of $n = a_i a_j$ is unbounded. Now I state an old conjecture of Turán and myself which is an additive analogue of this result (and which in fact lead me to this result) : Let $a_1 < \dots$ be an infinite sequence of integers, denote by $f(n)$ the number of solutions of $n = a_i + a_j$. Assume that $f(n) > 0$ for all sufficiently large n. Then $\limsup f(n) = \infty$. This conjecture if true is probably very deep.

The combinatorial theorem needed for the proof of (6) states as follows. Let r and t be given, $\epsilon = \epsilon(r , t)$ small and $n > n_0(\epsilon, r , t)$. Let

$$|\mathfrak{S}| = n \qquad \text{and} \qquad A_1, \dots, A_u, u > n^{r-\epsilon}$$

are r-tuples contained in $\mathfrak{S}$. Then there are rt distinct elements $\chi_i^{(j)}$, $i = 1, \dots, t$; $j = 1, \dots, r$ of $\mathcal{G}$ so that all the t^r sets $(\chi_{i_1}^{(1)}, \chi_{i_2}^{(2)}, \dots \chi_{i_r}^{(r)})$ occur among the A's. For $r = 2$ this is a theorem of Kóvári and the Turáns [19].

A well known theorem of Turán [20] states that in a graph of n vertices which has more than

$$(7) \quad f(n , r) = \frac{r-2}{2(r-1)}(n^2 - h^2) + \binom{h}{2}, n \equiv h (\mathrm{mod}\ r - 1) , 0 \leqslant h < r - 1$$

edges there is a complete r-gon. This theorem and its extensions have many applications in geometry and potential theory. A. Meir, the Turáns and I are publishing a series of joint papers on this subject. Here I state only one such application. Let $\mathfrak{S}$ be a set of diameter 1 and $\chi_1, \dots, \chi_n$, n points in $\mathfrak{S}$. Let the packing constants of $\mathfrak{S}$ be $p_2 = 1 \geqslant p_3 \geqslant \dots$. Then at most $f(n , r)$ of the distances $d(x_1 , x_j)$ are greater than p_r. Let C_r be the r-th covering constant of $\mathfrak{S}$, using a theorem of Moser and myself (Australien J. Math. XI (1970), 92-97), V.T. Sós obtained a lower bound for the number of distances $d(x_1 , x_j) > C_r$. This will also appear in the quadruple paper.

Another application of (7) is due to Katona [21]. Let $f_1(x), \dots, f_n(x)$ be n functions which satisfy $\int f_i(X)^2\, dx \geqslant 1$. Then there are at most $[n^2/4]$ pairs (i , j), $i < j$ for which $\int (f_i(X) + f_j(X))^2\, dx < 1$.

I investigated the following question : Let $\chi_1, \dots, \chi_n$ be n distinct points in k-dimensional Euclidean space. For how many pairs i , j $(i < j)$ can we have

$d(x_i, x_j) = 1$? Denote this maximum by $f(k, n)$. For $k = 2$ and $k = 3$ I have no good estimations for $f(k, n)$, e.g. for $k = 2$ I only know that

$$n^{1+c/\log\log n} < f(2, n) < c'n^{3/2} \tag{8}$$

It seems that in (8) the lower bound is close to being best possible.

For $k \geqslant 4$ one knows very much more. Lenz and I proved [23]

$$\lim_{n=\infty} f(k, n)/n^2 = \frac{1}{2}\left(1 - \frac{1}{[k/2]}\right)$$

and if $k = 4$, $n \equiv 0 \pmod 8$ I proved [23]

$$f(k, n) = \frac{n^2}{4} + n. \tag{9}$$

(9) follows from the following result of Simonovits and myself wich will soon appear in Acta Hungarica : Denote by $\mathfrak{G}(n; l)$ a graph of n vertices and l edges. $k(u_1, \ldots, u_r)$ denotes the complete r-chromatic graph where there are u_i vertices of the i-th colour and every two vertices of different colour are joined. We proved that every $\mathfrak{G}\left(n ; \left[\frac{n^2}{4}\right] + n + 1\right)$ contains a $k(1, 3, 3)$. This result is best possible, there is a $\mathfrak{G}\left(n \ \left[\frac{n^2}{4}\right] + n\right)$ which does not contain a $k(1, 3, 3)$.

A well known theorem of Sperner [24] states that if $|\mathfrak{S}| = n, A_i \subset \mathfrak{S}$, $1 \leqslant i \leqslant k$ is a family of subsets, no one of which contains any other, then

$$\max k = \binom{n}{[n/2]}, \tag{10}$$

This result and its generalisations and extensions has many applications. Using (10) Behrend [25] proved that if $a_1 < \ldots < a_k \leqslant n$ is a primitive sequence (a sequence of integers is called primitive if no a divides any other) then

$$\sum_{i=1}^{k} \frac{1}{a_i} < c \log n/(\log\log n)^{1/2}. \tag{11}$$

and Pillai proved that (11) is best possible.

Using a more complicated refinement of (10) Sárközi, Szemerédi and I [26] proved that if $a_1 < \ldots$ is an infinite primitive sequence then

$$\lim \sum_{a_i < x} \frac{1}{a_i} \left(\frac{(\log\log x)^{1/2}}{\log x}\right)^{-1} = 0.$$

We also proved that

$$\max \sum_{a_i < x} \frac{1}{a_i} = (1 + \sigma(1)) \frac{\log x}{(2\pi \log\log x)^{1/2}}$$

where the maximum is taken over all primitive sequences [27].

I made strong use of Sperners theorem in my papers on the distribution function of additive arithmetic functions [28].

Let $|\mathfrak{S}| = n$, $A_i \subset \mathfrak{S}$ $1 \leqslant i \leqslant k$. Assume that the union of two A's never equals a third. I conjectured that then $k < c\binom{n}{[n/2]}$. Kleitman [29] proved this conjecture as well as several other related conjectures, all of which have number theoretic applications [30].

Sharpening a result of Littlewood and Offord [31] I immediately deduced from (10) that if $\chi_i \geqslant 1$, $i = 1, \ldots, n$ then the number of sums

$$\sum_{i=1}^{n} \epsilon_i \chi_i, \epsilon_i = \pm 1$$

which fall inside an interval of length 2 is at most $\binom{n}{[n/2]}$. I conjectured that the same holds if the x_1 are vectors in a Banach space (the interval of length 2 has to be replaced by a sphere of radius 1). This was first proved for the plane independently by Katona and Kleitman [32] using an ingenious extension of (10). Very recently Kleitman proved my general conjecture in a surprisingly simple way without using Sperners theorem. Kleitman's proof is not yet published.

Rado and I proved the following theorem. Let $a \geqslant 2$ and $b > 1$ be integers. Then there is a smallest integer $f(a,b)$ so that if we have $f(a,b) + 1$ sets each having at most b elements there are always $a + 1$ of them which have pairwise the same intersection [33]. We proved

$$(12) \qquad f(a,b) \leqslant b\,!\,a^{b=1}\left(1 - \frac{1}{2\,!\,a} - \frac{2}{3\,!\,a^2} - \ldots - \frac{b-1}{b\,!\,a^{b-1}}\right)$$

(12) is far from being best possible and very likely

$$(13) \qquad f(a,b) < c^{b+1}\, a^{b+1}$$

We are very far from being able to prove (13), even for $a = 2$ (12) has not even been proved with $o(b!)$.

Sauer determined $f(a, 2)$ for every a. For $b > 2$ there are only relatively crude upper and lower bounds for $f(a,b)$.

(12) has many applications which could be significantly strengthened if (13) would be proved.

Denote by $f_t(n)$ the smallest integer so that if

$$1 \leqslant a_1 < \ldots < a_l \leqslant n, l = f_t(n)$$

is an arbitrary sequence of integers, one can always find t a's which have pairwise the same greatest common divisor. First I proved by number theoretical methods that

$$f_t(n) < \frac{n}{\exp((\log n)^{1/2-\epsilon})}$$

Later I observed that (12) implies that for every t and $\epsilon > 0$ there is an n_0 so that for all $n > n_0(t, \epsilon)$

$$\exp(c_t \log n/\log\log n) < f_t(n) < n^{3/4+\epsilon}$$

(13) would imply that the lower bound gives the right order of magnitude,

Using (12) I proved [35] that for every k there are squarefree integers satisfying ($V(n)$ denotes the number of distinct prime factors of n)

$$(a_i, a_j) = 1, \varphi(a_i) = \varphi(a_j), \quad V(a_i) = V(a_j), \quad 1 \leqslant i < j \leqslant k$$

If (13) would hold we could add $\sigma(a_i) = \sigma(a_j)$.

(12) has been improved by Abbot and others but as far as I know nobody came close to (13).

Dodson [36] investigated the following problem : Denote by $\Gamma^*(k, p^n)$ the smallest value of s for which for every choice of the integers $a_1, \ldots, a_s$ $\sum_{i=1}^{s} a_i x_i^k \equiv 0 \pmod{p^n}$ has a non trivial solution in integers x_i, $i = 1, \ldots, s$) (i.e not all the x_i are multiples of p). In one of the cases (12) was needed.

(12) has also many applications to combinatorial problems and set theory (see Engelking and others).

Before completing the paper I want to state a few miscellaneous combinatorial results which have applications in various branches of mathematics.

Let $|\mathfrak{S}| = n$, $A_i \subset \mathfrak{S}$, $1 \leqslant i \leqslant r$, $r \to \infty$ as $n \to \infty$, $|A_i| > cn$, $0 < c < 1$ for

$$1 \leqslant i \leqslant r, r \to \infty` \text{ as } \quad n \to \infty$$

Then there are two indices i and j for which $|A_i \cap A_j| > (c^2 + o(1))n$. This statement can be proved easily by using the characteristic functions of the sets A_i and it is easy to state various generalisations for the intersection of more than two sets, one can also reformulate the result for measurable sets [37]. This theorem has many applications to combinatorial analysis, number theory and analysis.

A theorem of Szekeres and myself states that if there are given 2^n points in the plane they always determine an angle $> \pi\left(1 - \frac{1}{n}\right)$. This result is best possible since Szekeres showed previously that to every $\epsilon > 0$ one can give 2^n points in the plane so that all the angles are $< \pi\left(1 - \frac{1}{n}\right) + \epsilon$; see [2].

The fact that 2^n points determine an angle $\geqslant \pi\left(1 - \frac{1}{n}\right)$ follows from the fact that the complete graph of $2^n + 1$ points is not the union of n bipartite graphs. The sharper result that one of the angles is $> \pi\left(1 - \frac{1}{n}\right)$ follows from a more careful study of the decompositions of the complete graph of 2^n vertices into n bipartite graphs.

Probability methods have often been applied successfully to solve combinatorial problems which seemed intractable by more direct methods and conversely combinatorial results often imply unexpectedly beautiful results in probability. e.g. the arc sine law of Andersen [38], see also my paper with Hunt [39]. Finally I want to mention that Davies and Rogers [40] rediscovered ans used a little known theorem of Hajnal and myself on chromatic graphs in the study of Hausdorff dimension of sets.

REFERENCES

[1] Erdös P and Szekeres G. — A combinatorial problem in geometry, *Compositio Math.* 2, 1935, p. 463-470.

[2] Erdös P. and Szekeres G. — On some extremum problems in elementary geometry, *Annales Univ. Sci. Budapest,* 3-4, 1960, p. 53-62.

[3] Erdös P., Hajnal A. and Rado R. — *Partition relations for cardinal numbers,* 16, 1965, p. 93-196.

[4] Croft H.T. — 9-point and 7 point configurations in three space, *Proc. London Math. Sci.* XII, 1962, p. 400-424.

[5] Greenwood R.E. and Gleason A.M. — Combinatoxial theorems and chromatic graphs, *Canad. J. Math.* 7, 1955, p. 1-7.

[6] Graham R.L. and Rotschild B.L. — Ramsey's theorem for *n-parameter* sets, *Colloquium, Comb. theory and applications* Balatonfüred, Hungary 1969.

[7] Csaszar A. — Sur une critère d'approximation uniforme, *Publ. Mat. Inst. Hung. Acad.* 8, 1963, p. 413-416.

[8] Berlekamp E.R. — A construction for partitions which avoid long arithmetic progressions. *Canad. Bull. Math.* 11, 1968, p. 409-414.

[9] Brauer A. — Über Sequenzen von Potenzresten I und II, *Sitrungsber. Preuss. Akad. Wiss. Phys. Math. Klass* 19, 1931, p. 329-341.

[10] Rado R. — Studien zur Kombinatorik, *Math. Zeitschrift* 1933, p. 424-480.

[11] Erdös P. and Turan P. — On some sequences of integers. *J. London Math. Soc.,* 11, 1936, p. 261-264.

[12] Behrend F.A. — On sets of integers which contain no three terms in arithmetical progression, *Proc. Mat. Acad. Sci., U.S.A.,* 32, 1946, p. 331-332.

[13] Roth K.F. — On certain sets of integers, *J. London Math. Soc.,* 28, 1953, p. 104-109.

[14] Szemerédi E. — On sets of integers containing no four elements in arithmetic progressions, *Acta Math. Acad. Sci. Hung.,* 20, 1969, p. 89-104.

[15] Erdös P. — On sequences of integers no one of which divides the product of two others and on some related problems, *Tomsk. Gos. Univ. Vcen Zap.,* 2, 1938, p. 74-82.

[16] Erdös P. — On some applications of graph theory to number theoretic problems, *Publ. Ramanujan Inst.* No. 1, p. 132-136.

[17] Brown W.G. — On graphs that do not contain a Thomsen graph, *Canad. Math. Bull.,* 9, 1966, p. 281-285, Erdos P., Rényi A. and Sos V.T., On a problem of graph theory, *Studia Sci. Math. Hung,* 1, 1966, p. 215-235.

[18] Erdös P. — On the multiplicative representation of integers, *Israel J. Math.,* 2, 1964, p. 251-261.

[19] Erdős P. — On extremal problems of graphs and generalised graphs, *Israel J. Math.*, 2, 1964, p. 183-190. Kovari T., Sos V.T. and Turan P. — On a problem of K. Zarankievicz, *Colloquium Math.*, 3, 1950, p. 50-57.

[20] Turan P. — Eine Extremalaufgabe aus der Graphentheorie (in Hungarian), *Mat. és Fiz. Lapok,* 48, 1941, p. 436-462.

[21] Katona Gy. — Graphs, vectors and inequalities in probability theory (in Hungarian), *Mat. Lapok,* 20, 1969, p. 123-127.

[22] Erdős P. — On sets of distances of n points, *Amer. Math. Monthly,* 53, 1946 p. 258-260.

[23] Erdős P. — On sets of distances of n points in an Euclidean space, *Publ. Math. Inst. Hung. Acad.,* 5, 1960, p. 165-169; Erdos P., On some applications of graph theory to geometry, *Canadian J. Math.,* 19, 1967, p. 968-971.

[24] Sperner E. — Ein Satz über Untermengen einer endlichen Menge, *Math. Zeitschrift,* 27, 1928, p. 544-548.

[25] Behrend F. — On sequences of numbers not divisible one by another, *J. London Math. Soc.,* 10, 1935, p. 42-49, see also chapter V of the book *Sequences* of H. Halberstam and K.F. Roth.

[26] Erdős P., Sarkozi A. and Szemerédi E. — On a theorem of Behrend, *J. Australian Math. Soc.* 7, 1967, p. 9-16.

[27] Erdős P., Sarkozi A. and Szemerédi E. — On an extremal problem concerning primitive sequences, *J. London Math. Soc.,* 42, 1967, p. 484-488.

[28] Erdős P. — On the density of some sequences of numbers I, II and III. *J. London Math. Soc.,* 11, 1935, p. 120-125; 13, 1937, p. 7-11; 14, 1938, p. 185-192.

[29] Kleitman D. — On a combinatorial problem of Erdos, *Proc. Amer. Math. Soc.,* 17, 1966, p. 139-141.

[30] Erdős P., Sarkozi A. and Szemerédi E. — On the solvability of the equations $[a_i, a_j] = a_r$ and $(a_i . a_j) = a_r$ in sequences of positive density, *J. Math. Anal. and Applications,* 15, 1966, p. 60-64.

[31] Erdős P. — On a Lemma of Littlewood and Offord, *Bull. Amer. Math. Soc.,* 31, 1945, p. 898-902.

[32] Katona G. — On a conjoncture of Erdos and a stronger form of spencer's theorem, *Studia Sci. Math. Hungar.,* 1, 1966, p. 59-63, Kleitman D. — On a lemma of Littlewood and Offord on the distribution of certain sums, *Math. Z.,* 90, 1965, p. 251-259.

[33] Erdős P. and Rado R. — Intersection theorems for systems of sets, *J. London Math. Soc.,* 35, 1960, p. 85-90.

[34] Erdős P. — On a problem in elementary number theory and a combinatorial problem, *Math. of Computation,* 18, 1964, p. 644-646.

[35] Erdős P. — Some remarks on the functions φ and σ, *Bull. Acad. Polon. Sci.,* X, 10, 1962, p. 617-619.

[36] Dodson M. — Homogeneous additive congruences, *Philos. Trans. Roy. Soc. London,* Ser. A, 261, 1967.

[37] Erdős P., Neveu J. and Rényi A. — An elementary inequality between the probabilities of events, *Math. Scand.,* 13, 1963, p. 99-104. See also papers by Gillis, Lorentz, Sucheston, Shapiro H.S.

[38] Andersen E.S. — On the number of positive sums of random variables, *Skandinavisk Aktuarietidsskrift,* 32, 1949, p. 27-36.

[39] Erdős P. and Hunt G.A. — Changes of sign of sums of random variables, *Pacific J. Math.*, 3, 1953, p. 673-687.

Mathematical Institute Hungarian Academy of Sciences
Realtanoda U. 13-15
Budapest V (Hongrie)

Actes, Congrès intern. Math., 1970. Tome 3, p. 211 à 215.

WEIGHT POLYNOMIALS OF SELF-DUAL CODES AND THE MacWILLIAMS IDENTITIES

by Andrew M. GLEASON

To Marshall Hall, Jr., on his sixtieth birthday.

Abstract

Many error correcting codes are known to be self-dual. Hence the MacWilliams identities put a considerable restriction on the possible weight distribution of such a code. We show that this restriction, for codes over GF(2) and GF(3), is that the weight polynomial must lie in an explicitly described free polynomial ring. To extend these results (in part) to self-dual codes over larger fields, we introduce more general weight polynomials and extend the MacWilliams identities to these.

1. Let F be a finite field with q elements and let F^n be the direct product of F with itself n times regarded as a vector space as usual. A k-dimensional linear subspace A of F^n is sometimes called an (n, k)-code. If $\mathbf{v} = < v_1, \ldots, v_n > \epsilon\ F^n$, then the weight of $\mathbf{v}$ is the number of indices i for which $v_i \neq 0$.

Suppose A is an (n, k)-code. For each $i = 0, \ldots, n$ let w_i be the number of vectors in A having weight i. Then the homogeneous polynomial $W_A \in \mathbf{C}[S, T]$ given by

$$W_A = \Sigma\, w_i S^{n-i} T^i$$

is called the *weight polynomial* of A.

Let (,) denote the usual inner product $F^n \times F^n \to F$. If A is a linear subspace of F^n, its *dual* is

$$B = \{\mathbf{v} \in F^n : (\forall \mathbf{x} \in A)\ (\mathbf{v}, \mathbf{x}) = 0\}.$$

The MacWilliams identities [2] connect the weight polynomial W_A of A with the weight polynomial of W_B of the dual B. They can be expressed in the single equation

$$(1) \qquad q^{\dim B}\, W_A = W_B\,(S + (q-1)\,T, S - T),$$

where the right hand side means the result of replacing S and T in W_B by $S + (q-1)\,T$ and $S - T$ respectively.

It may happen that A is identical with its dual B ; that is, A may be self-dual. This clearly implies that $n = 2k$. If n is even and $q \not\equiv -1 \pmod 4$ then F^n always contains self-dual subspaces. If $q \equiv -1 \pmod 4$, then self-dual subspaces appear if and only if n is divisible by 4.

When a code A is self-dual, the MacWilliams identity (1) becomes a non-trivial invariance property of W_A : namely, W_A is invariant under the linear substitution whose matrix is

$$(2) \qquad \frac{1}{\sqrt{q}}\begin{pmatrix} 1 & 1 \\ q-1 & -1 \end{pmatrix} .$$

Each vector in a self-dual code A is orthogonal to itself. When $q = 2$ this means that its weight is even. Hence W_A is also invariant under the substitution $T \to -T$. This substitution and that given by (2) for $q = 2$ generate the dihedral group D_8 with 16 elements and the ring of all polynomials invariant under this group is found to be a free polynomial ring generated by

$$S^2 + T^2 \quad \text{and} \quad S^2 T^2 (S^2 - T^2)^2$$

When A is a self-dual code over the two-element field, those vectors in A whose weights are divisible by 4 form a linear subspace of codimension at most 1. Hence either all the vectors in A have weight divisible by 4 or exactly half do. In the latter case, the weight polynomial must be divisible by $S^2 + T^2$. In the former, W_A is invariant under the substitution $T \to iT$. This substitution and that given by (2) with $q = 2$ generate a group of 192 elements. The ring of invariant polynomials is again a free polynomial ring. The generators may be taken to be $S^8 + 14S^4 T^4 + T^8$ and $S^4 T^4 (S^4 - T^4)^4$.

When F has three elements, that is $q = 3$, every vector in a self-dual code must have weight divisible by 3. Hence the weight polynomial will be invariant under the substitution $T \to \omega T$ where ω is a cube root of unity. This substitution and (2) for $q = 3$ generate a group with 48 elements and the ring of invariant polynomials is free with generators of degrees 4 and 12. We can take the generators to be $P = S^4 + 8ST^3$ and $Q = T^3(S^3 - T^3)^3$. Here P is the weight polynomial of the (essentially unique) self-dual (4,2)-code. The weight polynomial of the Golay (12,6) code is $P^3 - 24\,Q$.

To illustrate the application of these results we continue the study of self-dual codes over $GF(3)$. According to a theorem of Assmus and Mattson [1] a self-dual $(12t, 6t)$ code will produce 5-designs whenever its minimal weight exceeds $3t$. Pless [3] has recently given a construction of a self-dual $(12t, 6t)$ code whenever $6t - 1$ is a prime power. She showed that for $t = 1, 2, 3, 4, 5$ the minimal weight is indeed $3t + 3$. The weight polynomial for such a code must be a linear combination of P^{3t} ,$P^{3t-3} Q, \dots, P^3 Q^{t-1}$, and Q^t. Since these polynomials begin with successively higher powers of T^3, the weight polynomial is determined by its first $t + 1$ coefficients, namely $w_0 = 1, w_3, \dots, w_{3t}$. If the minimal weight is at least $3t + 3$, then all these w's are known and the weight polynomial can be directly computed. Peirce has computed these polynomials up to $t = 8$ on the assumption that the minimal weight exceeds $3t$. This gives the complete weight distribution for the first five Pless codes. For $t = 6$ and 8, one of the coefficients of the putative weight polynomial turns out to be negative. It follows that there is no self-dual (72, 36)-code with minimal weight exceeding 18 and no self-dual (96, 48)-code with minimal weight exceeding 24. It is not yet known whether or not an (84, 42)-code with minimal weight 24 or more exists.

2. Generalized weight polynomials and the MacWilliams identities.

When calculuating the weight of a vector, all non-zero components are lumped together. We introduce more refined weights and weight polynomials which distinguish all the different components.

Let G be a finite abelian group written additively. For each element $g \in G$, introduce an indeterminate X_g and consider the ring $\mathbf{C}[\{X_g : g \in G\}]$. Let G^n denote the direct sum of G with itself n times and consider the function $\varphi : G^n \to \mathbf{C}[\{X_g\}]$ given by

$$\varphi(g) = X_{g_1} \cdot X_{g_2} \cdot \ldots \cdot X_{g_n}$$

If A is any subset of G^n, the *generalized weight polynomial* of A is

$$V_A = \sum_{\mathbf{g} \in A} \varphi(\mathbf{g})$$

Let H be the dual group of G, that is, the set of all homomorphisms of G into $\mathbf{C}$. We identify H^n with the dual of G^n by defining

$$\mathbf{h}(\mathbf{g}) = h_1(g_1) h_2(g_2) \ldots h_n(g_n)$$

where $\mathbf{h} \in H^n$ and $\mathbf{g} \in G^n$.

Corresponding to elements of H we introduce indeterminates Y_h and define $\psi : H^n \to \mathbf{C}[\{Y_h\}]$ by

$$\psi(\mathbf{h}) = Y_{h_1}, Y_{h_2} \ldots Y_{h_n}$$

Then we define the generalized weight polynomial of a subset B of H^n to be

$$V_B = \sum_{\mathbf{h} \in A} \psi(\mathbf{h}).$$

THEOREM 1. – *The Fourier transform $\hat{\varphi}$ of φ is obtained from ψ by the substitution*

$$(3) \qquad Y_h \to \sum_{g \in G} h(g) X_g$$

If A is a subgroup of G^n and A^0 is its annihilator, the Poisson summation formula asserts that

$$\sum_{\mathbf{g} \in A} \varphi(g) = \frac{1}{|A^0|} \sum_{\mathbf{h} \in A^0} \hat{\varphi}(h)$$

Combining this with Theorem 1 we obtain.

THEOREM 2. – The generalized weight polynomial V_A for any subgroup A of G^n is obtained from the generalized weight polynomial V_{A^0} of its dual by the substitution (3) and division by $|A^0|$.

We are particularly concerned with the possibility that G is the additive group F^+ of a finite field F. Then G^n becomes the additive group of the vector space F^n.

Let χ be a fixed non-trivial character of F^+. We can identify F^n with the dual group of F^n by making the vector u correspond to the homomorphism

$$< g_1, g_2, \ldots g_n > \rightarrow \chi(\Sigma g_i u_i)$$

The dual B of a linear subspace A of F^n is then identified with the annihilator A^0. When these identifications are made, both a code A and its dual B have generalized weight polynomials in the ring $C[\{X_a : a \in F\}]$. Theorem 2 becomes the generalized MacWilliams identity

$$(4) \qquad q^{\dim B} V_A = j(V_B)$$

where j denotes the homomorphism $X_a \rightarrow \Sigma \chi(\mathrm{ba}) X_b$ of $C[\{X_a\}]$ into itself. The original MacWilliams identity can now be obtained by the substitution : $X_0 \rightarrow S$ and $X_a \rightarrow T$ for $a \neq 0$.

The weight polynomial for any linear subspace of F^n must be invariant under the substitution $X_a \rightarrow X_{\lambda a}$ for any fixed $\lambda \in F$. For any self-dual code (4) can be normalized by dividing by $q^{n/2}$. Allowing for the fact that when $q \equiv -1 \pmod 4$, n must be divisible by 4, we see that the weight polynomial for a self-dual code is invariant under the substitution whose matrix is

$$\frac{1}{\sqrt{\epsilon q}} (\chi(ij))$$

where $\epsilon = -1$ if $q \equiv -1 \pmod 4$ and $\epsilon = +1$ if $q \equiv +1 \pmod 4$. (For brevity we suppress the possibility q even, which is exceptional). Here the letters i and j refer to rows and columns, respectively, and range over the elements of the field F. Every vector in a self-dual code is orthogonal to itself and therefore the weight polynomial is invariant under the substitution

$$X_i \rightarrow \chi(ai^2) X_i$$

for any $a \in F$. All these substitutions together generate a group with $2q(q^2 - 1)$ elements which turns out to be a double-covering of the group $SL_2(q)$. (Recall that here q is odd. For q even, the group is solvable). The ring of all invariant polynomials is finitely generated, but for large values of q it is not a free polynomial ring.

These substitutions can be regarded as a group of linear operators on the space of all functions from F to C. There are two invariant subspaces, the space of even functions ($f(x) = f(-x)$), and the space of odd functions.

REFERENCES

[1] Assmus Jr. E.F. and Mattson Jr. H.F. — New 5-designs, *Jour. of Combinatorial Theory*, vol. 6, 1969, p. 122-151.

[2] MacWilliams F.J. — A theorem on the distribution of weights in a systematic code. *Bell System Tech. J.*, vol. 42, 1963, p. 79-84, MR 26 1963, 7462.

[3] PLESS V. — On a new family ot symmetry codes and related new five-designs, *Bull. Amer. Math. Soc.*, vol. 75, 1969, p. 1339-1342, MR 39 (1970) 6763.

Harvard University
Dept. of Mathematics,
2 Divinity Avenue,
Cambridge,
Massachusetts 02 138 (USA)

Actes, Congrès intern. Math., 1970. Tome 3, p. 217 à 222.

COMBINATORIAL DESIGNS AND GROUPS*

by Marshall HALL Jr

1. Introduction

Construction of designs by assuming a certain automorphism group is a powerful method first developed extensively by R.C. Bose. Recently the symmetric designs with $v = 56$, $k = 11$, $\lambda = 2$ and $v = 79$, $k = 13$, $\lambda = 2$ have been constructed by such methods. Examples of this method, including these two are given in section 2.

Section 3 studies groups generated by a class C of elements of order 3 in which any two non-commuting elements generate $SL(2, 3)$. Here the four groups of order 3 in $SL(2, 3)$ may be considered a block of 4 points, and the study of such groups may be related to designs with $k = 4$, $\lambda = 1$.

2. Applications of groups to combinatorial designs.

If designs with given parameters exist then usually there is such a design with a non-trivial group of automorphisms. Assuming the existence of a group of automorphisms may be of great help in the construction of the design, and also simplify the presentation of the design. The simplest case is that in which a design with v elements has a cyclic group of automorphisms of order v which permutes the elements cyclically. In such a case we may identify the elements with the residues modulo v, and let $\alpha : i \to i + 1 \pmod{v}$ be a generator of the automorphism group. The symmetric block design with $v = b = 73$, $r = k = 9$, and $\lambda = 1$ is such a design. Here the elements of one block B_1 may be taken as the set

$$(2.1) \qquad B = \{1, 2, 4, 8, 16, 32, 37, 55, 64\} \pmod{73}$$

and the mapping $i \to i + 1 \pmod{73}$ maps B_1 into all 73 blocks. This design is the projective plane of order 8. The fact that the residues in 2.1) do form a block of such a symmetric design is equivalent to showing that every residue $d \not\equiv 0 \pmod{73}$ is the difference $a_i - a_j$ of two residues a_i, a_j in 2.1) in exactly $\lambda = 1$ ways. Thus we refer to the set B in 2.1) as a difference set. In this example the design also has the further automorphism $u : i \to 2i \pmod{73}$ and we say that 2 is a *multiplier* of the difference set design. It has been proved [8] that a difference set $a_1, \dots, a_k \pmod{v}$ determining a design with parameters $b = v, r = k$, and λ has a multiplier p where p is a prime such that (i) $(p, v) = 1$, (ii) $p \mid k - \lambda$ and (iii) $p > \lambda$. In all known examples condition (iii) is unnecessary, but no proof has been found which does not use some variant of it. Elimination of condition (iii) is a challenging, but difficult problem.

(1) This research was supported in part by ONR Contract N00014-67-A0094-0010.

The example given in (2.1) has a further interesting property. The residues listed are the octic residues of the prime $p = 73$. This relates difference sets to the problems of cyclotomy, and new classes of designs based on this approach have been found by Whiteman [10] and others. It has been noted by the writer [5] that determining cyclotomic constants is equivalent to the calculation of certain group characters.

The 36 points (x, y) where x and y range independently over the residues modulo 6 under addition form a group G of order 36. The 15 points

$$\begin{array}{c} (1,1), (2,2), (3,3), (4,4), (5,5) \\ (0,1), (0,2), (0,3), (0,4), (0,5) \\ (1,0), (2,0), (3,0), (4,0), (5,0) \end{array} \tag{2.2}$$

if taken as a block B_1, together with its images under the action of G form a symmetric design with $v = b = 36$, $r = k = 15$, $\lambda = 6$.

There may be several orbits of points and also several orbits of blocks under the action of the group. This is the essence of R.C. Bose's "method of symmetrically mixed differences" [2]. For example with G the additive group of residues modulo 5, we may take three orbits of length 5 distinguishing the orbits by subscripts, having points i_1, i_2, i_3 with i modulo 5 to give $v = 15$ points. To obtain the Steiner triple system with $v = 15$, $b = 35$, $r = 7$, $k = 3$, $\lambda = 1$, the blocks fall into 7 orbits and representatives of these block orbits form "base blocks" which determine the rest : A set of base blocks is

$$\begin{array}{c} (0_1, 1_2, 4_2), (0_1, 2_2, 3_2), (0_2, 1_3, 4_3), (0_2, 2_3, 3_3), \\ (0_3, 1_1, 4_1), (0_3, 2_1, 3_1), (0_1, 0_2, 0_3). \end{array} \tag{2.3}$$

More recently the relationship between permutation groups and block designs has been studied [6]. If D is a design which has an automorphism group G transitive on the points of D and also ontthe blocks of G, then if we represent G as a permutation group on the points, then if H is the stabilizer of a block B_1, clearly B_1 consists of complete orbits of H. If H is also the stabilizer of a point we call D an orbital design. Any transitive permutation group will yield orbital designs which are partially balanced block designs in the sense of Bose and Shimamoto [3]. A case of particular interest is the study of rank 3 groups. The theory of these groups has been developed by D.G. Higman [9].

The group $G = PSL_3(4)$ is the little projective group of the plane π of order 4, which contains 21 points. An oval in π is a set of 6 points no three on a line, and an oval is determined by any 4 of these. The plane π contains 168 ovals which are permuted by G in three orbits of 56 ovals. Representing G as a permutation group on one set of 56 ovals, G is generated by the permutations

$$\begin{aligned}
a = {} & (1,2,3,4,5,6,7)\,(8,9,10,11,12,13,14)\,(15,16,17,18,19,20,21) \\
& (22,23,24,25,26,27,28)\,(29,30,31,32,33,34,35) \\
& (36,37,38,39,40,41,42)\,(43,44,45,46,47,48,49) \\
& (50,51,52,53,54,55,56) \\
c = {} & (1)\,(2,8,41)\,(3,27,28)\,(4,36,31)\,(5,20,53)\,(6,14,22)\,(7,42,54) \\
& (9,29,34)\,(10,52,17)\,(11,24,46)\,(12,30,48)\,(13,55,33) \\
& (15,26,32)\,(16,21,56)\,(18,40,35)\,(19,23,49)\,(25,50,44) \\
& (37,51,47)\,(38,39,43)\,(45)
\end{aligned} \tag{2.4}$$

G is a rank 3 group in which a stabilizer has orbits of lengths $1, 10, 45$. The letter 1 and the orbit of length 10 in G_1 are

$$B = \{1,12,19,23,30,37,45,47,48,49,51\}. \tag{2.5}$$

The set $B = B_1$ and its images under G form a symmetric block design with $v = 56$, $k = 11$, $\lambda = 2$. This construction [7] was the first for this design.

A symmetric block design with $v = 79$, $k = 13$, $\lambda = 2$ was constructed by Aschbacher [1] with an automorphism group G which is the Frobenius group of order 110. We define G by

$$G = <x\,,y\,,z \mid x^{11} = y^5 = z^2 = 1,\, y^{-1}xy = x^4, z^{-1}xz = x^{-1}, yz = zy> \tag{2.6}$$

We take $B_1\,, B_2\,, B_3\,, B_4$ as base blocks and $P_1\,, P_2\,, P_3\,, P_4$ as base points. Let H_i be the stabilizer of P_i and G_j be the stabilizer of B_j. These are defined by

$$\begin{aligned}
H_1 &= <x\,,y>, H_2 = H_3 = <y\,,z>, H_4 = <z> \\
G_1 &= G_2 = G, G_3 = <y>, G_4 = <z>.
\end{aligned} \tag{2.6}$$

Incidence on the base blocks is defined by

$$\begin{aligned}
B_1 &: \{P_1, P_1 z, P_2 G\} \\
B_2 &: \{P_1, P_1 z, P_3 G\} \\
B_3 &: \{P_1, P_2, P_3, P_4\, xy^i, P_4\, x^4\, y^j\} \\
B_4 &: \{P_2\, x^{\pm 2}, P_3\, x^{\pm 5}, P_4, P_4\, x^{\pm 1}\, y^2, P_4\, x^{\pm 1}\, y, P_4 x^{\pm 5} y, P_4\, x^{\pm 5}\, y^4\}
\end{aligned} \tag{2.8}$$

3. An application of the theory of designs to groups

The Conway group [4] contains a class of elements of order 3 such that any two which do not permute generate $SL(2, 3)$ of order 24, $SL(2, 5)$ of order 120 or their factor groups by a center of order 2 which are respectively A_4 and A_5. It is therefore of interest to determine those groups generated by such a class of elements of order 3. Here the more restricted case is examined in which there is a class of elements of order 3 in which any two elements either commute or generate $SL(2, 3)$ or its factor group A_4.

Defining relations for $SL(2,3)$ are

$$(3.1) \qquad a^3 = b^3 = abab^{-1}a^{-1}b^{-1} = 1.$$

We write $a \sim b$ as an abbreviation for these relations. We note that if $a \sim b$ then $a^{-1}ba = bab^{-1}$ so that a and b are conjugate. But we do *not* have $a^{-1} \sim b$, though we do have $a^{-1} \sim b^{-1}$. Thus in 3.1) there is a distinction between the generators a and a^{-1}, b and b^{-1} of the groups $<a>$ and $<b>$ of order 3.

Taking groups of order 3 such as $<a>$ generated by an element a of our class C as points, then we associate with $SL(2, 3) = <a, b>$ where $a \sim b$ a block of 4 points, namely $<a>, <b>, <a^{-1}ab>, <b^{-1}ab>$, these being the 4 conjugate subgroups of $SL(2, 3)$ and we note that $x \sim y$ where x and y are any two of $a, b, a^{-1}ba$. Thus we have a block of size $k = 4$ and $\lambda = 1$ as any two distinct points determine the block.

For three elements a, b, c of the class C in which $a \sim b$, the possibilities for $G = <a, b, c>$ are

1. $ca = ac, \text{-} cb = bc$. Here $<a, b, c> = <a, b> \times <c>$.

2. $ca = ac$, $c \sim b$. Here $|G| = 648$

and putting $h = a^{-1}c$, G has a normal subgroup $H = <h, b^{-1}hb>$ of exponent 3 and order 27 and $G/H = <a, b>$.

3. $c \sim a$, $c \sim b$, $c \sim a^{-1}ba$, $c \sim b^{-1}ab$.

Here $|G| = 768 = 2^8 \cdot 3$. G has 16 subgroups conjugate to $<a>$, and the center of G contains $(a^{-1}b)^2$, $(a^{-1}c)^2$, and $(b^{-1}c)^2$.

4. $c \sim a$, $c \sim b$, $c \sim a^{-1}ba$, $c^{-1} \sim b^{-1}ab$.

These relations make G collapse so that $G = 1$.

5. $c \sim a$, $c \sim b$, $c^{-1} \sim a^{-1}ba$, $c^{-1} \sim b^{-1}ab$.

Here $|G| = 6048$ and G has 28 groups conjugate to $<a>$.

In case 3 the 16 conjugates of $<a>$ form the block design with $v = 16, b = 20$, $r = 5$, $k = 4$, $\lambda = 1$, the affine plane of order 4. In case 5 the 28 conjugates of $<a>$ form a block design D with $v = 28$, $b = 63$, $r = 9$, $k = 4$, $\lambda = 1$. Here $G = U_3(3)$ the 3 dimensional unitary group over $GF(3^2)$. The points may be identified with the 28 isotropic points in the projective plane over $GF(3^2)$, and these lie in sets of 4 on 63 lines.

With 3 or more generators from C, unless case 5 arises, the elements of C are not conjugate to their inverses and G has a normal subgroup of index 3. We consider $G = <a, b, c, d>$ where $<a, b, c> = U_3(3)$ as in case 5 and d is a further element of C. If $<r>, <s>, <t>, <u>$ are one of the 63 blocks and r, s, t, u are conjugate in $<r, s>$, then if d does not permute with any one of these, we have $<d, r, s>$ a group of type 3 or type 5 above. Take $x_1 = a, x_2, \ldots, x_{28}$ as generators of the 28 groups conjugate to $<a>$ in $U_3(3)$ and choose the generator x_i of $<x_i>$ so that $x_i \sim a$ rather then $x_i^{-1} \sim a$. Then one of the 63 blocks r, s, t, u will consist of four of $x_1 \ldots x_{28}$ and their inverses. If d does not permute with any one of r, s, t, u, then in case 3 we have $d \sim r$, $d \sim s$, $d \sim t$, $d \sim u$ or $d^{-1} \sim r$, $d^{-1} \sim s$, $d^{-1} \sim t$, $d^{-1} \sim u$, and in case 5, $d \sim r$, $d \sim s$, $d^{-1} \sim t$, $d^{-1} \sim u$ or some other combination involving d twice and d^{-1} twice. If we have say d three times and d^{-1} once we are in case 4 and the group collapses. If d permutes with exactly one of r, s, t, u, say $dr = rd$, then we are in case 2 and $d \sim s$, $d \sim t$, $d \sim u$ or $d^{-1} \sim s$, $d^{-1} \sim t$, $d^{-1} \sim u$. Hence we must for each of $x_1, \ldots, x_{28}$, say that $dx_i = x_i d$, or $d \sim x_i$ or $d^{-1} \sim x_i$ so that with d and any one of the 63 blocks r, s, t, u we avoid case 4. This turns out to be a strong restriction. We summarize the results :

First case : d permutes with no one of $x_1, \ldots, x_{28}$. Two essentially different patterns arise. First d may mimic the relation of some x_i to the rest. By conjugation we may take this to be $x_1 = a$ and then $d \sim x_i$, all i. Here $d \sim x_i$, $i = 1, \ldots 28$. But then from case 3 for d with r, s, t, u, we have $(r^{-1}s)^2 d = d(r^{-1}s)^2$. There are enough of these $(r^{-1}s)^2$ so that $<(r^{-1}s)^2 \ldots>$ contains a and we conclude $da = ad$ contrary to assumption. In a second pattern the group $<a, c, d>$ is in case 3 and also $<(r^{-1}s)^2 \ldots>$ centralizing d contains aca. But $daca = acad$ together with $<a, c, d>$ in case 3 collapses to $a = c = d = 1$. Hence no group arises in this first case.

Second case : d permutes with exactly one of $x_1, \ldots, x_{28}$ which we take to be $x_1 = a$. The only permissible pattern is $d \sim x_i$, $i = 2, \ldots 28$. Here putting $d = a^{-1} d$. From computer studies by J. Cannon, the kernel presumably a 3 group, is of order at least 3^{10}.

Third case : d permutes with exactly two generators of $x_1, \ldots, x_{28}$, say $da = ad$, $db = bd$. Then G is of order 117, 573, 120, as calculated by J. Cannon. This must have $U_4(3)$ as a factor group, and has a normal subgroup of order 36 which will be central.

Of course a fourth case is that in which d permutes with all of $x_1, \ldots, x_{28}$ and here $G = <d> \times U_3(3)$.

REFERENCES

[1] Aschbacher M. — Collineation groups of symmetric block designs to appear in *J. Comb. Theory*.

[2] Bose R.C. — On the construction of balanced incomplete block designs *Ann. Eugenics*, 2, 1939, p. 353-399.

[3] Bose R.C. and Shimamoto T. — Classification and analysis of partially balanced incomplete block designs with two associate classes. *J. Amer. Stat. Assn.*, 47, 1952, p. 151-184.

[4] Conway J.H. — A group of order 8, 315, 553, 613, 086, 720, 000. *Bull. London Math. Soc.*, 1, 1969, p. 79-88.

[5] Hall Jr. Marshall. — Characters and cyclotomy. *Proc. Symposia in Pure Math. Amer. Math. Soc.*, 8, 1965, p. 31-43.

[6] Hall Jr. Marshall. — *Designs with transitive automorphism groups*, to appear.

[7] HALL Jr. M., LANE R. and WALES D. — Designs derived from permutation groups. *J. Comb. Theory*, 8, 1970, p. 12-22.

[8] HALL Jr. M. and RYSER H.J. — Cyclic incidence matrices. *Can. J. Math.*, 3, 1951, p. 495-502.

[9] HIGMAN D.G. — Finite permutation groups of rank 3, *Math. Z.*, 86, 1964, p. 145-156.

[10] WHITEMAN A.L. — *A family of difference sets, Ill. J. Math.*, 6, 1962, p. 107-121.

California Institute of Technology
Dept. of Mathematics,
Pasadena,
California 91 109 (USA)

Actes, Congrès intern. Math., 1970. Tome 3, p. 223 à 227.

RECENT DEVELOPMENTS ON COMBINATORIAL DESIGNS.*

by D.K. RAY-CHAUDHURI

(Dedicated to Prof. Marshall Hall, Jr. on his 60th birthday).

Let X be a finite set the elements of which will be called points or treatments. Let $\mathcal{B}$ be a list of subsets of X, i.e., a mapping from $P(X)$, (the set of all subsets of X) to nonnegative integers. A subset B of X with $\mathcal{B}(B) > 0$ is called a block or line. The pair $(X, \mathcal{B})$ is called a design. Let v, λ and t be positive integers and K be a set of positive integers. A design $(X, \mathcal{B})$ is said to be a (v, K, λ, t) – combinatorial design iff (i) $|X| = v$, (ii) $|B| \in K$ for every block B and (iii) for every t-subset X' of X, the number of blocks containing X' is λ, i.e. $\Sigma\, \mathcal{B}(S) = \lambda$ where the sum is taken over all subsets S of X that contain X'. Let k be a positive integer. Combinatorial designs with $K = \{k\}$ are also called tactical configurations. Let $v \geqslant k \geqslant t$. If a (v, k, λ, t) – tactical configuration exists, then

(1) $\lambda \binom{v-i}{t-i} / \binom{k-i}{t-i}$ must be an integer for $i = 0, 1, 2, \dots t-1$.

Very little is known about configurations with $t > 2$. Hanani [6] proved that for $k = 4$ and $t = 3$, the condition (1) is sufficient for the existence of the configurations. Tactical configurations with $t = 2$ are called balanced incomplete block designs (bibd). The condition (1) is not sufficient for the existence of bibd's. Results of Bruck and Ryser, Hall and Connor, Srikhande and Schützenberger establish the existence of infinite families of triples (v, k, λ) which satisfy (1) and for which the corresponding bibd does not exist. Bibd's with $k = 3$ and $\lambda = 1$ are called Steiner triple systems. Kirkman [1847], Riess [1859] and Moore [1893] proved that the condition (1) is sufficient for the existence of Steiner triple systems. In the 20th century first significant contribution to the problem of existence of bibd's was made by Bose. Bose [1] introduced the "difference method" which enabled one to construct several infinite families of (v, k, λ) – bibd's. Bose and Srikhande and Hanani introduced the "composition methods" which build up designs on larger number of treatments starting from designs on smaller number of treatments. Hanani proved that the necessary condition (1) is also sufficient for $k = 3$ and 4. For $k = 5$, he showed that the condition is sufficient except for $(v, k, \lambda) = (15, 5, 2)$ in which case the design does not exist.

Let $(X, \mathcal{B})$ be a (v, k, λ) – bibd. $(X, \mathcal{B})$ is said to be a resolvable bibd iff

[1] This work was supported in part by NSF Research Grant GP 9375.

there exist list of blocks $\mathcal{B}_1, \mathcal{B}_2, \ldots \mathcal{B}_r$ such that $\mathcal{B} = \mathcal{B}_1 + \mathcal{B}_2 + \ldots + \mathcal{B}_r$ and $(X, \mathcal{B}_i)$ is a $(v, k, 1, 1)$ – combinatorial design for $i = 1, 2, \ldots, r$. If a (v, k, λ)– resolvable bibd exists, then in addition to (1) we must have $v \equiv 0 \pmod{k}$. The necessary conditions for the existence of $(v, k, 1)$ resolvable bibd are equivalent to

$$(2) \qquad v \equiv k \ (\mathrm{mod}\ k\,(k-1)).$$

Resolvable bibds with $k = 3$ and $\lambda = 1$ are also called Kirkman designs or "Kirkman arrangement for school girls". In 1850 "query 6" of Reverend Thomas P. Kirkman published at page 48 of the Lad. and Gentleman's diary introduced the school girl problem as a practical puzzle. A teacher would like to take 15 school girls out for a walk, the girls being arranged in 5 rows of three. The teacher would like to insure equal chances of friendship between any two girls. Hence it is desirable to have different row arrangements for the 7 days of the week such that any pair of girls walk in the same row exactly one day of the week. Clearly such an arrangement for 15 girls is equivalent to a $(15, 3, 1)$ – resolvable Bibd, the 15 girls correspond to the 15 points, every row corresponds to a block of size 3 and 7 days correspond to the 7 parallel classes. In the general case one wants to arrange v girls in $v/3$ rows of three. The problem is to find different row arrangements for $(v-1)/2$ consecutive days such that every pair of girls walk in the same row exactly one day out of the $(v-1)/2$ days. Of course such an arrangement is equivalent to a $(v, 3, 1)$ – resolvable Bibd and hence a necessary condition is $v \equiv 3 \pmod{6}$. Kirkman's school girl problem generated great interest in the late 19th century and early 20th century. Between 1850 and 1947, more than 50 articles had been written on the problem. Celebrated Mathematicians like Burnside, Cayley, Moore, Sylvester and others contributed to the problem. It was proved that for several infinite families of integers n, Kirkman arrangements for $6n + 3$ girls exist. However, for an arbitrary integer n, no solution was known until 1968.

In 1968, Ray-Chaudhuri and Wilson [10] proved that for every positive integer n Kirkman arrangement for $6n + 3$ girls exist. Stated differently, the condition (2) is sufficient for the existence of $(v, 3, 1)$ – resolvable bibd. Hanani, Ray-Chaudhuri and Wilson [7] proved that the condition (2) is also sufficient for the existence of $(v, 4, 1)$ – resolvable bibd. Ray-Chaudhuri and Wilson [11] proved that the condition (2) is "asymptotically sufficient" for the existence of $(v, k, 1)$ – resolvable bibd, i.e., the following theorem is true : Let k be a fixed positive integer. Then there exists a constant $c(k)$ such that if $v \geqslant c(k)$ and $v \equiv k \pmod{k(k-1)}$, then a $(v, k, 1)$ – resolvable bibd exists.

Richard M. Wilson [13, 14] in his Ph.D. dissertation made significant contribution to the existence theory of bibd's. Let $B(K) = \{v \mid (v, K, 1, 2)$ – combinatorial design exists$\}$. The mapping $K \rightarrow B(K)$ is a closure operation on the subsets of positive integers, i.e., (1) $B(K) \supset K$, (ii) $B(B(K)) = B(K)$ and (iii) $K_1 \supset K_2 \Rightarrow B(K_1) \supset B(K_2)$. K is said to be closed iff $K = B(K)$. Let $\beta(K)$ denote the unique positive integer which generates the ideal generated by the set $\{k(k-1),\ k \in K\}$. A fiber f of a closed set K is a residue class f mod $\beta(K)$ such that $\exists\ k \in K$ with $k \equiv f \pmod{\beta(K)}$. A fiber f is said to be complete if and only if $\exists$ a constant M such that $\{v \mid v \geqslant M,\ v \equiv f \bmod \beta(K)\} \subseteq K$. If all fibers are complete, then the closed set K is said to be eventually periodic with period $\beta(K)$. Wilson [14] proved that every closed K is eventually periodic with period $\beta(K)$. A consequence of this theorem is that the condition (1) is "asymptotically sufficient" for the existence of (v, k, λ) – bibd whenever (i) $k/(k, \lambda)$ is 1 or a prime power or (ii) $\lambda \geqslant ([k/2] - 1)([k/2] - 2)$. However no $(v, k, 1)$ – bibd with $v \not\equiv 1$ or k (mod $k(k-1)$) is still known to exist. Jane W. Di Paola conjectured that such bibd's do not exist.

In the existence theory of designs, two different methods, the difference method and the composition method proved to be useful. The difference method assumes a "large" automorphism group for the design and produces a direct construction of the blocks of the design utilizing properties of some finite algebraic structures. The composition method composes several designs of small order into a design of large order. The composition methods depend heavily on orthogonal latin squares, orthogonal arrays and group divisible designs. $N(n)$ denotes the maximum number of mutually orthogonal latin squares of order n. Parker and Bose and Srikhande established the connection between orthogonal latin squares and combinatorial designs with $t = 2$ and proved that $N(n) \geqslant 2$ for all $n > 6$. Bose and Srikhande proved that if $m \leqslant N(t) + 1$ and $1 < u < t$, then

$$N(mt + u) \geqslant \text{Min}\,(N(m) - 1\,, N(m+1) - 1\,, N(t)\,, N(u)).$$

Chowla, Erdös and Strauss exploited this inequality to establish that for large n, $N(n) > \frac{1}{3} n^{1/91}$. This result plays an important role in the existence theory of bibd's. Recently Wilson [15] improved this result and showed that for large n, $N(n) \geqslant n^{1/17} - 2$.

Hoffman and Ray-Chaudhuri [8] proved an interesting and nontrivial characterization of symmetric bibd's. A (v, k, λ) – bibd is said to be symmetric iff any two blocks intersect in λ points. Let π be a (v, k, λ) – symmetric bibd. A loopless simple graph G is a design $(X, \mathscr{B})$ where each block (or edge) contains 2 points (or vertices) and $\mathscr{B}$ takes 0 and 1 as values. Eigenvalues of the graph G are defined to be those of the adjacency matrix of the graph. Hoffman and Ray-Chaudhuri's [8] theorem states that a (v, k, λ) – symmetric bibd exists iff there exists a regular connected simple loopless graph with distinct eigenvalues -2, $2k-2$, $k - 2 \pm \sqrt{k - \lambda}$. Similar result for affine plane i.e., for bibd's with $v = n^2$, $k = n$ and $\lambda = 1$ is also proved. Hoffman and Ray-Chaudhuri [9] proves that a bibd with $v = n^2$, $k = n$ and $\lambda = 1$ exists iff there exists a regular regular connected simple loopless graph with distinct eigen values

$$2n - 1, -2, n - 2, \frac{1}{2}(2n - 3 \pm \sqrt{4n + 1}).$$

Jean Doyen [3] is doing some important work on the number of nonisomorphic designs with a given set of parameters. Let $N(2, 3, n)$ denote the number of non isomorphic Steiner triple systems on n points where $n \equiv 1$ or 3 (mod 6). Doyen proved that for $n \geqslant 15$, $N(2, 3, n) \geqslant 2^{\log_3 \frac{n}{17}}$ and hence goes to ∞ as n goes to ∞.

Hall [4] developed a beautiful method of constructing symmetrie bibd's from finite permutation groups. One starts with a transitive finite permutation group on ν letters and considers the orbits of the stabilizer of a letter. Union of several orbits is taken as the base block. From a suitable representation of the group $LF(3, 4)$ Hall constructed the difficult symmetric bibd with $\nu = 56$, $k = 11$ and $\lambda = 2$.

Ryser [2] proved a striking result about $(\nu, K, \lambda, 2)$– combinatorial designs. Let $(X, \mathscr{B})$ be a $(\nu, K, \lambda, 2)$ – combinatorial design where $\mathscr{B}$ takes only 0 and 1 as values and the number of blocks b is equal to ν. Let r_x denote the number of blocks that contain the point x and let $R = \{r_x, x \in X\}$. Ryser proved that if $\forall x, r_x > \lambda$ and $|R| > 1$, then $|K| = 2$.

REFERENCES

[1] Bose R.C. — On the Construction of Balanced Incomplete Block Designs, *Am. Eugen.*, 9, 1939, p. 353-399.

[2] Bose R.C. and Srikhande S.S. — On the Construction of Sets of Mutually Orthogonal Latin Squares and the Falsity of a Conjecture of Euler, *Trans. Amer. Math. Soc.*, 95, 1960, p. 191-209.

[3] Doyen J. — Sur la Croissance du Numbre de Systems Triples de Steiner Non Isomorphes, *J. of Combinatorial Theory,* 8, 1970, p. 424-441.

[4] Hall Jr. Marshall. — *Construction of Designs from Permutation Groups,* Institute of Statistics Mimeo series No. 600.10, Chapel Hill, N.C., June, 1969.

[5] Hanani H. — The Existence and construction of Balanced Incomplete Block Designs, *Ann. Math. Stat,* 32, 1961, p. 361-386.

[6] Hanani H. — On Some Tactical Configurations, *Canadian J. of Mathematics,* 15, 1963, p. 702-722.

[7] Hanani H., Ray-Chaudhuri D.K. and Wilson Richard M. — On Resolvable Designs, submitted to *Canadian J. of Mathematics.*

[8] Hoffman A.J. and Ray-Chaudhuri D.K. — On the Line Graph of a Symmetric Balanced Incomplete Block Design, *Trans. Amer. Math. Soc.,* 116, 1965, p. 238-252.

[9] Hoffman A.J. and Ray-Chaudhuri D.K. — On the Line Graph of a Finite Affine Plane, *Canadian J. of Mathematics,* 17, 1965, p. 687-694.

[10] Ray-Chaudhuri D.K. and Wilson Richard M. — Solution of Kirkman's Schoolgirl Problem, *Proceedings of the Symposia in « Pure Mathematics », Combinatorics,* Vol. 19, *Amer. Math. Soc.*

[11] Ray-Chaudhuri D.K and Wilson Richard M. — *On the Existence of Resolvable Balanced Incomplete Block Designs, Combinatorial Structures and Their Applications,* Gordon and Breach, New York, 331-341.

[12] Ryser H.J. — An Extension of a Theorem of De Bruijn and Erdős on Combinatorial Designs, *J. Algebra,* 10, 1968, p. 246-261.

[13] Wilson Richard M. — Cyclotomy and Difference Families in Elementary Abelian Groups, submitted to *Journal of Number Theory.*

[14] Wilson Richard M. — An Existence Theory For Pairwise Balanced Designs, I & II, accepted for publication, *J. of Combinatorial Theory.*

[15] Wilson Richard M. — *Concerning the Number of Orthogonal Latin Squares,* Manuscript, Ohio State University, Department of Mathematics, Colombus, Ohio, U.S.A.

The Ohio State University
Dept. of Mathematics,
231 W. 18th Avenue,
Colunbus,
Ohio 43 210 (USA)

Actes, Congrès intern. Math., 1970. Tome 3, p. 229 à 233.

COMBINATORIAL THEORY, OLD AND NEW

by Gian-Carlo ROTA

1. Introduction.

Combinatorial analysis, or combinatorial theory, as it has come to be called, is currently enjoying an outburst of activity. This can be partly attributed to the abundance of new and highly relevant problems brought to the fore by advances in discrete applied mathematics, and partly to the fact that only lately has the field ceased to be the private preserve of mathematical acrobats, and attempts have been made to develop coherent theories, thereby bringing it closer to the mainstream of mathematics. By way of example of this second trend I shall sketch the outlines and prospects of perhaps one of the most promising of such infant theories : the theory of *combinatorial geometries*. This theory is historically rooted in the four-color conjecture much like algebraic number theory was born out of Fermat's conjecture. Indeed, its main achievement to-date is not only a substantial body of results with disparate applications, but, we hazard to state, its success in displaying the four-color problem in a new light, where it appears as merely one case of a general problem leading to a host of analogous problems of varing difficulty (the *critical problem*, §4). The interaction of techniques and insights which results from the simultaneous study of these problems offers –if we are to believe the so-called "lessons of the past"– the brightest hope for the future of the theory.

2. Basic Notions.

Combinatorial geometries are the creation of several mathematicians. Because of limitations of space, we omit all names, but we cannot pass under silence those of the two workers on whose pioneering shoulders the body of the theory rests : *Hassler Whitney and W.T. Tutte.* A *combinatorial geometry* (c.g.) $G = G(S)$ is a set S (hereafter assumed finite for simplicity) together with a closure relation $A \to \overline{A}$ defined on all subsets A of S (i.e., $\overline{\overline{A}} = \overline{A}$, $A \subseteq B$ implies $\overline{A} \subseteq \overline{B}$ and $A \subseteq \overline{A}$, but *not* $\overline{A} \cup \overline{B} = \overline{A \cup B}$) satisfying (1) $\overline{\emptyset} = \emptyset$ for the empty set, (2) $\overline{p} = p$ for every element $p \in S$, (3) the *exchange property* : if $p, r \in S$ and $A \subseteq S$, and if $r \in \overline{A \cup p}$ but $r \notin \overline{A}$, then $p \in \overline{A \cup r}$. The closed sets or *flats* $(A = \overline{A})$ form a *geometric lattice* (g.l.). A subset $I \subseteq S$ is *independent* if $p \notin \overline{I - p}$ for all $p \in I$; all maximal independent sets, or bases, have the same cardinality (hereafter denoted by n). There are equivalent axioms for a c.g. in terms of the family of independent sets and bases.

Dropping conditions (1) and (2) one obtains a pregeometry or matroid. To every pregeometry one can canonically associate a geometry, and we often use the two terms interchangeably. The *direct sum* of two geometries is defined in the obvious way, as is the *restriction* of a geometry to a subset A ; the cardinality of a basis of the restriction to A is the *rank* $r(A)$ of A, satisfying

$$r(A \cup B) + r(A \cap B) \leqslant r(A) + r(B) \qquad \text{and}$$

$$r(A \cup p) = r(A) + 0 \qquad \text{or} = r(A) + 1$$

(this gives another axiom system for c.g. 's). The numbers $W^{(k)}$ of closed sets of rank k are the *Whitney numbers* (*of the second kind*). An open problem is whether the Whitney numbers always form a *unimodal sequence*. Let μ be the Möbius function of the geometric lattice : the numbers

$$W_k = \{\Sigma\mu(0, \overline{A}) : A = \overline{A} \text{ and } r(A) = k\}$$

are the *Whitney numbers* (*of the first kind*). $p_G(\lambda) = \sum_{k=0}^{n} W_k \lambda^{n-k}$ is the *characteristic* (or *Birkhoff*) *polynomial* of the geometry. The *critical problem* (see §4) is the problem of locating the positive integer zeros (especially the largest) of the characteristic polynomial. A *strong map* f of $G(S)$ to $G(T)$ is a function from S to $T \oplus 0$ (where 0 is a "dummy element" added to T) such that the inverse image of a closed set is a closed set. Strong maps are the morphisms in the category of c.g. 's. A *contraction* by a subset A is the (pre-) c.g. on $S - A$ whose closed sets C are those for which $A \cup C$ is closed in $G(S)$. Every strong map can be factored into a monomorphism followed by a contraction. A *modular flat A* satisfies $(*) r(A \cup B) + r(A \cap B) = r(A) + r(B)$ for all flats B. The *orthogonal geometry* $G^{\perp}(S)$ of $G(S)$ is the unique geometry whose bases are those subsets of S whose complement is a basis of $G(S)$.

3. Typical Examples.

The notion of c.g. is abstracted from the properties of linear dependence in projective space P. In fact, the typical examples are obtained by taking a subset S of P and setting $\overline{A} = \text{sp}(A) \cap S$ for $A \subseteq S$, where $\text{sp}(A)$ is the linear variety in P spanned by the subset A. A host of notions of synthetic projective geometry can be carried over to c.g. 's by this analogy. Conversely, the *coordinatization* or *representation problem* (*v.* §5) asks for conditions on a c.g. that it be representable as a subset of projective space over a given field (or even more generally) in the manner just described. Most geometries are not representable over any field; : e.g. , take the direct sum of a projective space over $GF(2)$ and one over $GF(3)$.

Given an Abelian group A, a subgroup V of the Abelian group of all functions f from a set S to A can be used to define the *function space geometry* of V : for $A \subseteq S$, set $\overline{A} = \{p \in S : f(p) = 0$ for all $f \in V$ s.t. $f(a) = 0$ for all $a \in A\}$. For example, if S is the set of edges of an oriented linear graph, one can take V as the set of all 1-coboundaries ; in the resulting *bond geometry* independent sets are trees and bases are spanning trees. The orthogonal c.g. to the bond g. (the *circuit* g.) is obtained by taking V to be all 1-cycles. Much of classical graph theory can be obtained by applying general results on c.g. 's to these two special cases. If one starts with the complete graph on j vertices, the bond-geometric lattice is the lattice of partitions of a j-element set, whose Whitney numbers are the *Stirling numbers.* (of the first and second kind).

Passing from graphs to finite simplicial complexes, let $S = T_k$ be the set of all k-element subsets of a set T. For $A \subseteq T_k$, set $\Sigma(A) = T_0 \cup T_1 \cup \ldots \cup T_{k-1} \cup A$, and let β_i be the $(i - 1)$-dimensional Betti no. of the simplicial complex $\Sigma(A)$. Setting $r(A) = \binom{n-1}{k-1} - \beta_{k-1}(\Sigma(A)) = \nu(A) - \beta_k(\Sigma(A))$, where $\nu(A)$ is the cardinality of A, we obtain the rank function of a c.g. , the *simplicial geometry* of dimension k on T. It can be used to extend notions of graph theory to higher dimensions ; its orthogonal geometry is connected with Alexander duality. A slight extension of this construction leads to a class of geometries on triangulable manifolds whose Whitney nos. give a new set of topological invariants, not invariant under homotopy.

A *submodular set function* μ on S satisfies : (1) $\mu(A)$ is an integer for all $A \subseteq S$; (2) $\mu(A \cup B) + \mu(A \cap B) \leqslant \mu(A) + \mu(B)$; (3) If $A \subseteq B$, then $\mu(A) \leqslant \mu(B)$. The family of all sets I s.t. $\nu(A) \leqslant (A)$ for all non-empty subsets $A \subseteq I$ is the family of all independent set of a c.g. In this way, one can construct a wide variety of c.g. 's which are a long way from being representable as subsets of projective space. One can for example set μ as a linear combination with positive integer coefficients of rank functions of several geometries defined on the same set. These c.g. 's can be used to give a unified and widely extended treatment of matching, covering, minimax theorems, etc.

A connection with integer programming is obtained through *unimodular geometries*, which are integer-valued function space geometries V on S s.t. if $f \in V$ has minimal support, then there is a $g \in V$ with the same support as f and taking only the values ± 1 and 0.

4. The Critical Problem.

For c.g. 's $G(S)$ which are subsets S of projective space P of dim n over $GF(q)$, the smallest integer c (the *critical exponent*) such that $p_G(q^c) > 0$ equals the smallest number of hyperplanes $H_1, \ldots, H_c$ in P such that every $p \in S$ does not belong to at least one such H_i. If S is the set of all points with d or fewer non-zero coordinates, this is the classical problem of finding *linear codes*. Taking $G(S)$ to be all vectors in a vector space over $GF(5)$ whose coordinates are ± 1 or 0, the critical exponent is the *capacity* of the 5-graph. A host of other classical problems are thus seen to be critical problems. For a function space geometry with values in $A = GF(q)$, the critical exponent is the smallest number of a set of functions $f_1, \ldots, f_c$ such that for every $p \in S$, at least one $f_i(p) \neq 0$. For the bond geometry, the critical exponent is c whenever the graph is colorable in q^c colors. For a unimodular geometry, the smallest positive integer n such that $p_G(n) > 0$ is the smallest n for which there is a non-vanishing function $f \in V$ taking $n - 1$ values. For the circuit geometry of a graph with integer values, this reduces to the problem of minimum flows ; it is conjectured that $n = 6$ for all graphs (a far more interesting conjecture than the four-color conjecture).

A study of the characteristic polynomial for general c.g. 's has already yielded a crop of results, of which we shall quote only two. A c.g. is *supersolvable* when there exist modular flats $A_0 \subseteq A_1 \subseteq A_2 \subseteq \ldots \subseteq A_n = S$ s.t. $r(A_i) = i$. The roots of the characteristic polynomial of a supersolvable c.g. are positive integers, equal to $\nu(A_{i+1} - A_i)$

A *loop* of a pregeometry $G(S)$ is a point p s.t. $p \in \overline{\emptyset}$; a *link* is a point p s.t. $G(S)$ is the direct sum of $S - p$ and p. Take the free polynomial ring generated by all isomorphism classes of pregeometries, divide it by the ideal generated by : (a) $G(S) - G(T)\, G(V)$ whenever S is the direct sum of T and V ; (b) $G(S) - G(S-p) - G(S/p)$ whenever p is not a loop or a link, where $G(S-p)$ is the subgeometry on $S-p$ and $G(S/p)$ is the contraction by p. The resulting *Tutte-Grothendieck ring* has the remarkable property of being isomorphic to a polynomial ring in two generators without constant terms. By evaluating the polynomial $t_G(z, x)$ associated in this way to $G(S)$, one can compute all "invariants", for example $t_G(1 - \lambda, 0) = \pm\, p_G(\lambda)$ (the characteristic polynomial); $t(1, 1) =$ no. of bases ; $t(2, 1) =$ no. of independent sets $t(1, 0) =$ Möbius function $|\mu(\emptyset, S)|$, etc. It appears that classical "reductions" in the coloring problem can be systematically studied –and extended to the critical problem generally– by algebraic methods through the *T-G*-ring (and other categorical structures currently being developed). This leads to the conjecture (supported by the results in §5) that the critical exponent is related to the absence of certain "forbidden segments", or "obstructions" in the g.l. of $G(S)$ or in the g.l. of $G^{\perp}(S)$. Hadwiger's conjecture can be restated in these terms.

The *Bose-Segre problem* of finding in P over $GF(q)$ maximum-sized subsets B with the property that no k-subset of B is linearly dependent is also closely related to a suitable critical problem and opens interesting connections with algebraic geometry over finite fields.

5. Representation.

A c.g. is representable as a subset of a projective space over $GF(2)$ iff no segment $[A, B]$ of length 2 (i.e., $r(B) - r(A) = 2$) in the g.l. contains more than five flats. We abbreviate this by "the 4-point line is the only obstruction for representation over $GF(2)$".

In the same vein, the five-point line and its orthogonal c.g. and the Fano plane and its orthogonal c.g. are the only obstructions for representation over $GF(3)$.

There probably is a finite number of obstructions for representations over any given finite field, although their classification is far from complete. The obstructions for representation as a *unimodular* c.g. are the 4-pt. line, the Fano plane and its orthogonal c.g., and for representation as a bond-g. they are the same plus the circuit g. 's of the two Kuratowski graphs ; finally, for representation as the bond g. of a *planar* graph they are the above, plus the bond-g. 's of the two Kuratowki graphs (a crowning achievement of Tutte). Thus Kuratowski's theorem, which has not been satisfactorily explained by topology, finds itself in pleasing company in c.g.

The preceding results lead to the question : is there a "universal" algebraic structure which can "represent" every c.g. ? *An algebra of syzygies* of rank n

over an Abelian group A consists of a set S, together with a map T of *pairs of ordered n-tuples* of S into A, satisfying the following identities : (1) If σ and τ are permutations of $(1, 2, \ldots, n)$, then

$$T(x_{\sigma 1}, x_{\sigma 2}, \ldots, x_{\sigma n} \mid y_{\tau 1}, y_{\tau 2}, \ldots, y_{\tau n}) = \pm T(x_1, \ldots, x_n \mid y_1, \ldots, y_n)$$

according as the product $\sigma\tau$ is even or odd ($x_i \in S$ and $y_i \in S$).

$$(2)\ T(x_1, x_2, \ldots, x_n \mid y_1, y_2, \ldots, y_n)$$
$$= \sum_{i=1}^{n} T(x_1, x_2, \ldots, x_{i-1}, y_1, x_{i+1}, \ldots, x_n \mid x_i, y_2, y_3, \ldots, y_n)$$

Every algebra of syzygies defines a ***standard** c.g.* by taking the ***bases*** to be those n-tuples $(x_1, x_2, \ldots, x_n)$ for which $T(x_1, \ldots, x_n \mid y_1, \ldots, y) \neq 0$ or

$$T(y_1, \ldots, y_n \mid x_1, \ldots, x_n) \neq 0$$

for some n-tuple of y's. Every c.g. representable in projective space is standard : set $T(x_1, \ldots, x_n \mid y_1, \ldots, y_n) = \det(x_1, \ldots, x_n) \det(y_1, \ldots, y_n)$ relative to a fixed basis. More generally, to every c.g. one can canonically associate a standard c.g. which is, in a precise sense, universal relative to all representations. Several classical theorems of projective geometry, such as Desargues and Pappus, can be interpreted as conditions on the algebra of syzygies, allowing stronger representation theorems. The algebra of syzygies makes contact with the classical invariant theory of finite point sets in projective space, of which c.g. is an abstraction. The interaction and confluence of such manifold trends and problems is perhaps the most promising feature of this field of investigation.

BIBLIOGRAPHY

All bibliographical and historical references will be found in

CRAPO-ROTA. — *On the foundations of Combinatorial Theory : Combinatorial Geometries,* Cambridge, Mass., The M.I.T. Press, 1970.

In addition, I have drawn from unpublished work of T. BRYLAWSKI, R. REID, W. WHITELEY, N. WHITE, R. STANLEY, T. HELGASON and my own.

2-347/M.I.T
Cambridge, Massachusetts 02139
U.S.A

NEW TYPES OF COMBINATORIAL DESIGNS*

by H.J. RYSER

Dedicated to Marshall Hall, Jr.
on the Occasion of his Sixtieth Birthday

1. Introduction

Let $X = \{x_1, x_2, \dots, x_v\}$ be a set of v elements (a v-set) and let $X_1, X_2, \dots, X_v$ be subsets of X. These subsets of X are called a (v, k, λ)-*design* provided :

(1.1) Each X_i is a k-subset of X.

(1.2) Each $X_i \cap X_j$ for $i \neq j$ is a λ-subset of X.

(1.3) The integers v, k, and λ satisfy $0 < \lambda < k < v - 1$.

Now let $a_{ij} = 1$ if x_i is a member of X_j and let $a_{ij} = 0$ if x_i is not a member of X_j. Then

$$A = [a_{ij}] \tag{1.4}$$

is a (0,1)-matrix of order v that is called the *incidence matrix* of the (v, k, λ)-design. It follows directly from the definition of a (v, k, λ)-design that A satisfies the matrix equation

$$A^T A = (k\text{-}\lambda) I + \lambda J. \tag{1.5}$$

In (1.5) A^T denotes the transpose of the matrix A. The matrix J is the matrix of 1's of order v and I is the identity matrix of order v. The incidence matrix of a (v, k, λ)-design has a number of remarkable properties. One such property is the normality of A, namely,

$$AA^T = A^T A. \tag{1.6}$$

The (v, k, λ)-designs play a fundamental role in modern combinatorics, and they have been extensively investigated by way of their incidence matrices. We make no attempt to summarize the literature here. Such summaries are available in Dembowski [5], Hall [6], and Ryser [8]. Instead, we deal with "slightly modified" (v, k, λ)-designs. We state our main results in terms of (0,1)-matrices. This terminology turns out to be very economical. But throughout the discussion it should be understood that our basic interest is in the combinatorial configurations represented by these matrices.

(*) This research was supported in part by ONR Contract Number N00014-67-A-0094-0010.

2. An Extension of a Theorem of de Bruijn - Erdös

Let $S_1, S_2, \ldots, S_n$ be n subsets of an m-set S. Suppose now that each S_i and S_j with $i \neq j$ intersect in exactly one element of S. We further require that both n and the number of elements in each S_j be greater than one. Then the de Bruijn-Erdös theorem [4] asserts that $m \geqslant n$ and the theorem also gives a classification of the configurations in the case of equality $m = n$. Recently Ryser [9] and Woodall [11] attempted independently to extend the de Bruijn-Erdös theorem so that each S_i and S_j with $i \neq j$ intersect in exactly λ elements of S, where λ is a fixed but arbitrary positive integer. One of their main conclusions is the following. (Throughout the discussion a *line* of a matrix designates either a row or a column of the matrix, and the matrix J is the matrix of 1's of a specified order).

THEOREM 2.1. – *Let A be a* (0,1)*-matrix of size m by n such that*

$$A^T A = \operatorname{diag}[k_1 - \lambda, k_2 - \lambda, \ldots, k_n - \lambda] + \lambda J, \tag{2.1}$$

where $n > 1$ and $k_j > \lambda \geqslant 1$. Then

$$m \geqslant n, \tag{2.2}$$

and if equality holds in (2.2) *then A satisfies one of the following two requirements :*

(*X*) *Each line sum of A equals a positive integer k.*

(*Y*) *The matrix A has exactly two distinct row sums r_1 and r_2 and these numbers satisfy*

$$r_1 + r_2 = n + 1. \tag{2.3}$$

The configurations associated with (*X*) are the (v, k, λ)-designs with $m = n = v$ (apart from unimportant degeneracies that are entirely matters of definition). The configurations associated with (*Y*) are called *λ-designs* on n elements. The de Bruijn-Erdös theorem tabulates all λ-designs with $\lambda = 1$. But the λ-designs with $\lambda > 1$ are of a much more complicated structure.

Woodall [11] has shown that the number of λ-designs for each fixed value of $\lambda > 1$ is finite. Let a (v, k, λ')-design have parameters not of the form $v = 4\lambda - 1$, $k = 2\lambda - 1$, $\lambda' = \lambda - 1$. Then by an elementary but judicious modification of the (v, k, λ')-design one may construct a λ-design with $\lambda = k - \lambda'$ and row sums $v - k$ and $k + 1$. Bridges [1] has called the λ-designs constructed in this prescribed manner (including the λ-designs with $\lambda = 1$) *type* 1 *λ-designs*. All known λ-designs are of type 1, and it is conjectured that all λ-designs are of type 1. The combined efforts of Bridges [1], Bridges and Kramer [2], and Kramer [7] have verified the validity of this conjecture for $\lambda \leqslant 9$. But the possibility of the existence of highly exotic λ-designs remains open.

3. The Matrix Equation $XY = (k - \lambda)I + \lambda J$

The study of λ-designs was motivated by modifying slightly the right side of the matrix equation (1.5). Bridges and Ryser [3] have carried out investigations that involve a much more drastic modification of the left side of (1.5). Specifically, they prove the following theorem.

THEOREM 3.1. – *Let X and Y be nonnegative integral matrices of order $n > 1$ such that*

$$(3.1) \qquad XY = (k-\lambda)I + \lambda J,$$

where $k \neq \lambda$ and the integers k and λ are relatively prime. Then there exist positive integers r and s such that X has constant line sums r and Y has constant line sums s, where

$$(3.2) \qquad rs = k + (n-1)\lambda.$$

Moreover,

$$(3.3) \qquad YX = XY.$$

We remark that the arithmetical requirement $(k, \lambda) = 1$ in Theorem 3.1 is essential in the sense that the main conclusions of the theorem are no longer valid with this restriction removed. The following corollaries imply that the theorem has a direct bearing on the structure of certain (v, k, λ)-designs.

COROLLARY 3.2. – *Let X and Y be nonnegative integral matrices of order*

$$n^2 + n + 1 \ (n > 1)$$

such that

$$(3.4) \qquad XY = nI + J,$$

and let this factorization be proper in the sense that neither X nor Y is a permutation matrix. Then $X = Y^T$ is the incidence matrix of a projective plane of order n.

COROLLARY 3.3. – *Let the incidence matrix A of a (v, k, λ)-design be written as a product of $(0,1)$-matrices of order v*

$$(3.5) \qquad A = \prod_{i=1}^{t} A_i \qquad (t > 1),$$

where $k > \lambda^2$ and the integers k and λ are relatively prime. Then $t - 1$ of the factors of A are permutation matrices and the remaining factor A_j is of the form PAQ, where P and Q are permutation matrices.

4. A Generalization of (v, k, λ)-Designs and λ-Designs

Let A be a $(0,1)$-matrix of order $n \geqslant 3$ that satisfies the matrix equation

$$(4.1) \qquad A^T A = \operatorname{diag}[k_1 - \lambda_1, k_2 - \lambda_2, \ldots, k_n - \lambda_n] + [\sqrt{\lambda_i}\ \sqrt{\lambda_j}],$$

where $k_i - \lambda_i$ and λ_j are positive. We call a configuration whose incidence matrix A fulfills these requirements a *multiplicative design* on the parameters $k_1, k_2, \ldots, k_n$ and $\lambda_1, \lambda_2, \ldots, \lambda_n$. It is evident that our definition of a multiplicative design places heavy restrictions on the parameters $k_1, k_2, \ldots, k_n$ and λ_1, $\lambda_2, \ldots, \lambda_n$. But multiplicative designs may be regarded as a natural generalization of (v, k, λ)-designs and λ-designs. The basic generalization involves the replacement of the very critical matrix λJ by a symmetric matrix of rank 1. Ryser [10] has investigated various properties of multiplicative designs. For example, one may prove that the parameters $k_1, k_2, \dots, k_n$ and $\lambda_1, \lambda_2, \dots, \lambda_n$ of a multiplicative design satisfy

$$(4.2)\quad \left[\frac{k_1^2}{k_1-\lambda_1}+\dots+\frac{k_n^2}{k_n-\lambda_n}-n\right]\left[1+\frac{\lambda_1}{k_1-\lambda_1}+\dots+\frac{\lambda_n}{k_n-\lambda_n}\right] = \left[\frac{\sqrt{\lambda_1}}{k_1-\lambda_1}k_1+\dots+\frac{\sqrt{\lambda_n}}{k_n-\lambda_n}k_n\right]^2$$

A multiplicative design on the parameters $k_1, k_2, \dots, k_n$ and $\lambda_1, \lambda_2, \dots, \lambda_n$ is called a *uniform design* provided

$$(4.3)\qquad k_1-\lambda_1 = k_2-\lambda_2 = \dots = k_n-\lambda_n \equiv c.$$

Uniform designs are especially interesting because of the following duality theorem [10].

THEOREM 4.1. – *Let A be the incidence matrix of a uniform design on the parameters $k_1, k_2, \dots, k_n$ and $\lambda_1, \lambda_2, \dots, \lambda_n$. Then A^T is also the incidence matrix of a uniform design and satisfies the matrix equation*

$$(4.4)\qquad AA^T = cI + ct\,[x_i x_j],$$

where

$$(4.5)\qquad t = 1+\frac{1}{c}(\lambda_1+\dots+\lambda_n),$$

$$(4.6)\qquad tx_i = \frac{1}{c}(\sqrt{\lambda_1}\,a_{i1}+\dots+\sqrt{\lambda_n}\,a_{in}).$$

It is clear that (v, k, λ)-designs are examples of uniform designs. But in conclusion we remark that various families of uniform designs have been constructed that are not (v, k, λ)-designs.

REFERENCES

[1] BRIDGES W.G. — Some results on λ-designs, *J. Comb. Theory,* 8, 1970, p. 350-360.

[2] BRIDGES W.G. and KRAMER E.S. — The determination of all λ-designs with $\lambda = 3$, *J. Comb. Theory,* 8, 1970, p. 343-349.

[3] BRIDGES W.G. and RYSER H.J. — Combinatorial designs and related systems, *J. Algebra,* 13, 1969, p. 432-446.

[4] DE BRUIJN N.G. and ERDÖS P. — On a combinatorial problem, *Indagationes Math.,* 10, 1948, p. 421-423.

[5] DEMBOWSKI P. — *Finite Geometries,* Springer-Verlag, Berlin, 1968.

[6] Hall Jr. M. — *Combinatorial Theory,* Blaisdell, Waltham, Mass., 1967.

[7] Kramer E.S. — *On λ-designs,* Dissertation, Univ. of Michigan, 1969.

[8] Ryser H.J. — *Combinatorial Mathematics* (Carus Monograph 14), Wiley, New York, 1963.

[9] Ryser H.J. — An extension of a theorem of de Bruijn and Erdős on combinatorial designs, *J. Algebra,* 10, 1968, p. 246-261.

[10] Ryser H.J. — Symmetric designs and related configurations (to be published).

[11] Woodall D.R. — Square λ-linked designs, *Proc. London Math. Soc.* (3), 20, 1970, p. 669-687.

California Institute of Technology
Dept. of Mathematics,
Pasadena,
California 91109 (USA)

Actes, Congrès intern. Math., 1970. Tome 3, p. 241 à 247.

E6 - STATISTIQUE MATHÉMATIQUE

APPLICATIONS OF THE EMPIRICAL BAYES APPROACH*

by L.N. BOLSHEV

1. Introduction.

Let X and Y be random variables, $p(x|y)$ be a conditional density function of X given $Y = y$, $p(y)$ be a marginal density function of Y, and $q(y|x)$ be a conditional density function of Y given $X = x$. We have Bayes formulae

$$q(y|x) = p(x|y)\, p(y)/q(x) \quad , \quad q(x) = \int p(x|y)\, p(y)\, dy \ ;$$

$$E\{Y|X = x\} = \int y \cdot q(y|x)\, dy = \frac{1}{q(x)} \int y \cdot p(x|y)\, p(y)\, dy.$$

These formulae have played a fundamental role in the construction of statistical estimates of random parameters because

$$E\,(E\{Y|X\} - Y)^2 = \inf_{\varphi} E\,[\varphi(X) - Y]^2.$$

In usual statistical problems the a priori distribution $p(y)$ is unknown, in which case the above solution is not applicable and different procedure is called for. The new approach was proposed by Fisher (for point estimates and fiducial interval estimates) and Neyman (for confidence interval estimates). However, in 1941, Bernstein [1] constructed an example to demonstrate a situation in which an application of methods due to Fisher and Neyman may lead to paradoxical results.

EXAMPLE. – A producer has "boxes", $B_1, \ldots, B_m$, containing certain "objects". The objects in the box B_i are normally distributed with mean Y_i and variance unity. The values of $Y_1, \ldots, Y_m$ are unknown. A customer wants to receive s boxes with $Y \leqslant y^*$ (the value of y^* being known) but he is satisfied with a procedure which ensures probability P that, at least in s' boxes, $Y \leqslant y^*$ ($k \leqslant s'/s < 1$, k is a known positive constant). What is acceptance procedure when only one observation is available for each box ?

Let $X_1, \ldots, X_m$ be the observations from $B_1, \ldots, B_m$. Then the ordinary plan is : "If $X_i < y^* - c$ accept B_i (as good), if $X_i \geqslant y^* - c$ reject B_i (as bad)" ; c is a known constant. The customer may accept any s boxes from the set of boxes declared good ; if s = const and m is large, this is possible. However, this decision

[1] Prepared with the partial support of the Indian Statistical Institute (Calcutta).

may be incorrect. For example, let Y_i's be all equal, and $Y_1 = \dots = Y_m = Y > y^*$. Let $X^{(1)} \leqslant \dots \leqslant X^{(m)}$ be order-statistics from the sample $X_1, \dots, X_m$. In this case

$$P\{X^{(s)} < y^* - c\} = \frac{1}{B(s, m-s+1)} \int_0^z u^{s-1}(1-u)^{m-s}\,du, \quad z = \Phi(y^* - c - Y).$$

If $m \to \infty$ and $s = \text{const}$,

$$P\{X^{(s)} < y^* - c\} = \frac{1}{\Gamma(s)} \int_0^{mz} v^{s-1} e^{-v} dv + o(1) \to 1.$$

So that, in practice, all accepted boxes will be bad. Bernstein proposed an alternative procedure which uses an estimate of an apriori density for Y. However, he has not provided all the details for applying his method.

In 1955, Robbins [2] gave some examples of empirical Bayes procedure (e.B.p.) for point estimation without immediate estimation of the apriori distribution of Y.

1. Example. – $p(x|y) = y^x e^{-y}/x\,!$, $x = 0, 1, 2, \dots$, $y > 0$. Values of the random variables $X_1, \dots, X_m$ are known, and $Y_1, \dots, Y_m$ are unknown. What is an estimate of Y corresponding to a new observed value, $X = x$? We have

$$E\{Y|x\} = (x+1)\frac{P\{X = x+1\}}{P\{X = x\}} \quad \text{for any} \quad p(y).$$

Let ν_x be a sample density of the event $\{X = x\}$, $x = 0, 1, 2, \dots$, from the sample $X_1, \dots, X_m$. Then $(x+1)\, \nu_{x+1}/\nu_x$ is a consistent estimate of $E\{Y|x\}$.

2. Example. – $p(x|y) = \binom{n}{x} y^x (1-y)^{n-x}$, $x = 0, 1, \dots, n$, or

$$(1) \quad p(x|y) = \frac{\binom{n}{x}\binom{N-y}{n-x}}{\binom{N}{n}}, \quad \max(0, y+n-N) \leqslant x \leqslant \min(y, n).$$

In this case, for any apriori distribution $p(y)$ $(0 \leqslant y \leqslant 1$ in the binomial case, and $y = 0, 1, \dots, N$ in the hypergeometric case), the ratio

$$(x+1) P_{n+1}\{X = x+1\}/(n+1) P_n\{X = x\}$$

is equal $E\{Y|x\}$ (for the binomial distribution) or $E\{(Y-x)/(N-n)|x\}$ (for the hypergeometric distribution). $P_n\{X = x\}$ is the probability of the event $\{X = x\}$, n is "the sample size". In this example, consistent estimates of $E\{\cdot|x\}$ are functions of the vectors $(X_1, X_1'), \dots, (X_m, X_m')$ where X is an observation in the sample of the size n, and X' is one in the supplementary sample of the size 1. If ν_{x0} and ν_{x1} are the sample densities of the events $\{X = x, X' = 0\}$ and $\{X = x, X' = 1\}$ respectively then $(x+1)(\nu_{x+1,0} + \nu_{x,1})/(n+1)(\nu_{x,0} + \nu_{x,1})$ is a consistent estimate of $E\{\cdot|x\}$.

In 1963, Kagan [3] proved a theorem from which it follows that in second example a consistent estimate of $E\{\cdot|x\}$ does not exist. It is not, however, a contradiction. Robbins constructed his estimate with help of the sample $(X_1, X_1'), \ldots, (X_m, X_m')$; Kagan's theorem deals with the sample $X_1, \ldots, X_m$ only and demonstrates that first example is a special case. This may be the reason for e.B.p. being less popular in statistical applications, particulary, in statistical quality control.

The theoretical foundition of modern theory of acceptance sampling, on non-Bayesian lines, is due to Kolmogorov [4]. I shall be talk about this theory later. My essential purpose is a discussion of possibilities of using e.B.p. I wish to say

(1) Non-existence of consistence estimates is not an argument for rejecting the idea of the e.B.p. in statistical applications.

(2) All conclusions of the non-Bayesian theory of the acceptance sampling may be obtained from the general results of the e.B.p.

2. Essential formulae.

Let $Z_1, \ldots, Z_N$ be dichotomous random variables with values 0 or 1, and $X = Z_1 + \ldots + Z_n$, $Y = Z_1 + \ldots + Z_N$, $n < N$ (if $n = N$ then $X = Y$). Let $p(y)$, $y = 0, 1, \ldots, N$, be an apriori distribution of Y, and the conditional distribution of X given $Y = y$, $p(x|y)$, is the hypergeometric distribution (1). Then the marginal probabilities $r_x = P\{X = x\}$, $r_{x0} = P\{X = x, Z_{n+1} = 0\}$ and $r_{x1} = r_x - r_{x0} = P\{X = x, Z_{n+1} = 1\}$ satisfy the equations

$$(2) \qquad (n - x) r_{x,1} = (x + 1) r_{x+1,0} \quad , \quad x = 0, 1, \ldots, n - 1.$$

The equations (2) are a consequence of obvious combinatorial identities and relations :

$$(3) \quad \frac{y - x}{N - n} p(x|y) = P\{X = x, Z_{n+1} = 1 \mid Y = y\}$$

$$= \frac{x + 1}{n - x} P\{X = x + 1, \; Z_{n+1} = 0 \mid Y = y\}.$$

If N is size of a lot, n is the sample size, X and Y are the numbers of defective objects in the sample and in the lot respectively then

$$(4) \qquad E_x = E\left\{\frac{Y - X}{N - n}\middle| X = x\right\} = \frac{1}{r_x} \sum_y \frac{y - x}{N - n} p(x|y)\, p(y)$$

is the conditional expectation of defective proportion among unsampled objects given the number of defective objects among sampled objects, $X = x$. Also, from (2 - 4), it follows (for $s = 0, 1, \ldots, x$ and $t = 0, 1, \ldots, n - x$)

$$(5) \quad E_x = (-1)^x \frac{\binom{n}{x}}{r_x} \left[\sum_{k=x-s+1}^{x} (-1)^k \frac{r_k}{\binom{n}{k}} + (-1)^{x-s} \frac{r_{x-s,1}}{\binom{n}{x-s}} \right] ,$$

$$(6)\qquad E_x = 1 - (-1)^x \frac{\binom{n}{x}}{r_x} \left[\sum_{k=x}^{x+t-1} (-1)^k \frac{r_k}{\binom{n}{k}} + (-1)^{x+t} \frac{r_{x+t,0}}{\binom{n}{x+t}} \right].$$

The conditional distribution $p(x|y)$ is hypergeometric if and only if the equations (2) are true for all $n = 1, 2, \dots, N$.

The expectations, E_x's, are defined in terms of probabilities r_k and any one probability $r_{x-s,1}$ or $r_{x+t,0}$. If the r_k's are known but all r_{k0}'s and r_{k1}'s are unknown, we may not be able to difine the expectations E_x's. In this particular hypergeometric case, Kagan's theorem is, really, a consequence of this fact. However, we can construct inequalities for E_x's. We have, from (5) and (6) under the condition $s = t = 0$: $E_x = r_{x1}/r_x = 1 - r_{x0}/r_x$. These equations and the system (2) give the system of the equations for E_x's :

$$(7)\qquad (n-x)\,r_x E_x = (x+1)\,r_{x+1}(1 - E_{x+1}) \quad , \quad x = 0, 1, \dots, n-1.$$

The coeffitions of the system, r_x's, are non-negative constants satisfying the condition $r_0 + \dots + r_n = 1$ and a system of inequalities which is a consequence of the formulae (5) and (6). For example, if $0 \leqslant u < v \leqslant n$ then, from

$$(8)\qquad (-1)^u \frac{r_{u1}}{\binom{n}{u}} + \sum_{k=u+1}^{v-1} (-1)^k \frac{r_k}{\binom{n}{k}} + (-1)^v \frac{r_{v0}}{\binom{n}{v}} = 0\,,$$

it follows

$$(9)\qquad \sum_{k=u}^{v} (-1)^k \frac{r_k}{\binom{n}{k}} \begin{cases} \geqslant 0 \text{ if } u \text{ and } v \text{ are even,} \\ \leqslant 0 \text{ if } u \text{ and } v \text{ are odd.} \end{cases}$$

Geometricaly, the solutions $(E_0, \dots, E_n)$ of the linear system (7) define a straight line. We known the E_x's to satisfy obvius inequalities $0 \leqslant E_x \leqslant 1$, $x = 0, 1, \dots, n$. Intersect the straight line with this cube ?

THEOREM. – *The intersection is a necessary condition of the correctness of the formule* (1) *for* $p(x|y)$. In other words, the intersection is a necessary condition of the sample representation.

The system (7) and the theorem are true for the binomial distribution with the obvius replacements : $p(x|y) = \binom{n}{x} y^x (1-y)^{n-x}$ and $E_x = E\{Y|x\}$.

EXAMPLE. – $n = 3$, $r_0 = 0.10$, $r_1 = 0.58$, $r_2 = 0.30$, $r_3 = 0.02$. In this case the system (7) is $15\,E_0 = 29(1 - E_1)$, $29\,E_1 = 15(1 - E_2)$, $5\,E_2 = 1 - E_3$. We surmise firstly $0 \leqslant E_3 \leqslant 1$, i.e. $0 \leqslant E_2 \leqslant 1/5$, $12/29 \leqslant E_1 \leqslant 15/29$ and $14/15 \leqslant E_0 \leqslant 17/15$. However, $E_0 \leqslant 1$ therefore $E_1 \geqslant 14/29$, $E_2 \leqslant 1/15$ and $E_3 \geqslant 2/3$. Hence, $14/15 \leqslant E_0 \leqslant 1$, $14/29 \leqslant E_1 \leqslant 15/29$, $0 \leqslant E_2 \leqslant 1/15$ and $2/3 \leqslant E_3 \leqslant 1$.

The ordinary acceptance sampling plan is the principle : "If $X \leqslant c$ then accept the lot, if $X > c$ then reject the lot" $(0 \leqslant c < n)$. Practically, this is good in the case $\max(E_0, \dots, E_c) \leqslant \min(E_{c+1}, \dots, E_n)$. The expectation of the defective proportion in accepted lots is $W = w/(r_0 + \dots + r_c)$ where $w = r_0 E_0 + \dots + r_c E_c$. We have, from (6) under the condition $t = c - x + 1$,

$$n \cdot w = (1 \cdot r_1 + 2 \cdot r_2 + \cdots + c \cdot r_c) + (c+1) r_{c+1,0} \tag{10}$$

therefore

$$\frac{1}{n} \sum_{x=1}^{c} x r_x \leqslant w \leqslant \frac{1}{n} \sum_{x=1}^{c+1} x r_x \, . \tag{11}$$

The right hand side of (11) is the approximate value of w which follows from non-Bayesian theory of Kolmogorov (Kolmogorov [4] obtained this approximate value for $c = 0$, the general case was investigated by Sirazdinov [5]). The inequalities (11) are not unique. Other ones may be constructed with help of the formulae (10) and (8).

Analogous inequalities may be constructed for the variance, v, of the defective proportion in accepted lots too. For example, if $c = 0$ and $r_0 > 0$ then

$$v = \frac{1}{n(N-n)} \cdot \left[\frac{r_{1,0}}{r_0} + 2 \frac{N-n-1}{n-1} \cdot \frac{r_{2,0,0}}{r_0} - \frac{N-n}{n} \left(\frac{r_{1,0}}{r_0} \right)^2 \right]$$

where $r_{2,0,0} = P\{X = 2, Z_{n+1} = Z_{n+2} = 0\}$. Therefore $0 \leqslant v \leqslant V$ where

$$V = \begin{cases} \dfrac{1}{n(N-n)} \left[\dfrac{r_1}{r_0} + 2 \dfrac{N-n-1}{n-1} \cdot \dfrac{r_2}{r_0} - \dfrac{N-n}{n} \left(\dfrac{r_1}{r_0} \right)^2 \right] & \text{if } \dfrac{r_1}{r_0} \leqslant \dfrac{n}{2(N-n)}, \\ \dfrac{1}{n(N-n)} \left[\dfrac{n}{4(N-n)} + 2 \dfrac{N-n-1}{n-1} \cdot \dfrac{r_2}{r_0} \right] & \text{if } \dfrac{r_1}{r_0} > \dfrac{n}{2(N-n)}. \end{cases}$$

3. Statistical conclusions.

In second section it is assumed that the probabilities, r_k's, are known. But they are unknown in practice, of course, Let ν_x be a sample density of the event $\{X = x\}$ in a set of the independent random variables $X_1, \dots, X_m$ ($x = 0, 1, \dots, n$; m is the sample size).

The first problem of statistical control is a choice of a sampling plan. Usually, this problem is solved with help of some imitated characteristics (for example, with help of "the middle output equality"; see the book by Gnedenko a.o. [6]). The e.B.p. allows us to solve this one with help of the natural characteristics, E_x's. However, the usual statistical technique does not apply to E_x's. It is applicable for an estimation of their lower and upper limits only (see the second section).

Example. – $n = 8, \nu_0 = 825, \nu_1 = 4359, \nu_2 = 3832, \nu_3 = 67, \nu_4 = 345, \nu_5 = 429, \nu_6 = 135, \nu_7 = 8, \nu_8 = 0$. In this example, the sample size is large ($m = \Sigma \nu_x = 10{,}000$), and the densities, ν_x's, satisfy the inequalities (9) whichever are true for unknown r_x's. Let r_8 be an arbitrary probability, and $r_x \propto \nu_x$, $x = 0, 1, \dots, 7$. In this case, the system (7) has solutions (in the cube $0 \leqslant E_x \leqslant 1$, $x = 0, 1, \dots, 8$) if and only if $0 \leqslant r_8 \leqslant 0.0008$. Hence, we have, from the system (7) (under the conditions $r_x \propto \nu_x, x = 0, 1, \dots, 7, 0 \leqslant r_8 \leqslant 0.0008$) : $0.4945 \leqslant \widetilde{E}_0 \leqslant 0.4947$,

$0.2510 \leqslant \widetilde{E}_1 \leqslant 0.2512, 0 \leqslant \widetilde{E}_2 \leqslant 0.0008, 0.9104 \leqslant \widetilde{E}_3 \leqslant 1$, $0.7507 \leqslant \widetilde{E}_4 \leqslant 0.7790$, $0.4988 \leqslant \widetilde{E}_5 \leqslant 0.5128$, $0.1852 \leqslant \widetilde{E}_6 \leqslant 0.2074$, $0 \leqslant \widetilde{E}_7 \leqslant 0.1075, \widetilde{E}_8 = 0$. [In this example, the initial data are extracted from tables of random numbers ; the exact values are $E_0 = 0.5, E_1 = 0.25, E_2 = 0, E_3 = 1, E_4 = 0.75, E_5 = 0.5, E_6 = 0.25$, $E_7 = E_8 = 0$]. We see, in this case, the usual plan is bad. The correct plan is "If $X = 2$ or 7 accept the lot, otherwise reject", or the plan "Accept the lot if and only if $X = 2$".

We examined the large sample case. If the sample size, m, is not very large then estimates of E_x's are not very reliable. In this situation, it is expedient to construct a confidence set for the probabilities, r_x's, and to inspect the set of the stright lines (7) corresponding to this confidence set. The general question is : what is the confidence set for $(r_0, r_1, \ldots, r_n)$ if it known that $\nu_0, \nu_1, \ldots, \nu_n$ are multinomial random variables, and their likelihood function is $(\Sigma\ \nu_x)\,!\,\Pi(r_x^{\nu_x}/\nu_x!)$? This problem is difficult because some r_x's may be small. However, in the case of ordinary plans, we may be use the largest probabilities, r_x's, only.

Example (the ordinary plan). – $c = 1$, $n = 100$, $\nu_0 = 68$, $\nu_1 = 28$, $\nu_2 = 3$, $\nu_3 = \nu_4 = \nu_5 = 0$, $\nu_6 = 1$, $\nu_7 = \ldots = \nu_{100} = 0$. We have, from the system (7) under the condition $0 \leqslant E_2 \leqslant 1$: $0 \leqslant 49.5E_1 \leqslant r_2/r_1$, $0 \leqslant 100E_0 \leqslant r_1/r_0$. The conditional distribution of ν_0 (given $\nu_0 + \nu_1 = 96$) is binomial with the likelihood function $\binom{\nu_0 + \nu_1}{\nu_0} r_0^{\nu_0} r_1^{\nu_1} \Big/ (r_0 + r_1)^{\nu_0 + \nu_1}$. If λ_0 is a lower confidence limit for the probability $r_0/(r_0 + r_1)$ then $(1/\lambda_0 - 1)$ is an upper confidence limit for the ratio r_1/r_0 ; therefore, in our case, $E_0 < 0.01(1/\lambda_0 - 1)$. The upper confidence limit for E_1 may be constructed in an analogues manner. If the confidence coeffitions are equal 0.95 then we have, in this example, $E_0 < 0.00605$ and $E_1 < 0.00614$. In particular, if the lot size is $N = 1{,}000$ and controled objects are destroied then the number of accepted objects is equal $96 \cdot 900 = 86{,}400$, and the mean value of the number of accepted defective objects is not more than 525, i.e. the mean defective proportion (in the set of accepted objects) is not more than 0.61 % ; the confidence coefficient is not less than 0.9. An upper bound for the veritable number of accepted defective objects may be constructed with help of the usual technique of tolerance limits.

This paper deals with simplest cases of applications of the e.B.p. In particular, the questions connected with testing of the sample representation and as the problem of double samples so, generally, the sequential analysis were not considered. The purpose of this report is not complete describing of all cases but demonstrating that the e.B.p. is applicable and, therefore, is must be investigated.

Many thanks are due to Professor C.R. Rao for encouragement and discussion of this paper in the Indian Statistical Institute in February 1970.

REFERENCES

[1] BERNSTEIN S.N. — On the Fisher's fiducial probabilities, *Tidings of Acad. Sci. U.S.S.R.*, mathem. ser., 5, 1941, p. 85-94, (In Russian).

[2] ROBBINS H. — An empirical Bayes approach to statistics, *Proc. 3d Berkely Symp.*, 1955, p. 157-163.

[3] KAGAN A.M. — On the scheme by Robbins, *Reports of Acad. Sci. U.S.S.R.*, 150 : 4, 1963, p. 733-735. (In Russian).

[4] KOLMOGOROV A.N. — *The Statistical Sampling in the Case When the Admissible Number of Defective Objects is Zero,* Leningrad, House Sci. and Techn. Propaganda, 1951. (In Russian).

[5] SIRAZDINOV S.H. — Unbias estimates of the defective proportion among accepted objects; one sample case, *Proc. Inst. Mathem. and Mechan.*, Acad. Sci. Uzbek S.S.R., 20, 1957, p. 89-100. (In Russian).

[6] GNEDENKO B.V., BELJAEV Yu. K., SOLOVIEV A.D. — Mathematical Methods of the Reliability, Moscow, *Nauka,* 1966. (In Russian).

Steklov Mathematical Institute
Vavilov street 42,
Moscow V 333 (URSS)

Actes, Congrès intern. Math., 1970. Tome 3, p. 249 à 254.

OPTIMUM EXPERIMENTAL DESIGNS

by J. KIEFER

1. Introduction.

From the mathematical viewpoint, optimum design theory studies the geometry of generalized moment spaces $\mathfrak{M}$ of certain collections Ξ of probability measures. From the practical viewpoint, the consequence is often to find experimental designs which are more efficient than those which were used classically in statistics ; such usage was motivated by considerations of computation which are much less relevant with modern machines, or by such intuitively attractive properties of certain combinatorial designs as symmetries which need not guarantee optimality.

We will try to mention some developments obtained since the subject was revived almost 20 years ago by Mood, Elfving, and Chernoff after a 30-year pause from the early results of K. Smith, and will select results which lead to interesting and open mathematical questions. For brevity, we will not list all references in detail, but only a few whose bibliographies list most of the others. Even so, there is not space to mention many important areas of development, in particular designs for nonlinear models and sequential designs.

Let $f = (f_1, f_2, \ldots, f_k)$ where the f_i are continuous real functions on a compact space $\mathfrak{X}$. A point $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$ in R^k is unknown and its value is the concern of the statistician, who must choose an element (= "exact design") $x^* = (x_1, x_2, \ldots, x_N)$ in a specified subset $\mathfrak{X}^*$ of $\mathfrak{X}^N$. He then observes an N-vector $Y(x^*)$ of uncorrelated *r.v.*'s $Y_1, Y_2, \ldots, Y_N$ with common (known or unknown) variance σ^2, and with Y_i having expectation $(f(x_i), \theta) = \sum_{j=1}^{k} \theta_j f_j(x_i)$. For general statistical decision problems, the exact specification of the possible probability laws of $Y(x^*)$, not just of the possible values of the first two moments, is necessary ; moreover, in even very simple settings, the choice of design x^* and the possibility that randomization among designs may be called for, will depend strongly on the decision space (see [5]). However, we restrict attention here to problems of *point estimation* of a collection of linear forms $\{(c, \theta), c \in \mathcal{C}\}$, where $\mathcal{C}$ is a specified datum of the problem. Moreover, we will be concerned with nonrandomized linear estimators $(Y(x^*), h_c) = \sum_{1}^{N} h_{ci} Y_i$ of the (c, θ), and of the first and second moments of these estimators ; this can be justified (see [5], [8]) in Gaussian and certain nonparametric settings by an appeal to invariance, minimax, or various other principles for optimality criteria based on risk functionals of expected quadratic forms in the estimation errors. These restrictions reduce considerations to the first two moments of $Y(x^*)$.

The matrix A_{x*} of elements

$$(1) \qquad a_{x*rs} = \sum_{i=1}^{N} f_r(x_i) f_s(x_i)$$

is called the *information matrix* of the design x^*. *We shall treat here only problems where the performance of a design x^* is measured by a functional of A_{x*}*. This is motivated by the well known fact that the variance $\sigma^2 v(c, x^*)$ of the best linear unbiased estimator (= least squares or Gauss-Markov estimtor) of (c, θ) is $\sigma^2(c, A_{x*}^{-1} c)$ if A_{x*} is nonsingular, with an obvious analogue if A_{x*} is singular but (c, θ) has a linear unbiased estimator. (If no such estimator exists, we define $v(c, x^*) = +\infty$). Thus, duplicating a standard notion of statistical decision theory, we define x^* to be *at least as good as* x^{**} for $\mathcal{C}$, abbreviated $x^* \succ x^{**}$ $(\mathcal{C})$, if

$$(2) \qquad v(c, x^*) \leqslant v(c, x^{**}) \ \forall \ c \in \mathcal{C}.$$

In particular, for $\mathcal{C} = R^k$,

$$(3) \qquad x^* \succ x^{**}(R^k) \Leftrightarrow A_{x*} - A_{x**} \geqslant 0,$$

where we have written $A \geqslant 0$ if A is nonnegative definite. (In certain models considered by Box, Draper, and others, where restriction of the h_c to certain subspaces of R^N results in biased estimators, a functional of A_{x*} other than v is appropriate.)

Admissibility and complete classes of designs are defined as in decision theory, using maximality under the ordering $\succ$ of (2). If $\mathcal{C}$ is not a single element (or a set of multiples thereof), the partial ordering induced by (2) typically yields a "minimal complete class" consisting of more than one maximal A_{x*}, and in any particular application one must then choose the design x^* by reference to an optimality criterion induced by a real functional Φ on the space of variance functions $v(\cdot\,; x^*)$ on $\mathcal{C}$: We say x^* is *Φ-optimum for* $\mathcal{C}$ if it minimizes $\Phi(v(\cdot\,; x^*))$ over $\mathcal{X}^*$.

The above developments, and those which follow, allow considerable generalization. Results like those below (for example, Theorem 2) can still be obtained when Y_i and $(f(x_i), \theta)$ are replaced by vectors, often when $\mathcal{X}$ is not compact, and also when the cost per observation is not constant as implied above, and the covariance matrix of $Y(x^*_N)$ is not of the form $\sigma^2 I_N$. (Explicit characterization of Φ-optimum designs in terms of a given covariance function on $\mathcal{X} \times \mathcal{X}$, when $\mathcal{X}$ is a real interval, is obtained in a series of penetrating papers by Sacks and Ylvisaker.) But the most interesting questions are in fact often best understood in terms less of generalities than of some very particular examples we shall discuss in the next two sections.

2. The approximate theory.

Often there is no restriction among the x_i's and hence $\mathcal{X}^* = \mathcal{X}^N$. Defining the discrete probability measure ξ_{x*} on $\mathcal{X}$ by $N\xi_{x*}(x) =$ [number of i such that $x_i = x$], we see that $N^{-1}A_{x*}$ equals

$$M(\xi) = \int_{\mathscr{X}} f(x)' f(x) \, \xi(dx) \tag{4}$$

with $\xi = \xi_{x^*}$. Let Ξ denote the set of all probability measure on $\mathscr{X}$ relative to a σ-field which includes all one-point sets, and let Ξ_N denote the probability measures with range contained in the multiples of $1/N$. In the discussion of (2), (3), and the paragraph following them, we can clearly replace $\{A_{x^*}, x^* \in \mathscr{X}^N\}$ by $\{M(\xi), \ \xi \in \Xi_N\}$. If we replace Ξ_N by Ξ in these considerations, we have what is called the *approximate theory,* and an element of Ξ is called an *approximate design.* Whereas the exact design theory usually presents difficult combinatorial questions with solutions which exhibit a fine-structure dependence on N, the approximate theory often admits simple computational algorithms as discussed below ; and a single Φ-optimum approximate design ξ^* can easily be translated into an exact design for each N which, for suitably regular Φ (such as Φ_1 and Φ_2 below), minimizes $\Phi((\nu(\cdot, x^*))$ to within a relative error of $O(N^{-2})$.

The set $\mathfrak{M} = \{M(\xi), \xi \in \Xi\}$ can be viewed as the convex moment space associated with the functions $f_i f_j$, $1 \leqslant i \leqslant j \leqslant k$ on $\mathscr{X}$. The admissible designs are certain boundary points of $\mathfrak{M}$ which can generally be obtained (Elfving, also in [5] with misprints, and [3]) as measures supported by a subset of $\mathscr{X}$ on which $(f(x), f(x)B)$ attains its maximum for some matrix $B \geqslant 0$; as in other such characterizations, if B is not *positive* definite additional criteria are required. Of great interest is the question of the minimum number L such that, given any ξ, there is a ξ' which is (i) at least as good or (ii) at least as Φ-good as ξ, supported by at most L points. Simple examples [1] show that L can be as bad as the obvious bound $k(k+1)/2$ for quite common Φ. In the case of certain well known moment spaces it is of course known that a smaller L suffices, and sometimes one can be even more precise about admissibility :

THEOREM 1 [5]. – *If* $\mathscr{X} = \{x : -1 \leqslant x \leqslant 1\}$ *and* $f_i(x) = x^{i-1}$, *then* ξ *is admissible if and only if the open interval* $(-1, 1)$ *contains at most* $k-2$ *points of the support of* ξ.

This has been extended to weighted polynomials [3], splines, and other settings by Ehrenfeld, Karlin, Studden, Van Arman, Murty. But if f is a vector of polynomials on a Euclidean set $\mathscr{X}$ of dimension > 1, the problem is still far from solved [2].

Turning to specific optimality criteria Φ, among those most often encountered are (expressed as functionals on $\mathfrak{M}$) :

$$\begin{aligned} &\Phi_1(M) = tr\ BM^{-1} \qquad (B \geqslant 0), \\ &\Phi_2(M) = \det M^{-1} \qquad (\text{“}D\text{-optimality”}), \\ &\Phi_3(M) = \max_{c \in \mathcal{C}} \ (c, cM^{-1}). \end{aligned} \tag{5}$$

When B is of rank 1, a Φ_1-optimum approximate design yields optimum estimation of (c, θ) where $c' c = B$, and it was shown in [8] that this is equivalent to a Chebyshev approximation problem.

When $\mathcal{C} = \{f(x), \ x \in \mathscr{X}\}$, a Φ_3-optimum approximate design minimizes the maximum over $\mathscr{X}$ of the variance of the best estimator of $(f(x), \theta)$. In this case, we have [9].

THEOREM 2. – *If* $\mathcal{C} = \{f(x),\ x \in \mathscr{X}\}$, *then, in the approximate theory,*

(6) ξ^* is Φ_3*-optimum* $\Leftrightarrow$ ξ^* is Φ_2*-optimum* $\Leftrightarrow$ $\max\limits_{c \in \mathcal{C}} (c, c M^{-1}(\xi^*) = k.$

This turns out to yield a useful computational technique, especially the last equivalence. In [7] an analogous technique was obtained when $\Phi_2(M)$ is replaced by the determinant of a principal $s \times s$ minor of M^{-1}, corresponding to concern with s out of the k θ_i's. An alternative technique was given in [3] (corrected in [1]), but it remains to obtain a more useful algorithm than either of these, which are often much more difficult to apply than Theorem 2 to which they reduce when $s = k$. (When $s = 1$ they reduce to the Chebyshev problem mentioned above.)

There are many interesting optimality questions in even the simple model where f consist of polynomials on a d-dimensional ball, cube, or simplex [2]. In particular, the simplex yields the most elegant characterizations, Φ_2-optimum designs seeming to possess a regularity as d increases which is absent for the ball or cube. Uranisi and Atwood obtained related results.

If we think of $\mathscr{X}$ as a subset of $\bar{\mathscr{X}}$ on which f is defined, and if $\tilde{\mathscr{X}}$ is a subset of $\bar{\mathscr{X}}$, then Φ_3-optimality with $\mathcal{C} = \{f(x),\ x \in \tilde{\mathscr{X}}\}$ refers to *extrapolatory* estimation of $(f(x), \theta)$ on $\bar{\mathscr{X}}$. In [10] this was studied for the univariate polynomial model of theorem 1 when $\tilde{\mathscr{X}} = [-a, a]$, but the solution is difficult except when $a \to 0$ or $a \to 1$ or $a \to +\infty$. (This was recently extended to dimension $d > 1$). Hoel and Levine discovered the striking fact that the less symmetric problem $\tilde{\mathscr{X}} = \{a\}$ (or sometimes $\tilde{\mathscr{X}} = [-1, a]$) with $a > 1$ has a much more elegant solution, the Φ_3-optimum design being supported by the same "Chebyshev set" Λ_{k-1} (say) that supported the Φ_1-optimum design when $c = (0, \ldots, 0, 1)$. This result was extended to Chebyshev systems f in [11], where the set of all c for which the Φ_1-optimum design (with $B = c'c$) is supported by Λ_{k-1} is characterized. Among more recent results, we mention Studden's characterization for the univariate polynomial case (and certain other Chebyshev systems), that the optimum design for estimating θ_{k-i} is supported by Λ_{k-1} or Λ_{k-2} depending on whether i is odd or even $(i < k)$.

We mention finally, in the approximate theory, that there is a simple invariance result [5], [6] : Often there is a compact group G which operates on $\{\mathscr{X}, f, \Phi\}$ in such a manner that, because of the convexity in α of

$$M^{-1}(\alpha\xi_1 + (1-\alpha)\xi_2)$$

and of $-\Phi$ (or an increasing function of it), there is a Φ-optimum ξ which is G-invariant ($\xi(A) = \xi(gA)$, $g \in G$). Thus, for f consisting of all polynomials of degree $\leqslant m$ on the d-ball, one concludes that there is a Φ_2-optimum design consisting of multiples of Lebesgue measure on $(m+1)/2$ spheres of dimension $d-1$, where the origin counts as half a sphere [6]. What remains is then the more difficult problem of implementing this with a design of small finite support on those spheres, and with the same relevant moments [2].

3. Exact theory.

We illustrate the ideas with a setting in which combinatorial block designs are often used, the model of *2-way heterogeneity without interactions.* Here $N = k_1 k_2$ where the k_i are positive integers, and $k = k_1 + k_2 + u$ with $u > 1$. We rewrite $\theta = (\alpha_1, \ldots, \alpha_{k_1}, \beta_1, \ldots, \beta_{k_2}, \gamma_1, \ldots, \gamma_u)$. The space $\mathfrak{X}$ consists of triples (r, s, t) of integers with $1 \leqslant r \leqslant k_1$, $1 \leqslant s \leqslant k_2$, $1 \leqslant t \leqslant u$. However, in $\mathfrak{X}^*$ there is only one t, say $t(r, s)$, for each (r, s). One thinks of x^* as being represented by a $k_1 \times k_2$ array with entries $t(r, s)$. The expectation of the Y_i corresponding to position (r, s) is $\alpha_r + \beta_s + \gamma_{t(r,s)}$, a sum of "row, column, variety effects" where γ_t represents the contribution to Y_i of this t^{th} of u "varieties" (perhaps of grains being planted in the N positions of the rectangular array). Thus, $f_i(r, s, t(r, s))$ is an appropriate vector of 0's and 1's. The object in this experiment is usually to estimate the "variety contrasts" $\Sigma\, a_i\, \gamma_i$ with $\Sigma\, a_i = 0$. Let $c^{(1)}, c^{(2)}, \ldots, c^{(u-1)}$ be orthonormal u-vectors orthogonal to $(1, 1 \ldots, 1)$. Let $\sigma^2 H_{x*}$ be the covariance matrix of best linear estimators of the $u - 1$ contrasts $(c^{(i)}, \hat{\gamma})$. An invariance analysis related to that at the end of the last section shows easily, with $\Phi_i'(x^*)$ denoting the functionals of (5) with M^{-1} replaced by H_{x*},

THEOREM 3. – *If $\overline{x}^*$ maximizes tr H_{x*}^{-1} and $H_{\overline{x}*}^{-1} =$ const. I_{u-1}, then $\overline{x}^*$ minimizes $\Phi_1'(x^*)$, $\Phi_2'(x^*)$, and $\Phi_3'(x^*)$ with $\mathcal{C} = \{c : (c, c) = 1\}$.*

The computational importance of this is that *tr* H_{x*}^{-1} is easily computed from A_{x*} without inversion. In a less complex setting such as that where all α_i are assumed to be zero ("one way heterogeneity"), it is easily seen that a balanced block design with block size k_1 (appropriate generalization of balanced incomplete block design if $k_1 > u$) satisfies the hypothesis of Theorem 3. In our present context, though, the situation is much more complicated. A "generalized Youden square" (G Y S) is defined to be a balanced block design with respect to both rows and columns. In [4] we proved

THEOREM 4. – *If $u | k_1$ or $u | k_2$, and if a G Y S x^* exists, then it satisfies the hypothesis of Theorem 3.*

Recently we have shown that, although a G Y S is still Φ_3-optimum in the sense of Theorem 3 if the hypothesis of Theorem 4 is violated, *it need not be Φ_2-optimum.* The importance of this is that *exotic combinatorial designs with appealing symmetry properties need not be optimum for quite common criteria.*

In the above setting, even of one-way heterogeneity, there remains the important problem of design construction, on which we have made some progress over the work of Shrikhande and of Agrawal. The literature contains hundreds of papers on the case $k_1 < u$ for every one on $k_1 > u$.

Also of concern is the question of what to do when no balanced design exists. We have extended the idea of the first paragraph of Section 2 to give bounds on departure from optimality of certain unbalanced designs. This was also considered by Shah, and K. Takeuchi gave optimality proofs for certain partially balanced designs.

One of the most striking recent investigations concerns the possibility, in the settings of Section 2 (for example, of Theorem 1) that an exact design ξ' in Ξ_N with support identical to that of a Φ-optimum approximate design ξ^*, and with values appropriately "as close as possible" to those of ξ^*, is *exactly* Φ-optimum. The result is not always true, but Salaevskii showed for $\Phi = \Phi_2$ and the polynomial setting of Theorem 1 that it is always true for N sufficiently large ! Some special results for small N have been obtained by Granovskii.

REFERENCES

[1] ATWOOD C.L. — *Ann. Math. Statist.*, 40, 1969, 1570.

[2] FARRELL R.H., KIEFER J. and WALBRAN A. — *Proc. 5th Berkeley Symp.*, 1, 1965, p. 1-13.

[3] KARLIN S. and STUDDEN W.J. — *Ann. Math. Statist.*, 37, 1966, p. 783.

[4] KIEFER J. — *Ann. Math. Statist.*, 29, 1958, p. 675.

[5] KIEFER J. — *J.R.S.S. Series B*, 21, 1959, p. 272.

[6] KIEFER J. — *Proc. 4th Berkeley Symp.*, 1, 1960, p. 381.

[7] KIEFER J. — *Can. J. Math.*, 14, 1962, p. 597.

[8] KIEFER J. and WOLFOWITZ J. — *Ann. Math. Statist.*, 30, 1959, p. 271.

[9] KIEFER J. and WOLFOWITZ J. — *Can. J. Math.*, 14, 1960, p. 363.

[10] KIEFER J. and WOLFOWITZ J. — *Ann. Inst. Stat. Math.*, 16, 1964, p. 79-295.

[11] KIEFER J. and WOLFOWITZ J. — *Ann. Math. Statist.*, 36, 1965, p. 1627.

Cornell University
Dept. of Mathematics,
White Hall
Ithaca, N.Y. 14850 (USA)

Actes, Congrès intern. Math., 1970. Tome 3, p. 255 à 258.

SOME RECENT DEVELOPMENTS IN THE SEQUENTIAL ESTIMATION THEORY

by Yu V. LINNIK

Some new interesting facts about the sequential estimation processes were discoverd recently. Here one can note two directions : the asymptotic investigations (see for instance [1], [2], [3], [4]) and "exact formulas" (see [5], [6], [7], [8]). We shall indicate here some new results in both directions. We shall consider the simplest scheme of the sample with replacement (however, one of the results will relate to the samples without replacement). However, these results can be extended to the processes with independent increments and continuous time.

Thus, we consider the repeated scalar sample $X_1, X_2, \ldots$ from the family $\mathfrak{P}_\theta$ of distributions characterized by the density $f(x, 0)$ with respect to the Lebesgue or the counting measure. The parameter θ will be also a scalar one belonging to an interval Θ except in the multinomial case.

We shall consider a sequential estimation plan S as a pair $\{\tau, T_\tau\}$ consisting of a Markov stopping time τ and the statistic T_τ which is an unbiased estimate of a scalar function $g(\theta)$ of the parameter θ :

$$E_\theta(T_\tau) = g(\theta) \tag{1}$$

We consider the problem of choosing the plan $S = \{\tau, T_\tau\}$ minimizing the variance $D_\theta(\widetilde{T}_\tau)$ under condition

$$E_\theta \tau \leqslant n, \tag{2}$$

n being a given number.

There is no such optimal plan as a rule, but if we consider the asymptotic optimality for S, then the situation is changed. For the Bayesian set-up the corresponding results were proved by P. Bickel and I. Yahav ([1] - [4]) ; we shall indicate here the asymptotic results for the above mentioned non-Bayesian set-up). These results were obtained recently by the author and Professor I.V. Romanovsky [9]. We shall indicate also certain recent "small sample" results, and, in particular the sequential binomial estimation, a generalization of Bernstein polynomials in connection with the sequential analysis, which might be curious.

In regard to the asymptotic results, the principal point of view in this respect, which is illustrated in this communication, is that if there are no discontinuities in the information quantities (in the sense made precise below), the effects of introducing the sequential analysis will be relatively infinitely small as compared with the constant sample size method. If however, these discontinuities are present, the sequential estimation method may lead to a considerable gain in the estimate variance.

We shall now make precise the notion of the absence of the discontinuities in the information quantities. By this we shall mean the fulfillment of the conditions of I.A. Ibragimov and R.Z. Hasminski introduced in their interesting recent article on the generalized Bayes estimates for the constant sample size [10]. These conditions are the following :

1° The density $f(x, \theta)$ is integrable with respect to both arguments and $\int\int |\theta| \, f(x, \theta) \, f(x, \theta_0) \, dx \;\; d\theta < \infty$ for any $\theta_0 \in \Theta$

2° $\lim\limits_{|\theta| \to \infty} \int f(x, \theta) \, f(x, \theta_0) \, dx = 0$ if $\Theta \in (-\infty, \infty)$

3° $\lim\limits_{\epsilon \to 0} \int f^{1-\epsilon} (x, \theta_0) \, dx = 1$

4° For each set of real numbers $\theta_1, \ldots, \theta_s$ one can find such intervals $[0, T_1], \ldots, [0, T_s]$ that for all $t_i \in [0, T_j]$ for $\xi \downarrow 0$

$$\int \prod_{j=1}^{s} \left(\frac{f(x, \theta_0 + \xi \theta_j)}{f(x, \theta_0)}\right)^{t_j} f(x, \theta_0) \, dx = 1 + \xi^a a(t_1, \ldots, t_s, \theta_1, \ldots, \theta_s) + 0(\xi^a)$$

where the number α does not depend upon $\theta_1, \ldots, \theta_s$ and $a(\cdot) \neq 0$

5° For any θ_1, θ_2 with $|\theta_i| \leqslant H < \infty \, (i = 1, 2)$ we have for $\xi \to 0$

$$\int ln^r \left(\frac{f(x, \theta_0 + \xi \theta_1)}{[f(x, \theta_0 + \xi \theta_2)]}\right) f(x, \theta_0 + \xi \theta_2) \, dx| \leqslant |\xi|^a C_r(|\theta_2 - \theta_1|) \quad ; \quad r = 1, 2$$

where $[f(x, \theta_0 + \xi\theta_2)]$ concides with $f(x, \theta_0 + \xi\theta_2)$ for $f(\cdot, \cdot) \neq 0$ and equals 1 for $f(\cdot, \cdot) = 0$. The functions $C_r(h)$ may depend upon H and $C_r(h) \to 0$ for $h \to 0$ $(r = 1, 2)$.

We form now the Pitman estimate for the parameter θ :

$$\widetilde{\theta}_n = \int \theta \, \rho_n(\theta) \, d\theta$$

where $\rho_n(\theta) = \prod\limits_{i=1}^{n} f(x_i, \theta) \, d\theta \left(\int \prod\limits_{i=1}^{n} f(x_i, \theta) \, d\theta\right)^{-1}$

and, putting $l(x, \theta) = ln f(x, \theta)$, we require the finiteness of the moments

$$E_\theta \, |l^{(k)}(x_i, \theta_0)|^s \; ; \; k \leqslant s \; ; \; s \leqslant 20$$

$$E_\theta \max_{|\theta| \leqslant \epsilon} |l_\theta^5(x, \theta_0 - \theta)|^s \quad ; \quad s \leqslant 20. \quad \max_\theta E \widetilde{\theta}_n^2$$

Then the following theorems hold :

Theorem 1. – *Suppose θ to be a location parameter ($f(x, \theta) = f(x - \theta)$. Then $\widetilde{\theta}_{[n]}$ is an unbiased estimate of $y(\theta) = \theta$ and for any sequential estimation plan under conditions 1° and 2° we have :*

$$D(T_\tau) \geqslant D(\widetilde{\theta}_{[n]}) \left(1 + 0\left(\frac{1}{n}\right)\right) \tag{3}$$

THEOREM 2. – *In the general case, $\widetilde{\theta}_n$ is an estimate of θ which is asymptotically unbiased up to* $0\left(\frac{1}{n}\right)$ *and*

$$E(T_\tau - \theta)^2 \geqslant E(\widetilde{\theta}_{[n]} - \theta)^2 \left(1 + 0\left(\frac{1}{n}\right)\right) \tag{4}$$

Hence in case of the absence of discontinuities in the above indicated sense, the asymptotically inbiased sequential estimation cannot give relative gain in mean square deviation more than $0\left(\frac{1}{n}\right)$. This is not so when these discontinuities are present. Here is an interesting example due to A.I. Shalyt. Consider the distribution with the density $f(x - \theta)$ concentrated on the carrier $|x - \theta| \leqslant \frac{1}{2}$ which is continuous, symmetric and such that $f(x - \theta) = 1$ for $|x - \theta \pm \frac{1}{2}| \leqslant \epsilon_0$; $\epsilon_0 > 0$. Then the Pitman estimate is unbiased and optimal in the sense of variance among all regular estimates (i.e. estimates responding with a shift to the shift in the observations). We have here no analogae of the Rao-Cramer-Wolfowitz information unequalities because the information quantities are infinite. In using the sequential estimation plans under conditions 1° and 2° ($g(\theta) = \theta$), we can effectuate a threefold gain in variance in comparison with the fixed sample size method.

The investigation of the possibility of improvement of the estimate by applying sequential procedures in the presence of the discontinuities of the information quantities seems to be a very interesting subject for study.

Some new small sample ("exact") results were obtained recently in the case of binomial, multinomial and Poisson processes and first hit sequential estimation plans (see [7], [8]). In particular, some theorems on the determination of a bounded binomial plan by the values of $E_p\,\tau$ as the function of the probability of one step to the right p, were proved. Thus, a bounded complete binomial plan is determined by the values of $E_p\,\tau$. These results were only partly extended to multinomial plans (see [7]). Some problems of completeness of the sequential first hit binomial plans were solved. These are connected with the following generalization of Bernstein polynomials related to such a plan S :

$$B_f^s(\xi) = \sum_{(x\,,\,y)\in\partial s} K_{0,0}(x\,,y)\, f\left(\frac{K_{1,0}(x\,,y)}{K_{0,0}(x\,,y)}\right)\, \xi^x(1-\xi)^y \tag{5}$$

where: f is a continuous function ; $(x\,,y)$ the point on the plan boundary ∂S ; $K_{\alpha,\beta}(x\,,y)$, $(\alpha\,,\beta) = (0\,,0)$ or $(1{,}0)$, the number of trajectories "inside" the plan S starting at $(\alpha\,,\beta)$ and terminating at $(x\,,y)$. Thus, $\frac{K_{1,0}(x\,,y)}{K_{0,0}(x\,,y)}$ is an unbiased estimate for ξ. We obtain the usual Bernstein polynomials for the fixed sample size plan S with the boundary : $\partial S : x + y = n$, $n > 0$ an integer.

As regards the sampling plans without replacement we have the following interesting theorem :

THEOREM 3. – *For a sequential binomial plan without replacement to be complete it is necessary and sufficient that it is complete for an ordinary binomial sampling (the sampling with replacement).*

LITTERATURE

[1] BICKEL P.I., YAHAV I.A. — Asymptotically pointwise optimal procedures in sequential analysis. *Proc. of the V-th Berkeley Symposium on Math. Stat. and Probability,* v. I, 1965, p. 401-413.

[2] BICKEL P.I., YAHAV I.A. — Asymptotically optimal Bayes and minimax procedures in sequential estimation. *Ann. Math. Stat.,* v. 39, No. 2, 1968, p. 442-457.

[3] BICKEL P.I., YAHAV I.A. — Some contributions to the Asymptotic theory of Bayes Solutions. *Z. Wahrscheinlichkeitstheorie verw.* Geb. II, 1969, p. 257-276.

[4] BICKEL P.I., YAHAV I.A. — On an A.P.O. rule in sequential estimation with quadratic loss, *Ann. Math. Stat.* v. 40, No. 2, 1969, p. 417-427.

[5] DE GROOT M.N. — Unbiased sequential estimation for binomial population. *Ann. Math. Stat.,* 30, 1959, p. 80-101.

[6] TRYBULA S. — Sequential estimation in processes with independent increments. *Rozprawy matematyezne,* v. 50, Warszawa, 1968.

[7] ZAIDMAN R.A., LINNIK YU. V., ROMANOVSKY I.V. — Plans of sequential estimation and Markov Stopping times. *Doklady A.N. S.S.S.R.,* 185, 6, 1969, p. 1222-1225.

[8] ZAIDMAN R.A., LINNIK YU. V., SUDAKOV V.N. — *On sequential estimation and Markov stopping moments for the processes with independent increments.* U.S.S.R. — Japan Symposium on Probability, Habarovsk, August, 1969; Edited by the U.S.S.R. Academy of Sciences, Novosibirsk, (1969), 122-143 (in Russian).

[9] LINNIK YU. V., ROMANOVSKY I.V. — A contribution to the theory of sequential estimation. *Doklady A.N. S.S.S.R.,* 1970, (in Russian).

[10] IBRAGIMOV I.A., HASMINSKI P.Z. — On the asymptotic behaviour of generalized Bayes estimates. *Doklady A.N. S.S.S.R.,* 1970 (in Russian).

Mathematical Institute Academia Nauk
Nab. Fontanki 25,
Leningrad D 11 (URSS)

Actes, Congrès intern. Math., 1970. Tome 3, p. 259 à 263.

ASYMPTOTICALLY EFFICIENT TESTS AND ESTIMATORS

by J. WOLFOWITZ

1. Introduction.

This paper is a brief report on the general method of obtaining asymptotically efficient (a.e.) estimators and tests, due to L. Weiss and the present author. The estimators obtained by this method have been called maximum probability (m.p.) estimators, and the authors have concentrated on them. Their paper [1] contains the essential reference to the latter, except for the later paper [2]. Although the authors are critical of the practical value of tests of significance, they applied their method to the theory of tests in [3], in order to demonstrate the power of the method.

The now classical, widely employed, and most important method of obtaining a.e. estimators (and tests) is that of maximum likelihood (m.l.). This brilliant method nevertheless has the following inadequacies :

(1) The theory applies only under very onerous regularity conditions (the so-called "regular" cases of Cramér [5] and others) which exclude many of the most frequent problems of statistics. For example, the case where the density, at a point x, of a chance variable whose distribution depends upon a parameter θ, is $e^{-(x-\theta)}$ when $x \geqslant \theta$ and zero otherwise, is not "regular" !

(2) Only estimators which are asymptotically normally distributed are allowed to enter into competition with the m.l. estimator. This is convenient for the theory and allows comparison on the basis of variances, but does not correspond to practical application or necessity. This requirement begs the question whether estimators which are not asymptotically normally distributed may not sometimes be more efficient.

(3) The classical results are largely limited to the case $m = 1$, where m is the dimension of the unknown parameter.

(4) The theory applies mainly to the case of independent, identically distributed chance variables.

M.p. estimators are not subject to these inadequacies, and are applicable in wide generality to a tremendous variety of problems, which include all the problems of the statistical literature. In the special, so-called regular case, and for special loss functions, m.p. estimators are m.l. estimators. The theorem on the asymptotic efficiency of the m.p. estimator then implies the classical result on the efficiency of the m.l. estimator.

2. Description of the general estimator.

We limit ourselves to the case $m = 1$, for simplicity of exposition, but the results are valid for general m.

For each positive integer n let $X(n)$ denote the (finite) vector of (observed) chance variables of which the estimator is to be a function. $X(n)$ need not have n components (although the number of components will approach infinity), nor need its components be independently or identically distributed. Let $K_n(x \mid \theta)$, a Borel measurable function of both arguments jointly, be the density, with respect to a σ-finite (positive) measure μ_n, of $X(n)$ at the point x (of the appropriate space) when θ is the value of the (unknown to the statistician) parameter. The latter is known to be a point of the "parameter space" Θ. Any estimator T_n is a Borel measurable function of $X(n)$ with values in $\bar{\Theta}$; the set Θ is a closed region of m-dimensional Euclidean space and is contained in a closed region $\bar{\Theta}$ such that every (finite) boundary point of Θ is an inner point of $\bar{\Theta}$.

Let $L_n(z\,,\theta)$ be a non-negative loss function, i.e., when the value of the estimator (function of $X(n)$) is z, and the value of the parameter which determines the density of $X(n)$ is θ, the loss incurred by the statistician is $L_n(z\,,\theta)$. In many problems one will have

$$L_n(z\,,\theta) = k(n)\; L(z\,,\theta),$$

where $k(n)$ ($\to \infty$) is a normalizing factor for the distribution of the m.p. estimator Y_n. For any $y > 0$ define

$$s_n^*(y) = \sup L_n(z\,,\theta),$$

the supremum being taken over all z and θ such that $|z - \theta| \leqslant y$.

Let $\{k_1(n)\}, \{k_2(n)\}$ be sequences of positive numbers such that, as $n \to \infty$,

$$k_2(n) \to \infty, \quad \frac{k_2(n)}{k_1(n)} \to 0, \quad \frac{k_1(n)}{k(n)} \to 0.$$

Write for brevity

$$h_1(n) = \frac{k_1(n)}{k(n)}\,, \quad h_2(n) = \frac{k_2(n)}{k(n)}$$

and

$$s(n) = s_n^*(h_2(n)).$$

The m.p. estimator Y_n is defined as a value of d which maximizes

$$\int_{d-h_2(n)}^{d+h_2(n)} [s(n) - L_n(d\,,\theta)]\; K_n(X(n) \mid \theta)\, d\theta.$$

3. Statement of asymptotic efficiency. Idea of the proof.

Under weak regularity conditions, and reasonable conditions on the competing estimator T_n, it is proved that, for any $\theta \in \Theta$,

$$\lim_{n\to\infty} E_\theta\{L_n(Y_n , \theta)\} \leqslant \underline{\lim}_{n\to\infty} E_\theta\{L_n(T_n , \theta)\}.$$

This inequality is obviously the statement of asymptotic efficiency.

Why is the m.p. estimator a.e. ? This can be understood from the simple proof, whose idea we now proceed to describe. Assume, but only as a mathematical device, that the unknown (constant) parameter θ is actually a chance variable, uniformly distributed over an interval of half-length $h_1(n)$, centered at the true value (!), and compute the Bayes solution with respect to this a priori distribution ! On the face of it, this is absurd ; if the true value of the parameter were known, it would be unnecessary to estimate it. It is proved, however, that, except for an event whose probability approaches zero as $n \to \infty$, the m.p. estimator Y_n is actually a Bayes solution. It is now intuitively obvious that, even under relatively weak smoothness properties of $K_n(\cdot \mid \cdot)$, the Bayes solution Y_n will be a.e. even when the parameter θ is an unknown constant.

Consider the problem of asymptotically testing a null hypothesis which is simple in certain parameters, against an alternative hypothesis which is simple in these same parameters, in the presence of other (nuisance) parameters. The method of the present paper, applied in [3], uses the following basic idea : One constructs a Bayes test, using suitable a priori uniform distributions on the nuisance parameters, centered at the true values of the latter. The latter values are then estimated by m.l. The resulting test is asymptotically minimax. (See [3] for details).

4. Connection with maximum likelihood. Examples of the m.p. estimator.

Suppose the conditions of the regular case are fulfilled. Let $a > 0$ be an arbitrary constant, and suppose $L_n(z , \theta)$ is zero or one according as $\sqrt{n}|z - \theta| \leqslant a$ or $> a$. Then $Y_n = \hat{\theta}_n$, the m.l. estimator. From the displayed equation in Section 3 it follows that, for any θ, the variance of the limiting normal distribution of Y_n (after appropriate normalization) is not greater than the corresponding variance of an asymptotically normal competing estimator.

Consider again the regular case, but with $m > 1$. Let R be any convex set in the space of the parameter, symmetric about the origin. Then, by proper choice of the loss function L_n, one can prove the result of Kaufman [6] :

$$\lim P_\theta \{\sqrt{n}\,(\hat{\theta}_n - \theta) \in R\} \geqslant \overline{\lim}\, P_\theta\{\sqrt{n}\,(T_n - \theta) \in R\},$$

where T_n is a competing estimator.

From now on, assume that $X(n) = (X_1, \ldots , X_n)$, the X's independently and identically distributed with density $f(\cdot \mid \theta)$ (Lebesgue measure).

Example 1. – Let $f(x \mid \theta) = \exp\{-(x - \theta)\}$, $x \geqslant \theta$, and zero otherwise. Let $\Theta = (-\infty , \infty)$. Let $L_n(z , \theta) = 0$ when $n|z - \theta| \leqslant a$, and $= 1$ otherwise. Then $\hat{\theta}_n = \min(X_1, \ldots , X_n)$, but $Y_n = \hat{\theta}_n - \dfrac{a}{n}\cdot$

EXAMPLE 2. – Let $m = 2$, and

$$f(x \mid \theta_1, \theta_2) = \frac{1}{2} \exp\{-(x - \theta_1)\} + \frac{1}{2\theta_2} \exp\left\{-\frac{(x - \theta_1)}{\theta_2}\right\}, \quad x > \theta_1,$$

and zero otherwise. Let

$$\Theta = \{(\theta_1, \theta_2) \mid -\infty < \theta_1 < \infty, \quad 0 < \theta_2 < \infty\}.$$

The m.l. estimator is not consistent. Let $z = (z_1, z_2)$ and $L_n(z, \theta) = 0$ when $|z_1 - \theta_1| \leqslant a_1$ and $|z_2 - \theta_2| \leqslant a_2$, and 1 otherwise. Here a_1 and a_2 are arbitrary positive constants. Then $Y_n = (Y_{n1}, Y_{n2})$, where $Y_{n1} = \frac{-a_1}{n} + \min\{X_1, \ldots, X_n\}$, and Y_{n2} is described on page 88 of [6].

EXAMPLE 3. – Let $m = 1$ again, $\Theta = (-\infty, \infty)$,

$$f(x \mid \theta) = (2\pi)^{-1/2} \exp\left\{-\frac{1}{2}(x -- \theta)^2\right\}, \quad \text{and} \quad L_n(z, \theta) = n(z - \theta)^2.$$

Then $Y_n = \hat{\theta}_n = n^{-1} \sum_1^n X_i$. Any competing estimator T_n which satisfies the weak regularity conditions must also satisfy

$$\varliminf_{n \to \infty} E_\theta\{(T_n - \theta)^2\} \geqslant 1.$$

Such an estimator ***need not be unbiased or asymptotically normally distributed.*** Of course, if T_n *is* asymptotically normally distributed about θ it would be more intelligent to concern one's self with the variance of the limiting distribution ; this can be done using a suitable weight function. If the limiting distribution of T_n is not normal about θ the second moment may not be the appropriate measure of loss. However, this does not affect the validity of our illustration.

REFERENCES

[1] WEISS L. and WOLFOWITZ J. — Maximum probability estimators with a general loss function, *Proceedings of the International Symposium on Probability and Information Theory,* held at McMaster University, Hamilton, Ontario, Canada, April 4 and 5, 1968, pages 232-256. Springer-Verlag, Berlin-Heidelberg-New York, Lecture Notes in Mathematics, 89, 1969.

[2] WEISS L. and WOLFOWITZ J. — Maximum probability estimators and asymptotic sufficiency, *Ann. Inst. Stat. Math.,* 22, No. 2, 1970.

[3] WEISS L. and WOLFOWITZ J. — Asymptotically minimax tests of composite hypotheses, *Zeitschrift für Wahrscheinlichkeitstheorie,* 14 (161-168), 1969.

[4] WEISS L. and WOLFOWITZ J. — Generalized maximum likelihood estimators, *Teoriya Vyeroyatnostey,* 10, No. 2, 1965, p. 267-281.

[5] CRAMÉR H. — *Mathematical methods of statistics,* Princeton University Press, 1946, Princeton, N.J.

[6] KAUFMAN S. — Asymptotic efficiency of the maximum likelihood estimator. *Ann. Inst. Stat. Math.,* 18, No. 2, 1966, p. 155-178.

University of Illinois
Dept. of Mathematics,
Urbana,
Illinois 61 801 (USA)

Actes, Congrès intern. Math., 1970. Tome 3, p. 265 à 267.

E7 - PROBLÈMES MATHÉMATIQUES DE LA THÉORIE DE L'INFORMATION LANGAGE MACHINE

ALGEBRAIC ASPECTS OF AUTOMATA THEORY

by Samuel EILENBERG

Let S be a monoid. The following operations on subsets of S will be called *rational operations* :

(1) Union : $A \cup B$

(2) Product : $AB = \{ab \mid a \in A\ ,\ b \in B\}$

(3) Closure : $A^+ = A \cup A^2 \cup A^3 \cup \ldots$

The class of *rational subsets* of S is defined as the least class containing all the subsets of cardinality $\leqslant 1$ and closed under the rational operations. If S is a finitely generated free monoid with base Σ, then the rational subsets of S are exactly the sets recognized (or accepted) by finite state automata (Kleene's Theorem).

Let S and S' be monoids and let $f : S \to S'$ be a relation. The graph $f^{\#}$ of f is then a subset of the product monoid $S \times S'$. The relation f is called *rational* if its graph is a rational subset of $S \times S'$. If $g : S' \to S''$ is another rational relation, the composed relation $fg : S \to S''$ need in general not be rational. However fg is rational if the monoid S' is free.

The last fact leads to the following formal development. We shall consider a fixed countably infinite set Σ_0 and denote by Σ_0^* the free monoid generated by Σ_0. A subset A of Σ_0^* is called *finitary* if $A \subset \Sigma^*$ for some finite subset Σ of Σ_0. All rational subsets are finitary. Also for every rational relation $f : \Sigma_0^* \to \Sigma_0^*$ there exists a finite subset Σ of Σ_0 such that $f^{\#}$ is contained in $\Sigma^* \times \Sigma^*$. For finitary subsets A, B (of Σ_0^*) we define $B \leqslant A$ (*A rationally dominates B*) if $B = Af$ for some rational relation f. This relation $\leqslant$ is reflexive and transitive. Rational equivalence $A \equiv B$ is defined by $A \leqslant B$ and $B \leqslant A$. The empty set $\emptyset$ is the smallest rational equivalence class. All non-empty rational subsets (of Σ_0^*) form a single equivalence class **Rat**. We denote by $\mathcal{H}_0$ the (partially) ordered set of rational equivalence classes of finitary subsets excluding the class of the set $\emptyset$. Thus **Rat** becomes the smallest element of $\mathcal{H}_0$. If $\mathbf{A}_1, \mathbf{A}_2 \in \mathcal{H}_0$ we may choose representatives $A_1 \in \mathbf{A}_1$, $A_2 \in \mathbf{A}_2$ which are in general position i.e. $A_1 \subset \Sigma_1^*$, $A_2 \in \Sigma_2^*$ where Σ_1 and Σ_2 are disjoint finite subsets of Σ_0^*. Then $A_1 \cup A_2$ represents a unique class $\mathbf{A}_1 \vee \mathbf{A}_2$ of $\mathcal{H}_0$ which is the least upper bound of $\mathbf{A}_1$ and $\mathbf{A}_2$ in $\mathcal{H}_0$. Thus $\mathcal{H}_0$ is an upper semi-lattice.

Purely formally one defines the *completion* $\mathcal{H}$ of $\mathcal{H}_0$ as follows. A$_n$ *ideal I* in $\mathcal{H}_0$ is a non empty subset of $\mathcal{H}_0$ such that

$$\mathrm{B} \leqslant \mathrm{A} \quad \& \quad \mathrm{A} \in I \Rightarrow \mathrm{B} \in I$$

$$\mathrm{A}_1 , \mathrm{A}_2 \in I \Rightarrow \mathrm{A}_1 \vee \mathrm{A}_2 \in I$$

The ideals of $\mathcal{H}_0$ (partially) ordered by inclusion form a complete lattice $\mathcal{H}$. The upper semi-lattice $\mathcal{H}_0$ is imbedded into $\mathcal{H}$ by assigning to each $\mathrm{A} \in \mathcal{H}_0$ the principal ideal

$$I\mathrm{A} = \{\mathrm{B} \mid \mathrm{B} \in \mathcal{H}_0, \quad \mathrm{B} \leqslant \mathrm{A}\}$$

This imbedding preserves the order and least upper bounds. In $\mathcal{H}$ all ideals are principal.

An element $\mathcal{C}$ of $\mathcal{H}$ may be viewed as a class of subsets of Σ_0^* satisfying the following conditions

(4) All sets in $\mathcal{C}$ are finitary.

(5) $\mathcal{C}$ contains at least one non-empty set.

(6) If $A \in \mathcal{C}$ then $Af \in \mathcal{C}$ for every rational relation $f : \Sigma_0^* \to \Sigma_0^*$.

(7) If $A_1, A_2 \in \mathcal{C}$ then $A_1 \cup A_2 \in \mathcal{C}$.

Such a class $\mathcal{C}$ of sets will be called a *cone.* The cone $\mathcal{C}$ is *principal* with A as generator if

$$\mathcal{C} = \{B \mid B = Af\}$$

with f ranging over all rational relations $f : \Sigma_0^* \to \Sigma_0^*$.

The lattice $\mathcal{H}$ is called the *rational hierarchy*, and its study may properly be defined as one of the main objectives of the theory of automata and abstract langages. Various algebraic operations may be introduced in $\mathcal{H}$ which thus acquires a rich algebraic structure. Most of the interesting classes of sets turn out to be principal cones with "interesting" generators.

As an example of a new problem arising from the point of view presented here, consider a cone $\mathcal{C}$ and define

$$\mathcal{C}' = \{B \mid B < A \quad \text{for some} \quad A \in C\}$$

where $B < A$ means $B \leqslant A$ but not $B \equiv A$. One easily observes that if $\mathcal{C}$ is not a principal cone then $\mathcal{C}' = \mathcal{C}$. If $\mathcal{C}$ is principal with generator A then $\mathcal{C}'$ is a cone iff A is indecomposable i.e. if

$$A = A_1 \cup A_2, \quad A_1 \leqslant A, \quad A_2 \leqslant A$$

imply

$$A_1 \equiv A \quad \text{or} \quad A_2 \equiv A.$$

This is equivalent with $\mathcal{C}$ itself being indecomposable in the sense that

$$\mathcal{C} = \mathcal{C}_1 \vee \mathcal{C}_2$$

implies $\mathcal{C} = \mathcal{C}_1$ or $\mathcal{C} = \mathcal{C}_2$. A recent result of Schutzenberger (unpublished) shows that the cone of algebraic sets (= context free languages) is indecomposable.

Rational relations were first studied by Elgot and Mezie. In a different terminology, cones have been studied intensively by Seymour Ginsburg and his pupils and collaborators. Essentially there is nothing new in this article except for a hopefully convenient algebraic setting.

Columbia University
522/ Dept. of Mathematics,
New York, N.Y. 10027 (USA)

Actes, Congrès intern. Math., 1970. Tome 3, p. 269 à 274.

THE ANALYSIS OF ALGORITHMS

by Donald E. KNUTH

Some general aspects of algorithmic analysis are illustrated by discussing Euclid's algorithm. Euclid's method is extended in such a way that the gcd of two n digit numbers can be found in $0(n(\log n)^5 (\log \log n))$ steps as $n \to \infty$.

The advent of high-speed computing machines, which are capable of carrying out algorithms so faithfully, has led to intensive studies of the properties of algorithms, opening up a fertile field for mathematical investigations. Every reasonable algorithm suggests interesting questions of a "pure mathematical" nature ; and the answers to these questions sometimes lead to useful applications, thereby adding a little vigor to the subject without spoiling its beauty. The theory of queues, which analyzes a very special class of algorithms, indicates the potential richness of the theories which can be obtained when algorithms of all types are analyzed in depth.

The purpose of this paper is to illustrate some general principles of algorithmic analysis by considering an example which is interesting for both historical and mathematical reasons, the calculation of the greatest common divisor (gcd) of two integers by means of Euclid's algorithm. Euclid's procedure [2], which is one of the oldest nontrivial algorithms known, may be formulated as follows, given integers $U > V \geqslant 0$:

E1. If $V = 0$, stop ; U is the answer.

E2. Let R be the remainder of U divided by V, so that $U = AV + R, 0 \leqslant R < V$. Replace U by V, then replace V by R, and return to E1.

1. "Local" analysis.

Analyses of algorithms are generally of two kinds, "local" and "global". A local analysis consists of taking one particular algorithm (like Euclid's) and studying the amount of work it does as a function of the inputs ; a global analysis, on the other hand, considers an entire family of algorithms and investigates the "best possible" procedures in that class, from some point of view. In both types of analysis we can consider either the "worst case" of the algorithms, namely the work involved under least favorable choice of inputs, or the "average case", the expected amount of work under a given input distribution. More generally, we may be able to obtain the distribution of work given the distribution of inputs. "Work" may be measured in terms of the number of times each step of the algorithm is performed, or the amount of things which need to be remembered, etc.

The first local analysis of Euclid's algorithm was published in 1844 by G. Lamé [10], who showed that step E2 will never be performed more than five times the number of digits in the decimal representation of V. His analysis was based on the fact that the method is least efficient when U and V are consecutive Fibonacci numbers.

The *average* behavior of Euclid's algorithm is much more difficult to determine than the worst case, and it has been established only in recent years. Let $T(U, V)$ be the number of times step E2 is performed. J.D. Dixon proved [1] that, for all ϵ and $C > 0$, the probability that

$$|T(U, V) - (12\pi^{-2} \ln 2) \ln U| \geqslant (\ln U)^{1/2+\epsilon} \quad \text{is} \quad 0((\ln N)^{-C}),$$

given that $1 \leqslant V \leqslant U \leqslant N$. His proof is based on careful refinements of Kuz'min's study of continued fractions [9], showing that partial quotients which are far apart in the sequence are nearly independent.

At about the same time, H. Heilbronn introduced a new approach [6] to the study of continued fractions and Euclid's algorithm.

Let

$$T(V) = \lim_{N\to\infty} \frac{1}{N} \sum_{U=V+1}^{V+N} T(U, V) = \frac{1}{V} \sum_{U=V+1}^{2V} T(U, V)$$

be the average number of times when V is fixed. Heilbronn showed in effect that

$$nT(n) = \lfloor 3n/2 \rfloor + 2\Sigma \left[\left(\frac{n}{y+t} - t'\right) \frac{1}{y} \right]$$

where $\lfloor x \rfloor$ is the greatest integer $\leqslant x$, $\lceil x \rceil$ is the least integer $\geqslant x$, and the sum is over all positive integers y , t , t' such that $\gcd(t, y) = 1$, $t \leqslant y$, $t' \leqslant y$, $tt' \equiv n$ (modulo y). Evaluating this sum, he essentially found that $T(n) = (12\pi^{-2} \ln 2) \ln n + 0(\sigma_{-1}(n)^2)$. Indeed, somewhat more seems to be true, although proof is still lacking ; there is extensive empirical evidence [8, pp. 330-333] that $\Big(\sum_{1\leqslant k\leqslant V,\ \gcd(k, V)=1} T(V+k, V)\Big)\Big/\varphi(V) = (12\pi^{-2} \ln 2) \ln V + 1.47 + 0(1)$ as $V \to \infty$.

2. "Global" analysis.

Is Euclid's algorithm the "best" way to calculate greatest common divisors? Analyses of other gcd algorithms (cf. [8]) shows that, under certain conditions, Euclid's method is inferior ; and the average behavior of an interesting new algorithm discovered by V.C. Harris [4] is still unknown.

In searching for a "best" method, one way to measure the work is to consider the amount of time taken to perform the algorithm with pencil and paper, or with a conventional computer. Various abstract automata have been proposed by which the latter notions can be made precise (cf. [5, 7]). When we apply such models to Euclid's algorithm, it is not difficult to see [8, p. 526] that the amount of work is essentially proportional to the square of the number of digits in U, for both the average case and the worst case, analogous to the familiar method of long division. On the other hand, extremely fast methods of multiplication and division have recently been discovered ; A. Schönhage and V. Strassen have proved [13] that an m-digit number can be multiplied by an n-digit number in only $0\,(n\,(\log m)\,(\log\log m))$ units of time, when $n \geqslant m > 1$. It is therefore natural to ask whether the *gcd* of two n-digit numbers can be calculated in less than $0\,(n^2)$ steps. Section 3 of this paper shows that this is indeed possible, in $0\,(n^{1+\epsilon})$ steps for all $\epsilon > 0$, by suitably arranging the calculations of Euclid's algorithm. Obviously at least n steps are necessary in any event (we must look at the inputs), so this result provides some idea of the asymptotic complexity of *gcd* computation.

3. High-speed gcd calculation with large numbers.

If step E2 is performed t times, let $A_1, \ldots, A_t$ be the partial quotients obtained. It is well known that $U = Q_t(A_1, \ldots, A_t)D$, $V = Q_{t-1}(A_2, \ldots, A_t)D$, where $D = \gcd(U, V)$ and Q_t is the continuant polynomial defined by $Q_{-1} = 0, Q_0 = 1$, $Q_{t+1}(x_0, x_1, \ldots, x_t) = x_0 Q_t(x_1, \ldots, x_t) + Q_{t-1}(x_2, \ldots, x_t)$. We shall call $[A_1, \ldots, A_t, D]$ the *Euclidean representation* of U and V. After k iterations of step E2 we have $U = U_k = Q_{t-k}(A_{k+1}, \ldots, A_t)\,D$, $V = V_k = Q_{t-k-1}(A_{k+2}, \ldots, A_t)\,D$. Euler [3] observed that $Q_t(x_1, \ldots, x_t)$ is the set of all terms obtainable by starting with $x_1 \ldots x_t$ and striking out pairs $x_i x_{i+1}$ zero or more times. From this remark, it follows immediately that

$$(*)\quad Q_{s+t}(x_1, \ldots, x_{s+t}) = Q_s(x_1, \ldots, x_s)\,Q_t(x_{s+1}, \ldots, x_{s+t}) + Q_{s-1}(x_1, \ldots, x_{s-1})\,Q_{t-1}(x_{s+2}, \ldots, x_{s+t}),$$

an identity which forms the basis of Heilbronn's work cited above ; it was used on several occasions by Sylvester [14] and given in more general form by Perron [12, p. 14-15].

For convenience we shall write nonnegative integers N in binary notation, using $lN \equiv \ulcorner \log_2(N+1) \urcorner$ binary digits. It is easy to prove that

$$lQ_t(A_1, \ldots, A_t) \leqslant lA_1 + \cdots + lA_t + 1,$$

and Lamé's theorem implies that $lA_1 + \cdots + lA_t \leqslant lQ_t(A_1, \ldots, A_t) + t = 0(\log U)$ in Euclid's algorithm ; hence (except for a constant factor) it takes essentially as much space to write down the Euclidean representation $[A_1, \ldots, A_t, D]$ as it does to write U and V themselves in binary form. We shall show that it is possible to convert rapidly between these two representations of U and V.

THEOREM 1. – *Let $S(n) = n\,(\log n)\,(\log\log n)$ and $n = lA_1 + \cdots + lA_t$. There is an algorithm which, given the binary representations of $A_1, \ldots, A_t$, computes the binary representation of $Q_t(A_1, \ldots, A_t)$ in $0\,(S(n)\,(\log t))$ steps.*

Proof. – Consider four continuants associated with $(A_1, \ldots, A_t)$, namely $Q = Q_t(A_1, \ldots, A_t)$, $Q^{\cdot} = Q_{t-1}(A_1, \ldots, A_{t-1})$, ${}^{\cdot}Q = Q_{t-1}(A_2, \ldots, A_t)$, and ${}^{\cdot}Q^{\cdot} = Q_{t-2}(A_2, \ldots, A_{t-1})$. The four continuants associated with $(0, A_1, \ldots, A_t)$

are the same, in another order, so we add zeroes if necessary until t is a power of 2. Now let $L, L^{\cdot}, {}^{\cdot}L, {}^{\cdot}L^{\cdot}$ and $R, R^{\cdot}, {}^{\cdot}R, {}^{\cdot}R^{\cdot}$ be the continuants associated with $A_1, \ldots, A_t$ and $A_{t+1}, \ldots, A_{2t}$ respectively. By (*), $Q = LR + L^{\cdot}\ {}^{\cdot}R$, $Q^{\cdot} = LR^{\cdot} + L^{\cdot}\ {}^{\cdot}R$, ${}^{\cdot}Q = {}^{\cdot}LR + {}^{\cdot}L^{\cdot}\ {}^{\cdot}R$, ${}^{\cdot}Q^{\cdot} = {}^{\cdot}LR^{\cdot} + {}^{\cdot}L^{\cdot}\ {}^{\cdot}R^{\cdot}$. Choosing C so that we can evaluate the L's in $CS(lA_1 + \cdots + lA_t)k$ steps, the R's in $CS(lA_{t+1} + \cdots + lA_{2t})k$ steps, and the 8 multiplications and 4 additions in $CS(lA_1 + \cdots + lA_{2t})$ further steps by the Schönhage-Strassen algorithm, we can evaluate the Q's in at most $CS(lA_1 + \cdots + lA_{2t})\ (k + 1)$ steps.

Let $U = 2^m U' + U''$, $V = 2^m V' + V''$, where $0 \leqslant U''$, $V'' < 2^m$. D.H. Lehmer [11] has suggested that the partial quotients for (U, V) be found by first obtaining some of those for U' and V', stopping at A_s where s is maximal such that $(U' + 1, V')$ and $(U', V' + 1)$ have $A_1, \ldots, A_s$ in common. Then $A_1, \ldots, A_s$ are partial quotients for (U, V) also. We shall call $(A_1, \ldots, A_s)$ the "Lehmer quotients" for (U', V'). The example $(U', V') = (2^m, 2^{m-1})$ shows that Lehmer quotients might not amount to anything, but we can prove that four additional Euclidean iterations will always give a useful reduction.

LEMMA 1. – *Let* $U = 2^m U' + U'' \geqslant V = 2^m V' + V''$, *where* $0 \leqslant U''$, $V'' < 2^m$. *Let* $[A_1, \ldots, A_t, D]$ *be the Euclidean representation of* (U, V), *and let* $(A_1, \ldots, A_s)$ *be the Lehmer quotients for* (U', V'), *where* $t \geqslant s + 4$. *Then* $U_{s+4} < U/\sqrt{U'}$.

Proof. – Let $P_k = Q_{k-1}(A_2, \ldots, A_k)$, $Q_k = Q_k(A_1, \ldots, A_k)$, and let $\theta = V/U$. The well-known pattern of convergence of P_k/Q_k to θ, schematically

$$\frac{P_k}{Q_k} < \frac{P_k + P_{k+1}}{Q_k + Q_{k+1}} < \frac{P_k + 2P_{k+1}}{Q_k + 2Q_{k+1}} < \cdots < \frac{P_{k+2}}{Q_{k+2}} < \theta < \frac{P_{k+2} + P_{k+1}}{Q_{k+2} + Q_{k+1}} < \frac{P_{k+1}}{Q_{k+1}}$$

when k is even, shows that if θ and θ' are two real numbers whose continued fractions differ first at $A_{s+1} \neq A'_{s+1}$, either P_{s+1}/Q_{s+1} or P_{s+2}/Q_{s+2} lies between θ and θ'. Hence

$$\frac{1}{2} Q^2_{s+4} \geqslant Q_{s+2}(Q_{s+3} + Q_{s+2}) > 1/|\theta - P_{s+2}/Q_{s+2}| \geqslant 1/|(V' + 1)/U' - V'/(U' + 1)|$$
$$> \frac{1}{2}\ U',$$

using the well-known relation $|\theta - P_k/Q_k| > 1/Q_k(Q_{k+1} + Q_k)$. And by (*), $Q_{t-s-4}(A_{s+5}, \ldots, A_t) < U/Q_{s+4}$.

LEMMA 2. – *There is an algorithm which, given* $U \geqslant V \geqslant 0$ *with* $lU = n$, *finds all the Lehmer quotients for* (U, V) *in at most* $0(S(n)\ (\log n)^3)$ *steps.*

Proof. – For large n the algorithm first applies itself recursively to the leading $\frac{1}{2}$ n binary digits of U and V, finding r partial quotients ; then it computes

$$U_r = (-1)^r\ (Q_{r-2}(A_2, \ldots, A_{r-1})\ U - Q_{r-1}(A_1, \ldots, A_{r-1})V),$$
$$V_r = (-1)^r\ (Q_r(A_1, \ldots, A_r)\ V - Q_{r-1}(A_2, \ldots, A_r)U)$$

in $0(S(n)\log(n))$ steps by the method of Theorem 1. We can find A_{r+1} in $0(S(n)\log(n))$ further steps (see [8, p. 275]), so by Lemma 1 the algorithm performs a bounded number of Euclidean iterations until reaching U_{r+k} with at most $\frac{3}{4}n$ digits. Now the same process is repeated on the $\frac{1}{2}n$ leading digits of U_{r+k}, V_{r+k} ; after a bounded number of further Euclidean iterations, we have reduced U to less than $\frac{1}{2}n$ digits, and we have found quotients $A_1, \ldots, A_p$, where $p \geqslant s$ (since the proof of Lemma 1 can be readily modified to show that $Q_s < \sqrt{U'}$). Finally the value of s is located in approximately $\log_2 p = 0(\log n)$ iterations, using the well known "binary search" bisection technique ; each iteration tests some k to see whether or not $k < s$ or $k > s$. Such a test can rely on the fact that P_k/Q_k and P_{k+1}/Q_{k+1} are both "good" when $k \leqslant s$, while they are not both "good" when $k \geqslant s + 2$, where P_k/Q_k is called good when it is $< V_k/(U_k + 1)$, for k even, or $> (V_k + 1)/U_k$, for k odd. The running time $L(n)$ of this algorithm as a whole now satisfies $L(n) \leqslant 2\,L\left(\frac{1}{2}n\right) + 0\,(S(n)\,(\log n)^2)$.

THEOREM 2. – *There is an algorithm which, given $U > V \geqslant 0$ with $lU = n$, determines the Euclidean representation $[A_1, \ldots, A_t, D]$ in $0(n(\log n)^5(\log\log n))$ steps as $n \to \infty$.*

Proof. – Begin as in Lemma 2 to reduce n to $\frac{3}{4}n$ in $L\left(\frac{1}{2}n\right) + 0(S(n)\log n)$ steps, then apply the same method until $V_t = 0$. The running time $G(n)$ of this algorithm satisfies the recurrence

$$G(n) = G\left(\frac{3}{4}n\right) + 0(S(n)\,(\log n)^3) = G\left(\frac{9}{16}n\right) + 0(S(n)\,(\log n)^3)$$
$$+ 0\left(S\left(\frac{3}{4}n\right)(\log n)^3\right) = \cdots = 0(S(n)\,(\log n)^4).$$

In particular, we can find the gcd of n-digit numbers in $n^{1+\epsilon}$ steps, as $n \to \infty$, for all $\epsilon > 0$. The method we have used is rather complicated, but no simpler one is apparent to the author. In general, the idea of reducing n to αn for $\alpha < 1$ often leads to asymptotically efficient algorithms.

Note added in proof : A. Schönhage has recently discovered a somewhat simpler algorithm which requires only $0\,(S(n)\log n)$ steps.

REFERENCES

[1] DIXON John D. — *A.M.S. Notices,* 16, October, 1969, p. 958.
[2] EUCLID. — *Elements* (c. 300 B.C.), Book 7, Propositions 1 and 2.
[3] EULER Leonhard. — *Introductio in Analysin Infinitorum,* Lausanne : 1748, Section 359.
[4] HARRIS Vincent Crockett. — *Fibonacci Quarterly,* 8, 1970, p. 102-103.
[5] HARTMANIS Juris and STEARNS Richard Edwin. — *Transactions of the Amer. Math. Soc.,* 117, 1965, p. 285-306.
[6] HEILBRONN Hans. — *Abhandlungen aus Zahlentheorie und Analysis zur Erinnerung an Edmund Landau,* Berlin : V.E.B. Deutcher Verlag der Wissenschaften, 1968, p. 89-96.
[7] KNUTH Donald Ervin. — Fundamental Algorithms, *The Art of Computer Programming,* 1, Addison-Wesley, 1968, p. 462-463.
[8] KNUTH Donald Ervin. — Seminumerical Algorithms, *The Art of Computer Programming,* 2, Addison-Wesley, 1969.
[9] KUZ'MIN Rodion Osievich. — *Atti del Congresso internazionale dei matematici,* 6, Bologna, 1928, p. 83-89.
[10] LAMÉ Gabriel. — *Comptes Rendus, Acad. Sci. Paris,* 19, 1844, p. 867-870.
[11] LEHMER Derrick Henry. — *American Math. Monthly,* 45, 1938, p. 227-233.
[12] PERRON Oskar. — *Die Lehre von den Kettenbrüchen,* Leipzig : 1913.
[13] SCHÖNHAGE A. and STRASSEN Volker. — *Schnelle Multiplication grosser Zahlen,* Universität Konstanz, 1970, submitted for publication.
[14] SYLVESTER James J. — *Philosophical Magazine,* 6, 1853, p. 297-299.

Stanford University
Computer Science Dept.
Stanford,
California 94305 (USA)

Actes, Congrès intern. Math., 1970. Tome 3, p. 275 à 279

LES LANGAGES DE PROGRAMMATION

par S. S. LAVROV

Résumé : On s'intéresse à ce qui différencie les langages de programmation de la symbolique mathematique traditionnelle, on note certaines particularités de l'évolution et de l'application de langages de programmation et on expose les points principaux d'une nouvelle approche de l'élaboration d'un langage algorithmique universel.

1. — Bien qu'ils aient un grand nombre de points communs avec la symbolique habituelle appliquée par les mathematiciens dans leurs articles et leurs livres, les langages de programmation possèdent en même temps une série de traits spécifiques. Montrons le sur des exemples.

Comme premier exemple, prenons l'algorithme bien connu d'Euclide. La description de cet algorithme dans le langage habituellement employé par les mathématiciens est grosso modo de la forme suivante :

Soient u, v des entiers naturels. Si u est divisible par v, alors le plus grand commun diviseur (P.G.C.D.) des nombres u et v est égal à v. Dans le cas contraire, considérons la chaine de relations suivantes :

$$\begin{array}{ll} u = q_0 v + r_1 , & 0 < r_1 < v , \\ v = q_1 r_1 + r_2 , & 0 < r_2 < r_1 , \\ r_1 = q_2 r_2 + r_3 , & 0 < r_3 < r_2 , \\ \dots\dots\dots & \\ r_{n-2} = q_{n-1} r_{n-1} + r_n , & 0 < r_n < r_{n-1} , \\ r_{n-2} = q_n r_n & \end{array} \tag{1}$$

Une telle chaine existe toujours, et le PGCD des nombres u et v est égal à r_n.

Ce même algorithme, dans l'un des langages de programmation largement répandu (le langage ALGOL 60) est de la forme :

```
integer procedure gcd (u,v) ;
  value u , v ; integer u , v ;
begin integer r ;
  loop : if v = 0 then gcd : = u
    else begin
(2)          r : = u — (u ÷ v) x v ;
             u : = v  ;  v : = r ;
               go to loop
      end
  end
```

Pour comprendre la marche du procédé de calcul que décrit cette notation, il n'est pas nécessaire de savoir autre chose que ce qu'elle contient de façon manifeste. Et, bien entendu, il faut connaitre le langage dans lequel elle est faite. Ainsi, par exemple, il faut savoir que $u \div v$ signifie le quotient de la division (avec reste) des nombres entiers u et v, que la liste des valeurs **value** u , v permet de travailler dans cet algorithme avec comme n'importe quelles expressions, etc...

Sur cet exemple, on peut voir quelques différences fondamentales entre les langages de programmation et le langage traditionnel des mathématiciens :

a) Le caractère algorithmique ou dynamique de ces languages. Dans la notation (1), pour chaque valeur initiale ou intermédiaire, on utilise sa désignation. Dans la notation (2), les variables r et v prennent les valeurs de tous les restes obtenus pendant la division. Les quotients q_i , $i = 0,1, \ldots , n$, sont obtenus comme valeurs successives de l'expression $u \div v$ mais on n'introduit pour eux aucune autre désignation.

Les mathématiciens ont l'habitude d'éviter soigneusement les désignations dynamiques c'est-à-dire les désignations qui d'un moment à l'autre sont attachées successivement à des valeurs différentes. Même lorsqu'ils sont obligés d'avoir affaire à des quantités dynamiques, ils s'efforcent de les traiter sans les dénommer en aucune façon. Ainsi, dans les machines de Turing, les quantités dynamiques sont enregistrées dans les cellules d'un ruban. Dans les algorithmes normaux de Markov, la quantité dynamique est le mot auquel sont appliquées des règles de l'algorithme mais il n'est aucunement fait mention de ce mot dans ces règles.

Les inconvénients et les conséquences négatives d'une telle tradition sont tout à fait analogues à ceux dont parlait Littlewood [1].

b) Les langages algorithmiques sont plus descriptifs. Quand les mathématiciens veulent dire quelque chose au sujet des objets qu'ils considèrent, ils ont habituellement recours au langage naturel sans avoir une symbolique correspondante.

Par contre, dans les langages algorithmiques, il existe une telle symbolique qui permet d'écrire, par exemple,

integer u , v (u et v entiers),

array a [1 : m, 1 : n] (a matrice de dimensions m x n)

et etc. . .

c) Cela peut sembler étrange, mais, chez les programmateurs, il y a des standards de clarté et des précisions de description bien plus rigoureux que chez les mathématiciens. C'est compréhensible, car les mathématiciens peuvent compter sur le niveau suffisamment élevé de connaissances, la promptitude et l'intuition du lecteur, alors que les programmeurs ont, eux, affaire à une machine (ou à un programme) privée de ces qualités. C'est pour cette raison, et non par originalité, que les programmeurs travaillent sur des problèmes d'intelligence artificielle. Ils veulent avoir un interlocuteur le plus intelligent possible.

Un autre exemple est une version simplifiée de l'exemple proposé par J. McCarthy [2] pour illustrer ce qu'est un programme doué de raison.

On considère la situation :

Je suis à table chez moi. J'ai une automobile chez moi. Dans la région où est située ma maison il y a un aéroport. On demande : puis-je me rendre à l'aéroport ?

La description de cette situation dans un certain langage formel utilise les noms d'une série d'objets, de notions, de leurs propriétés et de leurs relations entre eux.

La situation elle-même se décrit sous la forme suivante

at (*I*, *desk*)

in (*desk*, *home*)

in (*car*, *home*)

in (*home*, *county*)

in (*airport*, *county*)

Le problème se pose sous forme de la question

possible (*at* (*I*, *airport*)) ?

Cette description est insuffisante pour une solution formelle du problème et il est indispensable de la compléter par une description des propriétés des objets et des notions qui figurent ici. Cette description peut, par exemple, être la suivante :

at (*x* , *y*), *in* (*y* , *z*) → *in* (*x* , *z*)

in (*x* , *y*), *in* (*y* , *z*) → *in* (*z* , *z*)

walkable (*home*)

drivable (*county*)

in (*I* , *x*), *walkable* (*x*), *in* (*y* , *x*) → *possible* (*at* (*I* , *y*))

in (*car*, *x*), *possible* (*at* (*I* , *car*)), *drivable* (*x*),
  in (*y* , *x*) → *possible* (*at* (*I* , *y*))

x → *possible* (*x*)

Pour cet exemple, nous ne donnerons aucun équivalent dans un langage plus proche de celui utilisé par les mathématiciens. Mais cela n'est pas nécessaire, puisqu'il s'agit d'un calcul entièrement formalisé bien qu'il ait une grande ressemblance apparente avec le langage naturel. Par manque de place, nous ne nous arrêterons pas sur la notion de formule construite correctement et de formule vraie (déductible) dans ce calcul.

Il existe un programme, écrit par F. Black [3], qui peut donner une réponse positive à la question posée dans l'exemple donné et aussi, bien entendu, résoudre une quantité d'autres problèmes.

Cependant, le langage employé ici permet de poser aussi des problèmes insolubles. Autrement dit, il existe une mesure triviale de la puissance résolutive des différents programmes affectés à la recherche de la démonstration des formules dans ce calcul. Cette mesure est égale à zéro pour tous ces programmes. Bien entendu une telle mesure n'est pas satisfaisante du point de vue pratique car, pour certains programmes concrets, il est relativement simple d'établir que l'un est capable de résoudre une classe plus vaste de problèmes que l'autre. Il est intéressant d'expliciter s'il existe une mesure non triviale effectivement calculable de la puissance résolutive de semblables programmes.

2. – On sait qu'il existe quelques dizaines de langages de programmation qui ont la prétention à être universels mais, parmi eux, moins d'une dizaine sont

effectivement utilisés de façon courante. Avec eux, on a proposé des milliers de langages spécialisés pour des problèmes particuliers, dont une écrasante majorité est restée inconnue de tous sauf de leurs auteurs. Quelle est la raison de cette situation ?

L'emploi des langages spécialisés se heurte aux difficultés suivantes :

a) Leurs ressources sont limitées et elles sont vite épuisées quand les problèmes à résoudre se compliquent,

b) Leur réalisation est complexe et c'est pourquoi, souvent, elle ne peut être menée à bien jusqu'à la fin,

c) Les auteurs de ces langages ont tendance à oublier qu'un langage doit permettre non seulement de décrire les algo..thmes mais aussi de contrôler la marche de la résolution d'un problème.

Voici les difficultés qui apparaissent lors de l'utilisation des langages "universels" :

a) Ils ne sont jamais suffisamment universels,

b) Les descriptions des algorithmes concrets sont par trop volumineuses,

c) Leurs possibilités d'extension sont limitées.

Autrement dit, ils sont affectés à la description d'algorithmes, et non à la création de nouveaux moyens permettant de décrire des algorithmes.

Jusqu'à présent, on n'a pas réussi à créer un langage de programmation universel suffisamment commode pour décrire tous les algorithmes. A mon avis, ce problème n'est pas désespéré mais il faut en trouver une approche qui soit vraiment nouvelle.

Le schéma typique de résolution d'un problème sur ordinateur est représenté par la figure 1. L'algorithme, décrit dans un certain langage qui n'est pas un langage machine, est traduit en un programme. Ce programme, une fois introduit dans la machine, est capable d'assimiler différentes données d'entrée et de fournir les résultats correspondant à ces données.

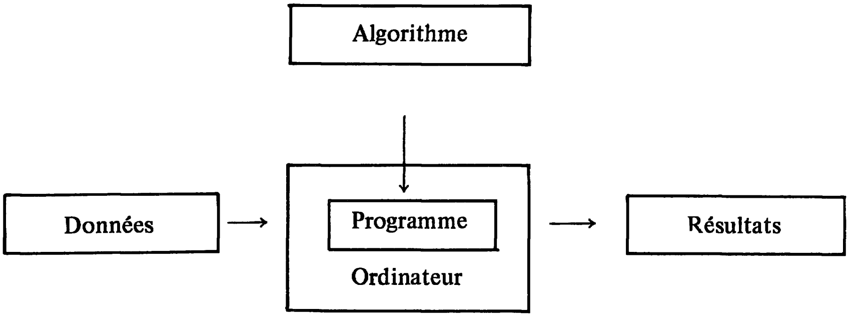


Figure 1

Les algorithmes les plus intéressants sont écrits en vue de résoudre une classe de problèmes ne différant que par leurs données initiales. La structure des données et des résultats peut être très complexe. C'est pourquoi, de pair avec le langage dans lequel l'algorithme est noté, il existe aussi un langage des données et un langage des résultats (y compris ceux qu'on sort au cours de la mise au point) et beaucoup de grands programmes doivent être considérés comme des traducteurs du langage des données en celui des résultats.

C'est là précisément la raison de l'apparition d'une multitude de langages spécialisés. Ce n'est pas un mal mais une réalité inévitable dont il faut tenir compte.

Voici une proposition d'approche : créer un langage qui ne soit pas directement affecté à la notation des algorithmes mais qui facilite la réalisation de nouveaux moyens pour décrire des algorithmes.

3. – Particularités principales d'une proposition d'approche :

a) Le langage doit être susceptible d'extension.

b) Il faut qu'il n'existe aucune limite à cette extension, en particulier la syntaxe d'une extension peut ne rien avoir de commun avec la syntaxe initiale.

c) Toute extension doit être définie exactement par des moyens formels.

d) On peut utiliser pour une telle définition une extension ou des extensions introduites précédemment mais, en fin de compte, toutes les extensions sont définies en termes d'un certain noyau.

e) Ce noyau n'est pas défini de manière formelle. Plus précisément, tout formalisme utilisé pour décrire le noyau ne peut que contribuer à une plus grande netteté de la description mais ne doit pas être un moyen de réalisation du noyau.

BIBLIOGRAPHIE

[1] Littlewood J. E. — *A mathematician's miscellany,* Methuen and Co., London.

[2] McCarthy J. — Programs with common sense, *Mechanization of thought processes,* Her Majesty's Stationery Office, London, 1959. Réimprimé dans 4.

[3] Black F. — *A deductive question-answering system,* PH. D. diss., Harvard, 1964. Réimprimé dans 4.

[4] *Semantic information processing,* Ed. M. Minsk, Cambridge, Mass., MIT Press, 1968.

Centre de Calcul
de l'Académie des Sciences d'U.R.S.S.
40, ul Vavilova
Moscou V 333 (U.R.S.S.)

Actes, Congrès intern. Math., 1970. Tome 3, p. 281 à 282.

PARTIES RATIONNELLES D'UN MONOIDE LIBRE

Par M. P. SCHUTZENBERGER

On résume certains résultats obtenus avec S. Eilenberg avec l'étude des *parties rationnelles* du monoïde libre X^* engendré par l'ensemble fini X. Par *partie*, A, on entend ici une fonction $A : X^* \to N$, c'est-à-dire une série formelle (à coefficients dans N) en les variables (non commutatives) $x \in X$. La famille des parties rationnelles **Rat** (X) est la plus petite famille **R** telle que :

$$\{0\} \in \mathbf{R} \; ; s \in X^* \Rightarrow \{s\} \in \mathbf{R} \tag{1}$$

$$A, B \in \mathbf{R} \Rightarrow A + B \in \mathbf{R} \quad \text{et} \quad A\,B \in \mathbf{R} \; ; \tag{2}$$

$$A \in \mathbf{R}, A(0) = 0 \Rightarrow A^* = 1 + \sum_{0 < n} A^n \in \mathbf{R}. \tag{3}$$

On sait que $A : X^* \to N$ appartient à **R** ssi il existe $k \in N$ et une représentation $\mu : X^* \to N^{k \times k}$, telle que pour chaque $s \in X^*$, la valeur $A(s)$ du coefficient de s dans A soit l'élément (s, k) de la matrice $s\mu$. On peut montrer que pour $A \in$ **Rat** (X) donné on peut choisir k et μ de telle sorte que tous les éléments de $x\mu$ $(x \in X)$ soient 0 ou 1. Le plus petit $k \in N$ pour lequel ceci est possible est le *nombre d'états* de A.

THEOREME 1. – *Soient donnés $A, B \in$* **Rat** *(X) de nombre d'états $\leqslant k$. L'égalité $A = B$ est décidable. L'inégalité $A \leqslant B$ (c'est-à-dire $s \in X^* \Rightarrow A(s) \leqslant B(s)$) est indécidable.*

Soient maintenant p un entier positif et $A \in$ **Rat** (X). Les relations

$$A = p\,B + C, \; C \leqslant p\, X^*$$

définissent de façon unique deux parties $B, C : X^* \to N$. Généralisant un théorème bien connu de Kronecker, on a :

THEOREME 2. – *B et C appartiennent à* **Rat** (X^*).

La démonstration utilise le résultat suivant :

THEOREME 3. – *Soient $F, G \in$* **Rat** *(X^*) où G est bornée (c'est-à-dire Sup $\{G(s) : s \in X^*\} < \infty$). Alors $F \dot{-} G \in$* **Rat** *(X^*) où $H = F \dot{-} G$ est définie par*

$$s \in X^* \Rightarrow H(s) = \text{Max}\,\{0, F(s) - G(s)\}.$$

Des contre exemples montrent que l'hypothèse G bornée est effectivement nécessaire dans cet énoncé, et qu'en particulier $F, G \in$ **Rat** (X), $G \leqslant F$ n'implique pas $F \dot{-} G \in$ **Rat** (X^*).

Généralisant la notion de produit de Hadamard, définissons maintenant pour $A, B : X^* \to N$, leur "intersection" $G = A \cap B$ par la condition

$$s \in X^* \Rightarrow G(s) = A(s)\,B(s).$$

On sait que $A\ ,B \in$ **Rat** $(X) \Rightarrow A\ \cap B \in$ **Rat** (X). Le “problème inverse” n’est pas résolu (même dans le cas classique des fonctions rationnelles dont la série de Taylor a ses coefficients dans Z) et nous proposons les

Conjectures (1). – Si $A, A \cap B \in$ **Rat** (X), il existe $C \in$ **Rat** (X) telle que $A \cap B = A \cap C$;

(2). – Si $A \cap A \in$ **Rat** (X) il existe $B \in$ **Rat** (X) telle que $A \cap A = B \cap B$.

Faculté des Sciences de Paris
Institut de Programmation
9, Quai Saint-Bernard,
Paris 5^e
France

Actes, Congrès intern. Math., 1970. Tome 3, p. 283 à 288.

ON THE ALGEBRAIC COMPLEXITY OF FUNCTIONS [*]

by S. WINOGRAD

1. Introduction.

The number of multiplications and additions required to evaluate a polynomial was first investigated by Ostrowski [1]. He showed that in order to evaluate a polynomial $P_n(x)$ of degree n, at least n multiplications and n additions are necessary, for $n = 1, 2, 3, 4$. The results were extended by Pan [2] and Balaga [3], who proved them for all n. Motzkin [4] introduced the notion of preconditioning of the co-efficients. He showed that if, in the course of evaluating $P_n(x) = \sum_{i=0}^{n} a_i x^i$, operations which depend only on the a_i's are not counted, then the minimum number of multiplications necessary is about $\frac{1}{2} n$, while the minimum number of additions is still n.

In this paper we will survey some of the recent results in this area. In particular, we will describe the extension of the results to matrix and vector operations. Because of the survey nature of the paper, we will omit all proofs, many of them appear in [5]. For all other results we will indicate where they can be found in the literature to the extent that they have already been published.

2. Definitions and notations.

DEFINITION 1. – Let F be a field and let $B \subset F$ be a subset of F. An *N-step algorithm* α over (F, B) is a mapping

$$\alpha : \{1, 2, \ldots, N\} \to B \cup \{p_1, p_2, p_3, p_4\} \times \{1, \ldots, N\}^2$$

subject to the restriction that if $\alpha(k) = (p_i, j_1, j_2)$, then $j_1, j_2 < k$. With each such algorithm, we associate a (partial) function e_α defined by :

(1) $e_\alpha(k) = \alpha(k)$ if $\alpha(k) \in B$;

(2) $e_\alpha(k) = p_i(e_\alpha(j_1), e_\alpha(j_2))$ when $\alpha(k) = (p_i, j_1, j_2)$ if $e_\alpha(j_1)$, $e_\alpha(j_2)$ and $p_i(e_\alpha(j_1), e_\alpha(j_2))$ are defined ; where

$$p_1(x, y) = x + y, \quad p_2(x, y) = x - y, \quad p_3(x, y) = xy, \quad p_4(x, y) = x/y \ ;$$

(3) $e_\alpha(k)$ is not defined otherwise.

[*] The research of this report was partially supported by the Office of Naval Research, Contract No. N00014-69-C-0023.

The function $\alpha(i)$ is the sequence of operation which the algorithm executes, while $e_a(i)$ is the sequence of partial results which the algorithm computes.

DEFINITION 2. – The algorithm α is said to *compute* $\psi_j \in F$, $j = 1, 2, \ldots, t$ if e_a is total and if there exist t integers $i_1, i_2, \ldots, i_t$ such that $e_a(i_j) = \psi_j$, $j = 1, 2, \ldots, t$.

The cardinality of $\alpha^{-1}(\{p_i\} \times \{1, 2, \ldots, N\}^2)$ is the number of times operation p_i appears in α. The number of multiplication/divisions (m/d) in an algorithm is $\alpha^{-1}(\{p_3, p_4\} \times \{1, 2, \ldots, N\}^2)$, and the number of additions/subtractions (a/s) is $\alpha^{-1}(\{p_1, p_2\} \times \{1, 2, \ldots, N\}^2)$.

Even though definitions 1 and 2 dealt with algorithms for constructing elements of a field F, the same definitions can be made for constructing elements in other algebraic structures. In particular, we will find it necessary, later in the paper, to discuss algorithms over commutative and noncommutative rings.

Let F be a field and $G \subset F$ a subfield of F. We will use F^t_G to denote the vector space F^t over the field G.

DEFINITION 3. – Let Φ be a $t \times n$ matrix with entries in F. We define $N(\Phi)$ as the number of linearly independent columns of Φ when viewed as vectors in F^t_G. We define $n(\Phi)$ as the number of linearly independent columns of Φ when viewed as representatives of vectors in F^t_G/G^t. (When G is not specified explicitly, we take it to be the rational subfield of F).

Throughout the paper we will use $F(x_1, \ldots, x_n)$ to denote the field extension of F by the indeterminates $x_1, x_2, \ldots, x_n$, and $F[x_1, \ldots, x_n]$ to denote the ring extension.

In particular, we will deal with elements of $F(x_1, \ldots, x_n)$ (or $F[x_1, \ldots, x_n]$) which are linear in the x_i's. We will, therefore, use Φx to denote the t elements $\psi_j = \sum_{i=1}^{n} \Phi_{j,i} x_i$, $j = 1, 2, \ldots, t$, where the $\Phi_{j,i}$'s are in F.

3. Results.

THEOREM 1. – *Every algorithm over* $(F(x_1, \ldots, x_n), F \cup \{x_1, \ldots, x_n\})$ *which is used to compute* $\Phi_{t \times n} x$ *requires at least* $n(\Phi)$ *m/s (where G is taken as the field of rationals of F).*

COROLLARY 1. – Let $P_j(x) = \sum_{i=0}^{n_j} a_{j,i} x^i$, $j = 1, 2, \ldots, t$. The minimum number of m/d needed to evaluate all the P_j's is $\sum_{j=1}^{t} n_j$, even when operations which depend on x alone are not counted.

COROLLARY 2. – Let A be an $m \times n$ matrix and b an n-vector. The minimum number of m/d required to compute Ab is mn, even when operations which depend on b alone are not counted.

THEOREM 2. – *Every algorithm over* $(F[x_1, \ldots, x_n], F \cup \{x_1, \ldots, x_n\})$ *which is used to compute* $\Phi_{t\times n}x$ *requires at least* $N(\Phi) - t$ *a/s (where G is taken as the field of rationals of F).*

COROLLARY 3. – The minimum number of *a/s* needed to compute $P_j(x) = \sum_{i=0}^{n_j} a_{j,i}x^i$, $j = 1, 2, \ldots, t$, when no divisions are allowed, is $\sum_{i=1}^{t} n_j$.

COROLLARY 4. – The minimum number of *a/s* needed to compute $A_{m\times n}b$, when no divisions are allowed, is $m(n-1)$.

As we will see later, the condition of no division could be dropped. As the corollaries indicated, the applications of theorems 1 and 2 arise when F is the field G extended by some indeterminates. In the case that G is a subfield of C, the field of complex numbers, then the requirement of no divisions can be relaxed. We will consider now the case that $F = G(y_1, \ldots, y_m)$ where G is a subfield of C.

THEOREM 3. – *Every algorithm over*

$$(G(y_1, \ldots, y_m)(x_1, \ldots, x_n), G(y_1, \ldots, y_m) \cup G(x_1, \ldots, x_n))$$

which is capable of computing $\Phi_{t\times n}x$ *requires at least* $\frac{1}{2}\, n(\Phi)$ *m/d.*

THEOREM 4. – *Every algorithm over*

$$(G(y_1, \ldots, y_m)(x_1, \ldots, x_n), G(y_1, \ldots, y_m) \cup G(x_1, \ldots, x_n))$$

which is capable of computing $\Phi_{t\times n}x$ *requires at least* $N(\Phi) - t$ *a/s.*

COROLLARY 5. – Every algorithm which computes

$$P_j(x) = \sum_{i=0}^{n_j} a_{j,i}x^i,$$

$j = 1, 2, \ldots, t$, requires at least $\frac{1}{2}\sum_{i=1}^{t} n_j$ *m/d and* $\sum_{j=1}^{t} n_j$ *a/s* even when operations involving only the coefficients $a_{j,i}$ are not counted.

This corollary was first discovered by Motzkin [4] in the case $t = 1$. He proved the slightly stronger result that at least $\frac{1}{2}(n+1)$ *m/d* and n *a/s* are necessary to compute $P(x) = \sum_{i=0}^{n} a_i x^i$. Motzkin discovered an algorithm for calculating $P(x)$ which requires $\frac{1}{2}(n+1)$ *m/d* and at most $n+1$ *a/s.* In Motzkin's algorithm, the function of the coefficients which have to be computed are, in general, not in $Q(a_0, a_1, \ldots, a_n)$, but in its algebraic closure ; moreover, it can happen that even if $a_0, a_1, \ldots, a_n$ are real numbers, the values of the functions are complex. Pan [2] devised an algorithm required $(n + 4)/2$ m/d and $n + 1$ a/s in which the functions of the coefficients which have to be computed are in the algebraic closure of $Q(a_0, a_1, \ldots, a_n)$, but are guaranteed to have real values whenever the a_i's are real. Rabin and the author [6] constructed an algorithm which requires $n/2 + o(n)$ m/d and $n + o(n)$ a/s with auxiliary functions which are in $Q(a_0, \ldots, a_n)$ (or even in $Q[a_0, \ldots, a_n]$).

COROLLARY 6. – Every algorithm which computes $A_{m \times n}b$ requires at least $\frac{1}{2}mn$ m/d and $m(n-1)$ a/s even when operations which depend only on the entries of A, or only on the elements of b, are not counted.

An algorithm for computing $A_{m \times n}b$ which requires only $\frac{1}{2}m \cdot n$ multiplication (for n even) appeared in [7]. This algorithm requires $\frac{3}{2}m\left(n + \frac{4}{3}\right)$ a/s. So even with the use of auxiliary functions, the algorithm succeeds only in "trading off" multiplications for additions. The next theorem, which appeared in [8], shows that at least for $m = 1$ there is no algorithm which minimizes the number of m/d and a/s simultaneously. The best that can be hoped for is a "trade-off" of multiplications and additions.

THEOREM 5. – *Every algorithm for computing* $\sum_{i=1}^{n} x_i y_i$ *over*

$$(Q(x_1, \ldots, x_n, y_1, \ldots, y_n), Q(x_1, \ldots, x_n) \cup Q(y_1, \ldots, y_n))$$

requires at least $2n - 1$ *binary operations.*

The algorithm in [7] was used there to construct an algorithm for multiplying two $n \times n$ matrices using only $\frac{1}{2}n^3 + n^2$ multiplications (for n even). This algorithm is not optimal as was shown by Strassen. Strassen [9] described an algorithm for multiplying two $n \times n$ matrices with number of m/d and number of a/s which grows as $n^{\log_2 7}$. Strassen's algorithm is based on an algorithm for multiplying 2×2 matrices which used only 7 multiplications and *does not involve commutativity.* It is not known whether Strassen's algorithm is optimal, but as the next theorem [10] shows, any improvement has to come from investigating the product of $n_0 \times n_0$ matrices for $n_0 \geqslant 3$.

THEOREM 6. – *Every algorithm for multiplying two* 2×2 *matrices requires at least* 7 *multiplications.*

A special case of theorem 6, when the algorithm does not use commutativity, was proved by Hopcroft and Kerr [11].

Let R be a ring with identity. We will use $R\{x_1, \ldots, x_n\}$ to denote the ring extension of R by the noncommuting indeterminants $x_1, x_2, \ldots, x_n$.

THEOREM 7. – *Let* $\psi_k = \sum_{i=1}^{m} \sum_{j=1}^{n} r_{i,j,k} x_i y_j$, $k = 1, 2, \ldots, t$, $(r_{i,j,k} \in R)$. *If an algorithm over* $(R[x_1, \ldots, x_n, y_1, \ldots, y_m], R \cup \{x_1, \ldots, x_n, y_1, \ldots, y_m\})$

exists which can compute the ψ_k's in N multiplications, then there exists an algorithm over $(R\{x_1, \ldots, x_n, y_1, \ldots, y_m\}, R \cup \{x_1, \ldots, y_m\})$ which can compute the ψ_k's in $\leqslant 2N$ multiplications.

Corollary 7. – Let $f(n)$ be the minimum number of multiplications required to multiply two $n \times n$ matrices, and let $g(n)$ be the minimum number of operations. If $f(n) \leqslant Cn^a$, then for every $\epsilon > 0$ there exists K such that $h(n) \leqslant Kn^{a+\epsilon}$.

Corollary 8. – Let $f(n)$ be as in Corollary 7, and $h(n)$ be the minimum number of m/d required to invert an $n \times n$ matrix. If $f(n) \leqslant cn^a$, then for every $\epsilon > 0$ there exists K such that $h(n) \leqslant KN^{a+\epsilon}$.

Theorem 7 shows that commutativity can cut down the number of multiplications needed to compute bilinear forms by at most a factor of 2. The following theorem, due to P. Ungar (private communication) shows that division does not help.

Theorem 8. – *Let $\psi_k = \sum_{j=1}^{m} \sum_{i=1}^{n} r_{i,j,k} x_i y_j$, $k = 1, 2, \ldots, t$ $(r_{i,j,k} \in G)$. If there exists an algorithm over $(G(x_1, \ldots, y_m), G \cup x_1, \ldots, y_m)$ which can compute the ψ_k's using N m/d, then there exists an algorithm over $(G[x_1, \ldots, y_m], G \cup \{x_1, \ldots, y_m\})$ which can compute the ψ_k's in N multiplications.*

Corollary 9. – Let $f(n)$ and $h(n)$ be as in Corollary 8. Then, $h(n) \geqslant \frac{1}{42} f(n)$.

BIBLIOGRAPHY

[1] Ostrowski A.M. — *On two problems in abstract algebra connected with Horner's rule,* Studies presented to R. von Mises, Academic Press, New York (1954), 40-48.

[2] Pan V. Ya. — Methods of computing values of polynomials, *Tussian Mathematical Surveys,* Vol. 21, 1966, p. 105-136.

[3] Balaga E.C. — Some problems in the computation of polynomials, *Dokl. Akad. Nauk. S.S.S.R.,* Vol. 123, 1958, p. 775-777.

[4] Motzkin T.S. — Evaluation of polynomials and evaluation of rational functions, *Bull. Amer. Math. Soc.,* Vol. 61, 1955, p. 163.

[5] Winograd S. — On the number of multiplications necessary to compute certain functions, *Comm. Pure and Appl. Math.,* Vol. 23, 1970, p .165-179.

[6] To appear.

[7] Winograd S. — A new algorithm for inner product, *I.E.E.E. Trans. on Computers,* Col. 17, 1968, p. 693-694.

[8] Winograd S. — On the algebraic complexity of inner product, *I.B.M. Research Report* R.C. 2729, December 1969.

[9] Strassen V. — Gaussian elimination is not optimal, *Numerische Mathematik,* Vol. 13, 1969, p. 354-356.

[10] Winograd S. — On multiplication of 2 $\times$ 2 matrices, *I.B.M. Research Report,* R.C. 2767, January 1970.

[11] Hopcroft J.E. and Kerr L.R. — Some techniques for proving certain simple programs optimal, *Proc. I.E.E.E. Symposium on Switching and Automata Theory,* 1969, p. 34-45.

I.B.M. Research
P.O. Box 218
Yorktown
Heights, N.Y. 10 598 (USA)

Actes, Congrès intern. Math., 1970. Tome 3, p. 289 à 295.

E8 - ANALYSE NUMÉRIQUE

ABOUT OPTIMISATION OF NUMERICAL METHODS

by N.S. BACHVALOV

The optimisation's problem of solution's methods arises when solving any complicated problem. That's why most works in numerical mathematics are connected with this problem.

Peculiarity of our approach is that we give some formalisation of conception of optimal methods and systematic investigation of optimisation problems. Our consideration is near to that in [1].

When choosing solution's method for any problem, researcher takes into account certain properties of solution and in accordance with them selects a suitable numerical algorithm.

Let P be the class of problems with the same properties. It is natural that the numerical method of solution would be chosen the same for all problems of this class. This method is chosed by the researcher from some set M of methods. We denote by $\epsilon(p, m)$ the error of the method m when solving the problem p.

The method's quality for every concrete class of problems can be described by some collection of parameters : the number of arithmetical or logical operations, the maximal volume of codes, which is simultaneously stored, the number of calculated values of certain function.

Let $\chi(p, m)$ be a vector, the components of which are these parameters. Let the set of these vectors χ be ordered, in particular, if every component χ_1 isn't less than corresponding component χ_2, then χ_1 preceeds χ_2 ($\chi_1 \prec \chi_2$). In the most typical case χ is scalor. Let's consider a vector χ and subset M_χ of methods $m \in M$ such, that $\chi(p, m) \prec \chi$ for every $p \in P$. The value $E(P, m) = \sup\limits_{p \in P} \epsilon(p, m)$ is called the error of method m for the class of problems P. The value

$$E_\chi(P, M) = \inf_{m \in M_\chi} E(P, m)$$

is called the optimal error's estimate for the class of problem P for methods whose quality characteristics are not worse then χ. If there exists a method $m \in M_\chi$ such that $E(P, m) = E_\chi(P, M)$, then let's call this method optimal.

The solution of a problem often is an operator acting on a function f and the set P is determined by some class F of these functions. Such determination of the class P is natural but not only possible.

Optimal methods have been constructed only for a few classes of problems. In the most developed cases the lower estimates of values $E_\chi\,(P\,,M)$ have been obtained and the methods $m \in M_\chi$ have been constructed for which

$$E\,(P\,,\,m) = \Theta(E_\chi\,(P\,,M)) \text{ if } E_\chi\,(P\,,M) \to 0.$$

Such methods may be called optimal with respect to the order of error's estimate.

The significance of upper estimates of the method's error is evident. The obtaining of lower estimates for $E_\chi\,(P\,,M)$ looks less natural. But this field has applied significance independently of the optimisation problem.

It may happen that we have obtained the lower estimate for $E_\chi\,(P\,,M)$, which shows that for any method $m \in M$ there exist such problem $p \in P$ that cannot be solved with acceptable accuracy if our expenses are acceptable. It means, that the considered class of methods M must be expanded or the considered class of problems P must be narrowed, that is some new properties of solved problem must be taken into account.

It is reasonable for many problems to take as χ the number of calculated values of some function f. Let $h_\epsilon(F)$ be minimal number of values of function f, which for arbitrary $f \in F$ is enough for finding the solution with accuracy ϵ. The following hypothesis seems to be probable : for reasonable classes F it is possible for every $c > 1$ to indicate an algorithm with the following properties : the function f is calculated not more than in $ch_\epsilon\,(F)$ points, additional number of arithmetical operations is

$$\Theta(h_\epsilon\,(F)\,ln^\gamma\,(\epsilon^{-1}))$$

the error isn't more than ϵ. The more detailed formulation of this hypothesis is given in [2].

In the examples, which are considered below and in [3] are constructed the methods for some problems, which satisfy the hypothesis for some $c > 1$.

It should be pointed that originaly on optimisation problem was posed by A.N. Kolmogorov and A.G. Vituschkin in connection with reconstruction of function $f \in F$ by given information.

All known lower estimates $E_\chi\,(P\,,M)$ for problems of mathematical physics are obtained from those in the problem of calculation of the integral

$$I\,(Kf) = \int_G K\,(P)\,f\,(P)\,dP$$

for fixed $K\,(P)$ and given class F of function f.

In some cases it is natural to consider $F = C_{r_1,\ldots,r_s}(A)(r_i$ – not necessarily integer) – a class of functions f for which $|f_{x_i l_i}| \leqslant A$ for $0 \leqslant l_i \leqslant r_i$. If $r_1 = \ldots = r_s = l$, then let us call this class $C^l_s(A)$. Let's consider the methods of computing $J(Kf)$. using as the information about f only the information about its values in N points.

THEOREM. — *For class of such methods the lower estimate is $d(r\,,K)\,AN^{-r} > 0$, where $r^{-1} = r_1^{-1} + \ldots + r_s^{-1}$.*

For the case, when G is a rectangular regions, the lower and upper estimates for integration problems differing only in a factor $\mathfrak{S}(ln^{\gamma}N)$ are received at present for any classes of function determined by a system of restrictions on derivatives norms [4]. In the same extent this problem is solved for nondetermined methods of integration in case of the probability estimates of the error.

Let's consider some other examples lower estimates. Let the solution be a nonlinear operator acting on a function f from some fixed class F. Then it is possible to apply the following procedure [3], which will be demonstrated on the example of elliptic equation. Let we have a problem

$$Lu = \sum_{i_1, i_2 = 1}^{2} a_{i_1 i_2} u_{x_{i_2} x_{i_2}} = f \quad \text{in } G$$

$$u|_{\Gamma} = 0, \Gamma \in \text{Л}_1 (B, 1), f \not\equiv 0, \quad f \in C_{q,q}(A_1), 0 < q,$$

Γ being a boundary of G.

The class of problems is determined by the conditions :

$$a_{i_1 i_2} \in C_{r,r}(A), \ r < 2,$$

$$0 < m \leqslant \Big(\sum_{i_1, i_2 = 1}^{2} a_{i_1 i_2} \xi_{i_1} \xi_{i_2} \Big) \Big/ \Big(\sum_{i=1}^{2} \xi_i^2 \Big)$$

Let

$$L^{\circ} u = \sum_{i_1, i_2 = 1}^{2} a_{i_1 i_2}(x_1, x_2) u_{x_{i_1} x_{i_2}},$$

Φ° — being the corresponding Green's function,

$$\rho(L, L^0) = \sup_{x_1, x_2, i_1, i_2} |a_{i_1 i_2}(x_1, x_2) - a^0_{i_1 i_2}(x_1, x_2)|$$

u° — the solution of the equation $L^0 u^0 = f$.

Let us fix some point $P \in G$. Then for all operators from that class :

$$u(P) = u^0(P) + \int_G \Phi^0(P, Q)(L - L^0) u^0 \, dQ + \sigma(P)$$

where $|\sigma(P)| \leqslant \text{const.} (\rho(L, L^0))^2$.

Let us consider an arbitrary method of calculating $u(P)$ using the information about values $a_{i_1 i_2}$ in N points. Using the standard methods it is possible to conduct two operator's $L^{(k)}$ of considering class, $k = 1, 2$, for which

$$\int_G \Phi^0 (P, Q) (L^{(1)} - L^{(2)}) u^0 \, dQ \geqslant \text{const}\, N^{-r/2} > 0, \rho(L^{(k)}, L^0) \leqslant \text{const}\, N^{-r/2}$$

and the values $a^{(k)}_{i_1 i_2}$ and $a^0_{i_1 i_2}$ in the points where they are calculated, coincide [4].

Thus we have obtained two equations, for which the approximate values of $u^{(k)}(P)$ coincide, but exact values differ more than on const. $N^{-r/2} > 0$; whence the needed estimate follows.

At present for some model classes of problem the upper and lower estimates are obtained, which differ at most on the factor $\Theta(ln^\gamma N)$.

Let's consider some such examples.

(1) For the Cauchy problem for system $y' = f(x, y)$ on $[0, 1]$ in the class $f \in C^l_{s+1}(A)$ the optimal method in that sense would be Adam's with upper and lower estimates const. N^{-l}.

(2) For some problems the nearness of the problem to the spectrum is important.

In [5] is considered the optimisation problem in the class of integral equations

$$u(P) = \int K(P, Q)\, u(Q)\, dQ + f(P)$$

with
$$K \in C^l_{2s}(A), f \in C^l_s(A).$$

under the condition, that the distance from 1 to the spectrum of K is not less $d > 0$. The solution is obtained with the error $\Theta(N^{-l/2s})$ by using N values of K and f und, besides, doing $\Theta(N)$ arithmetical operations. This result cannot be improved.

(3) Let's consider the boundary problem $y' = f(x,y)$ in $[0, 1]$,

$$B(y(0)) = 0, C(y(1)) = 0$$

The class of problem is described by the conditions :

(a) $f \in C^l_{s+1}(A)$

(b) B , C - the vector of fixed dimension, s_1 and s_2, belonging to class $C^2_{s_1}(A_1)$ and $C^2_{s_2}(A_2)$, accordingly.

(c) all solutions from considered class are bounded by some constant S.

(d) if y is a solution of some problem of this class, then for solutions of linear system

$$z' - \|\partial f/\partial y\| z = \gamma(x), \|\partial B/\partial y\|\, z(0) = \alpha, \|\partial C/\partial y\| z(1) = \beta$$

the next inequality is true

$$\|z\| \leqslant M(\|\alpha\| + \|\beta\| + \|\gamma\|);$$

the constant M is fixed.

The lower estimate const/N^l follows from the lower estimate for integration problem. The upper estimate of the same order may be obtained by next way. Let's take some difference approximation of the order $\Theta(h^l)$ and some $\epsilon_0 > 0$. The algorithm consists of successive obtaining the solutions of difference approximation with step $h_k = 2^{-k}$, satisfying the boundary condition with error $\epsilon_0 h^l_k$; this process begins from $k = 0$ and ends when $k \sim \frac{1}{l} \log_2(\epsilon^{-1})$, where ϵ is a given accurathy.

(4) For partial differential equations the class of problems may be given for example by the class of coefficients of the equation or boundary conditions.

The solution of optimisation problem for homogeneous equation with constant coefficients is specific. Solutions of homogeneous elliptic and parabolic equations with constant coefficients can be written as integrals with some weight of boundary condition nonhomogeneous term. That's why the lower estimate is of the same order as the one for integration of nonhomogeneous term in the boundary conditions. Using the fact, that the smoothness of solutions increases with moving away from the boundary, the difference approximation of the Laplas's and heat tranfer equations may be constructed with following properties. The volume of calculations for finding the solution is of the same order as for calculation the integral of nonhomogeneous term in the boundary conditions [6, 7]. In the case of heat transfer equations the solution will be found in all points of sufficiently compact set.

(5) For nonhomogeneous nonstationary problems the methods such as one of alternating directions are supposed to be optimal practically and theoretically. Let, for example, the next equation be solving $u_t = \Delta u + f(t, x_1, \dots, x_s)$ in the region $0 \leqslant x_1, \dots\ x_s \leqslant 1,\ 0 \leqslant t \leqslant T$ with $u|_\Gamma = 0$.

Let's consider the problem in the class $f \in C^r_{s+1}(A)$, $r < 1$.

One of the alternating directions methods in optimal in the following sense. Using the values of f in N points it is impossible for any methods to obtain the solution with error estimate better than $\mathcal{O}(N^{-\frac{r}{s+1}})$; by the mentioned method we obtained solution with error $\mathcal{O}(N^{-\frac{r}{s+1}})$ on all set with $\mathcal{O}(N)$ calculated values f and besides with $\mathcal{O}(N)$ arithmetical operations [3].

We think that such methods are optimal for wide classes of nonstationar problems with compatible boundary conditions.

(6) For stationary problems the optimisation of methods is more complicated. It isn't difficult to write the discrete approximation with minimal in order number of equations ; but some additional difficulties arise when solving it. However there some methods, especially in the case of rectangular regions, which are close to optimal ones. For example in [8, 9] some methods for solving difference elliptic problem in rectangle are proposed with $\mathcal{O}(h^{-2} \ln \epsilon^{-1})$ operations and accuracy ϵ. For the equation $\Delta u = f$ with $f \in W_2^2(A)$ we obtain the solution with optimal accuracy $\mathcal{O}(Ah^2)$ by $\mathcal{O}(h^{-2})$ operation.

In [10] the difference approximations for elliptical system in rectangle are solving with error ϵ by $\mathcal{O}(h^{-2} \ln h^{-1} \ln \epsilon^{-1})$ operations. In [11] the difference approximation for Laplas's equation in arbitrary region is solved with accuracy ϵ by $\mathcal{O}(h^{-2} \ln h^{-1} \ln \epsilon^{-1})$ operations.

Besides we can find the solution of Poisson's difference equation in rectangle with so cold fast Fourier transformation with $\mathcal{O}(h^{-2} \ln h^{-1})$ operations. All this gives a possibility to solve Poisson's difference equation in arbitrary region with $\mathcal{O}(h^{-2} \ln h^{-1} \ln \epsilon^{-1})$ operation.

With method based on using spectrum equivalent operators [10] this procedure gives possibility of fast solving of difference approximation with variable coefficients in variable region. Let's consider some possible directions of researches.

As we already mentioned, restricting of the class of problems improves the properties of optimal method. In some cases such restricting may be carried out, if we introduce some parameter in the optimisation problem. For example, the boundary problem for model equation

$$y'' - \mu^{-2} p(x) y = \mu^{-2} f(x) \qquad \text{by } p, f \in C_2(A)$$

may be considered in the class of problems mentioned above or as a problem for the class of equations

$$y'' - P(x) y = F(x) \qquad \text{by } P, F \in C_2(A\mu^{-2})$$

For μ small the consideration of restricted class essentially improves the optimal methods properties [12].

We have discussed optimisation problem for classes F of functions of finite smoothness. In all cases the difference methods have appeared to be optimal. In practice we work with piece analytical functions. That's why the consideration of problems with such classes of function must be an important branch of optimisation theory. It may be expected, that there some different methods would be optimal.

LITERATURE

[1] Соболев С. Л., Бабушка И. — Оптимизация численных методов. Aplikace mat., 1965, 10, p. 96-129.

[2] Бахвалов Н. С. — О свойствах оптимальных методов решения задач математической физики. Ж. Вычисл. матем. и матем. физ., 1970, 10, № 3, 555-568.

[3] Бахвалов Н. С. — Об оптимальных методах решения задач. Aplikace mat., 1968, 13, A, p. 27-38.

[4] Бахвалов Н. С. — Об оптимальных оценках сходимости квадратурных процессов и методов интегрирования типа Монте-Карло на классах функций. В сб. Численные методы решения дифференциальных и интегральных уравнений и квадратурные формулы. М., Наука, 1964, 5-63.

[5] Емельянов К. В., Ильин А. М. — О числе арифметических действий, необходимом для приближенного решения интегрального уравнения Фредгольма II рода. Ж. Вычисл. матем. и матем. физ., 1967, № 4, 905-910.

[6] Бахвалов Н. С. — О численном решении задачи Дирихле для уравнения Лапласа. Вестник МГУ, 1959, № 5, 171-195.

[7] Чжан - Гуан - Цзюань. — О минимальном числе узлов при численном интегрировании уравнения теплопроводности. Ж. Вичисл. матем. и матем. физ, 1962, 2, № 1, 80-88.

[8] Федоренко Р. П. — О скорости сходимости одного итерационного процесса. Ж. Вычисл. матем. и матем. физ., 1964, 4, № 3, 559-564.

[9] Бахвалов Н. С. — О сходимости одного релаксационного метода при естественных ограничениях на эллиптический оператор. Ж. Вычисл. матем. и матем. физ., 1966, 6, № 5, 861-883.

[10] Дьяконов Е. Г. — О построении итерационных методов на основе использования операторов, эквивалентных по спектру. Ж. Вычисл. матем. и матем. физ., 1966, 6, № 1, 12-34.

[11] Бахвалов Н. С. — Об одном способе приближенного решения уравнения Лапласа. ДАН СССР, 1957, 114, № 3, 455-458.

[13] Бахвалов Н. С. — К оптимизации методов решения краевых задач при наличии пограничного слоя. Ж. Вычисл. матем. и матем. физ., 1969, 9, № 4, 841-859.

University of Moscow
Dept. of Mathematics,
Moscow B 234
U.R.S.S

Actes, Congrès intern. Math., 1970. Tome 3, p. 297 à 309.

MÉTHODES NUMÉRIQUES NOUVELLES EN MÉCANIQUE DU CONTINU

par N. N. IANENKO

I. Dans notre conférence nous examinerons la structure et la construction des algorithmes numériques pour résoudre les problèmes de mécanique.

On peut distinguer trois parties essentielles d'un algorithme numérique se basant sur la discretisation des opérateurs différentiels, intégrodifférentiels ou intégraux et sur un maillage donné :

(1) *l'opérateur local de domaine*, associé à un point générique du domaine d'intégration qui est une approximation discrète de l'opérateur gouvernant l'écoulement du milieu. On peut le considérer comme un modèle discret indépendant du milieu continu,

(2) *l'opérateur local de la frontière* associé au point de la frontière qui est une discrétisation des conditions aux limites,

(3) *le maillage*, c'est-à-dire l'ensemble fini des points qui porte l'information sur le milieu.

Nous nous bornerons à une étude approfondie de l'opérateur local de domaine en supposant le maillage régulier et rectiligne et les conditions aux limites aussi simples que possible (périodiques, par example).

Soient

$$\frac{\partial u}{\partial t} = \mathfrak{L} u \tag{1}$$

l'équation différentielle originelle

$$\frac{u^{n+1} - u^n}{\tau} = \Lambda_1 u^{n+1} + \Lambda_0 u^n, \tag{2}$$

le schéma à deux niveaux correspondant,

(les opérateurs Λ_0, Λ_1 sont des opérateurs discrets locaux), qui peut s'écrire aussi sous forme résolue :

$$u^{n+1} = \sigma u^n, \; \sigma = (I - \tau \Lambda_1)^{-1} (I + \tau \Lambda_0) \tag{3}$$

Si $\Lambda_1 \equiv 0$ le schéma est dit *explicite*, dans le cas contraire *implicite*, σ est *l'opérateur de transition.* A quelles conditions doit satisfaire le schéma (2), approchant l'équation (1) ? Les deux premières conditions sont bien connues :

(1) la *consistance* ou l'approximation qui est responsable de la précision locale du schéma numérique,

(2) la *stabilité* qui est responsable de la croissance pas trop rapide des erreurs de calcul.

A ces deux conditions qui sont impératives, on doit ajouter une série de conditions qui sont moins impératives mais pratiquement indispensables. En premier lieu on place

(3) *l'économie du schéma* : le critère de l'économie étant le temps machine nécessaire mesuré dans un mode convenable.

Si l'équation (1) est une conséquence des lois de conservation nonlinéaires

$$\frac{\partial u_i}{\partial t} + \sum_k \frac{\partial \sigma_{ik}}{\partial x_k} = 0 \; ; \tag{4}$$

la forme conservative (4) du schéma est très importante comme montre l'exemple donné par Tikhonoff et Samarski [1].

Les conditions mentionnées ci-dessus ne sont pas seules, mais il est important de souligner ici, qu'elles sont contradictoires et pour les satisfaire on doit résoudre un problème d'optimisation.

II. Pour ce qui va suivre, il faut introduire quelques notions.

Definition. – Le schéma (2) est dit *absolument stable* s'il est uniformément stable pour un choix arbitraire des paramètres $\tau, h_1, \ldots h_m$ (> 0), c'est-à-dire les conditions

$$\|\sigma_{nm}\| \leqslant C(T) \tag{5}$$

ont lieu, où σ_{nm} est *l'opérateur de transfert* qui est défini par la relation :

$$u^n = \sigma_{mn}\, u^n, \qquad u^n = u(n\tau), \qquad u^m = u(m\tau) \tag{6}$$

et où la constante $C(T)$ ne dépend pas des paramètres τ, $h_1, \ldots, h_m$. Dans le cas contraire le schéma est dit *conditionnellement stable.*

Considerons l'opérateur $\dfrac{\sigma - I}{\tau}$ où σ est l'opérateur de transition du schéma (2) correspondant à une loi de passage à limite $h = h(\tau)$, $h \to 0$, $\tau \to 0$. Pour la fonction $h = h(\tau)$ l'opérateur $\dfrac{\sigma - I}{\tau}$ détermine l'opérateur infinitésimal $\mathcal{L}$ qui engendre un semi-groupe uniparamétrique $e^{\tau \mathcal{L}}$

Definition 2. — Si l'opérateur $\mathcal{L}$ ne dépend pas de la fonction $h = h(\tau)$ le schéma est dit *absolument consistant*, sinon *conditionnellement consistant.* Nous appelerons en général *adhérence du schéma* correspondant au schéma donné l'ensemble de tous les semi-groupes engendrés par le schéma (2). Maintenant on peut facilement voir la place qu'occupent les schémas absolument stables et absolument consistants. Pour le schéma absolument stable et absolument consistant on peut choisir le paramètre τ en ne considérant que les exigences de la précision parce que la stabilité est garantie pour chaque .τ, h.

Par contre, si le schéma est conditionnellement stable les restrictions sur τ, imposées par les conditions de stabilité, peuvent être très exigeantes et le schéma devient côuteux. Si le schéma est conditionnellement consistant ou ne peut pas dire avec certitude quelle équation il approche, en particulier dans le cas non-linéaire.

Ces considérations mettent en évidence les avantages des schémas absolument stables et consistants.

Maintenant on peut formuler une hypothèse :

Les schémas qui sont absolument stables et absolument consistants se trouvent dans –et seulement dans– la classe des schémas implicites. C'est presque évident pour les équations à coefficients constants (voir [3], [4]). Cette propriété met en relief la classe des schémas implicites comme la seule classe de schémas contenant les schémas absolument stables et consistants.

III. On peut citer beaucoup de problèmes dans la mécanique du continu pour lesquels l'application des schémas implicites est justifiée. L'utilisation des schémas implicites pour les équations paraboliques possédant un domaine de dépendance infini semble très naturelle parce que le schéma implicite a la même propriété. Pour la même raison ils sont peu répandus dans le cas d'équations hyperboliques.

Nous résumons brièvement les arguments en faveur des schémas implicites qui sont donnés dans le livre de Rojdestvenski, Ianenko [5]. Le critère de Courant qui apparaît dans les schémas explicites implique des restrictions sur le pas τ suivant la valeur de la vitesse du son. En même temps la précision ne dépend que des gradients des variables d'écoulements. Pour les écoulements lents et les milieux faiblement compressibles (la météorologie dynamique, les écoulements dans les conduits naturels ou artificiels), lorsque $\frac{|u|}{C} \leqslant 1$ (u-vitesse du fluide, C-celle du son) le désaccord entre les restrictions sur τ imposées par les critères de la stabilité et de la précision est très grand et le schéma implicite peut être recommandé.

Si, au contraire, le milieu est compressible et l'écoulement très variable c'est-à-dire si la condition $\frac{C_{\min}}{C_{\max}} \leqslant 1$ a lieu, le critère local de stabilité qui a la forme $\frac{\tau}{h} \leqslant \frac{1}{C_{\max}}$ est très exigeant, le pas τ est trop petit. Cela implique un temps machine excessif et augmente aussi l'effet de la viscosité artificielle.

La méthode de stationnarisation pour les problèmes stationnaires et *autosemblables* devient très efficace si l'on applique les schémas implicites. En fait dans ce cas les restrictions sur τ imposées par les conditions de la précision sont superflues et on peut choisir τ arbitrairement grand et accélérer la convergence.

Les propriétés de la précision et de la stabilité sont indépendantes, en plus elles sont en un certain sens en contradiction : en augmentant la précision on diminue la stabilité. Ainsi pour les schémas de haute précision l'application des schémas implicites devient nécessaire.

Enfin, notons l'application de plus en plus fréquente d'*interpolation globale* (à l'aide de spline fonctions) qui est en fait une variante du schéma implicite.

Notre “éloge” des schémas implicites ne doit pas vous amener à la fausse conclusion qu’ils soient les seuls à employer. Je voudrais seulement remarquer que les schémas explicites et implicites doivent constituer les parties égales d’un algorithme optimal.

Un rôle particulier et important appartient aux schémas absolument stables (explicites ou implicites) mais conditionnellement consistants. Ils engendrent plusieurs semi-groupes et cette circonstance les rend utiles pour la description des écoulements d’un type mixte laminaires-turbulents (voir [6]).

IV. Le schéma implicite absolument stable (2) (on peut l’appeler *le schéma de simple approximation*) est très efficace pour un problème unidimensionnel et inefficace dans le cas de plusieurs dimensions. En fait, l’inversion de l’opérateur $I - \tau\Lambda_1$ qui est nécessaire pour la résolution du schéma (2) nécessite const. $N^{\alpha(m)}$ opérations où N est le nombre de points dans une direction et $\alpha(m)$ croît rapidement avec la dimension d’espace m. Par exemple, pour l’équation de la chaleur $\alpha(1) = 1$, $\alpha(2) = 3$. Ainsi, le nombre d’opérations pour un nombre de points donné croît avec la dimension d’espace et le schéma (2) devient inefficace.

Nous dirons suivant Samarski que le schéma implicite est *économique*, si $\alpha(m) = m$. Il est possible de construire des schémas absolument stables, absolument consistants et économiques. C’est le point de départ de la méthode des pas fractionnaires [3], [7], [13], qui est basée sur les idées principales suivantes :

(1) *la décomposition* (*splitting up*) des schémas,

(2) *la factorisation approchée* d’un opérateur,

(3) *l’approximation faible* des équations différentielles.

Dans le cas de l’équation (1) et du schéma (2) des schémas correspondants à pas fractionnaires peuvent s’écrire sous la forme suivante (pour la simplicité nous n’utilisons que deux pas fractionnaires) :

$$(1) \quad \frac{u^{n+\frac{1}{2}} - u^n}{\tau} = \Lambda_{11} u^{n+\frac{1}{2}} + \Lambda_{01} u^n, \quad \frac{u^{n+1} - u^{n+\frac{1}{2}}}{\tau} = \Lambda_{12} u^{n+1} + \Lambda_{02} u^{n+\frac{1}{2}},$$

$$\Lambda_{11} + \Lambda_{12} = \Lambda_1, \ \Lambda_{01} + \Lambda_{02} = \Lambda_0$$

(schéma de splitting up),

$$(2) \qquad (I - \tau\Lambda_{11})(I - \tau\Lambda_{12}) u^{n+1} = (I + \tau\Omega) u^n,$$

$$\Lambda_{11} + \Lambda_{12} = \Lambda_1, \ \Omega \sim \Lambda_0,$$

(schéma de factorisation approchée),

$$(3) \qquad \frac{\partial u}{\partial t} = (\alpha_1 \mathcal{L}_1 + \alpha_2 \mathcal{L}_2) u = \mathcal{L} u, \ \mathcal{L}_1 + \mathcal{L}_2 = \mathcal{L},$$

$$\alpha_1(t, \tau) = 2, \ \alpha_2(t, \tau) = 0, \ t \in \left[n\tau, \left(n + \frac{1}{2}\right)\tau\right),$$

$$\alpha_2(t,\ \tau) = 0, \alpha_2\ (t, \tau) = 2, t \in \left[\left(n + \frac{1}{2}\right)\tau, (n + 1)\,\tau\right)$$

(schéma d'approximation faible).

Si les opérateurs considérés sont commutatifs, les schémas 1, 2 sont équivalents sous la condition que $\Omega \equiv \Lambda_{01} + \Lambda_{02} + \tau\Lambda_{01}\Lambda_{02}$. Dans ce cas l'inversion de la matrice $I - \tau\Lambda_1$ se réduit à l'inversion d'une matrice factorisée $(I - \tau\Lambda_{11})$ $(I - \tau\Lambda_{12})$, c'est-à-dire à l'inversion successive des matrices $I - \tau\Lambda_{11}, I - \tau\Lambda_{12}$ qui sont en général d'une structure plus simple. Si la condition $\Lambda_{11} + \Lambda_{12} \sim \Lambda_1$ est vérifiée, la relation $I - \tau\Lambda_1 \sim (I - \tau\Lambda_{11})\ (I - \tau\Lambda_{12})$ de la factorisation approchée a lieu.

L'interprétation 3 nous permet de traiter le schéma (1) de splitting up comme une approximation simple de l'équation 3.

V. Les schémas implicites ont une position particulière dans la méthode de stationnarisation. Soient

(7) $$\mathfrak{L}u = f, x \in \Omega, lu = g, x \in \partial\Omega$$

les équations qui déterminent un problème aux limites dans le domaine Ω, $\mathfrak{L}, l$ étant des opérateurs différentiels, nonlinéaires en général.

La méthode de stationnarisation est basée sur la relation

(8) $$U(x) = \lim u(x, t), t \to \infty$$

où $u(x, t)$ est la solution du problème nonstationnaire avec les mêmes conditions aux limites stationnaires. Si la relation (8) a lieu, $U(x)$ est une solution (pas obligatoirement unique) du problème stationnaire original.

D'habitude la relation entre le système (7) stationnaire et le système (1) nonstationnaire correspondant est naturelle, t étant le temps physique. Or si l'état transitoire est sans intêret et que seul l'état final nous intéresse, on peut considérer d'autres systèmes évolutifs

(9) $$\mathfrak{R}\left(x, t, u, \frac{\partial u}{\partial t}, \ldots, \frac{\partial^m u}{\partial t^m}\right) = \mathfrak{L}u - f$$

L'opérateur $\mathfrak{R}$ s'appelle l'*opérateur de la relaxation*, l'équation (9) n'a aucune signification physique, t étant le paramètre de relaxation. Or, comme dans le cas précédent si la solution $u(x, t)$ de (9) converge vers une fonction $U(x)$ celle-ci est la solution du problème original (7).

Si l'opérateur $\mathfrak{R}$ est linéaire par rapport à $\dfrac{\partial u}{\partial t}, \ldots, \dfrac{\partial^m u}{\partial t^m}$ le schéma itératif correspondant prend la forme suivante

(10) $$(B_1\Delta_0 + B_2\Delta_0^2 + \ldots + B_m\Delta_0^m)\,u^n = \Lambda u^n - f,$$

où $$\Delta_0 u^n = \frac{u^{n+1} - u^n}{\tau}$$

(schéma d'algorithme universel).

Si l'on applique la méthode de factorisation approchée à l'opérateur Bm on obtient le schéma factorisé :

$$(12) \qquad B_1 \Delta_0 + B_2 \Delta_0^2 + \ldots + B_1 \ldots B_p \Delta_0^m u^n = \Lambda u^n - f.$$

qui est facilement résoluble.

Comme nous l'avons déjà dit pour les schémas d'algorithme universel la condition de consistance

$$(13) \qquad B_1 \sim I, B_2 \ldots B_m \sim 0$$

est obligatoire, s'il est nécessaire de connaître les états intermédiaires du milieu.

Notons une propriété très intéressante des schémas itératifs. Les schémas de splitting up sont consistants avec le système évolutif mais à condition que $\tau \to 0$ (*approximation incomplète*). Les schémas d'algorithme universel garantissent par contre la convergence pour τ arbitraire (*approximation complète*) mais en général ne sont pas consistants avec le système évolutif. Pour cette raison nous considérons comme très bons les schémas de l'approximation complète qui en même temps sont consistants avec le système évolutif. Nous appelerons de tels schémas *harmonieux*. Par exemple le schéma

$$(I - \tau\Lambda_1)(I - \tau\Lambda_2) \frac{u^{n+1} - u^n}{\tau} = (\Lambda_1 + \Lambda_2) u^n - f$$

est un schéma harmonieux.

Les schémas harmonieux garantissent la convergence pour τ arbitraire en permettant en même temps de reproduire tout le processus physique de stationnarisation.

VI. La décomposition des opérateurs ne nous conduit pas toujours aux opérateurs unidimensionnels. En 1959 l'auteur a publié [11] une formulation général des schémas à pas fractionnaire pour les équations évolutives. Mais c'est seulement à présent qu'on peut voir assez clairement la liaison entre les schémas de Harlow [13], [14] et la méthode des pas fractionnaires. Nous discuterons ici très brièvement ce fait remarquable.

Un trait caractéristique des schémas numériques pour les problèmes à plusieurs dimensions de l'hydrodynamique est la nécessité de garder l'information simultanément sur les éléments matériels (particules) et ceux d'espace (cellules). L'application des coordonnées euleriennes ou lagrangiennes (ou de leur combinaison) ne résout pas complètement ce problème. Le schéma de Harlow garde l'information simultanément sur les éléments materiels, et ceux d'espace, en les unissant à l'aide d'une interpolation convenable. Si l'on applique le langage de la faible approximation pour les équations de l'hydrodynamique

$$\frac{\partial \rho}{\partial t} + \frac{\partial (\rho u)}{\partial x} + \frac{\partial (\rho V)}{\partial y} = 0$$

$$\frac{\partial(\rho u)}{\partial t}+\frac{\partial(\rho u^2)}{\partial x}+\frac{\partial(\rho u\,v)}{\partial y}+\frac{\partial p}{\partial x}=0,\ \frac{\partial(\rho\,v)}{\partial t}+\frac{\partial(\rho u\,v)}{\partial x}+\frac{\partial(\rho\,v^2)}{\partial y}+\frac{\partial p}{\partial y}=0,$$

(15)

$$\frac{\partial(\rho E)}{\partial t}+\frac{\partial(\rho uE)}{\partial x}+\frac{\partial(\rho\,vE)}{\partial y}+\frac{\partial(up)}{\partial x}+\frac{\partial(v\,p)}{\partial y}=0\ ,\ E=\epsilon+\frac{1}{2}(u^2+v^2)$$

le splitting up d'après Harlow se présente sous la forme suivante :

$$\frac{\partial\rho}{\partial t}=0\ ,\frac{1}{2}\frac{\partial(\rho u)}{\partial t}+\frac{\partial p}{\partial x}=0\ ,\frac{1}{2}\frac{\partial\rho\,v}{\partial t}+\frac{\partial p}{\partial y}=0,$$

(16)

$$\frac{\partial(\rho E)}{\partial t}+\frac{\partial(\rho u)}{\partial x}+\frac{\partial(v\rho)}{\partial y}=0$$

(le premier pas fractionnaire ou la première étape d'après Harlow),

$$\frac{1}{2}\frac{\partial\rho}{\partial t}+\frac{\partial(\rho u)}{\partial x}+\frac{\partial(\rho\,v)}{\partial y}=0,\ \frac{1}{2}\frac{\partial(\rho u)}{\partial t}+\frac{\partial(\rho u^2)}{\partial x}+\frac{\partial(\rho u\,v)}{\partial y}=0,$$

(17)

$$\frac{1}{2}\frac{\partial(\rho\,v)}{\partial t}+\frac{\partial(\rho u\,v)}{\partial x}+\frac{\partial(\rho\,v^2)}{\partial y}=0,\ \frac{1}{2}\frac{\partial(\rho E)}{\partial t}+\frac{\partial(\rho uE)}{\partial x}+\frac{\partial(\rho\,vE)}{\partial y}=0$$

(le deuxième pas fractionnaire ou la deuxième étape).

A la première étape on obtient le système des équations de l'hydrodynamique sans les termes de convection, dans la deuxième étape on obtient le système où tous les changements de variables de l'écoulement sont dus aux termes de convection. Si les deux systèmes étaient résolus dans le même maillage on obtiendrait le schéma de splitting up ordinaire. D'après Harlow le premier système est integré au moyen d'un schéma ordinaire dans le maillage eulerien, dans le second système les flux sont associés aux éléments matériels (particules) qui se meuvent dans le réseau eulerien. Ainsi, les particules se meuvent d'une cellule à l'autre en portant la masse, l'impulsion, l'énergie. Une telle interprétation formalisée a permis de diminuer les fluctuations des schémas PIC (voir [15]).

Notons que les équations (15-17) peuvent s'écrire sous la forme de lois de conservation et cela est à notre avis la formulation la plus complète de la méthode de la faible approximation pour les lois de conservation qui gouvernent l'écoulement d'un milieu.

VII. Maintenant, je voudrais bien attirer votre attention sur la liaison entre la théorie de splitting up et celle des semi-groupes. Comme l'a indiqué M. Temam [16] le théorème de Trotter [17] sur la décomposition des opérateurs infinitésimaux des semi-groupes est étroitement lié à la méthode de splitting up (d'éclatement). Dans les cas les plus simples les deux points de vue sont identiques. Or la méthode de splitting up est plus riche tant pratiquement que théoriquement, parce que la décomposition des opérateurs dans la méthode de splitting up s'effectue sous des conditions plus faibles que celles formulées dans les travaux de Trotter, de Chernoff et d'autres auteurs.

Dans la note [8] l'auteur a considéré les schémas de splitting up où les schémas intermédiaires ne sont pas obligatoirement stables ce qui est supposé dans les travaux de Trotter. Cependant à notre avis cette belle théorie de Trotter peut être étendue aux cas où les problèmes de Cauchy intermédiaires ne sont pas corrects.

VIII. A présent nous avons beaucoup de schémas efficaces qui nous permettent de représenter l'écoulement d'un milieu avec toutes les singularités qui peuvent se présenter : les ondes de choc, les ondes de raréfaction, les surfaces de contact (les frontières). Dans la plupart des cas ils ont une structure uniforme et traitent de la même manière les régions régulières et singulières de l'écoulement. Dans les premiers schémas de ce genre [19, 20] la viscosité artificielle imitait la viscosité physique et était introduite additivement dans la pression.

Ensuite ont été construits des schémas de structure uniforme, dont la viscosité artificielle était liée à la structure même du schéma (viscosité d'approximation). Tels étaient par exemple les schémas [21, 22] de Lax et de Godunoff. L'application des schémas de splitting up et en particulier des schémas PIC de Harlow rend la structure de la viscosité artificielle plus compliquée et l'analyse des propriétés dissipatives du schéma apparaît ainsi comme un problème assez difficile.

Dans les travaux de Shokin et de l'auteur [23, 26] et indépendamment dans une note de Hirt [27] des notions ont été introduites qui ont permis de décrire la structure de la viscosité artificielle et les propriétés dissipatives du schéma.

Nous introduisons la notion de *première approximation différentielle* (PAD) du schéma en prenant pour simplifier le cas d'un système hyperbolique, quasilinéaire à une variable d'espace :

(18) $$\frac{\partial u}{\partial t} + A_1 \frac{\partial u}{\partial x} = 0, \; A_1 = A_1 (x, t, u)$$

et le schéma explicite

(19) $$u^{n+1}(x) = \sum_{a} B_a u^n (x + \alpha h), \frac{\tau}{h} = \kappa = \text{const.}$$

En développant le schéma (19) en série de τ, h et en ne gardant que les termes du premier ordre on obtient la première approximation différentielle du schéma (19) sous forme hyperbolique

(20) $$\frac{\partial u}{\partial t} + \frac{\tau}{2} \frac{\partial^2 u}{\partial t^2} = A_1 \frac{\partial u}{\partial x} + C_1 \frac{\partial^2 u}{\partial x^2},$$

où

(21) $$C_1 = \frac{h^2}{2\tau} \sum_{a} \alpha^2 B_\alpha$$

On peut éliminer le terme $\frac{\partial^2 u}{\partial t^2}$ en l'exprimant à l'aide des dérivées d'espace en utilisant l'équation (18). Alors on obtient la première approximation différentielle sous forme parabolique

$$\text{(22)}\qquad \begin{gathered} \frac{\partial u}{\partial t} = A_2 \frac{\partial u}{\partial x} + C_2 \frac{\partial^2 u}{\partial x^2}, \\ A_2 = A_1 - \frac{\tau}{2}\left[\frac{\partial A_1}{\partial t} + A_1 \frac{\partial A_1}{\partial x}\right], C_2 = C_1 - \frac{\tau}{2} A_1^2 , \end{gathered}$$

Les systèmes différentiels $\frac{\partial u}{\partial t} = A_i \frac{\partial u}{\partial x_i}$ $(i = 1, 2)$ pour τ assez petit sont hyperboliques, le terme $C_2 \frac{\partial^2 u}{\partial x^2}$ donne la parabolicité et est responsable des propriétés dissipatives et de la stabilité du schéma. Dans le cas linéaire et pour quelques classes de schéma (les schémas simples, majorants, etc.) Shokin et l'auteur [23], [25] ont démontré que le schéma (19) est stable si, et seulement si, le système (20) est faiblement parabolique, c'est-à-dire $C \geqslant 0$, l'une des racines caractéristiques peut être nulle, les autres étant positives.

La notion de première approximation différentielle et quelques résultats sont valables dans le cas unidimensionnel. Des théorèmes de ce genre sont démontrés par Shokin [29], [28] pour les schémas de splitting up en terme de première approximation différentielle tant sous forme hyperbolique que parabolique. Pour quelques schémas les théorèmes restent valables aussi dans le cas quasilinéaire.

Des comparaisons ont été faites entre le critère local de stabilité (ânalyse locale de Fourier) et celui de PAD. Pour la plupart des schémas connus les deux critères coincident.

IX. C'est par le comportement de la solution numérique dans le voisinage d'une discontinuité que se manifeste le plus visiblement les propriété dissipatives d'un schéma. Les schémas conservatifs uniformes peuvent différer l'un de l'autre par les profils d'écoulement mais cela ne les empêche pas de donner des résultats identiques pour l'onde de choc, sa force et sa vitesse.

Des différences beaucoup plus grandes se manifestent dans le comportement des solutions numériques dans un voisinage de la surface de contact. Selon leur comportement sur la surface de contact les schémas peuvent être divisés en deux classes :

(1) Les schémas de la première classe remplacent surface de contact par une zone de contact. Dans cette zone de transition les profils numériques diffèrent beaucoup des solutions exactes, mais la largeur de la zone ne croît pas avec le temps.

Par exemple le schéma de la croix (le schéma de Neumann-Richtmyer) appartient à cette classe.

(2) Les schémas de la deuxième classe remplacent la surface de contact par une zone de transition qui s'élargit avec le temps. Une diffusion de masse par la surface et le réchauffement artificiel du milieu est caractéristique de ces schémas. Par exemple, le schéma de Lax appartient à cette classe. Dans les notes de Shokin et de l'auteur la notion de *K-propriété* est introduite qui permet à notre avis de distinguer les schémas des deux classes.

Par définition, le schéma (19) possède la propriété K, si la viscosité d'approximation ne fonctionne pas le long de la trajectoire, c'est-à-dire si la direction tangente à la trajectoire $X : \frac{\partial x}{\partial t} = u$ $(XA = uX)$ est un vecteur zéro de la matrice $C_2 : XC_2 = 0$.

Les schémas de Neumann-Richtmyer, Lax-Wendroff, prédicteur-correcteur, les schémas majorants possèdent contrairement au schéma de Lax, la propriété K. La propriété K définie ci-dessus est formulée au moyen de la PAD. On peut donner une autre formulation (plus forte) au moyen du schéma lui-même, en remplaçant la condition $XC_2 = 0$ par la condition $X(\sigma - I) = 0$, où σ est l'opérateur de transition du schéma. Les schémas ci-dessus qui possèdent la K-propriété faible possédent également la K-propriété forte.

X. La variété des schémas utilisés dans les calculs hydrodynamiques rend nécessaire une classification, en premier lieu selon leurs propriétés dissipatives. En fait, la PAD qui conserve l'information suffisante du schéma même est par définition une équation différentielle à laquelle s'applique entièrement la théorie des groupes continus de Lie.

C'est un fait bien connu, que le système d'équations de l'hydrodynamique correspondant à une équation d'état arbitraire admet un groupe continu de transformations dans l'espace x, y, t, u, v, ρ, p qui a comme base les sept groupes uniparamètriques suivants [31] :

1 - 3. Les déplacements le long des axes t, x, y,

4 - 5. Les transformations de Galilée dans les directions x, y,

6 . Les rotations dans le plan xy,

7 . Les homothéties dans l'espace t, x, y.

Nous dirons brièvement qu'un schéma numérique admet un groupe continu de transformations si sa PAD admet ce même groupe.

Il est bien naturel de poser la question s'il existe des schémas qui admettent le même groupe que l'équation originale. Pour cela il faut et il suffit que la dérivée de Lie pour les équations de la PAD, correspondant à chaque opérateur infinitésimal engendrant le groupe, soit identiquement nulle d'après les équations de la PAD.

Une telle analyse des schémas basée sur leur PAD est donnée dans les travaux de Shokin et de l'auteur [32]. On a démontré qu'il est possible de construire des schémas invariants c'est-à-dire des schémas qui admettent le même groupe de transformations que le système des équations de l'hydrodynamique.

XI. Jusqu'à présent nous avons considéré seulement les schémas du premier ordre de précision. La construction de schémas de haute précision est un problème qui est très compliqué et qui n'est pas encore résolu même pour les équations linéaires classiques de la physique mathématique.

On peut aisément construire des schémas de haute précision de simple approximation, mais leur résolution est très difficile parce que la contradiction entre la stabilité et la précision augmente parallèlement avec la croissance de la précision et de la dimension. Par exemple des schémas de haute précision sont depuis longtemps connus pour l'équation de Laplace, mais leur résolution efficace nécessite l'application des schémas à pas fractionnaires. Cela est démontré dans les travaux de Douglas et Gunn [33] et de Samarski [34]. Dans une note de Valiullin et de l'auteur [36] une construction générale des schémas de haute précision est donnée pour le cas d'équations polyharmoniques y compris les équations de l'élasticité.

La construction et l'analyse numérique des schémas de haute précision est une nécessité pressante pour les problèmes de l'hydrodynamique, en particulier dans les calculs des écoulements d'un fluide visqueux ayant un grand nombre de Reynolds et dans les calculs où la structure fine de l'écoulement doit être détectée (configuration de Mach, ondes de choc "suspendues", apparition des surfaces de contact, etc.).

Il est bien évident que les difficultés de construction de schémas de haute précision se multiplient pour les équations de l'hydrodynamique, à cause du manque des propriétés si agréables aux mathématiciens (commutativité, linéarité, symétrie des opérateurs considérés) et en particulier à cause de l'apparition de discontinuités dans l'écoulement.

De l'expérience numérique il découle que les schémas de haute précision en augmentant définitivement la précision dans la région régulière la perdent dans le voisinage d'une discontinuité. Ainsi la précision a un caractère local. Pour atteindre la précision globale on a besoin de schémas d'une précision uniforme.

XII. Jusqu'à présent nous avons parlé des schémas uniformes qui traitent de la même façon toutes les singularités de l'écoulement. Dans ces schémas les surfaces de discontinuité sont transformées en zônes de transition avec des profils réguliers. Dans un but pratique et pour l'accroissement de la précision il est nécessaire de détecter les surfaces de discontinuité disparues. Une telle détection s'effectue à l'aide d'un analyseur différentiel. De cette manière Kuropatenko [35] a appliqué l'analyseur différentiel pour la détection des ondes de choc dans le cas unidimensionnel. Dans son schéma le terme de viscosité est présent additivement dans la pression et l'on peut détecter l'onde de choc dans les endroits où ce terme atteint son maximum. Les analyseurs différentiels ont été utilisés dans les calculs par Wilkins et Moretti.

XIII. Cette revue que j'ai l'honneur de vous présenter est très brève et sans doute elle passe sous silence beaucoup d'autres problèmes qui apparaissent dans la construction des schémas numériques. On peut seulement constater que nous sommes au commencement de la route vers la construction d'une théorie des schémas numériques appliquées en mécanique des milieux continus. Je voudrais bien souligner que la construction des schémas optimaux dans un certain sens doit être basée sur les critères théoriques et sur un grand travail de classification.

Il est bien évident que les critères imposés aux schémas doivent avoir un caractère physique donc être invariants et cela nous conduit inévitablement à une application de la théorie des groupes continus.

BIBLIOGRAPHIE

[1] Тихонов А. Н., Самарский А. А. — ЖВМ и МФ, I, I, 1961, 5-64.

[2] Соболев С. Л. — Замыкание вычислительных алгоритмов и некоторые его применения. М., 1955.

[3] Ianenko N. N. — *La méthode à pas fractionnaires*. Armand Colin, 1968.

[4] Ильин А. М. — Доклады АН СССР, 164, 3, 1965, 491-494.

[5] Рождественский Б. Л., Яненко Н. Н. — Системы квазилинейных урав нений. М., 1968.

[6] Fromm J. E. — *The Physics of Fluids*, 12, part II, 1969, p. 3-12.

[7] Douglas Jim Jr. — *J. Soc. Ind. Appl. Math.*, 3, 1, 1955, p. 42-65.

[8] Peaceman D. W., Rachford H. H. Jr. — *J. Soc. Ind. Appl. Math.*, 3, 1, 1955, p. 28-42.

[9] Самарский А. А. — Лекции по теории разностных схем. М., 1969.

[10] Дьяконов Е. Г. — Доклады АН СССР, 138, 3, 1961.

[11] Яненко Н. Н. — Доклады АН СССР, 1959.

[12] Marchuk G. I., Ianenko N. N. — *I.F.I.P.*, 1965.

[13] Harlow F. H. — *J. Assoc. Comput. Math.*, 4, 2, 1967, p. 137-142.

[14] Anuchina N. N., Petrenko V. E., Shokin U. I., Ianenko N. N. — *Fluid Dynamics Transactions*, 5, 1, 1970.

[15] Яненко Н.Н., Анучина Н. Н., Петренко В. Е., Шокин Ю. И. — Информационный бюллетень « Численные мет. мех. сплошн. среды », I, I, 1970, 40-62.

[16] Temam R. — *La thèse doctorale*, 1968.

[17] Trotter H. F. — *Pacific J. Math.*, 8, 4, 1958, p. 887-919.

[18] Яненко Н. Н. — *Aplikace Matematiky*, 13, 1968, p. 148-151.

[19] Von Neumann J., Richtmyer R. D. — *J. Appl. Phys.*, 21, 1950, p. 232-237.

[20] Latter R. — *J. Appl. Phys.*, 26, 8, 1955, p. 955-960.

[21] Lax P. D. — *Comm. Pure Appl. Math.*, 7, 1, 1954, p. 159-193.

[22] Годунов С. К. — Математ. сборник, 47 (89), 3, 1959, 271-306.

[23] Яненко Н. Н., Шокин Ю. И. — Доклады АН СССР, 182, 4, 1968, 776-778.

[24] Яненко Н. Н., Шокин Ю. И. — Доклады АН СССР, 182, 2, 1968, 280-281.

[25] Яненко Н. Н., Шокин Ю. И. — Сибирский мат. журнал., 10, 5, 1969, 1174-1188.

[26] Яненко Н. Н., Шокин Ю. И. — The Physics of Fluids, 12, part II, 1969, 28-33.

[27] Hirt C. W. — *J. Comput. Phys.*, 2, 4, 1968, p. 339-355.

[28] Шокин Ю. И. — Труды всесоюзного семинара по числ. мет. мех. вязкой жидкости. Новосибирск, 1969.

[29] Шокин Ю. И. — Известия Сибирского отделения АН СССР, серия техн. наук., 8, вып. 2, 1970, 81-85.

[30] Lax P. D., Wendroff B. — *Comm. Pure Appl. Math.*, 13, 1, 1960, p. 217-238.

[31] Овсянников Л. В. — Групповые свойства дифференциальных уравнений. Новосибирск, 1962.

[32] Ianenko N. N., Shokin U. I. — (à paraître).

[33] Douglas J. Jr., Gunn I. E. — *Math. Comp.*, 17, 81, 1963, p. 71-80.

[34] Самарский А. А. — ЖВМ и МФ, 3, 6, 1963.

[35] Куропатенко В. Ф. — Труды МИАН, 74, 1966, стр. 107-137.

[36] Валиуллин А. Н., Яненко Н. Н. — Изв. СО АН СССР, сер. техн. наук, вып. 3, 1967, 88-96.

Centre de Calcul
Branche Sibérienne de l'Académie des
Sciences de l'U.R.S.S
Novosibirsk 90
U.R.S.S

Actes, Congrès intern. Math., 1970. Tome 3, p. 311 à 319.

QUELQUES MÉTHODES DE DÉCOMPOSITION EN ANALYSE NUMÉRIQUE

par R. TEMAM

Introduction.

On entend par "méthode de décomposition" en Analyse Numérique, les méthodes numériques qui ramènent la résolution d'un problème donné à la solution d'une succession de problèmes plus simples, exploitant pour cela une propriété de décomposition du problème initial. Les méthodes de décomposition les plus connues sont évidemment les méthodes de Directions Alternées [11-12-34], et de Pas Fractionnaires [28-30-37-46-47-38], où l'on utilise des décompositions d'ailleurs fort simples, des opérateurs différentiels impliqués. D'autres algorithmes numériques plus récents ou moins récents que les précédents se rattachent à cette idée ; en particulier la méthode d'éclatement en calcul des variations (ou optimisation) [26-27], et des techniques de décomposition en programmation linéaire [9] et non linéaire [20]. Les méthodes de décomposition ont fourni dans certains cas des algorithmes avantageux ; et plus encore, dans d'autres cas, elles ont rendu possible la résolution de problèmes dépassant les possibilités de l'ordinateur disponible.

Dans ce qui suit nous nous proposons d'évoquer quelques résultats récents sur ce sujet. Tous les travaux que nous évoquerons sont centrés, dans leur aspect théorique, autour des thèmes suivants : étude de problèmes de stabilité et de convergence pour des algorithmes de pas fractionnaires ou de directions alternées déjà connus ; applications nouvelles de ces méthodes à d'autres problèmes aux limites de la physique mathématique et de l'engineering, avec étude de ces mêmes problèmes de stabilité et de convergence ; enfin, étude d'algorithmes de décomposition pour des problèmes d'optimisation ou de théorie des jeux.

En raison du peu de place dont nous disposons nous nous bornerons à décrire succintement certains algorithmes et les résultats démontrés. Mais, le plus souvent, nous nous contenterons de références bibliographiques, renvoyant ainsi à des articles parus ou à paraître. Pour les mêmes raisons nous n'évoquerons pas les aspects "pratiques" de ces travaux : de très nombreux calculs ont maintenant été fait sur ces méthodes (cf. en particulier [5-7-8-13-14-17-21-32]) conduisant parfois à certaines améliorations empiriques mais efficaces.

Plan.

(1) Equations d'évolution,

(2) Equations elliptiques,

(3) Optimisation et théorie des jeux.

1. Equations d'évolution.

1.1. *Un exemple : les équations de Carleman.*

Soit Ω le rectangle $]a_1, a_2[\times]b_1, b_2[$ du plan $0xy$ et $Q = \Omega \times]0, T[$. Le système de Carleman (un système hyperbolique non linéaire de la théorie cinétique des gaz) est le suivant : il s'agit de trouver deux fonctions $u(x, y, t)$, $v(x, y, t)$, réelles *positives*, définies dans Q et vérifiant

$$\frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} + u^2 - v^2 = 0 \quad , \quad \frac{\partial v}{\partial t} + \frac{\partial v}{\partial y} + v^2 - u^2 = 0 \, , \tag{1.1}$$

$$u(a_1, y, t) = v(x, b_1, t) = 0 \, , \tag{1.2}$$

$$u(x, y, 0) = u_0(x, y) \quad , \quad v(x, y, 0) = v_0(x, y) \, . \tag{1.3}$$

On sait [41] que pour u_0 et v_0 donnés dans $H^1(\Omega) \cap L^\infty(\Omega)$, réels *positifs* et vérifiant

$$u_0(a_1, y) = v_0(x, b_1) = 0, \tag{1.4}$$

le système (1.1) - (1.3) possède une solution unique $\{u, v\}$, $u \geqslant 0$, $v \geqslant 0$, et $u, v \in L^\infty([0, T] \; ; \; H^1(\Omega)) \cap L^\infty(Q)$.

Dans le cas des équations de Carleman, les méthodes de décomposition permettent d'obtenir un schéma d'approximation particulièrement simple. Une interprétation de la méthode des pas fractionnaires conduit en effet au schéma suivant : soit N un entier fixé et $k = T/N$; on définit des $u^{n+\frac{i}{2}}$, $v^{n+\frac{i}{2}}$, fonctions réelles *positives* dans $H^1(\Omega) \cap L^\infty(\Omega)$, $i = 1, 2$, $n = 0, \ldots, N-1$.

On part de

$$u^0 = u_0 \quad , \quad v^0 = v_0 \, , \tag{1.4}$$

puis, lorsque $u^0, v^0, \ldots, u^n, v^n$, sont connus, on définit $u^{n+\frac{1}{2}}$, $v^{n+\frac{1}{2}}$ par

$$\frac{u^{n+\frac{1}{2}} - u^n}{k} + \left(u^{n+\frac{1}{2}}\right)^2 - \left(v^{n+\frac{1}{2}}\right)^2 = 0 \, , \tag{1.5}$$

$$\frac{v^{n+\frac{1}{2}} - v^n}{k} + \left(v^{n+\frac{1}{2}}\right)^2 - \left(u^{n+\frac{1}{2}}\right)^2 = 0 \, , \tag{1.6}$$

et ensuite u^{n+1} et v^{n+1} par

$$\frac{u^{n+1} - u^{n+\frac{1}{2}}}{k} + \frac{\partial u^{n+1}}{\partial x} = 0 \quad , \quad u^{n+1}(a_1, y) = 0 \tag{1.7}$$

$$(1.8)\qquad \frac{v^{n+1} - v^{n+\frac{1}{2}}}{k} + \frac{\partial v^{n+1}}{\partial y} = 0 \quad , \quad v^{n+1}(x\,,b_1) = 0\,.$$

Toutes ces équations se résolvent explicitement et très simplement :

$$(1.9)\qquad u^{n+\frac{1}{2}} = \frac{u^n + k(\sigma^n)^2}{1 + 2k\sigma^n} \quad , \quad v^{n+\frac{1}{2}} = \frac{v^n + k(\sigma^n)^2}{1 + 2k\sigma^n}\,,$$

où $\sigma^n(x\,,y) = u^n(x\,,y) + v^n(x\,,y)$, et,

$$(1.10)\qquad u^{n+1}(x\,,y) = \frac{1}{k}\int_{a_1}^{x} u^{n+\frac{1}{2}}(\xi\,,y)\exp\left(\frac{\xi - x}{k}\right)d\xi\,,$$

$$(1.11)\qquad v^{n+1}(x\,,y) = \frac{1}{k}\int_{b_1}^{y} v^{n+\frac{1}{2}}(x\,,y)\exp\left(\frac{\eta - y}{k}\right)d\eta\,.$$

On introduit les fonctions approximantes $u_{ik}\,, v_{ik}$:

$$(1.12)\qquad u_{ik}(t) = u^{n+\frac{i}{2}} \quad , \quad v_{ik}(t) = v^{n+\frac{i}{2}}\,,$$

pour $t \in [nk, (n+1)k[\,, i = 1, 2, n = 0, \ldots, N-1$.

On a alors le résultat de convergence suivant démontré dans [41] :

THEOREME 1.1. – *Pour* $k \to 0$, $u_{ik} \to u$, $v_{ik} \to v$ *dans* $\mathcal{C}([0\,,T]\ ;\ L^2(\Omega))$ *fort*, $L^\infty([0\,,T]\ ;\ H^1(\Omega))$ *et* $L^\infty(Q)$ *faible-étoile, où* $\{u\,, v\}$ *est la solution de* (1.1) - (1.3).

Remarque 1.1. – Dans le cas présent, la méthode d'approximation utilisée permet de démontrer l'*existence* d'une solution du problème (1.1) – (1.3).

Pour d'autres résultats d'existence pour le système de Carleman, cf. [3-10-19]. Un schéma d'approximation de (1.1) - (1.3), explicite à deux niveaux en t est étudié dans [35].

1.2. *Autres problèmes d'évolution.*

(a) *Equations générales.* Des théorèmes de stabilité et de convergence concernant l'approximation par la méthode des pas fractionnaires de certaines classes d'équations d'évolution linéaires et non linéaires, du premier et du second ordre en t, sont donnés dans [38], cf. aussi [1] et [22].

L'approximation par décomposition d'équations d'évolution couplées du type son-chaleur est étudiée dans [24]. Pour les équations de Schoedinger et du type de Schoedinger, cf [33]. L'approximation par pas fractionnaires de certaines inéquations d'évolution est développée dans [2]. On se reportera enfin à [16] et [22] pour certains résultats relatifs aux méthodes de directions alternées. Un autre aspect des méthodes de pas fractionnaires est la formule de Trotter des semi-groupes (cf. [44] pour le cas linéaire et [4] pour le cas non linéaire).

(b) *Equation de Riccati.* L'équation de Riccati de la théorie du contrôle est une équation d'évolution dans l'algèbre $\mathcal{L}(H)$ des opérateurs linéaires continus sur un espace de Hilbert H. Une variante de la méthode des pas fractionnaires nous a permis dans [43] de développer un schéma d'*approximation* simple du problème et d'obtenir également dans un cadre fonctionnel convenable, un résultat d'existence nouveau pour cette équation.

Une étude de l'approximation de l'équation de Riccati en dimension infinie est faite dans [31] par d'autres méthodes ; d'autres résultats d'existence pour cette équation sont donnés dans [3-10-25].

(c) *Equations de Navier-Stokes.* L'approximation des équations de Navier-Stokes des fluides visqueux incompressibles est étudiée à l'aide d'algorithmes du type pas fractionnaires dans [13-14-15-40]. Il n'est pas possible d'évoquer ici, même brièvement, ce sujet assez vaste.

L'approximation des équations de Navier-Stokes par des méthodes de pas fractionnaires est également étudiée par Chorin [7-8]. Pour d'autres études sur l'analyse numérique des équations de Navier-Stokes cf. [18-23-39-45].

(d) *Remarques diverses.* Parmi d'autres aspects intéressants de ces méthodes, citons

– certaines variantes des algorithmes de pas fractionnaires permettant le calcul en parallèle des étapes intermédiaires et pouvant peut être s'avérer utile à l'avénement de la prochaine génération d'ordinateurs [36-42].

– le problème du traitement des conditions aux limites, certaines précautions permettant d'améliorer sensiblement l'algorithme [21-37].

– divers recherches actuelles pour l'obtention de schémas "décomposés" à haute précision ; on trouvera dans [5] (resp. [29]) des schémas de pas fractionnaire d'ordre 2 (resp. 3) en t. Cela est lié à la formule de développement asymptotique de Campbell-Hausdorff [6] dont la formule de Trotter correspond au premier terme.

2. Equations elliptiques.

Soit A un opérateur non borné dans un espace de Hilbert H, admettant une décomposition

$$A = \sum_{i=1}^{q} A_i. \tag{2.1}$$

Pour résoudre l'équation [1]

$$Au = f\ , \tag{2.2}$$

on se donne une décomposition de f, $f = \sum_{i=1}^{q} f_i$, une famille de nombres $\tau_n > 0$, et on envisage alors le schéma de décomposition suivant :

[1] Pour des hypothèses plus précises, cf. [42],

$$(2.3) \qquad \frac{u^{n+\frac{i}{q}} - u^{n+\frac{i-1}{q}}}{\tau_n} + A_i u^{n+\frac{i}{q}} = f_i\,, \qquad (i = 1, \ldots, q\,, n > 0).$$

Pour un entier N donné on peut donner une majoration de l'erreur $\epsilon_N = |u^N - u|_H$ en fonction des paramètres $\tau_0, \ldots, \tau_{N-1}$, de l'algorithme et des données. Le calcul des paramètres τ_n optimaux (i.e. rendant l'erreur minimum) se fait par des relations de récurence simples ; pour tout cela cf. [42].

Il résulte des calculs d'erreur de [42] que l'erreur est d'autant plus faible que la quantité.

$$(2.4) \qquad \delta = \sum_{i=1}^{q} |f_i - A_i u|_H^2$$

est petite ; le schéma (2.3) est même exponentiellement convergeant si cette quantité était nulle, c'est-à-dire si par pur hasard, la décomposition de f était choisie en sorte que $f_i = A_i u$, $\forall i$. Cela nous a conduit à un algorithme dans lequel les $f_i = f_i^n$ étaient variables et "tendaient" vers $A_i u$:

$$(2.5) \qquad \frac{u^{n+\frac{i}{q}} - u^{n+\frac{i-1}{q}}}{\tau_n} + A_i u^{n+\frac{i}{q}} = f_i^n = A_i u^n + \frac{f - Au^n}{q}$$

Des calculs simples conduisent ici à des formules récurrentes donnant les paramètres optinaux τ_n et des bornes supérieures de l'erreur $\epsilon_N = |u^N - u|_H$; cette erreur ϵ_N vérifie une inégalité

$$(2.6) \qquad \epsilon_N \leqslant \alpha^N \epsilon_0\,,$$

où la constante $0 < \alpha < 1$ ne dépend que des données. Pour tout cela se reporter à [2] et [36].

Remarque 2.1. – L'algorithme (2.3) a été appliqué, sous une forme légèrement modifiée due à l'existence de dérivées croisées, à des problèmes d'élasticité bi- et surtout tridemensionnelle (cf. [17]).

3. Optimisation et théorie des jeux.

3.1. *Optimisation.*

Soient H un espace de Hilbert, V_i, $1 \leqslant i \leqslant q$, des espaces de Banach reflexifs, $V = \bigcap_{i=1}^{q} V_i$, avec

$$(3.1) \qquad V \subset V_i \subset H\,,$$

les injections étant continues et chaque espace étant dense dans le suivant.

Pour chaque i, soit K_i un ensemble convexe fermé de V_i, et soit $K = \bigcap_{i=1}^{q} K_i$.

Soit J_i une fonctionnelle réelle définie sur K_i, strictement convexe, semi-continue inférieurement et vérifiant

$$\lim_{u \in K_i, \|u\|_{V_i} \to \infty} \{J_i(u)\} = +\infty \tag{3.2}$$

Soit alors $J = \sum_{i=1}^{q} J_i$. Il est bien connu que le problème d'optimisation

$$\operatorname*{Inf}_{v \in K} J(v) \tag{3.3}$$

possède une solution unique u. On peut proposer un algorithme d'approximation de u par décomposition : soient $\tau > 0$ et N un entier fixé ; on définit par récurrence une famille d'éléments $u^{n+\frac{i}{q}}$, $u^0 \in H$ étant quelconque, et $u^{n+\frac{i}{q}}$ étant défini à partir de $u^{n+\frac{i-1}{q}}$ comme la solution dans K_i du problème d'optimisation

$$\operatorname*{Inf}_{v \in K_i} \left\{ \frac{1}{\tau} \left| v - u^{n+\frac{i-1}{q}} \right|^2 + J_i(v) \right\}. \tag{3.4}$$

Ayant ainsi défini les $u^{n+\frac{i}{q}}$, on introduit les moyennes (du type Cesaro)

$$w^{N+\frac{i}{q}} = \frac{1}{N} \sum_{n=0}^{N-1} u^{n+\frac{i}{q}} \in K_i. \tag{3.5}$$

On démontre alors le résultat suivant (cf. [27]) :

THEOREME 3.1. — *Sous les hypothèses précédentes, si* $\tau \to 0, N \to \infty$, *avec* $\tau N \to \infty$, *alors, pour tout* i, $1 \leqslant i \leqslant q$,

$$w^{N+\frac{i}{q}} \to u \tag{3.6}$$

solution de (3.3), *dans* V_i *faible.*

Remarque 3.1. — Dans certains cas on a des résultats de convergence ponctuelle : $u^N \to u$ dans H, $N \to \infty, \tau \to 0$.

3.2. *Théorie des jeux.*

On considère comme précédemment des espaces V_i, V, H et des convexes K_i et K ($i = 1, \ldots, q$). On se donne également des espaces W_i, W, G et des convexes L_i et L ($i = 1, \ldots, q$) vérifiant des hypothèses et relations analogues. Pour tout $i = 1, \ldots, q$, on se donne une fonctionnelle $J_i(v, w)$ définie sur $K_i \times L_i$, convexe s.c.i. en v, concave s.c.s. en w. On suppose également qu'il existe $\varphi \in K$ et $\psi \in L$, tels que pour $\forall i = 1, \ldots, q$.

$$\{J_i(v, \psi) - J_i(\varphi, w)\} \to +\infty, \tag{3.7}$$

$$\text{si} \quad v \in K_i \quad , \quad w \in L_i \quad , \quad \{\|v\|_{V_i} + \|w\|_{W_i}\} \to +\infty.$$

On introduit la fonctionnelle J :

$$J(v\,,w) = \sum_{i=1}^{q} J_i(v\,,w) \tag{3.8}$$

définie sur $K \times L$; la fonctionnelle J est convexe s.c.i. en v, concave s.c.s. en w et vérifie une propriété analogue à (3.7). Dans ces conditions la fonctionnelle J possède un point selle : il existe $\{v_*\,, w_*\} \in K \times L$ vérifiant

$$J(v_*\,, w_*) = \min_{v\in K}\max_{w\in L} J(v\,,w) = \max_{w\in L}\min_{v\in K} J(v\,,w). \tag{3.9}$$

L'algorithme de décomposition utilisant les décompositions précédentes de J, K et L est le suivant : pour $\tau > 0$ et N entier fixé on définit par récurrence des

$$v^{n+\frac{i}{q}} \in H \quad , \quad w^{n+\frac{i}{q}} \in G \quad , \quad i = 1,\dots,q \quad , \quad n = 0,\dots,N-1.$$

On part de $u^0 \in H$, $\;$, $w^0 \in G$ arbitraires ; lorsque $v^{n+\frac{i-1}{q}}$ et $w^{n+\frac{i-1}{q}}$ sont connus, on définit $v^{n+\frac{i}{q}}$ et $w^{n+\frac{i}{q}}$ comme réalisant le point selle de la fonctionnelle

$$\frac{1}{\tau}\left|v - v^{n+\frac{i-1}{q}}\right|_H^2 - \frac{1}{\tau}\left|w - w^{n+\frac{i-1}{q}}\right|_G^2 + J_i(v\,,w) \;; \tag{3.10}$$

(ce point selle existe grâce à (3.7)).

Ayant défini ces éléments on considère les moyennes ($i = 1,\dots,q$) :

$$\varphi^{N+\frac{i}{q}} = \frac{1}{N}\sum_{n=0}^{N-1} v^{n+\frac{i}{q}} \quad , \quad \psi^{N+\frac{i}{q}} = \frac{1}{N}\sum_{n=0}^{N-1} w^{n+\frac{i}{q}} \tag{3.11}$$

et on a le théorème d'approximation :

THEOREME 3.2. — *Sous les hypothèses précédentes, si $\tau \to 0$, $N \to \infty$, avec $\tau N \to \infty$, alors pour tout $i = 1,\dots,q$,*

$$\varphi^{N+\frac{i}{q}} \to v_* \quad \textit{dans } V_i \textit{ faible,} \tag{3.12}$$

$$\psi^{N+\frac{i}{q}} \to w_* \quad \textit{dans } W_i \textit{ faible} \tag{3.13}$$

$\{v_*\,, w_*\}$ étant le point-selle de la fonctionnelle J.

Pour cette application des méthodes de décomposition cf. [2].

BIBLIOGRAPHIE

[1] Bardos C., Sibony M. — *Journal A.F.I., R.O.*, 1969.

[2] Bensoussan A., Lions J.-L., Temam R. — A paraître.

[3] Brezis H. — A paraître.

[4] Brezis H., Pazy A. — A paraître.

[5] Burstein S.Z., Mirin A. — *Report NYU,* 1480-136, Courant Inst. of Math. Sc., New York Un. 1969.

[6] Campbell, Hausdorff. — Cf. M. Lutzky, *Journ. Math. Phys.*, 9, No. 7, 1968, p. 1125-1128.

[7] Chorin A.J. — *J. Comp. Phys.*, 2, 1967, p. 12-26.

[8] Chorin A.J. — *Math. Comput.*, 22, 1968, p. 745-762.

[9] Dantzig, Wolf. — *Programmation linéaire,* Dunod, Paris.

[10] Da Prato. — *J. Math. Pures Appl.*, 49, 1970, p. 289-348.

[11] Douglas Jr. J., Gunn J.E. — *Num. Math.*, 6, 1964, p. 428-453.

[12] Douglas Jr. J., Rachford H.H. — *Trans. A.M.S.*, 82, 1956, p. 421-439.

[13] Fortin M. — *Thèse de 3e cycle,* Université de Paris - Orsay, 1970.

[14] Fortin M., Peyret R., Temam R. — *Proceding of the Second Intern. Conf. on Num. Methods in Fluid Dynamics,* Berkeley, 1970, Lecture Note in Physics, Springer-Verlag, Berlin.

[15] Fortin M., Temam R. — *Première conférence intern. Méthodes num. en Dyn. des Fluides,* Novossibirsk, Août, 1969.

[16] Glowinski R. — *Thèse,* Université de Paris, 1970.

[17] Gorez J.P., Temam R., Touzot G. — *Conf. intern. de Mécanique Théor. et Appl.*, Liège, Août, 1970 et articles à paraître.

[18] Jamet, Lascaux, Raviart. — *Num. Math.*, à paraître.

[19] Kolodner I.I. — *Linear Problems,* The University of Wisconsin (Press 1963), p. 285-287.

[20] Lasdon L. — *Duality and Decomposition in mathematical programming, Report* SRC 119-c-67-52, Case Inst. of Technology, 1967.

[21] Lemarechal. — *Thèse Ing.-docteur,* Fac. Sc., Toulouse, 1969.

[22] Lieutaud. — *Thèse,* Faculté des Sciences de Paris (1968).

[23] Lions J.-L. — *Cours du C.I.M.E.*, Varenne, 1967.

[24] Lions J.-L. — *Rendic. di Mat.*, 1, 1968, p. 1-36.

[25] Lions J.-L. — *Sur le contrôle optimal de systèmes gouvernés par des équations aux dérivées partielles,* Dunod, Paris, 1968.

[26] Lions J.-L., Temam R. — *Comptes Rend. Ac. Sc. Paris,* Série A, 263, 1966, p. 563-565.

[27] Lions J.-L., Temam R. — *Comptes Rendus du Second symposium International d'Optimisation,* (Nice, 1970), Lecture Notes in Math., Springer-Verlag, Berlin.

[28] Marchuk G.I. — *Méthodes Numériques en Météorologie, Novossibirsk* (1966) (en russe) et Armand Colin, Paris, 1971 (en français) et la bibliographie de ce livre.

[29] Marchuk G.I. — *Conf. à l'Un. du Maryland,* Mai 1970.

[30] Marchuk G.I. — dans ce volume.

[31] Nedelec Ch. — *Thèse,* Faculté des Sciences de Paris, 1970.

[32] Nepomiatchy. — *Thèse ing.-doct.*, Fac. Sc. Toulouse, 1969.

[33] Pelissier M.-C. — *Comptes Rend. Ac. Sc. Paris,* Série A, 272, p. 1097-1100.

[34] Peaceman D.-W., Rachford H.H. — *Journ. S.I.A.M.*, 3, 1955, p. 28-41.

[35] Raviart P.-A. — A paraître.

[36] Saint-Pierre P. — *Thèse de 3e cycle,* Université de Paris-Orsay, 1971.

[37] Samarskii A.A. — *Soviet Math.*, 165, n° 6, 1965, p. 1601-1605, et la bibliographie de cette note.

[38] Temam R. — *Annali di Mat. Pura ed Appl.* (IV), LXXIX, 1968, p. 191-380.

[39] Temam R. — *Bull. Soc. Math. de Fr.*, 96, 1968, p. 115-152.

[40] Temam R. — *Arch. for Rat. Mech. and Anal.*, 32, n° 2, 1969, p. 135-153 et 33, No. 5, 1969, p. 377-385.

[41] Temam R. — *Arch. for. Rat. Mach. and Anal.*, 35, No. 5, 1969, p. 351-362.

[42] Temam R. — *Journ. of Comp. Syst. Sc.*, 4, No. 3, 1970, p. 250-259.

[43] Temam R. — *Comptes rendus, Ac. Sc. Paris,* Série A, 268, 1969, p. 1335-1338 et *Journ. of Fonct., Anal.*, 7, n° 1, 1971, p. 85-115.

[44] Trotter H.F. — *Proc. Am. Math. Soc.*, 10, 1959, p. 545-551.

[45] Viaud. — *Rapport I.R.I.A.*, 78 Rocquencourt, Fr.

[46] Yanenko N.-N. — *Méthode à Pas fractionnaire Novossibirsk* (1965) (en russe) et Armand Colin, Paris (1968) (traduction française).

[47] Yanenko N.-N. — dans ce volume.

Faculte des Sciences d'Orsay
Dept. de Mathématique
91. Orsay (France)

Actes, Congrès intern. Math., 1970. Tome 3, p. 321 à 329.

CONVERGENCE ESTIMATES IN DISCRETE INITIAL VALUE PROBLEMS

by Vidar THOMÉE

1. Preliminaries.

The purpose of this paper is to present a survey of recent results on the rate of convergence of finite difference schemes applied to initial-value problems for linear systems of partial differential equations. In particular, we shall consider the dependence of the rate of convergence upon the smoothness of the initial-values. Special attention will be given hyperbolic and parabolic systems. In the case of the heat equation early results were obtained by Juncosa, Young, and Wasow. The main sources of our presentation are the papers in the list of references. The reader is referred to these papers for a more complete presentation of the theory, its history and literature.

We shall begin by introducing the Banach spaces in terms of which our results will be expressed.

For $1 \leqslant p \leqslant \infty$ let L_p denote the space of measurable functions on euclidean d-space R^d with

$$\|v\|_p = \begin{cases} \left(\int_{R^d} |v(x)|^p dx\right)^{1/p} < \infty\,, & 1 \leqslant p < \infty\,, \\ \operatorname*{ess\,sup}_x |v(x)| < \infty\,, & p = \infty\,, \end{cases}$$

and let $\mathcal{C}$ denote the subspace of L_∞ consisting of uniformly continuous bounded functions. Set $W_p = L_p$ for $1 \leqslant p < \infty$ and $W_\infty = \mathcal{C}$. Further let W_p^m be the Sobolev space of order m, the space of $v \in W_p$ such that $D^\alpha v \in W_p$ for $|\alpha| \leqslant m$ $(D^\alpha = (\partial/\partial x_1)^{\alpha_1} \ldots (\partial/\partial x_d)^{\alpha_d}$, $\alpha = (\alpha_1, \ldots, \alpha_d)$, $|\alpha| = \Sigma_j \alpha_j)$ and set for $v \in W_p^m$,

$$\|v\|_{W_p^m} = \sum_{|\alpha| \leqslant m} \|D^\alpha v\|_p .$$

We shall write $\mathcal{C}^\infty = \bigcap_m W_\infty^m$, so that $v \in \mathcal{C}^\infty$ means that $D^\alpha v$ is bounded for any α.

We shall now define the Besov space $B_p^{s,q}$, $s > 0$, $1 \leqslant p$, $q \leqslant \infty$. For $v \in W_p$, $t > 0$, we introduce the modulus of continuity in W_p of order j, $j = 1, 2$,

$$\omega_{p,j}(t\,, v) = \sup_{|y| \leqslant t} \|(T_y - I)^j v\|_p\,,$$

where $T_y v(x) = v(x + y)$. Let $s = S + s_0$ where S is a non-negative integer and $0 < s_0 \leqslant 1$. Then $B_p^{s,q}$ is the subspace of W_p^S defined for $1 \leqslant q < \infty$ by

$$\|v\|_{B_p^{s,q}} = \|v\|_p + \begin{cases} \sum_{|\alpha|=S} \left[\int_0^\infty (t^{-s_0}\, \omega_{p,1}(t, D^\alpha v))^q \, \frac{dt}{t} \right]^{1/q} , & 0 < s_0 < 1, \\ \sum_{|\alpha|=S} \left[\int_0^\infty (t^{-1}\, \omega_{p,2}(t, D^\alpha v))^q \, \frac{dt}{t} \right]^{1/q} , & s_0 = 1 , \end{cases}$$

and for $q = \infty$ by

$$\|v\|_{B_p^{s,\infty}} = \|v\|_p + \begin{cases} \sum_{|\alpha|=S} \sup_{t>0} t^{-s_0}\, \omega_{p,1}(t, D^\alpha v) , & 0 < s_0 < 1 , \\ \sum_{|\alpha|=S} \sup_{t>0} t^{-1}\, \omega_{p,2}(t, D^\alpha v) , & s_0 = 1. \end{cases}$$

We shall write for short $B_p^s = B_p^{s,\infty}$. This space is then defined by a Lipschitz condition in W_p for the derivatives of order S (for $s_0 = 1$, i.e. s integer, a Zygmund condition).

The Besov spaces form a scale of spaces between the Sobolev spaces in the sense that $B_p^{s_1,q_1} \subset B_p^{s_2,q_2}$ if $s_1 > s_2$ or $s_1 = s_2$, $q_1 \leqslant q_2$, and $B_p^{s,1} \subset W_p^s \subset B_p^{s,\infty}$ if s is a natural number. Here inclusion stands for continuous embedding so that in each case a corresponding inequality between norms holds. The Besov spaces are also intermediate spaces between the Sobolev spaces in the sense of the theory of interpolation of Banach spaces. For instance, if m is a natural number and $0 < s < m$ then there is a constant C such that for any bounded linear operator A in W_p with

$$\|Av\|_p \leqslant C_j \, \|v\|_{W_p^{jm}} , \quad j = 0, 1,$$

we have

$$\|Av\|_p \leqslant C\, C_0^{1-\theta}\, C_1^\theta \, \|v\|_{B_p^s} , \quad \theta = \frac{s}{m} \cdot$$

2. A general convergence estimate.

Consider the initial-value problem

$$\text{(1)} \qquad \frac{\partial u}{\partial t} = P(x, D)u \equiv \sum_{|\alpha| \leqslant M} P_\alpha(x) D^\alpha u , \quad t \geqslant 0 ,$$

$$\text{(2)} \qquad u(x, 0) = v(x) ,$$

where P_α are $\mathcal{C}^\infty$ $N \times N$ matrices and $u = u(x, t)$ and $v = v(x)$ are N-vectors. The initial-value problem is said to be correctly posed in W_p if $P(x, D)$ generates a strongly continuous semi-group $E(t)$ for $t \geqslant 0$ in W_p. The solution $u(x, t) = E(t)v$ then satisfies

$$\|E(t)v\|_p \leqslant C \, \|v\|_p , \quad 0 \leqslant t \leqslant T .$$

We shall assume below that the following stronger condition holds, namely that for any $m \geqslant 0$, $v \in W_p^m$ implies $E(t)v \in W_p^m$ and

$$\|E(t)v\|_{W_p^m} \leqslant C \, \|v\|_{W_p^m} \, , \quad 0 \leqslant t \leqslant T .$$

In this case the initial-value problem is said to be strongly correctly posed in W_p. For systems with coefficients independent of x and for first order systems ($N = 1$) strong correctness follows from correctness.

The system is said to be parabolic in Petrovskii's sense if

$$\max Re \, \lambda_j(P(x \, , \xi)) \leqslant -c \, |\xi|^M + C \, , \quad c > 0.$$

Such a system is strongly correctly posed in W_p for all p with $1 \leqslant p \leqslant \infty$. In addition the solutions are smooth for $t > 0$; the following estimate holds, namely

(3) $$\|D^\alpha E(t)v\|_p \leqslant C t^{-|\alpha|/M} \, \|v\|_p \, , \; 0 < t \leqslant T.$$

Consider a finite difference operator which for simplicity we shall take to be explicit,

$$E_k v(x) = \sum_\gamma a_\gamma(x \, , h) \, v(x - \gamma h).$$

Here $h > 0$ is a small parameter tied to k by the relation $k/h^M = \lambda =$ constant where M is the order of the system (1). The summation is over a finite set of $\gamma = (\gamma_1, \ldots, \gamma_d)$, γ_j integer, and the a_γ are $\mathcal{C}^\infty$ $N \times N$ matrices in x and h.

The operator E_k shall be assumed to be consistent with (1) and accurate of order μ, so that for smooth solutions $u(x \, , t)$ of (1),

$$u(x \, , t + k) = E_k u(x \, , t) + k \, 0(h^\mu) \quad , \quad \text{as} \quad h \to 0.$$

This condition can be expressed more precisely as follows.

Lemma. – Assume that the initial-value problem (1), (2) is strongly correctly posed in W_p and that E_k is accurate of order μ. Then

$$\|(E_k - E(k))v\|_p \leqslant C \, kh^\mu \, \|v\|_{W_p^{M+\mu}} .$$

The operator E_k is said to be stable in W_p if for any $T > 0$,

$$\|E_k^n v\|_p \leqslant C \, \|v\|_p \quad , \quad nk \leqslant T.$$

It is well-known that for consistent operators E_k stability is the necessary and sufficient condition for the convergence of the solution of the discrete problem to the solution of the continuous problem in the sense that for any $v \in W_p$,

$$\lim_{h \to 0} \; \|E_k^n v - E(nk)v\|_p = 0 \, ,$$

uniformly in $0 \leqslant nk \leqslant T$.

We can now obtain the following more precise result (Peetre and Thomée [3]).

THEOREM 1. — *Assume that the initial-value problem (1), (2) is strongly correctly posed in W_p and that E_k is stable in W_p and accurate of order μ. Then for $nk \leqslant T$,*

$$\|E_k^n v - E(nk)v\|_p \leqslant \begin{cases} Ch^{\mu} \|v\|_{W_p^{M+\mu}}, \\ \\ C_s h^{s\mu/(M+\mu)} \|v\|_{B_p^s}, \quad 0 < s < M + \mu. \end{cases}$$

The proof of the first inequality follows from trivial estimates using the above lemma, the stability, and the strong correctness in

$$(4) \qquad E_k^n v - E(nk)v = \sum_{j=0}^{n-1} E_k^{n-1-j}(E_k - E(k))E(jk)v,$$

and the second inequality then follows by interpolation between the first and the following trivial estimate, namely

$$\|E_k^n v - E(nk)v\|_p \leqslant C\|v\|_p \quad , \quad nk \leqslant T.$$

3. The rate of convergence in the parabolic case.

In the case of a system (1) which is parabolic in Petrovskii's sense it is possible in estimating the different terms in (4) to use the smoothing inequality (3) for $E(jk)v$ to obtain the following more precise estimate, cf. [3].

THEOREM 2. — *Assume that (1) is parabolic in Petrovskii's sense and that E_k is stable in W_p and accurate of order μ. Then for $nk \leqslant T$,*

$$\|E_k^n v - E(nk)v\|_p \leqslant \begin{cases} Ch^{\mu}\left(\log \frac{1}{h}\right)^{1-1/q} \|v\|_{B_p^{\mu,q}} \quad , \quad 1 \leqslant q \leqslant \infty, \\ \\ C_s h^s \|v\|_{B_p^s} \quad , \quad 0 < s < \mu. \end{cases}$$

Investigations by Hedstrom, Löfström, and Widlund have shown that in particular cases, the factor $\log \dfrac{1}{h}$ can be removed from the first of these estimates for $q = \infty$. We shall describe the most far-reaching of these results. Let

$$E_h(x, \xi, h) = \sum_{\gamma} a_{\gamma}(x, h)\exp(i\langle \gamma, h\xi\rangle),$$

be the symbol of the difference operator E_k. This operator is said to be parabolic in the sense of F. John if for the spectral radius,

$$\rho(E_k(x, h^{-1}\xi, 0)) \leqslant 1 - c|\xi|^M \quad , \quad |\xi_j| \leqslant \pi \quad , \quad c > 0.$$

Parabolicity implies stability in W_p for $1 \leqslant p \leqslant \infty$ and we have the following result by Widlund [5].

THEOREM 3. – *Assume that (1) and E_k are parabolic in the sense of Petrovskii and John, respectively, and that E_k is accurate of order μ. Then for $nk \leqslant T$,*

$$\|E_k^n v - E(nk)v\|_p \leqslant Ch^\mu \|v\|_{B_p^\mu}.$$

The proof in [5] depends on estimates for a discrete fundamental solution.

The property of $E(t)v$ of being smooth for $t > 0$ and of satisfying the inequality (3) has the following analogue for parabolic difference operators. Let

$$\partial_h^\alpha v = \partial_{h,1}^{\alpha_1} \cdots \partial_{h,d}^{\alpha_d} v \ ,$$

where $\partial_{h,j} v(x) = (ih)^{-1}(v(x + he_j) - v(x))$ denotes forward difference quotients. Then

$$\|\partial_h^\alpha E_k^n v\|_p \leqslant Ct^{-|\alpha|/M} \|v\|_p \quad , \quad 0 < nk = t \leqslant T.$$

This result again depends on estimates for the discrete fundamental solution. Such estimates can also be used to prove convergence estimates for difference quotients. We have the following result, cf. [5].

THEOREM 4. – *Assume that (1) and E_k are parabolic in the sense of Petrovskii and John, respectively, and that E_k is accurate of order μ. Let*

$$Q_h v(x) = \sum_{|\alpha|=q,\gamma} q_{\alpha,\gamma} \partial_h^\alpha v(x + \gamma h)$$

be a finite difference operator which is consistent with the differential operator Q of order q and also accurate of order μ. Then for $0 < \tau \leqslant nk \leqslant T$,

$$\|Q_h E_k^n v - Q\,E(nk)v\|_p \leqslant Ch^s \|v\|_{B_p^s} \quad , \quad 0 < s \leqslant \mu.$$

In view of the fact that unsmooth initial-data give rise to lower rates of convergence it is natural to ask if the convergence can be made faster by first smoothing the initial-data. This is indeed the case for parabolic systems and we shall describe a result to this effect (Kreiss, Thomée, and Widlund [2]).

We shall consider operators of the form

$$M_h v = \Phi_h * v \ , \ \Phi_h(x) = h^{-d}\Phi(h^{-1}x),$$

where Φ is a function such that its Fourier transform satisfies

$$\hat{\Phi}(\xi) = 1 + \sum_{|\alpha|=\mu} \xi^\alpha b_\alpha^{(0)}(\xi),$$

$$\hat{\Phi}(\xi) = \sum_{|\alpha|=\mu} \left(\sin \frac{1}{2}\xi\right)^\alpha b_\alpha^{(1)}(\xi),$$

Here $b_\alpha^{(j)}$, $j = 0, 1$ are such that for some $\delta > 0$, $b_\alpha^{(0)}$ and $b_\alpha^{(1)}$ coincide with multipliers on $\mathcal{F}W_p$ for $|\xi| < 2\,\delta$ and $|\xi| > \delta$, respectively. Such an operator is said to be a smoothing operator of order μ in W_p. Since the multipliers on $\mathcal{F}L_2$ are simply the functions in L_∞, the above condition can be seen to be satisfied for $p = 2$ if

$$\hat{\Phi}(\xi) = 1 + 0(|\xi|^{\mu}) , \quad \text{as} \quad \xi \to 0 ,$$

and for any multi-index $\beta \neq 0$,

$$\hat{\Phi}(\xi) = 0(|\xi - 2\beta\pi|^{\mu}),, \text{ as } \quad \xi \to 2\beta\pi,$$

uniformly in β.

Special examples of smoothing operators of orders 1 and 2, respectively, in the case $d = 1$ are

$$M_h^{(1)} v(x) = h^{-1} \int_{-\frac{1}{2}h}^{\frac{1}{2}h} v(x-y)\,dy ,$$

$$M_h^{(2)} v(x) = h^{-1} \int_{-h}^{h} (1 - |h^{-1}y|)\, v(x-y)\,dy ,$$

and for general μ, a smoothing operator of order μ can easily be constructed in the form

$$M_h^{(\mu)} v(x) = h^{-1} \int \chi_{\mu}\, (h^{-1}y)\, v(x-y)\,dy ,$$

where χ_μ is a function which is piece wise a polynomial of degree $\mu - 1$ and which vanishes outside $\left(-\mu + \frac{1}{2}, \; \mu - \frac{1}{2}\right)$ for μ odd and $(-\mu + 1, \; \mu - 1)$ for μ even. For $p = 2$, the operator M_h corresponding to

$$\hat{\Phi}(\xi) = \begin{cases} 1, & |\xi| \leqslant \delta , \\ 0, & |\xi| > \delta , \end{cases}$$

where $0 < \delta \leqslant \pi$, is a smoothing operator of arbitrarily high order. In this case

$$\Phi(x) = \frac{\sin \delta x}{\pi x} \, .$$

Smoothing operators in higher dimensions can be obtained by taking products of one-dimensional operators

$$M_h v = \prod_{j=1}^{d} M_{h,j} v ,$$

where $M_{h,j}$ is a smoothing operator with respect to x_j.

The result on the rate of convergence is then the following.

THEOREM 5. – *Assume that (1) and E_k are parabolic in the sense of Petrovskii and John, respectively, that E_k is accurate of order μ, and that M_h is a smoothing operator of order μ. Then there is a constant $C = C_{p,T}$ such that for $0 < t = nk \leqslant T$,*

$$\|E_k^n M_h v - E(t) v\|_p \leqslant C h^{\mu} t^{-\mu/M} \|v\|_p .$$

4. The rate of convergence in hyperbolic cases.

Consider a first order system

$$\frac{\partial u}{\partial t} = \sum_{j=1}^{d} A_j(x) \frac{\partial u}{\partial x_j} + B(x)u.$$

Under the assymption that the corresponding initial value problem is correctly posed in W_p, the convergence estimates in § 2 give that

$$\|E_k^n v - E(nk)v\|_p \leqslant \begin{cases} Ch^{\mu} \|v\|_{W_p^{\mu+1}} , \\ Ch^{s\mu/(\mu+1)} \|v\|_{B_p^s} , \quad 0 < s < \mu + 1 , \end{cases}$$

for E_k accurate of order μ and stable in W_p.

It is well-known that sufficient conditions for correctness in L_2 are that

(a) the A_j are hermitian,

or (b) the A_j are constant and $\Sigma_j A_j \xi_j$ is uniformly equivalent to a real diagonal matrix when $\xi \in R^d$.

However, for $p \neq 2$ correctness in W_p is exceptional ; if A_j are constant and hermitian, the problem is then correctly posed if and only if the A_j commute, or equivalently, if they can be simultaneously diagonalized by a unitary transformation. In this case, after a corresponding change of dependent variables, the equations are only coupled in the lower order terms so that the system consists essentially of independent scalar equations.

We shall therefore consider separately the case of the initial-value problem for the scalar one-dimensional equation

$$(5) \qquad \frac{\partial u}{\partial t} = \rho \frac{\partial u}{\partial x} \quad , \quad \rho \text{ real constant.}$$

The solution operator is then $(E(t)\, v)\ (x) = v(x + \rho t) = T_{\rho t} v(x)$. We shall analyze corresponding difference schemes with constant coefficients,

$$E_k v(x) = \sum_{|j| \leqslant J} a_j v(x - jh) \quad , \quad k/h = \lambda .$$

The results will be expressed in terms of its characteristic polynomial,

$$a(\xi) = \Sigma_j a_j e^{ij\xi} .$$

It is well-known that E_k is stable in L_2 if and only if

$$(6) \qquad |a(\xi)| \leqslant 1 \quad , \quad \xi \text{ real} ,$$

and we shall consider only the case when this is satisfied. The results in § 2 then give error estimates in L_2 and in the case that E_k is stable in W_p for other p we get analogous convergence results in W_p. However, many operators E_k are stable only in L_2 and our purpose here is to give precise error estimates also in W_p.

We shall assume for simplicity that rather than (6),

$$|a(\xi)| < 1 \quad \text{for} \quad 0 < |\xi| \leqslant \pi. \tag{7}$$

For small ξ we can then write

$$a(\xi) = \exp(-i\lambda\,\rho\,\xi + \psi(\xi)),$$

where as $\xi \to 0$,

$$\psi(\xi) = \beta\xi^{\nu}(1 + o(1)) \ , \ \beta \neq 0.$$

Here $\nu = \mu + 1$ where μ is the order of accuracy of E_k. By (7) there is a smallest (even) number σ, the order of dissipation of E_k, such that

$$Re\,\psi(\xi) \leqslant -\gamma\,\xi^{\sigma} \ , \ |\xi| \leqslant \pi \ , \ \gamma > 0.$$

Clearly $\nu \leqslant \sigma$ and if $\nu < \sigma$, β is purely imaginary. It is known that E_k is stable in W_p, $p \neq 2$, if and only if $\nu = \sigma$. More precisely, under the above assumptions on the operator E_k there are positive c and C such that

$$cn^{\left|\frac{1}{2}-\frac{1}{p}\right|\left(1-\frac{\nu}{\sigma}\right)} \leqslant \|E_k^n\|_{W_p} \leqslant Cn^{\left|\frac{1}{2}-\frac{1}{p}\right|\left(1-\frac{\nu}{\sigma}\right)}, \tag{8}$$

where $\|\cdot\|_{W_p}$ denotes the operator norm in W_p (cf. [1]).

We now state the result on the rate of convergence [1].

Theorem 6. – *Under the above assymptions on the operator E_k, then for $s \geqslant 0$, $s \neq \nu$ and ν* $\left|\dfrac{1}{2} - \dfrac{1}{p}\right|$ and $nk \leqslant T$,

$$\|E_k^n v - E(nk)v\|_p \leqslant C_s h^{g(s)}\, \|v\|_{B_p^s}, \tag{9}$$

where $g(s) = \min\left\{\mu, s\left(1 - \dfrac{1}{\nu}\right), s\left(1 - \dfrac{1}{\sigma}\right) - \left|\dfrac{1}{2} - \dfrac{1}{p}\right| \left(1 - \dfrac{\nu}{\sigma}\right)\right\}$.

It can also be proved that this result is best possible in the sense that the function $g(s)$ above is the largest for which an estimate of the form (9) holds for all $v \in B_p^s$. In the stable cases, i.e. when $\nu = \sigma$ or $p = 2$, the order of convergence is $s(1 - 1/\nu) = s\mu/(\mu + 1)$ when $0 < s < \nu$ in agreement with theorem 1. In the opposite case the error is larger for $s < \nu \left|\dfrac{1}{2} - \dfrac{1}{p}\right|$. For small s, $g(s)$ is then negative and for $s = 0$ we recognize the exponent in (8).

It is interesting to note that if the irregularity of the initial-function stems from the behavior at isolated points then the result above may be improved so that for $p = \infty$ we obtain the same result as if E_k were stable in $\mathcal{C}$. We shall formulate this result in terms of the Banach space $\check{B}_M^s$ of functions v with support in $[-M, 0]$ such that

$$\|v\|_{\check{B}_M^s} = \|v\|_{B_\infty^{s+1/2}(R^-)} + \sup_x |x^{-s} v(x)|$$

is finite. Using the L_2 convergence result, Sobolev's inequality, and the fact that $\check{B}^s_M$ is continuously embedded in $B_2^{s+\frac{1}{2}}$ one can prove the following result [4].

Theorem 7. – *Consider a L_2 stable operator E_k for the equation (5). Then for given positive $s \neq \mu + 1$ and M, and for $nk \leqslant T$,*

$$\|E_k^n v - E(nk)v\|_\infty \leqslant Ch^{\min(s, s\mu/(\mu+1))} \|v\|_{\check{B}^s_M} .$$

This result also holds when ρ depends upon x.

REFERENCES

[1] Brenner Ph. and Thomée V. — Stability and convergence rates in L_p for certain differences schemes. *Math. Scand.*, 27, 1970, p. 5-23.

[2] Kreiss H.O., Thomée V. and Widlund O.B. — Smoothing of initial data and rates of convergence for parabolic difference equations. *Comm. Pure Appl. Math.*, 23, 1970, p. 241-259.

[3] Peetre J. and Thomée V. — On rate of convergence for discrete initial-value problems. *Math. Scand.*, 21, 1967, p. 159-176.

[4] Thomée V. — On the rate of convergence of difference schemes for hyperbolic equations. *Symposium on the Numerical solution of partial differential equations*, Maryland, 1970.

[5] Widlund O.B. — On the rate of convergence for parabolic difference schemes, II, *Comm. Pure Appl. Math.*, 23, 1970, p. 79-96.

Chalmers Institute of Technology and the
University of Göteborg
Dept. of Mathematics,
Fack S-40 220 Göteborg 5 (Suède)

F

HISTOIRE ET ENSEIGNEMENT

(Tome 3 : pages 331 à 367)

Actes, Congrès intern. Math., 1970. Tome 3, p. 333 à 334

F1 - HISTOIRE DES MATHÉMATIQUES

THE FOUNDATION OF ALGEBRAIC GEOMETRY FROM SEVERI TO ANDRÉ WEIL

by B.L. VAN DER WAERDEN

Summary.

(The full text of the lecture will be published in the Archive for History of Exact Sciences).

Algebraic geometry was created by Max Noether and the Italian school of geometers. The theory was admirable, but its foundations were shaky.

A great step towards a rigid foundation was taken by Severi in 1912. Severi gave a rigorous definition of specialization multiplicities and a correct enunciation of the "Principle of Conservation of Number". However, the algebraic tools necessary to prove his assertions were not yet available. The tools are : Theory of ideals, Elimination Theory, and Theory of Fields. They were developed in the schools of Dedekind, of Kronecker, and of Hilbert between 1882 and 1923.

The next step was my paper "Nullstellentheorie der Polynomideale", published in 1926, which was based upon the ideas of Emmy Noether. A basic notion of this paper was the notion "generic point of a variety".

In my next paper "Multiplizitätsbegriff" (1927) I introduced the notion "Specialization" and proved the correctness of Severi's assertions concerning the Principle of Conservation of Number.

Intersection multiplicities of a curve C and a hypersurface F can be defined in 3 ways : by using algebraic functions of one variable, or Ideal Theory, or Specializations of a generic form F^*. All three definitions are equivalent, and Bezout's Theorem can easiby be extended to this case.

In the more general case of an intersection of two varieties A and B of complementary dimensions in projective space, one can either use (in the classical case) the topological definition of Lefschetz, or bring A into a generic position with respect to B by a projective transformation T and then specialize T to identity. In the classical case the two definitions are equivalent.

If A and B are subvarieties of a variety U, of complementary dimensions, intersecting in a simple point of U, the multiplicity of such a point of intersection can either be defined (in the classical case) by the method of Lefschetz, or by a method given by Severi in 1933, which works for an arbitrary ground field, as I showed in 1938.

In 1946, André Weil's book appeared. Before Weil, the uniqueness of the specialization multiplicities had been proved only under the assumption that the specialized problem has a finite number of solutions. Weil succeeded in proving the uniqueness of any isolated solution y. This enabled him to define intersection multiplicities for all proper components C of an intersection $A \cdot B$, i.e. for all those components that have the normal dimension $r + s - n$.

University of Zürich
Bionstrasse 18,
8 006. Zürich (Suisse)

Actes, Congrès intern. Math., 1970. Tome 3, p. 335 à 345.

F2 - ENSEIGNEMENT DES MATHÉMATIQUES

MATHEMATICAL INSIGHT AND MATHEMATICAL CURRICULA

by H. Brian GRIFFITHS

The Organizing Committee have asked me to speak[1] on the general topic of Modern Mathematical Curricula, and so I suppose that I should include a survey for non-specialists as is common in lectures to the International Congress. It would be impossible to give a detailed account of all the various new curricula now being pursued in different parts of the world today, but I shall discuss certain philosophical attitudes behind them. Then I shall take up the question of whether or not they can cultivate mathematical insight. Finally, I shall attempt to be constructive by discussing an example within the discipline of Mathematical Education. Always, I shall be commenting from my limited experience as an academic mathematician in a English University, and from my knowledge as an external examiner in other Universities and institutions.

1. Philosophies

Nowadays, most mathematicians know that new Mathematical curricula are being devised, irr different countries and cultures. They are designed for different age-levels of student from Kindergarten through High School and beyond, to courses for parents and for in-service teachers. Few mathematicians, however, seem to know much in detail about such curricula, and this leads some to indulge in ignorant criticism. Such of it as does not merely come from hostility to change, often springs from forgetting that many of the curricula aim to teach mathematics to a wider band of the population than hitherto, and with the hope of making the subject attractive even to those highly intelligent students who formerly were repelled by the traditional approach. This is forgotten in the rather amateur critique [3] for example, where it is assumed that mathematical education is something to do with being clever and speedy ; in contrast to much of the professional hard work that has gone into the construction of many curricula. Many of the differences between these curricula reflect the relative weight given to mathematical, as opposed to teaching, arguments. There is no dogmatic formula for a unique "best mix", and we should expect intelligent adaptation by trial and error to evolve improvements.

[1] The author was not in fact able to deliver the lecture at the Congress, owing to the death of his youngest son Joe on August 23rd, 1970, aged six years. This paper is dedicated to his memory.

Also, few mathematicians realise how these new curricula relate to the changes that have occurred within University courses in the last twenty years. These University changes have not been so explicit, or carefully planned, as those for more elementary courses –perhaps because of the greater freedom from central direction that University teachers enjoy. But regardless of the level, the impetus for change has come from two directions ; and the relative emphasis in any planned course has depended on the mathematical level of the course.

First then, there is the tendency to change, which comes from the growth, internal logic, and reorganisation of mathematics itself ; this tends to be teacher-centred, because the need for change is seen by the experienced worker in mathematics, and he explains the subject accordingly. I shall refer to this tendency as the "Renewal" Tendency.

Second, there is the tendency to change, which springs from the spread of humanitarian, liberal, anti-authoritarian attitudes in schools and society at large ; this tends to be pupil-centred, in that mathematics (as any other subject) is developed with constant reference to, and respect for, the pupil's reactions. I shall refer to this tendency as the "Open" Tendency, choosing a neutral adjective to avoid political overtones.

This second tendency is always present to some extent with the first in the professional development of mathematics, where youthful minds are constantly creating new mathematics, and scepticism rather than authority is a guiding force. Within the education system proper however, the Open Tendency is associated above all with the teaching of very young children, through the names of Froebel, Montessori, Gattegno, etc. By contrast the first "Renewal" Tendency manifests itself most importantly at University level, especially through the work of Bourbaki.

2. The influence of Bourbaki.

In his youth, Bourbaki found that mathematics had grown in such a way that existing books were no longer suitable for the courses in mathematics that he wished to give ; so he developed his own course as recorded in his book "Elements de Mathématique", a record still incomplete, still under revision, still growing because it reflects the same trends in mathematics. To begin writing the book, Bourbaki had to devise mathematical goals which his readers were expected to reach, and then he had to devise appropriate mathematical paths to these goals. From existing mathematical treatments he had to devise suitable language, especially by selecting suitable definitions, so that he could then deduce all the mathematics he wanted, exposing it for his readers in a logical and beautiful way. To do all this involved immense toil and presumably some false starts, but these difficulties can hardly be inferred from the text itself, where Bourbaki tells his students the mathematical story, selecting for them the definitions and the main logical development ; the students do as he says. They may hear about his *human* story through gossip, or by noticing changes in successive editions, but the only official hint about the construction of the work is tucked away in the "Notes Historiques" at the ends of Chapters.

Bourbaki's approach has had immense influence among younger University teachers, who have (often unconsciously) modelled their style on his, giving lectures that unfold before their listeners with inevitable logic, and often with a clarity that is fine for a certain kind of listener. Unfortunately, most young teachers do not read historical notes on mathematics, and are pathetically ignorant of the history of the growth and motivation – even of their speciality. Consequently their exposition does not convey the mathematical insight that is should, and which Bourbaki has but fails to spell out. The result is that these University teachers fail to teach : they clarify their own minds in the process and may even get some students to pass examinations in the material, but their educational success is dubious because insight is failing.

3. Insight.

But what do we mean by "insight"? I obviously cannot define it formally, and I must convey the notion ostensively, by pointing to examples. I also believe it to be sufficiently important to be worth discussing, and I begin by saying that I shall be distinguishing between "insight" and "intuition". Let us then look at examples of insight (or lack of it).

(1) If a student takes a course in transfinite arithmetic, and cannot even begin to see how the theory might help him to explain why $3 + 5 = 8$, then his course has given him no real insight.

(2) A student who (as is common) is taught to grind out derivatives of functions, and is content to write down the derivative of arc sin $(1 + x^2)$ without further comment, has no insight into the matter.

(3) Herstein [5] quotes the case of a Ph. D. student who was well versed in the spectral theorem for normal operators and Hilbert Space, but the student could not simplify it to diagonalise a Hermitian form. That student lacked insight.

(4) When Kelvin said that a mathematician was "anyone to whom it is as obvious that $\int_{-\infty}^{\infty} e^{-x^2}\,dx = \sqrt{\pi}$, as that $2 + 2$ makes 4" he was talking about insight (although quite what he expected, I do not know).

(5) A course on homology groups which concludes by merely stating that "cohomology groups are just the same except that you reverse all arrows" is unlikely to yield insight ; a course on groups, which includes topological proofs of certain theorems, might well yield more insight than if a purely algebraic approach were given.

(6) (To choose a famous example). If we recall the famous occasion when Poincaré was tying his shoelace and there flashed into his mind the thought that the conformal transformations of the upper half plane were just the non-euclidean isometries, then we have an undoubted case of insight !

One could give many other examples. However, a common feature is that "insight" into a mathematical theory seems to be related to the realisation that the theory has a model in physics, geometry, or some more familiar or accessible part of mathematics. A computer cannot show insight in this sense, when it checks the steps of a formal proof for correctness, and this merely manipulative checking is unfortunately all that many students seem to get from preparing for traditional-type examinations, – regardless of whether or not the subject-matter is "modern". In case (2) above, we have the difficulty of teaching the algebraic manipulations of the calculus, without forgetting the motivating model of functions and their graphs (drawn on paper). In case (3) we see the typical research students failing of not building up a repertoire of examples against which to check the general theory (and he was ignorant of the way in which the general theory sprang from those very examples). In (4), I presume that either Kelvin was merely trying to be funny, or else he had some physical model in mind, with whose analysis he was thoroughly familiar. The insights in (5) are related to situations of the following kind, where *intuition* also plays a part. Suppose a person finds a book on analytic geometry which is purely algebraic in treatment, with no sketches. He will probably find the text easier to follow if he is given the insight that the algebra models geometry in the usual Cartesian way. The insight then makes it legitimate to say, for example : –

"It is intuitively obvious that two equations $f(x,y) = 0 = g(x,y)$, of the second degree, have no more than four common roots", but *only when the model is known to be a good one*[1]. Thus, in beginning courses of analysis, the definition of a continuous function is chosen to model certain properties of curves drawn on paper, and we often hear declarations that

"The continuity of the sum of two continuous functions is intuitively obvious".

On the contrary, such non-intuitive basic theorems of continuity give confidence that Weierstrass was showing the ability to model a subjective feeling (about graphs) in an objective language, – here in the language of mathematics. Also, it will be maintained that some algebraists, say, can work wholly within algebra and show great "insight" without reference to a geometrical model. Again I would prefer a different term, say "internal skill". Thus, the passage from the Weierstrass definition of pointwise continuity, to the one in general topology about inverse images of open sets, needed internal skill ; but we cannot be too precise because it could also be argued that the Weierstrass theory had become a familiar experience, from which the general topology was abstracted by the use of "poetic skill", a skill analogous to that of a poet changing experience into language.

It is obviously not worth arguing too closely about these distinctions. We surely agree, however, that the activities denoted by the terms are immensely important among the skills required of a mathematician. The Bourbaki account of mathematics does not mention these skills except in the Notes Historiques, with the educational result that students lose insight because they do not understand how the formal theory was (a) abstracted from some model *M* by someone's poetic skill, and (b) developed by internal skill, most often with the help of insight generated by knowledge that *M* exists. I am not blaming Bourbaki for this, since his purpose is to expound mathematics in an objective way ; but I am pointing out serious educational consequences of a slavish copying of Bourbaki as a teaching technique.

Consider again the important question of mathematical definitions. Bourbaki simply writes them out, and young enthusiasts hand them down to their classes.

[1] An algebraist with internal skill (see below) might say that it *is* intuitively obvious, meaning that he can see an easy proof. This uses "intuitive" in a different sense (see page 339).

Consequently, few students ever formulate a definition themselves, so that their "poetic skill" is never deveoloped in a creative way. As it is, the students are using other people's formulations, without real understanding, and they frequently write nonsense as a result. Thus, the definition handed to them may damage their primitive thought, instead of improving it as intended !

This lack of practice in "poetic skill" is influenced by the traditional imposition of timed examinations. For these, a student needs skill in giving answers to questions formulated and already answered by others. Such skill is certainly necessary up to a point ; and in real life, especially in industry or in a classroom, it is true that one is often expected to answer the questions of others, but then only when the questioner does *not* know the answer ! Even if the question is clear, discussion may be necessary as to an acceptable form of answer ; for example consider what has to be done to answer the question "Does $\sqrt{2}$ exist ?". Such questioning is hardly avoidable whenever the Open Tendency is strong in a curriculum, and one then sees less stress on timed examinations, with more stress on "looser" forms of assessment.

4. Intuition.

Now, it is often easy to ask important *questions* in a naïve language, whereas a thoroughgoing *answer* may require within mathematics the creation of a strict formal language. This is the root of the constant arguments between Curriculum designers about rigour versus intuition. To give an example, I recently saw, in a child's exercise book, a sequence of equations "$\frac{1}{2}$ of $2 = 1$", "$\frac{1}{3}$ of $3 = 1$", and so on, but then it caught his imagination and he worked out several more than required, jumping exponentially to "$\frac{1}{90{,}000}$ of $90{,}000 = 1$". Clearly, he realised the equation $y^{-1} \times y = 1$, but he was far too young to have that language imposed upon him : it might well have destroyed his insight, and at that stage he was best left to his intuitive beliefs. This seems to be what mathematicians mean when they say that things are "intuitively obvious", – that they are convinced but lack the formal language of a proof (and their internal skill gives them confidence that a formal proof can eventually be given).

Again, consider[1] the teacher who asks a class of children "Who can tell me what parallel lines are ?" No reply. "Who knows what parallel lines are ?". Everybody knows. A rigorist might say that none of the children really did know, because they could not convey their knowledge in an objective language ; but an "intuitive" would argue that they could be tested for it behavioristically, as one often does when reading applied mathematicians. This is the basis of curricula like the Nuffield Mathematics Project, with its emphasis on calculations associated with practical measurements ; and in many technical institutions in Britain, "mathematics" now means the derivation and application of algorithms, often quite advanced but expressed in non-repellent "commonsense" terms. See also, for example, Kapur [8].

[1] Due to my colleague, W.M. Brookes.

Certainly, such treatments have done much to make mathematics a popular subject.

It might nevertheless be maintained (because of the Renewal Tendency) that students ought to be trained to appreciate the power that comes from being able to make general statements like "For all x, ...", but then they need also to learn the notion of deductive proof (as distinct from the "natural" proof that arises in algorithmic manipulation). That is why I implied earlier that "definitions" were of vital importance. Suppose then that is was thought desirable that the schoolchildren I mentioned should have a definition of parallel lines. The Open Tendency would suggest that it might be better to get the children to formulate a definition themselves, cultivating their "poetic skill", than perhaps to repel them by imposing a definition. Such an imposed definition is likely to be given by a teacher who had "the" definition imposed on him as a student ; and if he himself got no insight at that stage of his own training, his pupils may be doubly deprived. Their reaction is likely to be more degenerate parroting. In any case few humans appreciate without explanation the point of proving anything at all. It should therefore not be taken for granted that pupils will immediately understand why definitions are made, yet this groundwork is taken for granted in a Bourbaki type of exposition, and is perhaps a strong reason for poor understanding by students. Even when a teacher attends to such points, there is *still* a grave danger that preoccupation with formal language and proof will drive insight out of the window ; for the motivating models are forgotten in the effort of acquiring the language.

Sometimes it is possible to use a model whose description is isomorphic to the formal one, but where the language is more natural. For example, the student of transfinite arithmetic in (1) of Section 2 might do better to use a calibrated number-line to explain why $3 + 5 = 8$, rather than talk of bijections (see for example [12] p. 19), but it needs a good insight on his part to give a non-confusing explanation. See also Section 7. For other interesting aspects of this type of modelling, see Steiner-Kaufman [9] and Crowell [1].

1. An unnecessary clash between the Renewal Tendency and the Open Tendency.

An important example of the difficulty with formal treatments arises in a new kind of curriculum, called the B.Ed. course, at present developing in Britain and which may well develop elsewhere. Let me take a moment to explain about it. Teachers of younger children in England usually train, not at Universities, but at "Colleges of Education". There, they take courses in Education and other subjects, and they are sent into schools at certain times for teaching practice. A student in such a College will often take extra courses in some speciality like Mathematics or Art, and after a three-year course he gets a Certificate which allows him to teach in schools but at a lower salary than a University graduate. About three years ago, for various reasons, the Government arranged that a good College student could be selected to stay for a fourth year of supervised study, planned in conjunction with the nearest University ; and if he is successful in the course, he is awarded a University degree, entitled Bachelor of Education (abbreviated to "B.Ed"). He is then presumed to be expert in *both* Education and his main technical subject, so that he can eventually teach at senior high school level, or become an Educational administrator. In particular it is hoped by this means to increase the supply of mathematics teachers, because the Universities are not supplying enough to meet the demand. The resulting attempts to plan curricula for the mathematics B.Ed. candidates form an interesting case-study in the discipline of Mathematical Education, (or rather in the non-application of some basic principles). In explaining it, I warn you that I must oversimplify for brevity.

The worst feature of the mathematical education of the B.Ed. candidates has been that they have been *ignorantly* subjected to both the Tendencies, Renewal and Open, I mentioned in Section 1, with disastrous results. In their three years of training for the Teacher's Certificate, the Open Tendency has been uppermost ; perhaps because the Colleges also train teachers for Primary Schools. The students have been taught in small classes with plenty of time for discussion because timed examinations, though used, have not had the supreme importance that a conventional University treatment would have given. Thus continuous assessment and project work have been usual, with an intuitive approach to practical or appealing problems without a great emphasis on formal language or deductive proof. The notions of logic and proof have been implicit, arising unobtrusively through the use of statistical techniques, algorithmic methods, and intuitive calculus. This way of doing things has gradually evolved, because the students are selected from the slightly less academic ones in High Schools ; if they had been good at formal examinations they would have automatically gone to a University rather than a College of Education. Such a student, then, has generated much of the actual work that he has submitted, and if he is selected for the fourth, B. Ed. year, this is because he promises to be a good teacher of mathematics. In the fourth year, however, he usually gets a course planned by the local University mathematics department, not by his College teachers. Thus he is handed out a formal treatment of some branches of mathematics, often very up to date because of the effect of the Renewal Tendency in the local University, and tested only by timed examinations. But the student, almost by definition, is bad at timed examinations ; and he has spent the previous three years doing a very different kind of mathematics which is usually left quite unrelated to the rigorous material that now is forced upon him. Consequently he loses interest : "modern" does not always mean "refreshing" ! Worse still, he loses confidence and is now in now in danger of promising to be a *poor* teacher of mathematics as a result of his extra year of training !

On the other hand, just as he is taught mathematics in a watertight compartment, he must spend half his time in another compartment called "Education" where he scores well, being basically very intelligent. The average of his low mathematics mark and his high education mark gives him his final, quite satisfactory, "career" mark ; and that mark will get him a teaching job with special pay allowances and seniority as an expert mathematician ! Mathematicians cannot quickly change the "Educational" compartment of these courses, but they can improve matters a lot by being more sophisticated about the Renewal Tendency. I should add that already corrective action is beginning, because the University mathematicians are being forced, by the bureaucratic processes for planning the B. Ed. degree, to consult with the mathematicians in Colleges. Some mutual sympathy and understanding is beginning to arise, and the academic mathematicians are losing arrogance when challenged by such questions as "How will the course you have laid down (say in Dynamics) help a teacher to explain to 14-year olds how a hawk hovers", or "How does your integration theory help with an explanation of area to 10-year olds ?" These very questions are also not suitable for testing by timed examinations nor for authoritarian lecturing techniques, so that the Open Tendency can be presented to the Universities as something for improving their own educational methods. In return, the College teachers (who have traditionally not concentrated on mathematical research) can be introduced to the Renewal Tendency. Also, moves are beginning, to integrate the "Education" requirement with the mathematics, taking into account such matters as History and Philosophy of mathematics, and the growth of concept-formation. The leaders in these moves see the B.Ed. course as a way of producing a better mathematics teacher than the conventional mathematics graduate, whose education as a potential teacher has been unsatisfactory for many years. (See, for example [11]).

6. Mathematical Education

The two questions I mentioned in the previous paragraph are typical of one aspect of what I called earlier the discipline of Mathematical Education. They have the characteristic form of asking for explanations *suitable for specified levels of understanding or experience*, by contrast with conventional mathematics in which explanation is given solely in a language suitable for mature professionals. In Mathematical Education, then, we have to take account of sociological factors(1) outside mathematics, but of course we need the mathematics first in order to process it appropriately. For this reason, insight is essential ; one cannot modify an accurate, "official" mathematical treatment, into an intelligible account with appropriate gaps and plausible jumps, if one has only a parrot's knowledge of the "official" version. This is why it is useless to teach mathematics to future mathematics teachers especially, without paying great attention to insight so that they may be able to prepare good "watered down" treatments of "official" mathematics. Such insight is likely to be lacking so long as the notion of a "curriculum" is retained, in the sense of nothing more than a planned syllabus or list of topics to be taught by the teacher and learned by the pupil. Consequently a striking change to be noticed in current work is the departure from the "global" standpoint of earlier schemes, and a move to "local" emphasis on special topics. For example, many of the papers in the journal "Educational Studies in Mathematics" are about ways of introducing one particular topic into a course, sometimes through a disguised model. In the ATM Journal "Mathematics Teaching" much of the material is about creating mathematical "situations" which are explored by classes in detail ; this is very much an example of the Open Tendency, but frequently the material is left in an "intuitive" state without a formal, summarised theory. Its primary aim is frequently to stimulate interest in both teacher and pupil, and it makes great demands on the teacher particularly, by placing him in positions where

(1) These factors are not solely concerned with the nature of the target audience ; they are explained in more detail in [2].

he must admit to his pupils that he may be more ignorant than they are. He must also be quick to respond appropriately to unforeseen reactions by his pupils as they develop the mathematical "situation" in class. An interesting sidelight on certain of such articles is the way in which they have sometimes stimulated an author's research within mathematics, exclusive of education. Two good examples are Hilton [6] and Steiner [10]. Throughout this lecture, I am neglecting a second important aspect of Mathematical Education, namely the development of "objective" methods of testing and evaluating the effects of new curricula, as distinct from the "subjective" judgements we all make. Such objective methods are being developed by the "big" projects, like SMSG in the United States, if only to find out whether the expenditure of money and effort is worthwhile : again we meet a sociological boundary-condition in the discipline ! I conclude this paragraph by drawing attention to the list of research problems in the paper [7], and I would like to point out the need for comparable efforts by Applied Mathematicians (who have a very hard task ahead of them, particularly in physics). See, for example, Heading [4].

7. An example from Topology.

I will end the lecture with an example, to illustrate both the "watering down" problem, and the "local" emphasis I mentioned previously. Thus I will show how a piece of "official" mathematical theory may be watered down in a rigorous way, commenting on the insight as we go along. The particular piece of theory is that which surrounds the statement that any two connected orientable surfaces are homeomorphic if and only if they have the same genus and the same number of boundary components. I wanted to teach it to second-year mathematics students in a University, partly because it might help develop their geometrical intuition in three-dimensional space better than more usual material, and partly to give them a good example of a classification theorem in mathematics. [Note that I give justifications for my choice of topic, other than "I simply wanted to teach it", because one competes with one's colleagues for time in a crowded programme. Note too that I was moved by the Renewal Tendency].

The class knew no topology, having taken previously only an elementary course of rigorous Analysis, and a course of linear algebra. Thus to give the "official" treatment in full, I would need to teach them about homeomorphisms, then define a surface, triangulate it, and either set to work on the surface symbol, or (as I prefer) follow handlebody theory. Of course I could race through the "official" theory as presented in the books and some graduate courses : this would require least energy from me, but would be unlikely to communicate much to my audience, *with a consequent total loss of academic standards* and a waste of all our time. Some mathematicians pride themselves on the speed with which they can empty a class-room, to be left in peace ; but the contemporary climate in Britain (not to mention social conscience) does not favor that type of solution.

To raise the academic standards, then, it was necessary to "water down" the official theory (this will seem paradoxical to a rigorist!).

As a first step in watering down, I could confine myself to triangulable surfaces and omit the triangulation theorem (which requires the Schoenfliess theorem in its proof). Knowing the difficulties all students have with notions of continuity, I wanted also to emphasise the combinatorial aspects of the proof ; the ideas involved are quite hard enough for them, and if these were mixed with notions of continuity, all insight might be obscured. I therefore had to begin with a combinatorial definition of surface, and homeomorphic surfaces would then need to be replaced by combinatorially equivalent surfaces. It was still necessary to avoid saying that the basic pieces of surface (which I called "panels") were homeomorphic to polygonal discs in the plane, and I did this essentially by ostensive definition, appealing to the "obvious" fact that the union of two such panels, with a single common edge, is still a panel of the same kind. Such appeals can be left as irritants to very critical students, to make them ready for the surprisingly hard formal proofs provided a year later when they can take a rigorous course including the Jordan curve theorem and the Schoenfliess extension of it.

Thus having got my definitions into a workable form, I demonstrated the appropriate version of the theorem. Unfortunately, they were *my* definitions, not the students ; as it was, the whole demonstration took several lectures if I include the time devoted to drawing pictures of particular surfaces and describing them in detail, with the associated calculations of Euler characteristics. I did begin, it is true, (using the Open Tendency) by *asking* the students to create a definition of a surface, but in spite of hearing very interesting discussions, I have never yet been able to get a class to devise a workable definition in the limited time available. Moreover, because they are twenty years old, they find it undignified to build paper models of surfaces to help their intuition. Their non-manual training makes them unable to give rapid replies to the off-hand question "How do engineers make surfaces ?". On the other hand, they find the subject interesting, but their lack of internal skill shows that it is difficult for them. After my experience so far, I now believe that the treatment I have mentioned should follow a course on the physical construction of surfaces, which might be practicable for much younger children. Thus, watering down further, the basic "panels" would be polygonal sheets of paper, and a "surface" would be an ordered set of panels successively taped together along edges according to obvious rules imposed to ensure simplicity. The resulting paper surfaces are then objects of study, with a theory[1] isomorphic to the abstract combinatorial theory. It is natural to compute Euler characteristics, to count boundary curves, and to recognise similarities by agreeing to ignore curvature, or by allowing sub-divisions. Here there is time and opportunity for the exercise of "poetic skill", to reach the stage where one can ask for the genus of real-life surfaces (like that of a spoked wheel) or a student can see that an annulus whose edges are joined by a bridge is similar to a punctured torus T ; or (quite hard) if we glue two opposite edges of a rectangular panel to the boundary of T, after a half-twist, we get the "same" result as glueing 3 such twisted rectangles to a disc. In principle, this part of the material is accessible to children without algebra ; but the statement of the classification theorem would have to follow training in making statements of the form '$\chi = 2 - 2n$' (rather than '$\chi = - 66$'), and a proof is impossible without the method of induction. It

[1] An expository book is in preparation.

may turn out to be a good *introduction* to induction, because students are often quite happy with algorithmic induction to evaluate simple sums and products, but they have great difficulty with bold inductive hypotheses of the form "Let every object with parameter n have property $P(n)$".

If we now imagine that this entire procedure is reversed, so that a student studies first paper surfaces, then combinatorial theory of surfaces, and finally the topological theory, then he ought to have a strengthened spatial intuition and a greater chance of possessing topological insight. I say "ought", but this statement is a pious hope, without weight until tested by experiment ; such a requirement is a third important aspect of recent work compared with the earlier, functionally ineffective, general statements and philosophies of traditional Education departments.

By these remarks, then, I hope to have indicated something of the problems involved in the design of modern curricula, using the principles of the emergent activity "Mathematical Education".

REFERENCES

[1] Crowell R.H. — *Knots and Wheels,* Enrichment Mathematics for High School, 28th Yearbook, 1963, Nat. Council of Teachers of Math.

[2] Griffiths H.B. — Mathematical curriculum studies for mathematicians, *Proc. U.N.E.S.C.O. Conf.,* 1968, Bucharest (to be pubished).

[3] Hammersley J.M. — On the enfeeblement of mathematical skills..., *Bull. Inst. Math. and Applics.,* 4, 1968, p. 66-85.

[4] Heading J. — *The present position of Applied Mathematics in the U.K.,* Inaugural Lecture, 1969 (University of Wales Press, Cardiff).

[5] Herstein I. — On the Ph. D. in Mathematics, *Amer. Math. Monthly,* 76, 1969, p. 818-824.

[6] Hilton P.J. — Correspondences and Exact Sequences, *Proc. Conf. on Categorical Algebra* (LaJolla), 1965, p. 254-272, Springer.

[7] Hilton P.J., Long R.S. and Meltzer N.S. — Research in Mathematics Education, *Educational Studies in Mathematics,* 2, 1970, p. 446-469.

[8] Kapur J.N. — Combinatorial analysis and school mathematics, *Ibid.,* 3, 1970, p. 111-128.

[9] Kaufman B.A. and Steiner H.G. — Checker games in operational systems as media for an inductive approach to teaching algebra, *Ibid.,* 1, 1969, p. 445-484.

[10] Steiner H.G. — Mathematisierung und Axiomatisierung einer politischen Struktur, *Der Mathematikunterricht,* 1967, p. 66-86, Klett, Stuttgart.

Pamphlets :

[11] *The development of the B. Ed. Degree in Mathematics,* First report of a working party of A.T.C.D.E. (Mathematics Section), 1970.

[12] *Goals for mathematical education of elementary school teachers,* Houghton-Miflin, 1966.

Dept. of Mathematics,
The University,
Southampton
Grande-Bretagne

Actes, Congrès intern. Math., 1970. Tome 3, p. 347 à 351.

PROBLÈMES DE LA FORMATION MODERNE DES PROFESSEURS DE MATHÉMATIQUES

par Zofia KRYGOWSKA

La réforme des études mathématiques des futurs maîtres a été proposée depuis longtemps. Les nouvelles tendences de l'enseignement mathématique ont mis en évidence l'urgence particulière de cette réforme, car les difficultés dans la modernisation de la mathématique élémentaire découlent non seulement des connaissances mathématiques surannées de beaucoup de maîtres plus âgés, mais aussi de l'indolence méthodologique de beaucoup d'autres maîtres qui ont reçu l'éducation mathématique moderne. On constate aussi les défauts du recyclage trop rapide et trop formel des maîtres en fonction, ce qui conduit chez certains enseignants à la fascination par la terminologie mathématique moderne couvrant souvent le manque d'une culture mathématique plus profonde.

C'est pourquoi le colloque international consacré aux problèmes de l'enseignement des mathématiques dans les écoles secondaires et supérieures des pays européens, qui à eu lieu a Bucarest en 1968, a souligné l'urgence des travaux concernant :

(1) l'analyse des structures actuelles des études scientifiques des futurs enseignants du point de vue de leur efficacité pour l'exercice de la profession de l'enseignant de la mathématique ;

(2) l'amélioration de ces structures au fur et à mesure des besoins ;

(3) leur adaptation au niveau réel du recrutement des futurs professeurs de la mathématique.

La discussion concernant tous ces problèmes devrait être basée sur les réponses aux trois questions suivantes :

(1) Quels sont les objectifs généraux de l'éducation mathématique au niveau scolaire envisagés actuellement et dans la perspective du développement de la science, de la technique et des structures sociales ?

(2) Quelles connaissances et quelles attitudes intellectuelles et culturelles des maîtres pourraient garantir la réalisation des buts ainsi définis ?

(3) Quelles sont les conditions réelles de la formation des maîtres de la mathématique dans la société donnée : recrutement à la profession, qualifications des cadres formant les futurs enseignants de la mathématique, etc. ?

Beaucoup de congrès internationaux et locaux se sont occupés de ces questions. Leurs conclusions expriment le plus souvent les mêmes postulats. Le colloque de Bucarest –précédemment mentionné– a exprimé certaines de ces idées générales commc suit : "l'objcctif social dc l'enscignement est double :

(1) Transmettre à la société la culture mathématique et développer au mieux le potentiel mathématique de chaque individu.

(2) Former les spécialistes qui assureront l'enseignement de la mathématique, feront progresser la recherche fondamentale et appliquée, développeront les applications en collaboration avec les spécialistes des disciplines interessées...

"L'attention portée à l'enseignement secondaire ne doit pas être restreinte aux orientations qui conduisent aux hautes études. Les autres orientations de l'enseignement secondaire méritent plus d'attention qu'elles n'en ont obtenu jusqu'alors" [1]. L'idée de "la culture mathématique pour tous" a été renforcée dans les conclusions et dans les recommandations du Premier Congrès International de l'Enseignement Mathématique qui a eu lieu à Lyon en 1969 et qui a recommandé à la CIEM : "En ce qui concerne la structure du prochain congrès, attribuer plus d'attention à l'enseignement pré-scolaire, à l'enseignement élémentaire, à l'enseignement mathématique pour la totalité de la jeunesse, à l'enseignement des adultes" [2]. Le colloque de Lausanne sur la coordination des enseignements de la mathématique et de la physique a admis comme principe des travaux concernant la modernisation de la mathématique élémentaire : "chaque enfant a le droit d'être introduit au monde des structures élémentaires et fondamentales de la mathématique" [3].

Les problèmes des études mathématiques des futurs maîtres ne se réduisent donc pas aux problèmes de la formation d'une élite des professeurs pour une élite des élèves. Au contraire, notre tâche très importante et très difficile, c'est la formation moderne de la masse des maîtres pour la masse des élèves, car ce n'est pas par l'élite des professeurs et par l'élite des élèves qu'on pourrait élever le niveau de la culture mathématique de toute la société.

Au cours de beaucoup de rencontres internationales on a analysé le concept même de cette "culture mathématique pour tous" en mettant en relief les traits suivants :

(1) connaissances de base structurées à l'aide des structures mathématiques considérées comme fondamentales à l'étape donnée du développement de la science ;

(2) prise de conscience –même approchée seulement– de la construction formelle de la mathématique, de son rapport avec la réalité, de ses relations multilatérales avec les autres sciences : concepts fondamentaux de la méthodologie scientifique de la mathématique, définition, théorème, démonstration, mathématisation, méthode axiomatique, interprétation et application de la théorie mathématique, etc. ;

(3) initiation aux procédés divers de la pensée mathématique en acte : abstraire, schématiser, mathématiser, déduire, rechercher et découvrir les connexions mathématiques, généraliser et spécialiser, se servir de la symbolique mathématique, des modèles concrets et pensés dans la solution des problèmes, organiser rationnellement les données du problème etc. ;

(4) aptitude à exprimer correctement la pensée mathématique propre : définir, présenter clairement le raisonnement, formuler d'une manière précise le problème etc. ;

(5) une certaine technique de l'étude individuelle en mathématique : savoir lire efficacement les textes mathématiques, contrôler les résultats du travail individuel, chercher et corriger les fautes commises au cours de ce travail, etc.

Evidemment l'interprétation concrète de cette conception de la "culture mathématique générale" devrait être pondérée quant au contenu des connaissances et au niveau de l'abstraction, compte tenu des possibilités des élèves et des conditions de l'enseignement. Mais cette restriction ne changera pas l'essentiel de cette conception : la culture mathématique de la société, ce n'est pas seulement l'habileté dans les simples calculs arithmétiques et la connaissance de certaines notions et théorèmes géométriques, mais avant tout la compréhension de la méthode mathématique de penser, de résoudre les problèmes, la compréhension du caractère relationnel et structurel de la mathematique contemporaine et la prise de conscience que c'est justement grâce à ce caractère abstrait et relationnel que la mathématique est devenue aujourd'hui omniprésente dans notre civilisation.

La transmission de cette prise de conscience aux larges couches de la société, c'est un des objectifs les plus importants de la modernisation de la mathématique élémentaire et un des objectifs primordiaux de l'éducation générale.

Cet objectif entraîne les implications inévitables concernant la formation adéquate des maîtres de la mathématique :

(1) Il faut préparer la masse de maîtres pour la masse d'élèves.

(2) Il n'est donc pas possible de restreindre le recrutement à la profession de l'enseignement de la mathématique aux étudiants très doués. Au contraire, il faut compter avant tout sur les étudiants moyens, et, dans certains pays même plus faibles, en tant que candidats représentatifs à cette profession.

(3) Ces étudiants moyens ou même plus faibles doivent être introduits, au cours de leurs études, dans le monde mathématique contemporain assez profondément, non seulement du point de vue de l'étendue de leurs connaissances, mais aussi du point de vue de leur culture mathématique générale, dans le sens précédemment défini.

La structure actuelle des études mathématiques des futurs maîtres n'est pas adaptée à ces conditions sociales et à ces besoins sociaux. Pour beaucoup d'étudiants moyens leurs études mathématiques c'est seulement la préparation permanente et forcée aux examens qui traitent des sujets isolés, une préparation qui néglige la structure du sujet, qui ne connaît aucune réflexion méthodologique, mais qui est avant tout concentrée sur la mémorisation du contenu présenté par les cours magistraux et sur l'assimilation de certains mécanismes dans la solution de problèmes typiques. Une telle étude, basée sur l'information forcée et non sur la formation de l'esprit mathématique, n'est pas favorable à l'assimilation active de la culture mathématique par les futurs maîtres de la mathématique.

Le futur chercheur passe aussi par cette période de l'information forcée. Mais (1) il est plus doué, il se prépare donc plus facilement et plus vite aux examens, il a donc le temps d'étudier la mathématique de base plus activement, il a le temps de réflechir ; (2) il développe son activité mathématique dans la suite de ses études, au cours des séminaires avancés, au cours de la recherche. C'est en faisant la mathématique qu'il assimile activement la culture mathématique de son temps. Les études d'un étudiant, futur professeur, s'arrêtent le plus souvent à cette étape de l'information de base. Il arrive donc trop souvent que le maître de la mathématique est obligé d'initier ses élèves activement aux rudiments d'une culture mathématique, à laquelle lui-même n'a pas été initié plus profondément.

Cet état de choses ne pourrait pas être efficacement amélioré sans changement de la conception actuelle des études mathématiques des futurs maîtres. L'analyse de cette conception faite au cours des discussions ferventes dans certains pays —p. ex. en Pologne— conduit aux conclusions suivantes :

(1) La formation intiale mathématique des étudiants qui s'orientent vers l'enseignement exige une spécialisation visant expressément les besoins de leur future profession.

(2) Les programmes surchargés freinent l'activité mathématique des étudiants moyens ou plus faibles, qui sont réprésentatifs pour la masse de futurs maîtres. L'expérience prouve que certaines réductions du contenu, l'élimination des détails au profit du traitement approfondi des idées et des techniques particulièrement importantes du point de vue de l'esprit mathematique moderne et de sa transmission au niveau élémentaire, s'avèrent favorables au développement de l'activité mathématique de ces étudiants.

(3) La présentation de la mathématique devrait être basée sur une conception synthétique de cette science. Au lieu d'étudier beaucoup de sujets lâchement liés, au lieu d'être obligé de passer par un grand nombre d'examens détaillés, l'étudiant — futur maître devrait être introduit dans le monde mathématique par des voies peu nombreuses et le conduisant vite et directement vers l'essentiel.

(4) La formation mathématique d'un bon maître exige son initiation à l'étude individuelle et à une véritable recherche mathématique, adaptée néanmoins à ses possibilités assez modestes. Ce postulat ne serait pas réalisé sans réduction du temps consacré aux cours magistraux au profit des séminaires et des consultations, sans une conception nouvelle des travaux dirigés. Les exercices traditionnels concernant l'application presque mécanique de la théorie présentée par le cours magistral à la solution des problèmes standards devraient être complétés et partiellement remplacés par la mise des étudiants dans les situations mathématiques ouvertes conduisant aux procédés de la pensée tels que : généraliser une notion ou un théorème, dégager une nouvelle notion au cours du processus de la mathématisation, construire les exemples et les contre-exemples, formuler les problèmes et les hypothèses etc. Il ne faut pas oublier que les situations analogues forment la base de l'enseignement mathématique moderne au niveau scolaire. Le maître ne sera pas apte à un tel enseignement sans expériences personnelles analogues faites au cours de ses études sous la direction des mathématiciens-créateurs.

(5) Du point de vue de l'étudiant moyen beaucoup de cours magistraux traditionnels ne sont que des chaînes formelles de définitions, de théorèmes, de démonstrations. Il ne voit pas derrière cette vitrine formelle une mathématique vivante, avec son passé et ses perspectives, avec les mécanismes de son développement. Souvent il ne connaît aucune motivation de telle ou telle généralisation et il arrive trop souvent qu'il ne voit aucune liaison entre les théories abstraites et leurs modèles les plus simples, élémentaires. Il ne comprend pas le sens intuitif

et la portée de telle ou telle méthode etc. Beaucoup de cours magistraux traditionnels négligent les commentaires méthodologiques et historiques qui mettraient en relief la pensée mathématique en acte, ce qui est particulièrement important du point de vue de l'éducation mathématique du futur maître. La réforme des études mathématiques des enseignants exige donc une nouvelle conception du cours magistral.

(6) La méthodologie de l'enseignement de la mathématique a dépassé déjà l'étape de l'apprentissage pratique du métier et devient plus en plus le domaine de la recherche étroitement liée à la mathématique et à la méthodologie scientifique. La formation des spécialistes dans l'enseignement ne peut pas passer sous silence cette discipline en devenir. Au cours de ses études le futur maître devrait être introduit aux travaux menés par les centres représentatifs de la recherche méthodologique de la même manière que le futur chercheur fait la connaissance des travaux menés par les centres représentatifs de sa propre spécialité. L'équipement de la Faculté devrait donc assurer les sources d'une telle information : ouvrages méthodologiques, revues consacrées aux problèmes de l'enseignement, manuels utilisés dans différents pays, programmes de l'enseignement mathématique dans différents pays, compte-rendu des travaux faits au cours des congrès sur les problèmes de l'enseignement mathématique, matériaux didactiques modernes etc. Le professeur qui ne connaît qu'un seul programme, qu'une seule conception de l'enseignement mathématique, qui ne s'intéresse pas aux expériences organisées dans les autres pays, qui n'a pas été introduit à l'étude comparative des conceptions diverses de la modernisation de la mathématique élémentaire, deviendra très facilement hostile aux réformes nouvelles dans son propre pays. La formation de son esprit méthodologique souple, ouvert à la recherche, résistant aux dangers de la routine est un des objectifs les plus importants de sa formation scientifique.

La création de certaines institutions telles que p. ex. les chaires de la méthodologie de l'enseignement mathématique au sein des Facultés ou les Instituts de la recherche sur l'enseignement mathématique liés aussi étroitement aux Facultés, prouve que la réforme de la formation des maîtres de la mathématique a fait certains progrés.

Ecole Normale Supérieure de Cracovie
Cracovie
Pologne

[1] Ce texte a été transmis directement aux participants du Congrès.

[2] Résolutions du Premier Congrès International de l'Enseignement Mathématique, Educational Studies in Mathematics, Volume 2, No 2/3, 1969.

[3] Dialectica, 21/1967/, fasc. 1-4.

Actes, Congrès intern. Math., 1970. Tome 3, p. 353 à 357.

ON TEACHING APPLICATION OF MATHEMATICS

by H.O. POLLAK

Mathematics curriculum reform in many countries during the past ten years has made major progress in three quite distinct areas : The subject matter that is taught in the schools, the emphasis on understanding, and the pedagogy of discovery and open-ended situations. In subject matter, all kinds of new things are finding their way into the curriculum. Thus, sets and functions, logic, absolute values, inequalities, linear algebra, matrices, probability, statistics, limits, calculus, computing, and modern algebra have all been added to varying degrees in diffërent situations. Furthermore, we have realized that it is essential to understand mathematics in addition to being able to carry out the procedures. In fact, it is surprising that we have survived for so many years without emphasizing understanding. After all, mathematics is a science. Just as, in physics, the student has the right to repeat any experiment rather than accepting its outcome on faith, so in mathematics the student has the right to understand when and how and why the mathematics works. With regard to discovery and open-endedness, people have come to believe in them without agreeing on what they mean. Sometimes a teacher will keep asking questions and wait for someone to say the right thing. Sometimes over a long period students and the teacher will agree adaptively on the right way to define a particular mathematical concept. Sometimes a teacher will take discovery to mean that you must never say "no" to youngsters but must always ride with whatever answers you get from the class. Sometimes discovery means finding a pattern out of examples. Sometimes it means working with a situation in mathematics, finding out what is going on, and then formulating conjectures. Whatever mixture of these is being carried on, it is clearly a change in pedagogy, and in most opinions an improvement.

It is perhaps worth noting in looking back that there is also some danger associated with each of these three aspects of curriculum reform. People sometimes tend to bring interesting highlights of subject matter down as early in the curriculum as possible. This means that the hard work associated with a particular topic sometimes doesn't get done. When this work must be done later its climax may thus have been spoiled. Secondly, if you emphasize understanding the students may get into the game by asking "why" every two minutes. This kind of disruption is, of course, easily handled by asking them to go home, think about it, and tell the class tomorrow. Finally, there is also some danger in discovery. Just because discovery of a pattern is better than straight rote learning, this doesn't mean that guessing a pattern is all there is to mathematics. It is necessary to prove or disprove your guesses after you make them.

The three kinds of changes which we have been discussing are also very valuable for the applications of mathematics. First of all every one of the topics which we listed above, and many others, are eminently useful applied mathematics as well as good pure mathematics. This statement of course does not relieve us of the responsibility for making choices. If, as often happens in the United States, there is some extra time for mathematics near the end of secondary school, should this be used, for example, for calculus or for probability and statistics or for computing ? An argument could be made in favor of each one. Calculus has the greatest number of applications to other subjects in the university, probability and statistics are probably the most useful in everyday life, and the computer will affect each of us more in the coming years than almost anything else. In the latest work on curriculum reform in the United States we have attempted to include something of all three. The emphasis on understanding is essential for applications because genuine applications of mathematics are often not exactly like problems in the textbook. You forever find yourself studying something new, and this has a chance of success only if previous mathematics has been understood. The open-ended discovery aspect of mathematics teaching is of course exactly what applied mathematics is all about. Whenever we apply mathematics in practice we are looking at a situation in some other field and trying to understand it through modeling it mathematically. I admit that we frequently do not teach applied mathematics in this way but this just means that we are teaching it badly. What I am considering here are real applications of mathematics. Differences between real and textbook applications of mathematics have been studied, for example, in Chapter 8 of the 69th Yearbook of the National Society for the Study of Education, E.G. Begle, editor, University of Chicago Press, Chicago, Illinois.

Real applications of mathematics may be applications to science or to the social sciences or to everyday life. It is important to have all kinds in the čurriculum for many reasons, but the most fundamental is that the mathematical experience is simply incomplete without them. We just haven't done our job of teaching mathematics if the student has not seen how it is really used. In this connection it must be emphasized that finding a mathematical problem that will help us understand a real situation is a definite achievement. This is true whether or not we can analytically solve the problem that we have found. Many times we lead the student to believe that only the answer is important. In real applications of mathematics finding the right question is every bit as important as finding the answer. It is of course not easy for students to take an ill-defined situation and pick a specific interesting aspect to study. If you simply assign a variable to every quantity in sight, you are not yet doing anything. There must be a specific question with a purpose, a question whose answer will add to the understanding of the situation.

Among the three major directions, I should like to examine particularly applications of mathematics to the social sciences. By these I mean, questions related to psychology, sociology, economics, geography, political science, etc. Applications to these areas are particularly valuable in the schools. First of all, they provide a rich variety and a change of pace for the students' interests. Secondly, questions in the social sciences whose mathematical modeling is interesting are often closer to the surface than questions in the physical sciences. It is perhaps particularly true in the United States that students –and teachers– of mathematics do not know much science. Against this kind of background, problems in the social sciences are often more readily understood than those in the physical sciences. Finally, and we might as well admit it, there is a definite decrease in the interest of students in the physical sciences and engineering. Students are now much more interested in biological and social problems. There is of course no reason why we should lose the student's interest in mathematics just because he has lost interest in physics. There are numerous excellent applications of mathematics throughout the social sciences which now excite students, the applications are excellent mathematics, and it would be wise for us to use them.

The latter argument applies at every level of education from kindergarten to the university. For the rest of this paper I should like to discuss in particular some examples of possible applications of mathematics to the social sciences in the elementary school. There are of course many obvious applications of arithmetic, and also of probability and statistics. Since these are more familiar let me emphasize instead a number of elementary applications of other areas of mathematics. There are many opportunities for all kinds of mathematical thinking in the elementary school. The three examples which I will give will deal with geometry in a broad sense, with algorithmic thinking, and with the mathematics of decision making. They are motivated by, and to a large extent taken from, a current experimental project entitled Unified Science and Mathematics for Elementary Schools.

In one experiment, the children were blindfolded and were given a series of samples of soft drinks. They knew that the six choices were grape, cherry, root beer, cola, orange and lemon, and were asked to identify which they were drinking. The outcome may be seen in Figure 1, which shows for each given sample what the students believed it was. We see, for example, in the twelve cases in which they were given grape soda they believed nine times that it was grape, two times that it was cherry and once that it was lemon soda. We now wish to make a geometric picture of the students' responses. We would like to represent the six drinks in a Euclidean space of suitable dimension so that two points will be close together if the substances were often confused, and will be far apart if they were rarely confused. More precisely, we deduce a confusion (similarity) matrix between each pair of drinks from Figure 1, and then wish to locate six points so that the distance between any two is monotonely related to their dissimilarity. One outcome of this procedure may be seen in two dimensions, in Figure 2. We now examine possible interpretations of the coordinates. We find that one coordinate is naturally interpreted as sweetness while the other is a discrete fruit-nonfruit variable. This procedure allows the students to see what kind of thinking they have actually used in order to make their decision, and is very interesting from the point of view of education in the social sciences. On the mathematical side, the students have practiced an approximate plotting of points in the plane so that the distances come out roughly in the right order. Such approximate thinking, and such a relation between geometric and arithmetic thinking, are mathematically very valuable.

A second example concerns a very popular kind of unit entitled "who am I ?" The purpose of this is to let the students see themselves characterized in many different ways and as members of many different groups. Thus, a student may be tall

		Guess					
		Grape	Cherry	Root beer	Cola	Orange	Lemon
Given	Grape	9	2				1
	Cherry	4	7				
	Root beer		1	5	1		
	Cola			1	9		2
	Orange	1			1	11	1
	Lemon						13

Figure 1

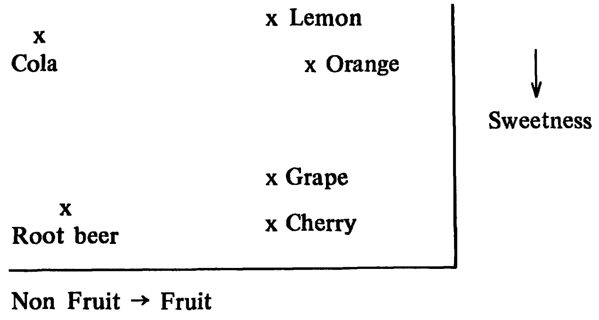


Figure 2

	Sex	Eyes	Weight	Height	Hair
Ann	F	Blue	Light	Tall	Brown
Dave	M	Blue	Heavy	Short	Brown
Jack	M	Blue	Heavy	Short	Blond
Bob	M	Brown	Light	Short	Blond

Figure 3

or short, or light or heavy, blue-eyed or brown-eyed, male or female, 10 or 11 years old, etc., etc. We can thus create a table of characteristics with one line for each student in the class. The mathematical interest now comes from the following kind of question. How would you identify one particular person uniquely ? Consider for example the simplified case in Figure 3. What series of questions would you ask to identify each student ? What is the *smallest* number of questions that can ever be successful ? Can you always expect to identify each student uniquely ? If so, what is the *largest* number of questions that it might take ? This is just the beginning – it is possible to develop both the mathematics and the applications of this kind of algorithmic thinking in many different directions.

As a third example let us consider the problem of group decision making. For example, the class may be told that they have five dollars to spend on any item they choose, provided they reach a decision before the end of the week. We may then study the alternative decision-making procedures and the distinctive properties of each. Other projects in the same spirit which the students might undertake are to lay out a playground or redesign the lunch line in the school, or study pedestrian crossings near their school and various methods for their protection. Another nice example in this general area is to study the best location for a new airport. Suppose, for example, that you wish to serve four cities of roughly equal population. In case these four cities lie at the vertices of a convex quadrilateral, the best location might well be the intersection of the diagonal. Now suppose you have built that airport and you find you need a second one. If you assume that every passenger will go to the airport which is closer to him, then the second location will be at the midpoint of one of the sides. If, on the other hand, you had decided from the beginning that you would have to build two airports rather than one, both of them would be located at the midpoints of sides and there would be none at the intersection of the diagonal. Sequential decisions can lead to solutions quite different from simultaneous decisions.

The above have been some samples of current experimental thinking on applications of mathematics to the social sciences in the elementary schools. We do not of course know how well they will work. As a matter of fact, there is not even any conclusive experimental evidence that motivating mathematics through applications helps in the learning of mathematics. Nevertheless, we are excited about the possibility that along with abstraction, generalization, puzzles and games, and classical applications of mathematics, these modern applications stand a good chance of success.

Bell Telephone Laboratories, Incorporated
Murray Hill, New Jersey 07974
U.S.A

Actes, Congrès intern. Math., 1970. Tome 3, p. 359 à 367.

QUELQUES ASPECTS DE L'ENSEIGNEMENT DES MATHÉMATIQUES EN U.R.S.S

par S.L. SOBOLEV

Comme dans tous les pays développés, les mathématiques sont enseignées en URSS á deux niveaux : celui de l'école moyenne et celui de l'école supérieure.

L'čeole moyenne qui correspond à la période de la scolarité obligatoire qui est de 10 ans en URSS reçoit tous les élèves de l'âge de 7 - 8 ans à celui 17 - 18, avec cette réserve que l'on peut quitter l'école moyenne après huit années d'études et fréquenter dans les deux dernières années une école professionnelle moyenne (école technique, école moyenne de médecine, école préparant directement un métier, . . .).

Sous le terme d'école supérieure il faut entendre : les Universités (au nombre de 47), les Instituts pédagogiques, les Instituts formant des ingénieurs pour les différentes branches de l'Industrie, les Conservatoires, les Ecoles supérieurs d'art. Les étudiants de ces écoles ont tous au moins 17 ans.

A ce niveau, la mathématique est enseignée dans les facultés mathématiques, mécanico-mathématiques, physico-mathématiques des Universités et des Instituts pédagogiques et d'autre part comme une des matières de base dans les autres Facultés et Ecoles Supérieures.

Au niveau de l'école moyenne, le contenu de l'enseignement mathématique est longtemps resté invariable et fidèle aux anciennes traditions. Après une étude assez sérieuse de l'arithmétique, comprenant la solution dite arithmétique de problèmes compliqués, on passait à un cours d'algèbre traditionnelle (transformation d'expressions contenant des lettres, solution d'équations et de systèmes d'équations algébriques avec application à divers problèmes). La géométrie, exposée en même temps que l'algèbre, était fondée sur la considération d'images concrètes et un système d'axiomes très proche de celui d'Euclide. Puis venait un cours traditionnel de trigonométrie.

Le développement du progrès technique et l'accroissement du rôle de la Science dans la Société ont imposé une révision du contenu et du style de l'enseignement des mathématiques (et des autres disciplines aussi) à l'école moyenne. Certaines notions essentielles de ce qui était la mathématique "Supérieure" : dérivée, intégrale, équation différentielles simples sont devenues un outil indispensable pour l'homme contemporain, indépendamment de son métier, s'il veut être au courant de ce qui se passe dans la vie moderne. Sans elles, il lui est difficile de dominer les acquisitions de la Science et de la Technique modernes (ou simplement de comprendre les articles de la presse ou de la littérature scientifique populaire). Ces idées sont également d'un grand secours pour l'ouvrier qui essaie d'inventer ou de rationaliser la production.

Pour être juste, il faut signaler que des tentatives d'introduction de l'Analyse infinitésimale au niveau de l'école moyenne, avaient été faites dès avant la Révolution dans certaines "écoles réelles", mais il ne s'agissait que de cas isolés.

Lors des dernières années, l'élargissement des programmes apparut comme une nécessité vitale, et de nouveaux programmes ont été élaborés avec l'aide de larges cercles de professeurs de mathématiques.

Le changement a concerné le contenu et le style, l'objectif étant de combler le fossé entre arithmétique et algèbre et de détruire la barrière qui séparait mathématique élémentaire et mathématique supérieure. A cet effet, on a introduit beaucoup plus tôt l'usage de lettres pour désigner des inconnues, ainsi que les nombres négatifs. Dès la première année, on étudie des figures géométriques et des problèmes simples d'origine géométrique.

Le point de vue fonctionnel intervient dans l'étude de l'algèbre ainsi que l'emploi de graphiques. Le cours d'algèbre des nouveaux programmes se terminera par l'introduction de la dérivée et de l'intégrale, mais sans donner une technique développée de la différentiation et de l'intégration. Les concepts les plus simples de la théorie des ensembles et de la logique mathématique sont introduits progressivement, et fournissent un langage commode, en particulier pour l'étude des systèmes d'équations et d'inéquations et pour celle des fonctions. Une plus grande importance est donnée à l'utilisation de coordonnées et aux représentations graphiques des fonctions. Enfin, les vecteurs sont utilisés en géométrie.

Les changements de programmes n'affectent à présent que les jeunes élèves ; qui utilisent de nouveaux manuels, choisis après un concours qui s'est étendu sur plusieurs années.

L'accent a été mis dans ces nouveaux manuels sur le contenu du cours, car le but principal a été d'élargir l'horizon des jeunes, mais les considérations purement mathématiques ou pédagogiques n'ont pas été laissées de côté.

Il y a peu de partisans en URSS d'une restructuration intégrale de l'enseignement mathématique à l'école moyenne, exclusivement fondée sur la théorie des ensembles, les notions topologiques etc. C'est un point de vue modéré qui a triomphé qui tient compte des idées modernes, mais considère que le plus important est le renouvellement des faits concrets enseignés et la pratique de la mathématique. Il n'y eut pas de lutte sérieuse sur ces problèmes.

L'introduction de nouvelles notions a pu être faite grâce à l'économie de temps qu'elles permettent. On a pu écarter les solutions dites purement arithmétiques, on a pu simplifier l'exposé de la géométrie. Cependant, on n'a pas pu éviter de faire quelques victimes : certains procédés algébriques trop compliqués, des méthodes particulières de résolution d'équations, certains détails géométriques.

Une nouvelle disposition permet cependant d'augmenter le volume des connaissances mathématiques : il s'agit des cours optionnels auxquels 4 h hebdomadaire sont réservées dans les dernières classes, et que les élèves choisissent selon leur gout. Beaucoup choisissent les mathématiques.

En outre, il existe des écoles spéciales dont les classes supérieures donnent une formation plus poussée en mathématiques et en physique. Ces écoles fonctionnent depuis plusieurs années et leur expérience a été utilisée pour mettre au point les nouveaux programmes et les nouveaux manuels. Les plus importantes sont des internats situés à proximité des grandes Universités : Moscou, Leningrad, Novosibirsk, Kiev.

Leur objectif était d'attirer vers les Universités la jeunesse la plus douée des petites villes et villages éloignés des grands centres culturels.

Certains des mathématiciens les plus éminents enseignent dans ces écoles. La recherche des élèves de talent est facilitée par les compétitions annuelles, dites olympiades, qui ont lieu en 3 phases : la première dans les écoles mêmes, la seconde dans certaines grandes villes, et la troisième à Moscou. Les olympiades sont maintenant assez populaires (la première a eu lieu à Leningrad en 1934). Grâce à elles, de nombreux jeunes gens ont trouvé leur vocation. Les olympiades, les écoles spéciales et d'autres institutions analogues font la propagande de la Science en expliquant aux jeunes gens ce que sont les Sciences et en particulier la mathématique, et ce qu'est leur rôle dans la société moderne, ce dont n'ont guère idée ceux qui vivent dans les régions éloignées des centres et n'ont eu aucune information sur la question. Elles ont joué un rôle certain dans le changement de mentalité que l'on constate maintenant.

Je ne décrirai pas en détail l'enseignement mathématique dans les écoles supérieures techniques et les facultés scientifiques non mathématiques où il joue un rôle auxiliaire, subordonné aux besoins des diverses professions et varie par suite d'un Institut à l'autre. Il faut noter toutefois dans les dernières décennies un accroissement du volume des mathématiques enseignées dans des directions non traditionnelles ; cet accroissement est dû à la nécessité de faire connaître aux futurs ingénieurs la technique mathématique contemporaine et son pouvoir. Mais l'introduction de nouveaux programmes ne se fait qu'assez lentement, à cause du manque de personnel apte à les enseigner.

Dans les Instituts pédagogiques, on enseigne les disciplines mathématiques fondamentales : analyse, algèbre, géométrie, théorie des probabilités, logique mathématique en insistant sur les questions qui sont le plus liées à l'enseignement à l'école moyenne.

La formation des mathématiciens a lieu dans des facultés mathématiques des Universités. Les cours durent 5 ans. L'admission dans toutes les Ecoles Supérieures a lieu sur concours. Les études sont gratuites et la plupart des étudiants sont boursiers d'Etat (pour être boursier, il faut obtenir à tous les examens la mention Bien ou Très Bien).

Les plans d'études et d'examens sont fixés pour la durée totale des études, mais les étudiants qui se sont montrés particulièrement brillants peuvent être autorisés à suivre un plan d'étude personnel.

Le couronnement des études mathématiques est la préparation de thèses de deux niveaux, l'un qui donne la grade de "Candidat en Sciences physico-mathématiques", l'autre plus élevé qui donne le grade de "Docteur ès-Sciences physico-mathématiques". Tandis que la deuxième thèse est toujours préparée sans aide financière, il existe des bourses d'Etat dont bénéficient une partie de ceux qui préparent la première thèse, et qu'on appelle "aspirants". On reste aspirant pendant 3 ans, et pour obtenir le grade de candidat, il faut, outre la présentation de la thèse, passer les examens çorrespondant à un enseignement approfondi suivi pendant ces 3 années.

A notre époque de progrès scientifique impétueux, d'accroissement du rôle des mathématiques et de mathématisation de la connaissance, les plans d'études des universités ne peuvent rester invariables. Lors de ces dernières années; une discussion a eu lieu dans beaucoup d'universités de l'URSS sur la question de savoir comment il faut maintenant enseigner la mathématique.

Dans notre pays, mais sans doute aussi dans le monde entier, depuis longtemps, les jeunes gens brillants qui essayaient d'avoir une conception large des choses et qui avaient du goût pour les idées générales et profondes aimaient mieux s'occuper des mathématiques en elles-mêmes que de leurs applications. Ceci était renforcé par le fait que dans les branches des mathématiques qui n'étaient pas en liaison directe avec les autres Sciences, il était plus simple d'obtenir des résultats nouveaux ou d'inventer de nouvelles méthodes de recherche. Il arrivait en outre que l'estime qu'obtenait l'auteur d'un résultat concret, mais étroit, bien que difficile, était moins élevée que celle de l'auteur d'une généralisation large, mais au fond plus facile. Il en est résulté chez les jeunes chercheurs une espèce de mépris pour les questions où la technique absorbe une part importante des forces et du temps et où les résultats ont un caractère concret.

Cette tendance n'a été évitée nulle part, si l'on excepte quelques écoles très conservatrices, où l'estime allait certes aux résultats concrets, malheureusement aussi souvent triviaux.

Cette évolution très nette pendant le 19$^{\text{ème}}$ siècle, était parfois masquée par l'apparition, comme à l'Université de Moscou, d'une grande école de mécaniciens avec Joukowski et Tchaplyguine, mais le penchant des chercheurs pour les problèmes abstraits demeurait le plus fort, même dans une Université comme celle de Leningrad, toujours fière de sa vieille tradition de recherche appliquée, illustrée par Tchebychev, Markov (le père), Liapounov, Steklov.

Cette rupture entre la mathématique et ses applications eut aussi des causes historiques : la plupart des applications passées des mathématiques concernaient la mécanique (mécanique des solides, des systèmes, des milieux continus) et il arrivait même au début du XX$^{\text{ème}}$ siècle que l'on considérât la mécanique comme partie intégrante des mathématiques, négligeant ainsi le rôle de l'expérience. Sans doute n'était-ce pas une opinion universelle et dès les premières années du XX$^{\text{ème}}$ siècle, une partie importante des mathématiciens actifs ont élargi le champ de leurs recherches, tandis que les mécaniciens se séparaient des mathématiciens. Mais dans les grandes Universités, comme celles de Moscou, de Leningrad et quelques autres on trouvait encore des facultés de "Mathématiques et Mécanique" tandis que la mécanique expérimentale se développait en dehors des universités dans les écoles supérieures techniques et certains instituts de recherche. Hors de la mécanique, aucune application n'était enseignée à l'Université, si ce n'est le calcul approché qui cependant se réduisait à des estimations, la plupart fort banales, des erreurs de calcul, et à une liste de méthodes de routine dont on recommandait l'usage dans les problèmes classiques de l'analyse. Rien de ce qui constitue la théorie contemporaine du calcul numérique n'était évoqué.

Après les découvertes remarquables de Liapounov, Joukowski et Tchaplyguine la mécanique universitaire connut une longue période de stagnation. Et si pendant ce temps, la mécanique expérimentale reprenait ses droits, la jeunesse étudiante marquait toujours un goût affirmé pour le travail théorique et d'une génération à l'autre la distance grandissait entre la mathématique et ses applications.

Cependant dès les années 30, dans les grandes universités, à Leningrad, puis à Moscou de nouvelles idées apparurent sur les liaisons entre le calcul numérique et l'analyse fonctionnelle. De nouvelles branches naquirent avec la programmation mathématique ; l'importance du calcul des probabilités et de la statistique mathématique fut reconnue. Ces tendances étaient tout à fait conformes aux grands courants de la Science mondiale et se faisaient jour dans les autres pays. C'est alors que les professeurs d'Université commencèrent à montrer aux étudiants que les applications des mathématiques aux Sciences de la nature, et le calcul numérique lui-même, ne sont pas autre chose que des branches de l'analyse mathématique contemporaine et de la nouvelle mathématique discrète. Néanmoins, cela n'a pas provoqué de changements notables dans la mentalité des jeunes chercheurs et l'apparition de nouvelles idées et de problèmes nouveaux comme la programmation linéaire, la théorie des jeux, la théorie de la gestion, etc. ne semblait pas devoir changer la situation.

Dans ces conditions est née chez beaucoup de mathématiciens l'idée de diviser l'enseignement en "mathématique pure", "mathématique appliquée", "mathématique pour ingénieurs", "calcul numérique", "cybernétique" avec l'espoir que si l'on faisait connaître le plus tôt possible aux étudiants les brillantes applications de la mathématique contemporaine et que si on ne leur faisait connaître les idées abstraites qu'après qu'ils aient acquis le goût des applications, on obtiendrait un changement de mentalité. Outre ces arguments psychologiques ou sociologiques, une autre raison milite pour la création de facultés spécialement consacrées aux applications : le contenu de la mathématique a été si fortement élargi, qu'il est devenu très difficile de faire connaître aux étudiants en 5 années d'étude toutes les idées et méthodes nouvelles, même en se limitant à leurs aspects essentiels, a fortiori si l'on veut conserver tout ce qui était enseigné auparavant. Une partie des mathématiciens souhaite séparer complètement la cybernétique, le calcul numérique, etc. de la mathématique étudiée à présent et qu'on peut appeler classique.

Après de longs débats, il fut décidé d'ouvrir de nouvelles Facultés en 69-70 dans les Universités de Moscou et Leningrad, tandis qu'à Novosibirsk on décidait de n'avoir qu'une seule Faculté avec deux sections : mathématique et mathématiques pour ingénieurs. L'expérience n'a pas duré assez longtemps pour que l'on puisse dès maintenant juger les résultats obtenus. Tout est en pleine évolution.

L'idée de séparer les études mathématiques en deux branches distinctes n'est nullement acceptée par la totalité des mathématiciens de notre pays, et elle a des adversaires, partisans de l'unité de la mathématique et de ses applications. Cependant la nécessité des réformes ne fait de doute pour personne. La mathématique a changé et s'est agrandie. Les problèmes nouveaux, les nouvelles manières de poser les problèmes doivent avoir leur place dans l'enseignement universitaire. D'ailleurs, parallèlement à la création de nouvelles facultés, des changements commencent à s'opérer dans celles des anciennes Facultés mathématiques, où l'on trouve des professeurs qui sont des mathématiciens créateurs travaillant sur les problèmes contemporains : de nouveaux programmes, de nouveaux cours, de nouvelles chaires apparaissent tandis que les anciens cours changent à leur tour.

L'argument essentiel de ceux qui défendent l'unité de l'enseignement mathématique est que cette unité doit correspondre à celle de la Science elle-même. Si le développement de la mathématique et l'élargissement de ses applications se poursuivent à leur vitesse actuelle, les élèves de nos universités rencontreront des situations et des problèmes totalement nouveaux. Ils devront être capables de les formuler, d'en élaborer les solutions et de créer de nouvelles disciplines mathématiques. Ils ne pourront le faire que s'ils connaissent la mathématique dans toute son étendue. C'est pourquoi de nombreux mathématiciens soviétiques pensent qu'il n'est pas raisonnable de donner aux étudiants un enseignement trop étroit et ne considèrent pas comme un fait définitivement acquis la création de facultés très spécialisées. Ils veulent au contraire consacrer leurs efforts à une reconstruction complète de l'enseignement mathématique universitaire, qui ne peut cependant s'opérer que lentement.

Les Facultés mathématiques et physico-mathématiques doivent progressivement changer de visage par l'introduction de nouvelles disciplines, comme cela s'est déjà fait dans le passé pour l'analyse, l'algèbre, la théorie des ensembles, la topologie, etc.

Les programmes des diverses Universités sont assez proches l'un de l'autre dans leur ensemble, les différences ne portent que sur des détails et proviennent souvent des goûts personnels des professeurs.

Prenons l'exemple typique de l'université de Léningrad dont je suis un ancien élève. La Faculté de mathématique et mécanique comporte trois sections, auxquelles s'ajoute, sans autre raison que la tradition, une section d'Astronomie. Je me bornerai aux programmes de mathématiques. Il faut y distinguer ce qui est obligatoire pour tous les étudiants, indépendamment de leur spécialisation, et ce qui est optionnel et que chaque étudiant choisit selon la chaire à laquelle il est rattaché à partir de la $3^{\text{ème}}$ année.

La partie commune a pour but de donner aux étudiants un volume de connaissances suffisant pour qu'ils puissent se perfectionner au-delà par un travail individuel. Ce volume de connaissance doit être assez large sans être écrasant. Aucun professeur ne doit considérer le sujet qu'il traite comme une fin en soi, dont l'importance l'emporte sur toute autre, mais comme une partie d'un tout. En revanche, cette restriction ne doit pas l'empêcher de montrer aux étudiants les parties les plus vivantes du sujet et les idées créatrices qui y sont à l'œuvre.

Cette partie commune comprend :

(1) L'Analyse mathématique qui est enseignée pendant les 5 premiers semestres et qui contient le calcul différentiel et intégral classique, la théorie des séries de Fourier, les intégrales multiples et curvilignes, la théorie des fonctions d'une variable complexe.

En dehors des cours, des séances d'exercices, où l'on résoud des problèmes d'analyse et où on étudie certaines applications, s'adressent à des groupes de 25 étudiants et sont animées par des assistants.

L'esprit des cours a été longtemps défini par le livre de Fichtenholz qui est très populaire chez nous. Mais dans les dernières années, les cours furent profondément modernisés, on a beaucoup plus insisté sur les idées de l'analyse fonctionnelle et l'influence de N. Bourbaki et des cours d'analyse de J. Dieudonné sont perceptibles.

(2) L'Algèbre qui est enseignée pendant les trois premiers semestres et qui comprend : calcul matriciel, déterminants, division des polynômes sur la droite réelle et dans le plan complexe, éléments de théorie des groupes, espaces vectoriels de dimension finie (y compris la forme canonique de Jordan), éléments d'algèbre tensorielle. On trouve dans ce cours des notions de théorie des nombres (divisibilité entre autres).

(3) La géométrie, qui est enseignée pendant les quatre premiers semestres. Le premier semestre est consacré à une révision, essentiellement sous forme d'exercices, de géométrie analytique et d'algèbre vectorielle. Au second semestre est étudiée la géométrie différentielle des courbes et des surfaces, au troisième des éléments de topologie générale et de topologie algébrique, au quatrième les espaces riémanniens.

(4) Les équations différentielles ordinaires sont enseignées pendant les $3^{\text{ème}}$, $4^{\text{ème}}$ et $5^{\text{ème}}$ semestres. Outre l'étude traditionnelle de l'intégration des principaux types d'équations, on expose la théorie qualitative et la théorie analytique des équations différentielles, et quelques éléments de la théorie des fonctions spéciales.

(5) L'analyse fonctionnelle est enseignée pendant les $5^{\text{ème}}$ et $6^{\text{ème}}$ semestre : espaces métriques, espaces hilbertiens, (y compris la théorie spectrale des opérateurs bornés).

(6) Les équations de la physique mathématique ($6^{\text{ème}}$ et $7^{\text{ème}}$ semestre). Etude des problémes fondamentaux concernant les équations aux dérivées partielles du $2^{\text{ème}}$ ordre et leurs applications à la physique mathématique. L'exposé utilise largement les méthodes de l'analyse fonctionnelle, dont certains aspects sont développés à cette occasion.

(7) Le calcul des probabilités ($5^{\text{ème}}$ et $6^{\text{ème}}$ semestres) fait l'objet d'un cours peu étendu consacré à l'étude des variables aléatoires, des théorèmes limites et des notions fondamentales concernant les processus stochastiques.

(8) Le calcul numérique est partagé entre les deux premiers semestres d'une part, le $5^{\text{ème}}$ et le $6^{\text{ème}}$ d'autre part.

Le premier cours présente les éléments de la programmation pour ordinateur, la langue Algol 60 avec exercices de programmation de questions d'algèbre ou d'analyse.

Le second, d'un niveau plus élevé, traite des méthodes numériques et des programmes concernant des équations différentielles, le calcul des probabilités, etc. Les programmes sont passés réellement sur ordinateur.

Au cours de calcul numérique est associé un cours spécial relatif à la programmation optimale, aux éléments de théorie des jeux et questions voisines. Entre le $4^{\text{ème}}$ et le $5^{\text{ème}}$ semestre, les étudiants suivent deux petits cours de mécanique théorique et de physique, que l'on a placé ainsi pour que l'on puisse s'y servir de l'outillage mathématique approprié.

Les cours généraux obligatoires pour tous les étudiants sont pratiquement terminés à la fin du $7^{\text{ème}}$ semestre. Dès le $5^{\text{ème}}$ (et quelquefois plus tôt), les étudiants commencent à suivre les cours spéciaux qui sont organisés par les diverses chaires, et participent à des séminaires spécialisés, qu'ils choisissent à leur gré, et qu'ils doivent fréquenter pendant deux ans. Chaque étudiant subit trois examens, chaque examen sanctionnant le travail d'une année.

Les chaires organisent en général assez de cours spécialisés et de séminaires, qui permettent à l'étudiant d'approfondir ses connaissances dans le domaine choisi, lui donnent l'habitude du travail personnel à partir de la littérature spécialisée et l'entrainent à faire des exposés.

La $5^{\text{ème}}$ année ($9^{\text{ème}}$ et $10^{\text{ème}}$ semestre), est consacrée, outre l'assistance aux cours et séminaires, à la préparation d'un diplome. L'étudiant doit présenter une petite thèse, qui est, en général, une étude personnelle dont il n'est pas exceptionnel qu'elle soit publiée dans un périodique.

La Faculté a les chaires suivantes :

1/ Analyse numérique. 2/ Algèbre et théorie des nombres. 3/ Géométrie et topologie. 4/ Probabilités. 5/ Equations différentielles. 6/ Physique mathématique. 7/ Calcul numérique.

Il résulte de ce qui précède qu'à la Faculté mathématico-mécanique de Léningrad il n'y a pas opposition entre mathématique pure et mathématique appliquée. Tous les étudiants quelle que soit la chaire à laquelle ils sont rattachés reçoivent à la fois une formation théorique et une connaissance des applications.

La nouvelle Faculté de mathématique appliquée et théorie de la gestion n'est que très peu liée à la Faculté de mathématique et mécanique. Les plans d'étude en ont été élaborés par le groupe des mathématiciens du conseil scientifique de cette faculté et portent en partie la marque des goûts personnels de ceux qui l'ont fondée.

A Moscou, il existe maintenant deux Facultés : mécanico-mathématique d'une part, calcul numérique et cybernétique d'autre part. Elles sont indépendantes l'une de l'autre, et chacune s'efforce de résoudre suivant sa propre voie les problèmes de l'enseignement mathématique contemporain. Une grande partie des applications sont traitées à la Faculté mécanico-mathématique et la nouvelle Faculté ne prétend traiter ni toutes les applications, ni même la plupart d'entre elles : on y trouve la programmation pour ordinateurs, la théorie de la gestion optimale bien developpée et certaines questions de physique mathématique.

Comme je l'ai déjà dit à Novosibirsk il n'existe qu'une seule Faculté comportant deux sections. La section "mathématique pour ingénieurs" est issue de la section de mécanique de l'ancienne Faculté. Ses programmes ont été élargis par l'inclusion de disciplines nouvelles. Mais les plans d'étude demeurent très proches de ceux de la section mathématique. L'unité du style d'enseignement a été préservée, et les deux premières années sont communes aux deux sections.

L'analyse est enseignée pendant deux ans, la suite de ce cours est le cours d'analyse fonctionnelle, auparavant dénommé Analyse III.

Dans les deux sections, le calcul numérique n'est pas négligé. On l'étudie la première année, et on y revient à la quatrième année.

Les programmes d'algèbre en revanche, sont différents. Ils sont plus étendus et sont présentés plus tôt dans la section mathématique, tandis que les applications sont beaucoup plus riches dans la section mathématique pour ingénieurs, où l'étude de la mécanique (même celle des milieux continus) commence plus tôt. Une plus grande importance y est également accordée aux méthodes concrètes de calcul numérique appliquées à divers problèmes de mécanique et de physique mathématique.

Cependant, certaines des spécialités que peuvent choisir les étudiants se trouvent sous le même nom dans les deux sections : par exemple, le calcul numérique (avec une chaire dans chaque section) la cybernétique, l'aérodynamique. Les différences entre les sections dépendent en partie des professeurs qui dirigent les chaires correspondantes.

Certaines spécialisations ne se trouvent en revanche que dans une seule section. La section mathématique possède des chaires d'analyse fonctionnelle, d'algèbre, de géométrie, d'équations différentielles, qui n'existent pas dans l'autre section, qui est inversement seule à posséder des chaires d'hydrodynamique, de théorie de l'élasticité, de géophysique mathématique, de programmation pour ordinateurs.

Je n'ai bien entendu esquissé dans ma conférence que les traits principaux et les tendances les plus importantes de l'enseignement mathématique en URSS. Mon but n'était pas de décrire l'état actuel, mais de montrer la situation dans son dynamisme. L'enseignement mathématique en URSS est actuellement en pleine évolution. Il y a des tentatives différentes et on cherche dans des directions variées. C'est l'esprit de notre temps.

Je dois remercier ici plusieurs de mes collègues qui m'ont aider à préparer cet exposé, et en premier lieu les Professeurs de l'Université de Léningrad D.K. Faddeev et S.V. Vallander, sans oublier d'exprimer ma profonde reconnaissance à tous ceux de mes collègues qui m'ont fait de précieuses remarques.

Institut des Mathématiques
Novosibirsk 90
URSS

TABLE DES MATIÈRES DU TOME 3

E - MATHÉMATIQUES APPLIQUÉES 1

E_1 – ASPECTS MATHEMATIQUES DE LA THEORIE QUANTIQUE DES CHAMPS

E_2 – THEORIE DE LA RELATIVITE

E_3 – PROBLEMES MATHEMATIQUES DE LA MECANIQUE DU CONTINU

E$_4$ – THEORIE DU CONTROLE OPTIMAL

E$_5$ – COMBINATORIQUE – ALGEBRE FINIE

E$_6$ – STATISTIQUE MATHEMATIQUE

E_7 – PROBLEMES MATHEMATIQUES DE LA THEORIE DE L'INFORMATION, LANGAGE MACHINE

E_8 – ANALYSE NUMERIQUE

F_1 – HISTOIRE DES MATHEMATIQUES

F_2 – ENSEIGNEMENT DES MATHEMATIQUES

Imprimerie LOUIS JEAN
Dépôt légal : 1847

Dépôt légal n°323-1971
N° Dossier : 70.30

Imprimé en France

Collection
Monographies internationales de Mathématiques modernes

Collection publiée sous la direction de S. Mandelbrojt, professeur au Collège de France. 11 volumes 16 x 25.

Derniers ouvrages parus :

LINNIK Y. V. **Décompositions des lois de probabilités.**
Traduit par *M.L. Gruel.* 300 pages, 1962, 58 F.

MIKUSINSKI J. et SIKORSKI R.
Théorie élémentaire des distributions.
Traduit par *S. Klarsfeld.* 114 pages, 1964, 21 F.

GELFAND I. M., RAIKOV D. A. et CHILOV G. E.
Les anneaux normés commutatifs.
Avec un appendice par *J.-P. Kahane* et *P. Malliavin.* Traduit par *J.-L.* et *M. Verley.* 259 pages, 1964, 55 F.

GELFOND A. O. et LINNIK Y. V.
Méthodes élémentaires dans la théorie analytique des nombres.
Traduit de l'anglais par *M.* et *J.-L. Verley.* 240 pages, figures, 1965, 48 F.

MITROPOLSKI Y. A. **Problèmes de la théorie asymptotique des oscillations non stationnaires.**
Avant-propos de *N. N. Bogolioubov.* Traduit par *G. Carvallo.* 547 pages, 117 figures, 1966, 106 F.

BOREVITCH Z. I. et CHAFAREVITCH I. R.
Théorie des nombres.
Traduit par *M.* et *J.-L. Verley.* 498 pages, figures tableaux, 1967, 83 F.

ARNOLD V. J. et AVEZ A.
Problèmes ergodiques de la mécanique classique.
243 pages, figures, 1967, 60 F.

LINNIK Y. V. **Leçons sur les problèmes de statistique analytique.**
126 pages, 1967, 25 F.

MANDELBROJT S. **Séries de Dirichlet.**
176 pages, 1969, 46 F.

GAUTHIER-VILLARS, Editeur, PARIS

Collection Cahiers Scientifiques

Collection publiée sous la direction de G. Julia, professeur à la Faculté des Sciences de Paris, membre de l'Institut. 36 volumes 16 x 25.

ARTIN E.

Algèbre géométrique.
Traduit par *M. Lazard.* 224 pages, 6 figures, nouveau tirage, 1967, 42 F.

DIEUDONNE J.

Eléments d'analyse. Tome I : Fondements de l'analyse moderne.
Traduit de l'anglais par *D. Huet.* Préface de *G. Julia*, nouvelle édition revue et corrigée. 417 pages, 1969, 66 F.

DIEUDONNE J.

Eléments d'analyse. Tome II.
420 pages, 1968, 66 F.

DIEUDONNE J.

Eléments d'analyse. Tome III.
392 pages, 1970, 66 F.

DIEUDONNE J.

Eléments d'analyse. Tome IV.
430 pages, 1971, 35 F.

DIXMIER J.

Les C*-algèbres et leurs représentations.
2^e édition. 407 pages, 1969, 80 F.

DIXMIER J.

Cours de mathématiques du premier cycle, première année.
480 pages, 149 figures, nouveau tirage, 1969, 42 F.

DIXMIER J.

Cours de mathématiques du premier cycle, deuxième année.
371 pages, 59 figures, nouveau tirage, 1969, 32 F.

McLANE S. et BIRKHOFF G.

Algèbre. Tome I.
436 pages, 1970, 49 F.

McLANE S. et BIRKOFF G.

Algèbre. Tome II.
358 pages, 1971, 49 F.

GAUTHIER-VILLARS, Editeur, PARIS

Œuvres de JORDAN

Œuvres publiées sous la direction de G. Julia par J. Dieudonné et R. Garnier. 4 volumes 16 x 25 se vendant séparément.

Les tomes I et II sont consacrés à la théorie des groupes.
Dans les « Notes sur les travaux de C. JORDAN relatifs à la théorie des groupes », qui sert d'introduction au premier volume, J. DIEUDONNE classe les travaux de JORDAN sous plusieurs rubriques : I. Généralités sur les groupes ; applications aux équations algébriques (Théorie de Galois). - II. Groupes linéaires sur un corps premier fini. - III. Sous-groupes de GLn (C.). - IV. Groupes transitifs et groupes primitifs.

Le tome III traite de l'algèbre linéaire et multilinéaire et de la théorie des nombres ; il est aussi présenté par J. DIEUDONNE qui répartit ainsi les articles : I. Géométrie à *n* dimensions. - II. Formes bilinéaires et formes quadratiques. - III. Théorie des invariants. - IV. Equivalence arithmétique des formes. - V. Autres travaux de théorie des nombres.

Le quatrième et dernier volume réunit des articles dispersés dans diverses revues sur des questions d'analyse relatives aux équations différentielles et à la topologie des polyèdres, ainsi que quelques travaux intéressant la mécanique, sur l'équilibre et sa stabilité et les petites oscillations autour d'une position d'équilibre. La mise au point des textes a été confiée à R. GARNIER et à J. DIEUDONNE. On y trouvera aussi, en tête du volume, deux portraits de JORDAN et sa biographie, composée par H. LEBESGUE lorsqu'il succéda à JORDAN à l'Académie des Sciences ; la fin du volume est consacrée à des extraits de lettres écrites à JORDAN par A. CLEBSCH, S. LIE, L. CREMONA, L. SYLOW, ainsi qu'aux principaux discours prononcés par JORDAN.

Tome I :	548 pages, 1961,	Broché :	95 F
		Cartonné :	110 F
Tome II :	562 pages, 1961,	Broché :	95 F
		Cartonné :	110 F
Tome III :	574 pages, 1962,	Broché :	95 F
		Cartonné :	110 F
Tome IV :	644 pages, 1964,	Broché :	125 F
		Cartonné :	140 F

GAUTHIER-VILLARS, Editeur, PARIS

Œuvres de H. POINCARÉ

Œuvres publiées sous les auspices de l'Académie des Sciences (Section de Géométrie), par G. Darboux, 11 volumes 23 x 28 se vendant séparément.

Tome I : *Analyse.* Publié sous les auspices de l'Académie des Sciences, par *P. Appell,* avec la collaboration de *J. Drach,* 382 pages, nouveau tirage, 1952.
Broché : 70 F — Cartonné : 85 F

Tome II : *Analyse.* Publié avec la collaboration de *N.-E. Norlund* et *E. Lebon,* 705 pages, 19 figures, un portrait de H. Poincaré, nouveau tirage, 1952.
Broché : 105 F — Cartonné :120 F

Tome III : *Analyse.* Publié avec la collaboration de *J. Drach,* 596 pages, nouveau tirage,1965.
Broché : 105 F — Cartonné : 120 F

Tome IV : *Analyse.* Publié avec la collaboration de *G. Valiron,* 632 pages, 1950.
Broché : 105 F — Cartonné : 120 F

Tome V : *Arithmétique et algèbre.* Publié avec la collaboration de *M. Chatelet,* 560 pages, 1950.
Broché : 105 F — Cartonné : 120 F

Tome VI : *Géométrie.* Publié avec la collaboration de *R. Garnier* et *J. Leray,* 548 pages, 1953.
Broché : 105 F — Cartonné : 120 F

Tome VII : *Mécanique céleste et astronomie.* Publié avec la collaboration de *J. Lévy,* 634 pages, 1952.
Broché : 105 F — Cartonné : 120 F

Tome VIII : *Mécanique céleste et astronomie.* Publié avec la collaboration de *P. Sémirot,* 695 pages, 1952.
Broché : 105 F — Cartonné : 120 F

Tome IX : *Physique mathématique.* Publié avec la collaboration de *G. Petiau,* 720 pages, 1954.
Broché : 105 F — Cartonné :120 F

Tome X : *Physique mathématique.* Publié avec la collaboration de *G. Petiau.* Préface de *G. Julia,* 642 pages, 1954.
Broché : 105 F — Cartonné : 120 F

Tome XI : Publié avec la collaboration de *G. Petiau.* Mémoires divers : *Hommage à H. Poincaré. Livre du Centenaire de H. Poincaré, 1854-1954,* 661 pages, 1956.
Broché : 105 F — Cartonné : 120 F

GAUTHIER-VILLARS, Editeur, PARIS

Œuvres de GALOIS

Ecrits et mémoires mathématiques. Edition critique intégrale de ses manuscrits et publications. Publiés par les soins de R. Bourgne et J.-P. Azra, avec le concours du C.N.R.S. Préface de J. Dieudonné.

Un volume 21 × 27, 561 pages, 1962 : 120 F.

Œuvres de CARTAN

Œuvres complètes publiées avec le concours du C.N.R.S. 6 volumes 16 x 25.

Première partie : *Groupes de Lie.* Deux volumes formant 1 356 pages, se vendant ensemble, avec un portrait, 1952.

Brochés : 150 F	Cartonnés : 170 F

Deuxième partie : Vol. I : *Algèbre, formes différentielles, systèmes différentiels.*
Vol. II : *Groupes finis, systèmes différentiels, théories d'équivalence.*

Ces deux volumes forment 1 384 pages et se vendent ensemble, 1953.

Brochés : 150 F	Cartonnés : 170 F

Troisième partie : Vol. I : *Divers, géométrie différentielle.*
Vol. II : *Géométrie différentielle* (fin).

Ces deux volumes forment 1 877 pages et se vendent ensemble, 1955.

Brochés : 175 F	Cartonnés : 195 F

GAUTHIER-VILLARS, Editeur, PARIS

Œuvres de LAGRANGE

Œuvres complètes, publiées par les soins de J.-A. Serret et G. Darboux. 14 volumes 23 x 28.

La première série comprend tous les mémoires imprimés dans les *Recueils des Académies de Turin, de Berlin et de Paris,* ainsi que les *Pièces diverses* publiées séparément. Cette série forme 7 volumes (Tomes I à VII, 1867-1877), qui se vendent séparément.

Chaque volume : 120 F.

La deuxième série se compose de 7 volumes qui renferment les ouvrages didactiques, la correspondance et les mémoires inédits :

Tomes VIII et IX : épuisés.

Tome X : *Leçons sur le calcul des fonctions,* 1884.

Tomes XI et XII : épuisés.

Tome XIII : *Correspondance inédite de Lagrange et de d'Alembert,* publiée d'après les manuscrits autographes et annotée par *Ludovic Lalanne,* 1822.

Tome XIV : *Correspondance de Lagrange avec Condorcet, Laplace, Euler et divers savants,* publiée et annotée par *Ludovic Lalanne,* avec deux fac-similés, 1892.

Chaque volume : 120 F.

GAUTHIER-VILLARS, Editeur, PARIS