ТРУДЫ МЕЖДУНАРОДНОГО КОНГРЕССА МАТЕМАТИКОВ

(Москва — 1966)

❊

PROCEEDINGS OF INTERNATIONAL CONGRESS OF MATHEMATICIANS

(Moscow — 1966)

❊

TRAVAUX DU CONGRÈS INTERNATIONAL DES MATHÉMATICIENS

(Moscou — 1966)

❊

BERICHTE DES INTERNATIONALEN MATHEMATIKER-KONGRESSES

(Moskau — 1966)

ИЗДАТЕЛЬСТВО «МИР»
МОСКВА 1968

ОРГАНИЗАЦИОННЫЙ КОМИТЕТ:

И. Г. Петровский — председатель
С. Н. Мергелян — зам. председателя
В. Г. Карманов — генеральный секретарь

ORGANIZING COMMITTEE:

I. G. Petrovsky — Chairman
S. N. Mergelyan — Vice-chairman
V. G. Karmanov — Secretary General

Под редакцией И. Г. Петровского

Edited by I. G. Petrovsky

ВВОДНАЯ ЧАСТЬ

*

INTRODUCTION

*

INTRODUCTION

*

EINLEITUNG

Международный конгресс математиков в Москве, как и предшествующие конгрессы, получил организационную и финансовую помощь со стороны Международного математического союза, ИКСУ и ЮНЕСКО. Эта помощь распространялась как на финансовую поддержку приглашенных докладчиков и некоторых молодых математиков—делегатов Конгресса, так и на подготовку к публикации трудов Конгресса.

The International Congress of Mathematicians in Moscow, as the previous congresses, received an organizational and financial support from the International Mathematical Union and from ICSU and UNESCO. This help consisted from the financial support to the invited speakers, to some young mathematicians—delegates of the Congress and from the help to the preparation of the printing of the Proceedings of the Congress.

РЕЧЬ ПРЕЗИДЕНТА АКАДЕМИИ НАУК СССР АКАДЕМИКА М. В. К Е Л Д Ы Ш А НА ОТКРЫТИИ КОНГРЕССА

Уважаемые коллеги!
Дамы и господа! Товарищи!

Мне доставляет большое удовольствие от имени Академии наук Советского Союза приветствовать всех участников Конгресса и передать пожелание успешной работы Конгрессу.

Математика, являющаяся самой древней из всех наук, вместе с тем остается вечно молодой, бурно развивающейся наукой, все время расширяющей области своего познания, все шире развивающей свои связи не только с естественными науками, но и с самыми разнообразными областями человеческой деятельности. Я думаю, что ценность математических теорий тем выше, чем теснее их корни связаны с явлениями мира, в котором мы живём, и вместе с тем, чем выше мы достигаем степени абстракции и общности точек зрения. Успех теории во многом зависит от того, находим ли мы адекватную изучаемому явлению степень общности и степень абстракции. Ценность теории определяется тем, насколько общие положения позволяют понимать конкретные явления и решать конкретные задачи. Общие математические теории позволяют нам глубоко понять взаимосвязи явлений. Внедрение математических методов преобразует области знания и не только ставит их на высшую ступень логического мышления, но открывает новые возможности, новые постановки задач, позволяет по-новому смотреть на явления. Достаточно вспомнить, какие революционные, принципиальные сдвиги в развитии естествознания дали анализ бесконечно малых, теория вероятностей, теория операторов и, наконец, в настоящее время бурно развивающееся познание логических процессов.

Развитие таких абстрактных областей математики, как теория множеств и топология, алгебра, функциональный анализ и др., недавно возникшие, привело не только к созданию теорий необыкновенной красоты, но и к созданию новых мощных методов во всей математике. Мне кажется, что мы переживаем эпоху, когда математический метод особенно стремительно завоевывает все новые области знания. Наряду с физикой дух математического мышления все большее значение приобретает в химии, биологии, геологии, широко проникает в общественные науки, и в первую очередь в экономическую науку. Изучение основ логических процессов и теории операций, методы дискретной математики, создание электронно-вычислительных устройств подготовили основы для новой

величайшей научно-технической революции во всей жизни человечества, новой ступени понимания многих процессов в природе и жизни и новой ступени в автоматизации процессов, которые до недавнего времени считались областью исключительно интеллектуальной деятельности человека, а также в реализации математических процессов, которые мы считали осуществимыми только в абстрактном мышлении.

Позвольте выразить надежду, что предстоящий Конгресс будет иметь большое значение в математической жизни. Область математики стала настолько широкой, что математики говорят не только на языках разных народов, но и на разных математических языках и их язык пока недоступен многим ученым других специальностей, но именно вследствие широты и силы математического метода Конгресс будет иметь большое значение для всей науки и развития человеческой культуры.

Позвольте открыть Международный конгресс математиков.

РЕЧЬ ПРЕЗИДЕНТА КОНГРЕССА АКАДЕМИКА И. Г. ПЕТРОВСКОГО НА ОТКРЫТИИ КОНГРЕССА

Уважаемые члены Конгресса!

Благодарю вас за оказанную мне честь. Позвольте от имени советского Организационного комитета приветствовать вас и пожелать успехов в работе.

При подготовке Конгресса Организационному комитету большую помощь оказал Консультативный комитет Международного математического союза. Все его предложения о количестве секций, о часовых и получасовых докладах были советским Организационным комитетом приняты полностью. Мы позволили себе только к рекомендациям Комитета добавить несколько докладов.

Я хочу от имени Организационного комитета выразить благодарность Исполкому Международного математического союза и избранному им Консультативному комитету под председательством профессора Р. Неванлинны. Мне хочется особенно поблагодарить профессора Р. Неванлинну.

А теперь позвольте предоставить слово профессору де Раму для сообщения о решениях Комитета по присуждению Филдсовских премий.

REPORT OF PROFESSOR G. DE RHAM, CHAIRMAN OF THE FIELDS MEDALS COMMITTEE, AT THE OPENING CEREMONY OF THE CONGRESS

Ladies and Gentlemen!

Professor J. C. Fields, President of the International Congress of Mathematicians held in Toronto in 1924, proposed that two gold medals be awarded at each International Congress, for outstanding achievements in Mathematics. He set up a fund for that purpose, from out of the balance left over at the end of the Toronto Congress. In 1932, after his death, the International Congress held at Zürich decided to accept his proposal with thanks. As is well known, two medals have been presented at every Congress since held: Oslo 1936, Harvard 1950, Amsterdam 1954, Edinburgh 1958 and Stockholm 1962.

Following a tradition which has become well established, the Executive Committee of the International Mathematical Union appoints, in advance of the Congress, a special Committee to select the candidates. This time the Committee consists of Professors H. Davenport, M. Deuring, W. Feller, M. A. Lavrentiev, J. P. Serre, D. C. Spencer, R. Thom, and I have been asked to be the chairman. Every one of the members has taken an active part in the deliberations. We have also consulted other experts. I thank them all for their valuable contribution.

The Memorandum of Fields says: "Because of the multiplicity of the branches of Mathematics and taking into account the fact that the interval between such Congresses is four years, it is felt that at least two medals should be available." In view of the vast development of Mathematics during the last forty years, it appears that this number could judiciously be increased to four. The Executive Committee of the International Mathematical Union has therefore viewed with sympathy the generous offer made by an anonymous donor to give this year two more medals. The Organizing Committee of this Congress having agreed to this and the Medals Committee having accepted the responsibility to select four names, four medals will be awarded today. The medals have been struck by the Royal Mint in Ottawa. The name of the recipient is engraved on each of them. The name of Fields does not figure on them. Fields himself proposed to call them: "International Medals for outstanding discoveries in Mathematics". Each of them carries with it a cash prize which, this year, amounts to 1,500 Canadian dollars.

The Memorandum of Fields also contains the following: "In coming to its decision, the hands of the International Committee should be

left as free as possible. It would be understood, however, that in making the awards, while it was in recognition of work already done, it was at the same time intended to be an encouragement for further achievements on the part of the recipients and a stimulus to renewed efforts on the part of the others."

On the basis of this text, and following precedent, we confine our choice to candidates under forty. We prepare a first list of about 30 names. We then looked for those whose work appeared to us the most important and the most striking, irrespective of any other consideration, setting aside any question of nationality. To our regret, we have had to give up several names which would have also deserved this distinction. Several young mathematicians of extraordinary brilliance were among them. But because they are so young, there will be many Congresses before they reach forty and if they continue in their course, they will have every chance of receiving a medal. The choice was thus not easy. Nevertheless, after serious consideration and reflexion, we arrive at a conclusion without undue difficulty. The following four names, in alphabetical order, constitute our choice:

Michael Francis Atiyah,
Paul J. Cohen,
Alexander Grothendieck,
Stephen Smale.

Unfortunately, A. Grothendieck, was unable to come. May I call Messrs. Atiyah, Cohen and Smale to come forward and receive these medals from the hands of Academician Keldysh. A brief account of their achievements will be given by noted mathematicians.

L'OEUVRE DE MICHAEL F. ATIYAH

HENRI CARTAN

Je parlerai très brièvement des travaux d'Atiyah dans trois domaines, d'ailleurs étroitement reliés entre eux : la K-théorie, la formule de l'indice, et la « formule de Lefschetz ». Je laisserai de côté d'autres contributions, fort intéressantes d'ailleurs, à la Géométrie algébrique ou à la théorie du cobordisme ; et je passerai aussi sous silence les résultats tout récents, encore inédits, dont l'auteur parlera lui-même dans sa conférence pendant ce Congrès.

1. La K-théorie. La plupart des travaux d'Atiyah en K-théorie ont été faits en collaboration avec F. Hirzebruch. C'est en 1956 que paraissait l'ouvrage fondamental de Hirzebruch (« Neue topologische Methoden in der algebraischen Geometrie ») dont le but ultime était le théorème fameux qui porte aujourd'hui le nom de « théorème de Riemann-Roch-Hirzebruch ». Il s'agissait de géométrie algébrique sur le corps complexe. Peu après, Grothendieck cherchait et obtenait une démonstration purement algébrique (valable sur tout corps de base algébriquement clos, de caractéristique quelconque) d'un théorème plus général [1], puisqu'au lieu de considérer une variété algébrique X il étudiait un morphisme $X \to Y$ (le cas traité par Hirzebruch étant celui où la variété algébrique Y est réduite à un point). C'est à cette occasion que Grothendieck introduisit un foncteur contravariant qui, à chaque variété algébrique X, associe un *anneau* construit à l'aide des classes d'isomorphie de fibrés vectoriels algébriques de base X. Atiyah et Hirzebruch [2] eurent l'idée de faire de même pour un espace topologique *compact* X et pour les classes de fibrés vectoriels *complexes* de base X (il s'agit de fibrés topologiques, localement triviaux). On définit ainsi un anneau $K(X)$ pour tout espace compact X, d'où le nom de K-théorie. Il y a aussi une KO-théorie pour les fibrés vectoriels *réels*, et une KSp-théorie pour les fibrés vectoriels quaternioniens.

Bornons-nous, pour simplifier, à la K-théorie. On définit des groupes relatifs $K(X, Y)$ (pour Y sous-espace fermé de X), puis, par suspension, des groupes $K^n(X, Y)$ pour n entier $\leqslant 0$, avec $K^0(X, Y) = K(X, Y)$. On a alors une suite exacte

$$\ldots \to K^n(X, Y) \to K^n(X) \to K^n(Y) \to K^{n+1}(X, Y) \to \ldots$$

analogue à la suite exacte de cohomologie. D'autre part, Atiyah observe que le célèbre théorème de périodicité de Bott [qui concerne

les groupes d'homotopie du groupe unitaire infini $U = \lim_{\rightarrow m} U(m)$ peut s'exprimer par un isomorphisme explicite

$$K^n(X) \approx K^{n+2}(X),$$

ce qui permet de définir le foncteur K^n aussi pour n entier >0. De cette façon on obtient une « théorie cohomologique » au sens d'Eilenberg-Steenrod, à cela près qu'un des axiomes d'Eilenberg-Steenrod (l'axiome « de dimension ») n'est pas vérifié. Cette théorie fut d'abord baptisée « cohomologie extraordinaire ».

Si on veut comparer la cohomologie extraordinaire à la cohomologie ordinaire, on peut dire en gros ceci : au lieu de considérer, comme en cohomologie ordinaire, les classes d'homotopie d'applications d'un espace X dans les espaces d'Eilenberg-MacLane $K(\pi, n)$, on envisage, dans la K-théorie, le groupe unitaire infini U (ou, ce qui revient au même, le groupe linéaire complexe infini), et son espace classifiant BU ; ce sont eux qui servent d'espaces de comparaison. Les relations existant entre les deux théories cohomologiques (ordinaire et extraordinaire) s'expriment par une suite spectrale, et le « caractère de Chern »

$$\text{ch: } K^*(X, Y) \longrightarrow H^*(X, Y; \mathbf{Q})$$

est un homomorphisme multiplicatif d'une théorie dans l'autre.

L'importance de la « cohomologie extraordinaire » fut vite mise en évidence par les applications qu'Atiyah et Hirzebruch en firent, en Topologie algébrique et ailleurs [3]. Citons quelques exemples qui illustrent ces applications de la K-théorie :

— un théorème du type « Riemann-Roch-Grothendieck », valable cette fois pour les variétés différentiables [4] ;

— le calcul de $K(X)$ pour certains espaces homogènes, et le lien de cette question avec la théorie des représentations des groupes de Lie compacts [2] ;

— des théorèmes de non-plongement [5] : par exemple, l'espace projectif complexe $P_n(\mathbf{C})$ ne peut pas être différentiablement plongé dans l'espace numérique $\mathbf{R}^{4n-2\alpha(n)}$, où $\alpha(n)$ désigne le nombre des chiffres 1 du développement dyadique de l'entier n ;

— des critères permettant de reconnaître si une classe de cohomologie d'une variété analytique complexe compacte peut être représentée par une sous-variété analytique [6].

Toutes ces applications sont dues à la collaboration d'Atiyah avec Hirzebruch. Il y en a d'autres ; par exemple, c'est grâce à la K-théorie et à l'introduction de certains foncteurs $K \rightarrow K$ (dont l'idée revient essentiellement à Grothendieck) que J. F. Adams [7] a pu résoudre complètement un problème classique, resté longtemps sans réponse : celui de la détermination exacte, en fonction de

l'entier n, du nombre maximum de champs de vecteurs linéairement indépendants sur la sphère S^n (voir la conférence d'Adams au Congrès de Stockholm en 1962).

2. Le théorème de l'indice. Mais la plus belle application de la K-théorie devait être faite par Atiyah lui-même : je veux parler du *théorème de l'indice* (1963), démontré en collaboration avec I. Singer [8].

Soit D un opérateur elliptique sur une variété différentiable compacte X (supposée sans bord), opérant de l'espace vectoriel $\Gamma(E)$ des sections différentiables d'un fibré vectoriel complexe E dans l'espace $\Gamma(F)$ des sections différentiables d'un fibré vectoriel complexe F. On sait que le noyau et le conoyau de l'application linéaire $D: \Gamma(E) \rightarrow \Gamma(F)$ sont de dimension finie ; l'*indice* $i(D)$ est l'entier défini par

$$i(D) = \dim(\text{Ker } D) - \dim(\text{Coker } D).$$

Les travaux de plusieurs mathématiciens soviétiques avaient mis en évidence le fait que $i(D)$ ne change pas quand D varie d'une façon continue, et I. M. Gelfand, en 1960 [9], avait conjecturé que $i(D)$ devait donc pouvoir être calculé au moyen d'invariants purement topologiques liés à la donnée de X et de D. C'est ce problème qu'Atiyah et Singer ont complètement résolu. Les termes homogènes de plus haut degré de l'opérateur D définissent un « symbole » $\sigma(D)$ qui permet d'abord de définir l'ellipticité de D, puis, par l'intervention de la K-théorie, du caractère de Chern, et de la classe de Todd du fibré cotangent à X (lequel a une structure presque complexe), de définir finalement une classe de cohomologie, élément de $H^*(X ; \mathbf{Q})$. Sa composante de degré $n = \dim X$ est un élément de $H^n(X ; \mathbf{Q}) \approx \mathbf{Q}$ (on suppose X orientable, pour simplifier). D'où un *nombre rationnel* $i_t(D)$ attaché à D (et à X), et défini au signe près ; on peut l'appeler l'« indice topologique » de D. Le théorème d'Atiyah-Singer dit alors que *l'indice topologique $i_t(D)$ est égal à l'indice $i(D)$* (moyennant des conventions convenables d'orientation). Ce théorème établit ainsi un pont entre deux vastes domaines des mathématiques : l'analyse des équations aux dérivées partielles d'une part, la topologie algébrique d'autre part.

Observons que, par définition, $i(D)$ est un entier. Il s'ensuit que le nombre rationnel $i_t(D)$ fourni par la Topologie algébrique est, en fait, un entier. On obtient par ce moyen, d'une façon naturelle et par le choix d'opérateurs elliptiques appropriés, tous les « théorèmes d'intégralité » relatifs aux classes caractéristiques des variétés (intégralité du L-genre, du genre de Todd, du $\hat{A}$-genre). Inversement, toute information fournie par la Topologie algébrique donne un résultat qui intéresse l'Analyse ; par exemple, on voit facilement que

l'indice topologique $i_t(D)$ est nul si la variété X est de dimension impaire.

La démonstration du théorème $i(D) = i_t(D)$ est laborieuse, mais fort intéressante, car elle conduit à introduire des opérateurs plus généraux que les opérateurs différentiels, à savoir les opérateurs intégraux singuliers de Calderon-Zygmund et de Seeley. La démonstration repose également sur une théorie du cobordisme qui constitue une généralisation (relativement facile) de celle due à Thom. En fait il existe une nouvelle démonstration, plus récente, de la formule de l'indice, qui évite le recours au cobordisme.

Au lieu de considérer un seul opérateur elliptique, on peut envisager une suite d'opérateurs différentiels

$$(D) \qquad \Gamma(E_0) \to \Gamma(E_1) \to \ldots \to \Gamma(E_k)$$

formant un « complexe » (i.e : le composé de deux opérateurs consécutifs est zéro). On définit l'« ellipticité » d'un tel complexe. A chaque complexe elliptique on attache encore un nombre rationnel $i_t(D)$. D'autre part les groupes d'homologie du complexe elliptique (D) sont des espaces vectoriels de dimension finie ; soit $\chi(D)$ la somme alternée de leurs dimensions (c'est une sorte d'invariant d'Euler-Poincaré). Alors on a le théorème :

$$\chi(D) = i_t(D).$$

Cette forme plus générale du théorème de l'indice est fort utile dans les applications. Par exemple, si on l'applique à une variété analytique complexe compacte X, et au « complexe » défini par l'opérateur différentiel d'' (noté aussi $\bar{\partial}$) des formes différentielles, on retrouve exactement l'énoncé du théorème de Riemann-Roch-Hirzebruch. Ce dernier n'était démontré auparavant que pour les variétés algébriques sans singularité ; il est désormais valable pour toute variété analytique compacte.

Je laisse de côté le théorème de l'indice pour les variétés à bord [10] ; il nécessite une nouvelle définition de l'ellipticité qui tienne compte des « conditions aux limites ». La question a été entièrement résolue par Atiyah en collaboration avec Bott et Singer.

3. Formules de points fixes. Le théorème de l'indice n'est, en réalité, qu'un cas extrême d'une situation dont un autre cas extrême est, lorsque le complexe elliptique est celui défini par l'opérateur de différentiation extérieure des formes différentielles, la formule de Lefschetz relative aux points fixes (supposés isolés) d'une transformation d'une variété compacte X en elle-même. Il y a de nombreux cas intermédiaires, dont l'étude est en cours. Les résultats déjà obtenus sont dûs à la collaboration d'Atiyah et de Bott [11]. Expliquons sur un exemple de quoi il s'agit : soit X une variété analytique complexe,

compacte, et soit $f : X \to X$ une application holomorphe ; on sait que les espaces vectoriels de cohomologie $H^q(X, \mathcal{O})$ à coefficients dans le faisceau $\mathcal{O}$ des fonctions holomorphes sont de dimension finie; soit $L(f)$ la somme alternée des traces

$$(-1)^q \operatorname{Tr}(f|_{H^q(X,\mathcal{O})}).$$

C'est un entier ; dans le cas où f est l'identité, cet entier n'est autre que le premier membre de l'égalité de Riemann-Roch-Hirzebruch. Dans le cas général, on se propose d'exprimer cet entier à l'aide des propriétés topologiques de f au voisinage de l'ensemble des points fixes de f. Si f est l'identité, on peut considérer que la formule de Hirzebruch (démontrée par Atiyah-Singer) résout le problème. Supposons au contraire que f n'ait qu'un nombre fini de points fixes P, et que la différentielle df_P n'admette pas la valeur propre 1 (c'est notamment le cas lorsque f est une transformation d'ordre fini). Alors le déterminant

$$\det_C(1 - df_P)$$

est un nombre complexe $\neq 0$; le théorème prouvé par Atiyah et Bott affirme que, sous ces hypothèses, l'entier $L(f)$ est égal à la somme des inverses de ces nombres complexes.

Nous bornant à cet exemple, nous ajouterons simplement que les résultats déjà obtenus fournissent une démonstration « sans calculs » de la formule de H. Weyl donnant le caractère d'une représentation d'un groupe semi-simple, et qu'ils permettent aussi de résoudre des problèmes de Conner-Floyd sur les variétés compactes où opère un groupe fini. Signalons aussi que, d'après Hirzebruch [12], on peut en déduire une formule de Langlands donnant la dimension des espaces vectoriels de formes automorphes pour un groupe discret à quotient compact.

En conclusion, l'on doit à Michael Atiyah plusieurs contributions majeures qui mettent en relation étroite la Topologie et l'Analyse. Chacune d'elles a été réalisée en collaboration ; sans diminuer en rien la part qui revient à des collaborateurs aussi éminents que Hirzebruch, Singer ou Bott, il ne fait aucun doute que dans chaque cas l'intervention personnelle d'Atiyah a été décisive. Il nous donne l'exemple d'un mathématicien chez qui la clarté des conceptions et la vision d'ensemble des phénomènes s'allie harmonieusement à l'imagination créatrice, et aussi à la persévérance qui conduit aux grands achèvements.

RÉFÉRENCES

[1] Borel A., Serre J. P., Le théorème de Riemann-Roch, *Bull. Soc. Math. France*, 86 (1958), 97-136.

[2] Atiyah M. F., Hirzebruch F., Vector bundles and homogeneous spaces, Symp. Pure Math., n° 3, A.M.S., 1961.

[3] Atiyah M. F., The Grothendieck ring in Geometry and Topology, Proc. Int. Congress Math., Stockholm (1962), 442-446.
[4] Atiyah M. F., Hirzebruch F., Riemann-Roch theorems for differentiable manifolds, *Bull. A.M.S.*, **65** (1959), 276-281.
[5] Atiyah M. F., Hirzebruch F., Quelques théorèmes de non plongement pour les variétés différentiables, *Bull. Soc. Math. France*, **89** (1959), 383-396.
[6] Atiyah M. F., Hirzebruch F., Analytic cycles on complex manifolds, *Topology*, **1** (1962), 25-46.
[7] Adams J. F., Vector fields on spheres, *Ann. Math.*, **75** (1962), 603-632.
[8] Atiyah M. F., Singer I., The index of elliptic operators on compact manifolds, *Bull. A.M.S.*, **69** (1963), 422-433. Voir aussi deux Seminaires tenus en 1963-64, l'un par S. Palais (Annals of Math. Studies, n° 57, Princeton Univ. Press, 1965), l'autre par H. Cartan et L. Schwartz (Secrétariat Math., Inst. H. Poincaré, Paris 1965).
[9] Гельфанд И. М., Об эллиптических уравнениях, *УМН*, **15**, №3(1960), 121-132. English translation: Gelfand I. M., On elliptic equations, Russian Math. Surveys, **15**, № 3, (1960), 113.
[10] Atiyah M. F., Bott R., The index problem for manifolds with boundary, Differential Analysis, Bombay, 1964.
[11] Atiyah M. F., Bott R., Report on the Woods Hole Fixed Point Theorem, Seminar (1964).
[12] Hirzebruch F., Elliptische Differentialoperatoren auf Mannigfaltigkeiten, Festschrift Weierstrass, Westdeutsche Verlag Köln u. Opladen, 1965, 583-608.

PAUL J. COHEN AND THE CONTINUUM PROBLEM

ALONZO CHURCH

On the occasion of the award of a prize to Paul Cohen, and in spite of significant contributions by him to analysis, to topological groups, and to the theory of differential equations, I believe that the audience will agree that it is appropriate to devote the entire time allowed for exposition to the continuum problem.

For here is another case, of the sort which arises from time to time in the history of mathematics, in which a mathematician who has done important work in other fields turns to a field not properly his own to solve an outstanding problem that has baffled the specialists. As a consequence of the tremendous growth of mathematics the universal mathematician of other days is no longer a possibility — David Hilbert was certainly the last of them. The next best thing is that abler men should not confine themselves too closely to one field or be afraid to turn to an area in which they may not have all the expert knowledge of those who have concentrated their work in it. Certainly Paul Cohen's results have been and will be greatly extended, and the method of his proof greatly improved, by the specialists in set theory. But we are concerned today with the initial break-through.

Number one of the Hilbert problems, placed even before the problem of the consistency of arithmetic which occupied so much of Hilbert's own attention in the latter part of his life, is "Cantor's problem of the cardinal number of the continuum." So it is titled in the contemporary English translation of Hilbert's famous paper. Hilbert himself in German uses Cantor's original term "Mächtigkeit," which has no good English translation. Hilbert does not say that the order in which the problems are numbered gauges their relative importance, and it is not meant to suggest that he intended this. But he does mention the arithmetical formulation of the concept of the continuum and the discovery of non-Euclidean geometry as being the outstanding mathematical achievements of the preceding century, and gives this as a reason for putting problems in these areas first.

Already in 1878 Cantor stated the continuum hypothesis as a conjecture. But there is a sense in which the continuum problem dates from Cantor's statement at the end of a paper which appeared in the *Mathematische Annalen* in 1884. Here it is proved that a closed

infinite subset of the (linear) continuum must have the cardinal number either of the natural numbers or of the whole continuum. Then it is said that the result can be extended to subsets which are not closed, and a proof will be provided. The paper closes with the words "Fortsetzung folgt." But the promised Fortsetzung never did folgen, and it seems clear that the proof Cantor believed he had broken down.

Cantor passes at once from the proposition that there is no cardinal number between that of the natural numbers and that of the continuum to the second form of the continuum hypothesis, that the cardinal number of the continuum is $\aleph_1$. Of course this depends on a tacit assumption of the axiom of choice, in particular as Sierpiński's result of 1947 deriving the axiom of choice from the generalized continuum hypothesis was not then available (or the background that made this result possible) [1]). Hilbert is more cautious and states as a separate problem, subsidiary to the continuum problem, the question whether the continuum can be well ordered. Zermelo's paper which explicitly states the axiom of choice for the first time (in the strong form which Zermelo later called "Prinzip der Auswahl"), and shows as a consequence of it that every set can be well ordered, followed Hilbert's paper on mathematical problems by only four years.

It was Cantor's original point of view that the transfinite cardinal and ordinal numbers are two different kinds of generalizations of the natural numbers and are to serve the same purposes for transfinite sets which the natural numbers do for finite sets. If this program is to be fulfilled, one evidently must be able to answer at least the simplest and most immediate questions that arise about the cardinal numbers of the most commonly used mathematical sets, among them the continuum. This is clearly the reason why the frustrating difficulties of the continuum problem acquired the importance that they did for the Cantor theory. Surely neither Cantor nor Hilbert could have surmised that the ultimate solution would take the negative form that it has. Yet Hilbert is quite explicit that in general it may happen that the solution of a problem must be in the form of an impossibility proof.

The antinomies of set theory, which first came to the attention of the mathematical public through Burali-Forti's paper of 1897, played an important role in the progress towards the ultimate solution, as it was the antinomies that forced the transition from the older naive and "genetic" use of sets in mathematics to an axiomatic basis for set theory. And it is of course only by the axiomatic method that a proof of the impossibility of a proof becomes possible.

[1]) Sierpiński points out that this result had been announced by Lindenbaum and Tarski in 1926. Their proof was never published.

Within axiomatic set theory there was a proof of the independence of the axiom of choice by Fraenkel as early as 1922. But this was unsatisfactory in that it referred to axioms of set theory so formulated as to admit a domain of Urelemente, or non-sets, of unspecified structure, and the possibility remained open that the axiom of choice would lose its independence upon adding axioms specifying the structure of the domain of Urelemente (or most simply, upon adding as an axiom that there are no Urelemente). Extensions of Fraenkel's result and improvements of his method, by Fraenkel himself, Lindenbaum, Mostowski, and more recently Shoenfield and Mendelson either did not remove this objection or only mitigated it (in the sense that independence from quite the full usual system of axioms for set theory was not yet proved).

A much more important step—which constitutes in fact the first half of the solution of the continuum problem, and on which subsequent work heavily depends—was taken by Kurt Gödel in 1938-40. Abstracts of Gödel's methods and results appeared in 1938 and 1939, and the monograph containing the full proofs, in 1940. Gödel's method is to set up what has since been called an inner model of set theory. I.e., set theory without axiom of choice is used to set up a model of set theory in which both the axiom of choice and the generalized continuum hypothesis hold. (The generalized continuum hypothesis is the proposition that the power set of a set of cardinal number $\aleph_\alpha$ has the cardinal number $\aleph_{\alpha+1}$, Cantor's original continuum hypothesis, in its second form, being the special case of this in which $\alpha = 0$.) The result of Gödel's procedure, setting up an inner model, is a relative consistency proof for the axiom of choice and for the generalized continuum hypothesis: If set theory without axiom of choice is consistent, it remains so upon introducing both the axiom of choice and the generalized continuum hypothesis as additional axioms.

After the (relative) consistency proof, the second half of the negative solution of the continuum problem is of course independence. A partial step in this direction was taken by Gödel, who in 1942 found a proof of the independence of the axiom of constructibility in type theory. According to his own statement (in a private communication) he believed that this could be extended to an independence proof of the axiom of choice; but due to a shifting of his interests toward philosophy, he soon afterwards ceased to work in this area, without having settled its main problems. The partial result mentioned was never worked out in full detail or put into form for publication.

These climactic results, the independence in set theory of the axiom of choice (even the weak form of the axiom of choice which concerns a countable set of pairs) and of the continuum hypothesis

from the axiom of choice, remained for Paul Cohen in 1963-64. It is no part of our present purpose to describe the details of his method. Let it only be said that, besides the now well-known notion of *forcing*, it depends on an adaptation of Gödel's method of 1940 for setting up models of set theory, on a modification of the earlier methods of Fraenkel, Mostowski, and others in connection with the independence of the axiom of choice, and on the result of Skolem that there exists a countable model of set theory (a model having the cardinal number of the natural numbers).

The feeling that there is an absolute realm of sets, somehow determined in spite of the non-existence of a complete axiomatic characterization, receives more of a blow from the solution (better, the unsolving [1]) of the continuum problem than from the famous Gödel incompleteness theorems. It is not a question of realism (miscalled "Platonism") versus either conceptualism or nominalism, but if one chooses realism, whether there can be a "genetic" realism without axiomatic specification. The Gödel-Cohen results and subsequent extensions of them have the consequence that there is not one set theory but many, with the difference arising in connection with a problem which intuition still seems to tell us must "really" have only one true solution.

I know of mathematicians who hold that the axiom of choice has the same character of intuitive self-evidence that belongs to the most elementary laws of logic on which mathematics depends. It has never seemed so to me. But how shall one argue matters of intuition? The point is, I know of no one who maintains such self-evidence for the continuum hypothesis.

The realist will expect that the reality independent of the human mind which he maintains must have many ramifications, and will take what has now become the classical mathematical view, dating from the nineteenth century discussions of non-Euclidean geometry, that all the ramifications equally demand exploration. The same view is possible also for him who takes the intermediate position between radical realism and conceptualism by holding that mathematical and physical objects alike, not excluding such basic logico-mathematical objects as sets, have their reality only relative to and within a certain theory. And if a choice must in some sense be made among the rival set theories, rather than merely and neutrally to develop the mathematical consequences of the alternative theories, it seems that the only basis for it can be the same informal criterion of simplicity that governs the choice among rival physical theories when both or all of them equally explain the experimental facts.

[1]) I borrow this whimsical term from W. W. Boone.

REFERENCES

[1] Cantor G., Ein Beitrag zur Mannigfaltigkeitslehre, *Journal für die reine und angewandte Mathematik*, 84 (1878), 242-258. [See p. 257.]

[2] Cantor G., Ueber unendliche, lineare Punktmannigfaltigkeiten, *Mathematische Annalen*, 23, № 6 (1884), 453-488.

[3] Burali-Forti Cesare, Una questione sui numeri transfiniti, *Rendiconti del Circolo Matematico di Palermo*, 11 (1897), 154-164.

[4] Hilbert David, Mathematische Probleme, *Nachrichten von der K. Gesellschaft der Wissenschaften zu Göttingen*, Math.-Phys. Kl., 1900, pp. 253-297. Reprinted with additions in *Archiv der Mathematik und Physik*, ser. 3, vol. 1 (1901), 44-63, 213-237. English translation in *Bulletin of the American Mathematical Society*, 8 (1901-2), 437-479.

[5] Zermelo Ernst, Beweis, daß jede Menge wohlgeordnet werden kann, *Mathematische Annalen*, 59 (1904), 514-516.

[6] Fraenkel A., Der Begriff «definit» und die Unabhängigkeit des Auswahlaxioms, *Sitzungsberichte der Preussischen Akademie der Wissenschaften*, Phys.-Math. Kl., 1922, pp. 253-257.

[7] Skolem Thoralf, Einige Bemerkungen zur axiomatischen Begründung der Mengenlehre, Wissenschaftliche Vorträge gehalten auf dem Fünften Kongress der Skandinavischen Mathematiker in Helsingfors vom 4, bis 7, Juli 1922, Helsingfors, 1923, 217-232.

[8] Lindenbaum A., Tarski A., Communication sur les recherches de la théorie des ensembles, *Comptes Rendus des Séances de la Société des Sciences et des Lettres de Varsovie*, Classe III, 19 (1926), see p. 314.

[9] Cödel Kurt, Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I, *Monatshefte für Mathematik und Physik*, 38 (1931), 173-198.

[10] Fraenkel A., Ueber eine abgeschwaechte Fassung des Auswahlaxioms, *The Journal of Symbolic Logic*, 2 (1937), 1-25.

[11] Lindenbaum Adolf, Mostowski Andrzej, Über die Unabhängigkeit des Auswahlaxioms und einiger seiner Folgerungen, *Comptes Rendus des Séances de la Société des Sciences et des Lettres de Varsovie*, Classe III, 31 (1938), 27-32.

[12] Gödel Kurt, The consistency of the axiom of choice and of the generalized continuum-hypothesis, *Proceedings of the National Academy of Sciences*, 24 (1938), 556-557.

[13] Gödel Kurt, Consistency-proof for the generalized continuum-hypothesis, *Proceedings of the National Academy of Sciences*, 25 (1939), 220-224.

[14] Mostowski Andrzej, Über die Unabhängigkeit des Wohlordnungssatzes vom Ordnungsprinzip, *Fundamenta Mathematicae*, 32 (1939), 201-252.

[15] Gödel Kurt, The consistency of the axiom of choice and of the generalized continuum-hypothesis with the axioms of set theory, Princeton, 1940, 66 pp.

[16] Sierpiński Wacław, L'hypothèse généralisée du continu et l'axiome du choix, *Fundamenta Mathematicae*, 34 (1947), 1-5.

[17] Shoenfield J. R., The independence of the axiom of choice, Abstract, *The Journal of Symbolic Logic*, 20 (1955), 202.

[18] Mendelson Elliott, The independence of a weak axiom of choice, *The Journal of Symbolic Logic*, 21 (1956), 350-366.

[19] Mendelson Elliott, The axiom of Fundierung and the axiom of choice, *Archiv für mathematische Logik und Grundlagenforschung*, 4 (1958), 65-70.

[20] Cohen Paul J., A minimal model for set theory, *Bulletin of the American Mathematical Society*, 69 (1963), 537-540.

[21] Cohen Paul J., The independence of the continuum hypothesis, *Proceedings of the National Academy of Sciences*, 50 (1963), 1143-1148, and 51 (1964), 105-110.

[22] Cohen Paul J., Independence results in set theory. The theory of models, Proceedings of the 1963 International Symposium at Berkeley, Amsterdam (1965), 39-54.

LES TRAVAUX DE ALEXANDER GROTHENDIECK

JEAN DIEUDONNÉ

Alexandre Grothendieck n'a pas 40 ans, et déjà l'ampleur de son œuvre et l'étendue de son influence sur les mathématiques contemporaines sont telles qu'il n'est pas possible d'en donner autre chose qu'une idée très déformée dans un aussi bref exposé.

Chacun sait que Grothendieck est le principal artisan de la rénovation de la Géométrie algébrique qui s'accomplit sous nos yeux. Bien entendu, cette rénovation a été préparée par les travaux de Weil-Zariski d'une part, fondant la Géométrie algébrique « abstraite » sur un corps quelconque, et d'autre part par ceux de Serre, introduisant dans la théorie les puissants outils que sont les faisceaux et l'algèbre homologique. Mais Grothendieck a su donner à ces idées toute leur portée en les développant sous leur forme générale, débarrassées des restrictions parasites qui en gênaient l'emploi ; et il y a ajouté de nombreuses idées entièrement originales.

Sous sa forme « affine », la Géométrie algébrique moderne se confond avec l'algèbre commutative ; déjà en Géométrie algébrique classique, à une variété affine était associé l'anneau des fonctions « régulières » sur la variété. Inversement, on fait maintenant correspondre biunivoquement à un anneau commutatif *quelconque* A (ayant un élément unité) un objet géométrique, le « schéma affine d'anneau A », qui est l'ensemble des idéaux premiers de A, muni d'une certaine topologie et d'un faisceau dont les fibres sont les anneaux locaux aux idéaux premiers de A. L'intérêt de cette formulation est : 1° de fournir une intuition géométrique qui est un guide très appréciable (en suggérant par exemple des analogies, avec les variétés différentiables ou analytiques) ; 2° de dépasser le point de vue « affine » pour aboutir à l'idée de « schémas » généraux (généralisant les «variétés abstraites» de Weil) par le simple procédé topologique de «recollement» des espaces topologiques munis de faisceaux (idée due à Serre).

Ce cadre est complété par deux idées nouvelles: 1° l'accent mis sur la notion de *morphisme*, qui, dans le cas affine, correspond à celle d'homomorphisme d'anneaux conservant l'élément unité, et qui permet de «relativiser» toutes les notions de la théorie classique; 2° la notion générale de «*changement de base*: étant donné un morphisme $X \to S$, pour tout morphisme «changement de base» $S' \to S$, on forme canoniquement un nouveau schéma $X' = X_{(S')}$

et un nouveau morphisme $X' \to S'$ par un procédé qui, dans le cas affine, correspond au produit tensoriel des anneaux, et englobe la classique «extension du corps de base» de la période Weil-Zariski.

Ces notions, ainsi que celle de *platitude* (due à Serre, mais dont Grothendieck a considérablement développé l'emploi) sont à la base d'une technique d'une puissance et d'une souplesse remarquables. Parmi les nombreux outils ainsi forgés, citons notamment:

I) Le passage à la *limite projective* dans les schémas, qui permet dans beaucoup de cas de ramener les problèmes au cas où les anneaux que l'on considère sont des algèbres de type fini sur **Z** (concrétisant ainsi la célèbre thèse kroneckérienne).

II) La théorie des anneaux *excellents*, qui systématise et complète des résultats profonds de Zariski-Nagata sur les anneaux locaux nœthériens, et peut être utilisée dans les problèmes généraux grâce au passage à la limite projective, les **Z**-algèbres de type fini étant des anneaux excellents.

III) La théorie de la *cohomologie relative* dans les schémas et de ses relations avec la notion de profondeur (due à Auslander-Buchsbaum et Serre).

IV) La théorie des *schémas formels* : ce sont ici des *limites inductives* de schémas, opération qui, dans le cas affine, correspond à la complétion des anneaux locaux, mais a une portée plus générale; elle permet par exemple, dans certaines questions, de ramener un problème en caractéristique $p > 0$ à un problème en caractéristique 0; c'est dans ce cadre aussi que se formule la théorie des «fonctions holomorphes» de Zariski (mise d'ailleurs sous une forme cohomologique beaucoup plus générale).

V) L'utilisation de la notion de *foncteur représentable*, qui remplace celle, plus limitée, de «problème universel»; il s'agit, étant donné un ensemble E attaché à un morphisme $X \to S$, d'une façon «compatible» avec les changements de base, de savoir s'il existe un schéma Z sur S tel que E s'identifie de façon naturelle à l'ensemble des *sections* du morphisme $Z \to S$.

VI) La théorie de la *descente*. De nombreux problèmes se simplifient lorsqu'on fait un «changement de base» approprié (par exemple, en Géométrie algébrique classique, lorsqu'on passe du corps de base à une clôture algébrique de ce corps). Il s'agit de pouvoir revenir à la situation initiale et y tirer des conséquences de ce qui se passe après le changement de base; c'est le but de la théorie de la «descente», qui fournit des critères permettant ce retour dans des cas assez généraux pour avoir de nombreuses applications.

VII) Poussant cette idée plus loin, Grothendieck a généralisé de façon très originale la notion de topologie et la cohomologie des faisceaux sur un espace topologique (théorie des « sites » et des « topos ») : le rôle joué par les ouverts d'un espace topologique est

tenu par des morphismes $X' \to X$ d'un type spécial, le plus souvent les morphismes *étales* (analogues des « revêtements non ramifiés » d'un ouvert d'une variété analytique).

Avant d'aller plus loin, il convient de remarquer qu'on ne rend pas justice aux théories précédentes en les qualifiant un peu dédaigneusement de « résultats techniques » ; pour certains d'entre eux, il s'agit bien d'extensions faciles de méthodes classiques, mais beaucoup nécessitent des méthodes d'attaque toutes nouvelles, s'appuyant sur des considérations subtiles d'algèbre commutative ou d'algèbre homologique, et à eux seuls constitueraient déjà une œuvre imposante et dont il est peu d'exemples.

Il est bien vrai cependant que, dans l'esprit même de Grothendieck, toutes ces méthodes ne sont pas développées pour elles-mêmes, mais en vue d'attaquer quelques problèmes fondamentaux de la Géométrie algébrique. Parmi ceux où il a réalisé (en partie avec la collaboration de ses élèves) des progrès sensibles, il faut citer :

1° La détermination, en caractéristique $p > 0$, de la partie première à p du groupe fondamental d'une courbe algébrique.

2° Les définitions du schéma formel de « modules » (classes de schémas isomorphes), du schéma de Picard (classes de diviseurs), du schéma de Hilbert (ensemble de sous-variétés d'une variété donnée, dont la structure de schéma est destinée à se substituer aux classiques « coordonnées de Chow »).

3° (En collaboration avec M. Demazure.) Une vaste théorie des « schémas en groupes », généralisant la théorie des groupes algébriques de Chevalley.

4° (En collaboration avec M. Artin et J. Verdier.) La définition de la « cohomologie étale » des schémas, où, grâce à la théorie des « sites », on dispose déjà en toute caractéristique de méthodes et résultats analogues à ceux que fournit la topologie algébrique pour les variétés algébriques sur le corps des complexes (théorèmes de finitude et de dualité, formule de Lefschetz, comparaison avec la cohomologie « topologique » dans le cas classique) ; grâce à ces résultats, Grothendieck a pu démontrer une partie des fameuses « conjectures de Weil » : rationalité des fonctions L attachées aux variétés sur un corps fini, et leur expression à l'aide d'invariants homologiques.

5° Enfin, le premier en date des travaux de Grothendieck en Géométrie algébrique, la généralisation du théorème de Riemann-Roch-Hirzebruch et sa démonstration purement algébrique en toute caractéristique. C'est à cette occasion que Grothendieck a introduit les premières notions de « K-théorie » (ou, comme on dit maintenant, les « groupes (ou anneaux) de Grothendieck »). Cette idée a beaucoup frappé notamment les topologistes et les algébristes, qui en ont tiré, dans de multiples domaines, les brillantes applications que l'on sait.

Il convient d'ailleurs de signaler aussi les travaux de Grothendieck en algèbre homologique, un peu antérieurs à sa démonstration du théorème de Riemann-Roch, et qui ont élargi et assoupli les résultats de Cartan-Eilenberg, notamment en donnant une « bonne » définition de la cohomologie des faisceaux sur un espace quelconque.

Enfin, je n'ai rien dit des premiers mémoires de Grothendieck sur les espaces vectoriels topologiques (1950-55), en partie parce qu'ils sont fort connus et de plus en plus utilisés en Analyse fonctionnelle, notamment la théorie des espaces nucléaires, qui « explique » les phénomènes rencontrés dans la théorie des distributions. J'ai eu personnellement le privilège d'assister de près, à cette époque, à l'éclosion du talent de cet extraordinaire « débutant » qui à 20 ans était déjà un maître ; et, avec 10 ans de recul, je considère toujours que l'oeuvre de Grothendieck de cette période reste, avec celle de Banach, celle qui a le plus fortement marqué cette partie des mathématiques.

S'il fallait chercher une parenté spirituelle à Grothendieck, c'est à Hilbert, me semble-t-il, qu'on pourrait le mieux le comparer : comme Hilbert, sa devise pourrait être : « simplifier en généralisant », en recherchant les ressorts profonds des phénomènes mathématiques ; mais, comme Hilbert aussi, lorsque cette analyse en profondeur a conduit à un point où seule l'attaque de front reste possible, il trouve presque toujours dans sa riche imagination le bélier qui enfonce l'obstacle. La comparaison est peut-être lourde à porter, mais Grothendieck est de taille à n'en pas être accablé.

SUR LES TRAVAUX DE STEPHEN SMALE

RENÉ THOM

Le premier travail scientifique de S. Smale est sa thèse de Ph.D. soutenue en 1956 à l'Université de Michigan (Ann Arbor). Faite sous la direction de Raoul Bott, elle témoigne déjà d'une éclatante maîtrise. Le résultat essentiel, maintenant bien connu, est le théorème du relèvement des homotopies des immersions d'une variété modulo une sous-variété. Etabli par des constructions géométriques raffinées, ce résultat témoignait chez son auteur de capacités d'intuition de tout premier ordre. Grâce à lui, on pouvait établir une conjecture—vieille alors d'une dizaine d'années—de C. Ehresmann sur la classification des immersions d'une variété dans une autre ; il en résultait qu'il était possible, par une déformation régulière (c'est-à-dire sans sortir des immersions) de transformer le plongement canonique de la « 2-sphère » dans l'espace euclidien R^3 à trois dimensions en un plongement antipodique ; ce résultat n'alla pas sans soulever la curiosité des topologues, dont beaucoup s'ingénièrent à préciser cette déformation. Mais, la thèse de Smale donnait plus que cette curiosité, elle ouvrait une voie d'attaque dans tout un domaine de questions jusqu'alors inabordables, et tout un chapitre de Topologie Différentielle, l'étude des immersions et plongements d'une variété différentiable dans une autre, marqué par les travaux de M. Hirsh, Haefliger, etc., en est plus ou moins directement sorti.

Avec les grands travaux de 1960 sur la conjecture de Poincaré, nous abordons la partie la plus connue de l'œuvre de Smale, celle qui, sans doute, nous vaut sa présence ici. On savait déjà, par suite de la théorie de Morse, que toute variété compacte se divise en cellules de gradient, et que, si l'on se donne à chaque niveau critique l'attachement de la cellule de gradient correspondante, on est en mesure de reconstituer la variété ; on avait déjà commencé,—à la suite des travaux de Kervaire, Milnor, Wallace—à pratiquer la « chirurgie» des variétés, c'est-à-dire la technique qui consiste à transformer une variété plus simple par résection d'un couple d'anses duales.

L'énorme mérite de Smale, en cette question, est d'avoir osé entreprendre ce que tout autre mathématicien du temps aurait considéré comme sans espoir : étant donnée une fonction de Morse sur une sphère d'homotopie, simplifier la présentation de cette variété en éliminant par chirurgie les couples d'index k, $k+1$ de points critiques excédentaires. Que cela fût possible, on connaissait trop

les difficultés dans les petites dimensions (trois ou quatre), pour l'espérer ; Smale osa, et réussit. Il comprit que les difficultés entrevues étaient un phénomène spécial aux petites dimensions : en se bornant aux dimensions supérieures à cinq, on se trouvait plus à l'aise pour travailler, et la chirurgie s'effectuait plus aisément ; par des constructions très ingénieuses, Smale vint alors à bout des dernières difficultés : il élimine sur une sphère d'homotopie tous les couples de points critiques d'indice k différent de zéro et n ; il obtient ainsi une variété sur laquelle il existe une fonction ne présentant que deux points critiques (minimum et maximum). D'après un théorème de Reeb, cette variété est une sphère topologique (mais non nécessairement difféomorphe à la sphère usuelle, comme l'ont montré les exemples dus à Milnor).

Ce résultat extrêmement brillant s'est trouvé complété peu après par le théorème dit du h-cobordisme, qui le généralise. Si deux variétés compactes M_1 et M_2 forment le bord d'une même variété à bord W, dont elles sont rétractées par déformation et si elles sont simplement connexes, alors M_1 et M_2 sont difféomorphes. Ce théorème permet, à l'aide d'un résultat ultérieur de Novikov et Browder, de ramener le problème de la classification des variétés différentiables à un pur problème d'homotopie, en fait à un problème d'algèbre (il est vrai, difficile). Les techniques usées dans la démonstration du théorème du h-cobordisme n'ont probablement pas donné tout leur fruit, et des travaux plus récents, comme ceux de J. Cerf, en ont élargi le champ d'application.

Avec le résultat contemporain de B. Mazur sur la conjecture de Schönflies, les travaux de Smale tournent une page en Topologie algébrique. On peut dire que la topologie des « espaces », des variétés différentiables est désormais quasi-achevée. Il subsiste certes beaucoup de questions non résolues : les structures algébriques définies par la classification ne sont pas élucidées —en général— ; mais le seront-elles un jour ? Il ne reste guère que la théorie—qui a fait d'ailleurs récemment de beaux progrès—de ces êtres malgré tout quelque peu pathologiques que sont les variétés semi-linéaires et les variétés purement topologiques. Dans ces conditions, si la Topologie veut se renouveler, et ne pas se cantonner en des problèmes ardus d'une vaine technicité, elle doit se préoccuper de renouveler ses matériaux et aborder des problèmes neufs. Avec les objets géométriques associés aux structures différentiables : formes différentielles, tenseurs, structures feuilletées, opérateurs différentiels, un champ immense est ouvert au topologue. On a vu d'ailleurs qu'un de nos lauréats s'est vu récompenser pour un résultat dans cette voie.

A côté de l'Analyse classique, essentiellement linéaire, il y a le domaine pratiquement inexploré de l'analyse non linéaire ; là, le topologue peut espérer encore mieux utiliser ses méthodes, et peut-être

à qualité essentielle, à savoir la vision intrinsèque des choses. C'est ce que comprendra très vite Smale ; dans un article en collaboration avec R. Palais, il définira les meilleures conditions possibles d'application de la théorie de Morse au Calcul des variations ; il en déduira ensuite des théorèmes relatifs à l'existence des solutions de problèmes elliptiques non linéaires. Mais, très tôt, Smale se tourne vers une théorie—alors bien délaissée— : la théorie qualitative des systèmes différentiels sur une variété. Quasiment seul, Smale lit Poincaré et Birkhoff ; devant l'inextricable du problème, il comprend très vite l'intérêt d'une notion essentielle, celle de « stabilité structurelle introduite par Andronov et Pontrjagin, cette notion vise à caractériser, parmi les champs de vecteurs sur une variété, ceux qui jouissent d'une propriété de stabilité qualitative, au sens suivant : tout champ (Z) assez voisin du champ donné (X) (avec la C^1-topologie) donne naissance à un champ de trajectoires homéomorphe au champ défini par (X). Le problème central est alors le problème de l'approximation: tout champ de vecteurs peut-il être approché par un champ structurellement stable? Ce problème, résolu positivement par Peixoto pour les variétés compactes de dimension inférieure à deux, était posé pour les dimensions supérieures. Smale construit alors une variété compacte M de dimension quatre, et un champ (X) sur M, tel qu'aucun champ (Z) assez voisin de (X) ne soit structurellement stable. Le problème général de la stabilité des systèmes différentiels est ainsi résolu par la négative. Cependant, la notion même de stabilité structurelle est loin d'avoir perdu tout son intérêt : d'abord, parce qu'il existe, dans l'espace fonctionnel des champs de vecteurs d'une variété M, un ouvert « relativement important » de champs structurellement stables : celui formé par les champs de vecteurs de type gradient génériques (sans récurrence) et, probablement, une classe de champs définis par Smale (les champs dits de Morse-Smale), qui présentent de la récurrence (avec des trajectoires fermées) mais sous une forme bénigne et sévèrement contrôlée. Mais, par l'étude des configurations de trajectoires associées aux points homocliniques de Poincaré, Smale se convainc bien vite que d'autres champs, à topologie complexe et rigide, sont structurellement stables. Il revenait aux brillants travaux de l'école soviétique (avec Sinai, Arnold, Anosov) d'établir l'existence d'une classe étendue de champs structurellement stables, du type du flot géodésique sur une variété riemannienne à courbure négative. Ces travaux ont exercé sur Smale une grande influence, et ont infléchi ses recherches dans la direction actuelle, à savoir la mise en évidence d'une « stabilité structurelle par morceaux », chaque « morceau » étant lié à une configuration rigide de trajectoires récurrentes (non-wandering) au sens de Birkhoff. Ces recherches sont en cours et semblent fort prometteuses.

Je m'en voudrais de ne pas insister sur un dernier point : si les oeuvres de Smale ne possèdent peut-être pas la perfection formelle du travail définitif, c'est que Smale est un pionnier qui prend ses risques avec un courage tranquille ; dans un domaine complètement inexploré, dans une jungle géométrique d'une inextricable richesse, il est le premier à avoir tracé la route et posé les premiers jalons. Et l'on peut prévoir que son oeuvre revêtira à l'avenir une importance fondamentale, comparable à celle des grands précurseurs, Poincaré et Birkhoff. Après tout bien des problèmes classiques, tels le problème de Fermat ou la conjecture de Riemann, peuvent attendre leur solution encore quelques années ; mais si la science veut user de l'outil différentiel pour décrire les phénomènes naturels, elle ne peut se permettre d'ignorer encore longtemps la structure topologique des attracteurs d'un système dynamique structurellement stable, car tout « état physique » présentant une certaine stabilité, une certaine permanence, est nécessairement représenté par un tel attracteur. Selon certaines vues de Smale, ces attracteurs seraient des espaces homogènes de groupes de Lie d'un type spécial. Si ces vues pouvaient s'étendre aux systèmes hamiltoniens, on pourrait peut-être s'expliquer l'apparition —jusqu'ici si incomprise—des groupes de Lie dans la Physique des particules élémentaires. En ce sens, le problème de Smale est —à mes yeux—d'une importance épistomologique essentielle.

ADDRESS DELIVERED BY PROFESSOR G. DE RHAM AT THE CLOSING CEREMONY OF THE CONGRESS

Mr. President, Ladies and Gentlemen!

As the retiring President of the International Mathematical Union, it is my pleasant duty to announce that the fifth General Assembly of the Union, which was held at Dubna on August 13-15, 1966, elected the following Executive Committee for a term of four years,

President:	Professor H. Cartan
Vice-Presidents:	Academician M. A. Lavrentiev and Professor D. Montgomery
Secretary:	Professor O. Frostman
Members:	Professors M. F. Atiyah, K. Chandrasekharan, G. Hajoš, G. Vesentini and K. Yosida.

I am sure that all of you would want me to wish the new Executive Committee every success.

The first object of the International Mathematical Union is to promote international cooperation in Mathematics. In respect to this, the most striking fact during the last years has been the progressive cooperation between mathematicians of the Soviet Union and those of other countries, especially of Western Europe and U.S.A. It is a particular pleasure for me to emphasize the important position occupied by Soviet Mathematicians in our Union. Their contribution to the development of our Science is of the highest significance. This will continue to increase, due to the abundance of brilliant young Soviet Mathematicians. Mathematicians of all countries welcome every opportunity to meet them. May I express the wish that such contacts will grow, for the benefit of all.

It is also my pleasant duty to express the warmest appreciation and thanks of the Union to the Organizing Committee of this magnificent Congress. It is the first International Congress of Mathematicians to be held in the Soviet Union and it is the largest of all our Congresses, with a record number of participants from the host country and rom abroad. The level of the lectures has been very high. A tremendous amount of work has gone into its organization. The Union owes special thanks to the President of the Congress, Academician Petrovsky, to the Chairman of the Soviet National Committee of Mathematicians,

Academician I. M. Vinogradov, to Academician Lavrentiev and to Professor Mergelyan, for the successful organization of this huge meeting. To all members of the Organizing Committee and to their assistants, we remain grateful, as well to the members of the Consultative Committee and his Chairman Professor R. Nevanlinna.

To Academician Keldysh, President of the Academy of Sciences of the Soviet Union, to the Government of the Soviet Union and to the Municipality of Moscow, we are deeply indebted for the hospitality and consideration we have all received.

Now, as Chairman of the Committee to decide the location of the next Congress, I have the pleasure to request the President, Academician Petrovsky, to call upon the delegate from France, Professor Dieudonné, Dean of the Faculty of Sciences of Nice, to address the Congress.

Thank you all.

DISCOURS DE CONCLUSION DU PROFESSEUR J. DIEUDONNÉ AU CONGRÈS

Au nom du Comité national français de mathématiques, j'ai l'honneur d'inviter le Congrès International des mathématiciens à tenir sa session de 1970 en France. La ville de Nice, par sa situation, son climat, son équipement touristique et l'existence d'une Université active, réunit les conditions requises pour le siège d'un Congrès scientifique. Je propose donc que le Congrès International des mathématiciens se tienne en 1970 dans la ville de Nice.

РЕЧЬ ПРЕЗИДЕНТА КОНГРЕССА АКАДЕМИКА И. Г. ПЕТРОВСКОГО НА ЗАКЛЮЧИТЕЛЬНОМ ЗАСЕДАНИИ КОНГРЕССА

Уважаемые члены Конгресса, уважаемые гости!

Позвольте мне еще раз поблагодарить вас за участие в работе Конгресса.

Мне хочется также еще раз поблагодарить Исполнительный комитет Международного союза математиков, с которым мы все время работали в контакте и который много нам помогал.

Позвольте пожелать всем вам успехов в работе и всего самого лучшего.

ЧАСОВЫЕ ДОКЛАДЫ

✻

ONE-HOUR REPORTS

✻

RAPPORTS D'UNE DURÉE D'UNE HEURE

✻

VORTRÄGE VON EINER STUNDE DAUER

A SURVEY OF HOMOTOPY-THEORY

JOHN F. ADAMS

First, I would like to explain that this lecture will be purely expository. It will be directed at the non-specialist; I shall merely try to explain what homotopy-theory is about and what you can expect it to do for you.

We need a definition. Let X and Y be two topological spaces, and let f_0, f_1 be two maps from X to Y, that is, two continuous functions. We say that f_0 is homotopic to f_1, and write $f_0 \sim f_1$, if there exist also maps f_t: $X \to Y$ for $0 \leqslant t \leqslant 1$ such that $f_t(x)$ is a continuous function of the two variables t, x for $0 \leqslant t \leqslant 1$, $x \in X$. In other words, f_0 is homotopic to f_1 if f_0 can be deformed continuously into f_1.

Homotopy is an equivalence relation, and so the maps from X to Y may be divided into equivalence classes. We write $[X, Y]$ for the set of homotopy classes of maps from X to Y.

For example, consider the case in which X and Y are both the unit circle S^1. Let f: $S^1 \to S^1$ be a map; as x moves once around S^1, $f(x)$ will move around S^1 some integer number of times. This integer is called the degree of f. Two maps f_0, f_1: $S^1 \to S^1$ are homotopic if and only if they have the same degree. The degree sets up a (1-1) correspondence between the set of homotopy classes $[S^1, S^1]$ and the set of integers Z.

Homotopy-theory studies those properties of spaces and maps which are not changed if we replace all the maps by homotopic ones.

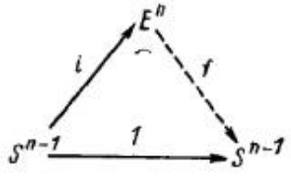
There are several ways of seeing that this is a justifiable restriction of our studies. First, in algebraic topology we assign invariants to spaces X, and also to maps f: $X \to Y$. For example, consider the case in which X and Y are both the unit sphere S^{n-1} in Euclidean n-space R^n, given by the equation $x_1^2 + x_2^2 + \ldots + x_n^2 = 1$. Then we assign to each f: $S^{n-1} \to S^{n-1}$ its degree $d(f)$, which is an integer. In general, we assign invariants which are algebraic in nature; typically they lie in discrete sets (like the integers). If there is any continuity about our proceedings, then homotopic maps (which can be continuously deformed into one another) will have their invariants equal. For example, $f \sim g$: $S^{n-1} \to S^{n-1}$ implies $d(f) = d(g)$. In fact, most of the classical invariants of algebraic topology are homotopy invariants.

Secondly, we can point out how many problems have been found to be amenable to study by the methods of homotopy-theory. Let me present a few examples.

(1) Questions on the existence or non-existence of spaces or maps with assigned properties.
(2) Questions on fiberings.
(3) Questions on cobordism.
(4) Other questions on manifolds.

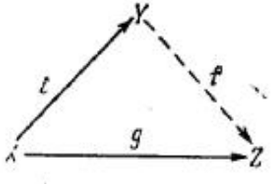
Let me present one question of type (1). Let E^n be the unit solid ball in R^n, given by the inequality $x_1^2+x_2^2+\ldots+x_n^2\leqslant 1$; then there is no map $f\colon E^n\to S^{n-1}$ whose restriction to S^{n-1} is the identity.

We can state this differently. Let $i\colon S^{n-1}\to E^n$ be the injection map; then there is no map $f\colon E^n\to S^{n-1}$ such that $fi=1$.



This result is intuitively very plausible, and is easy to prove. We have $i\sim c$: $S^{n-1}\to E^n$, where c is any constant map; so if f existed, we should have $1=fi\sim fc$: $S^{n-1}\to S^{n-1}$; but $d(1)=1$, $d(fc)=0$. The interest of the result is that from it one can easily obtain the Brouwer fixed-point theorem; see [14].

This particular result is one of a general class. In an "extension problem" we are given maps $i\colon X\to Y$, $g\colon X\to Z$ and we are asked if there is a map $f\colon Y\to Z$ such that $fi=g$.

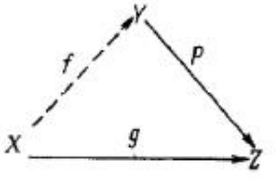


Sometimes there is such an f and sometimes there is not; it all depends on X, Y, Z, i and g. Many problems can be thrown into this form; and we can solve harder examples than the one I've given. See [2].

If the map i satisfies reasonable conditions, then replacing g by a homotopic map does not alter the answer to the problem; i.e., the extension problem falls within the scope of homotopy theory. See [15, 17, 27].

There is another class of problem which is formally very similar to the first. In a "lifting problem" we are given maps $p\colon Y\to Z$ and

$g\colon X\to Z$, and we are asked if there is a map $f\colon X\to Y$ such that $pf=g$.



Many problems can be thrown into this form. For example, take the theorem that any vector field on S^2 is somewhere zero. We take $X=Z=S^2$ and $g=1$; we take Y to be the space of non-zero tangent vectors to S^2, and let p assign to each tangent vector the point where it is a tangent vector. In this case the answer is that there is no such f.

We can also solve harder examples than the one I've given; see [3].

As before, if the map p satisfies reasonable conditions, then replacing g by a homotopic map does not alter the answer to the problem; i.e., the lifting problem falls within the scope of homotopy theory. See [15, 17, 27, 28].

The condition we need on p is that it should be a fibering, in some generalized sense. Now I need to talk about fiberings. Let $p\colon E\to B$ be a map. One example to bear in mind is that in which E is a Möbius band, B is a circle, and p identifies each meridian on the band to a point. We say that p is a locally trivial fibering, with fibre F, if there is an open cover $\{U_\alpha\}$ of B and homeomorphisms

$$\varphi_\alpha\colon F\times U_\alpha\to p^{-1}U_\alpha$$

such that $p\varphi_\alpha(f,u)=u$ for all α, all $f\in F$ and all $u\in U_\alpha$. That is, p must behave locally like the projection of a product onto one factor.

Numerous examples of fiberings arise in differential geometry. If we have a smooth manifold M of dimension m, then its space of tangent vectors gives a fibering over M, with fibre $F=R^m$. Similarly if we consider tensors of some specified type. Again, if we have a submanifold M embedded in some larger manifold N, with say a Riemannian structure, then we can consider the normal vectors to M in N. They form a fibering over M. For all these reasons, topologists who work with smooth manifolds have to be expert at manipulating fiberings.

Now suppose we are given B and F and wish to classify the possible fiberings over B with fiber F. This has been reduced to a problem

in homotopy-theory [28] and the answer can be calculated in practical cases; indeed this is a standard practice in the topology of smooth manifolds.

Next I must talk about cobordism. As originally posed by Thom [29], the problem went like this. We call two compact smooth m-manifolds M, M' cobordant if there is a compact smooth $(m+1)$-manifold with boundary whose boundary is the disjoint union of M and M'. This gives an equivalence relation, which divides manifolds into cobordism classes. The problem is to calculate how many cobordism classes there are. Using the theory of fiberings, Thom showed how to reduce his problem to a problem in homotopy-theory, and he also solved the resulting homotopy-theoretic problem. Since then, it has been shown that Thom's ideas apply to many related problems [4, 5, 8, 12, 20, 21, 33]. We may say that the solution of cobordism-type problems is now standard practice.

Finally, we may suppose that in any research into manifolds, if our investigation has any geometric quality, we are likely to run into questions about the position of submanifolds, their intersection and so forth; and in such questions, homotopy-theory will probably continue to be an indispensable tool. See [26, 34].

So far I have talked mainly about problems; now let me talk about methods. Suppose given a problem, say, on the existence of a map. It may happen that the situation is so favourable that we can calculate everything in sight. If not, then basically we face a choice of two methods.

(i) If we guess that the required map exists, then we can try to find an explicit geometrical construction which constructs the required map.

(ii) If we guess that the required map does not exist, then we can try to find some topological invariant which would be involved in a contradiction if the required map did exist.

As for explicit geometrical constructions, many such constructions have been introduced into homotopy-theory. I would like to mention the Whitehead product [16, 36], the Hopf construction [35] and the Toda bracket [30, 32].

As for topological invariants, many of them can be obtained by specializing the sets $[X, Y]$. So let us return to these sets $[X, Y]$. Sometimes I will speak as if the problem is to compute these sets $[X, Y]$; indeed this is something we often need to do. At least let us see what we can say about them.

First, "naturality". Given two maps $f: X \to Y$ and $g: Y \to Z$, we can compose them and obtain $gf: X \to Z$. The homotopy class of gf depends only on the classes of f and g. So composition with f gives a function

$$f^*: [Y, Z] \to [X, Z],$$

and composition with g gives a function

$$g_*: [X, Y] \to [X, Z].$$

Secondly, we can often give the set $[X, Y]$ some sort of algebraic structure. We have many cases in which we can usefully make $[X, Y]$ into a group (but in some cases, of course, there is no useful way of doing so). For example, suppose $X = S^1$; and suppose that we have base-points x_0, y_0 in X, Y and that we are only considering maps and homotopies which preserve the base-points. Then we can make $[S^1, Y]$ into a group: for a function $f: S^1 \to Y$, preserving base-points, is a closed path in Y starting and finishing at y_0; and given two such paths f, g, we define their product to be the path which first traces out f and then traces out g. In this way we get the classical fundamental group $\pi_1(Y)$. Similarly, by specializing to the case $X = S^n$, we obtain the homotopy group $\pi_n(Y) = [S^n, Y]$. For example, $\pi_n(S^n) = Z$.

Next I need to talk about Eilenberg-MacLane spaces. Suppose given an abelian group π and an integer n. We say that Y is an Eilenberg-MacLane space of type (π, n) if

$$\pi_r(Y) = \begin{cases} \pi & \text{if } r = n \\ 0 & \text{if } r \neq n,\ r > 0. \end{cases}$$

For any π and n there is a space of type (π, n). Some of them exist in nature; for example the circle S^1 is a space of type $(Z, 1)$; but mostly they have to be constructed artificially.

Now consider $[X, Y]$, where X satisfies reasonable restrictions and Y is of type (π, n). This set is a group, and it is independent of the choice of Y; in fact it is exactly the classical cohomology group $H^n(X; \pi)$ (which was originally defined rather differently). For example, we have

$$H^n(S^r; \pi) = [S^r, Y] = \begin{cases} \pi & \text{if } r = n \\ 0 & \text{if } r \neq n,\ r > 0. \end{cases}$$

But this was a digression. I need to talk about further properties which the sets $[X, Y]$ have. Suppose given a fibering $p: E \to B$, and let $i: F \to E$ be the injection of the fibre over the base-point of B. Then we have the following sets and functions

$$[X, F] \xrightarrow{i_*} [X, E] \xrightarrow{p_*} [X, B].$$

In $[X, B]$ we have a distinguished element 0, namely the class of the constant map at the base-point. We suppose that X satisfies reasonable

conditions. Then it is a basic theorem that this sequence is exact, in the sense that $p_*\alpha = 0$ if and only if $\alpha \in \operatorname{Im} i_*$. Indeed under suitable conditions this short exact sequence grows into a long exact sequence; for example

$$\ldots \to \pi_n(F) \xrightarrow{i_*} \pi_n(E) \xrightarrow{p_*} \pi_n(B) \xrightarrow{\partial} \pi_{n-1}(F) \xrightarrow{i_*} \ldots .$$

This sequence is exact at every group, in the sense just described.

Such exact sequence are very helpful in calculations and proofs. For this purpose, we find a small number of useful fiberings in nature. For example, we have a fibering $R^1 \to S^1$ with fiber Z and a fibering $S^3 \to S^2$ with fibre S^1. One can use these natural fiberings to deduce a few results, e.g. $\pi_3(S^2) = Z$. But the supply of natural fiberings soon runs out, so we turn to artificial ones.

I will illustrate this point from suspension-theory. Freudenthal originally showed that $\pi_{n+r}(S^n) \cong \pi_{n+r+1}(S^{n+1})$ if $r < n-1$. This is an archetypal theorem in homotopy-theory; we meet a lot of phenomena which are independent of some dimensional parameter n, provided that n is large enough. Such phenomena are called stable; the opposite word is unstable. For example, cobordism gives rise to problems of homotopy-theory which are stable.

Serre and G. W. Whitehead approached this subject as follows [24, 35]. Given a space X, we can consider the function-space LX of maps l: $[0, 1] \to X$ such that $l(0)$ is the base-point x_0 in X. We have a projection p: $LX \to X$ given by $p(l) = l(1)$ (take the end-point of each path). This projection is a fibering (in a suitable sense); its fibre is the loop-space ΩX, that is, the space of maps l: $[0, 1] \to X$ such that $l(0) = l(1) = x_0$. We easily see that $\pi_r(\Omega X) \cong \pi_{r+1}(X)$. So we can prove suspension theorems by comparing S^n with ΩS^{n+1}.

James and Toda went further [18, 19, 31]. It is a theorem of James that if n is odd, there is a fibering $F \to E \to B$ in which F is (near enough) S^n, E is ΩS^{n+1} and B is ΩS^{2n+1}. Thus we get an exact sequence valid for all dimensions, not merely a result for a restricted range of dimensions

$$\begin{array}{ccccccccc} \ldots \to \pi_r(S^n) & \xrightarrow{i_*} & \pi_r(\Omega S^{n+1}) & \xrightarrow{p_*} & \pi_r(\Omega S^{2n+1}) & \xrightarrow{\partial} & \pi_{r-1}(S^n) \to \ldots \\ & \searrow & \| & & \| & \nearrow & \\ & & \pi_{r+1}(S^{n+1}) & \to & \pi_{r+1}(S^{2n+1}) & & \end{array}$$

The other theorems of James and Toda in this direction are in a similar spirit.

This method, if combined with the method of explicit geometric constructions, is an exceedingly effective means of calculating

specific facts [32]. If, for example, one were forced to calculate $\pi_{25}(S^2)$, this would be the preferred method. On the other hand, the method has some disadvantages.

(1) It seems to be more suited to detailed calculation than to the discovery of systematic general phenomena.

(2) If one is interested in stable phenomena (which are in some sense the easier case), then this method (if followed in its original purity) only allows one to get at the stable groups by wading through the unstable ones first.

(3) Although the low-grade suspension theorems (which work for a certain range of dimensions) are available for any space X, the high-grade ones (which work for all dimensions) are restricted to the case of spheres. One problem will reveal the depth of our ignorance. A Moore space is a space with a single non-vanishing cohomology group (e.g., a sphere). Can one obtain exact sequences modelled on Toda's, but replacing the spheres by Moore spaces (say for the group Z_p)?

At the beginning of this section, I said that if we have a fibering $F \xrightarrow{i} E \xrightarrow{p} B$, we can get an exact sequence

$$[X, F] \xrightarrow{i_*} [X, E] \xrightarrow{p_*} [X, B].$$

There is a dual situation. Given a map f: $A \to X$, we can form a space $M = X \smile_f CA$ by taking the cone on A and attaching it to X using the map f. Then we have the following sets and functions.

$$[A, Y] \xleftarrow{f^*} [X, Y] \xleftarrow{i^*} [M, Y].$$

This sequence is exact. Under suitable conditions this short sequence grows into a long exact sequence; for example

$$\ldots \leftarrow H^n(A; \pi) \xleftarrow{f^*} H^n(X; \pi) \xleftarrow{i^*} H^n(M; \pi) \xleftarrow{\delta} H^{n-1}(A; \pi) \leftarrow \ldots,$$

This sequence is exact at each group.

To preserve the analogy, we call

$$A \xrightarrow{f} X \xrightarrow{i} X \smile_f CA$$

a *cofibering*. This concept is important, because many spaces can be built up in this way. For example, starting from spheres and building up by iterated cofiberings, we obtain the class of finite CW-complexes [37]. The fact that so many cofiberings exist in nature is one reason why we can calculate effectively with cohomology.

Originally, Eilenberg and Steenrod gave an axiomatization of cohomology theory, which included the exact sequence given above as a crucial axiom [14]. Later, there was some interest in the so-called extraordinary cohomology theories; these were theories which satisfied all but one of the Eilenberg-Steenrod axioms, including this exact sequence [13]. For example, if we start from the fiberings over X whose fibres are real or complex vector-spaces, we can construct certain groups $K^n(X)$ [7]; and with these groups we can calculate quite effectively [3, 4]. All these extraordinary cohomology theories have the form $[X, Y]$ for some fixed Y. Indeed, under reasonable restrictions we know necessary and sufficient conditions for a functor of X to have the form $[X, Y]$ for some fixed Y [10].

We have said that if we look at the homotopy groups of a fibering, or the cohomology groups of a cofibering, we get an exact sequence. We can also do something more difficult, and look at the cohomology groups of a fibering or the homotopy groups of a cofibering. In both cases we get something more complicated, a spectral sequence. We can approach matters like this. From one map

$$A \xrightarrow{f} X$$

we get an exact cohomology sequence. Now suppose given an infinite sequence of spaces and maps

$$X_0 \xrightarrow{f_0} X_1 \xrightarrow{f_1} X_2 \to \ldots \to X_n \xrightarrow{f_n} X_{n+1} \to \ldots .$$

We can form a lot of composites

$$f_n f_{n-1} \ldots f_{n-r},$$

and each such composite gives an exact cohomology sequence. So we get a large collection of exact sequences, which fit together in a certain way. However, it is possible to keep track of all the information and condense it into a usable form; and the result is an algebraic structure called a spectral sequence. For example, if $F \xrightarrow{i} E \xrightarrow{p} B$ is a fibering, then the cohomology groups $H^*(F)$, $H^*(E)$ and $H^*(B)$ are related by a spectral sequence [23].

A spectral sequence is a somewhat complicated structure, and does not provide a guarantee of effective calculation; but the experts succeed with it more often than not.

Our general philosophy would now run as follows. To calculate $[X, Y]$, we try to decompose X as an iterated cofibering, in which the successive factors are elementary from the point of view of cohomology; and we try to decompose Y as an iterated fibering, in which the successive factors are elementary from the point of view of homotopy. The first can always be done. The second can be done following ideas of Postnikov [22]. Using the methods of the French school [11], we always have enough fiberings; but since they are large function-space fiberings, the problem is to identify the spaces involved; this is usually done by cohomology calculations.

Using this sort of method, Serre has proved some very useful general theorems [25]. For example, the theorems show that under reasonable conditions, $[X, Y]$ is a finitely-generated group. They often allow us to show that $[X, Y]$ is finite, when that happens to be true; they may also allow us to say what primes can occur in the order of $[X, Y]$. At the worst we expect to be able to calculate $[X, Y] \otimes Q$, where Q means the ring of rational number.

For some purposes, variations of the classical French method may be useful [1, 20]. Recent work indicates that it may sometimes be profitable to calculate with K-theory, and so split off factors which are still comparatively elementary, but are not products of Eilenberg-MacLane spaces [4].

It is, of course, very dangerous to make pronouncements like this, but the state of affairs in homotopy-theory seems to be that most of the basic principles are known. The last really unexpected and unforeseeable advance was the Bott periodicity theorems [6, 9], and the consequent possibility of effective calculation with K-theory. It may be that a benevolent Nature has further such uncovenanted mercies up her sleeve; but if so, then in the nature of things nobody knows where to look for them. In their absence, the prospect is one of more complicated and strenuous use of the basic principles, and of honourable service in applications outside homotopy-theory. Inside homotopy-theory, the most exciting theorems to me are those which show the existence of systematic phenomena—orderly patterns existing in Nature. But when you contemplate some of the tabulated data, orderly patterns are hard to find. However, there are some results which keep up a small amount of hope in this direction. Outside homotopy-theory, the study of smooth manifolds already uses a fair amount of homotopy-theory; and any subject with a geometric flavour is a candidate for applications.

University of Manchester,
Manchester, England

REFERENCES

[1] Adams J. F., On the structure and applications of the Steenrod algebra, *Comment. Math. Helvetici*, **32** (1958), 180-214; *MR* 20 (1958), 2711.
[2] Adams J. F., On the non-existence of elements of Hopf invariant one, *Annals of Math.*, **72** (1960), 20-104; *MR* 25 (1963), 4530.
[3] Adams J. F., Vector fields on spheres, *Annals of Math.*, **75** (1962), 603-632; *MR* 25 (1963), 2614.
[4] Anderson D. W., Brown E. H., Peterson F. P., On spin cobordism, to appear.

[5] Atiyah M. F., Bordism and cobordism, *Proc. Cambridge Phil. Soc.*, **57** (1961), 200-208; *MR* 23A (1962), 4150.
[6] Atiyah M. F., Bott R., On the periodicity theorem for complex vector bundles, *Acta Math.*, **112** (1964), 229-247; *MR* 31 (1966), 2727.
[7] Atiyah M. F., Hirzebruch F., Vector bundles and homogeneous spaces, Proc. Symposia in Pure Math., Vol. 3, American Math. Soc. (1961), 7-38; *MR* 25 (1963), 2617.
[8] Авербух Б. Г., Алгебраическое строение групп внутренних гомологий, *Доклады АН СССР*, **125**, № 1 (1959), 11-14; *MR* 23A (1962), 2204.
[9] Bott R., The stable homotopy of the classical groups, *Annals of Math.*, **70** (1959), 313-337; *MR* 22 (1961), 987.
[10] Brown E. H., Cohomology theories, *Annals of Math.*, **75** (1962), 467-484; *MR* 25 (1963), 1551.
[11] Cartan H., Serre J.-P., Espaces fibres et groupes d'homotopie, I, II. *C.R. Acad. Sci. Paris*, **234** (1952), 288-290, 393-395; *MR* 13 (1952), 675.
[12] Conner P. E., Floyd E. E., Differentiable periodic maps, Springer, 1964; *MR* 31 (1966), 750.
[13] Dold A., Halbexacte homotopie Funktoren, Springer, 1966.
[14] Eilenberg S., Steenrod N. E., Foundations of algebraic topology, Princeton, 1952; *MR* 14 (1953), 398.
[15] Hilton P. J., An introduction to homotopy theory, Cambridge, 1953; *MR* 15 (1954), 52.
[16] Hilton P. J., On the homotopy groups of the union of spheres, *J. London Math. Soc.*, **30** (1955), 154-172; *MR* 16 (1955), 847.
[17] Hu S.-T., Homotopy theory, Academic Press, 1959; *MR* 21(1960), 3186. Русский перевод: Ху Сы-цзян, Теория гомотопий, М., ИЛ, 1964.
[18] James I. M., The suspension triad of a sphere, *Annals of Math.*, **63** (1956), 407-429; *MR* 18 (1957), 58.
[19] James I. M., On the suspension sequence, *Annals of Math.*, **65** (1957), 74-107; *MR* 18 (1957), 662.
[20] Milnor J. W., On the cobordism ring Ω^* and a complex analogue, I, *American Jour. Math.*, **82** (1960), 505-521; *MR* 22 (1961), 9975.
[21] Новиков С. П., О некоторых задачах топологии многообразий, связанных с теорией пространств Тома, *ДАН СССР*, **132** (1960), 1031-1034; *MR* **22** (1961), 12545.
[22] Постников М. М., Определение групп гомологий пространства с помощью гомотопических инвариантов, *ДАН СССР*, 76, № 3 (1951), 359-365. О классификации непрерывных отображений, *ДАН СССР*, **79**, № 4 (1951), 575-576. О гомотопическом типе полиэдров, *ДАН СССР*, **76**, № 6 (1951), 789-791.
[23] Serre J.-P., Homologie singuliere des espaces fibres, *Annals of Math.*, **54** (1951), 425-505; *MR* 13 (1952), 574.
[24] Serre J.-P., Sur la suspension de Freudenthal, *C. R. Acad. Sci. Paris*, **234** (1952), 1340-1342; *MR* 13 (1952), 675.
[25] Serre J.-P., Groupes d'homotopie et classes de groupes abeliens, *Annals of Math.*, 58 (1953), 258-294; *MR* 15 (1954), 548.
[26] Smale S., A survey of some recent developments in differential topology, *Bull. American Math. Soc.*, **69** (1963), 131-145; *MR* 26 (1963), 1896.
[27] Spanier E. H. Algebraic topology, McGraw-Hill, 1966.
[28] Steenrod N. E., The topology of fibre bundles, Princeton, 1951; *MR* 12 (1951), 522. Русский перевод: Стинрод Н., Топология косых произведений, ИЛ, М., 1953.
[29] Thom R., Quelques proprietes globales des varietes differentiables, *Comment. Math. Helvetici*, **28** (1954), 17-86; *MR* 15 (1954), 890.

[30] Toda H., Some relations in homotopy groups of spheres, *Jour. Inst. Polytech. Osaka City Univ.*, **2** (1952), 43-82; *MR* 14 (1953), 572.
[31] Toda H., On the double suspension E^2, *Jour. Inst. Polytech. Osaka City Univ.*, 7 (1956), 103-145; *MR* 19 (1958), 1188.
[32] Toda H., Composition methods in homotopy groups of spheres, Princeton, 1962; *MR* 26 (1963), 777.
[33] Wall C. T. C., Determination of the cobordism ring, *Annals of Math.*, **72** (1960), 292-311; *MR* 22 (1961), 11403.
[34] Wall C. T. C., Topology of smooth manifolds, *Jour. London Math. Soc.*, 40 (1965), 1-20; *MR* 30 (1965), 2524.
[35] Whitehead G. W., On the Freudenthal theorems, *Annals of Math.*, **57** (1953), 209-228; *MR* 14 (1953), 1110.
[36] Whitehead J. H. C., On adding relations to homotopy groups, *Annals of Math.*, **42** (1941), 409-428; *MR* 2 (1941), 323.
[37] Whitehead J. H. C., Combinatorial homotopy, *Bull. American Math. Soc.*, **55** (1949), 213-245; *MR* 11 (1950), 48.

THE ETALE TOPOLOGY OF SCHEMES

M. ARTIN

1. Introduction

Since Weil [56, 57] pointed out the need for invariants, analogous to topological ones, of varieties over fields of characteristic p, several proposals to define such invariants have been made, notably by Serre [47], Grothendieck [21], and Monsky and Washnitzer [38]. I would like to describe some of the recent work on one of these approaches, that of the *etale cohomology* of Grothendieck. This approach has yielded a proof of the rationality of Weil's zeta function for a variety over a finite field via the method suggested by Weil [57], and for generalized L-functions (Grothendieck [25]) [1]. The etale cohomology also provides a framework in which to state some beautiful conjectures of Tate [53] on algebraic cycles (now proved by him for divisors on abelian varieties over finite fields), and of Birch and Swinnerton-Dyer (cf. Tate [54]). Quite generally, it gives good results for coefficients prime to the characteristic p of the variety. In fact, the other proposals for a cohomology theory (Serre [47], Monsky and Washnitzer (cf. [37] or Lubkin [36]), Grothendieck's flat topology (cf. [12] or Shatz [51] for the case of a field)) all yield a cohomology with "mod p" or Witt vector coefficients, and it is not completely clear at present which of them will be the most fruitful. The problem of finding such a theory is obviously of great interest.

In this talk, I will restrict myself because of lack of time and competence to a description of some aspects of the etale theory, without going into detail on any of the applications mentioned above.

Let me begin by recalling that a morphism $X \to Y$ of schemes is called *etale* if it is flat and unramified. Those unfamiliar with the notion may get an intuitive understanding of its meaning from the fact that a map of schemes of finite type over the complex numbers is etale iff the map of associated analytic spaces is a local isomorphism.

---

[1] The methods also yield the functional equation and the explicit form of the zeta function as an alternating product (cf. [25]). Actually, the rationality was first proved for arbitrary varieties by Dwork [13]. For the rest, we prefer not to get involved in questions of priority. Suffice it to say that in addition, similar results have been obtained for a smooth proper variety which is a specialization from characteristic zero by Lubkin [35, 36], and that the rationality has been proved for arbitrary varieties by Monsky [38].

2. The etale topology

The first topology to be defined on an abstract variety or scheme was the *Zariski topology* (Zariski [59]). Recall that in this topology a closed set of the spectrum of a ring R is the set of zeros $V(S)$ of some subset S of R. Then Serre, in his fundamental paper FAC [45], showed that the Zariski topology could be used to define a cohomology theory of *coherent* sheaves on a variety, i.e., ones arising from modules over the coordinate rings. He also proved ([46] GAGA) that for a projective variety over the field of complex numbers, the theory thus obtained was the same as the analytic theory. These results left little doubt that the Zariski topology is a good one for the study of coherent sheaves.

For the purposes of our discussion, we may express this conclusion in a slightly different way by saying that ordinary localization in a ring R is a "sufficiently strong" process for most things in the study of modules over R. The conclusion is supported by the fact that if R is a local ring (say noetherian for simplicity), and if M, N are two finite modules over R which become isomorphic after any finitely generated faithfully flat extension of scalars $R \to R'$, then M and N are themselves isomorphic. Or, there are no twisted forms of a finite module M over a local ring R, relative to such extensions of scalars. (By descent theory (Grothendieck [23]), this can be interpreted as a generalized form of the famous Hilbert theorem 90.)

However, twisted forms of more complicated structures will usually not be locally trivial. For instance, central separable algebras over a field k are twisted forms of a full matrix algebra relative to the extension $k \to \bar{k}$ of k by its separable algebraic closure. Such examples led Serre [48] to introduce the notion of local isotriviality of a fibre space. A fibre space with given algebraic structure group G over a scheme X is called *locally isotrivial* if for every point $x \in X$ there is a Zariski open neighborhood U of x in X and a finite etale covering space U' of U such that the pull-back of the fibre space to U' is trivial. This definition yields a notion which includes the one studied by Weil [58] of fibre spaces which are locally trivial for the Zariski topology, and the one of structures over a field which become trivial after a separable algebraic field extension, studied by Lang and Tate [34] and others.

In 1958, Grothendieck found a general version of sheaf theory, which enabled him to define the notion of etale cohomology of schemes. This etale theory puts Serre's notion in a broad framework, and it provides an algebraic definition of the Betti numbers of an algebraic variety. We will describe briefly one version of the general theory. It is treated in detail in Verdier [55].

The necessary data for sheaf theory consist of the following:

(1) A category C and a collection of families of maps $\{X_i \to Y\}$ of C with common range, called *coverings* of the range Y. The following axioms are supposed to hold:
   (i) Isomorphisms are coverings.
   (ii) The composition of coverings is a covering, in the following sense: If $\{X_i \to Y\}$ cover Y, and $\{W_{ij} \to X_i\}_j$ is a covering of X_i for each i, then the compositions $\{W_{ij} \to Y\}$ cover Y.
   (iii) A pull-back of a covering is a covering: If $\{X_i \to Y\}$ is a covering, and $Y' \to Y$ is an arbitrary map, then the fibred products $X_i \times_Y Y'$ exist in C and they form a covering of Y'.

Actually, the exact phrasing of the axioms is not very important (cf. [3, 5, 55]). We will refer to such a collection of data as a *topology*.

Given a topology, a *sheaf* is a contravariant functor F from C to (say) sets, satisfying the *sheaf axiom*.

(2) If $\{X_i \to Y\}$ is a covering, then a "section" $s \in F(Y)$ is uniquely determined by a collection of sections $s_i \in F(X_i)$ such that for each i, j the sections of $F(X_i \times_Y X_j)$ induced by s_i and s_j via the projection maps are equal.

Cohomology with values in an abelian sheaf is defined as a derived functor, as in Grothendieck [20].

The Serre notion of local isotriviality was the starting point for Grothendieck's original definition of the etale topology, but it has turned out in the meantime to be more convenient to allow localization by an arbitrary etale morphism. Thus for the *etale topology* of a prescheme X, the category C above is taken to be the category of preschemes U *etale* over X, and a covering is a family of maps which is *surjective* in the sense that the range is covered by the images.

This definition is such that the first cohomology $H^1(X, G)$ of X with values in a linear group G classifies the fibre spaces with structure group G over X having the following property: For every $x \in X$ there is an etale map $U \to X$ (not necessarily finite over a Zariski open set) whose image contains x, such that the pull-back of the fibre space to U is trivial; or, as one says, which are locally trivial for the etale topology (cf. Giraud [16] for a general treatment of H^1).

3. Relations with Galois cohomology

If X is the spectrum of a field K, then the etale schemes of finite type over X are just spectra of separable (commutative) K-algebras, i.e., products of separable field extensions, and it is not difficult to show that the resulting cohomology theory is just Tate's cohomology of the galois group $G(\bar{K}K)$ w/here $\bar{K}$ is the separable algebraic closure of K. This theory has been treated in detail in various places (eg., Serre [50]).

The sheaf theory on an arbitrary noetherian scheme X can also be related to galois modules via the *specialization diagram* of X. We will describe it (per semplicità di discorso) only for an entire (this terminology is due to Lang [33]) normal scheme of dimension 1. This includes the case of a nonsingular algebraic curve and that of the spectrum of the ring of integers in a number field. The general case can be described in a similar way, but the specialization diagram is more complicated:

Let G be the galois group of the separable closure $\bar{K}$ of the function field K of X, and let $\bar{X}$ be the normalization of X in $\bar{K}$. For each $x \in X$, choose a point $\bar{x}$ of $\bar{X}$ above x, and let $D_x \subset G$ be the decomposition group of this point. (A change of the point $\bar{x}$ over x changes D_x by conjugation. It is an interesting feature of the etale cohomology, and one of its weaknesses, that the choices of the various points $\bar{x}$ are not important.) Let G_x^* be the galois group of the separable algebraic closure of the residue field $k(x)$ of X at x. Then there is a diagram of group homomorphisms

$$G \leftarrow D_x \to G_x$$

for each $x \in X$. The result is that the category of "constructible" abelian sheaves on X is equivalent with the category whose objects consist of

(1) (i) A G-module M and a G_x-module M_x for each $x \in X$, which are finitely generated abelian groups.
   (ii) A "specialization map" Φ_x: $M_x \to M$ for each $x \in X$ which is a homomorphism of D_x-modules,

satisfying the "continuity condition" that almost all of the maps Φ_x be isomorphisms.

Thus the cohomology of a sheaf on X in the etale topology can be described in terms of the galois cohomologies of the groups G, G_x, and of the relations between them. In this way one can interpret the results of Ogg [40] and Šafarevič [44] on the cohomology of abelian varieties over function fields as calculations in the etale cohomology. Their results contain implicitly a description of the cohomology of an algebraic curve over an algebraically closed field (cf. [6], exp. IX). The formula of Ogg [40] for the Euler characteristic of a sheaf has been generalized by Grothendieck (Raynaud [42]). Similarly, the exact sequences of Tate [52] for cohomology of a number field are closely related to the etale cohomology of the ring of integers of K, but there is a slight difference in the notion of local triviality used there.

4. Cohomology with values in the multiplicative group

The sheaf of units on a scheme X for the etale topology occupies a central role. We will denote this sheaf by $\mathcal{O}^*$. It contains as subsheaf the sheaf μ of all roots of unity, and for a regular X, the inclusion

$\mu \subset \mathcal{O}^*$ induces an isomorphism on cohomology in dimensions > 2, if one ignores p-torsion for the residue characteristics p of X. The sheaf μ is clearly a locally constant torsion sheaf (ignoring p) and so its cohomology can be treated by the theory discussed in the next section. But in dimensions $\leqslant 2$, the cohomology of $\mathcal{O}^*$ gives information of an arithmetic sort not contained in μ:

It follows from descent theory [23] that

$$H^1(X, \mathcal{O}^*) = \operatorname{Pic} X$$

is the group of isomorphism classes of locally free rank one sheaves on X. On the other hand, the group $H^2(X, \mathcal{O}^*)$ contains as subgroup the "Brauer group" of sheaves of Azumaya algebras on X (cf. [26]):

$$H^2(X, \mathcal{O}^*) \supseteq \operatorname{Br} X.$$

(The notion of *Azumaya algebra*, generalizing that of central simple algebra over a field, was first introduced for rings by Azumaya [10] and Auslander and Goldman [9], and its relation to cohomology theory was studied by various authors [2, 11, 43]. The theory is discussed in detail in Grothendieck [26].)

A most interesting and apparently difficult question of Auslander and Goldman is whether or not the Brauer group is all of $H^2(X, \mathcal{O}^*)$ when X is the spectrum of a regular ring (or more generally, when X is a regular scheme). This is true when X is of dimension $\leqslant 2$ (Auslander and Goldman) or when X is a semi-local ring of an algebraic variety in characteristic zero (cf. [26]). It is generally false if X is singular (Grothendieck [26]).

To see the importance of the Brauer group, suppose that X is a complete non-singular algebraic surface over an algebraically closed field k. For any X and n prime to the characteristics of X, the *Kummer sequence*

$$(1) \qquad 0 \to \mu_n \to \mathcal{O}^* \xrightarrow{n\text{-th power}} \mathcal{O}^* \to 0$$

is exact, where μ_n denotes the sheaf of n-th roots of unity. Applying (1) and the above facts to our surface X in the highest interesting dimension, we obtain an exact sequence

$$(2) \qquad 0 \to (\operatorname{Pic} X)/n \to H^2(X, \mu_n) \to (\operatorname{Br} X)_n \to 0$$

where the subscript n indicates the set of elements whose order divides n. Its middle term is the cohomology of X with values in the constant sheaf $\mu_n \approx \mathbf{Z}/n$ whose rank as a $\mathbf{Z}/n$-module is, up to a bounded term, the second Betti number B_2 of X (say by definition). The term on the left yields up to a bounded term the rank of the Neron-Severi group of X. Thus the Brauer group measures the algebraic analogue $\rho_0 = B_2 - \rho$ of the number of transcendental 2-cycles on a surface.

The inequality $B_2 \geqslant \rho$ was first proved in the abstract case by Igusa [31] with an ad hoc definition of B_2. His method of proof, using a pencil of curves on the surface and vanishing cycle theory, made no use of the etale cohomology, but a similar approach gives an expression for the Brauer group, and hence for the etale B_2, in terms of the pencil ([3], Tate [54]). The Brauer group is just the Šafarevič-Tate group of locally trivial principal homogeneous spaces of the Jacobian of the generic curve. This is also true for arithmetic surfaces (Tate [54]).

5. General cohomology theory

The approach to etale cohomology has been mostly via Grothendieck's generalized sheaf theory, as we have already indicated. Actually, the first case I know of in which etale coverings were used for cohomology theory of a variety is in Kawada and Tate [32].

The results of this section (due largely to Grothendieck and myself), together with proofs, may be found in [6]. The most important single result is the following:

Theorem (1) (proper base change theorem). Let $f\colon X \to Y$ be a proper map and F an abelian torsion sheaf on X. Let y_0 be a geometric point of Y, and X_0 the geometric fibre of f at y_0. Then the stalk at y_0 of the higher direct image $R^q f_* F$ is the cohomology $H^q(X_0, F|X_0)$ of the fibre. The assumption that F be a torsion sheaf is essential in all serious results.

For schemes of finite type over the complex numbers $\mathbf{C}$, one has

Theorem (2) (comparison with the classical cohomology). Let X be a scheme of finite type over $\mathbf{C}$, and F a constructible torsion sheaf on X for the etale topology. Then F includes a sheaf on X for the classical topology, and one has isomorphisms

$$H^q(X_{\text{etale}}, F) \approx H^q(X_{\text{class}}, F).$$

The condition of constructibility is the obvious finiteness condition in this context. The proof of (2) in the general case requires resolution of singularities (Hironaka [28]).

For passing from characteristic zero to characteristic p, the following is useful (cf. [6], also Lubkin [35]):

Theorem (3): (specialization theorem). Let $f\colon X \to Y$ be a smooth proper map, and let F be a locally constant torsion sheaf on X whose orders are prime to the residue characteristics of Y. Then the higher direct images $R^q f_* F$ are locally constant sheaves on Y (whose stalks are by (1) the cohomology of the geometric fibres of X/Y).

This result is one of a series reflecting the locally acyclic nature of smooth morphisms.

In a less definitive state are the results on cohomological dimension and finiteness of cohomology:

Theorem (4) (finiteness). Let $f\colon X \to Y$ be a morphism of finite type, and F a constructible torsion sheaf on X. Suppose either that f is proper or that Y is an excellent (cf. [27], IV) scheme of characteristic zero. Then the higher direct images $R^q f_* F$ are again constructible. Of course, this theorem gives in particular the finiteness of the cohomology groups when $Y = \operatorname{Spec} K$ is the spectrum of a separably closed field. Assuming resolution, one can prove (4) also in the case that Y is excellent and of equal characteristics, and that F is of orders prime to the characteristics (which is a necessary assumption). But very little is known about the cohomology in the unequal characteristic case, even in low dimensions where resolution is available (Abyankhar [1]), say when X has dimension 2.

The correct upper bound for cohomological dimension of a variety over a field can be proved from the theorems on cohomological dimension for fields of Grothendieck and Tate (cf. [50]). We denote by $\operatorname{cd}_l X$ the largest integer q such that $H^q(X, F) \neq 0$ for some l-torsion sheaf F:

Theorem (5): (cohomological dimension).

(i) Let X be a scheme of finite type over a separably closed field k. Then

$$\operatorname{cd}_l X \leqslant 2 \dim X.$$

If X is affine, then

$$\operatorname{cd}_l X \leqslant \dim X.$$

(ii) Let X be a scheme of finite type over the ring of integers of a number field K. Assume that either $l \neq 2$ or that K is totally imaginary. Then

$$\operatorname{cd}_l X \leqslant 2 \dim X + 1.$$

Here $\dim X$ is the Kronecker dimension. A much lower bound actually holds in (i) when l is equal to the characteristic, and for $l = 2$ in (ii) the totally imaginary number field can be replaced by $\mathbf{Q}$ if one adds to X in a formal way a "fibre at infinity" (cf. [8]). A number of other variants are treated in [6]. Again, little is known in the unequal characteristic case. Thus, for instance, the cohomology with values in $\mathbf{Z}/l$ of the scheme obtained from $\operatorname{Spec} \mathbf{Z}_p[[t]]$ by removing the locus $\{t = 0\}$ is not known, nor is the cohomological dimension of the field of fractions of $\mathbf{Z}_p[[t]]$.

6. The fundamental group

As usual, the fundamental group classifies covering spaces: Suppose that the scheme X is connected and *locally connected* for the etale topology, i. e., that every etale scheme U over X is a disjoint union of connected components. Suppose moreover that a geometric point of X is given (X is *pointed*). Then a *pro-group* (cf. [23]) $\pi_1(X)$ is defined in such a way that it classifies etale covering spaces of X, i. e., schemes X' over X which are twisted forms for the etale topology of the "trivial covering" $\coprod_S X$ of X (S is a set). It does so in the sense that such schemes X' correspond canonically to homomorphisms from $\pi_1(X)$ to the permutation group of S. If X is geometrically unibranch ([27], IV), a connected covering space X' of this type is necessarily finite over X, and so the fundamental group is pro-finite and equal to the one introduced by Grothendieck [22]. In general, a scheme may have infinite covering spaces, and the fundamental group obtained is somewhat larger (cf. Lubkin [35], Grothendieck [24]).

Almost all of our explicit information about the fundamental group still comes from the "Riemann existence theorem". It asserts that a finite topological covering of a scheme X of finite type over $\mathbf{C}$ has a unique algebraic structure, i. e., that the etale fundamental group of X (say X is geometrically unibranch) is the profinite completion of the fundamental group of X:

$$\pi_1(X_{\text{etale}}) = \widehat{\pi_1(X_{\text{class}})}.$$

The general form of this theorem requires the results of Grauert and Remmert [18] and GAGA [46] (cf. [6]). Its importance for the etale theory is indicated for instance by the fact ([6], X) that a nonsingular variety over $\mathbf{C}$ has a Zariski open covering by $K(\pi, 1)$ spaces (ones having π_1 as only non-vanishing homotopy group).

Although Grothendieck, in his beautiful paper GFGA [22], succeeded in computing the "tame" fundamental group of an algebraic curve over an arbitrary field, the proof made use of the Riemann existence theorem in the classical case, and there is still no algebraic proof known. The difficulties which present themselves are very interesting.

Suppose for instance that X is obtained from the affine line by removing some points p_i ($i = 1, \ldots, n$), so that the tame fundamental group is free on n generators. Then the freeness can be expressed by the assertion that

$$\text{(1)} \qquad \operatorname{Hom}(\pi_1, G) \approx \coprod_i \operatorname{Hom}(D_i, G)$$

where G is a finite "test group" of order prime to the characteristic, and where D_i is the decomposition subgroup of a suitably chosen point above p_i.

In fact, using Grothendieck's technique one can show by algebraic methods alone that in the above situation the ramification can be assigned arbitrarily at the points p_i, *up to inner automorphism*. By this we mean that the map

$$\text{Hex}(\pi_1, G) \to \coprod_i \text{Hex}(D_i, G) \tag{2}$$

is surjective, where Hex (A, B) is the set Hom (A, B) modulo conjugation by inner automorphisms of B. Here the choice of D_i no longer matters. This assertion is quite useful for cohomological questions, which are not very sensitive to conjugation (cf. section 3), but of course it is much weaker than (1). On the other hand, it has a chance to be true in characteristic p without restriction on the order of G.

7. Homotopy theory

In order to generalize the Riemann existence theorem to the higher homotopy groups, one needs a good way to associate something like a simplicial set to a scheme, and because of the difficulties inherent in the Čech procedure, it is not immediately clear how to do this. A way was first found by Lubkin [35]. Subsequently Verdier [55], using an idea of Cartier, found another method, and working along the lines suggested by the Cartier-Verdier approach, Quillen [41] has developed a homotopy theory for arbitrary categories.

The exact definitions are too technical to give here. Using them, one can associate to a connected and locally connected, pointed scheme X a *pro-object* in the homotopy category H of connected pointed CW-complexes, which represents the *homotopy type* of the scheme for the etale topology (cf. Lubkin [36]). Let us denote this pro-object by X_{et}. Its relation to the classical topology can be described as follows (cf. Artin and Mazur [7]):

Call a CW-complex *homotopy finite* if all of its homotopy groups are finite groups. Then one can associate to an object K in the homotopy category H a *pro-finite completion* $\hat{K}$ which is a formal inverse system of homotopy finite CW-complexes, i. e., a pro-object in the homotopy category of such complexes. The completion $\hat{K}$ is characterized by the property that any map from K to a homotopy finite complex factors through $\hat{K}$.

The natural extension of the Riemann existence theorem is the following result:

Theorem (1). Let X be a pointed geometrically unibranch scheme of finite type over the field of complex numbers. Denote by

X_{cl} its homotopy type for the classical topology. Then

$$X_{\text{et}} = \widehat{X_{\text{cl}}}.$$

It is in general not true that the homotopy groups of K are the pro-finite completions of the homotopy groups of K. However, this is true if K is simply connected and has finitely generated homotopy groups in each dimension [7]. Thus if in the above theorem X_{cl} is simply connected, then $\pi_q(X_{\text{et}}) = \widehat{\pi_q(X_{\text{cl}})}$ for each q.

The above method gives a definition of homotopy for quite general schemes X. For instance, the scheme Spec $\mathbf{Z}$ with a point at infinity added in a formal way has the homotopy type of a Moore space $K'(\mathbf{Z}/2, 2)$ with the single nonvanishing homology group $\mathbf{Z}/2$ in dimension 2. However, it is likely that a good homotopy theory for Spec $\mathbf{Z}$ should allow for homotopy groups of a twisted sort, such as roots of unity.

8. Henselian rings

The study of local properties of schemes for the etale topology leads to a series of questions of an interesting sort which I want to mention briefly. The local ring of a scheme X at a geometric point x, in the etale topology, is the ring

$$R = \varinjlim_{(X', x')} \Gamma(X', O_{X'})$$

where (X', x') runs through schemes X' etale over X with chosen geometric point x' over x. The most striking property of these rings is that they are *henselian* (i. e., that Hensel's lemma holds). It seems clear that the notion of henselian ring, introduced by Azumaya [10] and studied by Nagata [39], will play an important role in any detailed study of local phenomena.

Let me recall the general outline, proposed by Grothendieck in the introduction to his Elements [27], for treating certain types of questions about schemes. I have rephrased it slightly for my purposes:

Step 1. One compares a global problem with the corresponding local one for the etale topology.

Step 2. By a limit argument, the local problem is reduced to a question about henselian rings R.

Step 3. One may replace the henselian ring R by its completion $\hat{R}$.

Step 4. The complete local ring is related to the artinian rings $\hat{R}/\mathfrak{m}^n$ ($\mathfrak{m} = \text{rad } \hat{R}$), and the study of these rings is reduced by infinitesimal methods to a series of questions about the field $R/\mathfrak{m}$ (which

are perhaps "classical"). Actually, none of these steps is under complete control, exept for 2 which is generally trivial (cf. Grothendieck [27]). Step 1 may sometimes be treated by descent theory (Grothendieck [23]), and step 4 was discussed by Serre in his Stockholm talk [49].

The somewhat novel point I want to bring out is step 3, which has not received much attention, but which seems promising. It is one aspect of the general algebraization problem of relating henselian rings to their completions. This has been studied for the divisor class group by Hironaka [29] and for algebraic extensions of rings in [45]. An interesting example is obtained from the question of the existence of a section of a map $f: X \to Y$ of schemes. The corresponding algebraization problem, which would handle step 3 in this case, is the following:

Suppose that X is a scheme of finite type over a henselian ring R which has a "formal section", i. e., a point with values in $\hat{R}$. What assumptions are needed to assure that one can approximate the formal section by points with values in R?

For a discrete valuation ring R, this problem was solved recently by Greenberg [19] and Raynaud. Mild restrictions on R suffice.

In a slightly different direction is the theorem of Grauert, Hironaka and Rossi [17], [30] to the effect that analytic local rings with isolated singular points whose completions are isomorphic are themselves isomorphic. This theorem has an algebraic analogue in characteristic zero, which asserts that two algebraic varieties with isolated singular points whose local rings have isomorphic completions are locally isomorphic for the etale topology. (This is already a striking change from the Zariski topology in the case of simple points. For them it follows immediately from the Jacobian criterion.) The related conjecture of Grauert that any complete local ring with isolated singularity is "algebraic", i. e., is the completion of a local ring of an algebraic variety, is however still open except in low dimensions [4], [29].

Dept. of Mathematics,
Massachusetts Institute of Technology,
Cambridge, USA

REFERENCES

[1] Abhyankhar S., Resolution of singularities of arithmetical surfaces, Purdue conference on arithmetical algebraic geometry, Harpers (1965).
[2] Amitsur S. A., Simple algebras and cohomology groups of arbitrary fields, *Transactions Amer. Math. Soc.*, **90** (1959).

[3] Artin M., Grothendieck topologies, Mimeographed notes, Harvard (1962).
[4] Artin M., Algebraic extensions of local rings, *Rend. di Mat.* (to appear).
[5] Artin M., Etale coverings of schemes over Hensel rings, *Amer. J. Math.* (to appear).
[6] Artin M., Grothendieck A., Cohomologie étale des schémas, Sém. Inst. Hautes. Et. Sci., Bures-sur-Yvette (1963-64).
[7] Artin M., Mazur B. (in preparation).
[8] Artin M., Verdier J.-L., Etale cohomology of number fields, Mimeographed notes, Woods Hole (1964).
[9] Auslander M., Goldman O., The Brauer group of a commutative ring, *Transactions Amer. Math. Soc.*, **97** (1960).
[10] Azumaya G., On maximally central algebras, *Nagoya Math. J.*, **2** (1951).
[11] Chase S. U., Harrison D. K., Rosenberg A., Galois theory and cohomology of commutative rings, Memoir Amer. Math. Soc., **52** (1965).
[12] Demazure M., Séminaire de géométrie algébrique. Schémas en groupes, expose 4, Inst. Hautes Et. Sci., Bures-sur-Yvette (1963-64).
[13] Dwork B., On the rationality of the zeta function of an algebraic variety, *Amer. J. Math.*, **82** (1960).
[14] Gabriel P., Zisman M., Séminaire homotopique, Inst. Math. de Strasbourg (1963-64).
[15] Giraud J., Analysis situs, Sem. Bourbaki, No. 256 (1962-63).
[16] Giraud J., Cohomologie non abélienne (mimeographed notes), Columbia University (1966).
[17] Grauert H., Über Modifikationen und exceptionelle analytische Mengen, *Math. Ann.*, **146** (1962).
[18] Grauert H., Remmert R., Komplexe Räume, *Math. Ann.*, **136** (1958).
[19] Greenberg M., Rational points in Henselian discrete valuation rings, *Bulletin Amer. Math. Soc.* (to appear).
[20] Grothendieck A., Sur quelques points d'algèbre homologique, *Tohoku Math. J.*, **9** (1957).
[21] Grothendieck A., Cohomology theory of abstract algebraic varieties, Proc. Int. Congr. Math. Edinburgh (1958).
[22] Grothendieck A., Géométrie formelle et géométrie algébrique, Sem. Bourbaki, No. 182 (1958-59).
[23] Grothendieck A., Technique de descente et théorèmes d'éxistence en géométrie algébrique, Sem. Bourbaki, No. 190, 195 (1959-60).
[24] Grothendieck A., Séminaire de géométrie algébrique, Schémas en groupes, expose 10, Inst. Hautes Et Sci., Bures-sur-Yvette (1963-64).
[25] Grothendieck A., Formule de Lefschetz et rationalité des fonctions L., Sem. Bourbaki, No. 279 (1964-65).
[26] Grothendieck A., Le groupe de Brauer, I, II, Sem. Bourbaki, No. 290 and 297 (1964-65).
[27] Grothendieck A., Dieudonné J., Éléments de géométrie algébrique, Pub. Math. Inst. Hautes Et. Sci. No. 4, 7(1959).
[28] Hironaka H., Resolution of singularities of an algebraic variety over a ground field of characteristic zero, I, II, *Annals of. Math.*, **27** (1964).
[29] Hironaka H. (in preparation).
[30] Hironaka H., Rossi H., On the equivalence of imbeddings of exceptional complex spaces, *Math. Ann.*, **156** (1964).
[31] Igusa J.-I., Betti and Picard numbers of abstract algebraic surfaces, *Proc. Nat. Acad. Sci.*, **46** (1960).

[32] Kawada Y., Tate J., On the Galois cohomology of unramified extensions of function fields in one variable, *Amer. J. Math.*, 77 (1955).
[33] Lang S., Algebra, Addison-Wesley (1965). Русский перевод в печати.
[34] Lang S., Tate J., Principal homogeneous spaces over Abelian varieties, *Amer. J. Math.*, 80 (1958).
[35] Lubkin S., On a conjecture of A. Weil, *Amer. J. Math.* (to appear).
[36] Lubkin S., A p-adic proof of Weil's conjectures, Mimeographed notes, Inst. Adv. Study (1966).
[37] Monsky P., Washnitzer A., The construction of formal cohomology sheaves, *Proc. Nat. Acad. Sci.*, 52 (1964).
[38] Monsky P. (in preparation).
[39] Nagata M., Local rings, New York (1962).
[40] Ogg A. P., Cohomology of Abelian varieties over function fields, *Annals of Math.*, 76 (1962).
[41] Quillen D. (in preparation).
[42] Raynaud M., Caracteristique d'Euler-Poincaré d'un faisceau et cohomologie des variétés abéliennes, Sem. Bourbaki, No. 286 (1964-65).
[43] Rosenberg A., Zelinski D., On Amitsur's complex, *Transactions Amer. Math. Soc.*, 97 (1960).
[44] Шафаревич И. Р., Главные однородные пространства, определенные над полем функций, Труды Мат. института им. В. А. Стеклова, LXIV (1961).
[45] Serre J-P., Faisceaux algébriques cohérents, *Annals of Math.*, 61 (1955).
[46] Serre J-P., Géométrie algébrique et géométrie analytique, *Ann. Inst. Fourier Grenoble*, 6 (1955-56).
[47] Serre J.-P., Sur la topologie des variétés algébriques en caractéristique p, Sympos. top. alg., Mexico (1956).
[48] Serre J.-P., Espaces fibrés algébriques. Sem. Chevalley. Anneaux de Chow et applications, expose 1 (1958).
[49] Serre J.-P., Géométrie algébrique. Proc. Int. Congr. Math. Stockholm (1962).
[50] Serre J.-P., Cohomologie galoisienne, Lecture notes in math., No. 5, Springer (1965).
[51] Shatz S., The cohomological dimension of certain Grothendieck topologies, *Annals of Math.*, 83 (1966).
[52] Tate J., Duality theorems in Galois cohomology over number fields, Proc. Int. Congr. Math. Stockholm (1962).
[53] Tate J., Algebraic cycles and poles of zeta functions, Purdue conference on arithmetical algebraic geometry, Harpers (1965).
[54] Tate J., On the conjectures of Birch and Swinnerton-Dyer and a geometric analogue, Sem. Bourbaki, No. 306 (1965-66).
[55] Verdier J.-L., Cohomologie étale des schémas, Sem. Inst. Hautes Et. Sci., Exposés 1-6, Bures-sur-Yvette (1963-64).
[56] Weil A., Numbers of solutions of equations in finite fields, *Bulletin Amer. Math. Soc.*, 55 (1949).
[57] Weil A., Abstract versus classical algebraic geometry, Proc. Int. Congr. Math. Amsterdam (1954).
[58] Weil A., Fibre spaces in algebraic geometry, Mimeographed notes by A. Wallace, Chicago (1955).
[59] Zariski O., On the compactness of the Riemann manifold of an abstract field of algebraic functions, *Bulletin Amer. Math. Soc.*, 50 (1944).

GLOBAL ASPECTS OF THE THEORY OF ELLIPTIC DIFFERENTIAL OPERATORS

MICHAEL F. ATIYAH

Introduction

The subject matter of this talk lies in the area between Analysis and Algebraic Topology. More specifically, I want to discuss the relations between the analysis of *linear* partial differential operators of *elliptic* type and the algebraic topology of *linear* groups of *finite*-dimensional vector spaces. I will try to show that these two topics are intimately related, and that the study of each is of great importance for the development of the other.

The theory of elliptic differential equations has of course a long and rich history, with its origins in the study of the Laplace equation and of the closely-associated Cauchy-Riemann equations. Its connection with topology, via the theory of holomorphic functions and Riemann surfaces, is equally classical. Its development in the last fifty years or so has however followed two rather separate courses.

On the one hand there has been the purely *analytical development*, the qualitative study of general elliptic operators. Here the emphasis has been on extending the basic theory of the Laplace operator to general operators of the same type—what we now call elliptic operators. The questions studied include regularity of solutions, boundary conditions and more recently the extension to suitable classes of integro-differential operators—now called pseudo-differential operators. On the whole this sort of work was carried out for domains in Euclidean space, though the extension to more general manifolds presents nothing essentially new.

The second development has been the more detailed or quantitative study of the classical operators and their associated structures. This essentially includes the whole of algebraic geometry treated by topological and transcendental methods. The pioneering work in this field was of course done by Hodge some thirty years ago.

Roughly speaking we might say that the analysts were dealing with complicated operators and simple spaces (or were only asking simple questions), while the algebraic geometers and topologists were only dealing with simple operators but were studying rather general manifolds and asking more refined questions.

In recent times, the last five years or so, some serious attempts have been made to integrate these two different developments. Each

seems now to have reached such a stage of maturity that it can confidently offer its services to the other half. For example, some of the ideas and techniques developed in the general theory of partial differential equations have been very successfully applied to the study of complex manifolds (cf. [12]). My own interests, however, have been in the reverse direction and I would like to spend the rest of my time discussing the Riemann-Roch or Index Problem.

1. The Riemann-Roch theorem

The classical Riemann-Roch theorem is concerned with giving a formula for the dimension of the space of meromorphic functions on a compact Riemann surface having poles of orders $\leqslant \nu_i$ at points P_i. If $P = \Sigma \nu_i P_i$ then the dimension $l(P)$ is given by:

$$l(P) - i(P) = \deg P - g + 1$$

where $\deg P = \Sigma \nu_i$. Here $i(P)$ is in effect $l(Q)$ for a suitable Q so that what is computed is a difference of two numbers of the same sort. This is in the nature of the problem because as we vary P (i. e. if we vary P_i) the number $l(P)$ can jump, but the difference $l(P) - i(P)$ remains constant. Moreover one can, in certain circumstances, prove that $i(P) = 0$ (this happens if $\deg P > 2g - 2$) and one then has a genuine formula for $l(P)$.

The Riemann-Roch theorem is one of the basic theorems of algebraic geometry. It is an example of a quantitative or "refined" result. Considerable effort was devoted to extending it to higher dimensions, and success was achieved first by Hirzebruch [9] in 1954 and then (purely algebraically) by Grothendieck [6] in 1957.

On the other hand analysts had been independently studying the "index problem" for elliptic operators [8]. If D is an elliptic operator on a compact manifold (without boundary for simplicity) then the space of solutions of $Du = 0$ is finite-dimensional. If one wants a formula for this dimension $l(D)$ one finds that $l(D)$ can jump, but that

$$\text{index } D = l(D) - l(D^*)$$

(where D^* is the adjoint problem) is constant under continuous variation of D. The problem therefore is to find a formula for index D. The analogy with the Riemann-Roch theorem is obvious. Moreover since holomorphic functions are solutions of $\bar{\partial} u = 0$ we can easily set up the Riemann-Roch problem as an index problem. To solve the index problem in general is therefore to extend the Riemann-Roch theorem from the domain of holomorphic function theory to that of general elliptic systems.

In low dimensions, when the number of independent variables is 1 or 2, explicit answers were obtained by fairly elementary methods. In general, however, one is faced with two serious problems:

(A) We have to find suitable topological invariants of the pair (X, D) where X is the base manifold and D the elliptic operator.

(B) We have then to find the explicit formula for index D in terms of these invariants.

For example in the case of the classical Riemann-Roch theorem the topological invariants are just the genus g and $\deg(P)$.

2. Topology of the linear groups

For the classical structures (Riemannian and complex) an extensive theory of topological invariants, called characteristic classes, has been developed (cf. [9]). These classes are all generalizations of the Euler number, i. e. the number of singularities of a vector field. Roughly speaking, one considers the cycles where a given number of vector fields becomes linearly dependent. Fundamentally these homology invariants stem from the homology of the linear groups $GL(n, \mathbf{R})$ and $GL(n, \mathbf{C})$. The Riemann-Roch theorem of Hirzebruch gives a formula in terms of these characteristic classes, the actual formula being a very complicated one going back to Todd and involving the generating function $x/(1 - e^{-x})$ of the Bernoulli numbers.

It is a remarkably fortunate fact that an elliptic operator D also defines invariants of the characteristic class type. It is not difficult to see how these arise. Let us recall that a homogeneous constant coefficient $N \times N$ matrix of differential operators

$$P = \left[P_{ij}\left(\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_n} \right) \right], \quad i, j = 1, \ldots, N,$$

is elliptic if $\xi \neq 0$ (and real) $\Rightarrow \det P(\xi_1, \ldots, \xi_n) \neq 0$. Then P defines a map $\xi \to P(\xi)$ of $S^{n-1} \to GL(N, \mathbf{C})$, where S^{n-1} is the unit sphere in R^n. This shows at once that the homotopy and hence the homology of $GL(N, \mathbf{C})$ enters into the study of elliptic operators. Now according to a fundamental theorem of Bott [7] the homotopy groups

$$\pi_{n-1}(GL(N, \mathbf{C}))$$

are 0 for n odd and isomorphic to the integers for n even (provided $2N \geqslant n$). Thus if n is even and $2N \geqslant n$, P defines an integer which may be called its degree. This is a generalization of the obvious degree

in case $N = 1$, $n = 2$ and it may be computed explicitly as the value of an integral $\int_{S^{n-1}} \omega(P)$ where $\omega(P)$ is a differential expression in P generalizing the well-known formula $\frac{1}{2\pi i}\int_{S^1}\frac{dP}{P}$.

Using these invariants of a general elliptic operator D (together with the ordinary characteristic classes of X) one can then give an explicit formula for index D which is remarkably similar to that occurring in the Riemann-Roch formula. This is not accidental and has a quite deep significance. Very roughly one can say that, as far as the topology goes, the classical operators are just as complicated as the most general ones so that the Riemann-Roch formula gives a fair indication of the general case.

Let me make some very general remarks about the nature of the proof [1]) of the general index theorem. First of all, when X is a simple space like a sphere we can use Bott's theorem to deform D into a standard operator whose index may be computed directly. In the case of a complicated X we embed X in a sphere S and construct an elliptic operator D' on S with

$$\text{index } D = \text{index } D'.$$

We are then reduced to the preceding case. It is important to note that even if we start with a nice D on X (e. g., coming from a complex structure) it is not in general possible to get a nice D'. Thus it is not possible to restrict oneself to nice or classical operators. To prove the Riemann-Roch theorem for arbitrary compact complex manifolds one needs to go outside this category. The reason why the case of (projective) algebraic varieties is easier is because one can use the very special holomorphic embedding $X \subset P_N(\mathbf{C})$, whereas in general we have no alternative but to use the non-holomorphic embedding $X \subset S$.

3. Integral formulae

Characteristic classes have a representation by differential forms which is a generalization of the Gauss-Bonnet formula

$$E = \int K$$

expressing the Euler number as the integral of the scalar curvature (suitably normalized). The integral formula for the (local) degree

[1]) There are now two proofs of the index theorem. The first, modelled on Hirzebruch's proof of the Riemann-Roch theorem, appears in [4], [15]. The econd, which is closer to the work of Grothendieck, will appear in [5]. The emarks here refer to the second proof.

of an elliptic operator mentioned in § 2 is of the same type. Using such expressions it is possible in principle (though very complicated in practice) to express the index D as an integral $\int_X \omega(D)$. Here $\omega(D)$ depends on the coefficients of D and on a choice of Riemannian metric on X. Although this formula is algebraically complicated it is analytically fairly simple, in the sense that it involves only the first few derivatives of the coefficients of D (and the metric).

There is, however, an entirely different and purely analytical approach to the index problem which leads to another formula for index D as an integral. Unfortunately this formula is extremely complicated and it involves approximately n derivatives where n is the dimension of X. Only for low values of X it is easy to identify it with the curvature-type formula given by the other method.

This analytical method is essentially very classical except that no one seems to have considered using it on the index problem. The idea is as follows. Suppose first that Δ is a self adjoint positive elliptic operator. Then, by the spectral theorem, we can define Δ^{-s} for $s \in \mathbf{C}$ and consider the "Zeta-function"

$$\zeta(s) = \text{Trace } \Delta^{-s} = \Sigma\lambda^{-s}$$

where λ runs over the (discrete) eigenvalues of Δ. This converges if Re (s) is large and $\zeta(s)$ can be analytically continued as a meromorphic function on the entire s-plane. Moreover $s = 0$ turns out not to be a pole, and the value $\zeta(0)$ can be computed explicitly in terms of Δ. For the Laplace operator these results are due to Minakshisundaram and Pleijel [14]. Their extension to the general case can be done by pseudo-differential operator techniques (cf. [16]). Alternatively $\zeta(0)$ can be interpreted as the constant term in the asymptotic expansion of trace $(e^{-\Delta t})$ as $t \to 0$, where $e^{-\Delta t}$ (for $t > 0$) denotes the solution of the generalized Heat equation

$$\left(\frac{\partial}{\partial t} + \Delta\right) u = 0.$$

Suppose now that D is elliptic. Introducing metrics we can consider D^* the adjoint of D. Then

$$\Delta_0 = 1 + D^*D, \quad \Delta_1 = 1 + DD^*$$

are positive self-adjoint operators. Let

$$\zeta_i(s) = \text{Trace } \Delta_i^{-s}, \qquad i = 1, 2.$$

Then it is not difficult to show that

$$\zeta_0(s) - \zeta_1(s) = \text{index } D$$

is independent of s. Hence putting $s = 0$ and using the explicit formula mentioned we get a formula for index D. In fact one can use other integer values of s, besides 0. Each will give a formally different expression of index D as an integral. However $s = 0$ is the simplest.

4. Generalizations

4.1. Boundary problems. For elliptic operators with "coercive" boundary conditions [11] one also has an index problem. It turns out that these boundary conditions have a deep topological significance. For instance, if D admits any coercive boundary conditions then the local degree of D must be zero. In view of this it is not surprising that one ends up (cf. [2]) with an explicit index formula of much the same type as in the case of manifolds without boundary. Moreover, as a by-product, the examination of the topological meaning of the boundary conditions led to a new and elementary proof [1] of the basic periodicity theorem of Bott for $\pi_t(GL(N, \mathbf{C}))$.

4.2. Lefschetz fixed-point formula. Suppose $f: X \to Y$ is a differentiable map which "commutes" with a given D (one may think of a holomorphic map as the typical example). Then we can define [1]) a kind of "Lefschetz number":

$$L(f) = \mathrm{Trace}\,(f \mid \mathrm{Ker}\, D) - \mathrm{Trace}\,(f \mid \mathrm{Coker}\, D).$$

If f has isolated fixed points of multiplicity ± 1 one has a formula of the following type [3]

$$L(f) = \Sigma \nu(P)$$

where P runs over the fixed points of f and $\nu(P)$ is a complex number depending only on the differential $(df)_P$. This may be regarded as a generalization of the classical Lefschetz fixed-point formula. The proof is by the method of § 3, using Zeta-functions

$$\zeta(s) = \mathrm{Trace}\,(\Delta^{-s} \circ f^*)$$

which depend on D and f. The point is that because of the hypothesis on the fixed points of f it turns out that these Zeta-functions have no poles and $\zeta(0)$ is then very easily computed. In fact it depends only on f and not on Δ.

4.3. Group-situations. Assume G is a compact group of automorphisms of (X, D). Then Ker D and Coker D are G-modules and $g \to L(g)$ is a virtual character of G. Since the characters of G are a discrete set we can again use deformation methods and one

[1]) Coker $D \cong \mathrm{Ker}\, D^*$ but this involves a metric and f need not preserve a metric. Thus f acts naturally on Coker D, but not on Ker D^*.

obtains a theorem that expresses $L(g)$ as a sum over the fixed components of g, each term being an integral of similar type to that occurring in the index formula. These formulae (and also the formula of (4.2)) applied to a homogeneous space $X = G/H$ include the Hermann Weyl character formula. They also lead (cf. [10]) to the Langland's formula [13] for dimensions of spaces of automorphic forms.

4.4. Real operators. So far we have discussed only invariants which are integers or complex numbers and these may, if we wish, be expressed as integrals. It is therefore interesting to point out that there are analytical invariants which are not of this type. Thus let D be a *real skew-adjoint* elliptic operator. Then index $D = 0$ is not interesting but it is not difficult to see that dim (Ker D) mod 2 is a deformation invariant. This is a consequence of the fact that the non-zero eigenvalues of D occur in conjugate pairs $(\pm i\mu)$. It turns out that this invariant can be viewed as a kind of index and by methods similar to those above we can express this mod 2 invariant in terms of suitable invariants. Here what are relevant are the homotopy groups $\pi_t(GL(N, \mathbf{R}))$. For N large these are (cf. [7]) periodic with period 8 and π_t is of order 2 if $t \equiv 0$ or mod 8. These mod 2 homotopy invariants give what is required.

Conclusion

I have presented things so far in the form of topology being used to assist in computing an analytic invariant. The relation of the topology to the analysis however is very intimate indeed—as I have indicated in connection with boundary problems. There are in fact parts of the theory where the analysis has to be used to help the topology. For example the G-equivariant homotopy classes of maps

$$S(V) \to GL(W),$$

where V, W are representation spaces of G (with W "large") and $S(V)$ denotes the unit sphere of V can be determined by use of elliptic operators. So far there is no other method.

The situation here is analogous to that in Morse's theory of critical points. In the first place one would tend to say that the number of critical points of a real-valued function f on a manifold X could be estimated in terms of the topological invariants of X. The theory has however been used extensively the other way round: one constructs suitable functions and obtains information on X from the critical points of X. All the modern structure theory of differentiable manifolds is based on this point of view.

I would like to conclude with some philosophical remarks. In my view the close relation between topology of linear groups and the analysis of elliptic operators stems from the following three basic properties they have in common:

(i) Linearity
(ii) Stability (under deformation)
(iii) Finiteness.

University of Oxford,
Mathematical Institute,
Oxford, England

REFERENCES

[1] Atiyah M. F., Bott R., On the periodicity theorem for complex vector bundles, *Acta Math.*, **112** (1964), 229-247.
[2] Atiyah M. F., Bott R., The index problem for manifolds with boundary, Bombay Colloquium on Differential Analysis (Oxford University Press), 1964, 175-186.
[3] Atiyah M. F., Bott R., A Lefschetz fixed point formula for elliptic differential operators, *Bull. Amer. Math. Soc.*, **72** (1966), 245-250.
[4] Atiyah M. F., Singer I. M., The index of elliptic operators on compact manifolds, *Bull. Amer. Math. Soc.*, **69** (1963), 422-433.
[5] Atiyah M. F., Singer I. M., The index of elliptic operators, I (to appear).
[6] Borel A., Serre J.-P., Le théorème de Riemann-Roch, *Bull. Soc Math. France*, 86 (1958), 97-136.
[7] Bott R., The stable homotopy of the classical groups, *Ann. of Math.*, **70** (1959), 313-337.
[8] Гельфанд И. М., Об эллиптических уравнениях, *Успехи матем. наук*, **15**, № 3 (1960), 121-132. Russian Math. Surveys, **15**, № 3 (1960), 113-123.
[9] Hirzebruch F., Topological methods in algebraic geometry, Springer, 1966.
[10] Hirzebruch F., Elliptische Differentialoperatoren auf Mannigfaltigkeiten, Weierstrass Festband, Westdeutscher Verlag, Opladen, 1966.
[11] Hörmander L., Linear partial differential operators, Springer, 1964. (Русский перевод: Хёрмандер Л., Линейные дифференциальные операторы с частными производными, «Мир», М., 1965.)
[12] Hörmander L., Introduction to complex analysis in several variables, Van Nostrand, 1966. (готовится к печати русский перевод.)
[13] Langlands R. P., The dimension of spaces of holomorphic forms, *Amer. J. Math.*, 85 (1963), 99-125.
[14] Minakshisundaram S., Pleijel A., Eigenfunctions of the Laplace operator on Riemannien manifolds, *Canadian J. Math.* (1949), 242-256.
[15] Palais R., Seminar on the Atiyah-Singer index theorem, *Annals of Math. Studies*, **57** (Princeton, 1965).
[16] Seeley R. T., Fractional powers of elliptic operators (to appear).

DYNAMIC PROGRAMMING AND MODERN CONTROL THEORY

RICHARD BELLMAN

1. Introduction

One of the fundamental concepts in mathematics is that of transformation. The study of the unfolding over time of a physical process leads naturally to investigations of the effects of the repetition of a transformation, which is to say to the study of multistage processes. Much of classical and contemporary analysis stems from this source: iteration, ergodic theory, the theory of semigroups [1], the theory of branching processes [2], random transformations at fixed times and deterministic transformations at stochastic times [3, 4].

We wish to indicate still another direction of research, that of multistage decision processes. What happens when we allow a choice of the transformation to be employed at each time? As we shall see, in addition to raising many new questions and developing some new techniques to answer them, we shall also make repeated contacts with classical areas of analysis. The point is that many processes can be interpreted to be of the foregoing type. This is the fundamental "as if" property of mathematics.

Since the term "mathematical theory of multistage decision processes" is rather unwieldy, we have coined the shorter, but no less cryptic, "dynamic programming". Detailed accounts of what we sketch so briefly below will be found in the books [5, 6, 7, 8], as well as in the papers cited.

2. Dynamic programming

Let us now describe a dynamic programming process of discrete, deterministic type. Let p be a point in a space S, called the state space, and $T(p, q)$ be a transformation with the closure property that $p_1 = T(p, q)$ belongs to S whenever q belongs to a space D, called the decision, or control, space. The variable q is called the decision or control variable.

Consider now a sequence of q's, $q_1, q_2, \ldots, q_N$, which generate a corresponding sequence of points in S,

$$(1) \qquad p_1 = T(p, q_1), \; p_2 = T(p_1, q_2), \; \ldots, \; p_N = T(p_{N-1}, q_N).$$

We ask that these control variables be chosen so as to maximize a prescribed scalar function

$$(2) \qquad R_N = R(p, p_1, \ldots, p_N, q_1, q_2, \ldots, q_N).$$

In this general formulation, there is little that can be said profitably. To obtain some interesting problems, we must impose some structure upon R. Let us take it to have a separable form

$$R_N = h(p, q_1) + h(p_1, q_2) + \ldots + h(p_{N-1}, q_N) + k(p_N). \tag{3}$$

Fortunately, many significant problems in mathematics and applications arise naturally in this form which lends itself to analysis. In a number of interesting processes involving "stop rules", N itself depends upon the p_i and q_i. We shall not discuss these matters here; see [8, 9].

Let us assume either that all variables are discrete, which makes the existence of the maximum immediate, or that appropriate continuity conditions have been imposed. Observing that the maximum value depends on p, the initial state, and N, the number of transformations, or stages, let us introduce the sequence of functions $\{f_N(p)\}$, $p \in S$, $N = 1, 2, \ldots$, defined by the relation

$$f_N(p) = \max_{\{q_i\}} R_N. \tag{4}$$

This is an imbedding technique. We have imbedded a particular optimization problem within a family of related problems. If this is done correctly, we can obtain useful relations connecting different members of the family. In this way, we build a bridge between the simple problems and the difficult ones. We see that

$$f_1(p) = \max_{q_1} [h(p, q_1) + k(p_1)], \tag{5}$$

$$\begin{aligned} f_N(p) &= \max_{[q_1, q_2, \ldots, q_N]} [\ldots] = \\ &= \max_{q_1} \max_{[q_2, \ldots, q_N]} [\ldots] = \\ &= \max_{q_1} [h(p, q_1) + \max_{[q_2, \ldots, q_N]} [\ldots]] = \\ &= \max_{q_1} [h(p, q_1) + f_{N-1}(T(p, q_1))], \end{aligned}$$

for $N \geqslant 2$.

We thus possess an inductive technique for deducing the properties of f_N from that of the simpler function f_1. The case where $N = \infty$, an unbounded process, is particularly interesting. It leads to the novel nonlinear functional equation

$$f(p) = \max_{q} [h(p, q) + f(T(p, q))]. \tag{6}$$

3. Analytic aspects

There are now many possible directions to proceed. An important first step involves the study of the existence and uniqueness of solutions. In some cases, this can be done with relative ease [5]; in other cases, it is extremely helpful to introduce a carefully chosen metric, the Birkhoff metric [10]; in still other cases, it is "catch as catch can". Of interest are also questions concerning the convergence of the sequence $\{f_N(p)\}$, and the rate of convergence.

It is also important, both intrinsically and for approximation purposes, to find families of functions which satisfy (2.6). To begin with, it is useful to obtain functions $f(p, a)$, dependent on a vector parameter a, with the invariance property

$$f(p, a) = \max_{q} [h(p, q) + f(T(p, q), b)]. \tag{1}$$

For example, if p and q are N-dimensional vectors, $T(p, q) = Bp + q$, and $h(p, q)$ is a quadratic form in p and q, then we have a relation of the foregoing type with $f(p, a)$ a quadratic form in p,

$$f(p, a) = (p, ap), \tag{2}$$

and a a symmetric matrix. This relation occupies a central role in modern control theory; see [8, 11, 12, 32].

We can generate other families of functions with this reproducing property by use of the maximum transform. In the scalar version

$$M(u) = U(x) = \max_{y \geqslant 0} [u(y) - xy]. \tag{3}$$

This transform has the convenient unravelling property

$$M(u * v) = M(u) + M(v), \tag{4}$$

where

$$u * v = \max_{y_1 + y_2 = x} [u(y_1) + v(y_2)], \quad y_1, y_2 \geqslant 0. \tag{5}$$

This transform is closely related to duality aspects of dynamic programming processes we shall mention again below; see [13, 14].

4. Approximation in policy space

Referring to the equation of (2.6), we see that it involves two functions, $f(p)$, the maximum value, and $q(p)$, the decision function. We shall call this second function the policy function. If $f(p)$ is known, we can immediately determine $q(p)$ via the maximization. If $q(p)$

is known, we obtain $f(p)$ by iteration,

(1) $$f(p) = h(p, q) + h(T(p, q), q(T)) + \dots .$$

A solution of (2.6) can thus be effected by means of a determination of $f(p)$ or equally by a determination of $q(p)$, the optimal policy. This is an extremely important observation since it enables us to employ a type of approximation not available in classical analysis, approximation in policy space. Conventionally, we can approach the solution of (2.6) by means of a method of successive approximations, using a sequence $\{f_N(p)\}$ such as that defined by (1.5). However, as just noted, we possess the additional freedom of approximating to the policy function $q(p)$. If we guess $q_1(p)$, we obtain $f_1(p)$ as the soluton of

(2) $$f_1(p) = g(p, q_1) + f_1(T(p, q_1)),$$

obtained by direct iteration. One way to obtain a second approximation, $q_2(p)$, is to ask for the function which furnishes the maximum of

(3) $$g(p, q) + f_1(T(p, q)).$$

This has a very simple intuitive flavor. We try to improve our first decision, while reconciling ourselves to a continuation in the original fashion. Having obtained $q_2(p)$ in this fashion, we form $f_2(p)$ via

(4) $$f_2(p) = g(p, q_2) + f_2(T(p, q_2)),$$

where f_2 is again obtained via iteration. We now continue in this fashion, obtaining a sequence of policies $\{q_n(p)\}$ and a sequence of return functions $\{f_n(p)\}$.

Our reason for focussing upon policies is that the analytic structure of policies is quite often far simpler than that of the return functions. This is closely related to the Fermat principle of least time, Huygen's principle, and so forth [27]. Policies often possess simple intuitive flavor which can be rapidly ascertained from the original decision process.

5. Positive operators

From the derivation of q_2, we see that

(1) $$f_1(p) = g(p, q_1) + f_1(T(p, q_1)) \leqslant g(p, q_2) + f_1(T(p, q_2)).$$

Can we conclude from the equality of (4.4) and the inequality above that

(2) $$f_1(p) \leqslant f_2(p)?$$

This question establishes contact with the theory of positive and monotone operators. If f_1 and f_2 are two functions satisfying the relation

(3) $$T(f_1) \geqslant T(f_2)$$

where T is a given transformation, when can we conclude that $f_1 \geqslant f_2$?

A particular class of problems of this type involves ordinary and partial differential operators and is related both to the theory of differential inequalities inaugurated by Caplygin and Lyapunov [15, 16], and to the modern maximum principles of partial differential equations. Results of this type play an important role in approximation methods and computational approaches in general.

6. Quasilinearization

The flexibility inherent in an equation such as (2.6) prompts one to see whether or not it is possible to convert other types of nonlinear functional equations not at all related to dynamic programming into a similar form. If so, we can easily obtain sequences of successive approximants which converge monotonically.

For example, consider the Riccati equation

(1) $$u' = u^2 + a(t), \quad u(0) = c,$$

basic in the study of the second order linear differential equation and thus important in quantum mechanics. We can write this in the form

(2) $$u' = \max_v [2uv - v^2 + a(t)]$$

and consider $v(t)$ to be a pseudo-policy. That $v = u$ makes approximation particularly easy. Suppose we consider the associated linear equation

(3) $$U' = 2Uv - v^2 + a(t), \quad U(0) = c$$

where v is a fixed function of t. Using a simple result in the theory of differential inequalities, we can assert that

(4) $$u = \max_v U(v, t);$$

see [28]. This, and related results, have been systematically used by Calogero in connection with the study of quantum mechanical scattering [17].

The foregoing approach can also be used as a basis of successive approximations which coincides with the Newton-Raphson-Kantorovich technique in a number of cases. See [18] for a discussion of the theory of quasilinearization based upon the application of the foregoing ideas to general classes of nonlinear functional equations. A detailed discussion of differential inequalities based upon early work of Collatz will be found in [19].

Particularly interesting is the application of these ideas to the study of ordinary and generalized solutions of the equation

(5) $$u_t = g(u, u_x),$$

which we shall meet again below in connection with the calculus of variations. Results originally obtained in [29] essentially by divination can be obtained in a relatively straightforward way; see [18, 30]. This method has been extensively developed in some recent papers [31].

7. Realizations and applications

It is interesting and useful to apply these ideas to the study of planning and decision processes outside of mathematics, in the economic, engineering, and industrial worlds, and to the parts of mathematical analysis which can be interpreted in the foregoing fashion.

When we say "useful", we mean not only to the economist, engineer, and operations researcher, but also to the mathematician. There is never a simple one-to-one correspondence between a process in the real world and a mathematical process. Each real process contains features of novelty that both challenge and stimulate the mathematician, and thus serves a valuable purpose. Endogenous theories, like endogenous societies, contain the seeds of sterility.

Detailed accounts of applications will be found in [6, 20, 21]. The investigation of ways to capitalize on the abilities of the digital computer to obtain effective algorithms for the numerical solution of (2.6) and related equations raises many intriguing and formidable analytic problems involving the approximation of functions which we do not have the space to describe here. These are parts of modern disciplines of the storage, retrieval, and recognition of information; see [6] for some discussion of this.

One of the most important realizations of a multistage decision process of continuous type is furnished by the calculus of variations. This, in turn, is a particular aspect of the fact that control processes of quite general type can be interpreted to be dynamic programming processes. Let us now proceed to explain this.

8. The calculus of variations and control processes

Let x_n and y_n, $n = 1, 2, \ldots$, be N-dimensional vectors generated by the relation

(1) $$x_{n+1} = h(x_n, y_n), \quad x_0 = c,$$

and suppose that the y_n are to be chosen to maximize the scalar quantity

(2) $$\sum_{n=0}^{N} g(x_n, y_n) + k(x_N).$$

This is an example of a discrete control process [8]. If we consider a continuous version of the foregoing optimization problem, we are led in the usual fashion to the maximization of the scalar functional

(3) $$J(x, y) = \int_0^T g(x, y)\, dt + k(x(T)),$$

where x and y are connected by the differential equation

(4) $$\frac{dx}{dt} = h(x, y), \quad x(0) = c.$$

There is, of course, no difficulty in writing down an additional variational equation connecting x and y under suitable assumptions concerning g and h. If $g(x, y)$ and $k(x(T))$ are quadratic forms in the components of x and y, and if, further, $h(x, y)$ is linear in x and y, a detailed analytic investigation can be carried out due to the fortunate fact that the variational equation is linear. In general, we are confronted with the trials and tribulations of a nonlinear differential equation subject to two-point boundary conditions. Thorny questions of existence and uniqueness arise immediately. It is very difficult to deduce the analytic structure of the solution, to obtain efficient numerical algorithms, and even to fasten on the absolute maximum.

All of these difficulties are compounded if we allow the types of constraints which the control processes of the engineering and economic fields thrust upon us. In addition to (4), we add constraints such as

(5) $$r_i(x, y) \leqslant 0, \quad i = 1, 2, \ldots, M,$$

and even differential inequalities.

Nonetheless, the techniques of the classical calculus of variations are still available. With the aid of Lagrange multipliers, or equivalently Neyman-Pearson techniques, we can once again obtain variational equations [22, 23]. A particularly elegant version of these results is furnished by the "maximum principle" of Pontrjagin and his associates [24]; see the discussion in [22].

A problem which stimulated a great deal of interest in the field of control theory is that of "bang-bang" control. Here the differential equation is linear,

(6) $$\frac{dx}{dt} = Ax + y, \quad x(0) = c,$$

the constraints are of the form

(7) $$y_i = \pm k_i, \quad i = 1, 2, \ldots, N$$

(where the y_i are the components of y), and the problem is to choose them so as to drive x to the null vector, the origin, in minimum time. There are many variants of this [23, 25]. Many ingenious approximate

techniques to solve this apparently simple problem exist, but it has by no means been completely tamed.

In view of the difficulties cited above, it is reasonable to look for treatments of problems of this nature involving different principles. Problems of the foregoing type have been effectively treated using nonlinear programming, and by gradient techniques [26]. We wish here to describe an approach based on dynamic programming.

9. The calculus of variations as a continuous dynamic programming process

Consider the problem posed in (3) and (4) of Sec. 8. The usual approach regards the unknown control function as a point in a function space which is to be characterized by variational conditions. Instead, let us take advantage of the multistage aspects of the control process. We regard the choice of $y(t)$ over $[0, T]$ as a choice first over $[0, \Delta]$ and then $[\Delta, T]$. This point of view, optional in deterministic processes, becomes essential in the study of stochastic and adaptive processes. Take Δ to be an infinitesimal and let us boldly plunge ahead in a formal fashion. A choice of $y(t)$ over $[0, \Delta]$ is then a choice of $y(0)$ which we denote by v. We see that

$$x(\Delta) = c + h(c, v)\Delta, \tag{1}$$

$$J(x, y) = \int_0^{\Delta} + \int_{\Delta}^{T} + k(x(T)) =$$

$$= g(c, v)\Delta + \int_{\Delta}^{T} + k(x(T)),$$

to terms in $o(\Delta)$. Hence, if we write

$$f(c, T) = \max_v J, \tag{2}$$

we obtain via the same reasoning as before

$$f(c, T) = \max_v [g(c, v)\Delta + f(c + h(c, v)\Delta, T - \Delta)], \tag{3}$$

again to terms in $o(\Delta)$. Expanding in powers of Δ and letting $\Delta \to 0$, we readily obtain the nonlinear partial differential equation

$$f_T = \max_v [g(c, v) + (h(c, v), \operatorname{grad} f)]. \tag{4}$$

Here $(\cdot, \cdot)$ denotes the usual vector inner product. The initial condition is $f(c, 0) = 0$.

In deriving (2.6) and (3), we are using a simple, but basic, property of optimal policies which may be expressed as the

Principle of Optimality. *An optimal policy has the property that whatever the initial state and the initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision.*

10. Analytic aspects

At this stage, a number of questions come crowding on each other's heels. To begin with, we would like to understand the connection between the nonlinear partial differential equation derived in so cavalier a manner in the preceding lines, as well as the equation obtained from it for the policy function $v(c, T)$, and the Euler-Lagrange equation. One connection is by way of the theory of characteristics, an interrelation which once again displays the duality existing between the approach of the calculus of variations and dynamic programming [33]. The Euler equation can also be obtained directly from (9.4) by manipulations of partial derivatives [34]. In a simple fashion, a number of the other classical results of the calculus of variations, including Hamilton-Jacobi theory, can also be derived [34].

Having obtained all of these results so easily, the problem of a rigorous derivation of (9.4) becomes of greater interest. We can start either from a field theory of the calculus of variations, or from the original maximization problem [34, 35], or use an amalgam of methods.

When constraints are allowed, more difficult problems arise. The analogue of (9.4) is

$$f_T = \max_{v \in R} [g(c, v) + (h(c, v), \operatorname{grad} f)], \tag{1}$$

$f(c, 0) = 0$, with the region R determined by the relations

$$r_i(c, v) \leqslant 0, \quad i = 1, 2, \ldots, M. \tag{2}$$

From this equation the Pontrjagin maximum principle can be obtained formally in quite simple fashion [36], and indeed rigorously under appropriate assumptions. For the original derivation along different lines, see the book [24].

The presence of constraints gives rise to some novel analytic features, namely curves, and, more generally, surfaces, of singularity, "switching surfaces", along which grad f need not exist. These switching surfaces separate regions within which control policies of quite different type are employed. The situation is quite analogous to that in hydrodynamics where shocks can arise. Questions of existence and uniqueness of generalized solutions, and the matching of these solutions

along boundaries of the foregoing type are both intriguing and difficult [37, 30].

From both the analytic and computational side, it is important to investigate the limiting behavior of the solution of the difference equation

$$f(c, T+\Delta) = \max_{v \in R} [g(c, v)\,\Delta + f(c + h(c, v)\,\Delta, T)], \tag{3}$$

$f(c, 0) = 0$, $T = 0, \Delta, \ldots$, as $\Delta \to 0$ [38, 39, 23]. Convergence can hold under conditions which are weaker than those required for the existence of a maximum in the corresponding variational problem. The interesting point to note is that (3) is the exact equation for an approximating discrete process.

The idea of using (3) as a discrete approximation to (1) leads to the employment of the same techniques to partial differential equations not involving a maximization. Thus, for example, in place of any of the usual difference approximations to

$$u_t = g(u)\,u_x, \quad u(x, 0) = h(x), \tag{4}$$

we can write

$$u(x, t+\Delta) = u(x + g(u)\,\Delta, t), \tag{5}$$

$t = 0, \Delta, \ldots$. More complex difference schemes of similar type will lead to accuracy of a higher order in Δ. The function $u(x, t)$ is stored in the form of a Fourier series of orthogonal expansion rather than at grid points in x. This method has proved successful in practice. It is automatically stable according to the criteria of numerical analysis.

Another area which has come into prominence, is that of control processes associated with systems described by partial differential equations. Lumped parameters lead to the usual ordinary differential equations; distributed parameters lead to partial differential equations. For a detailed account of the great amount of work in this area done by the Russian school, see the book [44]. These optimization problems can be studied by the functional equation techniques of dynamic programming at the expense of introducing functionals and functional derivatives [45, 46].

A sample problem arising in the control of a heat process is that of minimizing the function

$$J(u, v) = \int_0^1 \int_0^T (u^2 + v^2)\,dx\,dT, \tag{6}$$

where

$$\begin{aligned} &u_t = u_{xx} + g(u, v), && 0 < x < 1, \quad t > 0, \\ &u(x, 0) = c(x), && 0 < x < 1, \\ &u(0, t) = u(1, t) = 0, && t > 0. \end{aligned} \tag{7}$$

Let us mention the new domain of asymptotic control theory in which the objective is to determine steady-state control policies, the behavior of $v(c, T)$ as $T \to \infty$ [41, 42]. This has connections with classical Poincare-Lyapunov theory [43]. The difference is that the differential equations are now subject to two-point boundary conditions.

There is also the inverse problem of control theory. Given an optimal policy $v(c, T)$, what relations must exist between g and h for this to be the case? A beginning of the study of this question may be found in [40]. Results of this nature are very useful in connection with the determination of approximate solutions of control processes.

Let us point out as a final note that it now seems feasible to unify both descriptive and control processes by means of a generalized theory of semigroups. In place of a fundamental equation such as

$$f_T = Af, \tag{8}$$

where A is an operator, we consider the more general

$$f_T = \max_q [A(q)\,f + b(q)], \tag{9}$$

where $A(q)$ is now an operator depending on q.

11. Dynamic programming processes of stochastic type

We can considerably enlarge the scope of our investigations by considering decision processes involving stochastic effects. So far we have supposed that the initial state, p, plus the control variable, q, uniquely determine the resultant state p_1. Let us now take the basic transformation to have the form

$$p_1 = T(p, q, r_1), \tag{1}$$

where r_1 is a random variable with a distribution which can depend upon p and q_1. Predictably, the q_i are to be chosen to maximize the expected value of a scalar function. Unpredictably, much more than this is required to make precise what we mean by a multistage decision process of stochastic type.

In studying deterministic control processes, we have the luxury of alternate approaches. We can focus either upon the set of maximizing variables as a point in a higher-dimensional space, or we can emphasize the concept of a policy, a rule for determining the decision in terms of the current state. The equivalence of these two approaches is a reflection of the fundamental duality of Euclidean geometry: a locus of points is an envelope of tangents.

When we turn to stochastic control processes, we find that the single spectral line has broadened into a continuous spectrum of pro-

cesses with each of the two approaches mentioned above representing extreme cases. It is now essential to make precise the amount of information available to the decisionmaker at each stage. One extreme is that where no information concerning the states of the system is available after the process has begun; the other corresponds to the case where complete information concerning the state of the system is available at each time.

A careful analysis of the possible types and degrees of information, together with the various kinds of uncertainty that can arise, discloses a richness in the area of stochastic control theory and a profusion of problems that will make it an attractive mathematical field for as far as one can see into the future.

Let us consider, as a mathematical formulation of the basic engineering idea of feedback control, that the process unfolds in the following way. The state p is observed, q_1 is chosen; p_1 is observed, q_2 is chosen, and so forth. Let the q_i be chosen to maximize the expected value over the r_i of the scalar function

$$R_N = g(p, q_1, r_1) + \ldots + g(p_{N-1}, q_N, r_N) + k(p_N, r_N), \tag{2}$$

and assume, for the sake of simplicity, that the r_i are independent random variables with the common distribution $dG(r)$.

If we write

$$f_N(p) = \exp_{r_i} R_N, \tag{3}$$

we obtain from the principle of optimality, or directly as before, the functional equation

$$f_N(p) = \max_{q_1} \left[\int (g(p, q_1, r_1) + f_{N-1}(T(p, q_1, r_1)))\, dG(r_1) \right], \tag{4}$$

$N \geqslant 2$, with

$$f_1(p) = \max_{q_1} \left[\int (g(p, q_1, r_1) + k(p_1, r_1))\, dG(r_1), \right. \tag{5}$$

see [7].

At the other extreme, we can consider the case where we are forced, a priori, to choose a set of vectors $[q_1, q_2, \ldots, q_N]$, and to maximize the expected value of the scalar function in (2) over this set. In certain simple cases, it makes no difference which approach we use. In general, it is clear that stage-to-stage information will materially improve the ability to control a system.

It is evident that all of the analytic and computational problems, arising in the deterministic case are present in the stochastic case, in addition to many more which have no counterparts in the more prosaic deterministic control theory. In particular, the concept of policy now assumes a dominant role. In the modern theory of stochastic control processes and other types of dynamic programming processes involving uncertainty, we find a merging of several streams of contemporary analysis: stochastic processes, control theory, ordinary and partial differential equations, information theory, in the broad sense, and numerical algorithms.

This theory has roots in a number of previous developments: sequential analysis [47], prediction theory [48], inventory theory [49], information theory [50], gambling systems and games of survival [5].

Any attempt to describe the recent work in dynamic processes of stochastic type would take us too far afield. In the area of discrete processes, let us mention Markovian decision processes; see the book [51]; in the area of continuous processes see the book [52] devoted to stochastic control theory and partial differential equations. The equations describing the system are now stochastic differential equations.

An interesting point is that the study of prediction processes by means of dynamic programming offers a new analytic and computational approach to the Wiener-Hopf integral equations of prediction theory. This is particularly important in the multidimensional case, where matrix factorization inhibits any routine solution of the corresponding Wiener-Hopf equation.

12. Dynamic programming processes of adaptive type

Let us now consider still more general decision processes of stochastic type where it is necessary both to learn and control at the same time. Processes of this type we call "adaptive" by analogy with the psychological phenomenon of adaptation.

In these areas, even the mathematical formulation is difficult. A worthwhile starting-point is a process where r_1, the random variable introduced in (11.1), possesses a distribution which has a known analytic structure, but some unknown parameters. For example, we may have a Gaussian distribution with unknown mean and variance.

We start with certain a priori estimates and then improve these estimates over time. Not only is there the problem of determining an optimal control policy, but also the intertwined problem of determining optimal estimation procedures. The technique of "sufficient statistics" plays a major role. That this is also a basic concept in the study of deterministic processes is not as appreciated as it should be.

Since two types of operations are involved, learning and control, these processes are occasionally called "dual control" processes, see the book [53]. For a discussion of applications to mathematical economics, see [54]. See also [7].

The analytic and computational problems in this field are particularly ferocious.

13. Intelligent machines, artificial intelligence, and combinatorics

The construction of a theory of intelligent machines appears to require the concept of hierarchies of decision processes, similar to the Russell theory of types. The concepts of dynamic programming and adaptive processes can be used to construct these hierarchies; see [55, 56]. The basic idea is that many types of thinking can be construed to be multistage decision processes.

This connects with the expanding area of the identification of systems, pattern recognition, and so forth. For some applications of dynamic programming to these fields, see [18].

Many of these problems are of combinatorial type and it is rather surprising that any analytic approaches based on functional equations are available. For some catch-as-catch-can applications of dynamic programming based on exploitation of structural features, see [57-61]. For extensions of dynamic programming to more general structures and order relations, see [62]; for an application to the idea of sequential computation, see [63].

14. Invariant imbedding and mathematical physics

From the standpoint of numerical analysis and digital computers, one of the paramount advantages of dynamic programming is that it permits formulations in terms of initial-value problems. It turns out that similar points of view, essentially invoking semigroups in both space and time, can be used to provide formulations of many of the fundamental processes of mathematical physics in terms of initial-value problems. The first extensive use of these ideas is due to Ambarzumian and Chandrasekhar [64] in the field of radiative transfer.

The theory of invariant imbedding represents an extension and abstraction of their "principles of invariance" [65, 66].

Dept. of Mathematics,
Electrical Engineering and Medicine,
Univ. of Southern California,
Los Angeles, California, USA

REFERENCES

[1] Hille E., Functional Analysis and Semi-groups, Amer. Math. Soc. Colloq. Publ., Vol. XXXI, 1948. Русский перевод: Хилл Э., Функциональный анализ и полугруппы, ИЛ, М., 1951.

[2] Harris T. E., The Theory of Branching Processes. Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1963. Русский перевод: Харрис Т., Теория ветвящихся случайных процессов, «Мир», М., 1965.
[3] Grenander U., Probabilities on Algebraic Structures, John Wiley & Sons, New York, 1965. Русский перевод: Гренандер У., Вероятности на алгебраических структурах, «Мир», М., 1965.
[4] Bellman R. (editor), Stochastic Processes in Mathematical Physics and Engineering. Proceedings of Symposia in Applied Mathematics, Vol. XVI, Amer. Math. Soc., Providence, Rhode Island, 1964.
[5] Bellman R., Dynamic Programming, Princeton University Press, Princeton, New Jersey, 1957. Русский перевод: Беллман Р., Динамическое программирование, ИЛ, М., 1960.
[6] Bellman R., Dreyfus S., Applied Dynamic Programming, Princeton University Press, Princeton, New Jersey, 1962. Русский перевод: Беллман Р., Дрейфус С., Прикладные задачи динамического программирования, «Наука», М., 1965.
[7] Bellman R., Adaptive Control Processes: A Guided Tour, Princeton University Press, Princeton, New Jersey, 1961. Русский перевод: Беллман Р., Процессы регулирования с адаптацией, «Наука», М., 1964.
[8] Bellman R., Kalaba R., Dynamic Programming and Modern Control Theory, Academic Press Inc., New York, 1966.
[9] Robbins H., Some Aspects of the Sequential Design of Experiments, *Bull. Amer. Math. Soc.*, **58** (1952), 527-536.
[10] Bellman R., Brown T., Projective Metrics in Dynamic Programming, *Bull. Amer. Math. Soc.*, 71 (1965), 773-775.
[11] Bellman R., Introduction to Matrix Analysis, McGraw-Hill Book Company, Inc., New York, 1960. Готовится русский перевод.
[12] Kalman R., Bertram J., General Synthesis Procedure for Computer Control of Single and Multiloop Linear Systems, Proc. Computers in Control Systems Conf., AIEE Publ. T-101, 1958.
[13] Bellman R., Karush W., On the Maximum Transform and Semigroups of Transformations, *Bull. Amer. Math. Soc.*, 68 (1962), 516-518.
[14] Bellman R., Karush W., On the Maximum Transform, *J. Math. Anal. Appl.*, 6 (1963), 67-74.
[15] LaSalle J. P., Lefschetz S., Stability by Liapunoy's Direct Method with Applications, Academic Press, Inc., New York, 1961. Русский перевод: Ла-Салль Ж., Лефшец С., Исследование устойчивости прямым методом Ляпунова, «Мир», М., 1964.
[16] Beckenbach E. F., Bellman R., Inequalities, Springer-Verlag, Berlin, 1961. Русский перевод: Беккенбах Э. Ф., Беллман Р., Неравенства, «Мир», М., 1965.
[17] Calogero F., Quantum Mechanical Scattering and Approximation Techniques, Academic Press, Inc., New York, 1967.
[18] Bellman R., Kalaba R., Quasilinearization and Nonlinear Boundary-Value Problems, American Elsevier Publishing Company, Inc., New York, 1965. Готовится русский перевод.
[19] Walter W., Differential- und Integral-Ungleichungen, Springer-Verlag, Berlin, 1964.
[20] Kaufmann A., Graphs-Dynamic Programming-Games: Theory and Applications, Academic Press, Inc., New York, 1967.
[21] Kaufmann A., Cruon R., Dynamic Programming, Academic Press, Inc., New York, 1967.
[22] Berkovitz L. D., An Optimum Thrust Control Problem, *J. Math. Anal. Appl.*, **3** (1961), 122-132.

[23] Bellman R., Glicksberg I., Gross O., Some Aspects of the Mathematical Theory of Control Processes, The RAND Corporation, R-313, 1958. Русский перевод: Беллман Р., Гликсберг И., Гросс О., Некоторые вопросы математической теории процессов управления, ИЛ, М., 1962.
[24] Понтрягин Л. С., Болтянский В. Г., Гамкрелидзе Р. В., Мищенко Е. Ф., Математическая теория оптимальных процессов, Физматгиз, М., 1961. English translation: Pontryagin L. S., Boltyanskii V. G., Gamkrelidze R. V., Mishchenko E. F., Mathematical Theory of Optimal Processes, John Wiley and Sons., New York., 1962.
[25] LaSalle J. P., Time Optimal Control Processes, *Proc. Nat. Acad. Sci. USA*, **45** (1959), 573-577.
[26] Bryson A. E. Jr., Denham W. F., Dreyfus S. E., Optimal Programming Problems with Inequality Constraints—I: Necessary Conditions for Extremal Solutions, *AIAA J.*, **1** (1963), 2544-2550. Русский перевод: *Ракетная техника и космонавтика*, 1963, № 11, 107-115.
[27] Fortet R., unpublished.
[28] Bellman R., Functional Equations in the Theory of Dynamic Programming—V: Positivity and Quasilinearity, *Proc. Nat. Acad. Sci. USA*, **41** (1955), 743-746.
[29] Lax P. D., Weak Solutions of Nonlinear Hyperbolic Equations and Their Numerical Computation, *Comm. Pure Appl. Math.*, **7** (1954), 159-193.
[30] Kalaba R., On Nonlinear Differential Equations, the Maximum Operation, and Monotone Convergence, *J. Math. Mech.*, **8** (1959), 519-574.
[31] Conway E. D., Hopf E., Hamilton's Theory and Generalized Solutions of the Hamilton-Jacobi Equation, *J. Math. Mech.*, **13** (1964), 939-986.
[32] Kalman R. E., Koepcke R. W., Optimal Synthesis of Linear Sampling Control Systems using Generalized Performance Indexes, *Trans. Amer. Soc. Mech. Eng.*, 1958.
[33] Bellman R., Dynamic Programming of Continuous Processes, The RAND Corporation, R-271, 1954.
[34] Dreyfus S., Dynamic Programming and the Calculus of Variations., Academic Press, Inc., New York, 1965.
[35] Dreyfus S., Berkovitz L. D., The Equivalence of some Necessary Conditions for Optimal Control in Problems with Bounded State Variables, *J. Math. Anal. Appl.*, **10** (1965), 275-283.
[36] Розоноэр Л. И., Принцип максимума Л. С. Понтрягина в теории оптимальных систем—III. *Автоматика и Телемеханика*, **20** (1959), № 12, 1561-1578.
[37] Bellman R., Fleming W. H., Widder D. V., Variational Problems with Constraints, *Annali di Matematica*, Ser. **4**, **41** (1956), 301-323.
[38] Bellman R., Functional Equations in the Theory of Dynamic Programming—VI: A Direct Convergence Proof, *Ann. Math.*, **65** (1957), 215-223.
[39] Bellman R., Cooke K. L., Existence and Uniqueness Theorems in Invariant Imbedding—II: Convergence of a New Difference Algorithm, *J. Math. Anal. Appl.*, **12** (1965), 247-253.
[40] Bellman R., Kalaba R., An Inverse Problem in Dynamic Programming and Automatic Control, *J. Math. Anal. Appl.*, **7** (1963), 322-325.
[41] Bellman R., On Analogues of Poincare-Lyapunov Theory for Multipoint Boundary-Value Problems—I, *J. Math. Anal. Appl.*, **13** (1966).

[42] Bellman R., Bucy R., Asymptotic Control Theory, *SIAM Control*, **2** (1964), 11-18.
[43] Bellman R., Stability Theory of Differential Equations, McGraw-Hill Book Company, Inc., New York, 1953. Русский перевод: Беллман Р., Теория устойчивости решений дифференциальных уравнений, ИЛ, М., 1954.
[44] Бутковский А. Г., Теория оптимального управления системами с распределенными параметрами, «Наука», М., 1965. English translation: Butkowski A., Control Processes Involving Distributed Parameters, Academic Press, Inc., New York, to appear.
[45] Bellman R., Functional Equations in the Theory of Dynamic Programming—VII: A Partial Differential Equation for the Fredholm Resolvent, *Proc. Amer. Math. Soc.*, **8** (1957), 435-440.
[46] Bellman R., Kalaba R., Dynamic Programming Applied to Control Processes Governed by General Functional Equations, *Proc. Nat. Acad. Sci. USA*, **48** (1962), 1735-1737.
[47] Wald A., Sequential Analysis, John Wiley & Sons, New York, 1947. Русский перевод: Вальд А., Последовательный анализ, Физматгиз, 1960.
[48] Wiener N., Cybernetics, John Wiley & Sons, New York, 1948. Русский перевод: Винер Н., Кибернетика, «Советское радио», М., 1958.
[49] Arrow K., Karlin S., Scarf H., Studies in the Mathematical Theory of Inventory and Production, Stanford University Press, Stanford, California, 1958.
[50] Shannon C., A Mathematical Theory of Communication, *Bell Systems Technical J.*, **27** (1948), 379-423, 623-656. Русский перевод: Шеннон К., Работы по теории информации и кибернетике, ИЛ, М., 1963, 243-322.
[51] Howard R., Dynamic Programming and Markov Processes, John Wiley & Sons, New York, 1960. Русский перевод: Ховард Р., Динамическое программирование и марковские процессы, «Советское радио», М., 1964.
[52] Стратонович Р. Л., Условные марковские процессы и их применение к теории оптимального управления, изд-во МГУ, М., 1966. English translation: Stratonovich R. L., Markov Processes and Optimal Control, American Elsevier Publishing Company, Inc., New York, to appear.
[53] Фельдбаум А. А., Основы теории оптимальных автоматических систем, изд. 2, «Наука», М., 1966. English translation of the first edition: Fel'dbaum A. A., Optimal Control Systems, Academic Press Inc., New York, 1965.
[54] Murphy R. E., Adaptive Processes in Economic Systems, Academic Press, Inc., New York, 1965.
[55] Bellman R., Dynamic Programming, Intelligent Machines, and Self-Organizing Systems, Proc. Symposium on Mathematical Theory of Automata, 1962.
[56] Bellman R., Adaptive Processes and Intelligent Machines, to appear.
[57] Bellman R., Dynamic Programming Treatment of the Travelling Salesman Problem, *J. Assoc. Computing Machinery*, **9** (1962), 61-63.
[58] Bellman R., Dynamic Programming and Markovian Decision Processes with Particular Application to Baseball and Chess, Applied Combinatorial Mathematics, John Wiley & Sons, New York, 1964, 221-236.
[59] Bellman R., On the Application of Dynamic Programming to the Determination of Optimal Play in Chess and Checkers, *Proc. Nat. Acad. Sci. USA*, **53** (1965), 244-247.

[60] Bellman R., An Application of Dynamic Programming to the Coloring of Maps, *I.C.C. Bull.*, 4 (1965), 3-6.
[61] Bellman R., Dynamic Programming, Pattern Recognition and Location of Faults in Complex Systems, *J. Appl. Prob.*, to appear.
[62] Brown T. A., Strauch R. E., Dynamic Programming and Multiplicative Lattices, *J. Math. Anal. Appl.*, 12 (1965), 364-370.
[63] Bellman R., Kalaba R., Lockett J., Numerical Inversion of the Laplace Transform with Applications to Biology, Economics, Engineering, and Physics, American Elsevier Publishing Company, Inc., New York, 1966.
[64] Chandrasekhar S., Radiative Transfer, Oxford, 1950.
[65] Bellman R., Kalaba R., Prestrud M., Invariant Imbedding and Radiative Transfer in Slabs of Finite Thickness, American Elsevier Publishing Company, Inc., New York, 1963.
[66] Bellman R., Kagiwada H., Kalaba R., Prestrud M., Invariant Imbedding and Time-Dependent Processes, American Elsevier Publishing Company, Inc., New York, 1964.

CONVERGENCE AND SUMMABILITY OF FOURIER SERIES

LENNART CARLESON

Let me first state quite explicitly that I do not intend to give in this lecture any survey of the very large field covered by the title. There is also no need for this since the Congress was presented such a survey quite recently. I rather want to present my personal interests which are concentrated on the almost everywhere behaviour of the partial sums. Also the subject of summability will only be touched upon.

1. Background

For a very long time, the outstanding result in the area of almost everywhere convergence has been the following result of Kolmogorov-Seliverstov-Plessner: *if for* $\lambda_n = \log n$

$$\sum_1^\infty (a_n^2 + b_n^2)\lambda_n < \infty, \tag{1.1}$$

then

$$s_n(x) = \frac{a_0}{2} + \sum_1^n (a_\nu \cos \nu x + b_\nu \sin \nu x) \tag{1.2}$$

converges a.e. The outstanding question was whether $\log n$ is a relevant sequence or not.

It has been known that conditions of the type (1.1) are related to capacities with respect to a kernel

$$K(x) \sim \sum \frac{\cos nx}{\lambda_n} \tag{1.3}$$

(Beurling [1]: $\lambda_n = n$, $K(x) \sim \log \frac{1}{|x|}$; Salem-Zygmund [4]: $\lambda_n = n^\alpha$, $K(x) \sim |x|^{\alpha-1}$). However, what they really prove is that the capacity of the divergence set vanishes for

$$K^*(x) = \frac{1}{|x|} \int_0^{|x|} K(t)\, dt$$

(see Temko [5]). When $\lambda_n = (\log n)^\beta$,

$$K(x) \sim |x|^{-1} \left(\log \frac{1}{|x|}\right)^{-1-\beta}$$

while

$$K^*(x) \sim |x|^{-1} \left(\log \frac{1}{|x|}\right)^{-\beta}.$$

Since $K^*(x) \in L^1$ only if $\beta > 1$, the result is meaningless unless $\beta > 1$ which could be considered as an indication of the relevance of the Kolmogorov factor $\log n$.

There is, however, a strong objection to this way of arguing: nothing better was known for summability (Abel or (C, 1), for example) either. But already the Fatou theorem here shows that we have a.e. summability for $f \in L^2$, $\lambda_n = 1$. We would then be faced with a possible interval $0 \leqslant \beta < 1$ with no distinction between the sizes of the exceptional sets for summability. This was of course, most unlikely and I recently proved [2] that the set where the Hardy-Littlewood maximal function

$$f^*(x) = \sup_t \int_x^{x+t} f(u)\, du \tag{1.4}$$

is infinite has K- (not only K^*-) capacity zero. But then, why could not the same be true for convergence?

The other aspect on the background is also quite subjective but it seems to me quite possibly to be of central importance. It depends on the following trivial observation.

If

$$\bar{\varphi}_j(x) = e^{2\pi i 2^j x}, \quad j = 0,\ 1,\ 2,\ \ldots,$$

and

$$n = \sum \varepsilon_j 2^j, \quad \varepsilon_j = 0,\ 1,$$

then

$$e^{2\pi i n x} = \bar{\psi}_n(x) = \prod \bar{\varphi}_j(x)^{\varepsilon_j}.$$

This means that the exponentials e^{inx}, $0 < n < 2^N$, are completely known by the knowledge of N functions.

To emphasize the point of view, let us replace $\bar{\varphi}_j(x)$ by the Rademacher functions $\varphi_j(x)$:

$$\varphi_j(x) = \operatorname{sign}(\operatorname{Im}(\bar{\varphi}_j(x)))$$

and $\psi_n(x)$ by the corresponding product—the Walsh functions.

φ_j and ψ_n can for j, $n \leqslant 2^N$ be represented by two matrices M_R and M_W respectively of -1's and $+1$'s where each column corresponds to a function and each row to a certain set. The number of columns are 2^N while the number of rows are 2^{2^N} and 2^N respectively. Let $M = (\varepsilon_{ij})$ be any such matrix. The divergence problem concerns the

existence of a_j, $\sum_{j=1}^{2^N} a_j^2 = 1$, such that

$$S_i^* = \max_{k < 2^N} \sum_{j=1}^{k} a_j \varepsilon_{ij}$$

is large for a large proportion of the possible i's. This is, of course, the more difficult to arrange the larger the number of possible i's are. This is in good correspondence with the fact that $\Sigma\, a_j \varphi_j(x)$ converges a.e. It should also be observed that if M is to correspond to an orthogonal system then the number of rows must be $\geqslant 2^N$. The Walsh system (and by analogy, the trigonometrical system) is therefore particularly advantageous to obtain divergence on sets of positive measure.

The following result is now quite surprising:

Given $\delta > 0$ there is a constant C so that a random square matrix $M = (\varepsilon_{ij})$, $1 \leqslant i, j \leqslant 2^N$, where $\varepsilon_{ij} = \pm 1$, with probability $> 1 - \delta$, has the property that for any $\{a_j\}$ and any $\lambda > 0$

$$S_i^* > \lambda \text{ only for at most } C \cdot \lambda^{-2} \cdot 2^N \text{ indices } i.$$

This means that—from this point of view—the existence of a divergent L^2-Walsh-series has probability zero.

If we introduce the possibility of changing the orders of the terms, i.e. permuting the columns of the matrix, the problem changes and the following may be true:

Given any square ± 1-matrix M there is a permutation of the columns so that for the new matrix M^* there exists a_j with S_i^* large for all i.

That purely combinatorial result would give the construction of a rearranged divergent L^2-Walsh-series and could most likely be used for a corresponding construction for the Fourier system. This line of work seems to me most promising and possibly the Kolmogorov $\log n$-factor could find its proper place here.

2. Two recent results

In a recent paper [3] I have proved the following result :

If $f \in L^2$ *then* $s_n(x)$ *converges a.e.*

If $\int |f| (\log |f|)^{1+\delta}\, dx < \infty$, *then* $s_n(x) = o(\log \log n)$ *a.e.*

I should like to try to give an idea of the method used to obtain these results.

We consider $f \in L^2(-\pi, \pi)$ and the dyadic intervals ω obtained by successive bisections of $(-2\pi, 2\pi)$. $\omega_0^* = (-2\pi, 2\pi)$ and generally ω^* is two neighbouring ω's of equal lengths. The first observation is

that $s_{n_0}(x)$, $n_0 < 2^N$, behaves as

$$(2.1)\qquad \int_{\omega_0^*} \frac{e^{-in_0 t} f(t)}{x-t}\,dt = I(p_0^*; x), \quad p_0^* = (n_0, \omega_0^*).$$

This is essentially a conjugate function which makes it possible to use a well-known theory, in particular that of maximal Hilbert transforms.

We now consider a suitable decomposition $\Omega(p_0^*)$ of ω_0^* into subintervals ω and find for a certain $\omega_1^* \subset \omega_0^*$,

$$I(p_0^*; x) = I(p_1^*; x) + R(p_0^*; x).$$

$R(p_0^*; x)$ is a remainder term and can be estimated outside a certain exceptional set $E(p_0^*)$ by means of *weak* norms such as

$$\|f; p_0^*\|^2 = \sum_{\nu=-\infty}^{\infty} \frac{1}{1+\nu^2}\left|\int_{\omega_0^*} e^{-in_0 t - i\frac{\nu}{3}t} f(t)\,dt\right|^2.$$

The estimate is such that e.g. $|R(p^*; x)| < \text{const}\,\|p^*; f\|^{\eta}\cdot\lambda$ outside a set $E(p^*)$ of measure $< e^{-\|p^*; f\|^{\eta-1}\lambda}$.

We next observe that $I(p_1^*; x)$ is of the type as $I(p_0^*; x)$ after a change of scale except that n_0 is then not an integer. However, the change by moving n_0 to the closest multiple of $|\omega_0^*|/|\omega_1^*|$ can also be estimated by $\|f; p_0^*\|$ and can be incorporated in R. We repeat the construction and find

$$(2.2)\qquad I(p_0^*; x) = I(p_k^*; x) + \sum_0^{k-1} R(p_j^*; x)$$

where we have stopped when $|\omega_k^*| \leqslant 2\pi\cdot 2^{-N}$ in which case $n_k = 0$. Therefore there is no mentioning of n_0 in the main term and this term is easily estimated.

Now the larger we allow $R(p^*; x)$ to be, the smaller we make the total exceptional set

$$(2.3)\qquad E = \bigcup_S E(p^*),$$

where the union comprises the set S of all p^* used in the different steps. On the other hand, if the R's are large, we get a bad estimate of $I(p_0^*; x)$ in the formula (2.2).

The log log n-result is obtained as follows. We let S be the set of *all* p^*'s whether used or not. For the estimation we use simply the Hausdorff-Young inequality corresponding to the integrability assumption, which gives for $\eta = \frac{\delta}{1+\delta}$,

$$\sum_S e^{-\|f; p^*\|^{-1+\eta}\cdot\log N}\cdot|\omega^*| < \varepsilon, \quad N \text{ large}.$$

The factor $\lambda = \log N$ ($n_0 \leqslant 2^N$) that is introduced here to give the small total measure ε enters as a factor of the R's and gives the estimate $\log N$, i.e. $\log\log n_0$.

To get the L^2-result the set S has to be controlled. This is quite involved and I here only want to say that the starting point for this is the modified partial sums

$$(2.4)\qquad S_a(x) = \sum_{|c_\nu|\geqslant a} c_\nu e^{i\nu x}, \quad c_\nu = \frac{1}{2\pi}\int f e^{-i\nu x}\,dx,$$

which give the best L^2-approximations using the least number of terms. These sums have, to my knowledge, not been studied and ought to be important for many problems.

3. Some open problems

(1) There is a classic example by Kolmogorov of an L^1-function whose Fourier series diverges everywhere. If one studies this example, one can quite easily see that for any $\varepsilon(n) \to 0$, there exists $f \in L^1$ with

$$s_n(x) \neq O(\varepsilon(n)\log\log n) \text{ a. e.}$$

This means that for a certain integrability between L and $L(\log L)^{1+\delta}$, $\delta > 0$, $\log\log n$ is the relevant order. The methods used above do not work, because the Hausdorff-Young inequality fails.

(2) By the result for L^2, the possible divergence sets are completely described for all classes between L^2 and C. Kahane and Katznelson have namely recently proved that for any set E of measure zero, there is a continuous function whose Fourier series diverges on E.

For L^p, $1 < p < 2$, the problem remains open. I can prove that $s_n(x) = o(\log\log\log n)$ a.e. This result obviously very strongly suggests that we actually have convergence a.e.

The L^2-proof fails for L^p because I cannot handle $S_a(x)$ in (2.4). $S_a(x)$ is an example of a function φ_a operating on L^p as a Banach algebra,

$$\varphi_a(z) = \begin{cases} z, & |z| \geqslant a \\ 0, & |z| < a. \end{cases}$$

However, it can be proved that φ_a is *not* a uniformly bounded operator as $a \to 0$, so some smoothing must be done. Also a complete solution of this problem would not immediately solve the convergence problem, but I think that important work in the general area of operators and multipliers that depend on the function could be done.

(3) In connection with $S_a(x)$ it is natural to ask for pointwise convergence. There is rather strong evidence that this fails in L^2.

(4) The study of the maximal partial sum

$$S^*(x) = \sup_n |s_n(x)| \tag{3.1}$$

is of very great interest. By a well-known theorem by Calderon, we have a weak (L^2, L^2) result, but is it true that

$$\|S^*(x)\|_q \leqslant C_q \|f\|_q, \quad 2 \leqslant q < \infty \text{ or possibly } 1 < q < \infty$$

and that

$$\text{meas}\{S^* > \lambda\} \leqslant Ce^{-c\lambda}, \quad |f| \leqslant 1?$$

The convergence proof is, in principle, constructive but it is not clear that it can give so strong results.

(5) In connection with the problem discussed in the first part, one should now be able to prove that we have convergence (not only summability) outside a set of K-capacity zero.

(6) By standard methods the convergence result for L^2 can be used to prove the pointwise convergence of Fourier integrals of functions in L^2 and of expansions of regular eigen-functions. Nothing follows, however, for several variables or for other systems such as the closely related Walsh system. As a last indication of how subtle these questions are, let me mention that there is a function $f \in L^\infty$ so that

$$\sup_n \left| \int \frac{f(t)\,\psi_n(t)}{x-t}\,dt \right| = +\infty \text{ a. e.} \tag{3.2}$$

The difference between this expression and the Dirichlet formula, for which (3.2) does not hold, is only that $\sin nt$ has been replaced in (3.2) by the function $\psi_n(t) = \text{sign}(\sin nt)$.

University of Uppsala,
Uppsala, Sweden

REFERENCES

[1] Beurling A., Ensembles exceptionnels, *Acta Math.*, **72** (1939), 1-13.
[2] Carleson L., Maximal functions and capacities, *Ann. Inst. Fourier*, **15** (1965), 59-64.
[3] Carleson L., On convergence and growth of partial sums of Fourier series, *Acta Math.*, **116** (1966), 135-157.
[4] Salem R., Zygmund A., Capacity of sets and Fourier series, *Trans. Amer. Math. Soc.*, **59** (1946), 23-41.
[5] Темко К. В., Выпуклая емкость и ряды Фурье, *ДАН СССР*, **110** (1956), 943-944.

HARMONIC ANALYSIS ON SEMISIMPLE LIE GROUPS

HARISH-CHANDRA

1. Introduction

The theory of semisimple Lie groups has, in recent years, become the meeting ground of several different branches of mathematics—differential geometry, topology, algebraic geometry, arithmetic and analysis. In this lecture I wish to speak about some recent progress in Fourier analysis on such groups. The results are far from complete. Although the case of real groups is beginning to be fairly well understood, our knowledge of the $\mathfrak{p}$-adic groups is still very rudimentary. Nevertheless there appears to be a deep-seated analogy between these two cases. In my opinion, one of the major tasks confronting us, is to try to discover and comprehend the reasons for this similarity. Once local Fourier analysis is well understood, one would have to globalize the problem by going over to the adèle group. It is in this global setting, which seems to provide the right frame-work for the understanding of the work of Hecke and Siegel, that the deeper connections between Fourier analysis and arithmetic are likely to emerge. This is indeed a big project which may take several decades to complete. All that one can say at present, is that this promises to be an extraordinarily rich and fruitful field.

2. The discrete series

Let G be a locally compact, separable and unimodular group. A unitary representation π of G on a Hilbert space $\mathfrak{H}$ is a mapping π which assigns to every $x \in G$ a unitary operator $\pi(x)$ on $\mathfrak{H}$ such that :

(1) $\qquad \pi(xy^{-1}) = \pi(x)\,\pi(y)^{-1} \qquad (x, y \in G),$

(2) The mapping $(x, \psi) \longrightarrow \pi(x)\,\psi$ of $G \times \mathfrak{H}$ into $\mathfrak{H}$ is continuous.

π is said to be irreducible if $\mathfrak{H} \neq \{0\}$ and no closed subspace of $\mathfrak{H}$, other than $\{0\}$ and $\mathfrak{H}$ itself, is stable under $\pi(x)$ for all $x \in G$. The equivalence of two representations is defined as usual. Let $\mathscr{E}$ be the set of all equivalence classes of irreducible unitary representations of G.

Fix a Haar measure dx on G and let r denote the right-regular representation of G on $L_2(G)$. A class $\omega \in \mathscr{E}$ is called discrete, if there exists a closed, invariant and irreducible subspace $\mathfrak{H}$ of $L_2(G)$ such that the restriction of r on $\mathfrak{H}$ lies in ω. Let $\mathscr{E}_d$ denote the set of all discrete classes. Then $\mathscr{E}_d$ is called the discrete series for G. Let π be

an irreducible unitary representation of G on a Hilbert space $\mathfrak{H}$ and ω the class of π. Then ω is discrete if and only if

$$\int_G |(\phi, \pi(x)\psi)|^2 dx < \infty$$

for all $\phi, \psi \in \mathfrak{H}$. Moreover if $\omega \in \mathcal{E}_d$, there exists a number $d(\omega) > 0$, called the formal degree of ω (or π), such that

$$\int_G |(\phi, \pi(x)\psi)|^2 dx = |\phi|^2 |\psi|^2 d(\omega)^{-1} \quad (\phi, \psi \in \mathfrak{H}).$$

Now suppose G is a connected (real) semisimple Lie group with finite center. Then the following theorem gives a criterion for the existence of the discrete series.

Theorem 1. *In order that $\mathcal{E}_d$ should not be empty, it is necessary and sufficient that G should have a compact Cartan subgroup.*

It is believed that a $\mathfrak{p}$-adic semisimple group always has a compact Cartan subgroup. Therefore, by analogy, we should expect it to have a discrete series. There are some indications that this is indeed so.

3. Characters

Let $\ell = \text{rank } G$ and D the coefficient of t^ℓ in $\det(t + 1 - \text{Ad}(x))$ $(x \in G)$, where t is an indeterminate. Then $D = D(x)$ is an analytic function which is not identically zero. Let G' be the set of all $x \in G$ where $D(x) \neq 0$. Then G' is an open and dense subset of G whose complement has measure zero.

Let u be a differential operator on G. Its adjoint u^* is the differential operator given by the relation

$$\int_G uf \cdot g\, dx = \int_G f \cdot u^* g\, dx \qquad (f, g \in C_c^\infty(G)).$$

Let T be a distribution on G. Then $f \to T(u^* f)$ $(f \in C_c^\infty(G))$ is also a distribution which we denote by uT. A locally summable function F on G defines a distribution T_F by the rule

$$T_F(f) = \int fF\, dx \qquad (f \in C_c^\infty(G)).$$

We say that a given distribution T is a function, if there exists a locally summable function F such that $T = T_F$. Then F is unique up to a set of measure zero and it is convenient to write $T = F$.

Fix a maximal compact subgroup K of G and let $\mathfrak{Z}$ denote the algebra of all differential operators on G which commute with both

left and right translations of G. A distribution T on G is said to be $\mathfrak{Z}$-finite if the space of all distributions of the form zT $(z \in \mathfrak{Z})$ has finite dimension. Similarly it is called K-finite if the left and right translates of T under K span a vector space of finite dimension. It is easy to see that if T is both $\mathfrak{Z}$-finite and K-finite, it satisfies an elliptic differential equation and so it is an analytic function. We say that T is invariant if it is left fixed by all inner automorphisms of G.

Theorem 2. *Let Θ be an invariant and $\mathfrak{Z}$-finite distribution on G. Then Θ is a function which is analytic on G'.*

Let π be an irreducible unitary representation of G and ω the class of π. For any $f \in C_c^\infty(G)$, define

$$\pi(f) = \int f(x)\pi(x)\, dx.$$

Then it can be shown that $\pi(f)$ is an operator of the trace class and there exists a distribution Θ_ω on G such that

$$\Theta_\omega(f) = \text{tr}\, \pi(f) \qquad (f \in C_c^\infty(G)).$$

Θ_ω is called the character of ω (or π). It is easy to show that Θ_ω is an invariant eigendistribution of $\mathfrak{Z}$. Therefore by Theorem 2, it is a function which is analytic on G'.

In the $\mathfrak{p}$-adic case one defines $C_c^\infty(G)$ to be the space of all locally constant functions with compact support. Then it seems plausible that $\pi(f)$ is still of the trace class. Put

$$\Theta_\omega(f) = \text{tr}\, \pi(f) \qquad (f \in C_c^\infty(G)).$$

Then again Θ_ω should turn out to be a locally summable function on G which is locally constant on the regular set G'.

4. The Selberg principle

For any $\gamma \in G$, let G_γ denote the centralizer of γ in G. It is not difficult to show that G_γ is unimodular and therefore the factor space $G = G/G_\gamma$ has an invariant measure $d\bar{x}$. Put

$$\gamma^{\bar{x}} = x\gamma x^{-1} \qquad (x \in G),$$

where $x \to \bar{x}$ is the projection of G on $\bar{G}$. An element $\gamma \in G$ is said to be elliptic, if it is contained in some compact Cartan subgroup of G.

Theorem 3. (The Selberg Principle.) *Let γ be a semisimple element of G and f a K-finite and $\mathfrak{Z}$-finite function in $L_2(G)$. Then*

the integral

$$\int_{G/G_\gamma} f(\gamma^x)\, d\dot{x}$$

is well defined and, if γ is not elliptic, its value is zero.

Let Γ be a discrete subgroup of G such that G/Γ is compact. Then G operates on G/Γ and so we get a representation λ of G on $L_2(G/\Gamma)$. It is easy to see that λ decomposes into a direct sum of irreducible representations. Moreover the multiplicity $m(\omega)$ of each class $\omega \in \mathcal{E}$, is finite in λ. Let π be an irreducible unitary representation on $\mathfrak{H}$ lying in a given class ω. We say that π (or ω) is integrable if

$$\int |(\phi, \pi(x)\psi)|\, dx < \infty$$

for any two K-finite vectors $\phi, \psi \in \mathfrak{H}$. (A vector $\phi \in \mathfrak{H}$ is called K-finite, if the space spanned by $\pi(k)\phi$ $(k \in K)$, is finite-dimensional.) The Selberg principle allows us to obtain an explicit formula for the multiplicity $m(\omega)$ corresponding to any integrable class ω.

I believe that the Selberg principle holds also for the $\mathfrak{p}$-adic groups after a slight reformulation. Let K be any open and compact subgroup of G and π a representation on $\mathfrak{H}$ of the discrete class. Take

$$f(x) = (\phi, \pi(x)\psi) \qquad (x \in G)$$

where ϕ and ψ are two K-finite vectors in $\mathfrak{H}$. Then the statement of Theorem 3 should remain true for f.

5. Formula for the characters

We return to the case when G is real and suppose that B is a Cartan subgroup of G contained in K. Let $\mathfrak{g}, \mathfrak{b}$ be the Lie algebras of G and B respectively and G_c the simply connected complex analytic group corresponding to the complexification $\mathfrak{g}_c$ of $\mathfrak{g}$. Let us assume that G is the real analytic subgroup of G_c corresponding to $\mathfrak{g}$. Let P be the set of all positive roots of $(\mathfrak{g}, \mathfrak{b})$ under some fixed order. For any $\alpha \in P$, we denote by H_α the element in $\mathfrak{b}_c$ such that

$$\operatorname{tr}(\operatorname{ad} H \operatorname{ad} H_\alpha) = \alpha(H) \qquad (H \in \mathfrak{b}).$$

Then

$$\tilde{\omega} = \prod_{\alpha \in P} H_\alpha$$

can be considered as a differential operator on B. There exists an analytic function Δ on B such that

$$\Delta(\exp H) = \prod_{\alpha \in P} (e^{\alpha(H)/2} - e^{-\alpha(H)/2}) \qquad (H \in \mathfrak{b}).$$

Let $\mathfrak{F}$ be the space of all real-valued linear functions on $(-1)^{1/2}\mathfrak{b}$ and L the lattice of those $\lambda \in \mathfrak{F}$ for which there exists a character ξ_λ of B given by $\xi_\lambda(\exp H) = e^{\lambda(H)}$ $(H \in \mathfrak{b})$. It is clear that $\tilde{\omega}$ can also be regarded as a polynomial function on $\mathfrak{F}$ and

$$\tilde{\omega}\xi_\lambda = \tilde{\omega}(\lambda)\, \xi_\lambda \qquad (\lambda \in L).$$

Put $W_G = \tilde{B}/B$ where $\tilde{B}$ is the normalizer of B in G. Then W_G is a subgroup of the Weyl group W of $(\mathfrak{g}, \mathfrak{b})$. Put $B' = B \cap G'$ and let L' denote the set of all $\lambda \in L$ such that $\tilde{\omega}(\lambda) \neq 0$.

Theorem 4. *For any $\lambda \in L'$, there exists exactly one invariant eigendistribution Θ_λ of $\mathfrak{Z}$ on G such that*

1) $$\sup_{x \in G'} |D(x)|^{1/2} |\Theta_\lambda(x)| < \infty,$$

2) $$\Delta(b)\Theta_\lambda(b) = \sum_{s \in W_G} \varepsilon(s)\, \xi_{s\lambda}(b) \qquad (b \in B').$$

Here D has the same meaning as in § 3 and we have to bear in mind Theorem 2. Moreover the character ε of W is defined as usual.

Put $\varepsilon(\lambda) = \operatorname{sign} \tilde{\omega}(\lambda)$ $(\lambda \in L')$ and $q = \frac{1}{2} \dim G/K$. Then q is an integer.

Theorem 5. *For each $\lambda \in L'$, there exists a unique class $\omega(\lambda) \in \mathcal{E}_d$ such that $\Theta_{\omega(\lambda)} = (-1)^q \varepsilon(\lambda)\Theta_\lambda$. The mapping $\lambda \to \omega(\lambda)$ of L' into $\mathcal{E}_d$ is surjective and $\omega(\lambda_1) = \omega(\lambda_2)$ $(\lambda_1, \lambda_2 \in L')$ if and only if $\lambda_2 = s\lambda_1$ for some $s \in W_G$.*

Define

$$F_f(b) = \Delta(b) \int_G f(xbx^{-1})\, dx \qquad (b \in B')$$

for $f \in C_c^\infty(G)$. Then there exists a number $c > 0$ such that

$$\lim_{b \to 1} (\tilde{\omega} F_f)(b) = (-1)^q c f(1)$$

for all $f \in C_c^\infty(G)$. Moreover

$$d(\omega(\lambda)) = c^{-1} |W_G|\, |\tilde{\omega}(\lambda)| \qquad (\lambda \in L')$$

where $|W_G|$ is the order of W_G and $d(\omega(\lambda))$ the formal degree of $\omega(\lambda)$ (see § 2). It is possible to compute c explicitly for a suitable normalization of dx.

Theorem 6. *Let f be a K-finite and $\mathfrak{Z}$-finite function in $L_2(G)$. Then the integral*

$$\int_G f\Theta_\lambda\, dx \qquad (\lambda \in L')$$

is well defined. Let $\Theta_\lambda(f)$ *denote its value. Then* $\Theta_\lambda(f) = 0$ *for all* $\lambda \in L'$ *except a finite number and*

$$(-1)^q cf(1) = \sum_{\lambda \in L'} \omega(\lambda)\Theta_\lambda(f).$$

We observe that f is automatically analytic (see § 3) and therefore $f(1)$ is well defined. The Selberg principle enters in an essential way in the proofs of Theorems 5 and 6.

For $\mathfrak{p}$-adic groups the number of conjugacy classes of compact Cartan subgroups is, in general, more than 1. Also there does not seem to be any analogue of the algebra $\mathfrak{Z}$. Therefore the problem of determining all the characters of the discrete series appears to be much more difficult in this case.

6. Concluding remarks

There are some striking similarities between the real and the $\mathfrak{p}$-adic groups. Let K be a maximal compact subgroup of G in the real case and any open compact subgroup in the $\mathfrak{p}$-adic case. Let $\mathfrak{E}_K$ be the set of all equivalence classes of irreducible unitary representations of K. For a given irreducible unitary representation π of G on $\mathfrak{H}$, let $m(\mathfrak{d})$ ($\mathfrak{d} \in \mathfrak{E}_K$) denote the multiplicity of the class $\mathfrak{d}$ in the restriction of π on K. In the real case it is known that $m(\mathfrak{d})$ is finite. The same is believed to be true in the $\mathfrak{p}$-adic case but so far no general proof for this has been found.

The function $|D|^{-1/2}$ is locally summable on G in the real case. I believe that this fact is also true in the $\mathfrak{p}$-adic case, provided we interpret the absolute value in the $\mathfrak{p}$-adic sense.

In the real case there is a simple connection between the asymptotic behaviour of the elementary spherical functions and the Plancherel measure for the space G/K. The same appears to be true also in the $\mathfrak{p}$-adic case. (For $G = SL(2)$, this follows from the results of Mautner.)

The algebra $\mathfrak{Z}$ of bi-invariant differential operators plays a very important role in the real case. However, as we have already observed, there does not seem to exist any $\mathfrak{p}$-adic analogue of $\mathfrak{Z}$. Nevertheless it appears likely that all the final results of Fourier analysis continue to hold, after some slight reformulation, for $\mathfrak{p}$-adic groups. The unravelling of this mystery would, in my opinion, be an important achievement.

The Institute for Advanced Study,
Princeton, N.J., USA

THÉORIE LOCALE DES FONCTIONS DIFFÉRENTIABLES

BERNARD MALGRANGE

Les résultats dont je vais parler ont une double origine : d'une part, deux problèmes posés par L. Schwartz à propos de la théorie des distributions, d'autre part l'étude des singularités des applications différentiables, developpée d'abord par H. Whitney et R. Thom.

Fixons d'abord quelques notations ; soit Ω un ouvert $\subset \mathbf{R}^n$, et soit $\mathscr{E}(\Omega)$ l'espace des fonctions de classe C^∞ dans Ω, à valeurs réelles, muni de sa topologie usuelle ; pour $a \in \Omega$ et $f \in \mathscr{E}(\Omega)$, désignons par $T_a f$ le développement de Taylor de f en a ; si l'on a $V \subset \mathscr{E}(\Omega)$, posons $T_a V = \{T_a f \mid f \in V\}$.

Cela posé, le premier problème est celui de la synthèse harmonique dans l'espace des distributions tempérées ; par dualité et transformation de Fourier, il est résolu par le théorème suivant, conjecturé par Schwartz et démontré par Whitney [7], [12] :

Théorème 1. *Soit* $\mathscr{I}$ *un idéal de* $\mathscr{E}(\Omega)$. *Pour que* f *appartienne à l'adhérence de* $\mathscr{I}$, *il faut et il suffit que l'on ait, pour tout* $a \in \Omega$: $T_a f \in T_a \mathscr{I}$.

Quant au second problème, celui de la division des distributions, il équivaut par dualité au suivant : on se donne $f \in \mathscr{E}(\Omega)$; à quelle condition l'idéal engendré par f est-il fermé ? Tout d'abord, on voit facilement que ce n'est pas toujours vrai : par exemple, si f est nulle en un point ainsi que toutes ses dérivées, sans être identiquement nulle au voisinage, l'idéal $f\mathscr{E}(\Omega)$ n'est pas fermé ; d'autre part, dans le cas d'une variable, on constate immédiatement que, si cette circonstance ne se produit en aucun point, notre idéal est fermé ; ces faits et quelques autres avaient amené Schwartz à conjecturer le résultat suivant [8] :

Théorème 2. *Si* f *est analytique,* $f\mathscr{E}(\Omega)$ *est fermé.*

Ce résultat a été démontré simultanement par Hörmander [1] (dans le cas où f est un polynôme) et par Łojasiewicz [2] (dans le cas général). Les deux méthodes reposent sur l'importante inégalité suivante :

Théorème 3. *Soit* f *une fonction analytique dans* Ω, X *l'ensemble de ses zéros (supposé non vide), et* K *un compact* $\subset \Omega$. *Il existe des constantes* $C > 0$ *et* $\alpha > 0$ *telles qu'on ait, pour tout* $a \in K$:

$$|f(a)| \geqslant C d(a, X)^\alpha.$$

Pour une fonction $f \in \mathscr{E}(\Omega)$ arbitraire, cet énoncé équivaut au suivant : toute $\varphi \in \mathscr{E}(\Omega)$, nulle ainsi que toutes ses dérivées sur X, appartient à $f\mathscr{E}(\Omega)$; ce fait, joint au théorème 1, montre que cette inégalité est nécessaire pour que cet idéal soit fermé. Dans le cas analytique, une fois le théorème 3 obtenu, la méthode consiste à prendre une stratification convenable de X : on l'écrit comme somme finie disjointe de sous ensembles X_i, avec $\overline{X}_i - X_i \subset \bigcup_{i<j} X_i$, et l'on établit un énoncé analogue au théorème 2 pour les « fonctions différentiables au sens de Whitney » sur $\overline{X}_i$, nulles à l'ordre infini sur $\overline{X}_i - X_i$ (ce qui exige d'appliquer le théorème 3 à d'autres fonctions que f...). Hörmander prend tout simplement la stratification dans laquelle X_i est l'ensemble des points où f s'annule exactement à l'ordre i ; Łojasiewicz en prend une plus fine, qui l'amène à une étude détaillée de l'ensemble des zéros des fonctions analytiques-réelles.

Quelles sont les extensions possibles de ces résultats ? Tout d'abord, il n'est pas nécessaire de se limiter aux idéaux principaux: le théorème 2 se généralise ainsi [3].

Théorème 2 bis. *Un idéal de $\mathscr{E}(\Omega)$ engendré par des fonctions analytiques est fermé.*

Joint au théorème 1, ceci a pour conséquence immédiate la démonstration d'une conjecture de Serre [9] : l'anneau des germes de fonctions C^∞ en un point est un module fidèlement plat sur l'anneau des germes de fonctions analytiques en ce point ; autrement dit : les relations à coefficients C^∞ entre des fonctions analytiques sont engendrées par les relations à coefficients analytiques.

Que peut-on dire maintenant dans le cas non analytique ? Nous avons déjà vu que les propriétés précédentes n'étaient pas toujours vraies ; mais nous allons voir, en utilisant une variante d'une idée de Thom [10], qu'elles le sont « en général » : il nous faut d'abord donner un précis à cette notion. Soit $\mathscr{E}_n$ l'espace des germes de fonctions C^∞ en 0 dans $\mathbf{R}^n$, et soit $\mathscr{J}$ un idéal de type fini de $\mathscr{E}_n$, engendré par $f_1, \ldots, f_n$. Nous dirons que $\mathscr{J}$ est fermé s'il existe un voisinage ouvert Ω de 0 et des représentants des f_i dans $\mathscr{E}(\Omega)$ tels que l'idéal qu'ils engendrent soit fermé.

Posons d'autre part, pour $f \in \mathscr{E}_n$: $\hat{f} = T_0 f$; l'espace $\hat{\mathscr{E}}_n$ s'identifie alors aux séries formelles $\mathbf{R}[[x_1, \ldots, x_n]]$, les x_i désignant les coordonnées dans $\mathbf{R}^n$. En considérant $\mathscr{E}_n$ comme la limite projective de ses quotients par les puissances de l'idéal maximal, on définit la notion de variété algébrique dans $\mathscr{E}_n$. Nous dirons alors qu'une propriété (P) des éléments de $\hat{\mathscr{E}}_n$ est « générale » s'il existe une variété algébrique $V \subset \hat{\mathscr{E}}_n$ de *codimension infinie*, telle que (P) soit vérifiée pour toute $f \in \mathscr{E}_n$ satisfaisant à $f \notin V$. (Cette notion ne doit pas être confondue avec celle de « propriété générique » au sens où l'entend Thom [10] : le fait, par exemple, pour une fonction d'être « de Morse » est générique, mais c'est une propriété bien trop restrictive pour être « générale ».) Nous définirons de même une propriété générale des systèmes $(f_1, \ldots, f_p) \in (\mathscr{E}_n)^p$.

Cette notion a été récemment étudiée systématiquement par Tougeron [11], qui a montré que, en général, les fonctions, et les systèmes de p fonctions différentiables ont des propriétés au moins aussi bonnes que les fonctions analytiques arbitraires. Parmi ses résultats, nous citerons l'un des plus simples, obtenu également par Mather [6].

Théorème 4. *Supposons $p < n$; alors, en général, l'idéal engendré par $(f_1, \ldots, f_p)$ dans $\mathscr{E}_n$ est une intersection complète admettant l'origine pour point singulier isolé (c'est-à-dire : l'idéal engendré par $f_1, \ldots, f_p$ et leurs jacobiens d'ordre p contient une puissance de l'idéal maximal de $\mathscr{E}_n$). Un tel idéal est équivalent, par difféomorphisme, à un idéal engendré par des polynômes ; en particulier, il est fermé.*

En réalité, on démontre davantage : étant donné $(f_1, \ldots, f_p)$, tel que l'idéal engendré soit une intersection complète avec point singulier isolé, il existe un entier k possédant la propriété suivante : si $(f'_1, \ldots, f'_p)$ a même développement de Taylor en 0 à l'ordre k de f, les deux idéaux sont équivalents par difféomorphisme ; il suffit alors de prendre pour f'_i des polynômes pour obtenir la seconde assertion du théorème précédent.

A l'usage des spécialistes de topologie différentielle, je me permettrai la suggestion suivante : il serait peut-être intéressant d'étudier, au point de vue global, les variétés différentiables admettant des singularités du type précédent.

Enfin, dans le cas où l'on a $p \geqslant n$, on voit facilement que, en général, l'idéal engendré par $(f_1, \ldots, f_p)$ est un idéal de définition de $\mathscr{E}_n$, c'est-à-dire contient une puissance de l'idéal maximal.

Comme je l'avais promis au début, j'en viens maintenant à la théorie des applications différentiables ; ici, nous ne nous intéresserons plus seulement à l'idéal engendré par $(f_1, \ldots, f_p)$, mais au germe d'application $f : \mathbf{R}^n \to \mathbf{R}^p$ que ce système de fonctions définit. Remarquons d'abord que, par composition des fonctions, f définit une application $f^* : \mathscr{E}_p \to \mathscr{E}_n$, qui fait de $\mathscr{E}_n$ un $\mathscr{E}_p$-module ; désignons respectivement par $\mathscr{M}_n$ et $\mathscr{M}_p$ l'idéal maximal de $\mathscr{E}_n$ et celui de $\mathscr{E}_p$. On a alors le théorème suivant :

Théorème 5. *Soit F un $\mathscr{E}_n$-module de type fini ; pour qu'il soit de type fini sur $\mathscr{E}_p$, il faut et il suffit que $F/\mathscr{M}_p F$ soit de type fini sur $\mathscr{E}_p/\mathscr{M}_p \cong \mathbf{R}$.*

Notons que l'on peut dire alors un peu plus : prenons en effet des générateurs de $F/\mathscr{M}_p F$ sur $\mathbf{R}$, et remontons-les dans F : d'après le « lemme de Nakayama », on aura des générateurs de F sur $\mathscr{E}_p$.

L'énoncé précédent est ce qu'on appelle le « théorème de préparation différentiable » ; la forme donnée ici, un peu plus générale que la forme habituelle [4], [5], est due à Mather. Je vais en donner quelques cas particuliers : prenons d'abord $F = \mathscr{E}_n$; l'hypothèse « $\mathscr{E}_n/\mathscr{M}_p\mathscr{E}_n$ fini sur $\mathbf{R}$ » signifie simplement que l'idéal (f) engendré par $f_1, \ldots, f_p$ est un idéal de définition de $\mathscr{E}_n$: ceci exige $p \geqslant n$, et, inversement, nous avons vu plus haut que, pour $p \geqslant n$, cette propriété sera généralement vérifiée.

Prenons ensuite $p = n - 1$, et prenons pour f la projection définie par $f_i = x_i$, $i = 1, \ldots, n-1$. Soit d'autre part Φ une fonction de $\mathscr{E}_n$, $\Phi(0, \ldots, 0, x_n)$ ayant un développement de Taylor commençant par x_n^q ; prenons enfin $F = \mathscr{E}_n/\Phi\mathscr{E}_n$. Il est facile de voir que le théorème s'applique, et que l'on peut prendre pour générateurs de F sur $\mathscr{E}_{n-1}$ les classes des x_n^h, $k = 0, \ldots, q-1$. Autrement dit, pour toute $g \in \mathscr{E}_n$, il existe une identité

$$(\mathrm{W}) \qquad g = \Phi Q + \sum_{j=0}^{q-1} x_n^j R_j, \text{ avec } Q \in \mathscr{E}_n, \quad R_j \in \mathscr{E}_{n-1}.$$

C'est la variante différentiable du théorème de préparation de Weierstrass, sous la forme que lui ont donnée Späth et Rückert (mais ici, contrairement au cas analytique, il n'y aura pas unicité) ; l'énoncé de Weierstrass lui-même, à savoir que Φ est équivalent à un polynome distingué, est donc encore vrai dans le cas différentiable : ce résultat avait été d'abord conjecturé par Thom.

La démonstration du théorème 5 se décompose en deux étapes : la première consiste, par une utilisation convenable du théorème des fonctions implicites, à ramener le cas général à un cas particulier : la démonstration de (W) dans le cas où Φ est un polynôme unitaire en x_n ; on peut même, en introduisant de nouvelles variables t_j, et en les substituant aux coefficients de ce polynôme, supposer que Φ est un polynôme par rapport à l'ensemble des variables. La seconde étape, plus difficile, consiste alors à établir le résultat particulier indiqué ; la méthode initiale du conférencier consistait à utiliser les techniques développées par Hörmander et Łojasiewicz, dont il a été question plus haut. Récemment, Mather [6] a donné une autre méthode, plus élémentaire, utilisant des développements en intégrales de Fourier.

On peut se poser, pour les applications différentiables, une question analogue à celle que nous avons examinée pour les idéaux : le germe d'application f étant donné, existe-t-il un entier k tel que tout germe f', ayant même développement de Taylor en 0 que f jusqu'à l'ordre k,

soit équivalent à f par un difféomorphisme de la source et du but ? Contrairement à ce qui se passe pour les idéaux, cette propriété n'est pas généralement vraie (toutefois, d'après Thom, elle le devient si l'on remplace « difféomorphisme » par « homéomorphisme » [10]). Mather, en utilisant le théorème de préparation, a donné une caractérisation des applications qui possèdent cette propriété. Je n'exposerai pas ce résultat, mais je parlerai d'un autre, très voisin, et qui se traite par la même méthode : la caractérisation des germes stables.

Donnons-en d'abord la définition ; soit f un germe d'application $\mathbf{R}^n \to \mathbf{R}^p$, avec $f(0) = 0$; en gros, f est stable si tout germe « suffisamment voisin » de f est équivalent à f par un difféomorphisme de la source et du but ; de façon plus précise, nous exigerons que, pour tout voisinage ouvert Ω de 0 dans $\mathbf{R}^n$, et tout représentant $\tilde{f}$ de f dans Ω, il existe un ouvert Ω', avec $0 \in \Omega' \subset \Omega$, et un ouvert $U \subset \mathbf{R}^p$, avec $\tilde{f}(\Omega') \subset U$ et $0 \in U$, possédant la propriété suivante : pour toute fonction $\tilde{f}' : \Omega \to \mathbf{R}^p$, suffisamment voisine de f dans la topologie de $\mathscr{E}(\Omega)$, il existe des plongements C^∞ $h : \Omega' \to \Omega$ et $h' : U \to \mathbf{R}^p$ tels qu'on ait

$$\tilde{h}' \circ (\tilde{f} \mid \Omega') = f' \circ h$$

(à noter que l'on ne suppose pas ici $f'(0) = 0$).

Une notion voisine est celle de « stabilité infinitésimale » : toute déformation infinitésimale g de f doit pouvoir être obtenue comme somme de deux déformations provenant de transformations infinitésimales de la source et du but ; de façon précise, il doit exister, pour tout germe g d'application en 0 de $\mathbf{R}^n$ dans $\mathbf{R}^p$ (on ne suppose pas nécessairement $g(0) = 0$), deux germes de champs de vecteurs X sur $\mathbf{R}^n$ et Y sur $\mathbf{R}^p$ tels qu'on ait

$$g = \langle X, df \rangle + Y \circ f.$$

Mather considère enfin, dans l'espace des jets (i.e. des développements de Taylor) d'ordre k d'applications de $\mathbf{R}^n$ dans $\mathbf{R}^p$, l'orbite $V(f)$ de $j^k f(0)$ sous l'action des difféomorphismes de la source et du but; son résultat est alors le suivant :

Théorème 6. *Les propriétés suivantes sont équivalentes :*

a) Le germe f est stable.

b) Le germe f est infinitésimalement stable.

c) Pour une (ou, pour toute) valeur de k suffisamment grande, $j^k f$ est transversal en 0 à $V(f)$.

L'équivalence de b) et c) se démontre en utilisant le théorème de préparation ; pour démontrer leur équivalence avec a), on utilise dans un sens un théorème de transversalité, et dans l'autre une intégration de champs de vecteurs qui permet de passer de transformations

infinitésimales à des transformations finies. J'indique pour terminer que Mather a réussi à traiter aussi par ses méthodes le cas global, c'est-à-dire à caractériser d'une manière analogue les applications C^∞-stables d'une variété compacte dans une autre variété.

Université de Paris,
Faculté des Sciences,
Orsay, France

RÉFÉRENCES

[1] Hörmander L., On the division of distributions by polynomials, *Arkiv för Mathematik*, 3 (1958), 555-568.
[2] Łojasiewicz S., Sur le problème de la division, *Studia Math.*, **18** (1959), 87-136.
[3] Malgrange B., Division des distributions, Séminaire L. Schwartz (1959-60), exposés 21-25.
[4] Malgrange B., Le théorème de prépération en géométrie différentiable, Séminaire H. Cartan (1962-63), exposès 11-12-13-22.
[5] Malgrange B., Ideals of differentiables functions, Tata Institute Bombay and Oxford University Press, 1966. (Готовится к печати русский перевод.)
[6] Mather J., On the preparation theorem of Malgrange, Notes miméographiées, Princeton, 1966, et travaux non publiés.
[7] Schwartz L., Analyse et synthèse harmonique dans les espaces de distributions, *Canadian Journal Math.*, 3 (1951), 503-512.
[8] Schwartz L., Théorie des distributions, Hermann, Paris, 1950-51.
[9] Serre J.-P., Un théorème de dualité, *Comm. Math. Helvet.*, **29** (1955), 9-26.
[10] Thom R., Local topological properties of differentiable mappings, Bombay Colloquium on Differential Analysis, Oxford University Press, 1964.
[11] Tougeron J. C., Equivalence des idéaux de fonctions différentiables, *C. R. Acad. Sc. Paris*, **262** (1966), 563-565, et travaux non publiés.
[12] Whitney H., On ideals of differentiables functions, *Amer. Journal Math.*, 70 (1948), 635-658.

UNGLEICHUNGEN UND FEHLERABSCHÄTZUNGEN

JOHANN SCHRÖDER

1. Einleitung

Gegeben sei eine Gleichung

$$Mu = r \tag{1.1}$$

mit einem Operator M, der eine Teilmenge D eines halbgeordneten Raumes $\mathfrak{R}$ in einen halbgeordneten Raum $\mathfrak{S}$ abbildet. Man möchte obere und untere Schranken für eine Lösung u^* dieser Gleichung ermitteln. Bei vielen Problemen ist dies mit Hilfe eines Schlusses folgender Art möglich:

$$Mw \leqslant r \leqslant Mv \Rightarrow w \leqslant u^* \leqslant v. \tag{1.2}$$

Ist die Existenz einer Lösung $u^* \in D$ gesichert, so genügt es, daß für alle $u \in D$ und die gewählten Elemente v, $w \in D$ gilt:

$$Mw \leqslant Mu \Rightarrow w \leqslant u, \quad Mu \leqslant Mv \Rightarrow u \leqslant v. \tag{1.3}$$

In vielen Fällen hat die gegebene Gleichung (1.1) die Form

$$u = Tu \tag{1.4}$$

(oder läßt sich auf eine solche Form zurückführen), und man kann als Schranken die Näherungen von Iterationsverfahren benutzen, welche monoton gegen eine Lösung u^* konvergieren. Dann erhält man also gleichzeitig eine Existenzaussage. Natürlich müssen dabei im allgemeinen stärkere Voraussetzungen gemacht werden, als zum Nachweis der Monotonie-Eigenschaft (1.3) allein.

Bereits Chaplygin [5] hat für bestimmte Operatoren die Eigenschaft (1.3) nachgewiesen und ferner für gewisse Gleichungen der Form (1.4) die Einschließung einer Lösung durch monotone Iterationsfolgen auch numerisch ausgenutzt. (Für mehr theoretische Zwecke entwickelte auch Perron [16] die "Methode der Ober- und Unterfunktionen".) In neuerer Zeit ist diese Methode der Einschließung durch Iterationsfolgen abstrakt formuliert ([12], [15], [2], [9], [20]) und vielfach angewendet worden. Insbesondere sind auf diesem Gebiet viele Arbeiten russischer Autoren erschienen, welche sich hauptsächlich mit gewöhnlichen Differentialgleichungen und Anfangswertaufgaben bei partiellen Differentialgleichungen (parabolischen und hyperbolischen) beschäftigen. Die sehr umfangreiche Literatur soll hier nicht aufgeführt werden. Eine Reihe von Autoren geht von einem Aufsatz von Azbelew und Tsaliuk [2] aus.

Hier werden wir uns mit hinreichenden (und notwendigen) Bedingungen für die Eigenschaft (1.3) beschäftigen, ohne genauer auf das Existenzproblem einzugehen. Die Eigenschaft (1.3) wurde für viele Typen von Operatoren mit z.T. sehr verschiedenartigen Methoden nachgewiesen und praktisch ausgenutzt. (Eine abstrakte Formulierung der Art (1.3) und mehrere Anwendungen findet man bei L. Collatz [7].) Wir berichten hier über eine sehr einfache abstrakte Theorie für Operatoren M mit der Eigenschaft (1.3). Diese Theorie gestattet es verschiedenartige Typen von Gleichungen in einheitlicher Weise zu behandeln.

Obwohl gerade diese verschiedenartigen Anwendungsmöglichkeiten ein charakteristischer Vorteil der abstrakten Theorie sind, wollen wir uns hier auf ein Anwendungsgebiet beschränken. Und zwar betrachten wir im wesentlichen nur Randwertaufgaben mit einer elliptischen Differentialgleichung zweiter Ordnung. Jedoch zeigt ein kurzer Abschnitt, wie sich parabolische Differentialgleichungen einordnen lassen, und ein anderer Abschnitt berichtet über einige Ergebnisse bei gewöhnlichen Randwertaufgaben vierter Ordnung. Mit den Ideen der abstrakten Theorie erhält man nicht nur die üblichen Aussagen über Ungleichungen bei elliptischen Differentialgleichungen, sondern auch diffizilere Aussagen wie etwa Ergebnisse von Redheffer [18]. (Über weitere Anwendungen der abstrakten Theorie siehe [21], [22], [23]. Andere zusammenfassenden Darstellungen über gewisse Typen von Ungleichungen findet man in [3], [8], [25], [26].)

Für die numerische Anwendung der Monotonie-Eigenschaft (1.3) gibt es verschiedenartige Beispiele. Wir berichten hier über ein Programm zur Lösung der ersten Randwertaufgabe mit einer Differentialgleichung

$$-\Delta u = f(x, y, u).$$

Das Programm liefert eine Näherungslösung und eine zugehörige Fehlerabschätzung.

2. Ungleichungen mit linearen Operatoren

Es seien $\mathfrak{R} = \{u, v, \ldots\}$ und $\mathfrak{S} = \{U, V, \ldots\}$ halbgeordnete lineare Räume. Die (transitive, reflexive und antisymmetrische) Ordnungsrelation werde in beiden Räumen mit $\leqq$ bezeichnet. Der Raum $\mathfrak{R}$ sei Archimedisch, d.h. für feste $u, v \in \mathfrak{R}$ gelte:

$$u \leqq nv \ (n = 1, 2, \ldots) \Rightarrow u \leqq O.$$

Ferner sei in $\mathfrak{R}$ und $\mathfrak{S}$ je eine zweite, mit $\preceq$ bezeichnete Ordnungsrelation definiert. Die zweite Ordnungsrelation in $\mathfrak{R}$ habe die folgende Eigenschaft σ.

Eigenschaft σ: *Ist $u \in \mathfrak{R}$ und $u \succ O$, so gibt es zu jedem $v \in \mathfrak{R}$ eine Nummer n mit $v \leqq nu$.*

In $\mathfrak{S}$ wird verlangt:

$$U \leqq V \prec W \Rightarrow U \prec W, \quad U \prec V \leqq W \Rightarrow U \prec W,$$
$$U \prec V, \ \lambda > 0 \Rightarrow \lambda U \prec \lambda V.$$

Mit M bezeichnen wir einen linearen Operator, der $\mathfrak{R}$ in $\mathfrak{S}$ abbildet.

Definition 2.1. *Der Operator M heiße inverspositiv, wenn für alle $u \in \mathfrak{R}$ gilt:*

$$Mu \geqq O \Rightarrow u \geqq O. \tag{2.1}$$

Satz 2.1. *Ist M inverspositiv, so existiert der inverse Operator M^{-1} und dieser Operator ist positiv, d.h.*

$$U \geqq O \Rightarrow M^{-1}U \geqq O \ \text{ für } \ U \in M\mathfrak{R}.$$

Satz 2.2 [21]. *Die folgenden zwei Voraussetzungen seien erfüllt.*

I.

$$\left.\begin{matrix} w \geqq O \\ Mw \succ O \end{matrix}\right\} \Rightarrow w \succ O \ (\text{für alle } w \in \mathfrak{R}). \tag{2.2}$$

II. *Es existiert ein $z \in \mathfrak{R}$ mit*

$$z \geqq O, \ Mz \succ O. \tag{2.3}$$

Dann ist der Operator M inverspositiv.

Satz 2.3 [22]. (*"Umkehrung" des Satzes 2.2*). *Ist der Operator M inverspositiv, so gelten die folgenden Aussagen:*

1. *Hat die zweite Ordnungsrelation in $\mathfrak{S}$ ebenfalls die Eigenschaft σ, so gilt (2.2).*

2. *Enthält der Wertebereich $M\mathfrak{R}$ mindestens ein Element $r \succ O$, so existiert ein $z \in \mathfrak{R}$ mit der Eigenschaft (2.3) (nämlich die Lösung der Gleichung $Mz = r$).*

Definition 2.2. *Eine für $0 \leqq t \leqq 1$ definierte Funktion $z(t) \in \mathfrak{R}$ heiße stetig in $t_0 \in [0, 1]$, wenn es zu jeder natürlichen Zahl n eine positive Zahl $\delta(n)$ gibt, derart daß*

$$-\frac{1}{n}e \leqq z(t) - z(t_0) \leqq \frac{1}{n}e \quad \text{für} \quad |t - t_0| \leqq \delta(n), \ t \in [0,1].$$

Dabei bedeute e ein beliebiges (festes) Element aus $\mathfrak{R}$.

Es sei nun $\mathfrak{M}$ eine Menge von Operatoren M: $\mathfrak{R} \to \mathfrak{S}$, derart daß für jedes $M \in \mathfrak{M}$ die Voraussetzung I in Satz 2.2 erfüllt ist. Dann gilt der folgende Satz.

Satz 2.4 [22]. *Für die genannte Menge $\mathfrak{M}$ von Operatoren M seien die folgenden Voraussetzungen erfüllt.*

1. a) *M^{-1} existiert für jedes $M \in \mathfrak{M}$.*

b) *Die Wertebereiche aller Operatoren $M \in \mathfrak{M}$ haben mindestens ein gemeinsames Element $r \succ O$.*

c) *Für jedes Paar M_0, $M_1 \in \mathfrak{M}$ existiert eine Schar $M(t) \in \mathfrak{M}$ $(0 \leqq t \leqq 1)$, derart daß $M(0) = M_0$, $M(1) = M_1$ und $M^{-1}(t)\, r$ in $t \in [0, 1]$ stetig ist.*

2. *Mindestens ein Operator $M \in \mathfrak{M}$ ist inverspositiv.*

Dann sind alle Operatoren $M \in \mathfrak{M}$ inverspositiv.

3. Lineare elliptische und parabolische Differentialoperatoren

3.1. Elliptische Differentialoperatoren. Es sei G ein offenes beschränktes Gebiet des n-dimensionalen Euklidischen Raumes $E^n = \{x, y, \ldots\}$, Γ dessen Rand und $\bar{G} = G \cup \Gamma$. Ableitungen einer auf G erklärten (reellwertigen) Funktion $u(x)$, $v(x)$, $w(x)$, $z(x)$ oder auch $\omega(x)$ nach den Koordinaten x_k $(k = 1, 2, \ldots, n)$ von x werden durch tiefgestellte Indices gekennzeichnet, z.B. $\partial^2 u/\partial x_k \partial x_l = u_{kl}$. Sonst, d.h. bei anderen Buchstaben, bedeuten Indices keine Ableitungen, sondern dienen nur zur Unterscheidung.

Es bedeute $L[u]$ einen Differentialausdruck

$$L[u](x) = -\sum_{k,l=1}^{n} a_{kl} u_{kl} + \sum_{j=1}^{n} b_j u_j + cu \tag{3.1}$$

mit auf G definierten Koeffizientenfunktionen $a_{kl}(x) = a_{lk}(x)$, $b_j(x)$, $c(x)$. Für jedes $x \in G$ sei die Matrix $(a_{kl})(x)$ positiv semidefinit:

$$(a_{kl}(x)) \geqq 0 \quad (x \in G), \tag{3.2}$$

und $\lambda_1(x)$ bezeichne ihren kleinsten Eigenwert.

Ferner sei auf Γ ein Randausdruck

$$R[u](x) = -\beta u_\sigma + \alpha u \tag{3.3}$$

mit auf Γ definierten Funktionen $\alpha(x)$, $\beta(x)$ gegeben, derart daß

$$\beta(x) \geqq 0 \quad (x \in \Gamma). \tag{3.4}$$

Dabei bedeute $u_\sigma = \dfrac{\partial u}{\partial \sigma}(x)$ die (einseitige) Ableitung längs einer stetig differenzierbaren Kurve $y = g(\sigma)$ mit $g(0) = x$ und $g(\sigma) \in \bar{G}$ für $0 \leqq \sigma \leqq \varepsilon$. (Diese Kurve kann etwa die innere Normale sein, falls diese definiert ist.) Eine solche Kurve soll für jedes $x \in \Gamma$ mit $\beta(x) \neq 0$ existieren.

Zur Anwendung der abstrakten Sätze definieren wir:

$\mathfrak{R}$ sei die Menge der Funktionen $u \in C_0(\bar{G}) \cap C_2(G)$, für welche die Ableitung u_σ in allen $x \in \Gamma$ mit $\beta(x) \neq 0$ existiert.

$$u \geqq O: \Leftrightarrow u(x) \geqq 0 \quad (x \in \bar{G}) \quad \text{für } u \in \mathfrak{R},$$

$$u \succ O: \Leftrightarrow u(x) > 0 \quad (x \in \bar{G}) \quad \text{für } u \in \mathfrak{R};$$

$\mathfrak{S}$ sei die Menge der Paare $U = (\varphi, \omega)$ mit auf G definierter Funktion $\varphi(x)$ und auf Γ definierter Funktion $\omega(x)$,

$$U \geqq O: \Leftrightarrow \varphi(x) \geqq 0 \quad (x \in G) \quad \text{und } \omega(x) \geqq 0 \quad (x \in \Gamma),$$

$$U \succ O: \Leftrightarrow \varphi(x) > 0 \quad (x \in G) \quad \text{und } \omega(x) > 0 \quad (x \in \Gamma);$$

$$Mu = (L[u],\ R[u]). \tag{3.5}$$

Dieser Operator M ist bei den obigen Definitionen genau dann inverspositiv, falls für $u \in \mathfrak{R}$ gilt:

$$\left.\begin{aligned} L[u](x) &\geqq 0 \quad (x \in G) \\ R[u](x) &\geqq 0 \quad (x \in \Gamma) \end{aligned}\right\} \Rightarrow u(x) \geqq 0 \quad (x \in \bar{G}). \tag{3.6}$$

Hilfssatz 3.1 [21]. *Der durch* (3.5) *definierte Operator M genügt der Voraussetzung* I *in Satz* 2.2.

Damit ergibt sich aus Satz 2.2:

Satz 3.1. *Existiert eine Funktion $z(x) \in \mathfrak{R}$ mit*

$$z(x) \geqq 0 \quad (x \in \bar{G}), \quad L[z](x) > 0 \quad (x \in G), \quad R[z](x) > 0 \quad (x \in \Gamma), \tag{3.7}$$

so gilt die Implikation (3.6) *für $u \in \mathfrak{R}$.*

Die Existenz einer solchen Funktion $z(x)$ ist auch in gewissem Sinne notwendig für (3.6), denn es gilt trivialerweise:

Satz 3.2. *Ist* (3.6) *für $u \in \mathfrak{R}$ richtig und besitzt die Aufgabe*

$$L[u] = r(x) \quad (x \in G), \quad R[u] = s(x) \quad (x \in \Gamma)$$

für irgendzwei Funktionen $r(x) > 0$, $s(x) > 0$ eine Lösung $u = z$, so genügt diese Lösung den Ungleichungen (3.7).

Sehr einfach ist die Anwendung des Satzes 3.1 wenn

$$c(x) > 0 \quad (x \in G), \quad \alpha(x) > 0 \quad (x \in \Gamma) \tag{3.8}$$

gilt, denn dann genügt $z(x) \equiv 1$ den geforderten Ungleichungen.

Um auch Fälle zu erfassen, in denen $c(x)$ den Wert 0 und negative Werte annehmen kann, konstruieren wir eine kompliziertere Funktion z.

Es sei $K \supset \bar{G}$ eine abgeschlossene n-dimensionale Hyperkugel mit Radius r_0. $r(x)$ sei der (euklidische) Abstand des Punktes x vom

Mittelpunkt der Kugel. Ferner bedeute

$$z = z(x) = br_0 + \int_r^{r_0} se^{\rho(s-r_0)}\,ds$$

mit $r = r(x)$ und Konstanten $b > 0$, $\rho \geqq 0$.

Bei Verwendung einer solchen Funktion folgt aus Satz 3.1 der folgende Satz 3.3 [21].

Satz 3.3. *Falls b und ρ so gewählt werden können, daß die Ungleichungen*

$$\alpha(x)\,b - \beta(x) > 0 \quad (x \in \Gamma)$$

und

$$e^{\rho(r-r_0)}\left\{\sum_{k=1}^{n} a_{kk}(x) + r\left(\rho\lambda_1(x) - \sqrt{\sum_{j=1}^{n} b_j^2(x)}\right)\right\} + c(x)\,z(x) > 0 \quad (x \in G)$$

erfüllt sind, gilt (3.6) *für* $u \in \mathfrak{R}$.

Zusatz. *Konstanten b, ρ mit den in Satz* 3.3 *geforderten Eigenschaften existieren z.B. falls die Ungleichungen*

$$\inf\{\alpha(x)\colon x \in \Gamma\} > 0, \quad c(x) \geqq 0 \quad (x \in G),$$

$$\inf\{\lambda_1(x)\colon x \in G\} > 0 \tag{3.9}$$

gelten und die Koeffizienten $b_j(x)$ auf G und $\beta(x)$ auf Γ beschränkt sind.

Die Ungleichung (3.9) fordert, daß der Operator L auf G gleichmäßig elliptisch ist. Gewöhnlich wird das Ergebnis des Zusatzes mit Hilfe des Randmaximum-Satzes [11] bewiesen (siehe etwa Collatz [8]).

3.2. Zusammenhängende Mengen von elliptischen Operatoren. Um die Eigenschaft (3.6) für umfassendere Klassen von Operatoren nachweisen zu können, benutzen wir nun Satz 2.4.

Es sei $\mathfrak{M}$ eine Menge von Operatoren M der Art (3.5). Über diese Menge setzen wir folgendes voraus.

Voraussetzungen: a) Die Koeffizienten a_{kl}, b_j und β sind für alle $M \in \mathfrak{M}$ dieselben, d.h. die Bestandteile

$$\tilde{L}[u] = -\sum_{k,l=1}^{n} a_{kl}u_{kl} + \sum_{j=1}^{n} b_j u_j \quad \text{und} \quad \tilde{R}[u] = -\beta u_\sigma \tag{3.10}$$

sind unabhängig von $M \in \mathfrak{M}$.

b) Die Vektoren $\mathfrak{k} = (c(x), \alpha(x))$ der zu $M \in \mathfrak{M}$ gehörigen Koeffizienten c und α bilden eine konvexe Menge $\mathfrak{K}$, und $\mathfrak{K}$ enthält

einen Vektor $\mathfrak{k} = (c(x), \alpha(x)) = (c_0, c_0)$ mit einer Konstanten $c_0 > 0$.

c) Für jedes $\mathfrak{k} \in \mathfrak{K}$ besitzt die Aufgabe

$$\tilde{L}[u] + c(x)\,u = 0 \quad (x \in G), \qquad \tilde{R}[u] + \alpha(x)\,u = 0 \quad (x \in \Gamma)$$

in $\mathfrak{R}$ nur die triviale Lösung $u(x) \equiv 0$ $(x \in \bar{G})$.

d) Für jedes $\mathfrak{k} \in \mathfrak{K}$ besitzt die Aufgabe

$$\tilde{L}[u] + c(x)\,u = 1 \quad (x \in G), \qquad \tilde{R}[u] + \alpha(x)\,u = 1 \quad (x \in \Gamma) \tag{3.11}$$

eine Lösung $u = z(x) \in \mathfrak{R}$. Ist ferner $\mathfrak{k}_t = t\mathfrak{k}_0 + (1-t)\,\mathfrak{k}_1 \in \mathfrak{K}$ für $0 \leqq t \leqq 1$, so hängen die zugehörigen Lösungen $z_t(x)$ stetig von t ab. (Für diese Voraussetzung d) muß man etwa fordern, daß der Rand Γ und die Koeffizienten in L und R "genügend glatt" sind.)

Unter diesen Voraussetzungen folgt aus Satz 2.4 der folgende Satz 3.4.

Satz 3.4. *Ist $M = (L, R) \in \mathfrak{M}$, so gilt* (3.6) *für $u \in \mathfrak{R}$.*

Zum Beweis mit Hilfe des Satzes 2.4 beachte man, daß $r = (\varphi(x), \omega(x))$ mit $\varphi(x) \equiv 1$, $\omega(x) \equiv 1$ zum Wertebereich aller Operatoren $M \in \mathfrak{M}$ gehört und $z_t(x) = M^{-1}(t)\,r$ ist. Ferner ist der zu $\mathfrak{k} = (c_0, c_0)$ gehörige Operator M inverspositiv, da die zugehörige Funktion $z(x) \equiv c_0^{-1}$ den Ungleichungen (3.7) genügt.

Beispiel. Es sei $\mathfrak{M}$ die Menge der Operatoren $M = (L, R)$ mit

$$L[u] = -\Delta u + c(x)\,u, \quad R[u] = u \tag{3.12}$$

und

$$c(x) > -\nu_1 \quad (x \in G). \tag{3.13}$$

Dabei bedeute ν_1 den kleinsten Eigenwert der Eigenwertaufgabe

$$\Delta\varphi + \nu\varphi = 0 \quad (x \in G), \quad \varphi = 0 \quad (x \in \Gamma).$$

Die Ungleichung (3.13) sichert, daß die obige Voraussetzung c) erfüllt ist, denn mit Hilfe des Gaußschen Satzes erhält man (wenn Γ genügend glatt ist) aus $L[u](x) = 0$ $(x \in G)$, $u(x) = 0$ $(x \in \Gamma)$:

$$0 = \int_G (-\Delta u + cu)\,u\,dx \geqq \int_G (\nu_1 + c)\,u^2\,dx$$

und damit $u(x) \equiv 0$ $(x \in \bar{G})$.

Wenn also Γ genügend glatt ist, gilt (3.6) *für die Operatoren in* (3.12) *unter der Voraussetzung* (3.13).

Ein solches Ergebnis bewies R. Bellman [4] mit Methoden der Variationsrechnung.

3.3. Positiv definite selbstadjungierte Probleme. Für selbstadjungierte Probleme

$$L[u](x)=r(x) \quad (x\in G), \qquad R[u](x)=0 \quad (x\in\Gamma), \tag{3.14}$$

bewiesen Aronszajn und Smith einen Satz, dessen (in diesem Zusammenhang) wesentlichste Aussage unter geeigneten Voraussetzungen folgendermaßen lautet.

Satz 3.5 [1]. *Ein selbstadjungiertes Problem* (3.14) *ist genau dann positiv definit, wenn dazu eine Greensche Funktion existiert und diese positiv ist*:

$$K(x,\xi)\geqq 0 \quad (x,\xi\in G). \tag{3.15}$$

Dabei wird insbesondere vorausgesetzt, daß Γ und die Koeffizienten in L und R "genügend glatt" sind, um u.a. die Existenz der Greenschen Funktion zu sichern. Das Problem heißt hier positiv definit, falls

$$\int_G L[u]\,u\,dx\geqq\mu\int_G u^2\,dx \quad \text{mit} \quad \mu>0$$

für alle genügend glatten $u(x)$, welche die Randbedingungen erfüllen.

Die schwierigere Richtung des Beweises, nämlich der Beweis von (3.15), geht bei Aronszajn und Smith über die Theorie der reproduzierenden (und pseudoreproduzierenden) Kerne. Eine entsprechende Aussage kann man jedoch recht einfach mit Hilfe des Satzes 3.4 erhalten. Wir skizzieren den Verlauf des Beweises, ohne alle erforderlichen Glattheitsvoraussetzungen genau zu formulieren.

Das gegebene Problem (3.14) sei selbstadjungiert und positiv definit, und u_σ bedeute die Ableitung in Richtung der inneren Normalen. Wir können als Normierungsvorschrift annehmen, daß $\beta(x)$ auf Γ nur die Werte 0 und 1 annimmt.

Es bedeute $c_0>0$ eine Konstante, derart daß $c(x)\leqq c_0$ $(x\in G)$, $\alpha(x)\leqq c_0$ $(x\in\Gamma)$, und es bezeichne M_0 den Operator $M_0=(L_0, R_0)$ mit

$$L_0[u]=\widetilde{L}[u]+c_0u, \qquad R_0[u]=\widetilde{R}[u]+c_0u,$$

wobei $\widetilde{L}$ und $\widetilde{R}$ durch (3.10) mit den Koeffizienten von L und R definiert sind. Dieser Operator M_0 ist inverspositiv, denn z.B. genügt $z\equiv\frac{1}{c_0}$ den Ungleichungen (3.7).

Durch Anwendung der Ergebnisse von Abschnitt 3.2 schließen wir, daß dann auch alle Operatoren $M_t=(L_t, R_0)$ $(0\leqq t\leqq 1)$ mit

$$L_t[u]=L[u]+[(1-t)\,c_0+tc(x)]\,u$$

inverspositiv sind, denn die Menge $\mathfrak{M}=\{M_t : 0\leqq t\leqq 1\}$ hat (unter geeigneten Glattheitsvoraussetzungen) alle geforderten Eigenschaften. Insbesondere ist die Voraussetzung c) erfüllt, da alle Operatoren (L_t, R_0) positiv definite Probleme bestimmen.

Insbesondere ist also $M_1=(L, R_0)$ inverspositiv, und wir betrachten nun die M_1 enthaltende Menge

$$\hat{\mathfrak{M}}=\{\hat{M}_s=(L, R_s):\ 0\leqq s\leqq 1\}$$

mit

$$R_s[u]=R[u]+[(1-s)\,c_0+s\alpha(x)]\,u.$$

Auch für diese Menge $\hat{\mathfrak{M}}$ statt $\mathfrak{M}$ gilt die Aussage des Satzes 3.4. Zwar braucht die zugehörige Vektormenge $\hat{\mathfrak{K}}$ kein $\mathfrak{k}=(c_0, c_0)$ mit $c_0>0$ zu enthalten, jedoch enthält $\hat{\mathfrak{K}}$ den Vektor $(c(x), c_0)$, für welchen das Problem (3.11) eine Lösung $z(x)\geqq 0$ besitzt, und dies ist hinreichend.

Insbesondere ist also $\hat{M}_1=(L, R)$ inverspositiv. Für die betrachteten L und R gilt also (3.6), und daraus folgt u.a. (3.15).

3.4. Parabolische Differentialoperatoren. Statt (3.3) sollen jetzt allgemeinere Randoperatoren zugelassen werden, derart daß auch parabolische Probleme erfaßt werden.

Es sei etwa $x_0\in\Gamma$, derart daß Γ in einer Umgebung von x_0 eine C_2-Mannigfaltigkeit bildet. D.h. Γ habe dort eine zweimal stetig differenzierbare Parameterdarstellung $x=x(\xi_1, \ldots, \xi_{n-1})$. Dann darf $R[u]$ in $x=x_0$ die verallgemeinerte Form

$$R[u](x)=-\sum_{k,\,l=1}^{n-1}\alpha_{kl}(x)\frac{\partial^2u}{\partial\xi_k\,\partial\xi_l}+\sum_{j=1}^{n-1}\beta_j(x)\,\frac{\partial u}{\partial\xi_j}-\beta(x)\,\frac{\partial u}{\partial\sigma}+\alpha(x)u \tag{3.16}$$

haben, wobei (α_{kl}) eine positiv semidefinite Matrix bedeutet und weiterhin (3.4) gefordert wird.

Wir betrachten nun Probleme, bei denen $L[u]$ wie bisher definiert ist und $R[u]$ auf Γ die verallgemeinerte Form (3.16) hat. In Punkten $x\in\Gamma$, für welche alle Koeffizienten α_{kl}, β_j verschwinden, braucht die oben beschriebene Glattheit des Randes nicht verlangt zu werden.

Die Definition des Raumes $\mathfrak{R}$ in Abschnitt muß dann dahingehend abgeändert werden, daß von $u\in\mathfrak{R}$ auch die Existenz aller Ableitungen verlangt wird, welche in (3.16) vorkommen.

Dann beweist man den folgenden Hilfssatz 3.2 mit entsprechenden Mitteln wie Hilfssatz 3.1.

Hilfssatz 3.2. *Der durch* (3.1), (3.16) *und* (3.5) *definierte Operator M genügt der Voraussetzung* I *des Satzes* 2.2.

Ein Spezialfall ist nun das folgende parabolische Problem. Es sei $\tilde{G}$ ein beschränktes Gebiet des $(x_1, \ldots, x_{n-1})$-Raumes, $I = \{x_n: 0 < x_n < T\}$ mit $0 < T < \infty$ und $G = \tilde{G} \times I$. Γ_1 bedeute den Randteil

$$\Gamma_1 = \{x: (x_1, \ldots, x_{n-1}) \in \tilde{G},\ x_n = T\}.$$

Auf $\tilde{G} \cup \Gamma_1$ sei ein Differentialoperator

$$\hat{L}[u] = -\sum_{k,l=1}^{n-1} a_{kl}(x)\, u_{kl} + \sum_{j=1}^{n-1} b_j(x)\, u_j + u_n + c(x)\, u$$

mit positiv semidefiniter $(n-1) \times (n-1)$-Matrix $(a_{kl})(x)$ gegeben, auf $\Gamma - \Gamma_1$ ein Randoperator der Art (3.3).

Dann hat man also auf G einen Differentialoperator $L[u] = \hat{L}[u]$ der Art (3.1), auf $\Gamma - \Gamma_1$ einen Randoperator (3.3) und ferner auf dem restlichen Randteil Γ_1 einen Randoperator der Art (3.16). Man kann auf Γ_1 nämlich $R[u] = \hat{L}[u]$ setzen. $\hat{L}[u]$ hat die Form (3.16), wenn man definiert:

$$\xi_i = x_i \quad (i = 1, 2, \ldots, n-1), \qquad \sigma = -x_n,$$
$$\alpha_{kl} = a_{kl} \quad (k, l = 1, \ldots, n-1), \qquad \beta_j = b_j \quad (j = 1, 2, \ldots, n-1),$$
$$\beta = 1, \quad \alpha = c.$$

Die Wahl einer Funktion z ist in diesem Spezialfall erheblich einfacher. Wir verwenden $z = e^{Nx_n}$ mit genügend großem $N > 0$ und setzen noch voraus, daß in $x \in \Gamma - \Gamma_1$ mit $\beta(x) \neq 0$ für die Kurve $y = g(\sigma)$ gilt: $\partial x_n / \partial \sigma = 0$. (Das ist z.B. der Fall, wenn u_σ für $x_n > 0$ die Normalen-Ableitung bedeutet und $\beta(x) = 0$ für $x_n = 0$.)

Satz 3.6 [21]. *Ist* $c(x)$ *auf* $G \cup \Gamma_1$ *nach unten beschränkt und*

$$\inf\{\alpha(x): x \in \Gamma - \Gamma_1\} > 0,$$

so gilt für $u \in \mathfrak{R}$:

$$\left.\begin{aligned} \hat{L}[u](x) &\geqq 0 \quad (x \in G \cup \Gamma_1) \\ R[u](x) &\geqq 0 \quad (x \in \Gamma - \Gamma_1) \end{aligned}\right\} \Rightarrow u(x) \geqq 0 \quad (x \in \bar{G}).$$

4. Ungleichungen mit nichtlinearen Operatoren

Es seien $\mathfrak{R}$ und $\mathfrak{S}$ wie in Abschnitt 2, jedoch bedeute M jetzt einen nichtlinearen Operator, welcher eine Teilmenge $\mathfrak{D} \in \mathfrak{R}$ in $\mathfrak{S}$ abbildet. Der Eigenschaft (2.1) entspricht jetzt die folgende:

$$Mw \leqq Mv \Rightarrow w \leqq v \quad (v, w \in \mathfrak{R}). \tag{4.1}$$

Gilt diese Implikation für alle $w, v \in \mathfrak{D}$, so läßt sich wieder die Existenz von M^{-1} und außerdem die Monotonie dieses Operators beweisen. Für Zwecke der Fehlerabschätzung braucht man (4.1) jedoch oft nur für bekanntes w oder bekanntes v zu beweisen. Ohne Beschränkung der Allgemeinheit nehmen wir im folgenden an, daß das rechtsstehende Element v gegeben sei.

Eine einfache Verallgemeinerung des Satzes 2.2 ist dann der folgende Satz.

Satz 4.1 [21]. *Die folgenden zwei Voraussetzungen seien erfüllt.*

I.

$$\left.\begin{aligned} u &\leqq \tilde{u} \\ Mu &\prec M\tilde{u} \end{aligned}\right\} \Rightarrow u \prec \tilde{u} \quad (u, \tilde{u} \in \mathfrak{D}). \tag{4.2}$$

II. *Es existieren eine Zahl* $\gamma > 0$ *und ein* $z \in \mathfrak{R}$, *derart daß*

$$\begin{cases} z \geqq 0,\ v + \lambda z \in \mathfrak{D} \text{ für } 0 \leqq \lambda \leqq \gamma, \\ Mv \prec M(v + \lambda z) \quad \text{für } 0 < \lambda \leqq \gamma, \end{cases} \tag{4.3}$$

Dann gilt für $w \in \mathfrak{D}$:

$$\left.\begin{aligned} w &\leqq v + \gamma z \\ Mw &\leqq Mv \end{aligned}\right\} \Rightarrow w \leqq v. \tag{4.4}$$

Bemerkung. Gilt Voraussetzung II für alle $\gamma > 0$, so kann $w \leqq v + \gamma z$ in (4.4) wegfallen.

Wir können insbesondere folgenden Spezialfall betrachten.

Spezialfall. M ist die Differenz

$$M = A - B \tag{4.5}$$

eines linearen Operators A: $\mathfrak{R} \to \mathfrak{S}$ und eines nichtlinearen Operators B: $\mathfrak{D} \to \mathfrak{S}$.

Hilfssatz 4.1. *Erfüllt* A *die Voraussetzung* I *des Satzes* 2.2 (*oder* 4.1) *und gilt*

$$u \leqq \tilde{u} \Rightarrow Bu \leqq B\tilde{u} \quad (u, \tilde{u} \in \mathfrak{D}),$$

so erfüllt M *in* (4.5) *die Voraussetzung* I *des Satzes* 4.1.

Um die folgende hinreichende Bedingung für Voraussetzung II formulieren zu können, setzen wir nun ferner voraus, daß $\mathfrak{R}$ und $\mathfrak{S}$ halbgeordnete Banachräume (oder lineare Teilmengen solcher Räume) sind und M eine stetige Fréchet-Ableitung besitzt. (Was genau wir an Axiomen benötigen, entnehme man dem kurzen Beweis des folgenden Hilfssatzes.)

Hilfssatz 4.2. *Die Voraussetzung* II *von Satz* 4.1 *ist erfüllt, falls* M *der folgenden Voraussetzung* II′ *genügt.*

II′. *Es existieren eine Zahl* $\gamma > 0$ *und ein Element* $z \in \mathfrak{R}$, *derart daß*

$$z \geqq O,\ v + \lambda z \in \mathfrak{D}\ (0 \leqq \lambda \leqq \gamma),$$
$$Mv \prec M(v + \gamma z),$$
$$M'(u)\, z \geqq M'(\tilde{u})\, z \quad \text{für} \quad v \leqq u \leqq \tilde{u} \leqq v + \gamma z.$$

Beweis. Für $0 < \lambda \leqslant \gamma$ gilt:

$$0 \prec \frac{1}{\gamma}\{M(v+\gamma z) - Mv\} = M'(v)\, z + \int_0^1 \{M'(v + t\gamma z) - M'(v)\}\, z\, dt \leqq$$

$$\leqq M'(v)\, z + \int_0^1 \{M'(v + t\lambda z) - M'(v)\}\, z\, dt = \frac{1}{\lambda}\{M(v+\lambda z) - Mv\}.$$

Im Spezialfall (4.5) brauchen wir nur vorauszusetzen, daß B eine stetige Fréchet-Ableitung besitzt und erhalten ein Hilfssatz 4.2 entsprechendes Ergebnis, indem wir $M'(u) = A - B'(u)$ einsetzen. Dies Ergebnis liefert zusammen mit Hilfssatz 4.1 das folgende Resultat.

Satz 4.2. *Der Operator* A *erfülle die Voraussetzung* I *des Satzes* 1, *und es sei*

$$0 \leqq B'(u) \leqq B'(\tilde{u}) \text{ für } u \leqq \tilde{u} \quad (u, \tilde{u} \in \mathfrak{D}).$$

Ferner existiere eine Zahl $\gamma > 0$ *und ein Element* $z \in \mathfrak{R}$, *derart daß*

$$z \geqq O,\ v + \lambda z \in \mathfrak{D}\ (0 \leqq \lambda \leqq \gamma),$$
$$Az \succ \frac{1}{\gamma}[B(v+\gamma z) - Bv]. \tag{4.6}$$

Dann gilt für $w \in \mathfrak{D}$:

$$\left.\begin{array}{c} w \leqq v + \gamma z \\ (A-B)w \leqq (A-B)v \end{array}\right\} \Rightarrow w \leqq v.$$

5. Nichtlineare elliptische Differentialoperatoren

5.1. Anwendung des abstrakten Satzes. Es sei G wie in Abschnitt 3.1. Statt (3.1) betrachten wir jetzt einen nichtlinearen Differentialausdruck

$$F_1[u](x) = F_1(x, u, u_j, u_{kl}),$$

der für $x \in G$ und alle Werte der u, u_j, u_{kl} $(j,k,l = 1, 2, \ldots, n)$ erklärt sei (wir sehen x, u, u_j, u_{kl} der Einfachheit halber auch als

unabhängige Variablen der "Funktion" F_1 an). Die Forderung (3.2) wird verallgemeinert zu:

$$F_1(x, u, u_j, u_{kl} + \zeta_{kl}) \leqq F_1(x, u, u_j, \zeta_{kl})$$

für jede Matrix $(\zeta_{kl}) \geqq O$.

Ferner wird auf Γ statt R ein Ausdruck

$$F_2[u] = F_2(x, u, u_\sigma)$$

mit

$$F_2(x, u, u_\sigma + \zeta) \leqq F_2(x, u, u_\sigma) \text{ für } \zeta \geqq 0$$

betrachtet, wobei $F_2(x, u, u_\sigma)$ für $x \in \Gamma$ und alle Werte der u, u_σ erklärt sei.

$\mathfrak{R}$ und $\mathfrak{S}$ seien definiert wie in Abschnitt 3.1, und M bedeute nun den auf $\mathfrak{D} = \mathfrak{R}$ erklärten Operator

$$Mu = (F_1[u],\ F_2[u]). \tag{5.1}$$

Dann bedeutet (4.1):

$$\left.\begin{array}{l} F_1[w](x) \leqq F_1[v](x)\ (x \in G) \\ F_2[w](x) \leqq F_2[v](x)\ (x \in \Gamma) \end{array}\right\} \Rightarrow w(x) \leqq v(x)\ (x \in \bar{G}). \tag{5.2}$$

Hilfssatz 5.1 [21]. *Der Operator* M *in* (5.1) *genügt der Voraussetzung* I *des Satzes* 4.1.

Beispiel 1. Wir betrachten die in den numerischen Beispielen des Abschnittes 7 vorkommenden Operatoren

$$L[u] = -\Delta u + e^u,\ R[u] = u$$

im Falle $n = 2$. Dabei sei G das Quadrat

$$G = \{x:\ |x_1| < 1,\ |x_2| < 1\}.$$

Für $z = 2 + \varepsilon - x_1^2 - x_2^2$ mit $\varepsilon > 0$ ist die Voraussetzung II des Satzes 4.1 bei beliebigem $\gamma > 0$ erfüllt.

Beispiel 2. Im Falle

$$L[u] = -\Delta u - e^u,\ R[u] = u$$

kann man Satz 4.2 mit

$$Au = \begin{pmatrix} -\Delta u \\ u \end{pmatrix}, \quad Bu = \begin{pmatrix} e^u \\ 0 \end{pmatrix}$$

anwenden. Die Bedingung (4.6) lautet dann

$$-\Delta z > e^v \frac{1}{\gamma}[e^{\gamma z} - 1]\ (x \in G),\ z(x) > 0\ (x \in \Gamma).$$

5.2. Zurückführung auf lineare Probleme. Sind etwa die Funktionen F_1 und F_2 für jedes $x \in G$ bzw. $x \in \Gamma$ nach den übrigen Variablen stetig differenzierbar, so kann man die Theorie für lineare Operatoren heranziehen, um hinreichende Bedingungen für die Aussage (5.2) zu erhalten.

Seien z.B. w, v feste Funktionen aus $\mathfrak{R}$ mit $F_1[w](x) \leqq F_1[v](x)$ $(x \in G)$, so folgt für $u = v - w$

$$L[u](x) \geqq 0 \quad (x \in G)$$

mit einem Operator L der Form (3.1) und

$$a_{kl} = \int_0^1 \frac{\partial F}{\partial u_{kl}}(x, w, w_j, \tilde{w}_{kl})\, dt,$$

$$(5.3) \quad b_j = \int_0^1 \frac{\partial F}{\partial u_j}(x, w, \tilde{w}_j, v_{kl})\, dt, \quad c = \int_0^1 \frac{\partial F}{\partial u}(x, \tilde{w}, v_j, v_{kl})\, dt.$$

Dabei ist z.B. $\tilde{w} = tw + (1-t)v$, und $\tilde{w}_j$, $\tilde{w}_{kl}$ sind entsprechend definiert.

Ist insbesondere F_1 ein quasilinearer Operator:

$$(5.4) \qquad F_1[u] = -\sum_{k,l=1}^{n} A_{kl}(x, u, u_j)\, u_{kl} + C(x, u, u_j),$$

so ergibt sich

$$(5.5) \qquad a_{kl}(x) = A_{kl}(x, w, w_j).$$

In entsprechender Weise läßt sich die zweite Ungleichung $F_2[w](x) \leqq F_2[v](x)$ $(x \in \Gamma)$ auf eine lineare Ungleichung $R[u](x) \geqq 0$ $(x \in \Gamma)$ zurückführen.

Beispiel. Für

$$F_1[u] = -(1+u_2^2)\, u_{11} + 2u_1u_2u_{12} - (1+u_1^2)\, u_{22}$$

erhält man:

$$(a_{kl}(x)) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} w_2^2 & -w_1w_2 \\ -w_1w_2 & w_1^2 \end{pmatrix} \geqq \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

und damit $\lambda_1(x) \geqq 1$, und ferner

$$b_1(x) = (w_2 + v_2)\, v_{12} - (w_1 + v_1)\, v_{22},$$
$$b_2(x) = -(w_2 + v_2)\, v_{11} + (w_1 + v_1)\, v_{12},$$
$$c(x) \equiv 0.$$

Um also etwa den Zusatz zu Satz 3.3 anwenden zu können, müßte man wissen, daß diese Koeffizienten $b_1(x)$, $b_2(x)$ auf G beschränkt sind.

Im Falle $F_2[u] = u$ genügt es aber bei diesem Beispiel—wie in anderen Fällen auch,—statt $\bar{G}$ abgeschlossene Teilgebiete $B \subset G$ zu betrachten (siehe etwa den letzten Absatz von Abschnitt 5.4). Auf solchen Mengen B ist die Beschränktheit der b_j gesichert.

5.3. Eine andere Behandlung quasilinearer Aufgaben. Die Ergebnisse, welche man mit dem Verfahren von Abschnitt 5.2 erhält, kann man auch herleiten, indem man die Beweisidee des Satzes 4.1 direkt auf die nichtlinearen Probleme anwendet, dabei jedoch die speziellen Eigenschaften der Differentialoperatoren stärker berücksichtigt als in Satz 4.1. Gleichzeitig bekommt man in dieser Weise auch andere Resultate, wenn man nämlich außerdem statt einer Funktionenschar $v(x) + \lambda z(x)$ eine in λ nichtlineare Schar $v(\lambda, x) = v(x) + z(\lambda, x)$ betrachtet [1]). Wir beschreiben dies für den Fall, daß F_1 ein quasilinearer Operator (5.4) ist und $F_2[u] = R[u]$ die Form (3.3), (3.4) hat.

Es bedeuten v, w wieder zwei feste Funktionen aus $\mathfrak{R}$, wobei

$$(5.6) \qquad (A_{kl}(x, w, w_j)) \geqq 0 \quad (x \in G).$$

Man berechnet

$$(5.7) \qquad F_1[w] - F_1[v] = -\sum_{k,l=1}^{n} A_{kl}(x, w, w_j)(w_{kl} - v_{kl}) +$$
$$+ [F_1(x, w, w_j, v_{kl}) - F_1(x, w, v_j, v_{kl})] +$$
$$+ [F_1(x, w, v_j, v_{kl}) - F_1(x, v, v_j, v_{kl})].$$

Mit $z(\lambda, x) \in \mathfrak{R}$ $(0 \leqq \lambda < \infty)$ bezeichnen wir eine in λ stetige Funktionenschar mit folgenden Eigenschaften:

$$(5.8) \qquad \begin{cases} z(\lambda, x) \begin{cases} > 0 \text{ für } \lambda > 0 \\ = 0 \text{ für } \lambda = 0 \end{cases} & (x \in \bar{G}), \\ z(\lambda, x) \leqq z(\lambda', x) \text{ für } \lambda \leqq \lambda' & (x \in \bar{G}), \\ \lim\limits_{\lambda \to \infty} z(\lambda, x) = \infty \text{ gleichmäßig auf } \bar{G}. \end{cases}$$

Es seien nun die beiden Ungleichungen auf der linken Seite von (5.2) erfüllt, jedoch sei $w(x) \leqq v(x)$ $(x \in \bar{G})$ nicht richtig. Dann gibt es eine kleinste Zahl $\lambda_0 > 0$ mit $w(x) \leqq v(x) + z(\lambda_0, x)$ $(x \in \bar{G})$ und ein $x_0 \in \bar{G}$ mit $w(x) = v(x_0) + z(\lambda_0, x_0)$.

Nimmt man an, daß

$$R[z(\lambda, x)] > 0 \quad \text{für} \quad \lambda > 0 \text{ und } x \in \Gamma$$

[1]) Eine solche in λ nichtlineare Schar $v(\lambda)$ kann man auch in der abstrakten Theorie benutzen und so den Satz 4.1 verallgemeinern.

gilt, so ist $x_0 \in \Gamma$ nicht möglich. Man beweist dies wie bei den linearen Problemen. Daher ist $x_0 \in G$, und man erhält für $x = x_0$ mit $z^0(x) = z(\lambda_0, x)$:

$$w = v + z^0, \quad w_j = v_j + z^0_j, \quad (w_{kl}) \leqq (v_{kl}) + (z^0_{kl}). \tag{5.9}$$

Aus den Beziehungen (5.7), (5.9) möchte man nun einen Widerspruch für eine geeignete Funktionenschar $z(\lambda, x)$ herleiten.

Sind die Funktionen A_{kl}, C bei festem x nach den übrigen Argumenten stetig differenzierbar, so erhält man mit (5.6), (5.7), und (5.9) für $x = x_0$:

$$F_1[w] - F_1[v] = -\sum_{k,l=1}^{n} A_{kl}(x, w, w_j)\, z^0_{kl} + \sum_{j=1}^{n} b_j(x)\, z^0_j + c(x)\, z^0$$

mit den in (5.3) genannten Koeffizienten. Man bekommt also einen Widerspruch, wenn die rechte Seite in der letzten Ungleichung für jedes $x \in G$ positiv ausfällt. Setzt man etwa $z(\lambda, x) = \lambda z(x)$, so ergibt sich die Bedingung $L[z](x) > 0$ $(x \in G)$, wobei L der lineare Operator mit den Koeffizienten (5.3), (5.5) ist. Auf diese Bedingung kommt man auch, wenn man wie in Abschnitt 5.2 linearisiert und dann etwa Satz 3.1 anwendet.

Eine andere Anwendung der hier beschriebenen Idee bringt der folgende Abschnitt.

5.4. Abschwächung der Differenzierbarkeitsvoraussetzungen. Wir betrachten nun einen quasilinearen Operator der speziellen Form

$$F_1[u] = -\sum_{k,l=1}^{n} A_{kl}(x, u_j)\, u_{kl} + C(x, u, u_j) \tag{5.10}$$

und auf Γ wieder den linearen Randoperator $F_2[u] = R[u]$ in (3.3), (3.4).

Es bedeuten w, v zwei feste Funktionen aus $\mathfrak{R}$, derart daß (5.6) gilt und die Ableitungen v_{kl} auf G beschränkt sind: $|v_{kl}(x)| \leqq \tilde{K}$. Wir setzen $K = n^2\tilde{K} + 1$.

Ferner nehmen wir an, daß es zwei für $0 \leqq t < \infty$ stetige isotone Funktionen $\psi_1(t)$ und $\psi_2(t)$ mit folgenden Eigenschaften gibt:

$$\psi_1(0) = \psi_2(0) = 0,$$

$$\begin{cases} |A_{kl}(x, w, v_j + \zeta_j) - A_{kl}(x, w, v_j)| \leqq \psi_1(\|\zeta_j\|) \\ |C(x, w, v_j + \zeta_j) - C(x, w, v_j)| \leqq \psi_1(\|\zeta_j\|), \end{cases} \tag{5.11}$$

mit

$$\|\zeta_j\|^2 = \sum_{j=1}^{n} |\zeta_j|^2;$$

$$C(x, v + \zeta, v_j) - C(x, v, v_j) \geqq -\psi_2(\zeta) \quad \text{für} \quad \zeta \geqq 0. \tag{5.12}$$

Ist z.B. $\psi_1(t)$ streng monoton und gilt

$$\int_0^1 \frac{dt}{\psi_1(t)} = \infty,$$

so heißt ψ_1 eine Osgood-Funktion.

Satz 5.1. *Es gebe eine in λ stetige Schar $z(\lambda) = z(\lambda, x) \in \mathfrak{R}$ $(0 \leqq \lambda < \infty)$ mit den Eigenschaften* (5.8), *derart daß für $\lambda > 0$ die folgenden Ungleichungen erfüllt sind:*

$$-\sum_{k,l=1}^{n} A_{kl}(x, w_j)\, z_{kl}(\lambda) - K\psi_1(\|z_j(\lambda)\|) - \psi_2(z(\lambda)) > 0 \quad (x \in G),$$

$$R[z(\lambda)](x) > 0 \qquad (x \in \Gamma).$$

Dann gilt die Aussage (5.2) *für die betrachteten Funktionen w, v.*

Beweis. Es sei x_0 der in den Ausführungen von Abschnitt 5.3 erwähnte Punkt aus G. Schätzt man dann die rechte Seite in (5.7) mit Hilfe von (5.11), (5.12) ab und benutzt dabei (5.6) und (5.9), so ergibt sich $F[w](x_0) > F[v](x_0)$. (Widerspruch!)

Satz 5.2. *Es seien folgende Voraussetzungen erfüllt:*

a) *$\psi_1(t)$ ist eine Osgood-Funktion,*

b) *$\psi_2(t) \equiv 0$,*

c) *$\beta(x)$ ist auf Γ beschränkt,*

d) *es existieren eine Funktion $\omega(x) \in C_1(\bar{G}) \cap C_2(G)$ und eine Konstante $\varkappa > 0$, derart daß auf G gilt:*

$$\sum_{k,l=1}^{n} A_{kl}(x, w_j)\, \omega_k \omega_l > 1, \qquad \sum_{k,l=1}^{n} A_{kl}(x, w_j)\, \omega_{kl} \leqslant \varkappa. \tag{5.13}$$

Dann gilt die Aussage (5.2) *für die betrachteten Funktionen w, v.*

Beweis. Es genügt zu zeigen, daß eine Funktionenschar $z(\lambda, x)$ existiert, welche die in Satz 5.1 verlangten Eigenschaften besitzt. Eine solche Schar ist hier

$$z(\lambda, x) = k(\lambda, \omega(x))$$

mit

$$k(\lambda, t) = \lambda b + \int_a^t h(\lambda, s)\, ds,$$

$$\int_{h(\lambda, t)}^{1} \frac{ds}{\varkappa s + K\psi_1(\mu s)} = t - a + \frac{1}{\lambda} \quad \text{für} \quad \lambda > 0, \quad h(0, t) = 0,$$

wobei b eine positive Konstante, a eine genügend kleine Konstante und $\| \omega_j(x) \| \leqq \mu \; (x \in \overline{G})$ ist.

Für den Fall $R[u] = u$ wurde der Satz 5.2 von R. Redheffer [18] mit Hilfe eines von ihm hergeleiteten Randmaximumsatzes [17] bewiesen. Redheffer fordert statt d) lediglich, daß für jedes abgeschlossene Teilgebiet $B \subset G$ eine Funktion $\omega \in C_2(B)$ und eine Konstante $\varkappa$ mit den Eigenschaften (5.13) existieren. Damit kommen wir auch hier aus, wie im folgenden gezeigt wird.

Es sei etwa $w(x) \leqq v(x) + m$ mit $m \geqq 0$ auf dem Rand einer solchen Teilmenge B und $F_1[w](x) \leqq F_1[v](x)$ auf G. Dann ist auch $F_1[w](x) \leqq F_1[v + m](x)$ auf B, und die Anwendung des Satzes auf B statt $\overline{G}$ ergibt: $w(x) \leqq v(x) + m$ auf B. Es gelte nun $w(x) \leqq v(x)$ auf Γ, und es bedeute $\{B_n\}$ eine Folge abgeschlossener Gebiete mit $B_n \subset B_{n+1} \subset G$, $\lim B_n = G$ und $w(x) \leqq v(x) + n^{-1}$ auf dem Rand von B_n. Dann hat man also auf B_n: $w(x) \leqq v(x) + \frac{1}{n}$ und damit auf ganz $\overline{G}$: $w(x) \leqq v(x)$.

5.5. Randmaximumsatz. Während sich die Monotonie-Eigenschaft (5.2) unter bestimmten Annahmen auch aus dem Randmaximum-Satz folgern läßt, erhält man umgekehrt aus der Monotonie-Eigenschaft Randmaximumsätze.

Im folgenden sei $F_2[u] = u$, und wir nehmen der Einfachheit halber an, daß die Aussage (5.2) für alle $v, w \in \mathfrak{R}$ richtig sei. Dann gilt der folgende Satz.

Satz 5.3. *Es seien $v, w \in \mathfrak{R}$ zwei Funktionen mit der Eigenschaft*

$$F_1[w](x) \leqq F_1[v](x) \quad (x \in G).$$

Ferner gebe es eine Funktion $z(x) \in \mathfrak{R}$, derart daß

$$z(x) > 0 \quad (x \in \overline{G}),$$

$$F_1[v](x) \leqq F_1[v + \lambda z(x)] \quad (x \in G,\ 0 \leqq \lambda < \infty). \tag{5.14}$$

Ist dann

$$m \geqq 0 \tag{5.15}$$

für

$$m = \max \left\{ \frac{v(x) - w(x)}{z(x)} : x \in \Gamma \right\},$$

so gilt

$$\frac{v(x) - w(x)}{z(x)} \leqq m \quad (x \in \overline{G}). \tag{5.16}$$

Hat man statt (5.14) *sogar*

$$F_1[v] = F_1[v + \lambda z] \quad (x \in G;\ -\infty < \lambda < \infty),$$

so gilt (5.16) *ohne die einschränkende Voraussetzung* (5.15).

Dieses Ergebnis erhält man, wenn man die Aussage (5.2) auf w und $v + mz$ statt v anwendet.

Für $z(x) \equiv 1$ ergibt sich der übliche schwache Randmaximumsatz. Um den starken Randmaximumsatz zu beweisen, nach dem das Maximum nur auf Γ angenommen wird, müßte man stärkere Monotonie Aussagen benutzen. In diesem Zusammenhang sei darauf aufmerksam gemacht, daß man die zweite Ordnungsrelation in $\mathfrak{S}$ auch etwa folgendermaßen definieren kann:

$$U = (\varphi, \omega) \succ O: \Leftrightarrow \begin{cases} \varphi(x) \geqq 0,\ \varphi(x) \not\equiv 0 & (x \in G), \\ \omega(x) > 0 & (x \in \Gamma). \end{cases}$$

6. Gewöhnliche Randwertaufgaben vierter Ordnung

Das Monotonie-Verhalten von Differentialoperatoren vierter Ordnung ist erheblich komplizierter als das von Operatoren zweiter Ordnung, und dementsprechend schwieriger ist auch die Anwendung des Satzes 2.2. Wir wollen hier nun einen gewissen Eindruck vermitteln, wie man vorgehen kann [24].

Wir betrachten Differentialoperatoren

$$L[u] = (au'')'' - (bu')' + \beta u' + cu$$

auf dem Intervall $[0, 1]$ und (der Einfachheit halber) nur Randbedingungen der Form

$$\begin{cases} u(0) = 0,\ \alpha_1 u'(0) + \alpha_2 u''(0) = 0 \text{ mit } |\alpha_1| + |\alpha_2| > 0, \\ u(1) = 0,\ \beta_1 u'(1) + \beta_2 u''(1) = 0 \text{ mit } |\beta_1| + |\beta_2| > 0. \end{cases} \tag{6.1}$$

(Die Überlegungen lassen sich aber weitgehend auf allgemeinere Randbedingungen übertragen.) Ebenfalls zur Vereinfachung wird angenommen, daß alle in diesem Abschnitt vorkommenden Funktionen von x auf $[0, 1]$ reellanalytisch sind.

Es bezeichne $\mathfrak{S}$ die Menge dieser analytischen Funktionen und $\mathfrak{R}$ die Teilmenge der Funktionen aus $\mathfrak{S}$, welche die Randbedingungen (6.1) erfüllen.

Wir fragen nach hinreichenden Bedingungen für die folgende Aussage:

(6.2) $L[u](x) \geqq 0 \; (0 \leqq x \leqq 1) \Rightarrow u(x) \geqq 0 \; (0 \leqq x \leqq 1)$ für $u \in \mathfrak{R}$.

Diese Aussage ist gleichbedeutend damit, daß eine zugehörige Greensche Funktion existiert und positiv ist.

Um Satz 2.2 anwenden zu können, untersuchen wir zunächst, wann die folgende Eigenschaft vorliegt:

$$(6.3)\quad L[u](x) \geqq 0\ (0 \leqq x \leqq 1) \Rightarrow \begin{cases} u(x) \geqq 0 \\ \qquad\qquad (0 \leqq x \leqq 1) \text{ für } u \in \mathfrak{R}. \\ N[u](x) \geqq 0 \end{cases}$$

Dabei bedeutet

$$(6.4)\qquad N[u] = -apu'' + ap'u' + Pu$$

einen Differentialoperator zweiter Ordnung mit geeigneten Funktionen p und P in den Koeffizienten.

Zur Anwendung der Theorie in Abschnitt 2 definieren wir $\mathfrak{R}$ und $\mathfrak{S}$ wie oben beschrieben,

$$Mu = L[u],$$

$$u \geqq 0: \Leftrightarrow \begin{cases} u(x) \geqq 0 \\ \qquad\qquad (0 \leqq x \leqq 1) \quad \text{für} \quad u \in \mathfrak{R}, \\ N[u](x) \geqq 0 \end{cases}$$

$$U \geqq 0: \Leftrightarrow U(x) \geqq 0 \qquad (0 \leqq x \leqq 1) \quad \text{für} \quad U \in \mathfrak{S}.$$

Ferner seien die jeweils zweiten Ordnungsrelationen in $\mathfrak{R}$ und $\mathfrak{S}$ mit Hilfe der Eigenschaft σ erklärt. D.h. also z.B., daß $u \in \mathfrak{R}$ genau dann $\succ O$ ist, falls zu jedem $v \in \mathfrak{R}$ eine Nummer n mit $v \leqq nu$ existiert. In $\mathfrak{R}$ hängt diese zweite Ordnungsrelation nicht nur von den gegebenen Randbedingungen, sondern auch von den gewählten Funktionen p, P ab.

Während bei Differentialoperatoren zweiter Ordnung die in Voraussetzung I geforderte Eigenschaft (2.2) unter schwachen Bedingungen recht einfach bewiesen werden kann, macht Voraussetzung I hier am meisten Schwierigkeiten. Sie ist erfüllt, falls p und P bestimmte Eigenschaften haben. Ein Beispiel gibt der folgende Satz.

Satz 6.1 [24]. *Es gebe zwei Funktionen $p(x)$ und $P(x)$, derart daß folgende Gleichungen und Ungleichungen erfüllt sind:*

$$(ap')'' - (bp)' + \beta p + 2P' = 0 \quad (0 \leqq x \leqq 1),$$

$$(aP'')' + bP' + \beta P - 2aCp' - (aC)'p \begin{cases} \geqq 0 \text{ für } 0 \leqq x \leqq \xi, \\ \leqq 0 \text{ für } \xi \leqq x \leqq 1, \end{cases}$$

$$p(x) > 0 \quad (0 < x < 1)$$

mit einer oberen Schranke C von c:

$$c(x) \leqq C(x) \quad (0 \leqq x \leqq 1)$$

und einer Zahl $\xi \in [0, 1]$;

$$(6.5)\qquad p = p' = ap'' + P = 0,\ p'' > 0 \text{ für } x = 0 \text{ und } x = 1.$$

$N[u]$ bedeute dann den mit diesen Funktionen p, P definierten Differentialausdruck (6.4).

Ferner existiere eine Funktion $z \in \mathfrak{R}$, derart daß

$$z(x) \geqq 0, \quad N[z](x) \geqq 0, \quad L[z](x) > 0 \quad (0 \leqq x \leqq 1).$$

Dann gilt (6.3) *für alle $u \in \mathfrak{R}$.*

Bemerkung. Bei allgemeineren Randbedingungen als (6.1) erhält man statt (6.5) kompliziertere Randbedingungen für p und P.

Wenn der Nachweis der Eigenschaft (6.2) das eigentliche Ziel der Untersuchungen ist, muß man nun fragen, ob es Funktionen p und P gibt, derart daß (6.2) und (6.3) im wesentlichen unter den gleichen Voraussetzungen gelten. Derartige Funktionen kann man in der Tat konstruieren. Und zwar haben sie die Form

$$p = \varphi\psi' - \varphi'\psi, \qquad P = -a(\varphi'\psi'' - \varphi''\psi'),$$

wobei φ und ψ gewisse Lösungen der homogenen Differentialgleichung $L[u](x) = 0$ sind, welche in notwendigen Bedingungen für die Eigenschaft (6.2) vorkommen. Darauf im Einzelnen einzugehen, würde an dieser Stelle zu weit führen.

Es sei darauf aufmerksam gemacht, daß jedoch auch die stärkere Eigenschaft (6.3) für einen festen Operator $N[u]$ von unmittelbarem Interesse sein kann.

Levin [13] und Čičkin [6] haben für Mehrpunktrandwertaufgaben n-ter Ordnung gezeigt, wie die Positivität der Greenschen Funktion mit dem oszillatorischen Verhalten der Differentialgleichung zusammenhängt. Diese Ergebnisse sind auf den Spezialfall $\alpha_2 = \beta_2 = 0$ der hier betrachteten Randbedingungen anwendbar.—Andere Bedingungen für die Positivität der Greenschen Funktion werden für spezielle Fälle von Problemen vierter Ordnung (auch mit partiellen Differentialoperatoren) von Aronszajn u. Smith [1] hergeleitet. Die abstrakte Theorie dieser Autoren bezieht sich allgemein auf Hilbert-Räume mit reproduzierendem Kern. Numerische Beispiele für die Anwendung der Monotonie-Eigenschaft bei Differentialgleichungen vierter Ordnung werden in [22] gegeben.

7. Fehlerabschätzung

7.1. Allgemeine Bemerkungen. Gegeben sei eine Gleichung

$$(7.1)\qquad Mu = r.$$

Dabei bedeutet M einen (nicht notwendig linearen) Operator, welcher eine Teilmenge $\mathfrak{D}$ eines halbgeordneten linearen Raumes $\mathfrak{R}$ in einen halbgeordneten linearen Raum $\mathfrak{S}$ abbildet, und es sei $r \in \mathfrak{S}$.

Der Operator M habe die Eigenschaft, daß für zwei feste Elemente $v, w \in \mathfrak{D}$ und alle $u \in \mathfrak{R}$ folgendes gilt:

$$Mu \leqq Mv \Rightarrow u \leqq v, \qquad Mw \leqq Mu \Rightarrow w \leqq u.$$

Hinreichende Bedingungen für die erste dieser Aussagen wurden in den vorangehenden Abschnitten besprochen. Die zweite läßt sich auf die erste zurückführen, indem man zur entgegengesetzten Ordnungsrelation übergeht.

Hat die Gleichung (7.1) eine Lösung $u^* \in \mathfrak{D}$, so gilt dann:

$$Mw \leqq r \leqq Mv \Rightarrow w \leqq u^* \leqq v. \tag{7.2}$$

Solche oberen und unteren Schranken v, w für die Lösung u^* kann man oft mit Hilfe einer Näherung $\varphi \in \mathfrak{D}$ für u^* konstruieren. Ist $v = \varphi + \rho z$, $w = \varphi - \rho z$ mit $z \geqq 0$ aus $\mathfrak{R}$ und reellem ρ, so geht die Aussage (7.2) über in:

$$\begin{cases} M(\varphi - \rho z) - M\varphi \leqq d[\varphi] \leqq M(\varphi + \rho z) - M\varphi \Rightarrow \\ \Rightarrow -\rho z \leqq u^* - \varphi \leqq \rho z, \end{cases} \tag{7.3}$$

wobei

$$d[\varphi] = -M\varphi + r$$

den "Defekt" der Näherung φ bedeutet. (Allgemeiner kann man den Ansatz $v = \varphi + \rho_1 z_1$, $w = \varphi - \rho_2 z_2$ machen.) Hat insbesondere M die Form (4.5) (mit linearem Operator A), so bedeutet (7.3):

$$\begin{cases} -\rho Az + [B\varphi - B(\varphi - \rho z)] \leqq d[\varphi] \leqq \rho Az - [B(\varphi + \rho z) - B\varphi] \Rightarrow \\ \Rightarrow -\rho z \leqq u^* - \varphi \leqq \rho z. \end{cases} \tag{7.4}$$

Wir haben für das in den Fehlerschranken auftretende Element z denselben Buchstaben verwendet, wie für das Element in der jeweils zweiten Voraussetzung der Sätze 2.2 und 4.1. Das geschah deshalb, weil man in der Tat in beiden Zusammenhängen oft dasselbe Element (oder doch sehr ähnliche Elemente) benutzen kann. Ist z.B. M linear, so folgt aus (7.4), daß wenigstens $Mz \geqq O$ gilt, während in (2.3) $Mz \succ O$, also "wenig mehr" gefordert wird.

Die Existenz einer Lösung $u^* \in \mathfrak{D}$ muß bei der obigen Fehlerabschätzung vorausgesetzt oder mit anderen Mitteln bewiesen werden. Dazu machen wir nun noch einige Bemerkungen.

Die obigen Ungleichungen sind gewissen Formeln sehr ähnlich, auf welche man kommt, wenn man einen Fixpunktsatz (etwa den Satz von Schauder) zum Existenzbeweis benutzt.

Gilt z.B.: $u \leqq \tilde{u} \Rightarrow Bu \leqq B\tilde{u}$ $(u, \tilde{u} \in \mathfrak{D})$, ist A inverspositiv und $Tu = A^{-1}(Bu + r)$ für $u \in \mathfrak{D}$ definiert, so hat (7.4) zur Folge, daß der Operator T die Menge $\mathfrak{K} = \{u \in \mathfrak{R} : \varphi - \alpha z \leqq u \leqq \varphi + \alpha z\}$ in sich abbildet (sofern $\mathfrak{K} \subset \mathfrak{D}$). Damit ist eine der wesentlichen Voraussetzungen der meisten Fixpunktsätze erfüllt.

Ist B nicht isoton, jedoch

$$-\tilde{B}(\sigma z) \leqq Bu - B\varphi \leqq \tilde{B}(\sigma z) \quad \text{für} \quad -\sigma z \leqq u - \varphi \leqq \sigma z$$

mit einem geeigneten Operator $\tilde{B}$, so bildet T (unter sonst gleichen Voraussetzungen) das Intervall $\{u \in R: \varphi - \sigma z \leqq u \leqq \varphi + \sigma z\}$ in sich ab, falls

$$-\sigma Az + \tilde{B}(\sigma z) \leqq d[\varphi] \leqq \sigma Az - \tilde{B}(\sigma z) \tag{7.5}$$

gilt.

Die Voraussetzung, daß A inverspositiv ist, erfordert dabei oft nur das Nachprüfen der Voraussetzung I von Satz 2.2, dann nämlich, wenn aus (7.4) bzw. (7.5) bereits $Az \succ O$ folgt.

Auch hinreichende Bedingungen für die Konvergenz des Iterationsverfahrens $u_{n+1} = Tu_n$ könnten oft in der Gestalt (7.4) oder (7.5) geschrieben werden.

Hat man nun auf irgendeine Art die Existenz einer Lösung $u^* \leqslant \varphi + \sigma z$ mit $\sigma > \rho$ nachgewiesen, so kann man in Satz 4.1 $\gamma = \sigma - \rho$ setzen.

7.2. Die erste Randwertaufgabe für die Differentialgleichung $-\Delta u = f(x, y, u)$. Das in Abschnitt 7.1 beschriebene allgemeine Abschätzungsprinzip haben wir auf Randwertaufgaben folgender Art praktisch angewendet:

$$\begin{aligned} -\Delta u &= f(x, y, u) \quad \text{für} \quad (x, y) \in G, \\ u &= s(x, y) \quad \text{für} \quad (x, y) \in \Gamma. \end{aligned}$$

Dabei ist (x, y) statt (x_1, x_2) geschrieben; sonst benutzen wir jedoch die Bezeichnungen der vorangehenden Abschnitte (im Falle $n = 2$). Die Funktion $f(x, y, u)$ sei etwa für $(x, y) \in G$, $-\infty < u < \infty$ stetig differenzierbar. Wir setzen zur Abkürzung:

$$A_0 u = -\Delta u, \quad B_0 u = f(x, y, u), \quad B_0'(u) = f_u(x, y, u).$$

Bedeutet nun φ eine Näherung für eine Lösung u^* und

$$\delta[\varphi] = -A_0\varphi + B_0\varphi$$

so läßt sich (7.3) bzw. (7.4) in der folgenden Form schreiben.

Die Ungleichungen

$$\begin{aligned} -\rho A_0 z + [B_0\varphi - B_0(\varphi - \rho z)] \leqq \delta[\varphi] \leqq \\ \leqq \rho A_0 z - [B_0(\varphi + \rho z) - B_0\varphi] \quad \textit{auf} \quad G, \end{aligned} \tag{7.6}$$

$$|\varphi - s| \leqq \rho z \quad \textit{auf} \quad \Gamma \tag{7.7}$$

implizieren die Fehlerabschätzung

$$|u^* - \varphi| \leqq \rho z \quad \text{auf} \quad \bar{G}. \tag{7.8}$$

Damit diese Abschätzung eine möglichst scharfe Aussage über die Lösung macht, möchte man zunächst eine Näherung φ mit "möglichst kleinem" Defekt $d[\varphi] = \begin{pmatrix} \delta[\varphi] \\ -\varphi+s \end{pmatrix}$ und dann eine möglichst kleine Konstante ρ mit der Eigenschaft (7.6), (7.7) ermitteln. Beide Probleme sind nichtlinear und werden für die praktische Durchführung linearisiert. Im Einzelnen verläuft das benutzte Verfahren folgendermaßen.

Schritt D (*Differenzenverfahren*; dieser Schritt entfällt bei linearen Problemen): Mit dem Differenzenverfahren werden in den Knotenpunkten (x_i, y_k) eines quadratischen Gitters Näherungswerte $\tilde{u}_{ik}$ berechnet. Die dabei auftretenden nichtlinearen Gleichungssysteme werden iterativ gelöst; und zwar haben wir eine Kombination aus dem Picardschen Iterationsverfahren und der "Successive Overrelaxation" verwendet.

Schritt A (*Approximation*): Statt des exakten Defektes $\delta[\varphi]$ wird in den Knotenpunkten der in φ lineare Näherungsdefekt

$$\tilde{\delta}[\varphi] = -A_0\varphi + B_0\tilde{u} + B_0'(\tilde{u})(\varphi - \tilde{u})$$

betrachtet. Die Parameter $\alpha_1, \ldots, \alpha_m$ in einem dem Problem entsprechenden linearen Ansatz

$$\varphi = \varphi_0 + \alpha_1\varphi_1 + \ldots + \alpha_m\varphi_m$$

werden so bestimmt, daß $\tilde{\delta}[\varphi]$ und $\tilde{\tilde{\delta}}[\varphi] = -\varphi + s$ im Sinne folgender Normen möglichst klein ausfallen.

Fall 1, *diskrete Gauß-Approximation*:

$$\left\| \begin{pmatrix} \tilde{\delta} \\ \tilde{\tilde{\delta}} \end{pmatrix} \right\|^2 = \sum\nolimits_1 \frac{|\tilde{\delta}(x_i, y_k)|^2}{W_{ik}^2} + \sum\nolimits_2 \frac{|\tilde{\tilde{\delta}}(x_i, y_k)|^2}{w_{ik}^2}. \tag{7.9}$$

Dabei erstreckt sich die erste Summe über das Innere und den Rand des Gitter-Gebietes und die zweite Summe über den Gitter-Rand. Als Gewichte W_{ik} und w_{ik} wurden benutzt: entweder

$$W_{ik} = 1, \quad w_{ik} = 1 \text{ (für alle vorkommenden Indices)} \tag{7.10}$$

oder

$$W_{ik} = (A_0 z - B_0'(\tilde{u})z)(x_i, y_k), \quad w_{ik} = z(x_i, y_k). \tag{7.11}$$

Die letzten Gewichte enthalten bereits die in der Fehlerabschätzung benutzte Funktion z.

Fall 2, *diskrete Tschebyscheff-Approximation*:

$$\left\| \begin{pmatrix} \tilde{\delta} \\ \tilde{\tilde{\delta}} \end{pmatrix} \right\| = \max\left\{ \max_1 \frac{|\tilde{\delta}(x_i, y_k)|}{W_{ik}}, \max_2 \frac{|\tilde{\tilde{\delta}}(x_i, y_k)|}{w_{ik}} \right\}, \tag{7.12}$$

wobei $\max_1$ die Punkte des Gittergebietes und $\max_2$ die des Gitterrandes betrifft.

Im 1. Fall ist ein lineares Gleichungssystem zu lösen, im 2. Falle ein Problem der linearen Optimierung.

Als Fall 3 wurde noch eine *diskrete Orthogonalitätsmethode* verwendet. Darüber wurde bereits an anderer Stelle [23] berichtet.

Schritt F (*Fehlerabschätzung*): Mit einer geeignet gewählten Funktion z wird zunächst eine Konstante ρ_0 bestimmt, derart daß

$$\begin{aligned} &|\delta[\varphi]| \leqq \rho_0 (A_0 z - B_0'(\varphi) z) && \text{auf } \bar{G}, \\ &|\varphi - s| \leqq \rho_0 z && \text{auf } \Gamma. \end{aligned}$$

Dann wird $\rho = (1 + \varepsilon)\rho_0$ gesetzt mit einer kleinen Zahl $\varepsilon > 0$ (z.B. $\varepsilon = 0{,}01$) und nachgeprüft, ob die Ungleichungen (7.6) mit dieser Zahl erfüllt sind. Ist dies der Falle, so gilt die Fehlerabschätzung (7.8).

7.3. Numerische Beispiele. Bei den folgenden Beispielen wurde das in Abschnitt 7.2 beschriebene Verfahren benutzt. Diese Beispiele wurden bereits früher in anderer Weise behandelt [23].

Beispiel 1a. Es sei G das Quadrat

$$G = \{(x, y): |x| < 1, \ |y| < 1\}, \tag{7.13}$$

und die Aufgabe laute

$$-\Delta u = -e^u \quad (x \in G), \quad u = 0 \quad (x \in \Gamma).$$

Es ist unzweckmäßig, die Abschätzungsmethode des Abschnittes 7.2 auf dies Problem direkt anzuwenden, da die zweiten Ableitungen der Lösung in den Eckpunkten des Quadrates nicht beschränkt bleiben. Eine Transformation $v = u - p$ mit geeigneter Funktion $p(x, y)$ behebt diese Singularitäten. Die Funktion p besteht aus vier Summanden vom Typ $\pi^{-1}\,\mathrm{Im}\,(z^2 \log z)$. Zu jeder Ecke gehört ein Summand; z. B. ist $z = 1 + x + i(1 + y)$ für die Ecke $(x, y) = (-1, -1)$.

Die Abschätzungsmethode wurde auf das transformierte Problem

$$-\Delta v = -e^{v+p} \quad (x \in G), \quad v = p \quad (x \in \Gamma)$$

angewendet.

Im Schritt D wurde die Maschenweite $h = 0{,}04$ benutzt.

Im Schritt A wurde eine Ansatzfunktion gewählt, welche die Randbedingungen erfüllt:

$$\varphi = \varphi_0 + (1-x^2)(1-y^2)\,[\alpha_1\psi_1(x) + \alpha_2\psi_2(x) + \ldots + \alpha_m\psi_m(x)]. \tag{7.14}$$

Dabei ist $\varphi_0 = -\psi_0$ mit

$$\pi\psi_0 = [H(x) + H(y) - H(1)], \quad H(x) = h(x) + h(-x),$$

$$h(x) = 2(1+x)\log[4+(1+x)^2] + [4-(1+x)^2]\operatorname{arctg}\tfrac{1}{2}(1+x) - \pi,$$

und die ψ_i bedeuten Polynome mit geeigneten Symmetrie-Eigenschaften: $1,\ x^2+y^2,\ x^2y^2,\ \ldots$.

Im Schritt F wurde $z = 2 - x^2 - y^2$ benutzt. Es ergab sich die Fehlerabschätzung:

$$|v^* - \varphi| = |u^* - p - \varphi| \leqq \rho(2 - x^2 - y^2).$$

Dabei entnehme man ρ und $\varphi(0,0)$ der folgenden Tabelle. In dieser Tabelle bedeutet "G-Appr." bzw. "T-Appr.", daß beim Schritt A im Sinne der Norm (7.9) bzw. (7.12) approximiert wurde, und die Formelnummern (7.10), (7.11) zeigen an, welche Gewichte benutzt wurden. m ist die Zahl der freien Parameter in der Ansatzfunktion φ.

Numerisches Verfahren	ρ	$\varphi(0,0)$
G-Appr. (7.10) $m=4$	0,000 729	−1,127 690
T-Appr. (7.10) $m=4$	0,000 462	−1,127 612
G-Appr. (7.10) $m=6$	0,000 382	−1,127 663
T-Appr. (7.10) $m=6$	0,000 314	−1,127 684

In allen Fällen war:

$$\tilde{v}(0,0) = -1{,}127\,655, \quad p(0,0) = 0{,}882\,454. \tag{7.15}$$

Beispiel 1b. Dieses Beispiel unterscheidet sich vom Beispiel 1a im wesentlichen dadurch, daß φ nicht die Randbedingungen erfüllt. Es wurden verwendet:

$$\varphi = \alpha_1\psi_1 + \alpha_2\psi_2 + \ldots + \alpha_m\psi_m \tag{7.16}$$

mit den ψ_i aus Beispiel 1a und

$$z = 2{,}1 - x^2 - y^2. \tag{7.17}$$

Numerische Resultate:

$$|v^* - \varphi| = |u^* - p - \varphi| \leqq \rho(2{,}1 - x^2 - y^2).$$

Numerisches Verfahren	ρ	$\varphi(0,0)$
G-Appr. (7.11) $m=9$	0,000 755	−1,127 664
T-Appr. (7.11) $m=9$	0,000 294	−1,127 635

$\tilde{v}(0,0)$ und $p(0,0)$ wie in (7.15).

Beispiel 2a. Es sei G wie in (7.13), jedoch laute die Randwertaufgabe jetzt:

$$-\Delta u = e^u \quad (x \in G), \quad u = 0 \quad (x \in \Gamma).$$

In diesem Falle wurde zunächst die Transformation $v = u + p$ durchgeführt (p wie in Beispiel 1a). Im übrigen unterschied sich das Vorgehen von dem in Beispiel 1a nur dadurch, daß in (7.14) $\varphi_0 = \psi_0$ gewählt wurde.

Numerische Resultate:

$$|v^* - \varphi| = |u^* + p - \varphi| \leqq \rho(2 - x^2 - y^2).$$

Numerisches Verfahren	ρ	$\varphi(0,0)$
G-Appr. (7.10) $m=4$	0,001 935	1,278 038
T-Appr. (7.10) $m=4$	0,003 644	1,277 976
G-Appr. (7.10) $m=6$	0,000 916	1,278 071
T-Appr. (7.10) $m=6$	0,001 198	1,277 953

$$\tilde{v}(0,0) = 1{,}278\,103, \quad p(0,0) = 0{,}882\,542. \tag{7.18}$$

Beispiel 2b. Hier handelte es sich um dieselbe Aufgabe wie in Beispiel 2a. Es wurden jedoch φ und z in (7.16), (7.17) verwendet.

Numerische Resultate:

$$|v^* - \varphi| = |u^* + p - \varphi| \leqq \rho\,(2{,}1 - x^2 - y^2).$$

Numerisches Verfahren	ρ	$\varphi\,(0,0)$
G-Appr. (7.11) $m=9$	0,002 931	1,277 948
T-Appr. (7.11) $m=9$	0,001 267	1,277 382

$\tilde{v}\,(0,\,0)$ und $p\,(0,\,0)$ wie in (7.17).

Bemerkung 1. Im Schritt F wurde die Ungleichung (7.6) lediglich für die Gitterpunkte eines Gitters mit der halben Maschenweite 0,002 nachgeprüft, und es wurden Höhenlinien der Defekt-Funktion $\delta\,[\varphi]\,(x, y)$ mit Hilfe der Maschine gezeichnet. Es ist jedoch ein Programm in Vorbereitung, mit dessen Hilfe auch die Prüfung der Ungleichungen (7.6) unter Verwendung einer Intervallarithmetik für das ganze Gebiet G exakt erfolgen soll.

Bemerkung 2. Bei allen Beispielen war die für Schritt A und F erforderliche Rechenzeit geringer, z. T. wesentlich geringer, als die für Schritt D benötigte.

Die numerischen Rechnungen wurden am "Institut für Instrumentelle Mathematik des Landes Nordrhein Westfalen" in Bonn durchgeführt. Ich danke meinen dortigen Mitarbeitern, den Herren Lautenbach, Schock und Schütz.

Mathematisches Institut
der Universität zu Köln,
Bundesrepublik Deutschland

LITERATUR

[1] Aronszajn N., Smith K. T., *Amer. J. Math.*, **79** (1957), 611-622.
[2] Азбелев Н. В., Цалюк З. Б., *Украинский математический журнал*, 10 (1958), 3-12.
[3] Beckenbach E. F., Bellman R., Inequalities, Berlin-Gottingen-Heidelberg, 1961. Русский перевод: Беккенбах Э., Беллман Р., Неравенства, «Мир», М., 1965.
[4] Bellman R., *Boll. Un. Math.*, **12** (1957), 411-413.
[5] Чаплыгин С. А., Собрание сочинений, ГИТТЛ, М., 1948.
[6] Чичкин Е. С., *Изв. Высш. Учебн. Завед., Математика*, № 2 (27) (1962), 170-179.

[7] Collatz L., *Arch. Math.*, **3** (1952), 365-376.
[8] Collatz L., The numerical treatment of differential equations, 3. Aufl., Berlin-Göttingen-Heidelberg, 1961. Русский перевод 1 изд.: Коллатц Л., Численные методы решений дифференциальных уравнений, ИЛ, М., 1953.
[9] Collatz L., Schröder J., *Numer. Math.*, **1** (1959), 61-72.
[10] Courant R., Hilbert D., Methods of mathematical physics, II, New York, 1962. Русский перевод: Курант Р., Уравнения с частными производными, «Мир», М., 1964.
[11] Hopf E., *Sitzungsber. Preuß. Akad. Wiss., Phys. Math. Kl.*, 1927, 147-152.
[12] Канторович Л. В., *Acta Math.*, **71** (1939), 63-97.
[13] Левин А. Ю., *Докл. Акад. Наук УССР*, **148** (1963), 512-515; **156** (1964), 1281-1284. Übersetzungen in: *Soviet Math.*, 4 (1963), 121-124; **5** (1964), 818-821.
[14] Meyer A., *Arch. Rational Mech. Anal.*, **6** (1960), 277-298.
[15] Morgenstern D., Beiträge zur nichtlinearen Funktionalanalysis, Diss. T.U., Berlin, 1952.
[16] Perron E., *Math. Ann.*, **76** (1915), 471-484.
[17] Redheffer R. M., *Monatsh. Math.*, **66** (1962), 32-42.
[18] Redheffer R. M., *J. reine angew. Math.*, **211** (1962), 70-77.
[19] Redheffer R. M., *Bull. Amer. Math. Soc.*, **69** (1963), 497-500.
[20] Schröder J., *Arch. Rational Mech. Anal.*, **4** (1959), 177-192.
[21] Schröder J., *Arch. Rational Mech. Anal.*, **8** (1961), 408-434; **14** (1963), 38-60.
[22] Schröder J., Beitrag in: Error in Digital Computation II, L. B. Rall (ed.), New York, 1965, 141-179.
[23] Schröder J., Proc. IFIP-Congress 65, New York, 1965, 187-194.
[24] Schröder J., *Math. Z.*, **90** (1965), 429-440; **92** (1966), 75-94; **96** (1967), 89-110.
[25] Szarski F., Differential Inequalities, Warszawa, 1965.
[26] Walter W., Differential- und Integral-Ungleichungen, Berlin-Göttingen-Heidelberg-New York, 1964.

NEUERE ERGEBNISSE DER BEWEISTHEORIE

KURT SCHÜTTE

Die Beweistheorie wurde von David Hilbert begründet, um die klassische Mathematik als widerspruchsfrei nachzuweisen. Das ursprüngliche Ziel des Hilbertschen Programms, Widerspruchsfreiheitsbeweise mit den elementarsten Methoden einer finiten Mathematik durchzuführen, erwies sich allerdings für die nichttrivialen Teile der Mathematik als grundsätzlich unerreichbar, seitdem die Unvollständigkeitssätze von Gödel bekannt wurden. Hiermit hat jedoch die Beweistheorie keineswegs ihre Bedeutung verloren, sondern sie ist vielmehr noch reichhaltiger in ihrer Aufgabenstellung geworden. Gerade mit der von Gödel stammenden Erkenntnis, daß sich höhere mathematische Theorien nicht mit den elementarsten Methoden als widerspruchsfrei begründen lassen, ergab sich für die Beweistheorie die Aufgabe, die einzelnen Teile der Mathematik hinsichtlich ihrer logischen Voraussetzungen und ihrer Beweiskraft gegeneinander abzugrenzen. Auch hierfür sind die Widerspruchsfreiheitsbeweise wichtig.

Es liegen jetzt bereits für verhältnismäßig starke Teile der Mathematik Widerspruchsfreiheitsbeweise vor, die zwar nicht auf streng finiten Methoden beruhen (wie es ja auch nach Gödel gar nicht möglich sein kann), aber doch noch als mehr oder weniger konstruktiv anzusehen sind. Solche Widerspruchsfreiheitsbeweise geben, wie insbesondere Kreisel gezeigt hat, konstruktive Interpretationen für nichtkonstruktive mathematische Theorien, und sie liefern konstruktive Verschärfungen für nichtkonstruktiv bewiesene mathematische Sätze. Hierzu ist es wichtig, sich genau darüber klar zu sein, auf welchen Methoden ein Widerspruchsfreiheitsbeweis beruht, um den Konstruktivitätsgrad dieses Beweises genau zu charakterisieren.

Ich möchte in diesem Vortrag versuchen, einen kurzen Überblick über beweistheoretische Abgrenzungen von einigen Teilen der Mathematik zu geben, wie sie sich aus den vorliegenden Widerspruchsfreiheitsbeweisen ergeben.

Lassen Sie mich von den Ergebnissen von Gentzen ausgehen, da auch die neueren Widerspruchsfreiheitsbeweise vielfach auf ähnlichen Methoden beruhen und zu ähnlichen Ergebnissen führen. Wie zuerst Gentzen erkannt hat, genügt es als methodisches Hilfsmittel zum Widerspruchsfreiheitsbeweis für die reine Zahlentheorie, neben streng finiten Mitteln eine Induktion zu verwenden, die stärker als die

gewöhnliche vollständige Induktion ist, nämlich eine Induktion über eine elementar definierte abzählbare Wohlordnung. Solche Wohlordnungen werden durch konstruktive Ordinalzahlen der 2. Zahlenklasse charakterisiert. Dabei handelt es sich bei der von Gentzen benutzten Induktion um die kleinste ε-Zahl ε_0. Gentzen bewies für ein formales System $\boldsymbol{Z}$ der reinen Zahlentheorie:

(1) Eine Wohlordnung R der natürlichen Zahlen vom Ordnungstyp ε_0 läßt sich in primitiv-rekursiver Weise so definieren, daß die Wohlordnung jedes echten Abschnittes von R im System $\boldsymbol{Z}$ beweisbar ist.

(2) Für keine Wohlordnung der natürlichen Zahlen vom Ordnungstyp $\geqslant \varepsilon_0$ ist die Wohlordnungseigenschaft im System $\boldsymbol{Z}$ beweisbar.

(3) Die Widerspruchsfreiheit des Systems $\boldsymbol{Z}$ ist durch Induktion über die unter (1) angegebene Wohlordnung R beweisbar.

Dieser Widerspruchsfreiheitsbeweis ist konstruktiv, sogar im strengsten Sinne prädikativ, da die hierzu benötigte Wohlordnung R in primitiv-rekursiver Weise definiert ist und die Wohlordnungseigenschaft von R zwar nicht mit streng finiten Mitteln, wohl aber auf einem einfachen konstruktiven Wege beweisbar ist. Man kann diesen Wohlordnungsbeweis ohne Benutzung des tertium non datur in einer konstruktiven Logik der 2. Stufe führen, in der nur in prädikativer Weise über die arithmetischen Grundprädikate quantifiziert wird.

Offensichtlich benutzt der Widerspruchsfreiheitsbeweis von Gentzen in optimaler Weise nur die schwächste Induktion, die zu einem solchen Beweis erforderlich ist. Die Induktion bis ε_0 liefert den Widerspruchsfreiheitsbeweis, und aus dem Unvollständigkeitssatz von Gödel geht hervor, daß hierzu keine schwächere Induktion ausreichen würde. Die reine Zahlentheorie wird somit beweistheoretisch durch die Ordinalzahl ε_0 charakterisiert. Ebenso werden stärkere mathematische Theorien durch höhere Ordinalzahlen charakterisiert, die sich zum Teil in ähnlicher Weise aus den betreffenden Widerspruchsfreiheitsbeweisen ergeben.

Bevor ich darauf eingehe, möchte ich jedoch noch weitere Methoden skizzieren, mit denen bisher konstruktive Widerspruchsfreiheitsbeweise durchgeführt wurden.

Die von Gentzen benutzte Zuordnung von Ordinalzahlen zu den formalen Beweisfiguren hat sich zwar als optimal erwiesen, ist aber durchaus nicht naheliegend, sondern verhältnismäßig kompliziert. Die beweistheoretischen Untersuchungen werden grundsätzlich einfacher und durchsichtiger, wenn man die zu untersuchenden endlichen Beweisfiguren in unendliche Bäume auflöst, indem man die formalisierte vollständige Induktion durch einen Schluß mit unendlich vielen Prämissen ersetzt, nämlich durch den Schluß von allen Formeln $A(z)$ für jede Ziffer z auf die Allformel $\forall x A(x)$. Diese sogenannte unendli-

che Induktion, die bereits von Hilbert vorgeschlagen wurde und die in ihrer nichtkonstruktiven Fassung als Carnap-Regel bekannt ist, wurde zuerst von Lorenzen zum Widerspruchsfreiheitsbeweis für die verzweigte Typenlogik herangezogen. Bei den Widerspruchsfreiheitsbeweisen, die mit Benutzung der unendlichen Induktion erfolgen, werden die endlichen Beweisfiguren in einer primitiv-rekursiven Weise in unendliche Formelbäume umgeformt, für die dann in konstruktiver Weise geeignete Rekursionen erklärt werden, mit denen sich die Widerspruchsfreiheit des zugrunde liegenden formalen Systems ergibt. Ordnet man den Formeln der unendlichen Bäume in der Weise Ordinalzahlen zu, daß jeder Konklusion die kleinste Ordinalzahl zugeordnet wird, die größer als alle Ordinalzahlen der Prämissen ist, so findet man für das formale System der reinen Zahlentheorie die kleinste ε-Zahl ε_0 als obere Grenze der benötigten Ordinalzahlen. Man kommt also auf dem Wege über die unendliche Induktion in natürlicher Weise zu denselben Ergebnissen, die Gentzen mit wesentlich komplizierteren Zuordnungen gefunden hat. Hiermit ist auch die Möglichkeit gegeben, Widerspruchsfreiheitsbeweise für stärkere Systeme als für die reine Zahlentheorie in verhältnismäßig durchsichtiger Weise zu gewinnen.

So wurde von Ackermann ein formales System der typenfreien Logik auf der Grundlage der unendlichen Induktion formuliert und als widerspruchsfrei nachgewiesen. In diesem typenfreien System ist das tertium non datur nicht allgemein herleitbar. Jedenfalls ist es aber für alle diejenigen Formeln herleitbar, für die sich eine prädikative Interpretation angeben läßt. Daher kann das gesamte System der verzweigten Typenlogik, das schon vorher von Lorenzen als widerspruchsfrei nachgewiesen wurde, in das typenfreie System von Ackermann eingebettet werden, womit sich ein zweiter Widerspruchsfreiheitsbeweis für die verzweigte Typenlogik ergibt. In dem typenfreien System sind die Begriffe, die zu den logischen Antinomien führen, formulierbar. Die Antinomien kommen jedoch in dem System nicht zustande, weil das tertium non datur für die betreffenden Formeln nicht herleitbar ist.

Um Teile der klassischen Analysis in dem typenfreien System von Ackermann darzustellen, müßte man solche Bereiche der Analysis abgrenzen, für die sich das tertium non datur in dem typenfreien System herleiten läßt. Wie weit dies über die Grenzen einer prädikativen Analysis hinaus möglich ist, ist bisher noch nicht genau untersucht worden. Ich möchte deshalb nicht näher auf die typenfreie Logik eingehen.

Es ist aber noch eine weitere Methode zu nennen, mit der konstruktive Widerspruchsfreiheitsbeweise durchgeführt wurden. Ich meine die Methode, die Gödel im Jahr 1958 zum Widerspruchsfreiheitsbeweis für die reine Zahlentheorie entwickelt hat. Die Überschreitung des

finiten Standpunktes erfolgt bei diesem Gödelschen Beweis nicht durch höhere Induktionen wie bei den bisher angegebenen Widerspruchsfreiheitsbeweisen, sondern durch Einführung von rekursiven Funktionalen höherer Typen. Es genügt also, sich auf gewisse abstrakte Objekte zu beziehen, die nicht mehr zum Gegenstand der finiten Betrachtung gehören, aber noch konstruktiv erklärt sind, um die reine Zahlentheorie als widerspruchsfrei einzusehen. Wie aus diesem Ergebnis hervorgeht, ist das Beweismittel der Funktionale höherer Typen nicht schwächer als die von Gentzen benutzte Induktion bis ε_0, aber es ist von grundsätzlich anderer Art und wohl auch einsichtiger als das Induktionsprinzip.

Zusammenfassend läßt sich sagen, daß bisher im wesentlichen drei verschiedene Methoden für konstruktive Widerspruchsfreiheitsbeweise verwendet wurden:

1. Die von Gentzen stammende Methode der Zuordnung von Ordinalzahlen zu endlichen Beweisfiguren mit einer konstruktiven Induktion über die betreffenden Ordinalzahlen.

2. Die Methode der unendlichen Induktion mit metamathematischen Induktionen über konstruktiv definierte unendliche Formelbäume.

3. Die Gödelsche Methode der Einführung von höheren Funktionalen.

Alle drei Methoden wurden nicht nur auf die reine Zahlentheorie, sondern auch auf wesentlich stärkere formale Systeme angewendet. Ich werde bei Nennung der neueren Widerspruchsfreiheitsbeweise im einzelnen angeben, welche der drei genannten Methoden dabei hauptsächlich verwendet wurden.

Eine Abgrenzung der prädikativen Mathematik ergab sich im wesentlichen mit der an zweiter Stelle genannten Methode der unendlichen Induktion. Feferman und ich fanden gleichzeitig und unabhängig voneinander eine Ordinalzahl, die für die prädikative Mathematik ebenso charakteristisch ist wie ε_0 für die reine Zahlentheorie. Diese Ordinalzahl, die ich als kleinste "stark kritische" Ordinalzahl $\varkappa_0$ bezeichne, läßt sich folgendermaßen beschreiben:

(1) Es sei $\varphi_0(\alpha) = \omega^\alpha$.

(2) Für $\nu \neq 0$ sei φ_ν die Ordnungsfunktion der Ordinalzahlen ξ mit der Eigenschaft $\varphi_\mu(\xi) = \xi$ für alle $\mu < \nu$. Das heißt, φ_ν sei eine monoton wachsende Funktion mit dem Wertebereich

$$\{\xi\colon\ \forall \mu < \nu\ (\varphi_\mu(\xi) = \xi)\}.$$

(3) Eine Ordinalzahl $\varkappa$ heiße "stark kritisch", wenn $\varphi_\varkappa(0) = \varkappa$ ist. Die Ordinalzahl $\varkappa_0$, die für die prädikative Mathematik charakteristisch ist, ist die kleinste stark kritische Ordinalzahl.

Die entsprechenden Eigenschaften, die ε_0 für die reine Zahlentheorie besitzt, lassen sich für $\varkappa_0$ bezüglich der prädikativen Mathematik nachweisen, nämlich:

(1) Eine Wohlordnung W der natürlichen Zahlen vom Ordnungstyp $\varkappa_0$ läßt sich in primitiv-rekursiver Weise so definieren, daß die Wohlordnung jedes echten Abschnittes von W mit prädikativen Mitteln beweisbar ist.

(2) Für keine Wohlordnung vom Ordnungstyp $\geqslant \varkappa_0$ ist die Wohlordnungseigenschaft prädikativ nachweisbar.

Feferman hat weiterhin gezeigt, daß sich diese Prädikativität ebenso wie durch die verzweigte Typenlogik auch durch eine ungeschichtete Folge formaler Systeme der Arithmetik 2. Stufe charakterisieren läßt, bei der die unendliche Induktion in prädikativer Weise zugelassen ist. Dieser Formulierung liegt nach dem Vorschlag von Kreisel eine hyperarithmetische Komprehensionsregel zugrunde, die die Anwendung des Komprehensionsaxioms in folgender Weise einschränkt: Angenommen, es sei in der Arithmetik 2. Stufe folgende Äquivalenz beweisbar:

$$\forall x\,[\forall Y A(x, Y) \leftrightarrow \exists Z B(x, Z)].$$

Dabei sei x eine Zahlenvariable, Y und Z seien Prädikatenvariablen, $A(x, Y)$ und $B(x, Z)$ seien arithmetische Formeln. Unter dieser Voraussetzung darf das Komprehensionsaxiom angewendet werden:

$$\exists Z \forall x\,[Z(x) \leftrightarrow \forall Y A(x, Y)].$$

Schränkt man das Komprehensionsaxiom in der Arithmetik 2. Stufe in dieser Weise ein und nimmt man die unendliche Induktion nur in prädikativer Weise hinzu, so erhält man eine prädikative Analysis als echten Teil der klassischen Analysis. Das heißt: Die Sprache dieser prädikativen Analysis ist dieselbe wie die der klassischen Analysis, und jeder Satz, der in der prädikativen Analysis beweisbar ist, gilt auch im Sinne der klassischen Analysis. Die kleinste stark kritische Ordinalzahl $\varkappa_0$ hat für diese prädikative Analysis dieselbe Bedeutung als Grenzzahl wie für die verzweigte Typenlogik.

Die Ordinalzahlen ε_0 und $\varkappa_0$ sind aber durchaus nicht die größten Ordinalzahlen, die sich in beweistheoretischen Untersuchungen als charakteristische Grenzzahlen für gewisse Bereiche der Mathematik ergeben haben.

Kreisel formulierte ein Teilsystem der intuitionistischen Analysis mit Bar-Induktion vom Typ 0 unter Zugrundelegung einer formalen Fixierung freier Wahlfolgen. Für dieses formale System gab W. Howard eine Interpretation durch Funktionale höherer Typen in Verallgemeinerung der von Gödel entwickelten Methode zum Widerspruchsfreiheitsbeweis der reinen Zahlentheorie. Hiermit fand Howard eine Ordinalzahl, die das formale System von Kreisel ebenso abgrenzt, wie die prädikative Analysis durch $\varkappa_0$ abgegrenzt wird. Die von Howard angegebene Ordinalzahl ist wesentlich größer als $\varkappa_0$. Benutzt man nach H. Bachmann Ordinalzahlen der 3. Zahlenklasse zur Kennzeichnung der Kritischkeitsgrade von Ordinalzahlen, so ergeben sich die "stark kritischen" Ordinalzahlen als die Ω-kritischen Zahlen, wobei Ω die Anfangszahl der 3. Zahlenklasse ist. Es ist also $\varkappa_0 = \varphi_\Omega(0)$. Die von Howard gefundene Grenzzahl des Teiles der intuitionistischen Analysis ist die kleinste Ordinalzahl $\varphi_{\varepsilon_{\Omega+1}}(0)$, deren Kritischkeitsgrad gleich der ersten ε-Zahl oberhalb Ω ist. Natürlich läßt sich diese Ordinalzahl nicht mehr prädikativ verstehen, wohl aber noch in einer gewissen konstruktiven Weise, auf die ich noch zu sprechen kommen werde.

Noch wesentlich höhere Ordinalzahlen werden in den Widerspruchsfreiheitsbeweisen gebraucht, die Takeuti für Teilsysteme der klassischen Analysis gegeben hat. Takeuti stellte die Fundamentalvermutung auf, daß der Gentzensche Hauptsatz über die Schnitt-Eliminierbarkeit auch in einem geeigneten System der einfachen Typenlogik gilt, und er bewies, daß aus dieser Fundamentalvermutung die Widerspruchsfreiheit der klassischen Analysis folgt. Es war lange Zeit unbekannt, ob auch umgekehrt die Fundamentalvermutung mit den Mitteln der klassischen Analysis beweisbar ist. Es gelang mir lediglich, ein semantisches Äquivalent für die syntaktische Aussage der Fundamentalvermutung abzuleiten. Dieses semantische Äquivalent bezieht sich auf partielle Wertungen der einfachen Typenlogik. Unter einer partiellen Wertung V verstehe ich eine Zuordnung von Wahrheitswerten w oder f zu Formeln der einfachen Typenlogik, bei der nicht jede Formel einen Wahrheitswert zu erhalten braucht. Eine partielle Wertung V soll folgenden Bedingungen genügen:

$$V \neg A = w \Leftrightarrow VA = f.$$
$$V \neg A = f \Leftrightarrow VA = w.$$

$V(A \vee B) = w \Leftrightarrow VA = w$ oder $VB = w$ (wobei nicht beide Formeln A und B einen Wahrheitswert bezüglich V zu besitzen brauchen).

$$V(A \vee B) = f \Leftrightarrow VA = f \text{ und } VB = f.$$

Für gebundene Variablen x vom Typ τ soll gelten:

$V\forall x^\tau A(x^\tau) = w \Leftrightarrow VA(t^\tau) = w$ für jeden Term t^τ vom Typ τ.
$V\forall x^\tau A(x^\tau) = f \Leftrightarrow VA(t^\tau) = f$ für mindestens einen Term t^τ vom Typ τ.

Entsprechende Bedingungen sollen für $V(A \wedge B)$ und $V\exists x^\tau A(x^\tau)$ gelten.

Eine solche Wertung V heiße *total*, wenn sie jeder Formel einen Wahrheitswert zuordnet. Es läßt sich dann beweisen:

Die Fundamentalvermutung von Takeuti trifft genau dann zu, wenn sich jede partielle Wertung zu einer totalen Wertung erweitern läßt.

Hiermit haben wir ein semantisches Äquivalent für die syntaktische Aussage der Fundamentalvermutung. Mit Hilfe dieses semantischen Äquivalents konnte W. Tait neuerdings die Fundamentalvermutung für die Prädikatenlogik 2. Stufe mit nichtkonstruktiven Methoden beweisen.

Für gewisse Teile der einfachen Typenlogik hat Takeuti konstruktive Beweise seiner Fundamentalvermutung gegeben, um auf diesem Wege zu konstruktiven Widerspruchsfreiheitsbeweisen für Teilgebiete der klassischen Analysis zu gelangen. Hiermit hat Takeuti die stärksten Widerspruchsfreiheitsbeweise geschaffen, die bisher überhaupt für Teile der Analysis gefunden wurden. Sie umfassen unter anderem die Arithmetik 2. Stufe mit dem Π^1_1-Komprehensionsaxiom, das heißt mit dem Komprehensionsaxiom

$$\forall x_1 \ldots \forall x_m \forall Y_1 \ldots \forall Y_n \exists Z \forall x [Z(x) \leftrightarrow A(x, x_1, \ldots, x_m, Y_1, \ldots, Y_n)]$$

für Zahlenvariablen $x, x_1, \ldots, x_m$ und Prädikatenvariablen $Y_1, \ldots, Y_n, Z$ unter der Voraussetzung, daß in der Formel $A(x, x_1, \ldots, x_m, Y_1, \ldots, Y_n)$ in pränexer Normalform nur Allquantoren für Prädikatenvariablen auftreten. Dieses Komprehensionsaxiom ist außerordentlich stark. Es gestattet, weite Teile der klassischen Analysis zu entwickeln.

Die Methoden von Takeuti schließen sich an Gentzen an, das heißt, sie beruhen auf Zuordnungen von Ordinalzahlen zu endlichen Beweisfiguren. Entsprechend der Stärke der behandelten formalen Systeme werden für die Widerspruchsfreiheitsbeweise Induktionen über außerordentlich große Abschnitte der 2. Zahlenklasse gebraucht. Takeuti verwendet hierzu die von ihm entwickelten "ordinal diagrams", die sehr umfangreiche Abschnitte der 2. Zahlenklasse in primitiv-rekursiver Weise definieren, aber nur mit verhältnismäßig starken Mitteln als wohlgeordnet nachweisbar sind. Es ist natürlich unmöglich, diese Wohlordnungsbeweise prädikativ zu führen. Vielmehr braucht man hierzu den Begriff der "Erreichbarkeit", wie er zuerst von Ackermann verwendet wurde, in einer imprädikativen Weise. Das imprädikative Element, das bei den Wohlordnungsbeweisen benutzt wird, läßt sich nach Kreisel durch das Prinzip der "verallgemeinerten induktiven Definition" kennzeichnen.

Dieses Definitionsprinzip läßt sich formal folgendermaßen fixieren: Angenommen, für eine Aussage $A(x, Y)$ gelte

$$Y \subseteq Z \rightarrow \forall x [A(x, Y) \rightarrow A(x, Z)],$$

wobei x eine Zahlenvariable und Y, Z Mengenvariablen sind. Unter dieser Voraussetzung darf nach dem verallgemeinerten induktiven Definitionsprinzip eine Menge M_A definiert werden mit den Eigenschaften

(1) $\forall x [A(x, M_A) \rightarrow x \in M_A]$,

(2) $\forall x [A(x, Y) \rightarrow x \in Y] \rightarrow M_A \subseteq Y$.

Das heißt, M_A soll die kleinste Menge mit der Eigenschaft (1) sein. Für die Wohlordnungsbeweise der ordinal diagrams braucht man außer finiten Methoden nur dieses verallgemeinerte induktive Definitionsprinzip. Man muß es aber in imprädikativer Weise anwenden. Das heißt, man muß die Eigenschaft (2) auch auf solche Mengen Y anwenden, deren Definition von der Menge M_A abhängt. Die Widerspruchsfreiheitsbeweise von Takeuti, die auf diesem verallgemeinerten induktiven Definitionsprinzip beruhen, sind also nicht mehr prädikativ, aber doch noch in einem gewissen Sinne konstruktiv. Mehr läßt sich aber auch nicht erreichen, denn die Ergebnisse von Takeuti sind für seine starken formalen Systeme ebenso optimal, wie es die Ergebnisse von Gentzen für die reine Zahlentheorie sind. Das heißt, die für die Widerspruchsfreiheitsbeweise gebrauchten Induktionen lassen sich durch keine schwächeren Induktionen ersetzen und können daher grundsätzlich nicht prädikativ begründet werden.

Die Ordinalzahlen, die sich aus den Widerspruchsfreiheitsbeweisen von Takeuti als charakteristisch für die von ihm behandelten Teile der klassischen Analysis ergeben, sind durch die ordinal diagrams fixiert. Näher können sie nicht angegeben werden, da sie sich einer einfachen Beschreibung entziehen.

Zusammenfassung. Für einzelne Teilgebiete der Mathematik haben sich folgende Ordinalzahlen als charakteristisch erwiesen:

1. Reine Zahlentheorie: $\varepsilon_0 = \varphi_1(0)$ (G. Gentzen)
2. Prädikative Mathematik: $\varkappa_0 = \varphi_\Omega(0)$ (S. Feferman, K. Schütte)
3. Intuitionistische Arithmetik 2. Stufe mit Bar-Induktion vom Typ 0: $\varphi_{\varepsilon_{\Omega+1}}(0)$ (W. Howard)
4. Klassische Arithmetik 2. Stufe mit Π^1_1-Komprehensionsaxiom: Eine Ordinalzahl von G. Takeuti.

Entsprechend gehören höhere Ordinalzahlen, die aber noch unbekannt sind, zu folgenden Systemen:

5. Klassische Arithmethik 2. Stufe mit dem Prinzip der verallgemeinerten induktiven Definition.
6. Formales System der klassischen Analysis, für das C. Spector einen nichtkonstruktiven Widerspruchsfreiheitsbeweis geführt hat.

Alle diese Ordinalzahlen liegen unterhalb Kleene's kleinster Ordinalzahl ω_1, die nicht mehr rekursiv definierbar ist.

Mathematisches Institut
der Universität München,
Bundesrepublik Deutschland

LITERATUR

[1] Ackermann W., Widerspruchsfreier Aufbau einer typenfreien Logik, I, *Math. Z.*, 55 (1952), 364-384; II, *Math. Z.*, 57 (1953), 155-166.
[2] Bachmann H., Die Normalfunktionen und das Problem der ausgezeichneten Folgen von Ordnungszahlen, *Vierteljschr. Naturforsch. Ges. Zürich*, 95 (1950), 115-147.
[3] Feferman S., Constructively provable well-orderings, *Notices Amer. Math. Soc.*, 8 (1961), 495.
[4] Feferman S., Provable well-orderings of and relations between predicative and ramified analysis, *Notices Amer. Math. Soc.*, 9 (1962), 323.
[5] Feferman S., Systems of predicative analysis, *Journal of symbolic logic*, 29 (1964), 1-30.
[6] Gentzen G., Die Widerspruchsfreiheit der reinen Zahlentheorie, *Math. Ann.*, 112 (1936), 493-565.
[7] Gentzen G., Neue Fassung des Widerspruchsfreiheitsbeweises für die reine Zahlentheorie, *Forschungen zur Logik und zur Grundlegung der exakten Wissenschaften*, Neue Folge, 4 (1938), 19-44.
[8] Gentzen G., Beweisbarkeit und Unbeweisbarkeit von Anfangsfallen der transfiniten Induktion in der reinen Zahlentheorie, *Math. Ann.*, 119 (1943), 140-161.
[9] Gödel K., Über eine bisher noch nicht benützte Erweiterung des finiten Standpunktes, *Dialectica*, 12 (1958), 280-287.
[10] Howard W., Transfinite induction and transfinite recursion, Reports of the seminar on foundations of analysis, vol. II, Stanford University, Summer 1963.
[11] Kleene S. C., On notation for ordinal numbers, *Journal of symbolic logic*, 3 (1938), 150-155.
[12] Kreisel G., Reports of the seminar on foundations of analysis, Stanford University, Summer 1963.
[13] Lorenzen P., Algebraische und logistische Untersuchungen über freie Verbände, *Journal of symbolic logic*, 16 (1951), 81-106.
[14] Schütte K., Beweistheorie, Berlin-Göttingen-Heidelberg, 1960.
[15] Schütte K., Syntactical and semantical properties of simple type theory, *Journal of symbolic logic*, 25 (1960), 305-326.
[16] Schutte K., Eine Grenze für die Beweisbarkeit der transfiniten Induktion in der verzweigten Typenlogik, *Archiv f. math. Logik u. Grundlagenforschung*, 7, 45-60.
[17] Schütte K., Predicative well-orderings. Formal systems and recursive functions, Proceedings of the 8th logic colloquium Oxford 1963, Amsterdam, 1965, 280-302.
[18] Spector C., Provably recursive functionals of analysis: A consistency proof of analysis by an extension of principles formulated in current intuitionistic mathematics, Proceedings of symposia in pure mathematics, vol.V, 1962, 1-27.
[19] Tait W., A non-constructive proof of Gentzen's Hauptsatz for second order predicate logic, to appear.
[20] Takeuti G., On a generalized logic calculus, *Jap. J. Math.*, 23 (1953), 39-64.

DIFFERENTIABLE DYNAMICAL SYSTEMS[1])

S. SMALE

Although motivated ultimately by ordinary differential equations and continuous flows, we concentrate mainly on studying the discrete dynamical system generated by a diffeomorphism $T: M \to M$ of a smooth manifold. One way of expressing our framework is saying that we aim to give a non-linear global spectral theory (finite dimensional) for T. In fact, we show that under fairly general conditions, M decomposes into a finite number of invariant indecomposable subspaces. These subspaces form a lattice under a boundary relation and generalize the cell decomposition of a generic gradient flow. The further study of these spaces, most remarkably, forces the introduction of group theory and compact homogeneous spaces.

Dept. of Mathematics,
University of California,
Berkeley, USA

---

[1]) This is a subset of survey article that will appear in the American Mathematical Society Bulletin.

SOME RECENT DEVELOPMENTS IN MATHEMATICAL STATISTICS

CHARLES M. STEIN

1. Introduction

In this report, I shall largely restrict my attention to statistical decision theory and its applications to classical statistics. I shall also discuss the theory of similar tests, partly because of its intrinsic interest, mathematical and at least potentially practical, but also because some of the methods that have been used in it seem likely to be applicable to the more decision-theoretic study of problems of testing hypotheses. Large sample theory will also be considered because its ideas are close to those of statistical decision theory, and it permits us to get more explicit approximate results than we can hope for in general problems without large samples.

Thus, somewhat unfortunately, I shall cover only a small part of the mathematical apparatus relevant to statistical practice. Probabilistic models for physical, biological and social phenomena will not be discussed although in serious statistical analysis of real data they tend to dominate the more properly mathematical statistical elements. See, for example, the applied papers in the Fourth and Fifth Berkeley Symposia. Nonparametric statistics and time series analysis will also be omitted. Both have been extensively studied in recent years, and like the rest of statistics, present many interesting problems and a few interesting results. However, their methods are so different from those of the branches considered here that I have been unable to find a way to include them in a reasonably short time, and so have omitted them. The design of experiments, both combinatorial and decision-theoretic, will not be considered either. The decision-theoretic aspects have been considered most recently by Karlin (1966) and Kiefer and Wolfowitz (1965).

Statistical decision theory was introduced and extensively studied by Abraham Wald (1950), inspired by the ideas of Neyman and Pearson (1933, 1938) on testing hypotheses, and by the theory of estimation in which decision-theoretic ideas can be traced back at least to Laplace and Gauss.

2. The theory of games in statistical decision theory

Wald, in his most extensive exposition of statistical decision theory (Wald, 1950) based his development on the more general framework of von Neumann's (1944) theory of games. It will be convenient to

follow the same path here, because it will enable me to indicate two of the basic tools, the minimax theorem and the notion of admissibility without the elaborate notation that is unavoidable in a precise formulation of the statistical applications. For brevity of presentation, I shall first discuss admissibility, reversing the historical and logical order.

Let K be a bounded real valued function on the Cartesian product $\mathcal{X} \times \mathcal{Y}$ of two sets. A point $y_0 \in \mathcal{Y}$ is said to be admissible if it is minimal in the partial order determined by the functions $K(x, \cdot)$. In other words, y_0 is admissible if, for every $x_0 \in \mathcal{X}$, the function $y \to K(x_0, y)$ is minimized, subject to $K(x, y) \leqslant K(x, y_0)$ for all x, by $y = y_0$. One way to prove admissibility of a point y_0 is by a Lagrange multiplier argument. Suppose K is pseudo-convex in its second argument in the sense that for every $y_1, y_2 \in \mathcal{Y}$ and $\alpha \in (0, 1)$ there exists y such that, for all $x \in \mathcal{X}$

$$K(x, y) \leqslant \alpha K(x, y_1) + (1-\alpha) K(x, y_2).$$

Suppose also that the space $\mathcal{Y}$ is compact in the smallest topology in which all the $K(x, \cdot)$ are lower semicontinuous. Then, in order that y_0 minimize $K(x_0, y)$ subject to $K(x, y) \leqslant K(x, y_0)$ for all x it is necessary and sufficient that for every $\varepsilon > 0$, there exist a finite positive measure ξ concentrated on a finite subset of $\mathcal{X}$ such that

$$\xi(x_0) = 1$$

and

$$K^*(\xi, y_0) \leqslant \inf_y K^*(\xi, y) + \varepsilon$$

where

$$K^*(\xi, y) = \sum_x K(x, y)\, \xi(x).$$

In statistical applications it is convenient to divide through by the total measure. Then ξ is required to be a probability measure, the condition $\xi(x_0) = 1$ is replaced by $\xi(x_0) > 0$, and the other condition becomes

$$K^*(\xi, y_0) \leqslant \inf_y K^*(\xi, y) + \varepsilon \xi(x_0).$$

We call a y_0 satisfying this inequality for particular ξ and ε, an $\varepsilon\xi(x_0)$-Bayes solution with respect to ξ.

This is a sharpening of results of Wald. The sufficiency of this condition, which, as far as I know, is the only part that has been

applied in statistics, is essentially a formalization of a method first used in a very special case by Blyth (1951). It is easy to prove and does not use the compactness or the boundedness above of K. The necessity of the condition is not completely trivial and requires an application of the minimax theorem to the function $(x, y) \to K(x, y) - K(x, y_0)$. In the present context, the minimax theorem (due in its original form to von Neumann but generalized by others to results much better than any mentioned here) asserts that, subject to the same conditions of boundedness, pseudo-convexity, and compactness

$$\sup_{\xi} \inf_{y} K^*(\xi, y) = \inf_{y} \sup_{\xi} K^*(\xi, y),$$

where ξ ranges over the set of probability measures concentrated on finite subsets of $\mathcal{X}$, and the infima are attained. In many applications the function K is not bounded above and, especially in the necessity part of the condition for admissibility, this seems to lead to complications that have not been fully explored.

As far as I know, no results very close to the necessary and sufficient condition for admissibility (or for conditional minimization) has appeared in the general mathematical literature, although it is certainly close to the Hahn-Banach theorem. Roughly speaking, it seems to be useful in analysis in the following way. It frequently happens that a property Π of a mathematical object is expressible either by definition or by a simple argument, at the inadmissibility (or failure of conditional minimization) of a certain y_0 with respect to a certain K. The condition of the first-mentioned theorem then transforms the negation of Π, into a condition that does not explicitly use negation. For example, a well-known condition for a connected stationary Markov process (with discrete time and denumerable state space) to be transient is thus transformed into a condition for the process to be recurrent. Such a process is reccurent if and only if for some state s_0 and every $\varepsilon > 0$ there exists a probability measure ξ concentrated on a finite set such that $\xi_{s_0} > 0$ and $\sum_j |\xi_j - \sum_i \xi_i p_{ij}| < \varepsilon \xi_{s_0}$ where the p_{ij} are the transition probabilities. In particular, the discrete case of a result of Chung and Fuchs on recurrence of sums of independently identically distributed random variables can be proved in this way by taking ξ to have an isosceles triangle with small slope as its graph, and it becomes clear that it is only required that the Markov process is qualitatively something like a sum of independent identically distributed random variables with nearly a first moment nearly equal to 0.

In statistical problems, the sufficient condition for admissibility is usually applied in a slightly different form because probability distributions concentrated on finite sets do not seem easy to handle. Let us look at an example.

3. Estimation with quadratic loss

Let p be a probability density on the real line such that

$$\int z p(z)\, dz = 0, \quad \int z^2 p(z)\, dz < \infty. \tag{4.1}$$

It is not difficult to prove, by a slight variant of the above method, that there does not exist a measurable function φ such that $\varphi(z) \neq z$ on a set of positive measure and

$$\int [\varphi(z) - \theta]^2 p(z-\theta)\, dz \leqslant \int [z-\theta]^2 p(z-\theta)\, dz, \tag{4.2}$$

for all real θ. In statistical language, if we observe a random variable Z distributed according to the density $p(z-\theta)$ with θ unknown, Z is an admissible estimate of θ with squared error as loss. The left-hand side of the above inequality plays the role of the function $K(\theta, \varphi)$ in the earlier general discussion. The proof is accomplished by showing that φ_0 defined by $\varphi_0(z) = z$ is for constant c and sufficiently large σ, $\frac{c}{\sigma^2}$-Bayes with respect to the prior density π_σ defined by

$$\pi_\sigma(\theta) = \frac{1}{\pi\sigma}, \quad \frac{1}{1+\frac{\theta^2}{\sigma^2}}, \tag{4.3}$$

that is, that φ_0 comes within $\frac{c}{\sigma^2}$ of minimizing

$$\int \pi(\theta) K(\theta, \varphi)\, d\theta, \tag{4.4}$$

with $\pi = \pi_\sigma$.

This seems like a lot of work to prove an intuitively obvious result. However, it turns out that the result does not generalize to the extent that at first seemed likely. If we consider the case where Z and θ are p-dimensional coordinate vectors and interpret the squared errors as sums of squares of errors, the result is no longer true for $p \geqslant 3$. The difficulty in the proof is that as we vary the scale σ in the prior density π_σ we can still prove that Z is $\frac{c}{\sigma^2}$-Bayes for some c, but need it to be $\frac{c}{\sigma^p}$-Bayes for all c. For $p = 2$, the proof can be modified to go through under slightly stronger hypothesis on p, but the result fails for $p \geqslant 3$.

It is trivial that the φ that minimizes (4.4) is given by

$$\varphi_\pi(x) = \frac{\int \theta p(x-\theta)\, \pi(\theta)\, d\theta}{\int p(x-\theta)\, \pi(\theta)\, d\theta},$$

the expectation of θ under the posterior distribution determined by π. If π is a probability density, it is easy to see that this φ_π is admissible. However, the formula for φ_π makes sense in many cases where π is an improper prior density, that is, where π is not integrable. In particular $\pi(z) \equiv 1$ yields $\varphi_\pi(z) \equiv z$. It then turns out that in the 1-dimensional case if p has enough moments and π is sufficiently regular (e.g., monotonic at ∞), φ_π is admissible if (and presumably only if) $\int_a^\infty \frac{d\theta}{(\pi\theta)} = \infty$ and $\int_{-\infty}^a \frac{d\theta}{\pi(\theta)} = \infty$ for all a.

4. General statistical decision theory

In describing the general statistical decision problem, which is not really a problem but rather a framework within which various problems related to statistical practice can be discussed, I shall postpone the treatment of sequential problems and the design of experiments. I shall also omit any mention of the σ-algebras involved since measure-theoretic subtleties seem not to be important in most of the problems considered here. Of course, it is not claimed that the statistical decision problem, as described here or elsewhere, is a realistic description of statistical practice, nor of the way statistics ought to be practiced. However, intuitive understanding of the basic notions is improved by pretending that we are talking about a real statistical problem.

We consider two sets $\mathcal{Z}$ and $\mathcal{H}$ and a function P on $\mathcal{H}$ to the set of probability measures in $\mathcal{Z}$. The triple $(\mathcal{Z}, \mathcal{H}, P)$ is called an experiment. Its interpretation is that we are going to observe a random point Z of the sample space $\mathcal{Z}$ distributed according to P_θ where θ is an unknown element of the parameter space $\mathcal{H}$. If I have time I shall talk later about the study of experiments without further structure, a subject which has many contacts with other branches of mathematics, and is of basic theoretical importance but limited immediate practical importance in statistics. To complete the description of a statistical decision problem, we add an action space $\mathcal{A}$ and a loss function L, a non-negative valued function on $\mathcal{H} \times \mathcal{A} \times \mathcal{Z}$. The intuitive interpretation is that after observing Z the statistician is required to choose an action $a \in \mathcal{A}$ and suffer the loss $L(\theta, a, Z)$. In order to be prepared to do this he chooses a decision function, that is, a function φ on $\mathcal{Z}$ to $\mathcal{A}$ and then, when he observes Z, takes the action $\varphi(Z)$. The statistician's aim is, roughly speaking, to choose φ so as to keep the expected loss

$$K(\theta, \varphi) = E_\theta L(\theta, \varphi(Z), Z) \tag{4.1}$$

small. Since he does not know θ, this aim is vague. More generally, he may choose a randomized decision function, that is, a function Δ on $\mathcal{Z}$ to the set of probability measures in $\mathcal{A}$ and then, when he observes Z, choose a random action in $\mathcal{A}$ distributed according to $\Delta(Z)$. His expected loss is then

$$K(\theta, \Delta) = E_\theta \int L(\theta, a, Z)\, \Delta(z)(da). \tag{4.2}$$

An important role in statistical decision theory is played by the notion of posterior distribution. A probability measure Π in $\mathcal{H}$, called a prior distribution, induces a joint distribution of θ and Z in which θ is distributed according to Π, and Z, given θ, according to P_θ. The conditional distribution of θ given $\mathcal{H}$, denoted here by Π_Z^* is called the posterior distribution of θ. If, as we shall assume

$$dP_\theta(z) = p_\theta(z)\, d\mu(z), \tag{4.3}$$

then

$$\Pi_Z^*(A) = \frac{\int_A p_\theta(Z)\, \Pi(d\theta)}{\int_{\mathcal{H}} p_\theta(Z)\, \Pi(d\theta)}. \tag{4.4}$$

The right-hand side of this formula is meaningful as a probability measure in many cases in which Π is an infinite measure. In such a case we call Π_Z^* the formal posterior distribution corresponding to the prior distribution Π. Confining our attention, for brevity of presentation, to non-randomized decision procedures, we define a Bayes solution with respect to the prior distribution Π to be a φ_Π that minimizes

$$\begin{aligned} E^\Pi \rho(\theta, \varphi) &= E^\Pi E_\theta L(\theta, \varphi(Z), Z) = \\ &= E^\Pi E^\Pi [L(\theta, \varphi(Z), Z) \mid Z] = \\ &= E^\Pi \int L(\theta, \varphi(Z), Z)\, d\Pi_Z^*(\theta). \end{aligned} \tag{4.5}$$

Thus, $\varphi_\Pi(z)$ is to be chosen such that minimizes

$$\int L(\theta, \varphi(z), z)\, d\Pi_Z^*(\theta). \tag{4.6}$$

If the risk, that is, expected loss

$$\int \int L(\theta, \varphi_\Pi(z), z)\, dP_\theta(z)\, d\Pi(\theta) < \infty, \tag{4.7}$$

then φ_Π is almost admissible (with respect to Π) in the sense that if φ is any decision function for which

$$\rho(\theta, \varphi) \leqslant \rho(\theta, \varphi_\Pi), \tag{4.8}$$

then strict inequality holds in (4.8) only on a set of Π measure 0.

More generally, as suggested dy the necessary and sufficient condition for admissibility, if, instead of (4.7) it is true that for each set C in a countable family of sets covering $\mathcal{H}$ and every $\varepsilon>0$ there exists a function r such that

(4.9) $$r(\theta)=1 \text{ for } \theta\in C$$

(4.10) $$\int\left[\int L(\theta, \varphi_\Pi(z), z)\, dP_\theta(z)\right] r(\theta)\, d\Pi(\theta) < \infty$$

(4.11) $$\int\int [L(\theta, \varphi_\Pi(z), z) - L(\theta, \varphi_{\Pi_r}(z), z)]\, dP_\theta(z)\, r(\theta)\, d\Pi(\theta) < \varepsilon$$

where

(4.12) $$d\Pi_r = r\, d\Pi,$$

then φ_Π is almost admissible. By an argument that has not been made precise, under suitable regularity conditions if $\mathcal{H}$ is a differentiable manifold and $d\Pi=\pi\, d\theta$ where $d\theta$ is the product of the coordinate differentials, the condition for admissibility becomes the following, related to an L_2 version of a general exterior Dirichlet problem: There should exist for every compact set C and every $\varepsilon>0$ a continuously differentiable function q such that

$$q(\theta)=1 \text{ for } \theta\in C$$

$$\int q^2(\theta)\,\pi(\theta)\, d\theta < \infty$$

$$\int g^{ij}(\theta)\,\frac{\partial q(\theta)}{\partial\theta^i}\,\frac{\partial q(\theta)}{\partial\theta^j}\,\pi(\theta)\, d\theta < \varepsilon.$$

Here g^{ij} is a positive definite contravariant tensor field, determined by the problem in a complicated way, and π transforms as a scalar density. The one-dimensional case is trivial, leading to the conditions

(i) $$\int_a^\infty \frac{d\theta}{g(\theta)\,\pi(\theta)} = \infty \text{ if } \int_a^\infty \pi(\theta)\, d\theta = \infty$$

and

(ii) $$\int_{-\infty}^a \frac{d\theta}{g(\theta)\,\pi(\theta)} = \infty \text{ if } \int_{-\infty}^a \pi(\theta)\, d\theta = \infty.$$

5. Problems invariant under a group of transformations

In statistics, as in many branches of mathematics, it is often possible to make considerable progress toward the solution of a special problem by observing that it is invariant under a certain group of

transformations. The experiment $(\mathcal{Z}, \mathcal{H}, P)$ is said to be invariant under the 1-1 transformation $g: \mathcal{X} \xrightarrow{\text{onto}} \mathcal{X}$ if there exists $\bar{g}: \mathcal{H}\to\mathcal{H}$ such that

$$P_{\bar{g}\theta} = P_\theta \circ g^{-1}.$$

A statistical decision problem, where we also have an action space $\mathcal{A}$ and a loss function L on $\mathcal{H}\times\mathcal{A}\times\mathcal{Z}$ is said to be invariant under g if there also exists $\hat{g}: \mathcal{A}\to\mathcal{A}$ such that

$$L(\bar{g}\theta, \hat{g}a, gz) = L(\theta, a, z).$$

The pure decision function $\varphi: \mathcal{Z}\to\mathcal{A}$ is invariant if $\varphi(gz) = \hat{g}[\varphi(z)]$ and the randomized decision function Δ invariant if $\Delta(gz, \hat{g}A) = \Delta(z, A)$. At least three fairly distinct problems arise in connection with these notions.

(i) How do we obtain good invariant procedures?
(ii) What justification is there for restricting our attention to invariant procedures?
(iii) Given a statistical problem, what are the transformations leaving it invariant?

Suppose we have a locally compact group $\mathcal{G}$ of transformations leaving a problem invariant, with the topological and measure theoretic aspects hanging together properly. In talking about Problem (i), I shall confine my attention to problems where the group $\bar{\mathcal{G}}$ of all $\bar{g}$: $\mathcal{H}\to\mathcal{H}$ corresponding to $g\in\mathcal{G}$ operates transitively on $\mathcal{H}$ and the subgroup leaving a particular θ invariant is compact. In this case, at least under regularity conditions that are commonly satisfied, the best invariant procedure is a formal Bayes procedure with respect to the prior measure Π in $\mathcal{H}$ induced by the right invariant measure in $\mathcal{G}$. Since Π enters only through the formal posterior distribution Π_z which is homogeneous of degree 0, this Π is invariant in the only sense relevant here, that is, relatively invariant.

Now let us look at the question of the justification for restricting our attention to invariant procedures. First we ask: If an invariant procedure is minimax among invariant procedures, is it necessarily minimax among all procedures? If $\mathcal{G}$ is finite, the affirmative answer is easily obtained by averaging over $\mathcal{G}$. Of course, we are talking about invariant randomized procedures, not only mixtures of pure invariant procedures. If $\mathcal{G}$ is compact, we get the same result by averaging with respect to the invariant probability measure in $\mathcal{G}$. If $\mathcal{G}$ is abelian, a limit of averages under Haar measure restricted to large compact sets works for testing problems and most other problems that arise. Examples can be found, somewhat artificial, where the result does not hold for the additive group of real numbers. Since a step by step reduction is possible, at least for testing problems,

the result also holds for solvable groups and for groups like the group of rigid motions in Euclidean space (or even similitudes). No other affirmative results are known to me. Numerous examples are known in multivariate analysis of problems invariant under the real or complex full linear group even in two dimensions, that do not possess an invariant minimax solution. There are, however, special problems invariant under the full linear group, where invariant minimax solutions do exist. In other problems, including Hotelling's T^2-test the answer is, as far as I know, still unknown. Even for simple, nonparametric problems, invariant under the group of all increasing homeomorphisms, the answer is unknown. Curiously, for the group of all homeomorphisms of the circle a negative answer is easy.

Next, we look at the question of whether an invariant procedure admissible in the class of invariant procedures is necessarily admissible in the class of all procedures. If the group $\mathcal{G}$ is finite, or more generally compact, an affirmative answer is obtained as before by averaging over $\mathcal{G}$. If the group $\mathcal{G}$ is the additive group of the real line and operates transitively on the parameter space, and the formal Bayes procedure with respect to Π is unique, and if, roughly speaking, the translation parameter has one more moment than is needed for the problem to be meaningful, the answer is again affirmative. This is not a precise result but a summary of many special results. For the additive group of the two-dimensional real linear space, a similar result holds, under stronger restrictions. Beyond this, results are negative except for very special problems.

Both the minimax problem and the admissibility problem considered here can sometimes be solved positively or negatively by considering procedures invariant under a proper subgroup of the given group $\mathcal{G}$. However, it seems likely that it will be more useful in the long run to study statistical problems having group theoretical aspects by the general methods of statistical decision theory, using the groups only to obtain more nearly explicit and understandable results.

6. Similar tests

Let $\mathcal{Z}$ and ω be sets, $\mathcal{B}$ a σ-algebra of subsets of $\mathcal{Z}$ and P a function on ω to the set of probability measures on $\mathcal{B}$. A $\mathcal{B}$-measurable function φ on $\mathcal{Z}$ to $[0,1]$ will be called a test and this test will be called a similar test of size α, where α is a real number between 0 and 1, if

$$\alpha = E_\theta \varphi(Z) = \int \varphi(z) P_\theta(dz) \text{ for all } \theta \in \omega.$$

The interpretation is that, being interested in testing the hypothesis that an observable random point Z of $\mathcal{Z}$ is distributed according to some P_θ with $\theta \in \omega$, we reject this hypothesis with conditional probability $\varphi(Z)$, and then our probability of rejecting the hypothesis if it is true is the constant α.

We shall need the notions of sufficient sub σ-algebra and sufficient statistic. A sub σ-algebra $\mathcal{C}$ of $\mathcal{B}$ is said to be sufficient with respect to the indexed family $\{P_\theta\}$, $\theta \in \omega$, introduced above, if there exists a determination of the conditional probability $P^{\mathcal{C}}$ that does not depend on θ. A measurable mapping $T: (\mathcal{Z}, \mathcal{B}) \to (\mathcal{Z}_1, \mathcal{B}_1)$ where $\mathcal{Z}_1$ is a set and $\mathcal{B}_1$ is a σ-algebra of subsets of $\mathcal{Z}_1$ is called a statistic. If the σ-algebra $T^{-1}\mathcal{B}_1$ is sufficient, the statistic T is said to be sufficient. It was observed, and successfully applied by Neyman, that if $E^{\mathcal{C}}\varphi = \alpha$ then φ is similar of size α and, if the family of $\{P_\theta|\mathcal{C}\}$, $\theta \in \omega$, is complete in the sense that there exists no bounded $\mathcal{C}$-measurable function differing essentially from 0 whose expectation is 0 for all $\theta \in \omega$, the condition $E^{\mathcal{C}}\varphi = \alpha$ is also necessary for φ to be similar.

In many common statistical problems, in particular, all those dealing with the normal distribution, we are concerned with an exponential family in the following sense. The sample space $\mathcal{Z}$ is a finite dimensional real linear space, the parameter space $\mathcal{H}$ is an open subset of the dual space $\mathcal{Z}'$, and

$$P_\theta(dz) = e^{-\chi(\theta)+\theta' z}\mu(dz),$$

where μ is a σ-finite measure and χ is the logarithm of the Laplace transform of μ. We may be interested in testing the hypothesis $\theta \in \omega$ where ω is a given subset of $\mathcal{H}$. If ω is an open subset of an affine subspace $\mathcal{Y}$ of $\mathcal{Z}'$, the class of all similar tests is easily characterized by the method of Neyman described above. The natural mapping T of $\mathcal{Z}$ onto the quotient space of $\mathcal{Z}$ modulo the annihilator of the linear space parallel to $\mathcal{Y}$ is a complete sufficient statistic for the family $\{P_\theta\}$, $\theta \in \omega$, so that the condition for φ to be similar is $E(\varphi(Z)|T(Z)) = \alpha$, where the left-hand side is an alternative notation for $E^{\mathcal{C}}\varphi$, with the sub σ-algebra induced by T. It is then not difficult to find, for example (by another Lemma of Neyman and Pearson), the similar test having maximum expectation at a given P_θ with $\theta \notin \omega$.

Except for a theorem of Lyapunov and similar tests obtained by considerations of group invariance, both of which will be considered later, this was essentially the status of the theory of similar tests until some recent work of Wijsman (1958) and more recent, more nearly definitive work of Linnik and his associates. In discussing this work in the case where ω is a curved submanifold of $\mathcal{Z}'$ it will be convenient to use the notion of предкотест introduced by Linnik. We suppose the σ-finite measure μ in the exponential family considered has the

form

$$\mu(dz) = f(z)\,dz.$$

A measurable function ρ on $\mathfrak{Z}$ is called a предкотест if its Laplace transform is absolutely convergent and vanishes on ω. If φ is a similar test of size α for ω, then ρ defined by

$$\rho(z) = [\varphi(z) - \alpha] f(z)$$

is a предкотест. It was observed by Wijsman that if $\omega = \{\theta : \Pi(\theta_1, \ldots, \theta_n) = 0\}$ where Π is a polynomial, then ρ defined by

$$\rho(t) = \Pi\left(\frac{\partial}{\partial t_1}, \ldots, \frac{\partial}{\partial t_m}\right) G(t)$$

is a предкотест, provided G is such that the appropriate integration by parts can be performed. In an interesting special case, the hypothesis that the ratio of the squared mean of a normal distribution to its variance is a given constant, he was able to show that, suitably interpreted, this yields all similar tests.

Linnik studied the class of предкотест as an ideal in the convolution algebra of functions possessing absolutely convergent Laplace transforms. In particular, he obtained in this way some interesting results about the Behrens-Fisher problem which I shall describe next.

Let $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ be independently normally distributed with unknown means and variances: $EX_i = \xi$, $EY_j = \eta$, and $E(X_i - \xi)^2 = \sigma^2$, $E(Y_j - \eta)^2 = \tau^2$. The Behrens-Fisher problem is that of testing the hypothesis that $\xi = \eta$. It was recognized by Neyman that the solution proposed by Behrens and later independently by Fisher does not have the property of similarity imposed here. For a long time it was an unsolved problem, attracting the attention of many statisticians, including Wilks and Wald, whether there existed non-randomized similar tests based only on the minimal sufficient statistic for the over-all problem which is $\sum_1^m X_i$, $\sum_1^n Y_j$, $\sum_1^m X_i^2$, $\sum_1^n Y_j^2$. Linnik has shown, roughly speaking, that there do exist such non-randomized tests (at least for a large class of pairs m, n) but no reasonable ones. The precise form of the latter result is as follows. We confine our attention to tests invariant under location and scale change, that is, tests of the form $\varphi\left(\frac{(\overline{X}-\overline{Y})^2}{S_2}, \frac{S_1}{S^2}\right)$ where

$$\overline{X} = \frac{1}{m}\sum X_i, \quad \overline{Y} = \frac{1}{n}\sum Y_j$$

$$S_1 = \sum (X_i - \overline{X})^2, \quad S_2 = \sum (Y_j - \overline{Y})^2.$$

Linnik showed that there exists no similar test of this form for which there exists $\varepsilon > 0$ such that $\varphi(t_1, t_2) = 0$ for all $t_1 < \varepsilon$.

7. Large sample theory

Classical large sample theory is concerned with the following sort of situation. There is an unknown parameter point θ whose value lies in a given differentiable manifold $\mathcal{H}$, a sample space $\mathcal{X}$ together with a σ-finite measure μ in $\mathcal{X}$, and a function p on $\mathcal{H} \times \mathcal{X}$ which is, for fixed θ a probability density function with respect to μ, and satisfies certain regularity conditions as a function on $\mathcal{H}$ to the set of probability density functions with respect to μ. We observe $X_1, \ldots, X_n$ independently identically distributed according to p_θ. Here n is large and we are interested in making inferences about θ. A positive definite covariant tensor field, Fisher's information matrix, is introduced by

$$I_{ij}(\theta) = E_\theta \frac{\partial \log p_\theta(X_1)}{\partial \theta_i} \frac{\partial \log p_\theta(X_1)}{\partial \theta_j},$$

and used to make $\mathcal{H}$ a Riemannian manifold. Then it is proved that there exist estimates $\hat{\theta}^{(n)}(X_1, \ldots, X_n)$ which, for large n are, in an appropriate local coordinate system approximately normally distributed with mean θ and covariance $\frac{1}{n} I^{-1}(\theta)$, for example, Bayes estimates with respect to a smooth prior density and, under stronger conditions, maximum likelihood estimation. Also, it is proved that there do not exist estimates much better than these. Problems such as testing the hypothesis that θ lies in a smooth submanifold are treated in the obvious way, but with considerable technical difficulties, by using the fact that the Riemannian geometry is to a first approximation Euclidean. Recent work in classical large sample theory has taken various directions. First, even when the problem is of the above sort, considerable ingenuity may be required to compute the estimates $\hat{\theta}^{(n)}$ for real data. Second, we may have independent but non-identically distributed observations, or observations forming a Markov process, or some still more general situation, where classical large sample methods are still applicable. It was recognized by Le Cam that this is the case, very roughly speaking, if the following conditions are satisfied. Let X be the entire sample and p_θ its density. If the random variability of $-\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log p_\theta(X)$ is small, and its variation as a function of θ is also small over appreciable distances in the Riemannian metric determined by

$$I_{ij}(\theta) = E_\theta\left[\frac{\partial}{\partial\theta_i} \log p_\theta(X) \frac{\partial}{\partial\theta_j} \log p_\theta(X)\right] =$$

$$= -E_\theta \frac{\partial^2}{\partial\theta_i\,\partial\theta_j} \log p_\theta(X),$$

the usual large sample results hold. Le Cam did not use second derivatives. Some interesting work on estimates and tests based on order statistics is also close in spirit to classical large sample theory.

Recently it has been recognized that there are many reasonable problems concerning large samples that are not included in the above framework. Hoeffding (1965) in "Asymptotically optimum tests for multinomial distributions", studies the behavior of tests at fixed alternatives when the sample size tends to ∞. Here classical large sample theory is useless because the normal approximation does not give an asymptotically correct evaluation of small probabilities. Instead, the Kullback-Leibler information numbers, related to Shannon's information, and the theory of large deviations, studied recently by Sanov (1957) and others, are important tools.

We consider $X_1, \ldots, X_n$ independently identically distributed according to a multinomial distribution

$$P\{X_i = j\} = p_j, \text{ for } j = 1, \ldots, k \quad (k \text{ fixed})$$

with

$$\sum p_j = 1,$$

and let

$$Z_j^{(n)} = \frac{1}{n} \text{ (number of } i \text{ for which } X_i = j).$$

Suppose we are interested in testing the hypothesis H_0: $p_j = \frac{1}{k}$ for $j = 1, \ldots, k$ with a very small level of significance α_n (probability of rejecting H_0 if true). Hoeffding considers a much more general problem, but the main ideas are brought out by this special case. Hoeffding shows that the likelihood ratio test, which in this case rejects H_0 when

$$\sum Z_i^{(n)} \log Z_i^{(n)} > c_n \text{ (an appropriate constant)}$$

has, for most alternatives p, a probability of rejecting H_0 that is about as large as you could hope for, even if you knew the p at which you were trying to approximate this probability. The χ^2-test, rejecting H_0 when

$$\sum_{j=1}^{k} \left(Z_j^{(n)} - \frac{1}{k}\right)^2 > c_n',$$

is for almost all p, quite poor if n is large. In classical large sample theory, which studies alternatives close to H_0 the likelihood-ratio test and the χ^2-test are asymptotically equivalent. Curiously, the χ^2-test is definitely superior to the likelihood-ratio test if k is large and $\frac{n}{k}$ moderate, for tests at moderate significance levels against not-too-distant alternatives. It is also true that for all k and n both the χ^2-test and the likelihood-ratio test are Bayes tests against many prior distributions covering the alternatives thoroughly. This and other results were obtained by a student, Carl Morris, in a dissertation written at Stanford. It would be desirable to have a good over-all picture of the behavior of symmetric tests for the hypothesis considered here subject only to the condition that n is large.

8. Multivariate analysis

Statisticians in both practical and theoretical work have long been interested in statistical procedures connected with the multivariate normal distribution. A random real coordinate vector X has a multivariate normal distribution with mean ξ and positive definite covariance matrix Σ if its density with respect to Lebesgue measure is given by

$$(1) \qquad p(x) = \frac{1}{(2\pi)^{p/2} (\det \Sigma)^{1/2}} e^{-\frac{1}{2} \operatorname{tr} \Sigma^{-1} (x-\xi)(x-\xi)'}.$$

For a combination of reasons, it is interesting to try to apply the general ideas of statistical decision theory to this class of problems. The class of non-singular multivariate normal distributions is an exponential family, that is, the exponential in the density is a bilinear form in the parameter $(\Sigma^{-1}, \Sigma^{-1}\xi)$ and the sample point $(x'x, x)$. The class of such distributions is also invariant under the full affine group (linear transformations combined with translations). Thus, we have a fair amount of general theory available for studying these problems. On the other hand, these problems are frequently non-trivial because of the large number of unknown parameters if the dimension p is large. A survey paper by Kiefer (1966) on optimality results in multivariate analysis will soon be published. For simplicity and definiteness, I shall speak only about Hotelling's T^2-test.

If we have a sample $X_1, \ldots, X_n$ (with $n \geqslant p+1$) random vectors independently identically distributed according to (1) we may be interested in testing the hypothesis H_0: $\xi = 0$. The standard test is Hotelling's T^2-test which rejects H_0 if $T^2 = \overline{X}' S^{-1} \overline{X} > c$ where

$$\overline{X} = \frac{1}{n} \sum_1^n X_i, \quad S = \sum_1^n (X_i - \overline{X})(X_i - \overline{X})'.$$

It is trivial that Hotelling's T^2-test is uniformly most powerful among tests invariant under the full linear group. We have seen that the full linear group is too large to be any assurance that Hotelling's T^2 has any absolute optimum properties. However, it is known that this test is admissible. The first proof by the author was obtained

as an extension of a result of Allen Birnbaum, who showed that any compact acceptance region is admissible for testing a simple hypothesis concerning an exponential family with unrestricted parameter space. This proof, although correct, is unsatisfactory because, roughly speaking, it shows that Hotelling's T^2-test cannot be improved upon for those parameter values where it is so powerful that we do not care much whether we can improve upon it. A recent proof by Kiefer and Schwartz is much more satisfactory. They observe that Hotelling's T^2-test is a Bayes test with respect to the prior distribution $\lambda\Pi_0 + (1-\lambda)\Pi_1$, where both Π_0 and Π_1 assign probability 1 to Σ's of the form $(1+\eta\eta')^{-1}$ with η ranging over the set of p-dimensional vectors, and

(i) Under Π_0, $\xi = 0$ and η is distributed according to the density

$$\pi_0(\eta) = (I+\eta\eta')^{-n/2}.$$

(ii) Under Π_1, $\xi = (1+\eta\eta')^{-1}\eta$ and η is distributed according to the density

$$\pi_1(\eta) = (1+\eta\eta')^{-n/2} e^{\frac{1}{2}\operatorname{tr}\eta'(1+\eta\eta')^{-1}\eta}.$$

Other prior distributions that work can be obtained by applying linear transformations to these, by using analogous formulas with $1 + c\eta\eta'$ rather than $1 + \eta\eta'$ and by taking convex combinations. Kiefer and Schwartz have extended both methods to a large class of problems.

It is still not known whether Hotelling's T^2-test is minimax against the alternatives $\xi^1\Sigma^{-1}\xi = \tau^2$ (a positive constant), except for the case $p = 2$, $n = 3$, where Giri, Kiefer, and the author proved that it is minimax. The proof was obtained by showing that among all tests invariant under the (solvable) multiplicative group of lower triangular matrices, Hotelling's T^2-test is Bayes with respect to a rather complicated explicitly computed prior distribution. Some approximate results on the more general problem have been obtained by Giri and Kiefer and by Linnik and his associates. It seems likely that methods analogous to those used by Linnik in the problem of similar regions for exponential families, but with the role of the sample and parameter point interchanged will be useful here.

A contrasting result for a problem that seems at first to be similar to the above concerns the problem of finding confidence sets for ξ. If, after observing $X_1, \ldots, X_n$ as above we assert that ξ is in the ellipsoid $(\xi - \bar{X})' S^{-1} (\xi - \bar{X}) \leqslant c$, with c suitably chosen depending on p and n, we can achieve a constant probability α of covering ξ, independent of ξ and Σ. However, it can be shown that the same probability of covering ξ can be obtained by a procedure that is not invariant under all affine transformations, but has an expected volume

smaller by a constant factor than that of the above random set based on Hotelling's T^2. This other procedure is obtainable as the optimum among all procedures invariant under affine transformations having a linear part whose matrix in a fixed coordinate system is lower triangular. This has not been brought to the point where it is applicable in practice.

People have recently started to study the complex analogues of the real multivariate normal distribution, which arise naturally in the spectral analysis of multiple time series. Complex analogues of all the common problems in the real case exist, and it is usually routine to develop the theory to the point that has been reached in the real theory. There is some indication that optimal properties of the complex analogues may be easier to study than the original real problems, because, for example, the distribution of the maximal invariant under the lower triangular group for the complex analogue of Hotelling's T^2 involves only polynomials times exponentials. Also, in the complex case, there are problems that have no real analogue.

A great deal of work has been done on devising tests and estimates for complicated practical problems in multivariate analysis. The work of Anderson and Rubin (1949) and others on structural equations is especially important. Also, I should mention the work of Alan James (1964), Constantine, and others using the theory of group representations to derive the distributions of certain multivariate statistics. Another interesting direction in multivariate analysis is suggested by some work of Wigner and other physicists, collected in Porter. Suppose we observe $X_1, \ldots, X_n$ independently normally distributed with mean 0 and the identity matrix as covariance matrix, and let L be the empirical cumulative distribution function of the roots of the equation

$$\det(\Sigma X_i X_i' - \lambda I) = 0.$$

Then, if I have not made a mistake, for p large and $\frac{n}{p}$ moderate (but $n \geqslant p$), L will, with high probability, be close to the integral of

$$p(x) = \begin{cases} 0 \text{ for } x \leqslant a \text{ or } x \geqslant b \\ \frac{c}{x}\sqrt{(x-a)(b-x)} \end{cases}$$

where $a = (\sqrt{n} - \sqrt{p})^2$, $b = (\sqrt{n} + \sqrt{p})^2$. The proof follows the method used by Wigner in an analogous, but slightly simpler problem. The assumption of normality is not really used. I believe it will be useful to make a systematic study of multivariate problems of high dimension.

9. Sequential problems

In sequential analysis, the mathematical treatment takes into account the possibility that the statistician may take observations, one at a time, deciding after each observation whether to take another observation or make a final decision. We shall consider only the case of independent identically distributed observations. We have a parameter space $\mathscr{H}$, a terminal action space $\mathscr{A}$ and a non-negative loss function L on $\mathscr{H} \times \mathscr{A}$, and we let $\mathfrak{X}$ be the space of the individual observations and P_θ the distribution of an observation when θ is the true parameter point. We have to decide on a rule for when to stop, which determines a random variable N, the number of observations taken, and we must also choose a sequence φ_n of functions, $\varphi_n : \mathfrak{X}^n \to \mathscr{A}$. After taking N observations $X_1, \ldots, X_N$ we take the terminal action $\varphi_N(X_1, \ldots, X_N)$. Our aim is to choose the stopping rule so as to keep $E_\theta N$ and $E_\theta L(\theta, \varphi_M(X_1, \ldots, X_N))$ small.

Again the Lagrange multiplier argument suggests that we look at the problem of minimizing

$$c \int E_\theta N \, d\Pi_1(\theta) + \int E_\theta L(\theta, \varphi_N(X_1, \ldots, X_N)) \, d\Pi_2(\theta)$$

where c is a constant and Π_1 and Π_2 are measures in $\mathscr{H}$ for which the integrals exist for some choice of stopping rule and $\{\varphi_n\}$. For simplicity we consider only the case where $\Pi_1 = \Pi_2 = \Pi$, say, a probability measure. Thus, we are interested in minimizing

$$E[cN + L(\theta, \varphi_N(X_1, \ldots, X_N))]$$

where θ is distributed according to Π and the X_i given θ are independently identically distributed according to P_θ. I shall assume the P_θ are dominated by a σ-finite measure μ: $dP_\theta = p_\theta \, d\mu$.

The sequence of posterior distributions $\Pi^*_{X_1 \ldots X_n}$ is a vector-valued martingale and a Markov process with temporally stationary transition probabilities. It is also spatially stationary with respect to the group of projective transformations of the simplex of prior probability measures leaving the vertices invariant. Subject to a continuity condition it seems to be the most general process with all these properties.

It has been proved under fairly general conditions that the Bayes procedures and the Bayes risk function

$$\rho^*(\Pi) = \inf E^\Pi E_\theta [cN + L(\theta, \varphi_N(X_1, \ldots, X_N))],$$

the infimum being over all stopping rules and terminal decision functions, are obtainable as follows. The function ρ^* satisfies the integral equation

$$\rho^*(\Pi) = \min [E^\Pi \rho^*(\Pi^*_{X_1}) + c, \ \psi(\Pi)]$$

where

$$\psi(\Pi) = \min_a E^\Pi L(\theta, a)$$

and the rule is to take another observation if

$$\psi(\Pi^*_{X_1 \ldots X_n}) \underset{(=)}{>} \rho^* \Pi^*_{X_1, \ldots X_n})$$

and otherwise to stop and take an action a_0 such that

$$E^{\Pi^*_{X_1 \ldots X_n}} L(\theta, a_0) = \psi(\Pi^*_{X_1 \ldots X_n}) \underset{(=)}{<} \rho^*(\Pi^*_{X_1 \ldots X_n}).$$

The formalism can easily be extended to include some aspects of the design problem, allowing the statistician to choose at each stage among a number of different kinds of observations.

The basic equation has been solved explicitly in only a few special cases. Some limiting results have been obtained in the case $c \downarrow 0$ with other features remaining fixed. Another case that can sometimes be handled is that where each observation provides very little information and c is reduced accordingly; in the limit the continuous time case. Under certain conditions the basic integral equation becomes a simple linear differential equation with a rather complicated free boundary condition. One example has been considered extensively by Chernoff.

Stanford University, USA

CHARACTERIZATIONS OF FINITE SIMPLE GROUPS

JOHN G. THOMPSON

1. Introduction

In the last four years, several results about finite groups have been obtained. The methods of proof are not easy to master, though in large measure they bear a striking fidelity to the foundations laid at the turn of the century by Frobenius, Schur, Burnside and Sylow.

At the moment we have no idea how much further effort will be necessary to classify the finite simple groups. Considering the time which has elapsed since Mathieu discovered his groups, and considering that no one pretends to understand these groups, any optimism must be guarded.

However, it may also be said that recent techniques have led to results which ten years ago seemed impenetrable, and that the power of these techniques is not yet exhausted.

2. Characterizations

Let $\mathcal{S} = \{L_2(q), L_3(q), S_z(q), U_3(q), A_7, M_{11}\}$. Most, but not all, of the recent characterization theorems deal with $\mathcal{S}$. For example,

Theorem. *If G is a non abelian simple group and $G \notin \mathcal{S}$, then*

(1) *G contains a solvable subgroup $\neq 1$ with non solvable normalizer.*

(2) (Gorenstein-Walter) [5]. *Sylow 2-subgroups of G are not dihedral.*

(3) (Suzuki) [9]. *G contains an involution whose centralizer does not have a normal Sylow 2-subgroup.*

In 1963, I announced that (1) held with finitely many exceptions. The complete proof of (1) has not yet appeared. The proof is complicated and it will take several years to determine the extent to which the various arguments admit of useful generalization.

One of the pregnant and technical parts of the proof of (1) is given by

Theorem ES. *$E_2(3)$ and $S_4(3)$ are the only simple groups G such that*

(i) $2,3 \in \pi_4(G)$.

(ii) *If $p \in \{2,3\}$, G_p is a S_p-subgroup of G and $A \in \mathcal{S}cn_3(G_p)$, then $\mathbf{N}(A)$ contains only 1.*

(iii) *The normalizer of every non identity 3-subgroup of G is solvable.*

(iv) *The centralizer of every involution of G is solvable.*

(v) $2 \sim 3$.

Gorenstein [4] has substantially generalized a portion of this theorem by characterizing the groups $E_2(3^n)$.

To explain the meaning of the various statements of the theorem requires a bit of notation. If π is a set of primes, π' is the complementary set of primes. A π-signalizer of a group X is a subgroup A such that $|A|$ and $|X : \mathbf{N}(A)|$ are π'-numbers. Let $\mathcal{S}cn(X)$ be the set of self centralizing normal subgroups of X and let $m(X)$ be the minimal number of generators of X. Let

$\mathcal{S}cn_m(X) = \{A \mid A \in \mathcal{S}cn(X),\ m(A) \geqslant m\}$. Let

$\pi_1(X) = \{p \mid$ a S_p-subgroup of X is a cyclic group $\neq 1\}$.

$\pi_2(X) = \{p \mid$ a S_p-subgroup X_p of X is non cyclic and $\mathcal{S}cn_3(X)_p = \varnothing\}$.

$\pi_3(X) = \{p \mid$ if X_p is a S_p-subgroup of X, then $\mathcal{S}cn_3(X_p) \neq \varnothing$ and 1 is not the only p-signalizer of $G\}$.

$\pi_4(X) = \{p \mid$ if X_p is a S_p-subgroup of X, then $\mathcal{S}cn_3(X_p) \neq \varnothing$ and 1 is the only p-signalizer of $X\}$.

If H is a subgroup of X, $\mathbf{M}(H) = \mathbf{M}_X(H)$ is the set of all subgroups K of X such that $K \cap H = 1$ and $H \subseteq \mathbf{N}(K)$. If X is a p-group, define $\mathcal{U}(X)$ as follows: if $\mathbf{Z}(X)$ is non cyclic, $\mathcal{U}(X)$ is the set of non cyclic subgroups of $\mathbf{Z}(X)$ of order p^2; if $\mathbf{Z}(X)$ is cyclic, $\mathcal{U}(X)$ is the set of non cyclic normal subgroups of X of order p^2. For general X, let $\mathcal{U}(p) = \mathcal{U}_X(p) = \bigcup \mathcal{U}(X_p)$, where X_p ranges over all the S_p-subgroups of X. If p is an odd prime, write $2 \sim p$ if and only if X has a solvable subgroup which contains a non cyclic abelian subgroup of order 8 and a non cyclic p-subgroup each element of which centralizes an element of $\mathcal{U}(p)$. These definitions explain the hypotheses of Theorem ES and serve to introduce some of the objects of current interest.

Another result deals with the groups $E_2^*(q)$ of Ree.

Theorem R. (Ward-Janko-Thompson) [10], [7]. *If G is a simple group with abelian S_2-subgroups and if G contains an involution i such that $\mathbf{C}(i) = \langle i \rangle \times L$, where $L \simeq L_2(q)$ and $q > 5$, then*

(a) *q is an odd power of 3.*

(b) $|G| = q^3(q-1)(q^3+1)$.

(c) *If P is a S_3-subgroup of G and N is its normalizer, then G is doubly transitive on the cosets of N in G and $G = N \cup NtP$ where t is an involution of G which inverts a S_3-subgroup of N.*

We would like to conclude that $G \simeq E_2^*(q)$. This is still open and gives rise to the conjecture

(C_1) The character table of a group determines the Brauer characters.

If the hypothesis $q > 5$ is replaced by $q \geqslant 5$ in Theorem R, Janko [6] has shown that precisely one further group arises. This group is new and is another tantalizing reason for studying simple groups.

Janko's work disclosed a lamentable error in one of my announcements which vitiates some results of Sah [8]. Janko's work contains a lovely application of the results of Brauer for blocks of defect 1.

Brauer and Fong [1] have characterized M_{12} and Wong [11] has characterized A_8. The isomorphism $A_8 \simeq L_4(2)$ places A_8 in the family of Chevalley groups. We have yet to take our first steps toward the characterization of A_n for $n \geqslant 9$.

3. Corollaries

(I) The group G is solvable if and only if every pair of elements generates a solvable subgroup.

(II) G is non solvable if and only if there are non identity elements a, b, c of G of pairwise coprime orders with $abc = 1$.

(III) (Gallagher [2]) G is non solvable if and only if there is a non principal irreducible character of G whose restriction to each Sylow subgroup contains the principal character.

(IV) If G is non solvable of order $p^a q^b r^c$, then one of the following groups is a subquotient of G:

$$L_2(q), \quad q = 5,\ 7,\ 8,\ 17, \quad L_3(3).$$

We may mention the conjectures

(C_2) There are only finitely many simple non abelian groups of order $p^a q^b r^c$.

(C_3) If G is a non abelian simple group and $3 \nmid |G|$, then G is a Suzuki group.

4. Techniques

If N_1, N_2, N_3 are subgroups of a group X and if for every permutation σ of $\{1, 2, 3\}$, $N_{\sigma(1)} \subseteq N_{\sigma(2)} N_{\sigma(3)}$, then $N_1 N_2$ is a subgroup. This elementary observation has numerous applications. When stripped of their group theoretic significance, these applications sometimes depend on the fact that the proposition $(p \vee q) \wedge (q \vee r) \wedge (r \vee p)$ is equivalent to the proposition obtained by interchanging $\vee$ and $\wedge$. I can illustrate this symmetry rather easily. Suppose P is a S_p-subgroup of a group X, A_1, A_2, A_3 are weakly closed subgroups of P, and $N_i = \mathbf{N}(A_i)$. Suppose also that

(a) $\mathbf{N}(P)$ is a maximal subgroup of X.

(b) $X = N_1 N_2 = N_2 N_3 = N_3 N_1$.

Then at least 2 of A_1, A_2, A_3 are normal in X.

There are many variations of this theme, and taken together, they are quite helpful. The difficulty is in finding subgroups A_i of P. This requires some discussion.

In working on the problem so brilliantly solved by Shafarevich and Golod, I was led to consider the following invariant of a p-group P: $d = d(P) = \max\{m(A)\}$, where A ranges over all the abelian subgroups of P. This invariant leads into thickets which I could not penetrate. However, my efforts were not without value, for I eventually realized that the related group $\mathbf{J}(P) = \langle A \mid A^* = 1, m(A) = d(P)\rangle$ plays an exploitable role in the structure of p-solvable groups. In particular, if X is a p-solvable group, $\mathbf{O}_{p'}(X) = 1$ and $SL(2, p)$ is not a subquotient of X, then $X = N_1 N_2$, $N_i = \mathbf{N}(A_i)$, $A_1 = \mathbf{Z}(X_p)$, $A_2 = \mathbf{J}(X_p)$, and where X_p is a S_p-subgroup of X.

When one considers the above result one is led to try to find a *uniformly normal* subgroup. This term requires some elaboration. Suppose P is a p-group $\neq 1$. Let $\mathscr{S}(P)$ be the set of all p-solvable groups X such that $\mathbf{O}_{p'}(X) = 1$ and P is a S_p-subgroup of X. The subgroup H of P is said to be uniformly normal provided $1 \neq H \lhd X$ for all X in $\mathscr{S}(P)$. It is a remarkable result of Glauberman [3] that if $p \geqslant 5$, $\mathbf{Z}(\mathbf{J}(P))$ is uniformly normal. This result has already led to a subtheory of finite groups with substantial ramifications. If $p \leqslant 3$, the "old" factorizations still appear indispensable, though much remains to be done.

A second technique involves transitivity theorems. As these have been discussed elsewhere, I need not elaborate.

The object of the factorizations and the transitivity theorems is to obtain information about $\mathscr{M}^*(G)$. To define $\mathscr{M}^*(G)$, we let $\mathscr{S}ol(G)$ be the set of solvable subgroups of G. $\mathscr{S}ol(G)$ is partially ordered by inclusion; $\mathscr{M}\mathscr{S}(G)$ is the set of maximal elements of $\mathscr{S}ol(G)$ and $\mathscr{M}^*(G)$ is the set of elements of $\mathscr{S}ol(G)$ which are contained in just 1 element of $\mathscr{M}\mathscr{S}(G)$. Thus, there is a map $\mathbf{M}: \mathscr{M}^*(G) \to \mathscr{M}\mathscr{S}(G)$ defined by $\mathbf{M}(H) =$ the unique element of $\mathscr{M}\mathscr{S}(G)$ which contains H. Several difficult results in the proof of (I) are of the shape $H \in \mathscr{M}^*(G)$.

A third technique has been introduced by Suzuki and has its roots in work of Zassenhaus. So far this technique has been used only for a limited class of doubly transitive groups. Suppose G is doubly transitive on Ω and $\alpha \in \Omega$. Let $H = G_\alpha$ and suppose $H = QK$, where $Q \lhd H$, $Q \cap K = 1$, and where Q is regular and transitive on $\Omega - \alpha$. Suppose also that t is an involution of G which normalizes K. Then $G = H \cup HtQ$ and if $x \in Q - \{1\}$, $txt = h(x)\, tf(x)$, where $h(x) \in H$, $f(x) \in Q$. These equations are the structure equations for G and G is completely determined by H, the structure equations and the automorphism of K induced by t. The groups $L_2(q)$, $U_3(q)$, $S_z(q)$ and $E_2^*(q)$ satisfy these hypotheses.

Elegant and subtle arguments of Suzuki deal with the structure equations. In the hope of extending Suzuki's ideas to obtain a characterization of $E_2^*(q)$, I have studied the structure equations of the groups which appear in Theorem R. One of the difficulties is that the structure-

equations for $E_2^*(q)$ have never been determined. If (C_1) is true, the difficulties can probably be avoided.

Dept. of Mathematics,
University of Chicago, USA

REFERENCES

[1] Brauer R., Fong P., A characterization of the Mathieu group M_{12}, to appear.
[2] Gallagher P. X., Group characters and Sylow subgroups, *J. London Math. Soc.*, **39** (1964), 720-722.
[3] Glauberman G., A characteristic subgroup of a p-stable group, to appear.
[4] Gorenstein D., Finite simple groups and the family $G_2(3^n)$, to appear.
[5] Gorenstein D., Walter J., The characterization of finite groups with dihedral Sylow 2-subgroups, *J. of Alg.*, **2** (1965), 85-151, 218-270, 354-393.
[6] Janko Z., A new finite simple group with Abelian Sylow 2-subgroups and its characterization, *J. of Alg.*, **3**, No. 2, 147-186.
[7] Janko Z., Thompson J., On a class of finite simple groups of Ree, *J. of Alg.*, to appear.
[8] Sah C. H., A glass of finite groups with Abelian 2-Sylow subgroups, *Math. Zeit.*, **82** (1936), 335-346.
[9] Suzuki M., Finite groups in which the centralizer of any element of order 2 is 2-closed, *Ann. of Math.*, **82** (1965), 191-212.
[10] Ward H. N., On Ree's series of simple groups, *Transactions of the Amer. Math. Soc.*, **121**, No. 1 (1966), 62-89.
[11] Wong W., A characterization of the alternating group of degree 8 *Proc. London Math. Soc.*, **13** (1963), 359-383.

О РАЗВИТИИ ЗА ПОСЛЕДНИЕ ГОДЫ АНАЛИТИЧЕСКОЙ ТЕОРИИ ЧИСЕЛ

И. М. ВИНОГРАДОВ, А. Г. ПОСТНИКОВ

Введение

Цель нашего доклада состоит в том, чтобы на примере новых исследований в теории чисел показать некоторые связи этой науки с другими областями математики, а также взаимодействие различных направлений внутри нее самой.

Наш доклад не является полным обзором. Ряда направлений аналитической теории чисел мы не затронем совсем, а по другим дадим лишь беглые сведения. В этом нет никакой принципиальной установки, дело в лимите времени, отпущенном на доклад.

В подготовке доклада нам оказали большую помощь многие советские математики. Мы приносим им нашу сердечную благодарность.

Диофантовы приближения и применения метода тригонометрических сумм

Классической задачей, относящейся к теории диофантовых приближений, является оценка количества целых точек (точек с целыми координатами) в областях евклидова пространства. При широких ограничениях на область это количество приближенно равно объему области. Поэтому предметом исследования явились оценки погрешности в этом соотношении. На этих задачах формировался и формируется метод тригонометрических сумм.

Мы остановимся сейчас на связях задачи о подсчете количества целых точек в областях с исследованием асимптотического поведения собственных чисел краевой задачи

$$\Delta u + \lambda u = 0, \qquad u \in D, \tag{1}$$

$$u|_B = 0 \quad \text{или} \quad \left.\frac{\partial u}{\partial n}\right|_B = 0, \tag{2) или (2')}$$

где D есть конечная область n-мерного пространства, B — ее граница, Δu — оператор Лапласа. Речь будет идти об асимптотическом поведении величины λ в зависимости от ее номера (собственные значения располагаем в порядке возрастания). Эту задачу можно формулировать как исследование при $T \to \infty$ асимптотического поведения величины

$$N(T) = \sum_{\lambda \leqslant T} 1.$$

Ограничимся плоским случаем. Г. Вейль получил свою асимптотическую формулу

$$N(T) \sim \frac{\text{mes } D}{4\pi} T,$$

аппроксимируя область D квадратами. Для квадрата со стороной, равной единице, собственные числа $\lambda = \pi^2(n^2 + m^2)$, где n и m — натуральные числа. Ввиду этого к изучению величины $N(T)$ можно применить теорию целых точек в круге. Для уравнения (1) при условии (2) получается

$$N(T) = \frac{T}{4\pi} - \frac{\sqrt{T}}{\pi} + o(\sqrt{T}).$$

Наличие в этой асимптотической формуле двух членов, по-видимому, дало повод Г. Вейлю высказать гипотезу, что в «общем случае» формула для $N(T)$ должна иметь вид

$$N(T) = \frac{\text{mes } D}{4\pi} T \pm \frac{L(D)}{4\pi} \sqrt{T} + o(\sqrt{T})$$

($L(D)$ — периметр области). Эта гипотеза до сих пор не доказана.

Н. В. Кузнецов (первая работа, совместная с Б. Ф. Федосовым) доказал гипотезу Г. Вейля для областей, в которых красвая задача допускает разделение переменных. Таких типов областей конечное число, и все они описаны Эйзенхартом. Следуя одной идее Титчмарша, собственные числа можно занумеровать двумя целочисленными индексами $\lambda = \lambda(n, m)$ таким образом, что количество целых точек в области

$$\lambda(n, m) \leqslant T$$

равняется количеству собственных чисел, не превосходящих T. Для оценки количества целых точек в этих областях применяются результаты И. М. Виноградова и Ван дер Корпута. При исследовании кривой

$$\lambda(n, m) = T$$

применяется так называемый модельный метод теории обыкновенных дифференциальных уравнений.

Задачи на целые точки имеют связи и с интегральной геометрией. В 1948 году Кендалл опубликовал работу о попадании целых точек в случайно брошенный круг. Пусть $A(x, \alpha, \beta)$ — количество целых точек, попавших в круг с центром в точке (α, β) радиуса $\sqrt{x}$. Грубо говоря, результат Кендалла таков: для почти всех точек

$$A(x, \alpha, \beta) - \pi x = O(x^{\frac{1}{4}+\varepsilon}).$$

Результат Кендалла не допускает существенного улучшения, а именно Ландау показал, что для любого положения центра

$$\overline{\lim_{x\to\infty}} \frac{|A(x, \alpha, \beta) - \pi x|}{x^{1/4}} > c > 0.$$

Недавно А. А. Юдин, применив в этой задаче соображения теории почти периодических функций, дал новое доказательство этой теоремы Ландау.

Переходим к вопросу о распределении дробных долей функций. Аналитическая теория чисел имеет также взаимно плодотворные связи с теорией вероятностей (как с метрическим, так и со статистическим направлениями). Особенно отчетливо видны эти связи в задаче о распределении дробных долей показательной функции. Пусть $g \geqslant 2$ — фиксированное целое число, $0 < \alpha < 1$ — вещественное число; рассматривается последовательность дробных долей $\{\alpha g^x\}$, $x = 0, 1, 2, \ldots$. Представим α бесконечной g-ичной дробью

$$\alpha = \frac{a_1}{g} + \frac{a_2}{g^2} + \ldots, \tag{3}$$

$0 \leqslant a_i \leqslant g - 1$, и тем самым сопоставим α бесконечную последовательность, составленную из знаков $0, 1, \ldots, g - 1$:

$$a_1, a_2, \ldots. \tag{4}$$

Вопрос о распределении дробных долей $\{\alpha g^x\}$ может трактоваться как исследование распределения знаков и групп знаков в последовательности (4). Такой метод будем называть методом знаков.

Пусть производится неограниченное количество независимых испытаний, при каждом из которых с равной вероятностью может произойти один из g возможных исходов $0, 1, \ldots, g - 1$. Результат этой серии испытаний запишем в виде строки (4). Естественно возникает вопрос о том, какие требования нужно наложить на последовательность (4), чтобы иметь основание сказать, что в ней отражено типичное течение случайного процесса. Как известно, Мизес потребовал, чтобы в последовательности (4) все знаки $0, 1, \ldots, g - 1$ встречались бы с одинаковой частотой (равной $1/g$). Но этого требования, очевидно, мало: течение процесса

$$0, 1, \ldots, g-1, 0, 1, \ldots, g-1, 0, 1, \ldots, g-1, \ldots$$

удовлетворяет требованию Мизеса, но такое течение процесса нельзя назвать типичным. Второе условие Мизеса состоит в том, чтобы в любой a priori заданной подпоследовательности сохранялась бы частота появления знаков $0, 1, \ldots, g - 1$. Это условие нельзя считать точно сформулированным, поскольку не уточнен класс рассматриваемых подпоследовательностей. Первое, что при-

ходит на ум,—это отнести требование Мизеса лишь к подпоследовательностям, номера которых образуют арифметические прогрессии, т. е. рассматривать последовательности (4), «случайность» которых имеет, так сказать, линейный характер. Такие последовательности будем называть линейными коллективами Мизеса.

Пусть число α задано разложением (3). Легко доказать, что для того, чтобы дробные доли $\{\alpha g^x\}$, $x = 0, 1, \ldots$, были равномерно распределены, необходимо и достаточно, чтобы соответствующая последовательность (4) была линейным коллективом Мизеса. Таким образом, задача о распределении дробных долей показательной функции оказывается связанной с традиционной натурфилософской темой: анализом понятия случайной последовательности.

Метрические результаты, относящиеся к распределению дробных долей показательной функции, можно получать как следствие классических теорем теории вероятностей. Например, для трактовки теорем о равномерном распределении дробных долей показательной функции надо рассмотреть как вероятностное пространство отрезок $[0, 1]$, на этом отрезке σ-алгебру измеримых по Лебегу множеств и меру Лебега в качестве вероятности. Пусть

$$\alpha = \frac{a_1(\alpha)}{g} + \frac{a_2(\alpha)}{g^2} + \ldots .$$

Основой теории является соображение, что функции $a_i(\alpha)$, $i = 1, 2, \ldots$, являются независимыми. Поэтому в метрической теории распределения дробных долей показательной функции можно применить теорию суммирования независимых и слабо зависимых случайных величин. В качестве примера такого результата приведем теорему, принадлежащую Форте (центральная предельная теорема для дробных долей показательной функции). Пусть $f(x)$ — вещественнозначная периодическая с периодом 1 функция, удовлетворяющая определенным ограничениям на модуль непрерывности. Тогда при любом вещественном λ

$$\lim_{P\to\infty} \operatorname{mes}\left\{\alpha:\ 0\leqslant\alpha\leqslant 1,\ \sum_{x=0}^{P-1} f(\alpha g^x) - P\int_0^1 f(t)\,dt < \lambda\sqrt{P}\right\} = \frac{1}{\sqrt{2\pi}\,\sigma_f}\int_{-\infty}^{\lambda} e^{-\frac{z^2}{2\sigma_f^2}}\,dz$$

(величина σ_f определяется по функции f).

Общим методом исследования равномерного распределения дробных долей является метод тригонометрических сумм. В применении к задаче о равномерном распределении дробных долей показательной функции он состоит в изучении тригонометрических

сумм вида

$$\sum_{x=0}^{P-1} e^{2\pi i a g^x}$$

и связанной с этим вопросом задаче об оценке количества решений диофантова уравнения

$$g^{x_1} + g^{x_2} + \ldots + g^{x_k} = g^{y_1} + g^{y_2} + \ldots + g^{y_k},$$
$$0 \leqslant x_1, \ldots, x_k, y_1, \ldots, y_k \leqslant P-1,$$

и более сложных уравнений такого же типа. По-видимому, перспективна работа по сравнению метода знаков и метода тригонометрических сумм. Чтобы проиллюстрировать работу метода тригонометрических сумм в задачах с показательной функцией, приведем две теоремы, доказанные недавно Л. П. Усольцевым. Первая теорема касается оценки тригонометрической суммы с показательной функцией: пусть $g \geqslant 4$ — фиксированное целое, $P \geqslant 3$ — натуральное число, α — вещественное число, такое, что

$$\frac{1}{g^P} \leqslant |\alpha| \leqslant \frac{1}{2(g-1)}.$$

Справедливо неравенство

$$\left|\sum_{x=0}^{P-1} e^{2\pi i \alpha g^x}\right| \leqslant P\left(1 - \frac{1}{9g^{2P}}\right).$$

Отметим, что это неравенство близко к достижимой границе. Из этой оценки выводится аддитивная теорема. Пусть $g \geqslant 4$ — фиксированное натуральное число, $A_k(P)$ — количество решений диофантова уравнения

$$g^{x_1} + \ldots + g^{x_k} = g^{y_1} + \ldots + g^{y_k}$$

в целых $0 \leqslant x_1, \ldots, y_k \leqslant P-1$. Пусть $k \to \infty$, $P \to \infty$, причем $P = o(\sqrt{k})$. Тогда

$$A_k(P) = \frac{(g-1)\sqrt{g^2-1}\,P^{2k+\frac{1}{2}}}{2\sqrt{\pi k}\,g^P}\left[1 + O\left(\frac{1}{\sqrt{P}} + \frac{\sqrt{P}\ln^{\frac{3}{2}} k}{\sqrt{k}}\right)\right].$$

Теорему Форте методом тригонометрических сумм доказал М. П. Минеев. Метод тригонометрических сумм удобен тем, что он применим к исследованию распределения дробных долей матричной показательной функции $\{A^x\bar{\alpha}\}$, где A — невырожденная целочисленная матрица, $\bar{\alpha}$ — вещественный вектор, дробная доля берется покомпонентно. В 1964 году опубликовано исследование В. П. Леонова, в котором методом тригонометрических сумм

(являющимся развитием метода М. П. Минеева) доказывается обобщение упомянутой выше теоремы Форте на дробные доли матричной показательной функции. Дальнейшие исследования в этом направлении были проведены Р. Х. Мухутдиновым.

Для нового развития теории диофантовых приближений характерно проникновение эргодических методов. Под эргодической теорией мы разумеем не только метрические теоремы, но и теоремы, относящиеся к отдельным траекториям динамических систем.

Эргодическая теория позволяет объединить в одной схеме как задачи, относящиеся к дробным долям многочлена, так и задачи, относящиеся к распределению дробных долей показательной функции. Рассмотрим целочисленную невырожденную матрицу $\overline{A}$ и вектор $\overline{b}$ с вещественными компонентами. Определим на n-мерном торе преобразование

$$T\overline{\alpha} = \{A\overline{\alpha} + \overline{b}\}.$$

Легко показать, что оно сохраняет меру Лебега, так что с помощью этого преобразования можно построить динамическую систему. Мы будем изучать последовательность степеней T, т. е. последовательность векторов

$$\overline{\alpha},\ T\overline{\alpha},\ T^2\overline{\alpha},\ \ldots.$$

Очевидно, что вопрос о дробных долях матричной показательной функции содержится в этой схеме: нужно положить $\overline{b} = \overline{0}$. Пусть γ — фиксированное иррациональное число. Определим на двумерном торе преобразование

$$T(\alpha_1,\ \alpha_2) = \left(\begin{Vmatrix} 1 & 2 \\ 0 & 1 \end{Vmatrix} (\alpha_1,\ \alpha_2) + (\gamma,\ \gamma)\right).$$

Преобразование T, повторенное x раз, дает

$$T^x(\alpha_1,\ \alpha_2) = (\{\alpha_1 + 2\alpha_2 x + \gamma x^2\},\ \{\alpha_2 + \gamma x\}).$$

Мы видим, что первая координата описывается дробными долями многочлена второй степени.

Значение эргодического подхода состоит в том, что он позволил распространить методы, применяющиеся в задаче о распределении дробных долей многочлена на другие объекты: в частности, эргодическая идея была ведущей при доказательстве ряда результатов о случайных последовательностях и дробных долях показательной функции.

Методы теории чисел нашли применение в задачах приближенного анализа, в частности в вопросах о квадратурных формулах для многомерных интегралов. Эти работы связаны с именами С. Л. Соболева, Н. М. Коробова, Е. Главки. Надеемся, что на Конгрессе это направление будет освещено в достаточной мере.

Теория мультипликативных функций

Мультипликативными функциями называются функции натурального аргумента, которые удовлетворяют функциональному уравнению

$$f(n_1,\ n_2) = f(n_1)\,f(n_2)$$

для взаимно простых чисел n_1 и n_2. Эти функции возникают в вопросах, связанных с разложением чисел на простые множители. С помощью мультипликативных функций могут быть изучены многие вопросы о распределении целых точек на кривых и поверхностях, вопросы распределения простых чисел.

Мы остановимся на трех связанных между собой аспектах теории мультипликативных функций.

а) *Вероятностная теория чисел.* По-видимому, Туран впервые обратил внимание на параллелизм между доказательством неравенства Чебышева в теории вероятностей и неравенства Рамануджана, касающегося распределения значений функции $\omega(n)$ — количества различных простых делителей числа n. Этот параллелизм, который мы будем называть вероятностной интерпретацией, обнаруживается в разнообразных задачах теории чисел и служит канвой, на которой развивается большое направление аналитической теории чисел. В СССР вероятностная теория чисел представлена литовской школой И. П. Кубилюса.

За исходный пункт возьмём аналогию между анализом и теорией чисел. Точнее говоря, здесь существует целая система аналогий, и мы остановимся на одной из них. Выпишем аналогичные понятия:

Анализ	Теория чисел
Интервал	Арифметическая прогрессия $Dx+l$, $D>0$
Мера множества	Асимптотическая плотность множества натуральных чисел
Кусочно постоянные функции	Периодические с целым периодом функции натурального аргумента
Интегральная сумма с равным разбиением	Выражение вида $\frac{1}{N}\sum_{n=1}^{N} f(n)$

Функцию $f(n)$ назовем целочисленно непрерывной, если для любого $\varepsilon > 0$ найдется такое N, что при $a \equiv b \pmod N$

$$|f(a) - f(b)| \leqslant \varepsilon.$$

Легко доказать, что для целочисленно непрерывных функций существует предел

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} f(n).$$

Это аналог теоремы о существовании интеграла.

В классе целочисленно непрерывных функций, точнее, в несколько более широком классе, содержатся некоторые популярные функции теории чисел. Например, $\frac{\varphi(n)}{n}$, где $\varphi(n)$ — функция Эйлера, и $\frac{\sigma(n)}{n}$, где $\sigma(n)$ — сумма делителей числа n. Указанная выше аналогия является ведущей идеей при изучении распределения значений многих арифметических функций. Укажем в качестве примера на следующую теорему. Пусть $\nu_N\left(\frac{\varphi(m)}{m} \leqslant \lambda\right)$ означает частоту натуральных чисел m, $m \leqslant N$, для которых $\frac{\varphi(m)}{m} \leqslant \lambda$. Существует непрерывная функция распределения $v(\lambda)$, такая, что

$$\nu_N\left(\frac{\varphi(m)}{m} \leqslant \lambda\right) = v(\lambda) + O\left(\frac{1}{\ln\ln\ln N}\right).$$

В связи с метрическими теоремами о распределении дробных долей показательной функции мы говорили о «теории вероятностей» на отрезке [0, 1]. Некоторые стороны вероятностной интерпретации, о которой шла речь, мы можем трактовать как аналогичную теорию, в которой вероятностным пространством является натуральный ряд чисел, или лучше кольцо целых чисел, а в качестве меры берется асимптотическая плотность множества. Для каждого простого p определим функцию целочисленного аргумента

$$\varepsilon_p(n) = \begin{cases} 1, & \text{если } p \mid n, \\ 0, & \text{если } p \nmid n. \end{cases}$$

Функции $\varepsilon_p(n)$ и $\varepsilon_q(n)$ для разных p и q являются независимыми.

Плотность является лишь конечно аддитивной, но не счетно аддитивной функцией множеств. Например, можно разложить натуральный ряд в бесконечную сумму непересекающихся арифметических прогрессий $\{D_i x + l_i\}$ так, что $\sum \frac{1}{D_i} < \varepsilon$. Из-за этого обстоятельства при предельных переходах нельзя пользоваться общими

теоремами теории вероятностей и приходится проводить выкладки. Есть несколько способов для преодоления этой трудности. Е. В. Новоселов предложил способ, основанный на вложении кольца целых рациональных чисел в некоторое большее кольцо. Рассматривается прямое произведение

$$\mathfrak{S} = \prod_p H_p,$$

где H_p — кольцо целых p-адических чисел, и в $\mathfrak{S}$ вводится тихоновская топология. Мера Хаара на аддитивной группе $\mathfrak{S}$ есть счетно-аддитивная функция множеств. В кольце целых рациональных чисел эта мера индуцирует плотность. Метод Новоселова позволяет дать новые доказательства основных теорем вероятностной теории чисел; с помощью этого метода были установлены новые асимптотические законы теории чисел и подсчитаны остаточные члены в некоторых формулах.

Заметим, что конструкция Е. В. Новоселова родственна конструкции аделей, предложенной Шевалле (более точно, конструкция аделей содержит конструкцию Е. В. Новоселова как промежуточный этап).

б) *Метод производящих функций.* Как известно, аналитическим методом в вопросах суммирования мультипликативных функций и в вопросе о распределении простых чисел является метод производящих функций, связанный с использованием дзета-функции Римана

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s} \quad \text{при} \quad \operatorname{Re} s > 1$$

и L-функций Дирихле

$$L(s, \chi) = \sum_{n=1}^{\infty} \frac{\chi(n)}{n^s}$$

($\chi(n)$ — характер Дирихле). С помощью этого аппарата устанавливаются асимптотические формулы для величины $\pi(x)$ — количества простых чисел, не превосходящих x, для $\pi(x, D, l)$ — количества простых чисел, не превосходящих x и сравнимых с l по модулю D, а также асимптотические формулы для сумм значений мультипликативных функций

$$M(x) = \sum_{1 \leqslant n \leqslant x} f(n) \quad \text{при} \quad x \to \infty.$$

Как известно, характер $\chi(n)$ есть вполне мультипликативная функция, т. е. для любых двух натуральных чисел n и m

$$\chi(nm) = \chi(n)\chi(m).$$

Это соотношение определяет некоторые свойства L-рядов. Успехи, достигнутые с помощью аппарата теории L-рядов, побудили ряд математиков, в первую очередь Н. Г. Чудакова, исследовать вопрос о том, для каких еще вполне мультипликативных функций $h(n)$ можно построить содержательную теорию рядов вида

$$\sum_{n=1}^{\infty} \frac{h(n)}{n^s}.$$

В этой связи отметим теорему Глазкова: пусть $h(n)$ — вполне мультипликативная функция, принимающая конечное количество значений и

$$\sum_{n \leq N} h(n) = \alpha N + O(1),$$

где $\alpha \neq 0$. Тогда $h(n)$ есть главный характер Дирихле по некоторому модулю.

Основную роль в теории L-рядов играют оценки для сумм характеров

$$\sum_{N_1 \leq n \leq N_2} \chi(n).$$

Мы сейчас остановимся на описании прогресса, который достигнут в этом вопросе.

Прежде всего в вопросе об оценке сумм характеров сыграло роль доказательство гипотезы Артина — Римана для дзета-функций кривых над конечным полем (теорема А. Вейля). С помощью теоремы Вейля Берджесс получил оценки сумм характеров. Оценки Берджесса находят новые применения в аналитической теории чисел. Недавно А. И. Виноградов и Ю. В. Линник получили приложения оценок Берджесса для сумм вещественных характеров Дирихле. Известна теорема Минковского: во всяком поле алгебраических чисел есть целый идеал, норма которого не превосходит корня квадратного из дискриминанта поля. Оказывается, что для квадратичных расширений поля рациональных чисел $R(\sqrt{D})$ этот результат усиливается. Именно для любого $\varepsilon > 0$ и любого $|D| \geqslant D_0(\varepsilon)$ в поле $R(\sqrt{D})$ есть целый идеал, норма которого не превосходит $|D|^{1/4+\varepsilon}$. Оценки количества идеалов с малой нормой, полученные А. И. Виноградовым и Ю. В. Линником, имеют значение для вопроса о распределении целых точек на алгебраических поверхностях. Как известно, Ю. В. Линник доказал, что целые точки на трехмерной сфере

$$x^2 + y^2 + z^2 = N$$

распределены равномерно по телесным углам. Ученики Ю. В. Линника А. В. Малышев и Б. Ф. Скубенко развили идеи Линника; тем

самым была создана теория распределения целых точек на невырожденных поверхностях второго порядка. Это — трудная теория. Новые результаты А. И. Виноградова и Ю. В. Линника позволяют значительно упростить теорию.

Далее, сочетая оценки Берджесса с оценками Зигеля числа классов, А. И. Виноградов и Ю. В. Линник получили оценку наименьшего простого квадратичного вычета по простому модулю D $p_{\min}(D)$: при любом $\varepsilon > 0$

$$p_{\min}(D) = O\left(D^{\frac{1}{4}+\varepsilon}\right).$$

Одной из основных творческих идей в теории L-рядов Дирихле является аналогия (замеченная Ландау и Литтлвудом) между поведением дзета-функции Римана

$$\zeta(s) = \zeta(\sigma + it)$$

при росте мнимой части аргумента $|t| \to \infty$ и теоремами о поведении L-рядов, когда стремится к бесконечности модуль характера D. Теория для $|t| \to \infty$ разработана лучше, чем для $D \to \infty$; исследователи стараются теоремы t-аспекта перенести на D-аспект. «t-D» аналогия помогает и в вопросе об оценках сумм характеров в случае, когда модуль есть степень простого числа $D = p^k$. В теории дзета-функций существенную роль играют оценки сумм

$$\sum_{n=N_1}^{N_2} e^{-it \ln n}.$$

Для получения этих оценок функцию $\ln x$ приближают с помощью строки Тейлора многочленом, и вопрос сводится к оценке тригонометрической суммы типа Вейля — здесь можно применять оценки, полученные по методу И. М. Виноградова. С другой стороны, в случае модуля, равного степени простого числа, сумма характеров имеет вид

$$\sum \chi(n) = \sum e^{2\pi i \frac{\operatorname{ind} n}{p^{k-1}(p-1)}}.$$

С помощью обрыва в подходящем месте ряда для p-адического логарифма получается полиномиальное выражение для индексов, и вопрос о сумме характеров опять редуцируется к оценкам сумм Вейля. Аналогия Ландау — Литтлвуда оказывается связанной с аналогией между вещественными и p-адическими числами. Особенно ясно трактуется эта идея в одной недавней работе Н. Г. Чудакова.

Полученные оценки сумм характеров в случае модуля, равного степени простого числа, имеют арифметические следствия.

Как широко известно, Ю. В. Линник установил существование абсолютной постоянной C, такой, что для наименьшего простого

числа $q_{\min}$, принадлежащего арифметической прогрессии

$$Dx+l, \quad (D, l)=1, \quad 1 \leqslant l \leqslant D-1,$$

справедлива оценка

$$q_{\min} \ll D^C.$$

Численный подсчет постоянной C, произведенный Чен Чжин Раном, дал неравенство

$$C \leqslant 777.$$

Сочетая метод Линника с оценками сумм характера по модулю, равному степени простого числа, М. Б. Барбан, Ю. В. Линник и Н. Г. Чудаков дали для наименьшего простого числа, лежащего в прогрессии с разностью $D = p^k$, оценку

$$q_{\min} \ll D^{\frac{8}{3}+\varepsilon}.$$

Гипотеза Римана дает для всех модулей оценку

$$q_{\min} \ll D^{2+\varepsilon}.$$

В работах Ю. В. Линника показано, что теория L-рядов имеет приложения к аддитивным задачам с простыми числами; особенное значение при этом приобретают плотностные теоремы: под этим понимают теоремы, относящиеся к оценкам количества нулей L-функций, лежащих в критической полосе. Ю. В. Линник открыл, что для многих аддитивных приложений нет необходимости в теоремах, относящихся к L-рядам с индивидуальным модулем D, достаточно теорем, в которых производится осреднение по D. Это направление теории будем называть теорией L-рядов в среднем. Ю. В. Линник создал основной метод этого направления, так называемый метод большого решета. Характер полученных здесь результатов можно продемонстрировать на так называемом усредненном законе распределения простых чисел в арифметической прогрессии

$$\sum_{D \leqslant N^{\nu-\varepsilon}} \max_{l} \left| \pi(N, D, l) - \frac{1}{\varphi(D)} \int_2^N \frac{dt}{\ln t} \right| = O\left(\frac{N}{\ln^A N}\right),$$

где $\pi(N, D, l)$ — количество простых чисел, принадлежащих прогрессии $Dx + l$ и не превосходящих границы N, $\varphi(D)$ — обозначение для функции Эйлера. М. Б. Барбан получил эту теорему с $\nu = \frac{3}{8}$. А. А. Бухштаб, внеся улучшение в метод решета и используя теорему М. Б. Барбана, доказал, что всякое достаточно большое четное число представимо в виде суммы простого числа

и числа, имеющего не больше трех простых сомножителей. А. И. Виноградов и далее Э. Бомбьери довели исследование в известном смысле до конца; это означает, что даже справедливость расширенной гипотезы Римана мало что прибавит к силе результата. А. И. Виноградов получил закон с $\nu = \frac{1}{2}$, а в работе Э. Бомбьери получено, грубо говоря, снятие ε. Следует отметить, что исследование Э. Бомбьери вносит упрощение в методы. Теорема А. И. Виноградова — Э. Бомбьери позволяет с единой точки зрения решать широкий класс задач, которые требовали для своего решения разных методов. Например, получено новое решение проблемы Харди и Литтлвуда о представимости достаточно больших чисел n в виде суммы простого и двух квадратов целых чисел:

$$n = p + x^2 + y^2,$$

и некоторых других задач, которые были ранее решены дисперсионным методом Линника, и дано новое доказательство упомянутой выше теоремы Бухштаба.

в) *Теория характеров Гекке.* Плодотворным аналитическим аппаратом теории чисел являются характеры величины или характеры Гекке. Поясним о чем идет речь. Пусть мы имеем поле гауссовых чисел $R(i)$, т. е. поле чисел вида $a + bi$, где a и b — рациональные числа. Характер Гекке первого рода в данном случае имеет вид

$$\Xi_k(a+bi) = \left(\frac{a+bi}{|a+bi|}\right)^k,$$

где k — фиксированное натуральное число.

С помощью оценок тригонометрических сумм вида

$$\sum_{x^2+y^2=N} \Xi_k(x+iy)$$

Г. Бабаев получил теорему, которую, грубо говоря, можно выразить так: если на окружности много целых точек, то они приблизительно равномерно распределены по углам.

Количество целых точек, лежащих на окружности, зависит от разложения числа N на простые множители. Отметим следующее обстоятельство, которое в свое время было замечено Б. И. Сегалом: если разложение N состоит из степеней фиксированных простых чисел, то можно применить теорему А. О. Гельфонда о приближении отношения логарифмов алгебраических чисел, что ведет к значительному усилению оценок; это новые применения теории трансцендентных чисел.

Характеры Гекке находят применение и в аддитивных задачах с простыми числами. Мы уже говорили о том, что Ю. В. Линник

получил асимптотическую формулу для количества представлений числа n в виде суммы простого и двух квадратов:

$$n = p + x^2 + y^2$$

(гипотетическую формулу Харди и Литтлвуда), главный член которой

$$Q^*(n) = \frac{\pi n}{\ln n} \prod_p \left(1 + \frac{\chi_4(p)}{p(p-1)}\right) \prod_{p/n} \frac{(p-1)(p-\chi_4(p))}{p^2-p+\chi_4(p)}$$

($\chi_4(n)$ — неглавный характер по модулю 4).

Недавно Б. М. Бредихин и Ю. В. Линник разработали метод доказательства того, что решения уравнения Харди — Литтлвуда асимптотически равномерно распределены по углам; было доказано, что для числа решений уравнения Харди — Литтлвуда, в которых точка $x + iy$ лежит в предписанном угле $\Delta\varphi$, $Q(n, \Delta\varphi)$, справедлива асимптотическая формула

$$Q(n, \Delta\varphi) = \Delta\varphi Q^*(n)\left(1 + O\left(\frac{1}{(\ln\ln n)^{1/2-\varepsilon}}\right)\right).$$

Доказательство этого соотношения может быть проведено в двух редакциях: это можно сделать эргодическим методом, а можно провести доказательство с помощью характеров Гекке.

Метод характеров Гекке, как показал Г. Бабаев, распространяется и на исследование распределения решений обобщенного уравнения Харди и Литтлвуда

$$n = p + q(x, y)$$

по углам, где $q(x, y)$ — целочисленная квадратичная форма, не обязательно одноклассная.

Математический институт им. В. А. Стеклова.
Москва, СССР

ГИПЕРБОЛИЧЕСКИЕ ЗАДАЧИ ТЕОРИИ ПОВЕРХНОСТЕЙ

Н. В. ЕФИМОВ

1. Предметом доклада являются задачи о связях между внутренней метрикой и внешними свойствами поверхностей трехмерного пространства E_3. При этом имеются в виду только те случаи, когда гауссова кривизна поверхности всюду отрицательна. Если считать, что метрика известна, а сама поверхность ищется или изучается, то задачи, о которых мы будем говорить, называют также задачами о погружении данной метрики в пространство E_3. Эти задачи можно выразить с помощью системы двух квазилинейных дифференциальных уравнений. Для метрики отрицательной кривизны они приводятся к гиперболической системе. Соответственно этому мы называем их гиперболическими задачами теории поверхностей.

2. В недавний период весьма интенсивно разрабатывались эллиптические задачи, отвечающие случаю положительной кривизны. Здесь достигнуты завершающие успехи (главным образом в исследованиях А. Д. Александрова и А. В. Погорелова).

Гиперболические задачи, составляющие другую половину теории поверхностей, разработаны в гораздо меньшей степени.

Однако теперь и здесь наметились заметные сдвиги. В работах Амслера, Стокера, Оссермана, Ефимова, Позняка, Розендорна, Вернера, Шефеля и других авторов получен ряд конкретных результатов, которые, как нам кажется, заслуживают внимания. Пожалуй, даже полученный материал уже настолько велик, что систематический его обзор трудно сделать хотя бы и в часовом докладе. Поэтому мы будем говорить только о немногих вещах. Заранее оговоримся, что их выбор существенно связан с математическими интересами докладчика. На характер доклада повлияет также и наше стремление сделать его возможно более доступным.

Мы хотим еще заметить, что излагаемые далее результаты удалось получить благодаря известному продвижению в разработке общих методов, которые позволяют получать и другие выводы. Но дать представление о методах трудно, и к этому мы будем стремиться в наименьшей мере.

3. В геометрии широкое изучение тех или иных областей часто определялось некоторым конкретным вопросом, который концентрировал усилия и интерес математиков. Для эллиптических задач

таким вопросом была проблема Вейля. В гиперболическом случае аналогичная роль принадлежит проблеме обобщения теоремы Гильберта.

Как известно, Гильбертом доказано, что регулярная поверхность постоянной отрицательной кривизны не может быть полной. Иначе говоря, если неограниченно продолжать регулярный кусок поверхности с отрицательной и постоянной гауссовой кривизной, сохраняя кривизну постоянной, то мы неизбежно натолкнемся на особенность в виде острия, или ребра возврата, или края, за который поверхность дальше нельзя продолжить, или какое-либо иное нарушение регулярности. Что касается условия регулярности, то после работ Хартмана и Уинтнера оно доведено до класса C^2 (т. е. предполагается непрерывность вторых производных).

Теорема Гильберта давно привлекала к себе внимание математиков. Вместе с тем давно возникло предположение, что в теореме Гильберта условие постоянства кривизны не существенно. В самом деле, если обратиться к любым примерам полных регулярных поверхностей отрицательной гауссовой кривизны K ($K < 0$), то на каждой такой поверхности легко усматривается последовательность точек, по которой $K \to 0$. Представляется вероятным, что это обстоятельство имеет место вообще и что именно оно является существенным. Таким образом, речь идет о том, что в пространстве E_3 невозможна полная регулярная поверхность с переменной гауссовой кривизной $K \leqslant \text{const} < 0$. То же можно высказать в виде следующей общей теоремы: *в пространстве E_3 на всякой полной регулярной поверхности* $\sup K \leqslant 0$. Мы будем называть ее теоремой А. Предположение об этой теореме публиковалось Кон-Фоссеном и другими авторами. Долгое время оно не поддавалось доказательству и было основным ориентиром в работах, которым посвящен наш доклад.

Заметим, что в данном случае сложность топологической природы поверхности не играет никакой роли, поскольку всегда можно от произвольной поверхности перейти к ее универсальной накрывающей. Тем самым ясно, что вопрос достаточно решить, предполагая поверхность односвязной. Его можно формулировать также в терминах теории изометрических погружений. Будем рассматривать плоскость (x, y) и зададим на ней произвольную гауссову метрику ds^2 с кривизной $K < 0$; плоскость с этой метрикой обозначим через R_2. Предположим, что метрика ds^2 полна на плоскости, т. е. что многообразие R_2 является полным метрическим пространством. Тогда теорема А равносильна утверждению: *если $K \leqslant \text{const} < 0$, то R_2 не допускает регулярного изометрического погружения в E_3.*

Теорема Гильберта соответствует частному случаю, когда R_2 есть полная плоскость Лобачевского.

4. Коэффициенты метрики ds^2 являются функциями от (x, y). Эти функции следует считать известными. Они входят в основную систему дифференциальных уравнений задачи об изометрическом погружении в качестве определяющих параметров. Таким образом, дело сводится к системе уравнений, заданной на обычной плоскости (x, y). Как мы уже говорили, эта система, будучи квазилинейной, имеет в случае $K < 0$ гиперболический тип.

Каждое решение основной системы, регулярное в некоторой области D плоскости (x, y), определяет изометрическое погружение этой области в E_3. Одновременно по данному решению в D устанавливается сеть характеристик основной системы.

5. Если кривизна поверхности постоянна, то основная система уравнений приводится к очень простому виду в локальных координатах (u, v) сети характеристик. Но при переходе к поверхности переменной кривизны дело весьма затрудняется. В общем случае получается система вида

$$e'_v = f_1(u, v, e, g, \omega), \quad g'_u = f_2(u, v, e, g, \omega),$$
$$\omega''_{uv} = F(u, v, e, g, \omega, e'_u, \ldots, \omega'_v),$$

где e, g — метрические параметры сети характеристик, ω — сетевой угол. Правые части этой системы сложны и плохо обозримы. Делать непосредственно из такой системы какие-либо выводы практически нельзя. Поэтому естественно было прежде всего направить усилия к ее обработке. Мы считаем серьезной удачей, что для нее оказалось возможным найти вполне обозримые канонические формы записи. Существенно при этом, что для некоторых важных внешних величин, входящих в уравнения, получились внутренние оценки (т. е. оценки, зависящие только от внутренней метрики). Об этом см. Н. В. Ефимов [1], Н. В. Ефимов и Э. Г. Позняк [2], [3], Н. В. Ефимов [4], Э. Р. Розендорн [5], [6], [7], [8]. Кроме того, см. Б. Л. Рождественский [9], Э. Г. Позняк [10], где основные уравнения приведены к римановым инвариантам r, s в координатах (x, y). В римановых инвариантах они получают вид

$$r'_x + sr'_y = f(x, y, r, s),$$
$$s'_x + rs'_y = f(x, y, s, r),$$

где f — многочлен относительно r, s, коэффициенты которого являются функциями от (x, y) и определяются метрикой.

Удачная обработка основных уравнений продвинула общие методы для гиперболических задач теории поверхностей. В цитированных статьях содержится много конкретных результатов, которые получены с помощью дифференциальных уравнений. Некоторые из них мы сейчас сообщим.

6. Мы будем говорить о трех циклах теорем.

I. Теоремы гильбертова типа. Здесь речь будет идти о некоторых глобальных явлениях, связанных с медленным изменением гауссовой кривизны. Если пытаться выразить их в самых общих чертах, то можно сказать, что медленное изменение кривизны предъявляет высокие требования к структуре сети характеристик и обременительно для поверхности. Эти явления в полной мере проявляются и легко усматриваются, когда кривизна постоянна. Они проявляются также и в случае переменной, медленно изменяющейся кривизны, но лишь до известного порога, т. е. если кривизна изменяется в некотором смысле очень медленно.

Пусть для простоты изложения всюду в D будет $K \leqslant -1$. Пусть, кроме того, $|K'| \leqslant C_1 = \text{const}$, $|K''| \leqslant C_2 = \text{const}$, где производные берутся по дуге любой геодезической. Последние два неравенства сами по себе означают медленное изменение кривизны. Но мы рассматриваем некоторый класс метрик при условии, что C_1 и C_2 подчинены определенному дополнительному ограничению. Приводить его здесь мы не станем, ограничиваясь ссылкой на статью [4]; если не стремиться к общности, то можно считать, например, $C_1 = 1/2$, $C_2 = 1$. Этот класс метрик полезно как-нибудь назвать; будем говорить, например, что он состоит из метрик с весьма медленно изменяющейся кривизной.

Условимся еще называть область D с метрикой ds^2 простой зоной, если ее метрическая граница относительно ds^2 состоит не более чем из двух связных некомпактных компонент. Для наглядности укажем примеры простых зон с евклидовой метрикой: полоса между параллельными прямыми, область между двумя ветвями гиперболы, полуплоскость, наконец, вся плоскость.

Справедливы следующие теоремы:

(1) *Всякая простая зона с весьма медленно изменяющейся кривизной при любом регулярном погружении в E_3 может включать не более одной полной характеристики.*

Если кривизна изменяется недостаточно медленно, то эффект этой теоремы теряется. В самом деле, на обыкновенном геликоиде, очевидно, имеются простые зоны (даже сколь угодно узкие), содержащие бесконечно много полных характеристик (см. рис. 1).

Из теоремы (1) сразу следует, что полное многообразие с весьма медленно изменяющейся отрицательной кривизной не допускает регулярного погружения в E_3. Это утверждение установлено независимо от теоремы (1) в совместной работе Н. В. Ефимова и Э. Г. Позняка [3]. В свое время оно было первым результатом, включающим теорему Гильберта как частный случай.

Приведем еще одну теорему того же цикла, где ограничение модулей вторых производных от кривизны не предполагается (другие теоремы см. в статье [4]).

(2) *Полное многообразие R_2, кривизна которого отрицательна и удовлетворяет условию*

$$|\operatorname{grad} 1/\sqrt{-K}| \leqslant 1/\sqrt{3},$$

не может быть регулярно погружено в E_3 (Н. В. Ефимов [12]).

Заметим, что здесь предполагается только, что $K < 0$ (отгороженность от нуля не требуется). Таким образом, для всех полных и регулярных поверхностей отрицательной кривизны имеет место универсальная оценка[1])

$$\sup |\operatorname{grad} 1/\sqrt{-K}| \geqslant 1/\sqrt{3}.$$

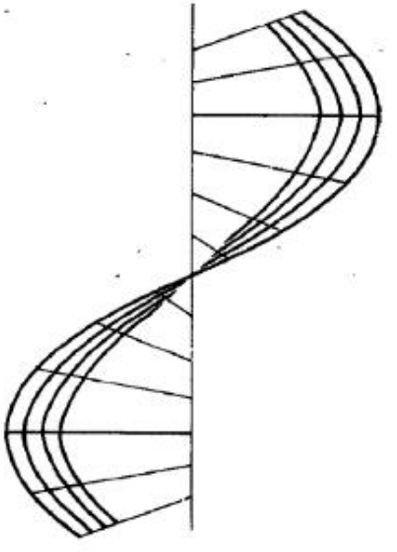


Рис. 1.

II. Теоремы существования. Уже на основании изложенных сейчас предложений видно, что *полное* многообразие R_2 с отрицательной кривизной во многих случаях *не допускает* регулярного погружения в E_3. Наряду с этим недавно Э. Г. Позняк (см. [10]) доказал ряд теорем существования, которые утверждают возможность регулярного погружения для некоторых частей такого многообразия. Приведем одну из них.

Если на R_2 всюду $K < 0$, то любая компактная часть R_2 может быть регулярно погружена в E_3.

Эта теорема получена ценою очень трудного анализа основной системы уравнений, как следствие большого числа вспомогательных теорем (некоторые из них могут быть полезными и в других вопросах).

Заметим, что для многообразия, которое, как и R_2, гомеоморфно плоскости, но имеет знакопеременную кривизну, утверждение теоремы, вообще говоря, неверно. Соответствующий пример построен Э. Г. Позняком [13]. Он очень прост по своей идее, и мы хотим его сообщить. Пусть S — поверхность с краем l, состоящая из частей S' и S'' (см. рис. 2). Здесь S' — коническая поверхность с вершиной O, регулярно продолженная в S'' через замкнутую кривую t. Полный внутренний угол S' в вершине O берется равным 2π. Поэтому внутренняя метрика поверхности S всюду регулярна. С другой стороны,

[1]) Недавно другими методами (отличными от методов статьи [12]) автор усилил этот результат и получил соотношение $\sup |\operatorname{grad} 1/\sqrt{-K}| = +\infty$. (Добавлено в корректуре.)

S' и S'' подбираются так, что любые две точки кривой t соединяются на S'' значительно более коротким путем, чем путь через точку O. Отсюда следует, что метрика поверхности не может быть регулярно погружена в E_3, так как в случае регулярного погружения на линии t нашлись бы две точки, расстояние между которыми в пространстве превосходило бы длину некоторого соединяющего эти точки пути на S''. Но это невозможно.

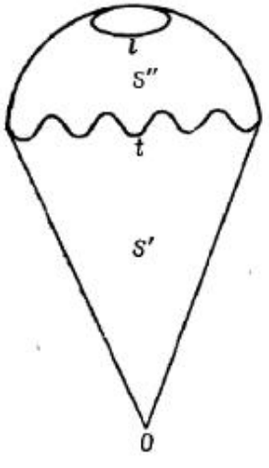


Рис. 2.

Поверхность S легко продолжить через l до полного многообразия. Тем самым строится гомеоморфное плоскости полное многообразие с регулярной метрикой, которое содержит компактную часть, не допускающую никакого регулярного погружения в E_3.

III. Теоремы о регулярности. Э. Р. Розендорном установлен ряд теорем, которые гарантируют для поверхностей отрицательной кривизны наличие определенных свойств регулярности в объемлющем пространстве E_3 в зависимости от регулярности внутренней метрики (см. [6], [7], [8], а также [11]).

Приведем для примера одну из теорем этого цикла:

Пусть S — компактный кусок поверхности отрицательной кривизны, S' — сколь угодно узкая его полоса вдоль границы (см. рис. 3). Пусть поверхность имеет регулярность класса C^2 и более высокую регулярность класса C^n $(n > 2)$ в S'. Тогда, если метрика ds^2 поверхности S принадлежит классу C^{n+1}, то сама поверхность имеет регулярность класса C^n всюду.

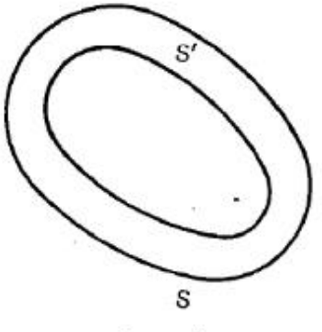


Рис. 3.

Следует обратить внимание, что теорема установлена при весьма скупых требованиях начальной регулярности — от поверхности требуется всего лишь принадлежность классу C^2. Поэтому Э. Р. Розендорну пришлось провести большую работу, прежде чем оказалось возможным применить аппарат дифференциальных уравнений. Еще более тонкими в смысле минимальности требований являются теоремы Э. Р. Розендорна об устранимости изолированных особенностей. К сожалению, ввиду недостатка времени мы не можем на них остановиться и отсылаем слушателей к цитированным выше статьям.

7. Теоремы гильбертова типа, о которых мы говорили выше, включают предложение Гильберта как частный случай. Однако нельзя считать, что они означают какое-либо приближение к теореме А, сформулированной в начале доклада. В самом деле, предметом этих теорем по существу является влияние скорости изменения кривизны на структуру сети характеристик; метод их доказательства непосредственно связан с основной системой дифференциальных уравнений. Что же касается теоремы А, то в ней обусловлена только верхняя грань кривизны, а характер локального поведения кривизны может быть каким угодно. Таким образом, здесь приходится иметь дело с явлениями совсем другой природы. Соответственно здесь потребовались и другие методы; точнее сказать, мы не видим подходов к этим вещам со стороны дифференциальных уравнений.

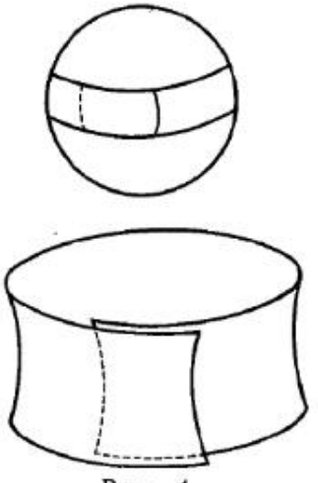


Рис. 4.

Теорема А была недавно доказана в результате некоторой специальной разработки геометрии отображений; см. Н. В. Ефимов [14], [15]. На более ранних стадиях методы отображений применялись также в других вопросах о поверхностях отрицательной кривизны; см. Н. В. Ефимов [16], [17], [18]. Указанные сейчас работы составляют отдельный цикл, мало связанный с тем, что излагалось раньше.

8. Мы постараемся хотя бы в самых общих чертах пояснить сущность дела в теореме А.

Ради наглядности обратимся к рис. 4, на котором изображена часть универсальной накрывающей однополостного гиперболоида вблизи его горловой линии. Вследствие отрицательности кривизны отображение рассматриваемой поверхности на гауссову сферу по параллельности нормалей является локально гомеоморфным. Из рисунка видно, что это отображение в целом не гомеоморфно: образ поверхности многолистно покрывает некоторую зону сферы вблизи экватора. Но такое обстоятельство можно было бы предвидеть и заранее, без всякого чертежа. В самом деле, с помощью подобного преобразования мы можем добиться для гауссовой кривизны в рассматриваемой части поверхности неравенства $K \leqslant -1$. Тогда площадь образа на сфере больше площади прообраза. А так как площадь всей поверхности бесконечна, то перекрытия ее образа на сфере неизбежны.

Совершенно такое же положение имело бы место, если бы в E_3 существовала полная поверхность с кривизной $K \leqslant \text{const} < 0$. Мы имеем в виду, что ее образ на гауссовой сфере был бы многолистным. Поэтому для доказательства теоремы достаточно установить, что в данной ситуации отображение поверхности на свой образ на сфере должно быть гомеоморфным в целом; тем самым получится нужное противоречие.

9. Дело приводится, следовательно, к общему вопросу об условиях, при которых локально гомеоморфное отображение является гомеоморфным глобально.

В цитированных статьях Н. В. Ефимова [14], [15] найдены предложения, которые приводят к дифференциальным признакам гомеоморфизмов. Мы хотим обратить внимание, что они могут применяться не только в дифференциальной геометрии. Укажем одно следствие этих предложений, лежащее вне теории поверхностей.

Пусть дано непрерывно дифференцируемое отображение всей плоскости (x, y) в плоскость (p, q) : $p = p(x, y)$, $q = q(x, y)$. Пусть $u = \{p, q\}$, $\operatorname{rot} u = p'_y - p'_x$, $\operatorname{Det} u$ — якобиан p, q по x, y. Будем предполагать, что $\operatorname{Det} u < 0$.

Имеет место теорема: *если* $|\operatorname{rot} u| \leqslant \text{const}$, $|\operatorname{Det} u| \geqslant \text{const} > 0$, *то отображение* $(x, y) \to (p, q)$ *гомеоморфно*, т. е. вся плоскость (x, y) взаимно однозначно отображается на свой образ в плоскости (p, q).

Заметим, что одного лишь сохранения знака $\operatorname{Det} u$ здесь недостаточно. Например, в случае $p = e^x \sin y$, $q = e^x \cos y$ имеем $\operatorname{rot} u = 0$, $\operatorname{Det} u < 0$; однако в этом случае образ всей плоскости (x, y) многолистен; нулевая точка $(p = 0, q = 0)$ не входит в образ и является для него точкой логарифмического ветвления.

10. Другое следствие тех же предложений относится к отображениям поверхностей на гауссову сферу и приводит к доказательству теоремы А так, как было указано выше. Теперь мы сформулируем эту теорему с точным выражением условия регулярности:

В пространстве E_3 *на всякой полной поверхности класса регулярности* C^2 *верхняя грань гауссовой кривизны не меньше нуля* ($\sup K \geqslant 0$).

Как мы видим, доказательство теоремы А получено не методами дифференциальных уравнений, а чисто геометрическими путями. Наоборот, из нее следует свойство всех решений основной системы уравнений терять регулярность.

11. В теореме А допустимые поверхности предполагаются регулярными локально, но могут иметь любые самопересечения. Попросту говоря, наличие самопересечений никак не отражается на доказательстве.

Труднее проанализировать значение других условий теоремы. Чтобы это сделать, нужно хорошо представлять себе, какими вообще бывают полные локально седловые поверхности. В связи с этим мы сообщим сейчас об интересных конструкциях Э. Р. Розендорна.

Прежде всего — по поводу условия полноты. Как в последней теореме, так и всюду раньше, когда мы говорили о полноте поверхности, мы имели в виду полноту в смысле внутренней геометрии. Заранее не ясно, в какой мере сильно ограничивает это условие внешнюю структуру поверхности, если ее кривизна не меняет знака. Заметим, что у многих геометров имелось суеверие, будто в этих случаях поверхность обязательно должна уходить в бесконечность объемлющего пространства. Э. Р. Розендорн построил в E_3 локально седловую поверхность с гауссовой кривизной $K \leqslant 0$ и с полной внутренней метрикой так, что вся эта поверхность лежит внутри данного шара; все ее бесконечно удаленные точки в смысле внутренней геометрии лежат на периферии шара (см. [19]). Заметим, что поверхность, построенная Розендорном, не имеет самопересечений. Этот пример показывает, насколько причудливыми могут быть седловые поверхности, даже при условии внутренней полноты, если $K \leqslant 0$. Вместе с тем ясно, что интуитивная убежденность в справедливости теоремы А вряд ли имела серьезные основания. Конечно, не исключено, что предположение $K < 0$ более существенно сужает класс рассматриваемых поверхностей. Но на этот счет никаких достоверных данных у нас нет.

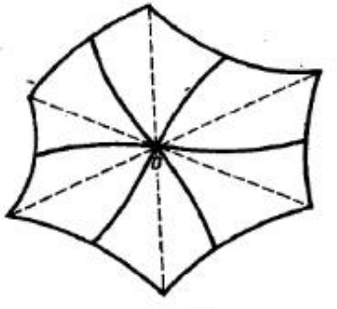


Рис. 5.

Обратимся теперь к условиям регулярности. В теореме А предполагается регулярность класса C^2. Оказалось, что в смысле обычной классификации это предположение является точным: теорема не верна уже в классе $C^{1,1}$ (т. е. при условии Липшица на первые производные). Соответствующий пример, построенный Э. Р. Розендорном (см. [7], [11]), прост по исходной идее, и нам хочется остановиться на нем подробнее.

Пусть $z = ax^2 - by^2$, $a, b > 0$,— гиперболический параболоид, где a и b подобраны с расчетом, чтобы в нулевой точке угол между прямолинейными образующими был равен $\pi/3$. Вырежем сектор между этими образующими и размножим его в шести экземплярах. Имеющиеся экземпляры можно расположить вокруг точки O так, что получится гладкая седловая поверхность S ($S \in C^1$); см. рис. 5. Порядок седла в точке O равен 2, т. е. здесь имеется так называемое обезьянье седло. С помощью специальной и довольно сложной

техники заглаживания нерегулярных седловых поверхностей Э. Р. Розендорн строит поверхность S', близкую к S ($S' \approx S$), так, что 1) $S' \in C^1$; 2) $S' \setminus O \in C^n$, где n может быть сколь угодно большим; 3) поверхность S' в точке O имеет гауссову кривизну в смысле предельного значения; 4) гауссова кривизна на S' всюду меньше нуля ($K < 0$), однако в точке O порядок седла остается

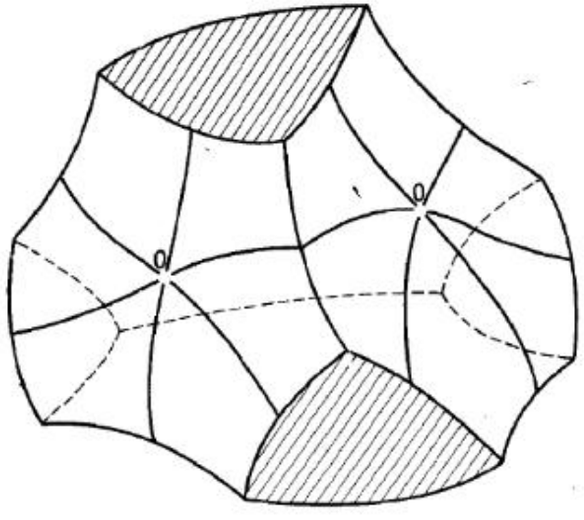


Рис. 6.

равным 2. Возьмем четыре экземпляра поверхности S'. Если краевой обрез поверхности S' сделать специальной формы, то из имеющихся экземпляров можно сложить поверхность с четырьмя отверстиями, напоминающую тетраэдр (см. рис. 6). Теперь к отверстиям можно пристроить четыре бесконечно сужающиеся трубки так, что получится полная седловая поверхность. После дополнительной закладки получается полная поверхность класса регулярности $C^{1,1}$, на которой $K \leqslant \text{const} < 0$.

Наконец, по поводу условия о размерности объемлющего пространства. Используя детали предыдущей конструкции, Э. Р. Розендорн построил в четырехмерном пространстве E_4 полную, даже замкнутую поверхность, на которой всюду $K \leqslant \text{const} < 0$. Степень регулярности этой поверхности может быть сколь угодно высокой.

Таким образом, как в отношении условий регулярности, так и в отношении условий размерности объемлющего пространства формулировка теоремы А является по существу окончательной.

В заключение мы хотим сказать, что круг задач, которых мы касались, допускает практически неограниченное и вместе с тем вполне естественное расширение. По существу речь идет об изучении седловых поверхностей вообще. К седловым поверхностям приводят многие вопросы математики. Сошлемся на работы В. П. Паламодова, где седловая поверхность встречается как естественная граница при продолжении решений некоторых систем дифференциальных уравнений, или на исследование изгибаний торообразных поверхностей, которые проводил Ниренберг. И в том и в другом случае возникли задачи о седловых поверхностях, не решенные до сих пор. Но и помимо возможных применений, седловые поверхности сами по себе заслуживают внимания. Они являются антиподами выпуклых поверхностей и составляют, вероятно, не менее интересный класс. Однако изучались они несравненно меньше, чем выпуклые поверхности. Может быть, это связано с тем, что выпуклые поверхности легко представляются наглядно, а седловые не поддаются столь непосредственному обозрению. Кстати, только что сообщенные примеры могут привести к мысли, что в разнообразии седловых поверхностей совсем нельзя разобраться. Однако такой вывод ошибочен. Дело в том, что конструкции, о которых сейчас говорилось, не получаются уже при сравнительно слабых дополнительных условиях. Иногда даже удается точно установить их невозможность. Например, невозможно осуществить розендорновскую конструкцию поверхности класса $C^{1,1}$, если только от метрики потребовать хотя бы скромной регулярности. Таким образом, здесь существуют очень строгие зависимости. Несомненно, имеется много не известных нам содержательных теорем о седловых поверхностях, о том, что здесь может быть и чего нет. Построение содержательной теории седловых поверхностей является общей геометрической проблемой. Нам хотелось бы своим докладом привлечь к ней внимание.

Московский университет,
Москва, СССР

ЛИТЕРАТУРА

[1] Ефимов Н. В., *ДАН СССР*, **136**, № 6 (1961).
[2] Ефимов Н. В., Позняк Э. Г., *ДАН СССР*, **137**, № 1 (1961).
[3] Ефимов Н. В., Позняк Э. Г., *ДАН СССР*, **137**, № 3 (1961).
[4] Ефимов Н. В., *УМН*, № 5 (131) (1966).
[5] Розендорн Э. Р., *ДАН СССР*, **145**, № 3 (1962).
[6] Розендорн Э. Р., *ДАН СССР*, **149**, № 4 (1963).
[7] Розендорн Э. Р., *УМН*, **21**, № 5 (131) (1966).
[8] Розендорн Э. Р., *Матем. сб.*, **70** (112), № 4 (1966).
[9] Рождественский Б. Л., *ДАН СССР*, 1963.

[10] Позняк Э. Г., *ДАН СССР*, **170**, № 4 (1966).
[11] Розендорн Э. Р., *Матем. сб.*, **58**, № 4 (1962).
[12] Ефимов Н. В., *ДАН СССР*, **146**, № 2 (1962).
[13] Позняк Э. Г., Вестник МГУ, 1960.
[14] Ефимов Н. В., *ДАН СССР*, **150**, № 6 (1963).
[15] Ефимов Н. В., *Матем. сб.*, 64 (106), № 2 (1964).
[16] Ефимов Н. В., *ДАН СССР*, **93**, № 3 (1953).
[17] Ефимов Н. В., *ДАН СССР*, **93**, № 4 (1953).
[18] Ефимов Н. В., *ДАН СССР*, **105**, № 4 (1955).
[19] Розендорн Э. Р., *УМН*, **16**, № 2 (98) (1961).

АНАЛИТИЧЕСКИЕ ПРОБЛЕМЫ И РЕЗУЛЬТАТЫ ТЕОРИИ ЛИНЕЙНЫХ ОПЕРАТОРОВ В ГИЛЬБЕРТОВОМ ПРОСТРАНСТВЕ

М. Г. КРЕЙН

В этом году исполняется шестьдесят лет с момента, когда в Göttingen Nachrichten появилось четвертое сообщение Давида Гильберта. В этом сообщении Гильберт подвел итог своим исследованиям по теории квадратичных форм с бесконечным числом переменных и по сути впервые получил для любого ограниченного самосопряженного оператора спектральное разложение.

Работа Гильберта определила развитие теории операторов на многие годы; это был один из тех могучих толчков, которыми современная математика обязана великому ученому.

Исследования Гильберта тотчас же привлекли внимание многих выдающихся математиков мира; вместе с тем лишь к концу 20-х годов Дж. Нейман и М. Стоун выработали точное понятие неограниченного самосопряженного оператора и в обобщение результата Гильберта получили для такого оператора спектральное разложение. Одновременно Дж. Нейман построил удивительно прозрачную теорию расширений эрмитовых операторов.

К этому моменту уже свершилось нечто непредвиденное и невероятное. Когда механики и физики прошлых времен при описании каких-либо явлений прибегали к линейным дифференциальным уравнениям, они делали это в глубоком убеждении, что истинные-то уравнения нелинейны, а их исследования являются исследованиями в первом приближении.

И вдруг новая механика — квантовая механика объявила: линейность операторов, описывающих явления микромира, есть закон природы. Более того, этими операторами оказались именно операторы, действующие в гильбертовом пространстве, — пространстве состояний микрообъекта.

Отныне к усилиям математического ума присоединилась физическая интуиция с ее смелостью и прозрением. Это не замедлило сказаться. В связи с различными вопросами квантовой статистики и теории квантованных полей в теории возмущений линейных операторов накопилось к настоящему времени множество с математической точки зрения полуфабрикатов, и трудно хотя бы приблизительно себе представить, во что переплавятся они в математическом горниле. В этой связи я укажу, например, на понятие S-матрицы Гейзенберга, которое как бы повторно рождается в математике. Это понятие обрело четкие математические контуры в теории воз-

мущения самосопряженных операторов: S-матрицу обнаруживают в классических вопросах математической физики, в вопросах, связанных с теорией стационарных случайных процессов, и самое удивительное — в различных вопросах теории несамосопряженных операторов.

Весьма значительные результаты в теории гильбертовых пространств получены в процессе встречного чисто внутреннего развития этой теории. В настоящее время мы наблюдаем, как на арене гильбертовых пространств разыгрываются события, свидетельствующие о том, что теория операторов в этих пространствах полна жизни и устремлена к новым крупным проблемам.

Я смогу рассказать здесь лишь об одной цепочке этих событий. Не знаю, удастся ли это передать, но все они принадлежат некоему связному множеству — некоему массиву, который имеет совершенно своеобразную архитектуру, свой особый аналитический аппарат и, можно даже сказать, свое особое исчисление. Сюда относятся различные вопросы теории эрмитовых, самосопряженных и несамосопряженных операторов, теории возмущений, теории рассеяния, разнообразные вопросы теории пространств с индефинитной метрикой, теория прямых и обратных спектральных задач для канонических уравнений, различные аналитические проблемы и многое другое.

1. Некоторые положения теории возмущений

Условимся относительно некоторых обозначений и терминологии. Через $\mathfrak{R} = \mathfrak{R}(\mathfrak{H})$ обозначим банахову алгебру всех линейных ограниченных операторов, действующих в сепарабельном гильбертовом пространстве $\mathfrak{H}$, а через $\mathfrak{S}_\infty$ — замкнутый идеал кольца $\mathfrak{R}$, состоящий из всех вполне непрерывных операторов.

Каждому оператору $A \in \mathfrak{S}_\infty$ соответствует последовательность s-чисел $\{s_j(A)\}_1^\infty$, которая определяется как последовательность всех собственных чисел неотрицательного оператора $|A| = (A^*A)^{1/2} \geqslant 0$, занумерованных с учетом их кратности в порядке убывания.

Через $\mathfrak{S}_p$ обозначаются идеалы Неймана — Шаттена:

$$\mathfrak{S}_p = \{A:\ A \in \mathfrak{S}_\infty,\quad \sum s_j^p(A) < \infty\}.$$

Особую роль играет идеал $\mathfrak{S}_1$ — идеал ядерных операторов. Ядерный оператор A среди других операторов из $\mathfrak{R}$ характеризуется тем, что в любом ортонормированном базисе $\{e_j\}_1^\infty$ у него существует абсолютно сходящийся матричный след, иными словами, ряд $\sum\limits_{j=1}^{\infty} (Ae_j, e_j)$ абсолютно сходится. Этот матричный след, по теореме В. Б. Лидского, совпадает со спектральным следом, т. е. с суммой собственных чисел оператора A:

$$\operatorname{sp} A = \sum (Ae_j,\ e_j) = \sum \lambda_j(A).$$

Для ядерных операторов A имеет смысл определитель

$$\det(I - A) = \lim \det \| \delta_{jk} - (Ae_j,\ e_k) \|_1^n = \prod_j (1 - \lambda_j(A)).$$

Известно, что в теории возмущений ядерные операторы выделяются среди прочих вполне непрерывных операторов своим сравнительно спокойным характером. Имеется в виду следующее: пусть H_0 и H_1 — два самосопряженных оператора, отличающиеся на ядерный $H_1 = H_0 + V$ $(V \in \mathfrak{S}_1)$. Тогда всегда абсолютно непрерывный спектр операторов H_0 и H_1 один и тот же; более того, абсолютно непрерывные части этих операторов унитарно эквивалентны. Это выяснилось в результате работ Розенблюма и Като. По поводу этих работ, как и ряда других предшествующих и последующих важных исследований М. Ш. Бирмана, Т. Като, Дж. Кука, С. Куроды, О. А. Ладыженской, К. Меллера, П. А. Рейто, Л. Д. Фаддеева, К. О. Фридрихса (основоположных), Я. М. Яуха и др. я могу лишь отослать к двум книгам по теории возмущений — к книге К. О. Фридрихса [1] и к книге Т. Като [2].

Как было установлено Г. Вейлем, Дж. Нейманом и С. Куродой, возмущениями V из идеалов более широких, чем $\mathfrak{S}_1$, можно всегда добиться того, чтобы непрерывный спектр превратился в точечный. Вместе с тем мы укажем сейчас идеал вполне непрерывных операторов, который содержит в себе все идеалы $\mathfrak{S}_p$, и все же операторы из этого идеала оказываются вполне благонадежными в других вопросах теории возмущений. Этот замечательный идеал $\mathfrak{S}_\omega$ был введен В. И. Мацаевым; его определение следующее:

$$\mathfrak{S}_\omega = \left\{A:\ A \in \mathfrak{S}_\infty,\ \sum \frac{1}{n} s_n(A) < \infty\right\}.$$

Оказывается, многие окончательные формулировки в спектральном анализе несамосопряженных операторов могут быть даны только с помощью этого идеала.

Приведем в качестве примера следующую теорему В. И. Мацаева [3а]:

Пусть оператор $H = H^*$ имеет дискретный спектр (т. е. спектр, состоящий из собственных чисел конечной кратности с единственной точкой сгущения на бесконечности). Тогда, каков бы ни был оператор $V \in \mathfrak{S}_\omega$, система корневых векторов оператора $A = H + V$ полна в $\mathfrak{H}$.

Если же вполне непрерывный оператор V не входит в $\mathfrak{S}_\omega$, то найдется самосопряженный оператор H с дискретным спектром, такой, что у оператора $A = H + V$ спектра не будет вовсе.

Первая часть этого утверждения допускает более сильную формулировку, дополняющую известную теорему М. В. Келдыша [4] о полноте.

Вообще вопросы полноты системы корневых векторов линейных операторов, как таковые, не входят в план доклада. В частности, вне его останутся многочисленные исследования (в основном советских) математиков, отправляющихся от основоположной работы М. В. Келдыша [4]. Некоторое представление об имеющихся здесь возможностях и достижениях можно получить из первой книги Гохра [6а], а также из обзорного доклада М. В. Келдыша и В. Б. Лидского [5].

При изучении спектральных свойств операторов со сложным спектром вопрос о полноте заменяется вопросом о существовании у оператора достаточно полного набора инвариантных подпространств.

Здесь также исключительная роль принадлежит операторам мацаевского класса.

Соответствующую теорему снова можно было бы сформулировать как теорему о несамосопряженных возмущениях самосопряженного оператора. Однако мы выиграем в общности и законченности, если перейдем от самосопряженных операторов к унитарным.

Пусть U — произвольный унитарный оператор, $V \in \mathfrak{S}_\omega$, и пусть весь спектр оператора $T = U + V$ находится на единичной окружности. Тогда каждой дуге единичной окружности

$$\delta_{\alpha,\beta} = \{e^{i\theta}\colon \alpha \leqslant \theta \leqslant \beta\}$$

отвечает инвариантное подпространство $\mathfrak{L}_{\alpha,\beta}$ оператора T со следующими свойствами:

1. *$\mathfrak{L}_{\alpha,\beta}$ является максимальным инвариантным подпространством оператора T, в котором спектр оператора T лежит на дуге $\delta_{\alpha,\beta}$.*
2. *Если конечная система $\{\delta\}$ открытых дуг δ покрывает единичную окружность, то алгебраическая сумма соответствующих подпространств $\{\mathfrak{L}_\delta\}$ плотна в $\mathfrak{H}$.*

Для оператора T, удовлетворяющего условиям теоремы, как показал В. И. Мацаев [3б], выполняется «билогарифмическое условие»:

$$\int_{1-\varepsilon}^{1+\varepsilon} \ln^+ \ln^+ M_T(\rho)\, d\rho < \infty \qquad (\varepsilon > 0), \tag{1}$$

где

$$M_T(\rho) = \max_{|\zeta|=\rho} \|(T-\zeta I)^{-1}\| \qquad (\rho \neq 1).$$

Условие (1), согласно исследованию Ю. И. Любича и В. И. Мацаева, уже обеспечивает указанные в теореме свойства оператора T. Из исследований этих же двух авторов следует, что билогарифмическое условие (1) эквивалентно условию

$$\sum_{n=-\infty}^{\infty} \frac{\ln^+ \|T^n\|}{1+n^2} < \infty. \tag{2}$$

Последнее условие в несколько усложненной форме впервые рассматривал Дж. Уэрмер.

Необходимо отметить, что к работе Ю. И. Любича и В. И. Мацаева [7] тесно примыкают выполненная параллельно работа Е. Бишопа, работа Ч. Фойяша, предшествующие работы Дж. Уэрмера и Ф. Вольфа, а также недавняя работа Л. де Бранжа [8а]. В работах В. И. Мацаева и де Бранжа доказывается точность этой теоремы. Для формулировки этого предложения заметим, что для обратимого оператора T условие $D_T = I - T^*T \in \mathfrak{S}_\omega$ эквивалентно возможности представления T в виде $T = U + V$, где U — унитарный оператор, а $V \in \mathfrak{S}_\omega$.

Пусть оператор $D \in \mathfrak{S}_\infty \setminus \mathfrak{S}_\omega$, тогда при некоторых ограничениях на размерность ядра оператора D можно построить обратимый оператор $T\,(\in \mathfrak{R})$, такой, что $D_T = D$, и в каждом инвариантном относительно T и T^{-1} подпространстве спектр оператора T совпадает со всей окружностью.

2. Диссипативные операторы и операторы сжатия

Всякая линеаризованная механическая система, в которой имеется диссипация энергии, описывается линейным оператором A, плотно определенным в $\mathfrak{H}$, со значениями формы (Af, f) в левой полуплоскости:

$$\operatorname{Re}(Af, f) \leqslant 0 \quad (f \in \mathfrak{D}_A).$$

В квантовой механике диссипация энергии характеризуется тем, что форма линейного оператора, описывающего физическую систему, лежит в верхней полуплоскости, т. е.

$$\operatorname{Im}(Af, f) \geqslant 0 \quad (f \in \mathfrak{D}_A).$$

Для определенности, говоря о диссипативных операторах, мы будем иметь в виду операторы последнего типа, т. е. диссипативные операторы квантовой механики.

Диссипативный оператор называется *максимальным*, если его нельзя расширить с сохранением свойства диссипативности. По теореме Ральфа Филлипса, для того чтобы диссипативный оператор был максимальным, достаточно, чтобы он имел хотя бы одну регулярную точку внутри нижней полуплоскости, и необходимо, чтобы все внутренние точки этой полуплоскости были регулярными.

Р. Филлипс показал также, что всякий диссипативный оператор допускает расширение до максимального.

Как и в теории Неймана расширения эрмитовых операторов, в исследованиях Р. Филлипса существенную роль играет преобразование Кэли

$$T=(A-iI)(A+iI)^{-1}.$$

Если оператор A является максимальным диссипативным, то в этом и только этом случае преобразование Кэли будет давать оператор, определенный на всем $\mathfrak{H}$ с нормой $\|T\|\leqslant 1$. Такие операторы именуются *сжатиями*. Как правило, всякий вопрос, касающийся максимальных диссипативных операторов, можно переформулировать как некоторый вопрос, относящийся к сжатиям.

Мы преимущественно будем говорить о сжатиях, поскольку понятие оператора сжатия является более элементарным.

Сжатие называется простым, если ни на каком из своих инвариантных подпространств оно не индуцирует унитарного оператора. По теореме Г. К. Лангера — Б. С. Надя — Ч. Фойяша, всякое сжатие распадается в ортогональную сумму унитарного оператора и простого сжатия.

В наших целях достаточно ограничиться только простыми сжатиями и соответственно только простыми диссипативными операторами, которые определяются аналогично.

Сжатие T будем называть *слабым*, если выполняются два условия:

а) оператор T обратим: $T^{-1}\in\mathfrak{R}$;

б) оператор отклонения D_T $(=I-T^*T)$ оператора T от унитарного является ядерным.

Условие а) можно было бы заменить более общим, а именно: у оператора T существует хотя бы одна регулярная точка внутри единичного круга.

Хотя условие б) является весьма стеснительным, все же при переходе от слабых сжатий к соответствующим диссипативным операторам получается класс операторов, содержащий, в частности, операторы, порождаемые радиальным уравнением Шредингера с комплексным потенциалом достаточного общего типа.

В исследованиях по теории линейных операторов общего типа задача о построении континуального аналога теории элементарных делителей, подобно синей птице, всегда оставалось неуловимой. Именно в поисках этой птицы И. М. Гельфанд пришел к теории коммутативных нормированных колец — одному из самых красивых творений современного анализа.

Эти же стремления привели Н. Данфорда, Дж. Шварца и их последователей к развитию известной теории спектральных операторов.

Однако после всего до недавнего времени трудно было указать сколько-нибудь общий класс операторов, для которого был бы построен некоторый аналог теории элементарных делителей. В настоящее время можно уже говорить о некоторых достижениях в этом направлении. А если речь идет о слабых сжатиях, то со всей категоричностью можно утверждать, что здесь мы уже достигли Мыса Доброй Надежды.

Прежде всего укажем, что для слабых сжатий можно построить некоторый эрзац векового определителя, а именно

$$d_T(\zeta)=\det\left(T^*(T-\zeta I)(I-\zeta T^*)^{-1}\right).$$

Легко показывается, что под знаком определителя стоит выражение типа $I+A$, где $A\in\mathfrak{S}_1$. Этот определитель допускает следующее представление:

$$d_T(\zeta)=\sqrt{\det(T^*T)}\,B_T(\zeta)\exp\left\{-\int_0^{2\pi}\frac{e^{it}+\zeta}{e^{it}-\zeta}\,d\omega(t)\right\}, \qquad (3)$$

где $B_T(\zeta)$ — произведение Бляшке:

$$B_T(\zeta)=\prod_j\frac{\lambda_j-\zeta}{1-\bar{\lambda}_j\zeta}\cdot\frac{|\lambda_j|}{\lambda_j},$$

а $\omega(t)$ — неубывающая функция, нормированная условием $\omega(0)=\omega(+0)=0$.

Все элементы в этом представлении допускают операторно-спектральное истолкование.

В частности, числа $\{\lambda_j\}$ суть собственные числа сжатия T, причем каждое из них фигурирует столько раз, какова его алгебраическая кратность (т. е. какова размерность соответствующего корневого подпространства).

Если в равенстве (3) положить $\zeta=0$, то получим

$$\sqrt{\det(T^*T)}=\prod|\lambda_j|\exp\left\{-\int_0^{2\pi}d\omega(t)\right\}.$$

Имеет место следующая группа утверждений:

I. *Пространство $\mathfrak{H}$ единственным образом распадается в квазипрямую сумму двух инвариантных относительно T и T^{-1} подпространств $\mathfrak{E}$ и $\mathfrak{H}_0$:*

$$\mathfrak{H}=\mathfrak{E}\dot{+}\mathfrak{H}_0 \quad (\mathfrak{E}\cap\mathfrak{H}_0=\{0\};\ \overline{\mathfrak{E}+\mathfrak{H}_0}=\mathfrak{H}),$$

где $\mathfrak{E}$ — замкнутая линейная оболочка всех корневых подпространств оператора T. Спектр оператора T в $\mathfrak{H}_0$ лежит на единичной окружности.

Система корневых векторов T будет полной, т. е. $\mathfrak{H} = \mathfrak{E}$, в том и только том случае, когда $d_T(\zeta) = B_T(\zeta)$, т. е. $\det(T^*T) = \prod |\lambda_j|^2$. Таким образом, следующие три утверждения эквивалентны:

$$\mathfrak{H} = \mathfrak{E} \iff d_T(\zeta) = B_T(\zeta) \iff \det(T^*T) = \prod |\lambda_j|^2.$$

Обозначим через T_0 сужение сжатия T на $\mathfrak{H}_0$. Пусть $\mathfrak{L}_t$ — максимальное инвариантное подпространство оператора T_0, в котором его спектр лежит на дуге $e^{i\theta}$ $(0 \leqslant \theta \leqslant t)$. Тогда

$$\omega(t+0) = -\frac{1}{2}\ln\det(I + P_t D_T P_t), \tag{4}$$

где P_t — ортопроектор, проектирующий $\mathfrak{H}$ на $\mathfrak{L}_T$, а $D_T = I - T^*T$.

Оператор $A \in \mathfrak{R}$ называется *одноклеточным*, если множество всех его инвариантных подпространств упорядочено по вложению. Для оператора в n-мерном пространстве одноклеточность означает, что оператор состоит из одной жордановой клетки.

Если у одноклеточного оператора T $(\in \mathfrak{R})$ оператор отклонения D_T принадлежит мацаевскому классу, то, как легко заключить из предыдущего, спектр оператора T состоит из одной точки, лежащей на единичной окружности.

У одноклеточного оператора в n-мерном пространстве спектр также состоит из одной точки. Но это еще его не характеризует полностью. Одна из полных характеристик такого оператора заключается в том, что его резольвента имеет полюс n-го (максимального возможного) порядка. Оказывается, что среди слабых сжатий одноклеточные также вполне характеризуются наивысшим порядком роста резольвенты. Точно теорема формулируется следующим образом:

II. *Для того чтобы простое слабое сжатие T было одноклеточным, необходимо и достаточно, чтобы* 1) *оператор T не имел собственных чисел внутри единичного круга и* 2) *выполнялось равенство*

$$\overline{\lim_{\rho \uparrow 1}}\{(1-\rho)\ln M_T(\rho)\} = -\frac{1}{2}\ln\det(T^*T), \tag{5}$$

где, как и выше,

$$M_T(\rho) = \max_{|\zeta|=\rho} \|(T-\zeta I)^{-1}\| \qquad (\rho \neq 1).$$

Если выполняется первое условие, но оператор T не является одноклеточным, то в (5) *имеет место знак $<$.*

В таком виде теорема приведена во второй книге Гокра [6б]. Она представляет собой некоторую переработку с уточнением критерия Бродского — Кисилевского одноклеточности вольтеррова оператора. Оператор A называется *вольтерровым*, если он вполне непрерывен и весь его спектр сосредоточен в нуле.

Не прибегая пока к понятию характеристической функции, теорему Бродского — Кисилевского [9] можно сформулировать следующим образом:

II′. *Пусть A — вольтерров диссипативный оператор с ядерной мнимой компонентой, т. е.*

$$A \in \mathfrak{S}_\infty,\ \sigma(A) = \{0\},\ A_J = \frac{1}{2i}(A - A^*) \geqslant 0,\ \operatorname{sp} A_J < \infty.$$

Тогда оператор A одноклеточен в том и только том случае, когда

$$\lim_{r \downarrow 0} r \ln\|(A - irI)^{-1}\| = 2\operatorname{sp} A_J.$$

В линейной алгебре линейный оператор A одноклеточен в том и только том случае, когда он имеет единственную точку спектра и цикличен, т. е. существует такой вектор f, что система $f, Af, \ldots, A^{n-1}f$ составляет базис пространства. Легко показывается, что всякий ограниченный одноклеточный оператор является циклическим, т. е. существует такой вектор f, что

$$\bigvee_{n=1}^{\infty} A^n f = \mathfrak{H}.$$

Однако в бесконечномерном гильбертовом пространстве, даже если ограничиться вольтерровыми операторами, цикличность оператора, вообще говоря, не влечет его одноклеточности. Вместе с тем, как недавно показал Г. Э. Кисилевский, всякий циклический простой диссипативный вольтерров оператор с ядерной мнимой компонентой одноклеточен.

Всякий линейный оператор, действующий в конечномерном пространстве, либо одноклеточен, либо распадается в прямую сумму одноклеточных. Г. Э. Кисилевский [10] распространил это предложение на диссипативные вольтерровы операторы, действующие в гильбертовом пространстве.

III. *Всякий простой вольтерров диссипативный оператор с ядерной мнимой компонентой либо одноклеточен, либо имеет место следующее:*

а) *пространство $\mathfrak{H}$ распадается в квазипрямую сумму подпространств $\mathfrak{L}_k$ $(k = 1, 2, \ldots, \omega \leqslant \infty)$, инвариантных относительно A:*

$$\mathfrak{H} = \Sigma \dot{+} \mathfrak{L}_k, \quad A\mathfrak{L}_k \subseteq \mathfrak{L}_k;$$

б) *сужение оператора A на $\mathfrak{L}_k$ $(k = 1, 2, \ldots)$ является одноклеточным оператором.*

Кардинальное число ω и положительные числа

$$\tau_k = \operatorname{sp}(A \mid \mathfrak{L}_k)$$

суть инварианты оператора A.

Таким образом, для простых диссипативных вольтерровых операторов с ядерной мнимой компонентой установлены аналоги многих основных предложений линейной алгебры. Однако следует особо подчеркнуть одно поразительное обстоятельство. Все эти аналоги установлены для операторов, для которых *нет никакого аналога* в линейной алгебре. Действительно, если в линейной алгебре $\sigma(A) = \{0\}$, то

$$\operatorname{sp} A = \operatorname{sp} A_R + i \operatorname{sp} A_J = 0,$$

следовательно, $\operatorname{sp} A_J^2 = 0$, а так как $A_J \geqslant 0$, то $A_J = 0$. Таким образом, оператор A самосопряжен, т. е. в конечномерном пространстве не существует простых диссипативных вольтерровых операторов.

Аналогичным образом показывается, что в линейной алгебре абсурдно понятие простого сжатия со спектром на единичной окружности.

Вместе с тем наиболее замечательные факты в теории сжатий относятся к простым сжатиям со спектром на единичной окружности.

Мы не исчерпали всех результатов, полученных для одноклеточных операторов. Отметим, что родственные, иногда эквивалентные, иногда дополняющие результаты для сжатий были недавно получены Б.С.-Надем и Ч. Фойяшем в их замечательной серии работ по сжатиям [1]).

В алгебраическом случае при исследовании спектральных свойств оператора невозможно обойтись одним вековым определителем. Тем более невозможно обойтись его суррогатом в бесконечномерном случае. В предыдущих утверждениях нам неоднократно приходилось обращаться к резольвенте оператора. Оказывается, имеется объект, который определяется значительно сложнее, чем резольвента, но который хранит в себе ключи к решению многих вопросов теории несамосопряженных операторов. В частности, значительная часть предыдущих утверждений была получена с помощью этого объекта.

Таким объектом является *характеристическая функция* несамосопряженного оператора. В настоящее время имеются разнообразные определения характеристической функции — этого спутника несамосопряженного оператора. Можно различать два типа характеристических функций. Функции первого типа характеризуют отклонение оператора от унитарного, а второго типа — от самосопряженного. Характеристическая функция $\theta_T(\zeta)$ первого типа

[1]) Эта серия публикуется в журнале *Acta Math. Szeged* начиная с 1953 г. и в настоящее время насчитывает 13 работ; в дальнейшем цитируются только некоторые из этих работ. (См. примечание, добавленное при корректуре, на стр. 209.)

оператора T определяется равенством

$$\theta_T(\zeta) = [\,|I - TT^*|^{-1/2}(T - \zeta I)(I - \zeta T^*)^{-1}|I - T^*T|^{1/2}\,]\,|\overline{D_T\mathfrak{H}},$$

где через $|A|$ обозначается операторный модуль самосопряженного оператора A, т. е. неотрицательный оператор, квадрат которого равен A^2, а через $A \mid \mathfrak{L}$ — сужение оператора A на подпространство $\mathfrak{L}$.

При каждом ζ, для которого $(I - \zeta T^*)^{-1} \in \mathfrak{R}$, значение $\theta_T(\zeta)$ является линейным ограниченным оператором, действующим из подпространства $\overline{D_T\mathfrak{H}}$ в подпространство $\overline{D_{T^*}\mathfrak{H}}$.

При весьма общих условиях характеристическая оператор-функция задает простой оператор с точностью до унитарной эквивалентности.

Впервые эта функция была введена в докторской диссертации М. С. Лившица в 1944 г. для случая, когда

$$\dim D_T\mathfrak{H} = 1.$$

В этом случае $\theta_T(\zeta)$ является, в сущности, скалярной функцией, а сам оператор T будет либо растяжением, либо сжатием. Вообще в случае, когда оператор отклонения D_T конечномерен, функцию $\theta_T(\zeta)$ можно рассматривать как матрицу-функцию. Для этого случая она по существу была получена в 1950 году М. С. Лившицем и В. П. Потаповым [13]. К сожалению, недостаток времени не позволяет изложить эволюцию этого важнейшего понятия — эволюцию, которая продолжалась в течение двадцати лет под влиянием и при участии автора понятия и которая продолжается в настоящее время.

В частности, я не имею возможности остановиться на характеристической функции второго типа, также впервые введенной М. С. Лившицем и получившей существенное развитие в работах М. С. Бродского [14, 15a].

С совершенно новых позиций к характеристической функции для сжатий пришли в 1963 г. Б. С.-Надь и Ч. Фойяш [11a]. В работах этих авторов она впервые изучалась без предположения полной непрерывности операторов отклонения D_T и D_{T^*}.

В случае сжатия T определение функции $\theta_T(\zeta)$ упрощается:

$$\theta_T(\zeta) = [\,T - \zeta D_{T^*}^{1/2}(I - \zeta T^*)^{-1}D_T^{1/2}\,]\,|\overline{D_T\mathfrak{H}}.$$

В этом случае оператор-функция $\theta_T(\zeta)$ голоморфна внутри единичного круга, причем

$$\|\theta_T(\zeta)\| \leqslant 1 \qquad (|\zeta| < 1).$$

Умножая произвольный линейный ограниченный оператор на достаточно малую константу, можно превратить его в сжатие со сколь угодно малой нормой. Затем для полученного сжатия можно

составить характеристическую функцию. Однако вряд ли можно рассчитывать на то, что характеристическая функция доставит в этом случае какую-либо новую ценную информацию. По-видимому, эта функция оказывается полезной, лишь когда оператор в каком-то смысле близок к унитарному. До недавнего времени в работах советских математиков мерилом такой близости служило условие полной непрерывности операторов отклонения D_T и D_{T^*}. Но возможна близость другого рода, а именно подобие оператора T унитарному, означающее существование ограниченного и обратимого S ($S^{-1} \in \mathfrak{R}$), такого, что

$$T = S^{-1}US, \tag{6}$$

где U — унитарный оператор. Сравнительно давно было обнаружено, что простое сжатие может быть подобно унитарному оператору. В алгебраическом случае всякое сжатие, подобное унитарному оператору, само является унитарным оператором. Большое достижение теории Надя — Фойяша составляет следующий общий критерий подобия сжатия унитарному оператору.

IV. *Сжатие T подобно унитарному оператору в том и только том случае, когда при любом ζ ($|\zeta| < 1$) оператор $\theta_T(\zeta)$ взаимно однозначно отображает подпространство $\overline{D_T\mathfrak{H}}$ на $\overline{D_{T^*}\mathfrak{H}}$ и*

$$\sup_{|\zeta|<1} \| \theta_T^{-1}(\zeta) \| < \infty.$$

Легко видеть, что если некоторый оператор T (не обязательно сжатие) подобен унитарному, то он обладает следующими двумя свойствами:

$$\alpha) \quad \sigma(T) \subset \{e^{i\varphi}: 0 \leqslant \varphi \leqslant 2\pi\}$$

и

$$\beta) \quad \sup_{|\zeta|\neq 1} (1 - |\zeta|) \| (T - \zeta I)^{-1} \| < \infty.$$

Как следствие критерия Надя — Фойяша получается, что в случае сжатия *T эти условия являются не только необходимыми, но и достаточными для подобия T унитарному оператору*; см. [6г].

Как заметил А. С. Маркус, можно построить операторы с наперед заданным спектром на единичной окружности, удовлетворяющие условиям α) и β) и не подобные никакому унитарному. Разумеется, всякий такой оператор уже не будет сжатием.

Критерий Надя — Фойяша для слабых сжатий дает следующий достаточный признак:

Слабое сжатие T со спектром на единичной окружности подобно унитарному оператору, коль скоро его функция спектрального сдвига $\omega(t)$ принадлежит классу $\mathrm{Lip}_1(0, 2\pi)$.

Если операторы отклонения конечномерны, то этот признак является также достаточным. Можно указать и другой случай, когда этот признак является достаточным. Для его получения требуется точный анализ структуры инвариантных подпространств оператора.

Если сжатие T подобно унитарному оператору, то

$$T = \int_0^{2\pi} e^{it}\, dF(t),$$

где $F(t)$ — некоторое *косое* разложение единицы, нормированное, например, условием $F(t-0) = F(t)$ $(0 < t \leqslant 2\pi)$, $F(0) = 0$, $F(2\pi) = I$). Обозначим через $P(t)$ ортопроектор, проектирующий все пространство $\mathfrak{H}$ на подпространство $F(t)\mathfrak{H}$ $(0 \leqslant t \leqslant 2\pi)$. Разложение единицы $P(t)$ $(0 \leqslant t \leqslant 2\pi)$ называется *спрямленной* спектральной функцией оператора T. Имеет место следующая теорема.

Пусть заданы неотрицательный оператор H с $\|H\| < 1$ и некоторое ортогональное разложение единицы $P(t)$ $(0 \leqslant t \leqslant 2\pi)$.

Для того чтобы существовало сжатие T, подобное унитарному оператору, с отклонением $D_T = H$ и со спрямленной спектральной функцией $P(t)$ $(0 \leqslant t \leqslant 2\pi)$, необходимо и достаточно, чтобы оператор-функция $H^{1/2}P(t)H^{1/2}$ удовлетворяла условию Липшица

$$\| H^{1/2}(P(t) - P(s))H^{1/2} \| \leqslant \mathrm{const}\, |t - s|.$$

При выполнении этого условия оператор $T = U(I - V)^{-1}$, где

$$U = \int_0^{2\pi} e^{it}\, dP(t)$$

и

$$V = \int_0^{2\pi} (I - P(t)HP(t))^{-1} P(t) H\, dP(t), \tag{7}$$

является единственным сжатием со свойствами:

1) *спектр T лежит на единичной окружности,*

2) *$P(t+0)\mathfrak{H}$ $(0 < t < 2\pi)$ является максимальным инвариантным подпространством оператора T, в котором его спектр лежит на дуге $e^{i\varphi}$ $(0 \leqslant \varphi \leqslant t)$.*

Здесь опускается формулировка аналогичной теоремы, дающей описание диссипативных операторов, подобных самосопряженным.

В качестве примера рассмотрим в пространстве n-мерных вектор-функций $L_2^{(n)}(0,1)$ оператор сжатия следующего вида:

$$(Tf)(t) = e^{i\alpha(t)} f(t) + \int_t^1 K(t, s) f(s)\, ds,$$

где $K(t, s)$ $(0 \leqslant t,\ s \leqslant 1)$ — (для простоты) непрерывное эрмитово матричное ядро, порождающее ядерный оператор. Тогда функция $\omega(t)$ вычисляется по формуле

$$\omega(t) = \int\limits_{0 \leqslant \alpha(s) \leqslant t} \operatorname{sp} K(s, s)\, ds.$$

Принадлежность функции $\omega(t)$ классу Lip_1 является необходимым и достаточным условием подобия оператора T унитарному.

Это же условие $(\omega \in \mathrm{Lip}_1)$ для диссипативного оператора

$$(Af)(t) = \alpha(t) f(t) + 2i \int\limits_t^1 K(t, s) f(s)\, ds \quad (f \in L_2(0, 1)),$$

где $K(t, s)$ — произвольное непрерывное эрмитово положительное ядро, является необходимым и достаточным для подобия оператора A самосопряженному оператору.

Последний результат для случая $n = 1$ и $K(t, s) \equiv 1$ был получен ранее в статье [11в]. При $\alpha(t) = t$ оператор A всегда будет подобен самосопряженному (см. [16 б]).

3. Треугольное представление операторов и мультипликативные представления оператор-функций

Сформулированная выше теорема, дающая полное описание сжатий, подобных унитарным, была получена Гокром [6 г] с использованием упоминавшихся результатов Надя — Фойяша, а также новых средств, связанных с теорией абстрактных треугольных и мультипликативных интегралов. В самой формулировке теоремы уже фигурирует треугольный интеграл (правая часть (7)).

Приведем ряд определений. Замкнутое (в смысле сильной сходимости) множество ортогональных проекторов $\mathfrak{P} = \{P\}$ называется *цепочкой*, если оно упорядочено (естественным образом) и содержит проекторы 0 и I. Пара проекторов (P^-, P^+) $(P^- < P^+;\ P^\pm \in \mathfrak{P})$ называется *разрывом* цепочки $\mathfrak{P}$, если в $\mathfrak{P}$ нет ни одного проектора, расположенного между P^- и P^+. Цепочка, не имеющая ни одного разрыва, называется *непрерывной*.

Цепочка $\mathfrak{P}$ называется *максимальной*, если она не является правильной частью никакой другой цепочки. Цепочка $\mathfrak{P}$ называется *собственной* цепочкой данного оператора A $(\in \mathfrak{R})$, если все подпространства $P\mathfrak{H}$ инвариантны относительно A, т. е. если $AP = PAP$ $(P \in \mathfrak{P})$.

Согласно теореме Гильберта, всякий линейный самосопряженный ограниченный оператор H допускает представление

$$H = \int\limits_a^b \lambda\, dE_\lambda,$$

где E_λ — разложение единицы, порождаемое оператором H. Пусть $\mathfrak{P}$ — произвольная собственная цепочка оператора H, содержащая цепочку $\{E_\lambda\}_{a<\lambda<b}$. Тогда последний интеграл можно переписать в виде

$$H = \int\limits_{\mathfrak{P}} \lambda(P)\, dP = \int\limits_{\mathfrak{P}} \lambda(P)\, P\, dP,$$

где $\lambda(P)$ — верхняя грань спектра оператора H в инвариантном подпространстве $P\mathfrak{H}$. Этот интеграл можно было бы назвать *диагональным интегралом* вдоль цепочки. Он является частным случаем *треугольного* интеграла вдоль цепочки, имеющего вид

$$J = \int\limits_{\mathfrak{P}} F(P)\, dP, \tag{8}$$

где $F(P)$ — функция на цепочке с операторными значениями.

Интеграл (8) понимается в смысле сходимости по норме операторов соответствующих частных сумм, причем предел понимается по направленному множеству всех разбиений цепочки $\mathfrak{P}$.

Если выполнено условие

$$F(P) = PF(P),$$

то интеграл (8) называется *треугольным*. В этом случае цепочка $\mathfrak{P}$ будет собственной для оператора J.

В конечномерном случае треугольный интеграл в базисе, расширяющемся вместе с цепочкой $\mathfrak{P}$, будет изображаться треугольной матрицей.

Простейшим треугольным интегралом, естественно, является интеграл

$$\int\limits_{\mathfrak{P}} PC\, dP, \tag{9}$$

где C — вполне непрерывный оператор. Этот интеграл впервые появился в 1958 г. в работе М. С. Бродского в результате некоторой переработки треугольной модели несамосопряженного оператора, предложенной М. С. Лившицем. М. С. Бродский показал, что

для всякого вольтеррова оператора A имеет место представление

$$A = \int_{\mathfrak{P}} PC\,dP,$$

где $\mathfrak{P}$ — какая-либо собственная максимальная цепочка оператора A, а $C = 2iA_I$ (существование такой цепочки вытекает из результатов Неймана — Ароншайна — Смита [17]).

Этим представлением, однако, не решался следующий вопрос: пусть наперед задан оператор C и цепочка $\mathfrak{P}$; при каких условиях существует интеграл (9), а если он существует, то какому классу операторов принадлежит этот интеграл, когда оператор C принадлежит тому или иному классу? В связи с этим вопросом были обнаружены весьма неожиданные связи между спектрами эрмитовых компонент вольтерровых операторов. Мацаевский класс операторов оказался и здесь исключительным по своей роли. Выяснилось, что интеграл (9) сходится вдоль любой непрерывной цепочки $\mathfrak{P}$ в том и только том случае, когда оператор $C \in \mathfrak{S}_\infty$. Самая трудная часть этой теоремы — достаточность — была доказана В. И. Мацаевым.

Теорию треугольного интеграла (9), вместе с ее приложениями к самосопряженным краевым задачам для канонических систем дифференциальных уравнений, можно найти во второй книге Гокра [6б]. Там же можно найти приложения теории этого интеграла в вопросах устойчивости решений канонических систем дифференциальных уравнений с периодическим гамильтонианом.

Мы не имеем возможности остановиться здесь на более общих треугольных представлениях операторов — представлениях, содержащих еще и диагональную часть операторов.

В основе теории треугольного интеграла лежит идея существования неких континуальных аналогов элементарной теоремы И. Шура о возможности приведения любой матрицы к треугольному виду с помощью унитарного преобразования.

Лагранж за целое столетие до Шура предложил метод приведения квадратичной формы к сумме квадратов. На языке теории матриц этот метод давал, в частности, решение задачи о представлении положительно определенной матрицы в виде

$$G = A^*A, \tag{10}$$

где A — треугольная матрица. Континуальный аналог этой задачи в общем виде был рассмотрен Гокром [6].

В операторном случае задача формулируется следующим образом: пусть, для простоты, G — положительно определенный оператор, а $\mathfrak{P}$ — некоторая цепочка; требуется представить оператор G

в виде (10), где A ($\in \mathfrak{R}$) — некоторый оператор с собственной цепочкой $\mathfrak{P}$: $PAP = AP$ ($P \in \mathfrak{P}$). Будем называть эту задачу задачей о факторизации оператора G вдоль цепочки $\mathfrak{P}$.

Если эта задача допускает решение, то оператор A определяется, очевидно, с точностью до правого унитарного множителя, коммутирующего с проекторами $P \in \mathfrak{P}$. Однако если оператор G имеет вид $G = I + H$, где $H \in \mathfrak{S}_\infty$, а множитель A разыскивается в таком же виде $A = I + X$, $X \in \mathfrak{S}_\infty$, то в факторизации (10), если она существует, оператор A определяется единственным образом.

Оператор A выражается через оператор H и цепочку $\mathfrak{P}$ треугольным интегралом

$$(A-I)^{-1} = I + \int_{\mathfrak{P}} (I-PHP)^{-1}\,PH\,dP \tag{11}$$

или мультипликативным [1]) интегралом

$$A = \overleftarrow{\int_{\mathfrak{P}}} \exp\left(dPHP\,(I-PHP)^{-1}\right).$$

Оказывается, что принадлежность H мацаевскому классу обеспечивает сходимость рассматриваемых интегралов вдоль любой непрерывной цепочки $\mathfrak{P}$. Если, однако, в случае простейшего треугольного интеграла (9) Гокру удалось показать, что условие $C \in \mathfrak{S}_\infty$ является также необходимым, то здесь аналогичный вопрос остается открытым.

Следует подчеркнуть, что для фиксированной непрерывной цепочки $\mathfrak{P}$ интеграл (9) (или (11)) может сходиться даже и в том случае, когда оператор C (соответственно H) не вполне непрерывен. Это будет, например, иметь место, если цепочка $\mathfrak{P}$ является достаточно гладкой по отношению к оператору $C(H)$. Именно с таким случаем мы и встретились в интеграле (7).

Не вдаваясь ни в какие детали, укажем лишь, что та факторизация, которая встречается в теории уравнений типа Винера — Хоп-

---

[1]) Отметим вообще, что при широких условиях треугольный интеграл

$$J = \int_{\mathfrak{P}} F(P)\,dP$$

связан с мультипликативным следующим равенством:

$$(J+I)^{-1} = \overleftarrow{\int_{\mathfrak{P}}} \exp\left(dPF(P)\right).$$

Последний интеграл понимается как предел по норме соответствующих частных произведений по направленному множеству разбиений цепочки.

фа, укладывается в приведенную общую схему абстрактной задачи факторизации вдоль цепочки.

В исследованиях Гокра задача факторизации не была самоцелью. Она возникла как некий промежуточный этап, необходимый для построения теории треугольных представлений операторов, близких к унитарным. Решение этой задачи сыграло важную роль при получении мультипликативного представления характеристической оператор-функции операторов, близких к унитарным. Ограничимся, для простоты, случаем сжатия T без спектра внутри единичного круга. Такое представление можно записать в следующем виде:

$$T^*\theta_T(\zeta) = |T| \overleftarrow{\int_{\mathfrak{P}}} \exp\left(-\frac{e^{i\varphi(P)}-\zeta}{e^{i\varphi(P)}+\zeta}\right) \mathfrak{R}^*(P)\, dP\, \mathfrak{R}(P), \qquad (12)$$

где $\mathfrak{P}$ — собственная максимальная цепочка оператора T, содержащая цепочку P_t, которая уже определялась, а $\mathfrak{R}(P)$ — некоторая оператор-функция на цепочке $\mathfrak{P}$, вычисляемая по вполне определенным правилам. В случае слабого сжатия или сжатия, подобного унитарному, интеграл определяется обычным способом. В других случаях, например, когда $D_T \in \mathfrak{S}_\omega$ — а в этом случае представление (12) всегда имеет место,— интеграл надо, вообще говоря, определять специальным образом.

Отметим, что в этих исследованиях Гокра принимал участие В. М. Бродский (младший). Они проводились под влиянием предшествовавших исследований М. С. Бродского [15а, б], которому удалось впервые получить мултипликативное представление характеристической функции (второго типа) для ограниченного оператора с вполне непрерывной мнимой компонентой как следствие треугольного представления операторов.

Отметим, что поскольку собственная цепочка $\mathfrak{P}_T$, вообще говоря, не определяется однозначно по оператору, то и представление (12) его характеристической функции не определяется единственным образом.

Аналитические проблемы, связанные с описанием и классификацией всех мультипликативных представлений характеристической оператор-функции данного оператора, по-видимому, и должны составить ту теорию, которая заменит теорию элементарных делителей при переходе от конечномерного случая к бесконечномерному.

Изложенным чисто операторным исследованиям предшествовали фундаментальные исследования В. П. Потапова [18] и его ученика Ю. П. Гинзбурга [19] о мультипликативном представлении аналитических матриц- и оператор-функций. Сформулируем теорему В. П. Потапова для так называемого дефинитного случая.

Пусть $\theta(\zeta)$ — голоморфная внутри единичного круга матрица-функция, значениями которой являются сжатия. Тогда $\theta(\zeta)$ допускает следующее представление:

$$\theta(\zeta) = U \overleftarrow{\int_a^b} \exp\left\{-\frac{e^{i\varphi(t)}+\zeta}{e^{i\varphi(t)}-\zeta}\, dF(t)\right\} B(\zeta), \qquad (13)$$

где

$$B(\zeta) = \overleftarrow{\prod_k} \left(\frac{\zeta_k-\zeta}{1-\bar{\zeta}_k\zeta}\cdot\frac{|\zeta_k|}{\zeta_k} P_k + Q_k\right),$$

U — постоянный унитарный оператор, $\varphi(t)$ — неубывающая функция $(0 \leqslant \varphi(t) \leqslant 2\pi$ при $a \leqslant t \leqslant b)$, $F(t)$ — монотонная эрмитовозначная матрица-функция, P_k — ортопроекторы, $Q_k = I - P_k$.

Эта теорема обобщает соответствующую теорему о мультипликативном представлении голоморфной внутри круга скалярной ограниченной функции. Последнюю мы, в сущности, использовали при записи мультипликативного представления определителя $d_T(\zeta)$ для слабого сжатия. Более общим классом скалярных функций является класс, введенный Р. Неванлинной,— так называемый класс функций ограниченного вида. Его известными подклассами являются классы Ф. Рисса, М. Рисса, Харди и др. Ю. П. Гинзбургу удалось определить матричные и операторные аналоги этих классов и получить для них мультипликативные представления.

Возникает вопрос, как связаны теоремы В. П. Потапова и Ю. П. Гинзбурга с теоремами о мультипликативном представлении характеристической оператор-функции? Связь самая тесная, хотя в настоящее время эти результаты не перекрывают друг друга. Дело в том, что всякая голоморфная внутри круга функция $\theta(\zeta)$, значениями которой служат сжатия в некотором пространстве $\mathfrak{R}$, допускает следующее представление:

$$\theta(\zeta) = V\theta_T(\zeta) \oplus U, \qquad (14)$$

где V, U — некоторые постоянные унитарные операторы, а $\theta_T(\zeta)$ — характеристическая функция некоторого простого сжатия T. Последнее восстанавливается с точностью до унитарной эквивалентности и действует в некотором новом пространстве $\mathfrak{H}$.

Представление (14) даже в случае скалярной функции $\theta(\zeta)$, когда $U = 0$, а V — число, равное по модулю единице, уже составляло некоторое открытие, сделанное М. С. Лившицем в упоминавшейся работе 1944 г. (см. [12а]). Впоследствии оно обобщалось этим автором, его учениками и сотрудниками. Наконец, в самом общем виде это представление было установлено в работах Надя и Фойяша. Ряд результатов и общая идея связи между мультипликативным представлением матрицы-функции или оператор-функции и структурой

инвариантных подпространств некоторого несамосопряженного оператора также принадлежит М. С. Лившицу.

Большим успехом явились установленные Ю. П. Гинзбургом теоремы единственности мультипликативных представлений (13) сжимающих матриц-функций, которые он распространил на так называемый ядерный случай. Характерно, что для получения определенной части этих результатов пришлось воспользоваться связью с теорией характеристических функций операторов и теоремами об одноклеточности слабых сжатий. Отметим, наконец, что В. П. Потапов получил также мультипликативное представление для матриц-функций, мероморфных внутри единичного круга, значениями которых являются операторы сжатия по отношению к некоторой индефинитной метрике. Эти результаты, а тем более последующие обобщения Ю. П. Гинзбурга до сих пор не удалось в полном объеме получить чисто операторными методами.

Следует надеяться, что в ближайшие годы будет ликвидировано расхождение между тем, что дают чисто аналитические методы в проблеме мультипликативного представления, с одной стороны, и операторные методы, опирающиеся на треугольные представления и теорию характеристических функций, с другой стороны.

Это расхождение в значительной мере объясняется тем, что для операторов, близких к унитарным и отличных от сжатий или растяжений, по-видимому, все еще не получено достаточно общего определения характеристической оператор-функции. Дело в том, что при имеющихся определениях этой функции в индефинитном случае нет разложения, аналогичного разложению (14).

Разумеется, когда речь идет о мультипликативном представлении оператор-функций, разделение методов на чисто аналитические и чисто операторные весьма условно. Ведь объект, подлежащий рассмотрению,— некий гибрид: это аналитическая функция, но ее значениями служат операторы.

В частности, этот гибрид имеет особо сложную операторную природу в индефинитном случае. Мы не имеем никакой возможности останавливаться на трудностях, порожденных индефинитностью. Они доставили много неприятностей даже в конечномерном случае.

В теории интеграла Стильтьеса — Коши играют важную роль формулы Сохоцкого — Племеля. В весьма широких предположениях аналоги этих формул для мультипликативных интегралов получил Л. А. Сахнович [16а, б]. Эти формулы позволили ему при достаточно общих предположениях о характере диссипативных возмущений построить теорию рассеяния, сходную во многом с теорией рассеяния Розенблюма — Като для самосопряженных операторов. Отмечу также перекрывающуюся с этими результатами недавнюю работу Тосио Като по теории уравнений Шредингера с комплексным потенциалом [2]. Я до сих пор ни разу не отметил,

что отправным пунктом исследований Секефальви-Надя — Фойяша служит теорема Б. Секефальви-Надя о существовании у каждого сжатия унитарного растяжения. Последовательное развитие теории растяжения сжатий выявило ее связи с задачей о рассеянии и теорией прогнозирования стационарных случайных процессов.

В результате выяснилось, что характеристическую функцию сжатия всегда можно трактовать как гейзенберговскую S-матрицу соответствующим образом сформулированной задачи рассеяния.

Все это обнаружилось, когда П. Лакс и Р. Филлипс построили новую теорию рассеяния акустических волн на препятствиях. Эта изящная теория замечательна единством методов классической математической физики, абстрактной теории операторов и идей, заимствованных из квантовой механики и теории прогнозирования стационарных случайных процессов. В этой теории и была предложена новая абстрактная схема задачи о рассеянии, которая сразу сблизила исследования этих авторов с исследованиями Надя — Фойяша. Связь между двумя теориями четко разъяснена в недавних работах двух молодых математиков: В. М. Адамяна и Д. З. Арова. В этих работах часть исследований четырех авторов, о которых шла речь, обобщена (в направлении дальнейшего их сближения с исследованиями по прогнозированию) в виде своеобразной теории сцеплений полуунитарных операторов [25а, б, в, г].

Насколько мне известно, в самом недалеком будущем мы будем иметь удовольствие ознакомиться в деталях с теориями Надя — Фойяша и Филлипса — Лакса по книгам этих авторов [1]).

Здесь же укажу на недавнюю книгу М. С. Лившица [12в], в которой изложены применения характеристической матрицы-функции в разнообразных задачах теории электрических сетей, волноводов и задачах ядерной физики.

4. Целые эрмитовы операторы и теория канонических представлений операторов

Все предыдущие рассуждения носили сугубо несамосопряженный характер. Вместе с тем, часть изложенного имеет кардинальное значение и в тех исследованиях по теории самосопряженных операторов, которые можно объединить под названием «Теория канонических представлений самосопряженных операторов».

К концу 30-х годов после работ Т. Карлемана, Дж. Неймана, М. Стоуна и М. А. Наймарка теория эрмитовых и самосопряженных операторов казалась завершенной; создавалось впечатление, что

[1]) Во время подготовки настоящего издания указанные книги вышли в свет [21, 26].

на долю последующих исследователей и поколений остается лишь внедрение законченной теории в смежные области математики и физики. На самом деле, впереди лежали трудные вопросы, требующие большого нового аналитического аппарата.

Чтобы дать первое представление об одной из столь поздно замеченных проблем, напомним следующий элементарный факт из линейной алгебры. Как известно, для всякого эрмитова оператора с простым спектром, действующего в конечномерном пространстве, существует бесконечное число ортонормированных базисов, в которых оператор изображается якобиевой (т. е. трехдиагональной) эрмитовой матрицей.

Этот результат непосредственно переносится на ограниченные операторы. Пусть G ($\in \mathfrak{R}$) — самосопряженный оператор с простым спектром. Простота спектра означает цикличность оператора G, т. е. наличие вектора $u \in \mathfrak{H}$, для которого последовательность $u, Gu, G^2u, \ldots$ полна в $\mathfrak{H}$. Если эту последовательность ортогонализовать по Шмидту, то в полученном базисе оператору G будет отвечать якобиева матрица

$$G \sim \begin{Vmatrix} a_1 & b_1 & 0 & & \\ \bar{b}_1 & a_2 & b_2 & 0 & \\ 0 & \bar{b}_2 & a_3 & \ddots & \\ & & \ddots & \ddots & \end{Vmatrix}.$$

Это представление, разумеется, не единственно и определяется выбором производящего вектора u.

А что произойдет, если наш оператор $G = H$ неограничен? Заведомо известно, что не для всякого оператора с простым спектром существует ортонормированный базис, в котором оператор будет изображаться якобиевой матрицей.

С другой стороны, давно известно, что дифференциальные операторы второго порядка следует рассматривать как континуальные аналоги операторов, задаваемых якобиевыми матрицами. Спрашивается, достаточен ли запас дифференциальных операторов для получения изображения произвольного самосопряженного оператора H по любому наперед заданному его производящему элементу u? Этот запас оказывается, вообще говоря, недостаточным, если даже допускать в качестве коэффициентов у дифференциального оператора обобщенные функции. Мы получим нужный нам класс, если введем в рассмотрение канонические дифференциальные операторы. С помощью этих дифференциальных операторов уже можно описывать любые самосопряженные операторы, причем со спектром любой кратности. Центральным в этой теории является понятие *целого эрмитова оператора*.

Приведем определение целого оператора, точнее $\mathfrak{L}$-*целого эрмитова оператора*. Пусть H — некоторый простой замкнутый эрмитов оператор с плотной областью определения $\mathfrak{D}(H)$. Обозначим через $\mathfrak{M}_z$ множество значений оператора $H - zI$:

$$\mathfrak{M}_z = (H - zI)\,\mathfrak{D}(H).$$

Пусть $\mathfrak{L}$ — некоторое подпространство в $\mathfrak{H}$. Данный простой эрмитов оператор H называется $\mathfrak{L}$-целым, если для любого z

$$1)\ \mathfrak{M}_z = \overline{\mathfrak{M}}_z \quad \text{и} \quad 2)\ \mathfrak{M}_z \dot{+} \mathfrak{L} = \mathfrak{H}.$$

Отметим, что уже из первого условия следует, что дефектные числа оператора H равны, т. е. $\dim(\mathfrak{H} \ominus \mathfrak{M}_z) = \dim \mathfrak{L}$. Обозначим через $\mathscr{E}(z)$ косой проектор, проектирующий пространство $\mathfrak{H}$ на $\mathfrak{L}$ параллельно $\mathfrak{M}_z$. Если оператор H является $\mathfrak{L}$-целым, то $\mathscr{E}(z)$ оказывается целой оператор-функцией. Это означает, что отображение $f \to f_{\mathfrak{L}}(z) = \mathscr{E}(z) f$ относит каждому вектору $f \in \mathfrak{H}$ целую функцию от z со значениями в $\mathfrak{L}$. Это отображение является одно-однозначным. При этом отображении оператор H переходит в оператор умножения на z, т. е. если $g = Hf$, то $g_{\mathfrak{L}}(z) = z f_{\mathfrak{L}}(z)$. Этим и объясняется название «$\mathfrak{L}$-целый оператор».

В полученном представлении скалярное произведение будет записываться следующим образом:

$$(f, g) = \int_{-\infty}^{\infty} g^*_{\mathfrak{L}}(\lambda)\, d\sigma(\lambda)\, f_{\mathfrak{L}}(\lambda) \qquad (f, g \in \mathfrak{H}), \tag{15}$$

где $\sigma(\lambda) = \sigma(\lambda - 0)$ $(\sigma(-\infty) = 0)$ — неубывающая ограниченная оператор-функция, значениями которой служат неотрицательные операторы, действующие в $\mathfrak{L}$. Если $\mathfrak{L}$ одномерно, то все стоящие здесь функции можно рассматривать как скалярные. Для придания правой части более изящного вида использовано следующее обозначение для скалярного произведения векторов из $\mathfrak{L}$:

$$(a, b) = b^* a \qquad (a, b \in \mathfrak{L}).$$

Формулой

$$\sigma(\lambda) = P_{\mathfrak{L}} E(\lambda) P_{\mathfrak{L}},$$

где $E(\lambda)$ — какая-либо обобщенная спектральная функция оператора H, а $P_{\mathfrak{L}}$ — ортогональный проектор на $\mathfrak{L}$, описывается множество $\Sigma(H; \mathfrak{L})$ всех спектральных функций $\sigma(\lambda)$, дающих представление (15). Существует чисто аналитическое описание множества $\Sigma(H, \mathfrak{L})$; оно получается с помощью так называемой резольвентной матрицы $W(z)$ второго порядка, элементы которой суть операторы, действующие в $\mathfrak{L}$. В недавней работе докладчика и Ш. Н. Саакяна

[23] было получено следующее компактное представление этой матрицы:

$$W(z)=\begin{pmatrix} I-zF(z)\,\mathscr{E}^*(0) & -zF(z)\,F^*(0) \\ z\mathscr{E}(z)\mathscr{E}^*(0) & I+z\mathscr{E}(z)\,F^*(0) \end{pmatrix}, \qquad (16)$$

где $F(z)$ — целая оператор-функция, *союзная* с функцией $\mathscr{E}(z)$; она определяется равенством

$$F(z)=P_{\mathfrak{L}}(H-zI)^{-1}(1-\mathscr{E}(z)).$$

С помощью этой матрицы-функции $W(z)$ находится общий вид интеграла

$$\int_{-\infty}^{\infty}\frac{d\sigma(\lambda)}{\lambda-z} \qquad \left(\sigma(\lambda)\in\Sigma(H;\mathfrak{L})\right).$$

Получаемый при этом результат следует рассматривать как обобщение классического результата Рольфа Неванлинны, дающего описание всех решений неопределенной проблемы моментов. В некотором отношении это не является неожиданным. На спектральную функцию $\sigma(\lambda)$ можно смотреть как на решение обобщенной проблемы моментов. В самом деле, в равенстве (15) в левой части стоит известная величина для любых f и $g\in\mathfrak{H}$, известными являются также функции $f_{\mathfrak{L}}(z)$ и $g_{\mathfrak{L}}(z)$. Этот результат примыкает к старым исследованиям М. А. Наймарка и докладчика по описанию обобщенных резольвент и последующим работам А. В. Штрауса и его учеников.

Неожиданным оказалось, что при весьма общих условиях $W\left(\frac{1}{z}\right)$ является характеристической функцией (в смысле М. С. Бродского) некоторого вольтеррова оператора. Поэтому на основании результатов В. П. Потапова и М. С. Бродского эта функция допускает мультипликативное представление

$$W(z)=\overrightarrow{\int_0^t}\exp(-zJ\mathscr{H}(t)\,dt),$$

где $\mathscr{H}(t)$ — функция, значениями которой являются ограниченные эрмитовы операторы, действующие в ортогональной сумме двух копий пространства $\mathfrak{L}$: $\mathfrak{L}\oplus\mathfrak{L}$, а J — вещественный квадратный корень из $-I$:

$$J=\begin{pmatrix} 0 & I_{\mathfrak{L}} \\ -I_{\mathfrak{L}} & 0 \end{pmatrix}.$$

Этому представлению отвечает каноническое дифференциальное уравнение

$$J\frac{d\varphi}{dt}=\lambda\mathscr{H}(t)\varphi \qquad (0\leqslant t\leqslant l).$$

Оказывается, что множество $\Sigma(H;\mathfrak{L})$ спектральных функций $\sigma(\lambda)$ совпадает с множеством спектральных функций неполной краевой задачи для приведенного канонического уравнения. Неполнота означает, что граничное условие задается только на левом конце. При этом сами функции $f_{\mathfrak{L}}(z)$ можно интерпретировать как обобщенные преобразования Фурье соответствующих функций из $L_2(0,\,l;\,\mathfrak{L})$ с помощью определенным образом нормированной фундаментальной функции $\varphi(t;\lambda)$.

Пусть теперь H — произвольный самосопряженный оператор,

$$H=\int_{-\infty}^{\infty}\lambda\,dE(\lambda)$$

— его спектральное разложение, а $\mathfrak{L}$ — некоторое минимальное воспроизводящее подпространство, т. е.

$$\bigvee_{-\infty<\lambda<\infty}\{E(\lambda)\,\mathfrak{L}\}=\mathfrak{H}.$$

Оказывается, оператор H можно рассматривать как предел расширяющейся системы $\mathfrak{L}$-целых операторов. Иначе это означает, что ему можно сопоставить каноническое уравнение типа (15), однако уже на полуоси, которое будет иметь в качестве своей единственной спектральной функции функцию $\sigma(\lambda)=P_{\mathfrak{L}}E(\lambda)P_{\mathfrak{L}}$.

Мы видим, что нам удалось обойтись уравнениями первого порядка, однако фазовая размерность этих уравнений равна удвоенной размерности $\mathfrak{L}$ (если $\mathfrak{L}$ конечномерно). Если оператор H положителен и имеет простой спектр, то отвечающее ему каноническое уравнение можно преобразовать в уравнение струны с произвольным распределением масс. К сожалению, гамильтониан $\mathscr{H}(t)$, вообще говоря, определяется неоднозначно. По-видимому, вопрос о том, какие дополнительные условия следует налагать на гамильтониан $\mathscr{H}(t)$, чтобы он определялся однозначно, является вопросом огромной трудности, и вряд ли можно надеяться, что он будет решен в ближайшие несколько лет.

Для случая $n=\dim\mathfrak{L}=1$, когда спектральная функция $\sigma(\lambda)$ — скалярная неубывающая функция, ей всегда можно сопоставить каноническую систему с вещественным гамильтонианом $\mathscr{H}(t)$ (со следом $=1$). Около пятнадцати лет тому назад я пришел к предположению, что при такой нормировке в этом случае гамильтониан будет определяться однозначно. В ряде важных случаев оно мною было подтверждено. Полное доказательство этого предположе-

ния недавно получил Луи де Бранж [8б] в его исследованиях по гильбертовым пространствам целых функций. Тем самым было доказано и второе мое предположение [24] о том, что *всякий вещественный простой вольтерров оператор с двумерной кососимметрической компонентой является вещественно одноклеточным*.

Я не имею возможности сформулировать различные предположения и проблемы, возникшие в теории канонических представлений самосопряженных операторов. Для развития этой теории весьма существенно распространить ее на тот случай, когда в качестве $\mathfrak{L}$ выбирается пространство обобщенных элементов первого порядка сингулярности относительно оператора H. Для случая конечномерного $\mathfrak{L}$ соответствующее обобщение получается без труда. В случае бесконечномерного $\mathfrak{L}$ здесь не выработана еще точная постановка проблемы. Вероятно, с этим связано, что до сих пор еще не построены целые эрмитовы операторы, изображаемые дифференциальными операторами в частных производных.

Заканчивая свой доклад, я хотел бы подчеркнуть, что исполнял обязанности не архитектора, а скорее гида, показывающего отдельные достопримечательности большого нового центра. Иногда мы шли медленнее, иногда мчались, пролетая стремглав целые магистрали. Естественно, за предоставленный короткий срок я не мог не огорчить руководителей целых районов, территории которых мы проскочили или даже обошли стороной. Приношу им мои извинения [1]).

Одесский инженерно-строительный институт, СССР

ЛИТЕРАТУРА

[1] Friedrichs K. O., Perturbation of spectra in Hilbert space, AMS, 1965.

[2] Kato T., Perturbation theory, Springer Verlag, 1966.

[3] Мацаев В. И., а) Теоретико-функциональные методы в некоторых вопросах теории линейных самосопряженных операторов, Докторская диссертация, Москва, 1966.

б) Об одном классе вполне непрерывных операторов, *ДАН СССР*, **139**, № 4 (1961), 810-814.

[4] Келдыш М. В., О собственных значениях и собственных функциях некоторых классов несамосопряженных уравнений, *ДАН СССР*, **74**, № 1 (1951), 11-14.

[5] Келдыш М. В., Лидский В. Б., Вопросы спектральной теории несамосопряженных операторов, Труды IV Всесоюзного математического съезда, т. 1, 1963, стр. 101-120.

[6] Гохберг И. Ц., Крейн М. Г., а) Введение в теорию линейных несамосопряженных операторов, изд-во «Наука», 1965.

б) Теория вольтерровых операторов в гильбертовом пространстве и ее приложения, изд-во «Наука», 1967.

[1]) Выражаю глубокую благодарность И. Ц. Гохбергу за неоценимую помощь, оказанную при подготовке доклада.

в) О факторизации операторов в гильбертовом пространстве, *Acta Sci. Math. Szeged*, **25**, 1—2 (1964), 90-123.

г) Об одном описании операторов сжатия, подобных унитарным операторам, *Функциональный анализ и его приложения*, **1**, 1 (1967).

[7] Любич Ю. И., Мацаев В. И., Об операторах с отделимым спектром, *Матем. сб.*, **56**, 4 (1962), 433-468.

[8] De Branges L., а) Some Hilbert spaces of analytic functions, II, III, *J. Math. Analysis and Applications*, **11** (1965), 44-72; **12** (1965), 149-186.

б) Some Hilbert spaces of entire functions, I—IV, *Trans. Amer. Math. Soc.*, **10** (1959), 840-846; **99** (1961), 118-152; **100** (1961), 73-115; **105** (1962), 43-83.

[9] Бродский М. С., Кисилевский Г. Э., Критерий одноклеточности вольтерровых операторов с ядерными мнимыми компонентами, *Изв. АН СССР*, сер. матем. **30**, № 6 (1966), 1213-1228.

[10] Кисилевский Г. Э., Инвариантные подпространства вольтерровых диссипативных операторов с ядерными мнимыми компонентами, *Изв. АН СССР*, сер. матем. (в печати).

[11] Sz.-Nagy B., Foiaş, C., а) Sur les contractions de l'espace de Hilbert, VIII. Fonctions caracteristiques, Modèles fonctionnels, *Acta Sci. Math. Szeged*, **25**, 1—2 (1964), 38-71.

б) Sur les contractions de l'espace de Hilbert, IX. Factorisations de la fonction caractéristique. Sous-espaces invariants, *Acta Sci. Math. Szeged*, **25**, 3—4 (1964), 283-316.

в) Sur les contractions de l'espace de Hilbert, X. Contractions similaires a des transformations unitaires, *Acta Sci. Math. Szeged*, **26**, 1—2 (1965), 79-91.

г) Sur les contractions de l'espace de Hilbert, XI. Transformations unicellulaires, *Acta Sci. Math. Szeged*, **26**, 3—4 (1965), 301-324.

[12] Лившиц М. С., а) Об одном классе линейных операторов в гильбертовом пространстве, *Матем. сб.*, **19** (1946), 236-260.

б) О спектральном разложении линейных несамосопряженных операторов, *Матем. сб.*, **34** (1954), 145-198.

в) Операторы, колебания, волны, изд-во «Наука», 1966.

[13] Лившиц М. С., Потапов В. П., Теорема умножения характеристических матриц-функций, *ДАН СССР*, **72** (1950), 625-628.

[14] Бродский М. С., Лившиц М. С., Спектральный анализ несамосопряженных операторов и промежуточные системы, *УМН*, **13**, 1 (1958), 3-85.

[15] Бродский М. С., а) Спектральный анализ линейных ограниченных операторов с вполне непрерывной мнимой компонентой, Докторская диссертация, Одесса, 1962.

б) О мультипликативном представлении некоторых аналитических операторфункций, *ДАН СССР*, **138**, № 4 (1961).

[16] Сахнович Л. А., а) О приведении несамосопряженных операторов к треугольному виду, *Изв. высших учебных заведений*, матем., **4** (11) (1959), 141-149.

б) О диссипативных операторах с абсолютно непрерывным спектром, *ДАН СССР*, **167**, № 4 (1966), 760-763.

[17] Aronszajn N., Smith R. J., Invariant subspaces of completely continuous operators, *Ann. Math.*, **60** (1954), 316-320 (имеется русский перевод: *Математика*, **2**, № 1 (1958)).

[18] Потапов В. П., Мультипликативная структура J-нерастягивающих матриц-функций, Труды Моск. матем. о-ва, № 4 (1955), 125-236.

[19] Гинзбург Ю. П., О J-нерастягивающих оператор-функциях, *ДАН СССР*, **117**, № 2 (1957).

[20] Sz.-Nagy B., a) Sur les contractions de l'espace de Hilbert, *Acta Sci. Math. Szeged*, **15** (1953), 87-92.
б) Sur les contractions de l'espace de Hilbert, II, *Acta. Sci. Math. Szeged*, **18**, 1—2 (1957), 1-14.
[21] Lax P. D., Phillips R. S., Scattering Theory, Academic Press, 1967.
[22] Крейн М. Г., а) Об одном замечательном классе эрмитовых операторов, *ДАН СССР*, **44** (1944), 191-195.
б) Основные положения теории представления эрмитовых операторов с индексом дефекта (m, m), *Укр. матем. журнал*, **1**, 2 (1949), 3-66.
[23] Крейн М. Г., Саакян Ш. Н., О некоторых новых результатах в теории резольвент эрмитовых операторов, *ДАН СССР*, **169**, № 6 (1966), 1269-1272.
[24] Бродский М. С., Гохберг И. Ц., Крейн М. Г., Мацаев В. И., О некоторых новых исследованиях по теории несамосопряженных операторов, Труды IV всесоюзного матем. съезда, т. II, 1964.
[25] Адамян В. М., Аров Д. З., а) Об одном классе операторов рассеяния и характеристических оператор-функциях сжатий, *ДАН СССР*, **160**, № 1 (1965), 9-12.
б) Об операторах рассеяния и полугруппах сжатий в гильбертовом пространстве, *ДАН СССР*, **165**, № 1 (1965), 9-12.
в) Об унитарных сцеплениях полуунитарных операторов, *Докл. АН Арм. ССР*, **43**, № 5 (1966), 257-263.
г) Об унитарных сцеплениях полуунитарных операторов, Математические исследования, **1**, вып. 1, Кишинев (1966), 3-64.
[26] Sz.-Nagy B., Foiaş C., Analyse harmonique des opérateurs de l'espace de Hilbert, Akadémiai Kiadó Budapest, 1967.

О НЕКОТОРЫХ ПОГРАНИЧНЫХ ВОПРОСАХ АЛГЕБРЫ И ЛОГИКИ

А. И. МАЛЬЦЕВ

В настоящем докладе мы хотим сделать обзор некоторых результатов и проблем, относящихся к математической дисциплине, возникшей в последние десятилетия на рубеже между математической логикой и классической абстрактной алгеброй и пока не получившей общепринятого наименования. Наиболее часто она называется теорией моделей и универсальной алгеброй, некоторые же говорят об общей алгебре.

Основными математическими структурами, изучаемыми в этой общей алгебре, являются алгебраические системы, т. е. ансамбли, состоящие из какого-то непустого множества и некоторого числа определенных на нем операций и отношений различных конечных арностей. Типичным примером алгебраической системы может служить упорядоченное кольцо $\langle A; -, \cdot, \leqslant; \tau\rangle$, состоящее из основного множества A элементов кольца, называемого также его носителем, символов $-, \cdot$, бинарных операций вычитания и умножения, символа $\leqslant$ бинарного отношения порядка и отображения τ, ставящего упомянутым символам $-, \cdot, \leqslant$ те конкретные операции и отношение, которые служат значениями символов в данном конкретном кольце. Совокупность символов $-, \cdot, \leqslant$, рассматриваемых вместе с их арностями 2, 2, 2, называется сигнатурой упорядоченного кольца.

В общем случае сигнатурой называется пара Ω_f, Ω_p непересекающихся множеств и отображение α: $\Omega_f \cup \Omega_p \to N$ объединения их в множество натуральных чисел $N = \{0, 1, 2, \ldots\}$. Элементы Ω_f называются функциональными, а элементы из Ω_p — предикатными сигнатурными символами. В дальнейшем сигнатура и множество всех сигнатурных символов будут обозначаться одной и той же буквой Ω. Отвечающее произвольному сигнатурному символу $\omega \in \Omega$ натуральное число $\alpha\omega$ называется арностью символа ω.

Алгебраической системой данной сигнатуры Ω называется ансамбль

$$\mathbf{A} = \langle A; \Omega; \tau\rangle,$$

состоящий из непустого множества A, сигнатуры Ω и отображения τ, ставящего в соответствие каждому $f \in \Omega_f$ некоторую функцию $f^\tau: A^{\alpha f} \to A$ и каждому предикатному сигнатурному символу $p \in \Omega_p$

некоторый предикат $p^\tau \subseteq A^{\alpha p}$. Функции f^τ и предикаты p^τ называются значениями соответствующих сигнатурных символов f, p в системе **A**, а отображение τ называется означивающим отображением или просто означиванием. Символ означивания τ при записи алгебраических систем обычно опускается и вместо $\langle A; \Omega; \tau\rangle$ пишется $\langle A; \Omega\rangle$. Алгебраическая система называется алгеброй, если ее сигнатура не содержит предикатных символов, и называется моделью, если ее сигнатура не содержит функциональных символов. Алгебраическая система **A** конечна, если конечно ее основное множество A.

Обычным путем определяются понятия подсистемы данной алгебраической системы и гомоморфного и изоморфного отображений одной алгебраической системы в произвольную систему той же сигнатуры. Произвольная совокупность алгебраических систем одной и той же сигнатуры Ω называется классом систем сигнатуры Ω. Класс систем называется абстрактным, если он вместе с каждой своей системой содержит и все ей изоморфные.

Идея о необходимости изучения свойств алгебр произвольной сигнатуры возникла еще в конце прошлого века. Однако в течение более трех последующих десятилетий она не получила сколько-нибудь существенного развития. Вместо этого, с одной стороны, были созданы глубокие теории частных классов алгебр—полей, колец, групп, решеток, а с другой стороны, в математической логике были проведены широкие исследования простейших формальных языков. Во второй половине 30-х годов было замечено, что объединение идей алгебраической системы и языка 1-й ступени позволяет сформулировать предложения, специализации которых для классических систем — полей, групп — не только дают нетривиальные уже известные теоремы теорий групп и полей, но дают ответ и на некоторые в то время открытые вопросы общей теории групп. Так на стыке классической абстрактной алгебры и математической логики возникла новая дисциплина — общая алгебра, в которой в отличие от классической алгебры видное место заняли проблемы зависимостей между структурными свойствами классов алгебр и свойствами формальных языков, на которых могут быть определены упомянутые классы. Полного развития исследования по общей алгебре достигли в послевоенные годы. Особенно значительные законченные результаты были получены в конце 50-х и 60-х годов. Достаточно назвать создание теории фильтрованных произведений и теории полных классов. Подробные обзоры этих теорий уже появились в журналах, и потому в дальнейшем мы будем рассматривать лишь другие направления исследований.

Напомним еще несколько понятий. Пусть задана какая-нибудь сигнатура Ω. Комбинируя по известным правилам сигнатурные символы, скобки, символы $x_1, x_2, \ldots$ предметных переменных,

символы логических связок $\&$, $\vee$, $\rightarrow$, $\neg$, $=$ и кванторы

$\forall x_i$ — «*для каждого элемента x_i носителя A системы*»,

$\exists x_i$ — «*существует такой элемент $x_i \in A$, что*»,

получим конечные последовательности символов, называемые *формулами 1-й ступени* сигнатуры Ω. Например, последовательности

$$(\forall x_1)(\forall x_2)(x_1 + x_2 = x_2 + x_1), \tag{1}$$

$$(\forall x_1)(\exists x_2)(x_2 \leqslant x_1 \,\&\, x_1 \neq x_2) \tag{2}$$

являются формулами 1-й ступени любой сигнатуры, включающей знаки $+$, $\leqslant$. Если теперь заданы какая-нибудь алгебраическая система **A** сигнатуры Ω и формула $\mathcal{A}$ той же сигнатуры, то в соответствии с содержательным смыслом символов, участвующих в записи формулы $\mathcal{A}$, определяется истинность или ложность формулы $\mathcal{A}$ в системе **A**. Например, если

$$\mathbf{A} = \langle\{0, 1, 2, \ldots\};\ +,\ \leqslant\rangle,$$

где символы $+$, $\leqslant$ имеют обычные арифметические *значения*, то формула (2) ложна, а формула (1) истинна в **A**.

Символом $\mathfrak{E}$ условимся обозначать класс всевозможных формул 1-ступени любой сигнатуры, а символом $\mathfrak{E}_\Omega$ будем обозначать множество формул 1-й ступени, сигнатура которых содержится в Ω. Помимо класса $\mathfrak{E}$ далее нам потребуются некоторые подклассы формул специального вида. Напомним, что формулы, в записи которых участвуют лишь функциональные символы и символы предметных переменных, называются термами (полиномами) от указанных переменных. Например, если $+$, $\wedge$ суть бинарные функциональные символы, то выражения

$$x + (x \wedge y), \qquad (x + x) + (x + x)$$

являются термами от x, y сигнатуры $\{+, \wedge\}$.

Введем теперь следующие специальные классы формул:

$\mathfrak{T}$ — *класс тождеств*, т. е. формул вида

$$(\forall x_1) \ldots (\forall x_n)\, P(f_1, \ldots, f_s),$$

где P—какой-нибудь предикатный сигнатурный символ или знак равенства, а $f_1, \ldots, f_s$ — термы от $x_1, \ldots, x_n$;

$\mathfrak{Q}$ — *класс квазитождеств*, т. е. формул вида

$$(\forall x_1) \ldots (\forall x_n)(P_1(f_1, \ldots, f_r) \,\&\, \ldots \,\&\, P_k(h_1, \ldots, h_s) \rightarrow P(l_1, \ldots, l_t)),$$

где $P_1, \ldots, P_k$, P — некоторые предикатные сигнатурные символы или знаки равенства, а $f_1, \ldots, f_r$, $h_1, \ldots, h_s$, $l_1, \ldots, l_t$— термы от $x_1, \ldots, x_n$;

$\forall$ — *класс общностных (универсальных) формул*, т. е. формул вида

$$(\forall x_1) \ldots (\forall x_n)\, \mathscr{B},$$

где $\mathscr{B}$ — формула, не содержащая кванторов;

$\forall\exists$ — *класс формул вида*

$$(\forall x_1) \ldots (\forall x_m)(\exists x_{m+1}) \ldots (\exists x_n)\, \mathscr{B},$$

где снова $\mathscr{B}$ — формула, не содержащая кванторов;

$\mathscr{D}$ — *класс диофантовых формул*, т. е. формул вида

$$(\exists x_1) \ldots (\exists x_m)(P_1(f_1, \ldots, f_r) \& \ldots \& P_k(h_1, \ldots, h_s)),$$

где $P_1, \ldots, P_k$ — предикатные символы или знаки равенства, а $f_1, \ldots, f_r$, $h_1, \ldots, h_s$ — термы от $x_1, \ldots, x_n$.

Пусть Γ — класс формул 1-й ступени какого-нибудь специального вида, например один из только что введенных классов $\mathscr{J}$, $\mathscr{Q}$, ..., $\mathscr{D}$, и пусть $\mathfrak{K}$ — какой-нибудь класс алгебраических систем сигнатуры Ω. Тогда Γ-теорией класса $\mathfrak{K}$ называется совокупность $\Gamma\mathfrak{K}$ всех тех (закрытых) формул из Γ, сигнатура которых содержится в Ω и которые истинны в каждой системе класса $\mathfrak{K}$. Напротив, если задано какое-то конкретное множество формул Γ, то через $K_\Omega\Gamma$ обозначается класс всех тех алгебраических систем сигнатуры Ω, на которых истинна любая формула из Γ, сигнатура которой содержится в Ω. В частности, $K_\Omega\varnothing$ является классом «всех» алгебраических систем сигнатуры Ω. Он будет более кратко обозначаться через $K\Omega$. Через $\mathfrak{K}_{fin}$ будет обозначаться класс всех конечных систем, принадлежащих $\mathfrak{K}$.

Множество $\mathscr{E}_\Omega$ формул 1-й ступени сигнатуры Ω является частью совокупности W_Ω всех конечных последовательностей, составленных из символов сигнатуры Ω, скобок, логических знаков &, ..., =, $\forall$, $\exists$ и символа x [1]), т. е. является подмножеством множества слов в алфавите, состоящем из упомянутых символов. Совокупность всех слов в фиксированном алфавите несет на себе известную структуру индуктивной алгебры, которая позволяет ввести для совокупностей слов понятия рекурсивности, рекурсивной перечислимости и т. п. Эти понятия хорошо исследованы в теории алгоритмов для конечных алфавитов, но недавно Р. Петер, Ф. Швенкель [9] изучили некоторые их свойства и для бесконечных алфавитов, что позволяет ныне говорить о рекурсивности или нерекурсивности теорий не только классов алгебраических систем конечной сигнатуры, но и классов систем бесконечной сигнатуры.

[1]) Как обычно, предметное переменное x_i обозначается при этом словом $(xx \ldots x)$.

В результате естественно возникают следующие 2 круга вопросов:

1) для наиболее важных классов $\mathfrak{K}$ алгебраических систем и наиболее интересных классов Γ формул найти алгоритмическую природу теории $\Gamma\mathfrak{K}$;

2) для наиболее интересных классов формул Γ найти общие алгебраические свойства классов алгебраических систем вида $K_\Omega\Gamma_1$, где $\Gamma_1 \subseteq \Gamma$.

Классы вида $K_\Omega\Gamma_1$ ($\Gamma_1 \subseteq \Gamma$) обычно называются Γ-классами. В частности, $\mathscr{J}$-классы называются многообразиями, $\mathscr{Q}$-классы — квазимногообразиями и $\forall$-классы — универсально аксиоматизируемыми классами (или универсалами) алгебраических систем.

Мы хотим теперь сделать небольшие обзоры новых результатов и открытых проблем, примыкающих к указанным направлениям.

1. Алгоритмическая природа теорий

1.1. $\mathscr{E}$-теории и теории тотальных классов. Вопрос о рекурсивности теории $\mathscr{E}K\Omega$ был известен как проблема разрешимости узкого исчисления предикатов. Для достаточно богатой сигнатуры Ω он был решен отрицательно А. Черчем (1939) [10]. Вопрос о рекурсивности теории $\Gamma K\Omega$, $\Gamma K\Omega_{fin}$ для различных типов формул Γ и различных сигнатур Ω привлекал внимание многих авторов. Один из наиболее замечательных результатов этого направления был получен Ван Хао, доказавшим нерекурсивность теории $\forall^1\exists^1\forall^1K\Omega$, где через $\forall^1\exists^1\forall^1$ обозначен класс формул вида $\forall x\exists y\forall z\mathscr{B}$ ($\mathscr{B}$ кванторов не содержит), а Ω состоит из одного бинарного и бесконечного числа унарных предикатных символов. Пользуясь методами Ван Хао, Гуревич [12] получил в конце прошлого года в каком-то смысле окончательный результат в указанном направлении. Вводя вместе с Гуревичем обозначения

Γ — некоторая совокупность слов вида $(Q_1x_1)(Q_2x_2) \ldots (Q_mx_m)$

$$(Q_i = \forall, \exists,\ Q_i \neq Q_{i+1},\ m = 1, 2, 3, \ldots);$$

Ω — сигнатура, не содержащая функциональных символов;

Γ_Ω — совокупность замкнутых формул пренексного вида, не содержащих знака равенства, имеющих сигнатуру Ω, кванторная часть которых принадлежит Γ;

Γ^*_Ω — то же, но допускается знак равенства;

$\Gamma_2 = \{\forall^m\exists^n\colon m, n = 0, 1, 2, \ldots\}$;

$\Gamma_3 = \{\exists^m\forall^i\exists^n\colon i = 1, 2;\ m, n = 0, 1, 2, \ldots\}$;

Ω_1 — сигнатура, состоящая лишь из одноместных предикатных символов,

можно представить его результаты в следующем виде:

$$\operatorname{Rec}\Gamma_{\mathfrak{Q}}\mathbf{K}\Omega \Leftrightarrow \operatorname{Rec}\Gamma^{*}_{\mathfrak{Q}}\mathbf{K}\Omega \Leftrightarrow \operatorname{Rec}\Gamma_{\mathfrak{Q}}(\mathbf{K}\Omega)_{\mathrm{fin}} \Leftrightarrow$$
$$\Leftrightarrow \operatorname{Rec}\Gamma^{*}_{\mathfrak{Q}}(\mathbf{K}\Omega)_{\mathrm{fin}} \Leftrightarrow \neg\operatorname{Creat}\Gamma_{\mathfrak{Q}}\mathbf{K}\Omega \Leftrightarrow$$
$$\Leftrightarrow \Omega \subseteq \Omega_1 \vee \Gamma \subseteq \Gamma_2 \cup \Gamma_3 \vee (\operatorname{Fin}\Omega \,\&\, \operatorname{Fin}(\Gamma \setminus (\Gamma_2 \cup \Gamma_3))),$$

где Rec M, Creat M, Fin M означают, что совокупность M соответственно рекурсивна, креативна, конечна.

По традиции эти исследования причисляются к «чистой логике». Напротив, вопросы о рекурсивности теорий $\mathcal{E}\mathfrak{K}$, $\mathcal{Q}\mathfrak{K}$, $\mathcal{J}\mathfrak{K}$, где $\mathfrak{K}$ — тот или иной специальный класс систем, например класс групп или класс конечных групп, чаще относятся уже к теории самого класса $\mathfrak{K}$ (теории групп, теории конечных групп и т. п.). Исходными результатами этого направления можно считать теоремы Россера о нерекурсивности арифметики $\mathcal{E}\langle N; +, \cdot\rangle$ и Пресбургера о рекурсивности $\mathcal{E}\langle N; +\rangle$.

Значение этой области исследований стало особенно ясно после работ А. Тарского, доказавшего рекурсивность $\mathcal{E}\langle \mathbf{K}; +, \cdot\rangle$, $\mathcal{E}\langle \mathbf{C}; +, \cdot\rangle$, и работы Ю. Робинсон [13], обнаружившей нерекурсивность $\mathcal{E}\langle \mathbf{R}; +, \cdot\rangle$, где $\mathbf{K}$, $\mathbf{C}$, $\mathbf{R}$ — поля комплексных, вещественных и рациональных чисел. В известной книге Тарского, Мостовского и Робинсона [14] были подведены первые итоги развития новой области. Однако для очень многих важных классов $\mathfrak{K}$ алгоритмическая природа теории $\mathcal{E}\mathfrak{K}$ оставалась в то время неизвестной. В течение следующего десятилетия была доказана нерекурсивность элементарных теорий многих классов систем и, в частности, элементарных теорий вида $\mathcal{E}\mathfrak{K}_{\mathrm{fin}}$, т. е. элементарных теорий совокупностей всех конечных систем, содержащихся в тех или иных более известных классах $\mathfrak{K}$. Были найдены также сравнительно немногие классы $\mathfrak{K}$, для которых теория $\mathcal{E}\mathfrak{K}$ оказалась рекурсивной. В обзоре Ершова, Лаврова, Тайманова и Тайцлина [15] дана еводка результатов, полученных к 1964 году.

В последние годы резко увеличилось внимание, уделяемое теориям вида $\Gamma\mathfrak{K}$ для различных типов Γ. С другой стороны, естественно спрашивать, не только рекурсивна или нет какая-нибудь теория $\Gamma\mathfrak{K}$, но и какова ее степень неразрешимости. Наконец, большой интерес представляет и вопрос о том, для каких $\mathfrak{K}$, $\mathfrak{L}$, Γ имеет место равенство $\Gamma\mathfrak{K} = \Gamma\mathfrak{L}$.

В связи с работами по структуре теорий был развит ряд общих методов. В частности, для доказательства неразрешимости теорий детально исследован метод интерпретаций. Для доказательства рекурсивности наряду с прямым методом исключения кванторов основное значение приобрел метод модельной полноты, открытый А. Робинсоном [16]. Методы доказательств совпадения элементарных теорий также известны ныне в различных формах: метод перекидывания, метод ультрапроизведений, метод стратегий. Тем не менее и сегодня существует большое число важных теорий, алгоритмическая природа которых остается совершенно неизвестной.

Ниже я позволю себе указать ряд важных новых результатов, полученных после прошлого конгресса, и напомнить в связи с ними о некоторых открытых проблемах.

1.2. Теория чисел. Одной из наиболее замечательных нерешенных проблем все еще остается так называемая 10-*я проблема Гильберта*:

а) *рекурсивна или нет теория* $\mathcal{D}\langle N; -\cdot\rangle$?

Не решен и тесно связанный с этой проблемой вопрос:

б) *можно ли для любого рекурсивно-перечислимого множества* M *натуральных чисел найти такой многочлен* $f(x_0, x_1, \ldots, x_n)$ *с целыми коэффициентами, что*

$$x \in M \Leftrightarrow (\exists y_1, \ldots, y_n)(f(x, y_1, \ldots, y_n) = 0)?$$

Интересен также следующий вариант проблемы Гильберта:

в) *существуют ли натуральное число* s *и целозначные многочлены* $f_{1i}(n), \ldots, f_{si}(n)$, *такие, что совокупность тех* n, *для которых разрешимо уравнение*

$$\sum_{i=1}^{t} \pm x_1^{f_{1i}(n)} x_2^{f_{2i}(n)} \ldots x_s^{f_{si}(n)} = 0,$$

является нерекурсивной? Если существуют, то каковы наименьшие значения s, t *и степеней* $f_{ki}(n)$?

Хотя эти проблемы остаются все еще открытыми, тем не менее очень близкая проблема существования алгоритма, распознающего по коэффициентам полинома $f(x_1, \ldots, x_m, y_1, \ldots, y_m)$ разрешимость показательного уравнения

$$f(x_1, \ldots, x_m, 2^{x_1}, \ldots, 2^{x_m}) = 0,$$

была решена отрицательно в замечательной работе Девиса, Путнама и Робинсон [17]. Но и здесь остался открытым вопрос о нахождении многочленов f по возможности простого вида, для которых все еще не существует алгоритма решения.

1.3. Теория полей. В каком-то смысле родственными проблеме Гильберта являются вопросы о рекурсивности элементарных теорий классов полей. Поскольку на данном конгрессе этому вопросу посвящен специальный доклад Ю. Л. Ершова, я ограничусь лишь некоторыми замечаниями. В течение почти 15 лет известны были по существу только 2 класса бесконечных полей с разрешимыми теориями: алгебраически замкнутые поля фиксированной характеристики и вещественно замкнутые поля. В то же время число известных классов полей с неразрешимыми теориями быстро росло.

Только в 1964—1965 гг. Аксом и Коченом [18] методом ультрапроизведений и независимо Ю. Л. Ершовым [19], [20], [21] методом модельной полноты была установлена разрешимость элементарной теории P-адического поля и элементарных теорий ряда других полей. Можно считать, что сегодня работами Ю. Л. Ершова и Акса — Кочена уже заложен фундамент теории полей, имеющих разрешимые элементарные теории. Отметим также, что, используя тонкий аппарат теории моделей, Акс, Кочен и Ершов смогли попутно доказать и некоторые гипотезы Ленга и Артина о формах.

Работы по проблеме Гильберта и работы Акса, Кочена и Ершова представляются особенно интересными, так как они открывают пути «внедрения» теории моделей в область классической теории чисел и, возможно, в алгебраическую геометрию.

Несмотря на большие успехи в изучении элементарных теорий классов полей, много очень простых по формулировке проблем, относящихся к этой области, остаются открытыми. Упомянем лишь, что сегодня не известно, рекурсивна или нет элементарная теория

а) *класса всех конечных полей* (Ю. Робинсон);

б) *поля рациональных функций от переменных* $x_1, \ldots, x_m$ *над произвольным полем коэффициентов*;

в) *поля тех комплексных чисел, которые могут быть построены при помощи циркуля и линейки, и его подполя вещественных чисел* (А. Тарский).

В самой теории чисел большой интерес представляют проблемы эффективизации в теоремах типа теоремы Туе. Применение методов теории моделей, возможно, окажется полезным и в этой области.

1.4. Теория групп и полугрупп. Неразрешимость элементарной теории класса всех групп $\mathfrak{G}$ была установлена А. Тарским. Более тонкая теорема о неразрешимости $\mathfrak{GS}$ доказана П. С. Новиковым и Буном. В 60-х годах доказана нерекурсивность элементарных теорий многих классов групп, в том числе класса всех конечных групп, всех n-ступенно ($n \geqslant 2$) разрешимых групп и т. п. Не известно, рекурсивны или нет

а) *теория* $\mathfrak{GS}_n$ (проблема эквивалентности слов в классе $\mathfrak{S}_n$ n-ступенно разрешимых групп ($n = 2, 3, 4, \ldots$));

б) *теории* $\mathfrak{D}\mathbf{F}_n$, $\exists\mathbf{F}_n$, $\mathfrak{E}\mathbf{F}_n$, где $\mathbf{F}_n$ — свободная группа ранга n ($n \geqslant 2$);

Недавно А. Д. Тайманов показал, что

$$\forall\exists\forall\exists\mathbf{F}_m = \forall\exists\forall\exists\mathbf{F}_n \qquad (m, n \geqslant 2).$$

Однако все еще неизвестно, верны ли утверждения

в) $\mathfrak{E}\mathbf{F}_m = \mathfrak{E}\mathbf{F}_n$ $\quad (m, n \geqslant 2)$;

г) $\mathfrak{E}G_0 = \mathfrak{E}G_1 \,\&\, \mathfrak{E}H_0 = \mathfrak{E}H_1 \Rightarrow \mathfrak{E}(G_0 * H_0) = \mathfrak{E}(G_1 * H_1)$,

где $G_i * H_i$ означает свободное произведение произвольных групп G_i, H_i ($i = 0, 1$).

Ю. И. Мерзляков показал, что в $\mathbf{F}_n$ нет неабелевых подгрупп, представимых позитивными формулами. Однако остается неясным,

д) *есть или нет в* $\mathbf{F}_n$ *неабелевы формульные подгруппы, отличные от* $\mathbf{F}_n$ ($n \geqslant 2$)?

Еще в 1949 г. В. Шмелева доказала разрешимость элементарной теории класса всех абелевых групп. В 1964 г. Ю. Ш. Гуревич [22] доказал, что элементарная теория класса всех упорядоченных абелевых групп также разрешима. Им же найдены и условия совпадения элементарных теорий двух упорядоченных абелевых групп. А. И. Кокорин и Н. Г. Хисамиев [23] изучили элементарные теории структурно-упорядоченных абелевых групп. Ими были найдены условия совпадения $\mathfrak{E}$-теорий структурно-упорядоченных групп, имеющих конечное число нитей. Н. Г. Хисамиевым (1966) [24] было показано, что $\forall$-теория абелевых структурно-упорядоченных групп рекурсивна. Вопрос,

е) *рекурсивна или нет элементарная теория класса всех абелевых структурно-упорядоченных групп*, остался пока открытым [1]).

Класс абелевых полугрупп более сложный, чем класс абелевых групп. М. А. Тайцлин и А. Тарский показали, что элементарная теория класса абелевых полугрупп с сокращением нерекурсивна. В текущем году М. А. Тайцлин [25] нашел серию классов абелевых полугрупп, имеющих разрешимые $\mathfrak{E}$-теории. В частности, он обнаружил, что $\mathfrak{E}$-теория каждой отдельной конечно-порожденной абелевой полугруппы рекурсивна.

Для теории абелевых полугрупп имеет принципиальное значение следующая проблема изоморфизма:

ж) *существует ли алгоритм, позволяющий для любых двух полугрупповых конечных систем определяющих соотношений узнать, определяют ли эти системы в классе всех абелевых полугрупп изоморфные полугруппы?*

Для соотношений с двумя порождающими упомянутый алгоритм был известен (ср. Редеи [26]). М. А. Тайцлин указал соответствующий алгоритм для полугрупп с 4 порождающими. В классе абелевых полугрупп с сокращением проблема изоморфизма была решена Е. А. Халезовым. М. А. Тайцлин нашел большое число других важных классов абелевых полугрупп, в которых проблема изоморфизма также решается положительно. Если бы 10-я проблема Гильберта

[1]) В сентябре 1966 г. Ю. Ш. Гуревич обычными методами решил эту проблему отрицательно.

имела положительное решение, то и проблема изоморфизма для абелевых полугрупп решалась бы положительно. Верно ли обратное — неизвестно.

Согласно П. С. Новикову, проблема изоморфизма в многообразии всех групп решается отрицательно. В многообразии абелевых групп эта проблема решается положительно. Она решается положительно и в нескольких других многообразиях групп, обладающих тем свойством, что все конечно-порожденные их группы конечны. Как решается проблема изоморфизма в других многообразиях — неизвестно. В частности, неизвестно, как решается эта проблема в неабелевых полинильпотентных многообразиях групп и даже в многообразии метабелевых групп.

1.5. Проблема тождеств. Рекурсивность теории $\mathcal{J}\mathfrak{K}$ для какого-нибудь многообразия $\mathfrak{K}$ означает, что свободные алгебры в классе $\mathfrak{K}$ допускают конструктивное описание (см. [27]). Во многих случаях такое описание было найдено, например, если $\mathfrak{K}$ — класс всех колец, ассоциативных колец, колец Ли, решеток или произвольное полинильпотентное многообразие групп. Однако существуют конечно-аксиоматизируемые многообразия коммутативных луп, $\mathcal{J}$-теория которых нерекурсивна [27]. Интересно было бы выяснить,

а) *существуют ли конечно-аксиоматизируемые многообразия групп, обладающие нерекурсивной $\mathcal{J}$-теорией, и*

б) *существуют ли конечно-аксиоматизируемые многообразия ассоциативных (или лиевых) колец, имеющих нерекурсивную $\mathcal{J}$-теорию.*

В настоящее время неизвестно даже,

в) *является ли каждое многообразие групп конечно-аксиоматизируемым* (Б. Нейман) и

г) *является ли конечно-аксиоматизируемым каждое многообразие ассоциативных (лиевых) колец* (Шпехт).

1.6. Степени неразрешимости теорий. Из теоремы полноты Геделя следует, что для каждого рекурсивно-аксиоматизируемого класса систем $\mathfrak{K}$ теории $\mathcal{E}\mathfrak{K}$, $\mathcal{J}\mathfrak{K}$, $\mathcal{Q}\mathfrak{K}$, $\forall\mathfrak{K}$ заведомо рекурсивно-перечислимы. Какие степени неразрешимости могут иметь эти теории? В частности, какие степени неразрешимости могут иметь упомянутые теории для конечно-аксиоматизируемых многообразий $\mathfrak{K}$? Совсем нетрудно для каждой рекурсивно-перечислимой степени неразрешимости построить бесконечно-аксиоматизируемый класс $\mathfrak{K}$, у которого теория $\mathcal{E}\mathfrak{K}$ имеет заданную степень неразрешимости. Для конечно-аксиоматизируемых многообразий аналогичный результат получен Ханфом [8]. Тем не менее у всех естественных (т. е. построенных независимо от данной проблемы) конечно-аксиоматизируемых классов $\mathfrak{K}$, у которых степень неразрешимости теории $\mathcal{E}\mathfrak{K}$ известна, эта степень оказалась или нулевой (множество $\mathcal{E}\mathfrak{K}$ рекурсивно), или наивысшей $0'$.

В связи с изложенным представляет особенный интерес следующий вопрос (Гжегорчик):

а) *какова степень неразрешимости теории*

$$\mathcal{E}\langle \mathbf{C};\ +, \cdot, \exp\rangle,$$

где $\mathbf{C}$ *— совокупность вещественных чисел?*

Пусть $A_{\mathbf{T}}$ — совокупность подмножеств некоторого топологического пространства $\mathbf{T}$ и $\cup$, $'$, $^{-}$ суть операции объединения, дополнения и замыкания множеств. Рассмотрим теорию

$$\mathcal{T} = \mathcal{E}\langle A_{\mathbf{T}};\ \cup, ', ^{-}\rangle.$$

Если $\mathbf{T}$ — квадрат $(0,1) \times (0,1)$, то, согласно Гжегорчику [28], теория $\mathcal{T}$ нерекурсивна. Однако до сих пор неизвестно,

б) *какова степень неразрешимости $\mathcal{T}$, если $\mathbf{T}$ — простой интервал* $(0,1)$ (Гжегорчик).

Было бы интересно найти возможные степени неразрешимости проблемы изоморфизма в конечно-аксиоматизируемых многообразиях, а также, пользуясь понятиями метрической теории алгоритмов, найти степени сложности теорий $\mathcal{E}\mathfrak{K}$, $\mathcal{J}\mathfrak{K}$ в тех случаях, когда эти теории рекурсивны.

2. Многообразия и квазимногообразия

2.1. Решетки подмногообразий. Пусть фиксирован какой-нибудь тип Γ формул. Совокупность всех Γ-подклассов произвольного Γ-класса $\mathfrak{K}$ является полной решеткой относительно теоретико-множественного отношения включения.

Эту решетку мы условимся обозначать через $\mathbf{L}_\Gamma\mathfrak{K}$.

Атомы решетки $\mathbf{L}_\Gamma K\Omega$ называются Γ-минимальными или Γ-полными классами. Ясно, что Γ-класс $\mathfrak{K}$ тогда и только тогда $\mathbf{L}$-минимальный, когда решетка $\mathbf{L}_\Gamma\mathfrak{K}$ двухэлементная. Заметим, что наименьшим элементом в $\mathbf{L}_\Gamma\mathfrak{K}$ может оказаться пустой класс. Создание общей теории $\mathcal{E}$-полных классов (см. [29]) было, по-видимому, одним из главных событий в общей алгебре в последние годы.

С чисто алгебраической точки зрения такое же большое значение имело бы и создание, по возможности, детальной теории $\mathcal{J}$, $\mathcal{Q}$ и $\forall$ классов, наиболее простых с точки зрения логического языка, на котором эти классы задаются. Хотя удобные теоретико-множественные характеристики указанных классов хорошо известны [34], тем не менее эти характеристики могут служить лишь отправным пунктом исследований.

Началом общей теории многообразий алгебр можно считать статью Г. Биркгофа (1935) [4], а началом теории многообразий

групп — статью Б. Неймана (1937) [30]. Уделяемое обоим этим аспектам теории многообразий внимание резко увеличилось в 50-х годах. Прямой вопрос о явном описании решетки $\mathbf{L}_{\mathcal{Y}}\mathfrak{K}$ для классических многообразий $\mathfrak{K}$ оказался весьма трудным и был решен только для очень простых многообразий. Например, обозначим через $\mathfrak{N}_k$ многообразие всех k-ступенно нильпотентных и через $\mathfrak{S}_k$ многообразие всех k-ступенно разрешимых групп ($k = 1, 2, 3, \ldots$; $\mathfrak{N}_1 = \mathfrak{S}_1$ — многообразие абелевых групп). Давно известно, что решетка $\mathbf{L}_{\mathcal{Y}}\mathfrak{N}_1$ состоит из подмногообразий $\mathfrak{A}_m$, выделяемых в $\mathfrak{N}_1$ тождествами $x^m = 1$ ($m = 0, 1, 2, \ldots$), причем $\mathfrak{A}_d \supseteq \mathfrak{A}_m$ равносильно условию $d \mid m$. Решетка $\mathbf{L}_{\mathcal{Q}}\mathfrak{N}_1$ недавно найдена А. А. Виноградовым [35]. В отличие от счетной решетки $\mathbf{L}_{\mathcal{Y}}\mathfrak{N}_1$ решетка $\mathbf{L}_{\mathcal{Q}}\mathfrak{N}_1$ оказалась мощности континуума. В настоящее время полностью описаны решетки $\mathbf{L}_{\mathcal{Y}}\mathfrak{N}_2$, $\mathbf{L}_{\mathcal{Y}}\mathfrak{N}_3$ (см. [36], [37]) и частично решетка $\mathbf{L}_{\mathcal{Y}}\mathfrak{S}_2$. В связи с некоторыми гипотезами о строении решеток представляется важным

а) *найти полное описание решетки* $\mathbf{L}_{\mathcal{Y}}\mathfrak{S}_2$ (Б. Нейман и Х. Нейман), *а также найти строение решетки* $\mathbf{L}_{\mathcal{Y}}\mathfrak{N}_4$.

Мало что известно о строении решетки $\mathbf{L}_{\mathcal{Y}}\mathfrak{L}$, где $\mathfrak{L}$ — многообразие всех решеток. Наиболее известные ее элементы — это многообразие всех модулярных решеток и многообразие дистрибутивных решеток. Последнее из них является единственным атомом в $\mathbf{L}_{\mathcal{Y}}\mathfrak{L}$. Другие многообразия решеток указаны Икбалуннизой [39], а также Лёвигом [40]. По-видимому, пока не потеряна надежда, что решетка $\mathbf{L}_{\mathcal{Y}}\mathfrak{L}$ или решетка $\mathbf{L}_{\mathcal{Y}}\mathfrak{M}$ ($\mathfrak{M}$ — класс модулярных решеток) не очень сложна и удастся

б) *найти описание* $\mathbf{L}_{\mathcal{Y}}\mathfrak{L}$ *или* $\mathbf{L}_{\mathcal{Y}}\mathfrak{M}$.

Из общих соображений вытекает, что в каждом многообразии содержится хотя бы одно $\mathcal{Y}$-полное многообразие и в каждом квазимногообразии содержится хотя бы одно $\mathcal{Q}$-полное квазимногообразие. В частности, каждое минимальное многообразие содержит минимальное квазимногообразие. Однако не каждое минимальное квазимногообразие содержится в подходящем минимальном многообразии. Поэтому число минимальных квазимногообразий в произвольном многообразии $\mathfrak{K}$ больше или равно числу минимальных многообразий, содержащихся в $\mathfrak{K}$.

Атомы решетки $\mathbf{L}_{\mathcal{Y}}\mathfrak{K}$ для многих классических многообразий $\mathfrak{K}$ были в явном виде найдены Калицким и Тарским. Калицкий доказал также, что многообразие всех группоидов содержит континуум минимальных подмногообразий. В текущем году этот результат был усилен Больботом (Новосибирск), показавшим, что многообразие группоидов, определяемое тождествами $x \cdot xy = yx \cdot x = x$, также содержит континуум минимальных подмногообразий.

Из теоремы компактности следует, что решетки $\mathbf{L}_{\mathcal{E}}\mathfrak{K}$, $\mathbf{L}_{\mathcal{Y}}\mathfrak{K}$, $\mathbf{L}_{\mathcal{Q}}\mathfrak{K}$ не могут иметь произвольное строение. Спрашивается,

в) *какие решетки могут быть реализованы в виде решеток* $\mathbf{L}_{\mathcal{Y}}\mathfrak{K}$, $\mathbf{L}_{\mathcal{Q}}\mathfrak{K}$ *для подходящих многообразий (квазимногообразий)* $\mathfrak{K}$?

2.2. Группоиды квазимногообразий. Новый подход к изучению многообразий групп нашла в 1956 г. Х. Нейман [31]. Она ввела ассоциативную операцию перемножения многообразий групп и предложила вместо решетки $\mathbf{L}_{\mathcal{Y}}\mathfrak{G}$ изучать полугруппу $\mathbf{G}_{\mathcal{Y}}\mathfrak{G}$ всех групповых многообразий относительно упомянутого умножения. Опираясь на результаты цитированной статьи [31], Б. Нейман, Х. Нейман и П. Нейман (1962) [32] и независимо от них А. Шмелькин (1963) [33] показали, что $\mathbf{G}_{\mathcal{Y}}\mathfrak{G}$ является свободной полугруппой с нулем и единицей. Однако мощность этой полугруппы (заведомо бесконечная и не превышающая мощности континуума) остается все еще неизвестной.

Вероятно, целесообразно по аналогии с умножением многообразий групп ввести следующее $\mathfrak{K}$-умножение любых подклассов произвольного фиксированного класса $\mathfrak{K}$ алгебраических систем. А именно для любых $\mathfrak{A} \subseteq \mathfrak{K}$, $\mathfrak{L} \subseteq \mathfrak{K}$ полагаем $\mathbf{A} \in \mathfrak{A} \mathbin{\mathring{}}_{\mathfrak{K}} \mathfrak{L}$, если $\mathbf{A} \in \mathfrak{K}$ и существует такая факторсистема $\mathbf{A}/\theta$, что $\mathbf{A}/\theta \in \mathfrak{L}$ и любой смежный класс $a\theta$ ($a \in \mathbf{A}$), являющийся $\mathfrak{K}$-подсистемой в $\mathbf{A}$, принадлежит $\mathfrak{A}$ (см. [41]). Легко показывается, что если $\mathfrak{K}$ есть $\mathcal{V}$-класс или $\mathcal{Q}$-класс конечной сигнатуры, то $\mathfrak{K}$-произведение любых двух его $\mathcal{V}$-подклассов (соответственно $\mathcal{Q}$-подклассов) является снова $\mathcal{V}$-подклассом ($\mathcal{Q}$-подклассом). Таким образом, наряду с решетками $\mathbf{L}_{\mathcal{V}}\mathfrak{K}$, $\mathbf{L}_{\mathcal{Q}}\mathfrak{K}$ в указанных случаях можно рассматривать группоиды $\mathbf{G}_{\mathcal{V}}\mathfrak{K}$, $\mathbf{G}_{\mathcal{Q}}\mathfrak{K}$. Интересно отметить, что даже в случае, когда $\mathfrak{K}$ — многообразие полугрупп, $\mathfrak{K}$-произведение подмногообразий может не быть подмногообразием. Однако если на всех алгебрах какого-нибудь многообразия $\mathfrak{K}$ конгруэнции перестановочны и существует терм, все значения которого совпадают и образуют одноэлементную подалгебру в каждой $\mathfrak{K}$-алгебре, то $\mathfrak{K}$-произведение подмногообразий будет подмногообразием в $\mathfrak{K}$, т. е. в этом случае группоид $\mathbf{G}_{\mathcal{Q}}\mathfrak{K}$ будет содержать $\mathbf{G}_{\mathcal{Y}}\mathfrak{K}$ в качестве своего подгруппоида. Можно указать и условия, при выполнении которых группоиды $\mathbf{G}_{\mathcal{Q}}\mathfrak{K}$, $\mathbf{G}_{\mathcal{Y}}\mathfrak{K}$ будут ассоциативны (см. [41]).

Пусть $\mathfrak{K}$ — квазимногообразие конечной сигнатуры и $\mathfrak{L} \in \mathbf{G}_{\mathcal{Q}}\mathfrak{K}$. Ясно, что в общем случае $\mathfrak{L}$-произведение двух подквазимногообразий из $\mathfrak{L}$ будет отличаться от $\mathfrak{K}$-произведения их и потому $\mathbf{G}_{\mathcal{Q}}\mathfrak{L}$ не будет подгруппоидом группоида $\mathbf{G}_{\mathcal{Q}}\mathfrak{K}$. Если же окажется, что $\mathfrak{L} \mathbin{\mathring{}}_{\mathfrak{K}} \mathfrak{L} = \mathfrak{L}$, то $\mathbf{G}_{\mathcal{Q}}\mathfrak{L}$ будет подгруппоидом в $\mathbf{G}_{\mathcal{Q}}\mathfrak{K}$. Это показывает,

что, помимо нахождения общей структуры группоидов $G_{\mathfrak{A}}\mathfrak{K}$, особый интерес представляет задача нахождения идемпотентов этих группоидов.

С последней задачей тесно связана задача Т. Тамуры [38] о нахождении так называемых достижимых подквазимногообразий в данном квазимногообразии. Действительно, легко заметить, что из достижимости подквазимногообразия $\mathfrak{A}$ в $\mathfrak{K}$ вытекает, что

$$(\mathfrak{X} \circ \mathfrak{A}) \circ \mathfrak{A} = \mathfrak{X} \circ \mathfrak{A}$$

для произвольного $\mathfrak{X} \in G_{\mathfrak{A}}\mathfrak{K}$. При некоторых условиях верно и и обратное.

Из сказанного выше вытекает, в частности, что если $\mathfrak{K}$ — многообразие колец, луп или коммутативных луп, то подмногообразия $\mathfrak{K}$ образуют группоид относительно $\mathfrak{K}$-умножения. Возможно, что некоторые из этих группоидов будут иметь не очень сложную структуру.

Сибирское отделение АН СССР,
Новосибирск, СССР

ЛИТЕРАТУРА

[1] Cohn P. M., Universal algebra, Harper & Row., N.Y., 1965.
[2] Gratzer G., Universal algebra (preprint, 1966).
[3] Conférence sur l'Algèbre générale, Varsovie 7. IX—11. IX, 1964, *Coll. Math.*, **14** (1966).
[4] Birkhoff G., On the structure of abstract algebras, *Proc. Cambridge Phill. Soc.*, **31** (1935), 433-454.
[5] Tarski A., Der Wahrheitsbergiff in der formalizierten Spachen, *Studia Philosophica*, **1** (1936), 261-404.
[6] Мальцев А. И., Untersuhungen aus dem Gebiete der mathematische Logik, *Матем. сб.*, **1** (1936), 323-336.
[7] Мальцев А. И., Об одном общем методе получения локальных теорем теории групп, Уч. зап. Ивановского пединст., **1** (1941), 3-9.
[8] The Theory of Models, Proceedings of the 1963 International Symposium at Berkeley, Amsterdam, 1965.
[9] Schwenkel F., Rekursive Wortfunktionen über Unendlichen Alfabeten, *Zeitschrift math. Logik und Grundl. d. Math.*, **11** (1965), 133-147.
[10] Church A., A note on the Entscheidungsproblem, *J.S.L.*, **1** (1936), 40-41, 101-102.
[11] Трахтенброт Б. А., Невозможность алгорифма для проблемы разрешимости на конечных классах, *ДАН СССР*, **70** (1950), 569-572.
[12] Гуревич Ю. Ш., Об эффективном распознавании выполнимости формул УИП, *Алгебра и логика*, **5**, № 2 (1966), 25-56.
[13] Robinson J., Definability and decision problems in arithmetics, *J.S.L.*, **14** (1946), 98-114.
[14] Tarski A., Mostowski A., Robinson R., Undecidable theories, Amsterdam, 1953.
[15] Ершов Ю. Л., Лавров И. А., Тайманов А. Д., Тайцлин М. А., Элементарные теории, *УМН*, **20**, № 4 (1965), 37-108.

[16] Robinson A., Introduction to Model Theory and to the Metamathematics of Algebra, Amsterdam, 1963.
[17] Davis M., Putnam H., Robinson J., The decision problem for exponential Diophantine equations, *Ann. of Math.*, **74** (1961), 425-436.
[18] Ax J., Kochen S., Diophantine problems over local fields, I, II, III (preprint, 1964).
[19] Ершов Ю. Л., Об элементарных теориях локальных полей, *Алгебра и логика*, **4**, № 2 (1965), 5-30.
[20] Ершов Ю. Л., Об элементарной теории максимальных нормированных полей, *Алгебра и логика*, **4**, № 3 (1965), 31-78.
[21] Ершов Ю. Л., Об элементарной теории максимальных нормированных полей, II, *Алгебра и логика*, **5**, № 1 (1966), 5-40.
[22] Гуревич Ю. Ш., Элементарные свойства упорядоченных абелевых групп, *Алгебра и логика*, **3**, № 1 (1964), 5-40.
[23] Кокорин А. И., Хисамиев Н. Г., Элементарная классификация структурно-упорядоченных абелевых групп с конечным числом нитей, *Алгебра и логика*, **5**, № 1 (1966), 41-50.
[24] Хисамиев Н. Г., Универсальная теория структурно-упорядоченных абелевых групп, *Алгебра и логика*, **5**, № 3 (1966), 71-76.
[25] Тайцлин М. А., Об элементарных теориях коммутативных полугрупп, *Алгебра и логика*, **5**, № 4 (1966), 55-89.
[26] Rédei L., Theorie der endlich erzeugbaren kommutativen Halbgruppen, Leipzig, 1963.
[27] Мальцев А. И., Тождественные соотношения на многообразиях квазигрупп, *Матем. сб.*, **69**, № 1 (1966), 3-12.
[28] Grzegorczyk A., Undecidability of some topological theories, *Fund. Math.*, **38** (1951), 137-152.
[29] Vaught R. L., Models of complete theories, *Bull. Am. Math. Soc.*, **69**, № 3 (1963), 299-313.
[30] Neumann B. H., Identical relations in groups, I, *Math. Ann.*, **114** (1937), 506-525.
[31] Neumann H., On varieties of groups and their associated near-rings, *Math. Z.*, **65**, № 1 (1956), 36-69.
[32] Neumann B. H., Neumann H., Neumann P. M., Wreath products and varieties of groups, *Math. Z.*, **80** (1962), 44-62.
[33] Шмелькин А. Л., Полугруппа многообразий групп, *ДАН СССР*, **146**, № 3 (1963), 543-545.
[34] Мальцев А. И., Несколько замечаний о квазимногообразиях алгебраических систем, *Алгебра и логика*, **5**, № 3 (1966), 3-10.
[35] Виноградов А. А., Квазимногообразия абелевых групп, *Алгебра и логика*, **4**, № 6 (1965), 15-20.
[36] Ремесленников В. Н., Два замечания о трехступенно нильпотентных группах, *Алгебра и логика*, **4**, № 2 (1965), 59-66.
[37] Jonsson B., Varieties of groups of nilpotency three, *Notices AMS*, **13**, № 4 (1966), 488.
[38] Tamura T., Attainability of systems of identities on semigroups, *Journal of Algebra*, **3**, № 3 (1966), 261-276.
[39] Iqbalunnisa, On types of lattices, *Fund. Math.*, **59**, № 1 (1966), 97-102.
[40] Löwig H., On the importance of the relation $[(A, B), (A, C)] < < (A, [(B, C), (C, A), (A, B)])$ between three elements of a structure, *Ann. of Math.*, **44** (1943), 573-579.
[41] Мальцев А. И., Об умножении классов алгебраических систем, *Сиб. матем. журнал*, **8**, № 2 (1967).

АВТОМОРФНЫЕ ФУНКЦИИ И АРИФМЕТИЧЕСКИЕ ГРУППЫ

И. И. ПЯТЕЦКИЙ-ШАПИРО

Введение

Теория автоморфных функций от многих переменных совершенно не аналогична классической теории автоморфных функций от одного комплексного переменного. Одна из причин этого в том, что все алгебраические кривые униформизируемы, в то время как в случае многих переменных алгебраические многообразия, как правило, не униформизируемы. Аналогичная ситуация имеет место для дискретных групп. В случае одного комплексного переменного дискретные группы сравнительно просто строятся и описываются с помощью геометрических соображений. В многомерном случае соответствующие дискретные группы строятся, как правило, с помощью арифметических соображений. Весьма вероятно, что, за небольшим числом исключений, все дискретные подгруппы полупростых групп Ли (не только вещественных, но даже и p-адических) являются арифметическими группами. Точная формулировка этой гипотезы, обсуждение исключений из нее и некоторые обобщения даются ниже в п. 3 и 11 настоящего доклада.

Совершенно не ясны в настоящее время истинное место и роль многомерных униформизируемых алгебраических многообразий, не ясен даже язык, на котором можно было бы описать эти многообразия. Складывается впечатление, что стандартный язык алгебраической геометрии не подходит для этой цели. Вполне возможно, что для решения этой проблемы нужно алгебраизовать теорию автоморфных функций, т. е. придать ей форму, в которой она бы имела смысл над произвольным полем констант.

Классическая теория абелевых функций полностью растворилась в современной алгебраической теории абелевых многообразий над произвольным полем (см. также работу Мамфорда [25]). Подобно этому современная теория комплексно-аналитических автоморфных функций, быть может, лишь должна послужить источником для какой-нибудь более алгебраической теории (например, теории алгебраических многообразий, обладающих квазиоднородными накрытиями; см. п. 9).

В классической теории автоморфных функций от одного комплексного переменного возможна и другая концепция, в которой на первое место выдвигается не процесс униформизации алгебраической кривой, а способы конструкции дискретной группы. Эта концепция также подверглась в последние годы значительному развитию и обобщению. Прежде всего отметим, что с точки зрения дискретных групп было неестественно ограничиться только случаями, связанными с комплексно-аналитическими автоморфными функциями; более естественной является задача исследования дискретных подгрупп всех групп Ли, в первую очередь полупростых групп Ли. По-видимому, нецелесообразно ограничиться только вещественными группами Ли; p-адические группы Ли должны занять равноправное место наряду с вещественными группами Ли. Не исключено, что наиболее интересные арифметические приложения будут связаны именно с p-адическими группами Ли.

Задача исследования свойств комплексно-аналитических автоморфных функций и автоморфных форм трансформировалась в задачу исследования всех функций на данной топологической группе G, которые инвариантны относительно данной дискретной группы Γ. В основном здесь интересно разложить на неприводимые представления представление группы G в пространстве $L^2(\Gamma \backslash G)$.

Наиболее содержательным здесь является случай, когда G — группа аделей некоторой редуктивной алгебраической группы, определенной над Q, а Γ — ее подгруппа главных аделей. В этом случае естественно возникают некоторые дзета-функции, тесно связанные с классической дзета-функцией Римана, а иногда даже явно выражающиеся через нее. Отметим также, что известная гипотеза Петерсона о собственных значениях операторов Гекке очень естественно интерпретируется в терминах свойств неприводимых представлений, входящих в представление в пространстве $L^2(\Gamma \backslash G)$; см. [9], [22]. Более подробно об этом будет сказано в п. 6.

1. Области существования автоморфных функций

Пусть D — ограниченная область в n-мерном комплексном пространстве, а Γ — дискретная группа аналитических автоморфизмов (взаимно однозначных аналитических отображений) области D на себя. Мероморфная в D функция, инвариантная относительно всех преобразований из группы Γ, называется автоморфной функцией. В случае когда D — единичный круг, накладывается дополнительное ограничение мероморфности в параболических вершинах границы.

Наиболее важный класс областей для теории автоморфных функций — это так называемые ограниченные симметрические области. Как известно, область D называется симметрической, если для каждой ее точки z_0 существует инволюция, т. е. автоморфизм φ_0 области D, квадрат которого равен единице и у которого нет в области D неподвижных точек, за исключением точки z_0. Всякая ограниченная симметрическая область однородна, но обратное неверно. Каждая ограниченная симметрическая область является одновре-

менно симметрическим римановым пространством. Пользуясь этим, Э. Картан [1] полностью расклассифицировал все ограниченные симметрические области.

Типичным примером симметрических областей является так называемый круг Зигеля. Пусть p — некоторое положительное целое число. Рассмотрим $\frac{p(p+1)}{2}$-мерное комплексное пространство, точками которого служат симметрические матрицы Z порядка p. Обозначим через K_p совокупность таких матриц Z, у которых все собственные значения матрицы Z меньше 1. K_p есть круг Зигеля.

Упомянем здесь также об ограниченных однородных областях. Полная классификация и исследование свойства однородных областей были получены сравнительно недавно [15], [16], [17]. Оказалось, что они обладают рядом свойств, сходных со свойствами симметрических областей. Например, все однородные области обладают реализацией в виде неограниченной аффинно однородной области, аналогичной реализации, которая существует для симметрических областей. Все однородные области гомеоморфны внутренности шара и т. д. Однако однородных областей значительно больше, чем симметрических, например число симметрических областей данной размерности конечно, а однородных областей, начиная с размерности 7, уже континуум [17].

Хотя однородные области обладают многими свойствами, аналогичными свойствам симметрических областей, их непосредственная роль в теории автоморфных функций, видимо, невелика. Это объясняется следующим. Наиболее важную роль играет класс дискретных групп с конечным объемом фундаментальной области. Таких дискретных групп нет в однородных несимметрических областях.

Насколько знает автор, в настоящее время неизвестно, могут ли существовать неоднородные ограниченные области с дискретными группами аналитических автоморфизмов с конечным объемом [1]) фундаментальной области.

Очень возможно, что пространство Тайхмюллера является такой областью. Эту гипотезу высказывал профессор Берс. Не исключено, что такие области можно построить, рассматривая подходящие подмногообразия симметрических областей, аналогично тому, что всякую ограниченную однородную область можно вложить в виде комплексно-аналитического подмногообразия в симметрическую область, например в круг Зигеля K_p при некотором достаточно большем p. Это вложение даже можно сделать однородным, т. е. таким, что подгруппа группы всех аналитических автоморфизмов области K_p, переводящая данное подмногообразие в себя, транзитивна на нем [15].

[1]) Имеется в виду объем в смысле меры, связанной с метрикой Бергмана.

2. Поля автоморфных функций

Совокупность всех автоморфных относительно данной дискретной группы функций, очевидно, образует поле. Какова структура этого поля? Это один из центральных вопросов классической теории комплексно-аналитических автоморфных функций. В типичных случаях это поле является полем алгебраических функций от такого же числа переменных, какова размерность области D.

Пусть D — единичный круг: $|z| < 1$, Γ — дискретная группа дробно-линейных преобразований D на себя. Обозначим через P поле автоморфных функций относительно группы Γ. Хорошо известно, что поле P является полем алгебраических функций от одного переменного тогда и только тогда, когда факторпространство D/Γ имеет конечную площадь в смысле меры Лобачевского. Весьма вероятно, что аналогичная теорема справедлива для любой ограниченной симметрической области или даже для любой ограниченной области D. Однако, насколько известно автору, это до сих пор не доказано.

3. Арифметические группы

Наиболее важный класс дискретных групп в симметрических областях — это арифметические группы. Типичным примером арифметической группы может служить модулярная группа, т. е. группа дробно-линейных преобразований верхней полуплоскости $\operatorname{Im} z > 0$ с целыми коэффициентами и определителем 1:

$$z \to \frac{az+b}{cz+d}.$$

Арифметические группы определяются как множества целых точек некоторой алгебраической группы, определенной над полем рациональных чисел Q.

Напомним вначале определение алгебраической группы. Пусть $GL(n, \mathbf{C})$ — совокупность всех комплексных невырожденных матриц порядка n. Алгебраической группой называется подгруппа группы $GL(n, \mathbf{C})$, выделяемая условиями обращения в нуль некоторого числа полиномов от элементов матриц. Если эти полиномы можно выбрать так, чтобы все их коэффициенты принадлежали полю рациональных чисел Q, то говорят, что эта группа определена над полем Q. Если k — некоторое кольцо, то через G_k обозначается множество точек группы G, определенных над k, определитель которых есть единица кольца k. Подгруппа $G_{\mathbf{Z}}$ ($\mathbf{Z}$ — кольцо целых чисел) называется арифметической подгруппой группы $G_{\mathbf{R}}$. Любая дискретная подгруппа Γ группы $G_{\mathbf{R}}$, пересечение которой с подгруппой $G_{\mathbf{Z}}$

представляет подгруппу, имеющую конечный индекс как в Γ, так и в $G_{\mathbf{Z}}$, также называется арифметической подгруппой. Отметим еще следующее. Пусть группа $G_{\mathbf{R}}$ представляет собой прямое произведение компактной группы K и некоторой группы G_1. Обозначим через φ естественную проекцию $G_{\mathbf{R}}$ на G_1. Нетрудно видеть, что φ переводит множество, дискретное в $G_{\mathbf{R}}$, в множество, дискретное в G_1. Следовательно, $\varphi(G_{\mathbf{Z}})$ представляет собой дискретную подгруппу группы G_1. Дискретные группы, получаемые такой конструкцией из арифметических групп, также называются арифметическими группами. Таким образом, окончательно для описания всех арифметических подгрупп данной вещественной полупростой группы Ли G_1 нужно найти все алгебраические группы G, определенные над Q, и такие, что $G_{\mathbf{R}} = K \times G_1$, где K — некоторая компактная группа Ли.

Фундаментальный результат принадлежит А. Борелю и Харишу-Чандра [4]. Они показали, что если G — полупростая группа Ли и Γ — ее арифметическая подгруппа, то объем факторпространства $\Gamma \backslash G$ конечен.

Давно известно, как построить дискретные группы в плоскости Лобачевского с конечной площадью фундаментальной области. В других случаях очень долго не удавалось построить такие примеры. А. Сельберг высказал гипотезу, что все дискретные подгруппы Γ полупростых вещественных групп Ли с конечным объемом факторпространства, за небольшим числом исключений, являются арифметическими группами. В пользу этого предположения говорят полученные недавно замечательные результаты А. Вейля [20] и А. Сельберга [3].

А. Вейль показал, что пространство несопряженных дискретных подгрупп вещественной простой группы Ли G с компактным факторпространством состоит из изолированных точек (если G отлично от группы вещественных матриц второго порядка).

А. Сельберг при тех же предположениях показал, что существует представление группы G, при котором все элементы группы Γ записываются матрицами с алгебраическими элементами. Последний результат с первого взгляда кажется эквивалентным гипотезе о том, что группа Γ арифметическая. Хотя это впечатление и обманчиво, все же надежда использовать этот результат для доказательства гипотезы А. Сельберга остается.

В настоящее время гипотезу А. Сельберга удалось доказать лишь при некоторых дополнительных предположениях. Наиболее интересный результат в этом направлении принадлежит самому А. Сельбергу, который рассматривал дискретные группы в произведении n $(n > 1)$ плоскостей Лобачевского. Этот результат был объявлен А. Сельбергом на международной конференции по теории функций комплексного переменного в Ереване в 1965 г.

Другой результат в этом направлении принадлежит докладчику. Он состоит в том, что всякая дискретная подгруппа типа Гекке простой группы Ли G, расщепляемой над $\mathbf{R}$, является арифметической, если ранг группы G больше чем единица.

Прежде чем привести определение группы типа Гекке, нам придется напомнить ряд понятий. Пусть G — простая вещественная группа Ли, A — максимальная коммутативная подгруппа группы G, диагонализуемая над полем вещественных чисел. Обозначим через $Z(A)$ и $N(A)$ централизатор и нормализатор группы A. Группу $S = N(A)/Z(A)$ принято называть группой Вейля группы G.

Дискретную подгруппу Γ группы G условимся называть группой типа Гекке, если существуют такая максимальная нильпотентная подгруппа Z и максимальная диагонализуемая над полем вещественных чисел $\mathbf{R}$ подгруппа A, что

1) факторпространство Z/Δ (где $\Delta = \Gamma \cap Z$) компактно,
2) A принадлежит нормализатору Z,
3) факторгруппа $N(A) \cap \Gamma / Z(A) \cap \Gamma$ совпадает с S.

Например, пусть G — группа вещественных матриц n-го порядка с определителем 1. Обозначим через Z верхнюю треугольную подгруппу группы G, а через A — диагональную подгруппу группы G.

Пусть дискретная подгруппа Γ содержит две подгруппы Δ и W, такие, что

1) $\Delta \subset Z$ и факторпространство Z/Δ компактно,
2) $W \in N(A)$, порядок группы $W/W \cap Z(A)$ равен $n!$

Тогда группа Γ является группой типа Гекке.

После того как мы привели и пояснили полную формулировку, скажем несколько слов о методе доказательства. Центральный пункт доказательства — это построение по дискретной группе Γ некоторой алгебры Ли над Q. Разумеется, такая конструкция не не имеет смысла в общей ситуации. Однако в условиях теоремы дискретные группы оказываются корневыми, т. е. их пересечение с некоторой системой соответствующих корневых подгрупп нетривиально. Пользуясь этим, можно определить некоторую алгебру Ли $\mathfrak{A}$ над Q. Конструкцию этой алгебры Ли можно провести, даже если в условиях теоремы отказаться от требования, чтобы ранг группы G был больше единицы. Однако в этом случае размерность алгебры $\mathfrak{A}$ над Q может оказаться даже бесконечной. В случае же когда ранг группы G больше чем единица, можно показать, что размерность алгебры $\mathfrak{A}$ над Q равна тому, чему ей положено быть равным, т. е. вещественной размерности группы G. Отметим, что идея конструкции алгебры Ли по дискретной группе, быть может, обобщается на случай, когда факторпространство $\Gamma \backslash G$ компактно.

4. Примеры неарифметических дискретных групп

Первый пример неарифметической дискретной группы движений в трехмерном пространстве Лобачевского с конечным объемом фундаментальной области принадлежит В. С. Макарову [28]. Э. Б. Винбергу [29] удалось обобщить конструкцию Макарова и построить аналогичные примеры в 4- и 5-мерных пространствах Лобачевского.

Все эти примеры представляют собой группы, порожденные отражениями. Общую теорию групп, порожденных отражениями в евклидовом и псевдоевклидовом пространстве, разработал Коксетер. Э. Б. Винбергу удалось успешно использовать ее для построения примеров неарифметических дискретных групп.

Отметим, что построенные дискретные группы, по-видимому, обладают жесткостью, хотя это и не доказано в настоящее время. Э. Б. Винбергу удалось также построить пример неарифметической дискретной группы с компактным факторпространством в 3- и 4-мерных пространствах Лобачевского.

5. Поля автоморфных функций относительно арифметических групп

Пусть D — ограниченная симметрическая область в n-мерном комплексном пространстве. Группа G всех аналитических автоморфизмов области D представляет собой, как хорошо известно, полупростую вещественную группу Ли. Условимся арифметические подгруппы группы G называть арифметическими группами аналитических автоморфизмов.

Обозначим через $P(\Gamma)$ поле всех функций в D, автоморфных относительно данной арифметической группы Γ. Сравнительно недавно было доказано [21], [22], [14], что это поле всегда представляет собой поле алгебраических функций от n неизвестных, где n — комплексная размерность области D.

Весьма вероятно также, что поле определения автоморфных функций относительно арифметической группы представляет собой поле алгебраических чисел. Однако, насколько известно докладчику, это в общем виде еще не доказано. Для многих важных случаев это доказано в работах Шимуры, Бейли и др.

Остановимся вкратце на способе доказательства теоремы о том, что поле всех автоморфных функций относительно данной арифметической группы является полем алгебраических функций. Эта теорема в дальнейшем будет называться теоремой об алгебраических соотношениях.

Как известно, поле мероморфных функций на n-мерном компактном комплексном многообразии всегда представляет собой поле

алгебраических функций от r неизвестных, где $r \leqslant n$. Из этой теоремы сразу следует нужное нам утверждение, в случае, когда факторпространство D/Γ компактно. Нужно воспользоваться тем, что всякая автоморфная функция может рассматриваться как мероморфная функция на D/Γ и обратно. В случае когда факторпространство D/Γ некомпактно, доказательство теоремы значительно сложней. Естественно было бы попробовать построить компактификацию, т. е. компактное комплексное многообразие, в которое D/Γ вкладывалось бы в виде всюду плотного аналитического подмногообразия. Однако пока такую компактификацию не удалось построить для любой арифметической группы. Удалось доказать лишь более слабое утверждение, которое, впрочем, достаточно для доказательства теоремы об алгебраических соотношениях, а именно что существует компактификация, являющаяся нормальным аналитическим пространством, причем коразмерность добавляемого многообразия не меньше 2.

Другой подход к доказательству теоремы об алгебраических соотношениях состоит в том, чтобы проверить условия псевдовыпуклости, которые были найдены Андреотти и Грауэртом [24]. Это было проделано в работе [14]. Поскольку, однако, такая проверка хотя и проще, но основана примерно на тех же соображениях, что и построение компактификации, мы ограничимся изложением лишь последней. Конструкция компактификации основана на правильном обобщении понятия параболической вершины.

Как известно, граница любой ограниченной симметрической области D состоит из аналитических кусочков (компонент) разной размерности. Каждая из компонент аналитически эквивалентна некоторой ограниченной симметрической области в пространстве меньшего числа измерений. Обобщением понятия параболической вершины служит понятие компоненты, рациональной относительно данной дискретной группы Γ или, более кратко, Γ-рациональной компоненты. В случае одного переменного компактификация, как хорошо известно, производится путем добавления к фундаментальной области некоторого числа параболических вершин. В случае многих переменных компактификация производится путем добавления некоторого числа Γ-рациональных компонент границы. Приведем теперь определение Γ-рациональной компоненты границы. Отметим, что это понятие Γ-рациональной компоненты может быть определено для любой дискретной группы Γ аналитических автоморфизмов области D. Можно также указать процесс аналитического расширения функциональной области и показать, что он приводит к некоторому аналитическому нормальному пространству. Однако показать, что это пространство является компактным, удалось лишь для арифметических групп [14], [21]. Ниже мы ограничимся определением Γ-рациональной компоненты лишь для случая, когда

Γ есть арифметическая группа. Пусть F — некоторая компонента границы области D. Без ограничения общности можно считать, что Γ есть множество всех целых точек группы $G_{\mathbf{Z}}$, где G — некоторая алгебраическая группа, определенная над Q, причем группа $\mathfrak{G}$ всех аналитических автоморфизмов области D совпадает с $G_{\mathbf{R}}$. Обозначим через $\mathfrak{G}(F)$ группу всех аналитических автоморфизмов области D, переводящих F в себя. Условимся говорить, что компонента F является Γ-рациональной компонентой, если группа $\mathfrak{G}(F)$ представляет собой подгруппу группы $G_{\mathbf{R}}$, определенную над полем рациональных чисел Q.

При проведении различных доказательств, связанных с процессом компактификации, весьма полезным является то, что с каждой компонентой границы связано однородное расслоение, базой которого является данная компонента границы. Это расслоение можно описать, например, следующим образом. Можно показать, что для каждой геодезической $z(t)$ существует $\lim_{t\to+\infty} z(t) = z(\infty)$, являющийся точкой границы области D. Оказывается, что любую точку области D можно соединить геодезической в точности с одной точкой данной компоненты границы. Тем самым возникает естественное отображение области D на данную компоненту границы. Это отображение и задает расслоение. Таким образом, слой, соответствующий данной точке компоненты, состоит из всех точек области D, которые можно соединить с этой точкой геодезическими. Можно показать, что каждый слой аналитически эквивалентен некоторой ограниченной однородной области, вообще говоря, не являющейся симметрической. Именно таким образом был построен первый пример [18] ограниченной несимметрической однородной области. Отметим еще, что указанные выше расслоения являются однородными, т. е. группа аналитических автоморфизмов, сохраняющих данное расслоение, транзитивна в области D. Можно показать, что все однородные расслоения области D получаются указанной выше конструкцией. Полное описание всех однородных расслоений всех ограниченных однородных областей было получено в [15].

Пользуясь понятием однородного расслоения, можно описать компактификацию еще другим способом: добавляемые точки можно интерпретировать как аналитические подмногообразия D, а именно как слои Γ-рациональных однородных расслоений. Существует и другое более прямое описание подмногообразий, являющихся слоями Γ-рациональных однородных расслоений. Для того чтобы его провести, нужно сформулировать два определения: а) ограниченной голоморфной оболочки, б) некоторого специального класса разрешимых подгрупп, называемых подгруппами Сатаке. Эти определения таковы.

Пусть X — некоторое множество точек области D. Ограниченной голоморфной оболочкой $O(X)$ множества X называется совокупность всех точек $z \in D$, таких, что для любой аналитической и ограниченной в D функции $\varphi(z)$ справедливо неравенство

$$|\varphi(z)| \leqslant \sup_{x\in X} |\varphi(x)|.$$

Подчеркнем, что требование ограниченности в D функции $\varphi(z)$ весьма существенно.

Приведем теперь определение подгруппы Сатаке. Пусть N — разрешимая подгруппа группы G, определенная над Q. Предположим, что подгруппа N расщепима над Q. Условимся называть подгруппу N подгруппой Сатаке, если она является в своем нормализаторе максимальным разрешимым и расщепимым над Q нормальным делителем.

После того как эти определения приведены, остается выполнить обещание и дать с их помощью иное описание слоев Γ-рациональных однородных расслоений. Оно состоит в том, что слои Γ-рациональных однородных расслоений суть ограниченные голоморфные оболочки орбиты подгруппы Сатаке.

6. Представление в пространстве $L^2(\Gamma \backslash G)$

Пусть G — некоторая топологическая группа, Γ — ее дискретная подгруппа, а $X = \Gamma \backslash G$ — факторпространство. Пусть $v(\Gamma \backslash G) < \infty$. Рассмотрим совокупность $H = L^2(\Gamma \backslash G)$ всех функций $f(x)$ на X с суммируемым квадратом. В H, естественно, действует унитарное представление группы G, определяемое следующим образом:

$$T_g f(x) = f(xg).$$

Как правило, это представление приводимо.

Вопросу, как разлагать это представление на неприводимое, посвящен обзорный доклад И. М. Гельфанда на прошлом конгрессе [5]. Поэтому мы здесь остановимся только на одном довольно частном вопросе. Петерсону принадлежит гипотеза, что собственные значения операторов Гекке T_p в пространстве модулярных форм не превосходят $2\sqrt{p}$. Оказывается, что эта гипотеза эквивалентна следующему факту, связанному со структурой представлений в пространстве $G_Q \backslash G_A$ для случая, когда G — группа матриц второго порядка. Именно каждое неприводимое представление группы G_A, в частности то, которое входит в $L^2(G_Q \backslash G_A)$, разлагается в тензорное произведение представлений групп G_p. Гипотеза Петерсона означает, что соответствующие неприводимые представления группы G_A, содержащиеся в $L^2(G_Q \backslash G_A)$, разлагаются в тензорное произведение представлений групп G_p и, таким образом, лишь конечное число может принадлежать дополнительной серии [9], [22].

7. Максимальные дискретные подгруппы

Вместе с данной группой Γ любая соизмеримая с ней группа также является арифметической. В связи с этим приобретает значение описание всех подгрупп, соизмеримых с данной арифметической группой Γ. В этом пункте речь идет об описании всех максимальных дискретных подгрупп, соизмеримых с данной группой Γ.

Если группа G не имеет центра, то можно доказать, что $\Gamma \subset G_Q$ [14], [22]. Обозначим через $\overline{\Gamma}^p$ замыкание группы Γ в p-адической топологии. В ряде случаев имеет место следующая теорема. Группа Γ является максимальной дискретной подгруппой группы $G_{\mathbf{R}}$ тогда и только тогда, когда группа $\overline{\Gamma}^p$ при любом конечном p является максимальной компактной подгруппой группы G_{Q_p}. Для ряда групп эта теорема проверена; насколько известно автору доклада, общего исследования ситуации пока нет.

8. Подгруппы конечного индекса арифметических групп

Пусть Γ — арифметическая группа, заданная своим матричным представлением. Без существенного ограничения можно считать, что это представление целочисленно. Пусть m — некоторое целое число. Обозначим через $\Gamma(m)$ совокупность всех матриц из Γ, сравнимых с единичной матрицей по модулю m. Подгруппа $\Gamma(m)$, очевидно, имеет конечный индекс в Γ. Условимся называть конгруэнц-подгруппой любую подгруппу Γ_1 группы Γ, содержащую подгруппу $\Gamma(m_1)$ при некотором m_1.

Естественно поставить вопрос, существуют ли у данной арифметической группы подгруппы конечного индекса, не являющиеся конгруэнц-подгруппами?

Ниже мы приводим пример, когда все подгруппы конечного индекса являются конгруэнц-подгруппами:

1) $\Gamma = SL(n, \mathbf{Z})$, где $n \geqslant 3$; см. [23];

2) Γ — группа целых точек некоторой простой односвязной алгебраической группы, расщепимой над полем алгебраических чисел k, ранг которой $r \geqslant 2$, k — не абсолютно мнимо; см. [22];

3) Γ — группа целых точек спинорной группы, связанной с некоторой квадратичной формой с пятью или большим числом переменных, в которой число отрицательных квадратов, так же как число положительных квадратов, не меньше двух (Л. Н. Васерштейн, готовится к печати).

Во всех рассмотренных случаях одновременно показывалось, что унипотентные элементы, принадлежащие данной подгруппе конечного индекса, порождают в ней всегда подгруппу конечного индекса. Рассмотрение унипотентных элементов играет важную роль в доказательстве перечисленных выше результатов.

Для арифметических групп, которые действуют в единичном круге $|z| < 1$, ситуация другая. Именно во всех них есть подгруппы конечного индекса, не являющиеся конгруэнц-подгруппами. Как заметил И. Р. Шафаревич, это можно доказать следующим образом. Каждая арифметическая группа Γ дробно-линейных преобразований единичного круга содержит подгруппу конечного индекса, которая или свободная, или с одним соотношением. Пользуясь этим, можно показать, что пополнение $\overline{\Gamma}$ группы Γ по топологии, индуцированной всеми подгруппами конечного индекса, не есть произведение p-адических групп Ли. В то же время легко проверить, что если бы все подгруппы конечного индекса были бы конгруэнц-подгруппами, то $\overline{\Gamma}$ была бы произведением p-адических групп Ли.

9. Накрытия над арифметическими многообразиями

Здесь будет изложена одна работа И. Р. Шафаревича и докладчика [19]. Условимся на протяжении настоящего пункта алгебраические многообразия, униформизируемые автоморфными функциями с арифметической дискретной группой, называть арифметическими многообразиями. В настоящем пункте описывается некоторое свойство арифметических многообразий в терминах, имеющих смысл над произвольным полем, а не только над полем комплексных чисел. Весьма правдоподобно, хотя это не доказано, что этим свойством обладают только арифметические многообразия.

Свойство арифметических многообразий, о котором мы говорим, тесно связано со следующим свойством арифметических подгрупп полупростых связных групп Ли.

Приведем вначале следующее определение. Пусть Γ — дискретная подгруппа топологической группы G. Назовем элемент $g \in G$ Γ-рациональным, если подгруппы Γ и $g\Gamma g^{-1}$ соизмеримы, т. е. их пересечение имеет конечный индекс в каждой из них. Легко показать, что совокупность всех Γ-рациональных элементов образует группу. Оказывается, все арифметические подгруппы Γ полупростых групп G обладают следующим свойством: подгруппа Γ', состоящая из всех Γ-рациональных элементов, всюду плотна в G.

Не доказано до сих пор, хотя и весьма вероятно, что это свойство имеет место только для арифметических групп.

Однако без труда можно показать следующее. Пусть D — ограниченная область в $\mathbf{C}^n$. Γ — дискретная подгруппа аналитических автоморфизмов области D, такая, что соответствующая фундаментальная область имеет конечный объем.

Обозначим через Γ' совокупность всех аналитических автоморфизмов g области D, таких, что группы Γ и $g\Gamma g^{-1}$ соизмеримы, т.е. их пересечение имеет конечный индекс в каждой из них. Пусть

группа Γ' обладает орбитой, которая всюду плотна в D, тогда область D является симметрической областью.

Однако в случае, когда D есть симметрическая область, не доказано, что указанное выше свойство выделяет только арифметические группы. Насколько известно автору, это не сделано даже в случае, когда D — единичный круг в плоскости одного комплексного переменного.

Свойство арифметических многообразий, о котором идет речь, состоит в следующем. Каждое арифметическое многообразие обладает накрытием, как правило не конечным, которое является квазиоднородным и которое разветвлено только на подмногообразии меньшей размерности.

Поясним это утверждение. Накрытие, о котором идет речь, представляет собой проективный предел конечных накрытий данного алгебраического многообразия. Представление в виде проективного предела позволяет ввести в него структуру кольцованного пространства и тем самым одновременно определить топологию и группу морфизмов. Это накрытие называется квазиоднородным, если существует орбита группы морфизмов, которая всюду в ней плотна.

Опишем теперь конструкцию квазиоднородного накрытия данного арифметического многообразия. Пусть D — симметрическая область, Γ — арифметическая группа аналитических автоморфизмов области D. Обозначим через X_Γ факторпространство D/Γ, если оно компактно, или его нормальную компактификацию, если оно некомпактно. Если Δ — подгруппа конечного индекса группы Γ, то X_Δ есть конечное накрытие X_Γ. Пусть $X(\Gamma)$ обозначает проективный предел X_Δ. Легко показать, что $X(\Gamma)$ есть квазиоднородное накрытие арифметического многообразия X_Γ. Отметим, что квазиоднородное накрытие также можно получить, если брать проективный предел X_Δ, когда Δ пробегает не все подгруппы конечного индекса, а лишь все конгруэнц-подгруппы или даже только конгруэнц-подгруппы, соответствующие данному множеству простых чисел. Интересно отметить, что группы морфизмов этих квазиоднородных накрытий обычно являются или p-адическими группами Ли, или их произведениями такого же типа, как в теории групп аделей. Весьма правдоподобно, что эта группа морфизмов квазиоднородного накрытия при естественных дополнительных предположениях является всегда p-адической группой Ли.

Аналогия между фактом существования однородного топологического накрытия и квазиоднородного алгебраического накрытия простирается довольно далеко. В частности, для алгебраического многообразия X, обладающего квазиоднородным накрытием, нетрудно определить понятие автоморфной формы и показать, что автоморфные формы данного веса образуют конечномерное линейное

пространство и т. д. Именно автоморфные формы веса m на X определяются как дифференциалы порядка m на X, которые регулярны на Ω.

В настоящее время исследование алгебраических многообразий, обладающих квазиоднородными накрытиями, находится в зачаточном состоянии. Даже самые первоначальные вопросы здесь неясны. Однако эта теория привлекает своим общим подходом и возможностью алгебраизации классической теории автоморфных функций.

10. Дискретные подгруппы групп Ли

Пусть $\mathfrak{G}$ — вещественная или p-адическая группа Ли. Тогда существует такая ее дискретная подгруппа Γ, что мера факторпространства $\Gamma \backslash \mathfrak{G}$ конечна. Как отметил Зигель, для существования такой дискретной подгруппы необходимо, чтобы группа $\mathfrak{G}$ была унимодулярна, т. е. чтобы левоинвариантная мера на $\mathfrak{G}$ совпадала с правоинвариантной мерой. Этого условия, как известно, не достаточно даже для линейных групп Ли. Дело по-разному обстоит для редуктивных и нильпотентных групп Ли. Для нильпотентных групп Ли таковые дискретные группы или не существуют, как в случае p-адических групп, или существуют при дополнительных условиях, найденных А. И. Мальцевым (рациональность структурных констант алгебры Ли). С другой стороны, во всех редуктивных группах Ли такие дискретные группы, по-видимому, существуют. Для вещественных редуктивных групп Ли это было доказано А. Борелем [26].

Отметим еще, что если $\mathfrak{G}$ есть p-адическая группа Ли, то, как нетрудно показать, если $\Gamma \backslash \mathfrak{G}$ имеет конечный объем, то $\Gamma \backslash \mathfrak{G}$ компактно. В то же время в вещественных группах Ли существуют дискретные подгруппы с некомпактным факторпространством конечного объема. Как хорошо известно, в любой вещественной некомпактной полупростой группе Ли $\mathfrak{G}$ есть такая дискретная подгруппа Γ, что $v(\Gamma \backslash \mathfrak{G}) < \infty$ и $\Gamma \backslash \mathfrak{G}$ не компактно. Существование таких дискретных подгрупп связано с существованием специального класса унипотентных подгрупп, так называемых орисферических подгрупп.

11. Дискретные подгруппы произведений групп Ли

Пусть S — некоторое конечное или бесконечное множество простых чисел, в частности S может содержать ∞. Пусть для каждого $p \in S$ указана p-адическая группа Ли G_p и для каждого $p \neq \infty$ в G_p отмечена некоторая открытая компактная подгруппа K_p.

Обозначим через G_S ограниченное прямое произведение групп G_p с отмеченными подгруппами K_p. Было бы интересно выяснить,

когда в группе G_S существует дискретная подгруппа Γ с конечным объемом факторпространства.

Разумеется, при этом интересен только случай, когда пара (Γ, G_S) в естественном смысле слова неприводима. Во всех известных примерах групп G_S, когда удавалось построить соответствующие дискретные группы, отдельные компоненты G_p оказывались согласованными. А именно существует такая алгебраическая группа G, определенная над полем рациональных чисел Q, что группы G_p представляют собой множество ее точек над полем p-адических чисел. Весьма вероятно, что это условие необходимо. При некотором уточнении оно является достаточным. Пусть G — линейная алгебраическая группа, определенная над Q. Обозначим через G_p множество ее точек, определенных над полем p-адических чисел, а через K_p — множество целых точек группы G_p. Пусть далее G_S — ограниченное прямое произведение групп G_p с отмеченными подгруппами K_p. Предположим, что или $\infty \in S$, или группа G_∞ компактна.

Рассмотрим группу Γ, состоящую из всех точек в G_Q, которые являются целыми при любом конечном $p \notin S$. Группу Γ можно естественным образом вложить в G_S в виде дискретной подгруппы и, как известно, $v(\Gamma \setminus G_S) < \infty$, если у группы G нет характеров, определенных над Q.

Дискретную подгруппу Γ группы G_S, полученную указанной конструкцией, естественно назвать арифметической подгруппой группы G_S. Отметим еще, что если группа G_S обладает компактным нормальным делителем $\mathfrak{R}$, то образ группы Γ в группе $\mathfrak{G} = G_S/\mathfrak{R}$ также дискретен и его также естественно назвать арифметической группой в $\mathfrak{G} = G_S/\mathfrak{R}$. Таким образом, чтобы найти все арифметические подгруппы данной группы $\mathfrak{G}$-ограниченного произведения p-адических групп Ли, надо перебрать все представления $\mathfrak{G}$ в виде $G_S/\mathfrak{R}$, где $\mathfrak{R}$ — некоторый компактный нормальный делитель.

Правдоподобно, что если S содержит не менее двух простых, то у группы $\mathfrak{G}$ все дискретные подгруппы Γ такие, что $v(\Gamma \setminus \mathfrak{G}) < \infty$, являются арифметическими.

В заключение автор считает своим приятным долгом поблагодарить профессора А. Бореля, ознакомившегося с предварительным текстом доклада и сделавшего ряд чрезвычайно ценных замечаний, учтенных автором при окончательном редактировании.

Автор считает своим приятным долгом поблагодарить также профессора И. Р. Шафаревича за ценное обсуждение предварительного текста доклада.

Институт прикладной математики АН СССР,
Москва, СССР

ЛИТЕРАТУРА

[1] Cartan E., Abh. Math. Sem. Univ. Hamburg, 11 (1935).
[2] Selberg A., *J. Ind. Math. Soc.*, **20** (1956), 47-87.
[3] Selberg A., Contr. Func. Th., Bombay (1960), 147-164.
[4] Borel A., Harish-Chandra, Arithmetic subgroups of algebraic groups, *Ann. Math.*, **75** (1962), 485-535.
[5] Гельфанд И. М., Proc. Inter. Cong. Math. Stockholm (1962), 74-85.
[6] Гельфанд И. М., Фомин С. В., *УМН*, 7: 1 (47) (1952), 118-137.
[7] Гельфанд И. М., Граев М. И., Труды Московского математического общества, **8** (1959), 323-390.
[8] Гельфанд И. М., Пятецкий-Шапиро И. И., *УМН*, XIV, (2) (1959), 172-194.
[9] Гельфанд И. М., Граев М. И., Пятецкий-Шапиро И. И., Теория представлений и теория автоморфных функций, М.—Л., 1966.
[10] Лапин А. И., *Изв. АН СССР*, **20**, № 3, 325 (1956).
[11] Igusa I., *Amer. J. Math.*, **84** (1962); **86**, № 2 (1964).
[12] Пятецкий-Шапиро И. И., Геометрия классических областей и теория автоморфных функций, М.—Л., 1961.
[13] Гиндикин С. Г., Пятецкий-Шапиро И. И., *ДАН СССР*, **162**, № 6 (1965), 1226-1229.
[14] Пятецкий-Шапиро И. И., *УМН*, XIX (6) (1964), 93-121.
[15] Пятецкий-Шапиро И. И., *УМН*, XX (2) (1965), 1-51.
[16] Винберг Э. Б., Гиндикин С. Г., Пятецкий-Шапиро И. И., Труды Московского математического общества, **12** (1963).
[17] Пятецкий-Шапиро И. И., Proc. Intern. Cong. Math. Stockholm (1962), 389-396.
[18] Пятецкий-Шапиро И. И., *ДАН СССР*, **124**, №2 (1959), 272-273.
[19] Пятецкий-Шапиро И. И., Шафаревич И. Р., *Известия АН СССР*, **30**, № 3 (1966), 671-705.
[20] Weil A., *Ann. Math.*, **75** (1962), 578-602.
[21] Baily W. L., Jr., Borel A., *Bull. A.M.S.*, **70** (1964), 588-593.
[22] Summer Institute on algebraic groups (1965).
[23] Bass H., Lazard M., Serre J. P., *Bull. A.M.S.*, **70** (3) (1964).
[24] Andreotti A., Grauert H., *Nachr. Akad. Wiss. Göttingen* (1961), 39-48.
[25] Mumford D., On the equations defining abelian varieties (preprint).
[26] Borel A., *Topology*, **2**, № 2 (1963), 111-123.
[27] Borel A., Density and maximality of arithmetic groups, *J. reine angew. Math.*, **224** (1966), 78-79.
[28] Макаров В. С., *ДАН СССР*, **67**, № 1 (1966), 30-33.
[29] Винберг Э. Б., *Матем. сб.*, **72** (114), № 3 (1967), 471-488.

ПОЛУЧАСОВЫЕ ДОКЛАДЫ

❊

HALF-HOUR REPORTS

❊

RAPPORTS D'UNE DURÉE D'UNE DEMI-HEURE

❊

VORTRÄGE VON EINER HALBEN STUNDE DAUER

MODEL THEORY AND SET THEORY

ROBERT L. VAUGHT

A number of results in model theory lie near the borderline of model theory and set theory. I shall discuss a group of such results, all of which can be considered as extensions in one direction or another of the well-known Löwenheim-Skolem-Tarski theorem.

For a more detailed summary of results and open problems in this area up to two years ago, see [15].

$\varkappa$, λ and μ will denote infinite cardinals (initial ordinals); α, β ordinals; m, n natural numbers. A (finitary) relational structure $\mathfrak{M} = \langle M, U^{\mathfrak{M}}, R_n^{\mathfrak{M}} \rangle_{n<\omega}$ (U unary) has type $\langle \varkappa, \lambda \rangle$ if $\varkappa = |M|$ (the power of M) and $\lambda = |U^{\mathfrak{M}}|$. We write $\varkappa, \lambda \Rightarrow \varkappa', \lambda'$ if every countable set of elementary sentences having a model of type $\langle \varkappa, \lambda \rangle$ also has one of type $\langle \varkappa', \lambda' \rangle$. For simplicity, we adopt throughout the generalized continuum hypothesis.

Theorem 1. (a) (R. Robinson) $\omega_{\alpha+n}, \omega_\alpha \not\Rightarrow \omega_{\beta+n+1}, \omega_\beta$.
(b) ([11]) If $\varkappa > \lambda$, then $\varkappa, \lambda \Rightarrow \omega_1, \omega_0$.
(c) (Chang-Keisler [2]) $\varkappa, \lambda \Rightarrow \varkappa', \lambda'$, if $\varkappa \geqslant \varkappa' \geqslant \lambda' \geqslant \lambda$.
(d) (Chang [1]) $\varkappa, \lambda \Rightarrow \mu^+, \mu$ if $\varkappa > \lambda$, $\mu > \omega$, μ regular.
(e) ([14]) $\omega_{\alpha+\omega}, \omega_\alpha \Rightarrow \lambda, \mu$ if $\lambda \geqslant \mu$.

It is conjectured that in general $\omega_{\alpha+n}, \omega_\alpha \Rightarrow \omega_{\beta+n}, \omega_\beta$. In [14] are several generalizations of (e), for example (extending the $\Rightarrow$ notation in an obvious way):

(e') $\omega_{\alpha_0}, \ldots, \omega_{\alpha_n} \Rightarrow \varkappa_0, \ldots, \varkappa_n$ if $\alpha_i \geqslant \alpha_{i+1} + \omega$ and $\varkappa_i \geqslant \varkappa_{i+1}$ for $i < n$.

A different proof of (e) was given by Morley [10].

The language $Q_\varkappa$ is obtained from the elementary language by adding a new quantifier Qx meaning "there are at least $\varkappa$ x such that." We write $\varkappa \to \lambda$ if every countable set of Q-sentences having a $Q_\varkappa$-model also has a Q_λ-model.

Theorem 2 (Furhken [3], [4]). (a) $\varkappa \not\to \lambda$ if $\varkappa$ is singular, λ not; or if $\varkappa$ is a successor cardinal, λ not; or if $\varkappa \neq \omega$, $\lambda = \omega$. (b) $\varkappa^+ \to \lambda^+$ if λ is regular. (c) $\varkappa \to \omega_1$ if $\varkappa$ is inaccessible. (d) $\omega \to \varkappa$.

The proof of 2 (d) makes use of a theorem of MacDowell and Specker [9] on extensions of models of arithmetic.

Fuhrken has conjectured that $\varkappa \to \lambda$ holds in all cases not blocked by 2 (a). Recently some more cases of this conjecture have been established.

Theorem 2. (e) (Helling [6]) $\varkappa \to \lambda$ if $\lambda \neq \omega$ and $\varkappa$ is weakly compact. (f) (J. Silver) $\varkappa \to \lambda$ if $\varkappa$ is inaccessible and λ singular[1]).

Recently Keisler [7] has obtained some results which much strengthen both 1 (b) and 2 (c). Let $Q_{\varkappa,\omega}$ be the language having two new quantifiers Q and Q' interpreted as in $Q_\varkappa$ and Q_ω, respectively. By Fuhrken's "normal form theorems" ([3]), Keisler's results imply:

Theorem 3. (Keisler) Let $\varkappa \neq \omega$ be regular. If a countable set of Q, Q'-sentences has a $Q_{\varkappa,\omega}$-model, then it has a $Q_{\omega_1,\omega}$-model.

Notice that in such results, Q_ω has the same expressive power as the so-called weak second order language.

Call a language L complete if its valid sentences are r.e. (recursively enumerable); ω-compact if a countable set of L-sentences must have a model if all its finite subsets do. Fuhrken [3] showed that $Q_\varkappa$ is ω-compact if $\varkappa = \omega_1$ or if $\varkappa \notin C_\omega = \{\lambda\colon \operatorname{cf}\lambda = \omega\} \cup \{\lambda^+\colon \operatorname{cf}\lambda = \omega\}$. The author showed in [13] that Q_{ω_1} is complete. An intermediate step is that the set T of all elementary sentences true in all models of type $\langle \omega_1, \omega\rangle$ is r.e. Recently Keisler [8] has given an explicit set of axioms for T.

The Hanf number $\nu_\lambda L$ of a language L is the smallest μ such that, for any set Σ of fewer than λ sentences, if Σ has a model of power at least μ, then Σ has arbitrarily large models.

Theorem 4. ([14], [15]) (a) $\nu_\omega Q_{\omega_1} = \nu_{\omega_1} Q_{\omega_1} = \omega_\omega$.

$$\nu_\omega Q_\varkappa = \nu_{\varkappa^+} Q_\varkappa = \omega_{\alpha+\omega} \text{ if } \varkappa = \omega_\alpha \notin C_\omega.$$

Morley discovered a powerful method for obtaining Hanf numbers which uses a transfinite version of Ramsey's theorem due to Erdös and Rado.

Theorem 5. (a) (Morley [10]) $\nu_{\omega_1} Q_\omega = \omega_{\omega_1}$.
(b) (Helling [5]) Indeed $\nu_{\varkappa^+} Q_\varkappa = \omega_{\varkappa^+}$ whenever $\operatorname{cf}\varkappa = \omega$.

Recently, J. Silver [12] obtained detailed information about the Hanf number of the very strong β-language (in which $<$ always denotes a well-ordering).

University of California,
Berkeley, USA

[1]) Silver's result is stated here for the first time, with his kind permission.

REFERENCES

[1] Chang C. C., A note on the two cardinal problems, *Proc. Amer. Math. Soc.*, **16** (1965), 1148-1155.
[2] Chang C. C., Keisler H. J., Applications of ultraproducts of pairs of cardinals to the theory of models, *Pacific J. Math.*, **12** (1962), 835-845.
[3] Fuhrken G., Skolem-type normal forms for first-order languages with a generalized quantifier, *Fund. Math.*, **64** (1964), 291-302.
[4] Fuhrken G., Languages with added quantifier "there exist at least $\aleph_\alpha$". The Theory of Models, Proc. 1963 Symposium held in Berkeley, 121-131.
[5] Helling M., Hanf numbers for some generalizations of first-order languages, *Notices Amer. Math. Soc.*, **11** (1964), 679.
[6] Helling M., Transfer and compactness properties of some generalized quantifiers, *Notices Amer. Math. Soc.*, **12** (1965), 723.
[7] Keisler H. J., Homogeneous theories, *Notices Amer. Math. Soc.*, **12** (1965), 600.
[8] Keisler H. J., First-order properties of pairs of cardinals, *Bull. Amer. Math. Soc.*, **72** (1966), 141-144.
[9] MacDowell R., Specker E., Modelle der Arithmetik. Infinitistic Methods, Proc. 1959 Symposium on Foundation of Mathematics held in Warsaw, 257-263.
[10] Morley M., Omitting classes of elements. The Theory of Models, Proc. 1963 Symposium held in Berkeley, 265-273.
[11] Morley M., Vaught R., Homogeneous universal models, *Math. Scand.*, **11** (1962), 37-57.
[12] Silver J., Some applications of model theory in set theory. Doctoral dissertation, University of California, Berkeley, 1966, Ch. 5.
[13] Vaught R., The completeness of logic with the added quantifier "there are uncountably many", *Fund. Math.*, **64** (1964), 301-304.
[14] Vaught R., A Löwenheim-Skolem theorem for cardinals far apart. The Theory of Models, Proc. 1963 Symposium held in Berkeley, 390-401.
[15] Vaught R., The Löwenheim-Skolem theorem. Proc. 1964 International Congress for Logic Methodology and Philosophy of Science held in Jerusalem, 81-89.

PECULIARITIES OF CONSTRUCTIVE MATHEMATICAL ANALYSIS

G. S. TSEYTIN, I. D. ZASLAVSKY, N. A. SHANIN

By constructive mathematics we understand the approach characterized by the following features (see [1], [2]):

(1) only constructive objects are considered; abstraction of potential realizability is used, while actual infinity is never considered;

(2) constructive understanding of mathematical statements and constructive rules of conclusion are used;

(3) Markov's principle of constructive choice is accepted.

A more detailed characterization of this approach and our views as to its role in mathematics are given in the abstract of the report [3] and we shall not repeat them [1]).

During the last decades a number of works appeared dedicated to mathematical analysis based on the constructive principles mentioned above. In this report we shall speak mainly of the works by professor A. A. Markov and his school.

There are works on the constructive theory of real numbers, theories of various kinds of functions of real variable (uniformly, continuous, differentiable, Riemann integrable, functions of bounded variation, etc.), theories of metric spaces, Banach and Hilbert spaces, etc. (See the reference list at the end of the paper.)

In all these theories were not merely constructive fundamental notions introduced in place of non-constructive notions of the corresponding classical theories, but preference was always given to notions based on algorithms in order to make clear the computational sense of the theorems.

On the basis of mathematicians' experience with these theories we can expect that other theories of constructive mathematical analysis analogous to various theories of classic analysis can be developed in a similar way. The aim of this report is to describe some general features of these constructive theories and their correspondence to the respective theories of classic analysis.

1. Types of constructive objects (i.e., constructive notions) are introduced by the following means:

1) basic types of objects (like integers, words in various alphabets, integer matrices) are introduced by means of generating rules;

2) subordinate types of objects are introduced by pointing out a basic type of objects and a selecting condition;

3) the definition of a type of objects may be supplied with a special equality predicate.

The word "set" means to us the same as the expression "type of constructive objects". If some basic type of objects and some language for selecting conditions and equality predicates is fixed, then "definable sets" can be regarded also as constructive objects of a certain type. No general notion of set valid for all constructive mathematics is introduced.

2. Many notions of constructive mathematics are introduced as "constructive translations" of some non-constructive notions of classic mathematics. They correspond to the same real situations as the classic notions and are similar in many respects to those classic notions. But the mathematical basis of these notions, the "material" they are made of is essentially different from that of classic notions; therefore serious differences in some properties between constructive and classic notions are inevitable.

[1]) This paper gives a more detailed exhibition of the item 7 of the abstract.

For example, for real numbers we use the notion of FR-number, based on algorithmic fundamental sequences of rational numbers (some details are given below). In terms of FR-numbers we can introduce, for instance, all the algebraic real numbers; the continuum of FR-numbers is also complete and connected (completeness and connectivity are understood here in the proper constructive sense). But the most essential difference from the classic continuum is the fact that for the FR-continuum Borel's theorem of coverings does not hold, namely, there exists a constructive sequence of intervals covering a segment and having no finite sub-covering [4]. In the theory of functions based on FR-numbers this fact leads to greater differences between local and integral properties of functions defined on a segment, e.g., there is a function continuous at every point of the segment (and even uniformly continuous in the neighbourhood of every point) but not bounded on the segment [4], [5]. On the other hand, every constructive function proves to be continuous at every point where it is defined [6], [7].

Perhaps we could achieve closer similarity to the classic analysis if we chose a different constructive version of the notion of real number. For example, we could consider fundamental sequences defined not by algorithms, but by predicates of some kind (cf. [8]). But in the constructive mathematics considered here the algorithmic notion has been deliberately chosen since we concern ourselves with problems of computability, and some "uncommon" properties of our real numbers are the price we pay for it.

In general, when in a classic definition such notions as "function", "sequence", "operator" occur we use "algorithm" as their standard translation for forming the definition of the respective constructive notion. But the classic "set" in such cases may be translated by some kind of "definable set", i.e., by a more general notion than "algorithm". For instance, constructive Hilbert space is defined as a s e t o f w o r d s with a l g o r i t h m s for addition, multiplication by constructive complex number and scalar product.

3. In many cases for one notion of classic analysis several constructive analogues are introduced that lead to different theories (these theories in general may be not equally interesting and fruitful). This "splitting of notions" is due to various reasons. We shall illustrate this again on real numbers.

1) Various classic definitions may be taken as a starting point. For example, "constructive Dedekind sections" may be considered, as well as FR-numbers based on constructive fundamental sequences.

2) Sometimes slight changes in the classic definition are made that are not essential for the classic notion but are very essential for its constructive translation. For example, we can define constructive Dedekind section directly as an algorithm attributing every rational number to one of the two classes. But under this definition there does not even exist any algorithm for the addition of two numbers. If we change the definition of the constructive Dedekind section to make the classifying algorithm applicable not to every rational number, but to every except at most one, then we obtain a notion, equivalent in a certain sense to FR-numbers [4].

3) Differences in completeness of information contained in the object can be essential. An FR-number is by definition a word consisting of codes of two algorithms:

the "base"—an algorithm defining a sequence of rational numbers, and

the "regulator"—an algorithm certifying its fundamentality (Cauchy's criterion of convergence).

The notion of F-number can be introduced. F-number is a code only of a "base" on the condition that a regulator does exist (see [9]). Every F-number by definition can be completed to an FR-number, but there is no algorithm for such completing [10]. Moreover it can be proved that there is no algorithm giving for each F-number an FR-number, n o t e q u a l to it.

4) "Splitting of notions" can arise from constructive non-equivalence of classically equivalent logical connectives (e.g., because of double negation). While the F-number is defined as a code of a "base" for which a "regulator" exists, a quasi-number is defined as a code of a "base" for which the double negation of existence of a "regulator" holds. Of course every F-number is a quasi-number and there is no quasi-number that is not an F-number. But these two notions are very different, which can be seen, for example, by considering sequences of F-numbers and of quasi-numbers. B. A. Kushner has constructed a fundamental sequence of quasi-numbers that has no limit [11].

For the notion of function we shall consider a still greater number of constructive versions than for real numbers. The variety of constructive notions of function arises not only because of the variety of notions of real number, but mainly because of the variety of aspects of functions that are considered in various theories. There are constructive theories for uniformly continuous functions, differentiable functions, Riemann and Lebesgue integrable functions, functions of bounded variation, etc. As we shall see later on, in many of such theories separate notions of function appear. In some of such theories functions are introduced not as number-to-number algorithms, but as elements of special metric spaces; these spaces are constructed by means of completing with respect to metric. Essential ideas of this

approach appear as early as in Goodstein's papers on uniformly continuous functions [12], [13], [14]. In the same way constructive measurable functions and Lebesgue integral are introduced [15]. It should be noted here that some theories of classic analysis also use notions of function that are not based on point-to-point correspondence (generalized functions). In constructive analysis the use of function notions of this type has increased.

4. We will now discuss some peculiarities connected with the use of certain types of sets and variables. Let x be a variable for objects of some basic type, $A(x)$ a selecting condition defining a subordinate type of objects, y a variable for such objects, and M the set of all such objects. By the rules of constructive interpretation [16], [17] the statement $A(x)$ can be reduced to one of the forms: $B(x)$ or

$$\exists z B(x, z),$$

where z is a variable for objects of a certain basic type and B (in both cases) is a so-called normal formula, i.e., a formula without disjunctions and existential quantifiers (every normal formula is equivalent to its double negation). In the first case the variable y and the set M are called normal. A normal set is equal to its "double complement", i.e., to the complement of its complement. This is not the case for sets that are not normal and this fact leads to some peculiarities. We will refer to an interesting example from the theory of constructive locally convex topologic spaces developed by Phan-đình-Diệu [18]. There is a space where every neighbourhood of zero is a set whose double complement is equal to the whole space. Thus no point can be separated from zero by a neighbourhood. However the intersection of all neighbourhoods of zero contains only zero.

Another peculiarity is connected with quantifiers on abnormal variables. A statement of the form

$$\forall x \exists u R(x, u),$$

x being a basic variable, is interpreted as the existence of an algorithm α applicable to every x and such that

$$\forall x R(x, \alpha(x)).$$

For a subordinate variable y the statement

$$\forall y \exists u R(y, u)$$

means that

$$\forall x (A(x) \supset \exists u R(x, u)),$$

so its interpretation depends on the character of $A(x)$. It can be seen that for y normal the same interpretation

$$\forall y R(y, \alpha(y))$$

as for basic variables can be used. But for the second case the correct interpretation of

$$\forall y \exists u R(y, u)$$

is that an algorithm β exists such that

$$\forall x \forall z \left(B(x, z) \supset \left(!\beta(x, z) \,\&\, R(x, \beta(x, z)) \right) \right)$$

($!\beta(x, z)$ means that $\beta(x, z)$ is defined). Thus in some cases the statement of the form $\forall y \exists u R(y, u)$ is true though no algorithm α such that $\forall y R(y, \alpha(y))$ exists.

A lot of examples of this kind are found in various constructive theories. The example of completing F-numbers to FR-numbers has already been mentioned. Here is another example. Every function uniformly continuous on a segment has a least upper bound (l.u.b.), but no algorithm exists giving the l.u.b. for any uniformly continuous function [4]. The following example is offered by G. E. Minz: there exists no algorithm giving a derivative for any function differentiable on a segment though the condition of differentiability means just that a derivative exists [19].

In classic analysis operators are often introduced in the following way. If a binary predicate $\mathscr{R}$ is defined such that for any Y there is no more than one U such that $\mathscr{R}(Y, U)$, then an operator $\mathscr{A}$ is immediately introduced such that $\mathscr{A}(Y)$ is defined if and only if $\exists U \mathscr{R}(Y, U)$, and in this case $\mathscr{R}(Y, \mathscr{A}(Y))$. In such cases the operator notation is usually preferred to the predicate notation. This is, for instance, the way of introducing the differentiation operator.

But in constructive analysis, if we want operators to be algorithmic, we cannot follow this way. We have an alternative: either we regard uniformly continuous functions, differentiable functions, etc., as particular cases of the general notion of function and then we have no operators for l.u.b., derivative, etc., and use only predicates, or else we introduce special normalized notions for these kinds of functions. For example, the normalization of the notion of uniformly continuous function is the notion of "uniformly continuous duplex". Uniformly continuous duplex is a pair consisting of a function and a special algorithm certifying its uniform continuity. Likewise the notion of FR-number is the normalization of the notion of F-number. In theories using normalized notions of differentiable function, uniformly continuous function algorithms for obtaining derivative, resp. l.u.b. do exist. This approach is usually preferred in the constructive mathematics now discussed because we are interested in problems of what information is really needed for a certain computational procedure. But as a result of this approach we have separate

normalized notions in each theory instead of a single notion of function.

5. Theorems proved in constructive analysis have different degrees of similarity to classic theorems.

1) Some theorems have exactly the same formulations as theorems of classic analysis, but the terms and logical connectives involved are understood constructively.

E x a m p l e s. Every fundamental sequence of real numbers has a limit. Every function differentiable on a segment and having a non-negative derivative is non-decreasing [10].

2) Sometimes a formulation of constructive theorem differs from the classic formulation in details that are unessential for the classic theorem. However the constructive statement obtained by literal translation of the original classic formulation does not hold. For instance, classically every bounded linear functional in the Hilbert space l^2 can be presented as a scalar product by a fixed element. Constructively, we must say "strictly bounded" instead of "bounded" [15], [19]. This means that the values of the functional on the unit sphere are not only bounded, but also have a l.u.b.; without this condition the theorem is not valid [15]. Another example is the Cauchy theorem saying that a continuous function taking values of opposite signs vanishes at some point. Constructively, only the double negation of existence of such a point is true [10].

3) Sometimes in a constructive theorem analogous to a certain classic theorem different constructive versions of the same classic notion are considered simultaneously. For example, for the Cauchy theorem just mentioned a constructive version can be proved stating that there exists a q u a s i - n u m b e r in which the function vanishes [10]. Another example is the classic theorem saying that every function of bounded variation is a difference of two monotonous functions, and vice versa. In constructive analysis we distinguish functions of weakly bounded variation for which the variation sums are bounded and functions of strongly bounded variation for which these sums have a l.u.b. It can be proved that every function of strongly bounded variation is a difference of two monotonous functions and that the difference of any two monotonous functions is a function of weakly bounded variation. The converse theorems in both cases are not valid [4].

4) Some constructive theorems can be regarded as approximate analogues of classic theorems.

E x a m p l e s. Every function differentiable on a segment is uniformly continuous in the neighbourhood of every point of this segment [20]. Every constructive function defined at least at one point is the limit of a sequence of polygonal functions (not necessarily

uniformly convergent) [21]. Every constructive function defined at least at one point is the limit of a uniformly convergent sequence of (so-called) pseudo-polygonal functions [6], [7]. For every constructive function f defined on a segment and taking values of opposite signs and for every positive ε there is a point x where $|f(x)|<\varepsilon$ [14].

5) Some theorems of constructive analysis show a substantial difference from classic analysis. We know a lot of examples of functions with combinations of properties that are impossible from the classic standpoint.

Examples. V. P. Orevkov's example of a continuous mapping of a circle into itself without a fixed point [22]. A continuous function defined on a segment and having no primitive function (and even no integral in a certain generalized sense) [23]. A. A. Markov's example of a function analytic on a segment and unbounded on it (see also [24]). A function, analytic on a segment, vanishing in infinitely many points, but not identical to zero.

Each of these functions is a counter-example disproving the literal constructive translation of a certain classic theorem. It should be noted that the literal constructive translations of some of the classic theorems mentioned above are not valid though they cannot be disproved by a counter-example (e.g., the Cauchy theorem).

6. But in general these differences between classic and constructive analysis usually appear in problems remote from concrete computation. The nearer we come to concrete computational problems the less are the differences. For example, tables of integration of elementary functions are valid for constructive analysis as well as for classic analysis. In some cases the difference between a classic theorem and its constructive analogue warns us that the classic theorem suggests an incorrect formulation of the computational problem, i.e., an algorithm with the properties suggested by the classic formulation proves to be impossible (consider, for instance, computation of zeros of a continuous function). Thus constructive analysis can be regarded as a theory of correctness of computational problems when we are interested only in computability in principle.

On the other hand, because of unrestricted use of potential realizability this approach is inadequate for problems of computation within given time and given memory capacity. This is of course a subject of special theories and perhaps also of some kind of "hyper-constructive" analysis still to be created.

Ленинградское отделение
математического института им. В. А. Стеклова,
Ленинград, СССР

REFERENCES

[1] Марков А. А., О конструктивной математике, Труды матем. инст. им. В. А. Стеклова, LXVII (1962), 8-15.
[2] Марков А. А., Теория алгорифмов, Труды матем. инст. им. В. А. Стеклова, XLII (1954).
[3] Цейтин Г. С., Заславский И. Д., Шанин Н. А., Особенности конструктивного математического анализа, Международный конгресс математиков, Тезисы докладов, Москва, 1966.
[4] Заславский И. Д., Некоторые свойства конструктивных вещественных чисел и конструктивных функций, Труды матем. инст. им. В. А. Стеклова, LXVII (1962), 385-457.
[5] Specker E., Der Satz vom Maximum in der rekursiven Analysis. Studies in logic and the foundations of mathematics. Constructivity in Mathematics. Proceedings of the Colloquium held at Amsterdam, 1957.
[6] Цейтин Г. С., Алгорифмические операторы в конструктивных полных сепарабельных метрических пространствах, *ДАН СССР*, **128** (1959).
[7] Цейтин Г. С., Алгорифмические операторы в конструктивных метрических пространствах, Труды матем. инст. им. В. А. Стеклова, LXVII (1962), 295-361.
[8] Weyl H., Das Kontinuum, Leipzig, 1918.
[9] Specker E., Nicht konstruktiv beweisbare Sätze der Analysis, *J. of Symbolic Logic*, **14**, (1949), 145-158.
[10] Цейтин Г. С., Теоремы о среднем значении в конструктивном анализе, Труды матем. инст. им. В. А. Стеклова, LXVII (1962), 362-384.
[11] Кушнер Б. А., Некоторые свойства квазичисел и операторов из квазичисел в квазичисла, *ДАН СССР*, 171, № 2 (1966), 275-277.
[12] Goodstein R. L., Mean value theorems in recursive function theory, I, *Proc. London Math. Soc.*, ser. 2, **52** (1950), 81-106.
[13] Goodstein R. L., Constructive formalism, University College Leicester, 1951.
[14] Goodstein R. L., Recursive analysis, North-Holland Publ. co., 1961.
[15] Шанин Н. А., Конструктивные вещественные числа и конструктивные функциональные пространства, Труды матем. инст. им. В. А. Стеклова, LXVII (1962), 15-294.
[16] Шанин Н. А., О конструктивном понимании математических суждений, Труды матем. инст. им. В. А. Стеклова, LII (1958), 226-311.
[17] Шанин Н. А., Об алгорифме конструктивной расшифровки математических суждений. *Zeitschr. f. math. Logik und Grundlagen d. Math.*, **4** (1958), 293-303.
[18] Фан Динь Зиеу (Phan-đình-Diệu), Конструктивные локально выпуклые топологические пространства, *ДАН СССР*, **162** (1965).
[19] Минц Г. Е., О предикатных и операторных вариантах построения теорий конструктивной математики, Труды матем. инст. им. В. А. Стеклова, LXXII (1964), 383-436.
[20] Заславский И. Д., О дифференцировании и интегрировании конструктивных функций, *ДАН СССР*, **156**, № 1 (1964), 25-27.
[21] Цейтин Г. С., Три теоремы о конструктивных функциях, Труды матем. инст. им. В. А. Стеклова, LXXII (1964), 537-543.
[22] Оревков В. П., О конструктивных отображениях круга в себя, Труды матем. инст. им. В. А. Стеклова, LXXII (1964), 437-461.
[23] Заславский И. Д., Цейтин Г. С., О сингулярных покрытиях и связанных с ними свойствах конструктивных функций, Труды матем. инст. им. В. А. Стеклова, LXVII (1962), 458-502.
[24] Кушнер Б. А., О существовании неограниченных аналитических конструктивных функций, *ДАН СССР*, **160**, № 1 (1965), 29-31.

2

Алгебра
Algebra
Algèbre
Algebra

WHITEHEAD GROUPS AND GROTHENDIECK GROUPS OF GROUP RINGS

HYMAN BASS

1. Introduction

A number of problems in algebraic and differential topology have led to the construction of invariants of an arithmetic nature, defined in terms of the group ring, $\mathbf{Z}\pi$, where π is some intervening fundamental group. These invariants live in groups which can, essentially, be defined over any ring A. One is K_0A, the "Grothendieck group" of finitely generated projective (left) A-modules. $\tilde{K}_0A$ denotes the quotient modulo the subgroup generated by the free modules. Another is the "Whitehead group", $K_1A = GL(A)/E(A)$. Here $GL(A) = \bigcup_n GL_n(A)$ is the infinite linear group, and $E(A)$ is the subgroup generated by elementary matrices; we have Whitehead's commutator formula, $E(A) = [E(A), E(A)] = [GL(A), GL(A)]$. The definition supplies homomorphisms $GL_n(A) \to K_1A$, and, for $n = 1$, we write $U(A) = GL_1(A) \to K_1A$. If A is commutative then determinant: $K_1A \to U(A)$ splits the latter, so $K_1A = U(A) \oplus SK_1A$, where $SK_1A = SL(A)/E(A)$. If π is a group then $\pm\pi \subset U(\mathbf{Z}\pi)$, and we write $\mathrm{Wh}(\pi) = \mathrm{coker}(\pm\pi \to K_1\mathbf{Z}\pi)$. Thus, if π is abelian, $\mathrm{Wh}(\pi) = (U(\mathbf{Z}\pi)/\pm\pi) \oplus SK_1\mathbf{Z}\pi$.

This report will survey the few cases in which relatively complete calculations of $\tilde{K}_0\mathbf{Z}\pi$ and of $\mathrm{Wh}(\pi)$ have been made. Some of this material is discussed in Milnor's very readable report [10] on Whitehead Torsion. The following contains both new results, and corrections of results that I prematurely announced to Milnor (see [10, p. 360]). Before giving the algebraic calculations, however, let me briefly indicate some of the sources of interest of these groups to topologists.

Around 1950 J. H. C. Whitehead, developing earlier ideas of Reidemeister, Franz, and de Rham, constructed his theory of simple homotopy types [18]. Whitehead attached to a homotopy equivalence, $f: X \to Y$, of finite complexes with fundamental group π, a "torsion"

invariant $\tau(f) \in \mathrm{Wh}(\pi)$. $\tau(f) = 0$ if and only if f is a "simple homotopy equivalence."

G. Higman's thesis (see [8]), written then under Whitehead, contains the first and, until 1962, the only non-trivial calculations of the groups $\mathrm{Wh}(\pi)$.

Since about 1963 Whitehead's torsion has begun to play an essential role in differential topology, due to the s-cobordism theorem of Barden-Mazur-Stallings (see, e.g. [10, § 10-11] or [11, exposé 7]). This asserts, in higher dimensions, that an h-cobordism $(W; M, M')$ is classified by M together with $\tau(f)$, where f is the inclusion of M in W, with a C^1-triangulation of (W, M).

C. T. C. Wall [17], around 1964, constructed, for a CW-complex which is dominated by a finite CW-complex, an invariant in $\tilde{K}_0\mathbf{Z}\pi$ $(\pi = \pi_1(X))$. It vanishes if and only if X has the homotopy type of a finite complex. In L. Siebenmann's 1965 thesis [13] he found an obstruction to putting a boundary on an end, ε, of an open manifold. This obstruction also lives in $\tilde{K}_0\mathbf{Z}\pi$, where, this time, $\pi = \pi_1(\varepsilon)$.

I will close this introduction by mentioning that recent work of Novikov, Wall, and others, some of it discussed at this congress, has led to the introduction of other Grothendieck and Whitehead groups, defined in terms of hermitian and anti-hermitian forms on $\mathbf{Z}\pi$-modules. These present a rich store of interesting, and difficult, algebraic problems that have yet hardly been touched.

2. Finite Groups

In this section π always denotes a finite group, and all modules are understood to be finitely generated. We shall describe $K_i\mathbf{Z}\pi$ $(i = 0, 1)$ with the aid of the homomorphisms

$$f_i: K_i\mathbf{Z}\pi \to K_i\mathbf{Q}\pi \qquad (i = 0, 1)$$

induced by the inclusion $\mathbf{Z}\pi \subset \mathbf{Q}\pi$.

Image f_0: "Rank"

Theorem 2.1 (Swan [15]). *If P is a projective $\mathbf{Z}\pi$-module then $\mathbf{Q} \otimes_{\mathbf{Z}} P$ is $\mathbf{Q}\pi$-free. Therefore* $\mathrm{Im}(f_0) \cong \mathbf{Z}$.

Kernel f_0: "Class number"

Theorem 2.2 (Swan [15]). *Every projective $\mathbf{Z}\pi$-module is the direct sum of a free module and a projective ideal. Therefore* (thanks to Jordan-Zassenhaus) $\mathrm{Ker}(f_0) \cong \tilde{K}_0\mathbf{Z}\pi$ *is a finite group.*

Example (Reiner-Rim, see [12] or [13, appendix]). If π is cyclic of prime order p then $\tilde{K}_0\mathbf{Z}\pi$ is isomorphic to the ideal class group of the field of p^{th} roots of unity.

Image f_1: "Unit theorem"

We will say π has exponent e if $x^e = 1$ for all $x \in \pi$, and write exp (π) for the smallest such $e > 0$. If π is an additive group, having exponent e means $e \cdot \pi = 0$.

We shall also write $r_0(\pi)$ (resp., $r_\infty(\pi)$) for the number of irreducible representations of π over $\mathbf{Q}$ (resp., $\mathbf{R}$). These numbers can be computed from π as follows:

Theorem 2.3 (see [7, Theorem 42.8]).

(E. Artin): *$r_0(\pi)$ is the number of conjugacy classes of cyclic subgroups of π.*

(S.D. Berman, E. Witt): *$r_\infty(\pi)$ is the number of conjugacy classes of unordered pairs $[x, x^{-1}]$ in π.*

The reason for introducing these numbers appears in the next theorem.

Theorem 2.4 (see [1] and [2]). *$GL_n(\mathbf{Z}\pi) \to K_1\mathbf{Z}\pi$ is surjective for all $n \geqslant 2$. Therefore* (thanks to Siegel), *$K_1\mathbf{Z}\pi$ is a finitely generated abelian group.*

Im (f_1) has rank $r_\infty(\pi) - r_0(\pi)$ and its torsion subgroup has exponent e (or $2e$, if e is odd) where $e = \exp(\pi)$.

Question. The units $\pm\pi$ in $\mathbf{Z}\pi$ map into the torsion subgroup of Im (f_1) (with kernel $[\pi, \pi]$); do they map *onto* this torsion subgroup? For abelian π this was proved by Higman [8], and Lam [9] has verified it for the symmetric group, S_3, and for the quaternion group.

The method of proof of Theorem 2.4 gives more precise information on the torsion free part of Im (f_1), as follows: Suppose $\rho : \pi \to GL_n(\mathbf{R})$ is a real representation of π. This induces $\rho : \mathbf{Z}\pi \to M_n(\mathbf{R})$ ($n \times n$ matrices), and therefore $K_1\mathbf{Z}\pi \xrightarrow{K_1\rho} K_1\mathbf{M}_n(\mathbf{R}) \xrightarrow{\text{"det"}} U(\mathbf{R}) \xrightarrow{\log|\cdot|} \mathbf{R}$. Write $g_\rho : K_1\mathbf{Z}\pi \to \mathbf{R}$ for the composite homomorphism. Now let $\rho_0, \ldots, \rho_{r_\infty}$ represent the distinct irreducible real representations of π, and $\sigma_1, \ldots, \sigma_{r_0}$ the distinct irreducible rational representations. Define $n_{ij} \in \mathbf{Z}$ by $\sigma_i \simeq \sum_{1 \leqslant j \leqslant r_\infty} n_{ij}\rho_j$ $(1 \leqslant i \leqslant r_0)$.

Theorem 2.5 ("Dirichlet", see [2, Theorem 1]). *The image of*

$$g = (g_{\rho_1}, \ldots, g_{\rho_{r_\infty}}) : K_1\mathbf{Z}\pi \to \mathbf{R}^{r_\infty}$$

is a lattice of rank $r_\infty - r_0$ in the subspace, V, defined by the equations

$$\sum_{1 \leqslant j \leqslant r_\infty} n_{ij}x_j = 0 \qquad (1 \leqslant i \leqslant r_0).$$

We can further give generators and relations for V. Suppose t and $s = t^a$ generate a cyclic group of order m in π. If h is a multiple of $\varphi(m)$ then

$$\alpha = (1 + t + \ldots + t^{a-1})^h + \frac{a^h - 1}{m}(1 + t + \ldots + t^{m-1})$$

is a unit in $\mathbf{Z}\pi$, not depending on a. α represents an element, say $[\alpha]$, of $K_1\mathbf{Z}\pi$, and

$$[s/t] = (1/h) \cdot g[\alpha] \in V$$

depends only on s and t, not h. The following theorem was conjectured, and first proved in special cases, by Milnor.

Theorem 2.6 (see [2, Theorem 5]). *With the notation above, $\{[s/t] \mid s$ and t generate the same subgroup of $\pi\}$ spans V. A complete set of linear relations between these generators is:*
For all s, t, u generating the same subgroup of π, and for all $x \in \pi$,

(R_1) $\qquad [s^{-1}/t] = [s/t]$,

(R_2) $\qquad [s/t] + [t/u] = [s/u]$,

(R_3) $\qquad [xsx^{-1}/xtx^{-1}] = [s/t]$.

Kernel f_1: "Congruence subgroup problem"

Suppose A is the ring of integers in an algebraic number field. $SL_n(A)$ is said to have the CSP (congruence subgroup property) if each subgroup of finite index contains all matrices $\equiv I$ modulo some non zero ideal. It was known that if this were always true (with n large) then it would follow, for finite abelian π, that Ker $(f_1) = SK_1\mathbf{Z}\pi = 0$ (see [10, appendix 1] for a proof of this implication). While this hypothesis had long been conjectured by several people, including the author, (see [2, § 5] and [10, appendix 1]), it was recently proved not to be the case (see [6]). As yet, however, no example of a non zero $SK_1\mathbf{Z}\pi$ is known, though the existence of one no longer seems unreasonable. We have, at least, the following:

Theorem 2.7 (Bass-Milnor-Serre [6]). *Let π be a finite abelian group of exponent e. Then* Ker $(f_1) = SK_1\mathbf{Z}\pi$ *has exponent e, and it is trivial if $e = 2$. If the Sylow p-subgroup of π is cyclic, then $SK_1\mathbf{Z}\pi$ has no p torsion.*

In the general case (π not necessarily abelian) we have:

Theorem 2.8 (Lam [9]). Ker (f_1) *is a finite group of exponent* $[\pi : 1]^2$.

Lam proves this by making a reduction to the abelian case, using the method of induced representations originated by E. Artin, and later extended by Swan to integral representations. The main point is to establish that $K_1\mathbf{Z}\pi$ is, functorially in π, a module over $G_0(\mathbf{Z}\pi)$,

the Grothendieck ring of integral representations of π, and that these structures obey a form of Frobenius reciprocity for induced representations. In fact f_1 is a morphism of such G_0-modules, so that Ker (f_1) is another. It then follows formally from an "induction argument", used already in other contexts by Swan (see, e.g. [16]) that Theorem 2.8 follows from the vanishing of Ker (f_1) for cyclic π (Theorem 2.7), and from Swan's generalization to $G_0(\mathbf{Z}\pi)$ [15, Theorem 4.1] of Artin's induction theorem on $G_0(\mathbf{Q}\pi)$. Lam's axiomatization of this argument yields a number of new results of the same genre.

3. Free Products

The situation when $\pi = \pi_1 * \pi_2$ is a free product is quite satisfactory.

Theorem 3.1 (Stallings [14]). $\mathrm{Wh}(\pi_1 * \pi_2) = \mathrm{Wh}(\pi_1) \oplus \mathrm{Wh}(\pi_2)$.

Theorem 3.2 (G. Higman [8]). $\mathrm{Wh}(\mathbf{Z}) = 0$.

Corollary 3.3 $\mathrm{Wh}(\pi) = 0$ *if* π *is a free group.*

Only the latter has been generalized to K_0.

Theorem 3.4 ([3]). *If* π *is free then projective* $\mathbf{Z}\pi$*-modules are free. Hence* $\tilde{K}_0\mathbf{Z}\pi = 0$.

Question. Is $\tilde{K}_0 Z[\pi_1 * \pi_2] \cong \tilde{K}_0\mathbf{Z}\pi_1 \oplus \tilde{K}_0\mathbf{Z}\pi_2$? This seems to be rather plausible [1]).

4. Abelian Groups

Here π will always denote an abelian group, and, modulo a simple direct limit argument, there is no loss in assuming that π is finitely generated. The theorems below on K_0 are rather complete, modulo the case of finite abelian groups. Those on K_1 are much less precise. Recall that $K_1\mathbf{Z}\pi = (U(\mathbf{Z}\pi)/\pm\pi) \oplus SK_1\mathbf{Z}\pi$, now that π is abelian.

Theorem 4.1. *Suppose* π *has rank* r *and torsion subgroup* τ.

(a) (See [2, § 11], [5, Proposition 5.12], and [8].)

$$U(\mathbf{Z}\tau)/\pm\tau \to U(\mathbf{Z}\pi)/\pm\pi$$

is an isomorphism of torsion free groups.

(b) ([1, Ch. II]) $SL_n(\mathbf{Z}\pi) \twoheadrightarrow SK_1\mathbf{Z}\pi$ *is surjective for* $n \geqslant r + 2$.

The next theorem can be construed as a strengthened analogue, for K_0, of 4.1(b) above on K_1. Let A be any commutative ring with no non-trivial idempotents, e.g. $\mathbf{Z}\pi$. Then a projective A-module P (always assumed to be finitely generated) has a well defined rank $n \geqslant 0$ (see, e.g. [1, § 15]). We write $\det P = \Lambda^n P$. Since $\det(P \oplus Q) \cong \det P \otimes_A \det Q$, we obtain a homomorphism, which is surjective,

$$\det: \tilde{K}_0 A \to \mathrm{Pic}(A),$$

where Pic (A) denotes the group, under $\otimes_A$, of projective A-modules of rank one.

Theorem 4.2 (Bass-Murthy [5]). *For any abelian group* π, $\det: \tilde{K}_0\mathbf{Z}\pi \to \mathrm{Pic}(\mathbf{Z}\pi)$ *is an isomorphism.*

When π is finite this is equivalent to a "stable" form of Swan's Theorem 2.2 above.

Henceforth we shall deal with a finitely generated abelian group, which we can write as a direct product, $\pi \times T$, where π is finite, say of order m, and where T is free abelian of rank $r > 0$. (The case $r = 0$ is treated in § 2.) If p is a prime we shall write π_p for the Sylow p-subgroup of π. We also denote by $r_p(\pi)$ the number of irreducible representations of π over $\mathbf{Z}/p\mathbf{Z}$. Recall that $r_0(\pi)$ and $r_\infty(\pi)$ were defined similarly above, by replacing $Z/p\mathbf{Z}$ by $\mathbf{Q}$ and $\mathbf{R}$, respectively.

Theorem 4.3 (Bass-Murthy [5, § 8]).

$\mathrm{Pic}(\mathbf{Z}[\pi \times T]) \cong \mathrm{Pic}(\mathbf{Z}\pi) \oplus (T \otimes_{\mathbf{Z}} M(\pi)) \oplus N_0(\pi, T)$, *where:*

(a) $\mathrm{Pic}(\mathbf{Z}\pi)$ *is a finite group* (Theorem 2.2);

(b) $M(\pi)$ *is a certain free abelian group whose rank is*

$$\Big(\sum_{\substack{p \mid m,\\ p \text{ prime}}} r_p(\pi)\cdot(r_0(\pi_p) - 1)\Big) - (r_0(\pi) - 1).$$

$M(\pi) = 0$ *if and only if* m *is a prime power.*

(c) $N_0(\pi, T)$ *is a torsion group of exponent some power of* m. $N_0(\pi, T) = 0$ *if* m *is square free, and otherwise it is not even finitely generated.*

Theorem 4.4 (Bass-Murthy [5, § 10]). *With the notation above,*

$$\mathrm{Wh}(\pi \times T) \cong \mathrm{Wh}(\pi) \oplus (T \otimes_{\mathbf{Z}} \mathrm{Pic}(\mathbf{Z}\pi)) \oplus (\Lambda^2 T \otimes_{\mathbf{Z}} M(\pi)) \oplus N_1(\pi, T),$$

where $N_1(\pi, T)$ *is a torsion group in which each element has order dividing a power of* m. *(Hence* $N_1(\pi, T) = 0$ *if* $\pi = 0$, *i.e. if* $m = 1$.*) If* m *is not square free and if* $r \geqslant 2$ *then* $N_1(\pi, T)$ *is not finitely generated. The rank of* $\mathrm{Wh}(\pi \times T)$ *is*

$$(r_\infty(\pi) - r_0(\pi)) + \frac{(r(r-1))}{2}\cdot\Big(\Big(\sum_{\substack{p \mid m,\\ p \text{ prime}}} r_p(\pi)\cdot(r_0(\pi_p) - 1)\Big) - (r_0(\pi) - 1)\Big).$$

The term $N_1(\pi, T)$ above remains somewhat inaccessible, whereas all the others are reasonably well understood. In particular one

[1]) Added in proof. This question has been answered affirmatively by Carsten, using the retults of [4] and [14].

doesn't know whether $N_1(\pi, T)$ has bounded exponent, nor when, precisely, it is finitely generated. It is plausible that the latter occurs whenever m is square free.

Theorems 4.3 and 4.4 are proved with the aid of theorems of Grothendieck and of Bass-Heller-Swan (see [4]) describing $K_1A[T]$ when A is a regular ring. However $\mathbf{Z}[\pi \times T] = A[T]$ where $A = \mathbf{Z}\pi$ is never regular unless $\pi = 0$. The theorems just mentioned intervene on passing from $\mathbf{Z}\pi$ to its integral closure, and on passing from $\mathbf{Z}\pi$ modulo the conductor to the latter's reduction modulo its nil radical. A variety of techniques are used to carry out the detailed analysis.

Institute of Advanced Study,
Princeton, USA

REFERENCES

[1] Bass H., K-theory and stable algebra, Publ. IHES, no. 22 (1964).
[2] Bass H., The Dirichlet unit theorem, induced characters, and Whitehead groups of finite groups, *Topology*, vol. 4 (1966), 391-410.
[3] Bass H., Projective modules over free groups are free, *J. of Algebra*, 4 (1964), 367-373.
[4] Bass H., Heller A., Swan R., The Whitehead group of polynomial extension, Publ. IHES, no. 22 (1964).
[5] Bass H., Murthy M. P., Grothendieck groups and Picard groups of abelian group rings, *Ann. of Math.*, 86 (1967), 16-73.
[6] Bass H., Milnor J., Serre J.-P., Solution of the congruence subgroup problem for SL_n ($n \geqslant 3$) and Sp_{2n} ($n \geqslant 2$), Publ. IHES (to appear).
[7] Curtis C. W., Renier I., Representation theory of finite groups and associative algebras, Interscience (1962).
[8] Higman G., The units of groups rings, *Proc. Lond. Math. Soc.*, 46 (1940), 231-248.
[9] Lam T. Y., Columbia University thesis (1967).
[10] Milnor J., Whitehead torsion, *Bull. A.M.S.*, 3 (1966), 358-426.
[11] Seminaire G. de Rham, Torsion et type simple d'homotopie, Univ. de Lausanne (1963-64).
[12] Rim D. S., Modules over finite groups, *Ann. of Math.*, 69 (1959), 700-712.
[13] Siebenmann L. G., Obstruction to finding a boundary for an open manifold of dimension greater than five, Princeton University thesis (1965).
[14] Stallings J., Whitehead torsion of free products, *Ann. of Math.*, 82 (1965), 354-363.
[15] Swan R., Induced representations and projective modules, *Ann. of Math.*, 71 (1960), 552-578.
[16] Swan R., The Grothendieck ring of a finite group, *Topology*, 2 (1963), 85-110.
[17] Wall C. T. C., Finiteness conditions for CW-complexes, *Ann. of Math.*, 81 (1965), 56-69.
[18] Whitehead J. H. C., Simple homotopy types, *Amer. J. Math.*, 72 (1950), 1-57.

SOME PROBLEMS IN DIFFERENTIAL ALGEBRA

E. R. KOLCHIN

Introduction

It is 36 years since the publication of Ritt's first paper in differential algebra [1], and almost 16 years since his death. The fact is that during the twenty years preceding his death the subject progressed at a much faster rate than afterward. This is in no way due to a lack of worthwhile problems, but rather is a reflection of the scanty attention these problems have received. The experience and techniques of algebrists and especially algebraic geometers are particularly suited to differential algebra, and it is surprising that so few of them have turned their talents in this direction. I hope that my description of a few of the more challenging unsolved problems will lure some into accepting the challenge.

Consider a differential field $\mathfrak{F}$, i.e., a field in the usual sense (which we shall suppose to have characteristic 0) together with a finite number of mutually commuting derivation operators $\delta_1, \ldots, \delta_m$; when $m = 1$ the differential field is *ordinary* (and we usually write a' instead of $\delta_1 a$), and when $m > 1$ it is *partial*. Our problems have to do with algebraic differential equations of the form $A = 0$, where A is a differential polynomial over $\mathfrak{F}$ in a finite number n of differential indeterminates (i.e. A is a polynomial, with coefficients in $\mathfrak{F}$, in the derivatives $\delta_1^{i_1} \ldots \delta_m^{i_m} Y_j$ ($0 \leqslant i_1 \leqslant \infty, \ldots, 0 \leqslant i_m < \infty$, $1 \leqslant j \leqslant n$). The set $\mathfrak{R} = \mathfrak{F}\{Y_1, \ldots Y_n\}$ of all differential polynomials over $\mathfrak{F}$ in $Y_1, \ldots, Y_n$ is a differential ring. A subset Σ of $\mathfrak{R}$ defines a system of differential equations: $A = 0$ ($A \in \Sigma$). A solution $(\eta_1, \ldots, \eta_n)$ of this system, having coordinates in an extension (i.e. differential field extension) of $\mathfrak{F}$, is also called a *zero* of Σ. It is convenient (albeit not essential) to work in a fixed *universal* extension $\mathfrak{U}$ of $\mathfrak{F}$. Such a $\mathfrak{U}$, which always exists, is, roughly speaking, so large that it contains enough elements for every purpose we can have.

An ideal $\mathfrak{A}$ of $\mathfrak{R}$ is *differential* if $\delta_i \mathfrak{A} \subset \mathfrak{A}$ ($1 \leqslant i \leqslant m$), and is *perfect* if the condition $A^2 \in \mathfrak{A}$ implies that $A \in \mathfrak{A}$. The smallest differential ideal containing the set Σ is denoted by $[\Sigma]$, and the smallest perfect differential ideal containing Σ is denoted by $\{\Sigma\}$; because the field characteristic is 0, $\{\Sigma\}$ is the radical of $[\Sigma]$. The zeros of Σ are the same as the zeros of $\{\Sigma\}$. A fundamental fact is that Σ always has a finite subset Φ such that $\{\Phi\} = \{\Sigma\}$; this is the Ritt-Raudenbush basis theorem. As a consequence, every perfect

differential ideal $\mathfrak{A}$ is the intersection of finitely many prime differential ideals none of which contains another; these are the minimal prime differential ideals containing $\mathfrak{A}$, and are called the *components* of $\mathfrak{A}$. These general facts from the Ritt theory, as well as others referred to below, can be found in his book [2].

1. Singular solutions

Ritt's theory of components is especially interesting when $\mathfrak{A}$ is the perfect differential ideal $\{A\}$ generated by an irreducible differential polynomial $A \in \mathfrak{R}$. Consider any total ordering of the set of all derivatives $\delta_1^{i_1} \ldots \delta_m^{i_m} Y_j$, such that for all such derivatives u and v and all δ_i

$$\text{(i)}\ u < \delta_i u, \qquad \text{(ii)}\ u < v \Rightarrow \delta_i u < \delta_i v;$$

such orderings exist (e.g. the lexicographic ordering relative to $(i_1 + \ldots + i_m, j, i_1, \ldots, i_m)$) but in general are not unique. The highest derivative u present in A is called the *leader* of A, and the partial derivative $\partial A/\partial u$ is called the *separant* of A; of course, a different choice of ordering may give to A a different leader and separant. A zero of A is called *singular* if it is a zero of every separant of A.

Among the components of $\{A\}$ there is one, which we denote by $\mathfrak{p}(A)$, that contains *no* separant, whereas each other component contains *every* separant: $\mathfrak{p}(A)$ is the *general* component, and the others are the *singular* components, of A. Every zero of a singular component is singular, but a singular zero may be a zero of the general component. It is remarkable that all the singular components of A are the general components of other irreducible differential polynomials in $\mathfrak{R}$; furthermore, there is an effective procedure for finding these others. Yet, given a zero $\eta = (\eta_1, \ldots, \eta_n)$ and a component, no general method is known for deciding whether η is a zero of that component. Because it is possible to translate by η (extending $\mathfrak{F}$ if necessary), this question reduces to the following problem posed by Ritt:

Given an irreducible differential polynomial $A \in \mathfrak{R}$, to determine whether $(0, \ldots, 0)$ is a zero of $\mathfrak{p}(A)$.

Only very special cases have been solved. When $m = 1$, $n = 1$ (*ordinary* differential equation in *one* unknown), and the order of A is $\leqslant 2$ the solution was given by Ritt [3]. Beyond this there are only certain sufficient conditions and other necessary conditions, which are far apart; these conditions, successively developed by Ritt, H. Levi, A. Hillman, and me, are described in [4] where references are given.

The condition that $(0, \ldots, 0)$ be a zero of $\mathfrak{p}(A)$ is equivalent to the condition that $\mathfrak{p}(A) \subset [Y_1, \ldots, Y_n]$. Thus, the above problem about A is a special case of the following: given a prime differential ideal $\mathfrak{p}$ of $\mathfrak{R}$, to determine whether or not $\mathfrak{p}(A) \subset \mathfrak{p}$. A second interesting special case is that in which $\mathfrak{p}$ is the general component of another irreducible differential polynomial B: *Given A and B, to determine whether $\mathfrak{p}(A) \subset \mathfrak{p}(B)$.* Here, too, the results are meager.

2. Extensions of differential specializations

Let $(\eta_1, \ldots, \eta_n)$ and $(\zeta_1, \ldots, \zeta_n)$ be points of $\mathfrak{U}^n$. We say that $(\zeta_1, \ldots, \zeta_n)$ is a *differential specialization of* $(\eta_1, \ldots, \eta_n)$ over $\mathfrak{F}$, and write $(\eta_1, \ldots, \eta_n) \underset{\mathfrak{F}}{\to} (\zeta_1, \ldots, \zeta_n)$, if every differential polynomial in $\mathfrak{F}\{Y_1, \ldots, Y_n\}$ that vanishes at $(\eta_1, \ldots, \eta_n)$ vanishes also at $(\zeta_1, \ldots, \zeta_n)$. When this is the case, and when k is an integer with $1 \leqslant k < n$, then obviously $(\eta_1, \ldots, \eta_k) \underset{\mathfrak{F}}{\to} (\zeta_1, \ldots, \zeta_k)$; we say that the former differential specialization is an *extension* of this one.

Let $(\eta_1, \ldots, \eta_n)$ and k be given, and let $B \in \mathfrak{F}\{Y_1, \ldots, Y_n\}$, $B(\eta_1, \ldots, \eta_n) \neq 0$. It is known ([5], [6], [7]) that there exists a nonzero $B_0 \in \mathfrak{F}\{Y_1, \ldots, Y_k\}$ with $B_0(\eta_1, \ldots, \eta_k) \neq 0$ such that every differential specialization $(\eta_1, \ldots, \eta_k) \underset{\mathfrak{F}}{\to} (\zeta_1, \ldots \zeta_k)$ with $B_0(\zeta_1, \ldots, \zeta_k) \neq 0$ can be extended to a differential specialization $(\eta_1, \ldots, \eta_n) \underset{\mathfrak{F}}{\to} (\zeta_1, \ldots, \zeta_n)$ with $B(\zeta_1, \ldots, \zeta_n) \neq 0$.

It might be expected, in analogy with the usual notion of specialization as used in algebraic geometry, that every differential specialization $(\eta_1, \ldots, \eta_k) \underset{\mathfrak{F}}{\to} (\zeta_1, \ldots, \zeta_k)$ can be extended to some differential specialization $(\eta_1, \ldots, \eta_n) \underset{\mathfrak{F}}{\to} (\zeta_1, \ldots, \zeta_n)$, provided that some of the elements $\zeta_{k+1}, \ldots, \zeta_n$ are allowed to be ∞ in an appropriate sense. It is noteworthy that this is not so, even in the simple case in which $m = 1$, $n = 2$, $k = 1$, and $\mathfrak{F}$ is the ordinary differential field $\mathbf{C}(x)$ of rational functions of a complex variable x. For example, if we take $\eta_1 = \Gamma(x)$, where $\Gamma(x)$ is the usual gamma function, and $\eta_2 = \wp\left(\int \Gamma(x)^{-1} dx\right)$, where $\wp$ is a doubly periodic function of Weierstrass, then $\eta_1 \underset{\mathfrak{F}}{\to} 0$ but there does not exist any ζ_2 such that $(\eta_1, \eta_2) \underset{\mathfrak{F}}{\to} (0, \zeta_2)$ or $(\eta_1, \eta_2^{-1}) \underset{\mathfrak{F}}{\to} (0, \zeta_2)$.

The problem, then, is *to find a criterion that a differential specialization* $(\eta_1, \ldots, \eta_k) \underset{\mathfrak{F}}{\to} (\zeta_1, \ldots, \zeta_k)$ *can be extended to a differential specialization* $(\eta_1, \ldots, \eta_n) \underset{\mathfrak{F}}{\to} (\zeta_1, \ldots, \zeta_n)$.

A special case is met in the problem of indeterminate forms. Let $F, G \in \mathfrak{F}\{Y_1, \ldots, Y_n\}$ be relatively prime, with $G \neq 0$, and suppose that F and G vanish at $(0, \ldots, 0)$. The problem is to assign a value to the quotient F/G at $(0, \ldots, 0)$. If elements $t_1, \ldots, t_n \in \mathfrak{U}$ are differentially algebraically independent over $\mathfrak{F}$ and we set $u = F(t_1, \ldots, t_n)/G(t_1, \ldots, t_n)$, it is natural to say that F/G admits the value α at $(0, \ldots, 0)$ when $(t_1, \ldots, t_n, u) \underset{\mathfrak{F}}{\to} (0, \ldots, 0, \alpha)$. Thus, the problem becomes that of finding the extensions of $(t_1, \ldots, t_n) \underset{\mathfrak{F}}{\to} (0, \ldots, 0)$ to $(t_1, \ldots, t_n, u)$. This is equivalent to determining the elements $\alpha \in \mathfrak{U}$ such that $(0, \ldots, 0, \alpha)$ is a zero of the general component of the differential polynomial $Y_{n+1} G - F \in \mathfrak{F}\{Y_1, \ldots, Y_{n+1}\}$. Ritt conjectured [8] that either α is unique (possibly ∞) or else α is completely arbitrary: he proved his conjecture in the case $m = 1$, $n = 1$, ord $(FG) = 1$.

3. Differential dimension polynomials

A prime differential ideal $\mathfrak{p}$ of $\mathfrak{F}\{Y_1, \ldots, Y_n\}$ has a *generic zero:* this is a zero $(\eta_1, \ldots, \eta_n)$ such that every element of $\mathfrak{F}\{Y_1, \ldots, Y_n\}$ that vanishes at $(\eta_1, \ldots, \eta_n)$ is an element of $\mathfrak{p}$. The extension $\mathfrak{G} = \mathfrak{F}\langle\eta_1, \ldots, \eta_n\rangle$ has a finite differential transcendence degree that we denote by $d(\mathfrak{p})$. This number $d(\mathfrak{p})$ is called the *differential dimension* of $\mathfrak{p}$, and is the "correct" definition of what in the classical literature is called the number of arbitrary functions of m complex variables on which the solution of the system $P = 0$ $(P \in \mathfrak{p})$ depends; it measures the size of the set of zeros of $\mathfrak{p}$. If $\mathfrak{q}$ is another prime differential ideal, and if $\mathfrak{p} \subset \mathfrak{q}$, then $d(\mathfrak{p}) \geqslant d(\mathfrak{q})$, but if the inclusion is strict, the inequality need not be. For this reason a finer measure of the size is desirable. This is provided by the *differential dimension polynomial* $\omega_\mathfrak{p}$ of $\mathfrak{p}$; we recall its definition [9]. For any natural number s the ring $\mathfrak{R}_s = \mathfrak{F}[(\delta_1^{i_1} \ldots \delta_m^{i_m} Y_j)_{i_1+\ldots+i_m \leqslant s,\ 1 \leqslant j \leqslant n}]$ is a polynomial algebra over $\mathfrak{F}$ in $\binom{s+m}{m} n$ indeterminates; the intersection $\mathfrak{p} \cap \mathfrak{R}_s$ is a prime ideal of $\mathfrak{R}_s$ and therefore has a dimension in the usual sense; there exists a unique polynomial $\omega_\mathfrak{p}$ in one variable over $\mathbf{Q}$ such that $\dim(\mathfrak{p} \cap \mathfrak{R}_s) = \omega_\mathfrak{p}(s)$ for all sufficiently big natural numbers s. This polynomial has degree $\leqslant m$, and therefore can be written in the form $\omega_\mathfrak{p}(s) = \sum_{0 \leqslant i \leqslant m} \alpha_i(\mathfrak{p}) \binom{s+i}{i}$. The coefficients $\alpha_i(\mathfrak{p})$ are then integers. We order these differential dimension polynomials by defining $\omega_\mathfrak{p} \geqslant \omega_\mathfrak{q}$ to mean that $\omega_\mathfrak{p}(s) \geqslant \omega_\mathfrak{q}(s)$ for all sufficiently big integers s, i.e. we order them lexicographically with respect to $(a_m(\mathfrak{p}), \ldots, a_0(\mathfrak{p}))$. Then the inclusion $\mathfrak{p} \subset \mathfrak{q}$ implies the inequality $\omega_\mathfrak{p} \geqslant \omega_\mathfrak{q}$, and when the inclusion is strict so is the inequality. The connection with the differential dimension is given by the equation $a_m(\mathfrak{p}) = d(\mathfrak{p})$.

It may happen that $a_m(\mathfrak{p}) = 0$. The biggest natural number $\tau = \tau(\mathfrak{p})$ such that $\alpha_\tau(\mathfrak{p}) \neq 0$ is called the *differential type* of $\mathfrak{p}$, and $\alpha_\tau(\mathfrak{p})$ is called the *typical differential dimension of* $\mathfrak{p}$. This terminology is justified by the fact that if we regard $\mathfrak{F}$ and $\mathfrak{G}$ as differential fields with respect to τ new derivation operators $\delta_{i'} = \sum_{1 \leqslant i \leqslant m} c_{ii'} \delta_i$ $(1 \leqslant i' \leqslant \tau)$, where the $c_{ii'}$ are constants in $\mathfrak{F}$ subject to a certain inequation, then $\mathfrak{G}$ becomes a finitely generated extension of $\mathfrak{F}$ of differential transcendence degree $\alpha_\tau(\mathfrak{p})$.

The differential dimension polynomial $\omega_\mathfrak{p}$ is a birational invariant but not a differential birational invariant; this means that if η and ζ are generic zeros of $\mathfrak{p}$ and $\mathfrak{q}$, respectively, then the condition $\mathfrak{F}(\eta) = \mathfrak{F}(\zeta)$ implies that $\omega_\mathfrak{p} = \omega_\mathfrak{q}$ but the weaker condition $\mathfrak{F}\langle\eta\rangle = \mathfrak{F}\langle\zeta\rangle$ does not. Nevertheless, $\omega_\mathfrak{p}$ carries certain differential birational invariants with it. An obvious example is the differential dimension $a_m(\mathfrak{p})$; two others are $\tau(\mathfrak{p})$ and $\alpha_{\tau(\mathfrak{p})}(\mathfrak{p})$. It would be interesting to discover other differential birational invariants, for any such invariant should have great significance for $\mathfrak{p}$. I should mention that an interesting connection between $\omega_\mathfrak{p}$ and the characteristic polynomials of Hilbert-Serre has recently been discovered by Joseph L. Johnson; his work will appear in his Columbia University dissertation.

Consider a subset Σ of $\mathfrak{F}\{Y_1, \ldots, Y_n\}$. If the elements of Σ have bounded orders then the differential dimension polynomials of the components of $\{\Sigma\}$ are subject to certain restriction. Specifically, *if for each* Y_j *the order in* Y_j *of every element of* Σ *is* $\leqslant e_j$, *then for any component* $\mathfrak{p}$ *of* $\{\Sigma\}$, *the condition* $a_m(\mathfrak{p}) = 0$ *implies the condition* $a_{m-1}(\mathfrak{p}) \leqslant \sum e_j$. (This generalizes to any m the result proved by Ritt [2, p. 135] for $m = 1$.) There are grounds for conjecturing that in general

$$\alpha_{\tau(\mathfrak{p})}(\mathfrak{p}) \leqslant \sum_{1 \leqslant j \leqslant n} \binom{e_j - m - \tau(\mathfrak{p}) - 1}{m - \tau(\mathfrak{p})}.$$

When $\tau(\mathfrak{p}) = m$ this states that $a_m(\mathfrak{p}) \leqslant m$, which is obvious, and when $\tau(\mathfrak{p}) = m - 1$ this reduces to the result just mentioned.

When the set $\Sigma \subset \mathfrak{F}\{Y_1, \ldots, Y_n\}$ consists of precisely n differential polynomials $F_1, \ldots, F_n$ then there are two further conjectures. Set $e_{ij} = \mathrm{ord}_{Y_j} F_i$ $(1 \leqslant i \leqslant n, \quad i \leqslant j \leqslant n)$ and $h = \max_{\pi} (e_{1\pi(1)} + \ldots + e_{n\pi(n)})$, where π runs through the symmetric group $\mathfrak{S}_n$. This number h, heuristically arrived at by Jacobi, is in general smaller than the number $\sum e_j$. The first of the conjectures then is: *For any component* $\mathfrak{p}$ *of* $\{F_1, \ldots, F_n\}$, *if* $a_m(\mathfrak{p}) = 0$ *then* $a_{m-1}(\mathfrak{p}) \leqslant h$. Ritt [10] proved this conjecture in two special cases: 1° $m = 1$, $n = 2$; 2° $m = 1$, each F_j linear. Aside from these cases, nothing seems to be known. The second of the conjectures is: *For any component* $\mathfrak{p}$ *of* $\{F_1, \ldots, F_n\}$, *if* $a_m(\mathfrak{p}) = a_{m-1}(\mathfrak{p}) = 0$ *then* $\omega_{\mathfrak{p}} = 0$. When the F_j are linear this may be regarded as a precise formulation of a conjecture made by Janet [11]. Even in this case the problem is open.

4. Picard-Vessiot theory

Now let $\mathfrak{F}$ be an ordinary differential field and denote its field of constants by $\mathfrak{C}$; denote the field of constants of $\mathfrak{U}$ by $\mathfrak{K}$. Consider a homogeneous linear differential polynomial $L = Y^{(n)} + a_1 Y^{(n-1)} + \ldots + a_n Y$ with coefficients in $\mathfrak{F}$. If $\mathfrak{C}$ is algebraically closed then L has a fundamental system of zeros $(\eta_1, \ldots, \eta_n)$ such that the extension $\mathfrak{G} = \mathfrak{F}\langle \eta_1, \ldots, \eta_n \rangle$ of $\mathfrak{F}$ has the same field of constants $\mathfrak{C}$, and then the set of all isomorphisms over $\mathfrak{F}$ of $\mathfrak{G}$ onto an extension of $\mathfrak{F}$ in $\mathfrak{U}$ has a natural group structure; by means of the fundamental system of zeros $(\eta_1, \ldots, \eta_n)$ this group can be identified with an algebraic subgroup G of $GL_h(n)$ defined over $\mathfrak{C}$. This G can be called the *Galois group of* L over $\mathfrak{F}$. Of course, a different choice of fundamental systems of zeros in general gives a different Galois group; G is determined only up to conjugation in $GL(n)$ by a matrix that is rational over $\mathfrak{C}$. (This ambiguity can be removed by taking for G an algebraic subgroup of the group $GL(V)$ of all automorphisms of the vector space V of zeros of L.)

As in ordinary Galois theory, there are two general problems:

1° *Given* L, *to determine* G.

2° *Given* G, *to find an* L *of which* G *is the Galois group*.

Of course, the nature of these problems depends on the differential field $\mathfrak{F}$ which is given. For example, when $\mathfrak{F} = \mathfrak{C}$ then it is easy to see that, for every L, G is reducible to triangular form (and hence is solvable). A. Białynicki-Birula [12] has shown that if $\mathfrak{F}$ has finite transcendence degree over $\mathfrak{C}$ and G is a connected and nilpotent algebraic subgroup of $GL(n)$ defined over $\mathfrak{C}$, then there exists

a homogeneous linear differential polynomial L with coefficients in $\mathfrak{F}$ such that the Galois group of L over $\mathfrak{F}$ is isomorphic to G. Aside from this and some work in a slightly different direction by Lawrence Goldman [13], even when $\mathfrak{F} = \mathbf{C}(x)$ very little of a general nature is known. For the second problem this is not unexpected, in view of the history of the analogous problem in classical Galois theory. The obstinacy of the first problem is somewhat of a surprise, and it now appears that we should be grateful for almost any kind of information about G; the difficulty stems in part from the fact that a zero of L that is not a zero of any *linear* differential polynomial over $\mathbf{C}(x)$ of lower order may very well be a zero of a nonlinear one.

5. Rational approximation

Let me close with my favorite problem. This is to prove for algebraic differential equations an analog of the Thue-Siegel-Roth theorem [14] on approximation to algebraic numbers by rational ones.

Let K be any field and x be an indeterminate over K. The field $K((x^{-1}))$ of formal power series in x^{-1} over K contains the field $K(x^{-1}) = K(x)$ of rational fractions in x over K. There is a discrete nonarchimedean valuation of $K((x^{-1}))$ such that $\left|\sum_{k \geqslant r} c_k (x^{-1})^k\right| = e^{-r}$ when $c_r \neq 0$; $K((x^{-1}))$ is the completion of $K(x)$ with respect to this valuation.

Relative to the derivation operator d/dx, $K((x^{-1}))$ is an ordinary differential field; $K(x)$ is a differential subfield of $K((x^{-1}))$, and $K[x]$ is a differential subring of $K(x)$ such that $|b| \geqslant 1$ for every nonzero $b \in K[x]$. Suppose henceforth that the characteristic of K is 0. Then the field of constants of $K((x^{-1}))$ is K, and $|b'| = |b|e$ for every $b \in K((x^{-1}))$ such that $|b| \neq 1$.

Consider any differential polynomial P in one differential indeterminate y, and let z be another differential indeterminate. There is a smallest natural number d such that $P(y/z)z^d$ is a differential polynomial in y, z; we call this smallest d the *denomination* of P. (When P is a polynomial, i. e. a differential polynomial of order 0, then the denomination of P is its degree.) An *element* $u \in K((x^{-1}))$ that is differentially algebraic over $K(x)$ is said to have denomination d, if u is a zero of a differential polynomial in $K(x)\{y\}$ of denomination d but is not a zero of one of denomination $< d$.

It is known ([15], [16]) that *if* $u \in K((x^{-1}))$ *is differentially algebraic over* $K(x)$ *and of denomination* d *then there exists a real number* $\alpha > 0$ *such that*

$$\left|u - \frac{a}{b}\right| > \frac{\alpha}{|b|^d}$$

for all $a, b \in K[x]$ *with* $b \neq 0$ *and* $u \neq a/b$. This generalizes the analog for algebraic functions of Liouville's simple precurser of the Thue-Siegel-Roth theorem.

The problem, then, is *to show that the exponent d here can be replaced by any number greater than* 2. This seems to be extremely difficult, and at present any improvement in the exponent would be of interest. One of the difficulties arises from the fact that in Ritt's process of reducing a differential polynomial A, by a differential polynomial B (analog of Euclidean division of polynomials), it is necessary to multiply A by a power of the separant of B; this spoils all estimates.

Dept. of Mathematics,
Columbia University,
New York, USA

REFERENCES

[1] Ritt J. F., Manifolds of functions defined by systems of algebraic differential equations, *Trans. Amer. Math. Soc.*, **32** (1930), 369-398.
[2] Ritt J. F., Differential algebra, Amer. Math. Soc. Colloquium Publications, **33** (1950), viii + 184 pp.
[3] Ritt J. F., On the singular solutions of algebraic differential equations, *Annals of Math.*, **37** (1936), 552-617.
[4] Kolchin E. R., Singular solutions of algebraic differential equations and a lemma of Arnold Shapiro, *Topology*, **3**, Suppl. 2 (1965), 309-318.
[5] Ritt J. F., On a type of algebraic differential manifold, *Trans. Amer. Math. Soc.*, **48** (1940), 542-552.
[6] Seidenberg A., An elimination theory for differential algebra, Univ. California Publ. Math. (N.S.), 3 (1956), 31-65.
[7] Rosenfeld Azriel, Specializations in differential algebra, *Trans. Amer. Math. Soc.*, **90** (1959), 394-407.
[8] Ritt J. F., Indeterminate expressions involving an analytic function and its derivatives, *Monatshefte für Math.*, **43** (1936), 97-104.
[9] Kolchin E. R., The notion of dimension in the theory of algebraic differential equations, *Bul. Amer. Math. Soc.*, **70** (1964), 570-573.
[10] Ritt J. F., Jacobi's problem on the order of a system of differential equations, *Annals of Math.*, **36** (1935), 303-312.
[11] Janet M., Sur les systèmes aux derivées partielles comprenant autant d'équations que de fonctions inconnues, *Comples Rendues Acad. Sci. Paris*, **172** (1921), 1637-1639.
[12] Białynicki-Birula A., On the inverse problem of Galois Theory of differential fields, *Bul. Amer. Math. Soc.*, **69** (1963), 960-964.
[13] Goldman Lawrence, Specialization and Picard-Vessiot theory, *Trans. Amer. Math. Soc.*, **85** (1957), 327-356.
[14] Roth K. F., Rational approximation to algebraic numbers, *Mathematika*, **2** (1955), 1-20.
[15] Maillet E., Nombres transcendents, Paris, 1906.
[16] Kolchin E. R., Rational approximation to solutions of algebraic differential equations, *Proc. Amer. Math. Soc.*, **10** (1959), 238-244.

CLASSES OF ELEMENTS OF SEMISIMPLE ALGEBRAIC GROUPS

ROBERT STEINBERG

Given a semisimple algebraic group G, we ask: how are the elements of G partitioned into conjugacy classes and how do these classes fit together to form G? The answers to these questions not only throw light on the algebraic and topological structure of G, but are important for the representation theory and functional analysis of G; conversely, of course, we can expect the representations and functions on G to contribute to the answers to the questions. In what follows, we discuss some aspects of these questions, present some known results, and pose some unsolved problems.

To do this, we first recall some basic facts about semisimple algebraic groups. Our reference for all this is [7]. A linear algebraic group G is a subgroup of some $GL_n(K)$ ($n \geqslant 1$, K an algebraically closed field) which is the complete set of zeros of some set of polynomials over K in the n^2 matric entries. The Zariski topology, in which the closed sets are the algebraic subsets of G, is used. G is semisimple if it is connected and has no nontrivial connected solvable normal subgroup; it is then the product, possibly with amalgamation of centers, of some simple groups. The simple groups have been classified, in the Killing-Cartan tradition. Thus there are the classical groups (unimodular, symplectic, orthogonal, spin, etc.) and the five exceptional types. The rank of a semisimple group is the dimension of a maximal torus (a subgroup isomorphic to a product of GL_1's). We also need the notions of simple connectedness and Lie algebra as they apply to linear algebraic groups, but we omit the definitions. Henceforth K denotes a fixed algebraically closed field, p its characteristic, G a simply connected semisimple algebraic group over K, and r the rank of G. The simplest example is $G = SL_{r+1}(K)$. L denotes the Lie algebra of G. If x is an element of G, we write G_x (resp. L_x) for the centralizer of x in G (resp. L). We shorten conjugacy class to class, and dimension to dim.

Now we take up the problem of surveying the conjugacy classes of G and finding for them suitable representative elements. Recall that an element of G is semisimple if it is diagonalizable, and unipotent if its characteristic values are all 1, and that every element can be decomposed uniquely $x = x_s x_u$ as a product of commuting semisimple and unipotent elements. An element x will be called regular if $\dim G_x = r$, or, equivalently, if $\dim G_x$ is minimal [12, p. 49]; here x need not be semisimple and may in fact be unipotent. Various characterizations, hence alternate possible definitions, of regular elements

may be found in [12]. The regular (resp. semisimple) elements form a dense open set with complement of codimension 3 (resp. 1) in G [12, 1.3 and 6.8]; thus most elements are regular (resp. semisimple). In [12, 1.4] we have constructed a closed irreducible cross-section, C, isomorphic to affine r-space, for the regular classes of G. Also, we have proved [12, 1.2]:

(1) *The map $x \to x$ sets up a* 1-1 *correspondence between the regular and the semisimple classes of G.*

Thus by choosing the elements of C and their semisimple parts we obtain a system of representatives for the regular classes and for the semisimple classes of G. If G is the group SL_n, the representatives of the regular classes chosen turn out to be in one of the Jordan canonical forms. Thus we are led to our first problem.

(2) Problem. *Determine canonical representatives, similar to those given by the Jordan normal forms in SL_n, for all of the classes of G, not just the regular ones.*

In proving the results described above, we are naturally led to consider the algebra of regular class functions, those regular functions on G that are constant on the conjugacy classes of G. This is a polynomial algebra generated by the characters $\chi_1, \chi_2, \ldots, \chi_r$ of the fundamental representations of G [12, 6.1]. (If $G = SL_{r+1}$, the χ's are just the coefficients of the characteristic polynomial with the first and last terms excluded.) First of all these functions give a very useful characterization of regular elements of G: x is regular if and only if the differentials of $\chi_1, \chi_2, \ldots, \chi_r$ are linearly independent at x [12, 1.5]. And secondly they are related to our classification problem by the following result [12, 6.17]:

(3) *The map $x \to (\chi_1(x), \chi_2(x), \ldots, \chi_r(x))$ on G induces a bijection of the regular classes, and of the semisimple classes, onto the points of affine r-space.*

(4) Problem. *Assume that x and y in G are such that for every rational representation (ρ, V) of G the elements $\rho(x)$ and $\rho(y)$ are conjugate in $GL(V)$. Prove that x and y are conjugate in G.*

By (3) this holds if x and y are both regular or both semisimple, and it can presumably be checked when G is a classical group, but we have not done this in all cases.

To continue our discussion, we will use the following result.

(5) *If y is a semisimple element of G, then G_y is a connected reductive group (i.e. the product of a semisimple group and a central torus).*

More generally one can show: The group of fixed points of every semisimple automorphism of G is connected. Such connectedness theorems are important in problems concerning conjugacy classes, as we shall see, and also in problems about cohomology [1, p. 224].

Assume now that x and x' in G have the same semisimple part y. Then x and x' are conjugate in G if and only if x_u and x'_u are conjugate in G_y, in fact, by (5), in the semisimple part of G_y. The latter group may not be simply connected, but this is immaterial for the study of unipotent elements. Thus the study of arbitrary classes is reduced to the study of unipotent classes, and these are the clases we now consider.

At present not much is known about the classification of the unipotent classes of G. Observe, though, that (1) implies that regular unipotent elements exist and that they in fact form a single conjugacy class.

(6) Problem. *Classify the unipotent classes and determine canonical representatives for them.*

If $p = 0$, the study of unipotent classes in G is equivalent to that of nilpotent classes in L, and a classification of these latter classes together with representative elements may be found in [3; 4]. The result, however, is in the form of a list, which, though finite, is very long, thus subject to error and inconvenient for applications. The main procedure in the construction of the list, that of imbedding each nilpotent element in an algebra of type sl_2 and then using the representation theory of this algebra, or the corresponding procedure with G in place of L, is not available if $p \neq 0$ [12, p. 60]. In studying the nilpotent classes of L (still for $p = 0$) one is naturally led to study the class H of G-harmonic polynomials on L: indeed these polynomials are in natural correspondence with the regular functions on the variety of all nilpotent elements of L, by results in [5]. To our knowledge, no one has succeeded in effectively relating H to the classification problem at hand.

(7) Problem. *Do this.*

(8) Problem. *Do the same for the unipotent classes of G (for p arbitrary), having first found an analogue for H.*

Now it follows from (5) that a unipotent element centralized by a semisimple element not in the center of G is contained in a proper semisimple subgroup of G. Thus an important special case of (6) is:

(9) Problem. *Same as* (6) *for the classes of unipotent elements x such that G_x is the product of the center of G and a unipotent subgroup.*

As is easily seen, the class of regular unipotent elements always satisfies this condition. If $p = 0$, then relatively few other classes do [3, Th. 10.6].

A more modest problem is as follows.

(10) Problem. *Prove that the number of unipotent classes of G is finite.*

Here only classes which satisfy (9) need be considered. Observe that a proof of (4) would yield the finiteness together with a bound on the number of classes. From the work of Kostant one obtains a rather lengthy proof of (10) together with the bound 3^r, in case $p = 0$. Richardson [6] has found a simple, short proof that works not only if $p = 0$, but, more generally, if p does not divide any coefficient of

the highest root of any simple component of G, this root being expressed in terms of the simple ones. In the sequel such a value of p will be called good. Richardson's method depends eventually on the fact that an algebraic set has only a finite number of irreducible components; thus it yields no bound on the number of unipotent classes, but it does yield the following useful result.

(11) *If p is good and G has no simple component of type A_n, then L_x is the Lie algebra of G_x for every x in G.*

(12) Problem. *Extend* (11) *to all p and G, with the conclusion replaced by: L_x is the sum of the Lie algebra of G_x and the center of L.*

This is easy if x is semisimple and known if x is regular (cf. [12,4.3]).

Finally let us mention an old problem which seems to be closely related to (6).

(13) Problem. *Determine, within a general framework, the conjugacy classes of the Weyl group W of G, and assuming that W acts, as usual, on a real r-dimensional space V, relate them to the W-harmonic polynomials on V.*

Next we consider some results and problems about centralizers of elements of G. We have already stated, in (5) and (11) above, two such results. Springer [10] has proved that for any x in G the group G_x contains an Abelian subgroup of dimension r. It follows that if x is regular, then the identity component G_x^0 of G_x is Abelian.

(14) Problem. *If x is regular, prove that G_x is Abelian.*

(15) Problem. *Prove conversely that if G_x is Abelian, then x is regular.*

These results would yield an abstract characterization of the regular elements, but other such characterizations are known [12, 3.14]. Both problems may easily be reduced to the unipotent case. If x is a regular unipotent element, then, as already mentioned, G_x is the product of the center of G and a unipotent group U_x. Springer [11,4.1] has proved:

(16) *If p is good, then U_x is connected (hence Abelian).*

Thus (14) holds in this case. If p is bad, (16) is false, in fact $x \notin U_x^0$.

(17) Problem. *For p bad determine the structure of U_x/U_x^0 and whether U_x is Abelian.*

A solution to (16) would complete the proof of (14), and quite likely would also have cohomological applications (cf. [11, § 3]). We do not know whether (15) is true even if $p = 0$.

(18) Problem. *Prove that $\dim G_x - r$ is always even. In other word, the dimension of each conjugacy class is even.*

This has been proved in the following cases: if $p = 0$ [5, Prop. 15], if p is good (because of (11) the same type of proof as for $p = 0$ can be given), if x is semisimple (this is easy); and presumably can be checked when G is a classical group.

All of the above problems are special cases of one big final problem about centralizers.

(19) Problem. *For every element x of G determine the structure of G_x.*

Next we discuss briefly the individual classes, and their closures. Each class is, of course, an irreducible subvariety of G (because G is connected). A class is closed if and only if it is semisimple [12, 6.13]. More generally, the closure of a class consists of a number of classes of which exactly one is semisimple [12, 6.11]. Whether the number is always finite or not depends on (10). The most important case occurs when we start with the class U_0 of regular unipotent elements; then the closure U is just the variety of all unipotent elements of G. Although some results about U are known, e.g. U is a complete intersection specified by the equations $\chi_i(x) = \chi_i(1)$ $(1 \leqslant i \leqslant r)$ (the χ's are still the fundamental characters), and the codimension of $U - U_0$ in U is at least 2 [12, 6.11], we feel that U should be studied thoroughly.

(20) Problem. *Study U thoroughly.*

Of course this problem is related to (8).

Using the properties of U just mentioned and also (16), Springer has proved the following result.

(21) *If p is good, then the variety U (of unipotent elements of G) is isomorphic, as a G-space, to the variety N of nilpotent elements of L.*

If p is bad, this is false, because (16) is. If $p = 0$, Kostant [5] has obtained results about the structure and cohomology of N, in particular that N and any Cartan subalgebra of L intersect at 0 with multiplicity equal to the order of the Weyl group W.

(22) Problem. *Find a natural action of W on affine $(\dim G - r)$-space so that the quotient variety is isomorphic to N.*

So far we have been considering G over an algebraically closed field K. From now on we shall assume that G is defined over a perfect field k which has K as an algebraic closure; thus the polynomials which define G as an algebraic group can be chosen so that their coefficients are in k. The problem is to study the classes of G_k, the group of elements of G which are defined over k, i.e. which have their coordinates in k. The main idea is to relate the classes of G_k to those of G with the help of the Galois theory. We will discuss mainly the problem of surveying, and finding canonical representatives for, the classes of G_k. Recalling our solution to this problem for the regular (and semisimple) classes of G, we naturally try to adapt our cross-section C of the regular classes to the present situation. Consider a regular class of G which meets G_k. This class, call it R, is necessarily defined (as a variety) over k. How can we ensure that the representative element $C \cap R$ of R is in G_k? The obvious way is to construct C so that

it also is defined over k. For this to be possible, the unique unipotent element of C must be in G_k, and then the unique Borel (i.e. maximal connected solvable) subgroup of G containing this element [12, 3.2 and 3.3] must de defined over k; i.e. G must contain a Borel subgroup defined over k. This last condition is a natural one which arises in other classification problems, and, although a bit restrictive, holds for a significant class of groups, e.g. for the so-called Chevalley groups [2], of which one is the group SL_n. Conversely, if this condition holds, then it turns out that with a few exceptions (certain groups of type A_m (but not SL_{m+1}) are forbidden as simple components of G) C can be constructed over k [12, § 9]. Thus if we form $C \cap G_k$, we obtain a set of canonical representatives for those regular classes of G which meet G_k; but not necessarily for the regular classes of G_k since a single class of G may intersect G_k in several classes. In other words, elements of G_k conjugate in G may not be conjugate in G_k. Looking for an idea to remedy this situation, we consider the case $G = SL_n$. We have the Jordan normal forms for all the classes of $SL_n(K)$, not just the regular ones, and, though this fails in $SL_n(K)$, it does hold in $GL_n(k)$. Now the action of $GL_n(k)$ by conjugation is that of $PGL_n(k)$, i.e. of $\hat{G}_k$, if we let $\hat{G}$ denote the adjoint group of G, i.e. the quotient of G over its center, which is PSL_n in the present case. Thus we are led to the following problem.

(23) Problem. *Asume that G contains a Borel subgroup defined over k (and perhaps that G has no simple component of type A_m). (a) Prove that every class of G defined over k meets G_k, i.e. contains an element defined over k. (b) Prove that two elements of G_k which are conjugate in G are conjugate under $\hat{G}_k$.*

As we have seen, (a) holds for regular classes. It holds also for semisimple classes, without any restriction on the components of type A_m [12, 1.7].

(24) *Assume that G contains a Borel subgroup defined over k. Then every semisimple class of G defined over k meets G_k. (And conversely.)*

This result has a number of applications to classification problems [12, p. 51-2], especially if (cohomological) dim $k \leqslant 1$. We recall [8, p. II-8] that dim $k \leqslant 1$ is equivalent to: every finite dimensional division algebra over k is commutative. This holds in several cases of interest in number theory, in particular if k is finite [8, p. II-10]. As a consequence of (24), we have [12, 1.9]:

(25) *If dim $k \leqslant 1$ and H is any connected linear group defined over k, then the Galois cohomology $H^1(k, H)$ is trivial.*

This result and an extension due to Springer [8, p. III-16] leap to the classification of semisimple groups defined over k, if dim $k \leqslant 1$ (we get [12, 10.2] that there is always a Borel subgroup defined over k), and answer most of the questions raised above. Concerning (23a), we get [12, 10.2]:

(26) *If k and H are as in* (25), *then every class of H defined over k meets H_k.*

Concerning (23b), we get [21, 10.3]:

(27) *If* dim $k \leqslant 1$, *then two semisimple elements of G_k which are conjugate in G are conjugate in G_k.*

This result is false for arbitrary elements, e.g. for regular ones. To show a typical kind of argument, we give the proof of (27). Let y and z be semisimple elements of G_k conjugate in G; thus $aya^{-1} = z$ with a in G. Combining this equation with the one obtained by applying a general element γ of the Galois group of K over k, we see that $a^{-1}\gamma(a)$, call it x_γ, is in G_y. As we check at once, x satisfies the cocycle condition $x_{\gamma\delta} = x_\gamma \gamma(x_\delta)$, hence represents an element of $H^1(k, G_y)$, which is trivial, by (25), because G_y is connected, by (5). Thus $x_\gamma = b\gamma(b^{-1})$ for all γ and some b in G_y. We conclude that y and z are conjugate in G_k, by the element ab in fact. The same argument, with (16) in place of (5), shows that regular unipotent elements of $\hat{G}_k$ which are conjugate in $\hat{G}$ (the adjoint group) are conjugate in $\hat{G}_k$, if p is good (here dim $k \leqslant 1$ is not necessary, but p good is).

By combining (26) and (27), we get a complete survey of the semisimple classes of G_k, if dim $k \leqslant 1$; in particular, these classes correspond to the rational points of affine r-space, suitably defined over k (cf. [12, 10.3]). Thus, if k is a finite field of q elements, so that G_k is a simply connected version of one of the simple finite Chevalley groups or their twisted analogues, the number of such classes is q^r, a useful fact in the representation theory of these groups [11]. A related result, whose proof, unfortunately, does not use the same ideas, states that the number of unipotent elements of G_k is $q^{\dim G - r}$ [13].

Finally we conclude our paper with another problem.

(28) Problem. *If G is defined over k (any perfect field), prove that every unipotent class of G is defined over k.*

This would yield a proof of (10). In fact, assuming $p \neq 0$, as we may, choosing k as the field of p elements, applying (28) with G suitably defined, and then using (26), we would get (10) with the number bounded by $|G_k|$.

Dept. of Mathematics,
University of California, Los Angeles, USA

REFERENCES

[1] Borel A., Sous-groupes commutatifs et torsion des groupes de Lie compacts connexes, *Tôhoku Math. J.*, **13** (1961), 216-240.
[2] Chevalley C., Sur certains groupes simples, *Tôhoku Math. J.*, **7** (1955), 14-66.
[3] Дынкин Е. Б., Полупростые подалгебры полупростых алгебр Ли, *Матем. сб.*, **30** (72) (1952), 349-462.

[4] Kostant B., The principal three-dimensional subgroup and the Betti numbers of a complex simple Lie group, *Amer. J. Math.*, **81** (1959), 973-1032.
[5] Kostant B., Lie group representations on polynomial rings, *Amer. J. Math.*, **85** (1963), 327-404.
[6] Richardson R. W., Conjugacy classes in Lie algebras and algebraic groups, to appear.
[7] Séminaire C. Chevalley, Classification des Groupes de Lie Algébriques (two volumes), Paris (1956-8).
[8] Serre J.-P., Cohomologie Galoisienne, Lecture Notes, Springer-Verlag, Berlin.
[9] Springer T. A., Some arithmetical results on semisimple Lie algebras, Publ. Math. I.H.E.S., **30** (1966), 115-141.
[10] Springer T. A., A note on centralizers in semisimple groups, to appear.
[11] Steinberg R., Representations of algebraic groups, *Nagoya Math. J.*, **22** (1963), 33-56.
[12] Steinberg R., Regular elements of semisimple algebraic groups, Publ. Math. I.H.E.S., No. 25 (1965), 48-80.
[13] Steinberg R., Endomorpisms of linear algebraic groups, Memoirs Amer. Math. Soc., to appear.

О НЕКОТОРЫХ ПРОБЛЕМАХ БЕРНСАЙДОВСКОГО ТИПА

Е. С. ГОЛОД

Целью доклада является описание приема, который позволяет строить контрпримеры для некоторых проблем бернсайдовского типа [1] в случае «неограниченного» показателя. Как известно, в случае «ограниченного» показателя большинство из этих проблем имеет положительное решение [2, 3, 4], за исключением собственно проблемы Бернсайда о периодических группах [5]. Далее будет доказана следующая

Теорема. Пусть k — произвольное поле. Существует k-алгебра A (без единицы) с $d \geqslant 2$ образующими, которая бесконечномерна как векторное пространство над k, в которой всякая подалгебра с числом образующих $< d$ нильпотентна и которая, такова, что $\bigcap_{n=1}^{\infty} A^n = (0)$.

Частными случаями этой теоремы являются результаты, полученные тем же методом в [6], а также следующие утверждения:

Следствие 1. Существует финитно-аппроксимируемая бесконечная p-группа с $d \geqslant 2$ образующими, в которой всякая подгруппа с числом образующих $< d$ конечна [1]).

[1]) Как доказал С. П. Струнков [7], если в группе с числом образующих $d \geqslant 3$ всякая собственная подгруппа конечна, то и сама группа также конечна.

Доказательство. Пусть A — алгебра над полем F_p из p элементов с d образующими $x_1, \ldots, x_d$, которая удовлетворяет теореме. Рассмотрим в алгебре $\hat{A}$, полученной из A присоединением единицы 1, мультипликативную подполугруппу, порожденную элементами $1 + x_i$ $(i = 1, \ldots, d)$ и их обратными $(1 + x_i)^{-1}$ (которые существуют, так как все элементы в A нильпотентны). Очевидно, что G — группа, причем p-группа. Если $1 + u_j$ $(j = 1, \ldots, s, s < d)$ — элементы из G, то порожденная ими подгруппа конечна, так как все ее элементы имеют вид $1 + u$, где u принадлежит подалгебре, порожденной элементами u_j, которая нильпотентна, следовательно, конечномерна над k и поэтому состоит из конечного числа элементов. Группа G бесконечна, поскольку бесконечномерная алгебра $\hat{A}$ является гомоморфным образом групповой алгебры группы G над F_p. Наконец, группа G финитно-аппроксимируема, так как если элемент $g = 1 + u \in G$ лежит в пересечении всех членов нижнего центрального ряда группы G, то $u \in \bigcap_{n=1}^{\infty} A^n$, а потому $u = 0$, $g = 1$.

Следствие 2. Существует финитно-аппроксимируемая нильпотентная группа G с числом образующих $d \geqslant 3$, которая удовлетворяет условию Энгеля, т. е. для любых $x, y \in G$ существует $n = n(x, y)$, такое, что

$$[\ldots \underbrace{[[x, y], y] \ldots y}_{n \text{ раз}}] = 1 \text{ [1])}.$$

Доказательство. Частный случай следствия 1.

Следствие 3. Пусть k — произвольное поле. Существует k-алгебра Ли L с $d \geqslant 2$ образующими, которая бесконечномерна как векторное пространство над k, в которой всякая подалгебра с числом образующих $< d$ нильпотентна и которая такова, что $\bigcap_{n=1}^{\infty} L^n = (0)$.

Доказательство. Пусть A — ассоциативная алгебра над k с образующими $x_1, \ldots, x_d$, удовлетворяющая условиям теоремы. Пусть L—алгебра Ли, порожденная элементами $x_1, \ldots, x_d$ относительно операции коммутирования $[x, y] = xy - yx$. Нужно проверить только, что алгебра L бесконечномерна над k. Это вытекает из того, что A как векторное пространство над k порождается произведениями вида $y_{i_1}^{\beta_1} \ldots y_{i_s}^{\beta_s}$, где $y_1, y_2, \ldots$ — некоторый

[1]) Аналогичное утверждение справедливо и в случае $d = 2$, однако тогда оно не является формальным следствием теоремы, а должно доказываться непосредственно тем же приемом, что и теорема.

k-базис в L и $i_1 < i_2 < \ldots < i_s$. Если бы алгебра L была конечномерна, то, поскольку все элементы в A нильпотентны, из элементов указанного выше вида можно было бы выбрать конечный базис для A — противоречие.

Следствие 4. Существует бесконечномерная алгебра Ли с числом образующих $d \geqslant 3$, удовлетворяющая условию Энгеля (так же как и в следствии 2 с показателем n, зависящим от x, y) [1]).

Доказательство. Частный случай следствия 3.

Доказательство теоремы основывается на признаке бесконечномерности алгебры, доказанном в [8]. Ниже этот признак формулируется в виде леммы и воспроизводится та часть из рассуждений в [8], которая необходима для ее доказательства.

Лемма. Пусть R — свободная ассоциативная алгебра от d образующих над некоторым полем k, I — идеал в R, порожденный последовательностью форм (т. е. однородных элементов) степени $\geqslant 2$, в которой для всякого i число форм степени i конечно и равно r_i. Пусть $r_i \leqslant s_i$. Если все коэффициенты степенного ряда

$$\left(1 - dt + \sum_{i=2}^{\infty} s_i t^i\right)^{-1}$$

неотрицательны, то факторалгебра $A = R/I$ бесконечномерна. В частности, это справедливо, если $r_i \leqslant \varepsilon^2 (d - 2\varepsilon)^{i-2}$ $(\varepsilon > 0)$.

Доказательство. Алгебра $A = R/I$ является градуированной. Обозначим через A_i ее i-ю составляющую: $A = \coprod_{i=0}^{\infty} A_i$. Пусть $x_1, \ldots, x_d$ — образующие свободной алгебры R, $\bar{x}_1, \ldots, \bar{x}_d$ — их канонические образы в A. Вообще, если x — элемент из R, то его канонический образ в A обозначаем через $\bar{x}$. Пусть $f_1, f_2, \ldots$ — формы, порождающие идеал I, а степень f_j равна k_j. Если $i < 0$, то полагаем $A_i = 0$. Пусть B_i — прямая сумма d экземпляров k-векторного пространства A_{i-1}: $B_i = \coprod_{d \text{ раз}} A_{i-1}$ и $C_i = \coprod_{j=1}^{\infty} A_{i-k_j}$. Определим отображения α: $B_i \to A_i$ и β_i: $C_i \to B_i$. Отображение α определяется по формуле

$$(a_1, \ldots, a_d) \to a_1\bar{x}_1 + \ldots + a_d\bar{x}_d, \quad a_\nu \in A_{i-1}, \ \nu = 1, \ldots, d.$$

Отображение β есть сумма отображений β_j: $A_{i-k_j} \to B_i$, где β_j определяется следующим образом: запишем форму f_j в виде

$$f_j = \sum_{\nu=1}^{d} f_j^\nu x_\nu,$$

[1]) См. сноску на предыдущей странице.

где f_j^ν — некоторые формы степени $k_j - 1$; тогда

$$\beta_j(z) = (z\bar{f}_j^1, z\bar{f}_j^2, \ldots, z\bar{f}_j^d), \qquad z \in A_{i-k_j}.$$

Покажем, что последовательность

$$C_i \xrightarrow{\beta} B_i \xrightarrow{\alpha} A_i \to 0 \tag{1}$$

точная, если $i > 0$. Очевидно, что α сюръективно и что $\alpha \circ \beta = 0$. Проверим, что если $a = (a_1, \ldots, a_d) \in \operatorname{Ker} \alpha$, то $a \in \operatorname{Im} \beta$. Представим элементы a_ν в виде $a_\nu = \bar{y}_\nu$, где y_ν — некоторые элементы степени $i - 1$ из R. Так как $\alpha(a) = 0$, то $y = \sum_{\nu=1}^{d} y_\nu x_\nu \in I$, а потому $\sum_{\nu=1}^{d} y_\nu x_\nu = \sum_{\nu=1}^{d} u_\nu x_\nu + \sum_{j=1}^{\infty} z_j f_j$, где $u_\nu \in I$ и $z_j \in R$ — однородные элементы степени $i - k_j$ соответственно. Тогда если $z = (\bar{z}_1, \bar{z}_2, \ldots)$, то, очевидно, $\beta(z) = a$.

Пусть теперь $b_i = \dim A_i$. Точная последовательность (1) дает неравенство

$$b_i \geqslant d b_{i-1} - \sum_{j=1}^{\infty} b_{i-k_j}.$$

Умножая его на t^i и суммируя по всем i от 1 до ∞, получим неравенство для формальных степенных рядов (которое нужно понимать как неравенство, выполняющееся для всех соответствующих коэффициентов в правой и левой частях):

$$\sum_{i=1}^{\infty} b_i t^i \geqslant dt \sum_{i=1}^{\infty} b_{i-1} t^{i-1} - \sum_{j=1}^{\infty} t^{k_j} \sum_{i=1}^{\infty} b_{i-k_j} t^{i-k_j}.$$

Пусть $P_A(t) = \sum_{i=0}^{\infty} b_i t^i$. Так как $b_0 = 1$, имеем

$$P_A(t) - 1 \geqslant dt P_A(t) - \left(\sum_{k=2}^{\infty} r_k t^k\right) P_A(t)$$

или

$$\left(1 - dt + \sum_{k=1}^{\infty} r_k t^k\right) P_A(t) \geqslant 1.$$

Если коэффициенты степенного ряда $\left(1 - dt + \sum_{k=2}^{\infty} r_k t^k\right)^{-1}$ неотрицательны, то

$$P_A(t) \geqslant \left(1 - dt + \sum_{k=2}^{\infty} r_k t^k\right)^{-1}$$

и алгебра A бесконечномерна. Наконец, если $s_i \geqslant r_i$ и неотрицательны коэффициенты степенного ряда $\left(1 - dt + \sum_{i=2}^{\infty} s_i t^i\right)^{-1}$, то, как легко видеть,

$$\left(1 - dt + \sum_{i=2}^{\infty} r_i t^i\right)^{-1} \geqslant \left(1 - dt + \sum_{i=2}^{\infty} s_i t^i\right)^{-1}.$$

Доказательство теоремы. Пусть снова R — свободная ассоциативная алгебра от d образующих $x_1, \ldots, x_d$ над полем k, R^+ — совокупность ее элементов, не имеющих свободных членов. В силу леммы для построения бесконечномерной алгебры A с d образующими, в которой любые $d-1$ элементов порождают нильпотентную подалгебру, достаточно составить в R последовательность форм $f_1, f_2, \ldots$ степени $\geqslant 2$, такую, что 1) $r_i \leqslant \varepsilon^2 (d - 2\varepsilon)^{i-2}$ для всякого $i \geqslant 2$; 2) если I — идеал, порожденный последовательностью форм $f_1, f_2, \ldots$, то для любых $d-1$ элементов $y_1, \ldots, y_{d-1}$ из R^+ существует показатель $n = n(y_1, \ldots, y_{d-1})$, такой, что всякий одночлен от $y_1, \ldots, y_{d-1}$ степени $\geqslant n$ лежит в I. Последовательность форм с указанными свойствами строится по индукции. Предположим, что уже построена конечная система форм $f_1, \ldots, f_{n_{k-1}}$, удовлетворяющая условию 1) и такая, что условие 2) выполняется для идеала I_{k-1}, порожденного формами $f_1, \ldots, f_{n_{k-1}}$, и элементов степени $\leqslant k-1$. Покажем, что эту систему можно расширить до конечной системы форм $f_1, f_2, \ldots, f_{n_k}$, по-прежнему удовлетворяющей условию 1), а условию 2) — для идеала $I_k = (f_1, \ldots, f_{n_k})$ и элементов степени $\leqslant k$. Рассмотрим $d-1$ общих элементов степени k:

$$g_i = c_{1i}^{(1)} x_1 + \ldots + c_{di}^{(1)} x_d + c_{1i}^{(2)} x_1^2 + \ldots + c_{d^k, i}^{(k)} x_d^k, \quad i = 1, \ldots, d-1,$$

и произвольное их произведение степени N:

$$g_{i_1} g_{i_2} \cdots g_{i_N}.$$

Это выражение есть многочлен от коэффициентов $c_{11}^{(1)}, \ldots, c_{d^k, d-1}^{(k)}$, которые рассматриваются как неизвестные. Коэффициентами при одночленах от $c_{11}^{(1)}, \ldots, c_{d^k, d-1}^{(k)}$ служат формы от $x_1, \ldots, x_d$ степени i, где $N \leqslant i \leqslant kN$. Именно эти формы мы и добавляем к построенным ранее для получения системы $f_1, \ldots, f_{n_k}$. Тогда, очевидно, условие 2) для элементов степени $\leqslant k$ выполняется. Выполнение условия 1) обеспечивается за счет выбора числа N достаточно большим. Прежде всего пусть N больше, чем степень любой из форм $f_1, \ldots, f_{n_{k-1}}$. В этом случае общее число форм степени $\geqslant N$ в системе $f_1, \ldots, f_{n_k}$ равно числу коммутативных одночленов степени N

от $q = (d-1)(d + \ldots + d^k)$ неизвестных, умноженному на $(d-1)^N$ — число некоммутативных одночленов $g_{i_1} \ldots g_{i_N}$ степени N от $d-1$ неизвестных. Это число, таким образом, есть

$$(d-1)^N \binom{N+q-1}{q-1} \leqslant (d-1)^N (N+q-1)^{q-1}.$$

Пусть $\alpha = \frac{d-2\varepsilon}{d-1}$. Если $\varepsilon < 1/2$, то $\alpha > 1$. Поэтому для достаточно больших значений N будем иметь

$$(N+q-1)^{q-1} \leqslant \frac{\varepsilon^2}{(d-2\varepsilon)^2} \alpha^N,$$

откуда для $i \geqslant N$ получаем

$$r_i \leqslant (d-1)^N (N+q-1)^{q-1} \leqslant \varepsilon^2 (d-2\varepsilon)^{N-2}.$$

При $i < N$ условие 1) также выполняется, ибо оно выполнялось для системы $f_1, \ldots, f_{n_{k-1}}$. Таким образом, система $f_1, \ldots, f_{n_k}$ построена. Объединение всех этих систем $f_1, \ldots, f_{n_k}$, $k = 1, 2, \ldots$, и является искомой последовательностью форм $f_1, f_2, \ldots$. Алгебра $A = R^+/I$ удовлетворяет условиям теоремы.

Московский текстильный институт,
Москва, СССР

ЛИТЕРАТУРА

[1] Курош А. Г., Проблемы теории колец, связанные с проблемой Бернсайда о периодических группах, *Известия АН СССР*, сер. матем., **5** (1941), 233-241.
[2] Levitzki J., On three problems concerning nil-rings, *Bull. Amer. Math. Soc.*, **51** (1945), 913-919.
[3] Jacobson N., Structure theory for algebraic algebras of bounded degree, *Ann. Math.*, **46** (1945), 695-707.
[4] Кострикин А. И., О проблеме Бернсайда, *Известия АН СССР*, сер. матем., **23** (1959), 1-34.
[5] Новиков П. С., О периодических группах, *Доклады АН СССР*, **127** (1959), 749-752.
[6] Голод Е. С., О ниль-алгебрах и финитно-аппроксимируемых p-группах, *Известия АН СССР*, сер. матем., **28** (1964), 273-276.
[7] Струпков С. П., Нормализаторы и абелевы подгруппы некоторых классов групп, *Известия АН СССР*, сер. матем., **31** (1967), 657-670.
[8] Голод Е. С., Шафаревич И. Р., О башне полей классов, *Известия АН СССР*, сер. матем., **28** (1964), 13-24.

3

Теория чисел
Theory of numbers
Théorie des nombres
Zahlentheorie

NUMBER FIELDS AND ZETA FUNCTIONS ASSOCIATED WITH DISCONTINUOUS GROUPS AND ALGEBRAIC VARIETIES

GORO SHIMURA

1. Introduction

Let Γ be an arithmetically defined discontinuous group operating on a bounded symmetric domain H, and φ a holomorphic mapping of H into a projective space which induces a biregular morphism of H/Γ onto a Zariski open subset V of a projective variety. Then one can set the following problems.

(A) Find a model V defined over an algebraic number field.

(B) With suitably chosen V and φ, characterize number-theoretically the field generated by the coordinates of $\varphi(z_0)$ for an elliptic fixed point z_0 of an analytic automorphism α of H such that $\alpha\Gamma\alpha^{-1}$ is commensurable with Γ.

(C) Find a connection between the zeta function of V and the Hecke operators defined for the automorphic forms with respect to Γ.

A typical example is the case where Γ is a principal congruence subgroup of the modular group $SL_2(\mathbf{Z})$ and H is the upper half plane. Then one finds a curve V so as to be defined over the rational number field $\mathbf{Q}$. The classical theory of complex multiplication solves Problem (B). The last problem (C) has been investigated by Eichler [1] and myself [8], [9].

The main purpose of this lecture is to answer the above questions for all possible arithmetic Γ (at least up to commensurability) acting on the upper half plane. This provides a complete analogue to the results obtained for $SL_2(\mathbf{Z})$. It turns out especially that the maximal abelian extension of a totally imaginary quadratic extension of a totally real algebraic number field F can be generated, over the maximal abelian extension of F, by the special values of automorphic functions of one variable. Although the principal part is thus concerned with the one-dimensional H, I shall first discuss Problems (A) and (B) in the higher dimensional case, because it will give a clear sight to the whole theory.

2. The moduli-variety for a family of abelian varieties

In many cases, we can attach to H and Γ a family $\Sigma = \{Q_z \mid z \in H\}$, parametrized by the points on H, of structures Q_z, each of which is formed by an abelian variety A_z, a polarization of A_z, endomorphisms of A_z, and points of finite order on A_z. Two members Q_z and Q_w of Σ are isomorphic if and only if $\gamma(z) = w$ for some $\gamma \in \Gamma$. We say that H/Γ is of *type* (P), if the members of Σ can be characterized only by the type of polarization, endomorphisms, and points of finite order. In such a case, one can approach Problems (A) and (B) as follows. First the algebro-geometric characterization of the members enables us to show the existence of an algebraic number field k with the property that, *for every* $Q_z \in \Sigma$ *and an automorphism* σ *of* $\mathbf{C}$, σ *is the identity mapping on* k *if and only if* $(Q_z)^\sigma$ *is isomorphic to a member of* Σ. Then one can choose V and φ so that the following conditions are satisfied [13].

(2.1) *V is defined over k.*

(2.2) *If Q and Q' are structures isomorphic to members Q_z and Q_w of Σ respectively, and if Q' is a specialization of Q over k, then $(\varphi(w), Q')$ is a specialization of $(\varphi(z), Q)$ over k.*

For a given Σ, such V and φ are unique up to biregular morphisms over k. The field $k(\varphi(z))$, for each point $z \in H$, has an invariant meaning for the structure Q_z, since it can be characterized by the following property.

(2.3) *An automorphism σ of $\mathbf{C}$ is the identity mapping on $k(\varphi(z))$ if and only if $(Q_z)^\sigma$ is isomorphic to Q_z.*

We call $k(\varphi(z))$ *the field of moduli of* Q_z.

Now if z is an isolated fixed point of an analytic automorphism of H described in (B), then A_z has sufficiently many complex multiplications. Therefore the determination of the number field $k(\varphi(z))$ can be essentially done by applying the theory of complex multiplication of abelian varieties [16] to A_z. However it is not always a simple task but actually an interesting problem to determine the structure of A_z. The study of some special ones among these A_z is very effectively employed to characterize the field k number-theoretically [12].

It should be observed that this recognition of H/Γ as moduli-variety does not necessarily give the ultimate answer to our problems, for some reasons which will be explained in the following section. Nevertheless, by this means, one can obtain at least the first approximation to the solution, which is really the best possible in a number of cases, as is seen in the classical case of $SL_2(\mathbf{Z})$ and elliptic curves.

3. Reflections on the idea of Section 2

There are some spaces H'/Γ' which are not of type (P), but can be embedded holomorphically into H/Γ of type (P) through injections $H' \to H$ and $\Gamma' \to \Gamma$. I found a non-trivial example for this in the case of quaternion unitary group defined by Siegel [17], and communicated the possibility of generalization to Kuga, who, together with Satake, has investigated such an embedding in a general framework (cf. [4], [6], [7]). In such a case, one can still show that H'/Γ' has a model defined over an algebraic number field. To show this, first we take V and φ for H/Γ so as to satisfy (2.1) and (2.2), and then choose a point z on H' so that A_z has sufficiently many complex multiplications. The points w on H' such that A_w is isogenous to A_z form a dense subset X of H'. Since $k(\varphi(w))$ is the field of moduli of Q_w, we see that the coordinates of the points in $\varphi(X)$ are all algebraic numbers. Therefore, if σ is an automorphism of $\mathbf{C}$ over the algebraic closure of $\mathbf{Q}$, then one has $\varphi(X) = \varphi(X)^\sigma \subset \varphi(H')^\sigma$, so that $\varphi(H') = \varphi(H')^\sigma$. It follows that $\varphi(H')$, a model for H'/Γ', is defined over an algebraic number field. The case of unitary groups of quaternion hermitian or anti-hermitian forms will be discussed in detail in [14]. In this way, we know that H'/Γ' has a model defined over an algebraic number field, so far as there is a family of abelian varieties attached to H'/Γ' in the sense of Kuga and Satake. A careful analysis of the special members Q_z and the fields $k(\varphi(z))$ may yield a more precise result.

The field k determined for the family Σ of § 2 is not necessarily the smallest field of definition for the field of automorphic functions with respect to Γ. In reality, k is a number field which has an essential meaning for the family Σ of abelian varieties, but not always so for H/Γ. For example, to a certain H/Γ, one can attach infinitely many distinct families of abelian varieties; the field k may depend on the choice of a family. These phenomena can happen both in the cases of type (P) and not of type (P). Therefore if we ask for "absolute" statements concerning H/Γ without reference to abelian varieties, the answer of § 2 to our questions may be considered incomplete in such a case. However, by studying these families more closely, it is possible to obtain an "absolute result" at least for the one-dimensional H, which is our object in the following sections.

4. Construction of class fields by automorphic functions with respect to an arithmetic fuchsian group

Let B be a quaternion algebra over a totally real algebraic number field F of degree g. If r is the number of archimedean primes of F unramified in B, one can identify $B_{\mathbf{R}} = B \otimes_{\mathbf{Q}} \mathbf{R}$ with the product

$M_2(\mathbf{R})^r \times \mathbf{K}^{g-r}$ of r copies of the total matrix ring $M_2(\mathbf{R})$ of degree 2 over the real number field $\mathbf{R}$, and $g - r$ copies of the division ring $\mathbf{K}$ of real quaternions. For every $\alpha \in B_{\mathbf{R}}$, let $\alpha = (\alpha_1, \ldots, \alpha_g)$ with $\alpha_\nu \in M_2(\mathbf{R})$ or $\mathbf{K}$ according as $\nu \leqslant r$ or $\nu > r$. Denote by B^+ the set of all α in B such that $\det(\alpha_\nu) > 0$ for $\nu \leqslant r$. If $\alpha \in B^+$, the action of α on the product $\mathfrak{H}^r$ of r copies of the upper half plane $\mathfrak{H} = \{z \in \mathbf{C} \mid \mathrm{Im}(z) > 0\}$ is defined by

$$\alpha(z_1, \ldots, z_r) = (w_1, \ldots, w_r),$$

$$w_\nu = (a_\nu z_\nu + b_\nu)/(c_\nu z_\nu + d_\nu), \qquad \alpha_\nu = \begin{bmatrix} a_\nu & b_\nu \\ c_\nu & d_\nu \end{bmatrix}.$$

Let $\mathfrak{o}$ be a maximal order in B, and $\mathfrak{c}$ an integral ideal in F. Let $\Gamma(\mathfrak{o}, \mathfrak{c})$ denote the set of all units γ in $\mathfrak{o}$, belonging to B^+, such that $1 - \gamma \in \mathfrak{c}\,\mathfrak{o}$. Then $\Gamma(\mathfrak{o}, \mathfrak{c})$ can be considered as a discontinuous group acting on $\mathfrak{H}^r$. It is well-known that $\mathfrak{H}^r/\Gamma(\mathfrak{o}, \mathfrak{c})$ is compact if B is a division ring, especially if $r < g$. Hereafter let us assume that $r = 1$, and the archimedean prime of F corresponding to the unique factor $M_2(\mathbf{R})$ of $B_{\mathbf{R}}$ is obtained from the identity mapping of F.

Lemma. *Let M be a totally imaginary quadratic extension of F which is isomorphic to a quadratic subfield of B over F. Then the following assertions hold.*

(1) *If f is an F-linear isomorphism of M into B, then $f(M) - \{0\} \subset B^+$, and every element in $f(M) - F$ has exactly one fixed point on $\mathfrak{H}$ which is common to all elements of $f(M) - F$.*

(2) *If $\mathfrak{r}_M$ denotes the ring of integers in M, then there exists an F-linear isomorphism f of M into B such that $f(\mathfrak{r}_M) \subset \mathfrak{o}$.*

For an algebraic number field K of finite degree and an integral ideal $\mathfrak{a}$ in K, we denote by $C(K, \mathfrak{a})$ the abelian extension of K in which a prime ideal $\mathfrak{p}$ of K, prime to $\mathfrak{a}$, decomposes completely if and only if $\mathfrak{p} = (b)$ with an element b of K which is congruent multiplicatively to 1 modulo $\mathfrak{a}$ and positive at every archimedean prime of K.

Theorem 1. *The notation and assumption being as above, there exists a complete non-singular algebraic curve V and a holomorphic mapping φ of $\mathfrak{H}$ into V, satisfying the following three conditions.*

(1.1) *V is defined over $C(F, \mathfrak{c})$.*

(1.2) *φ gives a biregular isomorphism of $\mathfrak{H}/\Gamma(\mathfrak{o}, \mathfrak{c})$ into V.*

(1.3) *Let M and f be as in (2) of Lemma, and z the fixed point of the elements of $f(M) - F$ on $\mathfrak{H}$. Then $C(M, \mathfrak{c})$ is the composite of $M(\varphi(z))$ and $C(F, \mathfrak{c})$.*

We call such a (V, φ) a *canonical model* for $\mathfrak{H}/\Gamma(\mathfrak{o}, \mathfrak{c})$. If (V, φ) and (V', φ') are two canonical models for $\mathfrak{H}/\Gamma(\mathfrak{o}, \mathfrak{c})$, then one can show the existence of a biregular morphism j of V to V', defined over

$C(F, \mathfrak{c})$, such that $j \circ \varphi = \varphi'$. In this sense, (V, φ) is uniquely determined for $\mathfrak{H}/\Gamma(\mathfrak{o}, \mathfrak{c})$.

The reciprocity-law for the extension $C(M, \mathfrak{c})/M$ can be described explicitly in terms of the special points $\varphi(z)$. For simplicity, let us consider the case $\mathfrak{c} = (1)$. Classify all the right $\mathfrak{o}$-ideals with respect to the left multiplication by the elements of B^+. Let $\{\mathfrak{x}_1, \ldots, \mathfrak{x}_h\}$ be a set of representatives for the classes of right $\mathfrak{o}$-ideals in this sense. Let $\mathfrak{v}_\lambda$ be the left order of $\mathfrak{x}_\lambda$. Set $\mathfrak{x}_{\lambda\mu} = \mathfrak{x}_\lambda \mathfrak{x}_\mu^{-1}$. For every right $\mathfrak{o}_\mu$-ideal $\mathfrak{x}$, let $N(\mathfrak{x})$ denote the ideal in F generated by the elements $N(\xi)$ for all $\xi \in \mathfrak{x}$, where $N(\xi)$ denotes the reduced norm of ξ to F. We notice that, for each μ, $\{N(\mathfrak{x}_{1\mu}), \ldots, N(\mathfrak{x}_{h\mu})\}$ is a set of representatives for the ideal-classes modulo the product $\mathfrak{u}$ of all archimedean primes of F, and $h = [C(F, 1): F]$.

Theorem II. *The notation being as above, there exists a system* $\{V_\lambda, \varphi_\lambda\ (\lambda = 1, \ldots, h)\}$ *satisfying the following conditions.*

(II.1) *For each* λ, $(V_\lambda, \varphi_\lambda)$ *is a canonical model for* $\mathfrak{H}/\Gamma(\mathfrak{o}_\lambda, 1)$.

(II.2) $V_\lambda = V_\mu^\sigma$ *if* $\sigma = \left(\frac{C(F, 1)/F}{N(\mathfrak{x}_{\lambda\mu})}\right)$.

(II.3) *Let* M *and* f *be as in Lemma, under the condition* $f(\mathfrak{r}_M) \subset \mathfrak{o}_\mu$. *Let* z *be the fixed point on* $\mathfrak{H}$ *of the elements of* $f(M) - F$, *and* $\mathfrak{b}$ *an ideal in* M. *Suppose that* f *is normalized in the sense defined below. Then there exist an element* α *of* B^+ *and a unique* λ *such that* $\alpha \mathfrak{x}_{\lambda\mu} = f(\mathfrak{b})\, \mathfrak{o}_\mu$. *With such an element* α *and* $\tau = \left(\frac{C(M, 1)/M}{\mathfrak{b}}\right)$, *one has* $\varphi_\mu(z)^\tau = \varphi_\lambda(\alpha^{-1}(z))$.

Here we say that f is *normalized* if $(d/dw)[f(a)(w)]_{w=z} = \bar{a}/a$ for every $a \in M$, $\neq 0$. If we define $\bar{f}$ by $\bar{f}(a) = f(\bar{a})$ for $a \in M$, then we see that either f or $\bar{f}$ is normalized.

Example. Let $d = 7$, 9 or 11, and let $F = F_d = \mathbf{Q}(\zeta + \zeta^{-1})$ with $\zeta = e^{2\pi i/d}$. Then there exists a quaternion algebra B over F which is unramified at exactly one archimedean prime of F corresponding to the identity mapping of F, and unramified at every finite prime spot of F. Such a B is unique up to isomorphisms over F. In this case we have $h = 1$ and $C(F, 1) = F$. Let $\mathfrak{o}$ be a maximal order in B. For every maximal order $\mathfrak{o}'$ in B, there exists an element β of B^+ such that $\mathfrak{o}' = \beta \mathfrak{o} \beta^{-1}$. One can show that $\Gamma(\mathfrak{o}, 1)$, modulo the center, is a triangle group with three classes of elliptic elements of order 2, 3, d, hence $\mathfrak{H}/\Gamma(\mathfrak{o}, 1)$ is of genus 0. Let w_2, w_3, w_d be corresponding elliptic points on $\mathfrak{H}$ which are unique modulo $\Gamma(\mathfrak{o}, 1)$. Then there is an automorphic function φ on $\mathfrak{H}$ with respect to $\Gamma(\mathfrak{o}, 1)$ which can be characterized by the property that $\varphi(w_2) = 1$, $\varphi(w_3) = 0$, $\varphi(w_d) = \infty$ and φ gives a biregular morphism of $\mathfrak{H}/\Gamma(\mathfrak{o}, 1)$ onto the complex projective line V. Then it can be shown that this (V, φ) is a canonical model for $\mathfrak{H}/\Gamma(\mathfrak{o}, 1)$. In this case, for *every* totally imaginary quadratic extension M of F, there exists an F-linear isomorphism f of M into B such that $f(\mathfrak{r}_M) \subset \mathfrak{o}$. Let $\{z_1, \ldots, z_q\}$ be a set of representatives for the $\Gamma(\mathfrak{o}, 1)$-equivalence classes of the fixed points of $f(M) - F$ for all such f. Then q is exactly the class number of M; and from the statements (I.3) and (II.3), one obtains the following assertion.

The values $\varphi(z_1), \ldots, \varphi(z_q)$ *form a complete set of conjugates of* $\varphi(z_1)$ *over* M, *and* $M(\varphi(z_1))$ *is the maximal unramified abelian extension of* M *for each* i.

Thus φ affords a complete analogue of the classical modular function $j(\tau)$, with the field F_d in place of $\mathbf{Q}$.

Remark 1. We can actually prove certain existence-theorems of a canonical model (V, φ) for $\mathfrak{H}^r/\Gamma(\mathfrak{o}, \mathfrak{c})$ with any $r \geqq 1$, which include the above theorems as particular cases. The cases where $r = 1$ and $r = g$ will be treated in [15]. We can also release the condition $f(\mathfrak{r}_M) \subset \mathfrak{o}$ in (I.3) by replacing $C(M, \mathfrak{c})$ by a suitably defined class field over M. I shall discuss the most general theorems including all these in a paper subsequent to [15].

Remark 2. It may be worth inserting a short history of the fuchsian group of our type. In the case $F = \mathbf{Q}$, Poincaré found that the integral transformations of an indefinite ternary quadratic form generate a fuchsian group [5]. He realized this when he *was walking on a cliff*, as he described it as an example of mathematical discoveries in his "Science et Méthode". Then Fricke treated the case of F of higher degree. A considerable part of two volumes of Fricke-Klein [2] is devoted to the investigation of this group and its transformation-equation. The possibility of transformation theory was already noticed in [5]. If $B = M_2(F)$, the group $\Gamma(\mathfrak{o}, 1)$ is essentially the so-called Hilbert modular group. The case where $r = g = 2$ was studied by Hecke in his dissertation [3]. In this work, discussing the analytic functions meaningful for the construction of abelian extensions, Hecke, curiously enough, referred to the results of Fricke mentioned above as being *without specific meaning in number theory* (Werke, p. 23). Actually, as I have shown, Fricke's automorphic functions provide a more transparent result for the construction of class fields than the functions of Hilbert-Hecke type, though we can regard both as special cases of a more general theory as noted in Remark 1. The discontinuous groups defined by Siegel [17] may be considered as a generalization of the Poincaré-Fricke group to the higher dimensional case. These are of course special cases of what recent authors call arithmetically defined discontinuous groups.

5. The zeta functions of B and V

Let V_λ, φ_λ and $\mathfrak{o}_\lambda$ be as in Th. II. Set $\Gamma_\lambda = \Gamma(\mathfrak{o}_\lambda, 1)$. We regard B as a subring of $M_2(\mathbf{R})$ through the projection of $B_\mathbf{R}$ to $M_2(\mathbf{R})$. Then

$N(\alpha) = \det(\alpha)$ for every $\alpha \in B$. For $\xi = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in GL_2(\mathbf{R})$ with $\det(\xi) > 0$, and for $z \in \mathfrak{H}$, set $j(\xi, z) = \det(\xi)^{1/2}(cz + d)^{-1}$. Let $S_m(\Gamma_\lambda)$, for a positive integer m, denote the vector space of all cusp forms of weight m with respect to Γ_λ. Every element of $S_m(\Gamma_\lambda)$ is a holomorphic function $f(z)$ on $\mathfrak{H}$ such that $f(\gamma(z)) = f(z)\, j(\gamma, z)^m$ for all $\gamma \in \Gamma_\lambda$. If $\alpha \in B^+$, one can define a linear map $(\Gamma_\mu \alpha \Gamma_\lambda)_m$: $S_m(\Gamma_\lambda) \to S_m(\Gamma_\mu)$ as follows. Let $\Gamma_\mu \alpha \Gamma_\lambda = \bigcup_{i=1}^{e} \alpha_i \Gamma_\lambda$ be a disjoint union. Then, for $f \in S_m(\Gamma_\lambda)$,

$$(\Gamma_\mu \alpha \Gamma_\lambda)_m f = \sum_{i=1}^{e} f(\alpha_i^{-1}(z))\, j(\alpha_i^{-1}, z)^m.$$

For an integral ideal $\mathfrak{a}$ in F, let $T_{\mu\lambda}(\mathfrak{a})_m$ denote the sum of all the $(\Gamma_\mu \alpha \Gamma_\lambda)_m$ such that $\alpha \mathfrak{x}_{\lambda\mu}$ is integral and $N(\alpha \mathfrak{x}_{\lambda\mu}) = \mathfrak{a}$. It should be observed that λ and $\mathfrak{a}$ determine μ uniquely. Further let $R_{\nu\lambda}(\mathfrak{a})_m = (\Gamma_\nu \beta \Gamma_\lambda)_m$ with an element $\beta \in B^+$ such that $\beta \mathfrak{x}_{\lambda\nu} = \mathfrak{a}\mathfrak{o}_\nu$. Then we let $T(\mathfrak{a})_m$ (resp. $R(\mathfrak{a})_m$) denote the linear transformation on

$$S_m = S_m(\Gamma_1) \oplus \ldots \oplus S_m(\Gamma_h)$$

whose restriction to $S_m(\Gamma_\lambda)$ coincides with $T_{\mu\lambda}(\mathfrak{a})_m$ (resp. $R_{\nu\lambda}(\mathfrak{a})_m$). Now define a Dirichlet series $D_m(s)$ by

$$D_m(s) = \sum T(\mathfrak{a})_m N(\mathfrak{a})^{-s},$$

where $\mathfrak{a}$ runs over all integral ideals in F. It can be shown that $D_m(s)$ converges for sufficiently large $\mathrm{Re}(s)$ and has an Euler product:

$$D_m(s) = \prod [1 - T(\mathfrak{p})_m N(\mathfrak{p})^{-s}]^{-1} \times \prod [1 - T(\mathfrak{p})_m N(\mathfrak{p})^{-s} + R(\mathfrak{p})_m N(\mathfrak{p})^{1-2s}]^{-1},$$

where the first product is taken over all the prime ideals $\mathfrak{p}$ ramified in B, and the second over all $\mathfrak{p}$ unramified in B. Moreover, $D_m(s)$ is holomorphically continued to the whole s-plane and satisfies a functional equation:

$$R_m(s) = \Gamma_m(s) D_m(s) = ZY \cdot R_m(2-s) \cdot Y^{-1},$$
$$\Gamma_m(s) = (2\pi)^{-gs} d^{s/2} \Gamma(s)^{g-1} \Gamma(s - 1 + m/2),$$

where Z and Y are certain invertible linear transformations on S_m, and d is a certain integer (for details, see [11]).

Let (V, φ) be as in Th. I. Let $\mathfrak{p}$ be a prime ideal in F, and $\mathfrak{P}$ a prime ideal in $C(F, 1)$ dividing $\mathfrak{p}$. Let us denote by $\mathfrak{P}(V)$ the reduction of V modulo $\mathfrak{P}$, and assume that $\mathfrak{P}(V)$ is a non-singular curve. The zeta function $Z_{\mathfrak{P}}(u)$ of the curve $\mathfrak{P}(V)$, over the residue

field mod $\mathfrak{P}$, has the form

$$Z_{\mathfrak{P}}(u) = Z^1_{\mathfrak{P}}(u)/[(1-u)(1 - N(\mathfrak{P})u)],$$

where $Z^1_{\mathfrak{P}}(u)$ is a polynomial in u of degree $2t$, if t denotes the genus of V.

Theorem III. *The notation being as above, for almost all $\mathfrak{p}$ which is unramified in B, one has*

$$\Pi_{\mathfrak{P}/\mathfrak{p}} Z^1_{\mathfrak{P}}(N(\mathfrak{P})^{-s}) = \det[1 - T(\mathfrak{p})_2 N(\mathfrak{p})^{-s} + R(\mathfrak{p})_2 N(\mathfrak{p})^{1-2s}].$$

It follows that if we define the global zeta function $\zeta(s, V)$ of V, in the sense of Hasse and Weil, by

$$\zeta(s, V) = \Pi'_{\mathfrak{P}} Z^1_{\mathfrak{P}}(N(\mathfrak{P})^{-s})^{-1},$$

the product being taken over all $\mathfrak{P}$ with good reduction, then $\zeta(s, V)$ differs from $\det[D_2(s)]$ only by a finite number of $\mathfrak{p}$-factors. Therefore $\zeta(s, V)$ can be continued holomorphically to the whole s-plane and satisfies a functional equation. We can actually obtain a more general result on certain L-functions for the coverings of V defined by the congruence subgroups of Γ_λ [15].

6. Ideas of proofs

One can attach two types of families of abelian varieties to the quotient $\mathfrak{H}^r/\Gamma(\mathfrak{o}, \mathfrak{c})$. Let K be a totally imaginary quadratic extension of F and let $\tau_1, \ldots, \tau_g$ be isomorphisms of K into $\mathbf{C}$ such that $\{\tau_1, \ldots, \tau_g, \tau_1\rho, \ldots, \tau_g\rho\}$ is the set of all isomorphisms of K into $\mathbf{C}$, where ρ means the complex conjugation. Let $L = B \otimes_F K$. We consider an abelian variety A belonging to one of the following two types.

(i) $\dim(A) = 2g$; there exists an isomorphism θ of K into $\mathrm{End}_{\mathbf{Q}}(A)$ such that the analytic representation (i.e. the representation in the tangent space of A at the origin) of $\theta(a)$ is equivalent to $\sum_{\nu=1}^{r}(\tau_\nu + \tau_\nu\rho) + 2\sum_{\nu=r+1}^{g}\tau_\nu$.

(ii) $\dim(A) = 4g$; and there exists an isomorphism θ of L into $\mathrm{End}_{\mathbf{Q}}(A)$ such that the analytic representation of $\theta(a)$ for $a \in K$ is equivalent to $2\sum_{\nu=1}^{r}(\tau_\nu + \tau_\nu\rho) + 4\sum_{\nu=r+1}^{g}\tau_\nu$.

With a suitable polarization and points of finite order, one obtains a family $\Sigma = \{Q_z \mid z \in \mathfrak{H}^r\}$ of the type described in § 1, with $\Gamma = \Gamma(\mathfrak{o}, \mathfrak{c})$. For the families of type (i), one has to assume that B splits over K. Let us assume $r = 1$, though one can investigate the general case $r \geqq 1$ by the same method. Let K' denote the field generated over $\mathbf{Q}$ by $\sum_{\nu=2}^{g} x^{\tau_\nu}$ for all $x \in K$. Define the field k as in § 1 for this Σ. It can be shown that $k \subset C(K', \mathfrak{c})$. Here and in the following we assume that $\mathfrak{c}$ is generated by a rational integer. (The general

case can be easily reduced to this case.) Take V and φ satisfying (2.1-2) for the present Σ. Let M and z be as in (I.3). Then either KM or $B \otimes_F KM$ is contained in the endomorphism algebra of the special member Q_z according as Q_z is of the above type (i) or (ii). In any case Q_z has sufficiently many complex multiplications. The general theory [16] tells us that the field of moduli of Q_z, which coincides with $k(\varphi(z))$ by virtue of (2.3), is an abelian extension E of $K'M$. The nature of the class field E has been investigated in [10]. Up to this point, we have chosen one K, and constructed (V, φ) with respect to this fixed K. Now there are infinitely many choices of K. The class field $C(M, \mathfrak{c})$ can be obtained as the intersection of the composite $E \cdot C(K', \mathfrak{c})$ for all K. This fact enables us to apply Weil's criterion [18] to lower the field of definition to $C(F, \mathfrak{c})$, and then to find a new model (V, φ) satisfying (I.1-3). Theorem II can be proved by investigating more closely the isogenies between Q_z and its transforms by Frobenius automorphisms of E.

To explain the proof of Th. III, let us first define some algebraic correspondences. Let $\{V_\lambda, \varphi_\lambda\}$ be as in Th. II, and let

$$X_{\mu\lambda}(\mathfrak{p}) = \{\varphi_\lambda(z) \times \varphi_\mu(\alpha(z)) \mid z \in \mathfrak{H}\} \subset V_\lambda \times V_\mu,$$
$$Y_{\mu\nu}(\mathfrak{p}) = \{\varphi_\nu(z) \times \varphi_\mu(\beta(z)) \mid z \in \mathfrak{H}\} \subset V_\nu \times V_\mu,$$

with elements α and β of B^+ such that $\alpha\mathfrak{r}_{\lambda\mu}$ is integral and $N(\alpha\mathfrak{r}_{\lambda\mu}) = \mathfrak{p}$, $\beta\mathfrak{r}_{\nu\mu} = \mathfrak{p}\mathfrak{o}_\mu$. Further let $\lambda \to \lambda'$ be a permutation of $\{1, \ldots, h\}$ such that $N(\mathfrak{r}_{\lambda'})$ and $N(\mathfrak{r}_\lambda^{-1})$ belong to the same ideal class modulo $\mathfrak{u}$ in F, and let δ_λ be an element of B^+ such that $\delta_\lambda \mathfrak{r}_\lambda = N(\mathfrak{r}_\lambda)\mathfrak{r}_{\lambda'}$ (cf. [11, p. 257]). Let

$$U_\lambda = \{\varphi_\lambda(z) \times \varphi_{\lambda'}(\delta_\lambda(z)) \mid z \in \mathfrak{H}\} \subset V_\lambda \times V_{\lambda'}.$$

Then $X_{\mu\lambda}(\mathfrak{p})$, $Y_{\mu\nu}(\mathfrak{p})$ and U_λ are algebraic correspondences defined over $C(F, 1)$, and U_λ is birational. The operators $T_{\mu\lambda}(\mathfrak{p})_2$ and $R_{\mu\nu}(\mathfrak{p})_2$ are nothing but the representations of $X_{\mu\lambda}(\mathfrak{p})$ and $Y_{\mu\nu}(\mathfrak{p})$ by differential forms of the first kind on the curves. Let $\mathfrak{P}$ be a prime ideal in $C(F, 1)$ dividing $\mathfrak{p}$. Denote by tilde the reduction modulo $\mathfrak{P}$. Then one has

$$\tilde{X}_{\mu\lambda}(\mathfrak{p}) = {}^t\Pi_\mu + \tilde{U}_\mu^{-1} \circ \Pi_{\lambda'} \circ \tilde{U}_\lambda, \tag{5.1}$$

$$\tilde{Y}_{\mu\nu}(\mathfrak{p}) = \tilde{U}_\mu^{-1} \circ \tilde{U}_\lambda^{N(\mathfrak{p})}. \tag{5.2}$$

Here Π_λ denotes the locus of $x \times x^{N(\mathfrak{p})}$ on $\tilde{V}_\lambda \times \tilde{V}_\lambda^{N(\mathfrak{p})}$, and ${}^t\Pi_\lambda$ its transpose. These congruence relations are obtained from the reciprocity law (II.3). By a simple computation, one can easily derive Th. III from (5.1-2).

Dept. of Mathematics, Princeton University,
Princeton, New Jersey, USA

REFERENCES

[1] Eichler M., Quaternäre quadratische Formen und die Riemannsche Vermutung für die Kongruenzzetafunktion, *Arch. Math.*, **5** (1954), 355-366.
[2] Fricke R., Klein F., Vorlesungen über die Theorie der automorphen Funktionen, I, II, Leipzig, Teubner, 1897-1912.
[3] Hecke E., Höhere Modulfunktionen und ihre Anwendung auf die Zahlentheorie, *Math. Ann.*, **71** (1912), 1-37 (Werke, 21-57).
[4] Kuga M., Fibre varieties over a symmetric space whose fibres are abelian varieties, lecture notes, Univ. of Chicago, 1963-64.
[5] Poincaré H., Les fonctions fuchsiennes et l'arithmétique, *Journ. de math.*, (4) **3** (1887), 405-464 (Oeuvre II, 463-511).
[6] Satake I., Holomorphic imbeddings of symmetric domains into a Siegel space, *Amer. J. Math.*, **87** (1965), 425-461.
[7] Satake I., Symplectic representations of algebraic groups satisfying a certain analyticity condition, to appear.
[8] Shimura G., Correspondances modulaires et les fonctions ζ de courbes algébriques, *J. Math. Soc. Japan.*, **10** (1958), 1-28.
[9] Shimura G., On the zeta-functions of the algebraic curves uniformized by certain automorphic functions, *J. Math. Soc. Japan.*, **13** (1961), 275-331.
[10] Shimura G., On the class-fields obtained by complex multiplication of abelian varieties, *Osaka Math. J.*, **14** (1962), 33-44.
[11] Shimura G., On Dirichlet series and abelian varieties attached to automorphic forms, *Ann. of Math.*, **76** (1962), 237-294.
[12] Shimura G., On the field of definition for a field of automorphic functions, I, II, III, *Ann. of Math.*, **80** (1964), 160-189; **81** (1965), 124-165; **83** (1966), 377-385.
[13] Shimura G., Moduli and fibre systems of abelian varieties, *Ann. of Math.*, **83** (1966), 294-338.
[14] Shimura G., Discontinuous groups and abelian varieties, *Math Ann.*, **168** (1967), 171-199.
[15] Shimura G., Construction of class fields and zeta functions of algebraic curves, *Ann. of Math.*, **85** (1967), 58-159.
[16] Shimura G., Taniyama Y., Complex multiplication of abelian varieties and its applications to number theory, *Publ. Math. Soc. Japan.*, No. 6, 1961.
[17] Siegel C. L., Symplectic geometry, *Amer. J. Math.*, **65** (1943), 1-86 (Gesammelte Abhandlungen II, 274-359).
[18] Weil A., The field of definition of a variety, *Amer. J. Math.*, **78** (1956), 509-524.

ТРАНСЦЕНДЕНТНОСТЬ И АЛГЕБРАИЧЕСКАЯ НЕЗАВИСИМОСТЬ ЗНАЧЕНИЙ *E*-ФУНКЦИЙ

А. Б. ШИДЛОВСКИЙ

В 1929—1930 гг. К. Зигель [1] опубликовал метод, который позволяет устанавливать трансцендентность и алгебраическую независимость значений в алгебраических точках одного достаточно

широкого класса целых функций, названных им E-функциями, являющихся решениями линейных дифференциальных уравнений с полиномиальными коэффициентами. Этот метод является непосредственным обобщением известных классических результатов Ш. Эрмита о трансцендентности числа e и Ф. Линдемана о трансцендентности и алгебраической независимости значений показательной функции в алгебраических точках, а также использует обобщение идеи А. Туэ из теории приближения алгебраических чисел рациональными дробями.

Зигель называет целую функцию

$$f(z)=\sum_{n=0}^{\infty} c_n \frac{z^n}{n!}$$

E-функцией, если: 1) все коэффициенты c_n функции $f(z)$ принадлежат алгебраическому полю K конечной степени над полем рациональных чисел; 2) при любом $\varepsilon > 0$

$$|\overline{c_n}| = O(n^{\varepsilon n}),$$

где $|\overline{\alpha}|$ — максимум модулей алгебраического числа α и всех его сопряженных относительно поля K; 3) существует последовательность натуральных чисел $\{q_n\}$, такая, что числа $q_n c_k$ $(k = 0, 1, \ldots, n)$ — целые алгебраические, а при любом $\varepsilon > 0$

$$q_n = O(n^{\varepsilon n}).$$

Нетрудно убедиться, что E-функции образуют кольцо функций, замкнутое по отношению к операциям дифференцирования, интегрирования в пределах от 0 до z и замены аргумента z на λz, где λ — алгебраическое число.

Любая алгебраическая постоянная, всякий многочлен от z с алгебраическими коэффициентами, функции e^z, $\sin z$, $\cos z$, функция Бесселя $J_0(z)$ являются простейшими примерами E-функций.

Свой метод Зигель применил к функциям

$$K_\lambda(z)=\sum_{n=0}^{\infty}{}' \frac{(-1)^n}{n!\,(\lambda+1)\ldots(\lambda+n)}\left(\frac{z}{2}\right)^{2n}, \quad \lambda \neq -1, -2, \ldots,$$

являющимся решениями линейного однородного дифференциального уравнения второго порядка и отличающимся только множителем $\frac{1}{\Gamma(\lambda+1)}\left(\frac{z}{2}\right)^{\lambda}$ от функций Бесселя с соответствующим индексом λ. Он доказал, что если λ — рациональное число, отличное от отрицательных целых и половины нечетного числа, то для любого алгебраического значения $\alpha \neq 0$ числа $K_\lambda(\alpha)$ и $K'_\lambda(\alpha)$ алгебраически независимы, а также обобщение этого предложения на случай совокупности функций $K_\lambda(z)$ с различными значениями параметра λ и различными значениями аргумента α.

В 1949 г. К. Зигель [2] изложил свой метод в форме общей теоремы об алгебраической независимости значений совокупности E-функций, составляющих решение системы линейных однородных дифференциальных уравнений первого порядка. Из этой теоремы следует теорема Линдемана, а при помощи дополнительных предложений — упомянутые выше результаты Зигеля о значениях функций Бесселя. Каких-либо новых результатов эта работа Зигеля не содержала, так как общая теорема сводит арифметическую проблему доказательства алгебраической независимости значений совокупности функций к проверке некоторого аналитического условия нормальности различных произведений степеней этих функций, а последняя весьма затруднительна и до сих пор удается только для совокупности E-функций, каждая из которых является решением линейного дифференциального уравнения порядка не выше второго. Поэтому общая теорема Зигеля имела мало приложений к конкретным функциям.

В 1953 г. автор [3] установил теорему, аналогичную теореме Зигеля, но при менее стеснительных предположениях и распространил ее на случай совокупности функций, удовлетворяющих системе линейных неоднородных дифференциальных уравнений. При помощи этой теоремы удалось установить трансцендентность и алгебраическую независимость значений некоторых E-функций, являющихся решениями линейных однородных и неоднородных дифференциальных уравнений не только 1-го и 2-го порядков, но 3-го и 4-го порядков.

В 1954 г. [4] путем обобщения метода Зигеля удалось найти необходимое и достаточное условие, при котором верна подобная теорема.

Первая основная теорема. *Пусть совокупность E-функций $f_1(z), \ldots, f_m(z)$, $m \geqslant 1$, является решением системы из m линейных дифференциальных уравнений первого порядка*

$$y'_k = Q_{k,0} + \sum_{i=1}^{m} Q_{k,i}\, y_i, \qquad k = 1, \ldots, m, \tag{1}$$

все коэффициенты которых $Q_{k,i}$ — рациональные функции от z, а α — любое алгебраическое число, отличное от нуля и полюсов всех функций $Q_{k,i}$. Тогда, для того чтобы m чисел $f_1(\alpha), \ldots, f_m(\alpha)$ были алгебраически независимы, необходимо и достаточно, чтобы функции $f_1(z), \ldots, f_m(z)$ были алгебраически независимы над полем рациональных функций.

Эту теорему легко переформулировать для случая E-функции $f(z)$, являющейся решением линейного дифференциального уравнения

$$P_m y^{(m)} + \ldots + P_1 y' + P_0 y = Q, \qquad (2)$$

и ряда ее последовательных производных.

Теорема 1. *Пусть E-функция $f(z)$ является решением линейного дифференциального уравнения m-го порядка (2), все коэффициенты которого — многочлены от z, а α — любое алгебраическое число, отличное от нуля и нулей многочлена P_m. Тогда, для того чтобы m чисел $f(\alpha), f'(\alpha), \ldots, f^{(m-1)}(\alpha)$ были алгебраически независимы, необходимо и достаточно, чтобы функции $f(z), f'(z), \ldots, f^{(m-1)}(z)$ были алгебраически независимы над полем рациональных функций.*

Из теоремы 1 при $m = 1$ следует

Теорема 2. *Пусть E-функция $f(z)$ трансцендентна и является решением линейного дифференциального уравнения первого порядка*

$$P_1 y' + P_0 y = Q, \qquad (3)$$

где P_1, P_0, Q — многочлены от z, а α — любое алгебраическое число, $\alpha P_1(\alpha) \neq 0$. Тогда $f(\alpha)$ трансцендентно.

В частности, показательная функция e^z удовлетворяет уравнению $y' = y$. Поэтому число e^α трансцендентно при любом алгебраическом значении $\alpha \neq 0$. Далее функция

$$\varphi_\lambda(z) = \sum_{n=0}^{\infty} \frac{z^n}{(\lambda+1)\ldots(\lambda+n)}, \qquad \lambda \neq -1, -2, \ldots, \qquad (4)$$

является при рациональном λ E-функцией и удовлетворяет уравнению

$$y' + \left(\frac{\lambda}{z} - 1\right) y = \frac{\lambda}{z}.$$

Значит, число $\varphi_\lambda(\alpha)$ трансцендентно при любом алгебраическом значении $\alpha \neq 0$ и любом рациональном λ.

Из первой основной теоремы сразу следует теорема Линдемана, которая обобщается на случай произвольной трансцендентной E-функции, удовлетворяющей уравнению (3), в частности функции (4) (см. [5]). Доказательство всех результатов Зигеля относительно функций $K_\lambda(z)$ при помощи этой теоремы также существенно упрощается. Если рассмотреть «неполную» гамма-функцию

$$F(z) = \int_0^z t^{p-1} e^{-t}\, dt$$

при любом дробном рациональном $p > 0$, то легко доказать, что числа $F(\alpha)$ и e^α алгебраически независимы при любом алгебраическом $\alpha \neq 0$.

Применение первой основной теоремы к конкретным функциям не представляет тех трудностей, как в случае Зигеля. При ее помощи можно устанавливать трансцендентность и алгебраическую независимость значений E-функций, являющихся решением линейных дифференциальных уравнений любых порядков. Например, если положить

$$\omega_k(z) = 1 + \sum_{n=1}^{\infty} \frac{z^n}{n!\, n^k}, \qquad k = 0, 1, \ldots, m,$$

$$\psi_k(z) = \sum_{n=0}^{\infty} \frac{z^{kn}}{(n!)^k}, \qquad k \geqslant 1,$$

то легко устанавливаются следующие утверждения: 1) при любом алгебраическом значении $\alpha \neq 0$ $m+1$ чисел $\omega_k(\alpha)$, $k = 0, 1, \ldots, m$, алгебраически независимы; 2) при любом $k \geqslant 1$ и любом алгебраическом значении $\alpha \neq 0$ числа $\psi_k(\alpha), \psi_k'(\alpha), \ldots, \psi_k^{k-1}(\alpha)$ алгебраически независимы.

Заметим, что функции $\omega_k(\alpha)$ и $\psi_k(\alpha)$ являются решениями линейных дифференциальных уравнений порядков $k+1$ и k.

В 1955 г. [6, 7] получен ряд теорем о трансцендентности и алгебраической независимости значений совокупности E-функций при наличии между ними одного алгебраического уравнения в поле рациональных функций, а в 1956 г. [8, 9] — общие теоремы о трансцендентности и алгебраической независимости значений в алгебраических точках у совокупности E-функций, связанных любым числом алгебраических уравнений в поле рациональных функций. Подробные доказательства последних результатов опубликованы в 1962 г. [11].

Вторая основная теорема. *Пусть E-функции $f_1(z), \ldots, f_m(z)$ и число α удовлетворяют всем условиям первой основной теоремы, а l — какое-либо целое число, $0 \leqslant l \leqslant m$. Тогда для того, чтобы среди чисел $f_1(\alpha), \ldots, f_m(\alpha)$ наибольшее число алгебраически независимых было равно l, необходимо и достаточно, чтобы среди функций $f_1(z), \ldots, f_m(z)$ наибольшее число алгебраически независимых над полем рациональных функций было равно l.*

Эту теорему также легко переформулировать для случая E-функции $f(z)$, являющейся решением дифференциального уравнения (2), и ряда ее последовательных производных.

При помощи этой теоремы доказываются следующие утверждения: 1) если E-функция $f(z)$ трансцендентна и удовлетворяет линей-

ному дифференциальному уравнению с полиномиальными коэффициентами, то она принимает трансцендентное значение в любой алгебраической точке, за исключением конечного числа таких точек; 2) если E-функции $f_1(z), \ldots, f_m(z)$ удовлетворяют условиям первой основной теоремы и функции $f_1(z), \ldots, f_l(z)$, $1 \leqslant l \leqslant m$, алгебраически независимы над полем рациональных функций, то при любом алгебраическом значении α, за исключением конечного числа таких значений, числа $f_1(\alpha), \ldots, f_l(\alpha)$ алгебраически независимы.

Исключительные алгебраические точки, в которых не имеет места трансцендентность или алгебраическая независимость значений функций в этих утверждениях, определяются из алгебраических уравнений, связывающих рассматриваемые функции в поле рациональных функций. Если дана конкретная совокупность функций, то, отыскав алгебраические уравнения, связывающие эти функции, можно найти все исключительные точки, в которых не выполняются сформулированные утверждения.

Устанавливаются и некоторые общие теоремы с аналогичными формулировками, в которых налагаются некоторые ограничения на алгебраические уравнения, связывающие рассматриваемые функции в поле рациональных функций, но зато утверждения их имеют место в точно оговоренных алгебраических точках.

При помощи всех указанных общих теорем можно устанавливать трансцендентность и алгебраическую независимость значений в алгебраических точках у конкретных E-функций, являющихся решениями линейных дифференциальных уравнений любых порядков и связанных любым числом алгебраических уравнений в поле рациональных функций. Несколько таких приложений к конкретным функциям опубликовано в [4, 10, 12, 13, 17].

Простейшим примером применения последних теорем являются функции $y_1 = \sin z$, $y_2 = \cos z$, связанные уравнением $y_1^2 + y_2^2 = 1$ и удовлетворяющие системе $y_1' = y_2$, $y_2' = -y_1$.

Если обозначить

$$\operatorname{si} z = \int_0^z \frac{\sin t}{t}\,dt, \qquad \operatorname{ci} z = \int_0^z \frac{1-\cos t}{t}\,dt,$$

то при любом алгебраическом значении $z \neq 0$ как числа $\operatorname{si} z$ и $\sin z$, так и числа $\operatorname{ci} z$ и $\cos z$ алгебраически независимы.

Значения каждого из «неполных» интегралов вероятностей и Френеля

$$\int_0^z e^{-t^2}\,dt, \qquad \int_0^z \sin t^2\,dt, \qquad \int_0^z \cos t^2\,dt$$

в любой алгебраической точке $z \neq 0$ алгебраически независимы со значением соответствующей из функций e^{z^2}, $\sin z^2$ и $\cos z^2$.

В приложениях теории трансцендентных чисел большее значение имеют количественные характеристики трансцендентности или алгебраической независимости чисел в виде неравенств, оценивающих снизу так называемую меру трансцендентности или меру взаимной трансцендентности чисел. Метод Зигеля и его обобщения позволяют получать подобные оценки. Так, Зигель получил оценку меры взаимной трансцендентности для функции Бесселя $J_0(z)$ и ее производной. Используя работу Зигеля [1] и первую основную теорему, нетрудно получить общую теорему об оценке меры взаимной трансцендентности для любой совокупности E-функций, алгебраически независимых над полем рациональных функций. Такая оценка была получена в 1962 г. С. Ленгом [14].

Указанные выше общие теоремы сводят проблемы отсутствия или наличия алгебраических связей между значениями рассматриваемых функций к отсутствию или наличию алгебраических связей между соответствующими функциями в поле рациональных функций. Поэтому важной задачей является разработка методов, позволяющих устанавливать алгебраическую независимость заданных функций, или отыскание всех алгебраических уравнений, связывающих их в поле рациональных функций. Общего метода, позволяющего решать эти проблемы для любых целых функций, удовлетворяющих линейным дифференциальным уравнениям, пока не имеется. В работах [3, 10, 12, 13] обобщаются ранее известные частные методы, применимые к некоторым классам функций, удовлетворяющих линейным дифференциальным уравнениям любых порядков. В последнее время В. А. Олейников [15, 16] разработал метод, который позволяет решать указанную задачу для решений линейных дифференциальных уравнений третьего порядка, и применил его к доказательству алгебраической независимости значений некоторых конкретных E-функций.

Недавно основную лемму метода удалось усилить и тем самым получить новые арифметические факты о значениях совокупности E-функций с коэффициентами из поля K, где K — поле рациональных чисел или мнимое квадратичное поле. Установлены критерии иррациональности, непринадлежности к полю K, отсутствия однородных и неоднородных алгебраических уравнений степени, не превосходящей заданного числа с коэффициентами из K. Первая основная теорема (для случая K) становится предельным случаем последних результатов.

Указанные в докладе общие теоремы позволяют устанавливать трансцендентность и алгебраическую независимость значений E-функций, удовлетворяющих линейным дифференциальным уравнениям, с коэффициентами из поля рациональных функций. Есте-

ственно встает вопрос об обобщении рассмотренного метода в таком направлении, чтобы его можно было применять к какому-либо более широкому классу функций, чем E-функции. Но эта задача является, по-видимому, очень трудной и требует привлечения каких-то новых идей. Заметим, что даже для такой простой функции, как функция (4), ничего не известно об арифметической природе значений в алгебраических точках для случая, когда λ иррационально.

Интересно было бы выяснить структуру класса E-функций, удовлетворяющих линейным дифференциальным уравнениям. Зигель высказал предположение, что всякая такая E-функция представляется в виде многочлена от z и конечного числа так называемых гипергеометрических E-функций и функций, получающихся из последних заменой z на λz при алгебраическом λ. Но эта проблема до сих пор не решена.

Московский университет,
Москва, СССР

ЛИТЕРАТУРА

[1] Siegel C., Über einige Anwendungen Diophantische Approximationen, *Abh. Preuss. Acad. Wiss.*, **81** (1929-30), 1-70.
[2] Siegel C., Transcendental numbers, Princeton, 1949.
[3] Шидловский А. Б., О трансцендентности и алгебраической независимости значений целых функций некоторых классов, *ДАН СССР*, **96**, № 4 (1954), 697-700; *Ученые записки МГУ*, вып. 186, математика, т. IX (1959), 11-70.
[4] Шидловский А. Б., О критерии алгебраической независимости значений одного класса целых функций, *ДАН СССР*, **100**, № 2 (1955), 221-224; *Изв. АН СССР*, сер. матем., **23**, № 1 (1959), 35-66.
[5] Шидловский А. Б., Об одном обобщении теоремы Линдемана, *ДАН СССР*, **138**, № 6 (1961), 1301-1304.
[6] Шидловский А. Б., О трансцендентных числах некоторых классов, *ДАН СССР*, **103**, № 6 (1955), 977-980.
[7] Шидловский А. Б., О трансцендентности и алгебраической независимости значений E-функций, связанных алгебраическим уравнением в поле рациональных функций, *Вестник МГУ*, № 5 (1960), 19-28.
[8] Шидловский А. Б., О трансцендентности значений одного класса целых функций, удовлетворяющих линейным дифференциальным уравнениям, *ДАН СССР*, **105**, № 1 (1955), 35-37.
[9] Шидловский А. Б., О новом критерии трансцендентности и алгебраической независимости значений одного класса целых функций, *ДАН СССР*, **106**, № 3 (1956), 399-400.
[10] Шидловский А. Б., Об алгебраической независимости трансцендентных чисел одного класса, *ДАН СССР*, **108**, № 3 (1956), 400-403.
[11] Шидловский А. Б., О трансцендентности и алгебраической независимости значений E-функций, связанных любым числом алгебраических уравнений в поле рациональных функций, *Изв. АН СССР*, **26**, № 6 (1962), 887-910.

[12] Шидловский А. Б., О трансцендентности и алгебраической независимости значений некоторых функций, *Труды Моск. матем. об-ва*, **8** (1959), 283-320.
[13] Шидловский А. Б., О трансцендентности и алгебраической независимости значений некоторых E-функций, *Вестник МГУ*, № 5 (1961), 44-59.
[14] Lang S., A transcendence measure for E-functions, *Mathematika*, **9** (1962), 157-161.
[15] Олейников В. А., О трансцендентности и алгебраической независимости значений E-функций, являющихся решением линейного дифференциального уравнения третьего порядка, *ДАН СССР*, **166**, № 3 (1966), 540-543.
[16] Олейников В. А., Об алгебраической независимости значений E-функций, удовлетворяющих линейным неоднородным дифференциальным уравнениям третьего порядка, *ДАН СССР*, **169**, № 1 (1966), 32-34.
[17] Шидловский А. Б., О трансцендентности и алгебраической независимости значений E-функций, удовлетворяющих линейным неоднородным дифференциальным уравнениям второго порядка, *ДАН СССР*, **169**, № 1 (1966), 42-45.

4

Классический анализ
Classical analysis
Analyse classique
Klassische Analysis

THE CONSTRUCTIVIZATION OF ABSTRACT MATHEMATICAL ANALYSIS

ERRETT BISHOP

The constructive point of view is that all mathematics should have numerical meaning. In other words, every mathematical theorem should admit an ultimate interpretation to the effect that certain finite computations within the set of positive integers will give certain results. In contrast, classical (that is, contemporary) mathematics is idealistic: there is no requirement that theorems and their proofs have a numerical meaning, or any predictive content whatever. For instance, the theorem that the real numbers can be well-ordered is evocative (or idealistic), rather than descriptive (or constructive). So is the theorem that a bounded monotone sequence of real numbers converges.

Brouwer has shown that the idealizations involved in classical mathematics, can, in most instances, be traced to the use of a certain logical principle—the principle of the excluded middle. It is perhaps more natural to trace them to a closely related principle—the principle of omniscience—which states that an arbitrary set A either has an element with a given property P or it does not. In case A is an infinite set this principle is not constructively valid, because the examination of each element of A to see whether one of them has property P is not something that can necessarily be done by a finite, routine process.

The constructivist replaces such transcendent logical principles as the principle of omniscience by common sense. The common sense, or operational, meanings of the standard mathematical quantifiers and connectives have been established by Brouwer. Brouwer undertook to develop mathematics along constructive lines. His development, which was not systematic, was impeded by a revolutionary, semi-mystical theory of the continuum. This theory, which in retrospect seems so unnecessary, was repellent to most mathematicians.

In addition to Brouwer, others have espoused more or less constructive points of view, usually less. There are the formalizers of constructivity, whose formal systems have little relevance to the constructivization of existing mathematics; the recursive-function theorists, who base constructivity on an ad hoc assumption which is more of an impediment than a tool; Hilbert, who believed the price of a constructive mathematics was too great; and various other groups, none of which is content to let constructive mathematics follow its natural course of development. This paper describes an attempt to redevelop abstract analysis along straightforward constructive lines, letting the material speak for itself, rather than forcing it to support a burden of philosophical preconceptions.

A real number x is defined by a sequence of rational numbers, the n^{th} term of which approximates x to within n^{-1}. More precisely, a real number is a sequence $\{x_n\}$ of rational numbers such that $|x_m - x_n| \leqslant m^{-1} + n^{-1}$ for all m and n. The theory of the real number system and the constructive development of calculus proceed along known lines. The same is true of elementary complex analysis, through at least the Riemann mapping theorem.

To go further we must face the question "What is a set?" A set is certainly *not* a preexistent object. A set exists only when it has been constructed. In other words, the question is better posed as "What must we do to define a set?" The answer is that to define a set we must (a) specify the process for constructing an element of the set, and (b) specify the process for proving that two elements of the set are equal. In the same vein, a function f from a set A to a set B is a rule which to each element a of A associates an element $f(a)$ of B, by a finite, routine process, in such a way that $f(a_1) = f(a_2)$ whenever $a_1 = a_2$. With these preliminaries, the proper constructive definitions for the various operations with sets and functions become clear, with the exception of set complementation. We do not wish to define $x \in -A$ to mean that the assumption $x \in A$ leads to a contradicition (in contrast to the approach of Brouwer). Complementation defined in terms of negation is too elusive. In addition, it leads to a loss of meaning. Therefore whenever a notion of set complementation is needed we introduce it affirmatively. (The same is true for inequality relations.) One way of doing this is through the very flexible notion of a complemented set. A *complemented set* relative to a family $\mathcal{F}$ of real-valued functions on a given set X, is an ordered pair (A, B) of subsets of X, such that for each a in A and b in B there exists f in $\mathcal{F}$ with $f(a) \neq f(b)$. The *union* of a family $\{(A_t, B_t)\}_{t \in T}$ of complemented sets is the complemented set (A, B), where $A \equiv \bigcup A_t$ and $B \equiv \bigcap B_t$. Intersections are defined similarly. An element a of X *belongs* to a complemented set (A, B) if x belongs to A. The *complement* of (A, B) is (B, A).

The constructive development of the theory of metric spaces is based on some fundamental definitions of Brouwer. A subset A of a metric space X is *located* if the distance from every point of X to A exists. A metric space is *compact* if it is complete and totally bounded. (Sequential or covering-property definitions of compactness have no value: for instance, there is no constructive proof that the closed unit interval has the Heine-Borel property.) A locally compact metric space is one in which every bounded set is contained in a compact set. With these definitions one can prove the standard results, such as Ascoli's theorem, the Baire category theorem, the Stone-Weierstrass theorem, and the Tietze extension theorem. For some of these results additional hypotheses are necessary. In the Tietze extension theorem, for instance, the set Y from which the function is extended must be assumed to be locally compact.

It is not true that a closed subset of a compact space is compact. Good constructive substitutes for this result can however be found. The following serves for most purposes: if f is a continuous (i.e., uniformly continuous) real-valued function on a compact metric space then the set $\{x: f(x) \leqslant a\}$ is compact for all except countably many real numbers a.

The core of measure theory is the Riesz representation theorem, which we examine in the following context. Let $C(X)$ be the set of all continuous real-valued functions on a compact metric space X (more generally, X could be locally compact). A *measure* μ on X is a bounded linear functional on $C(X)$, whose value at f is written $\int f\, d\mu$. We wish to show how μ can be used to assign a measure $\mu(A)$ to certain complemented sets A, to investigate the properties of the resulting measure function $A \to \mu(A)$, and show that $\int f\, d\mu$ can be interpreted as an integral with respect to the measure function. All this can be done, in case μ is a positive measure. In contrast to the classical theory, not every Borel set is measurable. However there do exist lots of measurable sets—in fact "most" compact sets are measurable, and every measurable set can be approximated from within by a compact measurable set of nearly the same measure. The measurable sets form an algebra, but not a σ-algebra in the classical sense. Every positive measure on $[0, 1]$ arises from a monotone function defined at all except countably many points. Every measure (not necessarily positive) on $[0,1]$ arises from a function of bounded variation defined almost everywhere with respect to Lebesgue measure.

The properties of the measure function constructed in connection with a measure on a locally compact space X can be abstracted, leading to the notion of a measure space. The theory of integration for functions on a measure space proceeds along classical lines. In particular one gets the Lebesgue monotone and dominated convergence theorems, Fubini's theorem, and the existence and completeness of the L_p spaces.

The circle of ideas containing the Radon-Nikodym theorem, the martingale theorem, the ergodic theorem, and Lebesgue's theorem on derivation of functions of bounded variation is fascinating from the constructive point of view. None of these theorems is constructively valid, but constructive substitutes can be found for them all; in fact there are many possibilities. For instance, by keeping track of the uses of the principle of omniscience that occur in the classical proof of the Radon-Nikodym theorem, we get a constructive substitute which states that if a certain bounded monotone sequence of real numbers (depending on the particular measures involved) converges, then the Radon-Nikodym theorem holds. Such an approach can be used to constructivize many classical results, but it has a very limited appeal. To constructivize the martingale theorem and Lebesgue's theorem, two additional approaches are possible. The first approach is based on "norms" of the form

$$\|f\|_\lambda \equiv \int \lambda(|f(x)|)\, d\mu(x),$$

where λ is a nonnegative convex even function on $\mathbf{R}$. The use of these "norms," instead of the usual L_1 norm, permits a simplified treatment which can be given constructive meaning. Another approach is by means of upcrossing inequalities. A sequence $\{x_n\}$ of real numbers *upcrosses* from a real number a to a real number $b > a$ at most N times if N is an upper bound for all integers k such that there exists a finite subsequence of length $2k$ whose terms are alternately $\leqslant a$ and $\geqslant b$. Classically the sequence $\{x_n\}$ converges if and only if it is bounded and upcrosses from an arbitrary real number a to an arbitrary real number $b > a$ at most finitely many times, whereas constructively convergence is much the stronger statement. Thus it is not surprising that many classical convergence results that fail constructively find constructive substitutes in terms of upcrossing inequalities; in fact these inequalities usually constitute considerable extensions of the classical results, from the classical point of view. For instance, one can give an upcrossing inequality for ergodic theory that not only implies (within the classical system) the Chacon-Ornstein and Dunford-Schwartz ergodic theorems and various generalizations thereof, but has the merit of a relatively simple proof. From the ergodic upcrossing inequality one can derive an upcrossing inequality that affords a constructive substitute for Lebesgue's theorem, and an upcrossing inequality that generalizes Doob's upcrossing inequality for martingales.

The theory of separable normed linear spaces can be given a constructive basis. In particular such theorems as the Hahn-Banach

theorem, the separation theorem, the spectral theorem, the Krein-Milman theorem, and various theorems on the forms of linear functionals all have satisfactory constructive versions. For instance, a bounded linear functional on L_p is induced by an element of L_q (where $p^{-1} + q^{-1} = 1$) if and only if it is normable. This implies that L_q is not the complete dual of L_p because (except in trivial situations) there always exist non-normable linear functionals. In addition to constructive versions of well-known classical results, there are results that, while trivial classically, have considerable constructive content. An instance is the theorem that the normable linear functionals are dense in the dual space relative to a certain norm (which corresponds to the weak* topology).

The theory of locally compact abelian groups (provided the underlying space is metric) can be constructivized. A constructive proof of the existence of Haar measure can be extracted from work of H. Cartan. The theory of the Fourier transform and Pontryagin duality theory are developed along classical lines, the difference being that certain least upper bounds, which trivially exist classically, must be proved to exist. For instance it is not trivial to give a constructive proof that an L_1 function acting by convolution induces a normable linear operator on L_2.

The theory of a commutative Banach algebra $\mathfrak{A}$ is a rare instance of a theory whose constructive version is substantially more complicated than its classial counterpart. The reason is that the classical techniques for building maximal ideals are not constructively valid. It is therefore necessary, instead of considering an ideal $\{x_1a_1 + \ldots + x_na_n : a_1, \ldots, a_n \in \mathfrak{A}\}$ generated by elements $x_1, \ldots, x_n$, to consider a *partial ideal* $\{x_1a_1 + \ldots + x_na_n : a_1, \ldots, a_n \in A\}$, where A is some compact subset of $\mathfrak{A}$. This makes the statements and proofs more awkward, but for the applications to analysis the constructive results seem to have the same force as their classical prototypes.

It is interesting to speculate whether we can relax the separability and metrizability assumptions that occur throughout most of the above development. Surprisingly, the way to do this does not seem to be to introduce larger spaces, but rather to examine functorial properties of mappings between the spaces already introduced. An example will suffice. Let $\mathfrak{A}$ be a commutative uniformly-closed algebra of normable Hermitian operators on a Hilbert space H. In case $\mathfrak{A}$ is separable the spectrum Σ of $\mathfrak{A}$ is compact, and the spectral theorem asserts that $\mathfrak{A}$ is isomorphic to $C(\Sigma)$. In case $\mathfrak{A}$ is non-separable, on the other hand, there seems to be no way of constructing even one element of Σ, so that the spectral theorem is apparently not constructively valid unless $\mathfrak{A}$ is separable. Nevertheless, the folowing theorem shows the situation in its true light: If $\mathfrak{A}_1$ and $\mathfrak{A}_2$ are separable commutative uniformly-closed algebras of normable Hermitian operators, with $\mathfrak{A}_1 \subset \mathfrak{A}_2$, and λ is the inclusion map, then the adjoint map λ^* from the spectrum Σ_2 of $\mathfrak{A}_2$ to the spectrum Σ_1 of $\mathfrak{A}_1$ has dense range (classically, the range therefore equals Σ_1). Thus the partially ordered family $\mathscr{F}$ of all uniformly closed separable subalgebras of a given (not necessarily separable $\mathfrak{A}$), together with the inclusion maps λ, gives rise to a dual family $\mathscr{F}^*$ of compact sets, with corresponding maps λ^*. Classically it would be a simple matter to use $\mathscr{F}^*$ and the maps λ^* to extend the spectral theorem to the algebra $\mathfrak{A}$, but constructively we can go no further.

University of California,
San Diego, USA

EXTENSION THEOREMS FOR QUASICONFORMAL MAPPINGS IN n-SPACE

FREDERICK W. GEHRING[1])

1. Quasiconformal mappings

Suppose that D and D' are domains in R^n, euclidean n-space, and that f is a diffeomorphism of D onto D'. For each point $P \in D$, the differential mapping $df(P)$ carries B^n, the unit ball $\{x: |x| < 1\}$, onto an ellipsoid E with center at the origin. Let B_I denote the largest ball contained in E and B_O the smallest ball containing E. Then the functions

$$H_I(P, f) = \frac{m(E)}{m(B_I)}, \quad H_O(P, f) = \frac{m(B_0)}{m(E)}$$

provide a natural way of measuring how far the mapping f is from being conformal at the point P, while the functionals

$$K_I(f) = \sup_{P \in D} H_I(P, f), \quad K_O(f) = \sup_{P \in D} H_0(P, f)$$

measure how far f differs from being a conformal mapping of D. We call $K_I(f)$ and $K_O(f)$ the *inner* and *outer dilatations* of f, respectively[2]).

We can calculate these dilatations by means of extremal lengths. Given a family Γ of arcs in $\overline{R}^n$, the one point compactification of R^n,

[1]) This research was supported in part by the National Science Foundation, Contract *GP*-4153.

[2]) Sometimes $K_I(f)$ and $K_O(f)$ are defined as the $(n-1)$-roots of these suprema, for example in [9]. This is clearly only a matter of notation.

we define the modulus of Γ as

$$M(\Gamma) = \inf_{\rho} \int_{R^n} \rho^n \, d\omega,$$

where the infimum is taken over all functions ρ which are nonnegative and Borel measurable in R^n and which satisfy the inequality

$$\int_{\gamma} \rho \, ds \geqslant 1 \tag{1}$$

for all $\gamma \in \Gamma$ [16]; the integral in (1) is taken with respect to linear measure whenever γ is not rectifiable. Then if f is the above mentioned diffeomorphism, it is easy to verify that

$$K_I(f) = \sup_{\Gamma} \frac{M(f(\Gamma))}{M(\Gamma)}, \qquad K_O(f) = \sup_{\Gamma} \frac{M(\Gamma)}{M(f(\Gamma))}, \tag{2}$$

where the suprema are taken over all families Γ of arcs in D with $M(\Gamma) \neq 0, \infty$.

Suppose next that D and D' are domains in $\overline{R}^n$ and that f is a homeomorphism of D onto D'. Then we may use (2) to *define* the inner and outer dilatations of f[1]). One can show that

$$K_I(f) \leqslant K_O(f)^{n-1}, \qquad K_O(f) \leqslant K_I(f)^{n-1},$$

and hence the dilatations $K_I(f)$ and $K_O(f)$ are both finite or infinite. In the former case, f is said to be *quasiconformal*; f is said to be K-quasiconformal if $K_I(f) \leqslant K$ and $K_O(f) \leqslant K$ where $1 \leqslant K < \infty$, [4], [9], [16].

It is important to identify the 1-quasiconformal mappings. When $n = 2$, f is 1-quasiconformal if and only if f or its complex conjugate is a univalent meromorphic function. When $n \geqslant 3$, f is 1-quasiconformal if and only if f is a Möbius transformation, [6] and [15].

2. Main results

This talk is concerned with the first of the following two basic problems.

Problem 1. *Characterize the domains $D \subset \overline{R}^n$ which can be mapped quasiconformally onto the unit ball B^n.*

Problem 2. *Given such a domain D, determine the inner and outer coefficients of D,*

$$K_I(D) = \inf_{f} K_I(f), \qquad K_O(D) = \inf_{f} K_O(f),$$

[1]) Sometimes $K_I(f)$ and $K_O(f)$ are defined as the $(n-1)$-roots of these suprema, for example in [9]. This is clearly only a matter of notation.

where the infima are taken over all quasiconformal mappings f of D onto B^n.

It is easy to give a complete solution for each of these problems when $n = 2$. For a domain $D \subset \overline{R}^2$ is quasiconformally equivalent to the unit disk B^2 if and only if its boundary ∂D is a nondegenerate continuum. The Riemann mapping theorem then implies that $K_I(D) = K_O(D) = 1$ for each such D.

The situation is more complicated when $n \geqslant 3$. For example, in the first problem, it is not possible to decide whether or not a given domain $D \subset \overline{R}^n$ is quasiconformally equivalent to B^n by looking only at ∂D. For consider the following pair of domains in R^n,

$$D_1 = \{x: x_n > \min(r^{1/2}, 1)\}, \qquad D_2 = \{x: x_n < \min(r^{1/2}, 1)\},$$

where $r = (x_1^2 + \ldots + x_{n-1}^2)^{1/2}$. Then D_1 and D_2 are Jordan domains with a common boundary, a trivial modification of the proof of Theorem 10.3 of [9] yields a quasiconformal mapping of D_1 onto B^n, while the n-dimensional analogue of Theorem 10.1 of [9] implies that D_2 is not quasiconformally equivalent to B^n.

Next because the functionals $K_I(f)$ and $K_O(f)$ are lower semicontinuous with respect to uniform convergence on compact sets, for each domain $D \subset \overline{R}^n$ quasiconformally equivalent to B^n there exist extremal mappings f_I and f_O of D onto B^n with

$$K_I(D) = K_I(f_I), \qquad K_O(D) = K_O(f_O).$$

Hence in the second problem, the coefficients are greater than 1 except when D is a ball or half space.

In [9], J. Väisälä and I considered Problem 2, with $n = 3$, and calculated the inner coefficient for a convex dihedral wedge and the outer coefficients for an infinite circular cylinder and for an infinite circular convex cone. In the present paper, I want to report on the following pair of extension theorems, each of which reduces the global mapping problem for a domain D of Problem 1 to a local mapping problem for the part of D near ∂D [8].

Theorem 1. *Suppose that D is a domain in $\overline{R}^n$, that U is a neighborhood of ∂D, and that f is a quasiconformal mapping of $D \cap U$ into B^n such that $|f(x)| \to 1$ as $x \to \partial D$ in $D \cap U$. Then there exists a neighborhood U^* of ∂D and a quasiconformal mapping f^* of D onto B^n such that $f^* = f$ in $D \cap U^*$.*

Theorem 2. *Suppose that D is a Jordan domain in $\overline{R}^3$ and that for each point $P \in \partial D$ there exists a neighborhood U_P of P and*

a homeomorphism f_P of $\overline{D} \cap U_P$ into $\overline{B}^3$ such that f_P is quasiconformal in $D \cap U_P$ and $f_P(\partial D \cap U_P) \subset \partial B^3$. Then for each point $Q \in \partial D$ there exists a neighborhood U^* of Q and a homeomorphism f^* of $\overline{D}$ onto $\overline{B}^3$ such that f^* is quasiconformal in D and $f^* = f_Q$ in $\overline{D} \cap U^*$

Thus roughly speaking, Theorem 1 implies that a domain $D \subset \overline{R}^n$ can be mapped quasiconformally onto B^n if and only if ∂D has a neighborhood U such that $D \cap U$ can be mapped quasiconformally into B^n with ∂D corresponding to ∂B^n, while Theorem 2 implies that a Jordan domain $D \subset \overline{R}^3$ can be mapped quasiconformally onto B^3 if and only if each point $P \in \partial D$ has a neighborhood U such that $D \cap U$ can be mapped quasiconformally into B^3 with $\partial D \cap U$ corresponding to a subset of ∂B^3.

Theorems 1 and 2 establish conjectures made independently by B. V. Shabat and by K. I. Virtanen.

3. Remarks

Theorems 1 and 2 are closely related to the following well known problem of topology.

Schoenflies problem. *Suppose that D is a domain in $\overline{R}^n$ and that f is a homeomorphism of ∂D onto ∂B^n. Under what circumstances can f be extended as a homeomorphism f^* of $\overline{D}$ onto $\overline{B}^n$?*

It is well known that when $n=2$, each such f has an extension f^*. On the other hand, when $n \geqslant 3$, there exist homeomorphisms f which have no such extensions f^*. The Alexander horned sphere yields an example in $\overline{R}^3$. Alternatively, by modifying a construction of R. Fox and E. Artin [5], one can obtain the following example.

Theorem 3. *There exists a domain $D \subset \overline{R}^3$, not homeomorphic to B^3, and a homeomorphism f of the complement $C(D)$ onto $\overline{B}^3$ such that*

$$\frac{1}{K}|x-y| \leqslant |f(x)-f(y)| \leqslant K|x-y| \tag{3}$$

for all $x, y \in C(D)$, where K is a constant, $1 < K < \infty$.

In 1959, B. Mazur [12] proved that the homeomorphism f, in the Schoenflies problem, has an extension f^* provided that for some neighborhood U of ∂D, f can be extended as a homeomorphism g^* of $\overline{D} \cap U$ into $\overline{B}^n$ so that g^* is linear in a neighborhood of a point

of $D \cap U$. In 1960, M. Morse [14] showed how to eliminate the linearity hypothesis on g^*, while M. Brown [2] gave an independent proof of the resulting theorem.

The proof for Theorem 1 is based upon explicit versions of the arguments of Mazur and Morse. Theorem 2 is established by means of a general sewing theorem which allows one to fit together certain pairs of quasiconformal mappings in $\overline{R}^3$ and hence eventually reduce the problem to one which Theorem 1 can handle. This sewing theorem is proved, in turn, by combining an argument due to M. Brown [3] with some extension theorems for plane quasiconformal mappings due to L. V. Ahlfors [1] and to O. Lehto and K. I. Virtanen [11].

4. Geometric conditions

Suppose that D is a domain in $\overline{R}^n$ and that f is a diffeomorphism of ∂D onto ∂B^n. Then f can be extended as a homeomorphism f^* of $\overline{D}$ onto $\overline{B}^n$ which is diffeomorphic in $\overline{D} - \{P\}$, where P is a prescribed point of D [10]. It is also well known that f may not have an extension f^* which is diffeomorphic in $\overline{D}$ when $n \geqslant 7$ [13]. On the other hand, from Theorem 1 it follows that f will always have an extension which is quasiconformal in D. Thus a domain $D \subset \overline{R}^n$ with ∂D diffeomorphic to ∂B^n is quasiconformally equivalent to B^n.

This sufficient condition is far from being necessary, since for example, it is easy to construct a domain $D \subset \overline{R}^3$ which is quasiconformally equivalent to B^3 and for which the set of points of ∂D, not accessible from D, has positive 3-dimensional measure [9]. Thus it is natural to seek a condition on ∂D, weaker than that of being diffeomorphic to ∂B^n, which will guarantee that D is quasiconformally equivalent to B^n. In particular, M. A. Lavrentieff has asked if the image of ∂B^3 under a Lipschitz homeomorphism bounds a pair of domains in $\overline{R}^3$ which can be mapped quasiconformally onto B^3. Theorem 3 shows that this is not the case.

By means of extremal lengths, one can find a simple geometric condition on $C(D)$ which is satisfied whenever a domain $D \subset \overline{R}^n$ is quasiconformally equivalent to B^n, $n \geqslant 3$. A closed set $E \subset \overline{R}^n$ is said to be *strongly locally connected* if there exists a constant c, $1 < c < \infty$, with the following property. For each point $P \neq \infty$ and each number r, $0 < r < \infty$, each pair of points in $E \cap \{x: |x-P| < r\}$ can be joined by a continuum in $E \cap \{x: |x-P| < cr\}$ and each pair of points in $E \cap \{x: |x-P| > r\}$ can be joined by a continuum in $E \cap \left\{x: |x-P| > \frac{r}{c}\right\}$. The n-dimensional form of

Lemma 1 of [7] implies that if $n \geqslant 3$ and if $D \subset \bar{R}^n$ is quasiconformally equivalent to B^n, then $C(D)$ is strongly locally connected. Unfortunately, Theorem 3 also shows that when $n = 3$, this necessary condition is not sufficient, since the image of $\bar{B}^3$ under a homeomorphism f satisfying (3) is easily seen to be strongly locally connected.

University of Michigan,
Ann Arbor, Michigan, USA

REFERENCES

[1] Ahlfors L. V., Extension of quasiconformal mappings from two to three dimensions, *Proc. Nat. Acad. Sci. USA*, **51** (1964), 768-771.

[2] Brown M., A proof of the generalized Schoenflies theorem, *Bull. Amer. Math. Soc.*, **66** (1960), 74-76.

[3] Brown M., Locally flat imbeddings of topological manifolds, *Ann. Math.*, **75** (1962), 331-341.

[4] Han-Lin Chen, Quasiconformal mappings in n-dimensional space, *Acta Math. Sinica*, **14** (1964), 93-102 (Chinese); translated in *Chinese Math. Acta*, **5** (1964), 101-111.

[5] Fox R. H., Artin E., Some wild cells and spheres in three-dimensional space, *Ann. Math.*, **49** (1948), 979-990.

[6] Gehring F. W., Rings and quasiconformal mappings in space, *Trans. Amer. Math. Soc.*, **103** (1962), 353-393.

[7] Gehring F. W., Extension of quasiconformal mappings in three-space, *J. d'Analyse Math.*, **14** (1965), 171-182.

[8] Gehring F. W., Extension theorems for quasiconformal mappings in n-space, *J. d'Analyse Math.*, **19** (1967), 149-169.

[9] Gehring F. W., Vaisala J., The coefficients of quasiconformality of domains in space, *Acta Math.*, **114** (1965), 1-70. Русский перевод: сб. *Математика*, **10**, № 6 (1966).

[10] Huebsch W., Morse M., An explicit solution of the Schoenflies extension problem, *J. Math. Soc. Japan*, **12** (1960), 271-289.

[11] Lehto O., Virtanen K. I., Quasikonforme Abbildungen, Springer-Verlag, Berlin-Heidelberg-New York, 1965.

[12] Mazur B., On embeddings of spheres, *Bull. Amer. Math. Soc.*, **65** (1959), 59-65.

[13] Milnor J., On manifolds homeomorphic to the 7-sphere, *Ann. Math.*, **64** (1956), 399-405.

[14] Morse M., A reduction of the Schoenflies extension problem, *Bull. Amer. Math. Soc.*, **66** (1960), 113-115.

[15] Решетняк Ю. Г., О конформных отображениях пространства, *ДАН СССР*, **130** (1960), 1196-1198.

[16] Väisälä J., On quasiconformal mappings in space, *Ann. Acad. Sci. Fenn.*, **298** (1961), 1-36.

QUASICONFORMAL MAPPINGS IN THE PLANE

OLLI LEHTO

During the past decade decisive progress has been made in the general theory of quasiconformal mappings in the plane. For many years the study of the relations between the various possible definitions for quasiconformality was one of the principal objects of research. Today, this part of the theory seems to be a fairly closed chapter. Therefore, the beginning of this survey, which deals with the definitions, is more of an historical nature. In the second part, attention is called to the important work of Ahlfors on quasiconformal reflection and to some new problems which have arisen in this connection. The concluding section contains some remarks on the parametric representation of quasiconformal mappings.

1. Definitions

The first quasiconformal mappings introduced by Grötzsch and Lavrentieff can be regarded as immediate generalizations of conformal mappings. In 1938 Morrey, on studying partial differential equations of elliptic type, defined a more general class of mappings. These were characterized as weak homeomorphic solutions of a Beltrami differential equation

$$w_{\bar{z}} = k w_z, \tag{1}$$

where k is measurable and $\sup |k(z)| < 1$.

Grötzsch mappings are continuously differentiable solutions of (1) with non-zero Jacobian, and so for Grötzsch mappings k in (1) is continuous. However, not every continuous k yields a Grötzsch mapping. It is a classical result that uniform Hölder-continuity of k is a sufficient condition, and it is also well known that Hölder-continuity can be replaced by weaker integral conditions ([7]). But it seems to be difficult to characterize Grötzsch mappings in terms of k.

In contrast to this, Bojarski proved that Lavrentieff mappings are weak homeomorphic solutions of exactly those equations (1) in which k is continuous. It is a beautiful result that, apart from constants, Morrey mappings constitute the closure of Grötzsch and Lavrentieff mappings under uniform convergence on compact sets.

In the early fifties, Ahlfors and Pfluger defined quasiconformality with the help of the conformal modulus $M(Q)$ of a quadrilateral Q:

A sense preserving homeomorphism f: $D \to D'$ is quasiconformal in D, if $M(f(Q))/M(Q)$ is bounded for all quadrilaterals $Q \subset D$. If the bound does not exceed K, f is called K-quasiconformal. It was one of the fundamental discoveries in the theory, about ten years ago, that this class of quasiconformal mappings coincides with the class of Morrey mappings ([7]).

Today many other definitions for quasiconformality are known. One way to obtain definitions is as follows: Take a conformal invariant, consider its change under a K-quasiconformal mapping, and study whether a homeomorphism with this property is K-quasiconformal. Thus quasiconformality can also be characterized in terms of the modulus of a ring domain, of the extremal length of a curve family, of the harmonic measure, of the hyperbolic measure, and of the angle. The case of an angle is not quite easy to handle, for the following reason: If a mapping is known to be quasiconformal, then for the value of K sets of two-dimensional measure zero are deletable. However, such null-sets are too large to be disregarded, if we want to decide whether a homeomorphism is quasiconformal or not. Therefore, the image of an angle must be studied at points at which the mapping is not necessarily differentiable. This difficulty has been recently overcome by Agard and Gehring, and by Taari.

For detailed results concerning various definitions, we refer to the survey [3] by Gehring, who lists 12 non-trivially different characterizations for quasiconformality and gives an extensive list of references.

2. Quasiconformal continuation

Let f be a quasiconformal mapping of a domain D, and let $F \subset D$ be a compact set. Then there always exists a quasiconformal mapping g of the plane such that $g \mid F = f$ ([7]). The extension g can be so constructed that in every component of the complement of the closure of D, g is a linear transformation ([6]).

In contrast to the above, it is not always possible to find a quasiconformal mapping g of the plane such that $g \mid D = f$. If D and $f(D)$ are bounded by a finite number of Jordan curves, such an extension is possible, if every boundary component of D and $f(D)$ is a quasiconformal curve, i.e. the image of a circle under a quasiconformal mapping of the plane ([7]). Ahlfors [1] has given a very simple characterization of quasiconformal curves in geometric terms: If $C \ni \infty$ and z_1, z_2, z_3 are any three successive finite points of C, then C is quasiconformal if and only if there exists a finite number M such that

$$\left|\frac{z_1 - z_2}{z_1 - z_3}\right| < M. \tag{1}$$

Let now C be a Jordan arc. We call C quasiconformal if it is the image of an interval under a quasiconformal mapping of the plane.

The problem of characterizing quasiconformal arcs has recently been solved by Rickman [8]. Suppose in the following that C is bounded. If C is a closed arc, the validity of the Ahlfors condition (1) is necessary and sufficient for C to be quasiconformal. If C is open and condition (1) is fulfilled, then C has two endpoints and the closure $\bar{C}$ satisfies the same condition (1) as C. Hence, again C is quasiconformal. In this case the converse is not true: there exist bounded open quasiconformal arcs which do not satisfy any global condition (1). For open arcs a geometric characterization is obtained in local terms: an open arc C is quasiconformal if and only if the condition (1) is locally valid with a uniformly bounded M.

There are still open problems concerning quasiconformal continuation, e. g., how the maximal dilatation of an extended mapping depends on the properties of the boundary of the original domain.

3. Continuous deformation

Let D be the unit disc and S_K the family of all K-quasiconformal homeomorphisms f: $D \to D$, such that $f(0) = 0$, $f(1) = 1$. Every $f \in S_K$ can be extended as a K-quasiconformal mapping to the whole plane by reflection. S_K is a metric space if the distance $\rho(f, g)$ of the mappings f, $g \in S_K$ is defined by $\rho(f, g) = \max |f(z) - g(z)|$, $z \in \bar{D}$.

Let us consider a mapping $f \in S_K$ and denote its complex dilatation by k. Let k_t be a measurable function in D such that $|k_t(z)| \leqslant (K-1)/(K+1)$ and such that $k_t(z)$ is continuous with respect to the real parameter t on the interval $I = \{t \mid 0 \leqslant t \leqslant T\}$, for every $z \in D$. Furthermore, we require that $k_0(z) = 0$, $k_T(z) = k(z)$. Let f_t be the (uniquely determined) mapping in S_K whose complex dilatation equals $k_t(z)$ for almost all $z \in D$. If $t, t' \in I$, we have for the complex dilatation $\tilde{k}$ of the mapping $f_{t'} \circ f_t^{-1}$,

$$|\tilde{k}(z)| = \left|\frac{k_t(z) - k_{t'}(z)}{1 - \overline{k_t(z)}\, k_{t'}(z)}\right| \tag{2}$$

a.e. This implies, by Teichmüller's distortion theorem, that $\rho(f_t, f_{t'}) \to 0$ as $t' \to t$. In other words, $\{f_t \mid 0 \leqslant t \leqslant T\}$ is a curve S_K from the identity mapping to the given mapping f. (For more details, see Ahlfors and Bers [2].)

There are, of course, many possibilities to construct the family $\{f_t \mid 0 \leqslant t \leqslant T\}$. For instance, we can choose k_t such that the point $k_t(z)$ moves with constant velocity along the radius from the point 0 to the point $k(z)$, as t moves from 0 to T with constant velocity. In the k-plane length can be measured either with respect to the euclidean metric or the non-euclidean metric of the unit disc. In the latter case, it follows from (2) that $\tilde{k}(z)$ depends on the difference $t' - t$ but not

on the value of t. In particular, if we take $t = T/2$, the absolute values of the complex dilatations of the mappings $f_{T/2}$ and $f \circ f_{T/2}^{-1}$ are the same. This implies the (well-known) result, that f admits a representation

$$f = f_2 \circ f_1, \quad \text{where} \quad f_1,\ f_2 \subset S_{\overline{K}}. \tag{3}$$

Bojarski has proved that for every $f \in S_K$ and for every measurable set $A \subset D$, $m(f(A)) = O(m(A)^{\delta})$, where m denotes the two-dimensional Lebesgue-measure and $\delta > 0$ ([7]). Application of the above deformation technique yields information about δ. In [5], the simple formula (3) alone was applied, while recently, Gehring and Reich [4] have made a more profound use of the parametric representation in this connection.

University of Helsinki,
Helsinki, Finland

REFERENCES

[1] Ahlfors L., Quasiconformal reflections, *Acta Math.*, **109** (1963).
[2] Ahlfors L., Bers L., Riemann's mapping theorem for variable metrics, *Ann. Math.*, **72** (1960).
[3] Gehring F. W., Definitions for a class of plane quasiconformal mappings, *Nagoya Math. J.*, **29** (1967).
[4] Gehring F. W., Reich E., Area distortion under quasiconformal mappings, *Ann. Acad. Sci. Fenn.* A I, **388** (1966).
[5] Lehto O., Remarks on the integrability of the derivatives of quasiconformal mappings, *Ann. Acad. Sci. Fenn.* A I, **371** (1965).
[6] Lehto O., An extension theorem for quasiconformal mappings, *Proc. London Math. Soc.* (3), **14** A (1965).
[7] Lehto O., Virtanen K. I., Quasikonforme Abbildungen, Springer-Verlag (1965).
[8] Rickman S., Characterization of quasiconformal arcs, *Ann. Acad. Sci. Fenn.* A I, **395** (1966).

О ВОЗМОЖНОСТИ ПРЕДСТАВЛЕНИЯ ФУНКЦИЙ СУПЕРПОЗИЦИЯМИ ФУНКЦИЙ ОТ МЕНЬШЕГО ЧИСЛА ПЕРЕМЕННЫХ

А. Г. ВИТУШКИН

С помощью преобразования Чирнгаузена общее алгебраическое уравнение n-й степени приводится к виду

$$t^n + a_4 \cdot t^{n-4} + \ldots + a_{n-1} \cdot t + 1 = 0.$$

В частности, уравнение 7-й степени приобретает вид

$$f^7 + xf^3 + yf^2 + z \cdot f + 1 = 0. \tag{1}$$

Дальнейшие попытки алгебраистов привести это уравнение к более простому виду по настоящее время остаются безуспешными. В своих «Математических проблемах» Д. Гильберт [1] по-новому подошел к этой задаче, сформулировав ее под № 13 в следующем виде: «Невозможность решения общего уравнения 7-й степени при помощи функций только двух переменных». Для этого Д. Гильберт считал нужным доказывать следующее: функция $f = f(x, y, z)$, являющаяся решением уравнения (1), не представима суперпозицией непрерывных функций двух переменных.

Подчеркнем, что задача состоит в том, чтобы или с помощью алгебраических подстановок свести решение уравнения (1) к решению алгебраических уравнений с двумя параметрами, т. е. доказать, что функция $f(x, y, z)$ является суперпозицией алгебраических функций двух переменных, или же доказать, что решение $f(x, y, z)$ уравнения (1) не является суперпозицией алгебраических функций двух переменных.

Д. Гильберт ожидал, что уравнение 7-й степени не разрешимо даже в непрерывных функциях двух переменных. Работы А. Н. Колмогорова [6, 8] и В. И. Арнольда [7] опровергают сформулированную гипотезу Д. Гильберта. Однако рассматриваемая проблема по существу остается открытой, так как остается возможность доказывать неразрешимость уравнения 7-й степени в каком-либо другом классе функций двух переменных, содержащем все алгебраические функции двух переменных.

Сформулируем теперь основные результаты, полученные в связи с 13-й проблемой Д. Гильберта.

I. Суперпозиции аналитических функций

Доказательство существования аналитических функций n переменных ($n \geqslant 2$), не представимых в виде суперпозиции аналитических функций меньшего числа переменных, может быть получено различными способами. Формулируя 13-ю проблему [1], Д. Гильберт прибавляет, что он «располагает строгим доказательством того, что существует аналитическая функция трех переменных, которая не может быть получена конечной суперпозицией функций только двух аргументов». Не указывая точно, о каких функциях двух аргументов идет речь, Д. Гильберт, по-видимому, имел в виду аналитические функции двух переменных.

Более сильные результаты в этом направлении получил в 1920 г. А. Островский [2], который доказал, в частности, что аналитическая функция двух аргументов $\zeta(x, y) = \sum_{n=1}^{\infty} x^n/n^y$ не есть конечная суперпозиция бесконечно дифференцируемых функций одного переменного и алгебраических функций любого числа переменных.

II. Проблема резольвент

Алгебраические уравнения до четвертой степени включительно разрешимы в радикалах, т. е. корни этих уравнений как функции коэффициентов представляются в виде суперпозиций арифметических операций и функций одного переменного вида $\sqrt[n]{t}$ $(n = 2, 3)$. Общее уравнение пятой степени в радикалах неразрешимо. Но поскольку общее уравнение 5-й степени алгебраическими подстановками приводится к виду $x^5 + t \cdot x + 1 = 0$, содержащему один параметр t, то мы можем сказать, что корень общего уравнения 5-й степени как функция коэффициентов также представляется в виде суперпозиции арифметических операций и алгебраических функций одного переменного.

Проблема резольвент в терминах суперпозиций может быть сформулирована следующим образом: указать для любого n такое наименьшее число k, что корень общего уравнения n-й степени как функция коэффициентов представляется в виде суперпозиции алгебраических функций k переменных. В работе [3] Д. Гильберт высказал предположение, что для n, равных 6, 7, 8, число k соответственно равно 2, 3, 4. Тем более неожиданным оказался результат Д. Гильберта 1926 г. [3], полученный для уравнения 9-й степени: корень общего уравнения 9-й степени представляется в виде суперпозиции алгебраических функций четырех переменных. А. Виман [12], обобщая результат Д. Гильберта, доказал, что при всяком $n \geqslant 9$ имеет место неравенство $k \leqslant n - 5$. Как заметил Г. Н. Чеботарев [13], тем же методом можно доказать, что для $n \leqslant 21$ $k \leqslant n - 6$, а для $n \leqslant 121$ $k \leqslant n - 7$. Проблеме резольвент был посвящен также цикл работ Н. Г. Чеботарева [14]. Однако доказательство основного результата оказалось ошибочным (см. [15]).

III. Суперпозиции гладких функций

В работе автора 1954 г. [4] было доказано, что в классе всех p раз непрерывно дифференцируемых функций n переменных существуют такие, которые не могут быть представлены в виде конечной суперпозиции функций, для которых отношение числа аргументов к числу имеющихся у них дифференциалов строго меньше, чем n/p.

Эта теорема по существу показывает, что характеристикой сложности p раз дифференцируемых функций n переменных может служить отношение n/p. Первоначальное доказательство этой теоремы использовало теорию многомерных вариаций множеств (см. [16]). А. Н. Колмогоров [5] показал, что тот же результат может быть получен, если основываться лишь на оценках числа элементов ε-сетей функциональных компактов.

Обозначим через F_p^n множество всех таких функций, заданных на n-мерном кубе, все частные производные которых до порядка p включительно непрерывны и ограничены некоторой константой. Пусть $N_\varepsilon(F_p^n)$ есть минимальное число шаров радиуса ε в пространстве всех непрерывных функций, которыми можно покрыть множество F_p^n.

Оказывается, что $\lim\limits_{\varepsilon \to 0} \log \log N_\varepsilon(F_p^n)/\log \frac{1}{\varepsilon} = \frac{n}{p}$. Отсюда следует, что если $(n/p) > (n'/p')$, то множество функций F_p^n в определенном смысле «массивнее» множества $F_{p'}^{n'}$.

Эти работы имеют интересные продолжения в работах различных авторов по оценкам сложности алгоритмов.

IV. Суперпозиции непрерывных функций

Крайне неожиданной была работа А. Н. Колмогорова 1956 г. [6], в которой доказывалось, что всякая непрерывная функция n переменных представима в виде суперпозиции непрерывных функций трех переменных. Затем В. И. Арнольдом в 1957 г. [7] в этой теореме число переменных было снижено с 3 до 2 и почти одновременно появилась теорема А. Н. Колмогрова [8] о представлении непрерывных функций n переменных в виде

$$f(x_1, \ldots, x_n) = \sum_{i=1}^{2n+1} \varphi_i\Big(\sum_{j=1}^{n} \alpha_{i,j}(x_j)\Big), \tag{2}$$

где все функции непрерывны, а внутренние функции $\alpha_{i,j}(x_j)$ заранее фиксированы, т. е. не зависят от разлагаемой функции f.

По теореме В. И. Арнольда решение уравнения 7-й степени представляется суперпозицией непрерывных функций двух переменных. Это опровергает сформулированную гипотезу Д. Гильберта [1]).

Следует отметить, однако, что участвующие в этом представлении функции заведомо не являются алгебраическими, поскольку они даже не дифференцируемы.

Теорему А. Н. Колмогорова [8] можно усилить следующим результатом Н. К. Бари [17], полученным еще в 1930 г. в связи с проблематикой рядов Фурье: всякая непрерывная функция одного

---

[1]) Говоря об этом крупном достижении, как правило, забывают, что решающий результат в опровержении гипотезы Д. Гильберта получен не В. И. Арнольдом, а А. Н. Колмогоровым. Я напоминаю об этом потому, что по кулуарным разговорам видно, что распространено неправильное представление об авторстве в данном цикле работ.

переменного $f(t)$ может быть представлена в виде

$$f(t)=f_1(\varphi_1(t))+f_2(\varphi_2(t))+f_3(\varphi_3(t)),$$

где все функции $\{f_i\}$ и $\{\varphi_i\}$ абсолютно непрерывны.

Из теоремы А. Н. Колмогорова и теоремы Н. К. Бари вытекает, что каждую непрерывную функцию n переменных можно представить в виде суперпозиции абсолютно непрерывных функций одного переменного и операции сложения.

Имеется ряд результатов, дополняющих теорему А. Н. Колмогорова (В. И. Арнольд [18], Л. А. Бассалыго [19], М. Л. Гервер, Р. Досс [20], Г. Г. Лоренц [21], П. А. Остранд [22], В. М. Тихомиров, Г. М. Хенкин, Д. Шпрехер [23, 24] и др.).

При рассмотрении суперпозиций гладких функций характер результатов существенно меняется. Один из таких результатов был указан в разделе III. Еще один результат связан с рассмотрением так называемых линейных суперпозиций.

V. Линейные суперпозиции

Одной из наиболее интересных задач в тематике суперпозиций в настоящее время является следующая: существует ли аналитическая функция двух переменных, не представимая в виде конечной суперпозиции непрерывно дифференцируемых (гладких) функций одного переменного и операции сложения.

Линейные суперпозиции появляются в результате следующих рассуждений. Пусть функция двух переменных $f(x, y)$ является суперпозицией некоторых гладких функций одного переменного $\{f_i(t)\}$ и операции сложения. Проварьируем эту суперпозицию, т. е. рассмотрим суперпозицию $\tilde{f}(x, y)$ того же вида, но составленную из функций $\{f_i(t)+\varphi_i(t)\}$, где $\{\varphi_i(t)\}$ — малые возмущения, которые тоже являются гладкими функциями одного переменного. Тогда разность этих суперпозиций можно записать в виде

$$\tilde{f}(x, y)-f(x, y)=\sum_{i=1}^{N} p_i(x, y)\,\varphi_i(q_i(x, y))+o\left(\max_i \sup_t |\varphi_i(t)|\right), \quad (3)$$

где функции $\{p_i(x, y)\}$ расписываются через исходные функции $\{f_i(t)\}$ и их производные, а потому про них можно сказать лишь то, что они непрерывны; $\{q_i(x, y)\}$ расписываются только через функции $\{f_i(t)\}$, а потому они непрерывно дифференцируемы; остаточный член есть бесконечно малая величина по сравнению с $\max_i \sup_t \{\varphi_i(t)\}$, если только функции $\{d\varphi_i/dt\}$ имеют некоторый фиксированный модуль непрерывности.

Равенство (3) дает некоторую надежду на сведение общей задачи о суперпозициях гладких функций к отысканию аналитических

функций, не представимых суперпозициями вида

$$\sum_{i=1}^{N} p_i(x, y)\,\varphi_i(q_i(x, y)), \quad (4)$$

где $\{p_i(x, y)\}$ — наперед фиксированные непрерывные функции, $\{q_i(x, y)\}$ — наперед фиксированные непрерывно дифференцируемые функции, а $\{\varphi_i(t)\}$ — произвольные непрерывные функции одного переменного.

Такие суперпозиции будем называть линейными, подчеркивая этим, что функции $\{p_i(x, y)\}$ и $\{q_i(x, y)\}$ фиксированы, а от переменных функций $\{\varphi_i(t)\}$ суперпозиция зависит линейным образом. Отметим здесь же, что суперпозиции А. Н. Колмогорова [2] также линейные, причем все $p_i=1$, а $q_i=\sum_{j=1}^{n}\alpha_{i,j}(x_j)$ $(i=1, 2, \ldots, 2n+1)$ являются фиксированными непрерывными функциями.

Когда появилось ощущение, что задачу о суперпозициях гладких функций имеет смысл начинать с рассмотрения линейных суперпозиций (4) с гладкими функциями $\{q_i(x, y)\}$, то была сделана попытка воспользоваться аппаратом линейных отображений банаховых пространств, используя при этом специальные свойства пространств непрерывно дифференцируемых функций. Однако в таком виде методы линейного функционального анализа оказались здесь непригодными. Г. М. Хенкин [25] построил изоморфное в равномерной метрике и линейное отображение (вложение) пространства $(4p+1)$ раз непрерывно дифференцируемых функций двух переменных в пространство p раз непрерывно дифференцируемых функций одного переменного.

Использование в дальнейшем более специальных свойств линейных суперпозиций [9,10] позволило автору доказать, что для любых непрерывных функций $\{p_i(x, y)\}$ и непрерывно дифференцируемых функций $\{q_i(x, y)\}$ существует аналитическая функция двух переменных, не представимая суперпозицией вида (4).

Применение теории линейных отображений привело к более сильному результату: Г. М. Хенкин [11] показал, что множество суперпозиций вида (4) является замкнутым и нигде не плотным в пространстве всех непрерывных функций двух переменных. Отсюда, в частности, следует, что существует даже многочлен, не представимый суперпозицией вида (4).

Отметим, что имеющиеся доказательства проходят также и для суперпозиций вида

$$\sum_{i=1}^{N} p_i(x_1, x_2, \ldots, x_n)\,\varphi_i(q_i(x_1, x_2, \ldots, x_n)).$$

Но, как оказалось, эти доказательства не проходят для суперпозиций вида

$$\sum_{i=1}^{N} p_i(x_1, x_2, \ldots, x_n)\,\varphi_i(q_{1,i}(x_1, x_2, \ldots, x_n), \ldots, q_{k,i}(x_1, x_2, \ldots, x_n)), \tag{5}$$

где $\{p_i\}$ — фиксированные непрерывные функции n переменных; $\{q_{1,i}\}, \ldots, \{q_{k,i}\}$ — фиксированные гладкие функции n переменных $(1 < k < n)$. Не решен, например, вопрос, существует ли аналитическая функция n переменных, не представимая суперпозицией вида (5).

Что касается возможности сведения задачи о суперпозициях гладких функций к задаче о линейных суперпозициях, то удается провести такое сведение в случае «устойчивых» суперпозиций. На этом пути удается доказать, что, например, все аналитические функции n переменных нельзя представить суперпозициями непрерывно дифференцируемых функций от k переменных $(k < n)$ так, чтобы малым изменениям (в равномерной метрике) разлагаемых функций соответствовали столь же малые изменения функций, составляющих суперпозицию.

Математический институт им. В. А. Стеклова, Москва, СССР

ЛИТЕРАТУРА

[1] Hilbert D., Mathematische Prob'eme, *Gesammelte Abhandlungen*, **3** (1935), 290-329.
[2] Ostrowski A., Über Dirichletsche Reihen und algebraische Differentialgleichungen, *Math. Zeitschrift*, **8**, 3-4 (1920), 241-298.
[3] Hilbert D., Über die Gleichung neunten Grades, *Gesammelte Abhandlungen*, **2** (1933), 393-400.
[4] Витушкин А. Г., К тринадцатой проблеме Гильберта, *ДАН СССР*, **95**, № 4 (1954), 701-704.
[5] Колмогоров А. Н., Оценки минимального числа элементов ε-сетей в различных функциональных классах и их приложение к вопросу о представимости функций нескольких переменных суперпозицией функций меньшего числа переменных, *ДАН СССР*, **101**, № 2 (1955), 192-194.
[6] Колмогоров А. Н., О представлении непрерывных функций нескольких переменных суперпозициями непрерывных функций меньшего числа переменных, *ДАН СССР*, **108**, № 2 (1956), 179-182.
[7] Арнольд В. И., О функциях трех переменных, *ДАН СССР*, **114**, № 4 (1957), 679-681.
[8] Колмогоров А. Н., О представлении непрерывных функций нескольких переменных в виде суперпозиций непрерывных функций одного переменного и сложения, *ДАН СССР*, 114, № 5 (1957), 953-956.
[9] Витушкин А. Г., Некоторые свойства линейных суперпозиций гладких функций, *ДАН СССР*, **156**, № 5 (1964), 1003—1006.

[10] Витушкин А. Г., Доказательство существования аналитических функций многих переменных, не представимых линейными суперпозициями непрерывно дифференцируемых функций меньшего числа переменных, *ДАН СССР*, **156**, № 6 (1964), 1258-1261.
[11] Хенкин Г. М., О линейных суперпозициях непрерывно дифференцируемых функций, *ДАН СССР*, **157**, № 2 (1964), 288-290.
[12] Wiman A., Über die Anwendung der Tschirnhausen Transformation auf die Reduktion algebraischer Gleihungen, *Nova Acta R. Soc. Sc. Uppsaliensis*, vol. extra ordin. editum, 1927, 3-8.
[13] Чеботарев Г. Н., К проблеме резольвент, *Уч. записки Казанского ун-та*, **114**, кн. 2 (1954), 189-193.
[14] Чеботарев Н. Г., Собрание сочинений, т. I, 1949, 255-340.
[15] Морозов В. В., О некоторых вопросах проблемы резольвент, *Уч. записки Казанского ун-та*, **114**, кн. 2 (1954), 173-187.
[16] Витушкин А. Г., О многомерных вариациях, Москва, ГИТТЛ, 1955.
[17] Бари Н. К., Mémoire sur la représentation finie des fonctions continues, *Math. Ann.*, **103** (1930), 185-248, 598-653.
[18] Арнольд В. И., О представимости функций двух переменных в виде $\chi(\varphi(x) + \psi(y))$, *УМН*, **12**, вып. 2 (1957), 119-121.
[19] Бассалыго Л. А., О представлении непрерывных функций двух переменных при помощи непрерывных функций одного переменного, *Вестник МГУ*, № 1, 1966.
[20] Doss R., On the representation of the continuous functions of two variables by means of addition and continuous functions of one variable, *Colloq. math.*, X, № 2 (1963), 249-259.
[21] Lorentz G. G., Metric entropy, widths and superpositions of functions, *Amer. Math. Monthly*, **69**, № 6, 1962, 469-485.
[22] Ostrand P. A., Dimension of metric spaces and Hilbert's problem 13, *Bull. American Math. Soc.*, **71**, № 4 (1965), 619-622.
[23] Sprecher D. A., On the structure of continuous functions of several variables, *Trans. Amer. Math. Soc.*, **115** (1964), 340-355.
[24] Sprecher D. A., On the structure of several variables as finite sums of continuous functions of one variable, *Proc. American Math. Soc.*, **17**, № 1 (1966).
[25] Хенкин Г. М., О вложении пространства s-гладких функций n переменных в пространство достаточно гладких функций меньшего числа переменных, *ДАН СССР*, **153**, № 1 (1963), 57-60.

СКОРОСТЬ ПРИБЛИЖЕНИЯ РАЦИОНАЛЬНЫМИ ДРОБЯМИ И СВОЙСТВА ФУНКЦИЙ

А. А. ГОНЧАР

Постановка задачи о наилучшем приближении функций вещественного переменного посредством рациональных функций заданного порядка принадлежит П. Л. Чебышеву. В классических работах П. Л. Чебышева [1] и Е. И. Золотарева [2] получен ряд принципиальных результатов в этом направлении: установлено характеристичес-

кое свойство вещественной рациональной функции наилучшего приближения (теорема Чебышева об альтернансе); найдены точные выражения для наилучших приближений некоторых конкретных функций, в частности решена важная задача о наилучшем приближении функции $\operatorname{sgn} x$, $x \in [-1,-k] \cup [k, 1]$, $0 < k < 1$, посредством рациональных функций (четвертая задача Золотарева) [1]).

Однако систематическое изучение связей между общими структурными свойствами функций и скоростью стремления к нулю их наилучших приближений рациональными функциями было начато лишь недавно. Интерес к этой проблематике, по крайней мере в Советском Союзе, возник под влиянием А. Н. Колмогорова и С. Н. Мергеляна, которым принадлежит, в частности, ряд важных постановок задач. Ниже приводится обзор некоторых результатов, полученных в этом направлении за последние 10—12 лет [2]).

Подчеркнем, что речь идет о приближениях рациональными функциями со *свободными* полюсами [3]); именно в этом случае в полной мере проявляются специфические особенности наилучших приближений рациональными функциями по сравнению с наилучшими приближениями многочленами. Имея в виду проследить за этими особенностями, мы ограничимся формулировкой основных результатов для функций, определенных на отрезке вещественной прямой. Заметим, однако, что многие из приведенных ниже теорем устанавливаются специфически «комплексными» методами и допускают естественные обобщения на случай функций комплексного переменного, определенных на множествах гораздо более общей природы; в некоторых случаях мы формулируем лишь весьма специальные следствия таких общих теорем. Краткие указания об этом делаются в тексте статьи.

1. Мы будем рассматривать, вообще говоря, комплекснозначные функции f, определенные на подмножествах вещественной прямой $\mathbf{R} = \{x\}$ или комплексной плоскости $\mathbf{C} = \{z = x + iy\}$.

Порядком рациональной функции $r(z) = p(z)/q(z)$ будем называть бо́льшую из степеней многочленов $p(z)$ и $q(z)$ (в предположении, что они не имеют общих нулей). Всюду в дальнейшем через $r_n(z)$ обозначается рациональная функция порядка не выше n.

Наилучшее приближение функции f, определенной и непрерывной на ограниченном замкнутом множестве $E \subset \mathbf{C}$, посредством рациональных функций порядка не выше n обозначим через $R_n(f, E)$. По определению, $R_n(f, E)$ есть нижняя грань чисел

$$\max_{z \in E} |f(z) - r_n(z)|$$

в классе всех рациональных функций $r_n(z)$ порядка $\leqslant n$, без каких-либо ограничений на расположение полюсов. Наилучшее приближение функции $f(z)$, $z \in E$, посредством многочленов степени $\leqslant n$ будем обозначать через $P_n(f, E)$.

Как было сказано выше, в основном мы будем рассматривать случай, когда $E = \Delta$ есть отрезок вещественной прямой. Для определенности, положим $\Delta = [-1, 1]$; в этом случае наилучшие приближения будем обозначать просто через $R_n(f)$ и $P_n(f)$.

Поскольку переход от многочленов к рациональным функциям связан с расширением класса аппроксимирующих функций, имеем

$$R_n(f) \leqslant P_n(f), \qquad n = 0, 1, 2, \ldots.$$

Возникают следующие вопросы: какова взаимосвязь между структурными свойствами функций f и скоростью стремления к нулю последовательности $R_n(f)$; в частности, насколько быстро — по сравнению с $P_n(f)$ — стремится к нулю последовательность $R_n(f)$ для функций f, принадлежащих тем или иным функциональным классам.

2. Прямые теоремы о наилучших приближениях многочленами справедливы, очевидно, и для наилучших приближений рациональными функциями; что касается обратных теорем, то это уже не так. Поэтому изучение интересующих нас вопросов естественно начать с выяснения того, какие свойства функции f влечет за собой та или иная быстрота убывания $R_n(f)$. Полученные на этом пути результаты позволят, в частности, понять, в каком направлении возможно усиление прямых теорем при переходе от многочленов к рациональным функциям.

Прежде всего ясно, что никакая скорость стремления к нулю $R_n(f)$ не может обеспечить аналитичность функции f ни в одной из точек отрезка Δ. Это нетрудно показать с помощью рядов вида

$$\sum_{n=1}^{\infty} \frac{A_n}{z - \alpha_n},$$

где точки α_n расположены в дополнении к Δ так, что множество предельных точек последовательности $\{\alpha_n\}$ содержит этот отрезок

---

[1]) Упомянутые выше результаты П. Л. Чебышева и Е. И. Золотарева отражены в известной монографии Н. И. Ахиезера [3] (см. гл. II и п. 35 в разделе «Дополнения и задачи»). В этой монографии приводятся также точные решения ряда других задач, связанных с наилучшими приближениями рациональными функциями; там же см. более подробную библиографию.

[2]) Некоторые из приведенных ниже результатов автора получены после выступления с докладом на конгрессе.

[3]) Этим объясняется тот факт, что здесь не упомянуты многие интересные результаты Дж. Уолша (см. монографию [4]), М. М. Джрбашяна, С. Н. Мергеляна, Г. Ц. Тумаркина и др., относящиеся к приближениям рациональными функциями с *фиксированными* полюсами.

(в частности, последовательность $\{\alpha_n\}$ может быть выбрана всюду плотной в $\mathbf{C}$), а коэффициенты A_n достаточно быстро стремятся к нулю. В неявной форме это утверждение содержится в работе С. Н. Мергеляна [5].

Более того, как бы медленно ни стремилась к нулю функция $\omega(\delta) > 0$ (при $\delta \to 0$) и как бы быстро ни убывала к нулю последовательность $\varepsilon_n > 0$ (при $n \to \infty$), существует непрерывная функция $f(x)$, $x \in \Delta$, для которой

$$R_n(f) < \varepsilon_n, \qquad n = 0, 1, 2, \ldots,$$

в то время как модуль непрерывности $\omega_f(\delta)$ этой функции на отрезке Δ удовлетворяет соотношению

$$\overline{\lim_{\delta \to 0}} \frac{\omega_f(\delta)}{\omega(\delta)} > 1$$

(см. [6]; можно добиться того, чтобы последнему соотношению удовлетворял и модуль непрерывности функции f на произвольном отрезке, принадлежащем Δ).

Таким образом, нельзя гарантировать не только аналитичности функции f, но и вообще нельзя получить никакой оценки для ее модуля непрерывности на Δ (или каком-либо отрезке, принадлежащем Δ), если исходить из скорости убывания последовательности $R_n(f)$.

Отсюда следует, в частности, что существуют функции, наилучшие приближения которых многочленами стремятся к нулю произвольно медленно, в то время как наилучшие приближения рациональными функциями убывают с произвольно высокой скоростью. Точнее, каковы бы ни были последовательности $E_n > 0$ и $\varepsilon_n > 0$, стремящиеся к нулю при $n \to \infty$, существует функция f, такая, что

$$P_n(f) \neq O(E_n),$$
$$R_n(f) < \varepsilon_n, \qquad n = 0, 1, 2, \ldots .$$

3. Однако ряд свойств функций f (в частности, свойств, формулируемых в дифференциальных терминах) можно гарантировать и исходя из скорости их приближения рациональными функциями. Приведем формулировки теорем, полученных в работе [6].

Если

$$R_n(f) < \frac{C}{n^{1+\delta}}, \qquad \delta > 0,$$

то функция f дифференцируема почти всюду на отрезке Δ.

Если

$$R_n(f) < \frac{C}{n^{k+\delta}}, \qquad \delta > 0,$$

где $k > 0$ — целое, то функция f почти всюду на отрезке Δ имеет k-ю асимптотическую производную.

Позже Е. П. Долженко [7] заметил, что в последней теореме можно гарантировать существование почти всюду на Δ не только k-й асимптотической производной, но и k-й производной в смысле Пеано (локального дифференциала k-го порядка) функции f.

Уже из приведенных теорем следует, что основное заключение хорошо известной обратной теоремы Бернштейна о наилучших приближениях многочленами сохраняет силу и для приближений рациональными функциями — с той существенной разницей, что теперь уже необходимо допускать исключительные множества произвольно малой меры. Исключительное множество точек недифференцируемости функции f может быть всюду плотным на Δ; этим объясняется наличие примеров, упомянутых в предыдущем пункте.

Аналогия с теоремой Бернштейна еще более выпукло выступает в следующей теореме [8]. Пусть P — произвольное совершенное подмножество вещественной прямой $\mathbf{R}$; $A = k + \alpha$, $k \geqslant 0$ — целое, $0 < \alpha \leqslant 1$. Будем говорить, что функция $f(x)$, $x \in P$, принадлежит классу $\operatorname{Lip} A$ или, короче, $f \in L(A)$, если функция f в каждой точке $x \in P$ имеет k-ю производную (по P), удовлетворяющую на множестве P условию Липшица порядка α. В обозначении класса $L(A)$ символ множества мы опускаем; запись $f \in L(A)$ означает, что функция f принадлежит классу $L(A)$ на том множестве, на котором она определена.

Если

$$R_n(f) < \frac{C}{n^{A+\delta}}, \qquad \delta > 0,$$

то, каково бы ни было $\varepsilon > 0$, существует совершенное множество $P_\varepsilon \subset \Delta$, $\operatorname{mes}(\Delta \setminus P_\varepsilon) < \varepsilon$, такое, что функция $f \mid_{P_\varepsilon}$ (сужение функции f на множество P_ε) принадлежит классу $L(A)$.

Обратные теоремы о наилучших приближениях рациональными функциями, очевидно, остаются справедливыми и после замены $R_n(f)$ на $P_n(f)$. В связи с этим интересно отметить, что все теоремы, приведенные в этом пункте, переносятся с отрезка на произвольные, по существу, подмножества вещественной прямой (и на широкие классы подмножеств комплексной плоскости). Эти теоремы показывают, что сложное влияние свойств множества E на зависимость дифференциальных свойств функции $f(x)$, $x \in E$, от скорости стремления к нулю ее наилучших приближений многочленами $P_n(f, E)$ (см. [5]) связано с желанием гарантировать определенные дифференциальные свойства функции f на всем множестве E. Пренебрегая множествами малой меры, можно избавиться от этого влияния и получить обратные теоремы, являющиеся непосредственными аналогами теоремы Бернштейна, например, для произвольных

замкнутых множеств $E \subset \mathbf{R}$. В частности, *если* $P_n(f, E) < Cn^{-(A+\delta)}$, $\delta > 0$, *то для любого* $\varepsilon > 0$ *существует совершенное множество* $P_\varepsilon \subset E$, $\operatorname{mes}(E \setminus P_\varepsilon) < \varepsilon$, *такое, что* $f|_{P_\varepsilon} \in L(A)$.

Возвращаясь к теоремам о наилучших приближениях рациональными функциями, формулируемым в дифференциальных терминах, напомним, что скорость стремления к нулю $R_n(f)$ порядка $\frac{1}{n^{1+\delta}}$, $\delta > 0$, обеспечивает дифференцируемость функции f почти всюду на Δ; бо́льшая скорость стремления к нулю $R_n(f)$ обеспечивает дополнительную гладкость функции f почти всюду на Δ (в обобщенном смысле, или «внутри» Δ — на множествах P_ε — в обычном смысле). Возникает вопрос, насколько окончательной при этом остается характеристика исключительного множества точек недифференцируемости функции f как множества лебеговой меры нуль. Следующие теоремы (некоторые из них сформулированы в [8]) показывают, что дополнительная скорость убывания $R_n(f)$ накладывает дополнительные ограничения и на множество точек недифференцируемости f, и устанавливают зависимость метрических свойств этого множества от скорости стремления к нулю $R_n(f)$.

Если

$$R_n(f) < \frac{C}{n^A}, \quad A > 1,$$

то, каково бы ни было $a > \frac{1}{A}$, *функция* f *дифференцируема почти всюду на* Δ *относительно* h*-меры Хаусдорфа, соответствующей функции* $h(r) = r^a$.

Другими словами, хаусдорфова размерность множества точек недифференцируемости функции f не больше $\frac{1}{A}$.

Если

$$R_n(f) < C\exp(-n^B), \quad B > 0,$$

то, каково бы ни было $b > \frac{1}{B}$, *функция* f *дифференцируема почти всюду на* Δ *относительно* h*-меры Хаусдорфа, соответствующей функции* $h(r) = \left(\log\frac{1}{r}\right)^{-b}$.

В условиях последней теоремы можно утверждать, что функция f не только имеет первую производную, но и *бесконечно дифференцируема* (в обобщенном смысле, например в смысле существования локального дифференциала любого порядка) почти всюду на Δ относительно h-меры, $h(r) = (\log\frac{1}{r})^{-b}$, $b > \frac{1}{B}$.

Справедливы также теоремы, одновременно учитывающие свойства гладкости более общего вида (условие Липшица, производные

высших порядков) и метрические свойства исключительных множеств, вне которых гарантируется соответствующая гладкость. Например:

Если

$$R_n(f) < \frac{C}{n^A}, \quad A > 0,$$

то, каковы бы ни были $A_1 > 0$ *и* $A_2 > 0$, $A_1 \cdot A_2 < A$, *найдется совершенное множество* $P_\varepsilon \subset \Delta$, *такое, что*

1) *множество* $\Delta \setminus P_\varepsilon$ *можно покрыть системой интервалов* δ_i, *удовлетворяющих условию* ($|\delta_i|$ — *длина* δ_i)

$$\sum_i |\delta_i|^{1/A_1} < \varepsilon;$$

2) $f|_{P_\varepsilon} \in L(A_2)$.

В условиях той же теоремы можно утверждать, что множество точек $x \in \Delta$, в которых функция f дифференцируема (в обобщенном смысле) менее, чем k раз, имеет хаусдорфову размерность, не превышающую $\frac{k}{A}$. В частности, *если*

$$\lim n^A \cdot R_n(f) = 0$$

при любом $A > 0$, *то функция* f *бесконечно дифференцируема (в обобщенном смысле) всюду на отрезке* Δ, *за исключением множества, хаусдорфова размерность которого равна нулю.* Теоремы, приведенные в начале этого пункта, позволяют утверждать только то, что лебегова мера этого множества равна нулю.

Приведем формулировку еще одной обратной теоремы.

Если

$$\sum_{n=1}^{\infty} R_n(f) < \infty,$$

то функция f *абсолютно непрерывна на отрезке* Δ.

Эта теорема принадлежит Е. П. Долженко [9]; он же показал, что, какова бы ни была последовательность $a_n > 0$, $\Sigma a_n = \infty$, существует функция f, не только не являющаяся абсолютно непрерывной, но и не имеющая почти всюду на Δ асимптотической производной, для которой $R_n(f) < a_n$.

Ограничения на $R_n(f)$, фигурирующие и в других теоремах настоящего пункта, сколько-нибудь существенно ослабить нельзя. В то же время эти теоремы не допускают обращения; гарантируемые ими свойства непрерывных функций f далеко не достаточны для того, чтобы можно было утверждать справедливость каких-либо оценок для $R_n(f)$. В частности, недавно Долженко привел простой пример монотонной и абсолютно непрерывной функции f,

для которой $R_n(f)$ стремятся к нулю произвольно медленно (см. также п. 5 ниже).

Тем не менее приведенные выше обратные теоремы позволяют сделать ряд выводов, полезных и с точки зрения прямых теорем; во всяком случае, они указывают некоторые из тех свойств, которые необходимо требовать от функций f, чтобы можно было гарантировать ту или иную скорость их приближения рациональными функциями.

4. Как уже отмечалось выше, прямые теоремы, установленные для наилучших приближений многочленами, справедливы и для наилучших приближений рациональными функциями. В частности, если функция f определена на отрезке Δ и принадлежит классу $L(A)$, то по теореме Джексона

$$P_n(f) < \frac{C}{n^A}, \quad C = C_f;$$

тем самым для $f \in L(A)$ и

$$R_n(f) < \frac{C}{n^A}, \quad C = C_f.$$

Приведенная в п. 3 обратная теорема (формулируемая в терминах множеств P_ε) показывает, что во всем классе $L(A)$ последнее неравенство сколько-нибудь существенно улучшить нельзя. Действительно, существуют функции $f \in L(A)$, такие, что, каково бы ни было $A' > A$, $f|_P \notin L(A')$ для некоторого множества $P \subset \Delta$ положительной меры. Для таких функций f имеем

$$R_n(f) \neq O\left(\frac{1}{n^{A+\delta}}\right),$$

где $\delta > 0$ — любое. Пусть

$$\pi(f) = \varlimsup_{n\to\infty} \frac{\log P_n(f)}{\log \frac{1}{n}};$$

$$\rho(f) = \varlimsup_{n\to\infty} \frac{\log R_n(f)}{\log \frac{1}{n}};$$

числа $\pi(f)$ и $\rho(f)$ характеризуют порядок (показатель степени при $\frac{1}{n}$) стремления к нулю $P_n(f)$ и $R_n(f)$ соответственно. При любом $A > 0$ имеем

$$\inf_{f\in L(A)} \pi(f) = \inf_{f\in L(A)} \rho(f) = A.$$

Таким образом, расширение класса аппроксимирующих функций от многочленов до рациональных функций не позволяет улучшить порядок стремления к нулю наилучших приближений для всего класса $L(A)$ на отрезке Δ. Переход от многочленов к рациональным функциям сказывается в другом — расширяется класс функций, допускающих приближение того же порядка.

Приведенные выше обратные теоремы показывают, что такое расширение может произойти за счет функций, «особенности» которых составляют достаточно «редкое» множество. И действительно, именно на этом пути недавно были выделены широкие классы функций, аппроксимируемых рациональными функциями существенно лучше, чем многочленами.

Основной результат в этом направлении был получен в работе Д. Ньюмена [10]. Д. Ньюмен нашел порядок приближения функции $|x|$ на отрезке Δ посредством рациональных функций. Соответствующую теорему коротко можно сформулировать так:

$$\frac{1}{2}e^{-9\sqrt{n}} < R_n(|x|) < 3e^{-\sqrt{n}};$$

напомним, что

$$P_n(|x|) \sim \frac{c}{n}, \quad n \to \infty.$$

Интересно отметить, что неравенства для $R_n(|x|)$, аналогичные приведенным выше, нетрудно вывести из теоремы Е. И. Золотарева (доказанной еще в 1877 году!) о наилучшем приближении функции $\operatorname{sgn} x$ на объединении отрезков $[-1, -k]$ и $[k, 1]$, $0 < k < 1$, посредством рациональных функций [1]). Более того, из теоремы Золотарева вытекают следующие, асимптотически более точные, неравенства:

$$e^{-(\pi\sqrt{2}+\varepsilon)\sqrt{n}} < R_n(|x|) < e^{-\left(\frac{\pi}{\sqrt{2}}-\varepsilon\right)\sqrt{n}},$$

где $\varepsilon > 0$ — любое, $n > n(\varepsilon)$ (подробности см. в [11]).

Важность решения задачи о скорости приближения функции $|x|$, $x \in \Delta$, хорошо известна. Более или менее простыми следствиями теоремы Ньюмена (и известных результатов конструктивной теории функций) являются следующие теоремы о скорости приближения кусочно «хороших» функций.

Будем говорить, что непрерывная функция $f(x)$, $x \in \Delta$, принадлежит классу $\hat{L}(A)$, если существует разбиение отрезка Δ на конечное число отрезков $\Delta_1, \ldots, \Delta_N$, такое, что при любом $i = 1, \ldots, N$ $f_i = f|_{\Delta_i} \in L(A)$.

[1]) В работе Д. Ньюмена [10] приводится очень интересное прямое доказательство оценок для $R_n(|x|)$; теорема Золотарева Ньюмену не была известна.

Если $f \in \hat{L}(A)$, *то*

$$R_n(f) < \frac{C}{n^A}, \quad C = C_f.$$

Следовательно, при любом $A > 0$

$$\inf_{f \in \hat{L}(A)} \rho(f) = A;$$

заметим, что, сколь велико бы ни было $A \geqslant 1$,

$$\inf_{f \in \hat{L}(A)} \pi(f) = 1$$

($\pi(f) = 1$ уже для функции $|x|$, $x \in \Delta$).

Непрерывную функцию f будем называть кусочно бесконечно дифференцируемой на Δ, если существует разбиение отрезка Δ на конечное число отрезков $\Delta_1, \ldots, \Delta_N$, такое, что каждая из функций $f_i = f|_{\Delta_i}$ бесконечно дифференцируема на соответствующем отрезке Δ_i. Если при этом для любого i, $1 \leqslant i \leqslant N$, и некоторого $\lambda > 0$

$$\max_{x \in \Delta_i} |f_i^{(p)}(x)| \leqslant Q^p \cdot p^{\lambda p}, \qquad p = 1, 2, \ldots,$$

где $Q = Q_f > 0$ — постоянная, зависящая только от f, то будем говорить, что функция f принадлежит классу $\hat{I}(\lambda)$.

Если $f \in \hat{I}(\lambda)$, *то*

$$R_n(f) < \exp(-c \cdot n^{\mu}),$$

где $c = c_f > 0$ *и*

$$\mu = \min\left(\frac{1}{\lambda}, \frac{1}{2}\right).$$

В частности, если f — кусочно аналитическая функция на Δ (f_i аналитичны на Δ_i, $i = 1, \ldots, N$), то $f \in \hat{I}(1)$. Получаем:

Если f — *кусочно аналитическая функция на* Δ, *то*

$$R_n(f) < e^{-c\sqrt{n}}, \quad c = c_f > 0.$$

Это обобщение теоремы Ньюмена (точнее, оценки сверху для $R_n(|x|)$) принадлежит П. Турану и П. Шюшу (см. [12]).

Оценка снизу для $R_n(|x|)$ также допускает обобщение [11]; наличие в некотором смысле правильного «излома» у графика функции или какой-либо ее производной в любом случае является препятствием для приближения рациональными функциями, лучшего чем $e^{-c\sqrt{n}}$. Если, например, в некоторой точке $x_0 \in (-1, 1)$ при каком-либо натуральном k справедливы разложения ($h > 0$; $\delta > 0$

и не зависит от h):

$$f(x_0 - h) - f(x_0) = a_1 h + \ldots + a_k h^k + O(h^{k+\delta}),$$
$$f(x_0 + h) - f(x_0) = b_1 h + \ldots + b_k h^k + O(h^{k+\delta})$$

и хотя бы при одном i, $1 \leqslant i \leqslant k$, $a_i \neq b_i$, то существует постоянная $c > 0$, такая, что

$$R_n(f) > e^{-c\sqrt{n}}, \quad n = 1, 2, \ldots.$$

Отсюда вытекает следующее утверждение.

Если f — *кусочно бесконечно дифференцируемая (в частности, кусочно аналитическая) функция на* Δ *и для некоторой последовательности* $n = n_i \uparrow \infty$ *при* $i \uparrow \infty$ *имеют место неравенства*

$$R_n(f) < e^{-A_n \cdot \sqrt{n}},$$

где $A_n \to \infty$ *при* $n \to \infty$, *то* f — *бесконечно дифференцируема (соответственно аналитична) на* Δ.

В связи с результатами о приближении кусочно гладких функций возникает вопрос, нельзя ли дать оценку для $R_n(f)$, где f — непрерывная функция на Δ, через $R_n(f_i, \Delta_i)$, $f_i = f|_{\Delta_i}$, $i = 1, \ldots, N$ (справедливую хотя бы в некотором интервале скоростей). Ответ на этот вопрос отрицательный: каковы бы ни были последовательности $\varepsilon_n > 0$ и $E_n > 0$, стремящиеся к нулю при $n \to \infty$, существует непрерывная (и монотонная) на Δ функция f, такая, что

$$R_n(f_i, \Delta_i) < \varepsilon_n, \quad \Delta_1 = [-1, 0], \quad \Delta_2 = [0, 1], \quad i = 1, 2,$$

но

$$R_n(f) > E_n, \quad n > n_0.$$

При этом функцию f можно построить так, что $f(0) = 0$ и для функции

$$\varphi(x) = \begin{cases} f(x), & x \in \Delta_1, \\ -f(x), & x \in \Delta_2, \end{cases}$$

также имеют место неравенства

$$R_n(\varphi) < \varepsilon_n$$

(функция f в точке 0 очень быстро растет — и приближается плохо, в то время как функция φ имеет в этой точке очень острый пик — и приближается хорошо; в остальных точках отрезка Δ обе функции аналитичны).

В приведенных выше прямых теоремах относительно $R_n(f)$ утверждается гораздо больше, чем можно утверждать — при тех же условиях на f — относительно $P_n(f)$. Туран и Шюш,

Фрейд и др. получили недавно ряд других теорем такого типа (не сводящихся к склеиванию «хороших» функций); доказательства этих теорем также используют — в той или иной форме — результат Ньюмена о скорости приближения функции $|x|$, $x \in \Delta$. Однако условия на функцию f, фигурирующие во всех этих теоремах, далеко не необходимы для того, чтобы имели место соответствующие оценки для $R_n(f)$ (хотя при данных условиях эти оценки и нельзя существенно улучшить). Между прямыми и обратными теоремами существует большой разрыв, причем пока не ясны термины, которые позволили бы его ликвидировать. В частности, остается открытым вопрос о том, какими внутренними свойствами характеризуется класс функций f, для которых $\rho(f) \geqslant A$ (или даже $0 < \rho(f) < \infty$).

5. Обратные теоремы, приведенные в п. 3, показывают, что препятствием для хорошего приближения рациональными функциями являются плохие дифференциальные свойства аппроксимируемых функций на достаточно массивных множествах (например, на множествах, имеющих положительную меру Лебега или h-меру Хаусдорфа, соответствующую той или иной функции h); в них содержится и количественное выражение этого факта. С помощью этих теорем нетрудно строить примеры функций, наилучшие приближения которых рациональными функциями стремятся к нулю примерно как $\frac{1}{n^A}$ или e^{-n^B} (при любых $A > 0$ и $B > 0$).

Покажем теперь, что поведение функции в одной внутренней точке отрезка (точнее, в окрестности этой точки) также может служить препятствием для хорошего приближения рациональными функциями. В известной мере это следует уже из результатов предыдущего пункта, в которых фигурирует скорость порядка $e^{-c\sqrt{n}}$, и в первую очередь из оценки снизу для $R_n(|x|)$. Однако эти результаты представляются несколько случайными, поскольку соответствующие препятствия сказываются лишь при желании приблизить функцию со скоростью, лучшей чем $e^{-c\sqrt{n}}$, и только в этом случае [1]). Возникает вопрос о том, нельзя ли построить непрерывную шкалу препятствий, связанных с локальным поведением функции, в которую вкладывалась бы оценка снизу для $R_n(|x|)$.

[1]) Заметим, что эти результаты нельзя объяснить тем, что для соответствующих функций может быть нарушено свойство единственности (в частности, это свойство нарушено для функции $|x|$, $x \in \Delta$). Общие теоремы единственности, приведенные в следующем пункте, справедливы только при скорости приближения порядка e^{-cn}, $c > 0$; каково бы ни было $B < 1$, в частности $\frac{1}{2} < B < 1$, скорость приближения порядка e^{-n^B} совместима с неединственностью.

Такую шкалу можно построить, если исходить, например, из скорости роста функций в фиксированной внутренней точке отрезка [11]. Мы приведем оценки снизу для наилучших приближений канонических функций [1]); можно было бы отвлечься от конкретной формы рассматриваемых функций и сформулировать соответствующие результаты в более общем виде.

Пусть

$$f_\varphi(x) = \begin{cases} 0, & x \in \Delta_1 = [-1, 0], \\ \varphi(x), & x \in \Delta_2 = [0, 1], \end{cases}$$

где φ — непрерывная возрастающая функция на отрезке $[0, 1]$, $\varphi(0) = 0$. Приведенные ниже оценки справедливы при всех $n \geqslant 1$; значения (положительных) постоянных c мы не выписываем, указывая только на те параметры, от которых они могут зависеть.

Переформулируем результат Ньюмена.

1°. *Если* $\varphi(x) = x$, *то*

$$R_n(f_\varphi) > e^{-c\sqrt{n}}.$$

Переход от функции x к функции x^α, $\alpha > 0$, еще не позволяет изменить показатель степени n в последней оценке.

2°. *Если* $\varphi(x) = x^\alpha$, $\alpha > 0$, *то*

$$R_n(f_\varphi) > e^{-c_\alpha \cdot \sqrt{n}}.$$

Непрерывная шкала получается при переходе к функциям, растущим существенно быстрее ($\beta < 1$) или медленнее ($\beta > 1$) любой из функций x^α, $\alpha > 0$.

3°. *Если*

$$\varphi(x) = \exp\left[-\alpha\left(\log\frac{1}{x}\right)^\beta\right], \quad \alpha > 0, \ \beta > 0,$$

то

$$R_n(f_\varphi) > \exp\left(-c_{\alpha,\beta} \cdot n^{\frac{\beta}{\beta+1}}\right).$$

При $\alpha = \beta = 1$ получаем оценку Ньюмена. Постоянную $c_{\alpha,\beta}$ можно выбрать зависящей только от α. Приведем еще две оценки — для функций φ, окаймляющих рассмотренные выше.

4°. *Если* $\varphi(x) = \left(\log\frac{1}{x}\right)^{-\gamma}$, $\gamma > 0$, $0 \leqslant x \leqslant \frac{1}{2}$ (*и в остальном произвольна*), *то*

$$R_n(f_\varphi) > c_\gamma \cdot n^{-\gamma}.$$

[1]) Соответствующие оценки сверху получены в статье автора [13].

В частности, если в окрестности начала координат

$$\varphi(x)=\frac{1}{\log\frac{1}{x}},$$

то

$$R_n(f_\varphi)>\frac{c}{n}.$$

5°. *Если*

$$\varphi(x)=\exp\left(-\frac{1}{x^\delta}\right),\quad \delta>0,$$

то

$$R_n(f_\varphi)>\exp\left(-\frac{c_\delta\cdot n}{\log n}\right).$$

Утверждения 1°—5° вытекают из следующей оценки, справедливой для любой функции вида f_φ:

$$R_n(f_\varphi)\geqslant \sup_{0<\varepsilon<1}[\varphi(\varepsilon)\cdot(e^{\frac{\pi^2 n}{\log 1/\varepsilon}}+1)^{-1}].$$

Эта оценка позволяет в каждом случае выписать зависимость постоянной от соответствующих параметров. Например, в 2° $c_\alpha=C\cdot\sqrt{\alpha}$, в 3° постоянная $c_{\alpha,\beta}$ может быть выбрана как $c+\alpha$ (неравенства с этими постоянными имеют место для $2R_n(f_\varphi)$, $n\geqslant 1$); c — абсолютные постоянные, значения которых также могут быть указаны.

Выбирая функцию φ растущей достаточно быстро, можно получить функцию f_φ со сколь угодно медленно убывающей последовательностью $R_n(f_\varphi)$; наоборот, выбирая ее растущей достаточно медленно (другими словами, имеющей высокий порядок касания с вещественной прямой в точке $x=0$), можно произвольно приблизиться к скорости геометрической прогрессии e^{-cn}.

6. Хорошо известно, что класс всех функций f, аналитических на отрезке Δ, характеризуется условием

$$\overline{\lim_{n\to\infty}}\sqrt[n]{P_n(f)}<1$$

(теорема Бернштейна). В п. 2 уже отмечалось, что любая скорость стремления к нулю последовательности $R_n(f)$ совместима с неаналитичностью функции f в каждой точке отрезка Δ. С. Н. Мергеляном было высказано предположение, что тем не менее класс функций f, для которых последовательность $R_n(f)$ стремится к нулю со скоростью геометрической прогрессии, должен обладать основным свойством класса аналитических функций на отрезке — свойством единственности. В работе [6] были доказаны несколько более слабые теоремы единственности; в полной общности соответствующие теоремы удалось доказать лишь недавно.

Обозначим через $R(\Delta)$ класс всех функций f, для которых выполнено условие

$$\overline{\lim_{n\to\infty}}\sqrt[n]{R_n(f)}<1.$$

Справедлива следующая теорема:

Если функции f и g принадлежат классу $R(\Delta)$ и $f(x)=g(x)$ на множестве $e\subset\Delta$ положительной лебеговой меры (или положительной гармонической емкости), то $f(x)\equiv g(x)$, $x\in\Delta$.

Теорема единственности справедлива и в следующей формулировке:

Если

$$\overline{\lim_{n\to\infty}}\sqrt[n]{R_n(f)}<1$$

и $f(x)=0$ на множестве $e\subset\Delta$ положительной меры (емкости), то $f(x)\equiv 0$, $x\in\Delta$.

Поэтому каждый из классов $R_{\{n_k\}}(\Delta)$, определяемых условием

$$\overline{\lim_{k\to\infty}}\sqrt[n_k]{R_{n_k}(f)}<1,$$

где $\{n_k\}$ — возрастающая последовательность натуральных чисел, также обладает свойством единственности и является тем самым квазианалитическим классом функций.

Класс $R(\Delta)$ является естественным расширением класса функций, аналитических на отрезке Δ, в котором сохраняется свойство единственности. Этот класс является линейным пространством и *кольцом*: если f и $g\in R(\Delta)$, то и $f\cdot g\in R(\Delta)$.

Переход от $R(\Delta)$ к $R_{\{n_k\}}(\Delta)$ совершенно аналогичен переходу от класса всех аналитических функций к квазианалитическим классам С. Н. Бернштейна $B_{\{n_k\}}(\Delta)$. Каждый из классов $R_{\{n_k\}}(\Delta)$ является расширением соответствующего класса $B_{\{n_k\}}(\Delta)$.

Аналогичные теоремы справедливы на произвольных *континуумах* в комплексной плоскости. Несколько видоизменяя постановку задачи, на этом пути можно получить одно из возможных решений задачи Э. Бореля о квазианалитическом продолжении аналитических функций через жорданову дугу (см. [14]). Заметим, наконец, что при несколько более сильных требованиях на скорость убывания $R_n(f, E)$ теоремы единственности справедливы и для несвязных замкнутых множеств $E\subset\mathbf{C}$. Например, если E — произвольное замкнутое множество, приведенное относительно емкости (каж-

дая порция E имеет положительную емкость), то в классе функций $f(z)$, $z \in E$, для которых

$$\lim_{n\to\infty} \sqrt[n]{R_n(f, E)} = 0,$$

имеет место свойство единственности по значениям на множествах положительной емкости.

7. При изучении различных вопросов, связанных с приближениями рациональными функциями, удобно основываться на оценках равномерного роста рациональных функций (оценках типа max-min). Задача ставится так.

Пусть E_1 и E_2 — непересекающиеся континуумы в расширенной комплексной плоскости $\overline{\mathbf{C}}$, $r_n(z)$ — произвольная рациональная функция порядка не выше n, удовлетворяющая условию

$$\max_{z\in E_1} |r_n(z)| \leqslant M < \infty.$$

Требуется дать оценку для

$$\min_{z\in E_2} |r_n(z)|,$$

правая часть которой зависела бы только от M, n и взаимного расположения E_1 и E_2.

Положим

$$\sigma(r_n; E_1, E_2) = \frac{\min\{|r_n(z)|,\ z \in E_2\}}{\max\{|r_n(z)|,\ z \in E_1\}}.$$

Пусть D_1 — компонента дополнения к E_1, содержащая E_2; D_2 — компонента дополнения к E_2, содержащая E_1; $D = D_1 \cap D_2$ (если E_1 и E_2 не разбивают плоскость, то $D = \overline{\mathbf{C}} \setminus (E_1 \cup E_2)$). Здесь D — двусвязная область; обозначим через $\rho = \rho(D) > 1$ конформный модуль области D (отношение радиусов кругового кольца, конформно эквивалентного области D). Решение поставленной выше задачи дается следующей теоремой:

Для любой рациональной функции $r_n(z)$ порядка не выше n ($n \geqslant 1$ — любое)

$$\sigma(r_n; E_1, E_2) \leqslant \rho^n, \qquad \rho = \rho(D).$$

Из одной теоремы Уолша (см. [4], § 8.7, теорема 9) вытекает, что число ρ в последней оценке нельзя заменить никаким меньшим числом. Точнее, каково бы ни было $\theta < 1$, для любого $n > n(\theta)$ существует $r_n(z)$, такая, что

$$\sigma(r_n; E_1, E_2) > \rho^{\theta n}, \qquad n > n(\theta).$$

Объединяя это утверждение с предыдущим, получаем

$$\limsup_{n\to\infty} \sqrt[n]{\sigma_{\{r_n\}}(r_n; E_1, E_2)} = \rho;$$

верхняя грань берется в классе всех рациональных функций порядка не выше n.

В случае когда $E_1 = [-1, -k]$, $E_2 = [k, 1]$, $0 < k < 1$ (величину $\sigma(r_n; E_1, E_2)$ в этом случае будем обозначать через $\sigma_k(r_n)$), получаем следующее утверждение [11]:

Для любой рациональной функции $r_n(z)$ порядка не выше n

$$\sigma_k(r_n) \leqslant \exp\left(2\pi n \frac{K}{K'}\right),$$

где K — полный эллиптический интеграл первого рода для модуля k, K' — соответствующий интеграл для дополнительного модуля.

При малых k удобнее пользоваться следующей оценкой:

$$\sigma_k(r_n) < \exp\left(\frac{\pi^2 n}{\log\frac{1}{k}}\right);$$

при этом, каково бы ни было $\theta < 1$, существует $r_n(x)$, такая, что

$$\sigma_k(r_n) > \exp\left(\frac{\theta\pi^2 n}{\log\frac{1}{k}}\right), \qquad n > n(\theta).$$

Результаты, приведенные в п. 5, являются непосредственными следствиями первого из этих утверждений; второе утверждение можно использовать при доказательстве оценок сверху для наилучших приближений рациональными функциями (в частности, при выводе верхней оценки для $R_n(|x|)$). Переходя от последних неравенств к асимптотическим оценкам для наилучших приближений функции $\operatorname{sgn} x$ на $[-1, -k] \cup [k, 1]$, $0 < k < 1$ (постоянная π^2 под знаком exp заменяется при этом на $\frac{\pi^2}{2}$), можно получить те же неравенства для $R_n(|x|)$, которые были отмечены в п. 4 в качестве следствий теоремы Золотарева.

Аналогичные оценки можно доказать для произвольных замкнутых множеств E_1 и E_2 положительной гармонической емкости, лежащих в $\mathbf{C}$. Предположим, что соответствующие этим множествам области D_1 и D_2 (см. выше) регулярны (в смысле задачи Дирихле); μ — емкость конденсатора, образованного множествами E_1 и E_2 (гринова емкость E_1 относительно D_2 или E_2 относительно D_1). Тогда *для любой $r_n(z)$ справедлива оценка*

$$\sigma(r_n; E_1, E_2) \leqslant e^{\frac{n}{\mu}}.$$

Эта оценка лежит в основе доказательства теорем единственности, приведенных в п. 6. Из нее вытекает следующее утверждение (которое достаточно для доказательства теорем единственности для случая отрезка Δ и mes $e > 0$).

Пусть непересекающиеся замкнутые множества e_1 и e_2 принадлежат некоторому отрезку длины 2δ, причем mes $e_i \geqslant \delta - \varepsilon$, $i = 1, 2$. Тогда для любой $r_n(x)$

$$\sigma(r_n; E_1, E_2) \leqslant \exp\left(\frac{\pi^2 n}{\log\frac{\delta}{\varepsilon}}\right).$$

В заключение отметим, что в классическом мемуаре Е. И. Золотарева [2] получено *точное* решение задачи типа max-min для случая двух отрезков на вещественной прямой. Е. И. Золотарев нашел величину наибольшего уклонения от нуля на множестве $\Delta_2 = \{x: |x| \geqslant \frac{1}{k}\}$, $0 < k < 1$, в классе рациональных функций порядка n, модуль которых ограничен единицей на отрезке $\Delta_1 = [-1, 1]$ (третья задача Золотарева). Соответствующий результат можно сформулировать так:

$$\sup_{\{r_n\}} \sigma(r_n; \Delta_1, \Delta_2) = \frac{1}{k^n \left(\operatorname{sn}\frac{K}{n} \operatorname{sn}\frac{3K}{n} \ldots \operatorname{sn}\frac{\nu K}{n}\right)^4},$$

где $\Delta_1 = [-1, 1]$, $\Delta_2 = \{x: |x| \geqslant \frac{1}{k}\}$, $0 < k < 1$, K — полный эллиптический интеграл первого рода для модуля k, ν — наибольшее нечетное число, меньшее чем n.

Доказательство приведенных выше оценок (формулируемых в терминах модуля области D) не представляет труда для канонического случая $E_1 = \{z: |z| \leqslant 1\}$, $E_2 = \{z: |z| \geqslant \rho\}$, $\rho > 1$; пример многочлена $r_n(z) = z^n$ показывает, что в этом случае соответствующая оценка точная. Свести вывод тех же оценок для произвольных континуумов E_1 и E_2 (тем более для множеств общей природы) к этому каноническому случаю не удается. Можно дать различные прямые доказательства соответствующих оценок сверху для $\sigma(r_n; E_1, E_2)$ (в частности, для случая двух отрезков). Конечно, для случая двух отрезков на вещественной прямой оценки в терминах модуля дополнительной к ним области в принципе содержатся в приведенной выше теореме Золотарева (хотя вывод их из этой теоремы вряд ли является простой задачей).

8. Приведем некоторые результаты, относящиеся к вопросу о скорости приближения аналитических функций рациональными.

Пусть φ — функция, конформно отображающая дополнение к Δ на внешность единичного круга, $\varphi(\infty) = \infty$; $D_R = D_R(\Delta)$ —

конечная область, ограниченная линией уровня $|\varphi(z)| = R > 1$ (в рассматриваемом случае D_R — область, ограниченная эллипсом с фокусами в точках ± 1 и суммой полуосей, равной R). Класс функций, однозначных и аналитических в области $D \supset \Delta$ (так же как и его сужение на отрезок Δ), будем обозначать через $A(D)$.

Согласно известной теореме Бернштейна, для того, чтобы $f \in A(D_R)$, необходимо и достаточно, чтобы

$$\overline{\lim_{n\to\infty}} \sqrt[n]{P_n(f)} \leqslant \frac{1}{R};$$

тем самым

$$\sup_{f\in A(D_R)} \left[\overline{\lim_{n\to\infty}} \sqrt[n]{P_n(f)}\right] = \frac{1}{R}.$$

В. Д. Ерохин [15] показал, что существует функция $f \in A(D_R)$, для которой

$$\overline{\lim_{n\to\infty}} \sqrt[n]{R_n(f)} = \frac{1}{R};$$

следовательно,

$$\sup_{f\in A(D_R)} \left[\overline{\lim_{n\to\infty}} \sqrt[n]{R_n(f)}\right] = \frac{1}{R}.$$

Таким образом, переход от многочленов к рациональным функциям не дает существенной выгоды в скорости аппроксимации всего класса функций, аналитических в области D_R (ср. п. 4, скорость аппроксимации в классе $L(A)$). Однако если вместо канонической области D_R рассмотреть произвольную область $D \supset \Delta$, то это уже не так.

Основные результаты в этом направлении получены Дж. Уолшем в 30-х годах; мы сформулируем здесь утверждение, являющееся лишь весьма специальным следствием важных теорем Уолша об интерполяции аналитических функций рациональными (см. [4], гл. VIII, в частности § 8.7).

Пусть D — односвязная область (в $\overline{\mathbf{C}}$), содержащая отрезок Δ, $\rho = \rho(D \setminus \Delta)$ — модуль двухсвязной области $D \setminus \Delta$. Тогда для любой функции $f \in A(D)$

$$\overline{\lim_{n\to\infty}} \sqrt[n]{R_n(f)} \leqslant \frac{1}{\rho}.$$

Очевидно, что в этом случае

$$\sup_{f\in A(D)} \left[\overline{\lim_{n\to\infty}} \sqrt[n]{P_n(f)}\right] = \frac{1}{\rho^*},$$

где $\rho^* = \sup\{R : D_R \subset D\}$. Если $D \neq D_\rho$, то $\rho^* < \rho$ (для любого $\varepsilon > 0$ существует односвязная область $D \supset \Delta$, такая, что $\frac{1}{\rho} < \varepsilon$, в то время как $\frac{1}{\rho^*} > 1 - \varepsilon$).

Для произвольного континуума E и односвязной области $D \supset E$ теорема Уолша справедлива в той же формулировке (ρ — модуль двухсвязной компоненты $D \setminus E$, один из граничных континуумов которой совпадает с границей D, другой принадлежит E).

Недавно В. М. Тихомиров показал, что число ρ в теореме Уолша нельзя заменить бо́льшим числом; тем самым

$$\sup_{f \in A(D)} [\overline{\lim_{n\to\infty}} \sqrt[n]{R_n(f)}] = \frac{1}{\rho}.$$

Сформулируем одну нерешеиную задачу, связанную с наилучшими приближениями аналитических функций. Пусть $f \in A(D)$, $D \supset \Delta$. Доказать (или опровергнуть) следующее утверждение: существует [1]) последовательность $r_n^*(z)$, $n = 1, 2, \ldots$, рациональных функций *наилучшего приближения* f на Δ, таблица полюсов $\{\alpha_{nk}^*\}$ которой не имеет предельных точек в области D:

$$\{\alpha_{nk}^*\}' \cap D = \varnothing.$$

Более общая задача относится к случаю произвольного континуума $E \subset \mathbf{C}$ и области D, $D \cap E \neq \varnothing$; формулируется она совершенно аналогично.

9. Любая последовательность многочленов $p_n(z)$, $n = 1, 2, \ldots$, степень $p_n(z) \leqslant n$, сходящаяся на Δ (к функции f) со скоростью, характеризуемой условием

$$\overline{\lim_{n\to\infty}} [\max_{x\in\Delta} |f(x) - p_n(x)|]^{\frac{1}{n}} \leqslant q < 1,$$

с необходимостью равномерно сходится внутри области D_R, $R = \frac{1}{q}$ (и, следовательно, $f \in A(D_R)$). То же явление (Уолш называет его *сверхсходимостью* последовательности $p_n(z)$) имеет место для произвольного континуума E (см., например, [4]). Обратная теорема о наилучших приближениях аналитических функций многочленами является непосредственным следствием этого утверждения.

[1]) Рациональная функция $r_n^*(z)$ наилучшего приближения f на Δ, вообще говоря, не единственна — даже для вещественной функции $f(x)$, $x \in \Delta$ (наилучшее приближение ищется в классе рациональных функций с *комплексными* коэффициентами).

Возникает вопрос, в какой мере это утверждение распространяется на последовательности рациональных функций с произвольными полюсами; в частности, при какой скорости сходимости последовательности $r_n(z)$, $n = 1, 2, \ldots$, на Δ можно утверждать, что эта последовательность с необходимостью сходится по крайней мере в одной точке вне Δ (какова бы ни была ее таблица полюсов).

Легко показать, что скорость порядка геометрической прогрессии для этого уже недостаточна. Более того (мы приводим сначала результаты, относящиеся к случаю $f(x) = 0$, $x \in \Delta$), *каково бы ни было* q, $0 < q < 1$, *существует последовательность* $r_n(z)$, $n = 1, 2, \ldots$, *сходящаяся на* Δ *со скоростью, характеризуемой условием*

$$\max_{x\in\Delta} |r_n(x)| \leqslant \left(\frac{q}{\sqrt{n}}\right)^n, \quad n = 1, 2, \ldots,$$

и расходящаяся в каждой точке плоскости, не принадлежащей Δ.

Однако если последовательность $r_n(z)$, $n = 1, 2, \ldots$, сходится (к нулю) на Δ с большей скоростью, то эта последовательность сходится на всей комплексной плоскости, за исключением достаточно «редких» множеств. Точнее, справедливы следующие утверждения:

Если

$$\max_{x\in\Delta} |r_n(x)| \leqslant n^{-\left(\frac{1}{2}+\delta\right)n}, \quad \delta > 0, \quad n = 1, 2, \ldots,$$

то $r_n(z)$, $n = 1, 2, \ldots$, *сходится* (к нулю) *почти всюду на плоскости относительно плоской меры Лебега.*

Если

$$\max_{x\in\Delta} |r_n(x)| \leqslant n^{-(\lambda+\delta)n}, \quad \delta > 0, \quad n = 1, 2, \ldots,$$

то $r_n(z)$, $n = 1, 2, \ldots$, *сходится* (к нулю) *почти всюду на плоскости относительно* h*-меры Хаусдорфа, соответствующей функции* $h(r) = r^{\frac{1}{\lambda}}$.

Последние неравенства с $\delta = 0$ (при любом λ) совместимы с расходимостью $r_n(z)$ на множестве бесконечной h-меры, $h(r) = r^{\frac{1}{\lambda}}$.

Аналогичные утверждения справедливы и в случае произвольной предельной функции $f(x)$, $x \in \Delta$. Например:

Если

$$\max_{x\in\Delta} |f(x) - r_n(x)| \leqslant n^{-An}, \quad n = 1, 2, \ldots,$$

то $r_n(z)$, $n = 1, 2, \ldots$, *сходится всюду в* $\mathbf{C}$, *за исключением множества* F, h*-мера Хаусдорфа которого равна нулю при любой функции* $h(r) = r^a$, $a > \frac{2}{A}$ (*размерность множества* F *не превышает* $\frac{2}{A}$).

Можно изучить характер сходимости $r_n(z)$ на $\mathbf{C} \setminus F$ (сходимость, равномерная на расширяющейся системе компактных множеств, исчерпывающих $\mathbf{C} \setminus F$) и свойства предельной функции $\tilde{f}(z)$, $z \in \mathbf{C} \setminus F$ (эти свойства во многом аналогичны свойствам функций, моногенных по Борелю; в частности, $\tilde{f}$ является в некотором смысле квазианалитическим продолжением функции f).

Вопрос о сверхсходимости последовательностей $r_n(z)$ возник в связи с изучением функций класса $R(\Delta)$ (см. п. 6). Представляется вероятным, что свойство единственности в классе $R(\Delta)$ связано с тем, что любая функция $f \in R(\Delta)$ допускает продолжение $\tilde{f}$ на достаточно массивное множество в окрестности Δ, на котором свойство единственности справедливо в классе всех функций, допускающих равномерное приближение рациональными функциями. Приведенные выше результаты показывают, что построить такое продолжение исходя из *произвольной* последовательности $r_n(z)$, сходящейся на Δ к f со скоростью геометрической прогрессии (и даже с существенно большей скоростью), не удастся; в этом существенное отличие от случая многочленов. В связи с этим возникает вопрос об изучении сходимости в дополнении к Δ последовательностей $r_n^*(z)$ рациональных функций *наилучшего приближения* f на Δ, или каких-либо их подпоследовательностей. Если для любой $f \in R(\Delta)$ существует последовательность $r_n^*(z)$ рациональных функций наилучшего приближения, такая, что $r_n^*(z)$ (или $r_{n_k}^*(z)$) сходится на достаточно массивном множестве, содержащем Δ, то продолжение $\tilde{f}$ можно было бы определить с помощью такой последовательности. . . .

Заметим, что вообще переход к лакунарной подпоследовательности позволяет строить продолжение функции f при меньших требованиях на скорость сходимости $r_n(x)$, $x \in \Delta$, чем указанные выше. Следующее утверждение интересно сравнить с предыдущей теоремой:

Если

$$\max_{x \in \Delta} |f(x) - r_n(x)| \leqslant n^{-\delta n}, \quad \delta > 0, \quad n = 1, 2, \ldots,$$

то, какова бы ни была последовательность $n_1, n_2, \ldots, \frac{n_{k+1}}{n_k} \geqslant Q > 1$, *последовательность* $r_{n_k}(z)$, $k = 1, 2, \ldots$, *сходится всюду в* $\mathbf{C}$, *за исключением множества* F, *h-мера Хаусдорфа которого равна нулю при любой функции* $h(r) = r^a$, $a > 0$ (*размерность* F *равна нулю*).

Вопрос о поведении рациональных функций наилучшего приближения функции $f \in R(\Delta)$, а также задача, сформулированная в конце п. 8, связаны с общим вопросом о расположении полюсов $\{\alpha_{nk}^*\}$ последовательностей рациональных функций наилучшего приближения для заданной непрерывной функции $f(x)$, $x \in \Delta$ (подчеркнем, что заранее на расположение полюсов не накладывается никаких ограничений). Можно предположить, что таблица $\{\alpha_{nk}^*\}$ (или множество $\{\alpha_{nk}^*\}'$) определяет в некотором смысле множество «особенностей» функции f в комплексной плоскости. Вопрос об уточнении этого высказывания представляется весьма сложным; здесь возникает целый ряд интересных конкретных задач. Например, если полюсы $\{\alpha_{nk}^*\}$ последовательности рациональных функций наилучшего приближения f на Δ не имеют предельных точек в фиксированной (пусть односвязной) области $D \supset \Delta$, то можно ли утверждать, что функция f аналитична в этой области? Если да, то как интерпретировать тот факт, что $\{\alpha_{nk}^*\}$ не имеют предельных точек в некоторой области D, не пересекающейся с отрезком Δ? Можно ли утверждать (в любом из этих случаев), что некоторая, заведомо не любая, последовательность $r_n^*(z)$ рациональных функций наилучшего приближения f на Δ (или какая-либо ее подпоследовательность) обязана равномерно сходиться внутри D? Если такая сходимость имеет место, то предельная функция $\tilde{f}(z)$, $z \in D$ (с необходимостью аналитическая в D), в случае $D \supset \Delta$ является аналитическим продолжением функции $f(z)$, $z \in \Delta$. Интересно было бы охарактеризовать внутреннюю связь между $\tilde{f}$ и f в случае $D \cap \Delta = \varnothing$ (в терминах, не связанных с $r_n^*(z)$; тот факт, что функции $f(z)$, $z \in \Delta$, и $\tilde{f}(z)$, $z \in D$, являются пределами одной и той же последовательности рациональных функций, не устанавливает — в случае $D \cap \Delta = \varnothing$ — между ними никакой внутренней связи: любые функции f_1 и f_2, одна из которых непрерывна на Δ, другая — аналитична в D, могут быть представлены в виде предела одной и той же последовательности $r_n(z)$, равномерно сходящейся на Δ и внутри D к f_1 и f_2 соответственно).

Нетрудно построить функцию $f(z)$, $z \in \Delta$, так, чтобы для любой последовательности $r_n^*(z)$, $n = 1, 2, \ldots$, ее рациональных функций наилучшего приближения имело место соотношение

$$\{\alpha_{nk}^*\}' = \mathbf{C}.$$

Например, таковой является функция

$$f(z) = \sum_{n=1}^{\infty} \frac{A_n d_n}{z - \alpha_n}, \quad z \in \Delta; \quad \alpha_i \neq \alpha_j, \quad i \neq j,$$

где $\alpha_n \notin \Delta$, $\{\alpha_n\}' = \mathbf{C}$, d_n — расстояние от α_n до Δ, $A_n \neq 0$, $n = 1, 2, \ldots$, $\lim \sqrt[n]{|A_n|} = 0$. Заметим, что наилучшие приближения $R_n(f)$ этой функции на отрезке Δ в классе всех рациональных функций, полюсам которых запрещено попадать в какой-либо круг произвольно малого радиуса, стремятся к нулю существенно медленнее, чем ее наилучшие приближения $R_n(f)$ в классе

всех рациональных функций. В известном смысле (не сводящемся к тому, что f не является аналитической ни в одной точке своей области определения Δ) множество «особенностей» f всюду плотно на комплексной плоскости; особенности ряда являются «особенностями» его суммы, рассматриваемой только на Δ.

Всюду в этом пункте вместо отрезка Δ можно было рассматривать произвольный континуум E.

Результаты о сверхсходимости, приведенные в этом пункте, в несколько ином виде сформулированы в [16]. Различные иллюстрации явления сверхсходимости в областях, не пересекающихся с E, и квазианалитического продолжения функций, определяемых равномерно сходящимися последовательностями рациональных функций, содержатся в заметках [17], [18], [14]. Оценки, приведенные в п. 7, позволяют получить более общие результаты в этом направлении.

10. В связи с результатами, приведенными в п. 3 и 6, А. Н. Колмогоров заметил, что теоремы о наилучших приближениях рациональными функциями естественно формулировать, не ограничиваясь классом непрерывных функций. Поскольку на полюсы аппроксимирующих рациональных функций не накладывается никаких ограничений, естественно допустить их и на отрезок Δ, накладывая ограничения на скорость приближения лишь вне множеств произвольно малой меры. Приведем некоторые результаты, полученные автором на этом пути.

Пусть f — измеримая функция, определенная и конечная почти всюду на Δ. Определением измеримой функции f может служить известное (C)-свойство Н. Н. Лузина: для любого $\varepsilon > 0$ найдется совершенное множество $P_\varepsilon \subset \Delta$, такое, что $\operatorname{mes}(\Delta \setminus P_\varepsilon) < \varepsilon$ и $f \mid_{P_\varepsilon}$ непрерывна на P_ε. Имея в виду теорему Вейерштрасса, требование непрерывности $f \mid_{P_\varepsilon}$ можно заменить любым из условий: $\lim\limits_{n\to\infty} P_n(f, P_\varepsilon) = 0$ или $\lim\limits_{n\to\infty} R_n(f, P_\varepsilon) = 0$.

С помощью дополнительных ограничений, с одной стороны, на дифференциальные свойства $f \mid_{P_\varepsilon}$, с другой — на скорость стремления к нулю наилучших приближений $P_n(f, P_\varepsilon)$ и $R_n(f, P_\varepsilon)$, выделим некоторые подклассы класса всех измеримых функций на Δ. Поскольку в первом случае эти подклассы определяются дифференциальными свойствами, во втором — конструктивными свойствами функций f, соотношения включения между ними носят характер прямых и обратных теорем о наилучших приближениях.

Обозначим через $D_A(\Delta)$ класс функций $f(x)$, $x \in \Delta$, обладающих тем свойством, что для любого $\varepsilon > 0$ найдется совершенное множество $P_\varepsilon \subset \Delta$, такое, что $\operatorname{mes}(\Delta \setminus P_\varepsilon) < \varepsilon$ и $f \mid_{P_\varepsilon} \in L(A)$ (см. п. 3).

С другой стороны, через $R_A(\Delta)$ обозначим класс функций $f(x)$, $x \in \Delta$, таких, что для любого $\varepsilon > 0$ существует совершенное $P_\varepsilon \subset \Delta$, $\operatorname{mes}(\Delta \setminus P_\varepsilon) < \varepsilon$, на котором

$$R_n(f, P_\varepsilon) < \frac{C}{n^A}, \quad n = 1, 2, \ldots,$$

где $C = C(f, \varepsilon)$ не зависит от n. Если вместо последнего условия выполнено условие

$$R_n(f, P_\varepsilon) < \frac{C}{n^{A+\delta}},$$

где C и $\delta > 0$ зависят от f и ε, но не зависят от n, то будем говорить, что $f \in R_{A+0}(\Delta)$.

Аналогично, исходя из наилучших приближений многочленами, определяются классы $P_A(\Delta)$ и $P_{A+0}(\Delta)$.

Справедливы следующие включения:

$$P_{A+0}(\Delta) \subset R_{A+0}(\Delta) \subset D_A(\Delta) \subset P_A(\Delta) \subset R_A(\Delta).$$

В этой цепочке включений содержатся прямые и обратные теоремы о наилучших приближениях измеримых функций как многочленами

$$P_{A+0}(\Delta) \subset D_A(\Delta) \subset P_A(\Delta),$$

так и рациональными функциями

$$R_{A+0}(\Delta) \subset D_A(\Delta) \subset R_A(\Delta).$$

Теоремы смыкаются с точностью до произвольно малого $\delta > 0$.

Отвлекаясь от дифференциальных свойств функций, интересно отметить включение

$$R_{A+0}(\Delta) \subset P_A(\Delta);$$

отсюда следует, что любая измеримая функция, допускающая приближение порядка $\frac{1}{n^{A+\delta}}$, $\delta > 0$, в классе всех рациональных функций, допускает приближение порядка $\frac{1}{n^A}$ многочленами. При этом существенно, что приближение осуществляется на «внутренних» множествах.

В предельном случае классы тождественны:

$$P_\infty(\Delta) = R_\infty(\Delta) = D_\infty(\Delta)$$

(класс с бесконечным индексом определяется как пересечение соответствующих классов с натуральными индексами).

Если, однако, в конструктивном определении классов измеримых функций на скорость стремления к нулю наилучших приближений наложить более сильные ограничения, то, несмотря на переход

к «внутренним» множествам, разница между многочленами и рациональными функциями оказывается существенной.

В самом деле, если функция f такова, что для любого $\varepsilon > 0$ существует совершенное множество $P_\varepsilon \subset \Delta$, удовлетворяющее условиям $\operatorname{mes}(\Delta \setminus P_\varepsilon) < \varepsilon$ и

$$\overline{\lim_{n\to\infty}} \sqrt[n]{P_n(f, P_\varepsilon)} \leqslant q < 1,$$

то f обязана быть аналитической на отрезке Δ (точнее, f эквивалентна некоторой функции, аналитической на Δ). С другой стороны, любая скорость стремления к нулю $R_n(f)$ (тем более $R_n(f, P_\varepsilon)$) совместима с неаналитичностью функции f в каждой точке отрезка Δ.

Тем больший интерес представляет изучение класса измеримых функций, наилучшие приближения которых рациональными функциями — на «внутренних» множествах P_ε — стремятся к нулю со скоростью геометрической прогрессии. Этот класс обладает рядом свойств, аналогичных важнейшим свойствам класса аналитических функций.

Обозначим через $\widetilde{R}(\Delta)$ класс измеримых функций $f(x)$, $x \in \Delta$, таких, что для любого $\varepsilon > 0$ существует совершенное множество $P_\varepsilon \subset \Delta$, удовлетворяющее условиям: $\operatorname{mes}(\Delta \setminus P_\varepsilon) < \varepsilon$ и

$$\overline{\lim_{n\to\infty}} \sqrt[n]{R_n(f, P_\varepsilon)} \leqslant q(f) < 1.$$

Этот класс является расширением введенного в п. 6 класса непрерывных функций $R(\Delta)$. В частности, произвольная функция вида

$$\sum_{n=1}^{\infty} \frac{A_n}{x - \alpha_n}, \qquad x \in \Delta,$$

где $\alpha_n \in \mathbf{C}$ (α_n могут попадать и на отрезок Δ), и

$$\overline{\lim_{n\to\infty}} \sqrt[n]{|A_n|} < 1,$$

принадлежит классу $\widetilde{R}(\Delta)$; конечно, далеко не любая такая функция непрерывна (эквивалентна непрерывной функции) на отрезке Δ.

Класс $\widetilde{R}(\Delta)$ замкнут относительно операции асимптотического дифференцирования: *если* $f \in \widetilde{R}(\Delta)$, *то почти всюду на* Δ *существует асимптотическая производная* $f^{[1]}(x)$, *причем* $f^{[1]} \in \widetilde{R}(\Delta)$. *Следовательно, любая* $f \in \widetilde{R}(\Delta)$ *почти всюду на* Δ *бесконечно (асимптотически) дифференцируема.*

Другим важным свойством класса $\widetilde{R}(\Delta)$ является свойство единственности: *если две функции,* f *и* g, *принадлежат* $\widetilde{R}(\Delta)$ *и совпадают на множестве* $e \subset \Delta$ *положительной меры, то эти функции совпадают почти всюду на* Δ *(эквивалентны).*

Заметим, наконец, что класс $\widetilde{R}(\Delta)$ является *полем*; не только линейные операции и умножение, но и деление (не на тождественный нуль) функций класса $\widetilde{R}(\Delta)$ не выводит за его пределы.

Если $\{n_k\}$ — произвольная последовательность натуральных чисел, то класс $\widetilde{R}_{\{n_k\}}(\Delta)$, определяемый условием

$$\overline{\lim_{k\to\infty}} \sqrt[n_k]{R_{n_k}(f, P_\varepsilon)} \leqslant q(f) < 1,$$

также обладает свойством единственности. Квазианалитические классы $\widetilde{R}_{\{n_k\}}(\Delta)$ интересны тем, что они позволяют полностью отделить свойства непрерывности и гладкости (даже в более слабом определении, чем обычно) от основного свойства аналитических функций — свойства единственности.

Отметим одно непосредственное следствие свойства единственности в классе $\widetilde{R}(\Delta)$. Пусть $\overline{D}$ — замкнутая область (D — множество внутренних точек $\overline{D}$), $\{\alpha_n\}$ — произвольное множество точек, лежащих в дополнении к $\overline{D}$ и таких, что $\{\alpha_n\}' = \partial D = \overline{D} \setminus D$,

$$f(z) = \sum_{n=1}^{\infty} \frac{A_n}{z - \alpha_n}, \qquad z \in D.$$

Если

$$\overline{\lim_{n\to\infty}} \sqrt[n]{|A_n|} < 1, \qquad A_n \neq 0, \qquad n = 1, 2, \ldots,$$

то каждая точка $z \in \partial D$ *является особой точкой аналитической функции* $f(z)$, $z \in D$.

Это утверждение нетрудно доказать и непосредственно, основываясь на оценках типа max-min для произвольных континуумов (см. п. 7). Несколько сложнее доказать тот же факт для произвольных $\alpha_n \in \mathbf{C} \setminus \overline{D}$ (без требования $\{\alpha_n\}' = \partial D$). Сформулируем одно общее утверждение в этом направлении.

Пусть $\{\alpha_n\}$, $n = 1, 2, \ldots$, — произвольное множество точек плоскости (в частности, возможен случай $\{\alpha_n\}' = \mathbf{C}$), E — континуум и f — функция, аналитическая в области $D \supset E$. *Если*

$$\sum_{n=1}^{\infty} \frac{A_n}{z - \alpha_n} = f(z), \qquad z \in E; \qquad \overline{\lim_{n\to\infty}} \sqrt[n]{|A_n|} < 1,$$

то $A_n = 0$ *для* $n \in N_D = \{n: \alpha_n \in D\}$; в частности, *если* $f(z) = 0$, $z \in E$, *то все* $A_n = 0$, $n = 1, 2, \ldots$. В случае $f(z) = 0$, $z \in E$, соответствующая теорема единственности была установлена Карлеманом [19] при условии

$$\sqrt[n]{|A_n|} < \frac{1}{n^{n+\delta}}, \qquad \delta > 0,$$

где $a > 0$ зависит от расположения E относительно $\{\alpha_n\}$ (требуется, чтобы континуум E лежал вне объединения кругов $|z - \alpha_n| < \varepsilon \cdot n^{-a}$ при некотором $\varepsilon > 0$).

Теоремы о приближениях измеримых функций, формулируемые в дифференциальных терминах, без каких-либо изменений переносятся на произвольные измеримые подмножества вещественной прямой; для справедливости теорем единственности в этом случае надо потребовать, чтобы

$$\lim_{n\to\infty} \sqrt[n]{R_n(f, P_\varepsilon)} = 0, \quad n = 1, 2, \ldots \text{ или } n = n_1, n_2, \ldots .$$

Наилучшие приближения измеримых функций рациональными рассматривались в работах [20], [21] (см. также [22]); приведенные здесь теоремы единственности усиливают известные ранее.

Математический институт им. В. А. Стеклова,
Москва, СССР

ЛИТЕРАТУРА

[1] Чебышев П. Л., Полное собрание сочинений, т. II, 151-235.
[2] Золотарев Е. И., Полное собрание сочинений, т. II, 1-59.
[3] Ахиезер Н. И., Лекции по теории аппроксимации, издание 2-е, М., 1965.
[4] Уолш Дж. Л., Интерполяция и аппроксимация рациональными функциями в комплексной области, М., 1961.
[5] Мергелян С. Н., Некоторые вопросы конструктивной теории функций, Труды Матем. ин-та им. В. А. Стеклова, **37**, 1951.
[6] Гончар А. А., *ДАН СССР*, **100**, № 2 (1955), 205-208.
[7] Долженко Е. П., *Известия АН СССР*, сер. матем., **26**, № 5 (1962), 641-652.
[8] Гончар А. А., *Известия АН СССР*, сер. матем., **25**, № 3, (1961), 347-356.
[9] Долженко Е. П., *Матем. сб.*, **56**, № 4 (1962), 403-434.
[10] Newman D. I., *Michigan Math. J.*, **11**, № 1 (1964), 11-14.
[11] Гончар А. А., *Матем. сб.*, **72**, № 3 (1967), 489-503.
[12] Turan P., сборник «Современные проблемы теории аналитических функций», изд-во «Наука», М., 1966, 296-299.
[13] Гончар А. А., *Матем. сб.*, **73**, № 4 (1967), 630-638.
[14] Гончар А. А., *ДАН СССР*, **166**, № 5 (1966), 1028-1031.
[15] Ерохин В. Д., *ДАН СССР*, **128**, № 1 (1959), 29-32.
[16] Гончар А. А., *ДАН СССР*, **143**, № 6 (1962), 1246-1249.
[17] Гончар А. А., *ДАН СССР*, **141**, № 5 (1961), 1019-1022.
[18] Гончар А. А., *ДАН СССР*, **141**, № 6 (1961), 1287-1289.
[19] Carleman T., *C. R. Acad. Sci. Paris*, **174** (1922), 588-591.
[20] Гончар А. А., *ДАН СССР*, **111**, № 5 (1956), 930-932.
[21] Гончар А. А., *Известия ВУЗов*, Математика, **6** (19), (1960), 74-81.
[22] Мергелян С. Н., [4], Приложение, 461-496.

5

Функциональный анализ
Functional analysis
Analyse fonctionnelle
Funktionalanalysis

ESPACE DUAL D'UNE ALGÈBRE, OU D'UN GROUPE LOCALEMENT COMPACT

JACQUES DIXMIER

1. Introduction

Soit G un groupe commutatif localement compact. On appelle *caractère* de G toute fonction continue complexe π sur G telle que $\pi(gg') = \pi(g)\pi(g')$ quels que soient $g, g' \in G$; si $|\pi(g)| = 1$ pour *tout* $g \in G$, le caractère est dit *unitaire*. L'ensemble des caractères unitaires de G se note $\hat{G}$; on le munit de la topologie *de la convergence compacte*, d'où un espace localement compact. (En fait, $\hat{G}$ est de manière naturelle un groupe localement compact, mais nous laisserons de côté ici cette structure de groupe.) Etant donnée une fonction plus ou moins arbitraire sur G, on cherche à l'écrire comme combinaison linéaire (en général intégrale) de caractères, et si possible de caractères unitaires. C'est là un problème typique de synthèse harmonique. Toute étude de synthèse harmonique sur G nécessite la connaissance détaillée de l'espace $\hat{G}$.

Choisissons une mesure de Haar sur G, et soit $L^1(G)$ l'algèbre de Banach (convolutive) des fonctions intégrables sur G. On appelle caractère d'une algèbre de Banach commutative un homomorphisme non nul de cette algèbre dans $\mathbf{C}$ (un tel caractère est de norme $\leqslant 1$). Tout $\pi \in \hat{G}$ définit un caractère de $L^1(G)$ par la formule

$$\pi'(f) = \int_G \pi(g) f(g)\, dg \tag{1}$$

et on obtient ainsi une correspondance bijective entre caractères unitaires de G et caractères de $L^1(G)$. La topologie de $\hat{G}$ correspond dans cette bijection à la topologie *faible* sur l'ensemble des caractères de $L^1(G)$.

Observons qu'il existe sur $L^1(G)$ une involution $f \to f^*$, définie par $\overline{f^*(g)} = f(g^{-1})$, et qu'un caractère π de $L^1(G)$ est automatiquement hermitien, c'est-à-dire tel que $\overline{\pi(f)} = \pi(f^*)$ pour tout $f \in L^1(G)$.

2. L'espace dual d'un groupe localement compact

Dans toute la suite, G désigne un groupe localement compact. Il semble raisonnable de définir l'ensemble $\hat{G}$ comme l'ensemble des *représentations unitaires topologiquement irréductibles* de G, deux représentations unitairement équivalentes étant identifiées. Nous allons maintenant définir une topologie sur $\hat{G}$.

Si $\pi \in \hat{G}$, nous noterons H_π l'espace hilbertien où opère π. Nous appellerons fonction de type positif associée à π toute fonction de la forme $g \to (\pi(g)\,\xi \mid \xi)$ sur G, où $\xi \in H_\pi$. Si $S \subset \hat{G}$, nous appellerons fonction de type positif associée à S une fonction de type positif associée à un élément de S. Ceci posé, nous écrirons $\pi \in \overline{S}$ si *l'une* des fonctions de type positif non nulles associées à π est limite uniforme sur tout compact de fonctions de type positif associées à S, ou, ce qui revient au même, si *toute* fonction de type positif associée à π est limite uniforme sur tout compact de fonctions de type positif associées à S. Ceci définit une opération d'adhérence dans $\hat{G}$, d'où une topologie sur $\hat{G}$ ([4], [10]). Quand G est commutatif, on retrouve la topologie du § 1. Dans le cas général, cette topologie n'est pas toujours séparée, ce qui suscite la méfiance. Nous espérons cependant montrer que cette topologie est «raisonnable». Nous allons d'abord généraliser la situation.

3. Passage aux C^*-algèbres

Soit A une algèbre de Banach involutive. On appelle représentation de A dans un espace hilbertien H une application linéaire π de A dans $\mathcal{L}(H)$ (algèbre involutive des endomorphismes continus de H) telle que $\pi(aa') = \pi(a)\,\pi(a')$, $\pi(a^*) = \pi(a)^*$ quels que soient $a, a' \in A$ (une telle représentation est automatiquement de norme $\leqslant 1$). On note $\hat{A}$ l'ensemble des *représentations topologiquement irréductibles de A dans des espaces hilbertiens*, deux représentations unitairement équivalentes étant identifiées.

On appelle C^*-algèbre une algèbre de Banach involutive telle que $\| x^*x \| = \| x \|^2$ pour tout élément x de l'algèbre. Reprenons l'algèbre A précédente. Il existe une C^*-algèbre A' et un homomorphisme (automatiquement continu) φ de A dans A' tels que l'application $\pi \to \pi \circ \varphi$ mette en correspondance bijective les représentations de A et celles de A' dans des espaces hilbertiens. En outre, A' et φ sont définis à un isomorphisme près. On dit que A' est la *C^*-algèbre enveloppante de A*. Les ensembles $\hat{A}$ et $\hat{A}'$ s'identifient. L'étude de $\hat{A}$ est ainsi ramenée au cas où A est une C^*-algèbre.

Choisissons une mesure de Haar à gauche sur G. On peut alors former l'algèbre de Banach involutive $L^1(G)$. Toute représentation unitaire π de G définit une représentation π' de $L^1(G)$ dans H_π par la formule (1) (où les deux membres représentent maintenant des endomorphismes continus de H_π). On obtient ainsi une bijection de $\hat{G}$ sur $L^1(G)\hat{}$. La C^*-algèbre enveloppante de $L^1(G)$ se note $C^*(G)$ et s'appelle la *C^*-algèbre de G*. D'après ce qui précède, il existe une bijection canonique de $\hat{G}$ sur $C^*(G)\hat{}$.

Dans toute la suite, A désigne une C^*-algèbre. L'étude de $\hat{G}$ est un cas particulier de celle de $\hat{A}$.

4. Le spectre d'une C^*-algèbre

Si $\pi \in \hat{A}$, appelons coefficient de π toute forme linéaire $a \to (\pi(a)\,\xi \mid \xi')$ sur A, où $\xi, \xi' \in H_\pi$. Si $S \subset \hat{A}$, appelons coefficient de S tout coefficient d'un élément de S. Ceci posé, nous écrirons $\pi \in \overline{S}$ si *l'un* des coefficients de π est limite faible de coefficients de S, ou, ce qui revient au même, si *tout* coefficient de π est limite faible de coefficients de S. Ceci définit l'espace topologique $\hat{A}$, appelé *spectre de A* [4]. En particulier, si $A = C^*(G)$, on retrouve la topologie introduite plus haut sur $\hat{G}$.

La topologie de $\hat{A}$ peut être définie par d'autres procédés équivalents :

A) Notons d'abord les propriétés suivantes : 1) si $\pi \in \hat{A}$, π est *algébriquement* irréductible ; 2) soient $\pi_1, \pi_2 \in \hat{A}$; si π_1 et π_2 sont équivalentes en tant que représentations de A dans les espaces vectoriels H_{π_1}, H_{π_2} (sans tenir compte de leurs structures hilbertiennes), alors π_1 et π_2 sont unitairement équivalentes ; 3) toute représentation algébriquement irréductible de A dans un espace vectoriel complexe est algébriquement équivalente à un élément de $\hat{A}$ [13]. Ainsi, $\hat{A}$ peut être considéré comme l'ensemble des classes de représentations irréductibles de A dans des espaces vectoriels complexes (tout étant pris en un sens purement algébrique). Or, si B est une algèbre complexe, l'ensemble $\hat{B}$ des classes de représentations algébriquement irréductibles de B dans des espaces vectoriels complexes peut être muni de la *topologie de Jacobson* [12] (la partie fermée la plus générale de $\hat{B}$ étant l'ensemble des $\pi \in \hat{B}$ qui s'annulent sur un sous-ensemble donné de B). Ceci posé, la topologie définie plus haut sur $\hat{A}$ n'est autre que la topologie de Jacobson [4].

B) Limitons-nous ici, pour simplifier, au sous-ensemble $\hat{A}_n$ de $\hat{A}$ formé des classes de représentations irréductibles dont la dimension hilbertienne est un cardinal fixé n. Soit H_n un espace hilbertien de dimension n. Soit $\mathrm{Irr}_n(A)$ l'ensemble des représentations irréductibles de A dans H_n ; on munit $\mathrm{Irr}_n(A)$ de la *topologie de la convergence*

simple forte ; autrement dit, $\pi_\lambda \to \pi$ dans $\mathrm{Irr}_n(A)$ si $\| \pi_\lambda(a)\xi - \pi(a)\xi \| \to 0$ pour tout $a \in A$ et tout $\xi \in H_n$. Si, à tout élément de $\mathrm{Irr}_n(A)$, on fait correspondre sa classe, on obtient une application de $\mathrm{Irr}_n(A)$ sur $\hat{A}_n$. Ceci posé, la topologie sur $\hat{A}_n$ induite par celle de $\hat{A}$ est la topologie *quotient* de celle de $\mathrm{Irr}_n(A)$ [5]. On observera le caractère concret de cette définition.

C) Supposons qu'il existe une sous-algèbre involutive A' de A, dense dans A, avec la propriété suivante : pour tout $x \in A'$, le rang de $\pi(x)$ reste borné quand π parcourt $\hat{A}$ (c'est le cas si $A = C^*(G)$ avec G linéaire semi-simple). Alors il y a des relations étroites entre la topologie de $\hat{A}$ et la notion de trace. Par exemple : soit $(\pi_1, \pi_2, \ldots)$ une suite d'éléments de $\hat{A}$, et $\nu_1, \ldots, \nu_r \in \hat{A}$; on suppose que, pour tout $x \in A'$, on a

$$\mathrm{tr}(\pi_n(x)) \to \mathrm{tr}(\nu_1(x)) + \ldots + \mathrm{tr}(\nu_r(x)) \quad \text{quand } n \to +\infty.$$

Alors les seules limites de la suite $(\pi_1, \pi_2, \ldots)$ dans $\hat{A}$ sont $\nu_1, \ldots, \nu_r$ [4].

5. Propriétés de la topologie du spectre

L'espace $\hat{A}$ est *localement quasi-compact* (c'est-à-dire que tout point de $\hat{A}$ admet un système fondamental de voisinages quasi-compacts) [6].

L'espace $\hat{A}$ est un *espace de Baire*.

Si A est séparable (par exemple si $A = C^*(G)$ avec G séparable) $\hat{A}$ est *séparable*.

Pour pouvoir énoncer d'autres propriétés, il faut nous limiter à des C^*-algèbres particulières (ou à des groupes particuliers) que nous allons maintenant définir. On dit que A est CCR (ou *liminaire*) si, pour tout $\pi \in \hat{A}$, $\pi(A)$ est l'ensemble des opérateurs compacts de H_π [14]. On dit que G est CCR (ou liminaire) si $C^*(G)$ est liminaire. Par exemple, si G est un groupe de Lie connexe réel semi-simple [11] ou nilpotent [3], [15], G est liminaire. On dit que A est GCR (ou *postliminaire*) si toute C^*-algèbre quotient non nulle de A possède un idéal bilatère fermé liminaire non nul [14]. Les C^*-algèbres liminaires sont postliminaires, mais la réciproque n'est pas vraie. Si A est séparable, les conditions suivantes sont équivalentes : 1) A est postliminaire ; 2) toute représentation factorielle de A est de type I (ceci veut dire par exemple que, si π est une représentation de A dans un espace hilbertien telle que l'adhérence faible de $\pi(A)$ ait son centre réduit aux scalaires, alors π est somme de représentations irréductibles équivalentes) ; 3) si $\pi \in \hat{A}$, $\pi(A)$ contient l'ensemble des opérateurs compacts de H_π ; 4) si $\pi, \pi' \in \hat{A}$ et si $\mathrm{Ker}\,\pi = \mathrm{Ker}\,\pi'$, alors $\pi = \pi'$ [9]. Un groupe G est dit GCR (ou postliminaire) si $C^*(G)$ est

postliminaire. Par exemple, si G est un groupe algébrique linéaire réel, G est postliminaire [2]. Beaucoup de groupes de Lie réels connexes résolubles sont postliminaires [17], mais pas tous.

Ces définitions et ces exemples étant donnés, voici les propriétés supplémentaires qu'on peut énoncer pour $\hat{A}$.

Si A est postliminaire, $\hat{A}$ est un T_0-espace. Si A est liminaire, $\hat{A}$ est un T_1-espace.

Supposons A postliminaire. Il existe une suite croissante transfinie $(U_\rho)_{0 \leqslant \rho \leqslant \alpha}$ de parties ouvertes de $\hat{A}$ possédant les propriétés suivantes : 1) $U_0 = \varnothing$, $U_\alpha = \hat{A}$; 2) si ρ est un ordinal limite, U_ρ est réunion des $U_{\rho'}$ pour $\rho' < \rho$; 3) pour tout $\rho < \alpha$, $U_{\rho+1} - U_\rho$ est une partie ouverte dense *séparée* (donc localement compacte) de $\hat{A} - U_\rho$. On voit que $\hat{A}$ est « presque » localement compact [14].

Cette suite (U_ρ) a le défaut de ne pas être canonique. Disons qu'un point fermé π de $\hat{A}$ est *séparé* si, pour tout $\pi' \in \hat{A}$ distinct de π, π et π' admettent des voisinages disjoints. Soit V_1 l'intérieur de l'ensemble des points fermés séparés de $\hat{A}$; soit V_2' l'intérieur de l'ensemble des points fermés séparés de $\hat{A} - V_1$, et $V_2 = V_1 \cup V_2'$, etc. On construit ainsi une suite croissante transfinie (V_ρ) de parties ouvertes canoniques de $\hat{A}$, et les $V_{\rho+1} - V_\rho$ sont localement compacts. Si $V_\rho = \hat{A}$ à partir d'un ordinal β, A est une C^*-algèbre liminaire d'un type particulier. Cette circonstance se présente, avec de plus $\beta < +\infty$, lorsque $A = C^*(G)$ avec G groupe de Lie nilpotent réel connexe, ou $G = SL(2, \mathbf{C})$ [4] (et probablement lorsque G est semi-simple réel connexe).

Si G est compact, $\hat{G}$ est *discret*. (Plus généralement, il y a des relations entre les points isolés de $\hat{G}$ et les représentations irréductibles intégrables de G.) Si G est discret, $\hat{G}$ est *quasi-compact*.

Comme les éléments de $\hat{G}$ sont souvent induits par des représentations unitaires irréductibles de sous-groupes fermés G' distincts de G, il est intéressant d'étudier les propriétés de continuité de l'opération d'induction. Il existe des résultats dans cette voie. Mais il est impossible ici de se limiter aux représentations irréductibles. Il faut donc définir une topologie dans l'ensemble des représentations unitaires quelconques. Nous n'aborderons pas cette question [7].

6. Exemples

1) Si G est le groupe diédral infini, $\hat{G}$ s'identifie à $]0, 1[\cup \{a\} \cup \cup \{b\} \cup \{c\} \cup \{d\}$, où un point t de $]0, 1[$ tend à la fois vers a et b quand t tend vers 0 au sens usuel, et où t tend à la fois vers c et d quand t tend vers 1 au sens usuel (cf. fig. 1).

2) Si $G = SL(2, \mathbf{C})$, $\hat{G}$ s'identifie au sous-ensemble de $\mathbf{R}^2$ formé des points suivants :

(i) les points (n, r), où n est un entier $\geqslant 0$, et où r est réel (et $\geqslant 0$ si $n = 0$) ; ces points correspondent à la série principale ;

Fig. 1.

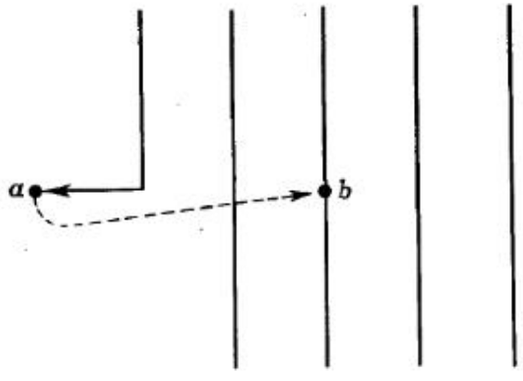


Fig. 2.

(ii) les points $(s, 0)$, où $-1 \leqslant s < 0$; ces points correspondent à la série supplémentaire si $-1 < s$, à la représentation triviale de dimension 1 si $s = -1$.

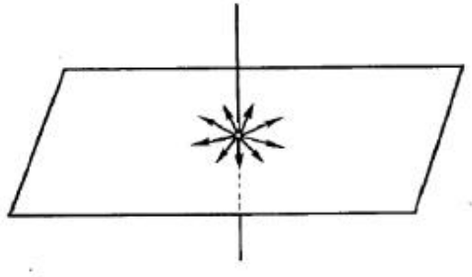


Fig. 3.

La topologie de $\hat{G}$ est la topologie induite par celle de $\mathbf{R}^2$, à ceci près que, si t tend vers a au sens usuel (cf. fig. 2), t tend à la fois vers a et b dans $\hat{G}$.

3) Dans le cas des groupes nilpotents, la non-séparation est un peu plus sensible. Par exemple, si G est le groupe simplement connexe réel nilpotent de dimension 3 non commutatif, $\hat{G}$ s'identifie à $(\mathbf{R} - \{0\}) \cup \mathbf{R}^2$; la topologie de $\hat{G}$ est celle de l'espace somme de

Fig. 4.

$\mathbf{R} - \{0\}$ et de $\mathbf{R}^2$, à ceci près qu'un point de $\mathbf{R} - \{0\}$ qui tend vers 0 au sens usuel tend dans $\hat{G}$ vers tous les points de $\mathbf{R}^2$ (fig. 3).

4) Dans le cas des groupes résolubles (même postliminaires), la non-séparation devient plus sévère. Par exemple, si G est le groupe résoluble réel de dimension 2 non commutatif, $\hat{G}$ s'identifie au sous-ensemble de $\mathbf{R}^2$ de la fig. 4, à ceci près que l'adhérence de a (resp. b) est $\hat{G} - \{b\}$ (resp. $\hat{G} - \{a\}$).

7. Structure borélienne sur $\hat{G}$ et $\hat{A}$

Supposons A séparable. Considérons sur $\mathrm{Irr}_n(A)$ la structure borélienne sous-jacente à la topologie, c'est-à-dire la structure borélienne définie par les fonctions $\pi \to \pi(a)\,\xi$ $(a \in A,\ \xi \in H_n)$. Pour cette structure, $\mathrm{Irr}_n(A)$ est un espace borélien standard. Munissons $\hat{A}_n$ de la structure borélienne quotient, et $\hat{A}$ de la structure borélienne somme de celles des $\hat{A}_n$ (comme A est séparable, on a $\hat{A}_n = \varnothing$ pour $n > \aleph_0$) [16]. Cette structure, appelée *structure de Mackey*, est plus fine que la structure borélienne sous-jacente à la topologie de $\hat{A}$. Les conditions suivantes sont équivalentes : 1) A est postliminaire ; 2) la structure de Mackey est la structure sous-jacente à la topologie de $\hat{A}$; 3) il existe une suite de parties de $\hat{A}$ boréliennes pour la structure de Mackey et qui séparent les points de $\hat{A}$ [9].

8. Cas d'un groupe de Lie; relations avec l'algèbre de Lie

Dans cette section, nous supposerons que G est un groupe de Lie réel connexe et nous noterons L son algèbre de Lie.

Le groupe G agit dans L par la représentation adjointe, donc dans l'espace vectoriel L^* dual de L. Si G est nilpotent [15] (ou, plus généralement, résoluble exponentiel [1]) simplement connexe, il existe une bijection canonique de L^*/G sur $\hat{G}$. On conjecture, au moins si G est nilpotent, que cette bijection est un homéomorphisme [15].

Soient $L_{\mathbf{C}}$ la complexification de L, U l'algèbre enveloppante de $L_{\mathbf{C}}$, $u \to u^*$ l'antiautomorphisme principal de U. Toute représentation unitaire topologiquement irréductible π de G définit une représentation π_∞ de U dans l'espace des vecteurs de H_π indéfiniment différentiables pour π. Soit Prim (U) l'ensemble des idéaux primitifs de U. Il est vraisemblable que Ker $(\pi_\infty) \in$ Prim (U). C'est en tous cas démontré quand G est semi-simple ou résoluble. Dans ces deux cas, on a donc une application Φ de $\hat{G}$ dans Prim (U), et Φ est continue si on munit Prim (U) de la topologie de Jacobson. Lorsque G est nilpotent simplement connexe, Φ est une bijection de $\hat{G}$ sur l'ensemble des $I \in$ Prim (U) tels que $I = I^*$; d'autre part, Prim (U) est en correspondance bijective canonique avec $L_{\mathbf{C}}^*/G$, et il est vraisemblable que la topologie de Jacobson sur Prim (U) est le quotient de la topologie de Zariski sur $L_{\mathbf{C}}^*$.

Ne supposons plus G nilpotent. Si z appartient au centre de U, $\pi_\infty(z)$ est scalaire pour tout $\pi \in \hat{G}$; posons $\pi_\infty(z) = \chi_\pi(z) \cdot 1$. Alors la fonction $\pi \to \chi_\pi(z)$ est continue sur $\hat{G}$.

9. Représentations non unitaires

Lorsque G est commutatif, $\hat{G}$ n'est qu'une partie de l'ensemble $\check{G}$ de tous les caractères (unitaires ou non) de G. Il serait intéressant de définir l'espace $\check{G}$ pour G non commutatif. Il y a divers résultats dans ce sens, mais la théorie ne semble bien établie que lorsque G possède un « grand » sous-groupe compact. Nous allons définir ce qu'il faut entendre par là.

Soit π une représentation continue de G dans un espace de Banach H. On dit que π est *complètement irréductible* (topologiquement) si le sous-espace vectoriel de $\mathcal{L}(H)$ engendré par $\pi(G)$ est fortement dense dans $\mathcal{L}(H)$. (Ceci entraîne que π est irréductible (topologiquement) ; on ignore si la réciproque est exacte.)

Soit K un sous-groupe compact de G. Pour tout $\delta \in \hat{K}$, soit H_δ la somme de tous les sous-espaces vectoriels V de H stables pour $\pi(K)$ et tels que la sous-représentation de $\pi|K$ induite dans V soit de classe δ. Chaque H_δ est un sous-espace vectoriel fermé de H, et ΣH_δ est dense dans H. Ceci posé, on dit que K est un *grand* sous-groupe compact si, pour tout $\delta \in \hat{K}$, il existe un entier n_δ tel que, pour toute représentation complètement irréductible π de G dans un espace de Banach H, on a dim $H_\delta \leqslant n_\delta$. Nous supposerons désormais que G possède un grand sous-groupe compact K. (C'est le cas si G est un groupe de Lie réel connexe semi-simple de centre fini, ou si G est le groupe des déplacements d'un espace euclidien.)

Considérons l'ensemble des représentations complètement irréductibles de G dans des espaces de Banach. Dans cet ensemble, définissons une relation d'équivalence : écrivons $\pi \sim \pi'$ s'il existe un sous-espace vectoriel H_0 (resp. H_0') dense dans H_π (resp. $H_{\pi'}$), stable par $\pi(G)$ (resp. $\pi'(G)$), et une application linéaire bijective T de H_0 sur H_0', fermée, telle que $\pi'(g) \circ T = T \circ (\pi(g) \mid H_0)$ pour tout $g \in G$. On note $\check{G}$ l'ensemble quotient. Nous allons définir une topologie sur $\check{G}$.

Soit $\mathcal{K}(G)$ l'algèbre convolutive des fonctions continues complexes à support compact sur G. Si $\pi \in \check{G}$, soit $\Phi(\pi)$ l'ensemble des formes linéaires $f \to \langle \pi(f)\xi, \xi' \rangle$ sur $\mathcal{K}(G)$, où $\xi \in H_\pi$, $\xi' \in H_\pi'$ (dual topologique de H_π). Soit $S \subset \check{G}$, et soit $\overline{\Phi(S)}$ l'adhérence de $\Phi(S)$ dans le dual de $\mathcal{K}(G)$ muni de la topologie faible. Les conditions $\Phi(\pi) \cap \bigcap \Phi(S)^- \neq \varnothing$ et $\Phi(\pi) \subset \Phi(S)^-$ sont équivalentes ; écrivons $\pi \in \overline{S}$ si elles sont remplies. Ceci définit une topologie sur $\check{G}$ [8].

L'ensemble $\hat{G}$ s'identifie à une partie de $\check{G}$, et la topologie induite sur $\hat{G}$ par celle de $\check{G}$ est la topologie considérée au § 2. Si G est un groupe de Lie connexe semi-simple linéaire, $\check{G}$ est localement quasi-compact, et il existe un entier p tel que toute suite (et même tout filtre) dans $\check{G}$ ait au plus p limites distinctes. L'espace $\check{G}$ a été calculé pour $G = SL(2, \mathbf{C})$ [8].

Université de Paris, Faculté des Sciences,
Département de Mathématiques, France

RÉFÉRENCES

[1] Bernat P., Sur les représentations unitaires des groupes de Lie résolubles, *Ann. Ec. Norm. Sup.*, **82** (1965), 37-99.
[2] Dixmier J., Sur les représentations unitaires des groupes de Lie algébriques, *Ann. Inst. Fourier*, **7** (1957), 315-328.
[3] Dixmier J., Sur les représentations unitaires des groupes de Lie nilpotents, V, *Bull. Soc. Math. France*, **87** (1959), 65-79.
[4] Fell J. M. G., The dual spaces of C^*-algebras, *Trans. Amer. Math. Soc.*, **94** (1960), 365-403.
[5] Fell J. M. G., C^*-algebras with smooth dual, *Ill. J. Math.*, **4** (1960), 221-230.
[6] Fell J. M. G., The structure of algebras of operator fields, *Acta Math.*, **106** (1961), 233-280.

[7] Fell J. M. G., Weak containment and induced representations of groups, II, *Trans. Amer. Math. Soc.*, **110** (1964), 424-447.
[8] Fell J. M. G., Non-unitary dual spaces of groups, *Acta Math.*, **114** (1965), 267-310.
[9] Glimm J., Type I C^*-algebras, *Ann. Math.*, **73** (1961), 572-612.
[10] Godement R., Les fonctions de type positif et la théorie des groupes, *Trans. Amer. Math. Soc.*, **63** (1948), 1-84.
[11] Harish-Chandra, Representations of semi-simple Lie groups, III, *Trans. Amer. Math. Soc.*, **76** (1954), 234-253.
[12] Jacobson N., A topology for the set of primitive ideals in an arbitrary ring, *Proc. Nat. Acad. Sci. U.S.A.*, **31** (1945), 333-338.
[13] Kadison R. V., Irreducible operator algebras, *Proc. Nat. Acad. Sci. U.S.A.*, **43** (1957), 273-276.
[14] Kaplansky I., The structure of certain operator algebras, *Trans. Amer. Math. Soc.*, **70** (1951), 219-255.
[15] Кириллов А. А., Унитарные представления нильпотентных групп Ли, *УМН*, **17**, № 4 (1962), 57-110.
[16] Mackey G. W., Borel structure in groups and their duals, *Trans. Amer. Math. Soc.*, **85** (1957), 134-165.
[17] Takenouchi O., Sur la facteur représentation d'un groupe de Lie résoluble de type (E), *Math. J. Okayama Univ.*, **4** (1955), 143-173.

NUCLEAR OPERATORS AND APPROXIMATIVE DIMENSION

B. S. MITIAGIN AND A. PELCZYNSKI

Introduction

In this report we would like to discuss some characteristics of linear operators and convex sets in linear topological spaces. We also will indicate some applications of these concepts mainly by introducing linear topological invariants—called *approximative dimensions* which are naturally induced by the characteristics. The origin of these concepts is going back to some approximation problems in the theory of functions of real and complex variables considered in earlier twenties by Bernstein, Favard and Kolmogorov. However the concepts themselves have been defined and investigated around 1960 in connection with two different topics:

first by Kolmogorov and Vitushkin and their colaborators Arnold, Erochin and Tichomirov in connection with the problem of compositions of smooth functions related to the 13-th Hilbert problem;

second by Bessaga, Mitiagin, Pelczynski, Pietsch and Rolewicz in connection with Gelfand's problem [4, p. 6] of characterization of Grothendieck's nuclear spaces [8] by degree of approximation by finite dimensional spaces.

In the present report we will restrict our attention to the second topic. The first one has been discussed by Kolmogorov in his address during the Stokholm Congress 1962.

1. Let $T: X \to Y$ be a linear operator (X and Y be normed linear spaces). The simplest characteristic seems to be the sequence of d-numbers of T defined as follows

$$d(T;\ n) = \inf_{E_n \subset Y} \sup_{x \in K_X} \inf_{y \in E_n} \|Tx - y\| =$$

$$= \inf_{E_n \subset Y} \inf \{\delta:\ TK_X \subset \delta K_Y + E_n\},$$

where K_X denotes the unit ball of X and E_n is an arbitrary n-dimensional subspace of X. Clearly $d(T;\ n)$ is the infimum over all n-dimensional subspaces $E_n \subset X$ of inclination of the image TK_X of the unit ball K_X from E_n. It expresses how «nicely» the image TK_X can be approximated by n-dimensional subspaces of Y.

The next inequalities show how the d-numbres depend on basic algebraic operations on linear operators.

(1) $$d(ATB;\ n) \leqslant \|A\| \cdot \|B\| \cdot d(T;\ n)$$

for arbitrary linear operators A and B.

(2) $$d(T \oplus S;\ n+m) \leqslant d(T;\ n) + d(S;\ m),$$

(3) $$d(T \otimes S;\ nm) \leqslant d(T;\ n) \cdot d(S;\ m);$$

$T \oplus S$ denotes the Cartesian product of operators and $T \otimes S$ is the projective tensor product of operators.

(4) $$d(T+S;\ n+m) \leqslant d(T;\ n) + d(S;\ m),$$

(5) $$d(TS;\ n+m) \leqslant d(T;\ n) \cdot d(S;\ m).$$

The most important and useful is the inequality (5) which is called the composition formula for d-numbers. In the case of operators acting in Hilbert spaces this formula is closely related to the classical inequalities on eigenvalues due to Horn and H. Weil [5], [20]. From the composition formula it is easy to derive the following corollary:

If $d(T;\ n) = O(n^{-a})$ *and* $d(S;\ n) = O(n^{-b})$, *then* $d(TS;\ n) = O(n^{-a-b})$ *for* $a,\ b > 0$.

Let us mention some "analytic" properties of d-numbers.

a) $d(T;\ n) \to 0$ iff T is compact,

b) if $T: H \to H$ is an operator in a Hilbert space then $d(T;\ n) = \lambda_n$ where λ_n denote the n-th eigenvalue of the operator $|T| = \sqrt{T \cdot T^*}$. In particular

T is a Hilbert-Schmidt operator iff $\sum d(T; n)^2 < +\infty$;
T belongs to the trace class iff $\sum d(T; n) < +\infty$.

The property b) can be generalized to the case of diagonal operators acting between symmetric sequence spaces. This is useful in computing of approximative dimension of Köthe spaces of sequences.

Furthermore let $T : X \to Y$ be an arbitrary linear operator (X, Y be linear normed spaces). Then

c) if T and $S : Y \to Z$ are nuclear (i.e. $T = \sum x_i^* (\cdot)\, y_i$ with $\sum \| x_i^* \| \cdot \| y_i \| < \infty$) then $d(ST; n) = O(n^{-1})$,

d) if $d(T; n) = O(n^{-3-\varepsilon})$ for some $\varepsilon > 0$ then T is nuclear.

2. Now we are ready to introduce the concept of *diametral approximative dimension* [2, 13]. (The analogous invariants in terms of ε-capacity have been introduced earlier in [10, 16].)

Let us recall that every locally convex linear topological space can be represented as an inverse limit of normed linear spaces

$$X = \lim \operatorname{inv} (X_\alpha;\ T_\alpha^\beta).$$

Let s_+ denote the set of all positive sequences. Let us set

$$\delta(X) = \{(t_n) \in s_+ : \underset{\alpha}{\forall}\, \underset{\beta}{\exists}\, t_n \cdot d(T_\alpha^\beta,\ n) \to 0\}.$$

The set $\delta(X)$ is called the *diametral approximative dimension* of X. It does not depend on the representation of X as an inverse limit of normed linear space. It is a linear topological invariant. Precisely:

1°. If the spaces X_1 and X_2 are linearly homeomorfic (=isomorphic) then $\delta(X_1) = \delta(X_2)$.

2°. If Y is a closed linear subspace of X, then

$$\delta(Y) \supset \delta(X) \text{ and } \delta(X/Y) \supset \delta(X).$$

3°. $\delta(X) = s_+$ iff X is the topological product of one dimensional spaces (of arbitrary power).

4°. $\delta(X)$ contains all bounded sequences iff X is a Schwartz space, i.e. X admits a representation

$$(*) \qquad X = \lim \operatorname{inv} (X_\alpha;\ T_\alpha^\beta)$$

where all $T_\alpha^\beta : X_\beta \to X_\alpha$ are compact operators.

The next result seems to be especially important (characterization of nuclearity [13]):

5°. Let X be a linear topological space. Then the following conditions are equivalent:

1) X is nuclear, i.e. X has a representation (*) with nuclear operators T_α^β;

2) for some $\varepsilon > 0$ the sequence $(n^\varepsilon) \in \delta(X)$;

3) $\delta(X)$ contains all sequences of power growth, i.e. $(n^\varepsilon) \in \delta(X)$ for each $\varepsilon > 0$.

For various special spaces the diametral approximative dimension can be computed. In particular

$$\delta(C^\infty(\Omega)) = \{(t_n) \in s_+ : \lim_n t_n \cdot n^{-k} = 0 \text{ for some } k > 0\},$$

where $C^\infty(\Omega)$ denotes the space of all infinitely differentiable functions on Ω;

$$\delta(A(D^m)) = \{(t_n) \in s_+ : \underset{b>0}{\forall}\, \lim_n t_n e^{-bn^{1/m}} = 0\},$$

$$\delta(A(\mathbf{C}^m)) = \{(t_n) \in s_+ : \underset{B>0}{\exists}\, \lim_n t_n e^{-Bn^{1/m}} = 0\},$$

where $A(D^m)$ (respectively $A(\mathbf{C}^m)$) denotes the space of all holomorphic (entire) functions on the m-polydisc (respectively on the m-dimensional complex Cartesian space).

The formulas for $\delta(A(D^m))$ and $\delta(A(\mathbf{C}^m))$ shows that for spaces of holomorphic functions the number of variables is a linear topological invariant [10] which is not the case for the spaces of infinitely differentiable functions.

Finally using diametral approximative dimension one can construct an example of infinite-dimensional Frechet space which is not isomorphic to any of its maximal hyperplanes [2], as well as to show that

every basis in a nuclear Frechet space is unconditional, i.e. every Frechet nuclear space with a basis is a nuclear Köthe space (Dynin-Mitiagin [13]).

3. One can define invariants analoguous to diametral approximative dimension starting from other sequences assigned to linear operators. Now we will discuss briefly some of them [11, 14, 15, 19, 21].

Let $T: X \to Y$ be a linear operator (X, Y be linear normed spaces). Let us put

$$e(T; n) = \inf_{y_1, \ldots, y_n \in Y} \sup_{x \in K_X} \inf_{1 \le i \le n} \| Tx - y_i \|;$$

$e(T; n)$ expresses the best approximation of TK_X by set consisting of n elements. $e(T; \cdot)$ as a function of n is inverse function of Hausdorf ε-capacity of the (compact) set TK_X.

$$d(T; n) = \inf_{E_n} D(TK_X;\ E_n);$$

$d(T; n)$ has been discussed previously;

$$c(T; n) = \inf_{\substack{V: X \to Y \\ \dim V \leqslant n}} \| T - V \|.$$

This characteristic have been investigated by Pietsch [19]. The first two characteristics are not trivial for all compact operators, i.e. $\lim e(T; n) = \lim d(T; n) = 0$ iff T is compact. Clearly $\lim c(T; n) = 0$ iff T is in the norm closure of finite dimensional operators [1]).

$$b(T; n) = \sup_{E_n} \inf_{x \in K_X \cap E_n} \| Tx \|,$$

$$a(T; n) = \sup_{E^n} \inf_{\|x\|=1} \| Tx \|_{Y/E^n}.$$

There are non-compact operators T for which $b(T; n) \to 0$. Hence the b-characteristic can be applied to the essentially wider class than the class of compact operators. The same situation holds for a-characteristic which is in fact dual to b-characteristic.

$$r(T; n) = \sup_{\dim TE_n = n} \left\{ \frac{\operatorname{Vol} T(K_X \cap E_n)}{\operatorname{Vol}(K_Y \cap TE_n)} \right\}^{1/n}.$$

The r-characteristic is of a different type than others; it cannot be expressed as a function of distance, i.e. in terms of "simple approximation".

The alphabetic order from a to e of the described characteristics is not casual. The a- and b-numbers depend on linear structure of both spaces X and Y and c-, d-numbers depend only on linear structure of Y. Since c is the value of the best linear approximation and d is the value of the best non-linear approximation, d is "less linear" than c. The most non-linear are e- and r-numbers.

Slightly modifying the definition of e-approximative dimension one can obtain an invariant of uniform spaces. For details we refer to the recent paper of Bessaga [1].

4. Now we will discuss briefly the relationships between different characteristics of operators described above. One can show that roughly speaking the n-th number of any type may be estimated by n-th number of each other type multiplied by some power of n. For example

$$d(T; n) \leqslant e(T; n) \leqslant n \cdot d(T; n)$$

$$b(T; n) \leqslant r(T; n) \leqslant n \cdot b(T; n).$$

The inequalities of this type follow immediately either from definition or applying the Auerbach lemma on the existence of orthonormal basis

[1]) Here E^n denotes an arbitrary n codimensional subspace of Y, i. e. the quotient space Y/E^n is n-dimensional.

in every finite dimensional normed linear space [18, lemma 8.4.1]. More sophisticated is the inequality

$$b(T; n) \leqslant d(T; n-1) \leqslant n^2 \cdot b(T; n).$$

The left side of this inequality is due to M. Krein, M. Krasnoselskii and D. Milman [12]; for the right-side inequality see [15]. We omitted here some non-trivial inequalities between e- and d-numbers [13].

We have very poor information on b-numbers which have not been studied systematically. In particular we have no satisfactory formula of composition for these numbers. Having such a formula one would be able to obtain the new composition theorem for (q, p) *absolutely summing* operators as well as new characterization of nuclearity.

A linear operator $T: X \to Y$ is said to be (q, p) absolutely summing, provided for every sequence $(x_n) \subset X$

$$\left(\sum \| Tx_n \|^q\right)^{1/q} \leqslant C \sup_{\|x^*\| \leqslant 1} \left(\sum |x^*(x_n)|^p\right)^{1/p}.$$

Using the Dvoretzky-Rogers lemma one can show that if $0 < 1/p - 1/q < 1/2$ and T is (q, p) absolutely summing, then

$$b(T; n) = O(n^{(2-p)/4p}) = O(n^{1/2+1/p-1/q}).$$

Hence the natural embedding $l_1 \to l_2$ which is 1-absolutely summing is an example of a non-compact operator the sequence of b-numbers of which tends to zero.

Futhermore we mention some relationship between b-numbers and projection constants. Namely, if $T: C(S) \to l_2$ is an arbitrary linear operator from the space $C(S)$ of all continuous functions on a compact space S to a Hilbert space l_2 then

$$b(T; n) \leqslant \frac{1}{\sqrt{p_n}},$$

where p_n denote the projection constants of n-dimensional Euclidian space.

Finally we observe that if $T: X \to Y$ is a linear operator and $b(T; n) \to 0$, then T is a strictly singular in the sense of Kato [9], i.e. the restriction of T to no infinite-dimensional closed subspace of X has a bounded inverse. Similarly if $a(T; n) \to 0$ then T is strictly cosingular [17]. Hence the linear operators with non-trivial a- and b-characteristics have the perturbation properties exhibited by Kato [9], Gohberg and Krein [6, 7].

University of Voronezh,
U.S.S.R,
University of Warsaw,
Poland

REFERENCES

[1] Bessaga C., On topological classification of complete linear metric spaces, *Fund. Math.*, **54**, № 3 (1965), 251—288.

[2] Bessaga C., Pelczyński A., Rolewicz S. On diametral approximatve dimension and linear homogeneity of *F*-spaces, *Bull. Acad. Pol. Sci.*, **9**, № 9 (1961), 677-683.

[3] Витушкин А. Г., Оценка сложности задачи табулирования, Физматгиз, М., 1959.

[4] Гельфанд И. М., О некоторых проблемах функционального анализа, *УМН*, **11**, № 6 (72) (1956), 3-12.

[5] Гохберг И. Ц., Крейн М. Г., Введение в теорию линейных несамосопряженных операторов, «Наука», М., 1965.

[6] Гохберг И. Ц., Крейн М. Г., Основные положения о дефектных числах, корневых числах и индексах линейных операторов, *УМН*, **12**, № 2 (74) (1957), 43-118.

[7] Гохберг И. Ц., Маркус А. С., Фельдман И. А., О нормально разрешимых операторах и связанных с ними идеалах, *Изв. Молдавского Филиала АН СССР*, **10** (76) (1960), 51-70.

[8] Grothendieck A., Produits tensoriels topologiques et espaces nucléaires, Memoirs AMS, 16, 1955.

[9] Kato T. Perturbation theory for nullity, deficiency and other quantities of linear operators, *Journ. d'Analyse Math.*, 6 (1958), 261-322.

[10] Колмогоров А. Н., О линейной размерности топологических векторных пространств, *ДАН СССР*, **120**, № 2 (1958), 239-241.

[11] Колмогоров А. Н., Тихомиров В. М., ε-энтропия и ε-емкость множеств в функциональных пространствах, *УМН*, **14**, № 2 (86) (1959), 3-86.

[12] Крейн М. Г., Красносельский М. А., Мильман Д. П., О дефектных числах линейных операторов в банаховых пространствах и о некоторых геометрических вопросах, Сб. трудов ин-та математики АН УССР, 11 (1948), 97-112.

[13] Митягин Б. С., Аппроксимативная размерность и базисы в ядерных пространствах, *УМН*, **16**, № 4 (100) (1961), 63-132.

[14] Митягин Б. С., Тихомиров В. М., Асимптотические характеристики компактов в линейных пространствах, Труды 4-го Всесоюзного математ. съезда, том 2, 299-308, «Наука», М., 1964.

[15] Митягин Б. С., Хенкин Г. М., Неравенства между различными *n*-поперечниками, Труды семинара по функциональному анализу, Воронеж, вып. 7, 1963, 97-103, изд-во ВГУ.

[16] Pelczynski A. On the approximation of *S*-spaces, *Bull. Acad. Pol. Sci.*, **5**, № 9 (1957), 879-881.

[17] Pelczynski A. On strictly singular and strictly cosingular operators, *Bull. Acad. Pol. Sci.*, **13**, № 1 (1965), I, 31-36; II, 37-41.

[18] Pietsch A., Nukleare lokalkonvexe Räume, Berlin, Academie-Verlag, 1965. Русский перевод: Пич А., Ядерные локально выпуклые пространства, «Мир», М., 1967.

[19] Pietsch A., Einige neue Klassen von kompakten linearen Abbildungen, *Revue de Math. Pure et Appl.* (Bucharest), **8** (1963), 427-447.

[20] Schatten R., Norm ideals of completely continuous operators, Berlin-Cottingen-Heidelberg, 1960.

[21] Тихомиров В. М., Поперечники множеств в функциональных пространствах и теория наилучших приближений, *УМН*, **15**, № 3 (93), (1960), 81-120; № 6 (96), 226.

ТЕОРИЯ ПРЕДСТАВЛЕНИЙ ГРУПП

М. И. ГРАЕВ, А. А. КИРИЛЛОВ

Введение

За четыре года, прошедшие после Стокгольмского конгресса, в теории представлений групп было получено много новых важных результатов. Некоторые из них изложены в докладах Хариш-Чандра, Диксмье и сообщениях Фелла и Блатнера.

В своем докладе мы хотим рассказать о результатах, полученных советскими математиками, и главным образом о тех задачах и направлениях в теории представлений, которые нам кажутся наиболее интересными и плодотворными.

I. Представления алгебраических групп

До сих пор в теории бесконечномерных представлений групп в основном занимались изучением комплексных и вещественных групп Ли. Однако в последнее время появляется все больше работ, посвященных представлениям групп других типов. Более того, как показывает опыт, многие факты теории представлений комплексных и вещественных групп Ли (конструкция дискретных серий, вычисление характеров и меры Планшереля) становятся понятными лишь после того, как их удается перенести на более широкий класс групп. Таким более широким классом, для которого теория представлений приобретает более естественный вид, является, по нашему мнению, класс алгебраических групп над локально компактными полями.

Отметим сразу же, что в этом направлении предстоит еще очень много работы и полученные результаты — лишь первые шаги в большую неизученную область.

Будущее этого направления представляется нам следующим образом. Пусть G — алгебраическая группа, определенная над полем k. Для всех групп G_K, где K — локально компактное расширение поля k, строится единая теория представлений, в которой поле K играет роль параметра. При этом возникают задачи нового типа: описать множество значений «параметра» K, при которых теория представлений группы G_K удовлетворяет тем или иным требованиям. В качестве примера можно привести такое предложение.

Пусть G — некоммутативная алгебраическая группа, определенная над полем k характеристики 0, и пусть K — локально компактное расширение поля k. Группа G_K принадлежит типу I

тогда и только тогда, когда поле K самодуально (т. е. аддитивная группа поля K двойственна самой себе в смысле Понтрягина).

Перечислим теперь имеющиеся результаты:

а) Для алгебраических нильпотентных групп над полем нулевой характеристики теория, построенная одним из авторов в [1], уже имеет тот вид, о котором мы говорили выше. А именно для любой нильпотентной группы G и самодуального поля K все основные вопросы теории представлений группы G_K получают простые и наглядные решения в терминах орбит группы G_K в пространстве $\mathfrak{g}_K^*$, дуальном к алгебре Ли $\mathfrak{g}_K$ группы G_K. Например, характеры неприводимых унитарных представлений группы G_K являются преобразованиями Фурье δ-функций, сосредоточенных на орбитах:

$$\chi_\omega(\exp x) = \int\limits_{\mathfrak{g}_K^*} e^{i\langle x, y\rangle}\,\delta_\omega(y)\,dy,$$

где $x \in \mathfrak{g}_K$, $y \in \mathfrak{g}_K^*$, ω — орбита группы G_K в пространстве $\mathfrak{g}_K^*$, exp — каноническое отображение $\mathfrak{g}_K$ на G_K.

Разумеется, методы вычисления этого интеграла и явные формулы для характеров будут специфическими для каждого поля.

На этом примере особенно наглядно проявляется роль поля K как «параметра».

б) *Классические представления редуктивных групп.* Известная конструкция Гельфанда — Наймарка, примененная ими для описания неприводимых унитарных представлений комплексных полупростых групп Ли, имеет на самом деле более широкую область применения. А именно для любой редуктивной расщепимой алгебраической группы можно определить следующие серии унитарных представлений, которые мы назовем классическими.

Основная невырожденная серия состоит из представлений T_χ, индуцированных унитарными одномерными представлениями χ подгруппы Бореля.

Основные вырожденные серии получаются, если вместо подгруппы Бореля рассматривать произвольные параболические подгруппы.

Дополнительные серии получаются из основных серий «методом аналитического продолжения», т. е. переходом к неунитарным характерам χ. При этом некоторые из представлений T_χ остаются унитарными.

Особые серии возникают вследствие того, что существуют «особые точки» χ, в которых представления T_χ оказываются приводимыми. При этом некоторые из неприводимых компонент могут быть унитарными.

Укажем некоторые возникающие здесь задачи.

1. Доказать неприводимость описанных представлений. Отметим, что доказательства известны лишь для представлений вещественных полупростых групп Ли (см. [2], [3], [4]), а в случае редуктивной группы над полем p-адических чисел — для невырожденных серий представлений [5].

2. Описать область неунитарных характеров χ, для которых представления T_χ принадлежат дополнительной серии. Окончательные результаты здесь получены только для групп над полем комплексных чисел. Описать представления особой серии.

3. Известно, что перечисленными сериями исчерпываются неприводимые унитарные представления комплексных полупростых групп Ли (см. [2, 3]), а также группы $SL(2, \mathbb{R})$. Недавно было установлено, что этот факт справедлив для групп $SL(3, \mathbb{R})$ и $GL(3, \mathbb{R})$ (см. [6]). Интересно выяснить, какова «область справедливости» этого факта. Наиболее вероятно, что это — все вещественные группы Шевалле — Диксона.

в) *Группа $SL(2, K)$.* Это первая (и пока единственная) из полупростых групп, для которых известны все неприводимые унитарные представления для всех самодуальных полей K (см. [7]). Если K — несвязное поле, то группа $SL(2, K)$ допускает представления, не входящие ни в одну из перечисленных серий. Эти новые представления мы будем называть аналитической серией, так как для вещественного поля они естественно реализуются в пространстве аналитических функций [1]).

Более точно, каждому нерасщепимому над K тору в $SL(2, K)$ отвечает своя серия представлений, которые нумеруются унитарными характерами π этого тора (и образуют, следовательно, дискретное множество, так как нерасщепимый тор в $SL(2, K)$ является компактной группой).

Получены явные формулы для матричных элементов этих представлений, в которых роль параметра играет квадратичное расширение поля K (поле расщепления соответствующего тора).

Получено также единое для всех групп $SL(2, K)$ интегральное представление для плотности меры Планшереля:

$$\mu(\pi) = -\int \pi(t)\,|1-t|^{-2}\,dt,$$

где в классическом случае интеграл берется по K, а в случае аналитической серии — по окружности $t\bar{t}=1$ в соответствующем квадратичном расширении K.

В случае комплексного и вещественного поля эти интегралы без труда вычисляются и приводят к совершенно различным окончательным формулам — факт, который до сих пор удивлял всех, кто с ним сталкивался.

[1]) Впрочем, для вещественного поля аналитическая серия совпадает с особой. Удачность нашего названия зависит от того, насколько случайно это совпадение.

Очень вероятно, что в высшей степени сложное выражение, полученное в [8] для меры Планшереля группы $SL(n, \mathbf{R})$, удается упростить после того, как соответствующий результат будет перенесен на все группы $SL(n, K)$.

II. Некоммутативная алгебраическая геометрия

Современная алгебраическая геометрия является по существу наукой о коммутативных нетеровых кольцах. Содержательность этой теории обеспечивается наличием классических примеров колец алгебраических функций и алгебраических чисел.

Теория представлений групп доставляет много примеров некоммутативных колец. Изучение алгебраической структуры этих колец интересно само по себе и существенно для многих вопросов анализа.

Один из наиболее интересных и важных примеров — обертывающие алгебры алгебр Ли.

Совокупность максимальных или простых идеалов этих колец является аналогом аффинной схемы и одновременно алгебраическим вариантом дуального пространства $\hat{G}$ для группы G. Это множество несет естественную структуру кольцованного пространства. Изучение такого рода пространств обещает быть очень интересным.

Отметим, что пространство $\hat{G}$ допускает естественную проекцию на аффинное многообразие Spec C, где C — центр алгебры $U(G)$. В простейших случаях это отображение оказывается взаимно однозначным, за исключением некоторых «особых точек».

По аналогии с обычной алгебраической геометрией можно поставить вопрос о «бирациональной классификации» возникающих здесь объектов.

Два кольцованных пространства, связанных с алгебрами $U(G_1)$ и $U(G_2)$, мы называем бирационально эквивалентными, если тела отношений этих алгебр изоморфны.

В работе И. М. Гельфанда и одного из авторов [9] был сделан первый шаг в бирациональной классификации дуальных пространств для алгебраических групп Ли. Полученные результаты позволяют сформулировать следующую гипотезу.

Пусть $D_{n,k}(L)$ — тело над полем L, порожденное образующими $p_1, \ldots, p_n, q_1, \ldots, q_n, z_1, \ldots, z_k$ и соотношениями $p_iq_j - q_jp_i = \delta_{ij}$, а остальные коммутаторы равны нулю. (Это тело, очевидно, изоморфно телу отношений кольца дифференциальных операторов с полиномиальными коэффициентами от n неизвестных и k параметров.)

Пусть G — алгебраическая группа над полем L, $\mathfrak{g}$ — ее алгебра Ли, $U(G)$ — обертывающая алгебра и $D(G)$ — тело отношений алгебры $U(G)$. Тогда тело $D(G)$ изоморфно «стандартному телу» $D_{n,k}(L)$, где числа n и k определяются из условий:

$$2n + k = \dim G,$$

k = степени трансцендентности центра $U(G)$.

Эта гипотеза доказана в [9] для групп $SL(n)$, $GL(n)$, полупростых групп ранга 2 и всех нильпотентных алгебраических групп.

Показано также (с помощью введенного авторами понятия размерности, обобщающего понятие степени трансцендентности), что тела $D_{n,k}(L)$ попарно неизоморфны.

По-видимому, справедливо также следующее утверждение.

Для любого простого идеала P алгебры $U(G)$ тело отношений факторалгебры $U(G)/P$ изоморфно одному из стандартных тел $D_{n,k}(L)$ [1].

Это направление имеет непосредственное отношение к теории бесконечномерных представлений групп Ли. Дело в том, что все сконструированные до сих пор представления вещественных групп Ли могут быть так реализованы в пространстве функций, что элементы обертывающей алгебры перейдут в дифференциальные операторы с полиномиальными коэффициентами.

Очень интересно было бы построить теорию представлений, в которой это свойство принято за определение: представлением называется гомоморфизм $U(G)$ в кольцо дифференциальных операторов с полиномиальными коэффициентами. Если перейти теперь от гомоморфизма алгебр к гомоморфизмам соответствующих тел и принять сформулированные выше гипотезы, то все задачи теории представлений сведутся к изучению алгебраической структуры стандартных тел $D_{n,k}(L)$.

Следует отметить, что в этом направлении пока возникло гораздо больше вопросов, чем получено ответов. Среди наиболее важных нерешенных задач отметим описание автоморфизмов тела $D_{n,k}(L)$ и, в частности, описание инволюций в этом теле.

III. Метод орисфер

Одна из основных задач теории представлений — описание спектров конкретных представлений, и в первую очередь — представлений, реализующихся в функциях на однородных пространствах.

В случае когда G — комплексная полупростая группа, для ряда однородных пространств эта задача эффективно решается с помощью

[1] Это утверждение не имеет смысла, если идеал P не является вполне простым (т. е. алгебра $U(G)/P$ содержит делители нуля). Как показал Диксмье, для разрешимых алгебр всякий простой идеал $P \subset U(G)$ будет вполне простым. Для неразрешимых алгебр простые, но не вполне простые идеалы существуют, но все они, по-видимому, связаны с конечномерными представлениями.

предложенного Гельфандом и Граевым [10] метода орисфер. Однако этот метод может быть полезен и в более общей ситуации.

Напомним коротко существо метода орисфер.

Пусть X — однородное пространство группы G. Орисферами в X называются замкнутые орбиты орисферических подгрупп группы G. С каждым транзитивным семейством орисфер Ω связано орисферическое отображение, которое функции $f(x)$ на X сопоставляет функцию $\varphi(\omega)$ на Ω:

$$\varphi(\omega) = \int_{\omega} f(x)\, dx.$$

Спектр представления, реализующегося в функциях на Ω, во многих случаях легко вычисляется. Таким образом, изучение спектра искомого представления сводится к исследованию орисферического отображения, в частности его ядра и коядра.

Интересные примеры однородных пространств возникают в теории чисел. Пожалуй, наиболее важным из них, интенсивно изучающимся в последнее время, является пространство $X = G_A/G_Q$, где G_A — группа аделей редуктивной алгебраической группы G, а G_Q — подгруппа главных аделей. Изучение спектра пространства X особенно интересно тем, что оно является мостом, связующим теорию представлений с теорией чисел, а получаемые здесь результаты можно интерпретировать как теоретико-числовые (см., например, [7] о связи с гипотезой Петерсона).

Изложим результаты, получаемые в этой задаче методом орисфер. Будем предполагать, что G расщепима над $\mathbf{Q}$. Пусть Z — орисферические подгруппы группы G, Z_A — их группы аделей. Орисферами в $X = G_A/G_Q$ называются компактные орбиты подгрупп Z_A. Первый результат: пересечение H ядер всех орисферических отображений пространства $L_2(X)$ имеет дискретный спектр. Для изучения спектра дополнительного подпространства рассмотрим пространство Ω орисфер максимальной размерности. Это пространство однородно. Соответствующее Ω орисферическое отображение переводит $L_2(X)$ в некоторое пространство H' функций на Ω. Второй результат: пространство $H'' = H' \cap L_2(\Omega)$ содержит все представления, входящие в спектр $L_2(\Omega)$, но с кратностью 1. Доказательство этого факта основано на изучении спектра пространства $L_2(\Omega)$.

Получающийся здесь очень красивый ответ может быть сформулирован следующим образом. Применим метод орисфер к самому пространству $L_2(\Omega)$. Оказывается, что в пространстве Ω имеется несколько транзитивных семейств орисфер, причем эти семейства находятся во взаимно однозначном соответствии с элементами группы Вейля W группы G. Пусть B_w — оператор орисферического отображения, отвечающий $w \in W$. Устанавливается, что операторы B_w можно с некоторого всюду плотного подмножества функций

продолжить до унитарных операторов $\overline{B}_w$ на $L_2(\Omega)$ и что операторы $\overline{B}_w$ образуют представление группы W. В силу самой конструкции операторы $\overline{B}_w$ перестановочны с операторами представления.

Кратность каждого неприводимого представления, входящего в спектр $L_2(\Omega)$, равна порядку группы W. Соответствующие подпространства можно выбрать так, что они переставляются операторами $\overline{B}_w$. Подпространство H'' состоит из векторов, инвариантных относительно всех $\overline{B}_w$, и, следовательно, совпадает с $M(L_2(\Omega))$, где $M = \sum_{w \in W} \overline{B}_w$.

Для полного описания спектра необходимо исследовать и другие орисферические отображения пространства $L_2(X)$. Эта задача еще не решена [1].

В заключение отметим, что с методом орисфер связано много красивых нерешенных задач теории представлений. Одна из основных задач — описание категории пространств $L_2(G/Z)$, где G — фиксированная редуктивная группа, а Z пробегает все попарно несопряженные орисферические подгруппы.

Другая задача состоит в выяснении «области справедливости» следующего утверждения: пересечение ядер всех орисферических отображений имеет дискретный спектр.

Московский университет,
Москва, СССР

ЛИТЕРАТУРА

[1] Кириллов А. А., *УМН*, **17**, вып. 4 (1962).
[2] Гельфанд И. М., Наймарк М. А., *Труды МИАН*, 1950.
[3] Березин Ф. А., *Труды Моск. матем. об-ва* (1957); **6** (1963), 12.
[4] Bruhat F., *Bull. Soc. Math. France*, **84**, № 2 (1956).
[5] Bruhat F., *Bull. Soc. Math. France*, **89**, № 1 (1961).
[6] Вахутинский И. Я., *ДАН СССР*, **170**, № 1 (1966).
[7] Гельфанд И. М., Граев М. И., Пятецкий-Шапиро И. И., Теория представлений и автоморфные функции, Москва, 1966.
[8] Ромм Б. Д., *Изв. АН СССР, сер. матем.*, **29**, № 5 (1965).
[9] Гельфанд И. М., Кириллов А. А., *ДАН СССР*, **167**, № 3 (1966) (подробное изложение см. в Publ. Math. IHES, № 31, 1966).
[10] Гельфанд И. М., Граев М. И., *Труды Моск. матем. об-ва*, № 8 (1959).

[1]) Другой подход к изучению спектра пространства содержится в работе Ленглендса (препринт).

6 Обыкновенные дифференциальные уравнения
Ordinary differential equations
Equations différentielles ordinaires
Gewöhnliche Differentialgleichungen

A CLASS OF LINEAR FUNCTIONAL EQUATIONS

JACK K. HALE[1]

This report is a summary of some unpublished results of K. Meyer and the author concerning a class of autonomous linear functional equations which includes as special cases autonomous linear functional differential equations of retarded and neutral type as well as functional difference equations.

Let R^n be a real or complex n-dimensional linear vector space of column vectors with norm $|\cdot|$ and let C_r ($[-r, 0]$, R^n) be the Banach space of continuous functions mapping $[-r, 0]$ into R^n with the norm $\|\varphi\|_r$ for φ in C_r defined by $\|\varphi\|_r = \max\{|\varphi(\theta)|, \theta \text{ in } [-r, 0]\}$. If g, f are continuous linear mappings of C_r into R^n, then there exist $n \times n$ matrices μ, η whose elements are of bounded variation on $[-r, 0]$ such that

$$(1) \qquad g(\varphi) = \int_{-r}^{0} [d\mu(\theta)]\,\varphi(\theta), \qquad f(\varphi) = \int_{-r}^{0} [d\eta(\theta)]\,\varphi(\theta)$$

for all φ in C_r. We shall suppose that the measure μ is nonatomic at 0 and more specifically that there is a continuous nondecreasing function $\delta(s)$, $0 \leqslant s \leqslant r$, such that $\delta(0) = 0$ and

$$(2) \qquad \left| \int_{-s}^{0} [d\mu(\theta)]\,\varphi(\theta) \right| \leqslant \delta(s)\,\|\varphi\|_s$$

for all φ in C_r.

For any φ in C_r, define $\gamma(\varphi) = \varphi(0) - g(\varphi)$. For any integrable function h mapping $[0, \infty)$ into R^n and any fixed element φ in C_r,

[1] This research was supported in part by National Aeronautics and Space Administration under Contract No. NGR 40-002-015, in part by the United States Army Research Office, Durham, under Contract No. DA-31-124-ARO-D-270 and in part by the Office of Aerospace Research, United States Air Force, under AFOSR Grant No. 693-66.

consider the functional integral equation

$$(3) \qquad x(t) = \varphi(t), \quad -r \leqslant t \leqslant 0,$$

$$x(t) = \gamma(\varphi) + g(x_t) + \int_0^t f(x_s)\,ds + \int_0^t h(s)\,ds, \quad t \geqslant 0,$$

where, for each fixed $t \geqslant 0$, x_t is in C_r and is defined by $x_t(\theta) = x(t+\theta)$, $-r \leqslant \theta \leqslant 0$. By a solution of (3), we will always mean a continuous function satisfying the above relation.

For $g \equiv 0$, equation (3) is equivalent to the functional differential equation of retarded type

$$(4) \qquad \dot{x}(t) = f(x_t) + h(t)$$

with the initial condition at $t = 0$ given by φ. If $f \equiv 0$ and $h \equiv 0$, equation (3) is a functional difference equation of retarded type and, in particular, includes difference equations. For both f and g not identically zero, equation (3) corresponds to a retarded equation of neutral type. In fact, formal differentiation of the equation yields

$$(5) \qquad \dot{x}(t) = g(\dot{x}_t) + f(x_t) + h(t)$$

where $\dot{x}_t$ is defined as $\dot{x}_t(\theta) = \dot{x}(t+\theta)$, $-r \leqslant \theta \leqslant 0$. Also, if one begins with (5) and defines a solution with initial function φ at 0 to be a continuous function satisfying (5) almost everywhere, then an integration yields (3) with $\gamma(\varphi) = \varphi(0) - g(\varphi)$.

This latter remark is precisely the reason for considering the equation (3) rather than (5). If one attempts to discuss equation (5) directly, then the first problem that is encountered is a precise definition of a solution and a precise definition of the topology to be induced on the space in which the solution will lie. Such a topology will necessarily include the first derivative of x in some way; whereas, if we consider equation (3), the simpler space C_r can be employed.

If h in (3) is identically zero, we will say equation (3) is homogeneous and, otherwise, it is nonhomogeneous.

Theorem 1. For any given φ in C_r, there is a unique function $x(\varphi)$ defined and continuous on $[-r, \infty)$ such that $x(\varphi)$ satisfies (3) on $[0, \infty)$. Furthermore, there is a constant $\beta > 0$ such that

$$(6) \qquad \|x_t(\varphi)\| \leqslant e^{\beta t}\left[\|\varphi\| + \int_0^t |h(s)|\,ds\right], \quad t \geqslant 0.$$

This theorem is proved by using the nonatomic property of μ at 0 together with the contraction mapping principle to first show that (3) has a solution on a small interval to the right of $t = 0$. An application

of a result on the continuation of the solution then allows one to obtain the estimate (6) for $t \geqslant 0$.

If h is identically zero and $x = x(\varphi)$ is a solution of the homogeneous equation

$$x_0 = \varphi$$

$$x(t) = \gamma(\varphi) + g(x_t) + \int_0^t f(x_s)\, ds, \qquad t \geqslant 0, \tag{7}$$

then it follows from the uniqueness of the solution that $x_t(\varphi)$ is a continuous linear mapping of C_r into C_r for each fixed $t \geqslant 0$, and $x_t(\varphi)$ satisfies the semigroup property. If we define the linear operator $T(t)$ by

$$x_t(\varphi) \overset{\text{def}}{=} T(t)\varphi, \qquad t \geqslant 0, \tag{8}$$

then we can prove

Theorem 2. The family of linear operators $\{T(t),\ t \geqslant 0\}$ mapping C_r into C_r is a strongly continuous semigroup on $[0, \infty)$ with $T(0) = I$. In addition, the infinitesimal generator A of $T(t)$ is given by

$$(A\varphi)(\theta) = \begin{cases} \dot{\varphi}(\theta), & -r \leqslant \theta < 0, \\ g(\dot{\varphi}) + f(\varphi), & \theta = 0 \end{cases}$$

and the domain of A, $\mathscr{D}(A)$, consists of all functions φ in C_r with a continuous first derivative and $\dot{\varphi}(0) = g(\dot{\varphi}) + f(\varphi)$.

It is interesting to note that if φ is in $\mathscr{D}(A)$, then $T(t)\varphi$ is actually a continuously differentiable solution of the functional differential equation (5) with $h \equiv 0$.

It is easy to show that the spectrum of A, $\sigma(A)$, consists of only point spectrum and that λ is in $\sigma(A)$ if and only if λ satisfies the *characteristic equation*

$$\det \Delta(\lambda) = 0, \tag{9}$$

$$\Delta(\lambda) = \lambda I - \int_{-r}^{0} \lambda e^{\lambda\theta} d\mu(\theta) - \int_{-r}^{0} e^{\lambda\theta} d\eta(\theta).$$

Also, because A is a closed operator and a root λ_0 of (8) has finite multiplicity, one can show that the resolvent operator $(A - \lambda I)^{-1}$ has a pole of finite order at λ_0 and, thus, the generalized eigenspace of λ_0 has finite dimension. If $\mathfrak{N}(A)$ and $\mathscr{R}(A)$ denote, respectively, the null space and range of an operator A and the generalized

eigenspace of λ_0 is given by $\mathfrak{N}(A - \lambda_0 I)^k$, then it can be shown that the space C_r is decomposed as a direct sum of the subspaces $P \overset{\text{def}}{=} \mathfrak{N}(A - \lambda_0 I)^k$, $Q \overset{\text{def}}{=} \mathscr{R}(A - \lambda_0 I)^k$ each of which is invariant under both A and $T(t)$, $t \geqslant 0$. When C_r is decomposed in this way, we shall say C_r is decomposed by λ_0 as $C_r = P \oplus Q$ and write any element φ in C_r as $\varphi = \varphi^P + \varphi^Q$, φ^P in P, φ^Q in Q. If Φ is a basis for P, then there is a matrix B such that $A\Phi = \Phi B$ and, thus, $\Phi(\theta) = \Phi(0)\, e^{B\theta}$, $-r \leqslant \theta \leqslant 0$. Also, one easily shows that $T(t)\Phi(\theta) = \Phi(0)\, e^{B(t+\theta)}$, $-r \leqslant \theta \leqslant 0$, which implies that the solutions of (3) on the generalized eigenspace of a solution of (9) can be defined on $(-\infty, \infty)$ and that the action of the semigroup $T(t)$ on this subspace is essentially the same as an ordinary differential equation. The decomposition outlined here plays the same role as the Jordan canonical form in ordinary differential equations.

In the applications, it is necessary to have an explicit representation for the projection operator E_{λ_0} associated with the above decomposition. This can be obtained from the formula

$$E_{\lambda_0}\varphi = \frac{1}{2\pi i} \int_C (A - \lambda I)^{-1} \varphi\, d\lambda$$

where C is a circle in the complex plane which contains no point in $\sigma(A)$ except λ_0. As in the case of retarded functional differential equations (see [1] or [2]) the projection operator E_{λ_0} can also be obtained in the following way. Let R^n be the n-dimensional real or complex space of row vectors and define the operator A^* with $\mathscr{D}(A^*) \subset C([0, r], R^n)$ given by all ψ in $C([0, r], R^n)$ which are continuously differentiable with $\dot{\psi}(0) = \int_{-r}^{0} \dot{\psi}(-\theta)\, d\mu(\theta) - \int_{-r}^{0} \psi(-\theta)\, d\eta(\theta)$ and for ψ in $\mathscr{D}(A^*)$,

$$(A^*\psi)(s) = \begin{cases} -\dot{\psi}(s) \text{ for } 0 < s \leqslant r, \\ \int_{-r}^{0} \dot{\psi}(-\theta)\, d\mu(\theta) + \int_{-r}^{0} \psi(-\theta)\, d\eta(0) \quad \text{for} \quad s = 0. \end{cases}$$

For any ψ in $C([0, r], R^n)$, $\dot{\psi}$ continuous, and any φ in $C([-r, 0], R^n)$, define

$$(\psi, \varphi) = \psi(0)\varphi(0) - \int_{-r}^{0} \left[\frac{d}{d\xi} \int_0^{\xi} \psi(s - \xi)\, d\mu(\theta)\, \varphi(s)\, ds \right]_{\xi=\theta} - \int_{-r}^{0} \int_0^{\theta} \psi(s - \theta)\, d\eta(\theta)\, \varphi(s)\, ds.$$

For ψ in $\mathcal{D}(A^*)$, φ in $\mathcal{D}(A)$, it follows that $(\psi, A\varphi) = (A^*\psi, \varphi)$. To obtain the projection operator E_{λ_0}, one proceeds as follows: if we let $\Phi = (\varphi_1, \ldots, \varphi_p)$ be a basis for the generalized eigenspace $P = \mathfrak{N}(A - \lambda_0 I)^k$ of λ_0 and let $\Psi = \operatorname{col}(\psi_1, \ldots, \psi_p)$ be a basis for $\mathfrak{N}(A^* - \lambda_0 I)^k$, then $E_{\lambda_0}\varphi = \Phi(\Psi, \varphi)$ for all φ in $C([-r, 0], R^n)$.

Another important relation in ordinary differential equations is the variation of constants formula. By using the fact the solution $x^*(t, h)$ of (3) with $\varphi = 0$ is a continuous linear mapping of $\mathcal{L}_1([0, t], R^n)$ into R^n, one can show that

$$x^*(t, f) = \int_0^t U(t-s)\, h(s)\, ds$$

where $U(t) = dV(t)/dt$ almost everywhere and $V(t)$ is the matrix solution of (3) with $\varphi = 0$ and f equal to the identity matrix. Because equation (3) is linear, it follows that the solution $x = x(\varphi)$ satisfies

$$x(t) = [T(t)\varphi](0) + \int_0^t [d_s V(t-s)]\, h(s).$$

Using the fact that $V(t) = 0$ for $-r \leqslant t \leqslant 0$, we also obtain

$$x_t(\theta) = [T(t)\varphi](\theta) + \int_0^t [d_s V_{t-s}(\theta)]\, f(s), \qquad -r \leqslant \theta \leqslant 0,$$

which can be written more compactly as

$$x_t = T(t)\varphi + \int_0^t [d_s V_{t-s}]\, h(s). \tag{10}$$

Equation (10) is called the *variation of constants formula* for (3).

If λ_0 is a solution of (8) and C is decomposed by λ_0 as $P \oplus Q$, then it can also be shown that

$$x_t^P = T(t)\varphi^P + \int_0^t [d_s V_{t-s}^P]\, h(s),$$

$$x_t^Q = T(t)\varphi^Q + \int_0^t [d_s V_{t-s}^Q]\, h(s). \tag{11}$$

If g is identically zero in (3), then we have seen that equation (3) is equivalent to (4) and the variation of constants formula (10) can be written as

$$x_t = T(t)\varphi + \int_0^t T(t-s)\, K_0 f(s)\, ds$$

where $T(t)K_0$ is the solution of (4) with initial value at 0 given by $K_0(\theta) = 0$ for $-r \leqslant \theta < 0$, $K_0(0) = I$, the identity matrix. This is the standard manner of writing the variation of constants formula for (4) as given in [3] and [4]. For the equation of neutral type, i.e., f, g not identically zero, this formula also coincides with the one given for some special cases in [3].

To apply these results in special situations, it is necessary to obtain precise exponential estimates of $T(t)\varphi^Q$ and V_t^Q. To obtain these estimates, further restrictions are imposed on the matrix function μ defined in (1). More specifically, we suppose that μ has no singular part; that is, $\mu = \mu_1 + \mu_2$, where μ_1 is a step function and μ_2 is absolutely continuous. Suppose γ is a real number such that there are only a finite number of roots $\lambda_1, \ldots, \lambda_p$ of (9) with real parts greater than γ. Let P be the algebraic sum of the generalized eigenspaces of the λ_j and decompose C as $P \oplus Q$ in the above manner. Using explicit representations of $T(t)\varphi$ and $\int_0^t [d_s V_{t-s}]\, h(s)$ as inverse Laplace transforms, one can then prove: for any $\varepsilon > 0$ there is a $K > 0$ such that

$$\|T(t)\varphi\| \leqslant K e^{(\gamma+\varepsilon)t}[\|\varphi\| + \|\dot{\varphi}\|], \qquad \varphi \in Q, \tag{12}$$

for all continuously differentiable functions φ and

$$\Big\| \int_0^t [d_s V_{t-s}^Q]\, h(s) \Big\| \leqslant K \Big[\int_0^t (e^{(\gamma+\varepsilon)(t-s)} |h(s)|)^2\, ds \Big]^{1/2}. \tag{13}$$

It is not clear that the estimate (12) actually reflects the true nature of an homogeneous equation (3) in the sense that precise estimates of this type can be obtained only by assuming some differentiability conditions on the initial function φ. In fact, if one could prove that

$$\sigma(T(t)) = e^{t\overline{\sigma(A)}} + \{0\}$$

then the estimate (12) could be sharpened. This question needs to be studied in greater detail.

Having developed the theory of linear systems as above, one can discuss stability and oscillations in certain classes of nonlinear equa-

tions. The equations that we have considered are

$$x(t) = \gamma(\varphi) + g(x_t) + \int_0^t f(x_s)\,ds + \int_0^t F(x_s)\,ds, \tag{14}$$

$$x(t) = \gamma(\varphi) + g(x_t) + \int_0^t f(x_s)\,ds + \varepsilon \int_0^t G(s, x_s)\,ds \tag{15}$$

where f, g are the same as before, $F(\varphi) = o(\|\varphi\|)$ as $\|\varphi\| \to 0$, $G(t, \varphi)$ is an arbitrary continuous function and ε is a small parameter. Using estimates (12) and (13) one can show that the solution $x = 0$ of (14) is asymptotically stable if all roots λ of (9) satisfy $\operatorname{Re}\lambda \leqslant -\delta < 0$. Of course, stability is defined relative to continuously differentiable perturbations in the initial data. For equation (15) we develop an analogue of the method of averaging to discuss the existence of periodic and almost periodic solutions.

Brown University,
Providence, USA

REFERENCES

[1] Hale J., Linear functional differential equations with constant coefficients, Contributions to differential equations, 2 (1963), 291-319.
[2] Шиманов С. М., К теории линейных дифференциальных уравнений с запаздыванием, *Дифференциальные уравнения*, 1 (1965), 102-116.
[3] Bellman R., Cooke K., Differential-difference equations, Academic Press, 1963.
[4] Halanay A., Differential equations, Academic Press, 1965.

ДИНАМИЧЕСКИЕ СИСТЕМЫ С ТРАНСВЕРСАЛЬНЫМИ СЛОЕНИЯМИ

Д. В. АНОСОВ[1])

Аксиоматизируя свойства неустойчивости, присущие геодезическому потоку на замкнутом римановом многообразии отрицательной кривизны, автор ввел понятие У-системы, т. е. динамической системы, поведение траекторий которой в окрестности любой фиксированной траектории напоминает поведение траекторий возле седла. Более общим является понятие динамической системы

[1]) Полное изложение доклада будет опубликовано в *Успехах математических наук*.

с трансверсальным слоением, т. е. инвариантным относительно этой системы слоением, в слоях которого под действием преобразований системы происходит сжатие или расширение. В докладе будет дан обзор работ автора по теории У-систем и смежных результатов других авторов о различных системах с трансверсальными слоениями.

Математический институт им. В. А. Стеклова,
Москва, СССР

ПРОБЛЕМА УСТОЙЧИВОСТИ И ЭРГОДИЧЕСКИЕ СВОЙСТВА КЛАССИЧЕСКИХ ДИНАМИЧЕСКИХ СИСТЕМ

В. И. АРНОЛЬД

Классическая динамическая система состоит из гладкого многообразия M и однопараметрической группы g^t его диффеоморфизмов $g^t : M \to M$.

Дифференциальные уравнения классической механики, например в задаче трех тел, доставляют немало примеров систем (M, g^t). Эти системы, впрочем, принадлежат более узкому классу *гамильтоновых систем* (M, g^t, ω^1); многообразие $M = M^{2n}$ четномерно, на нем задана каноническая структура, т. е. фиксирована некоторая 1-форма ω^1 (заданная с точностью до полного дифференциала однозначной функции), имеющая невырожденную производную $d\omega^1 = \sum_{i=1}^{n} dp_i \wedge dq_i$ в надлежащих локальных координатах p, q на M; диффеоморфизмы g^t канонические: для любой замкнутой кривой γ

$$\oint_{\gamma} \omega^1 = \oint_{g^t\gamma} \omega^1.$$

Промежуточное положение между общими и гамильтоновыми системами занимают *системы с инвариантной мерой* (M, g^t, τ): на M фиксируется невырожденная дифференциальная форма τ максимальной размерности, и g^t сохраняет эту форму.

Теория динамических систем ставит свой задачей изучение поведения типичной орбиты типичной системы (M, g^t), (M, g^t, τ) или (M, g^t, μ^1). Все три теории совершенно различны. Так, для общих систем одна из типичных возможностей — асимптотически устойчивые движения, притягивающие соседей. В системах же с инвариантной мерой асимптотически устойчивые движения невозможны. Менее известны дополнительные особые свойства

гамильтоновых систем; они проявляются, например, в консервативном противостоянии эволюции.

В настоящее время мы не располагаем сколько-нибудь удовлетворительной общей теорией ни в одном из трех случаев. Хорошо изучены лишь некоторые специальные системы. Можно пытаться разобраться в ситуации, переводя специальные системы в общие посредством малого возмущения.

1. Теория возмущений

Пусть $M = T^k \times \mathbf{R}^l$ — прямое произведение k-мерного тора T^k на евклидово пространство $\mathbf{R}^l$, $\varphi = \varphi_1, \ldots, \varphi_k \pmod{2\pi}$ — угловые координаты на T^k, $I = I_1, \ldots, I_e$ — декартовы координаты $\mathbf{R}^l$. Рассмотрим в качестве «невозмущенной системы» систему, определенную дифференциальными уравнениями

$$\frac{d\varphi}{dt} = \omega(I), \qquad \frac{dI}{dt} = 0, \qquad \omega = \omega_1, \ldots, \omega_k \in \mathbf{R}^k. \tag{1}$$

Очевидно, каждый тор $I = \text{const}$ инвариантен. Если частоты на нем несоизмеримы,

$$n_1\omega_1 + \ldots + n_k\omega_k \neq 0 \text{ для целых } n_i,\ n_1^2 + \ldots + n_k^2 \neq 0,$$

то орбита $\varphi(t)$ всюду плотна на торе и движение называется квазипериодическим с k частотами. Если частоты зависимы, то замыкание орбиты есть тор с числом измерений меньше k (резонанс).

Для исследования «возмущенной системы»

$$\frac{d\varphi}{dt} = \omega(I) + \varepsilon f(I, \varphi), \qquad \frac{dI}{dt} = \varepsilon F(I, \varphi) \tag{2}$$

классическая (т. е. нестрогая) теория возмущений предписывает составить среднее значение $\overline{F}(I) = (2\pi)^{-k} \oint\limits_{T^k} F(I, \varphi)\, d\varphi$ и «эволюционное уравнение»

$$\dot{J} = \varepsilon \overline{F}(J). \tag{3}$$

Считается, что для $\varepsilon \ll 1$ различие между решениями $I(t)$ и $J(t)$ систем (2) и (3) с одинаковыми начальными условиями невелико, по крайней мере на большом промежутке времени $0 < t < 1/\varepsilon$.

Проблема 1. *Как связаны $I(t)$ и $J(t)$, $0 < t < 1/\varepsilon$?*

Кроме особых случаев, когда система (3) имеет асимптотически устойчивое движение, известно лишь очень немногое:

Теорема 1. *Для $k = 1$, $\omega \neq 0$ имеем* $|I(t) - J(t)| < C\varepsilon$ [1]. *Для $k = 2$, $d(\omega_1/\omega_2)/dt \neq 0$ имеем* $|I(t) - J(t)| < C\sqrt{\varepsilon}\log^2 1/\varepsilon$ [2].

Случай $k > 2$ совсем плохо изучен. Известно лишь, что для систем с инвариантной мерой $dI\,d\varphi$ при малом ε для большинства начальных условий $|I(t) - J(t)|$ мал (см. работы Д. В. Аносова [3] и Т. Касуга [4]).

Пусть теперь система (2) гамильтонова: $k = l$, $\omega^1 = I\,d\varphi$.

2. Возмущения гамильтоновых систем

Гамильтонова система (2) имеет вид

$$\frac{d\varphi}{dt} = \frac{\partial H}{\partial I}, \qquad \frac{dI}{dt} = -\frac{\partial H}{\partial \varphi}, \qquad H = H_0(I) + \varepsilon H_1(I, \varphi). \tag{4}$$

Поэтому усреднение дает нуль:

$$-\overline{F}(I) = (2\pi)^{-k} \oint\limits_{T^k} \frac{\partial H_1}{\partial \varphi}\, d\varphi = 0.$$

Следовательно, эволюционная система нулевая и в первом приближении эволюции нет. Более того, справедлива

Теорема 2 (А. Н. Колмогоров [5]). *Пусть H_0, H_1 аналитичны при $I \in G$, $|\operatorname{Im}\varphi| < \rho$ и в G*

$$\det\left|\frac{\partial \omega}{\partial I}\right| = \det\left|\frac{\partial^2 H_0}{\partial I^2}\right| \neq 0.$$

Тогда бо́льшая часть $G \times T^k$ заполнена при достаточно малых ε инвариантными k-мерными торами системы (4).

Теорема 2 обобщается на некоторые важные для небесной механики случаи, когда $\partial^2 H_0/\partial I^2 = 0$. Таким путем, например, найдены квазипериодические движения в задаче многих тел [6]. Предположение аналитичности также можно ослабить — достаточно существования нескольких сот производных (Ю. Мозер [7]).

Множество инвариантных торов, о которых идет речь в теореме 2, имеет всюду плотное открытое дополнение.

Проблема 2. *Как ведут себя орбиты из этого дополнительного множества? В частности, верно ли, что для них нет эволюции в s-м приближении, т. е. $|I(t) - J(t)| \ll 1$, $0 < t < (1/\varepsilon)^s$?*

Недавняя работа о «формальной устойчивости» (Глимм [8]), видимо, оставляет этот важный вопрос открытым. В старой астрономической литературе он считался решенным положительно, однако математически строгое доказательство мне неизвестно, исключая лишь случай двух степеней свободы, когда об устойчивости можно судить по наличию инвариантных торов.

3. Инвариантные торы и устойчивость

В случае $k=2$ инвариантные торы теоремы 2 делят множество уровня энергии $H=\text{const}$ системы (4). Более того, если отношение частот меняется вдоль $H=\text{const}$, т. е.

$$\det\begin{vmatrix} \frac{\partial^2 H_0}{\partial I^2} & \frac{\partial H_0}{\partial I} \\ \frac{\partial H_0}{\partial I} & 0 \end{vmatrix} \neq 0,$$

то таких торов много на каждом уровне энергии. Эти торы запирают каждую орбиту в узком слое, ограниченном ими, поэтому движение устойчиво для всех начальных условий в том смысле, что

$$|I(t)-J(t)|<C(\varepsilon), \quad -\infty<t<+\infty, \quad C(\varepsilon)\to 0 \text{ при } \varepsilon\to 0.$$

При $k\geqslant 3$ торы T^k не делят $(2k-1)$-мерный уровень H.

Гипотеза. *«Общим случаем» для гамильтоновой системы* (4) *с $k\geqslant 3$ является такой, когда для любой пары окрестностей торов $I=I'$, $I=I''$ с общим $H_0(I')=H_0(I'')$ найдется при достаточно малом ε орбита, пересекающая обе окрестности.*

Что такие орбиты, соединяющие окрестности далеких торов, вообще возможны, показывает пример [9] системы, удовлетворяющей всем условиям теоремы об инвариантных торах и потому устойчивой для большинства начальных условий, но неустойчивой для «резонансных» начальных условий. Механизм «переходных цепочек», действующий в этом примере, носит, вероятно, весьма общий характер.

Для построения неустойчивой орбиты в [9] используется семейство инвариантных торов однократного резонанса (размерности $n-1$). Первые общие теоремы о существовании таких торов получены в самое последнее время В. К. Мельниковым, Ю. Мозером, А. М. Леонтовичем, Г. А. Красинским. Чтобы сделать следующий шаг к доказательству высказанной гипотезы, нужно разобраться в переходе между однократным и двукратным резонансом (резонанс порядка 3 и выше несущественен по топологическим соображениям). Модельной задачей здесь может служить построение переходной цепочки, соединяющей две периодические орбиты и положение равновесия в системе на $T^2\times\mathbf{R}^2$

$$H=\frac{I_1^2\pm I_2^2}{2}+U(\varphi_1,\varphi_2).$$

При построении этой цепочки полезна доказанная Е. В. Гайдуковым [18] элементарная

Теорема. *Через каждую точку риманова тора T^2 проходит геодезическая, асимптотическая к замкнутой геодезической, гомотопной данной.*

4. Геометрическая теорема Пуанкаре

Другим далеко не завершенным отделом теории многомерных гамильтоновых систем является теория периодических орбит. Кажутся правдоподобными, например, такие обобщения «последней теоремы» Пуанкаре.

А. *Пусть $A\colon q\to q+f(q)$ — диффеоморфизм тора $T^2=\{q_1, q_2 \bmod 2\pi\}$, сохраняющий меру $dq_1\wedge dq_2$ и центр тяжести $\left(\oint f(q)\,dq\wedge dq=0\right)$. Тогда A имеет по крайней мере 4 неподвижные точки, если считать с кратностями, и по крайней мере 3 геометрически различные неподвижные точки.*

Б. *Пусть $\Omega=T^k\times B^k$, $T^k=\{g \bmod 2\pi\}$, $B^k=\{p\in\mathbf{R}^k, |p|\leqslant 1\}$ — торовое кольцо с канонической структурой $\omega^1=pdq$ и $A\colon\Omega\to\Omega$ — канонический диффеоморфизм, гомотопный тождественному и такой, что каждая сфера $q\times\partial B^k$ зацеплена со своим образом на накрывающей края $T^k\times\partial B^k$. Тогда A имеет по крайней мере 2^k неподвижных точек, если считать с кратностями, в том числе $k+1$ геометрически различных.*

Доказательства основаны на теории Морса — Люстерника — Шнирельмана; их удается провести лишь при дополнительных ограничениях (см. [10]; в задаче А ограничение $\frac{\partial f}{\partial q}\xi\neq -2\xi$).

5. Эргодические свойства

Проблема. *Выяснить эргодические свойства движений в области, дополнительной к инвариантным торам системы* (4). *В частности, положительна ли энтропия этой системы?*

До сих пор доказана «типичность» только двух типов поведения динамических систем с инвариантной мерой: квазипериодических, рассмотренных выше, и неустойчивых, которым посвящен доклад Д. В. Аносова на этом конгрессе. В то же время известен ряд моделей промежуточного типа: орициклические потоки (см. Грин [11]), нильпотоки (Грин, Ауслендер, Хан [12]), системы с квазидискретным спектром (Абрамов [13]), перекладывания отрезков [14] и т. д. Для исследования этих «систем с медленным перемешиванием» создан специальный аппарат — метод периодических аппроксимаций (А. Б. Каток и А. М. Степин [14]), 2^n-энтропия (А. Г. Кушниренко [15]) и т. п. Однако неизвестно, как ведут себя подобные системы при возмущении и не распадаются ли они на компоненты с дискретным спектром и компоненты с положительной энтропией.

Было бы очень интересно, например, исследовать возмущения орициклических потоков, отделяющих в алгебраическом случае системы с дискретным спектром от систем с экспоненциальным разбеганием: это позволило бы проследить переход от системы с инвариантными торами к K-системе.

К этому же кругу вопросов примыкает вопрос о непрерывности энтропии классической динамической системы как функции системы; доказана лишь ограниченность энтропии диффеоморфизма компактного многообразия (А. Г. Кушниренко [16]).

Особый интерес представляет изучение эргодических свойств систем (4) в случае, когда ε не мало. В этом случае был бы полезен численный эксперимент, а он показывает, что мера инвариантных торов с ростом ε быстро убывает (см., например, работу Хэнона и Хейлса [17]).

Московский университет,
Москва, СССР

ЛИТЕРАТУРА

[1] Kruscal M., Asymptotic theory of Hamiltonian and other systems with all solutions nearly periodic, USA, 1961. Русский перевод: Крускал М., Адиабатические инварианты, ИЛ, М., 1962.
[2] Арнольд В. И., *ДАН СССР*, **161** (1965), 9-12.
[3] Аносов Д. В., *Изв. АН СССР, сер. матем.*, **24** (1960).
[4] Kasuga T., *Proc. Japan Acad.*, **37** (1961), № 7.
[5] Колмогоров А. Н., *ДАН СССР*, **98** (1954), № 4.
[6] Арнольд В. И., *ДАН СССР*, **145** (1962), 487-490; *УМН*, **18** (1963), № 5, № 6.
[7] Moser J., *Göttingen Nachr.*, (1962), № 1.
[8] Glimm J., *Comm. Pure Appl. Math.*, **17** (1963), № 4.
[9] Арнольд В. И., *ДАН СССР*, **156** (1964), 9-12.
[10] Арнольд В. И., *C. R. Acad. Sci. Paris*, **261** (1965), 3719.
[11] Green L. W., *Bull. Amer. Math. Soc.*, **72** (1966), 44-49.
[12] Auslander L., Hahn F., Green L., *Ann. of Math. Studies*, **53** (1963).
[13] Абрамов Л. М., *Изв. АН СССР, сер. матем.*, **26** (1962), 513-531.
[14] Каток А. Б., Степин А. М., Тезисы кратких сообщений ICM, Москва, 1966, секция 6.
[15] Кушниренко А. Г., Тезисы кратких сообщений ICM, Москва, 1966, секция 6.
[16] Кушниренко А. Г., *ДАН СССР*, **161** (1965), 37-38.
[17] Hénon M., Heiles C., *Astronom. J.*, **69** (1964), 73-79.
[18] Гайдуков Е. В., *ДАН СССР*, 1966.

7 Дифференциальные уравнения с частными производными
Partial differential equations
Equations différentielles aux dérivées partielles
Partielle Differentialgleichungen

ALGEBRAS OF SINGULAR INTEGRAL OPERATORS

A. P. CALDERÓN

In dealing with differential operator with non-smooth coefficients one is led to consider pseudo-differential operators and singular integral operators with non-smooth symbols and kernels. Such pseudo-differential operators, however, do not have desirable algebraic properties. This becomes evident if one observes that in particular differential operators cannot be composed freely unless one asssumes their coefficients to be infinitely differentiable. The situation is different with singular integral operators. Here it is possible to construct a theory, which is fairly general as far as regularity assumptions is concerned, preserving at the same time an algebra structure and a relatively simple approximate functional calculus. In what follows we will give a brief description of this algebra and of the tools employed in its construction.

The singular integral operators we have in mind are defined in terms of ordinary singular integral operators and a fractional integration operator I which acts on the space of temperate distributions and is given by

$$(If)^{\wedge} = \hat{f} d(x)^{-1}$$

where $\hat{f}$ is the Fourier transform of f and $d(x)$ is a positive, infinitely differentiable spherically symmetric function which coincides with $|x|$ for $|x| \geqslant 1$. This kind of fractional integration is closely related to Bessel integration, although some of its formal properties are somewhat simpler.

In what follows we will denote the spaces L^p, $1 \leqslant p \leqslant \infty$, by $L(1/p)$. The space of functions of the form $f + g$ with $f \in L(1/p)$ and $g \in L(1/q)$ will be denoted by $L(1/p) + L(1/q)$. This space has natural norm with respect to which it is complete.

Let now m be a positive integer, $m < q \leqslant \infty$ and also $q \geqslant n$ if $m > 1$, where n is the number of real variables of the functions under consideration, and consider the class $\mathscr{S}_m^{q,r}$, $r > n-1$, of operators of the form

$$K = \sum_{j=0}^{m-1} K_j I^j + S$$

where i) K_j is a singular integral operator

$$(K_j f)(x) = a_j(x) f(x) + \lim_{\varepsilon \to 0} \int_{|x-y|>\varepsilon} k_j(x, x-y) f(y)\, dy$$

where $a_j(x)$ has distribution derivatives of order i in $L(0) + L[(i+j)/q]$ for $0 \leqslant i \leqslant m-j$, and k_j has the property that

$$\Big[\int_{|z|=1} |\partial_x^\alpha \partial_z^\beta k_j(x, z)|^r dz \Big]^{1/r},$$

where $\partial_x^\alpha \partial_z^\beta k_j$ denotes the derivative of k_j of order α with respect to x and of order β with respect to z and dz denotes the surface area element of the sphere $|z| = 1$, belongs to $L(0) + L[(j + |\alpha|)/q]$ for $0 \leqslant |\alpha| \leqslant m-j$ and $0 \leqslant |\beta| \leqslant m-j-|\alpha|+1$;

ii) S is an operator on C_0^∞ which can be extended continuously to a bounded operator on $L(1/p)$ for $(m-1)/q < 1/p < 1$, and to a bounded operator from $L(1/p)$ to $L(1/p + m/q)$ for $0 < 1/p < 1 - m/q$, and such that $S\left(\frac{\partial}{\partial x}\right)^\alpha$ can also be extended to a bounded operator from $L(1/p)$ to $L(1/p + m/q)$ for $0 < 1/p < 1 - m/q$ and all α with $|\alpha| = m$.

The main result about $\mathscr{S}_m^{q,r}$ is that operators in this class can be extended to bounded operators in $L(1/p)$ for $(m-1)/q < 1/p < 1$, and that the class is an algebra under composition. The approximate functional calculus of singular integral operators as described in [1], section 8, also applies to this more general situation. For the sake of completeness, let us describe briefly this functional calculus. For this purpose let us consider differential operators in R^n whose coefficients are homogeneous functions of non-positive integral degree. The weight of a monomial differential operator of this kind is defined as its order minus the degree of its coefficient. Let now $\mathscr{A}$ be the class of differential operators which are sums of monomial differential operators of weight less than m, and let us define the product of two monomial differential operators to be their ordinary composition if the sum of their weights is less than m or to be zero otherwise. With this product $\mathscr{A}$ becomes an algebra. Now, with each operator K in $\mathscr{S}_m^{q,r}$ we associate its characteristic, which is a function of x with values in $\mathscr{A}$ given by

$$\chi(K) = \sum \frac{1}{\alpha!} \Big(\frac{1}{2\pi i} \frac{\partial}{\partial z}\Big)^\alpha [\partial_x^\alpha \sigma(K_j)(x, z) |z|^{-j}], \quad |\alpha| + j \leqslant m-1,$$

where $\partial_x^\alpha \sigma(K_j)(x, z)$ denotes the α-th derivative with respect to x of the symbol $\sigma(K_j)$ of K_j. The mapping $K \to \chi(K)$ is then a homomorphism of $\mathscr{S}_m^{q,r}$ into the algebra of functions of x with values in $\mathscr{A}$, and the kernel of the homomorphism consists of the m-times smoothing operators S described in ii).

The main tools employed in establishing the properties of $\mathscr{S}_m^{q,r}$ are the well known rotation method which affords a representation of an n-dimensional singular integral operator with odd kernel as an integral in one-dimensional Hilbert transforms and an extension of the results in [2] which can be stated as follows. Let $a(t)$ be a function on the real line with an m-th order derivative in $L(1/q)$, $1 > 1/q \geqslant 0$, and let $T(s, t)$ be the m-th remainder of the Taylor expansion of $a(t)$ at the point t. Then the integral

$$\int_{|t-s|>\varepsilon} (t-s)^{-m-1} T(s, t) f(s)\, ds$$

represents a bounded operator from $L(1/p)$ into $L(1/p + 1/q)$ with norm independent of ε, provided that $0 < 1/p \leqslant 1/p + 1/q < 1$.

References to the work on pseudo-differential operators by Kohn, Nirenberg, Hörmander and Seeley will be found in the bibliography of section 8 of [1] below.

The University of Chicago,
Chicago, USA

REFERENCES

[1] Calderón A. P., Singular integrals, *Bull. Am. Math. Soc.*, **72** (1966), 426-65.
[2] Calderón A. P., Commutators of singular integral operators, *Proc. Nat. Acad.*, **53**, n° 5 (1965), 1092-99.

HYPERSURFACES OF MINIMAL MEASURE IN PLURIDIMENSIONAL EUCLIDEAN SPACES

ENNIO DE GIORGI

In this talk I shall deal with some recent developments of the theory of minimal hypersurfaces which are embedded in the n-dimensional euclidean space (with $n > 3$). I shall not deal with the results for surfaces in R^3 (or, more generally, for 2-dimensional varieties in R^n)

which are dealt with, for example, in the recent papers [4], [22] and in [21], and in the numerous references there given.

However, I think it is worthwhile to observe that many typical methods in the theory of minimal surfaces (conformal representation,..;) do not appear to work well when we pass to the many-dimensional case. This, of course, has led many workers in this field to different approaches and methods.

The following exposition, even with the restrictions which I mentioned at the outset, will definitely lack of completeness and will be rather vague: I apologize for it. It is my hope, however, that what I am saying will prove to be useful as a brief indication for the non specialists of some lines along which the theory is developing.

I shall deal mainly with the Plateau problem in cartesian form, with the extension of Bernstein's theorem to the m-dimensional case, and with removable singularities of a minimal hypersurface.

I will add some concluding remarks on some recent formulations of the general Plateau problem which are interesting in themselves and have shown their soundness also for the problem in cartesian form.

The Plateau problem in cartesian form may be expressed as follows: let Ω be an open bounded set in R^n and let $g(x)$ be a real continuous function on the boundary $\partial\Omega$ of Ω; does there exist then a continuous functi on $u(x)$ on the closure $\overline{\Omega}$ of Ω with continuous and summable first derivatives on Ω and such that

$$g(x) = u(x), \text{ for every } x \in \partial\Omega, \tag{1}$$

which minimizes the integral

$$\int_{\Omega} \sqrt{1 + \sum_{h=1}^{m} \left(\frac{\partial u}{\partial x_h}\right)^2}\,? \tag{2}$$

It is easy to show that if such a function $u(x)$ does exist, then it is unique and, due to a result of Morrey (see [20]), it is continuous on Ω with its derivatives of every order and satisfies the Euler equation

$$\sum_{h=1}^{m} \frac{\partial}{\partial x_h} \left(\frac{\frac{\partial u}{\partial x_h}}{\sqrt{1 + \sum_{h=1}^{m} \left(\frac{\partial u}{\partial x_h}\right)^2}} \right) = 0. \tag{3}$$

Conversely, any solution of equation (3) with the boundary condition (1), if it has first derivatives summable on Ω, is also a solution

of the variational problem. Moreover, as Hopf has shown (see [15]), the solution is analytic on Ω.

Many theorems have been given ensuring the existence of the minimum for the functional (2) under fairly general hypothesis.

If Ω is convex and $g(x)$ is regular enough, Gilbarg, Hartman, Miranda, Stampacchia (see, respectively, [12], [14], [17], [26]) have given a solution which may be expressed as follows. We say that $g(x)$ satisfies the B.S.C. (bounded slope condition) hypothesis if two vector functions $a(\eta)$ and $b(\eta)$ exist on $\partial\Omega$ in such a way that there is a constant p such that

$$|a(\eta)| \leqslant p, \quad |b(\eta)| \leqslant p \tag{4}$$

and

$$\sum_{i=1}^{m} a_i(\eta)(x_i - \eta_i) \leqslant g(x) - g(\eta) \leqslant \sum_{i=1}^{m} b_i(\eta)(x_i - \eta_i) \tag{5}$$

for every $x, \eta \in \partial\Omega$; then, if $g(x)$ satisfies the B.S.C. hypothesis, then the functional (2) has a minimum which satisfies condition (1) (i.e., there exists a solution of equation (3) with condition (1)); moreover, the solution is uniformly lipschitzian on $\overline{\Omega}$. Hartman (see [13]) has proved that the B.S.C. hypothesis is equivalent to a $(m+1)$-points condition, natural extension of the 3-points condition. Miranda (see [17]) has dealt with the case of a continuous function $g(x)$ on a uniformly convex domain Ω: his formulation of the problem, in addition, is weaker. He has shown that there exists a continuous function on $\overline{\Omega}$, satisfying the boundary condition (1), which minimizes the Lebesgue area of the hypersurface

$$x_{m+1} = u(x_1, \ldots, x_m); \quad (x_1, \ldots, x_m) \in \overline{\Omega}. \tag{6}$$

His proof makes use of a uniform approximation of the boundary datum with functions satisfying the B.S.C. hypothesis. The last result has been sharpened by the author himself (see [18]) showing that, in the case $m = 3$, the solution $u(x)$ is analytic on Ω (so it turns out to be a solution of (3) with condition (1)).

It is worth noticing that while the usual methods for uniformly elliptic problems achieve the regularization of solutions which are lipschitzian on Ω, the same methods do not seem to be adequate to deal with continuous functions in the problem of minimizing the Lebesgue area: so, some results from the theory of the general Plateau problem had to be used (see [27]).

An extension of Miranda's results to the case $m = 4$ has been obtained by Almgren using some results of [2]. As far as I know, no result is available when $m > 4$.

Quite recently Jenkins and Serrin (see [16]) have considered the case of non convex domains; they have shown that the essential hypothesis on Ω is that it should have the boundary with mean curvature of constant sign.

As to the extension of the Bernstein's theorem, De Giorgi (see [6]) has shown that, if $m = 3$, every function which is solution of (3) on the whole of R^m is actually a first degree polynomial. The general setting of the problem follows the line of Fleming [10] and uses some results related to the general Plateau problem (see [27]). De Giorgi's result has been extended to the case $m = 4$ by Almgren (see [2]). To my knowledge, the possible extensions of the Bernstein's theorem are an open question.

As to the problem of removable singularities, Finn has proved (see [9]) that if a function solves the equation (3), then it cannot have isolated singularities. Using the same methods, De Giorgi and Stampacchia (see [7]) have extended this result showing that if A is an open set in R^m and if K is a compact subset of A whose $(m-1)$-dimensional Hausdorff measure is zero, then every function which solves equation (3) on $A - K$ may be analytically extended to the whole of A.

I believe that is still open the problem of removing singularities in the solutions of equation (3) when K is closed in A (not necessarily compact) and of zero $(m-1)$-dimensional measure, e.g. if K is the intersection of a straight line with A and $m > 2$.

I turn now from the cartesian form of the Plateau problem to the same problem in general form. I should say that the parametric approach to the problem for varieties of dimension higher than 2 seems to be rather difficult. Thus many authors have been led to the formulation of the general Plateau problem in various classes of generalized varieties; among the early papers along these lines I remember those of Caccioppoli (see [3]) and Young (see [28]).

The ideas of Caccioppoli have influenced in turn the work of De Giorgi, Miranda, Triscari. In the Miranda's most recent formulation of the problem (see [19]), the study of oriented minimal hypersurfaces has led to the search for the minimum of the integral of $|\operatorname{grad} g(x)|$ where $g(x)$ is a function whose derivatives are measures (in the sense of the distribution theory). More precisely, having denoted by $\int_K |Dg|$ the total variation on K of the vectorial measure $Dg = (D_1 g, \ldots, D_n g)$, gradient of the function $g(x)$, we consider an open set $A \subset R^n$ and a set $E \subset R^n$ whose characteristic function $\varphi(x, E)$ has derivatives which are measures. We define E as having oriented frontier of minimum measure on A if and only if, for every compact set $K \subset A$, $\varphi(x, E)$ is the minimum of the integral $\int_K |Dg|$

in the class of all functions $g(x)$ whose derivatives are measures and such that, for every $x \in R^n - K$, $g(x) = \varphi(x, E)$. Then the following regularization theorem holds: the part of the boundary of E contained in A is constituted of analytic hypersurfaces and of a singular set (which may be empty) whose $(n-1)$-dimensional Hausdorff measure is zero.

More generally, Federer and Fleming (see [8], [10]) have considered also the k-dimensional varieties in R^n with $0 \leqslant k \leqslant n$. They have considered the k-dimensional currents in R^n (that is, linear functionals on the space of k-th order differential forms) and they have defined the usual boundary operator as dual of the exterior differential. Moreover, they have defined a norm $M(T)$, called the mass of the current T, which, in the case of regular oriented varieties, coincides with the elementary measure. As rectifiable currents they have defined the elements of the closure with respect to the norm M of the set of all regular oriented varieties. In this setting, the general Plateau problem is reduced to the search for rectifiable currents with prescribed boundary and minimum mass, and the authors have given existence theorems. In the case $k = n - 1$, except on the boundary, a minimal current is decomposable into analytic hypersurfaces and a singular set (which may be vacuous) whose k-dimensional measure is zero.

Finally, Young (see [28]) has considered functionals of a more general type than the currents: precisely, his generalized varieties V are linear functionals defined on the real functions $f(x, y)$, where $x \in R^n$, y is a k-vector, f is continuous and homogeneous in the sense that $f(x, ty) = tf(x, y)$, for $t \geqslant 0$, such that, if $f(x, y) \geqslant 0$ for every (x, y), then $V(f) \geqslant 0$; then he considers minimizing and extremal varieties for many variational problems and gives a general cone-inequality.

All what has been said so far is related to the concept of oriented variety.

Non-oriented varieties have been considered by Reifenberg (see [23]). He calls k-dimensional variety a non-oriented set; he considers as its area the k-dimensional Hausdorff measure and defines the notion of "variety with given boundary" through the Čech homology. Following a theorem on the existence of the minimum area and some partial results on regularization (see [23]), Reifenberg proves (see [24], [25]) that a minimal variety, with the exception of the boundary, is decomposable into analytic varieties and a singular set (possibly empty) with zero k-dimensional measure.

A somewhat different approach to non-oriented varieties may be found in the papers by Ziemer (see [30]) and Fleming (see [11]). The latter author defines as flat chains the elements of the completion of polyhedral chains with coefficients in a metric finite abelian group,

with respect to the Whitney norm

$$W(P) = \inf \{M(Q) + M(R): P = Q + \partial R\}$$

where P, Q are polyhedral chains of dimension k, R is of dimension $k+1$, ∂R is the boundary of R, M is the elementary measure.

Finally, I wish to present briefly Almgren's theory of *varifolds*.

A k-dimensional varifold V is a non-negative functional defined over the space of k-th order differential forms which are continuous and have compact support, such that

(i) $V(r\varphi) = |r| V(\varphi)$, for every real r;
(ii) $V(\varphi + \psi) \leqslant V(\varphi) + V(\psi)$;
(iii) $V(f \wedge \varphi + g \wedge \varphi) = V(f \wedge \varphi) + V(g \wedge \varphi)$, when f and g are non-negative functions.

Almgren gives a solution to the general Plateau problem in the context of varifolds and relates it to the solutions obtained in other settings.

The same author also deals with the regularization problem for minimal varieties in the theory of the general Plateau problem, considered from various points of view. In [2] he proves that three dimensional minimal surfaces in R^4 are 3-dimensional real analytic submanifolds of R^4, except perhaps at their boundaries, where as minimal surface is meant:

(i) an oriented frontier of least 3-dimensional measure, see [5], p. 3;
(ii) a minimal 3-dimensional integral current, see [8], 9.1;
(iii) a minimal flat 3-chain over the group of integers mod 2, see [11];
(iv) a proper minimal surface of Reifenberg (see [23]) with boundary containing a cyclic subgroup of the 2-dimensional Čech homology group with coefficients in the group of integers mod 2 of the boundary set.

From these results the aforementioned regularization theorem for minimal varieties in cartesian form and the extension of Bernstein's theorem follow.

Scuola Normale Superiore,
Pisa, Italy

REFERENCES

[1] Almgren F. J., The theory of varifolds, Inst. Adv. Study, Princeton; Brown Univ.
[2] Almgren F. J., Some interior regularity theorems for minimal surfaces and an extension of Bernstein's theorem (to appear).
[3] Caccioppoli R., Misura ed integrazione sugli insiemi dimensionalmente orientati, *Rend. Accad. Naz. Lincei*, s. VIII, **XII**, 3-11, 137-146.
[4] Chern S. S., Minimal surfaces in an euclidean space of N dimensions, Symp. on diff. and combin. topology; ed. Cairns, Princeton, 1965.
[5] De Giorgi E., Frontiere orientate di misura minima, Scuola Normale Superiore, Pisa, 1961 (the results of this paper may be found, with complete proofs, also in paper [19]).

[6] De Giorgi E., Una estensione del teorema di Bernstein, *Ann. sc. Norm. Sup. Pisa*, **XIX** (1965), 79-85; 463.
[7] De Giorgi E., Stampacchia G., Sulle singolarita' eliminabili delle ipersuperficie minimali, *Rend. Accad. Naz. Lincei*, s. VIII, **XXXVIII** (1965), 352-357.
[8] Federer H., Fleming W. H., Normal and integral currents, *Ann. of Mathem.*, **72** (1960), 458-520.
[9] Finn R., On partial differential equations (whose solutions admit no isolated singularities), *Scripta Mathematica*, **26** (1963), 107-115.
[10] Fleming W. H., On the oriented Plateau problem, *Rend. Circ. Matem. Palermo*, **IX** (1962), 69-90.
[11] Fleming W. H., Flat chains over finite coefficients group, *Trans. Amer. Math. Soc.*, **121** (1966), 60-86.
[12] Gilbarg D., Boundary value problems for non-linear elliptic equations in n variables, Proc. Symp. on nonlinear problems, Univ. Wisconsin, Madison, 1962.
[13] Hartman P., On the bounded slope condition, *Pacific J. of Mathem.* (to appear).
[14] Hartman P., Stampacchia G., On some nonlinear elliptic differential functional equations, *Acta Mathem.*, **115** (1966), 271-310.
[15] Hopf E., Über den Funktionalen, insbesondere den analytischen Charakter der Lösungen elliptischer Differentialgleichungen zweiter Ordnung, *Math. Zeit.*, **34** (1932), 194-233.
[16] Jenkins H., Serrin J., The Dirichlet problem for the minimal surface equation in n dimensions (to appear).
[17] Miranda M., Un teorema di esistenza e unicita' per il problema dell' area minima in n variabili, *Ann. Sc. Norm. Sup. Pisa*, **XIX** (1965), 233-249.
[18] Miranda M., Analiticita' delle superfici di area minima in R^4, *Rend. Accad. Naz. Lincei*, **XXXVIII** (1965), 632-638.
[19] Miranda M., Sul minimo dell' integrale del gradiente di una funzione, *Ann. Sc. Norm. Sup. Pisa*, **XIX** (1965), 627-665.
[20] Morrey C. B., Second order elliptic systems of differential equations, Ann. of Math. Studies, no. 33, Princeton (1954), 101-159.
[21] Nitsche J. C. C., On new results in the theory of minimal surfaces, *Bull. Am. Math. Soc.*, (1965), 195-270.
[22] Osserman R., Le théorème de Bernstein pour des systèmes, *C. R. Acad. Sci. Paris*, **262** (1966), 571-574.
[23] Reifenberg E. R., Solution of the Plateau problem for m-dimensional surfaces of varying topological type, *Acta Mathem.*, **104** (1960), 1-92.
[24] Reifenberg E. R., An epiperimetric inequality related to the analyticity of minimal surfaces, *Ann. of Math.*, **80** (1964), 1-14.
[25] Reifenberg E. R., On the analyticity of minimal surfaces, *Ann. of Mathem.*, **80** (1964), 15-21.
[26] Stampacchia G., On some regular multiple integral problems in the calculus of variations, *Comm. Pure Appl. Math.*, **XVI** (1963), 383-421.
[27] Triscari D., Sulle singolarita' delle frontiere orientate di misura minima, Le Matematiche, Catania, **XVIII** (1963), 139-163.
[28] Young L. C., Surfaces paramétriques généralisées, *Bull. Soc. Math. France*, **79** (1951), 59-84.
[29] Young L. C., Contours in generalized and extremal varieties, *J. Math. and Mech.*, **11** (1962), 615-646.
[30] Ziemer W. P., Integral currents mod 2, *Trans. Am. Math. Soc.*, **105** (1962), 496-524.

DIFFERENTIAL COMPLEXES

JOSEPH J. KOHN

We are concerned with certain existence and regularity questions for systems of partial differential equations on manifolds. In considering such systems one has to take into account the compatibility conditions. Thus one of the important problems is to find all the compatibility conditions. This problem has been investigated (and solved in many important cases) by D. C. Spencer by means of the so called "Spencer resolution" (see [11]). Further study of the Spencer resolution has been done by D. G. Quillen (see [13]). In this lecture we assume that the system and the compatibility conditions are given, thus we have a differential complex. The guiding example of a differential complex is the case of the $\bar\partial$-complex; for functions this is the case of the Cauchy-Riemann equations.

The $\bar\partial$-complex on a compact manifold is a special case of an elliptic complex and can be treated by methods of classical elliptic theory. On non-compact manifolds, however, the $\bar\partial$-complex gives rise to questions in partial differential equations which cannot be answered by classical methods. The first decisive step in the study of the $\bar\partial$-complex on non-compact manifolds was the formulation of the $\bar\partial$-Neumann problem by D. C. Spencer. The study of the $\bar\partial$-Neumann problem has proved extremely fruitful and its solution has given rise to important new developments in partial differential equations as well as applications to several complex variables.

Spencer's original formulation of the $\bar\partial$-Neumann problem is given in his Paris lecture notes of 1955 (unpublished) and may also be found in [9]. The basic a-priori estimate for the $\bar\partial$-Neumann problem (in the case of forms of type (0,1) on strongly pseudo-convex manifolds) is due to C. B. Morrey (see [10]). However, to solve the problem, aside of the basic estimate some regularity results are needed. The author succeeded in proving regularity as well as establishing the basic estimate for forms of type (p, q) with $q > 0$ on strongly pseudo-convex manifolds, thus solving the $\bar\partial$-Neumann problem (see [4]). M. E. Ash found a new technique for deriving the basic estimate which has been very useful (see [1]). L. Hörmander found further extensions and applications of the $\bar\partial$-Neumann problem (see [3]). L. Nirenberg and the author have found a simpler proof of the regularity and have presented it in a much more general context in [6]; in order to establish local regularity it is necessary to study double commutators of fractional derivatives and this led to their work on pseudo-differential operators (see [7]). H. Rossi and the author have studied the problem of extending holomorphic forms from the boundary of a manifold (see [8]). This has suggested to the author the investigation of the "induced" $\bar\partial$-operator on real submanifolds of a complex manifold (see [5]). These provide examples of non-elliptic complexes on compact manifolds which still have "good" properties (i.e. regularity and finite cohomology). L. Hörmander has studied a much more general class of such problems in [3]. W. J. Sweeney has extended and generalized the methods of [4] and [8] to the case of the Spencer resolution and has applied Hörmander's characterizations given in [3] to the problem (see [13]).

For simplicity we will discuss systems of partial differential equations on a domain in $\mathbf{R}^n$ although our statements go over easily to differential operators on vector bundles over manifolds. Let M be a domain in $\mathbf{R}^n$ such that $\overline{M}$, the closure of M, is compact and bM, the boundary of M, is C^∞. Let $C^\infty(\overline{M})$ be the space of C^∞ complex valued functions on $\overline{M}$, let:

$$\mathscr{E} = \{u = (u^1, \ldots, u^p) \mid u^j \in C^\infty(\overline{M})\}$$

and

$$\mathscr{F} = \{f = (f^1, \ldots, f^q) \mid f^j \in C^\infty(\overline{M})\}.$$

We consider a first order differential operator $A: \mathscr{E} \to \mathscr{F}$ given by:

$$(Au)^i = \sum_{|\alpha| \leqslant 1,\, j} a^i_{\alpha j} D^\alpha u^j, \tag{1}$$

where $a^i_{\alpha j} \in C^\infty(\overline{M})$. We are interested in the following problem.

Problem I. Given $f \in \mathscr{F}$, when does there exist $u \in \mathscr{E}$, such that:

$$Au = f \tag{2}$$

and how does u depend on f.

Spencer's formulation of the $\bar\partial$-Neumann problem is easily generalized to the study of Problem I and we shall give a brief description of this. First we assume that we have given the compatibility conditions, that is, let

$$\mathscr{G} = \{g = (g^1, \ldots, g^r) \mid g^j \in C^\infty(\overline{M})\}$$

and let $B: \mathscr{F} \to \mathscr{G}$ be a linear differential operator such that:

$$BA = 0. \tag{3}$$

Then the sequence:

$$\mathscr{E} \xrightarrow{A} \mathscr{F} \xrightarrow{B} \mathscr{G} \tag{4}$$

is called a *differential complex*. If a solution of (2) exists we deduce that:

(5) $$Bf = 0.$$

Further, if we introduce L_2 inner products on $\mathscr{E}$, $\mathscr{F}$ and $\mathscr{G}$ denoted by $(\ ,\ )$ and defined by:

$$(u,\ v) = \sum \int_M u^i \bar{v}^i\, dV$$

and if we denote by A^* and B^* the L_2-adjoints of A and B respectively, we see that whenever (2) is satisfied the following condition holds:

(6) f is orthogonal to $\mathfrak{N}(A^*)$,

where $\mathfrak{N}(A^*)$ denotes the null space of A^*. Thus we see that (5) and (6) are necessary conditions for the existence of a solution of (2).

We will assume henceforth that the *operator B is of first order*. This is of course true in the case of the $\bar{\partial}$-complex, in general this assumption is not fulfilled; but it holds in the case of the Spencer resolution, by means of which many operators can be studied.

Let $\mathscr{D} \subset \mathscr{F}$ be defined by:

(7) $$\mathscr{D} = \{u \in \mathscr{F} \mid u \in \operatorname{dom}(A^*)\}.$$

For $u, v \in \mathscr{D}$ we define an inner product $Q(u, v)$ by:

(8) $$Q(u,\ v) = (A^*u,\ A^*v) + (Bu,\ Bv) + (u,\ v).$$

The analogue of the $\bar{\partial}$-Neumann problem (for the complex (4) instead of the $\bar{\partial}$-complex) is the following.

Problem II. Given $f \in \mathscr{F}$, does there exist a $u \in \mathscr{D}$ such that:

(9) $$Q(u,\ v) = (f,\ v)$$

for all $v \in \mathscr{D}$. How does u depend on f?

This problem always has a unique "weak" solution in the following sense. Denote by $\tilde{\mathscr{D}}$ the completion of $\mathscr{D}$ under Q and (extending Q to $\tilde{\mathscr{D}}$) we have:

(10) $$Q(u,\ u) \geqslant \|u\|^2, \text{ for all } u \in \tilde{\mathscr{D}},$$

hence given $f \in L_2$ there exist a unique solution of (9) in $\tilde{\mathscr{D}}$ and further the map of $f \to u$ is bounded since:

(11) $$\|u\|^2 \leqslant Q(u,\ u) = (f,\ u) \leqslant \|f\| \cdot \|u\|.$$

Thus Problem II always has a weak solution in the above sense. However, it is easy to see that this type of solution does not give any information on Problem I. What is needed is a smooth solution and regular behavior of the map $f \to u$ on the eigen-space corresponding to

the eigen-value 1. In particular the following theorem, proven in [6], gives conditions under which Problem II has a satisfactory solution.

Theorem 1. If bM is non-characteristic for Q and if Q is compact with respect to L_2 (i.e. a sequence bounded in Q has a convergent subsequence in L_2) then there exists a solution $u \in \mathscr{D}$ of Problem II for every $f \in \mathscr{F}$ and the map $f \to u$ is completely continuous in H_s for every s. (Here H_s denotes the Sobolev space of functions with square-integrable derivatives of order s).

It then follows that the solution u not only belongs to $\mathscr{D}$ but also $Bu \in \operatorname{dom}(B^*)$ and that the map $f \to u$ is the inverse of the operator $AA^* + B^*B + I$. Now let $\mathscr{H}$ be the eigen-space of this operator corresponding to the eigen-value 1 (i.e. $\mathscr{H} = \mathfrak{N}(AA^* + B^*B)$). Then $\mathscr{H}$ is a finite dimensional subspace of $\mathscr{D}$ and since for $u \in \mathscr{D}$

$$((AA^* + B^*B)u,\ u) = \|A^*u\|^2 + \|Bu\|^2$$

we conclude that:

(12) $$\mathscr{H} = \mathfrak{N}(A^*) \cap \mathfrak{N}(B).$$

Furthermore if f is orthogonal to $\mathscr{H}$ there exists a unique solution u of

(13) $$AA^*u + B^*Bu = f$$

with $u \in \mathscr{D}$ and $Bu \in \operatorname{dom}(B^*)$. Now we can solve Problem I. Observe that the necessary conditions (5) and (6) are equivalent to $Bf = 0$ and $f \perp \mathscr{H}$. So, assuming these conditions and applying B to (13), we obtain:

$$BB^*Bu = 0 \text{ hence } (BB^*Bu,\ Bu) = \|B^*Bu\|^2 = 0.$$

So that (13) reduces to

(14) $$AA^*u = f$$

and A^*u is the required solution of Problem I.

Definition. The differential complex (4) is called an *elliptic complex* if there exists $C > 0$ such that:

(15) $$\|u\|_1^2 \leqslant CQ(u,\ u)$$

for all $u \in \mathscr{D}$, where $\|\ \|_1$ denotes the norm in the Sobolev space H_1. When (15) holds we say that Q is a *coercive form* on $\mathscr{D}$.

Coercive forms have been studied extensively and the conclusions of Theorem 1 are well known in case Q is coercive. Thus, by the above argument, Problem I for elliptic complexes reduces to the standard theory of coercive forms. Theorem 1 is proven by approximating Q with coercive forms. For each $\delta > 0$ we define on $\mathscr{D}$ the form $Q^\delta(u, v)$ by:

(16) $$Q^\delta(u,\ v) = Q(u,\ v) + \delta P(u,\ s),$$

where P is coercive. Then Q^δ (for $\delta > 0$) is coercive and hence given $f \in \mathscr{F}$ there exists $u^\delta \in \mathscr{D}$ such that

$$Q^\delta(u^\delta, v) = (f, v) \tag{17}$$

for all $v \in \mathscr{D}$. The crucial step is to prove a-priori estimates which are independent of δ; in fact, the hypothesis of Theorem 1 imply that for each $s \geqslant 0$ there exists a constant $C_s > 0$ independent of δ such that

$$\| u^\delta \|_s \leqslant C_s \| f \|_s. \tag{18}$$

From (18) it is easily deduced that u^δ converges (as $\delta \to 0$) in H_s for every s and hence its limit u is in $\mathscr{D}$ and is the solution of (9). This is only a sketch of the argument; the complete proof may be found in [6].

It is important to express the notion of ellipticity directly in terms of the differential complex (4). For this purpose we define the *symbol sequence* which corresponds to (4). For each $x \in \overline{M}$ we let E_x and F_x stand for the sets of values of elements in $\mathscr{E}$ and $\mathscr{F}$ evaluated at x: thus E_x and F_x are vector spaces of dimensions p and q respectively. Now for each $\eta = (\eta^1, \ldots, \eta^n) \in \mathbf{C}^n$ and each $x \in \overline{M}$ we define the *symbol of* A at x, denoted by $\sigma_x(A, \eta)$ to be the linear map:

$$\sigma_x(A, \eta) : E_x \to F_x$$

given by:

$$(\sigma_x(A, \eta)e)^i = \sum_{|\alpha|=1,\, j} a^i_{\alpha j}(x)\, \eta^\alpha e^j, \quad i = 1, \ldots, q, \tag{19}$$

where $e = (e^1, \ldots, e^p) \in E_x$. Similarly we define the map

$$\sigma_x(B, \eta): \quad F_x \to G_x.$$

Thus for each η and each x we have the symbol sequence:

$$E_x \xrightarrow{\sigma_x(A, \eta)} F_x \xrightarrow{\sigma_x(B, \eta)} G_x. \tag{20}$$

We observe that the fact that $BA = 0$ implies that

$$\sigma_x(B, \eta)\, \sigma_x(A, \eta) = 0 \tag{21}$$

and the null space of $\sigma_x(B, \eta)$ contains the image of $\sigma_x(A, \eta)$.

The following theorem gives characterization of *interior* coerciveness of Q.

Theorem 2. The following are equivalent:

(a) For each $x \in \overline{M}$ and each $\eta \in \mathbf{R}^n - \{0\}$ the symbol sequence (20) is exact (i. e. null space of $\sigma_x(B, \eta)$ equals image of $\sigma_x(A, \eta)$).

(b) If $e \in E_x$, $\eta \in \mathbf{R}^n - \{0\}$, $\sigma_x(B, \eta)\, e = 0$ and $\overline{\sigma_x(A, \eta)^t} e = 0$ then $e = 0$.

(c) Q is coercive on *compactly supported* elements of $\mathscr{F}$.

(d) There exists $C > 0$ such that for all *compactly supported* u of $\mathscr{F}$:

$$\| u \|_2 \leqslant C \| AA^* u + B^* B u + u \|.$$

The $\bar\partial$-complex is an example of a complex for which Q is coercive on compactly supported elements but is not coercive on $\mathscr{D}$ (if $\dim_{\mathbf{C}} M > 1$). The following theorem gives conditions which characterize complexes for which Q is coercive on $\mathscr{D}$ (i.e. elliptic complexes).

Theorem 3. The following are equivalent:

(A) Q is coercive on $\mathscr{D}$.

(B) There exists a constant $C > 0$ such that for all $u \in \mathscr{D}$ with $Bu \in \operatorname{dom}(B^*)$ we have:

$$\| u \|_2 \leqslant C \| AA^* u + B^* B u + u \|.$$

(C) Q is coercive on compactly supported elements and if $x \in bM$, ν_x is the normal to bM at x, $\eta \in \mathbf{C}^n - \{0\}$, $e \in E_x$, and if

$$\overline{\sigma_x(A, \eta)^t} e = \overline{\sigma_x(A, \nu_x)^t}\, e = \sigma_x(B, \eta)\, e = 0$$

then $e = 0$.

We wish to study those differential complexes for which the conclusions of Theorem 1 hold but which are not elliptic we call them *subelliptic complexes*. For simplicity we will consider the case where M is a compact manifold without boundary. Then we take $\mathscr{E}$, $\mathscr{F}$ and $\mathscr{G}$ to be spaces of C^∞ sections of complex vector bundles over M. The operator A is still given by (1) but this expression makes sense only locally (i.e. relative to a local coordinate system and local bases for underlying bundles of $\mathscr{E}$ and $\mathscr{F}$). Now in the definition of symbol the vector η should be interpreted as an element of the complexified cotagent space at x, which we denote by $\mathbf{C}T^*_x$. Since M has no boundary Theorem 2 implies that ellipticity of the complex is equivalent to exactness of the symbol sequence for real cotagent vectors $\eta \in T^*_x - \{0\}$.

We are particularly interested in complexes for which there exists an estimate of the form:

$$\| u \|^2_s \leqslant \text{const. } Q(u, u) \tag{22}$$

for all $u \in \mathscr{F}$ and $0 < s < 1$. It is proven in [5] that the induced $\bar\partial$-complex in a submanifold of complex manifold of real co-dimension one satisfies (22) with $s = \frac{1}{2}$ whenever its Levi-form is positive definite. Hörmander, in [3], gives a characterization of very general Q which satisfy the above estimate for $s = \frac{1}{2}$. However, it is very difficult to interpret what this characterization says about the original complex. It should be remarked that when (22) is satisfied on a compact manifold then

Theorem I applies (the condition about the boundary being irrelevant). Furthermore it is proven in [6] that (22) implies local regularity of the solutions of (9); that is, if f is regular on an open set Ω then the solution u will also be regular on Ω. We define a subset $\dot{T}^*_x$ of T^*_x by

(23) $$\dot{T}^*_x = \{\eta \in T^*_x \mid \text{there exists } e \in E_x,\ e = 0 \text{ such that } \overline{\sigma_x(A, \eta)}^t e = 0 \text{ and } \sigma_x(B, \eta) e = 0\}.$$

We denote by S_x the anihilator of $\dot{T}^*_x$, i.e. S_x is the subspace of the tangent space T_x consisting of those elements whose contraction with every element of $\dot{T}^*_x$ is zero. Let $\mathscr{S}$ denote the set of local tangent vector fields which at each x belong to S_x. Since M is compact then, according to Theorem 2, the complex is elliptic (i.e. (22) holds for $s = 1$) if and only if $S_x = T_x$ for each x. Here we assume that the complex is subelliptic and hence for some x we have $S_x \neq T_x$. Let $\tilde{S}_x$ be the subspace of T_x consisting of those vectors which are generated by $\mathscr{S}$ under the Lie bracket [,]. Then we have:

T h e o r e m 4. If there exists a neighborhood $\mathcal{O} \subset M$ such that for each $x \in U$, $\tilde{S}_x \neq T_x$ then the form Q is not compact (and hence the complex is not sub-elliptic).

In fact the hypotheses of the above theorem imply that we can find a coordinate system $\{y^1, \ldots, y^n\}$ on U such that for each $x \in U$, $(dy^n)_x \in \dot{T}^*_x$. It follows that there exists an infinite dimensional subspace of $\mathscr{E}$ on which Q does not involve derivatives with respect to y_n.

It is then clear that the possibility of obtaining estimate (22) will depend on the Lie algebra generated by $\mathscr{S}$.

T h e o r e m 5. If $\dot{T}^*_x$ is a linear subspace of $\dot{T}^*_x$ and $\dim \dot{T}^*_x$ is independent of x then for each $L \in \mathscr{S}$ there exists a constant $C > 0$ such that:

(23) $$\| Lu \|^2 \leqslant C [Q(u, u) + \| u \|^2_{1/2}]$$

for all u whose support lies in the domain of definition of L.

From the above theorem we see that (under the same assumptions) if (22) holds with $s < 1$ then $s \leqslant \frac{1}{2}$. In fact, generically, the case $s = \frac{1}{2}$ corresponds to $\dim \dot{T}^*_x = 1$ and in that case inequality (22) depends on relations of single brackets $[L, L']$ with $L, L' \in \mathscr{S}$. In case $\dim \dot{T}^*_x = k$ the "generic" best value for s in (22) is 2^{-k} and the estimate will depend on k^{th} order brackets of elements of $\mathscr{S}$.

Dept. of Math., Brandies University,
Massachusetts, USA

REFERENCES

[1] A s h M. E., The basic estimate of the ∂-Neumann problem in the non-Kählerian case, *Amer. J. Math.* **86** (1964), 247-254.

[2] H ö r m a n d e r L., An introduction to complex analysis in several variables, Van Nostrand, 1966.

[3] H ö r m a n d e r L., Pseudo-differential operators and non-elliptic boundary value problems, *Ann. Math.*, **83** (1966), 129-209.

[4] K o h n J. J., Harmonic integrals on strongly pseudo convex manifolds, I and II, *Ann. Math.*, **78** (1963), 112-148 and **79** (1964), 450-472.

[5] K o h n J. J., Boundaries of complex manifolds, Proc. of the Conf. on Complex Analysis, Minneapolis, 1964 (Springer-Verlag, 1965), 81-94.

[6] K o h n J. J., N i r e n b e r g L., Non-coercive boundary value problems, *Comm. Pure and Appl. Math.*, **18** (1965), 443-492.

[7] K o h n J. J., N i r e n b e r g L., An algebra of pseudo-differential operators, *Comm. Pure and Appl. Math.*, **18** (1965), 269-305.

[8] K o h n J. J., R o s s i H., On the extension of holomorphic functions from the boundary of a complex manifold, *Ann. Math.*, **81** (1965), 451-472.

[9] K o h n J. J., S p e n c e r D. C., Complex Neumann problems, *Ann. Math.*, **66** (1957), 89-140.

[10] M o r r e y C. B., The analytic embedding of abstract real analytic manifolds, **68** (1958), 159-201.

[11] S p e n c e r D. C., Deformations of structures on manifolds defined by transitive, continuous pseudo-groups, I, II and III, *Ann. Math.*, **76** (1962), 306-445 and **81** (1965), 389-450.

[12] S w e e n e y W. J., The D-Neumann problem, Stanford thesis, 1966 (to appear).

[13] Q u i l l e n D. G., Formal properties of over-determined systems of linear partial differential equations. Harvard thesis, 1964 (to appear).

ЭЛЛИПТИЧЕСКИЕ УРАВНЕНИЯ В СВЕРТКАХ В ОГРАНИЧЕННОЙ ОБЛАСТИ И ИХ ПРИЛОЖЕНИЯ

М. И. ВИШИК

Настоящий доклад посвящен изложению содержания совместных работ Г. И. Эскина и моих по теории различных задач для эллиптических и параболических уравнений и систем в свертках в ограниченной области. Общие свойства операторов в свёртках или псевдодифференциальных операторов, а также их роль при исследовании ряда важных проблем анализа были выяснены в работах Михлина, Кальдерона и Зигмунда, Бицадзе, Вольперта, Сили, Дынина, Аграновича, Атья и Зингера, Кона и Ниренберга, Хёрмандера и других авторов.

I. Уравнения в свертках

Уравнение в свертках в ограниченной области G с гладкой границей Γ формально можно записать в виде

$$P^+Au_+ \equiv P^+\left(\int A(x,\ x-y)\,u_+(y)\,dy+\ldots\right)\equiv$$

$$\equiv P^+A_0u_+ +\ldots = f(x),\ x\in G, \tag{1}$$

где P^+ — оператор сужения функции на область G, $u_+(y)$ — обобщенная функция с носителем в $\overline{G}$, а $A(x, z)$ — обобщенная функция z, причем ее преобразование Фурье по z $\widetilde{A}(x, \xi) = F_{z\to\xi}A(x, z)$ является однородной функцией порядка α:

$$\widetilde{A}(x,\ t\xi)=t^{\alpha}\widetilde{A}(x,\xi)\quad (x=(x_1,\ \ldots,\ x_n),\quad \xi=(\xi_1,\ \ldots,\ \xi_n)). \tag{2}$$

В этом случае мы скажем, что уравнение (1) имеет порядок α, а функция $\widetilde{A}(x, \xi)$ называется *символом* этого уравнения или оператора Au. Многоточием в (1) обозначены члены более низкого порядка, чем α.

Уравнение (1) называется *эллиптическим в области* G, если $\widetilde{A}(x, \xi)\neq 0$, для $x\in G$, $\xi\neq 0$, или в случае систем уравнений

$$\det\widetilde{A}(x,\ \xi)\neq 0. \tag{3}$$

Символ $\widetilde{A}(x, \xi)$ мы будем считать заданным для всех $x\in R^n$, $\xi\in R^n\setminus 0$, причем $\widetilde{A}(x,\xi)\equiv\widetilde{A}(\infty,\xi)$ для $|x|\geqslant l-\varepsilon$. Тогда главную часть A_0u_+ оператора Au_+ можно определить с помощью разложения в тригонометрический ряд по формуле

$$A_0u_+ = A(\infty,\ x)*u_+(x)+\psi_0(x)\sum_k e^{i\pi kx/l}A_k(x)*u_+(x), \tag{4}$$

где

$$A_k(x)=F^{-1}\widetilde{A}_k(\xi),\qquad A(\infty,\ x)=F^{-1}\widetilde{A}(\infty,\xi),$$

$$A_k(\xi)=\frac{1}{(2\pi)^n}\int_{-l}^{+l}\left(\widetilde{A}(x,\xi)-\widetilde{A}(\infty,\xi)\right)e^{-i\pi kx/l}\,dx,$$

$$\psi_0(x)=\begin{cases}1 \text{ при } |x|<l-\varepsilon/2,\\ 0 \text{ при } |x|>l.\end{cases}$$

Свертни $A_k(x)*u_+(x)$ существуют, так как u_+ имеет компактный носитель.

Примеры. 1. Дифференциальные операторы являются частным случаем уравнений в свертках. В этом случае

$$A(x,\ z)=\sum_{|\beta|=m}a_\beta(x)\,\delta^\beta(z),\ \beta=(\beta_1,\ \ldots,\ \beta_n).$$

2. Интегральные уравнения в свертках первого рода отвечают тому случаю, когда $\operatorname{Re}\alpha<0$. В этом случае (1) является интегральным уравнением со слабой особенностью, и его можно представить в виде

$$\int\frac{K\left(x,\ \frac{x-y}{|x-y|}\right)}{|x-y|^{n+\alpha}}u_+(y)\,dy+\ldots=f(x),\ x\in G. \tag{5}$$

3. Сингулярные интегральные уравнения в ограниченной области — это те уравнения в свертках вида (1), для которых $\alpha=0$.

II. Факторизация символа

Основную роль при постановке задач для уравнений в свертках в ограниченной области играет факторизация символа $\widetilde{A}(x, \xi)$ в точках $x\in\Gamma$. Пусть $\widetilde{A}(y, \xi)$ ($y=(y_1,\ \ldots,\ y_n)$) — символ оператора Au_+ в локальной системе координат (л. с. к.) в окрестности некоторой точки $x'\in\Gamma$, причем $y_n=0$ — уравнение соответствующего куска Γ. Под *однородной факторизацией по переменной* ξ_n (двойственной к y_n) *эллиптического символа* $\widetilde{A}(y, \xi', \xi_n)$ ($\xi'=(\xi_1,\ \ldots,\ \xi_{n-1})$) подразумевается представление $\widetilde{A}(y, \xi', \xi_n)$ в виде следующего произведения:

$$\widetilde{A}(y,\ \xi',\ \xi_n)=\widetilde{A}_+(y,\ \xi',\ \xi_n)\,\widetilde{A}_-(y,\ \xi',\ \xi_n), \tag{6}$$

где $\widetilde{A}_+$ имеет порядок однородности $\varkappa=\varkappa(y)$ по ξ ($\operatorname{ord}\widetilde{A}_+=\varkappa(y)$) и $\operatorname{ord}\widetilde{A}_-=\alpha-\varkappa$, причем A_+ ($\widetilde{A}_-$) — аналитическая функция ξ_n при $\operatorname{Im}\xi_n>0$ ($\operatorname{Im}\xi_n<0$) и $\widetilde{A}_+\neq 0$ ($\widetilde{A}_-\neq 0$) для $\operatorname{Im}\xi_n\geqslant 0$, $\xi\neq 0$ ($\operatorname{Im}\xi_n\leqslant 0$, $\xi\neq 0$).

В [1] показано, что в скалярном случае такая факторизация всегда существует для эллиптических символов, удовлетворяющих условию Гёльдера по ξ при $\xi\neq 0$. Она однозначна с точностью до постоянных множителей.

Значение числа $\varkappa$, называемого *индексом* $\widetilde{A}$ (или индексом оператора A), весьма существенно при постановке задач для уравнений в свертках (1).

В том частном случае, когда $\widetilde{A}(x, \xi', \xi_n)$ — эллиптический многочлен от ξ степени $2m$, факторизация (6) хорошо известна. Тогда $\widetilde{A}_+(x, \xi', \xi_n)$ — та часть многочлена $\widetilde{A}$, которая по ξ_n имеет корни

в нижней полуплоскости, а $\widetilde{A}_-(x, \xi', \xi_n)$ — в верхней полуплоскости.

Факторизация символа оператора оказалась очень полезной при исследовании ряда проблем, таких, как решение интегральных уравнений на полупрямой (метод Винера — Хопфа; см. работы М. Г. Крейна и И. Ц. Гохберга [2]), исследование одномерных сингулярных интегральных уравнений (Мусхелишвили Н. И., Векуа И. Н., Гахов Ф. Д., Векуа Н. П. и др.), двумерных разрывных краевых задач (Петре [4] и др.).

III. Условие гладкости оператора свертки

Мы говорим, что оператор свертки A_0u_+ вещественного порядка α обладает свойством гладкости в области G, если для любого $s \geqslant 0$ имеет место оценка

$$\| P^+A_0u_+ \|_{s-\alpha} \leqslant C \| u_+ \|_s, \tag{7}$$

где $\| \ \|_r$ — норма Соболева — Слободецкого порядка r по области G. Как известно, для дифференциальных операторов оценка (7) всегда имеет место. Для операторов типа свертки свойство гладкости далеко не всегда имеет место. Однако если символ $\widetilde{A}(y, \xi', \xi_n)$ оператора A_0u_+ в любой л. с. к. $\{y\}$ на Γ удовлетворяет условию

$$D_{y_n}^{k_n} D_\xi^k \widetilde{A}(y, 0, -1) = (-1)^{|k|} e^{-i\alpha\pi} D_{y_n}^{k_n} D_\xi^k \widetilde{A}(y, 0, +1) \tag{8}$$

для всех $k = (k_1, \ldots, k_{n-1})$ и k_n, то оператор A_0u_+ — гладкий в области G.

Класс символов $\widetilde{A}$, для которых выполнено (8), обозначим через $D_\alpha^{(0)}$.

Отметим, что условие (8) при натуральном α означает, что символ $\widetilde{A}(y, \xi', \xi_n)$ ведет себя при $\xi_n \to +\infty$ и $\xi_n \to -\infty$, грубо говоря, как полином по ξ_n.

IV. Задачи для уравнений в свертках

Рассмотрим сначала случай уравнений (1), для которых $\widetilde{A} \in D_\alpha^0$ и подчиненные операторы, обозначенные в (1) многоточием, обладают свойством гладкости. В этом случае индекс $\varkappa$ символа $\widetilde{A}(x, \xi)$ — целое число. Задачи для уравнения в свертках (1) ставятся по-разному в зависимости от значения $\varkappa$.

а) $\varkappa \geqslant 0$. В этом случае решения однородного уравнения, соответствующего (1), грубо говоря, зависят от $\varkappa$ произвольных функций, заданных на Γ, аналогично тому, как это имеет место в случае однородного эллиптического дифференциального уравнения порядка $2\varkappa$. В этом можно убедиться, решив в явном виде для $G = R_+^n$ ($x_n > 0$) уравнение вида (1) с символом $\widetilde{A}(\xi)$, не зависящим от x. Преобразование Фурье общего решения этого уравнения выражается формулой

$$\widetilde{u}(\xi) = \frac{1}{\widetilde{A}_+} \Pi^+ \frac{1}{\widetilde{A}_-} \widetilde{f} + \frac{\sum_{k=1}^{\varkappa} c_k(\xi') \xi_n^{k-1}}{\widetilde{A}_+(\xi', \xi_n)}, \tag{8'}$$

где первое слагаемое является частным решением неоднородного уравнения (Π^+ — оператор типа Коши, отвечающий в x-представлении оператору умножения на $\theta(x_n) = \begin{cases} 1 \text{ при } x_n > 0, \\ 0 \text{ при } x_n < 0 \end{cases}$), а второе слагаемое представляет собой общее решение однородного уравнения. Так как множитель $\widetilde{A}_+(\xi)$ имеет порядок роста по ξ_n, равный $\varkappa$, то даже при наличии в числителе многочлена по ξ_n степени $\varkappa - 1$ последняя дробь в (8') является по ξ_n функцией из H_0 (при $\xi' \neq 0$). Отвечающая этой дроби в x-представлении функция $u_0(x', x_n)$ является по x_n гладкой функцией, убывающей при $x_n \to \infty$, если $c_k(\xi')$ удовлетворяют некоторым условиям. Таким образом, в общем решении имеется $\varkappa$ функциональных параметров $c_k(\xi')$ ($k = 1, \ldots, \varkappa$), если индекс $\widetilde{A}(\xi)$ равен $\varkappa$.

В связи с этим для корректной постановки задачи для уравнения (1) необходимо еще к нему добавить $\varkappa$ дополнительных условий. Эти дополнительные условия естественно записать в виде краевых условий с операторами типа свертки:

$$\gamma P^+ B_j u_+ = g_j \quad (j = 1, \ldots, \varkappa), \tag{9}$$

где γ — оператор сужения функции на границу и $B_j u_+$ — гладкий оператор типа свертки порядка a_j (см. [1]). Если в каждой точке границы Γ выполнено одно алгебраическое условие, аналогичное условию Шапиро — Лопатинского для дифференциальных эллиптических задач, то задача (1), (9) нормально разрешима, т. е. однородная задача имеет конечномерное ядро, а неоднородная разрешима при конечном числе условий на правые части. При этом имеет место априорная оценка

$$\| u \|_s \leqslant C \Big(\| Au_+ \|_{s-\alpha} + \sum_{j=1}^{\varkappa} \| \gamma P^+ B_j u_+ \|'_{s-\alpha_j-\frac{1}{2}} + \| u_+ \|_{s-1} \Big),$$

где $\| \ \|'_r$ означает норму порядка r по Γ.

б) $\varkappa < 0$. В этом случае даже при сколь угодно гладкой правой части $f(x)$ решение $u_+(x)$ уравнения в свертках (1) является, вообще говоря, обобщенной функцией. Например, в случае полупростран-

ства уравнение (1) с символом $\widetilde{A}(\xi)$ имеет решение следующей структуры:

$$u_+(x) = u_+^0(x) + \sum_{k=1}^{|\varkappa|} \rho_k(x')\,\delta^{(k-1)}(x_n),$$

где $u_+^0(x)$ — гладкая функция в R_+^n $(x_n > 0)$, а $\delta(x_n)$ — функция Дирака. Поэтому естественно при $\varkappa < 0$ поставить задачу о нахождении гладкой функции $u_+^0(x)$ и $|\varkappa|$ функций $\rho_k(x^i)$, заданных на границе Г.

В случае ограниченной области G корректно поставленной задачей для уравнения (1) является так называемая задача с дополнительными потенциалами, в которой ищутся $|\varkappa| + 1$ функций

$$u_+^0(x),\ \rho_1(x'),\ \ldots,\ \rho_{|\varkappa|}(x'),\ x' \in \Gamma,\ x \in G,$$

из уравнения, которое формально можно записать следующим образом:

$$P^+\left(A_0 u_+ + \sum_{i=1}^{|\varkappa|} \int_\Gamma G_i(x,\ x-y')\,\rho_i(y')\,d\Gamma + \ldots\right) = f(x),\ x \in G. \quad (10)$$

Если символы $\widetilde{A}(x, \xi)$, $\widetilde{G}_i(x, \xi)$ в каждой точке $x \in \Gamma$ удовлетворяют одному алгебраическому условию, то уравнение (10) нормально разрешимо относительно $(u_+^0(x), \rho_1(x'), \ldots, \rho_{|\varkappa|}(x'))$, причем имеет место соответствующая априорная оценка. Отметим, что задача (10) является в каком-то смысле двойственной к краевой задаче: в ней по заданной функции $f(x)$ ищутся, грубо говоря, плотность u_+ некоторого «объемного потенциала», и плотности $\rho_i(x')$ $(i = 1, \ldots, |\varkappa|)$ заданных «поверхностных потенциалов».

В [5] указаны также задачи, в которых дополнительные потенциалы стоят под знаком оператора в свертках. Главный член таких задач имеет вид $P^+ A\,(u_+ + \sum_{i=1}^{|\varkappa|} G_i \rho_i)$.

В общем случае уравнения (1), символ которого $\widetilde{A}(x, \xi)$ не удовлетворяет условию гладкости, корректно разрешимой при любом $\varkappa$, является первая однородная краевая задача для этого уравнения, которая состоит в следующем: ищется решение $u_+(x)$ этого уравнения, принадлежащее пространству $\overset{\circ}{H}_{\varkappa+\delta}(G)$, где $|\delta| < 1/2$ и $\varkappa$ — индекс символа $\widetilde{A}(x, \xi)$, который сейчас может быть и не целым числом. Для простоты допустим, что $\varkappa$ не зависит от x, $\overset{\circ}{H}_r(G)$ — пространство, полученное замыканием в метрике $\|\ \|_r$ финитных функций в G. Если правая часть $f(x)$ уравнения (1) принадлежит $H_{\varkappa+\delta-\alpha}(G)$, то задача о нахождении $u_+(x) \in \overset{\circ}{H}_{\varkappa+\delta}(G)$ из (1) нормально разрешима.

При увеличении гладкости правой части $f(x)$ решение $u_+(x)$ становится более гладким лишь внутри G, в то время как $u_+(x)$ при приближении к Г имеет гладкость только порядка $\varkappa$. Этот факт можно выразить с помощью норм. Вводится пространство $H_{r,N}(G)$ функций $u_+(x)$, в котором норма задается с помощью весовых множителей, причем конечность этой нормы означает, что $u_+(x)$ имеет гладкость порядка $r + N$ внутри области и гладкость порядка r на Г. Тогда первая краевая задача для уравнения (1) нормально разрешима в пространстве $\overset{\circ}{H}_{\varkappa+\delta,N}(G)$, если правая часть $f(x) \in H_{\varkappa+\delta-\alpha,N}(G)$. Аналогичный факт установлен также в случае переменного индекса $\varkappa = \varkappa(x)$. В этом случае факт нормальной разрешимости установлен в пространствах с «кусочно постоянными» нормами (см. [3]).

Отметим еще, что для общих уравнений в свертках можно поставить задачу с любым числом дополнительных потенциалов и с любым числом граничных условий:

$$P^+\left(A_0 u_+ + \sum_{j=1}^{M} \int_\Gamma G_j(x,\ x-y')\,\rho_j(y')\,d\Gamma_{y'} + \ldots\right) = f(x)\ \ (x \in G), \quad (11)$$

$$\gamma P^+ B_l u_+ + \sum_{j=1}^{M} \int E_{ij}(x',\ x'-y')\,\rho_j(y')\,d\Gamma_{y'} + \ldots = g_i(x')$$

$$(x' \in \Gamma;\ i = 1,\ \ldots,\ l). \quad (12)$$

При выполнении некоторых алгебраических условий эта задача нормально разрешима для $u_+ \in \overset{\circ}{H}_{\varkappa+M-l,N}(G)$,

$$\rho_l \in H'_{\varkappa-\alpha+M-l+\alpha_k+1/2}(\Gamma), \quad \alpha_k = \operatorname{ord}_\xi G_k(x, \xi).$$

Следовательно, каждое граничное условие снижает на единицу гладкость решения (u_+, ρ_l) задачи (11), (12), а каждый дополнительный потенциал повышает эту гладкость на единицу (число l входит в индекс пространства со знаком минус, а число M — со знаком плюс) (см. [3]).

V. Системы уравнений в свертках

Для случая системы уравнений в свертках вида (1) символ $\widetilde{A}(x, \xi)$ является квадратной матрицей размеров $p \times p$, имеющей порядок однородности α по ξ. Пусть $\widetilde{A}(y, \xi)$ — символ этой системы в л. с. к., отвечающей куску Г. При этом $\widetilde{A}(y, \xi) \in C^\infty$ по ξ и y при $\xi \neq 0$. Рассмотрим сначала тот частный случай, когда

$$\lim_{\xi_n \to \pm\infty} (\xi_n - i\,|\,\xi'\,|)^{-\alpha}\,\widetilde{A}(y, \xi', \xi_n) = I\ (I \text{ — единичная матрица}).$$

Тогда матрица $\widetilde{A}$ при фиксированных ξ' и y допускает следующую факторизацию:

$$\widetilde{A}(y, \xi', \xi_n) = \widetilde{A}_-(y, \xi', \xi_n) M_- M_+ \widetilde{A}_+(y, \xi', \xi_n), \qquad (13)$$

где $\widetilde{A}_+$ ($\widetilde{A}_-$) аналитически зависит от ξ_n в полуплоскости $\operatorname{Im} \xi_n > 0$ ($\operatorname{Im} \xi_n < 0$), причем $\det \widetilde{A}_+ \neq 0$ ($\det \widetilde{A}_- \neq 0$) при $|\xi| \neq 0$, $\operatorname{Im} \xi_n \geqslant 0$ ($|\xi| \neq 0$, $\operatorname{Im} \xi_n \leqslant 0$); M_+, M_- — матрицы вида

$$M_+ = \|(\xi_n + i|\xi'|)^{n_j} \delta_{ij}\|_{i,\, j=1,\, \ldots,\, p},$$

$$M_- = \|(\xi_n - i|\xi'|)^{\alpha - n_j} \delta_{ij}\|_{i,\, j=1,\, \ldots,\, p},$$

где n_j — целые числа, $n_1 \geqslant n_2 \geqslant \ldots \geqslant n_p$, называемые частными индексами матрицы $\widetilde{A}$. Как показано Крейном и Гохбергом [2], частные индексы, вообще говоря, неустойчивы при непрерывном изменении матрицы $\widetilde{A}$, однако они однозначно определяются матрицей $\widetilde{A}$, в то время как матрицы $\widetilde{A}_+$ и $\widetilde{A}_-$ неоднозначно определяются матрицей $\widetilde{A}$.

Мы исследуем сначала системы уравнений в свертках в предположении, что выполнено следующее условие (а): частные индексы матрицы $\widetilde{A}(y, \xi)$ не зависят от y и ξ'. Тогда, как доказывается, можно выбрать в (13) сомножители $\widetilde{A}_+(y, \xi)$ и $\widetilde{A}_-(y, \xi)$ однородными функциями ξ нулевого порядка и, кроме того, ограниченными и кусочно непрерывными (вместе с $\widetilde{A}_+^{-1}$ и $\widetilde{A}_-^{-1}$) функциями своих аргументов.

Для того чтобы существовала непрерывная факторизация вида (13) при $\xi \neq 0$, необходимо и достаточно, чтобы были выполнены некоторые топологические условия.

В общем случае при выполнении аналога условия (а) доказано существование кусочно непрерывной факторизации вида (13), только у матриц $\widetilde{A}_+$ и $\widetilde{A}_-$ появляются еще однородные по ξ множители $\widetilde{b}_+(y, \xi)$ и $\widetilde{b}_-(y, \xi)$, имеющие, быть может, при $\xi' = 0$ особенность (для краткости мы подробнее $\widetilde{b}_+$ и $\widetilde{b}_-$ не описываем). Эти множители $\widetilde{b}_+$ и $\widetilde{b}_-$ привносят к частным индексам n_j еще слагаемые $\gamma_j = \gamma_j(y)$, где $|\operatorname{Re} \gamma_l - \operatorname{Re} \gamma_j| < 1$, которые могут зависеть от y. Мы предположим здесь для простоты, что γ_j от y не зависит. Таким образом, в общем случае эллиптическая матрица $\widetilde{A}(y, \xi)$ допускает факторизацию с частными индексами $\varkappa_j = n_j + \gamma_j$, вообще говоря, комплексными.

Постановка задачи для общей эллиптической системы в свертках зависит от требуемой гладкости u_+ (ср. с концом п. IV). Если ищется решение u_+, обладающее гладкостью s, $u_+ \in \mathring{H}_s(G)$, где $s - \operatorname{Re} \gamma_l \neq \frac{1}{2} \pmod k$, k — целое число, то в систему вида (1) следует, вообще говоря, добавить некоторое число дополнительных потенциалов и граничных условий. Точнее, частные индексы $\varkappa_j$ символа $\widetilde{A}(y, \xi)$ системы вида (1) представимы в виде $\operatorname{Re} \varkappa_j = s + m_j + \delta_j$, где $|\delta_j| < \frac{1}{2}$ и m_j — целые числа. Пусть

$$m_+ = \sum_{j=1}^{l_0} m_j, \quad m_- = \sum_{j=l_0+1}^{p} |m_j|,$$

где $m_1 \geqslant m_2 \geqslant \ldots \geqslant m_{l_0} \geqslant 0 > m_{l_0+1} \geqslant \ldots \geqslant m_p$.

Тогда корректной является следующая задача, в которой неизвестными являются $u_+(x)$ и m_- плотностей потенциалов $(\rho_1(x'), \ldots, \rho_{m_-}(x')) = \rho(x')$:

$$P^+ A u_+ + P^+ G \rho = f(x)$$

$$(A = \|A_{ij}\|_{i,\, j=1,\, \ldots,\, p}, \quad G = \|G_{ik}\|_{i=1,\, \ldots,\, p;\ k=1,\, \ldots,\, m_-}), \qquad (14)$$

при m_+ граничных условиях на Γ

$$\gamma P^+ B u_+ = g(x') \qquad (B = \|B_{rj}\|_{r=1,\, \ldots,\, m_+;\ j=1,\, \ldots,\, p}); \qquad (15)$$

Au_+, Bu_+ — операторы в свертках, а $G\rho$ задается аналогично членам с дополнительными потенциалами в (10) или в (11). При выполнении некоторых алгебраических условий задача (14), (15) нормально разрешима в соответствующих пространствах, причем $u_+ \in \mathring{H}_s(G)$.

Если символы $\widetilde{A}(y, \xi)$, $\widetilde{G}(y, \xi)$, $\widetilde{B}(y, \xi)$ удовлетворяют условию гладкости (8), то естественно взять $s = 0$, и при гладких $f(x)$ и $g'(x)$ решение $u_+(x)$ также будет гладким вплоть до Γ.

Примечание. Недавно нам с Г. И. Эскиным удалось освободиться от сформулированного выше условия (а). Нами исследованы задачи вида (11), (12) для систем уравнений в свертках при достаточно больших M и l, $l - M = \sum m_i = \text{const}$, если символ $\widetilde{A}(x, \xi)$ удовлетворяет лишь следующему условию. Пусть в локальных координатах (y', y_n), связанных с Γ, где y_n — расстояние по нормали до границы, а ξ — двойственные с (y', y_n) координаты,

$$a_+(y') = A(y', 0; 0, 1), \quad a_-(y') = A(y', 0; 0, -1),$$

$$b(y') = a_+^{-1}(y')\, a_-(y').$$

Требуется, чтобы собственные числа $\gamma_l(y')$ матрицы $\gamma(y') = \frac{1}{2\pi i} \ln b(y')$ (которые из-за многозначности $\ln z$ определяются неоднозначно) можно было выбрать непрерывными функциями на всей границе Γ и притом так, что $|\operatorname{Re} \gamma_l(y') - \operatorname{Re} \gamma_j(y')| < 1$, $y' \in \Gamma$. При этом никаких других ограничений на частные индексы матрицы $\widetilde{A}(x, \xi)$ не налагается.

Найдены также необходимые и достаточные условия существования матриц $B(x, \xi)$ и $G(x, \xi)$, при которых соответствующая задача для оператора P^+Au_+ нормально разрешима.

VI. Приложения

1. Так как интегральные уравнения в свертках первого рода вида (5) и сингулярные интегральные уравнения и системы в ограниченной области являются частными случаями уравнений и систем в свертках, то к ним применимы приведенные выше предложения. В частности, для этих уравнений выяснены корректные постановки задач, построены левые и правые регуляризаторы, установлены априорные оценки.

2. Исследованы разрывные граничные задачи для эллиптического уравнения порядка $2m$:

$$A_{2m}(x, D)\, u(x) = f(x) \quad (x \in G), \tag{16}$$

$$B_{j1}(x, D)\, u|_{\Gamma^+} = g_{j1}(x') \quad (x \in \Gamma^+),$$

$$B_{j2}(x, D)\, u|_{\Gamma^-} = g_{j2}(x') \quad (x' \in \Gamma^-), \quad j = 1, \ldots, m, \tag{17}$$

где $\overline{\Gamma}^+ \cup \overline{\Gamma}^- = \Gamma$, $\overline{\Gamma}^+ \cap \overline{\Gamma}^- = \Gamma_{12}$ — гладкое $(n-2)$-мерное многообразие. Предполагается, что в каждой точке $\overline{\Gamma}^+$ и $\overline{\Gamma}^-$ выполнено условие Шапиро—Лопатинского. Эта задача сводится к исследованию некоторой системы так называемых парных уравнений в свертках на границе Γ. На такие системы переносится изложенная выше теория. В итоге при выполнении некоторых алгебраических условий получены теоремы о нормальной разрешимости задачи (16), (17), выяснен характер особенности u_+ на многообразии Γ_{12} разрыва граничных условий.

Найдены такие обобщенные постановки разрывных граничных задач, в граничные условия которых добавлены слагаемые типа потенциалов с плотностями, сосредоточенными на Γ_{12}, и добавлены краевые условия на Γ_{12}. Решения u_+ таких задач являются на соответствующее число единиц более или менее гладкими на Γ_{12}. В двумерном случае задача (16), (17) была исследована Петре [4].

VII. Параболические уравнения в свертках

Основные факты теории эллиптических уравнений в свертках обобщаются и на случай параболических уравнений в свертках в цилиндрической области $\Omega_T = (D \times [0 < x_0 < T])$, $D \in R^n$ $(x_1, \ldots, x_n)$, а также в нецилиндрической области Ω_T^*. Пусть $\widetilde{A}(x, \xi_0, \xi^1)$ $(x = (x_0, \ldots, x_n)$, $\xi^1 = (\xi_1, \ldots, \xi_n))$ — символ уравнения в свертках вида (1), где $G = \Omega_T$. Предполагается, что

1) $\widetilde{A}$ — аналитическая функция ξ_0 в полуплоскости $\operatorname{Im} \xi_0 > 0$;

2) $\widetilde{A}$ имеет порядок α с весом γ по ξ_0:

$$\widetilde{A}(x, t^\gamma \xi_0, t\xi^1) = t^\alpha \widetilde{A}(x, \xi_0, \xi^1), \quad t > 0, \ \gamma > 0;$$

3) $\widetilde{A}$ удовлетворяет условию параболичности: $\widetilde{A}(x, \xi_0, \xi^1) \neq 0$ при $|\xi_0| + |\xi^1| > 0$, $\operatorname{Im} \xi_0 \geqslant 0$, $\operatorname{Im} \xi^1 = 0$.

На боковой границе L цилиндра Ω_T (или Ω_T^*) производится факторизация символа $\widetilde{A}$ в трансверсальном к L направлении, и в зависимости от индекса символа для (1) ставятся те или другие задачи (краевые или с дополнительными потенциалами). Имеют место теоремы об однозначной разрешимости соответствующих задач для параболического уравнения (1) в свертках при нулевых начальных условиях при $x_0 = 0$. Важную роль при построении левого и правого обратных операторов играет понятие вольтерровских операторов по переменной x_0 (см. [6]).

VIII. Заключение

При исследовании уравнений в свертках во всем пространстве R^n или на замкнутом многообразии M^n основную роль играет изучение формул коммутации операторов в свертках и формул замены переменных. При исследовании задач для уравнений в свертках в ограниченной области фундаментальную роль играет факторизация символа на границе области в трансверсальном к ней направлении.

По всей видимости, в ближайшем будущем все основные факты, известные в теории краевых задач для основных типов дифференциальных уравнений, будут перенесены на уравнения и системы в свертках, причем для них, кроме краевых задач, возникают еще новые типы задач.

Московский университет,
Москва, СССР

ЛИТЕРАТУРА

[1] Вишик М. И., Эскин Г. И., Уравнения в свертках в ограниченной области, *УМН*, **20**, вып. 3 (1965), 89-152.

[2] Гохберг И. Ц., Крейн М. Г., Системы интегральных уравнений на полупрямой с ядрами, зависящими от разности аргументов, *УМН*, **13**, вып. 2 (1958), 3-72.

[3] Вишик М. И., Эскин Г. И., Уравнения в свертках в ограниченной области в пространствах с весовыми нормами, *Матем. сб.*, **69 (111)** : 1 (1966), 65-110.

[4] Peetre J., Mixed problems for higher order elliptic equations in two variables. I. *Ann. Scuola Norm. Sup. Pisa*, (1961), 337-353; II, 17 (1963), 1-12.

[5] Вишик М. И., Эскин Г. И., Сингулярные эллиптические уравнения и системы переменного порядка, *ДАН СССР*, **156**, № 2 (1964), 243-246.
[6] Вишик М. И., Эскин Г. И., Параболические уравнения в свертках в ограниченной области, *Матем. сб.*, **71 (113)** : 2 (1966), 162-190.
[7] Вишик М. И., Эскин Г. И., Эллиптические уравнения в свертках в ограниченной области и их приложения, *УМН*, **XXII**, вып. 1 (133), (1967), 17-76.

ОБЩИЕ СВОЙСТВА ЛИНЕЙНЫХ ДИФФЕРЕНЦИАЛЬНЫХ ОПЕРАТОРОВ С ПОСТОЯННЫМИ КОЭФФИЦИЕНТАМИ

В. П. ПАЛАМОДОВ

В докладе излагается экспоненциальное представление решений общих систем дифференциальных уравнений в частных производных с постоянными коэффициентами и некоторые следствия этого представления.

1°. Пусть $p(D)$ — произвольная прямоугольная матрица размера $t \times s$, образованная многочленами от дифференциального оператора $D = \left(i\frac{\partial}{\partial \xi_1}, \ldots, i\frac{\partial}{\partial \xi_n}\right)$, действующего в R^n. Рассмотрим соответствующую систему уравнений

$$P(D)\,u = 0, \tag{1}$$

где $u \in [D'(\Omega)]^s$, а Ω — некоторая область в R^n. Экспоненциальное представление есть разложение решения u в интеграл с некоторой мерой по множеству экспоненциальных полиномов, являющихся решениями той же системы. В простейшем случае $n = t = s = 1$ экспоненциальное представление совпадает с теоремой Эйлера о структуре решений однородного обыкновенного уравнения с постоянными коэффициентами.

В общем случае точная формулировка экспоненциального представления основывается на следующей алгебраической конструкции. Пусть P — кольцо многочленов от n переменных $z \in C^n$ с комплексными коэффициентами; для всякого целого $k > 0$, P^k — прямая сумма k экземпляров этого кольца. Пусть далее p' — матрица, транспонированная по отношению к p; ей отвечает P-отображение

$$p' \colon P^t \to P^s \tag{2}$$

(заключающееся в умножении векторов из P^t слева на матрицу $p(z)$). Несложные рассуждения из алгебраической геометрии дают следующее описание образа отображения (2): существует конечное число пар (I_λ, d_λ), $\lambda = 1, \ldots, l$, где I_λ для каждого λ есть простой

идеал в P, а

$$d_\lambda = d_\lambda(z, D) \colon P^s \to P^{s_\lambda}$$

— дифференциальный оператор в C^n с полиномиальными коэффициентами, такой, что пересечение ядер линейных отображений

$$d_\lambda \colon P^s \to [P/I_\lambda]^{s_\lambda}$$

совпадает с образом (2) (обобщение теоремы М. Нетера). Иными словами, последовательность линейных отображений

$$P^t \xrightarrow{p'} P^s \xrightarrow{d} \bigoplus [P/I_\lambda]^{s_\lambda}, \quad d = \bigoplus d_\lambda, \tag{3}$$

точна.

Отметим, что идеалы I_λ определяются последним условием однозначно (в предположении, что их число минимально): они суть радикалы примарных компонент подмодуля $p'P^t \subset P^s$. Пусть N_λ — множество общих корней в C^n многочленов из идеала I_λ. Алгебраическое многообразие $N = \bigcup N_\lambda$ совпадает с множеством решений неравенства $\operatorname{rang} p(z) < s$; это многообразие мы назовем характеристическим. Дифференциальные операторы, делающие последовательность (3) точной, мы назовем нетеровскими.

2°. *Теорема об экспоненциальном представлении.* Пусть d_λ — нетеровские операторы, а Ω — произвольная выпуклая область в R^n. Всякое решение системы (1), принадлежащее $[D'(\Omega)]^s$, может быть записано в виде

$$u = \sum_\lambda \int d'_\lambda(z, -i\xi) \exp(z, -i\xi)\, \mu_\lambda, \tag{4}$$

где μ_λ — некоторые векторные (s_λ компонент) комплексные аддитивные меры, такие, что $\operatorname{supp} \mu_\lambda \subset N_\lambda$, а интегралы (4) абсолютно сходятся в $[D'(\Omega)]^s$. Последнее означает, что для любого ограниченного множества B в $[D(\Omega)]^s$ интегралы

$$\int |d'^{t}_{\lambda}(z, -i\xi) \exp(z, -i\xi)\, \varphi(\xi)\, \mu_\lambda|, \qquad \varphi \in B, \tag{5}$$

сходятся равномерно.

Для полноты отметим обратное легко проверяемое утверждение: для любых мер μ_λ, обладающих описанными свойствами, правая часть (4) представляет обобщенную в Ω вектор-функцию, удовлетворяющую системе (1).

Если $\dim N = 0$, то каждое многообразие N_λ есть точка; следовательно, формула (4) содержит представление решения u в виде

конечной суммы экспоненциальных полиномов, удовлетворяющих системе (1). В частном случае $n = s = t = 1$ это представление сводится к упомянутой выше теореме Эйлера. Отметим, что в случае $\dim N > 0$ меры μ_λ не определены однозначно, однако можно дать описание всех мер μ_λ, приводящих к одному и тому же решению u.

Для решений, удовлетворяющих дополнительным условиям гладкости или ограничениям роста вблизи границы Ω, можно гарантировать существование представления (4), в котором меры μ_λ удовлетворяют соответствующим ограничениям роста (или убывания) на бесконечности. В частности, возможно адекватное представление бесконечно дифференцируемых решений системы (1) в выпуклой области Ω. Это представление имеет вид (4), где меры μ_λ таковы, что интегралы (5) сходятся равномерно на любом ограниченном множестве в $[\mathscr{E}'(\Omega)]^s$.

3°. *Из истории теоремы об экспоненциальном представлении.* Впервые теорема такого типа для случая $s = 1$ была анонсирована Л. Эренпрайсом в 1960 г. [1] как следствие его «фундаментального принципа». Независимо от Эренпрайса частный случай описанной выше теоремы (для $t = s = 1$) был получен автором [14]. Далее было обнаружено [17] что теорема, сформулированная Эренпрайсом, не совсем верна в случае $t > 1$. Дело в том, что, согласно его теореме, всякое решение (1) с $s = 1$ может быть записано в виде (4), где d_λ — некоторые дифференциальные операторы с постоянными коэффициентами, что, вообще говоря, неверно при $t > 1$. Теорема об экспоненциальном представлении, незначительно отличающаяся от сформулированной в 2°, была получена автором в 1960—1962 гг. в [14—17]. Независимо от автора ряд важных результатов в том же направлении был получен Мальгранжем [5, 6]. Один из результатов Мальгранжа будет сформулирован ниже.

4°. Представление (4) очень удобно для изучения локальных свойств решений (1) (для чего оно первоначально и предназначалось). В самом деле, если некоторая обобщенная функция может быть записана в виде

$$(f, \varphi) = \int \widetilde{\varphi}\mu, \qquad \varphi \in D(\Omega),$$

где $\widetilde{\varphi}$ — преобразование Фурье функции φ, а μ — некоторая комплексная мера, то чем меньше носитель этой меры, тем больше мы можем сказать о локальных свойствах самой функции f. Формула (4) дает аналогичное представление для всех решений системы (1), причем носитель соответствующей меры принадлежит характеристическому многообразию N. Этот метод позволяет получить любой из известных сейчас результатов, касающихся локальных свойств

решений системы (1), а также найти ряд новых локальных свойств (см. [15], [18]). Следствием этих локальных теорем является теорема единственности, которую мы сейчас опишем.

Разобьем переменные ξ на две группы: $\xi' = (\xi_1, \ldots, \xi_m)$ и $\xi'' = (\xi_{m+1}, \ldots, \xi_n)$; пусть (z', z'') — соответствующее разбиение двойственных переменных. Допустим, что на многообразии N выполнено неравенство

$$|z'| \leqslant B(|\operatorname{Im} z'|^{1/\gamma} + |z''|^{1/\beta} + 1)$$

с некоторыми $0 < \beta < 1$, $0 < \gamma \leqslant 1$ и $B > 0$. В таком случае оператор p является слабо гипоэллиптическим по переменным ξ'. Это означает, что всякое обобщенное решение системы (1) имеет сужение на подпространстве $\xi' = \text{const}$ и это сужение бесконечно дифференцируемо зависит от ξ'. Поэтому мы можем поставить обобщенную задачу Коши

$$D^i_{\xi'} u|_{\xi'=0} = 0 \qquad \forall i \tag{6}$$

для решений системы (1), определенных в окрестности подпространства $\xi' = 0$.

Теорема 1. Пусть $\gamma = 1$. Тогда задача Коши (6) для системы (1) в классе обобщенных функций, определенных в цилиндре $|\xi'| < \varepsilon$, растущих на бесконечности не быстрее, чем

$$C \exp(A|\xi''|^{\frac{1}{1-\beta}})$$

($\varepsilon > 0$, C и A произвольны), имеет лишь нулевое решение.

В случае $1 < \gamma < 0$ аналогичное утверждение справедливо для класса функций, определенных в R^n и растущих не быстрее, чем

$$C \exp(A(|\xi'|^{\frac{1}{1-\gamma}} + |\xi''|^{\frac{1}{1-\beta}})),$$

где A и C — произвольны.

Если $s = t = m = 1$, обобщенная задача Коши (6) эквивалентна обычной задаче Коши с конечным числом нулевых начальных данных, поскольку остальные соотношения (6) можно получить, выражая старшие производные по ξ_1 из системы (1). В этом случае теорема 1 сводится к известной теореме Гельфанда — Шилова [12] (в случае $m = 1$ всегда можно положить $\gamma = 1$).

5°. Специальный случай представления (4) содержит условия разрешимости неоднородной системы уравнений

$$p(D)u = w, \quad w \in [D'(\Omega)]^l. \tag{7}$$

Известно, что для всякой матрицы p мы можем подобрать матрицу q

того же типа, делающую последовательность

$$P^r \xrightarrow{q'} P^t \xrightarrow{p'} P^s \tag{8}$$

точной (нетеровость кольца P). Из точности этой последовательности, в частности, следует, что $p'q' = 0$, откуда $qp = 0$. Следовательно, для разрешимости системы (7) необходимо выполнение следующей однородной системы уравнений для правых частей:

$$q(D)w = 0. \tag{9}$$

С другой стороны, точность последовательности (8) означает, что матрица p' является нетеровским оператором по отношению к матрице q', а соответствующий идеал I нулевой. Поэтому, согласно теореме об экспоненциальном представлении при условии выпуклости области Ω, всякое решение системы (9) может быть записано в виде

$$w = \int p(z) \exp(z, -i\xi)\,\mu, \tag{10}$$

где мера μ такова, что интеграл

$$u = \int \exp(z, -i\xi)\,\mu \tag{11}$$

есть обобщенная функция в Ω. Сопоставляя (10) и (11), мы получаем соотношение (7). Таким образом, мы пришли к следующей теореме.

Т е о р е м а 2. Если область Ω выпукла, то для разрешимости системы (7) в пространстве $[D'(\Omega)]^s$ необходимо и достаточно выполнение системы (9).

В этой теореме пространство $D'(\Omega)$ можно заменить пространством $\mathcal{E}(\Omega)$, а также целым рядом других пространств (см. [17, 21, 22]). Впервые теорема 2 была сформулирована Эренпрайсом [1] как следствие «фундаментального принципа». Независимо от автора эта теорема была получена также Мальгранжем в несколько менее сильной форме [5, 6].

В частном случае $s = t = 1$, $p \neq 0$, мы, очевидно, имеем $q = 0$, что приводит нас к известной теореме Мальгранжа — Эренпрайса [4], [2] о разрешимости одного уравнения с одной неизвестной функцией.

Мальгранж в [5] отметил, что выпуклые области образуют наиболее широкий класс связных областей, для которых справедлива теорема 2 (и, следовательно, теорема об экспоненциальном представлении). Однако для каждого индивидуального оператора p класс областей, для которых справедлива теорема 2, называемых иногда p-выпуклыми областями, шире. Например, если $s = t$, а p — эллиптический оператор, то всякая область является p-выпуклой. Описание всех p-выпуклых областей для произвольного оператора p составляет одну из центральных проблем теории. Хорошо изучен скалярный случай $s = t = 1$ (Мальгранж [4], Хёрмандер [9]). В общей постановке имеется ряд частных результатов, которые в основном являются аналогами соответствующих теорем из теории функций многих комплексных переменных [7, 8, 19]. Отмечу теорему конечности типа Андреотти — Грауэрта, дающую условие p-выпуклости области в форме существования в ней вещественной функции, растущей при приближении к границе, гессиан которой имеет достаточно много положительных собственных значений. Аналог теоремы Серра о топологии областей Рунге содержит необходимое условие p-выпуклости в терминах обращения в нуль старших гомологий области. Более подробно эти результаты изложены в [22].

6°. *Переопределенные системы.* Этот термин употребляется довольно часто, но не имеет точного значения. Эвристический смысл этого термина заключается в том, что система тем более переопределена, чем меньше она имеет решений. Ввиду этого следующая теорема, характеризующая «количество» решений системы, может служить подходом к точному определению этого термина.

Т е о р е м а 3. Пусть $d = \dim N$, π — произвольный полиэдр в R^n, а Γ_d есть d-мерный остов этого полиэдра. Обобщенные решения системы (1), определенные в окрестности π, непрерывно зависят (в топологии обобщенных функций) от своих значений в сколь угодно малой окрестности Γ_d. Для остова Γ_{d-1} аналогичное утверждение неверно.

Говоря весьма приближенно, d есть число «свободных параметров», от которых зависят решения системы (1), и, следовательно, может быть использовано как мера переопределенности этой системы. Отметим, что для системы с квадратной матрицей, имеющей отличный от константы определитель, всегда $d = n - 1$. Ввиду этого естественно ввести следующее

О п р е д е л е н и е. Систему (1) назовем определенной, если $\dim N < n$, и переопределенной, если $\dim N < n - 1$.

Из теоремы 3, в частности, вытекает следующий закон распространения (бесконечной) дифференцируемости: всякое a priori обобщенное решение системы (1), определенное в окрестности π, дифференцируемое в окрестности Γ_d, является дифференцируемым в окрестности всего полиэдра π.

Т е о р е м а 4. Описанный закон распространения дифференцируемости справедлив при менее сильных предположениях: каждое

многообразие N_λ либо гипоэллиптично, либо его размерность не выше d.

Из теоремы 3 вытекает также следующая теорема единственности: если $u = 0$ в окрестности Γ_d, то $u = 0$ в окрестности всего π. Опишем более сильный результат. Пусть (ξ', ξ'') — разбиение переменных ξ на две группы (см. 4°). Скажем, что алгебраическое многообразие $M \subset C^n$ гиперболично по ξ', если для любой его несобственной точки $(0, w', w'')$ из $\operatorname{Im} w'' = 0$ следует $\operatorname{Im} w' = 0$.

Теорема 5. Предположим, что для каждого λ либо $\dim N_\lambda = n - m$ и многообразие N_λ не гиперболично по ξ', либо $\dim N_\lambda < n - m$. Тогда всякое решение (1), определенное в окрестности подпространства $\xi' = 0$, равное нулю вне некоторого компакта $K \subset R^n$, равно нулю в окрестности всего подпространства $\xi' = 0$.

В частном случае $s = t = m = 1$ эта теорема сводится к известным теоремам единственности Джона [10] и Бродды [11].

7°. *Продолжение решений переопределенных систем.* Рассмотрим наиболее простую задачу продолжения решений. Пусть дана область $\Omega \subset R^n$ и компакт $K \subset \Omega$; требуется описать множество решений системы (1), определенных в $\Omega \setminus K$, которые не продолжаются в Ω (как решения той же системы). Это удобно сделать в следующих терминах. Пусть S — частное пространства всех решений системы (1), определенных в $\Omega \setminus K$, по его подпространству, образованному теми решениями, которые, будучи исправлены в сколь угодно малой окрестности K, продолжаются в Ω. В случае когда компакт K выпуклый, пространство S допускает следующее описание.

Для этого рассмотрим последовательность

$$P^s \xrightarrow{p} P^t \xrightarrow{q} P^r, \tag{12}$$

полученную из (8) заменой матриц на транспонированные. Эта последовательность не обязательно точна, но полуточна. Применив к матрице p конструкцию 1°, мы построим для нее систему нетеровских операторов и соответствующих идеалов $(J_\alpha, \partial_\alpha)$, $\alpha = 0, 1, \ldots, a$, так что пересечение ядер операторов

$$\partial_\alpha\colon\ P^t \longrightarrow [P/J_\alpha]^t\alpha$$

совпадает с образом матрицы p в (12). Поскольку образ этой матрицы принадлежит ядру матрицы q, мы можем считать, что один из нетеровских операторов, пусть для определенности ∂_0, совпадает с q, а соответствующий идеал J_0 нулевой. В этом случае остальные идеалы J_α ненулевые. Пусть $M_\alpha \subset C^n$ — множество общих корней многочленов из J_α.

Рассмотрим далее векторы $f = (f_1, \ldots, f_a)$, образованные функциями f_α: $M_\alpha \to C^t\alpha$, удовлетворяющими условию: для любой точки $z \in \bigcup M_\alpha$ существует голоморфная в этой точке вектор-функция F, такая, что

$$\partial_\alpha F|_{M_\alpha} = f_\alpha, \qquad \alpha = 1, \ldots, a,$$

в окрестности z. Векторы f такого типа образуют линейное пространство. В этом пространстве выделим подпространство F, образованное векторами f, которые при каждом $\varepsilon > 0$ удовлетворяют неравенству

$$|f_\alpha(z)| \leqslant C(|z|+1)^q \exp(\varepsilon|\operatorname{Im} z|) \sup_{\xi \in K} \exp(-\operatorname{Im} z, \xi)$$

с некоторыми C и q.

Теорема 6. Имеется естественный изоморфизм между пространствами S и E.

В частном случае $s = t = 1$ этот изоморфизм был впервые отмечен В. В. Грушиным [13]. Рассмотрим подробнее специальный случай этой теоремы, когда $s = t = 1$, $n = 2$, а p — оператор Коши — Римана. Пространство S в этом случае можно отождествить с пространством аналитических в $C^1 \setminus K$ функций, убывающих на бесконечности. С другой стороны, мы имеем $q = 0$, $\partial_1 = 1$, $M_1 = \{z\colon z_1 = -iz_2\}$. Поэтому элементы пространства E мы можем рассматривать как целые функции от z_2 первого порядка роста, индикаторные диаграммы которых принадлежат компакту K^*, симметричному к K относительно начала координат. В такой ситуации изоморфизм теоремы 6 превращается в известный изоморфизм Бореля.

Отметим другой частный случай: последовательность (12) точна. Мы имеем в этом случае $a = 0$, откуда $E = 0$, и, следовательно, $S = 0$. Таким образом, всякое решение системы (1), определенное в $\Omega \setminus K$, может быть продолжено в Ω, будучи предварительно исправленным в сколь угодно малой окрестности K. Этот результат был найден Мальгранжем [6] и автором совместно с В. В. Грушиным [20]; несколько более слабая его форма была получена ранее Эренпрайсом [3].

Условие точности последовательности (12) тесно связано со свойством переопределенности системы (1), а именно справедливо следующее утверждение. Для того чтобы всякое решение (1) в $\Omega \setminus K$ единственным образом продолжалось в Ω, необходимо и достаточно, чтобы система (1) была переопределенной.

В общем случае изоморфизм теоремы 6 в комбинации с принципом Фрагмена — Линделёфа, примененным к элементам пространства E, позволяет получить более тонкие результаты о возможности продолжения [22].

Более сложные задачи продолжения не получили еще столь исчерпывающего решения. Отметим один результат, дающий достаточное условие продолжимости. Отнеся матрице p P-модуль $M = P^s/p'P^t$, мы получим соответствие между всеми матрицами p и всеми P-модулями конечного типа. Следовательно, со всяким P-модулем конечного типа мы можем связать набор алгебраических многообразий N_λ, $\lambda = 1, \ldots, l$ (этот набор зависит лишь от самого модуля M).

Теорема 7. Пусть $\mathrm{Ext}^i(M, P) = 0$, $i = 1, \ldots, m$, а каждое многообразие L_λ, связанное с модулем $\mathrm{Ext}^{m+1}(M, P)$, не гиперболично по переменным ξ'. Тогда всякое решение системы (1), определенное в окрестности $\varkappa \setminus \omega$, где ω — выпуклая область в подпространстве $\xi' = 0$, а $\varkappa \supset \omega$ — компакт, может быть продолжено в окрестность ω.

Московский университет,
Москва, СССР

ЛИТЕРАТУРА

[1] Ehrenpreis L., Proc. Int. Symp. on Linear Spaces, Jerusalem 1960.
[2] Ehrenpries L., *Amer. J. Math.*, 76, № 4 (1954), 883-903; *Amer. J. Math.*, 78, № 4 (1956), 685-715.
[3] Ehrenpreis L., *Bull. Amer. Math., Soc.*, 67, № 5 (1961), 507-509.
[4] Malgrange B., *Ann. Inst. Fourier*, 6 (1956), 271-356.
[5] Malgrange B., Séminaire Bourbaki, 1962.
[6] Malgrange B., Séminaire Leray, 1961/62.
[7] Malgrange B., Séminaire Leray, 1962/63.
[8] Malgrange B., Differential Analysis, Bombay Coll., 1964.
[9] Hörmander L., *Ann. Math.*, 76, № 1 (1962), 148-170.
[10] John F., *Comm. Pure Appl. Math.*, 10 (1957), 391-398.
[11] Brodda B., *Math. Scand.*, 9 (1961), 55-68.
[12] Гельфанд И. М., Шилов Г. Е., Обобщенные функции, вып. 3, Физматгиз, М., 1958.
[13] Грушин В. В., Труды Моск. мат. о-ва, т. 15.
[14] Паламодов В. П., *ДАН СССР*, 137, № 4 (1961), 774-777.
[15] Паламодов В. П., *ДАН СССР*, 140, № 5 (1961), 1015-1018.
[16] Паламодов В. П., *ДАН СССР*, 143, № 6 (1962), 1278-1281.
[17] Паламодов В. П., *ДАН СССР*, 148, № 3 (1963), 523-526.
[18] Паламодов В. П., *ДАН СССР*, 156, № 6 (1964), 1288-1291.
[19] Паламодов В. П., *ДАН СССР*, 161, № 5 (1965), 1015-1018.
[20] Паламодов В. П., *УМН*, 18, вып. 2 (1963), 164-167.
[21] Паламодов В. П., Диссертация, МГУ, 1965.
[22] Паламодов В. П., Линейные дифференциальные операторы с постоянными коэффициентами, изд-во «Наука», М., 1968.

8 Топология / Topology / Topologie / Topologie

ISOTOPIE ET PSEUDO-ISOTOPIE [1])

JEAN CERF

1. Définitions

Soit V une variété différentiable de classe C^∞ de dimension $n-1$.

Les difféomorphismes de V forment un groupe noté $\mathrm{Diff}\, V$; on va définir deux relations d'équivalence dans $\mathrm{Diff}\, V$.

Définition 1. Une *isotopie* de V est un chemin différentiable dans $\mathrm{Diff}\, V$, d'origine l'identité, i.e. une application:

$$I \ni t \to f_t \in \mathrm{Diff}\, V, \quad (\text{où } I = [0, 1])$$

telle que:

$$\begin{cases} f_0(x) = x \quad \forall x \in V \\ \text{l'application } (x, t) \to f_t(x) \text{ est différentiable.} \end{cases}$$

L'application:

$$(x, t) \to (f_t(x), t)$$

est alors un difféomorphisme du cylindre $V \times I$. L'ensemble des isotopies de V s'identifie donc au sous-groupe $\mathscr{H}$ de $\mathrm{Diff}\,(V \times I)$ formé des g tels que:

1°) $g(x, 0) = x \quad \forall x \in V$

2°) $p \circ g = p$ (où p désigne la projection $V \times I \to I$).

Le groupe $\mathscr{H}$ opère dans $\mathrm{Diff}\, V$ par la formule:

$$g \cdot f(x) = g(f(x), 1).$$

Deux éléments de $\mathrm{Diff}\, V$ qui sont dans la même orbite sont dits *isotopes*. Les orbites coïncident avec les *composantes connexes* de $\mathrm{Diff}\, V$

[1]) Résumé d'un travail à paraître aux «Publications de l'Institut des Hautes Etudes Scientifiques».

muni de la topologie C^∞ (car tout chemin continu dans Diff V peut être approché par un chemin différentiable).

Définition 2. Une *pseudo-isotopie* de V est un difféomorphisme de $V \times I$ qui vérifie la condition 1° (mais pas nécessairement 2°).

Les pseudo-isotopies de V forment un groupe qu'on note $\mathcal{G}$; $\mathcal{G}$ opère dans Diff V ; deux éléments qui sont dans la même orbite sont dits *pseudo-isotopes*.

Comme $\mathcal{H} \subset \mathcal{G}$, « isotope » implique « pseudo-isotope ».

Problème. Sous quelle condition ces deux classifications sont-elles les mêmes?

2. Résultats

Théorème. *Si V est compacte sans bord, si $\pi_1(V) = \pi_2(V) = 0$ et si $n \geqslant 10$ (où n est la dimension de V), alors $\mathcal{G}$ est connexe (pour la topologie C^∞).*

Corollaire 1. *Sous les hypothèses du théorème, les deux classifications de* Diff V *(isotopie et pseudo-isotopie) sont les mêmes.*

[En effet les orbites de $\mathcal{G}$ sont alors connexes ; comme elles contiennent les orbites de $\mathcal{H}$ qui sont les composantes connexes, ce sont les composantes connexes.]

Cas particulier où V est la sphère S^{n-1}. Alors « f pseudo-isotope à l'identité » équivaut à « f peut se prolonger en un difféomorphisme de la boule D^{n+1} ». On en déduit :

Corollaire 2. *Pour $n \geqslant 10$:*

$$\begin{cases} \pi_0(\mathrm{Diff}\, D^n) = 0; \\ \pi_0(\mathrm{Diff}\, S^{n-1}) \approx \Gamma_n; \\ \pi_1(\mathrm{Diff}\, S^{n-1}) \text{ est une extension de } \Gamma_{n+1}. \end{cases}$$

Remarques. 1) Je ne connais pas de contre-exemple lorsque $\pi_1(V)$ ou $\pi_2(V)$ sont différents de zéro.

2) Les premiers exemples de $\pi_i(\mathrm{Diff}\, S^n)$ non triviaux pour $i \geqslant 1$ sont dus à S. P. Novikov [1]; à ma connaissance, on ne sait rien sur $\pi_i(\mathcal{G})$ pour $i \geqslant 1$.

3. Principe de la démonstration

Pour démontrer :

$$\pi_0(\mathcal{G}) = 0 \tag{1}$$

on fait opérer $\mathcal{G}$ dans l'espace $\mathcal{F}$ des fonctions différentiables de classe C^∞ :

$$V \times (I, 0, 1) \to (I, 0, 1);$$

les opérations sont définies par la formule :

$$g \cdot f = f \circ g^{-1}.$$

La projection $p : V \times I \to I$ est un élément de $\mathcal{F}$. Soit $\mathcal{E}$ l'orbite de p. Le sous-groupe de $\mathcal{G}$ formé des difféomorphismes laissant p invariant n'est autre que $\mathcal{H}$; donc $\mathcal{E}$ est homéomorphe à l'espace homogène $\mathcal{G}/\mathcal{H}$. Cet espace homogène est localement trivial, et la fibre $\mathcal{H}$ est contractile (comme espace des chemins différentiables d'origine fixe dans Diff V) ; donc $\mathcal{G}$ est homéomorphe à $\mathcal{H} \times \mathcal{E}$; on est donc ramené à démontrer :

$$\pi_0(\mathcal{E}) = 0. \tag{2}$$

Tout élément f de $\mathcal{E}$ est une fonction sans point critique (puisque p a zéro point critique). La réciproque se montre facilement (à l'aide des lignes de gradient relatives à une métrique riemannienne de $V \times I$). Donc *$\mathcal{E}$ est le sous-espace de $\mathcal{F}$ formé des fonctions ayant zéro point critique.*

Puisque $\mathcal{F}$ est convexe, (2) équivaut à :

$$\pi_1(\mathcal{F}, \mathcal{E}) = 0 \tag{3}$$

ce qui apparaît comme une généralisation à 1 paramètre de la théorie du h-cobordisme de Smale. Les espaces analogues à $\mathcal{E}$ et $\mathcal{F}$ peuvent en effet être définis pour toute triade (W, V_0, V_1) (variété dont le bord a deux composantes connexes, V_0 et V_1). La théorie de Smale consiste à montrer que (moyennant des conditions homotopiques convenables et une condition de dimension) W est difféomorphe au cylindre $V_0 \times I$; or les cylindres sont caractérisés parmi toutes les triades par la condition $\mathcal{E} \neq \varnothing$, qui peut s'écrire :

$$\pi_0(\mathcal{F}, \mathcal{E}) = 0.$$

4. Subdivision co-cellulaire de $\mathcal{F}$

La notion de *codimension d'une singularité*, due à Thom, permet de munir l'espace $\mathcal{F}$ d'une *subdivision co-cellulaire* $\mathcal{F}^0, \mathcal{F}^1, \mathcal{F}^2$, etc.... La théorie de Smale nécessite la connaissance explicite de $\mathcal{F}^0$ et certaines informations sur $\mathcal{F}^1$. L'étude de $\pi_1(\mathcal{F}, \mathcal{E})$ nécessite la connaissance explicite de $\mathcal{F}^0$ et $\mathcal{F}^1$, et certaines informations sur $\mathcal{F}^2$.

$\mathcal{F}^0 \subset \mathcal{F}$ est l'espace des fonctions de Morse, i.e. les fonctions dont tous les points critiques sont quadratiques non dégénérés, et toutes les valeurs critiques distinctes. D'après un théorème classique de M. Morse, *$\mathcal{F}^0$ est ouvert et dense dans $\mathcal{F}$*. Les composantes connexes de $\mathcal{F}^0$ sont appelées *cellules de codimension* 0.

$\mathcal{F}^1 \subset \mathcal{F} - \mathcal{F}^0$ est la réunion (disjointe) de $\mathcal{F}^1_\alpha$ et $\mathcal{F}^1_\beta$ définis comme suit :

$f_i \in \mathscr{F}^1_\alpha \Leftrightarrow$ { f a un point critique du type $-x_1^2 - \ldots - x_i^2 + x_{i+1}^2 + \ldots + x_{n-1}^2 + x_n^2$; tous les autres points critiques sont du type de Morse ; toutes les valeurs critiques sont distinctes.

$f \in \mathscr{F}^1_\beta \Leftrightarrow$ { tous les points critiques sont du type de Morse ; toutes les valeurs critiques sont distinctes, sauf exactement deux d'entre elles.

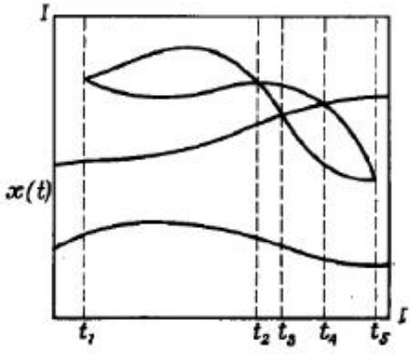
Propriétés de $\mathscr{F}^1$

(a) $\mathscr{F}^1$ est *ouvert et dense* dans $\mathscr{F} - \mathscr{F}^0$

(b) $\mathscr{F}^1$ est de *codimension* 1 dans $\mathscr{F}^0 \cup \mathscr{F}^1$ (au sens suivant: il existe des voisinages tubulaires locaux de fibre [0, 1]).

(c) $\mathscr{F} - (\mathscr{F}^0 \cup \mathscr{F}^1)$ est de *codimension* > 1.

On dit qu'un chemin α dans $\mathscr{F}$ est *bon* si $\alpha(t) \in \mathscr{F}^0$ sauf pour un nombre fini de valeurs de t, et si, pour ces valeurs exceptionnelles,



$\alpha(t)$ traverse $\mathscr{F}^1$ (ce qui a un sens en vertu de (b)). *L'espace des bons chemins dans $\mathscr{F}$ est dense dans l'espace de tous les chemins, muni de la topologie C^0.*

En particulier *l'espace des bons lacets relatifs de $(\mathscr{F}, \mathscr{C})$ définit un système de générateurs de $\pi_1(\mathscr{F}, \mathscr{C})$.*

Graphique d'un bon chemin.

Soit α un chemin dans $\mathscr{F}$; son *graphique* Γ est la partie de $I \times I$ définie par la condition :

$$(t, \gamma) \in \Gamma \Leftrightarrow \gamma \text{ est valeur critique de } \alpha(t).$$

Si α est bon, son graphique est du type de la figure 1 (où les valeurs exceptionnelles du paramètre sont t_1, t_2, t_3, t_4, t_5).

5. Lemmes sur la forme des cellules

Les cellules de codimension 0 et les cellules de codimension 1 (i. e. les composantes connexes de $\mathscr{F}^1$) *ne sont pas acycliques* en général. On a besoin d'informations sur leur forme, notamment sur le π_1 des cellules de codimension 0 modulo leur bord.

Soit σ^0 une cellule de codimension 0 ; soit σ^1 une réunion de cellules de codimension 1 telle que $\sigma^1 \subset \overline{\sigma^0}$. Soit $\mathscr{C}$ l'espace des bons chemins dans $\sigma^0 \cup \sigma^1$ dont l'origine soit un point donné $f \in \sigma^0$, et qui rencontrent σ^1 en un seul point. Soient c et c' deux points critiques consécutifs de f tels que $f(c) > f(c')$; on désigne par V' la « surface de niveau intermédiaire » $f^{-1}\left(\frac{f(c)+f(c')}{2}\right)$.

Lemme de croisement. *Si σ^1 est relatif au croisement de c et c', alors*

$$\pi_0(\mathscr{C}) = 0 \qquad \text{si indice } c < \text{indice } c'$$

$$\pi_0(\mathscr{C}) \approx \mathbf{Z} \qquad \text{si} \begin{cases} \text{indice } c = \text{indice } c', \\ \pi_1(V) = 0,\ 3 \leqslant \text{indice } c \leqslant n-3. \end{cases}$$

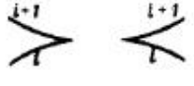
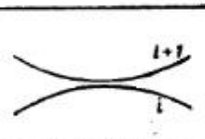
Lemme d'unicité des naissances. *Si σ^1 est relatif à la naissance de deux points critiques d'indices donnés i et $i+1$ entre les niveaux de c et c', alors:*

$$\pi_0(\mathscr{C}) \approx \pi_0(V').$$

Lemme d'unicité des morts. *Si c et c' peuvent être tués l'un par l'autre, et si σ^1 est relatif à cette mort, alors:*

$$\pi_0(\mathscr{C}) = 0 \quad \text{si} \quad \pi_1(V') = \pi_2(V') = 0,\ \dim V \geqslant 9.$$

Interprétation graphique:

	Lemme de croisement	Unicité des naissances	Unicité des morts
Concerne la possibilité de déformer le graphique ci-contre:			
en le graphique ci-contre:			

Principe de la démonstration de ces lemmes

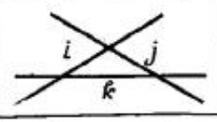
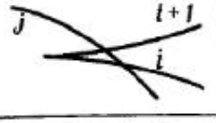
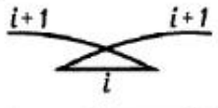
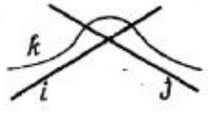
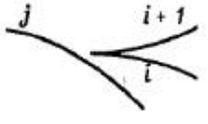
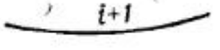
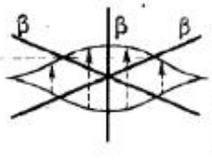
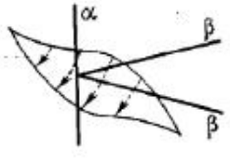
On définit une partie $\tilde{\mathscr{C}}$ de $\mathscr{C}$, dont les éléments sont dits « chemins élémentaires de traversée de σ^1 » ; ils sont définis dans chaque cas par transport d'une « déformation modèle » relative au modèle de la singularité correspondante ; on montre (en utilisant la transitivité des opérations de $\mathscr{G} \times \mathrm{Diff}\, I$ dans chaque cellule) que $\pi_0(\tilde{\mathscr{C}}) \approx \pi_0(\mathscr{C})$. Pour le a) du lemme de croisement (resp. pour le lemme d'unicité des naissances) il est immédiat que $\pi_0(\tilde{\mathscr{C}}) = 0$ (resp. $\pi_0(\tilde{\mathscr{C}}) \approx \pi_0(V')$). Les autres cas sont plus difficiles, en particulier le lemme d'unicité des morts (qui est une extension à 1 paramètre du « cancellation lemma » de Smale) ; sa démonstration utilise :

a) une généralisation à 1 paramètre du théorème de Whitney sur la suppression des points doubles ; c'est le seul point de toute la démonstration où intervient la nullité de $\pi_2(V)$.

b) une généralisation à k paramètres des théorémes de plongements de Haefliger ; cette généralisation a été faite par J. P. Dax (à paraître) ; c'est le seul point de toute la démonstration par lequel intervient la condition $n \geqslant 10$; il semble qu'elle puisse être affaiblie ($n \geqslant 8$ suffit vraisemblablement).

6. Lemmes sur l'agencement des cellules au voisinage de $\mathscr{F}^2$

Voici leur énoncé en termes de graphique:

	Lemme du triangle	Lemme du bec	Lemme de la queue d'aroude
Concerne la possibilité de déformer le graphique ci-contre:			
En le graphique ci-contre:			
Schéma dans l'espace fonctionnel (α-naissance, β-croisement)			

	Lemme du triangle	Lemme du bec	Lemme de laqueue d'aronde
Singularité de codimension 2 correspondante	point triple défini par l'égalité de 3 valeurs critiques	naissance à un niveau critique	singularité S^1 ($S^1(S^1)$) du type «queue d'aronde», i. e. $q(x_1,\dots,x_{n-1}) + xn^4$, où q est une forme quadratique
Conditions suffisantes de possibilité	$k > \inf(i, j)$ ou $i + j < n$ ou $i = j = k$, et $\pi_1(V) = 0$	$+ j < n-1$ ou $j < i$ ou $j = i$, $3 \leqslant i \leqslant n-3$ et $\pi_1(V) = 0$	conditions du lemme d'unicité des morts

Remarque sur le lemme de la queue d'aronde : on montre dans ce cas que la cellule σ^1_β de $\mathscr{F}^1_\beta$ relative au croisement des deux valeurs critiques d'indice $(i+1)$ est une *hypersurface à un seul côté* de $\mathscr{F}$. On en déduit que pour tout $f \in \sigma^1_\beta$, il existe un difféomorphisme de $V \times I$ laissant f invariante et échangeant les deux points critiques c et c' dont les valeurs sont égales; un bon exemple de ce fait est fourni par le cas particulier $n = 2$, $i = 1$.

7. Filtration de Smale de $\mathscr{F}^0$; fin de la démonstration

En théorie de Smale on caractérise la complexité d'un élément f de $\mathscr{F}^0$ par des invariants (ne dépendant que de la cellule de f) :

1) *le nombre d'inversions* v (une inversion est un couple (c, c') de points critiques tels que $f(c) > f(c')$ et indice $c <$ indice c').

2) *l'intervalle des indices* $[i, j]$: plus petit intervalle entier contenant les indices de tous les points critiques.

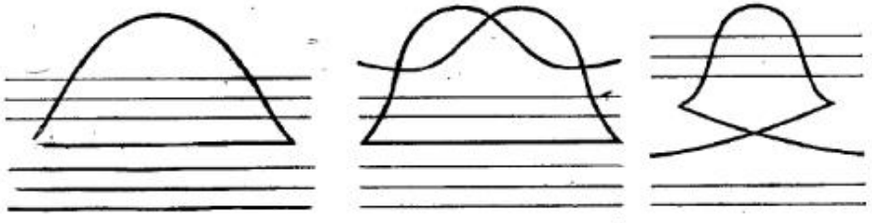
3) *le nombre total de points critiques* τ. A ces invariants est associée une *filtration* de $\mathscr{F}^0$; on montre en théorie de Smale qu'il existe pour tout $f \in \mathscr{F}^0$ un bon chemin *décroissant* d'origine f, d'extrémité dans $\mathscr{E}$. Il en résulte que pour toute paire $(\mathscr{A}, \mathscr{B})$ consécutive de cette filtration, $\pi_1(\overline{\mathscr{A}}, \overline{\mathscr{B}})$ est engendré par deux sortes de générateurs :

1°) *générateurs de* 1ère *espèce*, attachés à chaque 0-cellule σ^0 de $\mathscr{A} - \mathscr{B}$; ils sont définis par les paires de chemins décroissants issus d'un même point de σ^0.

2°) *générateurs de* 2ème *espèce*, attachés à chaque 1-cellule σ^1 de $\overline{\mathscr{A}} - \overline{\mathscr{B}}$: ils sont définis par un chemin de traversée de σ^1, et par les paires de chemins décroissants issus des extrémités de ce chemin.

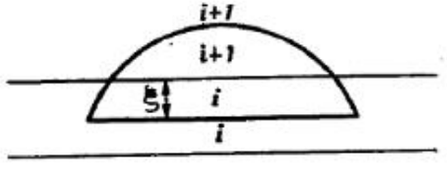
Une fonction pour laquelle $v=0$ est dite *ordonnée* ; on note $\mathscr{N}$ la partie de $\mathscr{F}^0$ formée par les fonctions ordonnées. On montre d'abord, dar récurrence sur V que $\pi_1(\mathscr{F}, \overline{\mathscr{N}})=0$ (ce qui a d'ailleurs pour conséquence le fait que $\overline{\mathscr{N}}$ *est connexe*) ; ce résultat est vrai *sans aucune condition* sur V.

Puis on se ramène à l'étude de $\pi_1(\overline{\mathscr{N}}_i, \mathscr{E})$ (où $\mathscr{N}_i$ est la partie de $\mathscr{N}$ formée des fonctions dont les points critiques sont tous d'indice i ou $i+1$). Par le procédé ci-dessus, on obtient des générateurs des types suivants pour les groupes d'homotopie relatifs associés à la filtration de $\mathscr{N}_i$ définie par l'invariant τ :



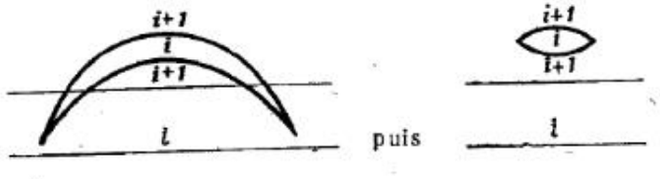
1ère espèce 2ème espèce

Voici par exemple comment on réduit les générateurs de 1ère espèce. Par un changement convenable de fonction, et par le choix d'un système de nappes descendantes convenable, on se ramène à la situation suivante :



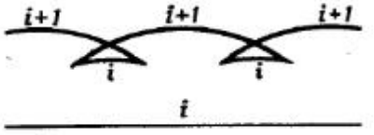
où le nombre d'intersection ξ des deux points critiques du milieu est soit zéro, soit 1, soit 2.

Dans le premier cas ($\xi = 0$) on peut déformer comme suit :



et on est ramené au lemme d'unicité des morts.

Dans le second cas ($\xi = 1$) on déforme en :



et on termine en appliquant deux fois le lemme de la queue d'aronde.

Dans le troisième cas ($\xi = 2$) on se ramène à une «chaîne» :

elle que la matrice d'intersection au point exceptionnel soit :

$$\begin{pmatrix} -1 & 1 \\ -2 & 1 \end{pmatrix}$$

On déforme la chaîne comme suit :

et on termine en appliquant deux fois le lemme de la queue d'aronde.

Université de Paris, Faculté des Sciences,
Dept. de Mathématiques, Paris, France.

RÉFÉRENCES

[1] Новиков С. П., Гомологические свойства группы диффеоморфизмов сферы, *ДАН СССР*, 148 (1963), 32-35.

KNOTTED SPHERES AND RELATED GEOMETRIC PROBLEMS

A. HAEFLIGER

The aim of this report is to show how the different kinds of groups of knotted spheres arise as obstructions in several non-stable problems of comparison between differentiable, piecewise linear and homotopy

categories. For the stable case, see the reports of C.T.C. Wall and M.W. Hirsch in these proceedings.

1. Comparison between differentiable and homotopy tubes

1.1. Classifying spaces. Let O_q be the group of orthogonal transformations of the $(q-1)$-sphere S^{q-1} and let G_q be the monoid of the homotopy equivalences of S^{q-1}. By suspension, we have homomorphisms $O_q \to O_{q+1}$ and $G_q \to G_{q+1}$; passing to the limits, we get the stable groups O and G, and the commutative diagramm

$$\begin{array}{ccc} O_q & \to & G_q \\ \downarrow & & \downarrow \\ O & \to & G \end{array}$$

We have a corresponding commutative diagramm for classifying spaces (see in particular Stasheff [20]):

$$\begin{array}{ccc} BO_q & \to & BG_q \\ \downarrow & & \downarrow \\ BO & \to & BG \end{array}$$

We can assume that all these maps are fibrations.

Homotopy classes of a complex K in BO_q (resp. BG_q) are in 1-1 correspondence with isomorphism classes of vector bundles of rank q over K (resp. $(q-1)$-spherical fiber spaces over K, i.e. whose fibers have the homotopy type of S^{q-1}).

1.2. Different kinds of tubes. A 1-connected Poincaré complex in dimension n is a 1-connected finite complex P with a fundamental integral class $p \in H_n(P, \mathbf{Z})$, such that the cup product with p defines an isomorphism $H^k(P, \mathbf{Z}) \to H_{n-k}(P, \mathbf{Z})$. This is the notion corresponding in the homotopy category to the notion of a closed oriented manifold of dimension n.

For a closed differentiable manifold V, *tubular neighbourhoods* of codimension q are classified by vector bundles of rank q (namely the normal bundle). By analogy, a *homotopy tube* of codimension q around a Poincaré complex P will be a $(q-1)$-spherical fiber space over P.

Following Wall [22], a *$(n+q)$-thickening* of a 1-connected finite complex K is a homotopy equivalence $\varphi: K \to X$ of K in a compact differentiable manifold X of dimension $n+q$, such that X and its boundary ∂X are 1-connected.

A thickening φ of a Poincaré complex P is a mixed notion of tube, because it defines a $(q-1)$-spherical bundle over P. Indeed one has

the following lemma whose proof was communicated to me by W. Browder and which is essentially due to Spivak [18].

L e m m a. *If P is a 1-connected Poincaré complex in dimension n and $\varphi: P \to X$ a $(n+q)$-thickening, where $q \geqslant 3$, then the inclusion $\partial X \to X \approx P$ is homotopy equivalent to a $(q-1)$-spherical fibration over P.*

1.3. The universal space BD_q for thickenings. BD_q will be the fiber product $BO \times_{BG} BG_q$ of the fibrations $BO \to BG$ and $BG_q \to BG$. One has

$$\pi_n(BD_q) = \pi_n(G, O, G_q, *)$$

the set of homotopy classes of quadruples $(D^n, D_+^{n-1}, D_-^{n-1}, S^{n-2})$ in $(G, O, G_q, *)$, where D_+^{n-1} and D_-^{n-1} are two opposite hemispheres in the boundary of the unit (n)-ball D^n, S^{n-2} being the intersection of D_+^{n-1} and D_-^{n-1} and $*$ a base point in G, for instance the identity map of S^{q-1}.

For $n+q \geqslant 6$ and $q \geqslant 3$, the $(n+q)$-thickenings of S^n are in 1-1 correspondence with the group FC_{n-1}^q of isotopy classes of embeddings of $S^{n-1} \times D^q$ in S^{n+q-1}. To such an embedding f is associated the thickening obtained in glueing to D^{n+q} the handle $D^n \times D^q$ by $f: \partial D^n \times D^q \to \partial D^{n+q}$ (cf. [5]). Moreover it was proved in [5] that FC_{n-1}^q is isomorphic to $\pi_n(G, O, G_q, *)$ so that the $(n+q)$-thickenings of S^n correspond to the elements of $\pi_n(\tau_q)$.

More generally, we have the following theorem whose proof is a straightforward application of a theorem of Sullivan [21] which formulates Browder-Novikov theory [1], [16], extended by Wall [23] to manifolds with boundary, under the form of an obstruction theory (see also Spivak [18]).

1.4. T h e o r e m A. *For $q \geqslant 3$ and $n+q \geqslant 6$, isomorphism classes of $(n+q)$-thickening of a compact differentiable n-manifold V correspond bijectively to the set $[V, BD_q]$ of homotopy classes of V in BD_q.*

If $\varphi: V \to X$ is a thickening, the corresponding map of V in τ_q is the fiber product of the map $V \to BG_q$ given by Lemma 1.2 and the map $V \to BO$ classifying the stable normal bundle of φ.

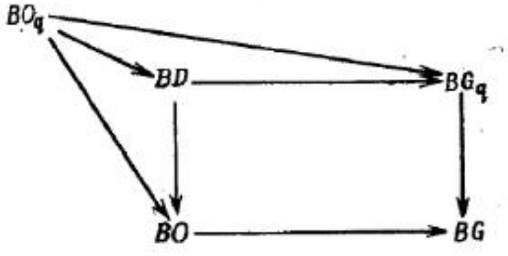
One has a natural pairing $BD_q \times BD_r \to BD_{q+r}$, hence a notion of Whitney sum for thickenings of a differentiable manifold which behave like vector bundles.

1.5. Remark. More generally, one proves that, for $q \geqslant 3$ and $n+q \geqslant 6$, the isomorphism classes of $(n+q)$-thickenings of a 1-connected Poincaré complex P of dimension n are in 1-1 correspondence with the homotopy classes of sections of a bundle over P with fiber BD_q.

1.6. Homotopy equivalences and embeddings. Let $\varphi: V \to X$ be a $(n+q)$-thickening of a differentiable closed n-manifold V. When is

φ homotopic to an embedding? For $n+q \geqslant 5$ and $q \geqslant 3$, this is equivalent by the h-cobordism theorem to the question: is X a differentiable tubular neighbourhood of V?

We have a unique map $BO_q \to BD_q$ making the diagramm commutative:



This is a fibration whose fiber we call C_q:

$$C_q \to BO_q \to BD_q.$$

One has $\pi_{n-1}(C_q) = \pi_n(G, O, G_q, O_q)$, the set of homotopy classes of $(D^n, D_+^{n-1}, D_-^{n-1}, S^{n-2})$ in (G, O, G_q, O_q). For $q \geqslant 3$, this group is isomorphic to the group C_{n-1}^q of isotopy classes of differentiable embeddings of S^{n-1} in S^{n-1+q}. This was suggested by the results of Levine [13] and proved in [5]. The following considerations generalize this theorem and help to explain it.

A deformation of φ: $V \to X$ in an embedding is a homotopy f_t: $V \to X$ such that $f_0 = \varphi$ and f_1 is a differentiable embedding. Two such deformations f_t^0 and f_t^1 are isotopic if they are connected by a two parameters family $f_t^{t'}$ such that $(t, t') \in I \times I$, $f_0^{t'} = \varphi$ and $f_1^{t'}$ is an isotopy of embeddings.

1.7. Theorem B. *Let φ: $V \to X$ be a $(n+q)$-thickening of a closed 1-connected differentiable n-manifold V. Let ϕ: $V \to BD_q$ be the associated map* (cf. 1.5). *For $q \geqslant 3$, the isotopy classes of deformations of φ in embeddings of V in X are in bijective correspondence with the homotopy classes of liftings of ϕ in BO_q.*

For $n+q > 6$, this is again a direct application of [1], [16], [21], [32]. For $n \leqslant 2$, the theorem is true by general position or by the Whitney process (cf. [25]). We also apply a theorem of Hudson [9] to replace pseudoisotopy by isotopy. Following a suggestion of Wall, we can apply the theorem of induced thickening [22] and get the following.

1.8. Corollary B. *Let φ: $V \to X$ be a map of a 1-connected differentiable n-manifold in a differentiable $(n+q)$-manifold X. If φ is $(n-q+2)$-connected, then the isotopy classes of deformations of φ in differentiable embeddings of V in X are in 1-1 correspondence with the homotopy classes of sections of a fiber space over V with fiber C_q.*

1.9. It follows from results of James [10] that

$$\pi_n(C_q) = 0 \quad \text{for} \quad n < 2q-3$$

so that in that range there is no obstruction to realize a homotopy equivalence by an embedding. However $\pi_{2q-3}(C^q)$ is $\mathbf{Z}$ or $\mathbf{Z}_2$ according to the case q odd or q even and the first obstruction to constructing a section of the universal fiber space $BO_q \to BD_q$ is a non zero element of $H^{2q-2}(BD_q)$, cf. [5].

2. Piecewise linear and homotopy tubes

2.1. In the category of piecewise linear (pl) manifolds, the notion of regular neighbourhood (or the notion of "block bundle" due to Morlet [14] and Rourke-Sanderson [17]) corresponds to the notion of differentiable tubular neighbourhood. If V is a piecewise linear n-manifold, then there is a bijective correspondence between isomorphism classes of regular neighbourhoods of V of dimension $n+q$ and homotopy classes of maps of V in the classifying space $\mathbf{BPl}_q$ of a simplicial group $\mathbf{PL}_q$ (this result is due independently to Morlet [14], Rourke-Sanderson [17] and the author (unpublished)).

The stable suspension of $\mathbf{Pl}_q$ is denoted by $\mathbf{Pl}$ and one has again a commutative diagramm

$$\begin{array}{ccc} \mathbf{BPl}_q & \to & BG_q \\ \downarrow & & \downarrow \\ \mathbf{BPl} & \to & BG \end{array}$$

All the considerations of § 1 are still valid in that case. One has just to replace everywhere differentiable by piecewise linear and O_q by $\mathbf{Pl}_q$.

In particular, the analogue of Theorem B (1.7) shows that for $q \geqslant 3$ the homotopy group $\pi_{n+1}(G, \mathbf{Pl}, G_q, \mathbf{Pl}_q)$ is isomorphic to the group of isotopy classes of deformations of the inclusion $S^n \to S^n \times D^q$ in pl-embeddings of S^n in S^{n+q}. But this group is trivial by the unknotting theorem of Zeeman [26]. Hence we get the

2.2. Theorem. *For $q \geqslant 3$, $\pi_{n+1}(G, \mathbf{Pl}, G_q, \mathbf{Pl}_q) = 0$. This is of course equivalent to each of the three properties:*

a) *$\pi_n(\mathbf{Pl}, \mathbf{Pl}_q) \to \pi_n(G, G_q)$ is an isomorphism.*

b) $\pi_n(G_q, \mathbf{PI}_q) \to \pi_n(G, \mathbf{PI})$ *is an isomorphism.*

c) *the natural map of* $\mathbf{BPI}_q$ *on the fiber product* $\mathbf{BPI} \times_{BG} BG_q$ *is a homotopy equivalence.*

On the other hand, one has

2.3. $$\pi_n(G, \mathbf{PI}) = \begin{cases} 0 & n \text{ odd} \\ \mathbf{Z} & n = 4k \\ \mathbf{Z}_2 & n = 4k+2 \end{cases}$$

This result is mainly due to Kervaire-Milnor [11] (see also Levine [13] and Sullivan [21]).

For $q = 1, 2$, it follows from Levine [12] and Wall [24] that $\mathbf{BPI}_q \approx \approx BO_q$.

In view of 2.2, the analogue of Theorem A is the

2.4. Theorem A'. *For* $n+q \geqslant 6$, *and* $q \geqslant 3$ *the isomorphism classes of piecewise linear* $(n+q)$*-thickenings of a pl, 1-connected* n*-manifold* V *are in 1-1 correspondence with the set of homotopy classes* $[V, \mathbf{BPI}_q]$.

Hence any $(n+q)$-thickening of V is a regular neighbourhood of V.

Note that $\pi_n(\mathbf{PI}_q)$ for $q \geqslant 3$ is isomorphic to the group of isotopy classes of *pl*-embeddings of $S^n \times D^q$ in S^{n+q}.

Using 2.2, the analogue of Theorem B is the

2.5. Theorem B'. *Let* $\varphi\colon V \to X$ *be a thickening, where* V *is a 1-connected closed pl,* n*-manifold and* X *a pl,* $(n+g)$*-manifold. For* $q \geqslant 3$, *then*

a) φ *is homotopic to an embedding;*

b) *two such embeddings are isotopic.*

Again using the theorem of induced thickening of Wall [22], or the equivalent theorem of Stallings [19], for a) one can assume φ to be $(n-q+1)$-connected and for b) that φ is $(n-q+2)$-connected.

It also follows immediately from Theorem B that the set of isotopy classes of embeddings of a closed *pl*, n-manifold V (1-connected) in a $(n+q)$-manifold X depends only on the homotopy type of V. This also follows from the embedding theorem of Browder [2].

These results have been also obtained independently by Sullivan and Casson, even without assuming that V is 1-connected.

Note that all these results can be generalized to the case of manifolds with boundary.

3. Differentiable and piecewise linear tubes

3.1. We also have a commutative diagramm of classifying spaces:

$$\begin{array}{ccc} BO_q & \to & \mathbf{BPI}_q \\ \downarrow & & \downarrow \\ BO & \to & \mathbf{BPI} \end{array}$$

If V is a *pl*-manifold of dimension n, a smooth regular neighbourhood N of V is a regular neighbourhood of V with a differentiable structure on N compatible with its *pl*-structure. This is the notion corresponding to thickenings.

We can make the same constructions as in § 1 and prove analogous theorems, using the smoothing theorems of Munkres [15] and Hirsch [8]. The fiber product $S_q = \mathbf{BPI}_q \times_{\mathrm{BPL}} BO$ is the universal space for smooth regular neighbourhoods of codimension q. Namely, Theorem A and its Corollary (1.4, 1.5) are still valid, without any restriction on n and q, if one replaces "Poincaré complex" by "*pl*-manifold" and "thickening" by "smooth regular neighbourhood".

Note that the natural fiber map of the fiber space $BO_q \to S_q$ in the fiber space $BO_q \to BD_q$ is a homotopy equivalence for $q \geqslant 3$ by 2.2. Hence for $q \geqslant 3$ and $n+q \geqslant 6$, smooth regular neighbourhoods of a *pl*-manifold V of dimension n correspond to smooth thickenings of V.

3.2. The analogue of Corollary 1.8 of Theorem B is the following. Let V be a compact piecewise linear n-manifold with a compatible differentiable structure. Let φ be a piecewise differentiable embedding of V in a differentiable $(n+q)$-manifold X (all embeddings are locally flat). A *smoothing* of φ is a piecewise differentiable isotopy connecting φ to a differentiable embedding of V in X.

The embedding φ defines up to isomorphism a smooth regular neighbourhood classified by a map $\phi\colon V \to S_q$.

Theorem B". *The piecewise differentiable embedding* $\varphi\colon V \to X$ *can be smoothed if and only if* ϕ *has a lifting in* BO_q.

As before, this leads to obstructions in the group $\pi_{n+1}(\mathbf{PI}, O, \mathbf{PI}_q, O_q)$ which can be interpreted as the group of concordance classes of differentiable embeddings of S^n in S^{n+q} which are piecewise differentiably isotopic to the inclusion. By 2.2, for $q \geqslant 3$, this group is isomorphic to $\pi_{n+1}(G, O, G_q, O_q)$.

In Theorem B", the homotopy classes of liftings of ϕ correspond to the concordance classes of smoothings of φ.

4. Classification of immersions

4.1. An *immersion* f of a differentiable (resp. piecewise linear) manifold V in a differentiable (resp. *pl*) manifold X is a map which can be expressed in local charts as a linear injective map.

If V is a *pl*-manifold and X a smooth manifold, a map f of V in X is a *smooth immersion* if f is a differentiable immersion with respect to a smooth structure on V compatible with its *pl*-structure.

Two differentiable (resp. *pl* or smooth) immersions $f_0, f_1\colon V \to X$ are *concordant*, if there is a differentiable (resp. *pl* or smooth) immer-

sion F: $V \times [0, 1] \to X \times [0, 1]$ such that $F(x, i) = (f_i(x), i)$, $i = 0, 1$.

To each map f: $V \to X$ is associated a stable normal "bundle", namely the Whitney sum of the stable tangent bundle to V and the inverse image by f of the stable normal bundle to X. This bundle is classified by a map φ: $V \to BO$ if V and X are smooth manifolds, and by a map φ: $V \to \mathbf{BPl}$ if V and X are piecewise linear manifolds.

4.2. Theorem. *Let V and X be manifolds, f: $V \to X$ be a continuous map. Let $q = \dim X - \dim V > 0$.*

If V and X are differentiable (resp. piecewise linear), then f is homotopic to an immersion if and only if the classifying map φ: $V \to BO$ (resp. $V \to \mathbf{BPl}$) for the normal bundle of f can be lifted in BO_q (resp. $\mathbf{BPL}_q$).

If V is piecewise-linear and X differentiable, then f is homotopic to a smooth immersion iff φ: $V \to \mathbf{BPL}$ can be lifted in BO_q.

The concordance classes of deformations of f in immersions are in 1-1 correspondence with the homotopy classes of liftings of φ.

In the differentiable case, this is the classical theorem of Hirsch-Smale (cf. [7]). For the *pl* and smooth case, see [4] and [6].

4.3. In the piecewise linear case, one has $\pi_n(\mathbf{PL}, \mathbf{Pl}_q) = \pi_n(G, G_q)$ for $q \geqslant 3$. On the other hand, let ψ: $V \to BG$ be a classifying map for the stable normal spherical fiber space of f (i.e. the "sum" of the stable tangent spherical fiber space to V with the normal one to X). The homotopy class of ψ depends only on the homotopy class of f, and the homotopy type of $(V, \partial V)$ and $(X, \partial X)$. (See Spivak [18]).

Corollary. *Let $q = \dim X - \dim V \geqslant 3$. A map f: $V \to X$ is homotopic to a piecewise linear immersion iff ψ: $V \to BG$ can be lifted in BG_q.*

The set of concordance classes of pl-immersions of V in X depends only on the homotopy type of $(V, \partial V)$ and $(X, \partial X)$.

Université de Genève,
Institut de Mathématiques,
Suisse

REFERENCES

[1] Browder W., Homotopy type of differentiable manifolds, Colloquium on algebraic topology, Aarhus (1962), 42-46.
[2] Browder W., Embedding 1-connected manifolds, *Bull. AMS*, **72** (1966), 225-231.
[3] Browder W., Hirsch M., Surgery on *Pl*-manifolds and applications (to appear).
[4] Haefliger A., Poenaru V., La classification des immersions combinatoires, Publ. Math. I.H.E.S., **23** (1964), 75-91.
[5] Haefliger A., Differentiable embeddings of S^n in S^{n+q} for $q > 2$, *Ann. of Math.*, **83** (1966), 402-436.

[6] Haefliger A., Lissage des immersions, I et II, Notes mimeographiées, Univ. de Geneve (1966).
[7] Hirsch M. W., Immersions of manifolds, *Trans. A.M.S.*, **93** (1959), 242-276.
[8] Hirsch M. W., Obstruction theories for smoothing manifolds and maps, *Bull. A.M.S.*, **69** (1963), 352-356.
[9] Hudson J. P., Concordance and isotopy of *Pl* embeddings, *Bull. A.M.S.*, **72** (1966), 534-535.
[10] James I. M., On the suspension sequence, *Ann. of Math.*, **65** (1957), 74-107.
[11] Kervaire M., Milnor J., Groups of homotopy spheres, I, *Ann. of Math.*, **77** (1963), 504-537.
[12] Levine J., Unknotting spheres in codimension 2, *Topology*, **4** (1965), 9-16.
[13] Levine J., A classification of differentiable knots, *Ann. of Math.*, **82** (1965), 15-50.
[14] Morlet C., Les voisinages tubulaires des variétés semi-linéaires, *C.R. Acad. Sc. Paris*, **262** (1966), 740-743.
[15] Munkres J., Obstructions to the smoothing of piecewise differentiable homeomorphisms, *Ann. of Math.*, **72** (1960), 521-524.
[16] Новиков С. П., О диффеоморфизме односвязных многообразий, *ДАН СССР*, **143**, № 5 (1962), 1046-1049. English translation: Diffeomorphisms of simply connected manifolds, *Soviet Math. Dokl.*, **3** (1962), 540-543.
[17] Rourke C., Sanderson B., Block Bundles, I, to appear.
[18] Spivak M., On spaces satisfying Poincaré duality, Thesis, Princeton University, 1964.
[19] Stallings J., The embedding of homotopy types into manifolds (to appear).
[20] Stasheff J., A classification theorem for fiber spaces, *Topology*, **2** (1963), 239-246.
[21] Sullivan D., Ph. D. Thesis, Princeton University (1966), (Mimeographed) Rice University.
[22] Wall C. T. C., Classification problems in differential topology — IV, Thickenings, *Topology*, **5** (1966), 73-94.
[23] Wall C. T. C., An extension of results of Novikov and Browder, *Amer. J. Math.*, **88** (1966), 20-32.
[24] Wall C. T. C., Locally flat *Pl*-submanifolds with codimension two, (to appear).
[25] Whitney H., The self-intersections of a smooth n-manifold in $2n$-space, *Ann. of Math.*, **45** (1944), 247, 293.
[26] Zeeman E. C., Unknotting combinatorial balls, *Ann. of Math.*, **78** (1963), 501-526.

ON SPACES CO-ABSOLUTE WITH METRIC SPACES

V. I. PONOMAREV

1. Necessary definitions

A one-valued continuous mapping f: $X \to Y$ is called perfect if it is closed and if for every point $y \in Y$ the set $f^{-1}y$ is bicompact. A one-valued continuous mapping f: $X \to Y$ is said to be irreducible if for

every closed proper subset $X_0 \subset X$ we have $fX_0 \neq Y$. In [15], [16] it is proved that a continuous one-valued mapping $f: X \to Y$ is irreducible and closed if and only if for every nonempty open set $U \subseteq X$ the set $f^{\#}U = \mathscr{E}(y, f^{-1}y \subseteq U)$ is nonempty and open in Y.

We shall also consider many-valued mappings. The many-valued mapping $f: X \to Y$ maps each point $x \in X$ onto a closed set fx of the space Y. The many-valued mapping $f: X \to Y$ is said to be continuous (see [12]) if for every $x \in X$ and every neighbourhood Ufx of the set $fx \subseteq Y$ there exists a neighbourhood Ux of the point x such that $fUx \subseteq Ufx$. For a given continuous many-valued mapping $f: X \to Y$, the inverse mapping $f^{-1}: Y \to X$ is defined as follows:

$$f^{-1}y = \mathscr{E}(x, fx \ni y).$$

A continuous many-valued mapping $f: X \to Y$ is said to be perfect if the following two conditions are satisfied: (a) for an arbitrary point $x \in X$ the set fx is bicompact and for an arbitrary point $y \in Y$ the set $f^{-1}y$ is bicompact, (b) the mapping $f^{-1}: Y \to X$ is also continuous, i.e. the mapping $f: X \to Y$ is closed.

In [13] the following theorem is proved: a necessary and sufficient condition for the continuous many-valued mapping $f: X \to Y$ to be perfect is that there exists a space Z and one-valued perfect mappings $p_x: Z \to X$, $p_y: Z \to Y$, such that $f = p_y p_x^{-1}$.

A perfect many-valued mapping $f: X \to Y$ will be said to be irreducible if there exists a space Z and one-valued perfect irreducible mappings $p_x: Z \to X$, $p_y: Z \to Y$, such that $f = p_y p_x^{-1}$.

Throughout the whole of our discussion, we shall be making use of the p-spaces introduced by A. V. Arhangel'skiĭ [1], [2]. Paracompact p-spaces are precisely those spaces that admit perfect mappings onto metric spaces. Every paracompact space which is complete in the sense of Čech is necessarily a paracompact p-space.

2. The concept of the absolute of a regular topological space [11], [15], [16], [17], [9], [8], [7]

The space X' will be called an irreducible preimage of the regular space X if there exists a one-valued perfect irreducible mapping $f: X' \to X$. The space $\dot{X}$ will be called the absolute of the regular space X if X is an irreducible preimage of $\dot{X}$ and if every irreducible preimage X' of X is homeomorphic to the space $\dot{X}$. The following facts are well known.

Every regular space X has a unique absolute $\dot{X}$, which is an extremally disconnected space. There exists a uniquely defined "canonical" perfect irreducible one-valued mapping $\pi_x: \dot{X} \to X$. If the two spaces X and Y are connected by a one-valued perfect irreducible mapping $f: X \to Y$, their absolutes $\dot{X}$ and $\dot{Y}$ are homeomorphic and there exists a homeomorphism $f: X \to Y$ between them such that $f = \pi_y \dot{f} \pi_x^{-1}$. Conversely, if the spaces X and Y have homeomorphic absolutes $\dot{X}$ and $\dot{Y}$, then every homeomorphism $h: \dot{X} \to \dot{Y}$ generates a perfect irreducible (in general, many-valued) mapping

$$f = \pi_y h \pi_x^{-1}.$$

Every extremally disconnected space X is completely regular and is its own absolute.

A space X is said to be extremally disconnected if the closure of every open set in it is open.

3. Statement of the problem

Two topological spaces will be said to be co-absolute if their absolutes are homeomorphic to each other, or in other words if they are related to each other by a perfect irreducible (generally speaking, many-valued) mapping. In this article, we attempt to give a systematic discussion of spaces co-absolute with metric spaces.

4. Fundamental definition and fundamental theorems

The fundamental definition [19] runs as follows: A system Σ of open sets of a space X is said to be dense in this space if every open $\Gamma \subseteq X$ contains a $U \in \Sigma$. The smallest cardinal number of a system dense in X is called the π-weight of the space X.

Fundamental Theorem 1 (see [20]). In order that the space X may admit a one-valued perfect irreducible mapping onto a metric space, it is necessary and sufficient that X be a paracompact p-space and that X satisfy one of the two following conditions:

1. In X there exists a σ-locally finite, everywhere dense system of open sets.
2. In X there exists a dense system of open sets which is the union of a countable number of disjoint and locally finite systems.

Fundamental Theorem 2. If the p-space X admits a many-valued perfect irreducible mapping onto a metrizable space, then it also admits a one-valued perfect irreducible mapping onto some (generally speaking, different) metrizable space.

Theorem 3. For the space X to admit a one-valued perfect irreducible mapping onto a metrizable space with countable base, it is necessary and sufficient that X be a finally-compact p-space with countable π-weight. If the p-space X admits a many-valued perfect

irreducible mapping onto some metrizable space with countable base, then the π-weight of X is countable and X admits a one-valued perfect irreducible mapping onto some (generally speaking, different) metric space with countable base.

Theorem 4. In order that the bicompactum X may admit a one-valued irreducible mapping onto some compactum, it is necessary and sufficient that X be of countable π-weight. If the space X admits a many-valued irreducible mapping onto some compactum, then X also admits a many-valued irreducible mapping onto some (generally speaking, different) compactum.

From Theorem 4 follows at once the answer to a problem of P. S. Aleksandrov:

Theorem 5. In order that a completely normal bicompactum X be co-absolute with a compactum (and thus even admit a one-valued irreducible mapping onto some compactum), it is necessary and sufficient that X be separable.

In connection with Theorem 5 let us note that no example is known of a completely normal nonseparable bicompactum.

Theorem 6. For the bicompactum X to be co-absolute with the Cantor perfect set, it is necessary and sufficient that X be of countable π-weight and the X have no isolated points.

Theorem 7. For the space X to admit a one-valued perfect irreducible mapping onto a complete metric space, it is necessary and sufficient that X be paracompact and complete in a sense of Čech and that X satisfy one of the conditions 1, 2 of Theorem 1. If X admits a many-valued perfect irreducible mapping onto some complete metric space, then X is paracompact and complete in the sense of Čech and admits a one-valued perfect irreducible mapping onto some (generally speaking, different) complete metric space.

Theorem 8. Every bicompactum X of π-weight τ admits a one-valued irreducible mapping onto some bicompactum τ. If the bicompactum X admits a many-valued irreducible mapping onto a bicompactum of π-weight τ, then the π-weight of the bicompactum X is equal to τ and there exists a one-valued irreducible mapping of X onto some bicompactum of weight τ.

Theorem 9. The π-weight of a dyadic bicompactum coincides with its weight.

Theorem 10. In the dyadic bicompactum X let there exist an everywhere dense subspace X_0 which is co-absolute with some metric space Y_0. Then X and Y_0 (and of course, X_0) have countable bases, and are therefore metrizable.

In connection with the theory of dyadic spaces I should like to mention the relatively recent publications [3], [4], [5].

5. A problem of G. Birkhoff

Let B be a complete Boolean lattice; by $K(B)$ let us denote the class of all Hausdorff spaces X such that the Boolean lattices $R_0(X)$ of canonical open sets of X are isomorphic to B. In his well-known book [24] G. Birkhoff stated the following problem (problem 82): to find a necessary and sufficient condition for a complete Boolean lattice to be represented in the form $R_0(X)$, where X is a suitable metric space. In [6] J. Flachsmeyer proved the following theorem: a Hausdorff space X belongs to $K(B)$ if and only if it is the image of a dense subspace of the space $M(B)$ (see [22]) of the lattice B under a mapping that is closed, irreducible, and θ-continuous.

Thus the problem of G. Birkhoff has the following topological analogue: find a necessary condition that must be satisfied by an extremally disconnected bicompactum X in order that the class of all continuous irreducible closed images of dense subspaces shall contain at least one metric space.

The following theorem gives the answer to this question:

Theorem 11. In order that in a given extremally disconnected bicompactum X there shall exist a dense subspace X_0 admitting a closed irreducible continuous mapping onto some metric space, it is necessary and sufficient that in the extremally disconnected bicompactum X there exist a dense σ-disjoint system of open sets.

On the other hand, if the extremally disconnected bicompactum $\dot{X}$ is the absolute of a dyadic nonmetrizable bicompactum X, then $\dot{X}$ does not contain an everywhere dense subspace admitting a closed irreducible mapping onto a metric space.

6. Borel sets

The following theorem gives an affirmative answer to a well known problem of P. S. Aleksandrov.

Theorem 12 (see [21]). Let X be an uncountable completely normal bicompactum. Then every class of the Borel classification of sets is nonempty.

The Moscow University,
Moscow, USSR

REFERENCES

[1] Архангельский А. В., *ДАН СССР*, **151** (1963), 751-754; English translation: Arhangel'skiĭ A. V., *Soviet Math. Dokl.*, **4** (1963), 1051-1055.
[2] Архангельский А. В., *Матем. сб.*, **67** (109), (1965), 55-88.
[3] Engelking R., Pelczyński A., *Colloq. Math.*, **11** (1963), 55-63.
[4] Engelking R., Efimov B., *Colloq. Math.*, **13** (1965), 181-197.
[5] Ефимов Б. А., *ДАН СССР*, **151** (1963), 1021-1024; Eglish translation: *Soviet Math. Dokl.*, **4** (1963), 1131-1134.
[6] Flachsmeyer J., *ДАН СССР*, **156** (1964), 32-34. English translation: *Soviet Math. Dokl.*, **5** (1964), 607-610.
[7] Flachsmeyer J., *Math. Nachr.*, **26** (1963), 1-4, 57-66.
[8] Илиадис С., Фомин С. В., *УМН*, **21**, № 4 (1966), 47-76.
[9] Илиадис С., *ДАН-СССР*, **149** (1963), 22-25. English translation: *Soviet Math., Dokl.*, **4** (1963), 295-298.
[10] Hewitt E., *Ann. of Math.*, (2) **47** (1946), 503-509.
[11] Gleason A. M., *Illinois Math. J.*, **2** (1958), 482-489.
[12] Пономарев В. И., *Матем. сб.*, 48 (90), (1959), 211-232.
[13] Пономарев В. И., *Матем. сб.*, **51** (93), (1960), 515-537. English transl.: Amer. Math. Soc. Transl. (2), **38** (1964), 119-140.
[14] Пономарев В. И., *Матем. сб.*, **52** (94), (1960), 847-862. English translation.: Amer. Math. Soc. Transl. (2), **38** (1964), 141-158.
[15] Пономарев В. И., *ДАН СССР*, **143** (1962), 46-49. English translation: *Soviet Math. Dokl.*, **3** (1962), 347-350.
[16] Пономарев В. И., *Матем. сб.*, 60 (102), (1963), 89-119. English translation: Amer. Math. Soc. Transl. (2), **39** (1964), 133-164.
[17] Пономарев В. И., *ДАН СССР*, **149** (1963), 26-29. English translation: *Soviet Math. Dokl.*, 4 (1963), 299-302.
[18] Пономарев В. И., *ДАН СССР*, **153** (1963), 1013-1016. English translation: *Soviet Math. Dokl.*, 4 (1963), 1777-1780.
[19] Пономарев В. И., *ДАН СССР*, 166 (1966), 291-294. English translation: *Soviet Math. Dokl.*, 7 (1966), 76-79.
[20] Пономарев В. И., *УМН*, **21**, № 4 (1966).
[21] Пономарев В. И., *ДАН СССР*, **170** (1966), 520-526.
[22] Stone M. H., *Trans. Amer. Math. Soc.*, 41 (1937), 375-481.
[23] Урысон П. С., т. I, М., изд-во АН СССР, 182-183.
[24] Birkhoff G., Lattice theory, Amer. Math. Colloq. Publ., vol **25**, 1940; rev. ed. 1948; reprint 1964. Русский перевод: Биркгоф Г., Теория структур, ИЛ., М., 1952.

HOMEOMORPHISM AND DIFFEOMORPHISM CLASSIFICATION OF MANIFOLDS

C. T. C. WALL

I want to compare the classification problems for four different kinds of manifolds.

(i) *Smooth* (or differential) *manifolds*. These are comparatively familiar mathematical objects: they are Hausdorff topological spaces with additional structure which can be specified by an atlas of local coordinate systems (homeomorphisms of open sets in the manifold onto open sets in Euclidean space) in which transformations between different coordinate systems are always given by smooth (i. e. C^∞) functions. Here we will only consider compact manifolds.

(ii) *PL* (or piecewise linear, or combinatorial) *manifolds*. These are defined analogously, with the modification that coordinate transformations must now be piecewise linear—i.e. linear on each simplex for some locally finite decomposition in simplices of the open set in question.

(iii) *Topological manifolds*: here no restriction is made on the coordinate transformations beyond their being homeomorphisms.

(iv) It is also convenient to have a homotopy-theoretic analogue of the above. The clue here is provided by the Poincaré duality theorem, which holds for all manifolds. We define a *Poincaré complex* to be a *CW* complex which satisfies a suitably strong form of the Poincaré duality theorem (the detailed definition is somewhat technical; see Wall [1] or [2]).

A manifold of any of these types determines one of each subsequent type, in an essentially unique manner:

(i) → (ii) by smooth triangulation (due to Whitehead [1], see also Munkres [1]),

(ii) → (iii) by just ignoring the *PL*-structure,

(iii) → (iv) by ignoring all but the homotopy type of our manifold

(this step ignores the local nice properties of a manifold to concentrate on the global structure). The problem I want to discuss is that of going in the opposite direction — i. e. imposing stronger structures. First, we must construct invariants of the various types of structure.

(i). As is well-known from differential geometry, a smooth manifold has tangent vectors, which are assembled in a vector bundle over M, the *tangent bundle*. We can describe this by saying that the vectors which form the fibre over a point $P \in M$ correspond diffeomorphically (by the exponential map) with a neighbourhood of P in M. This bundle has structure group the orthogonal group, O_m, and hence is classified by a (homotopy class of) maps from M to the classifying space, BO_m (see e. g. Milnor [1]).

(ii) and (iii). At the last international congress, J. Milnor introduced the theory of microbundles, which gives analogous results in cases (ii) and (iii) (Milnor [4], see also [3], [5]). There are several refinements and variants of his original definition (see Kister [1], Mazur [2], [3], Hirsch [1] and Kuiper and Lashof [1]): I choose the simplest, a bundle with fibre a Euclidean space of dimension m, and structure group known as PL_m or Top_m in the two cases: to be thought of as a group of homeomorphisms leaving the origin fixed. (The former has to be defined as a semi-simplicial group.) It is shown in the papers cited that

tangent bundles exist and are unique. Milnor also obtained corresponding spaces BPL_m, $BTop_m$, maps into which classify such bundles.

(iv). It has been shown recently by Spivak [1] that an analogous theory exists for this case also. The appropriate objects turn out to be fibre spaces (not bundles)

$$\begin{array}{ccc}\Sigma & \longrightarrow & E \\ & & \downarrow \pi \\ & & M\end{array}$$

in which Σ has the homotopy type of a sphere of some (usually large) dimension. Such fibrations are called spherical. The stable normal fibration of a Poincaré complex M is characterised by the requirement that its Thom space (the mapping cone of π) be reducible. It turns out also that given two such fibrations, and supposing (as we may, by suspension) that the fibres are homotopy equivalent to spheres of the same dimension $k-1$, and that $f_i : S^{m+k} \to M \cup_{\pi_i} CE_i$ $(i = 1, 2)$ have degree 1, then there exists a fibre homotopy equivalence (unique up to fibre homotopy) of π_1 on π_2 which carries f_1 to f_2. This result will be important below: it reduces complicated problems about homotopy groups of Thom spaces, which arose in the pioneering work of Novikov [1], to much simpler questions concerning equivalence of fibrations.

Instead of a structure group for spherical fibrations, one has a structure monoid G_k, the space of homotopy equivalences of S^{k-1} on itself: it too possesses a classifying space BG_k (Dold and Lashof [1], Stasheff [1].)

The next important remark is as follows: there exist maps

$$BO_m \longrightarrow BPL_m \longrightarrow BTop_m \longrightarrow BG_m$$

corresponding to natural transformations of bundle functors. Thus an m-vector bundle over a simplicial complex X is classified by a map $X \to BO_m$; we form the composite $X \to BO_m \to BPL_m$, and this induces a PL-bundle over X, which "triangulates" the vector bundle (see Lashof and Rothenberg [1], also Hirsch and Mazur [1]). Similarly $BPL_m \to BTop_m$ corresponds to forgetting the PL-structure: this is essentially due to Milnor [3, 4]. Finally given a bundle with fibre $\mathbf{R}^m$, by removing the section corresponding to 0 we change the fibre to $\mathbf{R}^n - 0$, with the homotopy type of S^{m-1}, and thus obtain a spherical fibration.

We have described these transformations in geometrical terms: it is now not difficult to see that if we take a smooth triangulation of a smooth manifold M, its tangent PL-bundle is given by triangulating the tangent vector bundle of M (proof in Lashof and Rothenberg [1]). Even more clearly, the tangent bundle of a PL-manifold is unchanged (as a bundle) by regarding the manifold as a topological manifold. The final step to BG_m is more complicated as we have only defined a normal bundle for that situation: it is an open problem whether one can characterise a tangent spherical fibration ξ for a Poincaré complex X, e. g. by requiring a map of degree 1 from $X \times X$ to the Thom space of ξ to satisfy some natural extra conditions. (With some restrictions on X, this has been solved by W. Sutherland.)

For this and other reasons, we now stabilise. On increasing m by 1 we obtain a commutative diagram

$$\begin{array}{ccccccc} BO_m & \longrightarrow & BPL_m & \longrightarrow & BTop_m & \longrightarrow & BG_m \\ \downarrow & & \downarrow & & \downarrow & & \downarrow \\ BO_{m+1} & \longrightarrow & BPL_{m+1} & \longrightarrow & BTop_{m+1} & \longrightarrow & BG_{m+1}; \end{array}$$

we denote the direct limit as $m \to \infty$ by

$$BO \longrightarrow BPL \longrightarrow BTop \longrightarrow BG.$$

Then the tangent bundle of a manifold M induces a map $M \to BTop \to BG$ which is homotopic to that induced by an inverse to the stable normal fibration of M regarded as Poincaré complex. This follows from Spivak's work, and is also closely related to the earlier paper of Milnor and Spanier [1].

The following may be regarded as the fundamental question in the subject.

P r o b l e m. Suppose given a manifold of one of our four types, and a reduction of the structural group of its stable tangent bundle to an earlier type: does the manifold similarly admit extra structure? And is this extra structure unique up to some natural equivalence relation? We can reformulate this using the classifying spaces above. For example: suppose M^m to be a topological manifold, with stable tangent bundle classified by $\tau : M \to BTop$. Suppose given a map $f : M \to BO$, and a homotopy of the composite map $M \xrightarrow{f} BO \to BTop$ to τ. Then can M be given a corresponding differentiable structure? Is this unique up to equivalence? (Note the explicit homotopy: this is an important part of the data.)

Before we go on to consider answers to this problem, I will mention some generalisations which can be treated by similar methods, but detailed consideration of which is outside the scope of this talk. First, we may consider manifolds M with boundary ∂M (M still compact): for case (iv) we insist that ∂M be a Poincaré complex, and that the CW pair $(M, \partial M)$ satisfy the Lefschetz duality theorem. We do not include the noncompact case, which can also be formulated, but probably not yet in the right form, and seems (except for the transition (i)→(ii)) to be apreciably harder. Next, we might consider submanifolds of a fixed larger manifold: e.g. the problem of smoothing PL-submanifolds of a smooth manifold. The normal bundle plays a dominant role here. For a discussion of the case (i) → (iv) c. f. the talk

immediately preceding this one (W. Browder [2]) and for the case (i) → (ii) compare the talk of A. Haefliger [1] at this congress, as also recent work of Morlet [1] and Rourke and Sanderson [1]. Finally, one can also classify automorphisms instead of objects, and even investigate the homotopy type of groups of homeomorphisms or spaces of embeddings. This problem is complicated by technical considerations of concordance and isotopy: see the talk of Cerf [1] at this congress, also a recent paper of Hudson [1].

We return to our own problem: the answer depends, not surprisingly, on the case investigated. The result is simplest for the case (i) → (ii): the problem of smoothing PL-manifolds. This was discussed in M. W. Hirsch's talk [3] (following Hirsch and Mazur [1]): the answer to our problem is yes. Furthermore, equivalence classes of smoothings of the PL-manifold M correspond bijectively to homotopy classes of homotopy factorisations

$$\begin{array}{ccc} & M & \\ & \swarrow \quad \searrow & \\ BO & \longrightarrow & BPL. \end{array}$$

We next consider the case (ii) → (iv): the problem here is to characterise the homotopy types of PL-manifolds. It is tackled as follows. We are given a Poincaré complex X, and a homotopy factorisation of the classifying map of its Spivak fibration

$$\begin{array}{ccc} & X & \\ & \swarrow \quad \searrow & \\ BPL & \longrightarrow & BG; \end{array}$$

or equivalently, a PL-bundle ν over X, and a fibre homotopy equivalence with the Spivak fibration; or equivalently again, the PL-bundle ν, and a homotopy class of maps of degree 1 from a sphere to its Thom space. A transversality argument due essentially to Williamson [1] and Browder [1] now shows that we can find at least a closed PL-manifold M, and a map $\varphi: M \to X$ of degree 1 such that $\varphi^*\nu$ is the stable normal bundle of M. Indeed, our data determine a bordism class of such maps (M, φ): we seek an (M_0, φ_0) in this class with φ_0 a homotopy equivalence, and also want to know about uniqueness of M_0 when it exists. The problem is now amenable to the method of surgery (initiated by Milnor [2] and first applied to this situation by Novikov [1]).

For technical convenience, it is more convenient to replace (iv) by a new class (iv)′ of finite Poincaré complexes (the definition, even more technical, involves Whitehead's theory of simple homotopy types, for which see Whitehead [2]). This is related to the other classes of "manifolds" by functors

$$\begin{array}{ccccc} \text{(i)} & \longrightarrow & \text{(ii)} & \longrightarrow & \text{(iii)} \\ & & \downarrow & & \downarrow \\ & & \text{(iv)}' & \longrightarrow & \text{(iv)}. \end{array}$$

(An outstanding problem is whether one can define a transformation of structures in the direction (iii) → (iv)′; this contains the problem whether a compact manifold has the homotopy type of a finite complex, also that of topological invariance of simple homotopy type.) The map (iv)′ → (iv) is related to the projective class group and the Whitehead group of the fundamental group $\pi_1(X)$ of the Poincaré complex X. We define w to be the homomorphism $w: \pi_1(X) \to \{\pm 1\}$ which takes the value -1 on orientation-reversing loops. Then surgery leads to the following result.

Theorem. *There exist functors L_m depending only on the integer m modulo 4, and defining abelian groups $L_m(\pi_1(X), w)$. Given a bordism class of maps (M, φ) of degree 1, as above, with $m \geqslant 5$, there is an obstruction in $L_m(\pi_1(X), w)$ to the existence of an element (M_0, φ_0) of it with φ_0 a simple homotopy equivalence, and an obstruction in $L_{m+1}(\pi_1(X), w)$ to its uniqueness, when it exists.*

We can formulate this as an exact sequence (of based sets), which can be extended in one direction

$$\ldots \longrightarrow L_{m+1}(\pi_1(X), w) \longrightarrow PL\text{-homeomorphism classes of } (M_0, \varphi_0) \longrightarrow$$
$$\longrightarrow \text{Bordism classes of } (M, \varphi) \longrightarrow L_m(\pi_1(X), w).$$

Example 1 : $X = S^m$. The transversality argument mentioned above allows an easy proof that the bordism classes form the group $\pi_m(G, PL)$. A result of Smale [1] shows that the PL-homeomorphism class is unique. The sequence (continued to the left) then provides an isomorphism

$$\pi_m(G, PL) \longrightarrow L_m(1)$$

for $m \geqslant 5$ (in fact this holds for $m \geqslant 1$ except for $m = 4$, when the image has index 2). The group $L_m(1)$ can be computed algebraically (see also Kervaire and Milnor [1], Levine [1], where it is called P_m): it is trivial for m odd, infinite cyclic for $m \equiv 0 \pmod 4$, and cyclic of order 2 for $m = 2 \pmod 4$.

Example 2 : $X = S^m \times S^1$. A similar argument can be used here, using results of Browder and Levine [1]. One finds that in the orientable case, $L_m(\mathbf{Z}) \cong [S^{m-1} \vee S^m : G/PL]$ $(m \geqslant 6)$. Similarly, using the nontrivial bundle over S^1 with fibre S^m, in the nonorientable case, $L_m(\mathbf{Z}) \cong [S^{m-1} \cup_2 e^m : G/PL]$ $(m \geqslant 6)$.

There are analogous results also for bounded manifolds, and in one important case, the corresponding result is simpler. Suppose (Y, X) a Poincaré pair, with X and Y connected, and such that inclusion induces an isomorphism of $\pi_1(X)$ on $\pi_1(Y)$. Then the group corresponding to L_m vanishes: provided $m \geqslant 6$, the corresponding M_0 exists and is unique. (So the answer to our problem is yes in this case also.)

Armed with this, we can give a complete answer in the closed, simply-connected case. Given a closed PL-manifold M, we write M^* for the manifold obtained by deleting the interior of an embedded disc D^m. One can invent a corresponding decomposition $X = X' \bigcup_f e^m$ for the case of Poincaré complexes: if $m \neq 2$, it is essentially unique (see Wall [2]). Now suppose the Spivak fibration for X' reduced to a PL-bundle:

$$\begin{array}{ccc} & X' & \\ & \swarrow \quad \searrow & \\ BPL & \longrightarrow & BG. \end{array}$$

As (X', S^{m-1}) is a Poincaré pair, and both are simply connected, we get a unique corresponding manifold M' $(m \geqslant 6)$: moreover, ∂M^* is homotopy equivalent to S^{m-1}, hence is PL-homeomorphic to it. We then attach a disc D^m along S^{m-1} (this process is unique) to give a closed PL-manifold M. Thus in the closed, simply-connected case the answer to our problem is: yes, provided we consider reductions over X' instead of ones for all of X. Our proof assumed $m \geqslant 6$: the result holds also for $m = 5$ and (trivially) for $m \leqslant 2$. In dimension 3, X is homotopy unique and S^3 proves existence: uniqueness is equivalent to the Poincaré conjecture. The statement given above is due to D. Sullivan [1].

In the non-simply-connected case, the results are more complicated. Let me cite by way of example that there are infinitely many PL-manifolds homotopy equivalent to $P_7(\mathbf{R})$, all corresponding to the same reduction from G to PL, but no two homeomorphic. This comes from the fact that $L_8(\mathbf{Z}_2) \cong \mathbf{Z} \oplus \mathbf{Z}$ (orientable case). We can describe explicitly an invariant that distinguishes our manifolds Q: note first that both they and their double covers are rational homology spheres, so if $\partial W = Q \bigcup Q'$ (W orientable) we can speak unambiguously of the signature of W. Now it is easy to show that for any of our manifolds, say Q_7, we can find $\partial W = Q_7 - P_7(\mathbf{R})$ with W oriented, of signature 0, and $\pi_1(Q) \cong \pi_1(W) \cong \pi_1(P_7(\mathbf{R}))$ by inclusion. The required invariant of Q is then the signature of the double cover of W.

Sullivan has used the methods of surgery to gain insight into the homotopy type of the quotient space G/PL. We have already seen how to calculate its homotopy groups. Now suppose given a closed, 1-connected PL-manifold M, and homotopy class of maps $f: M \to G/PL$.

Now the tangent bundle of M is classified by a map $\tau : M \to BPL$. Also, we have a principal fibration

$$\begin{array}{ccc} G/PL & \longrightarrow & BPL \\ & & \downarrow \\ & & BG: \end{array}$$

multiplication is induced by Whitney sum of bundles (see e. g. Lashof and Rothenberg [1]). Thus we can operate fibrewise on τ by f, obtaining a new map $f_*\tau : M \to BPL$ whose projection on BG is the same as that of τ. Proceeding as above, we then reach a surgery obstruction in L_m (1). This gives a retraction of the oriented PL-bordism group

$$\Omega_m^{SPL}(G/PL) \to \pi_m(G/PL).$$

In the case $m = 4k+2$ we need not suppose M 1-connected or even oriented, and can work with the unoriented bordism group. From these retractions, Sullivan deduces that all k-invariants of G/PL vanish mod 2 (though not modulo the class of finite groups of odd order: the first k-invariant already has order 2). Equivalently, there is a map

$$G/PL \to \prod_{k \geqslant 0} K(\mathbf{Z}_2,\ 4k+2)$$

inducing epimorphisms of homotopy groups. A slightly stronger result can in fact be obtained.

Work currently in progress is aimed at studying the behaviour of odd primes: it is conjectured that, modulo finite 2-groups, G/PL has the homotopy type of BO (though, of course, the natural maps $G/PL \to BPL \leftarrow BO$ do not correspond to a homotopy equivalence modulo 2-groups).

We come finally to the case (iii) of topological manifolds. Some information may be gathered here by observing that (iii) lies between (ii) and (iv) (but not—at present—(iv)'). In addition to this, we have Novikov's recent proof [2] of topological invariance of rational Pontrjagin classes. It is easy deduction from this that if M is compact, the kernel of the homomorphism

$$[M: BPL] \to [M: BTop]$$

is finite. We can sharpen the method to prove that for any n, $\pi_n(G, PL) \to \pi_n(G, Top)$ is a split monomorphism. What does this tell us about our problem? Unfortunately, it gives no method (and I know none) for constructing topological manifolds (other than PL ones), or for proving given topological manifolds homeomorphic. Newman's recent solution [1] of the Poincaré conjecture for topological manifolds is

a pointer in this direction, however. Thus we must abandon the case (iii) → (iv) of our problem.

For the case (ii) → (iii), one can obtain "stable" theorems by using the results of Milnor: see Mazur [2, 3] and Hirsch [2]. Thus, given a (not necessarily compact) topological manifold M, and a homotopy factorisation

$$\begin{array}{ccc} & M & \\ \swarrow & & \searrow \\ BPL & \longrightarrow & BTop \end{array}$$

then for some n, $M \times \mathbf{R}^n$ admits a PL-structure, and two such (both inducing the above factorisation) determine equivalent PL-structures on $M \times \mathbf{R}^{n+n'}$ for large enough n'. More interesting, however, are the recent "unstable" theorems of Sullivan and Wagoner. These concern only the uniqueness (Hauptvermutung) problem, and not the existence (Triangulation). Suppose given a homeomorphism h of compact PL-manifolds M and M'. Then we have a homotopy equivalence, and the two reductions of structural group from G to PL are equivalent in Top. Assume

(A) The structural groups are already equivalent in PL. Then we can apply the method of surgery to attempt to prove the manifolds PL-homeomorphic (by constructing an s-cobordism, and applying the s-cobordism theorem (Mazur [1], Kervaire [1])). We need to assume that h is a simple homotopy equivalence $M \to M'$; also (in the bounded case) $\partial M \to \partial M'$. If also

(B) Surgery can be performed, it follows that M and M' are PL-homeomorphic, as desired.

It follows from the discussion above that (B) is all right in the closed, simply-connected case, or if $\pi_1(\partial M) \cong \pi_1(M)$ by inclusion, or in certain other cases (e.g. closed, nonorientable, odd-dimensional, with fundamental group of order 2). This assumes that the dimension of the manifold M (and, if ∂M is nonempty, of ∂M) exceeds 4: if $\dim M \leqslant 3$ the Hauptvermutung and triangulation have also been proved, without further assumptions, by Moise [1] and Bing [1].

For (A) we encounter obstructions θ_i in $H^i(M; \pi_i(G, PL))$. For $i \equiv 0 \pmod 4$, their image in rational cohomology can be identified with the difference of the rational Pontrjagin classes (or rather, L-classes) of M and M'. By Novikov's result, this difference is zero. Thus θ_{4k} is a torsion element. Next consider θ_{4k+2} which, by the result on k-invariants of G/PL, is a well-defined obstruction. If all lower θ_i vanish, this is annihilated by the homomorphism induced by $G/PL \to G/Top$; on the other hand, we have seen that this homomorphism is injective. Thus (A) can be justified on the sole assumption that for all k, $H^{4k}(M; \mathbf{Z})$ is torsion free. I believe that the best known result (due, again, to Sullivan) is slightly stronger—but in any case, the main obstacles to further progress are now presented by the Whitehead group (simple homotopy equivalence) and by the surgery obstructions.

Dept. of Pure Mathematics,
University of Liverpool, England

REFERENCES

Bing R. H.
[1] An alternative proof that 3-manifolds can be triangulated, *Ann. of Math.*, **69** (1959), 37-65.

Browder W.
[1] Homotopy type of differentiable manifolds, Notes, Colloquium on Algebraic Topology, Aarhus (1962), 42-46.
[2] Embedding smooth manifolds, Proc. I.C.M., Moscow (1966), 712-719.

Browder W., Levine J.
[1] Fibering manifolds over S^1, *Comm. Math. Helv.*, **40** (1966), 153-160.

Cerf J.
[1] Isotopie et pseudo-isotopie, Proc. I.C.M., Moscow (1966), 429-437.

Dold A., Lashof R.
[1] Principal quasifibrations and fibre homotopy equivalence of bundles, *Illinois J. Math.*, **3** (1959), 285-305.

Haefliger A.
[1] Knotted spheres and related geometric problems, Proc. I.C.M., Moscow (1966), 437-445.

Hirsch M. W.
[1] On nonlinear cell bundles, *Ann. of Math.*, **84** (1966), 373-385.
[2] On tangential equivalence of manifolds, *Ann. of Math.*, **83** (1966), 211-217.
[3] Smoothings of manifolds, Abstracts of reports I.C.M., Moscow (1966).

Hirsch M. W., Mazur B.
[1] Smoothings of piecewise linear manifolds, notes, Cambridge University, 1964.

Hudson J.
[1] Concordance and isotopy of PL-embeddings, *Bull. Amer. Math. Soc.*, **72** (1966), 534-535.

Kervaire M. A.
[1] Le théorème de Barden-Mazur-Stallings, *Comm. Math. Helv.*, **40** (1965), 31-42.

Kervaire M. A., Milnor J. W.
[1] Groups of homotopy spheres, *Ann. of Math.*, **77** (1963), 504-537.

Kister J. M.
[1] Microbundles are fibre bundles, *Ann. of Math.*, **80** (1964), 190-199.

Kuiper N. H., Lashof R. K.
[1] Microbundles and bundles, I, Elementary theory, Inventiones Math., **1** (1966), 1-17.

Lashof R., Rothenberg M.
[1] Microbundles and Smoothing, *Topology*, **3** (1964), 357-388.

Levine J.
[1] A classification of differentiable knots, *Ann. of Math.*, **82** (1965), 15-50.

Mazur B.
[1] Differential topology from the point of view of simple homotopy theory, Publ. Math. I.H.E.S., No. 15 (1963).
[2] The method of infinite repetition in pure topology, I, *Ann. of Math.*, **80** (1964), 201-226.
[3] The method of infinite repetition in pure topology, II, *Ann. of Math.*, **83** (1966), 387-401.
Milnor J.
[1] Lectures on characteristic classes, notes, Princeton University (1957).
[2] A procedure for killing the homotopy groups of differentiable manifolds, Amer. Math. Soc. Symp. in Pure Math., III (1961), 39-55.
[3] Microbundles and differentiable structures, notes, Princeton University (1961).
[4] Topological manifolds and smooth manifolds, Proc. I.C.M. Stockholm (1962), 132-138.
[5] Microbundles, I, *Topology*, **3**, suppl. 1 (1964), 53-80.
Milnor J., Spanier E.
[1] Two remarks on fibre homotopy type, *Pacific J. Math.*, **10** (1960), 585-590.
Moise E. E.
[1] Affine structures in 3-manifolds. V. The triangulation theorem and Hauptvermutung, *Ann. of Math.*, **56** (1952), 96-114.
Morlet C.
[1] Les voisinages tubulaires des varietés semi-linéaires, *Comptes Rendus*, **262** (1966), 740-743.
Munkres J. R.
[1] Elementary differential topology, Ann. of Math. Studies, no. 54 (Princeton 1963).
Newman M. H. A.
[1] The engulfing theorem for topological manifolds, *Ann. of Math.*, 84 (1966), 555-571.
Новиков С. П.
[1] Гомотопически эквивалентные гладкие многообразия, I, *Изв. АН СССР*, сер. матем., **28** (1964), 365-474.
[2] О многообразиях со свободной абелевой фундаментальной группой и их применениях, *Изв. АН СССР*, сер. матем., **30** (1966), 207-246.
Rourke C. P., Sanderson B. J.
[1] Block bundles, I, II, to appear.
Smale S.
[1] Differentiable and combinatorial structures on manifolds, *Ann. of Math.*, 74 (1961), 498-502.
Spivak M.
[1] Spaces, satisfying Poincaré duality, *Topology*, **6** (1967), 77-102.
Stasheff J.
[1] A classification theorem for fibre spaces, *Topology*, **2** (1963), 239-246.
Sullivan D.
[1] Triangulating homotopy equivalences, to appear.
Wall C. T. C.
[1] Surgery of non-simply-connected manifolds, *Ann. of Math.*, 84 (1966), 217-276.
[2] Poincaré complexes, I, to appear.
Whitehead J. H. C.
[1] On C^1-complexes, *Ann. of Math.*, **41** (1940), 809-824.
[2] Simple homotopy types, *Amer. J. Math.*, **72** (1950), 1-57.
Williamson R. E.
[1] Cobordism of combinatorial manifolds, *Ann. of Math.*, **83** (1966), 1-33.

9 Геометрия Geometry Géométrie Geometrie

MORSE THEORIE AUF DEM RAUM DER GESCHLOSSENEN KURVEN

WILHELM KLINGENBERG

1. Sei X ein bogenweise zusammenhängender topologischer Raum. Dann ist diesem Raum zugeordnet ein Raum $\Lambda(X)$ *von parametrisierten geschlossenen Kurven*, etwa, indem man die Menge der stetigen Abbildungen $f\colon S^1 \to X$ des mit $[0, 1]/\{0, 1\}$ parametrisierten Kreises S^1 in dem Raum X mit der kompakt offenen Topologie versieht. Die natürliche Operation von $O(2)$ auf S^1 induziert eine stetige Operation auf $\Lambda(X)$, den Quotientenraum $\Pi(X)$ wird man als *Raum der unparametrisierten geschlossenen Kurven von X bezeichnen.*

Indem man einem $f = (f(t))$, $0 \leqslant t \leqslant 1$, aus $\Lambda(X)$ den Punkt $f(0) = f(1)$ in X zuordnet, erhält man eine Serre Faserung von $\Lambda(X)$ über X mit $\Omega(X)$, dem Schleifenraum, als Faser. Diese Bemerkung hat Шварц [7] benutzt, um die Homologie von $\Lambda(X)$ mit der Methode der spektralen Sequenzen zu untersuchen. Die Homologie von $\Pi(X)$ läßt sich dann mit den Hilfsmitteln der kompakten Transformationsgruppen studieren.

Wenn X speziell eine kompakte differenzierbare Mannigfaltigkeit M ist, dann läßt sich zu M ein ebenfalls mit $\Lambda(M)$ bezeichneter Raum von parametrisierten geschlossenen Kurven erklären, der die Struktur einer Hilbert-Mannigfaltigkeit trägt—wir verwenden wieder die Bezeichnung $\Lambda(M)$, da die beiden Räume denselben Homotopietyp haben. Eine Riemannsche Metrik auf M gibt Anlaß zu einer Riemannschen Metrik auf $\Lambda(M)$ und zu einer differenzierbaren Funktion, für welche die Voraussetzungen der von Palais [5] und Smale [6] entwickelten Übertragung der Theorie von Morse auf Hilbert-Mannigfaltigkeiten gelten. Damit haben wir also ein weiteres Hilfsmittel an der Hand, die Homologie von $\Lambda(M)$ und $\Pi(M)$ zu studieren. Um seine Durchschlagskraft zu demonstrieren, bestimmen wir die bislang nicht bekannte $\boldsymbol{Z}_2$-Homologie des Raumes $\Pi(S^n)$, mit dem sich schon Morse [4] und Bott [1] befaßt haben. Mit derselben Methode läßt sich auch die Homologie von $\Lambda(P)$ und $\Pi(P)$ bestimmen, wenn P ein

projektiver Raum über den reellen oder komplexen Zahlen oder über den Quaternionen oder Cayleyschen Zahlen ist.

2. Sei also jetzt M eine kompakte differenzierbare Mannigfaltigkeit. Unter $\Lambda(M)$ verstehen wir den Raum der absolut stetigen Abbildungen $f: S^1 \to M$, welche in lokalen Koordinaten quadrat-integrierbare erste Ableitungen besitzen. Die auf diesem Raum mit Hilfe von lokalen Koordinaten erklärte 1-Norm bestimmt auf $\Lambda(M)$ die Struktur einer Hilbert-Mannigfaltigkeit, vgl. Palais [5]. $\Pi(M)$ erklären wir wie eingangs als Quotientenraum $\Lambda(M)/\boldsymbol{O}(2)$ bezüglich der natürlichen Operation von $\boldsymbol{O}(2)$ auf $\Lambda(M)$. $\pi: \Lambda \to \Pi$ bezeichne die Projektionsabbildung.

Offenbar hängen der Homotopietyp von $\Lambda(M)$ und $\Pi(M)$ nur ab vom Homotopietyp von M.

Sei jetzt auf M eine Riemannsche Metrik, $\langle\ ,\ \rangle$, gegeben. Diese liefert auf $\Lambda(M)$ eine *Riemannsche Metrik*, $\langle\langle\ ,\ \rangle\rangle$, wie folgt:

$$\langle\langle X,\ X\rangle\rangle = \int_0^1 (\langle X(t),\ X(t)\rangle + \langle DX(t)/dt,\ DX(t)/dt\rangle\, dt,$$

wobei $X = (X(t))$, $0 \leqslant t \leqslant 1$, ein Tangentialvektor an $\Lambda(M)$ ist, das heißt, ein absolut stetiges Vektorfeld längs eines $f = (f(t))$, $0 \leqslant t \leqslant 1$, aus $\Lambda(M)$ mit quadrat-integrierbaren ersten Ableitungen. D/dt ist die kovariante Ableitung.

Man zeigt, daß $\Lambda(M)$ mit dieser Metrik *vollständig* ist.

Ferner liefert $\langle,\rangle$ eine differenzierbare Funktion $E: \Lambda(M) \to \boldsymbol{R}$, das sogenannte *Energieintegral*:

$$E(f) = \frac{1}{2}\int_0^1 \langle df(t)/dt,\ df(t)/dt\rangle\, dt.$$

Für das zugehörige negative Gradientenfeld $-\mathrm{grad}\ E$ gilt dann die Bedingung (C) von Palais [5] und Smale [6], unter welcher sich die Theorie von Morse übertragen läßt auf Hilbert-Mannigfaltigkeiten.

Für ein $c \geqslant 0$ setzen wir $E^{-1}([0, c]) = \Lambda^c$ und $E^{-1}([0, c[) = \Lambda^{c-}$. Für alle $t \geqslant 0$ sind die Integralkurven des Vektorfeldes $-\mathrm{grad}\ E$ erklärt. Wir erhalten eine Halbgruppe von Deformationen $\varphi_t: \Lambda \to \Lambda$, $t \geqslant 0$, indem wir einem f denjenigen Punkt zuordnen, den die in f mit $t = 0$ beginnende Integralkurve zur Zeit t erreicht hat. Da E und $\langle,\rangle$ mit der Operation von $\boldsymbol{O}(2)$ verträglich sind, wird durch $\boldsymbol{O}(2)$ das Feld $-\mathrm{grad}\ E$ in sich transformiert. Die Deformationen φ_t sind also verträglich mit der Operation von $\boldsymbol{O}(2)$ und sie induzieren daher eine Halbgruppe von Deformationen $\psi_t: \Pi \to \Pi$, $t \geqslant 0$, des Raumes der unparametrisierten Kurven.

Beachte, daß φ_t und ψ_t E-vermindernde Deformationen sind, welche kanonisch definiert sind und nicht von irgendwelchen Hilfskonstruktionen, wie etwa Unterteilungen der f, abhängen — sehr zum Unterschied von der bisherigen Methode, E-vermindernde Deformationen auf Kurvenräumen zu erklären.

3. Ein Element $f \in \Lambda(M)$ heißt *kritisch*, wenn $\mathrm{grad}\ E(f) = 0$. Mit einem Argument der klassischen Variationsrechnung zeigt man, *daß f genau dann kritisch ist, wenn es eine geschlossene Geodätische ist, mit Parameter proportional zur Bogenlänge.*

Insbesondere sind die konstanten Abbildungen $f: S^1 \to M$ kritische Punkte. Wir nennen diese auch die trivialen kritischen Punkte. Sie bilden eine *nicht-entartete kritische Untermannigfaltigkeit* $\Lambda^0(M) = E^{-1}(0)$ *vom Index* 0 in $\Lambda(M)$, im Sinne von Bott [1], welche natürlich isomorph ist zu M.

Wenn f kritisch ist, so ist auch der Orbit von f unter $\boldsymbol{O}(2)$ kritisch. f heißt *nicht-entartet*, wenn sein Orbit eine nicht-entartete kritische Untermannigfaltigkeit im Sinne von Bott [1] ist.

Für jedes $f \in \Lambda$ ist die Isotropiegruppe erklärt als diejenige Untergruppe von $\boldsymbol{O}(2)$, die auf f als Identität operiert. Wenn die Isotropiegruppe von f die zyklische Untegruppe $\boldsymbol{Z}_q$ der Ordnung q von $\boldsymbol{SO}(2)$ ist, dann nennen wir f eine Kurve der *Multiplizität* q. Kurven der Multiplizität 1 nennen wir auch *einfach*. Falls $f = (f(t))$ die Multiplizität q hat, dann ist auch $(f(t/q))$ ein Element von Λ, die sogenannte f *unterliegende einfache Kurve*.

4. Sei f eine nicht-entartete geschlossene Geodätische, sei F sein Orbit und sei $g = \pi(f) = \pi(F)$ die zugehörige unparametrisierte geschlossene Geodätische. Dann ist $E(f) = c$ positiv. Sei q die Multiplizität von f. F ist dann isolierte kritische Untermannigfaltigkeit, und es gibt eine unter $\boldsymbol{O}(2)$ invariante offene Umgebung $U = U(F)$ von F, deren Kurven ausschließlich eine Multiplizität q' haben, wo q' q teilt. Wir können sogar annehmen, daß U ein Hilbert Diskus-Bündel über F ist.

Der *Index* j *von* f bzw. F ist die Dimension des von den zu negativen Eigenwerten der Indexform (eingeschränkt auf den Normalenraum von F) gehörigen Eigenvektoren aufgespannten Raumes. Diese Eigenvektoren bilden das sogenannte *negative Bündel über* F. Durch die Exponentialabbildung ist ein Diskus-Bündel $D^j(F)$ dieses negativen Bündels in natürlicher Weise mit einem j-dimensionalen Diskus-Unterbündel des Hilbert Diskus-Bündels $U(F)$ über F identifiziert.

Setze $U(F) - F = U_0(F)$ und $D^j(F) - F = D_0^j(F)$. Wie in der klassischen Morse Theorie zeigt man, *daß $U(F)$ sich äquivariant* (d. h., auf mit der Operation von $\boldsymbol{O}(2)$ verträgliche Weise) *auf $U(F)^c = U(F) \cap \Lambda^c$ deformieren läßt und daß sich $(U(F)^c, U_0(F)^c)$ äquivariant auf $(D^j(F), D_0^j(F))$ deformieren läßt.*

Unter der Projektion π liefert dies eine Deformation von $V(g) = \pi(U(F))$ auf $V(g)^c = \pi(U(F)^c)$ und von $(V(g)^c, V_0(g)^c)$ auf $(\pi D^j(F), \pi D_0^j(F))$.

Wir wollen $\pi D^j(F)$ bestimmen. Dazu faktorisieren wir $\pi \mid D^j(F)$ wie folgt:

$$D^j(F) \xrightarrow{Z_q} D^j(F)/Z_q \xrightarrow{O(2)/Z_q} D^j(g)/Z_q.$$

Das heißt, wir dividieren zunächst durch die Operation der Isotropiegruppe Z_q und dann durch die induzierte Projektion. Der Quotient $D^j(F)/Z_q$ bedeutet, daß wir das D^j-Bündel $D^j(F)$ durch die Operation der Isotropiegruppe Z_q auf den Fasern D^j dividieren. Der zweite Quotient ist eine triviale Faserung, mit $O(2)/Z_q$, isomorph $O(2)$, als Faser. $\pi(D^j(F)/Z_q) = D^j(g)/Z_q$ *ist also isomorph dem Quotienten* D^j/Z_q *einer Faser* D^j *von* $D^j(F)$ *nach* Z_q.

Wir erkennen auf diese Weise, wie zuerst von Шварц [7] bemerkt wurde, daß für die relative Homologie $H_*(V(g), V(g)^{c-}) = H^*(D^j(g)/Z_q, D_0^j(g)/Z_q)$ die Operation der Isotropiegruppe Z_q auf dem "negativen" Diskus D^j in Betracht zu ziehen ist. Man findet für die Homologiegruppe mit Z_2-Koeffizienten damit folgendes (vgl. Шварц [7]):

Falls q ungerade, so $H_i(V(g), V(g)^{c-}) = Z_2$ für $i = j$, und $= 0$ sonst.

Falls $q = 2^{q_0}q_1$ gerade, mit q_1 ungerade, so $H_i(V(g), V(g)^{c-}) = Z_2$ für $i(q_1) + 2 \leqslant i \leqslant j$, und $= 0$ sonst. Hierbei ist $i(q_1)$ der Index der q_1-fachen Überlagerung der g unterliegenden einfachen Geodätischen.

5. Die Morse Theorie stellt eine Beziehung her zwischen der Homologie von $\Lambda(M)$ und $\Pi(M)$ einerseits und den geschlossenen Geodätischen g und ihren Typenzahlen (d. h., lokalen, relativen Homologiegruppen $H_i(V(g), V(g)^{c-})$) andererseits. Insbesondere kann man aus der Homologie von $\Lambda(M)$ und $\Pi(M)$ auf die Existenz einfacher geschlossener Geodätischer schließen, vgl. dazu [3].

Wir wollen hier umgekehrt aus den Eigenschaften der geschlossenen Geodätischen Information über die Homologie von $\Lambda(M)$ und insbesondere von $\Pi(M)$ gewinnen. Die Schwierigkeit hierbei ist, daß die geschlossenen Geodätischen und ihre Typenzahlen in hohem Maße abhängen von der Wahl der Metrik $\langle , \rangle$ auf M, während der Homologietyp von $\Lambda(M)$ und $\Pi(M)$ ja nur abhängen vom Homotopietyp von M.

Falls es jedoch in der Homotopieklasse von M eine Mannigfaltigkeit mit einer augezeichneten Metrik gibt, wie etwa wenn M in der Homotopieklasse einer Sphäre oder eines allgemeineren symmetrischen Raumes liegt, dann kann man erwarten, mit Hilfe der Morse

Theorie Information über die Homologie von $\Lambda(M)$ und $\Pi(M)$ gewinnen zu können.

6. Wir wollen dieses am Beispiel der Sphäre zeigen. Es sei also jetzt $M = S^n$, mit der Metrik konstanter Krümmung $K = 2\pi^2$. Wir beschränken uns auf die Z_2-Homologie.

Die nicht-trivialen kritischen Punkte auf $\Lambda = \Lambda(S^n)$ sind die q-fach durchlaufenen parametrisierten Großkreise. Das Energieintegral eines solchen Großkreises ist q^2. Die q-fach durchlaufenen parametrisierten Großkreise bilden eine kritische Untermannigfaltigkeit F_q, welche isomorph ist zur Stiefel-Mannigfaltigkeit $V(2, n-1)$ der orthonormalen 2-Beine im R^{n+1}.

Das Bild $G_q = \pi(F_q)$ in $\Pi = \Pi(S^n)$, also die q-fach durchlaufenen unparametrisierten Großkreise, ist isomorph zur Grassmann-Mannigfaltigkeit $G(2, n-1)$ der 2-Ebenen im R^{n+1}.

Über $G(2, n-1)$ haben wir einerseits das *kanonische* $(n-1)$*-Vektorraum-Bündel* ζ^{n-1} und ferner das *Tangentialbundel* τ^{2n-2}. Die durch π: $V(2, n-1) \to G(2, n-1)$ über $V(2, n-1)$ induzierten Bündel bezeichnen wir mit η^{n-1} bzw. σ^{2n-2}.

Wir bemerken, daß σ^{2n-2} eine komplexe Struktur trägt, d.h., die Strukturgruppe $O(2n-2)$ kann auf $U(n-1)$ reduziert werden. Dasselbe trifft zu für die 2-fache Überlagerung $\tilde{G}(2, n-1)$ von $G(2, n-1)$, also für die orientierte Grassmann-Mannigfaltigkeit. Es folgt, daß die Strukturgruppe von $G(2, n-1)$ reduziert werden kann auf die Erweiterung der Ordnung 2, $\tilde{U}(n-1)$, von $U(n-1)$ mit dem Element, welches den Übergang zum Konjugiert-Komplexen beschreibt.

Die kritische Untermannigfaltigkeit F_q der q-fach durchlaufenen parametrisierten Großkreise ist nicht-entartet. Das negative Bündel über $F_q = V(2, n-1)$ ist gegeben durch

$$(*) \qquad \begin{cases} \eta^{n-1}, & \text{für } q = 1, \\ \eta^{n-1} \oplus \underbrace{\sigma^{2n-2} \oplus \ldots \oplus \sigma^{2n-2}}_{(q-1)\text{Summanden}}, & \text{für } q > 1. \end{cases}$$

Es folgt, daß

$$H_*(\Lambda^{q^2}(S^n), \Lambda^{q^2-}(S^n)) = H_{*-(2q-1)(n-1)}(V(2, n-1)).$$

Ferner gilt:

Die Inklusion $\Lambda^c(S^n) \to \Lambda(S^n)$ induziert einen injektiven Homomorphismus.

Damit folgt:

$H_*(\Lambda(S^n), \Lambda^0(S^n))$ ist die direkte Summe, über alle $q \geqslant 1$ von $H_{*-(2q-1)(n-1)}(V(2, n-1))$. Vgl. Bott [1].

B e m e r k u n g: Wenn wir die Serre Faserung $\Omega(S^n) \to \Lambda(S^n) \to S^n$ betrachten, ist die zugehörige spektrale Sequenz trivial, und es gilt sogar multiplikativ $H^*(\Lambda(S^n)) = H^*(S^n) \otimes H^*(\Omega(S^n))$, vgl. Шварц [7].

Um das Bild des negativen Bündels (*) über F_q unter π zu bestimmen, müssen wir, wie schon in 4. bei der Bestimmung von $D^j(F)$, untersuchen, wie die Isotropiegruppe $\boldsymbol{Z}_q$ auf den Fasern des Bündels operiert. Es stellt sich heraus, daß diese wie die Identität auf η^{n-1} operiert und wie die Multiplikation mit $e^{2\pi i p/q}$ auf dem p-ten Summanden σ^{2n-2} von (*), $1 \leqslant p \leqslant q-1$. Dabei benutzen wir, daß das Zentrum der Strukturgruppe $\boldsymbol{U}(n-1)$ von σ^{2n-2} gleich $\boldsymbol{U}(1)$ ist.

Wir bezeichnen den Quotienten von σ^{2n-2} nach dieser Operation von $\boldsymbol{Z}_q$ mit $\sigma^{2n-2}/\boldsymbol{Z}_q$. Wir lassen hierauf die induzierte Projektion π wirken, d.h., wir bilden den Quotienten nach $\boldsymbol{O}(2)/\boldsymbol{Z}_q$.

Diese Gruppe wirkt effektiv auf $\sigma^{2n-2}/\boldsymbol{Z}_q$, das Bild $\tau^{2n-2}/\boldsymbol{Z}_q$ ist also die Basis einer Faserung von $\sigma^{2n-2}/\boldsymbol{Z}_q$ mit Faser $\boldsymbol{O}(2)/\boldsymbol{Z}_q$, isomorph zu $\boldsymbol{O}(2)$. Diese Operationen sind verträglich mit der Faserung $\sigma^{2n-2} \to F_q$, wir erhalten also das kommutative Diagramm

$$\begin{array}{ccccc} \sigma^{2n-2} & \xrightarrow{\boldsymbol{Z}_q} & \sigma^{2n-2}/\boldsymbol{Z}_q & \xrightarrow{\boldsymbol{O}(2)/\boldsymbol{Z}_q} & \tau^{2n-2}/\boldsymbol{Z}_q \\ \downarrow & & \downarrow & & \downarrow \\ F_q & \xrightarrow{\mathrm{id}} & F_q & \xrightarrow{\boldsymbol{O}(2)/\boldsymbol{Z}_q} & G_q \end{array}$$

$\tau^{2n-2}/\boldsymbol{Z}_q$ ist das Bündel über $G_q = G(2, n-1)$, das aus dem Tangentialbündel τ^{2n-2} durch Quotientenbildung mit der Operation (Multiplikation mit $e^{2\pi i p/q}$) von $\boldsymbol{Z}_q$ auf der Faser entsteht.

Wir finden also: *Das "negative" Bündel über dem Raum $G_q = G(2, n-1)$ der q-fach durchlaufenen unparametrisierten Großkreise ist gegeben durch:*

$$(**) \qquad \begin{array}{ll} \zeta^{n-1}, & \text{für } q = 1, \\ \zeta^{n-1} \oplus \tau^{2n-2} \oplus \ldots \oplus \tau^{2n-2}/\boldsymbol{Z}_q, & \text{für } q > 1, \end{array}$$

wobei $\boldsymbol{Z}_q$ auf dem p-ten Summanden τ^{2n-2} durch Multiplikation mit $e^{2\pi i p/q}$ operiert.

Die relative Kohomologie $H^*(\Pi^{q^2}(S^n), \Pi^{q^2-}(S^n))$ im kritischen Niveau q^2 ist also durch die Kohomologie des Thom Raumes T (**) des negativen Bündels (**) über $G_q = G(2, n-1)$ gegeben.

Für $q = 1$ findet man $H^*(T\zeta^{n-1}) = u^{n-1} \cup H^*(G(2, n-1))$, vgl. Morse [4], Bott [1].

Für $q > 1$ ist die Kohomologie des Thom Raumes von (**) gegeben durch

$$u^{(2q-1)(n-1)} \cup H^*(G(2, n-1)),$$

falls q ungerade, und

$$u^{(2q_1-1)(n-1)} + u^{(2q_1-1)(n-1)+3} + \ldots + u^{(2q-1)(n-1)} \cup H^*(G(2, n-1)),$$

falls $q = 2^{q_0} q_1$ gerade, q_1 ungerade. Hier ist $\dim u^i = i$.

Es gilt nun wielderum: *Die Inklusion $(\Pi^c(S^n) \to \Pi(S^n)$ induziert einen injektiven Homomorphismus in der Homologie.*

Folglich haben wir:

Die $\boldsymbol{Z}_2$-Homologie von $\Pi(S^n)$ mod $\Pi^0(S^n)$ ist die direkte Summe, über alle $q \geqslant 1$, von $H_*(\Pi^{q^2}(S^n), \Pi^{q^2-}(S^n))$, wobei jeder dieser Summanden gegeben ist durch die Homologie des Thom Raumes des negativen Bündels (**) über $G_q = G(2, n-1)$, die oben beschrieben wurde.

B e m e r k u n g: Damit ist ein altes Problem gelöst, das schon von Morse [4] und Bott [5] in Angriff genommen wurde, das aber, wie Шварц 1956 hier in Moskau bemerkte, von diesen Autoren nicht korrekt beantwortet wurde.

7. Wir schließen mit der Bemerkung, daß sich mit genau derselben Methode die Homologie von

$$\Lambda(P(\lambda)) \quad \text{und} \quad \Pi(P(\lambda))$$

bestimmen läßt, wenn $P(\lambda)$ ein proejktiver Raum ist über den reellen Zahlen ($\lambda = 1$), den komplexen Zahlen ($\lambda = 2$), den Quaternionen ($\lambda = 4$) oder den Cayleyschen Zahlen ($\lambda = 8$).

Man benutzt für diese symmetrischen Räume $P(\lambda)$ vom Range 1 wiederum die Tatsache, daß bei der kanonischen Metrik die nichttrivialen kritischen Punkte von $\Lambda(P(\lambda))$ aus den q-fach durchlaufenen Großkreisen bestehen, und daß diese eine nicht-entartete kritische Mannigfaltigkeit F_q bilden, $q = 1, 2, \ldots$ ..Großkreise von $P(\lambda)$ sind die Großkreise auf den 1-dimensionalen projektiven Unterräumen, welche isometrisch zu den λ-dimensionalen Sphären S^λ sind.

Das negative Bündel über F läßt sich wiederum explizit bestimmen, und die Inklusionen $\Lambda^c(P(\lambda)) \to \Lambda(P(\lambda))$ und $\Pi^c(P(\lambda)) \to \Pi(P(\lambda))$ induzieren wiederum einen injektiven Homomorphismus in der Homologie.

Es wäre interessant zu untersuchen, ob diese Tatsachen sich auch auf allgemeinere symmetrische Räume übertragen. Die Dissertation von Eliasson [2], wo die geschlossenen Geodätischen auf der Grassmann-Mannigfaltigkeit $G(2, n-1)$ untersucht werden, scheint darauf hinzudeuten, daß hier ein allgemeineres Prinzip herrscht.

Mathematisches Institut der Universität Bonn,
Bundesrepublik Deutschland

LITERATUR

[1] Bott R., Nondegenerate critical manifolds, *Ann. of Math.*, 60 (1954), 248-260.
[2] Eliasson H., Über die Anzahl geschlossener Geodätischer in gewissen Riemannschen Mannigfaltigkeiten, *Math. Ann.*, 166 (1966), 119-147.
[3] Klingenberg W., The Theorem of the three closed geodesics, *Bull. Amer. Math. Soc.*, 71 (1965), 601-605.
[4] Morse M., The calculus of variations in the large, Providence, R.I., 1934.
[5] Palais R., Morse theory on Hilbert manifolds, *Topology*, 2 (1963), 299-340.
[6] Smale S., Morse theory and a non-linear generalization of the Dirichlet problem, *Ann. of Math.*, 80 (1964), 382-396.
[7] Шварц А. С., Гомологии пространств замкнутых кривых, *Труды Московского математического общества*, 9 (1960), 3-44.

Алгебраическая геометрия и комплексные многообразия
Algebraic geometry and complex manifolds
Géométrie algébrique et variétés complexes
Algebraische Geometrie und komplexe Mannigfaltigkeiten

ON THE PROBLEM OF RESOLUTION OF SINGULARITIES

SH. SH. ABHYANKAR

1. The problem and its history

The resolution problem asks if every irreducible projective algebraic variety, defined over some ground field, can be birationally transformed into a nonsingular one. The same question can also be raised in the arithmetical case, i.e., for varieties defined over the ring of ordinary integers. To cover both these cases, the problem may be stated thus:

Resolution problem. Given a function field K over a pseudogeometric Dedekind domain k, does there exist a nonsingular projective model of K over k?

Before giving the history of the problem, let us recall the definitions of the terms used above.

Definition. By a pseudogeometric Dedekind domain we mean a noetherian integral domain k such that: k is integrally closed in its quotient field; every nonzero prime ideal in k is maximal; and the integral closure of k in any finite algebraic extension of the quotient field of k is a finite k-module. Note that every field is a pseudogeometric Dedekind domain, and so is the ring of ordinary integers. By a function field over a noetherian integral domain k we mean a finitely generated field extension K of the quotient field of k; by a projective model of K over k we mean a set V of local rings for which there exists a finite number of nonzero elements $x_0, \ldots, x_e$ in K such that K is the quotient field of $k[x_1/x_0, \ldots, x_e/x_0]$ and $V = V_0 \cup V_1 \cup \ldots \cup V_e$ where V_i is the set of all quotient rings of $k[x_0/x_i, \ldots, x_e/x_i]$ with respect to the various prime ideals in $k[x_0/x_i, \ldots, x_e/x_i]$; we define $\dim V$ to be $\max_{R \in V}(\dim R)$ where $\dim R$ denotes the (Krull) dimension of the local ring R; V is said to be nonsingular if every element in V is a regular local ring. Given a function field K over a pseudogeometric Dedekind domain k, let n^* be the transcendence degree of K over the quotient field of k, and let $n = n^*$ or $n^* + 1$ according as k is or is not

a field; n is called the (Kroneckerian) dimension of K over k; note that then dim $V = n$ for every projective model V of K over k.

History. Let K be a function field over a pseudogeometric Dedekind domain k, and let n be the dimension of K over k. The Resolution Problem has so far been settled affirmatively in the following cases: For $n = 1$ the solution is classical. For $n = 2$ and k = the field of complex numbers, after several geometric solutions by the Italians (such as Albanese, Levi, etc., see Chapter I of [20]), the first rigorous solution was given by Walker [19] in 1935. Walker's solution is function-theoretic and makes use of the local solution (i.e., solution of the local uniformization problem which is a localized version of the Resolution Problem) for $n = 2$ and k = the field of complex numbers given by Jung [18] in 1908. Then, in 1939-1944, Zariski introduced the tools of local algebra into algebraic geometry, and thereby obtained [21, 23, 24] a solution of the Resolution Problem for $n \leqslant 3$ and k = a field of characteristic zero; at that time he also obtained [22] a local solution for n arbitrary and k = a field of characteristic zero. In 1956, Abhyankar [4, 5, 9] gave a solution for $n = 2$ and k = a perfect field of nonzero characteristic. Finally, in 1964, Hironaka [17] settled the Resolution Problem for n arbitrary and k = a field of characteristic zero; Hironaka's solution is especially marked by a vigorous induction and by his ability to deal with several simultaneous equations. Recently, being prompted by Hironaka's outstanding work, I have extended my 1956 method and thereby obtained [2, 10, 11, 12, 13, 14, 15, 16] a solution of the Resolution Problem when (1) $n = 2$ and k/P is perfect for every maximal ideal P in k, and also when (2) $n = 3$ and k = an algebraically closed field of characteristic different from 2, 3, 5; note that the main novelty of (1) is for the case of an arithmetical surface, i.e., when $n = 2$ and k = the ring of ordinary integers. The chief reason why, after ten years, I was able to make some progress is that, in 1963—1964, I had the precious opportunity to spend some six months in Zariski's neighborhood (not a Zariski neighborhood). Thus I would like to thank Zariski for the inspiration, and Hironaka for the incentive.

2. Embedded resolution and dominance

Although the arithmetical case is genuinely interesting, in this lecture, so as to fix ideas, I shall say nothing more about it. Henceforth all varieties (and function fields) are assumed to be over an algebraically closed ground field k of characteristic p which may or may not be zero. Having once stated the problem precisely, henceforth I shall speak quite informally. A common feature of most of the proofs cited above is that to prove resolution for dimension n one needs a stronger result for dimension less than n which includes at least the following:

Embedded resolution. Given a nonsingular projective algebraic variety W of dimension n and a hypersurface H in W, there exists a composite monoidal transformation q: $W' \to W$ such that the total transform $q^{-1}(H)$ of H on W' has only normal crossings.

Concerning the definitions of the terms used above I shall only say this: Given any subvariety D of an irreducible algebraic variety W there exists a well-defined birational map q: $W' \to W$ called the monoidal transformation of W with center D; q is biregular on $W - D$; if W and D are nonsingular then so is W'; if D is a point then q is called the quadratic transformation with center D. We shall only consider monoidal transformations with nonsingular irreducible centers. If W is a nonsingular variety and $W_t \to W_{t-1} \to \ldots \to W_1 \to W$ is a sequence of monoidal transformations (with nonsingular irreducible centers) then the resulting birational map $W_t \to W$ is called a composite monoidal transformation. By a hypersurface we mean a subvariety (which may be reducible) of codimension one. A hypersurface H in an n-dimensional nonsingular variety W is said to have a normal crossing at a point $P \in H$, if there exist regular parameters $x_1, \ldots, x_n$ on W at P such that, at P, H is defined by $x_1 \ldots x_s = 0$ for some $s \leqslant n$; H is said to have only normal crossings if H has a normal crossing at each of its points.

Note that in the above formulation, Embedded Resolution for n subsumes Resolution for $n - 1$.

Usually, after Embedded Resolution for n and before Resolution for n one proves the following:

Dominance (or removal of points of indeterminacy). Given any two nonsingular projective models W and W^* of an n-dimensional function field, there exists a composite monoidal transformation $W' \to W$ such that W' dominates W^*.

The proof of Embedded Resolution and Dominance is quite easy for $n \leqslant 2$; in the works cited in § 1, these two statements have been proved: for $n = 3$ and $p = 0$ by Zariski; for $n = 3$ and $p \neq 0$ by Abhyankar; and for n arbitrary and $p = 0$ by Hironaka. Finally we note that, as an application of Dominance, we now have the birational invariance of the arithmetic genus of nonsingular projective algebraic varieties of dimension 3 (any p).

3. Peculiarities of nonzero characteristic

I shall now make various comments as to how the case $p \neq 0$ differs from the case $p = 0$ and what are the possible ways to make up for this difference.

(1). Algebraically speaking, the basic reason why the $p = 0$ proofs (of Zariski and Hironaka) fail for $p \neq 0$ is the BINOMIAL THEOREM

$$(Z+Y)^m = Z^m + c_1 Z^{m-1} Y + c_2 Z^{m-2} Y^2 + \ldots + Y^m;$$

to wit: $c_1 \neq 0$ if $m \not\equiv 0\ (p)$, but $c_1 = 0$ if $m \equiv 0\ (p)$. More generally, let

$$\Phi(Z) = Z^m + \Phi_1 Z^{m-1} + \ldots + \Phi_m$$

and make the translation

$$\Psi(Z) = \Phi(Z+Y) = Z^m + \Psi_1 Z^{m-1} + \ldots + \Psi_m;$$

then what is the relationship between the Φ_i and Ψ_j, i.e., which of the Φ_i affect a particular Ψ_j and how much? In any case, I believe that a better and better understanding of the Binomial Theorem will enable one to resolve more and more singularities.

(2). To illustrate what I have said above, consider an algebroid hypersurface

$$V\colon G(Z_1, \ldots, Z_{n+1}) = 0,$$

in the $(n+1)$-dimensional local space A_{n+1}, having a point of multiplicity $m > 1$ at its origin. Assuming Embedded Resolution (or rather, its algebroid analogue) for n we would like to resolve the singularity of V. Upon making a linear transformation and invoking the Weierstrass Preparation Theorem we get

$$V\colon g(Y_1, \ldots, Y_n, Z) = 0$$

with

$$g = Z^m + \sum_{1 \leqslant i \leqslant m} g_i Z^{m-i}$$

where $g_i = g_i(Y_1, \ldots, Y_n)$ are power series in $Y_1, \ldots, Y_n$ with coefficients in k such that $g_i(0, \ldots, 0) = 0$. Let h be the product of those g_i which are nonzero. Upon applying Embedded Resolution to the hypersurface

$$H\colon h(Y_1, \ldots, Y_n) = 0$$

in the n-dimensional local space A_n we find a composite monoidal transformation $q\colon B \to A_n$ such that $q^{-1}(H)$ has only normal crossings. Let $X_1, \ldots, X_n$ be suitable parameters at a point Q in B. This amounts to substituting certain power series $\sigma_1(X_1, \ldots, X_n), \ldots, \sigma_n(X_1, \ldots, X_n)$ for $Y_1, \ldots, Y_n$ in $h(Y_1, \ldots, Y_n)$ so that

$$h(\sigma_1(X_1, \ldots, X_n), \ldots, \sigma_n(X_1, \ldots, X_n)) = h' X_1^{a(1)} \ldots X_n^{a(n)}$$

where $h' = h'(X_1, \ldots, X_n)$ with $h'(0, \ldots, 0) \neq 0$, and $a(1), \ldots, a(n)$ are nonnegative integers. Then actually

$$g_i(\sigma_1(X_1, \ldots, X_n), \ldots, \sigma_n(X_1, \ldots, X_n)) = f_i X_1^{a(i,1)} \ldots X_n^{a(i,n)}$$

for all i with $g_i \neq 0$, where $f_i = f_i(X_1, \ldots, X_n)$ with $f_i(0, \ldots, 0) \neq 0$, and $a(i, j)$ are nonnegative integers. q induces $q^*\colon B \times A_1 \to A_{n+1}$

where A_1 is the space of the variable Z. Let V^* be the proper transform of V by q^{*-1}. Then at the point $(Q, 0)$, V^* is given by

$$V^*\colon Z^m + \sum_{1 \leqslant i \leqslant m,\ g_i \neq 0} f_i X_1^{a(i,1)} \ldots X_n^{a(i,n)} Z^{m-i} = 0.$$

For the sake of simplicity, suppose it has happened that $a(i, j) = 0$ whenever $j \neq 1$. Let d be the greatest integer such that $d \leqslant a(i, 1)/i$ for all i with $g_i \neq 0$. Then $a(i', 1) - i'd < i'$ for some i' with $g_{i'} \neq 0$. Make the composite monoidal transformation given by: $Z = Z^* X_1^d$. Let V' be the proper transform of V^* under this transformation. Let P be a point of V'. Then

$$Z^* = \gamma \text{ at } P \text{ for a unique } \gamma \in k.$$

Let $Z' = Z^* - \gamma$. Then $X_1, \ldots, X_n, Z'$ are local parameters at P, and at P we have

$$V'\colon f(X_1, \ldots, X_n, Z') = 0$$

where

$$f = (Z' + \gamma)^m + \sum_{1 \leqslant i \leqslant m,\ g_i \neq 0} f_i X_1^{a(i,1) - id} (Z' + \gamma)^{m-i}.$$

Let m' be the multiplicity of V' at P. Since $a(i', 1) - i'd < i'$ for some i', we get that if $\gamma = 0$ then $m' < m$; a reduction. So now suppose that $\gamma \neq 0$. At this point there are various essentially equivalent ways of arguing provided $m \not\equiv 0\ (p)$. For instance following Hironaka, we can make the initial coordinate transformation $Z \to Z - (1/m) g_1$ which will have the effect that the coefficient of Z^{m-1} will be zero, i.e., we will have $g_1 = 0$, and hence

$$f = Z'^m + m\gamma Z'^{m-1} + \text{terms of degree less than } m-1 \text{ in } Z'.$$

Then $m' < m$ because $m\gamma \neq 0$. However, if $m \equiv 0\ (p)$ then in the first place we cannot make the transformation $Z \to Z - (1/m) g_1$, and in the second place even if g_1 were zero to begin with we still cannot conclude that $m' < m$ because now $m\gamma = 0$. We shall now make several observations.

(3). Peculiarity arises when the multiplicity is divisible by p.

(4). The most primitive case of the above peculiarity is afforded by: $Z^p - h(Y_1, \ldots, Y_n) = 0$. The two-dimensional case (i.e., $n = 2$) of this was explicitly mentioned by Zariski in his 1950 address [25] and there he pronounced it "untractable".

(5). It is not necessary to kill g_1 completely, i.e., it is enough to kill the terms of low degree in g_1. Namely, if $g_1 \neq 0$ and $a(1, 1) > d$ then again the coefficient of Z'^{m-1} in f would be a unit and we would have $m' < m$, provided $m \not\equiv 0\ (p)$. The condition that either $g_1 = 0$,

or $g_1 \neq 0$ and $a(1, 1) > b$, is equivalent to the inequality

(*) $\quad (X_1\text{-value of } g_1) > \min_{1 \leq i \leq m} ((1/i)(X_1\text{-value of } g_i)).$

If $m \equiv 0 \ (p)$ then g_1 does not play the dominant role. But it turns out that in general (i.e., whether m is divisible by p or not) we would be in a reasonable shape if, say

(**) $\quad (X_1\text{-value of } g_i) \geqslant (i/m)(X_1\text{-value of } g_m)$ for $1 \leqslant i \leqslant m$.

In other words, g_m instead of g_1 (i.e., the norm instead of the trace) is to be given a more dominant role. It can easily be seen that if X_1 (i.e., the valuation given by X_1) does not split in the field extension given by g, then the inequalities (**) hold. So one should try to arrange that X_1 does not split.

(6). In the general case, i.e., when one does not assume $a(i, j) = 0$ for $j \neq 1$, one should accordingly try to arrange that each of the X_j, which occurs with a positive exponent, does not split in the field extension given by g. Moreover, one also tries to arrange matters so that for every $i \neq i^*$ either g_i divides g_{i^*} or g_{i^*} divides g_i; this really means that instead of only applying Embedded Resolution one also invokes Dominance.

(7). Instead of killing g_1, Zariski used differentiation arguments. But then after all the Binomial Theorem and differentiation are in essence one and the same thing.

In § 4, I shall further elucidate the nonsplitting business, and in § 5 I shall remark on the primitive case mentioned in (4).

4. Nonsplitting

Let W be a nonsingular projective algebraic variety of dimension n and let V be the normalization of W in a finite algebraic separable extension L of the function field K of W, i.e., we have a covering map $V \to W$. Let Δ be the branch locus on W, and let H be a hypersurface in W with $\Delta \subset H$. Assuming Embedded Resolution for n, we can find a composite monoidal transformation q: $W' \to W$ such that $q^{-1}(H)$ has only normal crossings. Let μ: $V' \to W'$ be the corresponding covering map and let Δ' be the branch locus on W'. Then $\Delta' \subset q^{-1}(H)$. It can be shown that if $p = 0$ (or more generally, if $V' \to W'$ is a tame covering) then $q^{-1}(H)$ does not split locally in L, i. e., if $P' \in V'$ lies above $Q' \in W'$ and H^* is any irreducible component of $q^{-1}(H)$ passing through Q', then only one irreducible component of $\mu^{-1}(H^*)$ passes through P' and it is analytically irreducible at P'. In other words, if $g = Z^m + g_1 Z^{m-1} + \ldots + g_m$ is a local equation of the covering $P' \to Q'$ and $X_1, \ldots, X_n$ are parameters at Q' such that at Q' we have $q^{-1}(H) \subset \{X_1 \ldots X_n = 0\}$, then for $j = 1, \ldots, n$ we have

that the valuation at Q' given by X_j does not split in the field extension given by g. Thus, in a sense, the same sort of thing which was achieved by Hironaka by killing g_i and by Zariski by using differentiation arguments can also be achieved by simplifying the branch locus. The idea of simplifying the branch locus to resolve (or rather, to uniformize) the singularities of V was actually used by Jung for $n = 2$ and $k =$ the field of complex numbers; recently Zariski [27] has used this idea to obtain a simpler proof of Embedded Resolution of surfaces when $p = 0$; moreover, already in 1954, Zariski [26] had proposed it as a possible method of Embedded Resolution for all n when $p = 0$. Both Jung and Zariski have used the simplifying of the branch locus only to have a nice structure for the local ring of P' and not for the nonsplitting business. However, we thus see that this Jungian method of simplifying the branch locus (i.e., transforming the discriminant into a monomial times a unit) and the Zariski-Hironaka method of transforming the coefficients into monomials times units are in essence very closely related, although they may not appear so at first sight.

Actually, in 1953, Zariski had suggested to me to study Jung's method and to see whether it could be used for resolution of singularities of a surface in $p \neq 0$. At that time I ended up by showing that in $p \neq 0$: the local Galois group above a simple point of the branch locus can be unsolvable, and a point lying above a simple point of the branch locus can be singular; so I surmised that Jung's method cannot be adapted. These examples were published in [1, 7]; there I also showed that, although the local nonsplitting above a normal crossing of the branch locus holds for tame coverings, it does not hold for nontame coverings even above a simple point of the branch locus, and I went on to comment: this "local splitting of a simple branch variety by itself" is the real reason behind the peculiarity in $p \neq 0$. Later on, I exploited [6] a similar splitting of a branch point on a curve to get results like the following: every curve in $p \neq 0$ can be projected onto the projective line so as to have only one branch point. On the other hand, in a series of papers [8] I used the nonsplitting for tame coverings to study the tame fundamental group of an algebraic variety.

Now after some ten years the circle is completed and I have reversed my belief about the nonadaptability of Jung's method for $p \neq 0$. The lesson which I have learned is this: do not stop applying quadratic transformations when the stage is reached at which the branch locus has only normal crossings; eventually you may reach a nonsplitting stage; moreover, and this is quite important, you may even reach a stable nonsplitting stage, i.e., when the nonsplitting is not destroyed by applying more quadratic transformations. This realization was forced upon me by working on the arithmetical case in which no single method seems to work by itself. The precise result which I can prove is this:

Theorem 1. If $n = 2$, $p \neq 0$, L is a Galois extension of K, and $[L : K]$ is a power of p, then there exists a composite quadratic transformation $W^* \to W$ such that for every composite quadratic transformation $W^{**} \to W^*$ we have that the total transform of H on W^{**} does not split locally in L.

Moreover:

Theorem 2. If $n = 2$, $p \neq 0$, L is a Galois extension of K, and $[L : K] = p$, then there exists a composite quadratic transformation $W^* \to W$ such that the normalization of W^* in L is Jungian (for definition see § 6).

I have also proved certain local versions of Theorems 1 and 2 in the arithmetical case. The said version of Theorem 2 is an important step in the proof of resolution of singularities of arithmetical surfaces. Also, Theorem 1 plays an important role in my proof of Embedded Resolution of algebraic surfaces.

All this leads me to pose the following conjectural supplement to Embedded Resolution.

Supplement 1. Let W be a nonsingular projective algebraic variety of dimension n, let V be the normalization of W in a finite algebraic separable extension L of the function field K of W, let Δ be the branch locus on W for the covering $V \to W$, and let H be a hypersurface in W with $\Delta \subset H$. You may assume that H has only normal crossings. Find a composite monoidal transformation $W^* \to W$ such that the total transform of H on W^* does not split locally in L. Do this in some stable sense.

It is proposed to use this as an inductive step in the general resolution problem.

Note that Theorem 1 settles Supplement 1 for $n = 2$ in a special case. Let me point out that the main step in the proof of Theorem 1 is:

Proposition. Let R be a two-dimensional regular local ring of characteristic $p \neq 0$ such that the residue field R/M of R is algebraically closed where M is the maximal ideal in R. Let L' be a Galois extension of the quotient field K' of R such that $[L' : K'] = p$. Let U be a real nondiscrete valuation of K' such that U dominates R, and U has only one extension to L'. Let R_0, R_1, R_2, ... be the unique sequence of two-dimensional regular local rings with quotient field K' such that $R_0 = R$, and R_i is the quadratic transform of R_{i-1} along U for $0 < i < \infty$. Then there exists a nonnegative integer i_0 such that for all $i \geqslant i_0$ we have that ord_{R_i} has only one extension to L'. [By ord_R we denote the discrete valuation of K' given by taking $\mathrm{ord}_R (x/y) = a - b$ for all $0 \neq x \in R$ and $0 \neq y \in R$ where a and b are the greatest integers such that $x \in M^a$ and $y \in M^b$.]

In turn, the proof of the Proposition follows from the following three lemmas; to state these lemmas we make a definition.

Definition. Given a polynomial $F(Z)$ in an indeterminate Z with coefficients in K', we say that $F(Z)$ is R-typical provided $F(Z)$ can be expressed in the form

$$F(Z) = Z^p - (\delta x^u y^v)^{p-1} Z + x^a y^b E$$

where: (x, y) is a basis of M; δ is a unit in R; u, v, a, b are nonnegative integers such that $a < p$, $b < p$, $u \neq 0$, and if $b \neq 0$ then $v \neq 0$; and E is an element in R with $E \notin xR$ such that upon letting $c = \mathrm{ord}_{\lambda(R)} \lambda(E)$, where $\lambda: R \to R/xR$ is the canonical epimorphism, we have that: $c \leqslant 1$; if $a = 0$ then $b + c \not\equiv 0 \ (p)$; and if $c = 1$ then $b = 0$.

Lemma 1. Let the assumptions be as in the Proposition. Then there exists a nonnegative integer i and a primitive element z of L' over K' such that, upon letting $F(Z)$ be the minimal monic polynomial of z over K', we have that $F(Z)$ is R_i-typical.

Lemma 2. Let R, K', L' be as in the Proposition. Assume that there exists a primitive element z of L' over K' such that, upon letting $F(Z)$ be the minimal monic polynomial of z over K', we have that $F(Z)$ is R-typical. Then ord_R has only one extension to L'.

Lemma 3. (Stability.) Let R be as in the Proposition and let $F(Z)$ be a polynomial in Z with coefficients in R such that $F(Z)$ is R-typical. Let $S_0, S_1, \ldots, S_j, R'$ be any sequence of two-dimensional regular local rings such that $S_0 = R$, S_i is a quadratic transform of S_{i-1} for $1 \leqslant i \leqslant j$, and $R' = S_j$. Then there exist elements α and β in S with $\alpha \neq 0$ such that, upon letting $F'(Z) = \alpha^{-p} F(\alpha Z + \beta)$, we have that $F'(Z)$ is R'-typical.

The proof of Lemmas 2 and 3 is quite easy. The proof of Lemma 1 is algorithmic. We remark that Lemma 1 is also the main step in the proof of Theorem 2.

5. Units cannot be neglected

Let us now consider the primitive case

$$V: Z^p - h(Y_1, \ldots, Y_n) = 0.$$

The nonsplitting business clearly has no bearing on this. Assuming Embedded Resolution for n we can achieve

$$h(Y_1, \ldots, Y_n) = h'(X_1, \ldots, X_n) X_1^{a(1)} \ldots X_n^{a(n)}$$

where $h'(0, \ldots, 0) \neq 0$, and $a(1), \ldots, a(n)$ are nonnegative integers. If at least one of the $a(i)$ is not divisible by p then we can do some-

thing. But if $a(i) \equiv 0 \ (p)$ for all i then, upon making the obvious composite monoidal transformation, we get

$$V'\colon\ Z'^{p} - h^*(X_1, \ldots, X_n) = 0$$

where

$$h^*(X_1, \ldots, X_n) = h'(X_1, \ldots, X_n) - h'(0, \ldots, 0).$$

So we achieved nothing because the order of $h^*(X_1, \ldots, X_n)$ may even be greater than the order of $h(Y_1, \ldots, Y_n)$. In other words, in $p \neq 0$ we cannot neglect the unit $h'(X_1, \ldots, X_n)$. A similar situation prevails for

$$V\colon\ Z^{p^w} - h(Y_1, \ldots, Y_n) = 0.$$

Thus we are led to another conjectural supplement to Embedded Resolution; this one cannot be formulated completely geometrically, i.e., we cannot talk of a hypersurface but we must deal with a power series, because we are not interested in a principal ideal but in a specific power series.

Supplement 2. Assume $p \neq 0$ and let $m = p^w$ where w is a positive integer. Let h be an element in the power series ring $k[[Y_1, \ldots, Y_n]]$ such that $h \notin k[[Y_1^m, \ldots, Y_n^m]]$. Find a composite monoidal transformation $q\colon B \to A_n$, where A_n is the local space of $Y_1, \ldots, Y_n$, such that at any $Q \in B$ there exist suitable parameters $X_1, \ldots, X_n$ such that upon considering h as an element in $k[[X_1, \ldots, X_n]]$ we have that

$$h = r^m + (X_1^{b(1)} \ldots X_n^{b(n)})^m h^*$$

where $b(1), \ldots, b(n)$ are nonnegative integers, and r and h^* are elements in $k[[X_1, \ldots, X_n]]$ such that $0 < (\text{order of } h^*) < m$.

Actually this is not entirely satisfactory because it is not a stable situation. One must ask for a stable situation. Here I shall not pursue this matter any further.

As such, the primitive case may not occur in practice because we can choose a separating basis, etc. However, in the separable case

$$Z^p + g_1(Y_1, \ldots, Y_n) Z^{p-1} + \ldots + g_p(Y_1, \ldots, Y_n)$$

the nonsplitting business will help only to see to it that $g_1, \ldots, g_{p-1}$ do not interfere too much. The game is still to be played with g_p. In other words, it is proposed that:

(general case) = (primitive case) + (nonsplitting).

Anyway, this is how I carry out things for surfaces.

6. Resolution for coverings

I would like to mention the following two stronger versions of the Resolution Problem which are of interest in themselves and some forms of which may very well be useful in an inductive set up for the original Resolution Problem.

Going up. Given a function field K and a finite algebraic extension L of K does there exist a nonsingular model of K whose normalization in L is Jungian? [A variety is said to be Jungian if the local ring of each point on it is Jungian in the following sense: an n-dimensional local domain R is said to be Jungian if there exists an n-dimensional regular local domain S lying above R, a basis $Y_1, \ldots, Y_n$ of the maximal ideal in S, and positive integers $m_1, \ldots, m_n$, such that $y_j^{m_j} \in R$ for $1 \leqslant j \leqslant n$; recall that a local domain S is said to lie above a local domain R if S is of the form $S = T_N$ where T is the integral closure of R in a finite algebraic extension of the quotient field of R, and N is a maximal ideal in T.]

Coming down. Given a function field K and a finite algebraic extension L of K, does there exist a Jungian model of K whose normalization in L is nonsingular?

Concerning the first question, note that, in view of Hironaka's work, the answer is yes for $p = 0$; also see Theorem 2 stated in § 4. Concerning the second question, note that the answer is not known even when $p = 0$ and we require the model of K to be only normal; some discussion of this may be found in [3, 11].

7. Concluding remark

Having praised the significance of the Binomial Theorem, let me continue in the same vein and extoll a concrete and very old-fashioned viewpoint of algebraic geometry according to which: a variety is something given by a finite number of polynomials (or power series) in several variables; to make a birational transformation means to substitute new variables for the old and see what effect this has on the original polynomials; a monoidal transformation is a particularly simple type of substitution; and so on. In other words, the interest is in polynomials, power series, and substitutions.

Needless to say that, without in any way criticizing the higher and higher abstractions of modern algebraic geometry, I am seeking converts to the algorithmic viewpoint.

Purdue University, Lafayette, Indiana, USA

REFERENCES

[1] Abhyankar S. S., On the ramification of algebraic functions, *American Journal of Mathematics*, 77 (1955), 575-592.

[2] Abhyankar S. S., On the valuations centered in a local domain, *American Journal of Mathematics*, 78 (1956), 321-348.

[3] Abhyankar S. S., Simultaneous resolution for algebraic surfaces, *American Journal of Mathematics*, 78 (1956), 761-790.

[4] Abhyankar S. S., Local uniformization on algebraic surfaces over ground fields of characteristic $p \neq 0$, *Annals of Mathematics*, 63 (1956), 491 526. Corrections: *Annals of Mathematics*, 78 (1963), 202-203.

[5] Abhyankar S. S., On the field of definition of a nonsingular birational transform of an algebraic surface, *Annals of Mathematics*, 68 (1957), 268-281.

[6] Abhyankar S. S., Coverings of algebraic curves, *American Journal of Mathematics*, 79 (1957), 825-856.

[7] Abhyankar S. S., On the ramification of algebraic functions, Part II, *Transactions of the American Mathematical Society*, 89 (1958), 310-325.

[8] Abhyankar S. S., Tame coverings and fundamental groups of algebraic varieties, Parts I to VI, *American Journal of Mathematics*, 81 (1959), 46-94 and 82 (1960), 120-178, 179-190, 341-364, 365-373, 374-388.

[9] Abhyankar S. S., Uniformization in p-cyclic extensions of algebraic surfaces over ground fields of characteristic p, *Mathematische Annalen*, 153 (1964), 81-96.

[10] Abhyankar S. S., Reduction to multiplicity less than p in a p-cyclic extension of a two dimensional regular local ring (p=characteristic of the residue field), *Mathematische Annalen*, 154 (1964), 28-55.

[11] Abhyankar S. S., Uniformization of Jungian local domains, *Mathematische Annalen*, 159 (1965), 1-43. Correction: *Mathematische Annalen*, 160 (1965), 319-320.

[12] Abhyankar S. S., Uniformization in p-cyclic extensions of a two dimensional regular local domain of residue field characteristic p, Festschrift zur Gedächtnisfeier für Karl Weierstrass 1815-1965, *Wissenschaftliche Abhandlungen des Landes Nordrhein-Westfalen*, 33 (1966), 243-317, Westdeutscher Verlag, Köln und Opladen.

[13] Abhyankar S. S., Resolution of singularities of arithmetical surfaces, Arithmetical Algebraic Geometry, 111-152, Harper and Row, New York, 1966.

[14] Abhyankar S. S., Nonsplitting of valuations in extensions of two dimensional regular local domains, *Math. Ann.*, 170 (1967), 87-144.

[15] Abhyankar S. S., An algorithm on polynomials in one indeterminate with coefficients in a two dimensional regular local domain, *Annali di Matematica Pura ed Applicata*, Serie IV, 71 (1966), 25-60.

[16] Abhyankar S. S., Resolution of singularities of embedded algebraic surfaces, Academic Press, New York, 1966.

[17] Hironaka H., Resolution of singularities of an algebraic variety over a field of characteristic zero, *Annals of Mathematics*, 79 (1964), 109-326.

[18] Jung H. W. E., Darstellung der Funktionen eines algebraischen Körpers zweier unabhängigen Veränderlichen in der Umgebung einer Stelle, *Journal für die reine und angewandte Mathematik*, 133 (1908), 289-314.

[19] Walker R. J., Reduction of singularities of an algebraic surface, *Annals of Mathematics*, 36 (1935), 336-365.

[20] Zariski O., Algebraic Surfaces, Ergebnisse der Mathematik und ihrer Grenzgebiete, 3 (1934).

[21] Zariski O., The reduction of singularities of an algebraic surface, *Annals of Mathematics*, 40 (1939), 639-689.

[22] Zariski O., Local uniformization on algebraic varieties, *Annals of Mathematics*, 41 (1940), 852-896

[23] Zariski O., A simplified proof for the reduction of singularities, *Annals of Mathematics*, 43 (1942), 583-593.

[24] Zariski O., Reduction of singularities of algebraic three dimensional varieties, *Annals of Mathematics*, 45 (1944), 472-542.

[25] Zariski O., The fundamental ideas of abstract algebraic geometry, Proceedings of the International Congress of Mathematicians, vol. II, (1950), 77-89.

[26] Zariski O., Le problème de la réduction des singularités d'un variété algébrique, *Bulletin Sci. Math.*, 78 (1954), 1-10.

[27] Zariski O., La resoluzione delle singolarita delle superficie algebriche immerse, *Rendiconti della Classe di Scienze fisiche, matematiche e naturali*, Accademia Nazionale dei Lincei, Serie 8, 31 (1962), 97-102 and 177-180.

QUELQUES PROBLÈMES DES MODULES EN GÉOMÉTRIE ANALYTIQUE COMPLEXE

ADRIEN DOUADY

I. Notations

Le corps de base est le corps **C** des nombres complexes.

Les anneaux locaux des espaces analytiques considérés peuvent avoir des éléments nilpotents. Si S est un espace analytique, on appelle *espace analytique au-dessus* de S un espace analytique X muni d'un morphisme $p: X \to S$. Si (X, p) et (X', p') sont deux espaces analytiques au-dessus de S, on appelle *morphisme au-dessus de S* tout morphisme h de X dans X' tel que $p = p' \circ h$.

Soit X un espace analytique au-dessus de S. Pour tout point $s \in S$, on appelle *fibre* de X en s et on note X_s l'espace analytique $p^{-1}(s)$. Pour tout ouvert U de S, on appelle *restriction* de X à U et on note X_U l'espace analytique $p^{-1}(U)$ au-dessus de U. Si $f: S' \to S$ est un morphisme, on note f^*X le produit fibré $S' \times_S X$, considéré comme espace analytique au-dessus de S', et on dit que f^*X se déduit de X par *changement de base*.

On dit que X est *propre* sur S si le morphisme p est propre ; c'est une condition topologique qui ne fait intervenir que l'application sous-jacente à p. On dit que X est *plat* sur S si le morphisme p est plat, c'est-à-dire si, pour tout point $x \in X$, l'anneau local $\mathcal{O}_{X,x}$ de X en x est un module plat sur l'anneau $\mathcal{O}_{S,s}$ où $s = p(x)$.

2. Le problème des modules locaux pour un espace analytique compact

Soient S un espace analytique muni d'un point de base s_0 et X_0 un espace analytique compact. On appelle *déformation* de X_0 au-dessus de S tout espace analytique X au-dessus de S, propre et plat sur S, muni d'un isomorphisme i: $X_0 \to X_{s_0}$. Soient U et U' deux voisinages ouverts de s_0 dans S, X et X' des déformations de X_0 au-dessus de U et U' respectivement; on dit que X et X' sont *localement isomorphes* s'il existe un voisinage V de s_0 dans S et un isomorphisme h de X_V sur X'_V au-dessus de V tel que $i' = h \circ i$.

On dira que X est une déformation localement semi-universelle (resp. universelle) de X_0 si, pour tout espace analytique pointé S' et pour toute déformation X' de X_0 au-dessus de S', il existe un germe de morphisme (resp. un germe de morphisme et un seul) f de S' dans S tel que X' soit localement isomorphe à $\dot{f}^*X$ pour un représentant $\dot{f}$ de f.

Conjecture 1. *Tout espace analytique compact X_0 admet une déformation semi-universelle.*

Conjecture 1'. *Si X_0 est un espace analytique compact dont le groupe des automorphismes infinitésimaux $H^0(X_0, G_{X_0})$ est réduit à 0, X_0 admet une déformation universelle.*

Ces conjectures ont été démonstrées dans le cas où X_0 est lisse, i.e. est une variété, par Kodaira—Nirenberg—Spencer [6] sous l'hypothèse $H^2(X_0, G_{X_0}) = 0$ (on trouve alors une variété de modules S lisse), puis par Kuranishi [8] sans cette hypothèse (S peut alors avoir des singularités). Dans le cas des courbes, i.e. X_0 lisse de dimension 1, la variété des modules a été étudiée par Teichmüller et L. Bers.

La question est toujours ouverte dans le cas général.

3. Le problème des modules pour les sous-espaces analytiques compacts d'un espace analytique donné

Etant donné un espace analytique X, nous appellerons famille analytique de sous-espaces analytiques compacts de X paramétrée par un espace analytique S tout sous-espace analytique Y de $S \times X$ propre et plat sur S. Si $Y \subset S \times X$ est propre et plat sur S, pour tout morphisme f: $S' \to S$, l'espace analytique f^*Y est propre et plat sur S' et s'identifie à un sous-espace analytique de $S' \times X$, donc constitue une famille analytique de sous-espaces analytiques compacts de X paramétrée par S'. On dit que Y est une *famille universelle* de sous-espaces analytiques compacts de X si, pour tout espace analytique S' et toute famille analytique Y' de sous-espaces analytiques compacts de X paramétrée par S', il existe un morphisme f: $S' \to S$ et un seul

tel que $Y' = f^*Y$. Si Y est une famille universelle, on voit, en prenant pour S' un espace réduit à un point, que l'ensemble sous-jacent à S s'identifie à l'ensemble des sous-espaces analytiques compacts de X.

Nous avons obtenu [1] le résultat suivant:

Théorème 1. *Pour tout espace analytique X, il existe une famille universelle de sous-espaces analytiques compacts de X.*

Le problème des modules pour les sous-variétés analytiques compactes d'une variété analytique donnée était beaucoup plus facile que le problème résolu par Kodaira—Nirenberg—Spencer et Kuranishi. Il était donc naturel de commencer par là l'étude des problèmes de modules pour des espaces analytiques (éventuellement avec singularités).

Pour résoudre ce problème, nous avons été amenés à sortir du cadre des espaces analytiques de dimension finie pour considérer des espaces analytiques « banachiques ». Cette méthode était déjà employée par Kuranishi [2], bien que cela n'apparaisse pas explicitement dans la première démonstration qu'il a publiée de son théorème. Les espaces analytiques banachiques ne sont que des intermédiaires, l'espace X d'où l'on part et l'espace de modules que l'on construit sont, bien entendu, de dimension finie.

4. Le problème des modules locaux pour les classes de fibrés analytiques principaux de base et de fibres données

Soit G un groupe de Lie complexe. Pour tout espace analytique X, notons G_X le faisceau des germes de morphismes de X dans G. L'ensemble $H^1(X; G_X)$ s'identifie à l'ensemble des classes de fibrés analytiques principaux de fibre G sur X. Un morphisme f: $Y \to X$ définit un morphisme $f^*G_X \to G_Y$, de faisceaux sur Y, d'où une application f^*: $H^1(X; G_X) \to H^1(Y; G_Y)$.

Soit X un espace analytique compact. Nous appellerons *famille analytique* d'éléments de $H^1(X; G_X)$ paramétrée par un espace analytique S tout élément de $H^0(S; R^1\pi_* G_{S\times X})$, où π est la projection $S \times X \to S$. Pour tout point $s \in S$, l'injection i_s: $X \to S \times X$ définit un morphisme i_s^*: $G_{S\times X} \to G_X$, d'où une application ε: $H^1(s \times X; G_{S\times X}) \to H^1(X; G_X)$. Si $\gamma \in H^0(S; R^1\pi_* G_{S\times X})$, on note γ_s le germe de γ en s; on a $\gamma_s \in (R^1\pi_* G_{S\times X})_s = H^1(s \times X; G_{S\times X})$, et on pose $\gamma(s) = \varepsilon(\gamma_s) \in H^1(X; G_X)$.

Si S est un espace analytique muni d'un point de base s_0, nous appellerons *famille analytique locale* d'éléments de $H^1(X; G_X)$ tout élément γ de $H^1(s_0 \times X; G_{S\times X})$. On dira que γ est une famille locale semi-universelle (resp. universelle) si, pour tout espace analytique S' muni d'un point de base s'_0 et tout élément $\gamma' \in H^1(s'_0 \times X; G_{S'\times X})$

tel que $\varepsilon(\gamma') = \varepsilon(\gamma)$, il existe un germe de morphisme (resp. un germe pe morphisme et un seul) f de S' dans S tel que $\gamma' = (f \times I_X)^*(\gamma)$.

Conjecture 2. *Pour tout espace analytique compact X, tout groupe de Lie G et tout élément $\gamma_0 \in H^1(X; G_X)$, il existe une famille locale semi-universelle γ d'éléments de $H^1(X; G_X)$ telle que $\varepsilon(\gamma) = \gamma_0$.*

Dans le cas où G est abélien, cette conjecture se démontre facilement, pour $G = \mathbf{C}^*$, on a mieux: il existe une famille analytique universelle paramétrée par le *groupe de Picard* de X.

Dans le cas où X est lisse, la conjecture 2 se démontre à partir d'une formule de Malgrange et Koszul [7], par des méthodes analogues à celles de Kuranishi. Des démonstrations ont été publiées indépendamment par Oniščik et Griffith.

Dans le cas général, la question est ouverte. Il est vraisemblable que, comme cela a lieu dans le cas lisse, ce problème est plus simple que le problème des modules locaux pour les espaces analytiques compacts. En un certain sens, le problème résolu par le théorème 1 revient à la détermination d'un H^0, tandis que les problèmes posés par les conjectures 1 et 2 reviendraient à la détermination d'un H^1. C'est pourquoi nous pensons qu'une démonstration de la conjecture 2 serait un pas vers celle de la conjecture 1.

Université de Nice,
Faculté des Sciences,
Nice, France

RÉFÉRENCES

[1] Douady A., Le problème des modules pour les sous-espaces analytiques compacts d'un espace analytique donné, *Annales de l'Institut Fourier*, 1966.
[2] Douady A., Le problème des modules pour les variétés analytiques compactes (d'après M. Kuranishi), Séminaire Bourbaki, n° 277, déc. 1964, Institut Henri Poincaré, Paris.
[3] Grothendieck A., Technique de construction et théorèmes d'existence en géométrie algébrique, IV: les schémas de Hilbert, Séminaire Bourbaki, n° 221, mai 1961, Institut Henri Poincaré, Paris.
[4] Grothendieck A., Techniques de constructions en géométrie analytique, IX, Quelques problèmes de modules, Séminaire Henri Cartan, 1960—1961, Exposé 16, Ecole Normale Supérieure, Paris.
[5] Kodaira K., Spencer D. C., On the deformation of complex analytic structures, *Annals of Math.*, **67** (1958), 328-460.
[6] Kodaira K., Nirenberg L. and Spencer D. C., On the existence of deformations of complex analytic structures, *Annals of Math.*, **68** (1958), 450-459.
[7] Koszul J. L. et Malgrange B., Sur certaines fibrés complexes, *Archiv der Mathematik*, **9** (1958), 102-109.
[8] Kuranishi M., On the locally complete families of complex analytic structures, *Annals of Math.*, **75** (1962), 536-577.

DEGRÉ D'INTERSECTION EN GÉOMÉTRIE DIOPHANTIENNE

A. NÉRON

1. Introduction

Soit K un corps. On notera *additivement* les valeurs absolues v de K. Autrement dit, si, pour $x \in K$, on représente par $|x|_v$ la valeur de v en x, avec la notation multiplicative habituelle, on posera $v(x) = \log |x|_v$.

On appellera *corps global* un corps K muni d'un système propre M de valeurs absolues de K au sens de [1] vérifiant la « formule du produit », i.e. tel qu'on ait $\sum_{v \in M} v(x) = 0$ quel que soit $x \in K$.

On a, en particulier, les deux exemples fondamentaux suivants de corps globaux:

(a) *Corps de nombres algébriques*: K est alors une extension algébrique de degré fini du corps $\mathbf{Q}$ des rationnels, et M est l'ensemble des valeurs absolues de K, normées de façon que

$$\begin{cases} v(x) = \log |x| \text{ pour } x \text{ réel, si } v \text{ est archimédienne} \\ v(p) = -\log p \text{ si } v \text{ prolonge la valeur absolue } p\text{-adique} \end{cases}$$

(b) *Corps de fonctions algébriques d'une variable sur un corps k* (on supposera pour simplifier que k est algébriquement clos) : K peut alors être regardé comme le corps des fonctions sur une courbe complète sans point multiple W définie sur k. On sait que toutes les valuations de K, triviales sur k, sont discrètes. On peut alors prendre pour M l'ensemble des opposées de ces valuations (chacune d'elles étant normée de façon que l'ensemble de ses valeurs soit $\mathbf{Z}$).

Rappelons la définition habituelle de la notion de *hauteur* d'un point. Soit $x = (x_0, \ldots, x_n)$ un point de l'espace projectif $\mathbf{P}_n$. Supposons d'abord x rationnel sur K (et prenons $x_i \in K$ pour tout i). On appelle hauteur de x le nombre réel

$$h(x) = \sum_{v \in M} \sup_i v(x_i).$$

Ce nombre ne dépend que de x, en vertu de la formule du produit. Plus généralement, supposons x algébrique sur K, i.e. $x_i \in K'$ pour tout i, où K' est une extension algébrique de degré fini d de K. On appelle alors *hauteur* de x le nombre réel

$$h(x) = \frac{1}{d} \sum_{w \in M'} n_w \sup_i w(x_i),$$

où w parcourt l'ensemble M' des valeurs absolues de K' prolongeant

les valeurs absolues $v \in M$, et où l'on note n_w le degré local de l'extension K'/K relativement à w. Dans le cas des corps de nombres, on montre qu'il n'existe qu'un nombre fini de points qui sont de degré borné et de hauteur bornée dans $\mathbf{P}_n$.

Si V est une variété algébrique définie sur K, et si φ: $V \to \mathbf{P}_n$ est un morphisme défini sur K, à valeurs dans $\mathbf{P}_n$, on pose $h_\varphi = h \circ \varphi$.

On dit d'autre part que deux fonctions f et g sur un ensemble E, à valeurs réelles, sont *équivalentes*, ce qu'on écrit $f \approx g$, si la différence $f(x) - g(x)$ est bornée.

L'une des propriétés essentielles des hauteurs est la suivante: h_φ ne dépend, modulo la relation $\approx$, que de la classe, pour l'équivalence linéaire, du système linéaire L sur V associé à φ. Pour cette raison, le symbole h_φ est également noté h_L, ou encore h_X, si X est un élément de L; on prolonge par linéarité la définition du symbole h_X au cas où X est un diviseur quelconque sur V, rationnel sur K.

La notion de hauteur joue, comme on sait, un rôle essentiel dans diverses questions de géométrie diophantienne, par exemple dans la méthode de descente infinie, intervenant dans la démonstration du théorème de Mordell-Weil et de ses variantes.

Cependant, on ne peut manquer d'observer le caractère artificiel de la définition de la hauteur h, et le fait que le symbole h_X est une notion «grossière», en ce sens qu'elle n'intervient que par sa classe modulo la relation $\approx$. Dans le but de mettre au point une théorie plus précise des hauteurs, il était naturel de commencer par approfondir les propriétés du symbole h_X dans le cas le plus simple, à savoir le cas (b), celui des corps de fonctions. En effet, pour $a \in V_K$, $h_X(a)$ est alors un nombre entier, qu'on peut interpréter par la théorie des intersections, au sens classique en géométrie algébrique. Dans certains cas où $V = A$ est une variété abélienne, on peut faire le calcul explicite de ce symbole pour tous les $x \in A$ rationnels sur K. Il en est ainsi, par exemple, lorsque A est la courbe générique d'un pinceau de courbes elliptiques convenable; certains résultats dans ce sens, implicitement contenus dans [5] et [6], ont été développés de façon détaillée, et complétés par d'autres analogues, dans un article récent de Manin [4]. On observe dans chaque cas que $h_X(a)$ est une fonction *quadratique* de x, à l'addition éventuelle près de termes linéaires, et de termes «périodiques».

On sait maintenant développer une théorie incluant ces derniers résultats, valable pour un corps global quelconque, et possédant, en un certain sens, le même degré de précision dans le cas des corps de nombres que dans celui des corps de fonctions.

On a, en premier lieu, le théorème fondamental suivant

Théorème 1 (caractère quadratique de la hauteur). *Soit A une variété abélienne définie sur un corps global K, et soit X un diviseur sur A, rationnel sur K. Notons $A_{\bar{K}}$ le groupe des points de A rationnels sur la clôture algébrique $\bar{K}$ de K. Il existe une forme quadratique q_X et une forme linéaire l_X sur $A_{\bar{K}}$, uniquement déterminées, à valeurs réelles, telles qu'on ait*

$$h_X \approx q_X + l_X$$

(on appelle «forme quadratique» une fonction de la forme $f(x, x)$, où f est bilinéaire).

La première démonstration de ce théorème, et de loin la plus simple, est due à Tate; pour cette démonstration, nous renvoyons à [2], [3], [4]. Une autre démonstration est donnée ci-dessous, au n°4. A titre d'application de ce théorème, on retrouve les résultats que j'avais énoncés ou conjecturés dans [7], concernant une valeur approchée du nombre des points rationnels sur A dont la hauteur admet une borne donnée (cf. [9], II, n°16, th. 6).

Compte tenu des propriétés des hauteurs, on déduit du théorème précédent que la fonction $g_X = q_X + l_X$ ne dépend que de la classe de X pour l'équivalence linéaire, et qu'elle est *birationnellement invariante* sur K. Autrement dit, g_X possède un caractère *intrinsèque*, et constitue en fait la «bonne» notion de hauteur, appelée sans aucun doute à supplanter l'ancienne notion h_X.

Mais il est possible également de généraliser l'interprétation signalée plus haut de la hauteur comme degré d'intersection. J'ai montré dans [9] que $g_X(a)$ peut, dans tous les cas, être exprimé sous forme d'une somme de termes locaux, i.e. respectivement associés aux différentes valeurs absolues $v \in M$. Désignant par A une variété abélienne définie sur K, par X un diviseur sur A, par $\mathfrak{a}$ un cycle de dimension et de degré nuls sur A, tous deux rationnels sur K, l'outil utilisé est un certain symbole local $(X, \mathfrak{a})_v$, appelé *degré d'intersection de X et $\mathfrak{a}$ relatif à v*; il s'agit d'un certain nombre réel attaché au couple $(X, \mathfrak{a})$ et à la valeur absolue v. La suite de cet exposé est essentiellement consacrée à l'énoncé de la définition de ce symbole, et à une étude de ses propriétés résumant les principaux résultats de [9]. D'une part, on peut définir ce symbole de façon axiomatique, et prouver son existence par une méthode de passage à la limite, valable pour un corps global quelconque, puis prolonger sa définition au cas d'une variété complète sans point multiple arbitraire. D'autre part, on peut interpréter ce symbole (et, en même temps, redémontrer son existence) grâce à l'introduction des *modèles minimaux* des variétés abéliennes au sens de [8] (dans le cas d'une valuation discrète) ou à celle des *fonctions thêta* (dans le cas d'une valeur absolue à l'infini, i.e. archimédienne).

L'introduction du symbole $(X, \mathfrak{a})_v$ semble constituer le point de départ d'une sorte de théorie globale des intersections valables pour des schémas sur $\mathbf{Z}$ (bien qu'à vrai dire la structure de schéma

ne soit pas seule en cause, en raison du rôle essentiel joué par les valeurs absolues à l'infini). Parmi les problèmes à résoudre, signalons en tout cas celui de l'extension de la définition du symbole pour les cycles de dimension intermédiaire.

2. Symbole $(X, \mathfrak{a})_v$ (cas d'une variété abélienne)

Dans ce n°, on considère un corps K, algébriquement clos, et muni d'une valeur absolue v.

Si V est une variété définie sur K, on désigne par V_K l'ensemble des points de V rationnels sur K, par $D(V)_K$ le groupe des diviseurs sur V rationnels sur K, par $D_a(V)_K$ (resp. $D_l(V)_K$) le sous-groupe de $D(V)_K$ composé de ceux de ses éléments qui sont algébriquement (resp. linéairement) équivalents à zéro. L'équivalence linéaire pour les diviseurs est représentée par le signe $\sim$. Si X est un diviseur sur V, l'ouvert complémentaire du support de X est noté $\mathcal{U}(X)$. Le groupe des cycles sur V qui sont de dimension 0 et rationnels sur K est noté $Z(V)_K$. L'élément $\mathfrak{a}$ de ce groupe ayant pour composants les points $a_i \in V_K$, respectivement affectés des coefficients $m_i \in Z$, est noté $\mathfrak{a} = \Sigma_i m_i (a_i)$. La somme $m = \Sigma_i m_i$ est appelée le *degré* de $\mathfrak{a}$. Le sous-groupe de $Z(V)_K$ composé de ceux de ses éléments qui sont de degré 0 est noté $Z_0(V)_K$. On dira que deux cycles sont *étrangers* si leurs supports sont sans point commun.

Soient $X \in D_l(V)_K$, et $\mathfrak{a} \in Z_0(V)_K$. Si V est complète, il existe, d'après [12], IX, 4, th. 8, coroll. 2, une fonction f sur V, définie sur K, telle que div $(f) - X$. Si $\mathfrak{a} = \Sigma_i m_i (a_i)$, l'élément $f(\mathfrak{a}) = \prod_i f(a_i)^{m_i}$ de K ne dépend que de X et de $\mathfrak{a}$, mais non de f. On le désigne par $X(\mathfrak{a})$.

Si A est une variété abélienne, et si X est un cycle sur A, le cycle déduit de X par la translation définie par $u \in A$ est noté X_u. Le transformé de X par la symétrie $x \to -x$ est noté X^-.

Théorème 2. *Soit A une variété abélienne définie sur K. A tout couple $(X, \mathfrak{a})$ composé d'un diviseur $X \in D(A)_K$ et d'un cycle $\mathfrak{a} \in Z_0(A)_K$, mutuellement étrangers, on peut, d'une et d'une seule façon, faire correspondre un nombre réel $(X, \mathfrak{a})_v$, de sorte que les conditions suivantes soient satisfaites:*

(i) *$(X, \mathfrak{a})_v$ dépend bilinéairement de X et de $\mathfrak{a}$.*

(ii) *Pour $X \sim 0$, on a $(X, \mathfrak{a})_v = -v(X(\mathfrak{a}))$.*

(iii) *$(X, \mathfrak{a})_v$ est invariant par toute translation sur A rationnelle sur K.*

(iv) *Pour $a \in \mathcal{U}(X)$ fixé, l'application λ: $\mathcal{U}(X)_K \to \mathbf{R}$ obtenue en posant $\lambda(a) = (X, (a) - (a_0))_v$ est localement bornée, i. e. est bornée sur tout sous-ensemble borné (pour la métrique définie par v) de l'ensemble $\mathcal{U}(X)_K$.*

(En abrégé: on pose $(X, \mathfrak{a})_v = -v(X(\mathfrak{a}))$[1] lorsque $X \sim 0$, et on affirme qu'il existe une et une seule manière «raisonnable» de prolonger cette définition pour $X \in D(A)_K$ quelconque).

Pour la démonstration de l'existence de $(X, \mathfrak{a})_v$, nous renvoyons à [9], ou, pour un bref résumé, à [3][2]). La méthode employée consiste à définir le symbole comme limite d'une certaine suite réelle; elle comporte l'introduction de la notion de *quasi-fonction*, constituant une version modifiée de la notion de *distribution* au sens de Weil.

Nous nous bornons ici à reproduire la démonstration de l'unicité de $(X, \mathfrak{a})_v$. Supposons qu'il existe deux symboles $(X, \mathfrak{a})_v$ et $(X, \mathfrak{a})'_v$ vérifiant les conditions du théorème, et posons $\xi(X, \mathfrak{a}) = (X, \mathfrak{a})'_v - (X, \mathfrak{a})_v$. On voit immédiatement que ξ est bilinéaire en X et $\mathfrak{a}$, et s'annule pour $X \sim 0$. Donc, si $X' \sim X$, le nombre $\xi(X', \mathfrak{a})$ ne dépend pas du choix de X', et ceci permet de définir $\xi(X, \mathfrak{a})$ même si X et $\mathfrak{a}$ ne sont pas étrangers, en faisant «bouger» X. On peut recouvrir V par un nombre fini d'ouverts de la forme $\mathcal{U}(X')$, avec $X'_i \sim X$ pour tout i. Compte tenu de la condition (iv), on en déduit que $\xi(X, (a) - (a_0))$ est une fonction bornée de a. Montrons que ξ s'annule lorsque $X \in D_a(A)_K$ et $\mathfrak{a} \in Z_0(A)_K$. En effet, supposons d'abord que $\mathfrak{a}$ appartient au noyau d'Albanese de A, i.e. qu'on a $\Sigma_i m_i a_i = 0$. Comme K est algébriquement clos, les a_i sont rationnels sur K; on se ramène, par linéarité, au cas où $\mathfrak{a}$ est de la forme $\mathfrak{a} = (a+b) - (a) - (b) + (0)$ avec a et $b \in A_K$. Compte tenu de (iii), on a alors $\xi(X, \mathfrak{a}) = \xi(X_{-b} - X, (a) - (0))$; comme $X_{-b} - X \sim 0$, on a donc bien $(X, \mathfrak{a}) = 0$, en vertu de (ii). Soit toujours $X \in D_a(A)_K$, et prenons $\mathfrak{a} \in Z_0(A)_K$ quelconque. Pour tout entier m, notons $m\delta$ l'homomorphisme de multiplication par m sur A. Le cycle $(m\delta)(\mathfrak{a}) - m\mathfrak{a}$ appartient au noyau d'Albanese, et on a donc $\xi(X, (m\delta)(\mathfrak{a})) = m\xi(X, \mathfrak{a})$. Pour $\mathfrak{a}$ fixé, le premier membre est borné quel que soit m, d'après ce qui précède. On a donc encore nécessairement $\xi(X, \mathfrak{a}) = 0$.

Pour traiter enfin le cas où $X \in D(A)_K$ est quelconque, on distingue à nouveau le cas où $\mathfrak{a}$ appartient au noyau d'Albanese, qu'on ramène à celui, plus particulier, où $\mathfrak{a}$ est de la forme $(a+b) - (a) - (b) + (0)$. On utilise cette fois la relation $X_{-b} - X \in D_a(A)_K$; pour passer de là au cas général, on introduit encore l'entier m, et on répète le raisonnement fait plus haut.

---

[1]) J'ai commis une faute de signe dans [9]: dans les prop. 3 et 5 du n° 6 du chap. III, il faut remplacer deg $(\overline{X}, \overline{\mathfrak{a}})$ par $-$ deg $(\overline{X}, \overline{\mathfrak{a}})$. Pour que la terminologie soit compatible avec celle de la théorie usuelle des intersections, il faut donc modifier la définition, et poser $(X, \mathfrak{a})_v = -v(X(\mathfrak{a}))$ comme ci-dessus, au lieu de $(X, \mathfrak{a})_v = v(X(\mathfrak{a}))$.

[2]) Dans [3], il n'était question que du cas où X est algébriquement équivalent à zéro; mais il est facile de passer de là au cas général.

3. Symbole $(X, a)_v$ (cas d'une variété complète sans point multiple quelconque)

On considère toujours un corps K, qu'on ne suppose plus ici nécessairement algébriquement clos, et une valeur absolue propre v de K.

Soit V une variété complète sans point multiple, définie sur K. Introduisons un morphisme canonique $\alpha: V \to A$ (défini sur la clôture algébrique $\bar{K}$ de K) de V dans sa *variété d'Albanese*. Il est nécessaire ici de considérer, au lieu du groupe $D(V)_K$ de tous les diviseurs sur V, rationnels sur K, le sous-groupe de ce dernier, qu'on note $\tilde{D}(V)_K$, et composé des diviseurs dont un multiple entier est linéairement équivalent à un diviseur de la forme $\alpha^{-1}(Y)$, avec $Y \in D(A)_K$. On voit facilement qu'on a toujours $D_a(V)_K \subset \tilde{D}(V)_K$. Lorsque V est une variété abélienne, ou lorsque V est une courbe de genre au moins égal à 1, on a $\tilde{D}(V)_K = D(V)_K$, mais ceci n'a pas lieu dans tous les cas: c'est faux, en particulier, pour les courbes de genre 0, car alors $\tilde{D}(V)_K$ est le groupe des diviseurs de degré nul, qui est distinct de $D(V)_K$.

Théorème 3. *A toute variété V complète sans point multiple, définie sur K et à tout couple $(X, \mathfrak{a})$ composé d'un diviseur $X \in D(V)_K$ et d'un cycle $\mathfrak{a} \in Z_0(V)_K$, mutuellement étrangers, on peut, d'une et d'une seule façon, faire correspondre un nombre réel $(X, \mathfrak{a})_v$, de sorte que les conditions suivantes soient satisfaites:*

(i) *$(X, \mathfrak{a})_v$ est bilinéaire en X et $\mathfrak{a}$.*

(ii) *Pour $X \sim 0$, on a $(X, \mathfrak{a})_v = -v(X(\mathfrak{a}))$.*

(iii) *Pour tout K-morphisme $\varphi: V' \to V$, et pour $X \in D(V)_K$, $\mathfrak{a}' \in Z_0(V')_K$, on a*

$$(\varphi^{-1}(X), \mathfrak{a}')_v = (X, \varphi(\mathfrak{a}'))_v,$$

toutes les fois que les deux membres ont un sens.

(iv) *Pour $a_0 \in U(X)_K$ fixe, l'application $\lambda: \mathcal{U}(X)_K \to \mathbf{R}$ obtenue en posant $\lambda(a) = (X, (a) - (a_0))_v$ est localement bornée, au sens introduit dans le th. 2.*

Ce théorème se ramène au précédent, par passage à la variété d'Albanese de V. On peut compléter son énoncé par les précisions suivantes:

(a) Le symbole $(X, \mathfrak{a})_v$ ainsi défini généralise celui du th. 2.

(b) Il est birationnellement invariant sur K (d'après (iii)).

(c) Il ne dépend pas du corps K, i.e. il est invariant lorsqu'on remplace K par un sous-corps et v par la valeur absolue induite sur ce sous-corps.

(d) Si K est un corps global, on a $(X, \mathfrak{a})_v = 0$ pour presque toute $v \in M$ (cf. les notations du n° 1), i.e. pour toute v n'appartenant pas à un certain sous-ensemble fini de M.

4. Symbole global $(X, \mathfrak{a})$. Lien avec la notion de hauteur

Supposons à nouveau que K est un corps global, et reprenons les notations du n°1. Pour $X \in \tilde{D}(V)_K$, et $\mathfrak{a} \in Z_0(V)_K$, mutuellement étrangers, considérons la somme

$$(X, \mathfrak{a}) = \sum_{v \in M} (X, \mathfrak{a})_v$$

Cette somme est définie, d'après la remarque (d) ci-dessus. En outre, le symbole $(X, \mathfrak{a})$ dépend bilinéairement de X et $\mathfrak{a}$, et s'annule lorsque $X \sim 0$. Ceci permet, en faisant «bouger» X, de prolonger la définition de $(X, \mathfrak{a})$ au cas où X et $\mathfrak{a}$ sont quelconques, non nécessairement étrangers.

Nous pouvons maintenant, en utilisant ce qui précède, retrouver le th. 1, et montrer que la fonction g_X intervenant dans ce dernier n'est autre que

$$g_X(a) = (X, (a) - (0)). \tag{1}$$

En effet, désignons le second membre par $g'(a) = g'_X(a)$. Nous allons d'abord montrer que g' est la somme d'une fonction quadratique et d'une fonction linéaire. Pour cela, remarquons que l'expression

$$\langle a, b \rangle_X = (X, (a+b) - (a) - (b) + (0))$$

est bilinéaire symétrique en a et b, donc que

$$g'(a) + g'(-a) = -\tfrac{1}{2}\langle a, -a \rangle_X$$

est quadratique en a. D'autre part, on a, d'après (iii),

$$g'(-a) = (X, (-a) - (0)) = (X^-, (a) - (0)).$$

On a donc

$$g'(a) = -\tfrac{1}{2}\langle a, -a \rangle_X + (X - X^-, (a) - (0)).$$

Or on sait que $X - X^-$ est algébriquement équivalent à zéro, donc linéairement équivalent à un diviseur de la forme $X_c - X$, avec $c \in A_{\bar{K}}$, et on en déduit que le second terme de cette expression est linéaire en a. Notre assertion est donc démontrée.

Il reste à prouver qu'on a $h_X \approx g'_X$. On peut supposer que A est plongée dans l'espace projectif $\mathbf{P}_n$, et que X est une section hyperplane de A. Pour $a \in A_K$, de coordonnées homogènes $a_0, \ldots, a_n$, on a $h(a) = \Sigma_v \sup_i v(a_i/a_{i_0})$, l'indice i_0 étant choisi tel que $a_{i_0} \neq 0$. Quitte à effectuer un changement de coordonnées linéaire, on peut supposer que l'origine sur A est le point de coordonnées $1, \ldots, 1$.

Pour tout i, notons X_i la section hyperplane de A obtenue en annulant la coordonnée d'indice i. On a, pour tout i,

$$v(a_i/a_{i_0}) = (X_i - X_{i_0},\ (a) - (0))_v,$$

d'où

$$h_X(a) = \sum_{v \in M} \sup_i (X_i,\ (a) - (0))_v + g'_X(a).$$

Les ouverts $\mathcal{U}(X_i)$ étant sans point commun, on déduit de la condition (iv) du th. 2 que le premier terme du second membre est borné. On a donc bien démontré le th. 1, ainsi que la formule (1).

5. Interprétation de $(X, \mathfrak{a})_v$ (cas d'une valuation discrète)

Dans ce n°, nous considérons un corps K muni d'une valeur absolue de la forme $v = -\omega$, où ω est une valuation discrète de K, normée de façon que l'ensemble de ses valeurs soit $\mathbf{Z}$. Nous désignons par $\mathbf{R}$ l'anneau de valuation correspondant, par $\mathfrak{p}$ son idéal maximal, et par K^0 le corps résiduel. Si V est une $\mathfrak{p}$-variété définie sur K, au sens de Shimura [10] (par exemple une variété projective définie sur K), on peut parler du cycle réduit $V^0 = \rho(V)$ de V (mod $\mathfrak{p}$). Considérons un point $a \in V_K$, tel que le point réduit $a^0 = \rho(a)$ soit simple sur V^0, et soit f une fonction sur V, définie sur K. On dit que X *est représenté par f en a^0* si a^0 n'appartient pas à l'ensemble réduit du support du diviseur $X - \operatorname{div}(f)$, et s'il n'appartient pas non plus au support du $\mathfrak{p}$-diviseur de f (au sens de [8], I, n°12), i.e. si f est génériquement inversible sur la composante de V^0 qui contient a^0. Si f représente X en a, le nombre entier $-v(f(a)) = \omega(f(a))$ ne dépend que de X et de a, mais non de f. On l'appelle *v-multiplicité d'intersection de X et a*, et on le note $i(X, a)$ ou $i_v(X, a)$. On note d'autre part $Z'(V)_K$ (resp. $Z'_0(V)_K$) le sous-groupe de $Z(V)_K$ (resp. de $Z_0(V)_K$) formé des cycles dont tous les composants sont rationnels sur K. Pour tout cycle $\mathfrak{a} \in Z'(V)_K$, on définit $i(X, \mathfrak{a}) = i_v(X, \mathfrak{a})$ en prolongeant par linéarité la définition précédente.

Remarque 1. Examinons le cas particulier où K est un corps de fonctions d'une variable sur un corps k, et prenons pour modèle de l'extension K/k une courbe W complète sans point multiple définie sur k (cf. n°1). On a $K = k(x)$, où x est un point générique de W sur k, et la valeur absolue v correspond à un point x^0 de W, rationnel sur k. La variété V peut être regardée comme l'élément générique d'une famille paramétrée par W. Si $\overline{V}$ est le graphe de cette famille, on a une projection canonique $\pi: \overline{V} \to W$. La variété V s'identifie à la «fibre générique» $\pi^{-1}(x)$, tandis que V^0 s'identifie à la «fibre spéciale» $\pi^{-1}(x^0)$. A toute sous-variété V' de V définie sur K (resp. à tout cycle X sur V rationnel sur K), de dimension r, il correspond une sous-variété $\overline{V}'$ de $\overline{V}$, définie sur k (resp. un cycle $\overline{X}$ sur $\overline{V}$, rationnel sur k, et n'admettant pas de composantes verticales), de dimension $r+1$, telle qu'on ait $V' = \overline{V}' \cdot V$ (resp. $X = \overline{X} \cdot V$). On voit, dans ces conditions, que $i_v(X, \mathfrak{a})$ *est le degré de la contribution des points de la fibre spéciale V^0 dans le cycle produit d'intersection $X \cdot \mathfrak{a}$*, ce dernier étant entendu au sens classique des «Foundations» de Weil.

Soit maintenant A une variété abélienne définie sur K, faiblement $\mathfrak{p}$-simple $\mathfrak{p}$-minimale au sens de [8], i.e. telle que les conditions suivantes soient satisfaites:

(a) A est faiblement $\mathfrak{p}$-simple, i.e. tout point rationnel $\mathfrak{p}$-adique de A (donc, en particulier, tout point $\in A_K$) se réduit en un point simple de A^0.

(b) A vérifie la propriété d'application universelle suivante: Si V est une $\mathfrak{p}$-variété définie sur K, toute application rationnelle $\varphi: V \to A$, définie sur K est $\mathfrak{p}$-morphique en tout point simple de $V^0 = \rho(V)$.

Rappelons que toute variété abélienne définie sur K admet un K-modèle de ce type ([8], II, th. 2). Rappelons aussi que l'ensemble G^0 des points simples de $A^0 = \rho(A)$ est alors canoniquement muni d'une structure de groupe algébrique sur le corps résiduel K^0. Dans ces conditions, quels que soient $X \in D(A)_K$ et $\mathfrak{a} \in Z'_0(A)_K$, le symbole $i_v(X, \mathfrak{a})$ est toujours défini.

Lorsque le groupe G^0 est connexe (i.e. ne possède qu'une composante), on a la formule

$$(X, \mathfrak{a})_v = i_v(X, \mathfrak{a}). \tag{2}$$

En effet, dans le cas particulier où $X \sim 0$, cela résulte trivialement des définitions; on passe facilement de là au cas général en montrant que le second membre $i_v(X, \mathfrak{a})$ vérifie toutes les conditions du th. 2.

Dans le cas où G^0 n'est pas connexe, la formule (2) n'est plus vraie en général, mais si on pose

$$(X, \mathfrak{a})_v = i_v(X, \mathfrak{a}) + j_v(X, \mathfrak{a}), \tag{3}$$

il est possible d'interpréter d'une manière simple le terme complémentaire $j(X, \mathfrak{a}) = j_v(X, \mathfrak{a})$. Pour cela, posons $\mathfrak{a} = \Sigma_i m_i (a_i)$ $(a_i \in A_K)$. Dans le cas où $X \sim 0$, X est le diviseur d'une fonction f définie sur K, et on trouve $j(X, \mathfrak{a}) = \Sigma_i m_i \nu_i$, en désignant par ν_i le coefficient, dans le $\mathfrak{p}$-diviseur de f, de la composante G^0_i de G^0 contenant a^0_i (dans le cas particulier des corps de fonctions (cf. remarque 1 ci-dessus), $j(X, \mathfrak{a})$ peut encore être interprété comme le degré du produit d'intersection $Y^0 \cdot \mathfrak{a}$, en désignant par Y^0 la contribution de la fibre spéciale dans le diviseur de la fonction $\overline{f}$ obtenue en étendant f à $\overline{V}$).

On voit ensuite que cette interprétation peut être prolongée de façon naturelle au cas où X est un élément quelconque de $D(A)_K$. On observe en même temps que $j(X, \mathfrak{a})$ est un nombre rationnel, et que c'est une fonction « périodique » de chacun des composants de $\mathfrak{a}$ (i.e. ne dépendant que de la classe de chacun d'eux modulo un sous-groupe d'indice fini du groupe de Mordell-Weil).

Dans le cas d'un corps de fonctions algébriques d'une variable, on en déduit que le symbole $(X, \mathfrak{a})$ (sur une variété V complète sans point multiple quelconque) est toujours un nombre *rationnel*, dont le dénominateur ne dépend pas en outre du couple $(X, \mathfrak{a})$, mais seulement de V et de K. Il en résulte en particulier que le nombre $g_X(a)$ défini au n°1 est alors aussi un nombre rationnel, dont la valeur ne dépend pas du couple (X, a).

Remarque 2. On peut, sous certaines hypothèses, considérer, dans les définitions qui précèdent, des cycles $\mathfrak{a} \in Z_0(V)_K$ n'appartenant pas nécessairement à $Z'_0(V)_K$, i.e. ayant des composants algébriques (non nécessairement rationnels) sur K (cf. [9], III). Les résultats exposés ci-dessus suffisent cependant pour l'interprétation que nous avions en vue : pour $\mathfrak{a}$ donné, on peut en effet agrandir K de sorte que les composants de $\mathfrak{a}$ deviennent rationnels sur K.

Remarque 3. Il existe différentes situations, concernant les variétés complètes sans point multiple, et dans lesquelles la formule (1) ci-dessus est encore valable ([9], III, 5, th. 3).

6. Interprétation de $(X, \mathfrak{a})_v$ (cas d'une valeur absolue archimédienne). Lien avec les fonctions thêta

On peut alors supposer que K est le corps des nombres complexes $\mathbf{C}$. Toute variété abélienne A définie sur $\mathbf{C}$ peut être identifiée à un tore, i.e. à un quotient de la forme $\mathbf{C}^n/\Delta$, où Δ est un sous-groupe discret maximal de $\mathbf{C}^n$. Notons alors μ l'application canonique $\mathbf{C}^n \to A$. On sait que, pour tout diviseur $X \in D(A)_{\mathbf{C}}$, on peut trouver une fonction thêta θ sur $\mathbf{C}^n$ admettant pour diviseur $\mu^{-1}(X)$. On sait en outre (cf. [13], IV, 5) qu'il existe une forme hermitienne H sur $\mathbf{C}^n$ telle que, pour $u \in \mathbf{C}^n$, l'expression

$$\psi(u) = \log|\theta(u)| - H(u, \bar{u})$$

soit invariante par Δ. Il existe donc une et une seule application $\varphi : A_{\mathbf{C}} \to \mathbf{R}$ telle que $\varphi \circ \mu = \psi$. Par linéarité, on peut étendre φ à une application $\varphi^* : Z(A)_{\mathbf{C}} \to \mathbf{R}$. Pour $\mathfrak{a} \in Z_0(A)_{\mathbf{C}}$, on voit de plus que le nombre réel $\varphi^*(\mathfrak{a})$ ne dépend que de X, et de $\mathfrak{a}$, mais non du choix de θ. On peut désigner ce nombre par $(X, \mathfrak{a})^*$. Les propriétés des fonctions thêta permettent de vérifier que ce symbole vérifie toutes les conditions du th. 2. On a donc dans ce cas $(X, \mathfrak{a})_v = (X, \mathfrak{a})^*$.

Université de Paris,
Faculté des sciences d'Orsay, France

RÉFÉRENCES

[1] Lang S., Diophantine geometry, Interscience Tracts, New York, 1959.
[2] Lang S., Diophantine approximation on toruses, *Amer. J. Math.*, 86 (1964), 521-523.
[3] Lang S., Conférence au Séminaire Bourbaki, n° 274, mai, 1964.
[4] Манин Ю. И. *Известия АН СССР*, 28, № 6 (1964), 1363-1390.
[5] Néron A., Une propriété des faisceaux linéaires de courbes de genre 1, *C. R. Acad. Sc. Paris*, mai, 1948.
[6] Néron A., Un théorème sur le rang des courbes algébriques dans les corps de degré de transcendance fini, *C. R. Acad. Sc. Paris*, mars, 1949.
[7] Néron A., Valeur asymptotique du nombre des points de hauteur bornée sur une courbe elliptique, Int. Congr. of Math., Edinburgh, 1958.
[8] Néron A., Modèles minimaux des variétés abéliennes sur les corps locaux et globaux, Publ. Inst. Hautes Et. Scientifiques, 21 (1964).
[9] Néron A., Quasi-fonctions et hauteurs sur les variétés abéliennes, *Ann. of Math.*, 82, 2 (1965), 249-331.
[10] Shimura G., Reduction of algebraic varieties with respect to a discrete valuation of the basic field, *Amer. J. Math.*, 77 (1955), 134-176.
[11] Weil A., Arithmetic on algebraic varieties, *Ann. of Math.*, 78, 5 (1956), 412-444.
[12] Weil A., Foundations of algebraic geometry, Amer. Math. Soc. Colloquium Publ., 29 (1962), 2e éd.
[13] Weil A., Variétés kählériennes, *Act. Sc. et Ind.*, n° 1267 (1958). Русский перевод: Вейль А., Введение в теорию кэлеровых многообразий, ИЛ, М., 1961.

RATIONAL SURFACES AND GALOIS COHOMOLOGY

Yu. I. MANIN [1])

1. Let k be an arbitrary field (which below we shall often take to be perfect), and let V be an n-dimensional algebraic variety over k. A variety is said to be rational if $V \otimes \bar{k}$ (where $\bar{k}$ is the algebraic closure of k) is birationally equivalent to the projective space $P^n_{\bar{k}}$. In other words, V is geometrically reduced and geometrically irreducible, and the field of rational functions $R(V \otimes \bar{k})$ is a purely transcendental extension of the field of constants $\bar{k}$.

For the most part we shall be interested in those properties of rational varieties which are invariant under a birational transforma-

[1]) English translation by S. H. Gould, American Mathematical Society.

tion over the ground field. To within this equivalence relation, the classification of rational curves is as follows. In each class there exists a proper regular model, isomorphic to a conic on a plane. A conic is birationally trivial if and only if it contains a k-point. The set of classes of rational curves is in one-to-one correspondence with the set of quaternion algebras over the field k. The splitting fields of a given quaternion algebra consist of those fields over which there is a point on the corresponding conic. In particular, the Minkowski-Hasse principle is valid for rational curves over number fields.

The next case is that of surfaces. The picture here is considerably more complicated and our information about rational surfaces is far from complete. The present report will discuss both known facts and unsolved problems.

The study of rational surfaces is motivated by problems in the theory of numbers and particularly in algebraic geometry.

The classical problem of representing a rational number as the sum of three cubes of rational numbers leads to the rational surface $x_0^3 + x_1^3 + x_2^3 = ax_3^3$. Geometric arguments enable us to find on the surface an infinite number of rational points, namely a family of them depending on two rational parameters (but of course, this family does not contain all the points). More generally, an arbitrary nonsingular cubic surface in P_Q^3 is rational; how are we to determine whether there exists a Q-point on it and how are we to describe the set of all Q-points?

In this case, the geometric study of the problem will enable us to make some progress.

If the ground field k is a function field, let us say the field of rational functions of one variable, the rational surfaces over k lead to an interesting class of three-dimensional algebraic varieties; namely, those varieties for which there exists a rational mapping $f\colon V \to C$ onto a rational curve C and the general fiber f is a rational surface. It is natural to study such varieties in connection with the problem of unirationality: a cubic hypersurface in P^4 is of this type. Among the varieties that can be explicity determined, the unirational ones are certainly included. But are all such varieties unirational?

The varieties V of such a type are also remarkable for their topology (either usual or Grothendieck). The only interesting group of cohomologies for them is the three-dimensional one, and the three-dimensional cycles are of "one-dimensional source", which allows us, for example, to calculate the zeta functions of such varieties (cf. E. Bombieri [1]).

The cycles of one-dimensional source occur on an arbitrary proper variety V if on it we carry out a monoidal transformation with center on a nonsingular curve of genus $\geqslant 1$. The study of such cycles appears to be of great importance for the problem of unirationality, but I shall not deal with them any further here.

2. Let us fix the ground field k. We shall be interested essentially in the following two questions: the problem of k-birational classification of rational surfaces, and the problem of determining k-points on them. The results are considerably more complete for the first problem than for the second. If the field k is perfect, the Galois theory allows us to give a standard cohomological description of the set of classes of rational surfaces to within birational equivalence over k. For let Cr be the Cremona group, i.e., the group of $\bar{k}$-automorphisms of the field $\bar{k}(x, y)$. The Galois group of the closure $\bar{k}/k$ acts in an obvious way on Cr. The set of classes of interest to us here can be naturally identified with $H^1(k, \mathrm{Cr})$. This reformulation is of little value for the classification of surfaces, but it is interesting because it allows us to restate the results of such a classification in terms of the little studied group Cr. For example, it turns out that even for a finite field k the set $H^1(k, \mathrm{Cr})$ is infinite. For algebraic groups of coefficients over finite k it is well known to be trivial (Lang [11]), but Šafarevič [12] showed that it is also trivial if instead of the group of automorphisms of the field $\bar{k}(x, y)$ we consider the group of the automorphisms of the ring $\bar{k}[x, y]$.

3. The determination of the set $H^1(k, \mathrm{Cr})$ consists of the following two distinct steps.

In the first step we begin with a given rational surface F and reduce it by birational transformations to a certain standard form. The fundamental method here is the classical "method of adjunction", which will be described briefly below. The second step consists of the birational classification of these standard forms.

In order to describe the method of adjunction, it is convenient to extend the problem slightly. Along with the cohomologies we will study the finite Abelian subgroups of the Cremona group.

We consider objects of one of the following two types:

a. *A rational regular proper surface F over an algebraically closed field k together with a finite Abelian group G acting on F as its group of automorphisms (the geometric case).*

b. *A rational regular proper surface F over a perfect field k. Let G be the Galois group of $\bar{k}/k$; then G acts on the surface of $F \otimes_k \bar{k}$ through the second factor (the algebraic case).*

Objects of this type will be called G-surfaces, where G and k are considered as fixed.

The definitions of a G-morphism and of a rational G-mapping are obvious (in the algebraic case they are the morphisms and rational mappings of the k-schemes).

The G-surface will be called G-minimal if an arbitrary birational G-morphism of it is an isomorphism.

For every G-surface there exists a birational morphism of it onto a G-minimal surface. In order to verify whether F is G-minimal we must consider how G acts on the system of exceptional curves of first kind on $F \otimes \overline{k}$. On a minimal surface there are no G-orbits of this system consisting of a collection of nonintersecting curves; if such a G-orbit exists, it can be retracted by a birational G-morphism. For example, on a nonsingular cubic surface the exceptional curves of first kind consist of 27 straight lines on the surface. B. Segre [6] noted the importance of considering the nonconnected sets of conjugate straight lines. In fact, we are dealing here with a general principle that allows us to simplify and generalize several of Segre's results.

Let us first consider the simplest case, with $G = \{1\}$ (which means that the field k is algebraically closed). Then every minimal surface, except the plane, is a fibering with base P^1 and with the projective line for fiber; all such fiberings are classified by their unique integer-valued invariant, namely the Chern number (see Nagata [5]). Thus there is a countable set of minimal surfaces here although all of them are, of course, birationally equivalent.

The general result, which goes back to Enriques, is as follows.

Let us denote by $N(F)$ the group $\mathrm{Pic}(F \otimes \overline{k})$ of classes of invertible bundles on $F \otimes \overline{k}$. This group is a free group with a finite number of generators, and it is obvious that G acts on it. Let $\Omega_F \in N(F)$ be the canonical class; Ω_F is G-invariant. We assume that in the algebraic case $P(F) = \mathrm{Pic}\, F$ and in the geometric case that $P(F)$ is the group of classes of G-invariant divisors.

Theorem 1. *Every G-surface is birationally G-equivalent to a G-surface F of one of the following types.*

1. *The rank of the group $P(F)$ is equal to unity, the bundle Ω_F^{-1} is ample and on the surface $F \otimes \overline{k}$ an arbitrary irreducible curve with negative index of self-intersection is an exceptional curve of first kind (nondegenerate case).*

The rank of the group $P(F)$ is greater than unity, on the surface $F \otimes \overline{k}$ there exist rational curves with index of intersection -2, and the bundle Ω_F^{-1} is ample on the complement of these curves (the degenerate case).

2. *The rank of the group $P(F)$ is equal to two; there exists a G-morphism $f\colon F \to C$ where C is a regular G-curve of genus zero over k; the general fibre f is geometrically reduced, is geometrically irreducible and is a curve of genus zero.*

G-surfaces determined in this way will be called standard. The surfaces of the first family will also be called surfaces of del Pezzo (nondegenerate and degenerate, respectively), and the surfaces of the second family will be called surfaces with a bundle of rational curves.

The plane and a cubic surface belong to the first family; $\overline{k}$-minimal fiberings belong to the second class.

Let us note that this result is somewhat weaker than the one obtained for the case $G = \{1\}$: here we do not assert that all G-minimal surfaces are standard.

Let us say just a word about the proof (for the details see [2]). Every G-surface, as is easily seen, has an ample bundle $L \in P(F)$. If the rank of $P(F)$ for the G-minimal surface F is equal to unity, it easily follows that the bundle Ω_F^{-1} is ample.

But if the rank of $P(F)$ is greater than unity, then we can find a very ample bundle $L \in P(F)$ which is linearly independent of Ω_F. A study of the zeros of G-invariant sections of the bundle $L \otimes \Omega_F^n$ with suitably chosen n then enables us to find on F a G-invariant curve of arithmetic genus zero, whereupon a G-pencil of such curves can be constructed without particular difficulty. The construction of such a curve can fail only if $(\Omega_F \cdot \Omega_F) > 0$; an analysis of this case shows that then the bundle Ω_F^{-1} is ample on the complement of a finite number of curves with index -2, which are retracted to singular points on an anticanonical model of the surface F. Moreover, it is possible to show that G-invariant singular points cannot exist in this case. The presence of singular points justifies the name "degenerate case".

In general the argument proceeds in the same way as for the Kodaira's proof of the criterion of rationality (see Serre [7]), and to some extent it clarifies the structure of the latter. Moreover, this method also enables us to obtain a complete classification of minimal surfaces for $G = \{1\}$.

4. Before proceeding further with our birational classification of the standard G-surfaces, let us introduce some information about their structure.

The surfaces of del Pezzo. We confine our attention here to the nondegenerate case.

Since on these surfaces the bundle Ω_F^{-1} is ample, we have $(\Omega_F \cdot \Omega_F) > 0$. For rational surfaces in general $(\Omega_F \cdot \Omega_F) \leqslant 9$; thus the number $(\Omega_F \cdot \Omega_F)$, which we shall call the degree of the del Pezzo surface F, may take any value from 1 to 9.

For $3 \leqslant (\Omega_F \cdot \Omega_F) \leqslant 9$ the bundle Ω_F^{-1} is very ample.

For $(\Omega_F \cdot \Omega_F) = 2$ the mapping determined by the bundle Ω_F^{-1} is of degree 2.

For $(\Omega_F \cdot \Omega_F) = 1$ the mapping determined by the bundle Ω_F^{-1} has one base point. By making a monoidal transformation with center

at this point, we obtain a fibering by elliptic curves over the projective straight line with k-section.

Let F be a surface of del Pezzo of degree n. There exists a k-morphism (not necessarily a G-morphism) $F \otimes \bar{k} \to P^2$, which is inverse to a monoidal transformation with center at $9-n$ closed points (the case $n=8$ is exceptional; here the surface F may itself be a minimal k-form $P^1 \times P^1$).

Nagata [5] gave a description of the irreducible exceptional curves of first kind on such surfaces. Here we shall confine ourselves to giving the following table, in which the first line gives the degree of the del Pezzo surface and the second line gives the number of exceptional curves on it:

9	8	7	6	5	4	3	2	1
0	0	or 1	3	6	10	16	27	56	240

Surfaces with rational pencil. The basic geometric fact here lies in the description of the degenerate fibers of G-morphism $f: F \to C$. Namely, every such fiber is the union of two irreducible exceptional curves of the first kind with index of intersection equal to unity. The two components belong to the same G-orbit.

5. Some simple corollaries of Theorem 1.

Corollary 1. An arbitrary rational surface over k either coincides up to birational equivalence with one of the surfaces of del Pezzo or can be defined by a system of equations of the form

$$\begin{cases} A(u, v)\, x^2 + B(u, v)\, y^2 = 1 \\ au^2 + bv^2 = 1 \end{cases}$$

where $a, b \in k$; $A, B \in k(u, v)$.

It should be noted here that the majority of the surfaces of del Pezzo can also be birationally defined by simple equations: for example, a surface of the fourth degree is the intersection of two hyperquadrics in four-dimensional projective space.

Corollary 2. An arbitrary Abelian finite subgroup of the Cremona group either acts biregularly on del Pezzo surfaces or is conjugate to a finite group which carries a tower of fields of the form $k(x, y) \supset k(x) \supset k$ into itself.

From this corollary we obtain the classical results of Cantor and Wiman (see [9]).

6. Another suggestive restatement of Theorem 1 is as follows. In the group $\mathrm{Cr} = \mathrm{Aut}\, \bar{k}(x, y)$ let us consider the "triangular" subgroup Cr_0 consisting of the automorphisms that take the field $k(x)$ into itself. The imbedding $\mathrm{Cr}_0 \to \mathrm{Cr}$ defines a mapping

$i: H^1(k, \mathrm{Cr}_0) \to H^1(k, \mathrm{Cr})$. Also let $P \subset H^1(k, \mathrm{Cr})$ be the elements represented by minimal del Pezzo surfaces. We have

$$H^1(k, \mathrm{Cr}) = (H^1(k, \mathrm{Cr}_0)) \cup P.$$

Let us first note that the set $H^1(k, \mathrm{Cr}_0)$ is comparatively easy to describe if we begin with the exact sequence

$$1 \to \mathrm{Aut}_{\bar{k}(x)} \bar{k}(x, y) \to \mathrm{Cr}_0 \to \mathrm{Aut}\, \bar{k}(x) \to 1.$$

Essentially, an element of $H^1(k, \mathrm{Cr}_0)$ is a pair of quaternion algebras over the field k and over the corresponding field of functions of genus zero over k. The standard surface F corresponding to such a pair of algebras has degenerate fibers at points of the base where the local invariants of the functional algebra are nontrivial. The number $(\Omega_F \cdot \Omega_F) = 8$ (the number of geometric points of the base nontrivial invariants).

Our problem now consists of determining the equivalence relation on $H^1(k, \mathrm{Cr}_0) \cup P$ defined the mapping of this set into $H^1(k, \mathrm{Cr})$.

Let us first formulate the basic known facts about the equivalence of interest to us here. In the geometric case these facts provide us with necessary conditions for the conjugacy of the Abelian subgroups of Cr.

As a preliminary, let us introduce certain concepts. Let F be a G-surface and let $f: F' \to F$ be a birational morphism. This morphism induces a monomorphism of the groups $f: N(F) \to N(F')$. In the projective system of birational morphisms on F the G-morphisms form a co-final subsystem. Thus G acts on the group

$$Z(F) = \varinjlim N(F').$$

Moreover, on the group $Z(F)$ the following structures are defined:

a. The nondegenerate scalar product $Z(F) \times Z(F) \to \mathbf{Z}$, induced by the index of intersection on $N(F')$.

b. The canonical homomorphism $\Omega: Z(F) \to \mathbf{Z}$ which is defined on $N(F')$ by the formula $L \to (L \cdot \Omega_F)$.

c. The cone of nonnegative elements $Z^+(F) = \varinjlim N^+(F')$, where $N^+(F')$ are the points corresponding to nonnegative divisors. In this cone, the subset $Z^{++}(F) = \varinjlim N^{++}(F')$ is distinguished, where $N^{++}(F')$ are classes of the bundles, corresponding to nonnegative divisors, whose linear systems do not have fixed components.

An arbitrary G-rational mapping of finite degree $f: F_2 \to F_1$ determines a G-homomorphism $f^*: Z(F_1) \to Z(F_2)$ which takes nonnegative elements into nonnegative elements and $Z^{++}(F_1)$ into $Z^{++}(F_2)$.

If the mapping f is birational, then f^* is a G-isomorphism preserving all three structures: the index of intersection, the canonical element Ω and the nonnegative elements, and also preserving Z^{++}.

We can now state the basic known results on birational classification of standard models.

The first theorem is due to V. Iskovskih (unpublished).

Theorem 2. *Let F be a standard minimal G-surface with pencil of rational curves and let $(\Omega_F \cdot \Omega_F) \leqslant 0$. Then in the set $(Z^{++}(F))^G$ there exists a unique element L satisfying the conditions $(L \cdot L) = 0$, $(L \cdot \Omega) = -2$. (This element corresponds to a fiber of the bundle of rational curves.)*

From this theorem it follows that in the class of standard models under consideration, every birational isomorphism must preserve the bundle. This fact enables us to determine all isomorphic surfaces.

For example, let the ground field be finite. Then the elements of the set $H^1(k, \mathrm{Cr}_0)$ are in one-to-one correspondence with the classes of quaternion algebras over $k(x)$. Let us consider only those classes in which the number of nontrivial local invariants is $\geqslant 8$; they are acted on by the group of automorphisms of the field $k(x)$: distinct orbits of this group correspond to distinct elements of $H^1(k, \mathrm{Cr})$. In particular, the set $H^1(k, \mathrm{Cr})$ is infinite. Moreover, we obtain in this way all elements of the set $H^1(k, \mathrm{Cr})$ except for a finite number of them: in fact, it can be shown that the standard surfaces with $(\Omega_F \cdot \Omega_F) > 0$ belong, up to birational equivalence, to a finite number of algebraic families, so that in the present case there is only a finite number of them altogether.

Second example. Let us consider birational automorphisms of order two of the field $k(x, y)$ (k is algebraically closed)

$$x \longrightarrow x, \quad y \longrightarrow \varphi(x)/y,$$

where $\varphi(x) \in k(x)$ is a rational function without multiple zeros or poles and such that $|k(x) : k(\varphi(x))| \geqslant 8$. Two such automorphisms, corresponding to the functions $\varphi(x)$, $\psi(x)$, are conjugate in the Cremona group if and only if they are conjugate in the group Cr_0.

For this it is easy to find sufficient conditions in the terms of the functions φ, ψ.

The theorem of Iskovskih gives a clear picture of the birational behavior of minimal surfaces with nonpositive $(\Omega_F \cdot \Omega_F)$. For the case $(\Omega_F \cdot \Omega_F) > 0$ the picture is less clear.

The surfaces of del Pezzo (nondegenerate) of degree $n \geqslant 4$ have been studied in [2]. In particular, it is proved there that for $n \geqslant 5$ the presence of a G-invariant point is equivalent to G-birational triviality of such a surface. For $n = 4$ this statement is no longer true, as is proved in outline below (see Theorem 4). For $n \leqslant 3$ the minimal

nondegenerate surfaces of del Pezzo admit a complete birational classification.

Theorem 3. *Let F be a minimal nondegenerate G-surface of del Pezzo of degree 1, 2 or 3. Then F is not G-birationally equivalent to any G-surface with rational pencil. Furthermore, for an arbitrary G-birational mapping $g: F' \to F$, where F' is an arbitrary del Pezzo surface of degree not less than the degree of F, there exists a birational G-mapping $f: F \to F$ such that $f \circ g$ is a G-isomorphism.*

Thus it follows, in particular, that *the birational classification of cubic G-minimal surfaces coincides with their projective classification.*

As part of the proof of this theorem we also obtain a complete determination of the group of birational G-automorphisms of the surface F in terms of the set of G-points. In order not to encumber the present discussion with technical details, let us state this result here for a very special case: on a cubic minimal k-surface F there exists a rational point if and only if F admits a birational automorphism that is not an isomorphism. (Of course, the necessity of this condition is almost trivial.)

Another supplementary remark: Let us consider the field of functions on the surface $x^3 + y^3 + z^3 = 1$ and on it the automorphism $\varphi: x \to \zeta x$, $y \to y$, $z \to z$, where $\zeta^3 = 1$. This automorphism is not conjugate in the Cremona group to any automorphism of the plane or of a minimal fibering. Thus, in particular, this automorphism cannot be included in any connected algebraic subgroup of the Cremona group, since every such subgroup acts on the plane or on a minimal fibering.

In order to state the following theorem we introduce the concept of a trivial G-module, by which we mean the direct sum of a finite number of free abelian groups, each of which is induced as a G-module by a trivial module $\mathbf{Z}$ with respect to some subgroup of finite index of the group G. Two G-modules are said to be equivalent if they become isomorphic to each other after the adjunction to each of them of a trivial module.

Theorem 4. *Let the surfaces F_1, F_2 be G-birationally equivalent. Then the G-modules $N(F_1)$ and $N(F_2)$ are equivalent. Thus any algebraic invariants $N(F)$ of an equivalence class of a G-module are G-birational invariants of the surface F.*

An example of an invariant that is easily calculated and leads to nontrivial results is the group $H^1(k, N(F))$[1]. In [2], this group

[1]) During the Congress M. Artin has pointed out to me that over finite fields k the group $H^1(k, N(F))$ is isomorphic to the Brauer group of surface F (cf. [13]). For proof consider the etale cohomology Leray sequence for the map $F \otimes k \to F$ and use rationality.

is calculated for all possible del Pezzo surfaces of degree four and turns out in a number of cases to be nontrivial, a fact which will be of essential importance to us below in our discussion of the examples given by Châtelet.

In the algebraic case a very interesting invariant over a number field k is Tamagawa number of the torus with the G-module of characters $N(F)$.

Since the zeta function of the surface F is closely related to the zeta function of this torus, the birational invariance of the Tamagawa number can probably be given an arithmetical interpretation.

7. In connection with all the remarks it is interesting to investigate the action of the group G on $N(F)$.

The index of intersection $N(F)$ is not a definite form. In order to obtain a (negative) definite form, we must confine our attention to the subgroup $N_0(F) \subset N(F)$ consisting of those L for which $(L \cdot \Omega_F) = 0$.

It is obvious that we have the following G-pairing of lattices

$$N_0(F) \times N(F)/(\Omega_F) \to \mathbf{Z}.$$

This pairing is degenerate if $(\Omega_F \cdot \Omega_F) \neq 0$. Moreover, the natural mapping $N_0(F) \to N(F)/(\Omega_F)$ under the same condition is an imbedding, therefore both lattices may be assumed to lie in a space $N_0(F) \otimes R$ which is euclidean with respect to this scalar product (with opposite sign). It is in this space that the group G acting on del Pezzo's surface is naturally represented.

For G-surfaces with rational pencil it is reasonable to consider a somewhat different representation. Namely, let $L_0 \in N(F)^G$ be an element determined by a bundle of rational curves. We set $N_1(F) = \{L \in N(F) \mid (L \cdot L_0) = (L \cdot \Omega_F) = 0\}$. Here the pairing $N_1(F) \times N_1(F) \to \mathbf{Z}$ is negative definite, since G again has an orthogonal representation in the space $N_1(F) \otimes R$.

Theorem 5. *The image of G in the group of orthogonal transformations of the space $N_0(F) \otimes R$ (or $N_1(F) \otimes R$, respectively) is contained in a certain finite group which is generated by reflections and which preserves lattices $N_0(F)$, $N(F)/(\Omega_F)$ ($N_1(F)$).*

The corresponding groups of reflections can easily be described in explicit form. In particular, for del Pezzo surfaces of degree 3, 2, 1, they will be the Weyl groups of the groups E_6, E_7, E_8, respectively. Surfaces with rational pencil give Weyl groups of one of the classical series.

The proof of this theorem depends on the symmetries of the configuration of exceptional curves of first kind, whose classes generate the corresponding lattices. The lattices $N_0(F)$ and $N(F)/(\Omega_F)$ are

related to the simply connected and non-simply connected forms of the corresponding Lie group. For surfaces with rational pencil it is sufficient to consider the components of the degenerate fibers which generate $N_1(F)$.

As was shown by Todd [8], there exists a three-dimensional variety and a morphism of it onto the curve C such that the general fiber F of this morphism is a del Pezzo surface of degree 3, and the fundamental group of the base (with excised images of degenerate fibers) acts on the space $N_0(F) \otimes R$ as the full Weyl group of type E_6. It would be interesting to establish the existence of the analogous fibering for the surfaces of degrees 2 and 1; for del Pezzo surfaces of degree > 3 this existence follows from the results of Todd.

The arithmetic analogue of this question, namely what are the properties of the splitting fields of a system of exceptional curves on minimal del Pezzo surfaces of degree $\leqslant 3$ over number fields, has not yet been solved; in particular, it is not known whether the Galois group of such fields can act on a corresponding space as the full Weyl group.

In this connection we should also make the following remark. The Cremona group acts on $\mathbf{Z}(P^2)$, preserving the scalar product. Let $\mathbf{Z}_0(P^2)$ be the kernel of the mapping Ω; then Cr acts on the group $\mathbf{Z}_0(P^2)$, which is a lattice in an "infinite-dimensional" euclidean space. A theorem of Noether states that the group Cr is generated by reflections and by the projective group (which is finite-dimensional). Apparently this fact points to a possible role for infinite-dimensional groups of reflections in the study of groups of Cremona type.

8. We shall now discuss some results concerning birational properties of surfaces and the presence of k-points on them. In order to simplify our statements, we shall confine our attention to the algebraic case; as usual, we consider only standard models.

The situation is simplest with del Pezzo surfaces of degree $\geqslant 5$: the presence of a k-point is equivalent to the birational triviality of such a surface. In particular, if the field k is infinite, it follows that the presence of one k-point implies the presence of infinitely many k-points. Moreover, surfaces of degree 5 and 7 always have a k-point. Also, over a number field it is sufficient to verify the presence of a point on the surfaces of degree 9, 8, and 6, everywhere locally; from the presence of such a point follows the existence of a point in the ground field, which means that all points are determined by two parameters. In particular, there follow the results of Selmer, Segre and Swinnerton-Dyer, to the effect that the Minkowski-Hasse principle is valid for cubic fields with a set of three conjugate pairwise nonintersecting straight lines: by retracting this set of three lines, we obtain a del Pezzo surface of degree 6. (For such a surface the Minkowski-Hasse prin-

ciple can be proved thus; by excising from it all exceptional curves, we obtain the principal homogeneous space over the two-dimensional torus, and for such a space the result is known.)

For del Pezzo surfaces F of degree $\leqslant 4$ the presence of a point does not yet imply birational triviality. The simplest example is $x^3 + y^3 + z^3 = a$. If $a \notin (k^*)^3$, the surface is minimal and therefore nontrivial (by the theorem of Segre or as a result of calculating $H^1(k, N(F)) = \mathbf{Z}_3 \oplus \mathbf{Z}_3$. But the existence of a rational mapping $f\colon P^2 \to F$ means the existence of an infinite set of points. In this case I shall say that the surface F is k-unirational.

Theorem 6. *Let F be a del Pezzo surface of degree* 2, 3 *or* 4 *and assume that it contains a k-point not lying on the exceptional curves. Then the surface F is k-unirational.*

Segre [6] showed that for a cubic surface the presence of a k-point, even of one lying on the exceptional curves, implies the existence of a point outside of them. I do not know whether this result is true for degrees 2 and 4. Finally, it is sufficient to consider only minimal surfaces, so that we may assume that the k-point lies on the intersection of at least two exceptional curves.

I do not know whether del Pezzo surfaces of degree 1 and surfaces with rational pencil are necessarily unirational if they have a k-point.

Let us return to the unirational surfaces F.

The structure of the G-module $N(F)$ sometimes allows us to obtain information on the possible degrees of the rational mappings $P^2 \to F$. Namely, we have the following theorem.

Theorem 7. *Let d be the least common multiple of the periods of the groups $H^1(K, N(F))$, where K runs through the finite extensions of field k. Then the degree of any rational k-mapping $P^2 \to F$ is divisible by d.*

An example. By the Todd construction, over suitable functional fields k there exist cubic surfaces F which have k-points and are such that the Galois group of $\bar{k}/k$ acts on $N(F) \otimes R$ as the full Weyl group of type E_6. Calculation shows that in this case $d = 6$. Consequently, an arbitrary general construction of a rational mapping $P^2 \to F$ which makes use only of the existence of k-points must produce a degree for the mapping which is divisible by 6. This fact more or less explains why the classical method gives a mapping of degree 6.

9. I have already touched several times on the question of determining the set of all k-points of rational surfaces.

If the surface is birationally equivalent to the plane, the problem is thereby solved. More interesting is the case when the surface F is only unirational. Then every rational k-mapping $P^2 \to F$ produces an infinite set of k-points of the surface F depending on two rational parameters. But it turns out that no finite number of such parametrizations can produce all the rational k-points if k is a number field and F is a surface with a "sufficiently large" number of elliptic curves. In a report by the author [4] this result was proved for nonsingular cubic surfaces F, but the proof is also valid without essential changes for del Pezzo surfaces of degree 2 and 4, and also for cubic surfaces with singularities, provided they are birationally nontrivial. The proof is quite simple but makes use of deep-lying facts in the arithmetic of algebraic curves over number fields. Under these circumstances there is considerable interest in the examples constructed by Châtelet on cubic surfaces F for which a finite number of parametrizations with four parameters produces all the rational points (at least, on a set that is open in the sense of Zariski). By means of Theorem 5 we may verify that the Châtelet surfaces enumerated in Sections III-V of the report [10] are birationally nontrivial (the group $H^1(k, N(F))$ is nontrivial of period 2), so that obviously the two-parameter representations are not sufficient. The standard models of Châtelet's surfaces have $(\Omega_F \cdot \Omega_F) = 4$; they include not only the del Pezzo surfaces but also the surfaces with rational bundle. They are birationally trivial over a quadratic extension (or the compositum of two quadratic extensions) of the ground field. The invariant significance of the Châtelet constructions is not clear and the domain of their natural applicability is unknown. The question certainly deserves further study.

My final remarks concern questions of basic theoretical importance that are still unsolved.

First, we are still completely ignorant of the terms in which we ought to describe the k-points on minimal standard models of F with rational bundle and with $(\Omega_F \cdot \Omega_F) \leqslant 0$. Even if the surface is unirational, it appears that the most we can count on is a result of the Châtelet type. But I do not believe that all such surfaces with rational points are unirational; in particular, not all surfaces with rational bundle and with $(\Omega_F \cdot \Omega_F) < 0$, although I have yet not been able to find any proof. But in a certain sense these surfaces are in a great majority among rational surfaces and this question relates to a number of the simplest surfaces in the theory of numbers. Perhaps it is badly stated? When rational surfaces are studied along with certain Kummer surfaces, the results seem to show that «the set of all k-points of a surface» is a much more complicated entity than in the case of curves (where the restriction to number fields and acceptance of the Mordell conjecture produces a reasonably clear picture from the qualitative point of view).

Again, practically no study has been given to the problem of effectively calculating whether or not there exists a k-point on a given rational surface (again we restrict our attention to number fields k). The examples given by Swinnerton-Dyer and Mordell for the failure of the

Minkowski-Hasse principle for cubic surfaces do not seem to point the way toward any theory. For standard models with rational bundle even examples of this sort are unknown. Such examples are also unknown for del Pezzo surfaces of degree four, which in a certain definite sense are the simplest nontrivial arithmetical class of surfaces. Here again the Châtelet surfaces are of considerable interest. The technique now being used to study them strongly recalls the "first descent" into the theory of elliptic curves; it will be very interesting to see whether this similarity is only a superficial one.

10. In conclusion, let us state a conjecture about k-points on rational surfaces. We recall that a field k is said to be a C_1-field if an arbitrary hypersurface in P^n over k which is of degree $\leqslant n$, has a k-point.

Conjecture. An arbitrary rational (regular, proper) surface over a C_1-field k has a k-point.

Obviously it would be enough to verify this conjecture for standard models, and in fact by different reasons that it is true for all types of standard models except possibly del Pezzo surfaces of degree two and degenerate del Pezzo surfaces. If the field k is finite then no such exceptions need be made.

This conjecture is the only simple general statement we are at present able to make, in a system of facts which distinctly lack a harmony.

Математический институт им. В. А. Стеклова,
Москва, СССР

REFERENCES

[1] Bombieri E., Nuovi resultati sulle geometria di una ipersuperficie cubica a tre dimensioni, Symp. Int. Geom. Alg. (1965), Roma, 1967, 22-27.
[2] Manin Yu. I., Rational surfaces over perfect fields, Publ. Math. I.H.É.S., 30 (1966), 55-113.
[3] Манин Ю. И., Рациональные поверхности над совершенными полями, II, *Матем. сб.* **72**, № 2 (1967), 161-192.
[4] Manin Yu. I., Two theorems on rational surfaces, Symp. Int. Geom. Alg., (1965), Roma, 1967, 198-202.
[5] Nagata M., On rational surfaces. I, II, *Mem. Coll. Sci. Univ. Kyoto*, Ser. A, XXXII, № 3 (1960), XXXIII, № 2 (1960).
[6] Segre B., The rational solutions of homogeneous cubic equations in four variables, *Math. Notae Univ. Rosario*, ano II, f. 1-2, 1-68.
[7] Serre J.-P., Critère de rationalité pour les surfaces algébriques, Sem. Bourbaki, Fev. 1957.
[8] Todd J. A., On the topology of certain algebraic three fold loci, *Proc. Edinburgh Math. Soc.*, Ser. 2; Vol. 4, III (1935), 175-184.
[9] Wiman A., *Math. Annalen*, 1895.
[10] Châtelet F., Points rationnels sur quelques surfaces cubiques, Coll. Int. Clermont-Ferrand 1964 (pré-print).

[11] Lang S., Algebraic groups over finite fields, *Amer. Journ. of Math.*, **23**, № 3 (1956), 555-563
[12] Šafarevič I. R. (to be published).
[13] Grothendieck A., Le groupe de Brauer, II, Sem. Bourbaki, № 297 (1965).

ON TAMAGAWA NUMBERS

TAKASHI ONO

1. Adele geometry

Let X be an algebraic variety defined over the field of rational numbers $\mathbf{Q}$. For each valuation v ($= \infty$ or p) of $\mathbf{Q}$, we get an analytic variety X_v consisting of points of X rational over the completion $\mathbf{Q}_v$. If $v = p$, X_p contains a compact set $X_{\mathbf{Z}_p}$, $\mathbf{Z}_p$ being the integers of $\mathbf{Q}_p$. An element $x = (x_v) \in \prod_v X_v$ is called an *adele* if $x_p \in X_{\mathbf{Z}_p}$ for almost all p, i. e. for all but a finite number of p. The set X_A of all adeles becomes a locally compact space and is called the adele space of X. We identify $X_{\mathbf{Q}}$ as a subset of X_A by the diagonal imbedding. If X is quasi-affine, $X_{\mathbf{Q}}$ is discrete in X_A.

The adele geometry is the study of the pair $(X_A, X_{\mathbf{Q}})$, together with the imbedding above. Therefore, one has to define all conceivable invariants of X in terms of the pair and study relations among them or connections with other invariants of X. The Tamagawa number $\tau(X)$ is an example of such invariants which is, so far, definable only when X is a connected linear algebraic group. It is quite desirable to find the true definition of $\tau(X)$ for arbitrary X just as the Euler number is defined not only for topological groups but for any topological manifold.

Having these in mind, we shall outline some known results on algebraic groups and an interpretation of the Siegel's mean value theorem as a statement about the Tamagawa number of a certain homogeneous space.

2. Convergence property of a variety

For a variety X defined over $\mathbf{Q}$, denote by $X^{(p)}$ the reduction of X modulo p. For almost all p, $X^{(p)}$ is a variety defined over the finite field $\mathbf{F}_p = \mathbf{Z}/p$ and $X^{(p)}_{\mathbf{F}_p} \neq \varnothing$. We put $\mu_p(X) = [X^{(p)}_{\mathbf{F}_p}]/p^{\dim X}$, where $[*]$ denotes the number of elements in a finite set $*$. We shall say that X is of type (C) if the product $\prod'_p \mu_p(X)$, taken over almost all p,

is absolutely convergent. When $X = G$, a connected linear algebraic group, we see that the following three properties for G are equivalent each other: (i) G is of type (C), (ii) $\hat{G} = 0$, where $\hat{G}$ is the character module of G, (iii) $\pi_1(G)$ is finite. (We denote by $\pi_i(X)$ the i-th homotopy group of X_C.)

We shall call G *special* if it satisfies any one of the above three conditions. By the decomposition of Levi-Chevalley, G is special if and only if it is a semi-direct product of a unipotent group and a semi-simple group. The condition (ii) can be replaced by (ii)′: $H^0(X, \mathcal{O}_X^*) = G_m$, i.e. the non-existence of non-constant everywhere holomorphic never zero rational functions on X. Thus, all three conditions (i), (ii)′, (iii) make sense for any variety. It will be interesting to study their mutual relations for X. For example, the hypersurface in affine $(r+1)$-space defined by $F(X) = \sum_{i=0}^{r} a_i X_i^d - b = 0$, $a_i \neq 0$, $b \neq 0$ in $\mathbf{Q}$, has all three properties, provided $r \geqslant 3$; actually X_C is simply connected.

Suppose that X is non-singular. Let ω be a gauge form on X defined over $\mathbf{Q}$, i.e. an everywhere holomorphic never zero differential form of highest degree defined over $\mathbf{Q}$. Such a form exists if $X = G$ or if X is a hypersurface in an affine space. For each v, ω induces a measure ω_v on X_v and we have, for almost all p, $\mu_p(X) = \int_{X_{\mathbf{Z}_p}} \omega_p$. If X is of type (C), the formal product $\prod_v \omega_v$ well-defines a measure on X_A. If, in addition, X has the property (ii)′, this measure on X_A is the unique one as is seen by the product formula in $\mathbf{Q}$, and is written dX_A.

3. Special groups

Let G be a special group defined over $\mathbf{Q}$. By the argument above, there is a unique measure dG_A on G_A. Since $\hat{G} = 0$, we have $(\hat{G})_{\mathbf{Q}} = 0$ and hence, by Borel—Harish-Chandra, the number $\tau(G) = \int_{G_A/G_{\mathbf{Q}}} dG_A$ is well-defined. A. Weil has conjectured that

(W) $$\pi_1(G) = 0 \Rightarrow \tau(G) = 1.$$

This has been proved for a large part of classical groups (Weil, Tamagawa), for some exceptional groups (Demazure, Mars) and for Chevalley groups (Langlands), but is not yet completely solved.

On the other hand, the author has determined $\tau(G)$ modulo (W), the so-called relative theory, as an application of his determination of the Tamagawa number of algebraic tori. Namely, since $\pi_1(G)$ is

finite, there is a universal covering group $\tilde{G}$ of G which is also algebraic and defined over $\mathbf{Q}$, unique up to isomorphisms over $\mathbf{Q}$. Since $\pi_1(G)$ can be identified with the kernel of the covering map $\tilde{G} \to G$, we can put a $\mathfrak{g} = \mathfrak{g}(\bar{\mathbf{Q}}/\mathbf{Q})$-module structure on $\pi_1(G)$, where $\mathfrak{g}(\bar{\mathbf{Q}}/\mathbf{Q})$ is the Galois group of the full extension $\bar{\mathbf{Q}}/\mathbf{Q}$. The relative theory tells that the ratio $\tau(G)/\tau(\tilde{G})$ depends only upon the $\mathfrak{g}$-module $\pi_1(G)$. More precisely, we have

(##) $$\tau(G)/\tau(\tilde{G}) = [\widehat{\pi_1(G)}^{\mathfrak{g}}]/[Ш(\widehat{\pi_1(G)})],$$

where $\widehat{\pi_1(G)} = \mathrm{Hom}(\pi_1(G), (\bar{\mathbf{Q}})^*)$, and $Ш(*) = \mathrm{Ker}\left(H^1(\mathbf{Q}, *) \to \prod_v H^1(\mathbf{Q}_v, *)\right)$. Here, it should be noticed that the quantity on the right hand side of (##) makes sense for any X, at least when $\pi_1(X)$ is finite.

4. Special homogeneous spaces

Let (G, X) be a homogeneous space defined over $\mathbf{Q}$. In particular, (G, G) can be identified with G in an obvious way, and all definitions for (G, X) will be consistent with respect to this identification. We shall say that (G, X) is *special* if (i) $X_{\mathbf{Q}} \neq \varnothing$ and (ii) G, G_ξ ($\xi \in X_{\mathbf{Q}}$) are special, where G_ξ is the isotropy group of ξ. When that is so, X is quasi-affine (hence $X_{\mathbf{Q}}$ is discrete in X_A) and satisfies the three properties (i), (ii)′, (iii) in § 2. Since G and G_ξ are unimodular, X has a G-invariant gauge form and, by the uniqueness, X has the canonical measure dX_A which is independent of the choice of G and its action on X that make X into a homogeneous space. We see that $G_A X_{\mathbf{Q}}$ is open and closed in X_A. For any f on $G_A X_{\mathbf{Q}}$ continuous with compact support, compare the following two integrals:

$$\int_{G_A X_{\mathbf{Q}}} f\, dX_A = (*)\, \tau(G)^{-1} \int_{G_A/G_{\mathbf{Q}}} \sum_{\xi \in X_{\mathbf{Q}}} f(g\xi)\, dG_A.$$

If the ratio $(*)$ is a constant independent of f, we shall call it the Tamagawa number of (G, X) and write $\tau(G, X)$. In case $X = G$, we have $\tau(G, G) = \tau(G)$, this being nothing else than the Fubini formula for $G_A/G_{\mathbf{Q}}$. By the *mean value theorem* we mean the statement $\tau(G, X) = 1$. For simplicity, assume that the universal covering groups of G, G_ξ satisfy (W) and that G_ξ satisfies the Hasse principle. Then, we can prove, via relative theory, that the mean value theorem is the consequence of

$$\pi_1(X) = \pi_2(X) = 0.$$

This can be thought as a natural generalization of $\pi_1(G) = 0$ because $\pi_2(G) = 0$. Here are two examples: (1) $X = \Omega^n - \{0\}$, $G = SL(n)$, $n \geqslant 2$, (2) $X = \{x \in \Omega_{m,n};\ {}^t xsx = t\}$, s, t non-singular symmetric matrices in $\mathbf{Q}$, $m > n \geqslant 1$, $m - n \geqslant 3$, $G = 0^+(s)$. For those, one checks all conditions above and obtains the mean value theorem, which is the adele version of the classical theorem due to Siegel.

(The details of the paper will be published in *J. of Math. Soc. Japan*, **20**, No. 1, dedicated to Prof. S. Iyanaga).

The University of Pennsylvania,
Philadelphia, USA

DIFFERENTIABLE MANIFOLDS IN COMPLEX EUCLIDEAN SPACE

HUGO E. ROSSI

Let M be a real C^∞ manifold of dimension d, and suppose we are given a collection $\{\xi_i\}$ of complex vector fields on M. The functions annihilated by this system; that is, the solutions of the first order differential equations

$$\xi_i f = 0$$

form a ring. This ring of functions is not dense in all functions in any reasonable function space topology and will give rise to various topological complex function algebras. The function algebraist hopes to find in this situation some natural relationship with holomorphic functions of several complex variables. In particular, following the direction of some natural examples, we would like to find, say, a complex space associated to the system on M in some canonical way so that these functions admit holomorphic extensions into this space. There is good reason to expect that some such relationship exists, but up to now the results are slight.

Of course, for there to be many solutions, it is necessary to assume that the sheaf of C^∞ vector fields generated by the $\{\xi_i\}$ forms a Lie algebra sheaf (under brackets). We also impose a nondegeneracy condition: we assume that this sheaf is locally free. What is the same, we assume that the linear span of the $\xi_i(x)$ has constant (complex) dimension m along the manifold. These conditions are sufficient in the real analytic case, to insure the existence of local solutions to the system of equations

$$\xi_i f = 0.$$

In fact, we can find (locally) $k = d - m$ solutions which embed M as a submanifold of C^k. However, in the C^∞ case nothing is known about the existence of these solutions.

Conversely, if M is a real d-dimensional submanifold of C^k, with $d > k$, then at each point of M there is a nonempty maximal space of tangent vectors which is invariant under the involution defining the complex structure of C^k. This space inherits a complex structure and has dimension at least $d - k$. The nondegeneracy assumption now requires this dimension to be $d - k$ everywhere. (In this sense then, a complex submanifold is degenerate.) We refer to this distribution of subspaces by S. If $\xi_1, \ldots, \xi_m$ are complex vector fields spanning S, then the holomorphic functions of C^k satisfy (on M) the system

$$\bar{\xi}_i f = 0,$$

and this is the system of equations with which we started. We are led to this definition.

D e f i n i t i o n. A *d-dimensional almost complex manifold* is a real C^∞ manifold M together with a $2m$-dimensional subbundle S of the tangent bundle, and a C^∞ tensor field $J: S \to S$ such that $J^2 = -I$.

Let Σ be the complexification of S, and H the subbundle (in Σ) of eigenvectors with eigenvalue $\sqrt{-1}$. We call a function on M *holomorphic* if its differential annihilates $\overline{H}$. We call the given structure *integrable* if the sections of H form a Lie algebra, and *completely integrable* if in addition the sections of $\Sigma = H + \overline{H}$ form a Lie algebra. We call the structure *induced* if locally it is derived from an embedding into C^k. The Frobenius theorem is enough to guarantee that a completely integrable induced structure is (locally) a $(d - 2m)$-parameter family of complex manifolds of dimension m. This is essentially a theorem of Sommer [7]; proved first for $k = 2$ by Krzoska [3]. L. Nirenberg [6] has proved this in the abstract case (the case $d = 2m$ is of course the Newlander-Nirenberg Integrability theorem for complex structure).

Both our initial constructions had integrable structures. We shall restrict attention to these, and refer to them as submanifolds of C^k. Presumably they can be so described locally. Now, if ξ, η are vector fields in H then $[\xi, \bar{\eta}]$ need not lie in Σ. It is the extent to which these brackets fill out the rest of the tangent bundle of M which will influence the description of the class of holomorphic functions on M, and the inherent structure. For hypersurfaces (the case $d = 2m + 1$) the Levi form[1]) translates this material into complete information on the

[1]) If $M = \{\varphi \in C^n;\ \varphi(x) = 0\}$, then the form $\partial\bar{\partial}\varphi$ acts as a Hermitian form on H; this is the Levi form.

structure, but in the case of higher codimension an analogous tool does not exist. However examples (the first of which are due to Hans Lewy [5]) indicate that more than the brackets $[\xi, \eta]$ need to be studied; the higher order derived systems will be needed to obtain full information.

For hypersurfaces, there is the following theorem.

Theorem. Let M be a hypersurface in C^k, and suppose that the origin is in M. If the Levi form is identically zero, then M is a one-parameter family of complex manifolds, and the ring of holomorphic functions are just those which are holomorphic on the sheets of the family.

If the Levi form at the origin is not zero, there is a domain D in C^k lying to one side of M such that every holomorphic function on M is the boundary value of a complex analytic function defined in D.

This theorem is due essentially to H. Lewy [42] ($k = 2$) and subsequently R. O. Wells [9] ($k > 2$). It can be strengthened to the following

Theorem. Suppose f is a square integrable (with respect to surface area) function defined on M (in a suitable neighborhood of 0). These statements are equivalent.

(i) f is the L^2 limit of functions complex analytic in a neighborhood of M.

(ii) f is the radial limit a.e. of a function complex analytic in D.

(iii) $f\,\bar{\partial}\alpha = 0$ for every compactly supported $(k, k-2)$-form α.

These theorems are local; there is only this very special global theorem.

Theorem. Let M be a compact d-dimensional real manifold with an integrable almost complex structure of rank $\frac{1}{2}(d-1)$ (i.e., a hypersurface locally). Suppose the Levi form is positive definite everywhere. If the structure is real-analytic, or if there are local holomorphic functions, then the structure is globally induced. More precisely, M can be represented as the boundary of a strongly pseudoconvex domain on a normal complex space. The functions holomorphic on M are precisely the boundary values of functions holomorphic on that space.

This theorem has application to the study of isolated singularities of normal complex spaces. The structure of the singularity is determined by the almost complex structure of the boundary of a suitable neighborhood. We can study the deformations of the singularity by the deformations of the boundary structure to which we can apply the harmonic analysis due to Kohn [2].

The local theorems discussed above are proved by first picking a suitable coordinate for C^k so that a large family of circles on M becomes visible; these circles bound analytic discs which fill out the domain D. Along a certain submanifold of M the circles degenerate to points. The extension is given by the Cauchy integral around the circles, and the conditions guarantee the proper behaviour of the integral.

For submanifolds of higher dimension, a similar construction does not exist in general; however in the Lewy examples it does exist and an analogous theorem follows. One Lewy example [5] is of a 4 dimensional manifold in C^3. There the smallest Lie algebra containing the sections of Σ is *all vector fields*, although the brackets of vectors in H and $\bar{H}$ do not span everything. Connected to M is a domain D in C^3 with M lying on its boundary, such that every function holomorphic on M is the boundary value of a function complex analytic in D. Here we can take holomorphic to mean $\int f\,\bar{\partial}\alpha = 0$ for all $(k, k-2)$-$((3, 1))$-forms α of compact support.

The general situation is still vague, primarily because of the difficulty in finding a suitable parametrization. However, there seems to be more than enough of the desired circles to choose from, thanks to this theorem of Bishop [1].

Theorem. Let M be a submanifold of dimension d in C^k, with $d > k$. We may choose coordinates for C^k so that the projection into C^{d-k} is open (as a mapping from M). Then any (small) family of analytic discs in C^{d-k} may be lifted to C^k, so that the boundaries are on M. This can be done in a unique way if we specify the real parts of the values of the other coordinates at the centers of the discs.

This theorem is used to show that functions, complex analytic in a neighborhood of M, have extensions to a larger set; one which includes the images of all such lifted families where at least one disc is a point. R. O. Wells [10] has shown that the extent to which this set is larger than M is dependent on the Levi form. B. Weinstock [8] has improved the theorem by showing that the lifted families are (almost) as differentiable as the given families. We hope to find a tractable coordinate system for M by an appropriate choice of family in C^{d-k}.

Finally, there is the question, what is the uniform closure on M of the space of complex analytic functions? For hypersurfaces (locally), that is answered above. Otherwise again little else is known. Even if the dimension of M is less than the dimension of C^k, and the distribution S is empty, the question is difficult; and involves recent techniques of the theory of partial differential equations. The following theorem has recently been obtained by R. Nirenberg and R. O. Wells.

Theorem. Suppose M is a compact d-dimensional C^∞ manifold and F is a family of complex-valued real-analytic functions on M which separates points on M and has this additional property: for

all $x \in M$ there are $f_1, \ldots, f_d$ in F such that $df_1 \ldots df_d \neq 0$ at x. Then the holomorphic functions in F approximate all continuous functions.

Dept. of Mathematics,
Braudies University, USA

REFERENCES

[1] Bishop E., Differentiable manifolds in complex Euclidean space, *Duke Mathematical Journal*, 1963.
[2] Kohn J. J., Boundaries of complex manifolds, Proc. of Conference on Complex Analysis, Springer-Verlag, 1965.
[3] Krzoska J., Über die natürlichen Grenzen der analytischen Funktionen mehrerer Veränderlicher, Dissertation, Griefswald, 1933.
[4] Lewy H., On the local character of the solutions of an atypical linear differential equation in three variables and a related theorem for regular functions of two complex variables, *Am. Math.*, 64 (1956), 514-522.
[5] Lewy H., On hulls of holomorphy, *Comm. Pure and Appl. Math.*, 1960.
[6] Nirenberg L., A complex Frobenius theorem, Seminars on analytic functions, Princeton, 1957.
[7] Sommer F., Analytische geometrie in C^M, Schriftensreihe Math. Inst. Munster, 11.
[8] Weinstock B., Thesis, M. I. T., 1966.
[9] Wells R. O., On the local holomorphic hull of a real submanifold in several complex variables, *Comm. Pure Appl. Math.*, 19 (1966), 145-165.
[10] Wells R. O., Locally holomorphic sets, *Journal d'Analyse Math.*, 17 (1966), 337-345.

RIGIDITY OF QUOTIENTS OF BOUNDED SYMMETRIC DOMAINS [1]

EDOARDO VESENTINI

The aim of this lecture is to give a brief account of certain results and problems which arise in the theory of deformations of quotients of bounded symmetric domains by properly discontinuous groups of automorphisms.

The theory of uniformization of Riemann surfaces of genus > 1 shows how to construct families of discontinuous groups acting on the unit disk, with compact quotients, which depend continuously, in a non-trivial way, on certain parameters. Any attempt to construct similar examples of non-trivial continuous families of properly discontinuous groups acting on higher dimensional irreducible bounded symmetric domains fails. "One is therefore led to suspect that, if one excludes certain product spaces, it might be true that there are no families of discontinuous groups with fundamental domain which is compact or of finite volume, and depend non-trivially on a continuous parameter" [2].

[1] This work was partially supported by the European Office of Aerospace Research under Grant AF-EOAR 65-42.

Several attempts to solve this problem have been carried out in recent years using different techniques, and some progress has been made toward that goal. A first general result, concerning compact quotients, was obtained in the framework of Kodaira-Spencer's theory of deformations of compact complex structures. Using this result, A. Selberg was able to prove in full generality his conjecture for compact quotients of bounded symmetric domains, and to answer some related questions. A refinement of Kodaira-Spencer's theory coupled with facts of the geometry of pseudoconcave spaces, led to some partial results for certain properly discontinuous groups with non-compact quotients.

In this talk I shall concentrate essentially on those results stressing some differential geometric aspects of their proofs (cf. appendix), and leaving aside some quite recent (unpunblished) results obtained by A. Borel, M. S. Raghunathan, A. Selberg and by I. I. Pjateckiĭ-Šapiro, by means of algebraic considerations, about the rigidity of certain arithmetic groups.

1. Families of uniformizable structures

Let $\mathscr{V}$ and M be differentiable [3] manifolds and let $\mathfrak{F} : \mathscr{V} \to M$ be a differentiable surjective mapping of maximal rank. Assume that for every point $x \in \mathscr{V}$ there exists:

a neighborhood W of x in $\mathscr{V}$,
a neighborhood U of $\mathfrak{F}(x)$ in M,
an open set S in some numerical space $\mathbf{C}^n$,
a diffeomorphism $\varphi : U \times S \to W$, such that

a) $\mathrm{pr}_U = \mathfrak{F} \circ \varphi$;

b) if $\varphi_i : U_i \times S_i \to W_i$ $(i = 1, 2)$ are any two such diffeomorphisms, then $\varphi_2^{-1} \circ \varphi_1$ is an isomorphism of $\varphi_1^{-1}(W_1 \cap W_2)$ onto $\varphi_2^{-1}(W_1 \cap W_2)$, structure sheaves being the sheaves of germs of complex valued C^∞ functions holomorphic on the fibers of the projections pr_{U_i} $(i = 1, 2)$.

If all these conditions are satisfied, the triple $(U, \mathfrak{F}, M)$ is called a *differentiable family of complex manifolds.*

[2] This conjecture of A. Selberg [17] belongs to the more general context of higher dimensional symmetric spaces. It has been given a general answer by A. Weil [24] for discontinuous groups with compact fundamental domain.

[3] Here differentiable means always C^∞.

For any $t \in M$, $\mathfrak{F}^{-1}(t) = X_t$ has a natural structure of a complex manifold. We assume as a structural sheaf on $\mathcal{V}$ the sheaf of germs of complex valued C^∞ functions whose restrictions to the fibers of $\mathfrak{F}$ are holomorphic.

Let X_0 be a complex manifold. A differentiable family $(\mathcal{V}, \mathfrak{F}, M)$ for which there exists a point $o \in M$ and an isomorphism $i : X_0 \to \mathfrak{F}^{-1}(o)$ is called a *differentiable deformation* of X_0.

In a similar way we define holomorphic families and holomorphic deformations of complex manifolds.

The above statements generalize Kodaira-Spencer's definitions of deformations of compact complex manifolds. One can introduce the notions of equivalent, locally equivalent deformations and classes of local deformations in the usual manner of deformations theory [12].

Any deformation $(\mathcal{V}, \mathfrak{F}, M)$ of X_0 which is equivalent to the deformation $(X_0 \times M, \mathrm{pr}_M, M)$ is called a *trivial deformation*.

Definition. A deformation $(\mathcal{V}, \mathfrak{F}, M)$ of X_0 is called *rigid at infinity* if there exists a compact set $K_0 \subset X_0$ and an isomorphism

$$g: (X_0 - K_0) \times M \to \mathcal{V}$$

onto an open subset of $\mathcal{V}$, such that

$$\mathfrak{F} \circ g = \mathrm{pr}_M,$$

and that $\mathfrak{F}_{\mathcal{V} - \mathrm{Im}\, g}$ is a proper map.

Let $(\mathcal{V}, \mathfrak{F}, M)$ be a family of complex manifolds over a connected and simply connected manifold M. Let $\pi : \tilde{\mathcal{V}} \to \mathcal{V}$ be the universal covering manifold of $\mathcal{V}$. Then $(\tilde{\mathcal{V}}, \mathfrak{F} \circ \pi, M)$ is a new family of complex manifolds over M.

Let D be a complex manifold. We say that $(\mathcal{V}, \mathfrak{F}, M)$ is a family of complex manifolds uniformizable on the manifold D, if there exists an isomorphism

$$\sigma: \tilde{\mathcal{V}} \to D \times M$$

(structure sheaves being the sheaves of germs of complex valued C^∞ functions holomorphic respectively on the fibers of $\mathfrak{F} \circ \pi$ and of pr_M) satisfying the condition

$$\mathrm{pr}_M \circ \sigma = \mathfrak{F} \circ \pi.$$

We shall always assume $\mathcal{V}$ to be connected. Hence $\tilde{\mathcal{V}}$ will be connected and simply connected. By consequence D must be connected and simply connected and for each $t \in M$

$$D \times \{t\} \xrightarrow{\pi \circ \sigma^{-1}} X_t = \mathfrak{F}^{-1}(t)$$

is the universal covering of X_t.

The fundamental group $\Gamma = \pi_1(\mathcal{V})$ can be considered as the group of automorphisms of the universal covering $\pi : \tilde{\mathcal{V}} \to \mathcal{V}$. It follows from our previous remark that $\Gamma = \pi_1(X_t)$ for all $t \in M$.

In a similar way we can define *holomorphic* families of complex manifolds uniformizable on D.

Theorem 1 [2]. *Let D be a bounded domain in $\mathbf{C}^n$ and let M be the unit ball of $\mathbf{C}^m$, $M = \{t = (t_1, \ldots, t_m) \mid \sum t_\alpha \bar{t}_\alpha < 1\}$. Any holomorphic family $(\mathcal{V}, \mathfrak{F}, M)$ uniformizable on D is trivial.*

2. Manifolds uniformizable on bounded symmetric domains

Theorem 1 shows that the condition that the family of complex manifolds uniformizable on a bounded domain be a holomorphic family is a very heavy requirement. Moreover Theorem 1 does not involve any restriction on the dimension of D; hence, for example, if $(\mathcal{V}, \mathfrak{F}, M)$ is the family of curves of genus $p > 1$ over the Teichmüller space M, then the uniformizing parameter of X_t on the Poincaré unit disk cannot depend holomorphically on t.

We shall consider *differentiable* families uniformizable on a bounded domain $D \subset \mathbf{C}^n$. We will first of all assume D to be homogeneous. Furthermore we will be interested in having quotients of finite volume by properly discontinuous groups [4]. Hence [11] we must assume D to be a bounded symmetric domain of $\mathbf{C}^n$.

Let G be the (Lie) group of all (holomorphic) automorphisms of D. Let Γ be a properly discontinuous group of automorphisms of D (i.e., a discrete subgroup of G), acting freely on D and such that $\Gamma \backslash G$ has finite invariant volume.

Let $(\mathcal{V}, \mathfrak{F}, M)$ be a differentiable family of complex manifolds uniformizable on D, which is a deformation of $X_0 = D/\Gamma = \mathfrak{F}^{-1}(o)$.

It has been conjectured by A. Selberg [5] that, if D has no irreducible component of complex dimension 1, then $(\mathcal{V}, \mathfrak{F}, M)$ is locally trivial at the point $o \in M$.

a) The compact case. Under the stronger assumption that D/Γ be compact, this conjecture has been proved to be true as a consequence of the following general rigidity theorem.

Theorem 2 [7]. *Let X be a compact complex manifold, whose universal covering space is a bounded symmetric domain D. If no compo-*

[4]) Volume is meant here in the sense of measure connected with the metric determined on the quotient by the Bergman metric of D.

[5]) Cf. [17], p. 164. Actually Selberg's conjecture concerns the more general setting which has been dealt with by A. Weil in [23], [24].

nent of X has complex dimension 1, then the complex structure of X is locally rigid.

This theorem was used by A. Selberg in [17] to prove in full generality that, if D contains no irreducible component of complex dimension 1, then for any discrete subgroup Γ of G such that $\Gamma\backslash G$ is compact, there exists a representation of the group G in which all the elements of the group Γ are represented by matrices with algebraic elements.

R e m a r k. Selberg's conjecture has been established by A. Weil [24] for the quotient $\Gamma\backslash G$, where G is any connected semisimple Lie group without compact component, whose Lie algebra has no simple factor of dimension 3, and where Γ is a discrete subgroup of G such that $\Gamma\backslash G$ is compact.

b) The non-compact case. The case where $\Gamma\backslash G$ is not compact seems much more diffficult to handle.

In an attempt to shed light on Selberg's conjecture in a joint paper with Andreotti [2] we have considered a particular class of differentiable families $(\mathscr{V}, \mathfrak{F}, M)$, *rigid at infinity*, of uniformizable structures on a bounded symmetric domain D.

In order to state exactly the results of [2] we shall briefly review some background material [2].

Let X be a complex manifold of pure complex dimension n. Given a real valued C^∞ function Φ on X, we denote by $L(\Phi)$ its Levi form, locally expressed by

$$L(\Phi) = \sum \frac{\partial^2 \Phi}{\partial z^\alpha \partial \bar{z}^\beta} dz^\alpha \overline{dz^\beta}.$$

The signature of $L(\Phi)$ at a point $x_0 \subset X$ depends only on Φ and x_0. We say that the C^∞ function Φ is *strongly q-pseudoconvex* at x_0 if the Levi form $L(\Phi)$ has at least $n-q$ positive eigenvalues at the point x_0. That amounts to saying that there exists a biholomorphic imbedding $\tau : B_{n-q} \to X$ of the unit ball of $\mathbf{C}^{n-q}$

$$B_{n-q} = \{t = (t_1, \ldots, t_{n-q}) \in \mathbf{C}^{n-q} \mid \sum t_\mu \bar{t}_\mu < 1\}$$

into X, such that $\tau(0) = x_0$ and that the C^∞ function $\Phi \circ \tau$ is strongly plurisubharmonic at $t = 0$, i.e., its Levi form is positive definite at $t = 0$.

More in general we say that a *continuous* function $\Phi : X \to \mathbf{R}$ is strongly q-pseudoconvex at x_0 if the following two conditions are satisfied:

there exists a neighborhood U of x_0 in X and finitely many real valued C^∞ functions $\Phi_1, \ldots, \Phi_k$ on X such that

$$\Phi(x) = \operatorname{Sup}(\Phi_1(x), \ldots, \Phi_k(x)) \text{ for all } x \in U;$$

there exists a biholomorphic mapping $\tau : B_{n-q} \to X$ such that $\tau(0) = x_0$ and that all the C^∞ functions $\Phi_1 \circ \tau, \ldots, \Phi_k \circ \tau$ are strongly plurisubharmonic at $t = 0$.

D e f i n i t i o n. The complex manifold X *is strongly q-pseudoconcave* if there exists a compact set $K \subset X$ and a real valued continuous function Φ on X such that:

i) Φ is strongly q-pseudoconvex at each point of $X - K$;

ii) for any constant $c > \operatorname{Inf} \Phi$ the set

$$B_c = \{x \in X \mid \Phi(x) > c\}$$

is relatively compact in X.

We are now in position to state the main result of [2].

T h e o r e m 3. *Let $(\mathscr{V}, \mathfrak{F}, M)$ be a differentiable family of deformations of a complex manifold $X = \mathfrak{F}^{-1}(0)$, over the unit ball of $\mathbf{R}^m$*

$$M = \{t = (t_1, \ldots, t_m) \in \mathbf{R}^m \mid \sum t_i^2 < 1\}.$$

We assume that

α) $\mathscr{V}$ is a family of uniformizable structures on a bounded symmetric domain D, containing no irreducible components of complex dimension 1;

β) the deformation is rigid at infinity;

γ) X is strongly q-pseudoconcave with $0 \leqslant q \leqslant \dim_{\mathbf{C}} X - 2$;

δ) the fundamental group Γ of X is finitely generated.

Then the whole family $(\mathscr{V}, \mathfrak{F}, M)$ is trivial.

Both conditions γ) and δ) are satisfied when Γ is an arithmetic group. Condition δ) has been established in general in [5], while γ) has been checked in particular cases by Andreotti and Grauert [1], by Spilker [19], by K. G. Ramanathan (unpublished), and finally has been established in general by A. Borel (unpublished) (see also [16]). Furthermore, if Γ is an arithmetic group, the $\Gamma\backslash G$ has finite invariant volume [5]. These facts relate theorem 3 to the conjecture of Selberg and to some other conjectures, for a brief account of which we refer to [2, 249-250].

R e m a r k. The notion of rigidity at infinity has been carried over by H. Garland [9] to deformations of the quotient $\Gamma\backslash G$ of any connected semisimple Lie group without compact components, whose Lie algebra contains no simple factor of dimension 3, by a discrete subgroup Γ such that $\Gamma\backslash G$ has finite invariant volume.

Appendix

This appendix is devoted to some comments on the differential geometry involved in the proofs of theorems 2 and 3.

a) W-ellipticity. Let Θ be the holomorphic tangent bundle to a complex manifold. We denote by $A^{pq}(X, \Theta)$ the complex vector

space of C^∞ (p, q)-forms on X with values in Θ, and by $\mathcal{D}^{pq}(X, \Theta)$ the subspace of compactly supported forms. We assume a positive definite hermitian metric of class C^∞ on X. Using this metric we introduce a positive definite inner product $(\ ,\ )$ on $\mathcal{D}^{pq}(x, \Theta)$.

Let $\|\ \|$ be the corresponding norm: $\|\varphi\|^2 = (\varphi, \varphi)$ $(\varphi \in \mathcal{D}^{pq}(X, \Theta))$. We denote by $\mathfrak{d}$ the formal adjoint of the $\bar\partial$-operator.

We say that Θ is $W^{p,q}$-elliptic with respect to the hermitian metric on X if there exists a positive constant c such that

$$\|\varphi\|^2 \leqslant c(\|\bar\partial\varphi\|^2 + \|\mathfrak{d}\varphi\|^2) \text{ for all } \varphi \in \mathcal{D}^{pq}(X, \Theta).$$

Proposition [3]. *If Θ is $W^{p,q}$-elliptic $(q > 0)$ with respect to a complete hermitian metric on X, then for any $\varphi \in A^{pq}(X, \Theta)$, with $\bar\partial\varphi = 0$, $\|\varphi\| < \infty$, there exists a form $\psi \in A^{p,q-1}(X, \Theta)$, such that $\|\psi\| < \infty$, and*

$$\varphi = \bar\partial\psi.$$

Letting $\Omega^p(\Theta)$ be the sheaf of germs of holomorphic Θ-valued p-forms on X, the above proposition implies that the canonical image of $H^q_k(X, \Omega^p(\Theta))$ (cohomology with compact supports) into $H^q(X, \Omega^p(\Theta))$ is zero.

If X is compact that means that $H^q(X, \Omega^p(\Theta)) = 0$.

b) The linear operator Q. We assume that the hermitian metric on X is a Kähler metric. Let $R_{\alpha\bar\beta\gamma\bar\delta}$ be the components of the Riemann curvature tensor, and let $R = -2R_\alpha{}^\beta{}_\beta{}^\alpha$ be the scalar curvature. Let, for $x \in X$, $\delta(x)$ be the smallest eigenvalue of the linear transformation

$$Q\colon \xi_{\alpha\beta} \to R_\sigma{}^{\alpha\beta}{}_\tau\,\xi_{\alpha\beta} \quad (\xi_{\alpha\beta} = \xi_{\beta\alpha}).$$

Lemma [7, 487] [2, 267]. *If X is an Einstein-Kähler manifold of complex dimension n, with $R < 0$, then*

$$\delta(x) \leqslant R/n(n+1).$$

If

$$\inf_{x\in X}\left\{\frac{R}{n\delta(x)} - (q+1)\right\} > 0,$$

then Θ is $W^{0,q}$-elliptic.

Let D be a bounded homogeneous domain and let X be the quotient of D by a properly discontinuous group of automorphisms of D acting freely on D. Then the Bergman invariant metric on D defines a Kähler-Einstein metric on X, whose scalar curvature is $R = -2n$ and for which $R/n\delta$ is a constant. We denote this constant by $\gamma(D)$.

In the case where D is an irreducible bounded *symmetric* domain the eigenvalues of the linear transformation Q, and hence $\gamma(D)$, have been evaluated in [4] and in [7]: in [7] leaving aside the two exceptional domains), by means of the classification of bounded symmetric domains and by a direct inspection of the Bergman metric; in [4] by means of the structure theory of simple Lie algebras. As is was pointed out by A. Borel in [4], the operator Q has at most two different eigenvalues. It turns out that under the hypothesis α) of theorem 3, the tangent bundle Θ is $W^{0,1}$-elliptic. This fact, coupled with the rigidity criterion of Frölicher-Nijnhuis [12], yields theorem 2.

Theorem 3 requires much more work. The main steps of the proof are the following.

1. Let B be a relatively compact subset of X. One first shows that, if B is sufficiently large then the inverse image of B in D has D its envelope of holomorphy.

2. As we said before, the tangent bundle of $X_t = \mathfrak{F}^{-1}(t)$ $(t \in M)$ is $W^{0,1}$-elliptic with respect to the hermitian metric induced on X_t by the Bergman metric of D. One shows that the same is also true if the metric is perturbed on a compact set of X_t, provided that $|t|$ is sufficiently small.

3. Condition β) of theorem 3 implies that the deformation cocycles can be represented on X_t by $\bar\partial$-closed, compactly supported Θ_t-valued (0, 1)-forms. Using the proposition stated in a) one shows that for any compact set $B \subset X = \mathfrak{F}^{-1}(0)$ $(0 \in M)$ there exists a trivial deformation of B in U. This trivial deformation of B extends to a trivialization of $\mathfrak{F}$ in view of 1.

We point out that the rigidity at infinity enters steps 2 and 3 of our proof.

c) The operator Q on some bounded homogeneous domains. We shall indicate the eigenvalues of Q for all bounded non-symmetric homogeneous domains D of complex dimension 4 and 5.

These eigenvalues have been evaluated with the help of an explicit representation of D as a Siegel domain of the first or of the second kind [14], by means of direct computations on the Bergman kernel function of D. The latter has been constructed by the methods of [10], while the computations have been partially performed by an IBM-7090 electronic computer using the formal language FORMAC.

According to Pjateckiĭ-Šapiro [15] (cf. also [22]) there exists exactly one non-symmetric bounded homogeneous domain of complex dimension 4. In $\mathbf{C}^5$ there exist 6 irreducible bounded homogeneous domains, 4 of which are non-symmetric.

1) Let D be the (non-symmetric) homogeneous Siegel domain of the second kind in $\mathbf{C}^4$ defined by

$$D = \{z = (z_1, z_2, z_3, u) \in \mathbf{C}^4 \mid (y_1 - |u|^2)\,y_2 - y_3^2 > 0,\ y_2 > 0\}$$

where $y_k = \operatorname{Im} z_k$ for $k = 1, 2, 3$.

The Bergman kernel function of D is given by

$$K(z, \bar{z}) = c\, y_2 [(y_1 - |u|^2)\, y_2 - y_3^2]^{-4}$$

where c is a numerical constant. The eigenvalues of Q are

$\frac{1}{3}$, 0, $-\frac{1}{2}$ (with multiplicity 6), $-\frac{2}{3}$ (with multiplicity 2).

Hence $\delta = -\frac{2}{3}$. Since $R = -2 \dim_{\mathbf{C}} D = -8$ we get

$$\gamma(D) = 3.$$

Thus *the holomorphic tangent bundle to D is $W^{0,q}$-elliptic for* $q = 0, 1$.

2) Let D be the (non-symmetric) homogeneous Siegel domain of the second kind

$$D = \{z = (z_1, z_2, z_3, u_1, u_2) \in \mathbf{C}^5 \mid (y_1 - |u_1|^2 - |u_2|^2)\, y_2 - y_3^2 > 0,\ y_2 > 0\}.$$

The Bergman kernel function is

$$K(z, \bar{z}) = c\, y_2^2 [(y_1 - |u_1|^2 - |u_2|^2)\, y_2 - y_3^2]^{-5},$$

c being always a numerical constant.

The eigenvalues of Q are

$\frac{1}{3}$, 0 (with multiplicity 2), $-\frac{2}{5}$ (with multiplicity 10), $-\frac{2}{3}$ (with multiplicity 2).

Hence

$$\delta = -\frac{2}{3}, \quad R = -2 \dim_{\mathbf{C}} D = -10, \quad \gamma(D) = 3.$$

The holomorphic tangent bundle to D is $W^{0,q}$-elliptic for $q = 0, 1$.

3) Consider now the (non-symmetric) homogeneous Siegel domain of the second kind

$$D = \{z = (z_1, z_2, z_3, u_1, u_2) \in \mathbf{C}^5 \mid (y_1 - |u_1|^2)(y_2 - |u_2|^2) - (y_3 - \mathrm{Re}\, u_1 \bar{u}_2)^2 > 0,\ y_1 - |u_1|^2 > 0\}.$$

The Bergman kernel function is

$$K(z, \bar{z}) = c\, [(y_1 - |u_1|^2)(y_2 - |u_2|^2) - (y_3 - \mathrm{Re}\, u_1 \bar{u}_2)^2]^{-4}.$$

The eigenvalues of Q are:

$\frac{1}{4}$ (with multiplicity 3), $-\frac{1}{4}$, $-\frac{1}{2}$ (with multiplicity 11);

$$\delta = -\frac{1}{2}, \quad R = -10, \quad \gamma(D) = 4.$$

The holomorphic tangent bundle to D is $W^{0,q}$-elliptic for $q = 0, 1, 2$.

4) Let D be the (non-symmetric) homogeneous Siegel domain of the second kind, defined by

$$D = \{z = (z_1, z_2, z_3, z_4, u) \in \mathbf{C}^5 \mid (y_1 - |u|^2)\, y_2 - y_3^2 - y_4^2 > 0,\ y_2 > 0\}.$$

The Bergman kernel function is given by

$$K(z, \bar{z}) = c\, y_2 [(y_1 - |u|^2)\, y_2 - y_3^2 - y_4^2]^{-5}.$$

The eigenvalues of Q are

$\frac{1}{2}$, 0, $-\frac{2}{5}$ (with multiplicity 10), $-\frac{1}{2}$ (with multiplicity 3);

$$\delta = -\frac{1}{2}, \quad R = -10, \quad \gamma(D) = 4.$$

The holomorphic tangent bundle to D is $W^{0,q}$-elliptic for $q = 0, 1, 2$.

5) The simplest example [21] of a non-selfadjoint convex homogeneous cone is the cone V of $\mathbf{R}^5$ defined by

$$V = \{y = (y_1, y_2, y_3, y_4, y_5) \in \mathbf{R}^5 \mid y_1 y_3 - y_4^2 > 0,\ y_2 y_3 - y_5^2 > 0,\ y_3 > 0\}.$$

Let D be the homogeneous tubular domain (i.e., Siegel domain of the first kind) over V; D is defined by

$$D = \{z = (z_1, z_2, z_3, z_4, z_5) \in \mathbf{C}^5 \mid y_1 y_3 - y_4^2 > 0,\ y_2 y_3 - y_5^2 > 0,\ y_3 > 0\}$$

where $y_k = \mathrm{Im}\, z_k$ ($k = 1, \ldots, 5$). Since V is non-selfadjoint, D is non-symmetric.

The Bergman kernel function of D can be computed using [13] (or [10]). We get

$$K(z, \bar{z}) = c\, y_3^2 (y_1 y_3 - y_4^2)^{-3} (y_2 y_3 - y_5^2)^{-3}.$$

The eigenvalues of Q are

$\frac{1}{3}$, $\frac{\sqrt{10}-2}{6}$, 0 (with multiplicity 4), $-\frac{1}{2}$ (with multiplicity 4), $-\frac{2}{3}$ (with multiplicity 4), $\frac{-2-\sqrt{10}}{6}$.

Hence

$$\delta = \frac{-2-\sqrt{10}}{6}, \quad R = -10, \quad \gamma(D) = 2(\sqrt{10} - 2).$$

The holomorphic tangent bundle to D is $W^{0,q}$-elliptic for $q = 0, 1$.

Scuola Normale Superiore,
Pisa, Italy

REFERENCES

[1] Andreotti A., Grauert H., Algebraische Körper von automorphen Funktionen, *Nachr. Akad. Wissensch. Göttingen* (1961), 39-48.

[2] Andreotti A., Vesentini E., On deformations of discontinuous groups, *Acta Mathematica*, **112** (1964), 249-298.
[3] Andreotti A., Vesentini E., Carleman estimates for the Laplace-Beltrami equation on complex manifolds, *Publ. Math. I.H.E.S.*, **25** (1965), 81-130; Erratum, ibid., **27**, 153-155.
[4] Borel A., On the curvature tensor of the hermitian symmetric manifolds, *Ann. of Math.*, **71** (1960), 508-521.
[5] Borel A., Harish-Chandra, Arithmetic subgroups of algebraic groups, *Ann. of Math.*, **75** (1962), 485-535.
[6] Calabi E., On compact Riemannian manifolds with constant curvature, I, Differential Geometry, Proceedings of symposia on pure mathematics, Vol. III (1961), 155-180.
[7] Calabi E., Vesentini E., On compact, locally symmetric Kähler manifolds, *Ann. of Math.*, **71** (1960), 472-507.
[8] Фукс Б. А., Специальные главы теории аналитических функций многих комплексных переменных, изд-во «Наука», М., 1963.
[9] Garland H., On deformations of discrete groups in the noncompact case (to appear).
[10] Гиндикин С. Г., Анализ в однородных областях, *УМН*, **19**, № 4 (1964). English translation: *Russian Mathematical Surveys*, **19**, № 4 (1964), 1-89.
[11] Hano J. I., On Kählerian homogeneous spaces of unimodular Lie groups, *Amer. J. Math.*, **79** (1957), 885-900.
[12] Kodaira K., Spencer D. C., On deformations of complex analytic structures, I and II, *Ann. of Math.*, **67** (1958), 328-460.
[13] Koranyi A., The Bergman kernel function for tubes over convex cones, *Pacific Journal of Math.* (1962), 1355-1359.
[14] Пятецкий-Шапиро И. И., Геометрия классических областей и теория автоморфных функций, Физматгиз, М., 1961.
[15] Пятецкий-Шапиро И. И., О классификации ограниченных однородных областей в n-мерном комплексном пространстве, *ДАН СССР*, **141**, № 2 (1961), 316-319. English translation: *Soviet Math. Dokl.*, **2** (1962), 1460-1463.
[16] Пятецкий-Шапиро И. И., Арифметические группы в комплексных областях, *УМН*, **19**, № 6 (1964), 93-121. English translation: *Russian Mathematical Surveys*, **19**, № 6 (1964), 93-109.
[17] Selberg A., On discontinuous groups in higher dimensional symmetric spaces, Contributions to function theory, Bombay, 1960, 147-164.
[18] Siegel C. L., Analytic functions of several complex variables, Mimeographed notes, Princeton, 1962.
[19] Spilker A., Algebraische Körper von automorphen Funktionen, *Math. Ann.*, **194** (1963), 341-360.
[20] Винберг Э. Б., Однородные конусы, *ДАН СССР*, **133** (1960), 9-12. English translation: *Soviet Math. Dokl.*, **1** (1960), 787-790.
[21] Винберг Э. Б., Теория однородных выпуклых конусов, *Труды моск. матем. общества*, **12** (1963), 303-358. English translation: *Transaction Moscow Math. Soc.*, **12** (1965), 340-403.
[22] Винберг Э. Б., Гиндикин С. Г., Пятецкий-Шапиро И. И., О классификации и канонической реализации комплексных однородных ограниченных областей, *Труды моск. матем. общества*, **12** (1963), 389-412. English translation: *Transactions Moscow Math. Soc.*, **12** (1965), 404-437.
[23] Weil A., On discrete subgroups of Lie groups, I, *Ann. of Math.*, **72** (1960), 369-384.
[24] Weil A., On discrete subgroups of Lie groups, II, *Ann. of Math.*, **75** (1962), 578-602.

Теория вероятностей и математическая статистика

Probability theory and statistics

Calcul des probabilités et statistique

Wahrscheinlichkeitsrechnung und mathematische Statistik

DER SATZ MIT DEM ITERIERTEN LOGARITHMUS

V. STRASSEN

Sei X_i die i-te Rademacherfunktion auf [0, 1] und $S_n = \sum_{i \leq n} X_i$. Borel's starkes Gesetz der großen Zahl ($S_n = o(n)$ außerhalb einer Menge vom Lebesguemaß 0) eröffnete 1909 eine Folge von Arbeiten von Hausdorff, Hardy-Littlewood, Steinhaus und Khintschin, die 1924 durch Khintschin's brillianten Satz mit dem iterierten Logarithmus abgeschlossen wurde:

$$\overline{\lim}\ (n \log\log n)^{-1/2} S_n = \sqrt{2} \tag{1}$$

außerhalb einer Menge vom Lebesguemaß 0.

Ich möchte hier über einen Teil der seitdem erzielten Fortschritte berichten.

In den ersten beiden Abschnitten werden einige Verschärfungen behandelt, wobei der oben betrachtete diskrete stochastische Prozeß $(S_n)_{n \geq 1}$ aus Gründen der leichteren Formulierung durch den kontinuierlichen Prozeß der Brownschen Bewegung $(\zeta(t))_{t \geq 0}$ ersetzt wird. Im letzten Abschnitt werden Verallgemeinerungen des zugrundeliegenden Prozesses diskutiert.

1. Der Kolmogorov-Petrovskij-Erdös Test

Khintschin's Satz lautet für die Brownsche Bewegung:

$$\overline{\lim}\ (t \log\log t)^{-1/2} \zeta(t) = \sqrt{2} \text{ fastsicher,}$$

also

$$\Pr\{\zeta(t) < (ct \log\log t)^{1/2} \text{ schließlich für } t \to \infty\} = 0 \text{ oder } 1,$$

je nachdem $c < 2$ oder $c > 2$. Sehr viel weiter geht der Kolmogorov-Petrovskij-Erdös Test (Kolmogorov's Beweis ist nicht veröffentlicht, Petrovskij [18], Erdös [5], Feller [6], Motoo [17]):

Sei $\varphi\colon R^+ \to R^+$ stetig und so, daß $t^{-1/2}\varphi(t)$ mit t wächst. Dann ist

$$\Pr\{\zeta(t) < \varphi(t) \text{ schließlich für } t \to \infty\} = 0 \text{ oder } 1,$$

je nachdem

$$\int_1^\infty t^{-3/2}\varphi e^{-\varphi^2/(2t)}\,dt = \infty \text{ oder } < \infty. \tag{2}$$

Man hat heute eine Reihe von 0-1-Kriterien ähnlicher Bauart für die Brownsche Bewegung in einer oder in mehreren Dimensionen. Hierzu sei auf Itô-McKean [12] verwiesen (S. 33-38, 161-164, 255-261, 266-269; siehe auch Chung [3], Barndorff-Nielsen [1], Shepp [19], V, 19). Konvergiert (2), so ist fastsicher schließlich $\zeta < \varphi$, von welcher Stelle an, hängt natürlich vom Zufall ab. Sei

$$T_\varphi = \sup\{t\colon \zeta(t) \geqslant \varphi(t)\}.$$

P. Lévy [16], S. 271-276 erhält eine Reihe von oberen Abschätzungen für $\Pr\{T_\varphi > s\}$, die sich als Spezialfälle der eleganten Ungleichung

$$\Pr\{T_\varphi \geqslant s\} \leqslant 2\int_s^\infty (2\pi t^3)^{-1/2}\varphi e^{-\varphi^2/(2t)}\,dt$$

in Itô-McKean [12], S. 34 erweisen (gültig unter einer zusätzlichen Monotoniebedingung an φ).

Ist nun φ sogar stetig differenzierbar mit

$$\varphi'(s) \sim \varphi'(t) \quad \text{für} \quad s \sim t, \quad t \to \infty,$$

und ist fastsicher $T_\varphi < \infty$, so hat T_φ außerhalb 0 eine stetige Dichte D_φ und es gilt

$$D_\varphi(t) \sim \varphi'(t)(2\pi t)^{-1/2}e^{-\varphi(t)^2/(2t)} \quad \text{für} \quad t \to \infty \tag{3}$$

([23], Theorem 3.6). Innerhalb der zugelassenen Funktionenklasse haben übrigens φ' und φ/t die gleiche Größenordnung, so daß man den Integranden in (2) auch durch die rechte Seite von (3) ersetzen kann. (3) liefert dann eine einfache wahrscheinlichkeitstheoretische Deutung (und einen neuen Beweis) für den Kolmogorov-Petrovskij-Erdös Test. Ein Beispiel für (3):

$$D_{(ct\log\log t)^{1/2}} \sim \left(\frac{c\log\log t}{8\pi}\right)^{1/2} t^{-1}(\log t)^{-c/2}, \tag{4}$$

wenn $c > 2$.

2. Das Verhalten der ganzen Pfade

Sei C der Banachraum der stetigen Funktionen x auf $[0, 1]$ mit $x(0) = 0$, und für $t \geqslant 0$ sei ζ_t die durch

$$\zeta_t(s) = \zeta(st) \quad (s \in [0, 1])$$

definierte Zufallsvariable mit Werten in C. ζ_t spiegelt das Verhalten von ζ bis zum Zeitpunkt t wieder. Khintschin's Satz kann (unwesentlich verschärft) so formuliert werden: Fastsicher ist $(t\log\log t)^{-1/2}\zeta_t(1)$ für $t \to \infty$ beschränkt und die Menge seiner Limespunkte ist $[-\sqrt{2}, \sqrt{2}]$.

Eine analoge Aussage läßt sich für die ζ_t als ganze machen ([21], Corollary 1, siehe auch Chover [2]): Fastsicher ist das Netz

$$((t\log\log t)^{-1/2}\zeta_{st})_{t>e}$$

relativ normkompakt, und die Menge seiner Normlimespunkte für $t \to \infty$ ist

$$\left\{x\colon x \in C,\ x \text{ ist absolutstetig, } \frac{1}{2}\int_0^1 \dot{x}^2\,d\mu \leqslant 1\right\}$$

(wobei μ das Lebesguemaß in $[0, 1]$ ist und $\dot{x} = dx/dt$ ist), besteht also aus genau den $x \in C$ mit "mittlerer kinetischer Energie $\leqslant 1$".

Dieses Ergebnis verhält sich zu Khintschin's Satz ähnlich wie Donsker's Invarianzprinzip [4] zum zentralen Grenzwertsatz (die für den Fall der Brownschen Bewegung allerdings trivial sind). Ebenso wie dort ergeben sich auch hier zahlreiche Folgerungen durch Anwenden von geeigneten Funktionalen, z.B. eine Art Gegenstück zu (4) für $c \leqslant 2$:

$$\overline{\lim_{t\to\infty}}\frac{1}{t}\mu\{s\colon s \in [0, t],\ \zeta(s) > (cs\log\log s)^{1/2}\} =$$

$$= 1 - \exp\left\{-4\left(\frac{2}{c} - 1\right)\right\} \text{ fastsicher}$$

([21], S. 223. Erdös [5], S. 434 hat schon bemerkt, daß die linke Seite fastsicher eine gewisse stetige, strikt monoton von 1 nach 0 fallende Funktion von $c \in [0, 2]$ ist, vgl. auch P. Lévy [16], S. 268-271). Oder ein Resultat, das kürzlich in etwas anderer Gestalt von Гапошкин für Summen unabhängiger Zufallsvariablen bewiesen wurde ([9], Theorem 3):

$$\overline{\lim_{a\downarrow 0}}\, a^{3/2}\left(\log\log\frac{1}{a}\right)^{-1/2}\int_0^\infty \zeta(s)\,e^{-as}\,ds = 1 \text{ fastsicher}$$

(die linke Seite ist fastsicher

$$= \lim_{n\to\infty} \overline{\lim_{a\downarrow 0}}\, a^{3/2}\left(\log\log\frac{1}{a}\right)^{-1/2}\int_0^{n/a}\zeta(s)\,e^{-as}\,ds =$$

$$= \lim_{n\to\infty} n^{3/2}\,\overline{\lim_{t\to\infty}}\int_0^1 (t\log\log t)^{-1/2}\,\zeta_t(s)\,e^{-ns}\,ds,$$

so daß man [21], S. 218 (ii) anwenden kann). Für andere Beispiele siehe [21].

Es liegt nahe, nach einer Verallgemeinerung des Kolmogorov-Petrovskij-Erdös Test's für das Verhalten der ganzen Pfade zu fragen. Vielleicht ist die folgende Problemstellung nützlich:

Sei für jedes $t>0$ $V_t \subset C$ offen und konvex so, daß $t^{-1/2}V_t \uparrow C$ (monoton), und sei

$$v_t = \inf\left\{\left(\int_0^1 \dot{x}^2\,d\mu\right)^{1/2}\colon x\in C-V_t,\ x \text{ ist absolutstetig}\right\}.$$

Welche zusätzlichen Bedingungen sind an die V_t zu stellen, damit:

$$\Pr\{\zeta_t\in V_t \text{ schließlich für } t\to\infty\} = 1 \Leftrightarrow$$

$$\Leftrightarrow \int_1^\infty t^{-3/2}v_t e^{-v_t^2/(2t)}\,dt < \infty.$$

3. Übertragung auf andere Prozesse

Kolmogorov [15] verallgemeinert Khintschin's Satz auf eine große Klasse von Prozessen $(S_n)_{n\geqslant 1}$ mit $S_n = \sum_{i\leqslant n} X_i$ und unabhängigen X_i, und P. Lévy überträgt diese Ergebnisse auf Martingale (siehe [16], S. 258-268).

Im Falle unabhängiger, identisch verteilter X_i zeigen Hartman-Wintner [11] etwas schärfer, daß $EX_1 = 0$, $EX_1^2 = 1$ für (1) hinreicht. Diese Bedingung ist auch notwendig, wenn man (1) durch

$$\underline{\lim}\,(n\log\log n)^{-1/2}S_n = -\sqrt{2} \quad \text{fastsicher}$$

ergänzt ([22], Corollary). Dagegen gibt es (D. Freedman, siehe [22]) unabhängige, identisch verteilte symmetrische X_i mit unendlicher Varianz und positive Konstante c_n so, daß

$$\overline{\lim}\, c_n^{-1}S_n = 1 \quad \text{fastsicher.}$$

Feller [6] und [7] dehnt den Kolmogorov-Petrovskij-Erdös Test auf sehr allgemeine Klassen von Prozessen $(S_n)_{n\geqslant 1}$ mit unabhängigen X_i aus, wo die S_n nicht einmal mehr asymptotisch normal zu sein brauchen (vgl. auch z.B. Гнеденко [10], Хинчин [14], Золотарев [24]).

In [21] und [23] werden die Ergebnisse der vorangehenden beiden Abschnitte mit Hilfe von Fastüberall-Invarianzprinzipien verallgemeinert. Diese sind fastsichere Gegenstücke zum Erdös-Kac-Donskerschen Verteilungsinvarianzprinzip (unter einem anderen Aspekt betrachtet als in Abschnitt 2). Ihr Beweis beruht auf einem wichtigen Satz von Скороход ([20], Theorem auf S. 163). Als Beispiel zitieren wir Theorem 1.3 (4.4) in [23] (ein Spezialfall hiervon stammt von Dubins und Freedman).

Sei $S_n = \sum_{i\leqslant n} X_i$ ein quadratisch integrierbares Martingal. Fastsicher gelte

$$V_n = \sum_{i\leqslant n} E(X_i^2 \mid X_1, \ldots, X_{i-1}) \uparrow \infty$$

für $n\uparrow\infty$ und

$$\sum_{n\geqslant 1} f(V_n)^{-1}\int_{x^2>f(V_n)} x^2\,d\Pr\{X_n\leqslant x \mid X_1, \ldots, X_{n-1}\} < \infty,$$

wobei $f\colon R^+\to R^+$ monoton wächst, aber schwächer als die identische Abbildung. Ist dann der zugrundeliegende Wahrscheinlichkeitsraum reichhaltig genug, so gibt es eine Brownsche Bewegung ζ mit

$$S_n = \zeta(V_n) + o((V_n f(V_n))^{1/4}\log V_n) \quad \text{fastsicher.}$$

Dies gestattet es leicht, den Kolmogorov-Petrovskij-Erdös Test und die Ergebnisse von Abschnitt 2 (auch z. B. des Resultat von Chung [3] und eine Version von Shepp's Dichotomie [19], V, 19) auf Martingale zu übertragen unter Bedingungen, die für Summen unabhängiger Zufallsvariablen kaum schärfer sind als Feller's Bedingung auf S. 399 von [6]. Es wäre interessant, fastsichere Invarianzprinzipien für Prozesse zu finden, die nicht asymptotisch normal sind.

Freedman [8] überträgt Abschnitt 2 auf Funktionale von Markoffschen Ketten.

Auch (3) kann mit einigen Abstrichen durch ein geeignetes Invarianzprinzip verallgemeinert werden ([23] Theorem 4.8 und Corollary 4.9), und zwar auf Irrfahrten mit endlicher erzeugender Funktion in einer Umgebung von 0.

Institut für mathematische Statistik und Wirtschaftsmathematik der Universität Göttingen und Mathematisches Institut der Universität Erlangen, Bundesrepublik Deutschland

LITERATUR

[1] Barndorff-Nielsen O., On the rate of growth of the partial maxima of a sequence of independent identically distributed random variables, *Math. Scand.* **9** (1961), 383-394.

[2] Chover J., On Strassen's version of the loglog law, *Zeitschrift für Wahrscheinlichkeitstheorie*, **8**, 2 (1967), 83—90.

[3] Chung K. L., On the maximal partial sum of sequences of independent random variables, *Trans. Amer. Math. Soc.*, **64** (1948), 205-233.

[4] Donsker M. D., An invariance principle for certain probability limit theorems, *Memoirs Amer. Math. Soc.*, **6** (1951), 1-12.

[5] Erdös P., On the law of the iterated logarithm, *Annals of Mathematics*, **43** (1942), 419-436.

[6] Feller W., The general form of the so-called law of the iterated logarithm, *Trans. Amer. Math. Soc.*, **54** (1943), 373-402.

[7] Feller W., The law of the iterated logarithm for identically distributed random variables, *Ann. of Math.*, II *Ser.*, **47** (1946), 631-638.

[8] Freedman D., Some invariance principles for functionals of a Markov Chain, *Annals of Mathematical Statistics*, **38**, 1 (1967).

[9] Гапошкин В. Ф., Теорема о суммировании по Абелю и Чезаро с помощью итерированных логарифмов, *Теория вероятностей и ее применения*, X, вып. 3 (1965), 449-459.

[10] Гнеденко Б. В., О росте однородных случайных процессов с независимыми однотипными приращениями, *ДАН СССР*, **40**, 3 (1943).

[11] Hartman P., Wintner A., On the law of the iterated logarithm, *Amer. J. Math.*, **63** (1941), 169-176.

[12] Itô K., McKean H. P., Diffusion Processes and their Sample Paths, Die Grundlehren der Mathematischen Wissenschaften in Einzeldarstellungen, Band **125**, Berlin, 1965. Русский перевод: Ито К., Маккин Г., Диффузионные процессы и их траектории, изд-во «Мир», 1968.

[13] Khintchin A., Über einen Satz der Wahrscheinlichkeitsrechnung, *Fund. Math.*, **6** (1924), 9-20.

[14] Хинчин А. Я., Две теоремы о стохастических процессах с однотипными приращениями, *Матем. сб.*, **3** (45), 3 (1938), 574-584.

[15] Kolmogorov A., Das Gesetz des iterierten Logarithmus, *Math. Annalen*, **101** (1929), 126-135.

[16] Lévy P., Théorie de l'Addition des Variables Aléatoires, Gauthiers-Villars, Paris, 1954.

[17] Motoo M., Proof of the law of the iterated logarithm through diffusion equation, *Ann. Inst. Statist. Math.*, **10** (1959), 21-28.

[18] Petrovskij I., Zur ersten Randwertaufgabe der Wärmeleitungsgleichung, *Compositio Math.*, **1** (1935), 383-419.

[19] Shepp L. A., Radon-Nikodym derivatives of Gaussian measures, Vervielfältigt an den Bell Telephone Laboratories, Murray Hill, 1965.

[20] Скороход А. Б., Теория случайных процессов, Киев, 1961.

[21] Strassen V., An invariance principle for the law of the iterated logarithm, *Zeitschrift für Wahrscheinlichkeitstheorie*, **3** (1964), 211-226.

[22] Strassen V., A converse to the law of the iterated logarithm, *Zeitschrift für Wahrscheinlichkeitstheorie*, **4** (1966), 265-268.

[23] Strassen V., Almost sure behaviour of sums of independent random variables and martingales, Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1965.

[24] Золотарев В. М., Аналог закона повторного логарифма для полунепрерывных устойчивых процессов, *Теория вероятностей и ее применения*, IX, вып. 3 (1964).

ОБ УСЛОВИЯХ СХОДИМОСТИ К ДИФФУЗИОННЫМ ПРОЦЕССАМ И АСИМПТОТИЧЕСКИХ МЕТОДАХ ТЕОРИИ МАССОВОГО ОБСЛУЖИВАНИЯ

А. А. БОРОВКОВ

Доклад по существу своему относится к предельным теоремам для случайных процессов и связан с двумя следующими вопросами:

I. Условия сходимости к марковской диффузии.

II. Асимптотические методы исследования систем обслуживания общего вида.

Краткости ради отдельно на первом вопросе мы останавливаться не будем, а проиллюстрируем имеющиеся здесь результаты на объектах раздела II, относящихся к теории обслуживания.

В названии раздела II сказано «систем обслуживания общего вида». Что при этом имеется в виду? Нам представляется, что существует необходимость иметь достаточно широкое определение основного объекта изучения в теории обслуживания, не связанное с большим количеством очень частных представлений и понятий. Здесь имеются очевидные препятствия, состоящие, например, в том, что можно придумать сколь угодно сложные системы. Поэтому мы ограничимся рассмотрением только «одноактовых» систем, когда вызовы неразличимы и каждый вызов обслуживается один раз.

Определение. Под «процессом обслуживания» на интервале времени $[0, T]$ мы будем понимать произвольный трехмерный ступенчатый вероятностный процесс

$$q(t) = \{v(t),\ r(t),\ s(t);\ 0 \leqslant t \leqslant T\}, \quad r(0) = s(0) = 0,$$

все компоненты которого неотрицательны, не убывают и обладают свойством

$$\zeta(t) = v(t) - r(t) - s(t) \geqslant 0.$$

Компонента $v(t)$ имеет смысл числа вызовов, поступивших в систему к моменту t; компонента $r(t)$ — числа вызовов, получивших отказ; $s(t)$ — числа вызовов, обслуженных системой к моменту времени t. $\zeta(t)$ есть «занятость» или «очередь» системы.

Процесс $q(t)$ можно рассматривать как заданный в пространстве $D^3(0, T)$ с σ-алгеброй борелевских множеств $\mathfrak{M}^3_{0,T}$ ($D(0, T)$ — пространство функций на $[0, T]$ без разрывов второго рода с полной метрикой). Все характеристики системы, которые принято изучать, можно рассматривать как измеримые функционалы от $\{q(t)\}$. Основ-

ные из них — это «процесс очереди» $\{\zeta(t)\}$ и «процесс отказов» $\left\{\pi(t)=\frac{r(t)}{v(t)}\right\}$.

В реальных системах распределение компонент v, r, s задается обычно путем указания их локальных свойств (распределение времени обслуживания, времени между приходами вызовов и др.), а также описанием некоторых алгоритмов, которые и определяют природу системы.

В сегодняшних условиях теория обслуживания, по-видимому, исчерпала почти все случаи «явной разрешимости» таких систем — скажем, случаи, когда удается найти (хотя бы предельные при $t\to\infty$) одномерные распределения $\zeta(t)$ или $\pi(t)$. Это весьма узкий класс систем, очерченный очень частными предположениями. Многочисленные попытки расширения этого класса сталкиваются с принципиальными трудностями.

В то же время оказалось, что существуют асимптотические законы (например, о распределениях $\{\zeta(t)\}$; $\{\pi(t)\}$), справедливые для весьма широких классов процессов обслуживания. Однако, чтобы выделить эти классы, необходимо предварительно выяснить общие свойства «входных» и «выходных» потоков $v(t)$ и $s(t)$.

Мы будем рассматривать «схему серий», т. е. последовательность

$$q_T(t)=\{v_T(t),\ r_T(t),\ s_T(t),\ 0\leqslant t\leqslant T\}$$

процессов обслуживания при $T\to\infty$. Обычно процесс $q(t)$ является функцией некоторого другого процесса $X(t)$, более полно описывающего работу системы и заданного, как правило, в фазовом пространстве $(X, \mathfrak{B})$, более богатом, чем трехмерное евклидово пространство (ср. с так называемым приемом дополнения до марковости). Процесс $\{X_T(t),\ 0\leqslant t\leqslant T\}$ мы будем называть «основным» и будем считать его заданным на некотором пространстве $(X^{(T)}, \mathfrak{B}^{(T)})$ функций $x(u)$ на $[0, T]$ со значениями в X.

Будем считать, что каждая выборочная траектория $X_T(u)$ на $[0, t]$ определяет реализацию $q_T(u)$ на $[0, t]$. Таким образом, чтобы задать $q_T(t)$, мы должны задать измеримое отображение

$$(X^{(T)}, \mathfrak{B}^{(T)})\to(D^3(0,\ T),\ \mathfrak{M}^3_{0,T}),$$

скажем, с помощью $\mathfrak{B}^{(t)}$-измеримых функционалов

$$q_T(t)=q_{T,t}(X_T(u),\ 0\leqslant u\leqslant t),$$

так, чтобы функции $q_T(t)$, $0\leqslant t\leqslant T$, с вероятностью 1 были ступенчаты.

Основное свойство входных потоков $v_T(t)$, которое в дальнейшем нами будет использоваться,— *это сходимость его распределений при* $T\to\infty$ (после некоторой нормировки процесса) *к некоторому предельному*. Я сформулирую условия на $v_T(t)$ которые представляются естественными и достаточно удобными и при которых будет иметь место сходимость к винеровскому процессу.

Предварительно введем некоторые обозначения. Пусть $\mathscr{D}\in\mathfrak{B}$ — некоторое множество из X;

$$n_{\mathscr{D}}(u)=\begin{cases}\inf(t\geqslant u:\{X_T(t)\in\mathscr{D}\}) & \text{на } \bigcup\limits_{t\geqslant u}^{T}\{X_T(t)\in\mathscr{D}\}\\ T & \text{на } \bigcap\limits_{t\geqslant u}^{T}\{X_T(t)\notin\mathscr{D}\}\end{cases}$$

— время первого после u попадания $X(t)$ в $\mathscr{D}$; $n_{\mathscr{D}}(n_{\mathscr{D}}(u)+1)$ — «время второго попадания»;

$$d_{\mathscr{D}}(u)=n_{\mathscr{D}}(n_{\mathscr{D}}(u)+1)-n_{\mathscr{D}}(u)\geqslant 1$$

— длина цикла между двумя «соседними» попаданиями $X_T(t)$ в $\mathscr{D}$. Пусть далее $\mathfrak{M}(u)$ есть σ-алгебра, порожденная событиями, связанными с траекториями $X_T(t)$ до момента времени $n_{\mathscr{D}}(u)$.

Определение. *Процесс* $\{X_T(t)\}$ *назовем* $(\mathscr{D}, \alpha)$*-возвратным, если при всех* u *п. в.*

$$\mathbf{M}_{\mathfrak{M}(u)}\,d^{\alpha}_{\mathscr{D}}(u)<c;\quad \lim_{T\to\infty}P(n_{\mathscr{D}}(0)>\Delta T)=0$$

для любого $\Delta>0$.

Обозначим, наконец,

$$V_T(t)=v_T(t)-\mathbf{M}v_T(t),$$
$$V_{u,t}=V_T(n_{\mathscr{D}}(u)+t)-V_T(n_{\mathscr{D}}(u)),$$

$\mu\{w(t)\}$ — распределение процесса $w(t)$ в $D(0, 1)$.

Теорема 1. *Пусть существуют* $\mathscr{D}\in\mathfrak{B}$ *и* $\gamma>0$, *такие, что* $\{X_T(t)\}$ $(\mathscr{D},\ 1+\gamma)$*-возвратен и при каждом* u *и* $\Delta>0$

$$\text{I. } \mathbf{M}_{\mathfrak{M}(u)}V_{u,t}=\frac{t}{\sqrt{T}}(a+\varepsilon_{t,T}),$$

$$\text{II. } \mathbf{M}_{\mathfrak{M}(u)}V^2_{u,t}=t(b+\delta_{t,T}),$$

$$\text{III. } \mathbf{M}_{\mathfrak{M}(u)}\sup_{0\leqslant t\leqslant d_{\mathscr{D}}(u)}|V_{u,t}|^{2+\gamma}<c,$$

$$\lim_{T\to\infty}\mathbf{P}\Big(\sup_{0\leqslant t\leqslant n_{\mathscr{D}}(0)}|V_T(t)|>\Delta\sqrt{T}\Big)=0,$$

где $\varepsilon_{t,T}$, $\delta_{t,T}\to 0$ *п. в. при* $t>T^{\theta}$, $T\to\infty$ *и некотором* $\theta>0$. *Тогда при* $T\to\infty$

$$\mu\{T^{-1/2}(v_T(uT)-\mathbf{M}v_T(uT))\}\Rightarrow\mu\{w(t)\},$$

где $w(t)$ — *винеровский процесс со сносом* a *и коэффициентом диффузии* b.

Замечание. Это утверждение может быть усилено до сходимости всех $(D(0,1), \mathfrak{M}_{0,1})$-измеримых и $C(0,1)$-непрерывных функционалов. Коэффициенты a и b могут зависеть от значений $V_T(n_{\mathscr{D}}(u))$. Можно ввести условия на границе, обеспечивающие сходимость к процессам с отражением, поглощением и др. (см. [1], [2]).

Условиям теоремы 1 удовлетворяет широкий класс входных потоков, включая многоканальные системы входа с разного рода зависимостью между временами поступления, между временами поступления и объемами приходящих партий вызовов и др. ([2]). Если положить $\mathscr{D} = X$, то $n_{\mathscr{D}}(u) = u$, $d_{\mathscr{D}}(u) = 1$, а условия I—II превращаются в условия обычной слабой зависимости моментов, которые мы считаем, однако, слишком ограничительными.

Сформулируем теперь некоторые результаты, относящиеся непосредственно к теории обслуживания. При этом для наглядности мы будем останавливаться главным образом на «промежуточных» случаях, когда весьма общие условия на входной процесс $v(t)$ комбинируются с конкретным заданием системы обслуживания.

Здесь нам будет удобнее рассматривать интенсивные потоки $v_T(t)$, полученные, например, из исходных сжатием времени в T раз.

Теорема 2. *Пусть*

I. *Существуют* $t_0 > 0$, *неубывающая функция* $m(t)$ *на* $[0, t_0]$ *и функция* $B(T) \to \infty$ *при* $T \to \infty$, *такие, что*

$$\mu\left\{\frac{v_T(t) - Tm(t)}{B(T)}\right\} \Rightarrow \mu\{\xi(t)\} \text{ в } D(0, t_0),$$

когда $T \to \infty$. *Процесс* $\xi(t)$ *с вероятностью* 1 *непрерывен.*

II. *Система обслуживания представляет собой* n *независимых линий обслуживания с функцией распределения времени обслуживания* $F(x)$. *Функция* $G(x) = 1 - F(x)$ *имеет на* $[0, t_0]$ *конечное число скачков.*

Тогда если $\frac{B(T)}{\sqrt{T}} \to \sigma \geqslant 0$, $n = n_T > T\int_0^{t_0} G(t_0 - u)\, dm(u) + T^{1/2+\nu}$ *при некотором* $\nu > 0$, $\zeta_T(0) = 0$, *то в* $D(0, t_0)$

$$\mu\left\{T^{-1/2}\left(\zeta_T(t) - T\int_0^t G(t-u)\, dm(u)\right)\right\} \Rightarrow$$

$$\Rightarrow \mu\left\{\Gamma(t) + \sigma\int_0^t G(t-u)\, d\xi(u)\right\},$$

где $\{\Gamma(t)\}$ *— центрированный гауссовский процесс, не зависящий от* $\{\xi(t)\}$,

$$\mathbf{M}\Gamma(t)\,\Gamma(t+u) = \int_0^t G(t+u-z)\, F(t-z)\, dm(z).$$

Теорема 3. *Пусть в условиях* I, II *теоремы* 2 $\frac{B(T)}{\sqrt{T}} \to \infty$,

$$n = n_T > T\int_0^{t_0} G(t_0 - u)\, dm(u) + B^{1+\nu}(T),\ \zeta_T(0) = 0.$$

Тогда

$$\mu\{B^{-1}(T)(\zeta_T(t) - T\int_0^t G(t-u)dm(u))\} \Rightarrow \mu\{\int_0^t G(t-u)\, d\xi(u)\}.$$

Если например, $m(t) = mt$, процесс $\xi(t)$ однороден, а $G(t_1) = 0$, $t_1 < t_0$, то уже при $t > t_1$ система будет работать в «почти стационарном режиме».

Если траектория $\zeta_T(t)$ будет часто задевать уровень n (например, когда n меньше указанных в теоремах 2, 3 пределов; в условиях теорем 2, 3 поведение вызовов, пришедших, когда $\zeta_T(t) \geqslant n$, несущественно), то возникают качественно иные задачи. В некоторых из них само конструктивное задание предельных процессов, которые получаются для $\zeta_T(t)$, вызывает значительные трудности. Но эти трудности оказываются устранимыми, если $G(t) = e^{-\alpha t}$. Предположим, что вызов, заставший систему занятой, теряется.

Теорема 4. *Пусть процесс* $v_T(t/T)$ *удовлетворяет условиям теоремы* 1 *и, кроме того,*

$$\mathbf{M}_{\mathfrak{M}(u)}\left[v_T\left(\frac{n_{\mathscr{D}}(u)+t}{T}\right) - v_T\left(\frac{n_{\mathscr{D}}(u)}{T}\right)\right] = Tmt + o(\sqrt{T}),$$

$$n = \frac{mT}{\alpha} + c\sqrt{T},\quad \zeta_T(0) = \frac{mT}{\alpha} + d\sqrt{T}\ (d \leqslant c).$$

Тогда $\mu\left\{T^{-1/2}\left(\zeta_T(t) - \frac{T_m}{\alpha}\right)\right\} \Rightarrow \mu\{w(t)\}$ *в* $D(0, 1)$, *где* $w(t)$ *— диффузионный процесс,* $w(0) = d$, *с отражением на границе* $x = c$ *и инфинитезимальным оператором*

$$(a - \alpha x)\frac{d}{dx} + \left(\frac{b+m}{2}\right)\frac{d^2}{dx^2}.$$

Замечание. Здесь также слабую сходимость в $D(0, 1)$ можно усилить до сходимости функционалов, непрерывных в равномерной метрике.

Рассмотрена задача «управления», когда коэффициенты a, b и α зависят от длины очереди ζ.

Теорема 4 является следствием более общих результатов, когда на процессы $v_T(t)$ и $s_T(t)$ накладываются условия, аналогичные условиям I—III теоремы 1, с коэффициентами a, b, зависящими от значения $\zeta_T(n_{\mathcal{T}}(u))$. Кроме того, добавляются условия, характеризующие поведение процесса $s_T(t)$ вблизи границ $\zeta = 0$, $\zeta = n$.

Если n существенно меньше пределов, указанных в теоремах 2, 3, то основной интерес начинает представлять поведение процесса отказов $\{\pi(t)\}$. Относительно $\pi(t)$ найдены эргодические теоремы также для весьма широких классов систем.

Институт математики Сибирского отделения АН СССР, Новосибирск, СССР

ЛИТЕРАТУРА

[1] Боровков А. А., Некоторые предельные теоремы о распределении функционалов от процессов, асимптотически близких к марковским, *ДАН СССР*, **169**, № 3 (1966), 507-510.

[2] Боровков А. А., О сходимости слабозависимых процессов к винеровскому, *Теория вероятностей и ее применения*, **12**, № 2 (1967), 193-222.

2 Прикладная математика и математическая физика
Applied mathematics and mathematical physics
Mathématiques appliquées et physique mathématique
Angewandte Mathematik und mathematische Physik

THE EFFECT OF GEOMETRY ON ELASTIC BEHAVIOUR

FRITZ JOHN

In this report I shall confine myself to some general observations that have motivated my work.

The classical theory of elasticity dealt with infinitesimal transformations of materials characterised essentially by a single physical parameter, Poisson's ratio. It provided clear cut answers to many concrete questions, usually by reduction to some kind of boundary value problem for the biharmonic equation. This contrasts with the more general materials considered nowadays, whose physical properties require a number of functions of several variables for their description and whose motion is determined by non-linear laws. Using modern methods of analysis, existence and uniqueness theorems have been established for the standard boundary value problems for some of these more general materials and for finite deformations. One of the basic uses of these general theories is to demonstrate the validity of the classical linear theory for sufficiently small deformations. What constitutes "sufficiently small" in this context depends, however, strongly on the geometry of the solid; it is different for a bulky solid like a sphere or cube than for a solid that is thin like a rod or plate. It is this limitation on linear behavior imposed by the geometry of the region that I should like to comment on in more detail.

The classical linear theory is based mainly on the assumption that we deal with *infinitesimal transformations*

$$x = f(X)$$

from unstrained state X to strained state x. Let $p = \frac{dx}{dX}$ be the matrix of the first derivatives of the mapping functions. The mapping is *infinitesimal* when it is close to a rigid motion in the sense that

$$p - c \ll 1,$$

where c is a *constant* orthogonal matrix. Generally the forces associated with the deformation depend on the *strain matrix* ε associated with

the mapping f. Denoting by I the unit matrix and by a superscript T transposition, the strain matrix can be defined by the non-linear relation

$$\varepsilon = \frac{1}{2}(p^T p - I).$$

For infinitesimal transformations this becomes the linear relation

$$\varepsilon = \frac{1}{2}(c^T(p-c) + (p^T - c^T)c),$$

which leads to linear equations of motion or equilibrium if we add the *physical* assumption of a linear stress-strain law for small strains.

An interesting semi-classical portion of the theory of elasticity deals with transformations for which the strains ε are so small that the classical linear stress-strain law ("physical" linearity)

$$\tau = \lambda(\operatorname{trace}\varepsilon)I + 2\mu\varepsilon$$

can be assumed to hold approximately, without, however, postulating beforehand that the transformation is infinitesimal, or that the connection between the matrices ε and p is linear ("geometric" non-linearity). In addition to the classical theory of elastic solids this area includes the theory of rods, plates, and shells.

The first question that arises is the extent to which a transformation with small strain automatically is close to a rigid motion. This can be considered purely a question in differential geometry concerning "quasi-isometric" mappings. We can measure here the *degree of rigidity* of a mapping f in various ways. One could consider the relative changes in distance between points produced by the mapping, or one could consider how close the Jacobian matrix p is to a constant orthogonal matrix.

We consider transformations $x = f(X)$ defined in an open set R of euclidean space which locally are homeomorphisms. The maximum relative change in distance between any two points of R under the transformation can be measured by the quantity

$$\eta(f) = \sup_{X, Y \in R} \left| \log \frac{|f(Y) - f(X)|}{|Y - X|} \right|,$$

which vanishes for rigid motions and is infinite for transformations that are not 1-1 in the large. The maximum relative change in distance between *neighbouring* points, that is the *maximum strain*, can be defined similarly by

$$\varepsilon(f) = \sup_{X \in R} \overline{\lim_{Y \to X}} \left| \log \frac{|f(Y) - f(X)|}{|Y - X|} \right|,$$

where $\varepsilon(f) = 0$ for isometric mappings. We have

$$0 \leqq \varepsilon(f) \leqq \eta(f) \leqq \infty.$$

We can measure *flexibility* of the region R by the flexibility function

$$\varphi_R(t) = \sup_{\substack{f \\ \varepsilon(f) \leqq t}} \left(\frac{\eta(f)}{\varepsilon(f)} - 1 \right) \quad \text{for } t \geqslant 0.$$

This function expresses how much a solid R can be deformed by transformations with maximum strain t. The corresponding function formed for an interval in 1-dimensional space is 0 for all $t \geqslant 0$ by the mean value theorem of Calculus. Little precise information is available when R is a region in 3-space. We still have

$$\varphi_R(t) = 0 \quad \text{for} \quad t \geqq 0 \quad \text{when } R \text{ is the whole space.}$$

For bounded convex R it can be proved that

$$\varphi_R(0) = 0 \text{ ("isometric mappings are rigid")}$$

$$\varphi_R(t) > 0 \text{ for } t > 0$$

$$\varphi_R(t) = \infty \text{ for } t > \frac{1}{2}\log 2.$$

For a solid sphere R we have actually

$$\varphi_R(t) < \infty \text{ for } t < \frac{1}{4}\log 2.$$

Convex sets should be less flexible for small strains than nonconvex ones. For a convex R the stiffness function rapidly becomes infinite once the maximum strain t reaches a certain order of magnitude given by the square of the (width/diameter) ratio. This probably also is the order of magnitude for the strain required to bring about compressive buckling.

These results imply that for sufficiently small strain the transformation f is approximately linear. This would suggest that its first derivatives are approximately constant, and that the Jacobian matrix p is close to a constant orthogonal matrix. Actually for uniformly small strain the matrix p does not have to lie close to one and the same constant matrix everywhere; however a lemma of Nirenberg and John on functions of bounded mean oscillation implies that p differs little from its average if the difference is measured in the L_p-norm for any $p > 1$.

If we assume that our transformation f corresponds to a position of elastic equilibrium then strains that are uniformly sufficiently small imply even *pointwise* that p is approximately constant (except possible near the boundary of R). This follows from a priori estimates for solutions of elliptic equations without reference to quasi-isometric mappings. Again, however, "sufficiently small" depends on the geomet-

ry of R. For thin bodies and "medium-sized" strains the classical linear equations for 3-dimensional bodies are no longer applicable. Combining the nonlinear equations of equilibrium for solids with the boundary conditions we obtain in the limit for vanishing thickness a different type of non-linear equilibrium equations. These two-dimensional "interior equations" can be derived with precise error estimates from suitable a priori estimates for derivatives of solutions of elliptic equations. They are similar to the plate or shell equations obtained ordinarily by formal asymptotic expansions or on the basis of special "Kirchhoff" hypotheses about the nature of the transformation.

New York University,
Courant Institute of Mathematical Sciences, USA

REFERENCES

[1] John F., Rotation and strain, *Comm. Pure Appl. Math.*, **14** (1961) 391-413.
[2] John F., Nirenberg L., On functions of bounded mean oscillation, *Comm. Pure Appl. Math.*, **14** (1961), 415-426.
[3] John F., Quasi-isometric mappings, Seminari dell'Istituto Nazionale di Alta Matematica, 1962-63 (1964), 462-473.
[4] John F., Quasi-isometric mappings in Hilbert space, New York University, IMM-NYU 336 (1965).
[5] John F., Estimates for the derivatives of the stresses in a thin shell and interior shell equations, *Comm. Pure Appl. Math.*, 18 (1965), 235-267.

SCATTERING THEORY

P. D. LAX AND R. S. PHILLIPS

Scattering theory compares the asymptotic behaviour of an evolving system as t tends to $-\infty$ with its asymptotic behaviour as t tends to $+\infty$. Scattering theory is especially fruitful for studying systems constructed from a simpler one by the imposition of a disturbance (also called perturbation or scatterer) provided that the influence of the disturbance on motions at large $|t|$ is negligible, i.e . if any motion of the *perturbed system* for large $|t|$ is indistinguishable from a motion of the unperturbed system. Thus if $U(t)$ and $U_0(t)$ denote the operators relating the states of the perturbed and unperturbed systems at time zero to their respective states at time t, then to each state f of the perturbed system there correspond two states f_- and f_+ of the unperturbed system such that $U(t)f$ behaves like $U_0(t)f$ as

$t \to -\infty$ and like $U_0(t) f_+$ as $t \to +\infty$. The scattering operator is defined as the mapping:

$$S : f_- \to f_+.$$

The aim of scattering theory is to prove the existence of such a scattering operator and to link its properties to the nature of the scatterer. In situations where the scattering operator constitutes the only physically observable data of motion the main task is the *inverse problem* of reconstructing the scatterer from the scattering operator.

This notion of scattering is meaningful for systems described by nonlinear operators. However most work on scattering theory, including the present lecture, deals with linear time-invariant systems in which case $\{U(t)\}$ form a one-parameter group of linear operators.

In our approach we deal with systems described by a group of unitary operators $\{U(t)\}$ acting on a Hilbert space H in which there are two distinguished subspaces D_- and D_+, with the property that, as t varies from $-\infty$ to ∞, the subspaces $U(t) D_-$ and $U(t) D_+$ increase (decrease) monotonically from the zero subspace to the whole space H; we call D_- and D_+ the incoming and outgoing subspaces, respectively. It is not difficult to show that with each subspace D_- and D_+ we can associate a special spectral representation of the group $\{U(t)\}$; in the one D_- is represented by functions analytic in the lower half-plane, in the second D_+ is represented by functions analytic in the upper half-plane. The two representations are related by a unitary, operator-valued multiplicative factor $\mathcal{S}(\sigma)$, $-\infty < \sigma < \infty$, which we call the *scattering matrix*. If D_- and D_+ are orthogonal then $\mathcal{S}(\sigma)$ is the restriction to the real axis of a bounded operator-valued analytic function holomorphic in the lower half-plane.

We apply this theory to systems governed by hyperbolic differential equations. The unit form of the Hilbert space is defined as energy; D_- consists of all initial states f such that $U(t) f$ is zero in some backward cone $|x| < -ct + \rho$, D_+ of states f for which $U(t) f$ is zero in some forward cone $|x| < ct + \rho$. Here ρ is so chosen that all scatterers, i.e. obstacles, potentials and inhomogeneities, are contained in the ball $\{|x| < \rho\}$. We show that in an odd number of space dimensions D_+ and D_- are orthogonal.

Denote by P_+ and P_- the operators which remove the D_+ and D_- components, i.e. project onto the orthogonal complements of D_+ and D_-. Since incoming motions are not influenced by the scatterer for $t < 0$, and outgoing motions are not influenced for $t > 0$, they can be discarded without losing any information about the scatterer. This suggests looking at the operators

$$Z(t) = P_+ U(t) P_-, \quad t > 0.$$

We show that these operators form a semigroup closely connected with the scattering matrix: the set of points in the lower halfplane at which the scattering matrix is not invertible is $i\overline{\Sigma}$, where Σ denotes the spectrum of the infinitesimal generator B of $Z(t)$.

Since $\mathcal{S}$ is unitary in the real axis, it can be continued into the upper half-plane by Schwarz reflection:

$$\mathcal{S}(z) = \{\mathcal{S}^{-1}(\bar{z})\}^*.$$

If $\mathcal{S}$ is not invertible at $\bar{z}$, z is a singularity of the analytic continuation of $\mathcal{S}$. Using the theory of elliptic and hyperbolic equations we can show that for $\operatorname{Re} k > 0$ the operator $Z(2\rho)(B - kI)^{-1}$ is compact. It follows then from the foregoing that the analytic continuation of $\mathcal{S}(z)$ into the upper half-plane is meromorphic.

Further information on the location of the poles of $\mathcal{S}$ can be obtained by using more refined properties of solutions of hyperbolic equations. From the generalized Huygens principle, according to which sharp signals propagate along rays, we can deduce that for scattering by potentials and by inhomogeneities the operators $Z(t)$ are compact for t large enough. This implies that any half-plane $\text{const} < \operatorname{Re}\mu$ contains only a finite number of eigenvalues of B; this implies that $Z(t)$ has an asymptotic expansion of the form

$$Z(t)f = \Sigma a_j e^{\mu_j t}$$

and that $\mathcal{S}$ has only a finite number of poles in each horizontal strip. We conjecture that the generalized Huygens principle holds also for solutions of mixed problems if we include reflected rays as well. If so, the above results are valid also for scattering by obstacles, provided that no ray is reflected for more than some fixed amount of time.

Independently of this conjecture C. S. Morawetz and the authors have proved that for scattering by a starshaped obstacle, governed by the wave equation, $Z(t)$ decays exponentially. This implies that $\mathcal{S}$ has no poles in some horizontal strip in the upper half-plane.

The foregoing analysis shows the role played by the eigenvalues and eigenfunctions of B. These are most naturally studied by performing the following passage to the limit:

Let a be any positive number; define $D_{\pm}^a$ to be

$$D_+^a = U(a)D_+, \quad D_-^a = U(-a)D_-,$$

and denote the corresponding projections by P_+^a and P_-^a. The spaces $D_{\pm}^a$ also are incoming and outgoing respectively and therefore we can form the corresponding semigroup $\{Z^a(t)\}$ acting on the space K^a. It is not hard to show that the scattering operator associated with the pair $D_{\pm}^a$ depends in an inessential fashion on a and that the spectrum of B^a is independent of a. The eigenelements f^a of B^a, satisfsfying

$$B^a f^a = \mu f^a,$$

depend in the following simple fashion on a:

$$P_+^a f^b = f^a, \quad \text{for } a < b.$$

P_+^a is an increasing family of projections, tending to the identity as $a \to \infty$. The relations above imply that the eigenfunctions f^a have a projective limit f. This limit f can be characterized in the following intrinsic fashion:

a) f satisfies locally the eigenvalue equation

$$Af = \mu f$$

where A is the infinitesimal generator of $U(t)$;

b) f grows exponentially as $x \to \infty$;

c) denote by $U_0(t)$ the group of operators associated with the unperturbed system, i. e. the one governing the system with all scatterers removed; then f satisfies the following version of the radiation condition: $U_0(t)f$ is eventually outgoing, i.e. is zero in some cone $|x| < t - \text{const}$.

Furthermore we can show: If μ is not a pole of the scattering matrix then for any g with compact support the equation

$$(A - \mu I)f = g$$

has a uniquely determined solution which satisfies the radiation condition.

In deriving these results we make essential use of the incoming outgoing translation representation of data with respect to the unperturbed group U_0; this is intimately connected with the Radon transform. Using the Birman-Kato principle of the invariance of the scattering operator we can relate the scattering matrix for the Schrödinger equation to that for the wave equation.

An outline of our results is described in [2]: a detailed treatment is contained in [3]. A Russian translation will be published by the Publishing House Mir.

New York University, Courant Institute of Mathematical Sciences, New York, USA

Stanford University, Stanford, USA

REFERENCES

[1] Lax P. D., Morawetz C. S., Phillips R. S., Exponential decay of solutions of the wave equation in the exterior of a star-shaped obstacle, *Comm. Pure Appl. Math.*, **16** (1963), 477-486.

[2] Lax P. D., Phillips R. S., Scattering theory, *Bull. Amer. Math. Soc.*, **70**, No. 1 (1964), 130-142.

[3] Lax P. D., Phillips R. S., Scattering theory, Academic Press, New York and London, 1969.

THÉORIE DES GROUPES ET PARTICULES ÉLÉMENTAIRES

LOUIS MICHEL

1. Rappel historique et notations

Moins de trois ans après la création de la mécanique quantique (1) paraissait le livre de H. Weyl « Gruppentheorie und Quantenmechanik (Huzel Leipzig 1928) (2) suivi par celui de E. P. Wigner « Gruppentheorie und ihre Anwendung auf die Quantenmechanik » der Atomspektren » (Vieweg, Braunschweig 1931) (3) et celui de Van der Warden « Die gruppentheoretische Methode in der Quantenmechanik » (Springer, Berlin 1932) (4).

Non seulement deux des auteurs étaient des mathématiciens, mais une partie importante des travaux du troisième sur ce sujet avait été faite en collaboration avec J. Von Neumann (5). Ces trois livres traitaient surtout des spectres atomiques, mais ils ont fortement influencé les autres applications de la théorie des groupes aux particules élémentaires.

Avant de rappeler un théorème essentiel utilisé dans ces livres, commençons par donner un exemple banal simplifiant à l'extrême : seule la partie paire $f_+(\vec{r}) = (1/2)(f(\vec{r}) + f(-\vec{r}))$ [respectivement $f_+(\vec{r}_1, \vec{r}_2) = (1/2)(f(\vec{r}_1, \vec{r}_2) + f(\vec{r}_2, \vec{r}_1))$] de l'intégrand contribue à l'intégrale sur tout l'espace $\int f(\vec{r})\, d^3\vec{r}$ (resp. $\int f(\vec{r}_1, \vec{r}_2)\, d^3\vec{r}_1 d^3\vec{r}_2$) ; ceci explique respectivement la règle de sélection de Laporte pour les spectres atomiques et la séparation du spectre de l'hélium en deux spectres distincts (6) (ortho et parahélium). Ces deux phénomènes étaient inexplicables avant l'avènement de la mécanique quantique.

Dans les trois remarquables livres cités, il s'agissait entre autres de généraliser au groupe des rotations de l'espace (SO_3) et au groupe symétrique S_n (permutations de n objets) les considérations établies dans l'exemple pour le groupe de deux éléments (Z_2 ou S_2). Comme nous le verrons, à cause du spin de l'électron, c'est le groupe SU_2, recouvrement universel de SO_3, qui intervient. Les physiciens dénotent traditionnellement D_j la représentation unitaire irréductible de dimension $2j+1$ de SU_2 (à une équivalence près). On rappelle la réduction :

$$(1) \qquad D_{j_1} \otimes D_{j_2} \sim \bigotimes_{j=|j_1-j_2|}^{j_1+j_2} D_j, \quad j + j_1 + j_2 \text{ entier} \geqslant 0.$$

Rappelons encore ce qui nous est nécessaire sur les représentations de S_n et U_n. Les représentations irréductibles du groupe symétrique

S_n sont étiquetées par des partitions de n : $[\lambda_1^{\alpha_1} \ldots \lambda_i^{\alpha_i} \ldots \lambda_k^{\alpha_k}]$ avec $\lambda_1 > \lambda_2 \ldots > \lambda_k > 0$, $\sum_{i=1}^{k} \alpha_i \lambda_i = n$. Pour tout $n > 1$, deux représentations seulement sont de dimension un, les représentations complètement symétrique $[n]$, antisymétrique $[1^n]$. On sait que

$$(2) \quad [n] \subset [\ldots \lambda_i^{\alpha_i} \ldots] \otimes [\ldots \lambda_i'^{\alpha_i'} \ldots] \Leftrightarrow \lambda_i = \lambda_i', \ \alpha_i = \alpha_i',$$

$$(2') \quad [1^n] \subset [\quad] \otimes [\quad] \Leftrightarrow \lambda_i' = \hat{\lambda}_i = \sum_{j=1}^{k+1-i} \alpha_j,$$

$$\alpha_i' = \hat{\alpha}_i = \lambda_{k-i+1} - \lambda_{k-i+2}.$$

Dans ce dernier cas, les représentations seront dites complémentaires. Parfois, nous noterons simplement $[\ ]_\lambda$ une représentation irréductible de S_n, en notant alors $[\ ]_\lambda^c$ la complémentaire.

Nous appelons représentation primaire une somme directe de représentations unitaires irréductibles équivalentes.

Soit $\mathscr{H}^{(1)}$ un espace d'Hilbert ; nous notons $\mathscr{H}^{(n)} = \overset{n}{\otimes} \mathscr{H}^{(1)} = \mathscr{H}^{(1)} \otimes \ldots \otimes \mathscr{H}^{(1)}$ le produit tensoriel de n copies de $\mathscr{H}$. Par permutation des facteurs le groupe S_n agit sur $\mathscr{H}^{(n)}$. Décomposons cette représentation, que nous noterons $[\ ]_{\mathscr{H}^{(n)}}$, en somme directe de représentations primaires, et notons $\mathscr{H}^{(n)}_{[\ ]_\lambda}$ le sous-espace de $\mathscr{H}^{(n)}$ sur lequel agit la représentation primaire $\oplus\, [\ ]_\lambda$. Pour le sujet des livres cités, le théorème suivant est fondamental :

Théorème : Si $\mathscr{H}^{(1)}$ est de dimension finie k, l'action du groupe unitaire U_k sur $\mathscr{H}^{(1)}$ est transportée sur $\mathscr{H}^{(n)}$, où il agit par $\overset{n}{\otimes} U_k$. La décomposition de cette représentation de U_k en somme directe de représentations primaires fournit les *mêmes* espaces $\mathscr{H}^{(n)}_{[\ ]_\lambda}$. On peut donc noter par les mêmes symboles $[\ ]_\lambda$ les représentations irréductibles de U_k ainsi obtenues (ce sont toutes les représentations continues). Plus précisément

$$(3) \qquad \text{pour } S_n \qquad [\ ]_{\mathscr{H}^{(n)}} \sim \oplus_\lambda u_\lambda [\ ]_\lambda,$$

$$(3') \qquad \text{pour } U_k \qquad \overset{n}{\otimes} U_k \sim \oplus_\lambda s_\lambda [\ ]_\lambda$$

où u_λ et s_λ sont les dimensions de la représentation notée $[\ ]_\lambda$ respectivement pour le groupe U_k et le groupe S_n.

Remarque 1 : Si $k < n$, seules interviennent les représentations de U_n ou de S_n telles que $\Sigma \alpha_i \leqslant k$. Par exemple les représentations de U_2 sont $[\lambda_1, \lambda_2]$ où les entiers λ_1, λ_2 satisfont $\lambda_1 \geqslant \lambda_2 \geqslant 0$.

Remarque 2 : La restriction d'une représentation irréductible de U_k sur le sous-groupe SU_k, est irréductible. Ainsi la restriction de $[\lambda_1, \lambda_2]$ de U_2 sur SU_2 est D_j avec $2j = \lambda_1 - \lambda_2$, et dim $[\lambda_1, \lambda_2] = \lambda_1 - \lambda_2 + 1 = 2j + 1$.

Nous noterons encore $\overset{n}{\Lambda}\mathscr{H}^{(1)}$ pour $\mathscr{H}^{\{n\}}_{[1^n]}$ (resp. $\overset{n}{V}\mathscr{H}^{(1)}$ pour $\mathscr{H}^{\{n\}}_{[n]}$). C'est l'espace des tenseurs de rang n complètement antisymétriques (resp. symétriques) sur $\mathscr{H}^{(1)}$ (de dimension quelconque).

Il nous faut aussi expliquer en quelques phrases ce qu'est la mécanique quantique [7]. Dans cette mécanique, un état physique est représenté par un vecteur $x\rangle$, normé : $\langle x, x\rangle = 1$, d'un espace d'Hilbert. Une grandeur physique $\mathfrak{A}$ est représentée par A, un opérateur self adjoint sur $\mathscr{H}$ [8]. Le résultat de la mesure de $\mathfrak{A}$ pour l'état $x\rangle$ appartient au spectre de A. La mécanique quantique ne permet de prédire que l'espérance mathématique de ce résultat : $\langle x, Ax\rangle = \mathrm{Tr}\, AP_x$ où P_x est le projecteur sur le sous-espace à une dimension engendré par $x\rangle$. Notons que deux vecteurs propres normés de P_x, qui ne diffèrent donc que par une phase, représentent le même état physique, puisqu'ils impliquent les mêmes prédictions. On appelle les A des observables, et il est utile de considérer les différentes algèbres associatives qu'ils engendrent [9]. Les P_x eux-mêmes sont des observables. Ainsi

$$\mathrm{Tr}\, P_x P_y = |\langle x, y\rangle|^2 \tag{4}$$

est la probabilité d'observer dans l'état $x\rangle$ (resp. $y\rangle$) le système qu'on savait être dans l'état $y\rangle$ (resp. $x\rangle$) [10].

2. Invariance relativiste

Dès 1928, Dirac [11] créa pour l'électron une mécanique quantique relativiste. Dans une telle théorie, la composante connexe $\mathscr{P}_0$ du groupe de Lorentz inhomogène [12] — que les physiciens appellent groupe de Poincaré, puisque le premier il en montra l'intérêt physique — est un groupe d'automorphismes de l'algèbre des observables qui agit donc sur les P_x tout en laissant invariantes les probabilités de transitions $\mathrm{Tr}\, P_x P_y$. Le groupe $\mathscr{P}_0$ agit donc par isométries sur $\mathscr{H}$, et par conséquent [13] cette action est réalisée par une représentation unitaire projective (que l'on admet continue). Par définition de la notion de particule (système qu'on peut isoler et dont on peut négliger la composition interne) sur l'espace $\mathscr{H}^{(1)}$ des états d'une particule, la représentation est irréductible. La caractérisation des représentations unitaires projectives continues irréductibles de $\mathscr{P}_0$ correspondant aux particules élémentaires a été accomplie par Wigner [14] en 1937 ; elles sont données par des représentations de $\overline{\mathscr{P}}_0$ [15], le

recouvrement universel de $\mathscr{P}_0$ et sont définies par les invariants

$$m > 0 \text{ et } s(s+1) \text{ avec } 2s \text{ entier} \geqslant 0$$

(5) ou

$$m = 0 \text{ et } 2\lambda \text{ entier.}$$

Ils correspondent respectivement à la valeur de la masse $m \geqslant 0$ et du spin s ou $|\lambda|$ de la particule.

Les générateurs (= éléments de l'algèbre de Lie multipliés par $i = \sqrt{-1}$) de $\mathscr{P}_0$ sont des observables : l'impulsion pour les translations d'espace, l'énergie pour celle de temps. Ce dernier observable est l'Hamiltonien H. Pour un système physique, il détermine son évolution ; les observables commutant avec H sur l'espace des états du système sont des constantes du mouvement. Les états stationnaires sont des états propres de H, la valeur propre étant leur énergie.

La cinématique des particules élémentaires (conservation de l'énergie, de l'impulsion, du moment cinétique, règles de sélection et corrélations angulaires de réactions successives, effets de polarisation, etc...) n'est autre qu'une étude détaillée, géométrique, du groupe $\mathscr{P}_0$. D'une dizaine il y a vingt ans, le nombre des différentes particules connues dépasse aujourd'hui largement la centaine. La découverte de nouvelles particules, la détermination de leur masse et spin (ainsi que d'autres caractéristiques), leur classification, sont parmi les principales activités de la physique actuelle des particules fondamentales.

Le groupe complet de Poincaré est engendré par $\mathscr{P}_0$, par une réflexion d'espace, telle que « P » : $t, \vec{r} \to t, -\vec{r}$, et par une réflexion de genre temps, telle que « T » : $t, \vec{r} \to -t, \vec{r}$. Depuis le XIX^e siècle les physiciens ont conscience de l'invariance des lois physiques par rapport aux réflexions spatiales. C'est vrai à l'échelle macroscopique où n'interviennent que les interactions électromagnétiques et la gravitation (si l'on excepte la production d'énergie nucléaire, soit par l'homme, soit au sein des étoiles). C'est encore vrai pour les interactions nucléaires, beaucoup plus intenses que les interactions électromagnétiques, mais de faible portée (10^{-13} *cm*). Il existe dans la nature un quatrième type d'interactions, aussi de très courte portée, mais beaucoup plus faibles en intensité. On les appelle interactions faibles (ou encore de Fermi). On n'a pu provoquer des réactions entre particules par leur intermédiaire que depuis 10 ans. Toutes les particules connues, à l'exception du photon, semblent douées d'interactions faibles. Aussi ces interactions sont responsables de la désintégration spontanée de la plupart des particules découvertes avant 1960. Plusieurs contradictions apparentes entre certains résultats expérimentaux concernant ces désintégrations furent résolues en 1957 par la confirmation de l'hypothèse de Lee et Yang [16] que les

interactions faibles violent la parité, c'est-à-dire ne sont pas invariantes pour les réflexions spatiales « P » $\mathcal{P}_0$. Disons encore que ces réflexions ne sont des automorphismes de l'algèbre des observables que dans l'approximation (souvent valide) où l'on peut négliger les interactions faibles. Les mêmes résultats expérimentaux impliquaient aussi la violation de « C », l'involution qui échange particules et antiparticules et qui permit à Dirac en 1931 de prédire l'existence de l'électron positif découvert l'année suivante. Jusqu'en 1957, « C » avait paru être aussi une symétrie fondamentale des lois physiques. Cette dissymétrie pour « P » et « C » peut être attribuée en partie aux neutrinos, particules de masse nulle, de spin 1/2, les seules qui n'ont que des interactions faibles (et gravifiques) ; conformément à des prédictions théoriques, il fut établi en 1962 qu'il existe deux espèces de neutrinos. Chacune d'elles a des particules et des antiparticules, mais les états des unes et des autres ne se correspondent ni par « P », ni par « C » mais seulement par le produit « PC » [17]. Cependant la violation de « P » et de « C » apparaît aussi dans des désintégrations dues aux interactions faibles, où n'interviennent pas les neutrinos. Le produit « PC » semblait être une invariance fondamentale de la physique, lorsqu'en 1964 fut découvert un mode rare (fréquence 2.10^{-3}) de la désintégration des K_2^0 qui est une manifestation de violation de « PC » et qui ne semble pas due à la non-symétrie de l'environnement (terre, et même galaxie) pour cette involution.

Nous ne savons pas réaliser des réflexions temporelles, mais « T » peut être interprété comme renversement du mouvement et aucun résultat expérimental n'a encore infirmé l'hypothèse que « T » est un antiautomorphisme involutif de l'*-algèbre des observables. Mais il faut souligner que les différentes explications théoriques proposées pour la violation observée de « PC » violent toutes aussi « T » de façon à préserver le produit «PCT» comme symétrie de la physique. En effet, en théorie quantique des champs, le meilleur cadre de pensée que nous ayons actuellement pour étudier les particules fondamentales, la covariance par rapport à la composante connexe $\mathcal{P}_0$ du groupe de Poincaré entraîne l'équivalence entre « PCT » , antiautomorphisme involutif de la théorie , et « bonne relation entre spin et statistique ».

Il nous faut expliquer cette dernière phrase. Si $\mathcal{H}_1$ et $\mathcal{H}_2$ sont les espaces des états de deux systèmes physiques, le produit tensoriel $\mathcal{H}_1 \otimes \mathcal{H}$ est l'espace des états de leur réunion si ces systèmes sont différents. Mais l'espace des états d'un système de n particules identiques, dont $\mathcal{H}^{(1)}$ est l'espace des états pour chacune d'elles est soit $\overset{n}{\Lambda}\mathcal{H}^{(1)}$ (statistique de Fermi-Dirac) soit $\overset{n}{\mathrm{V}}\mathcal{H}^{(1)}$ (statistique de Bose-Einstein) suivant que ces particules aient un spin s demi-entier ou entier [18].

Les effets de la «statistique» (utilisation de $\overset{n}{\Lambda}\mathcal{H}$ ou $\overset{n}{\mathrm{V}}\mathcal{H}$ au lieu de $\overset{n}{\otimes}\mathcal{H}$) sont remarquables (superfluidité de l'hélium, superconductivité, électrons dans les solides, règles d'intensité des spectres moléculaires, etc...). Limitons-nous à les décrire rapidement dans les atomes (où ils furent découverts empiriquement par Pauli : principe d'exclusion [19]) et dans les noyaux.

3. Physique atomique et nucléaire

Pour étudier la structure des atomes et de leurs noyaux, la limite non relativiste est une excellente approximation. L'espace $\mathcal{H}^{(1)}$ des états d'une particule est alors le produit tensoriel

$$\mathcal{H}^{(1)} = \mathcal{H}_r \otimes E_{2s+1} \tag{6}$$

où $\mathcal{H}_r = \mathcal{L}^2(R^3)$ et E_{2s+1} l'espace à $2s+1$ dimensions sur lequel agit la représentation D_s de SU_2, s étant le spin de la particule. Explicitement, les éléments de $\mathcal{H}$ sont les fonctions à valeur complexe $\psi(\vec{r}, \sigma)$, ($\vec{r} \in R^3$, $\sigma \in$ d'un ensemble de deux éléments) normées à $1 = \langle \psi, \psi \rangle$

$$\langle \psi, \psi \rangle = \sum_{\sigma} \int \bar{\psi}(\vec{r}, \sigma)\, \psi(r, \sigma)\, d^3\vec{r}. \tag{7}$$

192

Tous les observables des états de n particules indiscernables sont invariants pour S_n ; ils laissent donc stables les sous-espaces $\mathcal{H}^{(n)}_{[\lambda]}$ de $\mathcal{H}^{(n)}$. Considérons un atome (ou ion) de n électrons. A une excellente approximation, l'Hamiltonien H ainsi qu'entre autres les opérateurs d'absorption et d'émission de photons, sont indépendants du spin, c'est-à-dire sont de la forme :

$$A = A_r \otimes I_s \quad \text{sur} \quad \mathcal{H}^{(n)} = \mathcal{H}_r^{(n)} \otimes E_{2s+1}^{(n)} \tag{8}$$

(où I_s est l'opérateur identité sur E_{2s+1}).

L'espace des états de l'atome de n électrons est

$$\overset{n}{\Lambda}\mathcal{H}^{(1)} = \oplus_{\substack{\lambda_1+\lambda_2=n \\ \lambda_1 \geqslant \lambda_2}} (\mathcal{H}_r^{(n)})_{[\lambda_1, \lambda_2]^c} \otimes (E_2^{(n)})_{[\lambda_1, \lambda_2]} \tag{9}$$

puisque les électrons ont spin 1/2 et obéissent donc à la statistique de Fermi (voir aussi notre remarque 1 ; ici $k = \dim E_2 = 2$). Chacun des $\lambda = \text{Entier}\ (n/2+1)$ sous-espaces dans (9) correspondant à un couple λ_1, λ_2 est stable pour les opérateurs de (8). Pour l'atome d'hélium, $n = 2$ et les deux sous-espaces sont ceux des états de l'orthohélium ([2]) et du parahélium [1^2] dont nous avons déjà parlé.

Faute de temps je ne rappellerai pas ici comment ces considérations amènent à prévoir la structure « en couche » des électrons dans les atomes. Je me bornerai à affirmer que pour un niveau atomique la symétrie $[\lambda_1, \lambda_2]$ est définie (à une bonne approximation) et $\lambda_1 - \lambda_2$ est la valence chimique de ce niveau. Les différentes valences d'un atome appartiennent à des niveaux d'énergie différente. A cause de la répulsion électrostatique des électrons, les états d'énergie les plus bas sont ceux qui ont le plus antisymétrique possible d'espace (pour $\mathscr{H}_r^{(n)}$). Ils ont donc un $\lambda_1 - \lambda_2$ le plus grand possible en général.

Grâce au théorème cité en introduction, nous aurions pu atteindre les mêmes conclusions dans le langage suivant, les opérateurs de type (6) commutent avec ceux de la représentation $I \otimes (\overset{n}{\otimes} U_2)$ du groupe unitaire U_2 de l'espace des spins E_2. A cette approximation, en physique atomique, il y a conservation séparée du moment cinétique orbital et du moment cinétique de spin pour l'ensemble des n électrons. D'après la remarque 2, le spin total d'un état de symétrie $[\lambda_1, \lambda_2]$ est $(1/2)(\lambda_1 - \lambda_2)$.

Bien que notre connaissance des forces nucléaires soit plus imprécise, la statistique de Fermi nous permet une étude assez détaillée du noyau atomique qu'on étudie en plaçant ses particules constituantes, les nucléons, dans le potentiel sphérique attractif moyen qu'ils créent. Il y a deux sortes de nucléons, les protons, qui ont une charge électrique $+$, et les neutrons, électriquement neutres. Leur masse est égale au millième près. Ils ont tous deux spin 1/2. L'espace des états d'un noyau de p protons, n neutrons, donc $a = p + n$ nucléons est donc

$$\mathscr{H}^{(a)} = (\overset{p}{\Lambda}\mathscr{H}^{(1)}) \otimes (\overset{n}{\Lambda}\mathscr{H}^{(1)}), \tag{11}$$

les espaces $\overset{p}{\Lambda}\mathscr{H}^{(1)}$ et $\overset{n}{\Lambda}\mathscr{H}^{(1)}$ pouvant chacun être décomposés selon (8). Comme pour les atomes les noyaux présentent une structure en couche, séparément pour les protons ou les neutrons, successives étant pour 2, 8, 20, 50, 82, 126 protons ou neutrons. Les forces résiduelles entre nucléons étant attractives, l'état fondamental d'un noyau sera le plus symétrique possible d'espace pour les protons et les neutrons séparément, la valeur correspondante de chaque $\lambda_1 - \lambda_2$ (avec $\lambda_1 + \lambda_2 =$ soit p, soit n) étant alors minimum. C'est zéro lorsque p et n sont deux nombres pairs, ce qui semble bien vérifié, puisque pour tous les noyaux p pair, n pair dont on a mesuré le spin dans l'état fondamental, on a trouvé $s = 0$.

Les protons et neutrons ont les mêmes propriétés nucléaires. A l'approximation où l'on ne tient compte que des forces nucléaires (à l'échelle des noyaux les forces électromagnétiques sont bien moins intenses et peuvent être négligées) les observables de la physique nucléaire sont symétriques par rapport à *tous* les nucléons et ils laissent stables les sous-espaces $\mathscr{H}_{[\lambda]}^{(a)}$ de l'espace $\mathscr{H}^{(a)}$ de a nucléons. Parce qu'il n'y a que deux sortes de nucléons, les protons et les neutrons qui satisfont indépendamment à la statistique de Fermi, seuls les espaces de symétries $[\lambda_1, \lambda_2]^c$, $\lambda_1 + \lambda_2 = a$ représentent des noyaux. Les états d'une même représentation irréductible $[\lambda_1, \lambda_2]^c$ seront identiques (même énergie, même spin, etc...) à cette approximation, bien qu'ils appartiennent à des noyaux différents dits isobares (différents n et p mais même $a = n + p$).

Dès 1932 Heisenberg [19] employa un raisonnement équivalent, mais plus élégant, pour traiter de cette question: si on néglige leur légère différence de masse et les propriétés électromagnétiques qui les différencient, tous les nucléons deviennent indiscernables ; ils satisfont la statistique de Fermi, à condition cependant de tenir compte de leur degré supplémentaire de liberté, la variable τ qui peut prendre deux valeurs : « proton » et « neutron ». Les états d'un nucléon sont alors les fonctions de l'espace $\mathscr{H}_N^{(1)}$ (N = nucléon)

$$\mathscr{H}_N^{(1)} \ni \psi(\vec{r}, \sigma, \tau) = \mathscr{H}^{(1)} \otimes E_2(\tau) = \mathscr{H}_r \otimes E_2(\sigma) \otimes E_2(\tau) \tag{12}$$

où $\mathscr{H}^{(1)}$ a été défini en (6) et (7). L'analogie avec (6) et tout ce qui précède est complète et la variable τ est appelée par analogie isospin [20]. Les observables nucléaires sont indépendantes de τ et le groupe U_1 de E_2 (C) est un groupe d'invariance de la théorie son action commutant avec celle de toutes les observables. Pour a nucléons tous les états d'une même représentation irréductible $[\lambda_1, \lambda_2]$ sur $E_2^{(a)}(\tau)$ sont nucléairement identiques. Nous disons encore qu'ils forment un « multiplet » de $\lambda_1 - \lambda_2 + 1$ états de *même* « isospin » $t = (1/2)(\lambda_1 - \lambda_2)$ (cf. remarque 2). Le nucléon ($a = 1$) est un doublet d'isospin 1/2. Les physiciens choisissent une base usuelle pour l'algèbre de Lie de SU_2

$$[T_0, T_+] = T_+, \quad [T_0, T_-] = -T_-, \quad [T_+, T_-] = T_0, \tag{13}$$

T_0 étant choisi tel que l'état proton et l'état neutron soient ses deux vecteurs propres.

Ayant l'expérience du spin ordinaire, les physiciens préfèrent le langage de l'isospin à celui, équivalent, des permutations.

Répétons alors le raisonnement fait à propos du spin. Les forces nucléaires étant attractives, les états de $\overset{N}{\Lambda}\mathscr{H}_N$ les plus stables sont les plus symétrique possible pour les variables ordinaires $(\vec{r}, \sigma)$ et donc les plus antisymétrique possibles pour τ, c'est-à-dire l'isospin pes états les plus stables $t = (1/2)(\lambda_1 - \lambda_2) \geqslant 0$ est minimum. En effet, les noyaux légers ont à peu près même nombre de protons et de neutrons. Pour les noyaux plus lourds, l'excès de neutrons ($n - p/a \sim$

$\sim 20\%$ pour $a > 200$) est dû à la répulsion électrostatique des protons. De plus, à de rares exceptions près, l'état le plus stable d'un noyau a un isospin t minimum (le t pouvant encore être défini et mesuré pour des états de noyaux beaucoup plus lourds ($a \sim 50$)) qu'on ne le pensait il y a quelques années.

Dès 1937, Wigner [21] étudia une approximation, qui bien que beaucoup plus grossière, n'est pas sans intérêt pour les noyaux légers et la classification des désintégrations β. Si on admet que les forces nucléaires sont non seulement indépendantes d'isospin, mais aussi de spin, alors le groupe d'invariance est le groupe U_4 agissant sur $E_2(\sigma) \otimes E_2(\tau)$. Chaque représentation irréductible : $[\lambda_1, \lambda_2, \lambda_3, \lambda_4]$, avec $\Sigma\lambda = a$, $\lambda_1 \geqslant \lambda_2 \geqslant \lambda_3 \geqslant \lambda_4$ de dimension k, forme un «supermultiplet» de k états identiques dans cette approximation. Pour trouver le contenu en spin et isospin du supermultiplet, on réduit en somme directe de représentations irréductibles la restriction de la représentation de U_4 au sous-groupe $SU_2 \otimes SU_2$ agissant sur $E_2(\sigma) \otimes E_2(\tau)$.

L'introduction de l'isospin est-elle purement formelle et équivalente à la permutation des nucléons? On peut répondre maintenant par la négative. Dès 1935, Yukawa avait prédit l'existence d'une particule, le méson π, qui jouerait pour les forces nucléaires le rôle que joue le photon pour les interactions électromagnétiques [22]. Dès 1938, son isospin t_π fut prédit : $t_\pi = 1$; cela signifie l'existence de 3 particules nucléairement identiques, mais de charge électrique différente. Les états chargés π^+ et π^- furent découverts en 1947, l'état neutre π° (vie moyenne 10^{-16} sec) en 1949. Si on ne veut parler qu'en termes de permutations, il faut alors considérer les π comme des états liés d'un nucléon et d'un antinucléon. Mais depuis 1950 de nombreuses particules ont été découvertes. A l'exception des neutrinos, toutes ces particules ont des intéractions nucléaires (on dit encore fortes), et sont toutes instables (10^{-8} à 10^{-23} sec de vie moyenne). Leur mode de production et de désintégration révèle une loi de conservation aussi fondamentale que la conservation de la charge électrique, la conservation de la charge nucléaire [23], qu'on appelle encore charge baryonique b. Les nucléons font partie de la famille des baryons (états fondamentaux p^+, n, Λ^0, Σ^+, Σ^0, Σ^-, Ξ^0, Ξ^- et de nombreux états excités) qui ont $b = 1$ et le spin demi-entier ; les π font partie de la famille des mésons (π^+, π^0, π^-, K^+, K^0, $\overline{K}^0$, K^-, η^0, etc...) qui ont $b = 0$ et spin entier [24]. A toutes ces particules on peut attribuer un isospin ; $t = 0$ pour Λ^0, η^0 ; $t = (1/2)$ (n, p) (Ξ^0, Ξ^-), (K^+, K^0) ; $t = 1$, $\Sigma^+\Sigma^0\Sigma^-$, etc... les membres des multiplets ayant différentes charges électriques q (indiquées en indice), même charge baryonique, même spin, des masses dont les rapports diffèrent au plus de quelques 10^{-2} de l'unité ; à l'approximation où l'on peut négliger les interactions faibles on peut encore définir une parité ± (caractérisant

la variance pour les réflexions spatiales). L'involution « PCT » dont nous avons parlé implique l'existence (jamais infirmée) d'antiparticules de même masse, même spin, même isospin que les particules, mais de charges q et b opposées. Il y a même isomorphisme mathématique entre le groupe engendré par SU_2, recouvrement des rotations et « P » (réflexion d'espace) et le groupe engendré par SU_2, groupe d'isospin et « C », la conjugaison de charge, d'où un nouvel invariant, « l'isoparité », en physique des interactions fortes [25].

Une nouvelle étape fut franchie en 1958. En dénotant $j^\mu(x)$, la densité (dans l'espace temps) du courant électromagnétique, l'opérateur « charge électrique totale » est $Q = \int_\sigma j_\mu(x)\, d\sigma^\mu(x)$, où σ est une surface de genre espace ; Q est indépendant de σ, sa conservation étant assurée par $\frac{\partial}{\partial x^\mu} j^\mu(x) = 0$. Cette conservation peut encore se traduire par l'invariance par rapport à un groupe U_1 : $\{e^{i\alpha Q}\}$. De même on peut concevoir une densité de courant $b^\mu(x)$ pour la charge baryonique. (Serait-ce $\hat{j}^\mu(x)$ la partie du courant $j^\mu(x)$ invariante par le groupe SU_2 des transformations d'isospin?) Feynman et Gell-Mann [26] découvrirent l'existence de la densité du courant vectoriel $v_i^\mu(x)$ d'isospin 1, dont la composante neutre $v_0^\mu(x) = j^\mu(x) - \hat{j}^\mu(x)$ et les composantes chargées $v_\pm^\mu(x)$ sont source des interactions faibles, et $T_i = \int_\sigma v_i^\mu(x)\, d\sigma_\mu(x)$ sont les générateurs infinitésimaux du groupe SU_2 d'invariance d'isospin. On se sent à la racine d'une profonde découverte, les générateurs T_i d'un groupe abstrait introduit pour les interactions fortes, étant des observables des interactions électromagnétiques et faibles! Insistons cependant encore sur le fait que ce ne sont que des lois approchées ; par exemple $\frac{\partial}{\partial x^\mu} v_\pm^\mu(x) = 0$ seulement si on néglige les interactions faibles et éléctromagnétiques. Poursuivant sa ligne de pensée, Gell-Mann [27] fut amené à proposer que les 3 intégrales $A_j = \int_\sigma a_j^\mu(x)\, d\sigma_\mu(x)$, de la densité pseudo-vectorielle d'isospin 1, $a_j^\mu(x)$, une autre source des interactions faibles engendraient avec les trois T_i l'algèbre de Lie de $SO_4 = SU_2 \times SU_2$ (base $A_i + T_i$, $A_i - T_i$) ; cette hypothèse permit, l'an dernier, le calcul théorique d'une constante connue expérimentalement depuis plus de dix ans, ainsi que le calcul de plusieurs autres effets [28]. Nonobstant des difficultés conceptuelles, beaucoup de physiciens ont essayé depuis un an d'agrandir cette algèbre pour incorporer les autres sources connues des interactions faibles et calculer d'autres phénomènes. Il est trop tôt pour faire un bilan.

5. L'invariance SU_3

Terminons enfin par un succès récent des plus spectaculaires : la prédiction en 1962, basée sur le groupe SU_3, de l'existence et des propriétés d'une nouvelle particule, qui fut trouvée deux ans après!

Les générateurs T_i de l'algèbre SU_2 de l'isospin et la charge électrique Q engendrent l'algèbre de Lie de U_2, un générateur du centre étant [29] [30]

$$\frac{Y}{2} = Q - T_0.$$

L'observable Y est appelé hypercharge, c'est donc une quantité conservée, à l'approximation nucléaire, les particules de la nouvelle génération découverte depuis 1950 ont $y = 1, 0, -1$. Des symétries apparaissent. On sent qu'il faut agrandir le groupe U_2. On peut croire un moment refaire avec p, n, Λ ce qui avait été fait avec p, n pour l'isospin et obtenir ainsi U_3. Cela échoue. Cependant une famille de 8 baryons de spin 1/2 apparaît clairement. Les physiciens interrogent les groupes de Lie compacts, contenant U_2 et ayant une représentation irréductible de dimension 8. Indépendamment Gell-Mann et Ne'eman [31] nous convainquirent tous que le bon choix était SU_3. On put comprendre alors, bien que le succès fut au prime abord étonnant, la relation entre les masses des 8 baryons

$$m = m_0 + m_1\left(t(t+1) - \frac{1}{4}y^2\right) - m_2 y \tag{15}$$

appartenant à la représentation adjointe [2, 1] de SU_3 [32]. Gell-Mann montre qu'un groupe de sept baryons (résonances de vie moyenne 10^{-22} sec) de spin 3/2, de parité +, parmi lesquels quatre ont $t = 3/2$, $y = 1$ et trois autres ont $t = 1$, $y = 0$ doit appartenir à la représentation [3] de dimension 10, la formule (15) se simplifiant alors en $m = m_0' - m'y$. Le doublet prédit $t = 1/2$, $y = -1$ fut vite trouvé, avec la bonne masse et le spin 3/2. La particule $t = 0$, $y = -2$ se fit attendre deux interminables années [33]. C'est encore la seule particule connue dont $|y| > 1$. Elle a la bonne masse, mais le spin n'a pû être mesuré puisque cinq exemplaires seulement de cette particule (Ω^-) ont pu être observés jusqu'à ce jour parmi plusieurs millions de clichés photographiques puis avec des chambres à bulles exposées aux faisceaux des accélérateurs géants.

Plusieurs octets (représentations [2,1] de SU_3) (ainsi que deux singlets [0, 0]) ont été identifiés parmi les mésons et d'autres pour les baryons. Seul le groupe adjoint SU_3/Z_3 intervient-il?

6. S'arrêtera-t-on à U_6 ?

Une plus grande symétrie put encore être trouvée [34] en étendant la théorie des supermultiplets de Wigner de l'isospin (SU_2) au « spin unitaire» SU_3. Les états des particules sont alors classés par les représentations de U_6, des relations apparaissant alors entre spin, masse, t, y. L'octet de baryon de spin 1/2 et le décuplet de spin 3/2 forment la représentation [3] de U_6 de dimension $56 = 8\,[(2 \times 1/2) + 1] + 10\,[(2 \times 3/2) + 1]$ tandis que tous les mésons de plus basses masses connues, 8 de spin et parité 0^-, $8 + 1$ de spin et parité 1^- forment la représentation adjointe [2, 1^4] de U_6, de dimension $35 = 8 + 9\,(2 + 1)$.

L'utilité de cette symétrie U_6 est certaine, mais son domaine de validité est encore mal circoncis. Il est plus facile d'étudier des symétries exactes que des symétries approchées. L'art du physicien est de débrouiller nos observations de la nature en simplifiant, en idéalisant, en un mot, en faisant les approximations nécessaires. Loin de rebuter le mathématicien, ce passage d'une approximation à une autre devrait l'intéresser, puisqu'il s'agit d'une « déformation » de structure, entre autres pour notre sujet, de déformation de groupes et d'algèbres [35], [36], [37].

7. Conclusion

Nous connaissons bien les forces électromagnétiques : les équations de Maxwell quantifiées forment l'électrodynamique quantique [38]. Elles sont invariantes par le groupe de Poincaré. Nous connaissons mal les forces nucléaires, ignorant encore les équations qui les régissent. Cependant avec l'aide de différents groupes de Lie, les physiciens ont quand même pu faire, et avec une certaine efficacité, de la physique des noyaux atomiques et même, un peu de la zoologie pour cette faune inattendue de particules subnucléaires.

Institut des Hautes Études Scientifiques,
France

REMARQUES

1. Heisenberg W., *Z. Phys.*, **33** (1925), 879. Dès 1927, von Neumann, Göttinger Nachrichten, p. 245, donnait un exposé cohérent de la mécanique quantique et y introduisait la notion de matrice densité. Voir aussi réf. 8 pour les traités de mécanique quantique.
2. En 1931 parut une 2e édition très augmentée et une traduction anglaise «The theory of groups and quantum mechanics, Methuen, London, 1931»; édition livre de poche, Dover, New York, 1949.
3. Une traduction anglaise par S. J. Griffin, augmentée de 3 chapitres, a été éditée par Academic Press, New York, 1959.

4. Signalons aussi l'excellent, mais plus élémentaire, livre de E. Bauer, «Introduction à la théorie des groupes et ses applications à la physique quantique», Presses Universitaires de France, Paris, 1933.
5. J. von Neumann et Wigner E. P., «Zur Erklärung einiger Eigenschaften der Spektren aus der Quantenmechanik des Drehelektrons I.II.III»; *Z. Physik*, **47**, 203, **49**, 73, **51**, 844 (1928). Ces deux auteurs ont collaboré durant toute la vie du premier et ont publié sept articles ensemble sur des problèmes de physique mathématique.
6. Heisenberg W., «Über die Spektra von Atomsystem mit zwei Elektronen», *Z. Phys.*, **39** (1926), 499.
7. Parmi les traités classiques de mécanique quantique: Pauli W., Handbuch der Physik V. 1, Springer (1958) (réédition d'un livre écrit avant 1933); Dirac P. A. M., The Principles of Quantum Mechanics, Clarendon Press, Oxford, 1e éd. (1930), 4e édition (1958); J. von Neumann, Mathematische Grundlagen der Quantenmechanik (1930), Traduction anglaise, Princeton University Press, Princeton (1955). Des axiomatisations de la mécanique quantique ont notamment été proposées par: G. Birkhoff, J. von Neumann, The logic of Quantum Mechanics, *Ann. Math.*, **37** (935) 1936 (voir aussi Piron C., *Helv. Phys. Acta*, **37** (1964), 439 et par Segal I. E., «Postulates for general quantum mechanics», *Ann. Math.*, **48** (1947), 930. Citons enfin Mackey G. W., The Mathematical Foundation of Quantum Mechanics, Benjamin, New York, 1963.
8. L'espace $\mathscr{H}$ est sur le corps des complexes. Dès le premier chapitre de son traité (Cf. 7.) Dirac implique ce choix. Les physiciens ont aussi considéré le choix du corps des réels ou des quaternions.
9. En fait les algèbres de Jordan apparaissent plus naturelles comme algèbres d'observables et furent crées à cet effet. Jordan P., J. von Neumann, Wigner E., «On an algebraic generalization of the quantum mechanical formalism», *Ann. Math.*, **35** (1934), 29. Les physiciens préfèrent les algèbres associatives! Quelle topologie choisir? Plusieurs. Voir par exemple l'intéressant article de R. Haag et D. Kastler sur le rôle respectif de l'algèbre de von Neumann et de la C^*-algèbre d'observables: «An algebraic approach to quantum field theory», *J. Math. Phys.*, **5** (1964), 848. Haag, Araki et, à leur suite, Borchers, Ruelle, Doplicher, Dell'Antonio, etc., ont rénové cette approche algébrique.
10. En général notre information sur un système n'est que partielle et de nature probabiliste. Elle peut alors être représentée par un opérateur self-adjoint positif, de trace 1.
11. Dirac P. A. M., The Quantum Theory of the Electron, *Proc. Roy. Soc.*, **A117**, 510, **A178**, 351 (1928).
12. Le groupe de Lorentz est le sous-groupe du groupe linéaire réel à 4 dimensions laissant invariante la forme quadratique $t^2-r_1^2-r_2^2-r_3^2$. C'est un groupe de Lie simple, non compact, à 6 paramètres.
13. Wigner, réf. 5, p. 251-254; Bargmann V., *J. Math. Phys.*, **5** (1959), 852.
14. Wigner E., *Ann. Math.*, **40** (1939), 149.
15. Ce fut la première fois que fut caractérisée toute une famille de représentations linéaires, unitaires, continues, irréductibles, d'un groupe de Lie non compact. Complété par les travaux de I. M. Gelfand, M. A. Naimark, *Acad. Sci. USSR, J. Phys.*, **10**, 93 (1946), *Izv. Acad. Nauk USSR*, Ser. Mat. **11**, 911 (1947), et de Bargmann, *Ann. Math.*, **48** (1947), 568, le travail de Wigner permet d'obtenir toutes les représentations de $\overline{\mathscr{P}}_0$.
16. Lee T. D., Yang C. N., «Question of parity conservation in weak interactions», *Phys. Rev.*, **104** (1956), 254. Pour une anthologie d'articles originaux sur les interactions faibles (ed. Kabir) «The development of Weak Interaction Theory», Gordon and Breach, New York, 1963.

17. Pour les neutrinos, toutes les particules (resp. les antiparticules) sont polarisées circulairement à gauche (droite). La fixation des états «particule» par rapport à ceux «antiparticule» est conventionnelle et historique.
18. Tout cela est principalement l'œuvre de Pauli; cf. son dernier article sur ce sujet, «Exclusion Principle, Lorentz Group and Reflection of Space-Time and Charge», p. 30 in Niels Bohr, and the Development of Physics, Pauli (éd.), Pergamon Press, New York 1955, mais un assez grand nombre d'auteurs y participèrent et les idées générales ne se dégagèrent que lentement. Nous conseillons R. F. Streater, A. S. Wightman, «*PCT*, Spin and Statistics, and all that», Benjamin, New York, 1964 (et sa bibliographie) écrit dans le cadre de la théorie axiomatique des champs. Récemment H. Epstein (à paraître *J. Math. Phys.*) a prouvé le théorème «*PCT*» dans le cadre algébrique plus général de Haag et Araki.
19. Heisenberg W., *Z. Phys.*, **77** (1932), 1. Voir aussi Cassen B., Condon E. V., *Phys. Rev.*, **50** (1936), 846.
20. Heisenberg (réf. 24) l appela 5e degré de liberté du nucléon, Wigner (réf. 26) «isotopic spin». L'expression spin isobarique aurait mieux convenu. Le raccourcissement en isospin provient de l'évolution normale du langage.
21. Wigner E., *Phys. Rev.*, **51** (1937), 106.
22. Yukawa H., *Proc. Phys. Math. Soc. Japan*, **17** (1935), 48.
23. Wigner E., «On the law of conservation of heavy particles», *Proc. Nat. Ac. Sci. U.S.A.*, **38** (1952), 449.
24. Pour la relation entre charge et spin du point de vue de la théorie des groupes, voir Lurçat F., Michel L., *N. Cim.*, **27** (1951), 574.
25. Michel L., *N. Cim.*, **10** (1953), 319.
26. Feynman R. P., Gell-Mann M., *Phys. Rev.*, **109**, (1958), 193. Gell-Mann M., *Phys. Rev.*, **111** (1958), 362.
27. Gell-Mann M., *Phys. Rev.*, **125** (1962), 1067. Voir aussi *Physics*, **1** (1964) 63; voir aussi réf. 34.
28. Adler S. L., *Phys. Rev. Lett.*, **14** (1965), 1051; Weisberger W. I., *Ibid.*, 1047.
29. La rélation (14) fut dégagée par Gell-Mann M., *Suppl. N. Cim.*, **2** (1956), 848, et Nishijima K., *Prog. Theor. Phys.* (*Japan*), **13** (1955), 285.
30. Parmi les groupes qui ont cet algèbre comme algèbre de Lie, c'est U_2 qui convient; e. g. L. Michel dans «Group theoretical concepts and methods in elementary particle physics,» (éd. F. Gursey), Gordon Breach, N. Y., 1964.
31. Les principaux articles originaux sur ce sujet on été rassemblés dans l'anthologie: Ed. Gell-Mann M., Ne'eman U., «The Eightfold Way», Benjamin, New York, 1961.
32. Gell-Mann et Okubo indépendamment; voir réf. 35.
33. *Phys. Rev. Letters*, **12** (1964), 204, 33 noms d'auteurs! Typique du travail en équipe pour la physique expérimentale des hautes énergies.
34. Gürsey F. et Radicati L., *Phys. Rev. Letters*, **13** (1964), 299; Sakita B., *Phys. Rev.*, **136B** (1964), 1756. Voir aussi l'anthologie d'articles originaux, Dyson F., Ed., «Symmetry groups in nuclear and particle physics», Benjamin, New York, 1966.
35. Voir par exemple Segal I. E., *Duke Math. J.*, **18** (1951), 221, et Inonu E., Wigner E., *Proc. Nat. Acad. Sci. U.S.A.*, **39** (1953), 510 «contractant» les groupes.
36. Nous voulons signaler une conférence sur le même sujet que celle-ci et faite par un physicien à une audience de mathématiciens: Salam A., *J. London Math. Soc.*, **41** (1966), 49.
37. Beaucoup de questions n'ont pu être traitées ici; il me semble utile cependant de signaler la ligne suivante de travaux de physique utilisant les groupes.

La prédiction théorique des quantités mesurées en physique atomique ou nucléaire requiert le calcul multiple de produits de fonctions ayant une variance déterminée pour SU_2. Les physiciens ont alors dû créer un algorithme efficace pour la réduction des produits tensoriels $D_{j_1} D_{j_2} \dots D_{j_n}$. Pour les phénomènes à symétrie sphérique (ex.: calcul de l'énergie d'un niveau, probabilité de transition, corrélations angulaires successives à partir d'un état polarisé, etc.) les valeurs prédites sont des polynomes d'expression appelées coefficients de Racah et dont le type le plus simple est

$$\begin{Bmatrix} j_1 j_2 j_3 \\ j_4 j_5 j_6 \end{Bmatrix} = \int_{SU_2} \chi_1(\alpha)\chi_2(\beta)\chi_3(\gamma)\chi_4(\beta\gamma^{-1})\chi_5(\gamma\alpha^{-1})\chi_6(\alpha\beta^{-1})d\mu(\alpha)d\mu(\beta)d\mu(\gamma)$$

où $\chi_i(\alpha) = \operatorname{Tr} D_{j_i}(\alpha)$, caractère de la représentation D_{j_i}... On peut donc définir les coefficients de Racah pour toute une famille de groupes. On ne sait s'ils caractérisent le groupe. Sur l'ensemble de ce sujet, et sur d'autres questions apparentées d'états polarisés, donne plus d'informations. Toutes les prédictions s'expriment en termes des coefficients de Wigner, définis en fonction des éléments de matrices $(D_j(\alpha))_{mn}$ voir l'anthologie L. C. Biedenharn, H. Van Dam (éd.), «Quantum Theory of Angular Momentum», Academic Press, New York, 1965.

38. L'électrodynamique quantique permet de calculer à la précision relative de 10^{-6} tous les effets atomiques actuellement mesurés (on ne peut calculer effectivement au delà, mais il faudrait de toute façon sortir alors du cadre de la théorie pour tenir compte des effets nucléaires).
Cette excellente théorie est un défi aux mathématiciens, car elle n'est pas bien définie à leur sens. De nombreux traités exposent cette théorie. Citons simplement une anthologie d'articles originaux: Schwinger (éd.), Quantum Electrodynamics, Dover Press, New York, 1958.

О НЕКОТОРЫХ НЕЛИНЕЙНЫХ ЗАДАЧАХ ТЕОРИИ СПЛОШНЫХ СРЕД

О. А. ЛАДЫЖЕНСКАЯ

1. Об уравнениях параболического типа

Установлено, что при хороших коэффициентах существуют предельно точные оценки для решений весьма широкого класса линейных параболических систем и граничных условий. Первой оценкой такого типа было неравенство

$$\max_{0\leqslant t\leqslant T}\int_\Omega u_x^2(x,t)\,dx+\int_{Q_T}(u_t^2+u_{xx}^2)\,dx\,dt\leqslant$$
$$\leqslant c_1\int_\Omega u_x^2(x,0)\,dx+c\int_{Q_T}(\mathscr{L}u)^2\,dx\,dt, \qquad (1)$$

где

$$\mathscr{L}u=u_t-a_{ij}(x,t)\,u_{x_ix_j}+a_i(x,t)\,u_{x_i}+a(x,t)\,u,\quad Q_T=\Omega\times[0,T],$$

справедливое для любой функции $u(x, t)$, удовлетворяющей одному из основных однородных краевых условий. Затем были доказаны предельно точные оценки для $\mathscr{L}$ в пространствах $W_p^{2l,l}(Q_T)$, $p>1$ (В. А. Солонников) и в пространствах Гёльдера $C^{\alpha,\frac{\alpha}{2}}(\overline{Q}_T)$ (А. Фридман). Далее такого типа оценки доказывались для все более и более широких классов параболических систем и граничных условий. Наиболее общие результаты в этом направлении получены В. А. Солонниковым. Одновременно с этими оценками на их основе устанавливалась однозначная разрешимость начально-краевых задач.

В основе всех доказательств (кроме первоначального доказательства неравенства (1), данного автором в начале 50-х годов) лежит идея склейки Шаудера, позволяющая сводить эти проблемы к изучению канонических задач для уравнений с постоянными коэффициентами. Но для ее проведения *необходима непрерывность коэффициентов при старших производных*.

Исследования параболических уравнений и систем с *разрывными коэффициентами* базировались на энергетическом неравенстве. Оно имеет место для более узкого класса систем — для так называемых сильно параболических систем, в число которых входят параболические уравнения вида

$$u_t-\mathscr{L}^{(2m)}u=f(x,t),$$

где $\mathscr{L}^{(2m)}$ — эллиптический оператор порядка $2m$, $m\geqslant 1$, с дивергентной главной частью $\mathscr{L}^{(2m)}u=D_x^{(m)}(a_{(m)}(x,t)\,D_x^{(m)}u)+\dots$. Это неравенство позволяет доказывать существование обобщенных решений, имеющих производные $D_x^{(m)}u$ из $L_2(Q_T)$ и непрерывных по t в норме $L_2(\Omega)$ (свободный член при этом может быть довольно плохим, а $u_0(x)=u(x,0)\in L_2(\Omega)$). Другой информации о решении оно не дает, даже если f и u_0 суть очень гладкие функции.

Методы, которыми мы располагали до 1956—1957 гг., не позволяли сказать что-либо об улучшении дифференциальных свойств решений при улучшении свойств f и u_0 (при недифференцируемых старших коэффициентах $a_{(m)}(x,t)$) даже применительно к одному уравнению 2-го порядка. Начиная с известных работ Нэша и Де Джорджи стали вырабатываться новые методы, и это привело к открытию ряда новых закономерностей в линейных уравнениях 2-го порядка, а также способствовало изучению квазилинейных уравнений.

В совместных работах Н. Н. Уральцевой и автора исследованы уравнения

$$u_t-\frac{\partial}{\partial x_i}(a_{ij}(x,t)\,u_{x_j}+a_iu)+b_iu_{x_i}+au=f+\frac{\partial f_i}{\partial x_i} \qquad (2)$$

при условии, что старшие коэффициенты a_{ij} разрывны, но удовлетворяют условию

$$\nu\xi^2 \leqslant a_{ij}\xi_i\xi_j \leqslant \mu\xi^2, \qquad \nu, \mu = \text{const} > 0, \tag{3}$$

а коэффициенты при младших членах и свободные члены f и f_i принадлежат пространствам вида $L_{q_k, \infty}(Q_T)$ с разными q_k, т. е. имеют конечные нормы $\| u \|_{q, \infty, Q_T} = \sup\limits_{0 \leqslant t \leqslant T} \left(\int\limits_\Omega | u |^q dx \right)^{1/q}$. Выявленные при этом зависимости гладкости решений u уравнений (2) от значений q_k являются точными, что подтверждают специально построенные примеры. Возникшие при этом методы исследования уравнений (2), а также эллиптических уравнений 2-го порядка после некоторых модификаций позволили изучить и случаи более общей характеристики коэффициентов и свободных членов, а именно когда они суть элементы $L_{q_k, r_k}(Q_T)$; норма в $L_{q, r}(Q_T)$ определяется равенством $\| u \|_{q, r, Q_T} = \left[\int\limits_0^T \left(\int\limits_\Omega | u |^q dx \right)^{r/q} dt \right]^{1/r}$. Автором совместно с Н. Н. Уральцевой и двумя молодыми ленинградскими математиками А. В. Ивановым и А. Л. Трескуновым установлено, при каких q_k и r_k решения u уравнений (2) принадлежат $L_{q, r}$ с теми или иными q и r, когда для них конечен интеграл $\int \exp \{\lambda u(x, t)\}\, dx\, dt$, $\lambda > 0$, когда $\max | u | < \infty$ и когда конечна норма Гёльдера $| u |^{(\alpha, \alpha/2)}$.

Все это сделано для обобщенных решений уравнений (2) из пространства $V_2^{1, 0}(Q_T)$, получаемого пополнением множества гладких функций по норме

$$| u |_{Q_T} = \max_{0 \leqslant t \leqslant T} \| u \|_{2, \Omega} + \| u_x \|_{2, 2, Q_T}, \tag{4}$$

и подтверждено примерами, что доказанные закономерности являются точными (т. е. понижение показателей q_k, r_k в общем случае недопустимо). Как часто бывает, переход от классов $L_{q, \infty}$ ко всей шкале пространств $L_{q, r}$ позволил сделать доказательства и результаты более обозримыми и законченными. В качестве одной из доказанных теорем приведем следующий результат:

Теорема 1. Пусть для уравнения (2) выполнено условие (3) и

$$\sum_{i=1}^n a_i^2, \quad \sum_{i=1}^n b_i^2, \quad a \in L_{q, r}(Q_T), \quad \frac{1}{r} + \frac{n}{2q} \leqslant 1. \tag{5}$$

Тогда при $\sum\limits_{i=1}^n f_i^2 \in L_1(Q_T)$, $f \in L_{q_1, r_1}$, $\frac{1}{r_1} + \frac{1}{2q_1} \leqslant 1 + \frac{n}{4}$, $u_0 = u |_{t=0} \in L_2(\Omega)$ и $u |_S = 0$ для уравнения (2) однозначно разрешима первая начально-краевая задача в пространстве $V_2^{1, 0}(Q_T)$ (точнее в $V_2^{1, 1/2}(Q_T)$).

Если f_i и f обладают лучшими свойствами, т. е.

$$f \in L_{q_2, r_2}(Q_T), \quad \frac{1}{r_2} + \frac{n}{2q_2} \leqslant 1 + \frac{n}{4}\theta,$$
$$\sum_i f_i^2 \in L_{q_3, r_3}(Q_T), \quad \frac{1}{r_3} + \frac{n}{2q_3} \leqslant 1 + \frac{n}{2}\theta, \tag{6}$$

где $\theta \in (0, 1)$, то лучшими свойствами обладает и решение u, а именно:

$$u \in L_{\hat{q}, \hat{r}}, \quad \frac{1}{\hat{r}} + \frac{n}{2\hat{q}} = \frac{n}{4}\theta$$

(заметим, что из принадлежности u к $V_2^{1, 0}(Q_T)$ следует лишь принадлежность его к $L_{q, r}(Q_T)$ с $\frac{1}{r} + \frac{n}{2q} = \frac{n}{4}$). Если q и r из (5) удовлетворяют неравенству

$$\frac{1}{r} + \frac{n}{2q} < 1, \tag{7}$$

а θ в (6) равно 0, то $\int\limits_\varepsilon^T \int\limits_\Omega \exp [\lambda u(x, t)]\, dx\, dt < \infty$ с некоторым $\lambda > 0$ и любым $\varepsilon > 0$. Если $\sum\limits_i a_i^2$, $\sum\limits_i b_i^2$, a, $\sum\limits_i f_i^2$, $f \in L_{q, r}(Q_T)$ и q, r удовлетворяют (7), то решение u будет непрерывно по (x, t) в смысле Гёльдера.

При формулировке теоремы 1 не были указаны допустимые диапазоны изменения q и r; они зависят от размерности n.

Упомянем об одном из примеров, показывающих необходимость ограничений (5) на сингулярности коэффициентов. Функция

$$u(x, t) = e^{-|x|^2/4t}$$

является решением задачи Коши, равным нулю при $t = 0$, для уравнений

$$u_t - \Delta u + n \sum_{i=1}^n \frac{x_i}{|x|^2} u_{x_i} = 0, \tag{8}$$

$$u_t - \Delta u - \frac{n}{4t} u = 0. \tag{9}$$

Она заведомо принадлежит $V_2^{1, 1/2}$. Более того, из u можно сглаживанием по t и x построить почти классические решения (8) и (9).

Это говорит о недопустимости особенностей «силы» $1/|x|$ в коэффициентах b_i и $1/t$ в коэффициенте a. Условия (5) устраняют такие особенности.

Эти примеры дают основание думать, что слабое решение Хопфа для системы Навье — Стокса в случае трехмерной задачи неединственно. К этому мы вернемся в конце доклада. По уравнениям (2) с неограниченными коэффициентами имеются также работы Гуглielьмино и С. Н. Кружкова. Однако все упомянутые и сформулированные только что закономерности установлены лишь для уравнений и некоторых классов параболических систем 2-го порядка. В основе их получения лежит принцип максимума (хотя и в очень завуалированном и непривычном виде). Было бы интересным понять, имеют ли место аналогичные закономерности для уравнений высоких порядков. Кроме того, все, о чем мы говорили, относится к уравнениям вида (2). Для недивергентных же уравнений

$$u_t - Mu \equiv u_t - \sum_{i,j=1}^{n} a_{ij}(x, t)\, u_{x_i x_j} + a_i u_{x_i} + au = f \qquad (10)$$

с разрывными недифференцируемыми коэффициентами a_{ij} при $n \geqslant 2$ почти ничего неизвестно. В книге автора и Н. Н. Уральцевой по эллиптическим уравнениям на примере оператора

$$M_0 u \equiv a_{ij} u_{x_i x_j}, \quad a_{ij} = \delta_i^j + b\frac{x_i x_j}{|x|^2} \qquad (11)$$

выявлен ряд особенностей таких уравнений. Отметим еще одно их «неприятное» свойство: оператор M_0 при $n > 2$ не допускает замыкания в $L_2(\Omega)$, $\Omega = \{x : |x| > 1\}$. Действительно, для последовательности функций

$$u_{\varepsilon,\eta} = (|x|^2 + \eta)^{1-\frac{n}{4}+\varepsilon} - (|x|^2+\eta)^{1-\frac{n}{4}+\frac{\varepsilon}{2}}$$

отношение

$$\frac{\|u_{\varepsilon,\eta}\|_{2,\Omega}}{\|M_0 u_{\varepsilon,\eta}\|_{2,\Omega}} \to 0, \text{ а } \frac{M_0 u_{\varepsilon,\eta}}{\|M_0 u_{\varepsilon,\eta}\|_{2,\Omega}} \rightrightarrows v \in L_2(\Omega)$$

при $\varepsilon \to 0$ и $\eta = \eta(\varepsilon) \to 0$, если $b = \frac{n}{n-2}$. Пример этот построен автором совместно с А. Л. Трескуновым. В совместных работах Н. Н. Уральцевой и автора рассмотрены также квазилинейные уравнения вида

$$u_t - \frac{d}{dx_i}(a_i(x, t, u, u_x)) + a(x, t, u, u_x) = 0, \qquad (12)$$

$$u_t - a_{ij}(x, t, u, u_x)\, u_{x_i x_j} + a(x, t, u, u_x) = 0. \qquad (12')$$

Для уравнений (12) рассмотрены обобщенные и классические решения, для уравнений (12′) — в основном классические решения; при

этом исследования проведены и в направлении изучения гладкости всей совокупности решений уравнений, и в направлении доказательства однозначной разрешимости «в целом» основных краевых задач. По переменным u и p функции $a_i(x, t, u, p)$ и $a_{ij}(x, t, u, p)$ предполагаются гладкими, по переменным же (x, t) они могут иметь особенности — быть элементами разных $L_{q,r}$. Мы не будем перечислять полученные нами результаты. Они совместно с упомянутыми выше результатами по линейным уравнениям составили бóльшую часть книги О. А. Ладыженской, В. А. Солонникова и Н. Н. Уральцевой «Линейные и квазилинейные уравнения параболического типа», которая должна выйти в свет в начале 1967 г. Сформулируем лишь для примера один из результатов, относящихся ко всему классу квазилинейных уравнений (12′).

Теорема 2. Пусть u — произвольное обобщенное решение класса $\mathfrak{M}$ уравнений (12′), т. е. непрерывно, имеет обобщенные производные u_t и u_{xx} из $L_2(Q_T)$, его производные u_x ограничены по модулю и непрерывно зависят от t в норме $L_2(\Omega)$, и оно почти всюду удовлетворяет уравнению (12′). Пусть функции $a_{ij}(x, t, u, p)$ дифференцируемы по x, u и p в окрестности $u = u(x, t)$, $p = u_x(x, t)$, а на решении (т. е. при $u = u(x, t)$, $p = u_x(x, t)$) удовлетворяют условиям

$$\nu\xi^2 \leqslant a_{ij}(x, t, u, p)\,\xi_i\xi_j \leqslant \mu\xi^2, \quad \nu, \mu = \text{const} > 0,$$

$$\max_{Q_T}\left|\frac{\partial a_{ij}}{\partial p_k}\right| \leqslant \mu_1, \quad \left|\frac{\partial a_{ij}}{\partial u}, \frac{\partial a_{ij}}{\partial x_k}, a\right| \leqslant \varphi(x, t),$$

где $\|\varphi\|_{2q,2r,Q_T} \leqslant \mu_1$, причем $\frac{1}{r} + \frac{n}{2q} < 1$. Тогда u_x будут непрерывными по (x, t) в смысле Гёльдера и постоянная Гёльдера

$$\langle u_x\rangle_{Q'}^{(\alpha,\alpha/2)} \leqslant c(\max_{Q_T}|u_x|, n, \nu, \mu, \mu_1, q, r, d),$$

где d — расстояние Q' до боковой поверхности и нижнего основания Q_T, а

$$\alpha = \alpha(\max_{Q_T}|u_x|, n, \nu, \mu, \mu_1, q, r).$$

Этот результат, как было сказано, относится ко всему классу уравнений вида (12′), причем ограничения, наложенные на решение u и функции a_{ij} и a, вызваны существом дела (например, как показано в нашей книге по эллиптическим уравнениям, нельзя отбросить требование ограниченности $|u_x|$) и совместно с известными результатами по линейным уравнениям с гладкими коэффициентами сводит всю проблему получения априорных оценок для решений уравнений (12′) к получению оценок для $\max|u|$ и $\max|u_x|$.

Мы изучили уравнения (12), (12′) в предположениях: 1) их равномерной эллиптичности; для (12′) оно имеет вид

$$\nu(|u|)(1+|p|)^m\xi^2 \leqslant a_{ij}(x, t, u, p)\xi_i\xi_j \leqslant \mu(|u|)(1+|p|)^m\xi^2, \qquad (13)$$

и 2) непрерывной (а большей частью и гладкой) зависимости функций $a_i(x, t, u, p)$, $a_{ij}(x, t, u, p)$ и $a(x, t, u, p)$ от u и p. Именно такие уравнения были основным предметом исследований в нелинейных задачах до последнего времени. Их удалось изучить довольно хорошо. Результаты и методы, возникшие при этом, позволили исследовать и ряд задач механики, в которых имеются некоторые нарушения свойств 1) и 2), например задачи типа нестационарной фильтрации и уравнение Прандтля для пограничного слоя (в них есть вырождение формы $a_{ij}\xi_i\xi_j$, т. е. отклонение от (13)), задачи Стефана, в которых не выполнено свойство 2) (в них образующие уравнения функции разрывны по u) (см. в связи с этим работы А. Фридмана, О. А. Олейник, С. Л. Каменомостской, А. С. Калашникова, Е. С. Сабининой, Чжоу Юн-линь и др.). Некоторые другие задачи гидродинамики с неизвестными границами раздела разных сред (или течений) тоже могут быть, подобно задаче Стефана, переформулированы как задачи для уравнений вида (12) с разрывными по u функциями $a_i(x, t, u, p)$ и $a(x, t, u, p)$. Нам кажутся интересными исследования в направлении ослабления условий 1) и 2). Это может привести к открытию новых эффектов для параболических уравнений и потребовать новых приемов и методов. Интересно получить и обобщения указанных результатов по уравнениям 2-го порядка на уравнения высоких порядков и на системы. На результатах такого типа Н. Н. Уральцевой и автора по линейным и квазилинейным системам 2-го порядка мы останавливаться не будем.

Вместо этого обратимся к уравнениям Навье — Стокса, точнее к общей трехмерной нестационарной начально-краевой задаче для них.

2. Об уравнениях Навье — Стокса

Трехмерная начально-краевая задача для уравнений Навье — Стокса состоит в нахождении вектора скорости v и давления p из системы

$$\begin{aligned} &v_t - \nu\Delta v + v_k v_{x_k} = -\operatorname{grad} p + f, \\ &\operatorname{div} v = 0, \end{aligned} \qquad (14)$$

и условий

$$v|_{t=0} = v^0(x), \qquad v|_S = 0, \qquad (14')$$

где $x = (x_1, x_2, x_3)$, $v = (v_1(x, t), v_2(x, t), v_3(x, t))$, $p = p(x, t)$.

Несмотря на усилия многих математиков, начиная с исследований Ж. Лере в 30-х годах, вопрос о ее однозначной разрешимости «в целом» без каких-либо упрощающих предположений или предположений малости остается открытым. Десять лет тому назад мною и А. А. Киселевым была доказана однозначная разрешимость задачи (14), (14′) в двух случаях: во все моменты времени, если числа Рейнольдса малы в начальный момент времени и силы потенциальны (или мало отклоняются от потенциальных), и на малом интервале времени, если v^0 и f обладают некоторой гладкостью. Из первого результата легко выводится, что если при каком-либо v^0 и f решение существует на интервале $[0, T]$, то оно существует на этом интервале и при всех достаточно близких к v^0 и f начальных векторах и силах. Ввиду этого *для решения задачи* (14), (14′) *надо установить разрешимость (всегда имеется в виду разрешимость в таком пространстве, в котором есть единственность) лишь для какого-либо плотного множества начальных векторов и сил.*

Возможно, для доказательства такой разрешимости окажутся полезными рассмотрения всей совокупности траекторий — решений задачи (14), (14′) в духе теории динамических систем.

На задаче (14), (14′) математиками разных стран в последние годы были испытаны все методы решения начально-краевых задач, созданные при исследовании линейных уравнений. С каждым из них наиболее естественно связываются свои функциональные пространства, в терминах которых характеризуются решение и данные задачи. Делалось это, как мне кажется, в основном не для того, чтобы дать еще один вариант упомянутых выше теорем об однозначной разрешимости задач (14), (14′), а в надежде найти то функциональное пространство, в котором удастся доказать однозначную разрешимость «в целом» в общем случае. Надежды эти пока не оправдались. В каждом из вариантов доказана разрешимость задачи (14), (14′) в лучшем случае так же, как и в первом варианте, т. е. или для всех t, но при малых числах Рейнольдса R в начальный момент, или при произвольных R, но вблизи гладких начальных условий v^0 (форма этих условий в разных методах разная).

В связи с этим мне хочется высказать свою точку зрения относительно этих разнообразных попыток. В них стремятся «забить» нелинейные члены $v_k v_{x_k}$ с помощью главных линейных членов, т. е. $v_t - \nu\Delta v$, имея лишь единственную априорную оценку

$$\sup_{0\leqslant t\leqslant T} \|v(x, t)\|_{2,\Omega} + \|v_x\|_{2,2,Q_T}, \qquad (15)$$

которая следует из энергетического неравенства. Нам кажется, что это обречено на неудачу. В случае двух пространственных переменных такая попытка мне в свое время удалась, но это специфический двумерный эффект. При трех пространственных переменных

сопоставление «силы» линейных и нелинейных членов приводит к заключению, что нелинейные члены не подчиняются линейным. В. А. Солонниковым были доказаны предельно точные оценки в пространствах $W_p^{2l,\,l}(Q_T)$ и Гёльдера для оператора линейной части задачи (14), (14′). Их же можно доказать и в других пространствах с так называемыми дробными нормами. Во всех этих пространствах невозможно оценить нелинейные члены $v_k v_{x_k}$ через $v_t - \nu\Delta v$, имея (15) как единственную оценку «в целом». Более того, даже если взять упрощенную систему

$$v - \nu\Delta v + (v_k v)_{x_k} = 0, \tag{16}$$

отбросив давление и уравнение несжимаемости, и считать известной оценку (15), то и результаты последнего времени по параболическим уравнениям и системам, о которых говорилось выше и которые в определенных отношениях точны, не позволяют погасить влияние нелинейных членов и сделать вывод о разрешимости задачи Коши для (16) «в целом».

Во всех же работах по разрешимости задачи (14), (14′) был расчет на такое погашение, а оно, я думаю, невозможно в обычно употребляемых пространствах; и выбор метода, и выбор соответствующих ему функциональных пространств тут не спасет. Мне кажется, что усилия надо направить не в сторону анализа разных методов, а в сторону поисков новой априорной оценки «в целом», дополняющей оценку (15). Возможно, это будет «смесь» из принципа максимума для какой-то функции v и p и интегральных соотношений, в которых исчезают все нелинейные члены. Это удается сделать в ряде случаев, в которых вводится какое-либо дополнительное предположение о решении системы (14). Поясним это.

1) Предположим, что все компоненты скорости $v = (v^x, v^y, v^z)$, p и f не зависят от координаты z. Тогда система (14) эквивалентна системе

$$\Delta\psi_t - \nu\Delta^2\psi + \frac{\partial(\psi,\,\Delta\psi)}{\partial(x,\,y)} = F, \tag{17}$$

$$v_t^z + v^x v_x^z + v^y v_y^z - \nu\Delta v^z = f^z, \tag{18}$$

где $\psi = \psi(x, y, z)$ связана с v^x и v^y равенствами $v^x = \frac{\partial\psi}{\partial y}$, $v^y = -\frac{\partial\psi}{\partial x}$, а $\frac{\partial(u,v)}{\partial(x,y)} = u_x v_y - u_y v_x$. Уравнение (18) можно рассмотреть как линейное параболическое уравнение для v^z с коэффициентами v^x и v^y, в силу чего для v^z справедлив принцип максимума, причем существенно то, что величина $\max|v^z|$ зависит лишь от $\max|f^z|$ и $\max|v^z(x, y, 0)|$ и не зависит от коэффициентов v^x и v^y. Только через известные величины оценивается также и $\|v^z\|_{p,\,Q}$. Из уравнения (17) и граничных и начальных условий оцениваются, как показано в моих прежних работах, различные нормы ψ. Этих оценок достаточно, чтобы доказать однозначную разрешимость в целом задачи (14), (14′).

2) Пусть v^r, v^φ, v^z — цилиндрические компоненты вектора скорости v. Предположим, что все величины не зависят от угла φ. Система (14) эквивалентна системе

$$D\psi_t - \nu D^2\psi - r\frac{\partial\left(\psi,\,\frac{D\psi}{r^2}\right)}{\partial(r,\,z)} - \frac{\partial}{\partial z}\left(\frac{w}{r}\right)^2 = F_1, \tag{19}$$

$$w_t - \nu Dw - \frac{1}{r}\frac{\partial(\psi,\,w)}{\partial(r,\,z)} = F_2, \tag{20}$$

где F_i — известные функции, а $D\psi = \psi_{zz} + r\frac{\partial}{\partial r}\left(\frac{1}{r}\frac{\partial\psi}{\partial r}\right)$. Компоненты v выражаются через ψ и w так:

$$v^z = -\frac{1}{r}\frac{\partial\psi}{\partial r}, \quad v^r = \frac{1}{r}\frac{\partial\psi}{\partial z}, \quad w = rv^\varphi.$$

Величина $D\psi = r\omega^\varphi$, а $\omega = (\omega^r, \omega^\varphi, \omega^z)$ — ротор вектора v. Для решений системы (19), (20), удовлетворяющих условиям

$$w|_S = v^\varphi|_S = 0, \quad \psi|_S = 0, \quad D\psi|_S = \omega^\varphi|_S = 0, \tag{21}$$

на границе S ограниченной области Ω, являющейся телом вращения вокруг оси z, можно, помимо известного энергетического неравенства, дать оценки более сильных норм, которых уже достаточно для доказательства однозначной разрешимости в целом задачи (19)—(21). Так из (20) и (21) оцениваются разные интегральные нормы w и $\max|w|$, а из (19) и (21) оценивается величина

$$\max_{0\leqslant t\leqslant T}\int_\Omega \frac{(D\psi)^2}{r^3}\,dr\,dz + \int_{Q_T}\left(D\frac{\partial\psi}{\partial z}\right)^2\frac{1}{r^3}\,dr\,dz.$$

Второй случай в отличие от первого нельзя считать двумерным, если ось z входит в область, заполненную жидкостью. Но упрощающее предположение о независимости от φ было существенно использовано. Я не буду приводить здесь другие случаи, ибо уже на этих двух я пояснила свою мысль, какого рода априорные оценки следует искать. При этом *для доказательства однозначной разрешимости задачи* (14), (14′) *в целом при* $n = 3$ *нужно сравнительно небольшое усиление оценки* (15), *например достаточно иметь априорную оценку*

$$\max_{0\leqslant t\leqslant T}\int_\Omega |v|^q dx \text{ с } q > 3, \text{ или } \|v\|_{q,\,r,\,Q_T} \text{ с } \frac{1}{r} + \frac{3}{2q} \leqslant \frac{1}{2},\ q \in (3,\ \infty],$$

$r \in [2, \infty)$.

Для обобщенных решений Хопфа, имеющих конечной одну из только что указанных норм, имеет место теорема единственности. Она справедлива также и в классе обобщенных решений v из пространства

$$L_{q,r}(Q_T), \quad \frac{1}{r}+\frac{3}{2q}\leqslant\frac{1}{2}, \quad 4\leqslant q\leqslant\infty, \quad 4\leqslant r\leqslant\infty.$$

Существование производных v_x для таких решений не предполагается.

Вернемся к высказанному выше предположению о неединственности слабого решения Хопфа в задаче (14), (14′). Относительно этих решений известна лишь конечность нормы v из (15). Разность u двух возможных таких решений v' и v'' удовлетворяет однородной линейной системе, в которой имеются члены вида $u_k w_{x_k}$, где $w=\frac{1}{2}(v'+v'')$. В этих членах компоненты w играют роль коэффициентов. О них известно лишь, что они суть элементы V_2 (т. е. имеют конечную норму из (15)). Но, как показано на примерах по параболическим уравнениям (см. § 2, гл. I упомянутой выше книги по параболическим уравнениям), таких свойств коэффициентов недостаточно для теорем единственности решений из $V_2^{1,0}$ (и даже почти классических) в случае трехмерного пространства. Это и есть довод в пользу моего предположения о неединственности решений Хопфа при $n=3$.

Укажем на совместную работу А. Кржевицкого и автора, в которой построены две конечно-разностные схемы для определения решений задачи (14), (14′) и доказана их сходимость к слабому решению Хопфа. За неимением времени я не буду приводить их здесь, поскольку эта работа появится в Трудах Математического института АН СССР им. В. А. Стеклова, обращаю только внимание на то, что в соответствующей заметке в ДАН СССР имеются опечатки в схеме.

3. О новых уравнениях для описания движений вязких несжимаемых жидкостей

Мне кажется, что движения вязких несжимаемых жидкостей разумно описывать одной из следующих систем:

$$\left.\begin{aligned} &v_t-\frac{\partial}{\partial x_k}\left[(\nu_0+\nu_1 v_x^2)\,v_{x_k}\right]+v_k v_{x_k}=-\operatorname{grad}p+f,\\ &\operatorname{div}v=0,\end{aligned}\right\} \tag{22}$$

$$\left.\begin{aligned} &v_t+\operatorname{rot}\left[(\nu_0+\nu_1\operatorname{rot}^2 v)\operatorname{rot}v\right]+v_k v_{x_k}=-\operatorname{grad}p+f,\\ &\operatorname{div}v=0,\end{aligned}\right\} \tag{23}$$

или

$$\left.\begin{aligned} &Mv\equiv v_t-\nu(v_x)\,\Delta v+v_k v_{x_k}=-\operatorname{grad}p+f,\\ &\operatorname{div}v=0,\end{aligned}\right\} \tag{24}$$

где ν_0 и ν_1 — положительные постоянные, а

$$\nu(v_x)=\nu_0+\nu_1\int\limits_{\Omega}v_x^2(x,t)\,dx \quad\text{или}\quad \nu(v_x)=\nu_0+\nu_1\int\limits_{\Omega}\operatorname{rot}^2 v(x,t)\,dx.$$

Для этих систем имеет место однозначная разрешимость в целом начально-краевых задач. Так, например, для системы (24) справедливо следующее предложение:

Теорема. Задача

$$\left.\begin{aligned} &Mv=-\operatorname{grad}p+f, \quad \operatorname{div}v=0,\\ &v|_S=0, \quad v|_{t=0}=a(x)\end{aligned}\right\} \tag{25}$$

имеет единственное обобщенное решение v с конечной нормой $\max\limits_{0\leqslant t\leqslant T}\|v\|_{2,\Omega}+\|v_x\|_{2,4,Q_T}$, непрерывное по t в норме $L_2(\Omega)$ и такое, что $h^{-1}\|v(x,t+h)-v(x,t)\|^2_{2,2,Q_{T-h}}\to 0$ при $h\to 0$, если только $a(x)\in\overset{\circ}{J}(\Omega)$, а $f\in L_{2,1}(Q_T)$. Если $a(x)\in W_2^1(\Omega)$, $\operatorname{div}a=0$, $a|_S=0$ и $f\in L_{2,1}(Q_T)$, $f_t\in L_{2,1}(Q_T)$, то это решение v имеет конечную норму $\max\limits_{0\leqslant t\leqslant T}\|v_t\|_{2,\Omega}+\|v_{tx}\|_{2,2,Q_T}$. Если к тому же f непрерывно по (x,t) в Q_T в смысле Гёльдера, а граница S дважды ограниченно дифференцируема, то это решение v будет классическим, т. е. непрерывным в $\bar{Q}_T$ и имеющим наравне с p непрерывные в Q_T производные, входящие в систему (25).

Для стационарных систем, соответствующих системам (22)—(24), разрешимы в целом краевые задачи в ограниченных областях при произвольном граничном режиме $a|_S=v|_S$, удовлетворяющем лишь необходимому условию $\sum\limits_{i=1}^{N}\int\limits_{S_i}(a,n)\,ds=0$, где $\bigcup\limits_{i=1}^{N}S_i=S$. Если Ω неограничена и содержит полную окрестность бесконечно удаленной точки, то в Ω разрешима в целом такая же краевая задача при условии, что на бесконечности скорость стремится к заданному значению v_∞.

Решения v начально-краевых задач для систем (1)—(3) устойчивы на любом конечном промежутке времени по отношению к возмущениям внешних воздействий f, а также начальных и граничных значений. Из предложений о поведении решений v этих задач при $t\to\infty$ приведем для примера два: 1) если $\int\limits_0^\infty\|f(x,t)\|_{2,\Omega}\,dt<\infty$ и $v|_S=0$, то для

решений v норма $\| v(x, t) \|_{2,\Omega} \to 0$ при $t \to 0$; 2) пусть v' и v'' решения системы (24), равные нулю на S и соответствующие $f = f'$ и $f = f''$. Если, начиная с некоторого t_1

$$\nu_0 + \frac{\nu_1}{2} \| v_x''(x, t) \|_{2,\Omega}^2 - \frac{2}{\sqrt{\beta}} \| v_x''(x, t) \|_{2,\Omega} \geqslant \alpha > 0,$$

где $\beta = \min\limits_{v \in \mathring{W}_2^1(\Omega)} \dfrac{\| v_x \|_{2,\Omega}}{\| v \|_{2,\Omega}}$, и $\int\limits_0^\infty \| f' - f'' \|_{2,\Omega}\, dt < \infty$, то $\| v'(x, t) - v''(x, t) \|_{2,\Omega} \to 0$ при $t \to \infty$.

В пользу систем (22)—(24) говорит несколько фактов:

1) они ненамного сложнее системы Навье — Стокса и при $\nu_1 = 0$ совпадают с ней;

2) в противоположность системе Навье — Стокса для них удается доказать однозначную разрешимость в целом начально-краевых задач, т. е. то их свойство, которым должно обладать любое математическое описание детерминированного физического процесса;

3) они инвариантны по отношению к галилеевым преобразованиям x, t;

4) при выводе уравнений Навье — Стокса из уравнений Максвелла — Больцмана постулируют постоянство температуры T в потоке. Если этого не делать, то в случае несжимаемой жидкости и отсутствия внешних воздействий получается система пяти уравнений

$$v_t - \frac{\partial}{\partial x_k}\left(c \sqrt{T}\, v_{x_k}\right) + v_k v_{x_k} = -\operatorname{grad} p, \tag{26}$$

$$\operatorname{div} v = 0, \tag{27}$$

$$T_t - \frac{\partial}{\partial x_k}\left(c_1 \sqrt{T}\, T_{x_k}\right) + v_k T_{x_k} - \frac{c}{2} \sqrt{T} \sum_{k,\, l=1}^{3} (v_{kx_l} + v_{lx_k})^2 = 0, \tag{28}$$

для пяти неизвестных величин v, p и T. В ней c и c_1 — положительные постоянные, характеризующие среду. Система (26)—(28) довольно сложная. Для нее я не умею доказывать однозначную разрешимость в целом начально-краевых задач. Естественно возникло желание упростить ее (кстати заметим, что уравнению (28) функция $T = \text{const}$, вообще говоря, не удовлетворяет).

Используя принцип максимума, нетрудно доказать, что если уравнение (28) удовлетворяется во всем пространстве, если $T(x, 0) > 0$ и T, T_x и v ограничены в полосе $\Pi_{t_1} \{(x, t) : x \in E_3$,

$t \in [0, t_1]\}$, то T положительно в Π_{t_1} и для любой точки (x, t) из Π_{t_1}

$$\min_{E_3} \sqrt{T(x, 0)} + t \min_{\Pi_{t_1}} \frac{c}{4} \sum_{k,\, l=1}^{3} (v_{kx_l} + v_{lx_k})^2 \leqslant \sqrt{T(x, t)} \leqslant$$

$$\leqslant \max_{E_3} \sqrt{T(x, 0)} + t \max_{\Pi_{t_1}} \frac{c}{4} \sum_{k,\, l=1}^{3} (v_{kx_l} + v_{lx_k})^2. \tag{29}$$

Эта оценка говорит в пользу того, что вместо $c\sqrt{T}$ в уравнении (26) естественно взять или $\nu_0 + \nu_1 v_x^2$, $\nu_0, \nu_1 > 0$, или усредненную по всем x величину

$$\nu_0 + \nu_1 \int\limits_{\Omega} v_x^2\, dx = \nu_0 + \frac{\nu_1}{2} \int\limits_{\Omega} \sum_{k,\, l=1}^{3} (v_{kx_l} + v_{lx_k})^2\, dx.$$

Система (23) написана по аналогии с системой (22): она обладает свойствами 1) — 3), и нелинейная добавка $\nu_1 \operatorname{rot}^2 v$ имеет тот же характер, что и v_x^2.

Отметим следующее: мы доказали предложения, аналогичные перечисленным в этом параграфе, для систем, несколько более общих, чем системы (22) — (23), а именно для систем вида

$$\left.\begin{aligned} &v_t - \frac{\partial}{\partial x_k}\left[(\nu_0 + \nu_1 |v_x|^{2\mu})\, v_{x_k}\right] + v_k v_{x_k} = -\operatorname{grad} p + f, \\ &\operatorname{div} v = 0, \end{aligned}\right\} \tag{30}$$

$$\left.\begin{aligned} &v_t + \operatorname{rot}\left[(\nu_0 + \nu_1 |\operatorname{rot} v|^{2\mu}) \operatorname{rot} v\right] + v_k v_{x_k} = -\operatorname{grad} p + f_1, \\ &\operatorname{div} = 0, \end{aligned}\right\} \tag{31}$$

которые при $\mu = 1$ совпадают с системами (22) и (23) соответственно. Для них однозначная разрешимость начально-краевых задач доказана при $\mu \geqslant 1/4$, а разрешимость стационарных краевых задач при любом $\mu > 0$.

Ленинградский университет,
Ленинград, СССР

13 Математические проблемы управляющих систем
Mathematical problems of control systems
Problèmes mathématiques des systèmes de contrôle
Mathematische Probleme der Regelungssysteme

NETWORKS OF GAUSSIAN CHANNELS WITH APPLICATIONS TO FEEDBACK SYSTEMS[1])

PETER ELIAS

Summary. This paper discusses networks (directed graphs) having one input node, one output node, and an arbitrary number of intermediate nodes, whose branches are noisy communications channels, in which the input to each channel appears at its output corrupted by additive Gaussian noise. Each branch is labeled by a non-negative real parameter which specifies how noisy it is. A branch originating at a node has as input a linear combination of the outputs of the branches terminating at that node.

The channel capacity of such a network is defined. Its value is bounded in terms of branch parameter values and procedures for computing values for general networks are described. Explicit solutions are given for the class D_0 which includes series-parallel and simple bridge networks and all other networks having r paths, b branches and v nodes with $r = b - v + 2$, and for the class D_1 of networks which is inductively defined to include D_0 and all networks obtained by replacing a branch of a network in D_1 by a network in D_1.

The general results are applied to the particular networks which arise from the decomposition of a simple feedback system into successive forward and reverse (feedback) channels. When the feedback channels are noiseless, the capacities of the forward channels are shown to add. Some explicit expressions and some bounds are given for the case of noisy feedback channels.

Introduction

The min-cut max-flow theorem [1, 2, 3] gives the capacity of a network made up of branches of given capacity. It applies to networks of noisy communications channels if the assumption is made that arbitrarily large delays and arbitrarily complex encoding and decoding operations may take place at each interior node.

This paper presents the theory of networks of another kind of channel — a channel with additive gaussian noise, for which the only operation which takes place at a node is linear combination of the arriving signal and noise voltages, with no significant delay and no decoding or recoding.

[1]) The work was supported by the Joint Services Electronics Program (Contract DA36-039-AMC-03200 (E)), and the National Aeronautics and Space Administration (Grant NsG-334).

The problem

Consider the class D of two-terminal networks like that shown in Fig. 1, in which there are no cycles, each of the b branches B_i is directed, and each branch lies on one of the r paths R_i which go from the input terminal on the left to the output terminal on the right. A signal voltage e_0 of mean square value P_0 (the *signal power*) is applied to the input terminal, node V_1 at the left. At each interior node, the output (signal plus noise) of each branch B_i arriving at the node is given a (positive or negative) weight, the *branch transmission* t_i, and the resulting linear combination of signal and noise voltages is supplied as input to all the branches leaving that node.

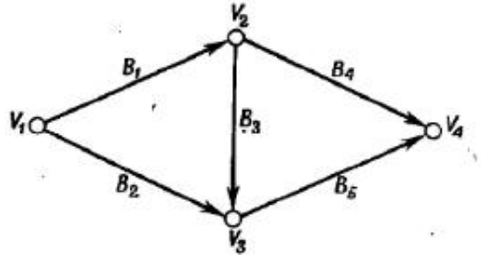


Fig. 1. A network in class D.

$$\text{PATHS} \begin{cases} R_1 = B_1B_4 \\ R_2 = B_1B_3B_5 \\ R_3 = B_2B_5 \end{cases} \qquad \text{PATH TRANSMISSIONS} \begin{cases} T_1 = t_1t_4 \\ T_2 = t_1t_3t_5 \\ T_3 = t_2t_5 \end{cases}$$

$$\|G_{ij}\| = \begin{Vmatrix} (1+N_1)(1+N_4) & (1+N_1) & 1 \\ (1+N_1) & (1+N_1)(1+N_3)(1+N_5) & (1+N_5) \\ 1 & (1+N_5) & (1+N_2)(1+N_5) \end{Vmatrix}$$

Each branch B_i adds to its input voltage e_i a gaussian noise voltage n_i whose mean square value (the *noise power*) is a constant N_i (the *noise to signal power ratio*, also called the *parameter* of the branch) times the mean square value P_i (the *input power*) of its input voltage. The noise voltage in each branch is statistically indepen-

dent of the noise voltages in other branches and of the signal voltage:

$$\overline{e_i^2} = P_i,\ 0 \leqq i \leqq b;\ \overline{n_i^2} = N_i P_i,\ 1 \leqq i \leqq b, \tag{1}$$
$$\overline{n_i n_j} = 0,\ i \neq j;\ \overline{n_i e_0} = 0.$$

Since the branch input voltage and its noise are uncorrelated, the mean square value of the branch output voltage (the *output power*) is just

$$\overline{(e_i + n_i)^2} = \overline{e_i^2} + \overline{n_i^2} = P_i + N_i P_i = P_i(1 + N_i). \tag{2}$$

The power output of each branch generator depends on the power level at its input, and thus on the power level of the signal and of all other noise generators which affect its input power, as well as on the values of the branch transmissions. However once the power levels of the signal and of all noises and the values of the branch transmissions are fixed, the network is linear. The final output at the right hand output terminal V_v is a linear combination of the b branch noise generator voltages and the signal voltage e_0. We constrain the values of the branch transmissions t_i by requiring that the coefficient of e_0 in this sum be unity.

The network is equivalent to a single branch (noisy channel) of the same kind as the component branches, since the linear combination of the b branch noise voltages which appears in the output is a gaussian noise voltage, and the overall action of the two-terminal network is to receive an input signal and to produce at its output the input signal plus an independent gaussian noise. The ratio of output noise power to signal power, N_{b+1}, is a function of the branch transmissions as well as the parameters N_i of the network branches: the *optimum noise to signal power* of the network, N_{opt}, is defined as the minimum value of N_{b+1} which can be obtained by varying the branch transmissions. The problem is to find N_{opt} as a function of the given N_i.

Series and parallel networks

To express the results most simply in important special cases it is convenient to associate with each branch, not only the parameter N_i, but the *signal to noise ratio*,

$$S_i = 1/N_i, \tag{3}$$

and the *capacity* per use of the channel

$$C_i = \frac{1}{2}\log(1 + S_i). \tag{4}$$

Equivalent quantities are defined for the network: S_{opt} is the maximum signal to noise ratio attainable by varying the branch transmissions, and C_{opt} is the largest channel capacity so attainable.

We can then state three results.

Series Networks. A network in D in which all b branches are in series has N_{opt} given by

$$1 + N_{\text{opt}} = \coprod_{i=1}^{b} (1 + N_i). \tag{5}$$

Parallel Networks. A network in D in which all b branches are in parallel has S_{opt} given by

$$S_{\text{opt}} = \sum_{i=1}^{b} S_i. \tag{6}$$

Duality. Given two channels of capacities C_1 and C_2. Let the optimum capacity of the network consisting of the two channels in series be C_s. Let the optimum capacity of the two channels connected in parallel be C_p. Then

$$C_1 + C_2 = C_s + C_p. \tag{7}$$

The result on series networks expressed by Eq. (5) does not seem to have been published. The result for parallel branches expressed by Eq. (6) is known as optimum diversity combining, or the ratio squarer [4,5] and was discovered independent of the general theory. Both follow directly from the general results below. The duality relationship of Eq. (7) follows directly from Eqs. (4), (5) and (6), and also seems not to have been published. We have

$$\begin{aligned} C_s &= \frac{1}{2}\log\left(1 + \frac{1}{N_s}\right) = \frac{1}{2}\log\left(1 + \frac{1}{(1+N_1)(1+N_2)-1}\right) = \\ &= \frac{1}{2}\log\left(\frac{(1+N_1)(1+N_2)}{N_1+N_2+N_1N_2}\right) = \frac{1}{2}\log\left(\frac{(1+S_1)(1+S_2)}{1+S_1+S_2}\right) = \\ &= \frac{1}{2}\log(1+S_1)(1+S_2) - \frac{1}{2}\log(1+S_1+S_2) = \\ &= C_1 + C_2 - C_p. \end{aligned}$$

Eq. (7), incidentally, also holds for other pairs of channels, such as two binary symmetric channels with different crossover probabilites p_1 and p_2, or a binary symmetric channel in series with a binary erasure channel and in parallel with it. The interpretation of parallel channels is different in those cases however: it involves having the receiver observe the outputs of both channels when a common input symbol is applied to both. Since the output symbols of the two channels cannot be combined into an input symbol for the same kind of channel without loss of information, there is no tidy network theory for such channels and we discuss them no further.

Feedback networks

The next results apply to a subset F of networks in D which represent a dissection in space of the time sequence of forward and return signal flows encountered in a feedback system, as shown in Figs. 2 and 3. The transmitter applies a signal voltage to the input node V_1.

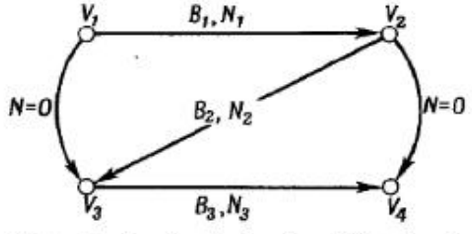


Fig. 2. A network in class F for $k=2$.

It proceeds over a noisy branch B_1 to node V_2 at the receiver. The receiver sends it back over B_2 to V_3 at the transmitter. The transmitter forms a linear combination of the original signal and the noisy version

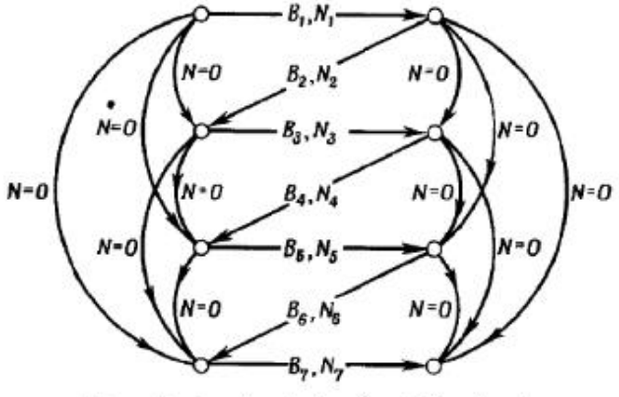


Fig. 3. A network in class F for $k=4$.

of it received at V_3 and transmits it over B_3 to V_4. In Fig. 2 the receiver then takes a linear combination of the outputs at V_2 and V_4 as the output voltage: in Fig. 3 the process continues. In both Figures, and for all nets in F, the branches on the left connecting odd-numbered nodes and the branches on the right connecting even-numbered nodes are noiseless, and serve only to provide linear combinations of previously received values for the next transmission and to provide the

requisite delay. Odd-numbered branches, from odd to even nodes, are called *forward* channels; even-numbered branches, from even to odd nodes, are *feedback* channels.

The Uniform Delay Property. In practice delays will be introduced by the forward and feedback channels. In order to avoid having signal voltage samples applied at different times getting mixed up at intermediated nodes, we assume that the noiseless branches on the left and on the right have delays so selected as to give the network the *uniform delay property* that all paths connecting any two nodes have the same delay, so that at any node only one signal sample and one sample of the output of each earlier noise generator will arrive at a given time over different paths. This can be accomplished for any network in F, or indeed in D, if an initial set of delay values d_i are given for the branches B_i, by increasing some of them in the following fashion. Assign a delay value to each node V_j equal to the maximum delay obtained by adding the delay values of the branches along each path from V_1 to V_j. Then assign to B_i the new delay value d_i' which is the difference between the delay values of its terminal and initial nodes; $d_i' \geqq d_i$.

We will henceforth assume that this process has been carried out for all networks in F or D, and that all have the uniform delay property. It is then not necessary to keep track of the delay values of networks or branches. We now state results for feedback networks.

Noiseless Feedback. For a network in F, if all feedback branches are noiseless, and the k forward branches have capacities C_{2j-1}, $1 \leqq j \leqq k$, then the optimum capacity of the network is given by

$$C_{\text{opt}} = \sum_{j=1}^{k} C_{2j-1} \tag{8}$$

and the optimum signal-to-noise ratio S_{opt} by

$$1 + S_{\text{opt}} = \prod_{j=1}^{k} (1 + S_{2j-1}). \tag{9}$$

In particular, if

$$S = \sum_{j=1}^{k} S_{2j-1}$$

is fixed, but an arbitrarily large k is available, we have

$$1 + S_{\text{opt}} = \lim_{k \to \infty} \prod_{j=1}^{k} \left(1 + \frac{S}{k}\right) = e^S, \text{ or} \tag{10}$$

$$S_{\text{opt}} = e^S - 1.$$

Noisy Feedback, $k=2$. For a network in F with two noisy forward channels B_1 and B_3 and one noisy feedback channel B_2, the optimum signal to noise ratio is

$$S_{\text{opt}} = S_1 + S_3 + \frac{S_1 S_2 S_3}{(1+S_1)(1+S_3)+S_2}. \tag{11}$$

Unfortunately a general formula like Eq. (11) for a net in F with $k>2$ is not available, although the computation of S_{opt} for any particular case is a straightforward numerical analysis problem. However we do have some inequalities which hold for all nets in F and which provide some insight.

Noisy Feedback, general case. For a network in F with k noisy forward branches B_{2j-1}, $1 \leqq j \leqq k$, and $k-1$ noisy feedback branches B_{2j}, $1 \leqq j \leqq k-1$, the optimum signal to noise ratio S_{opt} is bounded by

$$S_{\text{opt}} \geqq \sum_{j=1}^{k} S_{2j-1} \tag{12}$$

$$1 + S_{\text{opt}} \leqq \prod_{j=1}^{k} (1 + S_{2j-1}) \tag{13}$$

$$S_{\text{opt}} \leqq \sum_{j=1}^{2k-1} S_j. \tag{14}$$

If signal to noise ratio costs c_1 per unit for forward channels and c_2 per unit for feedback channels, so that the total cost for a network in F is

$$c = c_1 \sum_{j=1}^{k} S_{2j-1} + c_2 \sum_{j=1}^{k-1} S_{2j}$$

then for sufficiently large S_{opt}, the cost per unit of S_{opt} may be made arbitrarily close to c_2:

$$\frac{c}{S_{\text{opt}}} \leqq c_2 (1+\delta). \tag{15}$$

The results for noiseless feedback and for noisy feedback with $k=2$ were published by the author some time ago [6,7]. Schalkwijk and Kailath have recently investigated the noiseless case from the point of view of error probability for the transmission of discrete messages [8, 9, 10]. Turin [14] has also dealt with a closely related question. The noiseless feedback results of Eqs. (8) and (9) are remarkable, since they permit the transmission of a continuous signal of fixed bandwidth over a noisy channel at a rate equal to channel capacity no matter how large the bandwidth of the forward channel, with no coding or decoding needed, provided that a noiseless feedback channel is available. Furthermore they do so without introducing any of the discontinuities which must occur when a continuous signal is mapped onto a space of higher dimensionality — discontinuities which were pointed out by Shannon [11] and Kotelnikov [12] and have recently been discussed by Wozencraft and Jacobs [13]. Eq. (10) implies that a signal to noise ratio of 10 in bandwidth W is equivalent to a signal to noise ratio of e^{10}:1, or about 22,000 if the available forward channel is wideband and has white noise and a noiseless feedback channel is available: see [6,7] for further discussion.

The inequality (12) follows from the parallel network result of Eq. (6). The result of setting all feedback channel transmissions at zero and using the forward channels in parallel gives the right side of Eq. (12), and the optimum choice of branch transmissions must do at least as well. The second inequality, Eq. (13), says that noise in the feedback channels does not help: the right side is just the noiseless feedback result of Eq. (9). It is a consequence of a more general result which is given below, and which shows that increasing N_i in any branch cannot decrease N_{opt}. The third result, Eq. (14), says that given a choice, it is better to use signal to noise ratio in the forward rather than the feedback channels: the total S_{opt} attainable by feedback is less than would be attained by taking all of the feedback channels, turning them around, and using them in parallel with the forward channels, which gives the right side of Eq. (14) by Eq. (6). It will also be derived later. The final result, Eq. (15), shows why feedback is interesting even if it does not do as well as the same amount of signal to noise ratio in the forward direction would do, by Eq. (14). Signal to noise ratio in the feedback direction may be cheaper, as when a satellite is communicating to Earth, and if it is, it is possible by means of feedback to buy forward signal to noise ratio at the same cost, if one wants enough of it. Eq. (15) is a direct consequence of Eq. (11): it is necessary only to choose S_1 equal to S_3, and S_2 so large that it is possible to have $S_1 \ll S_2$ and $S_1^2 \gg S_2$ at the same time. For $k>2$ the result will be of the same character but better — i.e. a smaller δ will do, or a smaller amount of S_{opt} can be bought at the same unit cost but the absence of a formula makes the demonstration harder.

General results

To state and prove the theorem from which the above results follow we need some further definitions. For each pair of paths R_i, R_j from V_1 to V_2 in a network in D, we define G_{ij} as a product which contains one factor $(1+N_k)$ for each branch B_k which lies in both paths; if R_i and R_j share no branches, $G_{ij}=1$. Formally, if we treat the symbol R_i as denoting the set of branches which are contained in the i-th

path, then $R_i \cap R_j$ is the set of branches which the two paths have in common, and

$$G_{ij} = \prod_{k:\, B_k \in R_i \cap R_j} (1 + N_k), \quad G_{ij} = 1 \text{ for } R_i \cap R_j \text{ empty.} \tag{16}$$

We also define the *path transmission* T_i of path R_i as the product of the branch transmissions t_k for those branches which lie on R_i:

$$T_i = \prod_{k:\, B_k \in R_i} t_k. \tag{17}$$

The *network transmission* $T_{0,\,b+1}$ is the sum of all path transmissions: by the assumption made in the discussion following Eq. (2), the branch transmissions t_k are constrained so that the network transmission, which is the coefficient of the signal voltage e_0 in the output, is unity:

$$T_{0,\,b+1} = \sum_{i=1}^{r} T_i = 1. \tag{18}$$

T h e o r e m. For any network in D, we have

$$1 + N_{\text{opt}} = \min_{t_k} \Big\{ \sum_{i=1}^{r} \sum_{j=1}^{r} G_{ij} T_i T_j \Big\} \geqq 1 \Big/ \Big\{ \sum_{i=1}^{r} \sum_{j=1}^{r} G_{ij}^{-1} \Big\} \tag{19}$$

and

$$S_{\text{opt}} = 1 \Big/ \min_{t_k} \Big\{ \sum_{i=1}^{r} \sum_{j=1}^{r} (G_{ij} - 1)\, T_i T_j \Big\} \leqq \sum_{i=1}^{r} \sum_{j=1}^{r} [G_{ij} - 1]^{-1}, \tag{20}$$

where the T_i are given in terms of the t_k by Eq. (17) and are subject to the constraint (18), and G_{ij}^{-1} and $[G_{ij} - 1]^{-1}$ are elements of the inverses of the matrices $\| G_{ij} \|$ and $\| G_{ij} - 1 \|$. The inverses of $\| G_{ij} \|$ and $\| G_{ij} - 1 \|$ always exist unless there is at least one noiseless path from input to output, so that some $G_{ii} = 1$. In this case $N_{\text{opt}} = 0$ and $S_{\text{opt}} = \infty$: these values are attained by setting $T_i = 1$ and all other $T_j = 0$, $j \neq i$.

Equality holds on the right in Eqs. (19) and (20) for networks in the set D_0, which includes any network in D with r paths, b branches and v nodes for which

$$r = b - v + 2, \tag{21}$$

and for networks in the set D_1 which includes the networks in D_0 and, inductively, any network which is constructed from a network in D_1 by replacing any branch by another network in D_1.

Note that D_0 contains simple series networks, for which $r = 1$ and $b = v - 1$, and simple parallel networks, for which $r = b$ and $v = 2$. D_1 therefore contains all series-parallel networks, but it contains others as well — for example the (topologically equivalent) networks of Figs. 1 and 2, for which $b = 5$, $v = 4$ and $r = 3$, but not the network of Fig. 3, for which $b = 11$, $v = 6$ and $r = 8$, or any network in F with $k > 2$.

P r o o f. For the proof we need one more definition. T_{ij}, the *transmission from branch i to branch j*, is just the network transmission as defined in Eq. (18) for the subnetwork consisting of branch i and all other branches which lie on some directed path which goes through branch i to the initial node of branch j. (Thus B_i is included in the subnetwork but B_j is not, and t_i is a common factor of all of the terms in the sum T_{ij}.) If there are no paths through B_i and B_j, or if B_j precedes B_i on such a path, then $T_{ij} = 0$. $T_{0,\,j}$ is the transmission of a subnetwork with input node V_1 and output node the initial node of B_j, and $T_{i,\,b+1}$ is the transmission of the subnetwork of paths through branch i to the output node V_v.

We now derive an expression for P_{b+1}, the output power of the network. By the statistical independence of the noise voltage generators from one another and from the signal source, the output power at the right hand node is the sum of the powers transmitted to that node by these $b + 1$ separate sources. The source in branch i contributes an amount of power equal to its generated power $P_i N_i$ times the square of the transmission from B_i to the output. Thus

$$P_{b+1} = \sum_{i=0}^{b} P_i N_i T_{i,\,b+1}^2 = P_0 (1 + N_{b+1}), \tag{22}$$

where the rightmost equality follows from the fact that by the constraint of Eq. (18), Eq. (2) holds for $i = b + 1$, and where the signal power contributed to the output is represented in the sum by the term for $i = 0$, with $N_0 = 1$ and $T_{0,\,b+1} = 1$.

Similarly the input power to any branch B_i may be expressed as the sum of the contributions of the generators which lie to its left:

$$P_i = \sum_{k=0}^{i-1} P_k N_k T_{ki}^2. \tag{23}$$

Here we have assumed that the branches are numbered in an order such that if B_i precedes B_j on some directed path, $i < j$.

By successive substitution of Eq. (23) in Eq. (22) and in the resulting expressions, the subscripts on the P's appearing on the right can all be reduced to zero. The result is a sum of terms, all of which have P_0 as a factor. There is one term for each of the 2^b subsets W_m of the b branches which has the property that all of the branches in W_m are included in a path from input to output—i.e. that there is an integer f with $R_f \supseteq W_m$. If W_m is such a subset, say $W_m = (B_i, B_j, B_k)$ with $i < j < k$, then the corresponding term is

$$P_0 T_{0i}^2 N_i T_{ij}^2 N_j T_{jk}^2 N_k T_{k,\,b+1}^2 = P_0 (T_{0i} T_{ij} T_{jk} T_{k,\,b+1})^2 N_i N_j N_k. \tag{24}$$

The product of the transmission terms which appears on the right is just the sum of the transmissions of all paths from input to output

which include all three of the branches B_i, B_j, B_k. If there are no such paths then one or more of the T_{ij} in Eq. (24) will vanish. Thus the output power is expressed in terms of the path transmissions T_i and the branch parameters N_i: dividing through by P_0 gives an expression for $1 + N_{b+1}$

$$1+N_{b+1}=\sum_{k=0}^{2^b-1}\Big\{\sum_{i:\ R_i\supseteq W_k} T_i\Big\}^2 \prod_{j:\ B_j\in W_k} N_j, \tag{25}$$

where W_0 is the null set, for which the product is taken to be 1. The sum is also 1 for $k = 0$, since it is just the square of the network transmission of Eq. (18), so excluding the term for $k = 0$ gives an expression for N_{b+1} as a sum of products of positive terms, which is monotone nondecreasing in each N_j. We thus have proved

L e m m a 1. For any given set of path transmissions T_i, the network noise to signal ratio N_{b+1} is a monotone nondecreasing function of each branch noise to signal ratio N_j.

This lemma provides the proof of Eq. (13), which was referred to above.

We have also proved that N_{b+1} can vanish for a nonvanishing set of path transmissions only if there is some path R_i along which every branch is noiseless, so that setting that $T_i = 1$ and $T_j = 0$, $j \neq i$, gives a right-hand side in Eq. (25) in which only the term for W_0 remains. The matrix $\| G_{ij} - 1 \|$ will be singular if and only if there is such a noiseless path, since it will then map the transmission vector T with $T_i = 1$ and $T_j = 0$, $j \neq i$, into the null vector. The matrix $\| G_{ij} \|$ can be singular only under the same circumstances, but may not be even when noiseless paths exist.

We next show the equivalence of the right side of Eq. (25) to the quadratic form

$$\sum_{i=1}^{r}\sum_{j=1}^{r} G_{ij}T_iT_j, \tag{26}$$

where the T_i are still subject to the constraint (18). Substituting in (26) the definition (16) of G_{ij} gives

$$\sum_{i=1}^{r}\sum_{j=1}^{r} T_iT_j\Big\{\prod_{m:\ B_m\in R_i\cap R_j}(1+N_m)\Big\}. \tag{27}$$

Expanding the product gives

$$\sum_{i=1}^{r}\sum_{j=1}^{r} T_iT_j \sum_{k:\ W_k\subseteq R_i\cap R_j}\Big\{\prod_{m:\ B_m\in W_k} N_m\Big\} \tag{28}$$

Inverting the order of summation to sum over all W_k,

$$\sum_{k=0}^{2^b-1}\Big\{\sum_{i,\,j:W_k\subseteq R_i\cap R_j} T_iT_j\Big\}\Big\{\prod_{m:B_m\in W_k} N_m\Big\}. \tag{29}$$

We then recognize that the parentheses enclose a term which is just the square of the sum of T_i over the i for which W_k is included in R_i:

$$\sum_{k=0}^{2^b-1}\Big\{\sum_{i:\ W_k\subseteq R_i} T_i\Big\}^2 \prod_{m:\ B_m\in W_k} N_m, \tag{30}$$

which is just the right side of (25).

We have thus proved that for T_i constrained by Eq. (18),

$$1+N_{b+1}=\sum_{i=1}^{r}\sum_{j=1}^{r} G_{ij}T_iT_j. \tag{31}$$

Squaring Eq. (18) gives

$$1=1^2=\sum_{i=1}^{r} T_i^2=\sum_{i=1}^{r}\sum_{j=1}^{r} T_iT_j, \tag{32}$$

and subtracting (32) from (31) gives

$$N_{b+1}=\sum_{i=1}^{r}\sum_{j=1}^{r}[G_{ij}-1]\,T_iT_j, \tag{33}$$

or

$$S_{b+1}=1/\sum_{i=1}^{r}\sum_{j=1}^{r}(G_{ij}-1)\,T_iT_j. \tag{34}$$

Now N_{opt}, by definition, is the minimum value of N_{b+1} as the branch transmissions are varied, and S_{opt} is its reciprocal. We have therefore proved the first part of the Theorem: namely the equalities on the left in Eqs. (19) and (20).

To obtain the inequalities on the right in Eqs. (19) and (20), we minimize Eqs. (31) and (33) by varying the path transmissions T_i independently, subject only to the constraint imposed by Eq. (18). The additional constraints imposed by the topology of the network and by Eq. (17), which expresses the T_i in terms of the real independent variables t_k, are ignored. The results are lower bounds to the minima which Eqs. (31) and (33) can actually attain in the network.

Using a Lagrange multiplier $2M$, we set the derivative of

$$\sum_{i=1}^{r}\sum_{j=1}^{r} G_{ij}T_iT_j-2M\sum_{i=1}^{r} T_i \tag{35}$$

with respect to T_i equal to zero. This gives

$$\sum_{j=1}^{r} G_{ij}T_j = M, \quad 1 \leqq j \leqq r. \tag{36}$$

Using the minimizing T_j which satisfy Eq. (36), we multiply by T_i and sum, using the constraint of Eq. (18) and attaining a lower bound to $1+N_{\text{opt}}$:

$$\sum_{i=1}^{r}\sum_{j=1}^{r} G_{ij}T_iT_j = M \sum_{i=1}^{r} T_i = M \leqq 1+N_{\text{opt}}. \tag{37}$$

Solving the equations of Eq. (36) for the minimizing T_i gives

$$T_j = M \sum_{i=1}^{r} G_{ij}^{-1}. \tag{38}$$

Summing on j and using Eq. (18),

$$1 = \sum_{j=1}^{r} T_j = M \sum_{i=1}^{r}\sum_{j=1}^{r} G_{ij}^{-1}, \tag{39}$$

or from Eq. (37),

$$1+N_{\text{opt}} \geqq M = 1/\{\sum_{i=1}^{r}\sum_{j=1}^{r} G_{ij}^{-1}\}. \tag{40}$$

This completes the proof of Eq. (19) in the Theorem. The derivation of Eq. (20) is strictly parallel and will be ommitted. It remains only to prove the assertions made for networks in D_0 and in D_1. To prove that equality holds on the right in Eqs. (19) and (20) for networks in D_0, it is necessary to show that for such networks it is possible to vary path transmissions independently by varying branch transmissions. In fact we prove a stronger result.

L e m m a 2. A network in D which has $r = b - v + 2$ has a cutset of r branches each of which is included in just one path. Removal of this cutset divides the network into two parts: a tree connected to V_1 (which may reduce to V_1 alone) and a tree connected to V_v (which may reduce to V_v alone)

Given the lemma, we can set the r transmissions of the branches in the cutset as the r desired path transmissions and set the transmissions of all other branches equal to unity.

To prove the lemma, assign weights to nodes and branches from the left, assigning weight 1 to node V_1 and then assigning to each branch the weight of its initial node and to each node the sum of the weights of its incoming branches. With this assignment the weight of a node or a branch is clearly the number of routes from the input node V_1 to that node or branch.

Choose from each of the r paths the rightmost branch of weight 1. This set of branches, c in number, is a cutset, since it interrupts each patch. We have $c \leqq r$: $c = r$ if and only if no branch is selected more than once.

Deleting the cutset of c branches divides the network into two parts, M_1 connected to V_1 and M_2 connected to V_v. M_1, which contains b_1 branches and v_1 nodes, is a tree, since it is connected and since all of its nodes are of weight 1, so that there is only one path from V_1 to each node. Thus $b_1 = v_1 - 1$, as for any tree.

M_2 is connected to V_v and thus includes at least one tree. Let one of the trees included in M_2 have b_2 branches and v_2 nodes, with $b_2 = v_2 - 1$. Then there are two possible situations. (i) M_2 is a tree. In that case $r = b - v + 2$. Or (ii) M_2 is larger than a tree, and includes b_3 branches beyond the b_2 branches in a tree which it includes. In that case $r > b - v + 2$. We will prove the labelled statements.

(i) If M_2 is a tree, then $b = c + b_1 + b_2 = c + (v_1 - 1) + (v_2 - 1) = c + v - 2$, or $c = b - v + 2$. Since each branch in the cutset connects two trees, it completes just one path, so the number of paths $r = c$, and $r = b - v + 2$, QED_i.

(ii) If M_2 contains b_3 branches beyond those contained in a tree, then $b = c + b_1 + b_2 + b_3 = c + (v_1 - 1) + (v_2 - 1) + b_3 = c + v - 2 + b_3$, or

$$c = b - v + 2 - b_3. \tag{41}$$

Now each branch among the b_3 has weight $\geqq 2$ by construction, so it lies on at least two paths. Without these b_3 branches, V_v has weight at least c, since the c branches in the cutset have weight 1 each and are connected to V_v. Adding each of the b_3 additional branches adds a weight $\geqq 2$ to V_v, since each of them is conneted to V_v through the tree included in M_2. Thus the total weight r of V_v is $r \geqq c + 2b_3$. Combining this with Eq. (41) gives

$$r \geqq c + 2b_3 = b - v + 2 + b_3 > b - v + 2, \quad QED_{ii}. \tag{42}$$

For a network M which is in D but not in D_0, $r > b - v + 2$ and it is impossible to independently vary the path transmissions. For $b - v + 2$ is the cyclomatic number of the graph M' obtained from M by adding a branch B_{b+1} directed from V_v to V_1, and is thus the maximum number of linearly independent cycles in a graph-theoretic sense. Thus the set of r cycles in M', each of which consists of a path R_i from V_1 to V_v followed by the branch B_{b+1} from V_v to V_1 are linearly dependent in the graph-theory sense, and so therefore is the set of the paths themselves in M.

The linear dependence of the R_i implies, by taking logarithms in Eq. (17), one or more linear relations between the logarithms of the

path transmissions $\log T_i$, leading to constraints of the form

$$\log T_i + \log T_j = \log T_m + \log T_n, \text{ or } T_i T_j = T_m T_n \tag{43}$$

and no selection of values for the branch transmissions t_k can provide independent control of all path transmissions.

It may still be possible to achieve equality in Eqs. (19) and (20) for a network in D which is not in D_0, however, if the optimizing values of the path transmissions happen to satisfy the additional constraints of the form (43) imposed by the network topology. This happens in particular for the networks which are in D_1 but not in D_0.

Lemma 3. Given a network M in D, and a network M' in D_0. Let M'' be constructed by replacing branch B_j in M by the network M'. Then the value of the parameter N''_{opt} of M'' will be the same as the value of the parameter N_{opt} of M if the latter is evaluated using the parameter value N'_{opt} of M' for B_j. And the path transmissions obtained in computing N''_{opt} will lead to the same set of transmissions for the subnetwork M' as are obtained directly in the computation of N'_{opt}.

For the network M'' is equivalent to the network M with some value of the parameter N_j for branch B_j by the argument following Eq. (2) — i.e. the subnetwork M' is equivalent to *some* noisy branch B_j, and the only question is what its parameter value is. The optimum set of path transmissions for M'' must lead to the same transmissions inside M' as does the direct optimization of M', for any other choice would give a larger value to the parameter of M'' by lemma 1

Lemma 3 completes the proof of the theorem. Lemma 2 covers networks in D_0 and Lemma 3 justifies the extension of the results to networks in D_1. More practically, it permits the solution of network problems of large order by local reductions— the combining of series or parallel branches, etc.—which greatly reduces the computation. Unfortunately the other tool used for the local reduction of resistive networks— the star-mesh transformation—cannot be used for gaussian channels, since it leads to transformed branches which have correlated generators. This takes us outside of our present model. Networks with correlated noise present problems which are discussed briefly in a later section.

Proof of earlier results. The result of Eq. (5) follows from the Theorem by noting that for a series network $r = 1$, and $\|G_{ij}\| = \|G_{11}\|$. Thus

$$G_{11} = \prod_{i=1}^{b} (1 + N_i) = 1/G_{11}^{-1} \tag{44}$$

Eq. (6) follows by noting that for a parallel network, $r = b$ and $\|G_{ij} - 1\|$ is diagonal with elements $G_{ii} = N_i$, so that

$$[G_{ij} - 1]^{-1} = S_i, \quad \sum_{i=1}^{b} \sum_{j=1}^{b} [G_{ij} - 1]^{-1} = \sum_{i=1}^{b} S_i. \tag{45}$$

Eq. (11) follows from the evaluation of Eq. (20) for the network of Fig. 1. Eq. (9) follows by letting S_2 approach infinity in Eq. (11), for $k = 2$. For larger k, the first three branches are combined into an equivalent forward branch of capacity $C_1 + C_2$ and it is combined with the next noisy forward branch and the next noiseless feedback branch in the same way, etc.

Eqs. (12) and (13) have already been justified above. Eq. (14) follows by throwing away all but the linear terms—i.e. terms having a single N_i as a factor—in Eq. (33). By Eq. (25) this reduces the right side and provides a lower bound to N_{opt} or an upper bound to S_{opt}. The resulting equations are those for a set of resistors—the noisy branches—with resistance $= N_i$, all in parallel—both the forward and the feedback branches—with the noiseless branches acting as short circuits at the two ends and the conductances $S_i = 1/N_i$ adding.

Reduction of another problem to the above. A more general problem concerning networks of Gaussian channels can be reduced to the

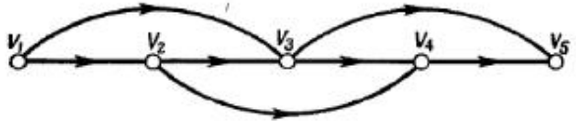


Fig. 4. A network not in class D_1.

above results. Consider the class of two-terminal networks as in D above, but in which each node may supply a different linear combination of the voltages on its incoming branches to each outgoing branch. This model still leaves the operation at the node simple and linear, and provides an increased number of independently controllable path transmissions, so that it enlarges the class of networks for which explicit solution is possible and for which equality holds in Eqs. (19) and (20).

As an example, the network shown in Fig. 4 consisting of five vertices connected by four branches forming a directed path from V_1 to V_2 to V_3 to V_4 to V_5, with three additional branches from V_1 to V_3, V_3 to V_5, and V_2 to V_4, has $b = 7$, $v = 5$ and $r = 5$, and is thus not in D_0: it has no two-terminal subnetworks, and is thus not in D_1.

The reduction to the former case replaces each node V_j which has $I_j > 1$ incoming branches and $O_j > 1$ outgoing branches by I_j nodes at each of which one of the incoming branches arrives and O_j nodes from each of which one of the outgoing branches leaves, together with I_jO_j noiseless branches connecting each of the I_j arrival nodes to each of

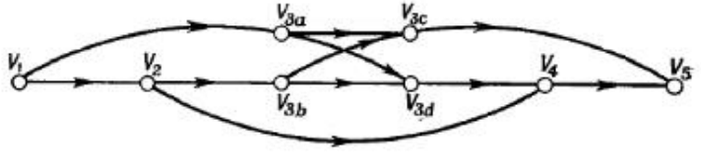


Fig. 5. A reduction of the network of Fig. 4 to a network in class D_1.

the O_j departure nodes. The added noiseless branches permit the formation of the desired different linear combinations of input branch voltages for each output branch. In the case of the five-node network described above, replacing V_3 by 4 nodes and 4 branches, as shown in

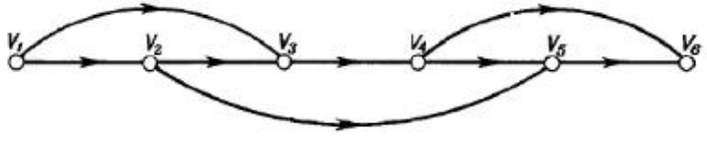


Fig. 6. A network not in class D_1 which cannot be reduced.

Fig. 5, adds 3 nodes, 4 branches and no paths, so that $b - v + 2 = 7 + 4 - (5 + 3) + 2 = 5 = r$, and the resulting net is in D_0.

The simplest network which cannot be expanded by the above substitution, has no two-terminal subnetworks, and is not in D_0, is shown in Fig. 6. It consists of six nodes V_1 to V_6 connected in order by five branches, with three additional branches from V_1 to V_3, V_4 to V_6 and V_2 to V_5.

Unfortunately the additional control provided by the change in rules provides no help for networks in F, which remain outside D_0 for $k > 2$.

Networks with correlated noise. One can consider networks in which $\overline{n_i n_j} \neq 0$, although the noise and the signal remain uncorrelated. For parallel branches, if w etake $\overline{n_i n_j} = G_{ij}$ and T_i as the branch transmision subject to the constraint of Eq. (18), then minimizing the mean

square value of the sum

$$\overline{\left[\sum_i T_i (e_0 + n_i)\right]^2} \tag{46}$$

feads to precisely the result of the Theorem, with equality in Eqs. (19) and (20), by precisely Eqs. (35) to (40). In fact the proof of the Theorem may be taken as a proof that the voltages transmitted to the output node V_v by the different paths R_i have the average product matrix $\|G_{ij}\|$.

For series branches the situation is different, however. Correlated series branches do not commute unless their parameter values are equal. Even the validity of the branch model breaks down. The definition in Eq. (1) of the added branch noise power $\overline{n_i^2} = P_i N_i$ is valid as a model of a physical channel so long as the channel is always used at the maximum possible input power. This is always advantageous when branch noises are uncorrelated, so the restriction is not felt in the optimization problem of the theorem. However a more realistic model of a physical channel has an input power limit P_i, and adds a noise of power $N_i P_i$ to any input signal whose power is $\leqq P_i$. With correlated noise in series branches it will sometimes be advantageous to use less than the maximum input power to a branch: no analog to Lemma 1 holds.

As an example consider two identical channels in series. Each accepts inputs of power $\leqq 2$ and adds to them the *same* noise voltage n, of power $\overline{n^2} = 1$. If we apply only 1 unit of signal power to the first channel, invert its output using a branch transmission of -1 and apply the result to the second channel, the output of the second channel has no noise voltage, and therefore an infinite signal-noise ratio. If, however, we apply 2 units of signal power to the first channel, and scale its output voltage by $-\sqrt{\frac{2}{3}}$ to provide 2 units of input power to the second channel, we cannot get a signal-noise power ratio at the output of the second channel which is better than $4/(5-2\sqrt{6}) \cong 40$.

Further bounds on networks in F

The open questions of greatest interest for applications concerns networks in F with $k > 2$ and with noisy feedback. In a feedback system it is reasonable to assume that the transmitter has a limited amount of signal to noise ratio S_{odd} available, and that the receiver has a limited amount S_{even}, given by

$$S_{\text{odd}} = \sum_{j=1}^{k} S_{2j-1}, \quad S_{\text{even}} = \sum_{j=1}^{k-1} S_{2j}, \tag{47}$$

and that they are free to allocate their limited resources between the different forward and feedback channels in the way which maximizes the resulting S_{opt} of F. This freedom may even extend to deciding how large k should be, if the available forward and feedback channels have infinite bandwidth.

In the case of noiseless feedback, $k = \infty$ is best, and gives the result of Eq. (10) above. When the feedback is noisy, evaluating what S_{opt} the best division of limited power gives, and how S_{opt} depends on k, involves a great deal of numerical solution of linear equations subject

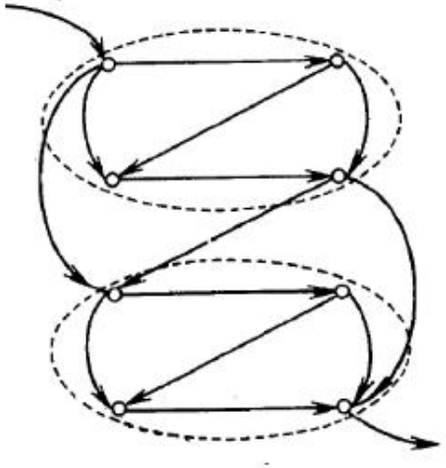


Fig. 7. An iteration of networks in class F for which $k=2$.

to constraints of the form of Eq. (43). Even evaluating the upper bound to S_{opt} of Eq. (20) is not easy. Lower bounds to S_{opt} which are more meaningful than that of Eq. (12) can be computed, however, by making use of iteration of networks for which $k = 2$, as illustrated in Fig. 7.

For the first level network, we assume that the two forward branches have equal signal to noise ratio, since this maximizes S_{opt} in Eq. (11), for fixed S_{odd}. Denoting their common signal to noise ratio as S_1, the feedback branch as S_2, and the resulting S_{opt} as S_3, we have from Eq. (11)

$$S_3 = 2S_1 + \frac{S_1^2 S_2}{(1+S_1)^2 + S_2}. \tag{48}$$

We now consider the second-level network to consist of two forward branches of ratio S_3 and a feedback branch of ratio S_4: the

resulting S_{opt} is denoted by S_5, and we have for the k^{th} level

$$S_{2k+1} = 2S_{2k-1} + \frac{S_{2k-1}^2 S_{2k}}{(1+S_{2k-1})^2 + S_{2k}}, \tag{49}$$

$$S_{odd} = 2^k S_1,$$

$$S_{even} = S_{2k} + 2S_{2(k-1)} + \cdots + 2^{k-1} S_2.$$

For this network the optimum allocation of S_{odd} among the 2^k forward branches has already been made: each receives an equal amount. S_{even} is divided unequally, however, with more for higher-numbered branches, in the optimum case. The optimum allocation can be determined by solving Eq. (49) for S_{2k}:

$$S_{2k} = \frac{S_{2k+1} - 2S_{2k-1}}{1 - \dfrac{1+S_{2k+1}}{(1+S_{2k-1})^2}}. \tag{50}$$

Now differentiating $S_{2k} + 2S_{2k-2}$ with respect to S_{2k-1} for fixed S_{2k+1} and S_{2k-3} and setting the result equal to zero gives an equation:

$$\frac{S_{2k-3}^2 (1+S_{2k-3})^2}{[(1+S_{2k-3})^2 - (1+S_{2k-1})]^2} = \tag{51}$$

$$= (1+S_{2k-1}) \frac{(1+S_{2k-1})^3 - (1+S_{2k+1})(2-3S_{2k-1}) + (1+S_{2k+1})^2}{[(1+S_{2k-1})^2 - (1+S_{2k+1})]^2}.$$

For given S_{2k-3} and S_{2k-1}, this equation is quadratic in S_{2k+1}: solving it enables us to start with a desired S_1 and S_3 and generate S_{2k+1} for any k. Alternatively, we may fix S_{2k+1} and S_{2k-1} and solve for S_{2k-3}: taking the pos. square root of each side of Eq. (51) gives a quadratic in S_{2k-3} and we can proceed from given values of S_{2k+1} and S_{2k-1} down to S_1. In either case the resulting set of values is optimum in the sense that by keeping the end points fixed, and fixing k, any other division of S_{odd} will take more of it: choosing all combinations of values for, e.g., S_1 and $S_3 > 2S_1$, generates the full set of optimum curves.

The result, unfortunately, must be displayed as a set of curves rather than an equation. A much weaker lower bound to S_{opt} can be given as an equation. Although it is not the best strategy, we may pick a division of S_{even} which gives us a fixed c such that

$$1 + S_j = c(1+S_{j-2})^2, \text{ odd } j. \tag{52}$$

Then from Eq. (50),

$$S_{2k} = \frac{S_{2k+1} - 2S_{2k-1}}{1-c} \tag{53}$$

and from Eqs. (49) and (52),

$$S_{\text{even}} = \frac{S_{2k+1} - 2^k S_1}{1-c} \leqq \frac{S_{\text{opt}} - S_{\text{odd}}}{1-c}, \tag{54}$$

since $S_{2k+1} \leqq S_{\text{opt}}$. We also have from repeated application of Eq. (52)

$$c(1+S_{\text{opt}}) \geqq c(1+S_{2k+1}) = c^{2k}(1+2^{-k}S_{\text{odd}})2^k. \tag{55}$$

Together, Eqs. (54) and (55) provide a useful analytic lower bound to S_{opt}.

Department of Electrical Engineering
and Research Laboratory of Electronics,
Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

REFERENCES

[1] Dantzig G. B., Fulkerson D. R., On the Max-Flow Min-Cut Theorem of Networks, in «Linear Inequalities», *Ann. Math. Studies*, No. 38, Princeton, New Jersey, (1956).
[2] Ford L. R., Fulkerson D. R., Maximal Flow Through a Network, *Canadian J. Math.*, **8** (1956), 399-404.
[3] Elias P., Feinstein A., Shannon C. E., A Note on the Maximum Flow Through a Network, *IRE Transactions on Information Theory*, **IT-2** (1956), 117-119.
[4] Kahn L. R., Ratio Squarer, *Proc. IRE*, **42** (1954), 1704.
[5] Brennan D. G., On the Maximum Signal-Noise Ratio Realizable from Several Noisy Signals, *Proc. IRE*, **43** (1955), 1530.
[6] Elias P., Channel Capacity Without Coding, Quarterly Progress Report, Research Laboratory of Electronics, M.I.T., October 15, 1956, 90-93.
[7] Elias P., Channel Capacity Without Coding, in E. J. Baghdady, Editor, Lectures on Communication System Theory, McGraw-Hill, New York (1961), 363-368.
[8] Schalkwijk J. P. M., Kailath T., A Coding Scheme for Additive Noise Channels with Feedback, Part I, *IEEE Transactions on Information Theory*, **IT-12**, No. 2 (1966), 177-182.
[9] Schalkwijk J. P. M., A Coding Scheme for Additive Noise Channels with Feedback, Part II, *IEEE Transactions on Information Theory*, **IT-12**, No. 2 (1966), 183-189.
[10] Schalkwijk J. P. M., Center of Gravity Information Feedback, Research Report No. 501, Applied Research Laboratory, Sylvania, Waltham, Mass.
[11] Oliver B. M., Pierce J. R., Shannon C. E., The Philosophy of PCM, *Proc. IRE*, **36** (1948), 1324-1331.
[12] Kotelnikov V. A., The Theory of Optimum Noise Immunity, McGraw-Hill, New York (1959).
[13] Wozencraft J. M., Jacobs I. M., Principles of Communication Engineering, John Willey Sons, New York (1965).
[14] Turin G. L., Signal Design for Sequential Systems with Feedback, *IEEE Transactions on Information Theory*, **IT-11**, No. 3 (1965), 401-408.

АВТОМАТНО-АЛГЕБРАИЧЕСКИЕ АСПЕКТЫ ОПТИМИЗАЦИИ МИКРОПРОГРАММНЫХ УПРАВЛЯЮЩИХ УСТРОЙСТВ

В. М. ГЛУШКОВ

Теория конечных автоматов является достаточно эффективным средством решения многих практических задач, возникающих при проектировании электронных вычислительных машин. Однако для так называемых последовательностных схем (схем с памятью) область применения этой теории ограничивается устройствами с относительно небольшим числом состояний (порядка нескольких тысяч). В то же время хорошо известно, что число состояний многих

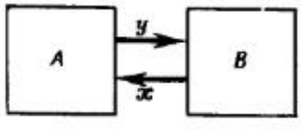


Рис. 1.

устройств современных электронных вычислительных машин выражается единицей с сотней и даже с несколькими сотнями нулей.

Развившаяся за последние годы теория бесконечных и растущих автоматов также не решает вопроса, поскольку эта теория имеет дело прежде всего с абстрактными проблемами представимости языков, а не с практическими задачами оптимизации схем электронных вычислительных машин.

Целью настоящего доклада является новая постановка задачи в теории автоматов (см. [1]), возникшая из практических применений и обобщающая ряд известных ранее постановок. Рассмотрим схему взаимодействия двух автоматов A и B (см. рис. 1), где A — конечный автомат Мили, а B — автомат Мура (вообще говоря, бесконечный) с некоторой структурой на множестве его состояний.

Будем называть автомат A управляющим, а автомат B операционным. Сигналы x и y, которыми обмениваются между собой автоматы A и B, составляют конечные алфавиты $\mathcal{X}$ и $\mathcal{Y}$. При этом выходные сигналы x автомата B обычно представляются в виде конечных кортежей булевых (двоичных) сигналов: $x = \langle \alpha_1, \alpha_2, \ldots, \alpha_h \rangle$. В целом изображенную на рис. 1 схему будем называть микропрограммным управляющим устройством.

При различных дополнительных предположениях об автоматах A и B мы получаем ряд известных ранее постановок задач:

1) Если переходы в автомате B не зависят от входных сигналов (y-ов), то автомат B можно рассматривать просто как генератор входных сигналов для автомата A. В этом случае схема рис. 1 превращается в обычную схему отдельно рассматриваемого конечного автомата Мили.

2) Если в качестве автомата A выбрать головку машины Тьюринга, а в качестве автомата B — ленту этой машины, то в схему рис. 1 укладывается описание функционирования машины Тьюринга.

3) Еще одна интерпретация схемы рис. 1 получается, если в качестве автомата A рассматривать произвольный алгоритм, а в качестве автомата B — перерабатываемую этим алгоритмом информацию.

4) Важное значение при проектировании электронных вычислительных машин имеет интерпретация схемы рис. 1, при которой в качестве автомата A фигурирует устройство управления вычислительной машины, а в качестве автомата B — ее арифметическое устройство.

5) Возможна, наконец, и такая интерпретация, при которой в автомат A объединяются все логические цепи электронной вычислительной машины (как устройство управления, так и арифметическое устройство). В этом случае под автоматом B понимается запоминающее устройство машины.

Из приведенного (неполного) перечня примеров видно, что класс микропрограммных управляющих устройств достаточно широк и важен в прикладном отношении.

В автомате A в общем случае схемы рис. 1 оказывается целесообразным выделять начальное и заключительное состояния, которые мы будем обозначать a_0 и a^* соответственно. Если теперь фиксировать автомат B, то с каждым автоматом A естественным образом связывается некоторое частичное отображение φ_A множества M состояний автомата B в себя. Мы полагаем, что автомат A начинает работу в состоянии a_0 и останавливается, когда первый раз попадает в состояние a^*. Если при этом автомат B переходит из состояния b_0 в состояние b_1, мы принимаем, что $\varphi_A(b_0) = b_1$.

При фиксированной паре A, B и фиксированном начальном состоянии a_0 автомата A выходное слово l этого автомата до момента первого его перехода в заключительное состояние a^* будет определяться лишь начальным состоянием b автомата B. Тем самым для тех состояний b, для которых такой переход возможен, определено отображение $b \to l_b$ в множество слов в алфавите $\mathcal{Y}$. Обозначим это отображение через R_A и условимся называть его *представлением* ранее рассмотренного отображения φ_A.

При практических приложениях рассматриваемой схемы взаимодействия двух автоматов важное значение приобретают два вида эквивалентности автоматов A по отношению к заданному автомату B:

1. автоматы A и A' называются *слабо B-эквивалентными*, если $\varphi_A = \varphi_{A'}$;
2. автоматы A и A' называются *сильно B-эквивалентными*, если $R_A = R_{A'}$.

Сильная эквивалентность автоматов — понятие, близкое к равносильности схем программ в смысле Ю. И. Янова [2]. От обычной эквивалентности автоматов она отличается тем, что в рассматриваемой схеме на входе автомата A могут появляться, вообще говоря, не любые слова. Благодаря этому некоторые переходы в автомате A не используются и могут рассматриваться как неопределенные. Минимизационные проблемы сводятся, таким образом, в большей своей части к минимизации частично определенных автоматов. Задача состоит лишь в том, чтобы по автомату B находить переходы в автомате A, которые могут быть сделаны неопределенными.

Один из удобных подходов к решению этой задачи (см. [3]) состоит в том, что вместо автомата B, вообще говоря бесконечного, рассматривается конечный недетерминированный автомат X_B, множество состояний которого совпадает с множеством выходных сигналов автомата B. Каждый выходной сигнал x автомата B можно отождествить с множеством отмечаемых им состояний этого автомата. Тогда для любого входного сигнала y естественно принять, что состояние x автомата X_B переводится этим сигналом во все те состояния $x_1, x_2, \ldots, x_k$, которые, рассматриваемые как множества состояний автомата B, имеют непустые пересечения с множеством $x \cdot y$.

Рассматривая схему взаимодействия автомата A и автомата X_B, мы можем последовательно, отправляясь от начального состояния a_0 автомата A, определять множества M_a состояний автомата X_B, в которые он может попасть, когда автомат A попадает в соответствующее состояние a (множество M_0 предполагается при этом заданным). При повторных прохождениях состояния a множество M_a может возрастать, однако в силу конечности автоматов A и X_B процесс возрастания множеств M_a заканчивается через конечное число шагов. Легко понять, что переход в автомате A, соответствующий паре (a, x), не может встретиться при взаимодействии автоматов A и B, если $x \notin M_a$. Делая неопределенными все такие переходы, мы увеличиваем возможности минимизации автомата A.

Нетрудно видеть, что указанный прием представляет собою автоматную интерпретацию результатов Ю. И. Янова [2]. Подобная интерпретация позволяет намного проще, чем раньше, доказать многие результаты о равносильности схем программ. С практической же точки зрения особенно важно то, что предлагаемая интерпретация дает не только методику формальных преобразований схем программ, но и методику их оптимизации. Оптимизация здесь

понимается лишь в смысле уменьшения числа состояний управляющего автомата A, или, что то же самое, уменьшения объема памяти, необходимой для хранения программы.

Сказанным, разумеется, не исчерпывается проблема сильной эквивалентности автоматов, однако для таких практических целей, как минимизация управляющих устройств электронных вычислительных машин (в смысле уменьшения числа их состояний), приведенных результатов оказывается часто достаточно.

Понятие слабой эквивалентности с математической точки зрения является более тонким. С одной стороны, оно обобщает понятие сильной эквивалентности, а с другой — такое чисто алгебраическое понятие, как понятие эквивалентности слов в полугруппе. Чтобы убедиться в последнем, достаточно рассмотреть случай, когда в схеме рис. 1 переходы в автомате A не зависят от входных сигналов x. В отличие от сильной эквивалентности проблема слабой эквивалентности автоматов в общем случае алгоритмически неразрешима. Вместе с тем использование слабо эквивалентных преобразований позволяет осуществлять наиболее глубокие преобразования автоматов. Оказывается, что алгоритмическая неразрешимость проблемы не является помехой для построения содержательной теории, находящей важные практические приложения.

Работы в области слабой эквивалентности автоматов развертываются в настоящее время в трех основных направлениях.

Первое направление связано с нахождением необходимых и достаточных условий слабой эквивалентности. Такие условия, разумеется, не могут быть конструктивными. Однако представляют интерес и различные теоретико-множественные условия, если их формулировать в круге понятий, достаточно далеких от исходных. Исследования в этом направлении ведутся учеником автора А. А. Летичевским [4]. Им же получено достаточное условие существования в классе слабо B-эквивалентных автоматов быстрейшего автомата A (у которого образ любого состояния b из B при отображении R_A имеет наименьшую длину по сравнению с образами при отображениях $R_{A'}$, когда A' пробегает совокупность автоматов, слабо B-эквивалентных автомату A).

Второе направление исследований (см. [1]) связано с построением исчислений, позволяющих фактически выполнять слабо B-эквивалентные преобразования. Имея в виду указанную выше возможность интерпретации пары автоматов A и B как алгоритма и перерабатываемой им информации, можно применять получаемые здесь результаты для эквивалентных преобразований алгоритмов (в частности, программ и микропрограмм электронных вычислительных машин).

Исчисление, о котором идет речь, строится на основе пары специальных алгебр $(\mathfrak{A}, \mathfrak{B})$, определяемых автоматом B и называемых

соответственно *операционной алгеброй* и *алгеброй условий*. Элементами операционной алгебры $\mathfrak{A}$ являются частичные отображения множества M состояний автомата B в себя, называемые нами *операторами*. Элементами алгебры $\mathfrak{B}$ служат частично определенные булевы функции на множестве, называемые ниже *условиями* или *логическими условиями*.

Кроме обычной операции умножения отображений, для каждого условия β из $\mathfrak{B}$ в алгебре $\mathfrak{A}$ определяются еще две операции, называемые β-*дизъюнкцией* и β-*итерацией* операторов. Результатом β-дизъюнкции $(P \vee Q)_\beta$ двух операторов P и Q является оператор R, такой, что $bR = bP$ для любого состояния $b \in M$, если условие β истинно на состоянии b, $bR = bQ$, если β (b) ложно и, наконец, оператор R считается неопределенным на состоянии b, если β (b) не определено. Результатом β-итерации $\{P\}_\beta$ оператора P является оператор S, такой, что bS для любого $b \in M$ принимается равным первому из состояний ряда b, bP, (bP) $P = bP^2$, bP^3, . . ., для которого истинно условие β. Если же такого состояния в указанном ряду не окажется, то результат применения оператора S к состоянию b считается неопределенным.

Операциями в алгебре условий $\mathfrak{B}$ служат дизъюнкция, конъюнкция, отрицание и (левое) умножение на операторы из $\mathfrak{A}$. Для первых трех операций пояснения требуют только те ситуации, когда наряду с обычными значениями условий и, л встречается неопределенное значение н. Правила выполнения операций даются в этом случае следующими соотношениями: и $\wedge$ н = н $\wedge$ и = н, л $\wedge$ н = = н $\wedge$ л = л, н $\wedge$ н = н; и $\vee$ н = н $\vee$ и = и, л $\vee$ н = н $\vee$ л = н, н $\vee$ н = н; $\neg$ н = н.

Если P — произвольный оператор алгебры $\mathfrak{A}$, β — произвольное условие из $\mathfrak{B}$, а b — любое состояние автомата B, то результат умножения $P \cdot \beta$ есть условие γ, истинное или ложное на состоянии b в зависимости от того, истинно или ложно условие β на состоянии $c = bP$. Если же последнее состояние не определено или не определенным на нем является условие β, то результат применения условия γ к состоянию b (т. е. γ (b)) также считается неопределенным.

В схеме рис. 1 имеем дело с некоторым исходным (конечным) множеством элементарных операторов (микроопераций) $y_1, y_2, \ldots, y_n$ и некоторым (также конечным) множеством элементарных условий $\alpha_1, \alpha_2, \ldots, \alpha_k$. Построим минимальную пару алгебр $(\mathfrak{A}, \mathfrak{B})$, порожденную указанными элементами как образующими. Алгебры $\mathfrak{A}$ и $\mathfrak{B}$ полностью определяются заданием автомата B, в связи с чем мы будем употреблять нижний индекс B в обозначениях этих алгебр.

С другой стороны, при фиксированном автомате B выбор автомата A (с выделенными в нем начальным и заключительным состоя-

ниями) определяет некоторый оператор A на множестве состояний автомата B. Имеет место следующая основная

Теорема. *Для любого конечного автомата A в схеме рис. 1 соответствующий ему оператор A^* является элементом алгебры* $\mathfrak{A}$.

Любое представление данного оператора алгебры $\mathfrak{A}$ через образующие элементы пары алгебр $(\mathfrak{A}, \mathfrak{B})$ называется *регулярной программой* (или микропрограммой) этого оператора. Значение только что сформулированной основной теоремы состоит в том, что она утверждает возможность регуляризации (т. е. представления в регулярной форме) любого алгоритма, в частности любой микропрограммы или программы электронной вычислительной машины.

После же такого представления мы получаем возможность использовать обычную технику эквивалентных преобразований слов в алгебрах с конечным числом образующих для эквивалентных преобразований алгоритмов. Разумеется, для этой цели необходимо иметь не только систему образующих, но и систему определяющих соотношений в паре алгебр $(\mathfrak{A}_B, \mathfrak{B}_B)$.

Для одного из наиболее важных (с точки зрения теории проектирования электронных вычислительных машин) операционного автомата B автором [1] была выписана некоторая (неполная) система соотношений, позволившая осуществить ряд практически важных преобразований алгоритмов. Как сообщил мне проф. Мак-Карти, после выхода в свет работы [1] им совместно с проф. Д. Скотт (Стэнфорд, США) были предприняты исследования в области поиска полных систем определяющих соотношений в алгебрах $\mathfrak{A}$ и $\mathfrak{B}$. Однако до последнего времени даже для упоминавшегося выше частного случая автомата B не был решен вопрос о существовании в алгебре $\mathfrak{A}_B$ конечной системы определяющих соотношений. Профессором Мак-Карти было предложено также новое доказательство основной теоремы о возможности регуляризации произвольного алгоритма.

Третьим направлением исследований в связи с рассматриваемой в докладе постановкой задачи является решение различного рода конкретных задач, возникающих при проектировании электронных вычислительных машин. Одной из таких задач является проектирование устройства управления ЭВМ. Как известно, значительная часть искусства проектировщика уходит на составление микропрограмм выполнения различных операций машины, особенно если эти операции не сводятся лишь к традиционному набору, а включают в себя, например, вычисление элементарных функций или куски программы-диспетчера.

Предлагаемая методика позволяет при составлении микропрограмм писать их в простейшей (с точки зрения удобства записи) форме, не заботясь о достижении наилучших характеристик. Оптимизация же написанных микропрограмм может проводиться формальными методами и даже быть автоматизирована на основе описанной выше техники.

В качестве примера укажем одну из первых простейших задач, решенных таким образом.

В этой задаче состояниями операционного автомата B служат тройки (a, b, c) целых неотрицательных чисел. В качестве элементарных операторов выбираются следующие операторы:

$$Q_1\colon (a, b, c) \to (0, b, c); \qquad Q_2\colon (a, b, c) \to (a, 0, c);$$
$$Q_3\colon (a, b, c) \to (a, b, 0); \qquad S_{12}\colon (a, b, c) \to (a, a+b, c);$$
$$l_1\colon (a, b, c) \to (2a, b, c); \qquad r_3\colon (a, b, c) \to \left(a, b, \frac{1}{2}c\right);$$
$$p_3^{-1}\colon (a, b, c) \to (a, b, c-1).$$

Предпоследний оператор r_3 определен только тогда, когда число c — четное, а оператор p_3^{-1} — когда число c отлично от нуля. Тождественный оператор будем обозначать через e. Наконец, через α_3 обозначим условие, истинное на состояниях вида $(a, b, 0)$ и ложное на всех остальных состояниях, а через β_3 — условие, истинное на состояниях вида $(a, b, 2c)$ и ложное на всех прочих условиях.

Обозначим через Q оператор $(a, b, c) \to (0, ac, 0)$, вычисляющий произведение двух чисел. Непосредственно из определения умножения как последовательного сложения легко получается следующая регулярная микропрограмма оператора Q:

$$Q = Q_2 \underset{\alpha_3}{\{} S_{12}p_3^{-1}\} Q_1 Q_3. \tag{1}$$

С помощью некоторой (заведомо неполной) системы определяющих соотношений алгебры $\mathfrak{A}_B$ (включающей в себя соотношения $S_{12}^2 l_1^2 = l_1 S_{12}$, $p_3^{-2} r_3 = r_3 p_3^{-1}$ и ряд других, несколько более сложных соотношений) выписанная микропрограмма легко преобразуется в следующую микропрограмму:

$$Q = Q_2 \underset{\alpha_3\ \beta_3}{\{} (l_1 \vee S_{12}p_3^{-1}) l_1 r_3\} Q_1 Q_3. \tag{2}$$

Эта микропрограмма представляет собой, как легко понять, хорошо известный алгоритм позиционного умножения в двоичной системе счисления. Хотя он записывается сложнее, чем алгоритм [1], однако он несравненно эффективнее с точки зрения скорости работы. Число шагов (микротактов), затрачиваемое первым алгоритмом на вычисление произведения двух чисел a и c, имеет порядок c, а вторым алгоритмом — порядок $[\log_2 c]$.

Когда автомат B представляет собою арифметическое (операционное) устройство электронной вычислительной машины, имеется возможность дальнейшей формализации указанных построений.

Дело в том, что как элементарные операции, так и элементарные условия в арифметических устройствах обычно задаются функциями, имеющими достаточно простую периодическую природу, что связано с простотой реализации соответствующих схем (см. [5, 6]). Задаваясь такими функциями, можно получать некоторые из соотношений алгебр $\mathfrak{A}_B$ и $\mathfrak{B}_B$ (аналогичные приведенным выше примерам) строго формально и в конечном счете автоматически.

Институт кибернетики АН СССР,
Киев, СССР

ЛИТЕРАТУРА

[1] Глушков В. М., Теория автоматов и формальные преобразования микропрограмм, *Кибернетика*, № 5 (1965), 1-9.
[2] Янов Ю. И., О логических схемах алгоритмов, *Проблемы кибернетики*, 1 (1958), 75-127.
[3] Глушков В. М., К вопросу о минимизации микропрограмм и схем алгоритмов, *Кибернетика*, № 5 (1966), 1-3.
[4] Летичевский А. А., Эквивалентность автоматов с заключительным состоянием, *Кибернетика*, № 4 (1966), 18-24.
[5] Глушков В. М., Теория автоматов и вопросы проектирования структур цифровых машин, *Кибернетика*, № 1 (1965), 3-11.
[6] Глушков В. М., Капитонова Ю. В., Летичевский А. А., Язык для описания алгоритмических структур вычислительных машин и устройств, *Кибернетика*, № 5 (1966).

ЧИСЛЕННЫЕ МЕТОДЫ, ИСПОЛЬЗУЮЩИЕ ВАРИАЦИЮ В ПРОСТРАНСТВЕ СОСТОЯНИЙ

Н. Н. МОИСЕЕВ

1. Сведение задачи отыскания оптимального управления к задаче нелинейного программирования

1. Основное место в докладе занимают методы решения задачи оптимального управления. Она формулируется следующим образом:

Определить минимум (максимум) функционала $I(x, u)$, если $x \in G_x$ и $u \in G_u$, при условии, что

$$\dot{x} = X(x, u),$$
$$x(0) \in \varepsilon_0;\ x(T) \in \varepsilon_T. \tag{1}$$

Здесь ε_0, ε_T, G_x, G_u — заданные множества, x называется фазовым вектором, u — управляющий вектор, или управление.

Задачи этого типа в последнее время стали играть большую роль в различных областях естествознания и техники, и для их решения разработан целый ряд эффективных приемов.

Для этой цели можно, например, использовать необходимые условия экстремума (уравнения Эйлера). С их помощью широкий класс вариационных задач может быть сведен к краевым задачам для обыкновенных дифференциальных уравнений.

Другая группа методов основана на идее «уточнения управления». Типичным представителем этой группы является известный метод Шатровского — Брайсона — Келли. Существуют и другие идеи построения решения задач оптимального управления. Например, Р. Беллман и В. Ф. Кротов предлагают сводить проблему отыскания оптимального управления к некоторой задаче Коши для уравнения в частных производных. Известное количество задач удалось решить прямыми методами (метод Ритца), которые сводят вариационную задачу к алгебраической. Методы, о которых шла речь, используют понятие вариации управления; но поскольку допустимость вариации определяется не только условием $u \in G_u$, но и условием $x \in G_x$ (принадлежности траектории системы (1) допустимой области фазового пространства), то применение этих методов обычно требует преодоления ряда трудностей. Поэтому эффективная реализация большинства из них существенно усложняется при необходимости учитывать ограничения на фазовые координаты.

2. Настоящий доклад посвящен изложению методов, относящихся к другому кругу идей, основанных на исследовании вариации в пространстве (x, t) — пространстве состояний. Эти методы возникли в связи с необходимостью построения эффективных вычислительных схем для решения задач оптимального управления с фазовыми ограничениями.

Заметим, что определение вариации управления δu по заданной вариации фазовой траектории δx является процедурой не более сложной, чем определение функции $\delta x(t)$ по заданной вариации $\delta u(t)$. Построение вариации δx проводится в согласии с условием $x + \delta x \in G_x$. Вопрос же о допустимости этой вариации δx решается проверкой условия $u + \delta u \in G_u$. Некоторые трудности появляются при переходе от непрерывной задачи к дискретной, так как при этом возникает неоднозначность в определении δu по δx. Эта трудность преодолевается введением элементарной операции.

3. В пространстве состояний построим гиперплоскости Σ_i, определяемые условиями

$$t = C_i. \tag{2}$$

Пусть P_i и P_{i+1} — две точки пространства состояний, лежащие на гиперплоскостях Σ_i и Σ_{i+1} соответственно, и пусть $P_i \in G_x$, $P_{i+1} \in G_x$. Будем говорить, что задана *элементарная операция* $B(P_i, P_{i+1})$, если точкам P_i и P_{i+1} поставлены в соответствие управление u, переводящее систему из состояния (P_i, C_i) в состояние (P_{i+1}, C_{i+1}), и траектория системы (1), соединяющая эти две точки.

Если не существует такого $u \in G_u$, которое переводит систему из состояния (P_i, C_i) в состояние (P_{i+1}, C_{i+1}), то мы говорим, что точка P_{i+1} не достижима из точки P_i; в противном случае говорим, что точка P_i достижима из точки P_{i+1}.

Траектория, проведенная при помощи данной элементарной операции $B(P_i, P_{i+1})$ через совокупность последовательно допустимых точек $P_0, P_1, \ldots, P_N$, и управление, реализующее эту траекторию, определяются теперь только этими точками.

Таким образом, элементарная операция определяет некоторый класс траекторий и управлений $\{(x, u) \in G^B\}$, зависящих от конечного числа параметров — точек пересечения траектории $x(t)$ с поверхностями (2). Обозначая $x(C_i) = x_i$, мы найдем, что функционал, определенный на множестве G^B, является функцией конечного числа переменных

$$I = I(x_0, x_1, \ldots, x_N). \tag{3}$$

Основная идея, которая лежит в основе данной работы, состоит в замене исходной задачи задачей отыскания

$$\min_{(x, u) \in G^B} I(x, u), \tag{4}$$

т. е. сведении исходной задачи к задаче отыскания минимума функции конечного числа переменных $I(x_0, \ldots, x_N)$.

Такая замена при известных условиях, наложенных на структуру элементарной операции, для некоторых классов функционалов приводит к построению приближенного решения, которое при $N \to \infty$ стремится к точному.

Заметим, что задача (4) в общем случае представляет собой некоторую задачу нелинейного программирования и для ее решения могут быть использованы известные методы этой теории. Эта задача значительно проще исходной, если эффективно построена элементарная операция.

4. Остановимся теперь несколько подробнее на структуре элементарной операции, от которой в первую очередь зависит эффективность численной реализации изложенной схемы решения вариационных задач. В настоящее время разработаны различные способы построения элементарной операции, учитывающие специфику конкретной задачи, что приводит к экономным численным процедурам (с точки зрения количества машинных операций).

В рамках доклада трудно дать подробное изложение этих приемов. Поэтому мы ограничимся только некоторыми общими замечаниями.

Обозначим $\tau = C_{i+1} - C_i$ и на отрезке $[C_i, C_{i+1}]$ заменим уравнение (1) некоторой конечно-разностной схемой. Ее особенность состоит в том, что неизвестные заранее значения управлений должны быть определены по значениям фазового вектора на концах интервала. Поэтому структура разностной аппроксимации будет существенным образом зависеть от соотношения размерностей векторов u и x. Поясним это на примере. Пусть размерности векторов u и x совпадают. Тогда простейшая разностная аппроксимация уравнения (1) может быть записана, например, в следующем виде:

$$\frac{x_{i+1} - x_i}{\tau} = X(x_i, u_i). \tag{5}$$

Векторное равенство (5) — это система n уравнений относительно n компонент вектора u.

Если размерности векторов u и x различны, то мы вынуждены использовать схемы дробных шагов. Например, если размерность вектора u в два раза меньше размерности вектора x, то простейшая разностная аппроксимация, аналогичная (5), будет иметь следующий вид:

$$\frac{x_{i+1} - x_i}{\tau/2} = X(x_i, u_{i1}) + X\left(x_i + \frac{\tau}{2} X(x_i, u_{i1}), u_{i2}\right). \tag{5'}$$

Мы снова пришли к системе нелинейных уравнений относительно n неизвестных: $\frac{n}{2}$ — компонент вектора u_{i1} и $\frac{n}{2}$ — компонент вектора u_{i2}.

В построение элементарной операции может быть внесен целый ряд упрощений, если имеется дополнительная информация относительно природы управления. Например, источником разнообразных упрощений является возможность применения принципа максимума при переходе системы с Σ_i на гиперплоскость Σ_{i+1}.

Примечание. Редукция задачи оптимального управления к некоторой задаче нелинейного программирования может быть осуществлена и другими способами, отличными от изложенного. Например, заменив уравнение (1) конечно-разностным уравнением

$$x(k+1) = x(k) + \tau X(x(k), u(k)),$$

мы можем рассматривать исследуемый функционал как функцию переменных $u(k)$ — значений управления на k-м интервале. Однако при этом мы снова столкнемся со всеми трудностями, свойственными методам, использующим вариации в пространстве управлений.

2. Методы динамического программирования

1. Будем рассматривать теперь возможные способы отыскания экстремума функции (3). Выбор того или другого метода численного решения должен производиться с учетом особенностей этой функции. Широкий класс задач приводит к функциям I, которые обла-

дают следующим свойством:

$$I(x_0, x_1, \ldots, x_N) = \sum_{i=1}^{N} I_i(x_{i-1}, x_i). \qquad (6)$$

Эти задачи мы условимся называть аддитивными. Для них справедлив принцип оптимальности.

Класс аддитивных задач оптимального управления значительно доступнее для построения численных методов. Для него элементарную операцию можно ввести одним специальным способом. Будем говорить, что элементарная операция построена, если двум точкам x_{i-1} и x_i, лежащим на смежных гиперплоскостях, поставлено в соответствие приближенное решение следующей вариационной задачи: определить управление u, переводящее систему из состояния x_{i-1} в состояние x_i за время τ таким образом, чтобы функционал I_i достигал наименьшего значения.

2. До сих пор мы фиксировали шаг только по независимому переменному t. Фиксируем теперь шаг по фазовым переменным. Тогда в пространстве (x, t) мы получим некоторую сетку. Соединив отрезками узлы этой сетки, расположенные на соседних гиперплоскостях $t = c_i$, мы получим граф некоторого специального вида. Так как каждой паре узлов, лежащих на соседних поверхностях, элементарная операция ставит в соответствие управление u и отрезок траектории, соединяющей их, и так как функционал имеет форму (6), то в качестве длины отрезка, соединяющего два узла, мы можем принять значение функционала I_i вдоль траектории, соединяющей эти узлы.

Теперь мы можем нашу задачу сформулировать в терминах теории графов: разыскать ломаную кратчайшей длины, соединяющую узлы, принадлежащие множествам ε_0 и ε_T.

3. Для решения этой задачи можно использовать существующие методы отыскания кратчайшего пути на графе. В настоящее время известен целый ряд способов решения такой задачи. С точки зрения количества необходимых операций эти методы равноценны: необходимое число операций пропорционально числу звеньев графа. Поэтому вопрос о выборе метода, который следует рекомендовать для массовых расчетов, носит в известном смысле субъективный характер.

В практической деятельности Вычислительного центра АН СССР использовалась модификация схемы динамического программирования, которая известна под названием «киевский веник». Она предложена В. С. Михалевичем и Н. З. Шором и ее схема описывается рекуррентным соотношением

$$S(x_i^k) = \min_r \{S(x_{i-1}^r) + I_i(x_{i-1}^r, x_i^k)\}. \qquad (7)$$

В равенстве (7) приняты следующие обозначения:

x_j^s — узлы номера s на гиперплоскости Σ_j,

$S(x_j^s)$ — длина кратчайшего пути, ведущего из множества ε_0 к узлу x_j^s.

4. Изложенный выше метод является методом глобального перебора. Он дает возможность определить абсолютный минимум исследуемого функционала на множестве G^B. Однако его реализация в случае задач высокой размерности требует большой затраты машинного времени и большой памяти машины.

Последнее обстоятельство особенно затрудняет использование метода динамического программирования. В Вычислительном центре АН СССР был разработан метод последовательных приближений. На графе вводилась топология и назначалось некоторое первое приближение. Каждый шаг итерационного процесса состоял в реализации алгоритма «киевского веника» (7) на подграфе, который является окрестностью предыдущего приближения.

И. А. Крылов показал, что метод последовательных приближений требует тем меньше машинных операций, чем меньшее число узлов входит в окрестность данного приближения. Следовательно, на каждом шаге итерационного процесса должен использоваться минимальный подграф. Таким будет тот, в котором варьируется положение одного узла. Такую схему метода последовательных приближений разработал Ф. Л. Черноусько, и она получила название метода локальных вариаций. Эта схема описывается следующим рекуррентным соотношением:

$$S(x_N) = S(x_{i-1}^l) + S(x_{i+1}^s, x_N) + \min_r \{I_i(x_{i-1}^l, x_i^r) + I_{i+1}(x_i^r, x_{i+1}^s)\}. \qquad (8)$$

Здесь $x_i^r \in \Sigma_i$ — узлы, находящиеся в тем или иным образом определенной окрестности узла $P_i \in \Sigma_i$, через который проходила траектория предыдущего приближения. Через $S(x_N)$ обозначено уточненное значение функционала, через $S(x_j^h, x_N)$ — длина ломаной предыдущего приближения, соединяющей точку x_j^h с точкой $x_N \in \varepsilon_T$.

Все варианты метода последовательных приближений в отличие от метода глобального перебора дают возможность отыскивать только локальные экстремумы в смысле топологии, вводимой на множестве G^B.

Применение изложенных методов позволило эффективно решить целый ряд практически важных и трудных задач оптимального управления с фазовыми ограничениями. Приведем некоторые из них:

а) задача выбора оптимального (по быстродействию или по затратам топлива) пути корабля, пересекающего океан по заданным

прогнозам ветра, волнения и течений. Фазовые ограничения в этой задаче определяются наличием запретных зон: острова, области штормов, области, в которых встречаются плавающие льды, и т. д. Эта задача решена Н. К. Буровой.

б) Задача определения оптимального по энергетике управления, которое обеспечивает за заданное время достижение космическим аппаратом, снабженным двигателем малой тяги, второй космической скорости при условии, что его траектория не пересекает поясов радиации ван Аллена. Эта задача была решена Н. Я. Багаевой.

в) Задача о стабилизации спутника, первоначальное движение которого является хаотическим, которую решил И. А. Крылов.

5. Реализация метода динамического программирования, как это уже говорилось, требует большой машинной памяти. Поэтому стратегия построения решения состоит в следующем. Сначала строится «грубый граф» с малым числом узлов, т. е. с большим шагом по пространственным и временным переменным. Глобальный перебор на этом графе позволяет определить область, где может находиться решение. Следующий граф имеет меньшие шаги и охватывает уже только эту область фазового пространства и т. д. Итак, возможности машины и необходимость обеспечения высокой точности решения приводят к необходимости изучения влияния дробления сетки на результат численного решения. Таким образом, мы естественным образом приходим к проблеме устойчивости. Этого следовало ожидать, поскольку проблема устойчивости типична для любых разностных методов.

Уравнение (7) можно трактовать как конечно-разностный аналог уравнения Беллмана, с функцией S в качестве функции Беллмана. Тогда развиваемая теория может рассматриваться как теория конечно-разностных методов решения уравнения Беллмана. Вывод этого уравнения из рекуррентного соотношения (7) не представляет труда.

6. Значения функционала $I_i(x_{i-1}, x_i)$ определяются структурой элементарной операции. Поэтому значение функционала вдоль траектории, соединяющей точки x_{i-1} и x_i, полученной применением операции $B(x_{i-1}, x_i)$, будем обозначать через I_i^B. Тогда

$$I_i^B \geqslant S(x_{i-1}, x_i),$$

где $S(x_{i-1}, x_i)$ — минимальное значение функционала на всем множестве допустимых траекторий, соединяющих точки x_{i-1} и x_i. Обозначим через τ_n и h_n шаги некоторого разбиения пространства состояний, через Z_n и U_n — ломаные, отобранные методом динамического программирования из множества ломаных, проведенных в некоторой допустимой области Ω фазового пространства; пусть $I(n)$ — значение функционала, соответствующее (Z_n, U_n).

Далее, предположим, что

а) функция $S(x^*, x^{**})$ удовлетворяет условию Гельдера по обоим аргументам с показателем q;

б) погрешности элементарной операции ограничены для любого разбиения:

$$|I_i(x_{i-1}, x_i) - S(x_{i-1}, x_i)| \leqslant C(\tau_n, h_n);$$

в) в Ω существует единственная фазовая траектория Z, которая является решением исходной задачи, и она реализуется при управлении U, которое является измеримой функцией; S^* — значение функционала, соответствующее решению.

В этих условиях имеет место следующая

Теорема. Для того чтобы при $\tau_n \to 0$ и $h_n \to 0$ последовательность Z_n сходилась равномерно к Z и $I(n) \to S^*$, необходимо и достаточно выполнение следующих условий:

$$h_n \leqslant \alpha_1 \tau_n^c, \quad c \geqslant \frac{1}{q} + \beta_1, \tag{8}$$

$$C(\tau_n, h_n) \leqslant \alpha_2 \tau_n^{1+\beta_2}, \tag{9}$$

где α_1, β_1, α_2 и β_2 — некоторые положительные числа. Последовательность U_n при этом сходится слабо к U.

Необходимость здесь понимается в том смысле, что при невыполнении одного из условий (8) или (9) можно построить пример, в котором

$$\lim_{n \to \infty} I(n) \neq S^*.$$

Теорема легко обобщается на случай, когда шаг по каждой из компонент может выбираться произвольно, а функция S по каждому из переменных удовлетворяет условию Гельдера со своим показателем.

7. Сформулированная теорема дает условия, гарантирующие устойчивость предложенной разностной схемы решения задачи оптимального управления. Следовательно, достаточно правильно распорядиться отношением шагов по фазовым и временным переменным и уметь с требуемой точностью строить элементарную операцию. В этом случае мы можем быть уверенными в том, что описанная процедура решения задачи приведет нас к цели. Однако характер сформулированного результата вряд ли может полностью удовлетворить математика, так как условия теоремы требуют некоторой априорной информации о свойствах функции Беллмана и природе решения.

Поэтому мне казалось интересным поставить следующий вопрос: каким образом можно получить информацию о дифференциальных

свойствах функции Беллмана, располагая сведениями только о свойствах функционала и правых частей системы (1)?

В настоящее время этот вопрос еще очень мало продвинут и существующие результаты носят элементарный характер. Например, можно показать, что если правые части системы (1) аналитические функции, функционал имеет вид

$$I(x,\ u)=\int_0^T F(x,\ u)\,dt,$$

где F — нелинейная функция своих аргументов, множества G_x и G_u совпадают со всем пространством и существует единственное решение задачи, то функция Беллмана имеет ограниченные производные.

В теории, которая излагается в этом параграфе, имеется еще целый ряд нерешенных вопросов. Например, до сих пор речь шла только о классических решениях. В то же время большой интерес представляет изучение скользящих и особых режимов. Решения подобного рода (в особенности скользящие режимы) часто возникают в прикладных задачах. Они не являются исключительно редкими, как это может показаться сначала. Практика вычислений показывает, что изложенные методы служат достаточно надежным средством построения решений, которые являются хорошей аппроксимацией скользящих режимов (в смысле близости функционалов). В связи с этим возникает потребность создания новой математической теории — теории аппроксимации скользящих режимов, т. е. аппроксимации функций, которые не являются измеримыми.

8. До сих пор мы рассматривали только аддитивные задачи. Однако схемы динамического программирования могут быть использованы и для решения некоторых неаддитивных задач. К их числу относится, например, задача Майера.

Пусть нам надо определить вектор $(u,\ x)\in G^B$ так, чтобы заданная функция $F_N(x_N)$ достигала своего минимального значения. Здесь каждому узлу x_N^k поставлено в соответствие значение $F_N(x_N^k)$.

Рассмотрим теперь точку x_{N-1}^s. Обозначим через H_N^s множество узлов, которые лежат на Σ_N и достижимы из точки x_{N-1}^s. Тогда этой точке мы поставим в соответствие число

$$F_{N-1}(x_{N-1}^s)=\min_{x_N^k\in H_N^s} F_N(x_N^k).$$

Это равенство определяет на множестве узлов, лежащих на Σ_{N-1}, функцию $F_{N-1}(x_{N-1})$.

Общая схема алгоритма тогда будет такой:

$$F_r(x_r^s)=\min_{x_{r+1}^k\in H_{r+1}^s} F_{r+1}(x_{r+1}^k).$$

и решением задачи будет число

$$F_0(x_0)=\min_{x_1^k\in H_1^0} F_1(x_1^k).$$

Если в качестве функционала выступает функция от начальных значений, то алгоритм нужно строить слева направо.

Примечания. 1. Заметим, что подобные рассуждения приводят нас к синтезу, поскольку мы сразу получаем целый пучок траекторий.

2. Рассматривая в качестве функционала квадрат невязки краевых условий, можно использовать изложенную процедуру для численного решения краевых задач.

3. Градиентный метод в пространстве состояний

1. Предположим, что при помощи элементарной операции мы построили фазовую траекторию, целиком лежащую внутри допустимой области (см. рис. 1). Тогда значение функционала

$$I=I(x_0,\ x_1,\ \ldots,\ x_N)$$

будет определяться точками пересечения фазовой кривой с гиперплоскостями Σ_i. Для определения экстремума функции (3) на мно-

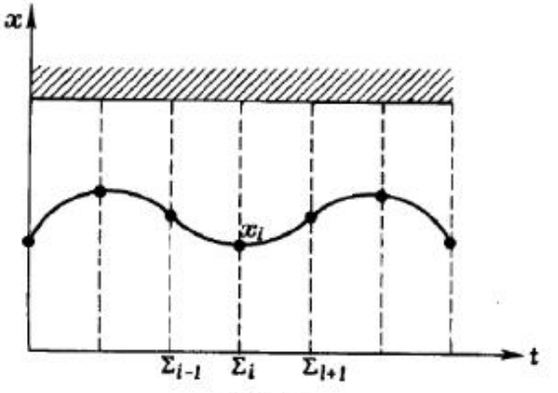


Рис. 1.

жестве G^B можно использовать многие методы нелинейного программирования. В Вычислительном центре АН СССР изучалась возможность использования различных вариантов градиентного метода. Этот метод формально может быть применен к функционалам весьма общего вида, однако следует иметь в виду, что в прикладных задачах число N бывает весьма велико. Если, кроме того, и размерность вектора x также достаточно велика, то мы приходим к задаче определения экстремума функции многих сотен и тысяч переменных.

Поэтому на первый взгляд применение подобных методов не имеет особенной перспективы. В действительности оказывается, что природа многих прикладных задач такова, что градиентные методы в пространстве состояний позволяют весьма эффективно провести вычисления. Многие трудные задачи были решены в Вычислительном центре применением именно градиентного метода или его модификаций.

Примечание. Предположим, что мы тем или иным способом построили траекторию. Вопрос о том, можно ли ее принять в качестве приближенного решения, зависит от величин производных $\frac{\partial I}{\partial x_i}$ и требований точности. Наличие блока проверки $\left(\frac{\partial I}{\partial x_i} = 0\right)$ делает естественным включение в схему расчета алгоритма наискорейшего спуска.

2. Градиентный метод отыскания экстремума в отличие от метода динамического программирования никак не связан с условием аддитивности задачи. Однако аддитивность задачи значительно облегчает его применение.

Предположим, что задача аддитивная, т. е. функционал I представим в форме (6). Тогда выражение производной $\partial I/\partial x_i$ содержит два слагаемых:

$$\frac{\partial I}{\partial x_i} = \frac{\partial I_{i+1}(x_i, x_{i+1})}{\partial x_i} + \frac{\partial I_i(x_{i-1}, x_i)}{\partial x_i}.$$

Это значит, что для вычисления производной нам достаточно вычислить изменение функционала за счет вкладов участка фазовой траектории, заключенной между гиперплоскостями Σ_{i-1} и Σ_i.

Наряду с аддитивными задачами естественно рассматривать квазиаддитивные задачи (или задачи с наследственностью). Этим термином мы будем называть такие задачи, в которых в представлении (6) функционалы I_i являются функциями вида

$$I_i = I_i(x_{i-s_i}, x_{i-s_i+1}, \ldots, x_i, \ldots, x_{i+l_i}). \tag{10}$$

Примером такой задачи является классическая задача динамики тела переменной массы: определить управление и траекторию ракеты, достигающей заданного положения с минимальным расходом массы. В задаче типа (10) мы должны вычислять изменение функционала за счет изменения траектории на участке, заключенном между плоскостями Σ_{i-s_i} и Σ_{i+l_i}.

3. Предположим, что мы определили некоторую траекторию. Тогда новая фазовая траектория определяется по формуле

$$x_i = \tilde{x}_i - k\left(\frac{\partial I}{\partial x_i}\right)_{x=\tilde{x}}. \tag{11}$$

Число k здесь произвольно. Для его определения можно использовать разнообразные способы. После вычисления новых значений x_i по формуле (11) функционал I становится функцией одного переменного k. Тогда число k можно определить из условия

$$I^* = \min_k I(k). \tag{12}$$

При такой схеме реализации градиентного метода мы не пользуемся сеткой в пространстве состояний. Фиксация сетки делает множество возможных значений параметра k дискретным. Следует иметь в виду, что выбор k (величины шага в направлении, противоположном градиенту) связан также с ограничениями по управлению: варьированная траектория должна состоять только из последовательно достижимых точек.

4. Метод градиента обеспечивает максимальную скорость сходимости в смысле числа итераций. Однако при выборе метода необходимо прежде всего учитывать возможности машины: необходимый объем памяти и время реализации алгоритма. Поэтому при решении конкретных задач часто более выгодно применять «градиентный метод с закрепленными узлами», при котором часть переменных не варьируется. Так, например, если производные по некоторым из переменных x_i малы, то естественно ту часть траектории, которая им соответствует, считать фиксированной. Если зафиксировать все точки x_i, кроме одной, то мы получим тот частный случай градиентного метода с закрепленными узлами, который называется методом координатного спуска. Для случая аддитивных задач он уже был описан в предыдущем параграфе. Координатный спуск в пространстве состояний оказался очень эффективным средством решения целого класса задач и получил название метода локальных вариаций. Он был предложен Ф. Л. Черноусько и И. А. Крыловым, которые применяли некоторую специальную организацию координатного спуска с фиксированным шагом по фазовым переменным.

Заметим, что метод локальных вариаций может рассматриваться одновременно и как вариант метода последовательных приближений в динамическом программировании и как специальный случай градиентного метода. Это обстоятельство позволяет строить гибкие вычислительные схемы, сочетающие варианты метода глобального перебора с градиентными методами.

5. До сих пор мы говорили о градиентных методах, которые определяются формулой (11). Но реализация такой процедуры возможна лишь тогда, когда отсутствуют ограничения на фазовые координаты.

В том случае, когда имеются фазовые ограничения, градиентный метод переходит в метод возможных направлений:

$$x_i = \tilde{x}_i - kp_i, \tag{13}$$

где p_i — вектор возможных направлений. Поясним особенности этого метода на простом примере.

Предположим, что движение происходит на поверхности

$$F(x,\ t) = \text{const}. \tag{14}$$

Тогда пересечение поверхностей (14) и Σ_t есть многообразие

$$F_t(x_t,\ t_t) \equiv F_t(x_t^{(1)},\ \ldots,\ x_t^{(n)},\ t_t) = \text{const}. \tag{15}$$

Здесь через $x_t^{(j)}$ обозначены компоненты вектора x_t. Метод проекции градиента в этом случае описывается формулой (13), где вектор p_t определяется следующим образом (см. рис. 2, где даны обозначения):

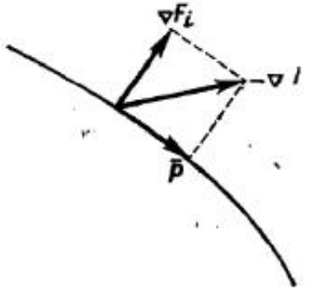


Рис. 2.

$$p_t = \nabla I - \frac{(\nabla I \cdot \nabla F_t)\,\nabla F_t}{(\nabla F_t)^2}.$$

Часто встречаются задачи с ограничениями вида $x^{(j)}(t) = \text{const}$. В этом случае метод возможных направлений дает

$$x_t^{(j)} = \tilde{x}_t^{(j)},$$

т. е. соответствующие координаты просто не варьируются.

Примечания. 1. При использовании формулы (13) предполагается, что угол между направлением векторов ∇I и ∇F_t острый. В противном случае следует использовать формулу (11).

2. Если в пространстве состояний фиксирован шаг по фазовым переменным, то процедура расчета должна опираться на возможность предварительного отбора узлов, лежащих в допустимой области.

6. В связи с развитием градиентных методов в пространстве состояний возникает целый ряд вопросов как прикладного, так и теоретического характера, которые могут составить предмет дальнейших исследований. Укажем лишь на один из таких вопросов.

Градиентные методы отыскания экстремума функций многих переменных уже давно являются предметом тщательных исследований. В частности, подробно изучены различные условия сходимости процесса. Эти результаты могут быть без особого труда перенесены на изучаемые задачи и переформулированы в терминах теории оптимальных управлений. Такая работа представляет сама по себе безусловный интерес. Однако в этом случае речь может идти только об условиях, накладываемых на функции $I(x_0, x_1, \ldots, x_N)$, и о существовании экстремума функционала на множестве G^B. Таким образом, известные факты из теории градиентных методов

легко переносятся лишь на тот случай, когда фиксирована система поверхностей Σ_t и элементарная операция. Наибольший интерес представляет изучение процессов градиентного спуска при $N \to \infty$. При этом должны появиться определенные требования к элементарной операции, необходимые для сходимости процесса к точному решению. Как оценить скорость сходимости? Эти вопросы имеют важное прикладное значение.

7. Мы рассмотрели общий подход к численному решению вариационных задач, основанный на применении понятий варьирования в пространстве состояний и элементарной операции. Сведение задач к отысканию экстремума функций многих переменных позволило применить общие идеи, которые используются в анализе при отыскании экстремума функций. В течение последних лет в Вычислительном центре АН СССР подробно изучались возможности, которые открывает этот подход для численного решения задач теории оптимального управления. Схемы численного расчета, разработанные здесь, использовали соображения динамического программирования и градиентных методов. Этим путем удалось решить целый ряд трудных вариационных задач с фазовыми ограничениями, для которых использование вариаций в пространстве состояний доставляет наиболее простой способ эффективного решения. При этом оказалось, что наличие фазовых ограничений может иногда служить средством упрощения задачи, поскольку сужение допустимой области фазового пространства может сократить перебор возможных вариантов.

4. Приложения к задачам теории управления и планирования

1. В предыдущих параграфах были изложены методы решения вариационных задач, развитые в связи с задачами теории оптимального управления. Однако идеи, которые лежат в основе этих методов, могут быть использованы для решения целого ряда других задач, допускающих вариационную постановку. Среди таких задач, например, оказываются многие задачи теории больших систем.

В настоящее время в термин «большие системы» вкладывают в разных случаях самое различное содержание. В настоящей работе за этим термином не скрывается никакого особого смысла. У нас установилась привычка называть большими системами задачи управления столь большого объема, что применение для их решения регулярных (неэвристических) методов требует очень большой затраты машинного времени.

Формулируя задачу большого объема, часто можно указать «цену» решения. Например, время для решения задачи распределения судов, прибывающих в данный порт, по его причалам для раз-

грузки должно занимать не более одного часа, так как 1 час — это по техническим условиям время между окончанием процесса сбора информации и выдачей команды капитану судна. К числу больших систем мы относим, например, задачи линейного программирования для десятков и сотен тысяч переменных. Точное решение такой задачи можно получить симплекс-методом, но оно потребует при современном уровне вычислительной техники затраты времени, стоимость которого намного превосходит сумму, которую фирма, заинтересованная в решении задачи, может выделить для этой цели.

Разумеется, понятие большой системы не является четким. Точно так же я не знаю, как в математических терминах описать понятие «эвристическое программирование»— понятие, которое сейчас широко используется. Я вкладываю в него приблизительно следующий смысл. Пусть среди лиц, имеющих отношение к данной задаче, имеется некто — условимся называть его диспетчером. Это лицо неизвестным (ему и нам) способом формулирует решение, которое условимся называть диспетчерским решением. Процесс формирования этого решения — некоторый иррациональный процесс адаптации диспетчера к данной задаче — я и называю эвристическим программированием. Разумеется, диспетчерское решение с математической точки зрения не имеет ничего общего со строгим решением исходной математической задачи. Тем не менее мы обязаны считаться с существованием и огромной ролью диспетчерских решений.

Диспетчер Балтийского пароходства в течение часа принимает решение о распределении по портам кораблей, приходящих в Балтийское море. Эта задача целочисленного программирования с большим количеством сложных ограничений и очень сложным видом функционала. Ее строгое решение требует произвести полный перебор вариантов и многих часов работы электронной машины среднего класса.

Офицер штаба, который в течение вечера составляет план передислокации большого армейского соединения, также является «специалистом по эвристическому программированию». Строгое решение его задачи возможно. Но оно требует многих недель работы машины, а в его распоряжении — вечер и он принимает решение без электронной машины.

Подобных примеров можно привести очень много. Включение «биологического звена»— диспетчера — в процесс принятия решения на много порядков сокращает время, необходимое для решения задачи, и математик должен с этим считаться.

Методы, которые мы рассматриваем в докладе, как раз и являются удобным средством улучшения диспетчерских решений. При относительно малой затрате машинного времени они могут дать заметное уменьшение значения функционала. Применение развиваемых методов также требует введения понятий «состояние системы» и «элементарной операции». Поясним эти возможности на примере двухиндексной задачи нелинейного программирования.

2. Рассмотрим задачу отыскания минимума функционала:

$$I(u_j^i) \equiv I(u_1^{(1)}, u_1^{(2)}, \ldots, u_M^{(N)}) \; (i = 1, 2, \ldots, N) \; (j = 1, 2, \ldots, M) \tag{16}$$

при следующих ограничениях:

$$\sum_j \alpha_{ji} u_j^{(i)} = a_i \quad (i = 1, 2, \ldots, N), \tag{17}$$

$$\sum_i \beta_{si} u_s^{(i)} \leqslant b_s \quad (s = 1, 2, \ldots, R). \tag{18}$$

$$u_j^{(i)} \geqslant 0. \tag{19}$$

Введем сначала новые обозначения: $u_j^{(i)} = u_j(t_i)$, $\alpha_{ji} = \alpha_j(t_i)$ и т. д. В этих обозначениях условия (17) и (18) будут иметь следующий вид:

$$\sum_j \alpha_j(t_i)\, u_j(t_i) = a(t_i) \quad (i = 1, 2, \ldots, N), \tag{17'}$$

$$\sum_i \beta_s(t_i)\, u_s(t_i) \leqslant b_s \quad (s = 1, 2, \ldots, R). \tag{18'}$$

Условимся еще считать задачу аддитивной, т. е. будем считать, что

$$I = \sum_{i=1}^{N} I_i[u_1(t_i), \ldots, u_M(t_i)]. \tag{20}$$

Ниже мы увидим, что это определение аддитивности совпадает с тем, которое мы дали в начале доклада.

Введем новые переменные, которые назовем фазовыми:

$$x_s(t_i) = \sum_{k=1}^{i} \beta_s(t_k)\, u_s(t_k) \quad (s = 1, 2, \ldots, R). \tag{21}$$

Для того чтобы эти переменные удовлетворяли условию (18′), необходимо, чтобы

$$x_s(t_N) \leqslant b_s. \tag{22}$$

Легко видеть, что задачу (16—19), сформулированную в терминах фазовых переменных, можно рассматривать как конечно-разностный аналог некоторой специальной задачи оптимального управления.

Для того чтобы определить множество G^B и применить методы, развитые в этом докладе, нам осталось ввести понятие элементарной операции.

3. Уравнение (21) можно записать в следующем виде:

$$\Delta x_s(t_i) \equiv x_s(t_i) - x_s(t_{i-1}) = \beta_s(t_i)\, u_s(t_i),$$

и, следовательно, «управление» $u_s(t_i)$ определяется однозначно, если заданы значения $x_s(t_i)$ и $x_s(t_{i-1})$:

$$u_s(t_i) = \frac{\Delta x_s(t_i)}{\beta_s(t_i)}. \qquad (23)$$

Из этого выражения видно, что структура элементарной операции существенно зависит от соотношения размерностей вектора x (число R) и вектора u (число M).

Рассмотрим сначала случай $R = M - 1$. Так как компоненты вектора u связаны ограничением (17′), то независимых компонент у этого вектора будет $M - 1$. Следовательно, уравнения (17) и (23) однозначно определят все компоненты управляющего вектора и значения функционала I_i — вклад, который вносит отрезок «фазовой траектории» при переходе из состояния $\{x_s(t_{i-1})\}$ в состояние $\{x_s(t_i)\}$.

Если $R < M - 1$, то в качестве элементарной операции мы можем рассматривать следующую задачу нелинейного программирования: определить $\min I_i(u_1, \ldots, u_M)$, где функции $u_j = u_j(t_i)$ удовлетворяют ограничениям (19), причем u_s ($s = 1, 2 \ldots R$) определяются равенствами (23) при ограничениях (17′) и (19).

В рассматриваемом случае могут быть использованы и другие способы построения элементарной операции. Например, функции $u_k(t)$ ($k = R + 1, \ldots, M - 1$) могут быть отнесены к фазовым переменным. Такой подход приведет к более простому выражению для элементарной операции. Однако размерность задачи при этом возрастет. Каждый из этих способов имеет свои достоинства и недостатки, и выбор метода должен производиться всякий раз с учетом свойств данной конкретной задачи.

Случай $R > M - 1$ соответствует той ситуации, когда размерность фазового вектора больше размерности управляющего вектора. Когда с подобной ситуацией мы сталкиваемся в непрерывных задачах, то при построении элементарной операции используется разностная схема с дробным шагом. Так как в данной задаче шаг фиксирован, то для построения элементарной операции используется несколько последовательных шагов, т. е. при построении элементарной операции фиксируются не все значения $x_s(t_i)$. Некоторые из этих величин считаются неизвестными. Структура элементарной операции будет зависеть от величины разности $R - M$.

После того как определена элементарная операция, дальнейший ход расчета проводится по схеме, изложенной выше. В зависимости от характера задачи и параметров вычислительной машины, которая

находится в нашем распоряжении, мы можем воспользоваться стандартной программой киевского веника или другими вариантами метода динамического программирования, а также градиентным методом и его модификациями. В последнем случае мы получаем весьма экономные алгоритмы. Однако в общем случае мы ничего не можем сказать о соответствии найденного решения точному: мы можем отыскать этим способом только локальный экстремум на множестве G^B.

В предыдущих параграфах, когда мы рассматривали непрерывные задачи, мы уже сталкивались с подобным обстоятельством. Если возможности машины таковы, что мы не можем провести глобального перебора, то нам приходится ограничиваться отысканием локального экстремума.

Заметим, что понятие локального экстремума связано с введением топологии на множестве G^B. В случае задач оптимального управления топология вводится естественным образом: мы говорим, что две траектории близки, если они во все моменты времени (которое принимает дискретные значения при переходе к множеству G^B) близки в смысле евклидовой метрики. В задачах, которые рассматриваются в данном параграфе, у нас возникает дополнительная трудность, так как время введено формально и зависит от принятой нумерации величин $u_j^i = u_j(t_i)$. Поэтому понятие локального экстремума в рассматриваемых задачах весьма условно. Изменив нумерацию неизвестных, мы получим совершенно иную систему окрестностей и, следовательно, отправляясь от данного решения, придем к другому «локальному решению».

Я уже обратил внимание на важность диспетчерских решений. Изложенный подход, несмотря на все его дефекты, позволяет очень легко улучшить такое решение. Таким образом, при исследовании больших систем мы имеем прием, позволяющий сочетать эвристическое программирование с методами вычислительной математики и уточнять эвристическое (диспетчерское) решение.

Выбор топологии на множестве G^B также может относиться к сфере эвристики. С другой стороны, определив упорядоченную систему топологий, мы можем при помощи изложенного аппарата провести дальнейшее улучшение найденного «локального решения». При организации такого поиска, по-видимому, окажутся полезными идеи отыскания экстремумов функции, определенных на подмножествах, которые легли в основу метода В. П. Черенина.

Вопросы, которые возникают в связи с решением подобных задач, еще не носят вполне отчетливого математического характера, а использование метода очень часто напоминает поиски экспериментатора. Появление вычислительных машин делает неизбежным появление экспериментального направления в вычислительной математике. Экспериментальная практика, с которой мне пришлось

иметь дело, показывает, что изложенные подходы к решению дискретных задач имеют не только право на существование, но и определенную перспективу.

5. Применение к задачам математической физики

1. Методы, развиваемые в этой работе, могут быть использованы для решения краевых задач математической физики. К числу подобных задач относятся все те, для которых существует вариационный принцип. Пусть речь идет о решении уравнения

$$Ax = 0, \quad x \in G. \tag{24}$$

Если A — потенциальный оператор, то функционал I, градиентом которого он является, записывается так:

$$I = \int_0^1 [A(x_0 + \varepsilon(x - x_0)),\ x - x_0]\, d\varepsilon, \tag{25}$$

и мы можем перейти к рассмотрению задачи отыскания экстремума функционала (25). Потенциальные операторы играют большую роль в приложениях. К ним сводятся, например, все задачи о равновесии механических систем.

Формально любую краевую задачу можно свести к вариационной, причем эта редукция связана со способом введения обобщенного решения.

Условимся говорить, что x есть обобщенное решение задачи (24), если он обращает в нуль функционал

$$I = \| Ax \|. \tag{26}$$

Тогда задачу отыскания обобщенного решения мы можем заменить задачей отыскания минимума функционала (26). Множество решений последней задачи содержит множество обобщенных решений, поэтому среди найденных решений вариационной задачи мы должны отобрать те, которые обращают в нуль функционал (26).

В Вычислительном центре АН СССР такой подход к решению краевых задач неоднократно применялся с использованием метрики L_2.

2. Сведение краевой задачи к вариационной часто служит средством для применения численных методов и, в частности, позволяет применить метод Ритца. Однако этот метод требует предварительного построения полной системы координатных функций. Методы, излагаемые в данной работе, лишены этого недостатка.

Если вектор x является функцией одной переменной, т. е. уравнение (24) есть система обыкновенных дифференциальных уравнений, то использование методов, изложенных в первых параграфах этой работы, проводится по стандартной схеме и не требует никаких комментариев.

Если вектор x — функция двух и более переменных, то методы типа динамического программирования перестают быть применимыми, а градиентные методы сохраняют свою силу. В то же время методы динамического программирования позволяют разыскивать экстремумы функционалов на множествах «большого объема». Поэтому, располагая вычислительной машиной с большой памятью и быстродействием, целесообразно иметь в своем распоряжении алгоритмы, подобные алгоритмам киевского веника, приспособленные для решения краевых задач в случае функций многих переменных.

Для этого достаточно свести исходную задачу к системе обыкновенных дифференциальных уравнений. В вычислительной практике последнего десятилетия подверглись апробации разнообразные способы такой редукции. Среди методов этого рода я хотел бы отметить прежде всего метод прямых и метод интегральных соотношений А. А. Дородницына.

Рассмотрим два примера, поясняющих технику решения задачи.

3. Пусть нам требуется найти решение уравнения

$$\ddot{x} + F(x,\ t) = 0, \tag{27}$$

удовлетворяющее следующим краевым условиям:

$$x(0) = x(1) = 0. \tag{28}$$

Запишем функционал (26) в виде

$$I = \int_0^1 (\ddot{x} + F(x,\ t))^2\, dt \tag{29}$$

и заменим его интегральной суммой. Например, мы можем взять следующее аппроксимирующее выражение:

$$I = \tau \sum_i \left[\frac{x_{i+1} + x_{i-1} - 2x_i}{\tau^2} - F(x_i,\ t_i) \right]^2. \tag{30}$$

Если мы решили разыскивать минимум выражения (30) методом динамического программирования, то к числу фазовых переменных удобно добавить $y_i = \frac{x_i - x_{i-1}}{\tau}$. Тогда вместо выражения (30) мы будем иметь

$$I = \tau \sum_i \left\{ \left[\frac{x_{i+1} - x_i}{\tau} - y_i \right]^2 + \left[\frac{y_{i+1} - y_i}{\tau} - F(x_i, t_i) \right]^2 \right\}. \tag{31}$$

В зависимости от характера функции F функционал (30) может иметь то или другое количество локальных экстремумов. Поэтому, прежде чем применять методы наискорейшего спуска, целесообразно применить сначала (с большим шагом τ) алгоритм киевского веника. Найденную фазовую траекторию мы примем затем в качестве первого приближения и, уменьшив шаг τ, используем один из вариантов градиентного метода, например метод локальных вариаций.

Заметим, что весь процесс решения производится по стандартным программам и не требует квалифицированного персонала.

В рассматриваемой задаче нет фазовых ограничений. Поэтому для проверки точности решения нам достаточно контролировать значения величин $\frac{\partial I}{\partial x_i}$.

Краевые условия (28) приводят к тому, что значения x_0 и x_N не варьируются — они фиксированы.

Примечание. При решении нелинейных краевых задач общепринятыми методами (метод нелинейной прогонки, метод подбора недостающих условий) возникает целый ряд трудностей, связанных с неустойчивостью решения задачи Коши. Особые трудности реализации вычислительной процедуры встречаются в том случае, когда решение имеет характер пограничного слоя. Для изложенного метода эти обстоятельства не вызывают трудностей. Однако процесс итераций идет значительно быстрее, если мы используем априорную информацию об асимптотическом поведении решения для выбора первого приближения или той области фазового пространства, где мы предполагаем отыскать решение.

4. Выше мы уже отмечали, что методы типа наискорейшего спуска в пространстве состояний могут быть использованы в задачах произвольной размерности. Увеличение размерности задачи приводит только к увеличению ее объема. В настоящее время в Вычислительном центре АН СССР проводятся исследования возможностей подобных методов при решении краевых задач для уравнений в частных производных. Приведем один из примеров подобного рода. Рассмотрим плоскую задачу для уравнений Навье — Стокса:

$$\frac{\partial^2\psi}{\partial\xi^2}+\frac{\partial^2\psi}{\partial\eta^2}=\omega,$$

$$\frac{\partial^2\omega}{\partial\xi^2}+\frac{\partial^2\omega}{\partial\eta^2}=F\left(\frac{\partial\psi}{\partial\xi};\ \frac{\partial\psi}{\partial\eta};\ \frac{\partial\omega}{\partial\xi};\ \frac{\partial\omega}{\partial\eta}\right), \qquad (32)$$

где

$$F=\frac{D(\psi,\ \omega)}{D(\xi,\ \eta)}.$$

Для того чтобы применить метод наискорейшего спуска, мы должны составить выражение функционала и выписать его конечно-

разностную аппроксимацию. Если ограничиться простейшими схемами, то мы получим

$$I=h_\xi h_\eta\sum_{i,\ j}\left\{\left[\frac{\psi_{i+1,\ j}+\psi_{i-1,\ j}-2\psi_{ij}}{h_\xi^2}+\frac{\psi_{i,\ j+1}+\psi_{i,\ j-1}-2\psi_{ij}}{h_\eta^2}-\omega_{ij}\right]^2+\right.$$

$$+\left[\frac{\omega_{i+1,\ j}+\omega_{i-1,\ j}-2\omega_{ij}}{h_\xi^2}+\frac{\omega_{i,\ j+1}+\omega_{i,\ j-1}-2\omega_{ij}}{h_\eta^2}-\right.$$

$$\left.\left.-F\left(\frac{\psi_{i+1,\ j}-\psi_{ij}}{h_\xi};\ \frac{\psi_{i,\ j+1}-\psi_{ij}}{h_\eta};\ \frac{\omega_{i,\ j+1}-\omega_{ij}}{h_\eta};\ \frac{\omega_{i+1,\ j}-\omega_{ij}}{h_\xi}\right)\right]^2\right\}. \qquad (33)$$

Здесь h_ξ и h_η — шаги по обоим пространственным переменным ξ и η. Для того чтобы обеспечить достаточную точность аппроксимации, эти величины должны быть достаточно малыми. Это значит, что число переменных в выражении

$$I(\psi_{ij},\ \omega_{ij})$$

должно быть очень велико (порядка 10^4 и более). Однако значения производных

$$\frac{\partial I}{\partial\psi_{ij}},\ \frac{\partial I}{\partial\omega_{ij}}$$

выписываются в явном виде и их вычисление не занимает много времени. В результате каждый шаг итерационного процесса занимает всего лишь несколько секунд машинного времени. Если предварительные качественные исследования позволяют написать удовлетворительное первое приближение, то замена краевой задачи для системы (32) задачей отыскания минимума функционала (33) позволяет развивать эффективные способы численного решения.

5. Примеры, которые мы привели, относились к краевым задачам и не содержали ограничений на фазовые координаты. Однако существует много важных задач, содержащих фазовые ограничения и допускающих вариационную постановку. К их числу относятся, например, задачи с освобождающими связями, контактные задачи теории упругости и многие другие.

В этих случаях вместо градиентного способа употребляются методы проекции градиента (либо другие модификации метода возможных направлений). В задачах этого типа удобно использование фиксированной сетки не только по независимым, но и по фазовым переменным. Ограничения могут быть учтены заданием узлов только в допустимой области. Фиксация шага по фазовым переменным снижает, разумеется, эффективность метода наискорейшего спуска. Спуск с фиксированными узлами в этой ситуации может оказаться более удобным для машинной реализации, нежели одновременное варьирование всей траектории, как в случае градиентного спуска.

Ф. Л. Черноусько вместе со студентами дал решение ряда подобных задач, используя координатный спуск с фиксированным шагом по фазовым переменным (метод локальных вариаций).

Рассмотрим мембрану под действием распределенных сил, действующих перпендикулярно к ее поверхности (рис. 3). В невозмущенном состоянии мембрана находится в плоскости $\xi = 0$. Точки

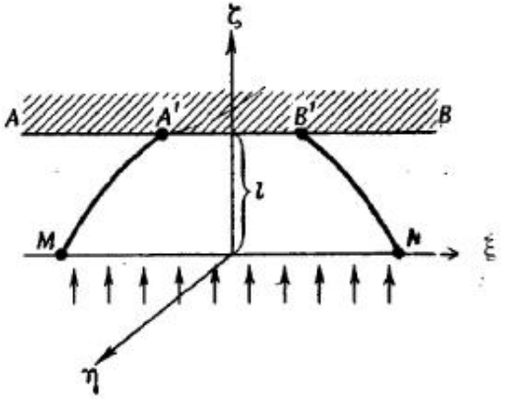


Рис. 3.

M и N — это точки пересечения ее контура с плоскостью $\eta = 0$. Задача о прогибе мембраны под действием распределенных сил допускает вариационную формулировку.

Предположим теперь, что существует абсолютно жесткая стенка AB, ограничивающая прогиб мембраны. Это обстоятельство делает задачу существенно нелинейной.

При сравнительно небольшом числе итераций, потребовавшем 15—20 минут машинного времени на машине класса БЭСМ-2, авторы получили решение этой задачи, причем с большой точностью указали положение точек контакта A^1 и B^1.

Заключение

В докладе мы дали описание цикла работ, целью которых было отыскание рациональных вычислительных методов решения вариационных задач. В этих работах принимали участие многие сотрудники и аспиранты Вычислительного центра АН СССР, которые внесли в них большой вклад[1]). Изложенные методы решения задач

[1]) В 1965 г. Гриншпан (Greenspan) опубликовал метод решения простейших вариационных задач, использующий примерно те же идеи. Автор, по-видимому, не знал о наших работах. Во всяком случае, в его статье отсутствуют ссылки на наши более ранние публикации.

возникали в связи с необходимостью решать задачи оптимального управления с фазовыми ограничениями. Варьирование в пространстве состояний позволяет избежать трудностей, которые появляются в тех случаях, когда для численного решения используют необходимые условия экстремума. Та численная процедура, которая была в результате разработана, по существу является дальнейшим развитием метода ломаных Эйлера.

В процессе совершенствования техники решения вариационных задач выяснилось, что метод может быть применен к более широкому кругу задач, нежели те, для которых он был разработан, и наши усилия были направлены прежде всего на изучение новых возможностей его применения. В докладе было показано, что идея варьирования в пространстве состояний может применяться и в задачах планирования, и в математической физике. По-видимому, она может быть использована и в целом ряде других задач, например в задачах управления системами с распределенными параметрами и теории многошаговых игр. Что касается последней, то ей следует посвятить самое пристальное внимание. Вычислительные методы теории операций (анализ конфликтных ситуаций) разработаны очень плохо, в то же время ее значение непрерывно растет.

Несмотря на значительное количество задач разной природы, которые были решены с использованием вариаций в пространстве состояний, изложенные методы не претендуют на универсальность. Идея создания универсальных численных методов решения задач оптимизации в настоящее время совершенно утопична. Успешное решение каждой трудной задачи этого рода всегда является следствием предварительного тонкого аналитического (или качественного) исследования ее природы. Такое исследование позволяет сделать правильный выбор численного метода.

Теория, которая была развита в докладе, дает один из возможных подходов к численному решению вариационных задач. Ее полезность подтверждается нашим опытом, и доклад имеет своей целью привлечь внимание к рассматриваемым проблемам. В ней есть много нерешенных вопросов не только прикладного, но и теоретического характера.

Вычислительный центр АН СССР,
Москва, СССР

ЛИТЕРАТУРА

[1] Багаева Н. Я., Моисеев Н. Н., Об одном способе численного решения задач оптимального управления, *ДАН СССР*, **153**, № 4 (1963), 347-750.

[2] Моисеев Н. Н., Методы динамического программирования в теории оптимальных управлений — I, *Ж. вычисл. математики и матем. физики*, **4**, № 3 (1964), 485-494.

[3] Моисеев Н. Н., Методы динамического программирования в теории оптимальных управлений — II, *Ж. вычисл. математики и матем. физики*, 5, № 1 (1965), 44-56.
[4] Багаева Н. Я., Решение двух неклассических задач оптимального управления. Доклад на II съезде по теоретической и прикладной механике, Москва, 1964.
[5] Михалевич В. С., Шор Н. З., О численных методах решения многовариантных плановых и технико-экономических задач. Научно-методические материалы экономико-математического семинара, ВЦ АН СССР, вып. 1, Киев, 1962.
[6] Черноусько Ф. Л., Метод локальных вариаций для численного решения вариационных задач, *Ж. вычисл. математики и матем. физики*, 5, № 4 (1965), 749-754.
[7] Крылов И. А., Черноусько Ф. Л., Решение задач оптимального управления методом локальных вариаций, *Ж. вычисл. математики и матем. физики*, 6 № 2 (1966), 203-218.
[8] Черенин В. П., Решение некоторых комбинаторных задач оптимального планирования методом последовательных расчетов. Материалы к конференции по опыту и перспективам применения математических методов и электронных вычислительных машин в планировании, Новосибирск, 1962.

4

Вычислительная математика
Numerical mathematics
Mathématique numérique
Numerische Mathematik

COMPUTER EXPERIMENTS WITH THE BIEBERBACH CONJECTURE

PAUL R. GARABEDIAN

It is a classical problem to look for necessary and sufficient conditions on the coefficients $a_2, a_3, \ldots$ of an analytic function

$$f(z) = z + a_2 z^2 + a_3 z^3 + \ldots$$

defined in the unit circle in order that it be schlicht there. In 1916 Bieberbach made the famous conjecture that

$$|a_n| \leqslant n$$

for any such function, with the equality sign holding only for the Koebe mapping

$$\frac{z}{(1-z)^2} = z + 2z^2 + 3z^3 + \ldots$$

and its rotations. Recently inequalities for expressions involving the coefficients a_k and certain parameters λ_k have been found which imply that

$$|a_2| \leqslant 2, \quad |a_3| \leqslant 3, \quad |a_4| \leqslant 4$$

without restriction, and which moreover give

$$|a_5| \leqslant 5, \quad |a_{2m}| \leqslant 2m$$

for schlicht functions $f(z)$ close enough to the Koebe function. It is our intention to describe here the role that numerical computation plays in this analysis.

We introduce matrices $a_n^{(l)}$ and c_{jk} by means of the generating functions

$$f(z)^l = \sum_{n=l}^{\infty} a_n^{(l)} z^n,$$

$$\log \frac{\sqrt{f(z^2)} - \sqrt{f(w^2)}}{z-w} = \frac{1}{2} \sum_{j,k=0}^{\infty} c_{jk} z^j w^k.$$

It can be shown by Schiffer's variational method that

$$\mathrm{Re}\left\{\lambda_0^2 \log\frac{\alpha}{4} + \sum_{j,k=0}^{2m} c_{jk}\lambda_j\lambda_k + \right.$$

$$\left. + \sum_{j,k=0}^{m} \Gamma_{jk}\alpha^{j+k} \sum_{l=j}^{m} \lambda_{2l}a_l^{(j)} \sum_{l=k}^{m} \lambda_{2l}a_l^{(k)}\right\} \leqslant 2\sum_{j=1}^{2m} \frac{|\lambda_j|^2}{j},$$

where α is any root of the polynomial equation

$$\sum_{k=0}^{m} \Gamma_k\alpha^k \sum_{l=k}^{m} \lambda_{2l}a_l^{(k)} = 0,$$

where the parameters $\lambda_0, \ldots, \lambda_{2m}$ are confined to a specified range, and where Γ_k and Γ_{jk} are known numerical factors. This inequality gives $|a_2| \leqslant 2$ with $m = 1$ and $\lambda_0 = \lambda_2 = 0$, it gives $|a_3| \leqslant 3$ with $m = 1$ and $\lambda_1 = 0$, and it gives $|a_4| \leqslant 4$ with $m = 2$ and $\lambda_0 = \lambda_2 = \lambda_4 = 0$. Moreover, in appropriate neighborhoods of the Koebe function it yields $|a_5| \leqslant 5$ with $m = 2$ and it yields $|a_{2m}| \leqslant 2m$ for general m and $\lambda_0 = \lambda_2 = \ldots = \lambda_{2m} = 0$.

The local theorem we have just stated is relatively easy to obtain by expanding our fundamental coefficient inequality in a Taylor series about the conjectured extremal point $a_n = n$. To establish that the inequality ought to imply Bieberbach's conjecture in the large, too, we have had to resort to computer experiments because of the nonlinearity of the problem. The purpose of the computations has been to ascertain which inequalities may be expected to contain the conjecture and should therefore be subjected to a more penetrating analysis.

By tabulating at 40^5 points a simplified version of our inequality depending on only five variables and corresponding to $m = 3$ and $\lambda_0 = \lambda_2 = \lambda_4 = \lambda_6 = 0$, George Ross has compiled convincing evidence that $|a_6| \leqslant 6$. The question of a_5 is harder because it involves no less than eleven independent variables. Using a Monte-Carlo technique, Neal Friedman has just finished calculations that indicate the success of our method in this case as well. We hope that such data will stimulate interest in developing a complete proof of the Bieberbach conjecture along the lines we have suggested.

New York University,
Courant Institute of Mathematical Sciences, USA

A PRIORI ERROR ANALYSIS OF ALGEBRAIC PROCESSES

J. H. WILKINSON

Introduction

Modern error analysis of matrix processes may be said to have started with the work of Goldstine and von Neumann in their analysis of the fixed-point elimination of a positive definite matrix [4, 6]. Since then a considerable number of analyses of algorithms for the inversion of matrices and the calculation of eigensystems have been made, covering both fixed-point and floating-point computation. The stage has now been reached at which this work may be reviewed as a whole and the significance of *a priori* analyses reassessed.

1. Mathematical properties of Cholesky factorization

As an example of the general type of analysis we have in mind we reconsider the Cholesky factorization of a positive definite matrix, using floating-point arithmetic without accumulation of inner-products. The result proved here has been well known to the author for several years and I have quoted it from time to time but there does not appear to be an explicit proof in the literature.

We consider first the simple mathematical properties of the factorization. Let $A^{(1)}$ be a positive definite matrix of order n and $L^{(1)}$ its Cholesky factor. Writing

$$A^{(1)} = \begin{bmatrix} a_{11}^{(1)} & a_1^T \\ a_1 & A^{(2)} \end{bmatrix}, \quad L^{(1)} = \begin{bmatrix} l_{11} & 0 \\ l_1 & L^{(2)} \end{bmatrix} \tag{1.1}$$

we have

$$l_{11}^2 = a_{11}^{(1)}, \quad l_{11}l_1 = a_1, \quad l_1l_1^T + L^{(2)}(L^{(2)})^T = A^{(2)} \tag{1.2}$$

and we may define $B^{(2)}$ by the relation

$$B^{(2)} = L^{(2)}(L^{(2)})^T = A^{(2)} - l_1l_1^T. \tag{1.3}$$

Theorem 1. $B^{(2)}$ is positive definite and

$$\|B^{(2)}\|_2 \leqslant \|A^{(2)}\|_2 \leqslant \|A^{(1)}\|_2,$$

$$\|l_1l_1^T\|_2 = \|l_1\|_2^2 = \frac{\|a_1\|_2^2}{a_{11}^{(1)}} \leqslant \|A^{(2)}\|_2 \leqslant \|A^{(1)}\|_2.$$

The final inequality in each group is an immediate consequence of the classical separation theorem for symmetric matrices.

If we assume that $B^{(2)}$ is not positive definite then there exists an $x \neq 0$ such that

$$x^T B^{(2)} x \leqslant 0, \tag{1.4}$$

giving

$$x^T A^{(2)} x - x_1^T l_1 l_1^T x = x^T A^{(2)} x - \frac{(a_1^T x)^2}{a_{11}^{(1)}} \leqslant 0. \tag{1.5}$$

Taking $y^T = (\alpha \mid x^T)$ we have

$$y^T A^{(1)} y = a_{11}^{(1)} \alpha^2 + 2\alpha (a_1^T x) + x^T A^{(2)} x. \tag{1.6}$$

With $\alpha = -(a_1^T x)/a_{11}^{(1)}$ this gives

$$y^T A^{(1)} y = -(a_1^T x)^2 / a_{11}^{(1)} + x^T A^{(2)} x \leqslant 0 \tag{1.7}$$

by (1.5), contradicting the positive definiteness of $A^{(1)}$, and hence the hypothesis that $B^{(2)}$ is not positive definite is false.

Now for any z we have

$$z^T B^{(2)} z = z^T A^{(2)} z - (l_1^T z)^2 \leqslant z^T A^{(2)} z. \tag{1.8}$$

Hence $\| B^{(2)} \|_2 \leqslant \| A^{(2)} \|_2$. Finally if x is defined by

$$x = \frac{l_1}{\| l_1 \|_2} \quad (\| x \|_2 = 1), \tag{1.9}$$

then $x^T B^{(2)} x = x^T A^{(2)} x - \| l_1 \|_2^2 \leqslant \| A^{(2)} \|_2 - \| l_1 \|_2^2$ giving

$$\| l_1 \|_2^2 \leqslant \| A^{(2)} \|_2 - x^T B^{(2)} x \leqslant \| A^{(2)} \|_2 \tag{1.10}$$

since $B^{(2)}$ is positive definite.

We observe that the Cholesky decomposition of $A^{(1)}$ of order n consists of

(i) the determination of the first column of $L^{(1)}$,

(ii) the Cholesky decomposition of the matrix $B^{(2)}$ of order $(n-1)$.

In the same way after finding the first column of $L^{(2)}$ we required the Cholesky decomposition of a matrix $B^{(3)}$ of order $(n-2)$ etc.

2. Errors in floating-point arithmetic

We shall make certain assumptions about the rounding errors made in the basic operations. These differ a little from one computer to another [12] but not in such a way as to alter materially any of our results. Specifically we assume

$$\left.\begin{aligned} \mathrm{fl}(a \pm b) &= a(1+\varepsilon_1) \pm b(1+\varepsilon_2), \\ \mathrm{fl}(a \times b) &= ab(1+\varepsilon_3), \\ \mathrm{fl}(a/b) &= a(1+\varepsilon_4)/b \end{aligned}\right\} |\varepsilon_i| \leqslant 2^{-t} \tag{2.1}$$

where a mantissa of t binary digits is used. The notation $\mathrm{fl}(a + b)$ etc. denotes the result of adding two floating-point number a and b using standard floating-point arithmetic.

An analysis of repeated additions, multiplications etc. leads to bounds involving terms of the type $(1 \pm 2^{-t})^r$ which are somewhat inconvenient. If $r2^{-t} < 0.1$ it is easy to show that

$$(1+2^{-t})^r < 1 + (1.06) r 2^{-t} = 1 + r2^{-t_1}, \tag{2.2}$$

$$(1-2^{-t})^r > 1 - (1.06) r 2^{-t} = 1 - r2^{-t_1} \tag{2.3}$$

where $t_1 = t - \log_2(1.06)$.

Notice that t_1 is only marginally different from t. In practice storage considerations and time limitations always ensure that the restriction on r is met in any case.

The only other error bound we shall require is that for the square root. This bound depends to some extent on the procedure which is used; we shall assume that if

$$x = \mathrm{fl}(\mathrm{sqrt}\, a) \tag{2.4}$$

then

$$x^2 = a(1+\varepsilon) \quad \text{with } |\varepsilon| \leqslant 2.2^{-t_1}. \tag{2.5}$$

This result is not critical in our analysis.

3. The practical Cholesky process

We now turn to the computational process itself. For simplicity l_{ij}, $b_{ij}^{(2)}$ will be used to denote the actual computed quantities. No confusion need arise because we do not compare these quantities with those arising in the exact Cholesky factorization of $A^{(1)}$. The elements of $L^{(1)}$ may be determined column by column the relevant equations being

$$l_{jj} = \mathrm{fl}\,[\mathrm{sqrt}\,(a_{jj}^{(1)} - l_{j1}l_{j1} - l_{j2}l_{j2} - \ldots - l_{j,j-1}l_{j,j-1})], \tag{3.1}$$

$$l_{ij} = \mathrm{fl}\,[(a_{ij}^{(1)} - l_{i1}l_{j1} - l_{i2}l_{j2} - \ldots - l_{i,j-1}l_{j,j-1})/l_{jj}] \tag{3.2}$$

where, of course, the previously computed l_{pq} are used when evaluating the right-hand sides.

We may express the computation of the right-hand side of (3.2) say in the form

$$\left.\begin{aligned} b_{ij}^{(2)} &= \mathrm{fl}(a_{ij}^{(1)} - l_{i1}l_{j1}), \quad i,j = 2,\ldots,n, \\ b_{ij}^{(3)} &= \mathrm{fl}(b_{ij}^{(2)} - l_{i2}l_{j2}), \quad i,j = 3,\ldots,n, \\ b_{ij}^{(k+1)} &= \mathrm{fl}(b_{ij}^{(k)} - l_{ik}l_{jk}), \quad i,j = k+1,\ldots,n. \end{aligned}\right\} \tag{3.3}$$

where the $b_{ij}^{(k)}$ are the computed elements of the $B^{(k)}$ discussed at the end of section 1.

The full practical factorization of $A^{(1)}$ is therefore computationally equivalent (including equivalence of rounding errors) with the computation of the first colum of $L^{(1)}$ and the computation of $B^{(2)}$ followed by the practical factorization of the computed $B^{(2)}$, a matrix of order $(n-1)$.

Theorem 2. If A is a positive definite digital matrix of order n then provided

$$\lambda_{\min} = \frac{1}{\| A^{-1} \|_2} \geqslant 20n^{3/2}2^{-t_1} \| A \|_2 \tag{3.4}$$

the Cholesky factor L can be computed without breakdown and the computed L satisfies the relation

$$LL^T = A + E,$$

$$\| E \|_2 \leqslant 2.2^{-t_1}[1+(n^{1/2}+2.2)2^{-t_1}]^n \times \left[\frac{2}{3}(n+1)^{3/2}+1.1n\right] \| A \|_2. \tag{3.5}$$

In other words L is the exact Cholesky factor of some matrix $A+E$ where E satisfies the given bound.

The condition (3.4) ensures that $A+E$ is positive definite. It may be written in the form

$$2n^{3/2}2^{-t_1} \| A \|_2 \| A^{-1} \|_2 \leqslant 0.1, \tag{3.6}$$

i.e.

$$2n^{3/2}2^{-t_1} \varkappa(A) \leqslant 0.1. \tag{3.7}$$

Since $\varkappa(A) \not< 1$ for any matrix relation (3.7) gives *a fortiori*

$$2n^{3/2}2^{-t_1} \leqslant 0.1. \tag{3.8}$$

The condition $n2^{-t} < 0.1$ is therefore certainly satisfied.

The proof is by induction. Assume that the result is true for matrices of order $n-1$.

Consider now the computation of the first column of $L^{(1)}$ and of the matrix $B^{(2)}$ for the matrix $A^{(1)}$ of order n. We have

$$l_{11} = \mathrm{fl}\,[\mathrm{sqrt}\,(a_{11}^{(1)})] \quad \text{giving } l_{11}^2 = a_{11}^{(1)}(1+\varepsilon_{11}) \quad (|\varepsilon_{11}| < 2.2^{-t_1}) \tag{3.9}$$

from (2.5). The remaining elements of the first column of $L^{(1)}$ are defined by

$$l_{i1} = \mathrm{fl}\,(a_{i1}^{(1)}/l_{11}) = a_{i1}^{(1)}(1+\varepsilon_{i1})/l_{11} \quad (|\varepsilon_{i1}| < 2^{-t_1}), \tag{3.10}$$

$$l_{i1}l_{11} = a_{i1}^{(1)}(1+\varepsilon_{i1}) \quad (i = 2, \ldots, n). \tag{3.11}$$

Equations (3.9) and (3.11) state that $l_{11}, \ldots, l_{n1}$ are exact for the matrix $A^{(1)}+F^{(1)}$ where $F^{(1)}$ is defined by

$$f_{11}^{(1)} = a_{11}^{(1)}\varepsilon_{11}, \quad f_{11}^{(1)} = f_{1i}^{(1)} = a_{11}^{(1)}\varepsilon_{i1}, \quad f_{ij}^{(1)} = 0 \text{ otherwise.} \tag{3.12}$$

We therefore have certainly

$$\| F^{(1)} \|_2 \leqslant \| \, | F^{(1)} | \, \|_2 \leqslant 2.2^{-t_1} \| \, | A^{(1)} | \, \|_2 \leqslant 2n^{1/2}2^{-t_1} \| A^{(1)} \|_2. \tag{3.13}$$

It follows from (3.4) that $A^{(1)}+F^{(1)}$ is positive definite and hence from theorem 1

$$\| l_1 l_1^T \|_2 = \| l_1 \|_2^2 \leqslant \| A^{(1)} + F^{(1)} \|_2 \leqslant (1 + 2n^{1/2}2^{-t_1}) \| A^{(1)} \|_2. \tag{3.14}$$

The elements of $B^{(2)}$ are defined by

$$b_{ij}^{(2)} = \mathrm{fl}\,(a_{ij}^{(1)} - l_{i1}l_{j1}) = a_{ij}^{(1)}(1+\varepsilon_{ij}) - l_{i1}l_{j1}(1+\eta_{ij}) =$$

$$= (a_{ij}^{(1)} + a_{ij}^{(1)}\varepsilon_{ij} - l_{i1}l_{j1}\eta_{ij}) - l_{i1}l_{i2} \quad (|\varepsilon_{ij}| < 2^{-t_1}, \ |\eta_{ij}| < 2.2^{-t_1}). \tag{3.15}$$

Hence we can say that the computed first column of $L^{(1)}$ and the computed $B^{(2)}$ would be obtained exactly with the matrix $A^{(1)}+E^{(1)}$ where $E^{(1)}$ is defined by

$$e_{11}^{(1)} = a_{11}^{(1)}\varepsilon_{11}, \quad e_{1i}^{(1)} = e_{i1}^{(1)} = a_{i1}^{(1)}\varepsilon_{i1}, \tag{3.16}$$

$$e_{ij}^{(1)} = a_{ij}^{(1)}\varepsilon_{ij} - l_{i1}l_{j1}\eta_{ij} \quad (i, j = 2, \ldots, n). \tag{3.17}$$

From the bound on the ε_{pq} and the η_{pq}

$$|e_{11}^{(1)}| \leqslant 2.2^{-t_1}|a_{11}^{(1)}|, \quad |e_{1i}^{(1)}| = |e_{i1}^{(1)}| \leqslant 2^{-t_1}|a_{i1}^{(1)}|, \tag{3.18}$$

$$|e_{ij}^{(1)}| \leqslant 2^{-t_1}|a_{ij}^{(1)}| + 2.2^{-t_1}|l_{i1}||l_{j1}|, \tag{3.19}$$

and hence

$$\| E^{(1)} \|_2 \leqslant \| \, | E^{(1)} | \, \|_2 \leqslant 2.2^{-t_1} \| \, | A^{(1)} | \, \|_2 + 2.2^{-t_1} \| \, | l_1 | \, | l_1^T | \, \|_2 =$$

$$= 2.2^{-t_1} \| \, | A^{(1)} | \, \|_2 + 2.2^{-t_1} \| \, | l_1 | \, \|_2^2 \leqslant$$

$$\leqslant 2.2^{-t_1}(n^{1/2} + 1 + 2n^{1/2}2^{-t_1}) \| A^{(1)} \|_2 \tag{3.20}$$

since $\| \, | l_1 | \, \|_2 = \| l_1 \|_2$.

Again applying Theorem 1 we have

$$\| B^{(2)} \|_2 \leqslant \| A^{(2)} + G^{(1)} \|_2 \tag{3.21}$$

where $G^{(1)}$ is the matrix $E^{(1)}$ without its first row and column. Clearly

$$\| B^{(2)} \|_2 \leqslant \| A^{(2)} \|_2 + \| \, | G^{(1)} | \, \|_2 \leqslant$$

$$\leqslant \| A^{(2)} \|_2 + 2^{-t_1}\left[\| \, | A^{(2)} | \, \|_2 + 2(1 + 2n^{1/2}2^{-t_1}) \| A^{(1)} \|_2\right] \leqslant$$

$$\leqslant \| A^{(1)} \|_2 \{1 + [n^{1/2} + 2(1 + 2n^{1/2}2^{-t_1})] \, 2^{-t_1}\}. \tag{3.22}$$

We may use the overall bound given in (3.8) to simplify these results. The relations (3.20), (3.22) become certainly

$$\| E^{(1)} \|_2 \leqslant 2.2^{-t_1}(n^{1/2}+1.1)\| A^{(1)} \|_2 \tag{3.23}$$

and

$$\| B^{(2)} \|_2 \leqslant \| A^{(1)} \|_2 [1+(n^{1/2}+2.2)2^{-t_1}]. \tag{3.24}$$

From our inductive hypothesis we know that $L^{(2)}$ which is the computed Cholesky factor derived from the computed $B^{(2)}$ satisfies the relation

$$L^{(2)}(L^{(2)})^T = B^{(2)} + E^{(2)} \tag{3.25}$$

where

$$\| E^{(2)} \|_2 \leqslant 2.2^{-t_1}\{1+[(n-1)^{1/2}+2.2]2^{-t_1}\}^{n-1} \times \times \left[\frac{2}{3}n^{3/2}+(1.1)(n-1)\right]\| B^{(2)} \|_2. \tag{3.26}$$

Combining the first column of $L^{(1)}$ with $L^{(2)}$ we obtain the computed $L^{(1)}$ and this satisfies

$$L^{(1)}(L^{(1)})^T = A^{(1)} + E^{(1)} + \bar{E}^{(2)} \tag{3.27}$$

where $\bar{E}^{(2)}$ is $E^{(2)}$ augmented by a null first row and column and

$$\| E^{(1)} + \bar{E}^{(2)} \|_2 \leqslant \| E^{(1)} \|_2 + \| \bar{E}^{(2)} \|_2. \tag{3.28}$$

A little manipulation gives certainly

$$\| E^{(1)} + \bar{E}^{(2)} \|_2 \leqslant 2.2^{-t_1}\left[\frac{2}{3}(n+1)^{3/2}+1.1n\right] \times \times [1+(n^{1/2}+2.2)2^{-t_1}]^n \| A^{(1)} \|_2 \tag{3.29}$$

which is the required result.

For values of n greater than 10 satisfying condition (3.8) the result (3.29) can be expressed in the simpler form

$$LL^T = A+E, \quad \| E \|_2 \leqslant 2.5n^{3/2}2^{-t_1}\| A \|_2. \tag{3.30}$$

The constant 2.5 has not been computed with any great care but could not be reduced to less than (say) 2.3.

4. Accuracy of a computed inverse

To invert A the most convenient procedure is to compute L^{-1} and then $(L^{-1})^T L^{-1}$, the full computation including that of L requiring $\frac{1}{2}n^3$ multiplications. Error analyses of the two remaining steps have been given elsewhere [12, 14]. They show that when A is not well-conditioned the error made in the Cholesky decomposition is

easily the most damaging. If the other two parts were performed exactly the computed inverse would be $X=(A+E)^{-1}$ and we have

$$\frac{\| X-A^{-1} \|_2}{\| A^{-1} \|_2} \leqslant \frac{\| E \|_2 \| A^{-1} \|_2}{1-\| E \|_2 \| A^{-1} \|_2}. \tag{4.1}$$

provided $\| E \|_2 \| A^{-1} \|_2 < 1$. Since we have

$$\| E \|_2 \| A^{-1} \|_2 \leqslant 2.5n^{3/2}2^{-t_1}\| A \|_2 \| A^{-1} \|_2 < 0.125 \tag{4.2}$$

by condition (3.6) the existence of $(A+E)^{-1}$ is assured. Provided (3.6) is satisfied (4.1) can be written in the form

$$\frac{\| X-A^{-1} \|_2}{\| A^{-1} \|_2} \leqslant 2.86n^{3/2}2^{-t_1}\varkappa(A). \tag{4.3}$$

To put the result in perspective the result obtained by Goldstine and von Neumann [6] for the fixed-point inversion of a positive definite matrix was effectively

$$\frac{\| X-A^{-1} \|_2}{\| A^{-1} \|_2} \leqslant 14.24n^2 2^{-t_1}\varkappa(A) \tag{4.4}$$

so that (4.3) is significantly sharper.

As far as I know it is the sharpest bound so far attained though greater attention to trivial details would obviously reduce the factor 2.86 to some extent. This seems scarcely worth considering.

A similar analysis based on the use of floating-point arithmetic but with accumulation of inner-products gives the bound

$$\frac{\| X-A^{-1} \|_2}{\| A^{-1} \|_2} \leqslant 3.14n^{1/2}2^{-t_1}\varkappa(A). \tag{4.5}$$

5. Comments on the error bound

The first comment one might make is that in one sense they can scarcely be regarded as *a priori* bounds since one cannot make practical use of them without a knowledge of $\varkappa(A)$. This is certainly true but I feel that this objection reduces to a mere argument about words. My claim is that the bound does relate the accuracy of the computed result with the inherent sensitivity of the original problem and we cannot expect a general result to do more than that.

The analysis has been carried out in the now classical tradition. The result is completely rigorous and covers all errors arising from second and higher order effects. In this respect it has followed the model established by Goldstine and von Neumann in their pioneering paper. In my opinion the time has now arrived when this policy might well be reconsidered.

At the time when the Goldstine-von Neumann paper was written very pessimistic views were widely held on the cumulative effect of

rounding errors in matrix processes. In order to dispel this pessimism it was essential that the analysis should be completely above suspicion. Moreover it was desirable that the final result should give a clear concise and unequivocal statement concerning the overall limitations of the algorithm since few would have time to read the details of the analysis.

In the light of experiences over the last twenty years I doubt whether present-day error analysts ever carry out such a meticulous analysis for their own benefit. For example, the bare essentials of the result given above were established on a single sheet of paper and left no doubt that a bound of the form

$$\| E \|_2 \leqslant k n^{3/2} 2^{-t} \| A \|_2 \tag{5.1}$$

(where k was some suitable constant) could be rigorously established. The rigour was added afterwards and was a tedious but, by now, routine process.

Again, in my opinion, it is quite common for too much weight to be attached to the precise bounds that have been obtained. Before elaborating on this point it is illuminating to make the following comments on the bounds given in (4.3) and (4.5) respectively.

Consider the upper bounds of the errors caused by replacing A by $A + F$ where F satisfies

$$\text{(i)} \quad |f_{ij}| \leqslant 2.86n\,|a_{ij}|\,2^{-t}, \tag{5.2}$$

$$\text{(ii)} \quad |f_{ij}| \leqslant 3.14\,|a_{ij}|\,2^{-t}. \tag{5.3}$$

It is immediately obvious that in the first case we have

$$\frac{\| (A+F)^{-1} - A^{-1} \|_2}{\| A^{-1} \|_2} \leqslant \frac{2.86 n^{3/2} \varkappa(A)}{1 - 2.86 n^{3/2} \varkappa(A)} \tag{5.4}$$

and in the second

$$\frac{\| (A+F)^{-1} - A^{-1} \|_2}{\| A^{-1} \|_2} \leqslant \frac{3.14 n^{1/2} \varkappa(A)}{1 - 3.14 n^{1/2} \varkappa(A)}. \tag{5.5}$$

These are essentially the same bounds as in (4.3) and (4.5) respectively. Notice that in (5.2) and (5.3) the bound on each element of F is proportional to the corresponding element of A.

If the matrix A is not exactly representable by t digit binary numbers then *ab initio* we must work with $A + G$ where

$$|g_{ij}| \leqslant |a_{ij}|\,2^{-t}. \tag{5.6}$$

Comparing (5.6) with (5.3) we see that the effect of rounding errors made during the solution are unlikely to be significantly more important than the effect of the initial digital representation. In (5.2) we have the additional factor n which is not surprising since the a_{nn} element at least is involved in n independent operations involving roundings. These considerations suggest that it is rather unlikely that either of our overall bounds can be improved upon in any significant way.

6. Inadequacy of the overall bounds

Experiments in which the computed errors are compared with the upper bounds given above seem to be becoming increasingly common. Inevitably they lead to the conclusion that even the bounds given above are much too pessimistic. This should occasion no surprise. Indeed a study of the error analysis itself as distinct from a concentration on the final result leads one to expect this.

For example, in the analysis no account is taken of the statistical distribution of the rounding errors. Unless n is small one might expect that for this reason alone the factors $n^{3/4}$ and $n^{1/4}$ in (4.3) and (4.5) could safely be replaced by $n^{3/4}$ and $n^{1/4}$ respectively. Again in order to obtain a succinct result the quantities $\|B^{(r)}\|_2$ have been replaced by $\|A^{(1)}\|_2$ for all relevant r though this will obviously be a severe overestimate in general.

When we turn to special classes of matrices further reductions can often be made. For example if A is positive $|A| = A$ and therefore $\| |A| \|_2 = \|A\|_2$. In the analysis of the first step $\| |A| \|_2$ has been replaced by $n^{1/2} \|A\|_2$ so that there is an unnecessary factor of $n^{1/2}$. Although this argument cannot be extended directly to the remaining steps it is likely that this factor persists. When A is a band matrix with bandwidth s, L is also a band matrix and the error matrix E is of the same width as A. A simple modification of the analysis shows that the factor $n^{3/2}$ in (4.3) can certainly be replaced by $n^{1/2}s$ and if s is small compared with n this is a substantial reduction.

These remarks are of course trivial and yet the errors made when working with matrices belonging to these special classes are often compared with those given in (4.3) and (4.5), or rather with the somewhat poorer overall bounds currently available in the literature! Comparisons of this kind are extremely misleading and may give the impression that the analysis leading up to the quoted bounds is suspect.

The finite segments of the Hilbert matrix are often used for test purposes. For such matrices the errors in the computed inverses compare particularly well with the bounds we have given especially when accumulation of inner-products is used. A study of the error analysis shows immediately why this should be. In the first instance such matrices are positive and hence there is probably a superfluous factor of $n^{1/2}$ in the bounds. Secondly, ignoring higher order effects, the analysis shows that when inner-products are accumulated the computed L satisfies the relation

$$LL^T = A + E \tag{6.1}$$

where

$$|e_{ij}| \leqslant 2^{-t} |l_{ij}||l_{jj}| \quad (i > j). \tag{6.2}$$

Hence if many of the elements of L are very small the corresponding elements of E are also particularly small. Now in the factorization of a Hilbert segment the l_{ij} do indeed get progressively smaller as i and j increase. This compensates to some extent the ill-conditioned nature of A. Indeed the l_{ij} become progressively smaller precisely because A is so ill-conditioned. Careful study of the details of the error analysis enables us to forecast the behaviour quite accurately and the overall bound in terms of $\|A\|_2$ and $\|A^{-1}\|_2$ is clearly seen to be irrelevant. Finally it is easy to see that it is the minimum value of $\varkappa(DAD)$, where D is any positive diagonal matrix, rather than $\varkappa(A)$ itself which should appear in the error bound. This minimum condition number has been discussed by Bauer [1] and Forsythe and Straus [2]. For Hilbert segments the minimum condition number is usually appreciably smaller than the condition number itself.

7. General conclusions

The following tentative conclusions may be drawn from the above discussion. The main purpose of an *a priori* analysis is to reveal the strengths and weaknesses of an algorithm. It is not worthwhile expending too much effort on the production of concise overall bounds since there are certain to be severe overestimates in general.

The essential points revealed by the analysis of sections 1—4 is that the Cholesky factorization will give satisfactory inverses for a useful range of condition numbers and values of n and that this is achieved without any form of pivoting. Although the accuracy attainable may be expected to vary from one choice of pivot to another this is in general a secondary effect, though with special types of matrices a certain choice of pivots might give an inverse of unexpectedly high accuracy having regard to the condition of the matrix. This contrasts with the situation for general matrices for which elimination methods without pivoting may fail to give any accuracy even for perfectly conditioned matrices (i.e. matrices for which $\varkappa(A) = 1$) of low order! Finally it shows that for *all* positive definite matrices the effect of rounding errors is unlikely to be more important than the initial rounding of the coefficients when inner-products are accumulated. For matrices of Hilbert type the error when inner-products are accumulated will generally be far less than that resulting from the initial roundings.

The overall upper bound itself should not be taken too seriously and it is important to study details of the error analysis in order to determine whether matrices of special classes can be expected to give exceptionally accurate results.

Only in rare cases will an *a priori* bound be of practical use. The best examples of such algorithms are those based on the unitary reduction of hermitian matrices. For several of these it may shown that the effect of rounding errors is equivalent to an initial hermitian perturbation E of A where the bound for $\|E\|_2$ is of the form $2^{-t} f(n) \|A\|_2$. Here $f(n)$ represents a simple function of n which depends on the algorithm and the type of arithmetic being used. The bound for E implies that if λ_i and λ_i' are the computed and true eigenvalues of A then

$$|\lambda_i - \lambda_i'| \leqslant 2^{-t} f(n) \|A\|_2 = 2^{-t} f(n) \max |\lambda_i|. \tag{7.1}$$

For several algorithms of this type $f(n)$ is sufficiently small for the bound given in (6.1) to be quite usable as it stands for many practical purposes [3, 5, 11, 13, 14, 15].

However, it is my opinion that in practice one should use *a posteriori* bounds for error estimates and not the overall *a priori* bounds, since the former take full advantage of the special nature of the matrix involved and of the distribution of rounding errors that have occurred. Discussion of such bounds has been given in [12, 14] for the linear equation problem and in [10, 14] for the algebraic eigenproblem.

Acknowledgements

The work described here has been carried out as part of the research programme of the National Physical Laboratory and is published by permission of the Director of the Laboratory.

Mathematical Division,
National Physical Laboratory, England

REFERENCES

[1] Bauer F. L., Optimally scaled matrices, *Numerische Math.*, **5** (1963), 73-87.
[2] Forsythe G. E., Straus E. G., On best conditioned matrices, *Proc. Amer. Math. Soc.*, **6** (1955), 340-345.
[3] Givens W., Numerical computation of the characteristic values of a real symmetric matrix, Oak Ridge National Laboratory, ORNL-1574, 1954.
[4] Goldstine H. H., von Neumann J., Numerical inverting of matrices of high order, II, *Proc. Amer. Math. Soc.*, **2** (1951), 188-202.
[5] Householder A. S., Generated error in rotational tridiagonalization, *J. Ass. Comp. Mech.*, **5** (1958), 335-338.
[6] von Neumann J., Goldstine H. H., Numerical inverting of matrices of high order, *Bull. Amer. Math. Soc.*, **53** (1947), 1021-1099.
[7] Newman M., Todd J., The evaluation of matrix inversion programs, *J. Soc. Industr. and App. Math.*, **6** (1958), 466-476.

[8] Ortega J. M., An error analysis of Householder's method for the symmetric eigenvalue problem, *Numerische Math.*, **5** (1963), 211-225.
[9] Turing A. M., Rounding errors in matrix processes, *Quart. J. Mech.*, **1** (1948), 287-308.
[10] Wilkinson J. H., Rigorous error bounds for computed eigensystems, *Computer J.*, **4** (1961), 230-241.
[11] Wilkinson J. H., Error analysis of eigenvalue techniques based on orthogonal transformations, *J. Soc. Industr. and Appl. Math.*, **10** (1962), 162-195.
[12] Wilkinson J. H., Rounding errors in algebraic processes, Notes on applied science, № 32, Her Majesty's Stationery Office, London; Prentice-Hall, New Jersey (1963).
[13] Wilkinson J. H., Plane rotations in floating-point arithmetic, *Proc. Amer. Math. Soc.*, Symposium in Applied Mathematics, **15** (1963), 185-198.
[14] Wilkinson J. H., The algebraic eigenvalue problem, Oxford University Press, London, 1965.
[15] Wilkinson J. H., Error analysis of transformations based on the use of matrices of the form $I - 2\,ww^H$. Error in digital computation (edited by L. B. Rall), John Wiley and Son, New York, 1965, 77-101.

ВЫЧИСЛИТЕЛЬНЫЕ МЕТОДЫ В ТЕОРИИ ПЕРЕНОСА

Г. И. МАРЧУК

Теория переноса излучения — одна из ведущих проблем современной науки, бурно развивающаяся на основе достижений теоретической физики и быстро проникающая в самые различные области естествознания и техники. В настоящее время трудно назвать область, которая могла бы обходиться без представлений и методов, выработанных в теории переноса и в которой эта теория не играла бы прогрессивной роли.

Теория переноса уже с начала XX столетия прочно вошла в теоретическую астрофизику (Шварцшильд, 1906 г.). Позднее она начала проникать в физику атмосферы, атмосферную оптику, оптику моря и даже физику земли — нейтронный и гамма-каротаж. Параллельно с этим процессом шло проникновение теории излучения в технику. Раньше других областей методы теории переноса излучения начали разрабатывать в светотехнике (20-е годы), позднее в теплотехнике.

Нам особо хотелось бы подчеркнуть, что начиная с 40-х годов ведущая роль в разработке математических методов теории переноса излучения перешла к атомной технике, где теория переноса столкнулась не только с принципиально новыми задачами, но и с новыми физическими и математическими проблемами.

Необходимо отметить также, что именно в связи с проблемами атомной физики были разработаны и мощные математические методы

решения задач теории переноса, в частности машинные методы.

Мы не имеем возможности перечислить в докладе различные применения теории переноса излучения — настолько они разнообразны и специальны. В дальнейшем мы остановимся только на вычислительных методах в теории переноса.

Использование в расчетах мощных цифровых вычислительных машин позволило поставить и решить новые задачи науки и техники, решение которых казалось раньше практически неосуществимым. В значительной степени это относится именно к задачам теории переноса излучения.

1. Весьма общая математическая формулировка нестационарных задач теории переноса излучения дается с помощью линейного интегро-дифференциального уравнения Больцмана с соответствующими краевыми условиями и начальными данными:

$$\frac{1}{v}\frac{\partial\varphi}{\partial t}+\Omega\,\mathrm{grad}\,\varphi+\Sigma\,\varphi-\int dv'\int d\Omega'\varphi\cdot w(\Omega,\ \Omega';\ v,\ v')=f, \qquad (1)$$

$$\varphi=g \text{ при } \mathbf{r}\in S \text{ и } (\Omega,\ \mathbf{n})<0, \qquad (2)$$

$$\varphi(\mathbf{r},\ \Omega,\ v,\ 0)=\varphi^0, \qquad (3)$$

где $\varphi(\mathbf{r}, \Omega, v, t)$ — поток частиц в точке с радиусом-вектором $\mathbf{r}$, имеющих скорость $\mathbf{v} = v\Omega$, Ω — единичный вектор направления полета частицы, Σ — сечение взаимодействия частицы с веществом. Функция w является дифференциальным сечением процесса взаимодействия частицы с веществом, f — функция, описывающая мощность источников излучения, $\mathbf{n}$ — внешняя нормаль к поверхности S, g — заданная функция, характеризующая поток частиц, проникающих в область D через поверхность S, φ^0 — заданное начальное распределение частиц.

Обычно решение сформулированной задачи ищется в вещественном гильбертовом пространстве. Классы функций и области определения решений в различных задачах формулируются в соответствии с физическим и математическим содержанием (Дэвисон, Владимиров, Биркгоф и Варга, Марек, Кейс и др.). Теория переноса, вообще говоря, имеет дело с решением интегро-дифференциального уравнения, которое является функцией шести независимых переменных.

Важное значение в теории переноса имеют сопряженные в смысле Лагранжа задачи

$$-\frac{1}{v}\frac{\partial\varphi^*}{dt}-\Omega\,\mathrm{grad}\,\varphi^*+\Sigma\,\varphi^*-\int dv'\int d\Omega'\varphi^*\cdot w(\Omega',\ \Omega;\ v',\ v)=p, \qquad (1')$$

$$\varphi^*=g \text{ при } \mathbf{r}\in S \text{ и } (\Omega,\ \mathbf{n})>0, \qquad (2')$$

$$\varphi^*(\mathbf{r},\ \Omega,\ v,\ 0)=\varphi_0^*, \qquad (3')$$

где $\varphi^*(\mathbf{r}, \boldsymbol{\Omega}, v, t)$ — сопряженная функция, а p, φ_0^*, g — некоторые заданные функции.

Сопряжённые уравнения первоначально рассматривались для однородных задач теории ядерных реакторов и были успешно применены к теории возмущений (Вигнер, Фукс, Усачев и др.). В дальнейшем теория сопряженных уравнений была разработана для неоднородных уравнений. В этом случае объектом исследования явились линейные функционалы от решений задач, которые интерпретировались как результаты измерения приборов, регистрирующих излучение. Введение сопряженных уравнений позволило сформулировать теорию возмущений применительно к исследованию функционалов задач (Каломцев, Орлов, Марчук и др.). Следует подчеркнуть, что развитая в связи с решением задач ядерной физики теория возмущений в настоящее время используется в общей теории линейных измерений физики и техники.

Аппарат сопряженных функций во многих случаях позволяет провести эффективную замену задачи с непрерывным аргументом v — скоростью частиц — многогрупповой системой, каждое из уравнений которой в пределах фиксированного интервала скоростей не зависит от этой переменной. Таким образом, переход к многогрупповым задачам в конечном итоге является математической идеализацией, с помощью которой осуществляется переход к более простым односкоростным задачам. Указанная редукция имеет большое значение для численного решения задач теории переноса. Теоретическое обоснование метода замены задач с непрерывным аргументом многогрупповыми системами по-прежнему остается важной и в значительной мере нерешенной проблемой теории интерполяции в пространстве операторов.

Учитывая это обстоятельство, в дальнейшем мы более подробно остановимся именно на вычислительных методах решения односкоростных задач теории переноса. Мы будем рассматривать только стационарные проблемы, поскольку именно они представляют наибольший интерес в приложениях.

2. Активный прогресс атомной науки и техники стимулировал развитие простых приближенных методов решения уравнений переноса. К настоящему времени наиболее продвинутой в алгоритмическом и теоретическом аспектах является диффузионная теория, основы которой были заложены еще в 40-х годах. Как известно, сущность диффузионного приближения состоит в том, что решение уравнений переноса ищется в форме линейной функции относительно угловой переменной $\boldsymbol{\Omega}$:

$$\varphi(\mathbf{r}, \boldsymbol{\Omega}, v) = \frac{1}{4\pi}[\varphi_0(\mathbf{r}, v) + 3\boldsymbol{\Omega}\boldsymbol{\varphi}_1(\mathbf{r}, \mathbf{v})],$$

где φ_0 — полный поток частиц через единичную сферу с центром в точке $\mathbf{r}$, а $\boldsymbol{\varphi}_1$ — векторный ток частиц в точке $\mathbf{r}$.

В этом случае односкоростная задача теории переноса приводится к следующей:

$$\operatorname{div}\boldsymbol{\varphi}_1 + \Sigma_C\varphi_0 = f_0,$$

$$\frac{1}{3}\operatorname{grad}\varphi_0 + \Sigma\boldsymbol{\varphi}_1 = \mathbf{f}_1,$$

$$2(\boldsymbol{\varphi}_1, \mathbf{n}) - \varphi_0 = g_0 \quad \text{при } \mathbf{r} \in S,$$

где f_0, $\mathbf{f}_1$ — первые два члена разложения функции f в ряд по сферическим функциям, а $g_0 = \int \boldsymbol{\Omega} g d\Omega$, причем интегрирование проводится по $\boldsymbol{\Omega}$, удовлетворяющим условию $(\boldsymbol{\Omega}, \mathbf{n}) < 0$.

Если $\mathbf{f}_1$ равно нулю, система приводится к уравнению

$$-\operatorname{div} D\operatorname{grad}\varphi_0 + \Sigma_C\varphi_0 = f_0,$$

где D и Σ_C выражаются через элементарные константы и функционалы от ω.

Таким образом, в этом простейшем случае задача свелась к хорошо изученному классу проблем математической физики. Численные методы решения этого уравнения к настоящему времени развиты достаточно хорошо.

В общем случае, когда изучается многоскоростная диффузионная проблема, приходим к задаче следующего вида:

$$A\varphi_0 = B\varphi_0 + f_0,$$

где оператор A описывает диффузию и поглощение, а B — рассеяние частиц в процессе переноса.

Аналогичная задача на собственные числа, связанная с расчетом критического режима ядерных реакторов, имеет вид

$$A\varphi_0 = B\varphi_0 + \frac{1}{\lambda}C\varphi_0,$$

где λ — формальный параметр задачи и оператор C описывает размножение частиц. Существенные результаты по теоретическому обоснованию методов решения задач теории переноса в диффузионном приближении были получены Варга и Биркгофом на основе теории неотрицательных матриц Перрона — Фробениуса.

Однако диффузионное приближение далеко не во всех случаях приводило к удовлетворительному решению задачи, поэтому разработка более точных методов решения уравнений переноса уже давно являлась важной задачей вычислительной математики.

С самого начала казалось естественным искать решение в виде ряда Фурье по сферическим функциям, учитывая более высокие их порядки, чем в диффузионном приближении. На этом пути были получены существенные результаты по крайней мере для задач теории переноса в одномерных геометриях — плоской, сферической и цилиндрической.

Для иллюстрации ограничимся рассмотрением этого метода, называемого методом сферических гармоник, для случая плоской геометрии, когда поток частиц зависит только от одной геометрической координаты x и угловой переменной $\boldsymbol{\Omega} = \Omega_x \mathbf{i}$, где $\mathbf{i}$ — орт вдоль оси x, а $\Omega_x = \mu$. В этом случае односкоростное уравнение переноса имеет вид

$$\mu \frac{\partial \varphi}{\partial x} + \Sigma \varphi = \frac{1}{2} \int_{-1}^{1} \varphi(x, \mu') \varpi(\mu, \mu') d\mu' + f(x, \mu).$$

В соответствии с идеей метода сферических гармоник решение уравнения ищется в форме ряда Фурье

$$\varphi(x, \mu) = \sum_{l=0}^{\infty} \frac{2l+1}{2} \varphi_l(x) P_l(\mu),$$

где $P_l(\mu)$ — полиномы Лежандра. В результате приходим к системе обыкновенных дифференциальных уравнений для коэффициентов Фурье

$$l \frac{d\varphi_{l-1}}{dx} + (l+1) \frac{d\varphi_{l+1}}{dx} + (2l+1) \Sigma_l \varphi_l = (2l+1) f_l \qquad (l = 0, 1, 2, \ldots),$$

а Σ_l и f_l находятся через коэффициенты Фурье функций $\varpi(\mu\mu')$ и $f(x, \mu)$.

Предположим, что на внешней границе поток излучения извне отсутствует. В этом случае граничные условия формулируются в виде

$$\int_{-1}^{0} \mu^{2i+1} \varphi \, d\mu = 0 \text{ при } \mathbf{r} \in S \qquad (i = 0, 1, 2, \ldots).$$

Эти условия, введенные в рассмотрение Маршаком, в дальнейшем получили обоснование в работах Владимирова на основе специального вариационного принципа. Если ряд Фурье для φ подставить в условие Маршака, то приходим к граничным условиям

$$\sum_{l=0}^{\infty} a_{il} \varphi_l = 0 \text{ при } \mathbf{r} \in S \quad (i = 0, 1, 2, \ldots),$$

где a_{il} — заданные числа.

Решение уравнений методом сферических гармоник проводится численным путем, который изложен ниже (Ляшенко, Кочергин, Марчук и др.). Теория метода сферических гармоник достаточно полно изучена в работах Дэвисона, Маршака, Вика, Марка, Владимирова, Николайшвили и др.

Введем векторы

$$\boldsymbol{\varphi} = \begin{vmatrix} \varphi_0 \\ \varphi_2 \\ \ldots \end{vmatrix}, \quad \mathbf{j} = \begin{vmatrix} \varphi_1 \\ \varphi_3 \\ \ldots \end{vmatrix}, \quad \mathbf{f} = \begin{vmatrix} f_0 \\ f_2 \\ \ldots \end{vmatrix}, \quad \mathbf{g} = \begin{vmatrix} f_1 \\ f_3 \\ \ldots \end{vmatrix}$$

и рассмотрим отдельно уравнения для четных и нечетных индексов. Тогда приходим к следующей системе векторно-матричных уравнений:

$$\xi \frac{d\mathbf{j}}{dx} + \mathbf{a}\boldsymbol{\varphi} = \mathbf{f},$$

$$\eta \frac{d\boldsymbol{\varphi}}{dx} + \mathbf{b}\mathbf{j} = \mathbf{g},$$

где ξ, η, $\mathbf{a}$ и $\mathbf{b}$ — известные матрицы, определяемые коэффициентами исходной системы уравнений. Граничное условие может быть записано в виде

$$\mathbf{A}\boldsymbol{\varphi} + \mathbf{B}\mathbf{j} = 0 \text{ при } \mathbf{r} \in S.$$

Дальнейшая задача состоит в решении полученных векторно-матричных уравнений с соответствующими граничными условиями. Разумеется, для этой цели удобнее всего воспользоваться конечно-разностным методом. Соответствующие разностные уравнения ищутся в классе кусочно непрерывных коэффициентов $\mathbf{a}$ и $\mathbf{b}$ и функций $\mathbf{f}$ и $\mathbf{g}$. При достаточной гладкости функций $\boldsymbol{\varphi}$ и $\mathbf{j}$ методом, аналогичным методу Тихонова и Самарского, можно прийти к формальной трехточечной векторно-матричной системе уравнений

$$\mathbf{A}_k \boldsymbol{\varphi}_{k+1} - \mathbf{B}_k \boldsymbol{\varphi}_k + \mathbf{C}_k \boldsymbol{\varphi}_{k-1} = -\mathbf{F}_k,$$

где $\boldsymbol{\varphi}_k$ — значение вектор-функции $\boldsymbol{\varphi}$ в точке $x = x_k$, коэффициенты $\mathbf{A}_k$, $\mathbf{B}_k$ и $\mathbf{C}_k$ — специальным образом определенные матрицы, а $\mathbf{F}_k$ — известный вектор.

Для решения полученных уравнений вместе с граничными условиями применяется метод матричной факторизации, развитый Гельфандом и Локуциевским, а также Бабенко, Ченцовым и Русановым.

Метод матричной факторизации в применении к решению уравнений сферических гармоник в плоской, сферической и цилиндрических геометриях оказался весьма эффективным численным методом решения широкого класса задач теории переноса излучения.

Однако попытки обобщения метода сферических гармоник на решение неодномерных проблем до последнего времени не приводили к успеху.

Можно отметить, что принципы построения численных алгоритмов решения уравнений сферических гармоник оказались весьма общими и позволили в дальнейшем сформулировать ряд новых численных алгоритмов решения задач методом дискретных ординат, развитым Виком и Чандрасекаром, и другими методами (Мертенс, Курганов, Гермогенова, Романова и др.).

Интерес представляет другой численный метод решения уравнений сферических гармоник, основанный на ортогонализации собственных векторов задачи и разработанный Годуновым. Указанный метод состоит в решении краевой задачи для системы уравнений сферических гармоник с помощью задач с начальными данными. Устойчивый алгоритм счета достигается применением специального метода ортогонализации ошибок, возникающих в процессе счета.

3. Одновременно с развитием численных методов решения уравнений сферических гармоник имел место интенсивный поиск новых путей решения уравнений переноса. В этом аспекте внимание исследователей было привлечено к различным итерационным методам. Такие методы были предложены уже в 50-х годах. Это метод характеристик, разработанный Владимировым и Нейманом, а также S_n-метод, разработанный Карлсоном. Эти методы развивались в связи с потребностью решения задач теории переноса в нейтронной физике для областей сферической и цилиндрической геометрий. Общим в этих методах является то, что в обоих методах решения задачи представляются в виде ряда Неймана на основе реализации следующего итерационного процесса:

$$\mathbf{\Omega}\,\mathrm{grad}\,\varphi^{(n)}+\Sigma\,\varphi^{(n)}=\int\varphi^{(n-1)}(\mathbf{r},\ \mathbf{\Omega}')\,w(\mathbf{\Omega},\ \mathbf{\Omega}')\,d\mathbf{\Omega}'+f(\mathbf{r},\ \mathbf{\Omega})$$

или в формальном виде

$$L\varphi^{(n)}=S\varphi^{(n)}+f,$$

где n — номер последовательного приближения. Различие методов состоит в способах реализации обратного оператора L^{-1}. В методе характеристик при реализации обратного оператора уравнение приводится к характеристическому виду, а в S_n-методе используется для этой цели специальным образом определенный метод дискретных ординат. Оба указанных метода, в особенности S_n-метод, были обобщены на решение широкого класса одномерных задач. В последние годы S_n-метод был применен к решению многомерных задач теории переноса (Карлсон, Белл, Вендроф, Гельбард и др.). Различные аспекты применения S_n-метода и его обоснования изложены в ряде докладов на второй и третьей Женевских

конференциях по использованию атомной энергии и в настоящее время хорошо известны.

Несмотря на известные успехи в построении вычислительных алгоритмов для задач теории переноса, проблема новых численных алгоритмов для решения многомерных задач по-прежнему требовала большого внимания со стороны исследователей. Научный поиск, продолжавшийся около десяти лет, в настоящее время завершился созданием ряда новых численных методов, к изложению которых мы и переходим.

4. Первый из таких методов основан на идее расщепления сложных интегро-дифференциальных операторов переноса на дифференциальный и интегральный. После известных работ Писмана, Рэкфорда и Дугласа, сформулировавших метод переменных направлений, появилась возможность формулировки универсальных и общих алгоритмов (принципы построения которых были описаны Фаддеевыми), охватывающих различные схемы реализации. В Советском Союзе был предпринят широкий научный поиск в этом направлении, начиная с работы Яненко, в которой был предложен метод расщепления некоторых классов нестационарных задач. Методы расщепления в дальнейшем были развиты в работах Яненко, Самарского, Дьяконова и др. В результате были сформулированы итерационные методы решения стационарных задач и схемы расщепления нестационарных задач математической физики. Сущность этих методов состоит в следующем.

Требуется найти решение уравнения

$$\Lambda\varphi=f,$$

где Λ — линейный оператор, а f — заданная функция. Сформулируем итерационный процесс

$$\prod_{\alpha=1}^{n}\left(E+\frac{\tau_\alpha}{2}A_\alpha\right)\frac{\varphi^{j+1}-\varphi^j}{\tau}+\Lambda\varphi^j=f.$$

Здесь E — единичный оператор.

Операторы A_α, вообще говоря, произвольные, так же как и $n+1$ параметров τ_α и τ. Операторы A_α и параметры τ_α, τ выбираются из условия простоты реализации алгоритма и быстроты сходимости итерационного процесса. В частности, если

$$\Lambda=\sum_{\alpha=1}^{n}\Lambda_\alpha,$$

где Λ_α — элементарные операторы, из которых состоит Λ, то мы приходим к простейшей формулировке метода. Схема реализации

рассматриваемого алгоритма имеет простой вид:

$$\left(E+\frac{\tau_1}{2}A_1\right)\varphi^{j+1/n}=-F^j,$$

$$\left(E+\frac{\tau_2}{2}A_2\right)\psi^{j+2/n}=\psi^{j+1/n},$$

$$\cdots\cdots\cdots\cdots\cdots$$

$$\left(E+\frac{\tau_n}{2}A_n\right)\psi^{j+1}=\psi^{j+\frac{n-1}{n}},$$

$$\varphi^{j+1}=\varphi^j+\tau\psi^{j+1},$$

где

$$F^j=\Lambda\varphi^j-f$$

—невязка релаксационного процесса. Предполагается, что операторы A_α допускают простую реализацию полученной системы уравнений. Таким образом, основная задача приводится к реализации ряда задач более простой структуры.

Для задач нестационарных

$$\frac{\partial\varphi}{\partial t}+\Lambda\varphi=f$$

предполагается следующая схема второго порядка аппроксимации относительно $\Delta t=\tau$, по форме аналогичная рассмотренной выше итерационной схеме:

$$\prod_{\alpha=1}^{n}\left(E+\frac{\tau}{2}A_\alpha\right)\frac{\varphi^{j+1}-\varphi^j}{\tau}+\Lambda\varphi^j=f^{j+1/2},$$

где операторы A_α удовлетворяют условию

$$\Lambda=\sum_{\alpha=1}^{n}A_\alpha+O(\tau).$$

В этом случае на гладких решениях схема имеет второй порядок точности по τ. Операторы A_α выбираются из условия простоты реализации алгоритма и требования счетной устойчивости численного алгоритма. Таким образом, принципиальная схема реализации для решения разностного аналога нестационарного уравнения будет та же, что и для решения стационарных задач.

Эти соображения были положены в основу решения задач теории переноса излучения. В работе Яненко и автора, а также Пененко, Султангазина и др. дано дальнейшее развитие и теоретическое обоснование метода расщепления для решения задач теории переноса. Идея метода может быть проиллюстрирована на простейшем примере односкоростной задачи в плоско-параллельной геометрии.

Оператор интегро-дифференциального уравнения

$$\Lambda\varphi=f$$

представляется в виде суммы двух операторов — дифференциального

$$\Lambda_1=\mu\frac{\partial}{\partial x}+\Sigma_C$$

и интегрального

$$\Lambda_2=\Sigma_s-\int_{-1}^{1}w(\mu,\ \mu')\,d\mu'.$$

Можно показать, что оператор Λ_1 положителен, а Λ_2 положительно полуопределен. Для решения задачи формулируется итерационный процесс

$$\left(E+\frac{\tau_1}{2}\Lambda_2\right)\left(E+\frac{\tau_2}{2}\Lambda_2\right)\frac{\varphi^{j+1}-\varphi}{\tau}+\Lambda\varphi^j=f.$$

Указанный итерационный процесс имеет три произвольных параметра τ_1, τ_2 и τ, которые выбираются так, чтобы процесс сходился быстро. Доказательство сходимости итерационного процесса проводится с использованием известной техники доказательства (Келлог, Дуглас и Пирс, Самарский).

Проблема выбора оптимальных параметров является весьма сложной. Частное решение этой проблемы было дано Пененко, который разработал алгоритм выбора параметров на основе анализа диффузионного приближения задачи. Оценка сходимости процесса в асимптотических случаях показывает, что спектральная норма итерируемого оператора оказывается величиной, малой по сравнению с 1.

Главное применение метод расщепления имеет в связи с решением многомерных задач теории переноса. В работах Султангазина, Пененко и других теоретически установлена возможность расщепления оператора на дифференциальную и интегральную составляющие в общем случае конечных связных областей в трехмерном евклидовом пространстве и доказана сходимость итерационного процесса в метрике пространства L_2. Расщепление оператора на еще более простые одномерные дифференциальный и интегральный операторы является очередной проблемой этого направления.

Вторым направлением работ, связанных с методом расщепления в теории переноса, является проблема решения уравнений методом сферических гармоник для многомерных областей. Даже в простейшем случае двумерных цилиндрических областей с круговой симметрией метод сферических гармоник приводит к сложной систе-

ме дифференциальных уравнений в частных производных, порядок которой определяется номером приближения. Решение этой системы связано с большими математическими трудностями. Лишь в последнее время Бояринцеву и Узнадзе, развивая метод расщепления, удалось представить матричный оператор системы уравнений методом сферических гармоник в виде суммы двух операторов специального вида и для решения системы применить итерационный процесс в форме универсального алгоритма. Эта работа, по-видимому, открывает ряд новых возможностей в использовании метода сферических гармоник для решения многомерных задач теории переноса.

5. Оригинальный подход к решению уравнений переноса с помощью итерационных методов предложен в работе Морозова. Сущность метода состоит в том, что метод простых итераций

$$\mathbf{\Omega}\,\mathrm{grad}\,\varphi^{(n)}+\Sigma\varphi^{(n)}=\int d\mathbf{\Omega}'\varphi^{(n-1)}w(\mathbf{\Omega},\ \mathbf{\Omega}')+f(\mathbf{r},\ \mathbf{\Omega}),$$

схема реализации которого возможна как по методу характеристик, так и по S_n-методу, дополняется вторым этапом: уточнением решения с помощью решения приближенного уравнения для погрешности

$$\mathbf{\Omega}\,\mathrm{grad}\,\varepsilon^{(n)}+\Sigma_C\varepsilon^{(n)}=\int d\mathbf{\Omega}'\,(\varphi_0^n-\varphi_0^{n-1})\,w(\mathbf{\Omega},\ \mathbf{\Omega}'),$$

$$\varepsilon^{(n)}(\mathbf{r},\ \mathbf{\Omega})=0\ \text{при}\ \mathbf{r}\in S,\ (\mathbf{\Omega},\ \mathbf{n})<0.$$

Такие алгоритмы в дальнейшем будем условно называть двухшаговыми.

В работах Лебедева сформулирован более общий метод решения уравнения переноса на этом пути (КР-метод). Основываясь на простой итерации, как на первом этапе решения, Лебедев построил специальные уравнения для погрешности на основе теории аппроксимации оператора. Им была восстановлена общая структура оператора уравнения для погрешности

$$Q\varepsilon^{(n)}=P\left[\varepsilon^{(n)}-\int d\mathbf{\Omega}'\,(\varphi^{(n)}-\varphi^{(n-1)})\,w(\mathbf{\Omega},\ \mathbf{\Omega}')\right],$$

где Q и P — некоторые операторы полиномиальной структуры. В случае изотропного рассеяния эти операторы действуют на функции только от $\mathbf{r}$. Показано, что полученная итерационная схема является в известном смысле оптимальной и во многих случаях просто реализуется на ЭВМ. Этот метод может быть применен для решения многомерных задач.

Интересный метод решения уравнений переноса в рамках двухшаговой схемы сформулировал Гольдин. Он также использует на первом шаге простую итерацию, а на втором шаге осуществляет

уточнение решения на основе решения системы уравнений

$$\mathrm{div}\ \boldsymbol{\varphi}_1+\Sigma\varphi_0-\int d\Omega\,\varphi\,w(\mathbf{\Omega},\ \mathbf{\Omega}')=f_0,$$

$$\nabla D^{(n)}\boldsymbol{\varphi}_0+\Sigma\boldsymbol{\varphi}_1-\int d\Omega\,\mathbf{\Omega}\,\varphi\,w(\mathbf{\Omega},\ \mathbf{\Omega}')=\mathbf{f}_1,$$

где $D^{(n)}$ — симметричный тензор с компонентами

$$D_{ik}^{(n)}=\frac{\int\Omega_i\Omega_j\varphi^{(n)}\,d\Omega}{\int\varphi^{(n)}\,d\Omega}$$

при соответствующих граничных условиях.

Теоретическое обоснование этого метода пока отсутствует, однако метод Гольдина представляет большой интерес, поскольку он связан с реализацией нелинейного итерационного процесса.

6. К вычислительным методам теории переноса тесно примыкают аналитические методы решения уравнений Больцмана, основанные на разложении решения в ряд по собственным функциям оператора переноса. Уже давно было замечено, что оператор переноса, кроме дискретного спектра, имеет спектр непрерывный (Вигнер, Лефар и Милот и др.). Далее Кейсу удалось построить полную систему собственных функций оператора переноса и сформулировать теорему о разложимости для некоторого класса функций. С помощью этого метода Кейс решил проблему Милна и задачу об альбедо для полупространства и построил в явном виде функции Грина для бесконечной и полубесконечной среды в случае односкоростной модели переноса при изотропном рассеянии частиц. Эта работа явилась началом большого цикла исследований, которые привели к серии интересных в теоретическом и практическом аспектах результатов. Была исследована полнота системы собственных функций и получены соотношения ортогональности для различных моделей, дано распространение результатов на случай ограниченных плоско-параллельных сред, анизотропного рассеяния частиц, энергетической зависимости и т. д. (Мика, Желязны, Кужель, Митсис, Фердзигер, Леонард, Мак Кормик, Кучер и Зоммерфельд, Инерней и др.). Вычислительные основы метода рассмотрены в работах Митсиса и Ковальской. В последнее время Лалетин дал обобщение метода Кейса на случай задач в сферической и цилиндрической геометриях. Для n-мерного пространства и самосопряженной формулировки задачи теории переноса метод Кейса развит Лебедевым.

7. Оригинальный подход к решению уравнения переноса излучения, основанный на принципах инвариантности, был предложен Амбарцумяном и развит во многих направлениях (Чандрасекар,

Беллман и Калаба, Соболев, Седельников и др.). Этот метод позволяет свести уравнение переноса излучения к решению нелинейных и линейных уравнений простой конструкции. Этот метод нашел применение в задачах астрофизики и нейтронной физики.

8. Существенный интерес для теории переноса имеют вариационные методы решения задач. Впервые примененные к решению уравнения Пайерлса — интегральному аналогу уравнения переноса, они позволили получить ряд важных результатов как в нейтронной физике, так и в других научных направлениях (Дэвисон, Лё-Кен, Брудно и др.). Важный этап достигнут при формулировании вариационного принципа непосредственно для кинетического уравнения (Владимиров).

9. Большое развитие в последние годы получили статистические методы решения уравнений переноса — методы Монте-Карло. Благодаря использованию все более совершенной вычислительной техники методы Монте-Карло приобретают все большее значение в решении сложных задач теории переноса. К настоящему времени уже накоплен ценный опыт в применении методов Монте-Карло для решения задач математической физики вообще и теории переноса в особенности. Работы Улама и Неймана нашли дальнейшее развитие во многих направлениях науки, и к настоящему времени мы располагаем системой алгоритмов, эффективно реализуемых на машинах для широкого класса задач (Бергер, Фано, Спенсер, Гельфанд, Ченцов, Фролов, Михайлов, Золотухин, Гольдштейн, Ермаков и др.). Можно предположить, что уже в близком будущем методы Монте-Карло окажутся мощным математическим аппаратом для решения наиболее сложных задач теории переноса. Успехи, достигнутые в способе уменьшения дисперсии, а также новые подходы к моделированию случайных величин позволяют приблизиться к формированию универсальных и экономичных методов решения задач теории переноса.

10. Прогресс в развитии вычислительных методов теории переноса ставит перед математиками комплекс теоретических проблем, развитие которых стимулирует развитие новых вычислительных методов. К настоящему времени уже достигнуты существенные успехи в области математической теории переноса как для односкоростных задач (Дэвисон, Владимиров, Гермогенова, Кучер, Масленников, Ленер, Цинг, Йоргенс и др.), так и для задач с энергетической зависимостью (Марек, Хабетлер и Мартно, Биркгоф и Варга, Масленников, Шихов и др.). Особое значение имеет принцип максимума, сформулированный для уравнений переноса Гермогеновой.

11. Развитие вычислительных методов теории переноса и практическое их осуществление на цифровых вычислительных машинах всегда были тесно связаны с уровнем вычислительной математики. Это определялось тем, что решение практических задач, как правило, было связано с проблемами развития новой техники. Вместе с тем нельзя не отметить, что сама вычислительная математика зачастую обогащалась методами и идеями, возникавшими при разработке проблем теории переноса излучения. К числу таких методов можно отнести численные методы решения уравнений диффузии, различные методы факторизации, теорию решения больших алгебраических систем, ускорение сходимости итерационных процессов и др. На основе разработанных вычислительных алгоритмов решение задач теории переноса в настоящее время составлен большой комплекс программ для цифровых вычислительных машин, которые используются в практических расчетах во многих областях науки и техники.

12. В заключение остановимся на некоторых актуальных проблемах, решение которых крайне важно для дальнейшего развития теории переноса и ее приложений.

Наиболее важной задачей, по нашему мнению, является разработка численных методов решения многомерных кинетических уравнений переноса для областей сложной геометрии. Существующие методы решения таких задач указывают лишь направление научного поиска. Можно предположить, что развитие новых алгоритмов решения многомерных задач и теоретическое их обоснование явится основным направлением в развитии вычислительных методов в теории переноса на ближайшие годы.

Большое значение в задачах теории переноса уже сейчас имеет метод Монте-Карло. В связи с совершенствованием цифровых вычислительных машин роль этого метода, по-видимому, будет непрерывно возрастать. Можно ожидать, что на первом этапе метод Монте-Карло будет комбинироваться с разностными методами в тех частях алгоритмов, где их применение окажется наиболее целесообразным. Это позволит уменьшить дисперсию, а также количество информации, перерабатываемой машиной в процессе решения задач. Разумеется, пути широкого использования метода Монте-Карло для решения задач теории переноса могут быть самыми различными. Однако несомненно, что вычислительная техника и вычислительная математика в настоящее время созрели для широкого применения этих методов для решения задач науки и техники в рассматриваемой области.

В последние годы сформулировалось новое научное направление в теории переноса излучения, связанное с решением обратных задач. Решение обратных задач нейтронной физики, атмосферной

оптики и т. д. давно привлекает внимание исследователей. Такие задачи ставились и решались в основном в связи с экспериментальным изучением сечений взаимодействия частиц с веществом, когда по эффекту ослабления потока частиц необходимо определить элементарные константы физического процесса. Решение таких задач, как правило, опиралось на теорию возмущений для однородных и неоднородных задач теории переноса. В настоящее время появился новый объект исследования, где обратные задачи теории переноса имеют большое значение. Речь идет об интерпретации данных с метеорологических спутников. Поле радиации земной атмосферы, регистрируемое приборами на метеорологических спутниках, существенно зависит от метеорологических условий в атмосфере (температуры, влажности, давления и т. д.), и задача состоит в восстановлении этих величин по измеряемым на спутниках характеристикам поля излучения. В связи с постановкой и решением таких задач возникает необходимость в изучении сопряженных уравнений по отношению к функционалам задач, которыми являются показания приборов, регистрирующих поле уходящей из атмосферы радиации, а также в развитии теории возмущения по отношению к указанным функционалам задач.

Большое внимание исследователей в настоящее время привлекают экстремальные задачи теории переноса, связанные с определением конструкций систем, для которых реализуется минимум существенных функционалов задачи. Такие проблемы, за редким исключением, изучены еще очень слабо, хотя отдельные интересные результаты в этом направлении уже получены в ряде областей и в особенности в атомной физике.

Методы решения линейных уравнений переноса к настоящему времени развиты достаточно хорошо. Это обстоятельство позволяет надеяться, что эти методы могут проникать в более сложные области нелинейных кинетических уравнений, интенсивно развиваемые в теории разреженных газов, теории плазмы и других. По-видимому, переход к решению нелинейных задач явится естественным этапом в развитии теории переноса.

В кратком докладе не представляется возможным остановиться на всех аспектах вычислительных методов в теории переноса. Мы отметили лишь некоторые тенденции, которые оказались в сфере внимания автора доклада, и потому естественно, что в настоящем докладе несколько большее внимание было уделено исследованиям, проводимым в Советском Союзе.

Вычислительный центр Сибирского отделения АН СССР, Новосибирск, СССР

ЛИТЕРАТУРА

[1] Амбарцумян В. А., Рассеяние и поглощение света в планетных атмосферах, *Уч. зап. ЛГУ*, **82** (1941).

[2] Амбарцумян В. А., Мустель Э. Р., Северный А. Б., Соболев В. В., Теоретическая астрофизика, гл. 8, Гостехиздат, 1952.

[3] Bednarz R. J., Mika J. R., Energy dependent Boltzmann equation in plane geometry, *J. Math. and Phys.*, **4**, № 10 (1963).

[4] Bellman R., Kalaba R., Prestrud M., Invariant imbedding and radiative transfer in slabs of definite thickness, New York, 1963.

[5] Birchoff Y., Varga R. S., Reactor criticality and non-negative matrices, *J. Soc. Indust. and Appl. Math.*, **6**, № 4 (1958).

[6] Weinberg A. M., Wigner E. P., The physical theory of neutron chain reactors, Chicago, 1958.

[7] Вигнер Е., Математические проблемы теории ядерных реакторов, сб. «Теория ядерных реакторов», М., Госатомиздат, 1963, 103-119.

[8] Варга Р., Численные методы решения многомерных многогрупповых дифференциальных уравнений, сб. «Теория ядерных реакторов», Госатомиздат, 1963.

[9] Varga R. S., Matrix iterative analysis, New Jersey, 1963.

[10] Владимиров В. С., Численное решение кинетического уравнения для сферы, *Вычислительная математика*, **3** (1958).

[11] Владимиров В. С., Математические задачи односкоростной теории переноса частиц, Труды математического института им. В. А. Стеклова АН СССР, XI, 1961.

[12] Владимиров В. С., О некоторых вариационных методах приближенного решения уравнения переноса, *Вычислительная математика*, **7** (1961).

[13] Гибетлер Т. И., Мартино М. А., Теоремы существования и теория спектров для многогрупповой диффузионной модели, сб. «Теория ядерных реакторов», М., Госатомиздат 1963, 145-159.

[14] Gelbard E., An iterative method for solving the Pe equations in slab geometry, *Nucl. Sci. and Engng*, **3**, № 4 (1958).

[15] Гельфанд И. М., Локуциевский О. В., Метод «прогонки» для решения разностных уравнений в книге С. К. Годунова и В. С. Рябенького «Введение в теорию разностных схем», М., Физматгиз, 1962.

[16] Гельфанд И. М., Фейнберг С. М., Фролов А. С., Ченцов Н. Н., О применении метода случайных испытаний (метода Монте-Карло) для решений кинетического уравнения, Докл. № 2141, Труды Второй международной конференции по мирному использованию атомной энергии, Докл. советских ученых, т. 2, Атомиздат, 1959.

[17] Гермогенова Т. А., Экстраполированная длина и плотность вблизи границы в сферической задаче Милна, сб. «Некоторые математические задачи нейтронной физики», изд-во МГУ, 1960, 80-119.

[18] Гермогенова Т. А., Принцип максимума для уравнения переноса, *Журнал вычислительной математики и математической физики*, **1**, № 1 (1962).

[19] Годунов С. К., Метод ортогональной прогонки для решения систем разностных уравнений, *Журнал вычислительной математики и математической физики*, **2**, № 6 (1962).

[20] Гольдин В. Я., Характеристическая разностная схема для нестационарного кинетического уравнения, *Докл. АН СССР*, **133** (1960), 748-751.
[21] Гольдин В. Я., Квазидиффузионный метод решения кинетического уравнения, *Журнал вычислительной математики и математической физики*, **4**, № 6 (1964).
[22] Davison B., Neutron transport theory, Oxford, 1957.
[23] Davison B., Remark on the variational methods, *Phys. Rev.*, **71** (1947), 694.
[24] Douglas J., Rachford H. H., On the numerical solution of heat conduction problems in two and three space variables, *Trans. Amer. Math. Soc.*, **82** (1956), 421-439.
[25] Douglas J., Pearcy C., On convergence of alternating direction procedures in the presence of singular operators, *Numer. Math.*, **5** (1963).
[26] Дьяконов Е. Г., О некоторых разностных схемах для решения краевых задач, *Журнал вычислительной математики и математической физики*, **2**, № 1 (1962).
[27] Дьяконов Е. Г., О некоторых итерационных методах решения систем разностных уравнений, возникающих при решении методом сеток уравнений в частных производных эллиптического типа, сб. «Вычислительные методы и программирование», изд-во МГУ, 1965, 191-222.
[28] Ермаков С. М., Золотухин В. Г., сб. «Вопросы физики защиты реакторов», М., Атомиздат, 1962, стр. 171.
[29] Желязны Р., Метод разложения по собственным функциям в теории транспорта нейтронов, Третья международная конференция по использованию атомной энергии в мирных целях, Доклад А. (conf. 28) p/498.
[30] Zelazny R., Kuszell A., Two-group approach by neutron transport theory in plane geometry, *Ann. Phys.*, **16**, № 1 (1961).
[31] Jörgens K., An asymptotic expansion in the theory of neutron transport, *Communs Pure and Appl. Math.*, **11** (1958), 219-242.
[32] Case K. M., Elementary solution of the transport equations and their applications, *Ann. Phys.*, **9**, № 1 (1960).
[33] Кадомцев Б. Б., О функции влияния в теории переноса лучистой энергии, *Докл. АН СССР*, **113**, № 3 (1957).
[34] Kaplan S., Some new methods of flux syntesis, *Nucl. Sci. and Engng.*, **13** (1962), 22-38.
[35] Kuszell A., The critical problems for multilayer slab systems, *Acta Phys. Polon.*, **20** (1961), 567.
[36] Kuščer I., McCormick N. I., Summerfield G. C., Orthogonality of Case's eigenfunctions in one-speed transport theory, *Ann. Phys.*, **30**, № 3 (1964).
[37] Kuščer I., Milne's problem for anisotropic scattering, *J. Math. and Phys.*, **34** (1956), 256-266.
[38] Kowalska Kr., Critial calculation of the slab reactors, *Nucl. Sci. and Engng.*, **24**, № 3 (1966).
[39] Kellog R. B., Another alternating direction implicit method, *J. Soc. Indust. and Appl. Math.*, **11**, № 4 (1963).
[40] Карлсон Б., Белл Дж., Решение транспортного уравнения S_n-методом, Труды Второй международной конференции по мирному использованию атомной энергии, Избранные доклады иностранных ученых, т. 3.— Физика ядерных реакторов, М., Атомиздат, 1959.
[41] Лалетин Н. И., Элементарные решения уравнения переноса нейтронов, сб. «Физика ядерных реакторов», т. 1, 1966, 3-11.

[42] Lehner J., Wing G. M., On the spectrum of an unsymmetric operator arising in the transport theory of neutrons, *Communs Pure and Appl. Math.*, **8** (1955), 213-234.
[43] Лебедев В. И., О КР-методе ускорения сходимости итераций при решении кинетического уравнения, *Журнал вычислительной математики и математической физики*, **6**, № 2 (1966).
[44] Le Caine J., Application of a variational method to Milne's problem, *Phys. Rev.*, **72** (1947), 564.
[45] Ляшенко Е. И., Николайшвили Ш. С., Обзор численных методов и программ расчета малогабаритных реакторов, сб. «Физика ядерных реакторов», т. 1, 1966, 173-192.
[46] Марек И., Некоторые математические задачи теории ядерных реакторов на быстрых нейтронах, *Aplikace matematiky*, **8**, № 6 (1963).
[47] Марчук Г. И., Методы расчета ядерных реакторов, Госатомиздат, 1961.
[48] Марчук Г. И., Орлов В В., К теории сопряженных функций, сб. «Нейтронная физика», Госатомиздат, 1961.
[49] Марчук Г. И., Яненко Н. Н., Применение метода расщепления (дробных шагов) для решения задач математической физики, Доклад на конгрессе ИФИП, июль 1965, Нью-Йорк.
[50] Марчук Г. И., Яненко Н. Н., Решение многомерного кинетического уравнения методом расщепления, *Докл. АН СССР*, **157**, № 6 (1964).
[51] Марчук Г. И., Султангазин У. М., К обоснованию метода расщепления для уравнений переноса излучения, *Журнал вычислительной математики и математической физики*, **5**, № 4 (1965).
[52] Марчук Г. И., Пененко В. В., Султангазин У. М., О решении кинетического уравнения методом расщепления, Труды семинара по прикладной и вычислительной математике, Новосибирск, 1965.
[53] Marshak I. R. E., Theory of the slowing down of neutrons by elastic collision with atomic nuclei, *Rev. Mod. Phys.*, **19** (1947), 185-238.
[54] Mika J., The thermalization theory with a simple scattering kernel. *Nucl. Sci. and Engng.* **22**, № 2 (1965).
[55] Mitsis G. J., Transport solutions to the one-dimensional critical problem, *Nuclear Sci. and Engng.*, **17**, № 1 (1963).
[56] Михайлов Г. А., Расчеты критических систем методом Монте-Карло, *Журнал вычислительной математики и математической физики*, **6**, № 1 (1966).
[57] Морозов В. Н., К вопросу о решении кинетических уравнений с помощью S_n-метода, сб. «Теория и методы расчета реакторов», Госатомиздат, 1962.
[58] Масленников В. В., Проблема Милна с произвольной индикатрисой, *Докл. АН СССР*, **118**, № 2 (1958).
[59] Николайшвили Ш. С., Односкоростная задача об угловых распределениях нейтронов, сб. «Теория и методы расчета ядерных реакторов», Госатомиздат, М., 1962.
[60] Peaceman D. W., Rachford H. H., The numerical solution of parabolic and elliptic differential equations, *J. Soc. Indust. and Appl. Math.*, **3** (1955), 28-41.
[61] Пененко В. В., Об алгоритмах и системе программирования задач расчета двумерных ядерных реакторов и некоторых задач теории переноса, Диссертация, Новосибирск, 1965.

[62] Романова Л. М., Задача Милна для полупространства с анизотропным рассеянием и захватом нейтронов, сб. «Некоторые математические задачи нейтронной физики», изд-во МГУ, 1960, 8-27.
[63] Рихтмайер Р. Д., Разностные методы решения краевых задач, М., ИЛ, 1960.
[64] Рихтмайер Р. Д., Методы Монте-Карло, сб. «Теория ядерных реакторов», Госатомиздат, 1963.
[65] Русанов В. В., Об устойчивости метода матричной прогонки, *Вычислительная математика*, **6** (1960).
[66] Самарский А. А., Об одном экономичном разностном методе решения многомерного параболического уравнения в произвольной области, *Журнал вычислительной математики и матем. физики*, **2**, № 5 (1962).
[67] Самарский А. А., О разностных схемах для многомерных дифференциальных уравнений математической физики, *Aplikace matematiky*, **10**, № 2 (1965), 146-163.
[68] Саульев В. К., Интегрирование уравнений параболического типа методом сеток, М., Физматгиз, 1960.
[69] Сб. «Теория ядерных реакторов» под ред. Г. Биркгофа и Э. Вагнера, М., Госатомиздат, 1963.
[70] Соболев В. В., Новый метод в теории рассеяния света, *Астрономический журнал*, **28**, № 5 (1951).
[71] Тихонов А. Н., Самарский А. А., Об однородных разностных схемах, *Докл. АН СССР*, **122**, № 4 (1958).
[72] Усачёв Л. Н., Уравнение для ценности нейтронов, кинетика реакторов и теория возмущений, сб. «Реакторостроение и теория реакторов», изд-во АН СССР, 1955.
[73] Фаддеева Д. К., Фаддеев В. Н., Вычислительные методы линейной алгебры, М., Физматгиз, 1962.
[74] Fano U., Spencer L. V., Berger M. J., Penetration and diffusion of *X*-rays, Handbuch der Physik, Band XXXVIII/2, Neutronen und Vervandte Gammastrahlprobleme, Berlin — Göttingen — Heidelberg, 1959.
[75] Ferziger J. H., Leonard A., Energy-dependent neutron transport theory. I. Constant cross sections, *Ann. Phys.*, **22**, № 2 (1963).
[76] Ferziger J. H., Robinson A. H., A transport theoretic calculation of the disadvantage factor, *Nucl. Sci. and Engng.*, **21**, № 3 (1965).
[77] Fucks K., Perturbation theory in neutron multiplication problems, *Proc. Phys. Soc.*, **62** (1949), 791.
[78] Hammersley J. M., Handscomb D. C., Monte-Carlo methods, London, New York, 1964.
[79] Шихов С. Б., Шишков Л. К., Существование и единственность положительного решения однородного уравнения переноса нейтронов, сб. «Физика ядерных реакторов», т. 1, 1966, 50-55.
[80] Jacobs A. M., McInerney J. J., On the Green's function of monoenergetic neutron transport theory, *Nucl. Sci. and Engng.*, **22**, № 1 (1965).
[81] Яненко Н. Н., Об одном разностном методе счета многомерного уравнения теплопроводности, *Докл. АН СССР*, **125**, № 6 (1959).
[82] Яненко Н. Н., Метод дробных шагов решения многомерных задач математической физики, Новосибирск, 1966.
[83] Chandrasekhar S., Radiative transfer, Oxford, 1950. Русский перевод: Чандрасекар Ш., Перенос лучистой энергии, М., ИЛ, 1953.

ТЕОРИЯ ПРИБЛИЖЕНИЯ ИНТЕГРАЛОВ ФУНКЦИЙ МНОГИХ ПЕРЕМЕННЫХ

С. Л. СОБОЛЕВ

Интеграл от функции n переменных $\varphi(x)$ по области Ω

$$\int \mathcal{E}_\Omega(x)\,\varphi(x)\,dx$$

приближенно выражается в виде

$$\sum C_k \varphi(x^{(k)}) = \int \sum C_k \delta(x - x^{(k)})\,\varphi(x)\,dx.$$

Ошибка приближения

$$(l,\ \varphi) = \int \left(\mathcal{E}_\Omega(x) - \sum C_k \delta(x - x^{(k)})\,\varphi(x)\right) dx$$

является линейным функционалом над соответствующим линейным пространством функций φ.

При $m > n/2$ автор рассматривает функционалы (l, φ) из $L_2^{(m)*}$, определенные над пространством $L_2^{(m)}$ с нормой

$$\|\varphi\|_{L_2^{(m)}} = \left\{\int \sum_{|\alpha|=m} (D^\alpha \varphi)^2\,dx\right\}^{1/2},$$

инвариантной относительно ортогональных координатных преобразований. Условие, что $l \in L_2^{(m)*}$, означает, в частности,

$$(l,\ x^\alpha) = 0 \text{ при } |\alpha| < m.$$

В докладе рассматриваются функционалы с узлами в точках решетки Γ:

$$\Gamma = E(x : x = x^{(0)} + hH\gamma),$$

где H — матрица с определителем, равным единице, x — произвольный вектор, γ — произвольный целочисленный вектор, h — малый параметр. Исследуются оптимальные формулы, т. е. формулы с наименьшей нормой функционала погрешности. Основные результаты следующие.

1. Оценка погрешности

а) При интегрировании функций $\varphi(x)$, периодических с матрицей периодов ω, кратной матрице hH, где

$$\omega = hHK,\ K = \begin{pmatrix} k_1 & 0 & \dots & 0 \\ 0 & k_2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & & k_n \end{pmatrix},\ k_j \text{ — целые числа,}$$

оптимальными будут постоянные коэффициенты $C = h^n$.

Норма функционала погрешности в этом случае имеет вид

$$\|l\|_{L_2^{(m)*}} = \sqrt{\Omega}\left(\frac{h}{2\pi}\right)^m \sqrt{\zeta(H^{-1}\,|\,2m)}. \qquad (1)$$

Здесь

$$\zeta(H^{-1}\,|\,2m) = \sum_{\gamma \neq 0} \frac{1}{r_\gamma^{2m}},$$

где r_γ — расстояния от начала координат до переменной точки решетки $H\gamma$. Функция

$$\zeta(H^{-1}\,|\,2m)$$

называется функцией Римана для решетки H^{-1}, взаимной с решеткой H.

б) Для всех финитных функций $\varphi(x)$ с фиксированным носителем Ω имеет место неулучшаемая оценка

$$|(l, \varphi)| \leqslant \sqrt{\Omega}\left(\frac{h}{2\pi}\right)^m \sqrt{\zeta(H^{-1}\,|\,2m)}\,\|\varphi\|_{L_2^{(m)}} + O(h^{m+1}). \qquad (2)$$

в) Для произвольной области Ω с достаточно гладкой границей справедлива формула

$$\inf_C \|l\|_{L_2^{(m)*}} = \sqrt{\Omega}\left(\frac{h}{2\pi}\right)^m \sqrt{\zeta(H^{-1}\,|\,2m)} + O(h^{m+1}). \qquad (3)$$

2. Нахождение оптимальных коэффициентов

а) Найден класс формул, называемых формулами с регулярным пограничным слоем, для которых

$$\|l\|_{L_2^{(m)*}} = \sqrt{\Omega}\left(\frac{h}{2\pi}\right)^m \sqrt{\zeta(H^{-1}\,|\,2m)} + O(h^{m+1}).$$

Они аналогичны формулам Грегори, являющимся асимптотически оптимальными в случае одной независимой переменной.

В этих формулах все коэффициенты C_k, относящиеся к точкам, удаленным от границы больше чем на Lh, где L — некоторая постоянная, называемая толщиной слоя, равны h^n, а в остальных точках, принадлежащих к «пограничному слою», вычисляются с помощью определенного алгорифма. Их вычисление требует Kh^{-n+1} действий, что на порядок меньше числа действий, необходимых для подсчета самого интеграла по внутренности области.

б) Указан алгорифм с конечным числом действий, не зависящим от h, для нахождения коэффициентов регулярного пограничного слоя в случае, когда Ω есть многогранник с гранями, рациональными относительно решетки.

Исследование наилучших решеток приведено, таким образом, к задаче теории чисел: исследованию минимумов ζ-функции при различных m и n.

Асимптотически для больших m

$$\zeta(H^{-1}\,|\,2m) = \frac{K}{r_{\min}^{2m}},$$

где K—постоянная и, следовательно, наилучшими будут решетки, где H^{-1} — решетка наиплотнейшей упаковки. Иными словами, решетка H будет взаимной с решеткой наиплотнейшей упаковки.

Вопрос о возможности получения лучших формул при x^h, не образующих решетку, остается открытым.

3. Сходимость на классах Жеврея бесконечно дифференцируемых функций при переменной толщине пограничного слоя

Для периодических функций из классов Жеврея $\mathcal{G}(A, \beta)$ ($\beta \geqslant 1$; $A > 0$), удовлетворяющих соотношению

$$\left|\frac{D^\alpha \varphi}{\alpha!}\right| < KA^{|\alpha|}\,|\alpha|^{(\beta-1)|\alpha|}$$

при любом α, дана оценка погрешности формул с регулярным пограничным слоем

$$|(l, \varphi)| \leqslant Kh^{-1/2} \exp\left[-\frac{\beta}{e}\left(\frac{Ah}{2\pi e r_{\min}}\right)^{-1/\beta}\right].$$

Толщина слоя порядка $L \simeq \left(\frac{Ah}{2\pi e r_{\min}}\right)$. Мы видим, что на этих классах сходимость значительно лучше степенной.

Метод автора представляет собой дифференциальное исследование функционалов из $L_2^{(m)}$. Он состоит в представлении функционалов суммами слагаемых, имеющих малые носители.

Пусть $l_\gamma(y)$ такой функционал, что

$$\operatorname{supp} l_\gamma(y) \subset E(y: |y| < L), \quad \|l_\gamma(y)\|_{L_2^{(m)}} < A.$$

После изменения масштаба и переноса начала в точку $hH\gamma$ получим функционал

$$l_\gamma\left(\frac{x}{h} - H\gamma\right)$$

с малым носителем. Для него справедливы оценки

$$\operatorname{supp} l_\gamma\left(\frac{x}{h} - H\gamma\right) \subset E(x: |x - hH\gamma| < Lh);$$

$$\left\| l_\gamma\left(\frac{x}{h} - H\gamma\right)\right\|_{L_2^{(m)*}} \leqslant Ah^{2m+n}.$$

Оценка показывает быстрое убывание скалярных произведений двух функционалов по мере удаления их носителей друг от друга. Отсюда для каждого функционала вида

$$l_0(x) = \sum_{hH\gamma \in \Omega} l_\gamma\left(\frac{x}{h} - H\gamma\right) \tag{4}$$

получается оценка нормы

$$\|l_0(x)\|_{L_2^{(m)*}} \leqslant Kh^m \sqrt{\Omega},$$

где K зависит только от L и A, Ω — носитель функционала l. Невозможность улучшения этой оценки показана Н. С. Бахваловым.

Такое же представление в виде суммы функционалов с малыми носителями допускает периодический функционал

$$l(x) = 1 - \sum_\gamma h^n \delta(x - hH\gamma), \tag{5}$$

неограниченный в $L_2^{(m)}$, но определенный над всеми финитными элементами $L_2^{(m)}$. Доказывается оптимальность этого функционала для всех финитных функций. Вычисление его минимума с помощью преобразования Фурье дает формулу (1).

Отыскание оптимальных кубатурных формул равносильно отысканию проекции функционала $\mathscr{E}\Omega(x)$ на многообразие всех линейных комбинаций

$$\sum C_\gamma l(x - hH\gamma),$$

образующих линейное подпространство $\mathfrak{H} \subset L_2^{(m)}$. Доказывается эквивалентность $\mathfrak{H}$ и $l_2^{(m)}$, пространства коэффициентов C_γ.

Элементы $\mathfrak{H}$ с малым носителем аналогичны дифференциальным операторам в частных производных порядка m. Для них справедливы правила суммирования по частям, позволяющие преобразовывать кубатурные формулы с различными дифференциальными элементами.

Формулы с регулярным пограничным слоем строятся как формулы, у которых все внутренние дифференциальные функционалы, входящие в выражение (4), совпадают с точностью до сдвига.

Из формулы (5), переписанной в виде

$$l_0(x) = \sum_{hH\gamma \in \Omega} l_*^!\left(\frac{x}{h} - H\gamma\right) + \sum_{hH\gamma \notin \Omega} l_*\left(\frac{x}{h} - H\gamma\right), \tag{6}$$

получим представление для формул с регулярным пограничным слоем

$$l(x) = l_0(x) - l_1(x),$$

где $l_1(x)$ — некоторый функционал с регулярным пограничным слоем для внешности области Ω.

Подсчет квадрата нормы $l(x)$ в виде

$$\{l(x),\ l_0(x) - l_1(x)\}, \tag{7}$$

где $\{\ \}$ — реализация скалярного произведения в $L_2^{(m)}$, методом дифференциальных функционалов с малым носителем приводит к формуле (2).

Асимптотическая оптимальность формул с регулярным пограничным слоем устанавливается сравнением погрешности «формул с регулярным пограничным слоем» с погрешностью оптимальных формул. Для этого используется теорема Бабушки.

Если формула (4) оптимальна при заданных узлах, то экстремальная функция, на которой функционал $l(x)$ достигает своего максимального значения, имеет нули во всех узлах решетки.

Скалярное произведение функционалов вида (6) порождает в пространстве C_k некоторую квадратичную форму. Взаимная с ней разностная форма позволяет построить разностный оператор перехода от значений экстремальной функции $\psi(x^{(k)})$ кубатурной формулы в узлах решетки к коэффициентам C_k.

Вместе с теоремой Бабушки, представленной (7), это позволяет провести оценки, доказывающие формулу (3).

При эффективном нахождении коэффициентов пограничного слоя для многогранников доказывается сначала, что эти коэффициенты определяются только локальными свойствами границы и, следовательно, совпадают в точках, одинаково отстоящих от граней. Отыскание их величин удобно производить при помощи преобразования Фурье.

Институт математики Сибирского отделения АН СССР, Новосибирск, СССР

15

История и вопросы преподавания
History and pedagogical questions
Histoire et questions d'enseignement
Geschichtliche und pädagogische Fragen

ИССЛЕДОВАНИЯ ПО ИСТОРИИ МАТЕМАТИКИ В СТРАНАХ ВОСТОКА В СРЕДНИЕ ВЕКА: ИТОГИ И ПЕРСПЕКТИВЫ

А. П. ЮШКЕВИЧ

В последние десятилетия значительно вырос интерес к средневековой науке, и в частности к науке средневекового Востока. Это связано не только с значительностью соответствующих научных проблем, выявившейся в ходе исследований, но и с тем огромным политическим, хозяйственным и культурным подъемом, который ныне переживает большинство государств Азии и Африки, еще недавно находившихся в колониальной или полуколониальной зависимости. Конечно, и ранее историки науки обращались к науке древних Китая и Индии и тем более к науке арабского мира. Однако никогда работы в этой области не были столь интенсивными, как в наше время, и никогда еще не был произведен столь радикальный пересмотр традиционных концепций и оценок, как теперь. Короче говоря, этот пересмотр состоит в отказе от господствовавшего ранее европоцентризма, согласно которому современная мировая наука явилась созданием одних только европейских народов начиная с древних греков. Другие народы объявлялись лишенными творческой изобретательности, по крайней мере в сфере науки. В лучшем случае им отводилось место неоригинальных эпигонов, способных лишь к пассивному усвоению наследия эллинов и передаче его новым европейским нациям,— так трактовали арабоязычную индийскую науку. В худшем случае им вовсе не отводилось места в созидании современной международной науки, а их отдельные достижения считали лежавшими в стороне от основного потока человеческой мысли и потому лишенными исторического значения. Так оценивали науку Средней Азии XIV — XV вв. и древнего Китая. Приверженцы европоцентризма и сейчас имеются среди историков науки.

Сказанное в полной мере относится к математике, которая наряду с астрономией и в значительной мере на службе у нее была на протяжении средних веков одной из двух ведущих наук.

Изучение истории средневековой математики сопряжено с особенными трудностями. Здесь, в отличие от математики древней Греции или нового времени, все еще не рассмотрено огромное число сочинений, более всего рукописей, хранящихся в библиотеках многих стран. К тому же далеко не все рукописи выявлены и целый ряд собраний ожидает своего описания. Понятно, что усилия многих исследователей направляются прежде всего на разыскание и прочтение первоисточников, на их публикацию, переводы на современные языки и на комментирование.

В этом направлении за последнее время были достигнуты весьма существенные успехи. Если ограничиться даже только тем, что было издано после второй мировой войны, то число таких публикаций достигает нескольких десятков. Так появились русские переводы древнекитайской «Математики в девяти книгах» и «Математического трактата» Сунь цзы, «Патиганиты» Шридхары, арифметического и алгебраического трактатов ал-Хорезми, «Достаточного об индийской арифметике» ан-Насави, «Книги измерения фигур» Бану Муса, сочинения о составных отношениях Сабита ибн Корры, книги о геометрических построениях Абу-л-Вафы, работ по теории параллельных Сабита ибн Корры, ибн ал-Хайсама, Насир ад-Дина ат-Туси, ас-Самарканди, а также Льва Герсонида, трактата об изопериметрических фигурах ибн ал-Хайсама, всех научных трудов Омара Хайяма, сочинений о тройных правилах и об определении хорд в круге ал-Бируни, тригонометрического труда и отрывков из арифметики Насир ад-Дина ат-Туси, «Ключа арифметики» и «Трактата об окружности» Джемшида ал-Каши, «Трактата об определении синуса одного градуса» Кази-заде ар-Руми. К этому следует добавить, что подготовлены русские переводы «Канона Масуда» ал-Бируни, всех математических трудов Сабита ибн Корры, многочисленных комментариев к десятой книге «Начал» Евклида, «Трактата об измерении шара» ибн ал-Хайсама, анонимного тригонометрического трактата, написанного, вероятно, в XI веке, и еще некоторых других арабских сочинений, а также древнекитайских «Математического трактата пяти ведомств» и «Математического трактата» Чжан Цю-цзяня. На английском языке вышли переводы сочинений Бхаскары I, «Патиганиты» Шридхары, фрагмента алгебры ибн Турка, алгебраического трактата Абу Камила, «Начал индийской арифметики» Кушияра ибн Лаббана, «Книги измерения фигур» Бану Муса, фрагментов по теории отношений от ал-Джаухари до Хайяма, комментариев Хайяма к первой и пятой книгам «Начал» Евклида. На немецком был опубликован перевод «Трактата об окружности» ал-Каши [1]. Наконец содержание многих еще не опубликованных в современном переводе рукописей было изложено и проанализировано в ряде книг и журнальных статей [2].

Наряду с переводами чисто математических сочинений появлялись также переводы астрономических трудов, содержащих специальные математические отделы или, во всяком случае, широко применяющих тот или иной математический аппарат. Некоторые из таких сочинений уже были названы, к ним можно добавить английские переводы астрономических таблиц ал-Хорезми, отдельных трудов ал-Бируни, ал-Каши и других ученых [3].

Изучение развития математики в странах средневекового Востока тесно переплетается, как известно, с изучением истории этой науки в Европе. Поэтому здесь нельзя не упомянуть и выполненные в последнее время переводы латинских рукописей, часть которых сама представляла собой переводы или обработки арабских сочинений, а иногда арабских переводов и обработок греческих оригиналов. Так, в английском переводе вышли латинские сочинения, примыкающие к архимедовой традиции в арабской литературе, обширные собрания фрагментов по теории калькуляций Суайнсхеда и других представителей Оксфордской школы XIV в. и по теории широт форм Н. Орема, «Трактат о пропорциях» Брадвардина, «Вопросы к геометрии Евклида» Орема; в русском переводе появились алгебраический трактат «О данных числах» Иордана Неморария, сочинения Орема «О конфигурациях качеств» и «О соизмеримости или несоизмеримости движений неба», а также фрагменты из «О континууме» Брадвардина и т. д. [4].

Приведенное перечисление, отнюдь не полное, дает все же представление о размахе публикаций двадцати последних лет. Значение всех этих изданий чрезвычайно велико. Каждое содействовало более полной и точной характеристике развития математики в средние века, и именно в ходе работы над новыми первоисточниками были получены наиболее существенные новые результаты, порой совершенно неожиданные и раскрывавшие весьма важные стороны исторического процесса, о которых ранее мы и не подозревали. Современные переводы средневековых трудов тем более необходимы, что только благодаря им становятся доступными широкому кругу исследователей сочинения, написанные на языках, известных самому малому числу специалистов. Это, разумеется, нисколько не лишает значения публикаций на языке только оригинала, как были изданы некоторое время назад одна из рукописей Сурья-Сиддханты, «Канон Масуда» ал-Бируни, ряд трудов Насир ад-Дина ат-Туси и т. д.

В докладе на IV съезде математиков Советского Союза (июль 1961 г.), представленном покойным В. П. Зубовым, Б. А. Розенфельдом и мною [5], в качестве важнейшей задачи в изучении средневековой математики была поставлена публикация переводов сочинений этого времени и названы некоторые издания, по нашему мнению, первоочередные. Несколько высказанных тогда пожеланий

за истекшие пять лет были выполнены. В печати появились упоминавшиеся книги по арифметике Кушияра и ан-Насави, новое издание латинского перевода арифметики ал-Хорезми [6], алгебра Абу Камила (по версии М. Финци на древнееврейском языке), геометрия Бану Муса, отдельные работы ал-Бируни, изопериметрический трактат ибн ал-Хайсама. Однако гораздо бо́льшая часть работы, в том числе намеченной нами в 1961 г., впереди.

Конечно, издать все неопубликованные средневековые сочинения по математике практически немыслимо и даже нецелесообразно. Ограничиваясь кругом восточных стран, я предложил бы на ближайшие годы следующую программу изданий, реализация которой, конечно, предполагает дальнейшую подготовку историков математики, владеющих соответствующими языками.

Математика народов Индии: 1) математические главы (14 и 15) «Мага-Ариабхатия» Ариабхаты II; 2) математическая (13) глава «Сиддханта шекхара» и арифметический трактат Шрипати; 3) «Ганита Каумуди» и «Биджа-ганита» Нарайяны; 4) «Тантрасангграха» Нилаканты и комментарий к ней «Юкти-Бхаша», а также «Каранападдхати»,— все эти сочинения представляют исключительный интерес для истории инфинитезимальных методов и специально бесконечных степенных рядов.

Математика Китая: 1) комментарии Лю Хуэя к «Математике в девяти книгах» и другие еще не изданные тексты из собрания «Десяти математических классиков»; 2) математические отделы «Рассуждений» Мэн-си Шэнь Ко; 3) серия алгебраических трактатов Цинь Цзю-шао, Ли Е, Ян Хуэя, Чжу Ши-цзе.

Математика стран ислама: 1) новое издание алгебраического трактата ал-Хорезми (с учетом не использованной Ф. Розеном арабской рукописи и латинских версий, изданных и исследованных Ч. Карпинским); 2) полное собрание математических трактатов Сабита ибн Корры; 3) полное собрание математических сочинений ал-Бируни, а также его «Канона Масуда», содержащего большие математические отделы; 4) полное собрание математических сочинений Насир ад-Дина ат-Туси, включая обе его редакции «Начал» Евклида; 5) полное собрание математических трудов ибн ал-Хайсама, а также его книги по оптике, в которой широко использовалась теория конических сечений; 6) сочинения представителей Самаркандской школы — Кази-заде ар-Руми, ал-Кушчи и др.

К сожалению, я не могу дать каких-либо рекомендаций, относящихся к литературе на древнееврейском языке, тесно связанной с арабской литературой и служившей одним из существенных звеньев, соединявших ее с наукой средневековой Европы. Здесь открывается поле весьма перспективных исследований, о чем свидетельствуют хотя бы труды Льва Герсонида, работавшего в Провансе в первой половине XIV в.

Точно так же я не беру на себя составление программы изданий астрономических трудов (кроме названного уже «Канона Масуда») — это дело историков астрономии.

Намеченная программа, в которую, разумеется, жизнь внесет немалые изменения, мне представляется осуществимой в течение десятилетия, особенно если удастся достичь большего согласования в деятельности ученых различных стран.

Было бы целесообразно создать при одной из наших международных организаций, например Международной Академии истории наук или Международном союзе истории и философии наук, комиссию, которая уточнила бы программу действий и содействовала бы наиболее рациональному распределению усилий между специалистами различных стран. Эта комиссия могла бы также вынести рекомендации относительно желательной формы публикаций. Мне лично представляется наиболее правильным одновременное издание перевода на один из распространенных языков — русский, английский, французский или немецкий — и оригинального текста с указанием основных разночтений в различных рукописях и с достаточно солидным историко-научным комментарием. Такой тип изданий становится все более распространенным, однако наряду с ним продолжают выходить издания либо оригиналов без перевода, доступные немногим, либо одних переводов, которые не могут быть критически проверены учеными, владеющими языком оригинала, а также издания без необходимого комментаторского аппарата.

В различных справочных трудах имеются обширные библиографические сведения о рукописях, хранящихся во многих библиотеках и музеях мира. Сведения эти все же далеки от полноты и богатства многих собраний, или вовсе не описаны, или же зарегистрированы только в их собственных каталогах. При современном размахе исследований остро необходимо подготовить сводные библиографические каталоги всех средневековых математических рукописей, какие можно выявить в настоящее время. То же относится к рукописям по астрономии и механике, а отчасти по другим частям физики и философии, поскольку соответствующие сочинения иногда содержат математические отделы. Между прочим, такой каталог, составленный по данным многих хранилищ Советского Союза, в скором времени выйдет в свет [7]. Подготовка всеобъемлющего единого каталога может мыслиться лишь как международное мероприятие, и организацию его следовало бы поручить той же комиссии, о которой я говорил несколько ранее.

Разумеется, помимо первоисточников, в послевоенное время вышло большое число оригинальных исследований, перечислять которые здесь нет возможности. Я могу лишь упомянуть, что обширные циклы исследований проведены были главным образом

английскими, голландскими, индийскими, китайскими, немецкими, североамериканскими и советскими учеными, причем в Советском Союзе и Соединенных Штатах возникли целые научные школы, которые и внесли основной вклад в разработку истории средневековой математики вообще и в восточных странах в частности. В результате было установлено множество новых отдельных фактов, существенно уточнена хронологическая последовательность в развитии тех или иных направлений и проблем и установлены новые связи между различными странами Азии, Африки и Европы. Все это позволило, а вернее сказать, заставило приступить к упомянутому в начале доклада пересмотру господствовавших ранее концепций [8]. Здесь возникает немало вопросов, и далее я остановлюсь лишь на некоторых. Это проблемы, во-первых, единства и своеобразия средневековой математики, взятой как в целом, так и в основных региональных течениях, во-вторых, ее научного уровня и оригинальности и, в-третьих, значения математики средневекового Востока в общем развитии нашей науки.

До недавнего времени большинство историков математики — Г. Вилейтнер, Дж. Лориа, Э. Белл и многие другие — рассматривали математику средних веков как совокупность нескольких совпадавших во времени, но по существу разнородных и преимущественно независимых друг от друга циклов развития. Совершенно особняком стояла замкнутая математика в Китае, казавшаяся полностью отделенной от прочего мира и не достигшей еще уровня подлинной науки. Весьма примитивной представлялась и математика в Индии; ее отдельные яркие достижения выступали на общем сером фоне как необъяснимое чудо. За математикой стран ислама признавалась только заслуга перевода, изложения и комментирования древних греков и отчасти индийцев и дальнейшей передачи заимствованных знаний европейцам. Столбовая дорога идейного развития, согласно этой концепции, пролегала только по европейским странам и еще их внеевропейским форпостам; она вела от тех же греков к средневековому «роду латинян», как выразился некогда Леонардо Пизанский. Все это, однако, весьма далеко от исторической действительности.

Прежде всего средневековая математика, взятая в целом — я имею в виду время примерно с III по XVI в. н. э.,— была в главном единой по преобладающему в ней предмету исследований, и по своим внутренним связям, причем в ее развитии на протяжении всего этого долгого времени происходили хотя и медленные, но глубокие реальные взаимодействия между различными странами Востока, а также между Востоком и Европой.

В советской литературе распространено деление истории математики на четыре периода: период образования простейших понятий и господства практической математики, следующий период эле-

ментарной математики или учения о постоянных величинах с VII до XVII в., затем период математики переменных величин примерно до последней трети XIX в. и, наконец, современный период математики переменных отношений. Однако если в основу периодизации положить, как сделано здесь, преобладающие объекты исследования, то средние века более естественно выделить в самостоятельный период.

Характеристика греческой и эллинистической математики, как элементарной или как науки о постоянных величинах, является недостаточной, хотя понятия переменной величины и функции в то время в общем виде выделены не были. Уже начиная с V в. до н. э. инфинитезимальные проблемы и методы стали существенным элементом, роль которого, наряду со своеобразными аналитико-геометрическими приемами, еще несколько столетий заметно возрастала и после того осталась значительной.

Архимед не менее типичен для греческой математики, чем Евклид; к тому же античная форма теории пределов является неотъемлемой составной частью «Начал», автор которых вместе с тем написал и один из первых трудов по геометрико-алгебраической теории конических сечений. Древнегреческая математика в этом смысле не была ни элементарной, ни даже преимущественно элементарной. Такое наименование гораздо более адекватно характеризует средневековую математику, задачи и методы которой принято относить к так называемой «высшей» части нашей науки и которые находились на втором или еще более далеком плане. Наоборот, первый план математики средних веков повсеместно образуют элементарно-математические дисциплины, и ведущую роль в развитии здесь приобретают вопросы, которые вовсе не доминировали в классической математике греков, по крайней мере в пору расцвета.

Я постараюсь теперь в нескольких словах указать основные общие черты, присущие математике в средние века. Крайние границы этого периода для разных стран несколько отклоняются в ту или иную сторону от принимаемых обычно для Европы V и XVII вв.

Для математика средних веков характерна на протяжении всего этого времени преимущественная разработка вычислительных методов, связанных с решением сначала сравнительно простых практических задач, а затем и более сложных выросших на этой основе теоретических вопросов. Математика выступает прежде и более всего как собрание разнообразных расчетных алгоритмов, которые, объединяясь, образуют более обширные ее отделы, вступающие в плодотворное взаимодействие. Ряд особенно тонких и сложных алгоритмов создается в непосредственной или косвенной зависимости от астрономии, и, например, задачи учения о календаре порождают первые проблемы того отдела теории чисел, который позднее назвали диофантовым анализом.

В фундаменте средневековой математики лежит арифметика целых и дробей, а в ней господствуют приемы алгоритмического решения нескольких типов задач на пропорции и иных линейных задач. Это известные тройные правила, правила одного и двух ложных положений.

Выше располагается совокупность алгебраических приемов выражения и решения задач. Прогресс числовой алгебры завершается выделением ее в самостоятельную науку, иногда вместе с диофантовым анализом. В разных местах, хотя и не повсюду, изобретаются приемы решения систем линейных уравнений, вычисления корней уравнений высших степеней с любой степенью точности, создается символика, алгебра применяется в геометрии, а геометрия — в алгебре.

Развитие вычислительной математики и алгебры влечет за собой расширение понятия о числе и усовершенствование нумерации.

В геометрии преобладают, как правило, элементарные приемы точного или приближенного измерения наиболее употребительных плоских и пространственных фигур — элементарные в том смысле, что не применяются какие-либо формы предельного перехода. Особенно успешно развивается учение о решении треугольников на плоскости и на шаре, и в конце концов подобно алгебре выделяется как особая наука плоская и сферическая тригонометрия, становящаяся основным аппаратом астрономии.

Это существенное единство проблематики и методов во многих странах Азии, Северной Африки и Европы объяснялось в конечном счете тем, что общей социальной основой развития всей идеологии в них был феодальный строй. Весьма важное значение в формировании науки играли также международные торговые, политические и вместе с ними культурные связи. Конечно, духовные контакты не могли быть столь регулярными, какими они стали в новое время. И все же за длительные промежутки времени, соответственно невысоким темпам всего общественного развития, задачи и методы, представлявшие общий интерес, переносились на огромные расстояния, повсеместно усиливая сродство математических исследований. Более известен процесс распространения арабского наследия в Европе и затем арабско-индийские научные связи. Но целый ряд фактов свидетельствует о весьма интенсивном, хотя и неравномерном идейном обмене также между Китаем и Индией, с одной стороны, и Китаем и странами ислама — с другой. И если в одних случаях одни и те же открытия производились в разных местах независимо (так, быть может, обстояло дело с теоремой Пифагора), то во многих других имело место прямое взаимодействие. Так, способы извлечения корней и правило двух ложных положений распространились

из Китая в арабские страны, а тригонометрические приемы — из арабских стран в Китай.

Должен подчеркнуть, что научное взаимодействие до сих пор весьма неполно исследовано даже в лучше всего изученной области связей между арабскими и европейскими странами. Его дальнейшее исследование, которое затрудняется тем, что в сочинениях тех времен почти не встречаются ссылки на источники, является одной из важных задач.

Единству средневековой математики в главных направлениях, проблематике и методах не противоречат специфические особенности ее в различных странах и группах стран, определявшиеся различием в общих условиях их развития, точно так же как единству математики в Европе нового времени не противоречит относительное своеобразие ее, скажем, в Англии или России XIX в. Конечно, в средние века такого рода различия были значительнее и устойчивее, ибо возможности взаимодействия были гораздо слабее.

Математика Китая, который менее других восточных стран был подвержен влияниям греко-римской культуры, по структуре и стилю, особенно стилю изложения, напоминает математику древнего Вавилона, хотя и далеко превзошла его в алгебре, теории чисел и интерполяционных методах, применявшихся в астрономии. Геометрия в греческом ее построении оставалась Китаю чуждой, по крайней мере до XIII в. Впрочем, Китай не остался, видимо, полностью в стороне от некоторых идей Архимеда, проникших сюда, вероятно, через посредство Индии. Я имею в виду серию китайских работ II—V веков по уточненному вычислению отношения длины окружности к диаметру. Однако этот пример пока остается единственным.

В тесной связи и родстве с китайской была математика Индии. Здесь все же гораздо сильнее отразилось влияние греческой и затем эллинистической науки, стимулировавшей развитие астрономии и тригонометрии в собственном смысле слова, а быть может, и теории чисел. Но столь характерные для Индии арифметические, алгебраические и теоретико-числовые проблемы и методы идейно ближе к китайским, чем к греческим.

Наиболее высокого уровня на Востоке математика достигла в странах ислама на протяжении IX—XV веков. В этом сыграла свою роль большая поддержка, которую оказывали астрономии многие правители; но в значительной мере блестящие успехи науки определялись здесь, по-видимому, благоприятными «начальными условиями». На территориях арабского халифата имелась возможность быстрого усвоения как греческой и эллинистической культуры частично в ее сплаве с местными вавилонскими и персидскими традициями, так и достижений индийцев и китайцев. Это позволило математикам стран ислама быстро приобрести обширные фактические познания и затем применить к решению поставленных задач более мощные методы исследования, чем те, какими располагали ученые других восточных стран, а потому достичь во многих случаях более общих и глубоких результатов. На Востоке только в странах ислама была воспринята и обогащена новыми открытиями античная дедуктивная система геометрии. Здесь же некоторое время успешно применялись и совершенствовались античные методы квадратур и кубатур и получила обширные приложения в алгебре и оптике теория конических сечений. Впрочем, разработка метода исчерпывания продолжалась сравнительно недолго, а теория конических сечений не получила нового развития: весь соответствующий круг проблем лежал вне основной области интересов ученых стран ислама.

Наконец, математика в Европе после первоначального резкого упадка вновь становится на путь ускоряющегося прогресса в XII в., прежде всего отправляясь от изучения арабской литературы. Довольно продолжительное время в Европе процветает латинский по языку вариант науки стран ислама. В XIII в. начинается и непосредственное изучение греческих источников, первое знакомство с которыми происходило по их арабским переводам и изложениям. И в том же XIII в. в математике европейских стран впервые появляются новые идеи, которым суждено было впоследствии произвести подлинную революцию во всем математическом и научном мышлении. Я имею в виду первые опыты построения буквенной алгебры, ставшей прообразом всех последующих исчислений, и первые же попытки разработки механики неравномерных движений, ставшей главным источником математики переменных величин.

Что касается вопроса, была ли средневековая математика, и особенно восточная математика, наукой, то ответ на него зависит от смысла, который придается этому слову. Если называть наукой только дедуктивные системы, построенные на явно сформулированных аксиомах и определениях, подобно геометрии в «Началах» Евклида, то ни в Китае, ни в Индии математика еще не стала наукой. Но в указанном смысле наукой не были ни арифметика, ни алгебра вплоть до XIX в., хотя к этому времени они уже более двух тысяч лет представляли собой развитые математические теории, т. е. системы, в которых из тех или иных данных предложений получаются с помощью определенных правил вывода другие ранее неизвестные предложения. Если же принять, что математическое познание, развивающееся как теория, есть научное познание, то математика Китая и Индии в средние века безусловно являлась наукой, а не собранием разрозненных эмпирических рецептов решения задач. Конструкция правил или алгоритмов этой науки опиралась на разветвленную систему понятий и операций арифметики, алгебры, геометрии. Дискурсивное математическое мышление

играло активную творческую роль; справедливость многих приемов обнаружилась в самом ходе конструкции алгоритма и они как бы содержали в самих себе проверку — доказательство своей истинности. Потребность в теоретическом обосновании общих предложений, впервые обнаруженных, вероятно, эмпирически и на частных примерах, очень давнего происхождения. Так, в Китае еще в древности доказывали теорему, за которой закрепилось имя Пифагора. В общем математика средневековой Индии и Китая была не в меньшей мере научной, чем, скажем, алгебра и теория чисел в «Арифметиках» знаменитого Диофанта.

Я уже касался по ходу дела другого поставленного вопроса об оригинальности математики средневекового Востока. Для более полного ответа я просто перечислю, не претендуя на исчерпывающую полноту, наиболее важные принадлежащие ей открытия. Это: десятичная позиционная нумерация без знака нуля (Китай, до начала н. э.); десятичная позиционная система с нулем (Индия, не позднее VII в.); десятичные дроби (Китай, начиная с IV в.; страны ислама, начало XV в.); полная шестидесятеричная позиционная система целых и дробей, ранее известная в древнем Вавилоне (страны ислама, около 1000 г.); отрицательные числа с применениями в алгебре (Китай, около начала н. э.; Индия, не позднее начала VII в.); общая «антифайретическая» теория отношений, опирающаяся на разложение отношения в непрерывную дробь и ранее известная, вероятно, грекам до Эвдокса; трактовка иррациональных величин как чисел (страны ислама с IX в.); тройное правило и его обобщения (Китай, около начала н. э.; Индия, не позднее V в.); правило двух ложных положений (Китай, около начала н. э.) и его доказательство (страны ислама, IX в.); алгоритм извлечения квадратного и кубичного корней, основанный на разложениях квадрата и куба двучлена (Китай, около начала н. э.); алгоритм извлечения корня с любым натуральным показателем (Китай, I тысячелетие; страны ислама, XI в.); применение разложения произвольной натуральной степени двучлена к приближенному извлечению корней (страны ислама, не позднее середины XIII в.); алгоритм решения определенной линейной системы n уравнений с n неизвестными (Китай, около начала н. э.); введение произвольных натуральных степеней неизвестной величины (Индия, VII в.; Китай, не позднее XIII в.); так называемый метод Руффини — Горнера нахождения (положительных) корней численных алгебраических уравнений любой степени с применением линейных преобразований корней (Китай, VII — XIII вв.); символическая запись алгебраических выражений и уравнений (Индия, не позднее VII в.; Китай, не позднее XIII в.); преобразование алгебры в самостоятельную науку; геометрическая теория кубических уравнений, отделение корней; различные численные методы решения уравнений 3-й степени, в частности трисекции угла (страны ислама, IX — XV вв.); применение алгебры к геометрии (Китай, около начала н. э.; Индия, не позднее III в. до н. э.; страны ислама с IX в.); решение так называемого трансцендентного уравнения Кеплера по методу итерации (страны ислама, IX в.); начала тригонометрии (Индия, не позднее V — VI вв.); создание плоской и сферической тригонометрии как самостоятельной науки; усовершенствованные методы вычисления тригонометрических таблиц; вычисление синуса одного градуса и числа π с 17 верными десятичными цифрами (страны ислама, IX — XV вв.); разработка теории параллельных (страны ислама, IX — XIV вв.); суммирование арифметических рядов (Китай, XI — XIII вв.; Индия VI и следующих веков); квадратичные и кубичные интерполяционные формулы (Китай, VII — XIII вв.); квадратичное интерполирование (страны ислама, XI в.); интегрирование $\sqrt{x}$ путем деления промежутка интегрирования на неравные части в арифметической прогрессии (страны ислама, IX в.); интегрирование x^4 с помощью суммирования ряда четвертых степеней натуральных чисел (страны ислама, XI в.); исследование ускоренного движения и поведения величины в окрестности экстремума (страны ислама, XI в.); разложения в степенные ряды арктангенса, синуса и косинуса и вычисление π с 10 десятичными цифрами (Индия около 1500 г.).

Список этот говорит сам за себя. Оригинальность упоминаемых в нем открытий не умаляется тем, что первый толчок в некоторых случаях исходил от древних греков (как в случае тригонометрии, выросшей на почве эллинистического исчисления круговых хорд) и что отдельные вещи были открыты заново (как давно забытая антифайретическая теория отношений).

Значительная часть только что перечисленных открытий и достижений, хотя и не все, вошла в состав математики стран Европы. Насколько же велико было их значение для последующего прогресса математики?

На одной из дискуссий по истории науки, состоявшихся в Оксфорде в 1961 г., мне довелось услышать из уст Б. ван дер Вардена характеристику роли древнегреческой науки в становлении науки XVII и следующих веков, которая не оставляла никакого существенного места для средневекового Востока. Аргументация выдающегося алгебраиста, ставшего не менее выдающимся историком античной математики и астрономии, очень типична для европоцентрической концепции, и я ее воспроизведу. Основой большей части современной науки является, говорил Б. ван дер Варден, механика Ньютона, а в ней сошлись три нити, каждая из которых ведет из Греции. Это планетарная система астрономии, развитая Коперником и Кеплером, а без законов Кеплера Ньютон не мог бы построить свою механику. Далее это аксиоматическая структура

греческой геометрии — модель аксиоматики механики Ньютона и конические сечения Аполлония, широко примененные в «Математических началах натуральной философии». Наконец, это греческая механика, оказавшая, вероятно, влияние на Галилея и его открытие закона инерции, а значит, на Ньютона. Впрочем, возможно, что Галилей пришел к закону инерции независимо от греческих влияний. Во всяком случае, остаются две первые нити, а на долю восточной науки приходится, очевидно, только передача европейцам греческих знаний [9].

В колоссальном значении греческой науки для творчества Ньютона сомневаться не приходится, как и в колоссальном влиянии, оказанном великим английским ученым на последующее развитие научной мысли. Однако наука нового времени развивала не только те идеи, которые входили в состав греческого наследия, в творчестве Ньютона сходились не только указанные три нити, да и не одна лишь механика Ньютона послужила основой большей части современной науки.

Ведущей математической дисциплиной в Европе средних веков, прогресс которой был предпосылкой и аналитической геометрии Декарта — Ферма, и исчисления бесконечно малых Ньютона — Лейбница, а позднее теории групп, математической логики и т. д., была алгебра. А отправным пунктом ее в Европе средних веков была алгебра ал-Хорезми и его преемников в странах ислама, т. е. прежде всего алгебра численных уравнений, содержавшая элементы алгебраического исчисления и развивавшая подобно алгебре Диофанта, но независимо (насколько известно) от нее древневосточную традицию, выраженную уже в вавилонских клинописях.

Другой ведущей математической наукой средневековой Европы, тесно связанной с алгеброй, была тригонометрия, с которой европейцы познакомились также по арабской литературе. Тригонометрия была основным и незаменимым аппаратом астрономии, который постоянно применяли те же Коперник и Кеплер, да и сам Ньютон. Лишь первые начала тригонометрии были положены в греческой геометрии хорд, все остальное сделали индийцы и особенно ученые стран ислама.

Напомню, что вся наша десятичная позиционная арифметика имеет индийско-арабское происхождение.

Я не думаю, что следует пытаться взвесить, что было важнее для создания современной науки: аксиоматическая геометрия или же числовая буквенная алгебра, конические сечения или тригонометрия, античная теория пропорций или принцип поместного значения. Я не думаю также, что стоит размышлять, как это иногда делают, над тем, что было бы, если бы ученые Европы к началу нового времени не располагали той или иной из греческих

или восточных теорий. Такие мысленные эксперименты, которые призваны доказать, что не будь тригонометрии, то Ньютон ее тут же бы и придумал, а вот не будь теории конических сечений, то ее изобретение далось бы ему с трудом, лишены ценности, ибо ненаблюдаемы и непроверяемы. История учит не тому, что могло бы быть в прошлом, при условиях, которых на деле вовсе не было, а тому, что было, есть и может случиться в действительных условиях. В действительности математика, подобно всей науке нового времени, включила в себя как существенные элементы открытия и древних греков, и народов Востока; она была и остается результатом международного творчества ученых многих стран. Разумеется, нарисованная мною картина — неполная и нуждается в различных добавлениях и уточнениях. Я уже отметил, например, что мы слишком мало знаем еще о взаимных связях между Китаем, Индией и странами ислама. Еще менее изучены связи науки этих стран с традициями, восходящими к древнему Вавилону, с одной стороны, и к Византии — с другой. Практически ничего неизвестно о состоянии научных знаний на территории Средней Азии, в частности Хорезма, перед арабским завоеванием. Перечень таких «белых мест» можно было бы продолжить. Здесь более всего следует ожидать от привлечения новых источников, притом не только самих математических или астрономических рукописей, но и данных общей истории культуры, торговых и политических отношений.

Мне представляется, что высказанные соображения о характере и особенностях средневековой математики имеют существенное значение для решения другой категории проблем, которые я почти не затрагивал и здесь могу лишь упомянуть. Это проблемы анализа того воздействия общественного базиса на научную надстройку, которое в конечном счете определяло прогресс — а временами регресс — математических наук. Здесь требуется более глубокое, чем было дано до сих пор, исследование коренных причин, обусловивших как общие черты математики эпохи господства феодализма, так и отличия ее истории в Азии и Северной Африке, с одной стороны, и в Европе — с другой, или в различных азиатских странах. Почему, скажем, научные достижения в мавританских государствах оказались гораздо менее значительными, чем в восточных странах ислама? Почему в Китае после выдающихся успехов науки и техники вплоть до XIII в. наступает затем многовековой научный застой [10]? Чем вообще объясняется упадок науки, наблюдаемый почти в одно время в странах Востока — в Китае с XIV в., в странах ислама с середины XV в., в Индии, по-видимому, еще ранее, если не считать недолгого подъема в отдельных ее районах на рубеже XVI — XVII вв.? И почему именно в Европе, которая еще в XII в. была всего лишь ученицей арабов и древних греков, темп культурного и научного развития оказы-

вается гораздо более быстрым, чем на Востоке, и, взятый для всего круга европейских стран, бесперебойным, так что к концу феодального периода здесь закладываются прочные основы науки нового времени?

Решение всех этих больших вопросов не может быть дано в рамках одной только истории науки, оно требует глубокого марксистского исследования исторического процесса в целом. Надо признать, что в этом направлении сделано пока немногое. Однако уже при нынешнем состоянии исторических знаний можно ожидать больших успехов в решении поставленных вопросов, если к ним будет привлечено достаточное внимание. Школы, работающие над историей средневековой науки, многочисленны и активны; рядом с учеными старших поколений мы видим среди их участников и молодых исследователей. В этой преемственности творческих усилий верный залог дальнейших достижений в решении стоящих перед нами задач.

Институт истории естествознания и техники АН СССР, Москва, СССР

ЛИТЕРАТУРА

[1] «Математика в девяти книгах». Перевод, статья, примечания Э. И. Березкиной. Историко-математические исследования (ИМИ), X, 1957.

Суньцзы, Математический трактат. Перевод, статья, примечания Э. И. Березкиной. Из истории науки и техники в странах Востока (ИНТВ), III, 1963.

Шридхара, Патиганита. Перевод О. Ф. Волковой и А. И. Володарского, статья А. И. Володарского. Физико-математические науки в странах Востока (ФМНСВ), I (IV), 1966.

Мухаммад ал-Хорезми, Математические трактаты. Перевод Ю. Х. Копелевич и Б. А. Розенфельда, комментарии Б. А. Розенфельда, Ташкент, 1964.

Абу-л-Хасан ан-Насави, Достаточное об индийской арифметике. Перевод и примечания М. И. Медового, ИМИ, XV, 1963.

Бану Муса, Книга измерения фигур. Перевод и примечания Джамаля ад-Даббаха, ИМИ, XVI, 1965.

Сабит ибн Корра, Книга о составлении отношений. Перевод и примечания Б. А. Розенфельда и Л. М. Карповой, ФМНСВ, I (IV), 1966.

Абу-л-Вафа, Книга о том, что необходимо ремесленнику из геометрических построений. Перевод и примечания С. А. Красновой, ФМНСВ, I (IV), 1966.

Сабит ибн Корра, Книга о доказательстве известного постулата Евклида. Перевод Б. А. Розенфельда, статья и примечания Б. А. Розенфельда и А. П. Юшкевича, ИМИ, XIV, 1961.

Сабит ибн Корра, Книга о том, что две линии, проведенные под углами, меньшими двух прямых, встретятся. Перевод и примечания Б. А. Розенфельда, ИМИ, XV, 1963.

Хасан ибн ал-Хайсам, Книга комментариев к введениям книги Евклида «Начала» (отрывок). Перевод, статья, примечания Б. А. Розенфельда, ИМИ, XI, 1958.

Насир ад-Дин ат-Туси, Трактат, исцеляющий сомнение по поводу параллельных линий. Перевод Б. А. Розенфельда, статья и примечания Б. А. Розенфельда и А. П. Юшкевича, ИМИ, VIII, 1960.

Шамс ад-Дин ас-Самарканди, Основные предложения (отрывок). Перевод Б. А. Розенфельда, ИМИ, XIV, 1961.

Герсонид Лев, Комментарии к введениям книги Евклида. Перевод И. Г. Польского, примечания Б. А. Розенфельда, ИМИ, XI, 1958.

Ибн ал-Хайсам, Трактат об изопериметрических фигурах. Перевод и примечания Джамаля ад-Даббаха, ИМИ, XVII, 1966.

Омар Хайям, Трактаты. Перевод Б. А. Розенфельда, статья и комментарии Б. А. Розенфельда и А. П. Юшкевича, М., 1961.

Омар Хайям, Первый алгебраический трактат. Перевод и примечания Б. А. Розенфельда и С. А. Красновой, ИМИ, XV, 1963.

Ал-Бируни, Трактат об определении хорд в круге при помощи ломаной линии, вписанной в него. Перевод С. А. Красновой и Л. А. Карповой, примечания Б. А. Розенфельда и С. А. Красновой, ИНТВ, III, 1963.

Ал-Бируни, Книга об индийских рашиках. Перевод и примечания Б. А. Розенфельда, ИНТВ, III, 1963.

Мухаммед Насирэддин Туси, Трактат о полном четырехстороннике. Перевод под редакцией Г. Д. Мамедбейли и Б. А. Розенфельда, Баку, 1952.

Насир ад-Дин ат-Туси, Сборник по арифметике с помощью доски и пыли (отрывок). Перевод и примечания С. А. Ахмедова, ИМИ, XV, 1963.

Джемшид Гиясэддин ал-Каши, Ключ арифметики. Трактат об окружности. Перевод Б. А. Розенфельда, редакция В. С. Сегаля и А. П. Юшкевича, комментарии А. П. Юшкевича и Б. А. Розенфельда, М., 1956.

Кази-заде ар-Руми, Трактат об определении синуса одного градуса. Перевод Б. А. Розенфельда, статья и примечания Б. А. Розенфельда и А. П. Юшкевича, ИМИ, XIII, 1960.

Bhaskara I., The Maha-Bhaskariya, ed. and transl. into English with notes and comment by K. S. Shukla, Lucknow, 1960.

Sridharacarya, The Patiganita with an ancient Sanskrit commentary, ed. with introduction, notes and English transl. by K. S. Shukla, Lucknow, 1959.

Sayili A., Logical necessities in mixed equations by Abd al Hamid ibn Turk and the algebra of his time, Ankara, 1962.

Levey M., The Algebra of Abu Kamil in a Commentary by Mordecai Finzi. Hebrew text, translation and commentary with special reference to the Arabic text, The University of Wisconsin Press, 1966.

Kyshyar ibn Labban, Principles of Hindu reckoning. Translation with introduction and notes by M. Levey and M. Petruck, The University of Wisconsin Press, 1965.

The Verba filiorum of the Banu Musa (M. Clagett, Archimedes in the Middle Ages, vol. I, The University of Wisconsin Press, 1964).

Plooij E. B., Euclid's conception of ratio and his definition of proportional magnitudes as criticized by arabian commentators, Rotterdam, 1950.

Amir-Moez A. R., Discussion of difficulties in Euclid by Omar Khayyam, *Scripta mathematica*, **24**, № 4, 1959.

Luckey P., Der Lehrbrief über den Kreisumfang von Ǧamšid b. Mas'ud al-Kaši, Berlin, 1953.
[2] См. библиографию в книге: Юшкевич А. П., История математики в средние века, М., 1961 и в дополненном немецком издании: Juschkewitsch A. P., Geschichte der Mathematik im Mittelalter, Leipzig, 1964.
[3] The astronomical tables of al-Khwarizmi. Translation with commentaries by O. Neugebauer, Kobenhavn, 1962.
Al-Biruni on transits... translated by M. Saffouri and A. Iram with a commentary by E. S. Kennedy, American University of Beirut, 1959.
The Planetary equatorium of ...al-Kashi. With translation and commentary by E. S. Kennedy, Princeton University Press, 1960.
[4] Clagett M., Archimedes in the Middle Ages, vol. I, The Arabo-Latin tradition, The University of Wisconsin Press, 1964.
Clagett M., The Science of Mechanics in the Middle Ages. The University of Wisconsin Press, 1959.
Th. of Bradwardike, Tractatus de proportionibus... edited and translated by H. L. Crossby, Jr. The University of Wisconsin Press.
Oresme N., Quaestiones super Geometriam Euclidis, ed. by H. L. L. Busard, Leiden, 1961.
Иордан Неморарий, О данных числах. Перевод С. Н. Шнейдера, ИМИ, XII, 1959.
Орем Н., О конфигурации качеств. Перевод, статья и примечания В. П. Зубова, ИМИ, XI, 1958.
Орем Н., О соизмеримости или несоизмеримости движений неба. Перевод, статья и примечания В. П. Зубова. Историко-астрономические исследования, VI, 1960.
Зубов В. П. Трактат Брадвардина «О континууме», ИМИ, XIII, 1960.
[5] Зубов В. П., Розенфельд Б. А., Юшкевич А. П., Об исследованиях по истории математики средних веков, ИМИ, XV, 1963.
[6] Mohammed ibn Musa, Alchwarizmi's Algorismus, von K. Vogel, Aalen, 1963.
[7] Розенфельд Б. А., Арабские и персидские физико-математические рукописи в библиотеках Советского Союза, ФМНСВ, I (IV), 1966.
[8] Юшкевич А. П., О математике народов Средней Азии в IX—XV вв. ИМИ, IV, 1951.
Юшкевич А. П., О достижениях китайских ученых в области математики, ИМИ, VIII, 1955.
Юшкевич А. П., История математики в средние века, М., 1961.
Юшкевич А. П., Розенфельд Б. А., Математика в странах Востока в средние века, ИНТБ, 1, 1960.
Needham J. (with collaboration of Wang Ling), Science and Civilisation in China, Vol. 3, Cambridge, 1963.
Crombie A. C., Augustine to Galileo, Vol. I—II, London, 1959—1961.
[9] Scientific Change, ed. by A. C. Crombie, London, 1963, p. 168.
[10] Needham J., Poverties and triumphs of the Chinese Scientific tradition, в книге: Scientific Change, ed. by A. C. Crombie, London, 1963.

Секция 5 Section 5

NON-LINEAR RELATIVISTIC PARTIAL DIFFERENTIAL EQUATIONS

IRVING E. SEGAL

Introduction

As is well known, evolutionary partial differential equtions involving non-linear local interactions are of importance in many diverse connections. The theory of the Navier-Stokes (and related) equations has had great impact both in pure and applied mathematics; these equations are typical of non-linear parabolic equations with a local interaction. In relativity and quantum field theory one has primarily to deal with hyperbolic equations, with similarly local interactions. We present here an account of recent developments concerning fairly typical such equations, with emphasis on global aspects relevant to the cited applications; these are, more specifically, the temporal asymptotics and the phase space structure on the solution manifold.

The Cauchy problem

The basic partial differential equations of quantum mechanics are conveniently taken, for purposes of global analysis as well as for generality, in the evolutionary form

$$u' = Au + K(u, t), \tag{1}$$

where $u = u(t)$ has values in a Banach space $\mathbf{L}$, A is a given unbounded operator in $\mathbf{L}$, and $K(u, t)$ is a given non-linear function of u and t, quite commonly not everywhere defined on $\mathbf{L}$. For many theoretical purposes, the corresponding integral equation

$$u(t) = W(t - t_0)\, u(t_0) + \int_{t_0}^{t} W(t - s)\, K(u(s), s)\, ds, \tag{1'}$$

where $W(s)$ denotes the one-parameter semi-group generated by A, is more fundamental. From this abstract standpoint the basic distinction between the parabolic and hyperbolic cases is that in the former

A has typically a real semibounded spectrum, while in the latter A has a pure imaginary spectrum; in the hyperbolic case, $W(s)$ is a full group, not merely a semi-group.

An important and relatively typical hyperbolic equation is the time-independent (i.e. autonomous) second-order equation

$$\Phi''(t) + B^2\Phi(t) = J(\Phi), \tag{2}$$

where B is a given non-negative self-adjoint operator in a Hilbert space $\mathbf{H}$, and J is a given non-linear operator in $\mathbf{H}$. This equation is readily subsumed under equation (1′) by taking $\mathbf{L}$ as the set of all pairs $[f, g]$ with $f \in \mathbf{H}^a$ and $g \in \mathbf{H}^{a-1}$, where a is an adjustable parameter, and $\mathbf{H}^a$ denotes the Hilbert space completion of the domain of B^a relative to the norm $\|f\|_a = \|B^a f\|$, $\|\cdot\|$ denoting the norm in $\mathbf{H}$; $W(t)$ as the unitary operator on $\mathbf{L}$ whose matrix decomposition in the representation $\mathbf{L} = \mathbf{H}^a + \mathbf{H}^{a-1}$ is

$$\begin{pmatrix} \cos(tB) & \frac{\sin(tB)}{B} \\ -B\sin(tB) & \cos(tB) \end{pmatrix}$$

(these operators being extended from $\mathbf{H}$ to the $\mathbf{H}^b$ for arbitrary b in the obvious manner); and $K(u, t)$ as the mapping $[f, g] \to [0, -J(f)]$. The choice of a is influenced by physical considerations, and strongly affects the existence of energy-inequalities and the continuity of K as an operator in $\mathbf{L}$. The "scalar relativistic equation with local interaction", $\Box\varphi = m^2\varphi + F(\varphi)$, where F is a given function of a complex variable such that $F(0) = F'(0) = 0$, is readily subsumed, for example, the value $a = 1$ corresponding to the conventional energy norm, but in more than two space dimensions, the mapping K will not in general be continuous when F is a polynomial, unless a larger value of a is taken.

Developing classical differential equation methods in functional analytic form, the basic theory of such equations may be developed, a representative result being (with a taken as 1).

Theorem 1. *If J is locally lipschitzian from $\mathbf{H}^1$ to $\mathbf{H}$, then the Cauchy problem for equation (2), in its integrated form (1′), has a unique local solution.*

If in addition there exists a non-positive function E on $\mathbf{H}$ whose Frechet differential exists at every point $\Phi \in \mathbf{H}^1$ and has the form: $\Psi \to \mathrm{Re}\,\langle \Phi, \Psi \rangle$, then the solution exists and is unique globally in time.

If J is Frechet differentiable, then the differential equation is satisfied in the strict form (1) if and only if the initial data $[f, g]$ are in the domain of the generator of the one-parameter group W, i.e. if $f \in \mathbf{H}^2$ and $g \in \mathbf{H}^1$.

For example, the equation

$$\Box\varphi = m^2\varphi + g\varphi^p \qquad (g > 0,\ p \text{ odd and} > 0) \tag{3}$$

has unique global solutions in one and two space dimensions, and also in three dimensions in case $p = 3$, to the Cauchy problem, as first shown by K. Jörgens [4] by a classical treatment, inasmuch as $E(\Phi)$ may be taken as $-\text{const} \cdot \int \Phi^{p+1}$. The partial conflict between the non-negativity and regularity requirements on the functional $J(\Phi)$ is, illustrated by the circumstnace that the mapping $\Phi \to \Phi^p$ is not continuous if the number of space dimensions n exceeds 3 and $p \geqslant 3$, or even if $n = 3$ and $p > 3$, from $\mathbf{H}^1$ to $\mathbf{H}$; with a change in the value of a, these mappings may be made continuous, and a local existence theorem obtained, but the non-negativity feature is lost, and thereby the global existence. Nevertheless, in this case, Theorem 1 may be applied to "cut-off" equations, involving modified functionals J, which are spatially non-local, and combined with compactness arguments to show

Theorem 2. *The Cauchy problem for equation (3) with arbitrary finite-energy initial data has a global weak solution, for arbitrary n and (odd) p.*

It is difficult to determine whether these solutions are unique, or whether they are globally regular if initially so. This is however the case as regards the solutions primarily relevant to dispersion theory, in the case $n = 3$, as indicated below.

Dispersion of solutions

Dispersion theory concerns the temporal asymptotics of two different evolutionary equations, having the same Cauchy data spaces, in relation to one another. To illustrate the concepts in a representative and fairly general case, the *forward wave operator* for equation (1′) is the transformation $u_0 \to u$ from a solution $u_0(t)$ of the "free equation" $u_0(t) = W(t - t_0)u(t_0)$ to a solution $u(t)$ of the given equation (1′), which is asymptotic as $t \to -\infty$ to u_0 in the sense that

$$u(t) = u_0(t) + \int_{-\infty}^{t} W(t-s)K(u(s), s)\,ds. \tag{4}$$

In general, especially for the time-independent case in which $K(v, t)$ is independent of t, the infinite integral in equation (4) will fail to exist; although the wave operator is widely used in theoretical applied work, it is only with material restrictions on K that it has mathematical existence. Closely related to the wave operator, and still more impor-

tant in theoretical applied work, is the "dispersion" (also called "scattering" or "collision") operator; this is the transformation $u_0 \to u_1$, where u_1 is the solution of the free equation which is asymptotic as $t \to +\infty$ to the same solution of equation (1') to which $u_0(t)$ is asymptotic as $t \to -\infty$.

Fairly representative of results on the wave operator is the

Theorem 3. *The forward wave operator exists for the equation*

$$\Box\varphi = m^2\varphi + F(\varphi) \qquad (m > 0,\ F \text{ of class } C^\infty)$$

and is unique within a specified regularity class, provided:

(a) *the free solution φ_0 is sufficiently regular (e.g. has Cauchy data which are infinitely differentiable and of compact support);*

(b) $F^{(j)}(\lambda) = O(|\lambda|^r)$ *for some* $r > 2 + \frac{2}{n} - j$ *and* $j = 0, 1, \ldots, j_n$, *near* $\lambda = 0$.

Corollary 3.1. *The wave operator exists in the indicated sense for equation* (3) *provided:* (a) $n \geqslant 3$, $p \geqslant 3$, *and* $m > 0$; (b) $n = 1$ *or* 2, $p \geqslant 5$, *and* $m > 0$.

A considerably more general result may be given, but the statement is quite long. In the case $m = 0$ there are similar results provided $n \geqslant 3$, but the requirements on F are more stringent. The proof depends on the use of auxiliary norms in the Cauchy data space, and of suitable estimates of the temporal decay of the solutions of the free equation relative to these norms, as indicated below.

The question of the existence of an asymptotic free solution φ_1 to the solution φ of the non-linear equation is of a different nature; unlike Theorem 3, which is essentially a local existence theorem (in a neighborhood of the time $t = -\infty$) and does not depend on energy boundedness, such bounds are essential in deriving results such as

Theorem 4. *Any finite-energy weak solution of the integrated form of equation* (3) *is asymptotic weakly, as $t \to +\infty$, to a unique finite-energy solution of the free equation, provided $n \geqslant 3$ and $m > 0$.*

This shows the existence of dispersion when $n \geqslant 3$ and $m > 0$, for a suitable given solution of the non-linear equation, but the univalence, regularity, etc. properties of the dispersion operator are difficult to treat in the case of weak asymptotics. Strong asymptotics depend on estimates of temporal decay of auxiliary norms of solutions of the non-linear equation. In the case of the important equation $\Box\varphi = g\varphi^3$ ($g > 0$, $n = 3$), a good estimate was obtained by W. Strauss [9] by refined energy integral methods, which however are not readily generalized. Results for more general equtions, of a somewhat different nature, may be obtained relatively systematically by combining

sharp estimates of the decay of auxiliary norms of solutions of the associated linear relativistic equations (these are obtained by a method due essentially to A. R. Brodsky [1]) with a series of applications of conventional inequalities. In general, larger values of n, p and m lead to more rapid decay and stronger dispersion results, although the situation is somewhat complex; a representative result, for the case of greatest quantum-mechanical interest is

Theorem 5. *If $n = 3$ and $m > 0$, the dispersion operator exists strongly for equation* (3), *for any given free solution $u_0(t)$ whose Cauchy data have a sufficient number of integrable derivatives, if g is sufficiently small.*

More specifically, there exists a solution $u(t)$ of the equation

$$u(t) = u_0(t) + \int_{-\infty}^{t} W(t-s)\, K(u(s))\, ds,$$

and a free solution $u_1(t)$ such that

$$u(t) = u_1(t) - \int_{t}^{\infty} W(t-s)\, K(u(s))\, ds,$$

both integrals being absolutely convergent in the energy norm.

The general method may be sketched briefly as follows. Setting $\Phi(t) = \varphi(x, t)$ as a function of x, and defining $\Phi_0(t)$ similarly, equation (4), with $K(u, s)$ as in equation (3), is equivalent to the equation

$$\Phi(t) = \Phi_0(t) + \int_{-\infty}^{t} \sin((t-s)B)\, B^{-1}(g\Phi(s)^p)\, ds, \qquad B = (m^2 I - \Delta)^{1/2}.$$

Taking L_r norms, writing $\sin(tB)\, B^{-1}\Psi$ as $(\sin(tB)\, B^{-b})\,(B^{b-1}\Psi)$, and noting that if b is sufficiently large, $\sin(tB)\, B^{-b}$ is convolution with a function $G_{t,b}(x)$, it follows, using the Hausdorff-Young inequality, that

$$\|\Phi(t)\|_r \leqslant \|\Phi_0(t)\|_r + g\int_{-\infty}^{t} \|G_{t-s,\,b}\|_q \,\|B^{b-1}\Phi(s)^p\|_{q'}\, ds$$

for certain q and q'. Now $\|\Phi(t)\|_r$ and $\|G_{t-s,b}\|_q$ are norms of free solutions, and may be relatively sharply majorized (explicit bounds are given in [7]; another variant of Brodsky's approach is indicated in [8d, Remark 1]); inequalities of Sobolev type will bound $\|B^{b-1}\Phi(s)^p\|_{q'}$ in terms of $\|B^c\Phi(s)^p\|_{q''}$ for an integer c; $\|B^c\Phi(s)^p\|_{q''}$ may now be bounded in terms of the norms of derivatives of $\Phi(s)^p$, e.g. $\|B\Phi(s)^p\|_{q''} \leqslant \text{const}\,(\|\Phi(s)^p\|_{q''} + \|\Phi'(s)\,\Phi(s)^{p-1}\|_{q''})$; Hölder's inequality now bounds the last expression by $E(s)^d\,\|\Phi(s)\|_r^l$, where $E(s)$ is the energy

at time s, and d, e, f are certain constants. The free parameters may be adjusted in a number of ways to arrive at an inequality of the form

$$\| \Phi(t) \|_r \leqslant N(\varphi_0)(1+|t|)^{-e'} + gE(\varphi_0)^{d'} C \int_{-\infty}^{t} (1+|t-s|)^{-e''} \| \Phi(s) \|_s^f ;$$

where N is a certain norm on the free solution at time $-\infty$, $E(\varphi_0)$ is the energy of the free solution, and C is a constant.

An estimate adequate for the proof of the indicated theorem may be obtained by taking $r = \infty$, $b = 2$, and $q > 4$; with these values, $e'' < 1$ and $f > 1$, corresponding to a bound which is in itself insufficient, but for sufficiently small g, and in combination with related estimates used in the proof of Theorem 3, it follows that $\varphi(x, t) = O(|t|^{-1+\varepsilon})$ uniformly in x, for every $\varepsilon > 0$, implying the convergence of the integrals in question. It is likely that further application of the same method will show that $\varphi(x, t) = O(|t|^{-\frac{3}{2}+\varepsilon})$, which is virtually best possible.

Phase space structure of the solution manifold

The solution manifold of a hyperbolic equation may be regarded physically as a phase space which differs from conventional phase spaces: fundamentally only in being infinite-dimensional. This analogy may be made mathematically explicit through the introduction in the solution manifolds of suitable equations of the structures which are well known in the finite-dimensional case. Fundamental among these is the symplectic structure in the solution manifold of a second-order hyperbolic equation; this is relatively simply described in the case of equation (2).

Theorem 6. *The manifold of all finite-energy solutions of equation* (2) *is of class* C^∞ *(as a Frechet-Banach manifold) provided* J *is of class* C^∞, *and its tangent vectors correspond naturally to solutions of the first-order variational equation,* $\Psi''(t) + B^2\Psi(t) = (\partial_{\Phi(t)} J)\,\Psi(t)$.

On defining, for the tangent vectors represented by the solutions $\Psi_1(\cdot)$ *and* $\Psi_2(\cdot)$ *of the foregoing equation, the form* $\Omega_{\Phi(\cdot)}$ *by the equation*

$$\Omega_{\Phi(\cdot)} = \langle \Psi_1(t),\ \Psi_2'(t) \rangle - \langle \Psi_1'(t),\ \Psi_2(t) \rangle,$$

Ω *is a closed, non-degenerate, differential form of class* C^∞ *which is invariant under the one-parameter group of transformations on the solution manifold defined by the differential equation.*

In the case of a relativistic equation (e.g. that of Theorem 3), Ω *is invariant under the Poincaré group.*

In general the form of Ω is more complicated, although similar results are valid. This is illustrated in the treatment by A. Lichnerowicz of the quantization of linear relativistic local equations in curved space-time manifolds; the field commutator distribution $D(x, x')$ is closely related to the form Ω, and is the difference of the advanced and retarded elementary solutions provided by the general theory of Leray. The relation to the given expression for Ω is visible from the alternative, non-relativistic, definition of $D(x, x')$ as the solution of the differential equation in question such that relative to a particular Lorentz frame, it has Cauchy data:

$$D(x, x')|_{t=t'} = 0, \qquad \partial_t D(x, x') = \delta(\vec{x} - \vec{x}') \qquad (x = (\vec{x}, t)).$$

The theory may in part be extended to non-linear equations, as may be illustrated by the case of the scalar equation $\Box\varphi = F(\varphi)$ on a Lorentzian space-time manifold M, where F is a given C^∞ function. A tangent vector to the solution manifold at the solution φ may be identified with a solution λ of the first-order variational equation, $\Box\lambda = F'(\varphi)\,\lambda$. According to a result of Y. Fourès-Bruhat [2], every suitably regular such function λ is of the form $\lambda(x) = \int D_\varphi(x, x') \times f(x')\,dx'$ for some infinitely differentiable function f of compact support on M, where $D_\varphi(x, x')$ is the commutator distribution for the indicated tangential equation. The form

$$\Omega_\varphi(\lambda, \lambda') = \int\int D_\varphi(x, x')\, f(x)\, f'(x')\, dx\, dx'$$

depends only on λ, λ', and φ, and not at all on the choices of f and f'. In part extending and rigorizing developments indicated in [8g], it follows from a study of the Frechet differential $\delta_\varphi D_\varphi(x, x')$, which may be expressed in terms of $D_\varphi(x, x')$, that

Theorem 7. Ω *is a closed, non-degenerate,* C^∞, *second-order differential form on the manifold of all local solutions of the equation* $\Box\varphi = F(\varphi)$, *in the vicinity of any fixed point on the given Lorentzian manifold.*

The vector field X_f on the solution manifold which assigns at the solution φ the tangent vector λ given above represents an integral, over the times involved in the support of f, of the vector fields corresponding to infinitesimal vector displacements of the Cauchy data at each time. Relatively explicit expressions involving the X_f, such as the commutator $[X_f, X_g]$ for any two functions f and g of the indicated type, may be given in terms of the $D(x, x')$ functions above; these provide an algebraic interpretation for domains of influence, etc.; e.g. two points P and Q are outside each others' domains of influence if and

only if $[X_f, X_g] = 0$ whenever the supports of f and g are contained in sufficiently small neighborhoods of P and Q respectively.

Another important structure in a phase space is an invariant integral. In a phase space of finite dimension $2n$, Ω^n provides such an integral; when $n = \infty$, Ω^n has no meaning; however, for an important class of linear hyperbolic equations (e.g. relativistic wave equations in Minkowski space) there is a natural canonical measure, of a generalized nature (a "weak distribution" as defined in [8j]); this measure is temporally invariant, and the corresponding flow is ergodic and in fact mixing. In the case of a non-linear equation, no temporally invariant measure in the solution manifold can be given explicity, but if the wave operator exists and is regular in a sufficiently strong sense, it will induce according to results of L. Gross [3] a transformation of an invariant weak (probability) distribution on the free solution manifold into an invariant weak distribution in the solution manifold of the non-linear equation. This amounts to the construction of a (generalized) stationary stochastic solution to the non-linear equation which is asymptotic at time $-\infty$ to a given stationary stochastic solution of the associated linear equation.

Stochastic and quantized equations

Stochastic equations are those for which the unknown function $\varphi(\vec{x}, t)$ is for each time t a (possibly generalized) random-variable-valued function on space; quantized equations are those in which $\varphi(x, t)$ is similarly (possibly generalized) operator-valued, and indeed it is commonly postulated that the commutator between the Cauchy data at two points x and x' on a space-like surface is $iD(x, x')$, and so can not vanish. Such equations are in general, when non-linear, not a priori mathematically meaningful, since they involve non-linear functions of generalized (i.e. weak) functions; indeed this is always the case for quantized equations satisfying commutation relations of the indicated nature.

Weak solutions of stochastic non-linear equations may be obtained by the method just indicated, when the non-linear term is sufficiently regular, but the question remains open (even in the case of linear equations, in part) of whether these solutions are strong, in the sense that with probability one suitably differentiable versions of the generalized solutions obtained exist satisfying the given differential equation in the classical sense. When the non-linear term is irregular relative to the initial probability distribution in function space,—e.g. if $\varphi(\vec{x}, 0)$ is not well-defined at almost all points $\vec{x}$, but a local non-linear function such as $\varphi(x)^p$ is involved, one is in the indicated mathematically ambiguous situation; however, there exist natural definitions, connected with the Wick products arising in the case of quantized fields.

Two basically different interpretations of such quantized equations have been treated mathematically (exclusive of conventional perturbation theory as practiced in theoretical physics, which there is reason to doubt can be given comprehensive mathematical meaning). The first of these interprets polynomials in $\varphi(\vec{x}, t)$ at a fixed time t in terms of so-called Wick products, either in their conventional form applicable to the "free field", or in a generalized form based on an intrinsic characterization ([8e, 8k]). The Cauchy problem, for the quantized equation can then be made mathematically well-defined. There are, analytically speaking, two major questions in addition to the solution of this problem: (a) the existence (and uniqueness) of a regular positive linear functional E on the operator algebra $\mathbf{A}$ generated by the (bounded functions of) the field operators, such that for any element $A \in \mathbf{A}$, $E(A^t A^*)$ has a non-negative frequency spectrum, as a function of the time t, where A^t denotes the temporal displacement of the operator A through the time t (such a functional E is called a physical vacuum); (b) the existence, and nature (especially, relation to symmetric or skew-symmetric tensor algebras over the Hilbert space of free (classical) solutions, i.e. solutions of the first-order variational equation at the solution $\varphi = 0$), of asymptotic linear fields as $t \to \pm\infty$. While a substantial variety of results in these directions now exist, especially in the case of two space-time dimensions, in which case the free-field Wick products at a sharp time are strictly operator-valued, crucial aspects of these problems remain unresolved.

The second interpretation is concerned primarily with the propagation of the quantized field from time $-\infty$ to $+\infty$, rather from one finite time to another, i.e. with the dispersion of the field. The (linear) quantum-theoretic dispersion operator is taken as the induced action of the (non-linear) dispersion operator S_e previously considered, acting on the classical solution manifold of the given equation, on a certain implicitly-defined S_e-invariant class of functionals over the solution manifold. In the simplest case, this class consists of the holomorphic functionals relative to a complex structure on the solution manifold which is determined by the condition that it be invariant under S_e and extend the symplectic structure described earlier to a Kählerian one. For a quantized field interacting with an external source or potential, this formalism gives results equivalent to the more conventional, first-indicated formalism; unlike this conventional formalism, it is adaptable to the association of a quantum field with a suitably structured infinite manifold which is not necessarily defined by a partial differential equation.

Massachusetts Institute of Technology, USA

44—1220

REFERENCES

[1] Brodsky A. R., Asymptotic decay of solutions to the relativistic wave equation and the existence of scattering for certain non-linear hyperbolic equations, Ph. D. Thesis, M.I.T., Cambridge, Mass., 1964.

[2] Fourès-Bruhat Y., Propagateurs et solutions d'équations homogènes hyperboliques, *C. R. Acad. Sci. Paris*, **251** (1960), 29-31.

[3] Gross L., Integration and non-linear transformations in Hilbert space, *Trans. Amer. Math. Soc.*, **105** (1960), 404-440.

[4] Jörgens K., Das Anfangswertproblem im Grossen für eine Klasse nichtlinearer Wellengleichungen, *Math. Zeits.*, **77** (1961), 295-307.

[5] Leray J., Hyperbolic partial differential equations, Institute for Adv. Study, Princeton, N.J., U.S.A., 1951-52.

[6] Lichnerowicz A., (a) Theorie quantique des champs sur un espace-temps courbe. Cours de l'École d'Été de Physique théorique des Houches (France), 1963;

(b) Propagateurs et commutateurs en relativite generale, *Inst. Hautes Études Sci. Publ. Math.*, **10** (1961), 1-56;

(c) Champs spinoriels et propagateurs en relativité générale, *Bull. Soc. Math. France*.

[7] Nelson S., Asymptotic behavior of certain fundamental solutions to the Klein-Gordon equation, Ph. D. Thesis, M.I.T., Cambridge, Mass., 1966.

[8] Segal I., (a) Non-linear semi-groups, *Ann. Math.*, **78** (1963), 339-364;

(b) The global Cauchy problem for a relativistic scalar field with power interaction, *Bull. Soc. Math. France*, **91** (1963), 129-135;

(c) Differential operators in the manifold of solutions of a non-linear differential equation, *Jour. Math. pur. appl.*, **44** (1965), 71-132;

(d) Quantization and dispersion for non-linear relativistic equations. Proc. Conf. on Math. Th. El. Particles, M. I. T. Press, Cambridge, Mass., 1966, 79-108;

(e) Interpretation et solution d'équations non linéaires quantifiées, *C. R. Acad. Sci. Paris*, **259** (1964), 301-303, sec. 1-3 (heuristic).

(f) Conjugacy to unitary groups within the infinite-dimensional symplectic group, Argonne Nat. Lab. report ANL-7216, 1966, 1-11;

(g) Quantization of non-linear systems, *Jour. Math. Phys.*, **1** (1960), 468-488, sec. 4A-B (heuristic);

(h) Explicit formal construction of nonlinear quantum fields, *Jour. Math. Phys.*, **5** (1964), 269-282;

(i) La variété des solutions d'une équation hyperbolique, non linéaire, d'ordre 2, C.I.M.E. lectures on non-linear partial differential equations at Varenna, 1964;

(j) Abstract probability spaces and a theorem of Kolmogoroff, *Amer. Jour. Math.*, **76** (1954), 721-732;

(k) Non-linear functions of random distributions and generalized normal products. Proc. Conf. on Funct. Integration and Constr. Quant. Fld. Theory, M.I.T., Cambridge, Mass., April, 1966.

[9] Strauss W., (a) Les operateurs d'onde pour les equations d'onde non-linéaires independantes du temps, *C.R. Acad. Sci. Paris*, **256** (1963), 5045-5046;

(b) La deccroissance asymptotique des solutions des equations d'onde non-lineaires, *C.R. Acad. Sci. Paris*, **256** (1963), 2749-2750.

Секция 13. Section 13

ЭКСТРАПОЛЯЦИОННЫЕ ЗАДАЧИ АВТОМАТИЧЕСКОГО УПРАВЛЕНИЯ И МЕТОД ПОТЕНЦИАЛЬНЫХ ФУНКЦИЙ

М. А. АЙЗЕРМАН, Э. М. БРАВЕРМАН, Л. И. РОЗОНОЭР

1. Введение

При проектировании и исследовании так называемых самонастраивающихся и обучающихся систем (в частности, обучающихся распознаванию образов) возникают проблемы, которые можно понимать как проблемы экстраполяции и интерполяции функций. В задачах такого рода часто оказывается возможным ввести пространство X (определяемое условиями конкретной задачи) и некоторую функцию $f(x)$, заданную на X так, что процесс обучения можно интерпретировать как появление точек $x^1, x^2, \ldots, x^n, \ldots$ из X с одновременным сообщением некоторой информации о значениях $f(x)$ в этих точках. Результатом обучения при этом является построение функции, в том или ином смысле близкой к $f(x)$.

Обычные методы экстраполяции часто оказываются практически неприменимыми для решения таких задач главным образом потому, что точки $x^1, \ldots, x^n, \ldots$ не могут быть выбраны по нашему желанию, а появляются независимо от нас некоторым нерегулярным образом (например, в соответствии с неизвестным заранее вероятностным законом). Кроме того, пространство X часто имеет высокую размерность, и это также затрудняет применение обычных методов экстраполяции. Наконец, сообщаемая при обучении информация о значениях функции $f(x)$ в точках x^i может быть неполной (например, может сообщаться лишь знак $f(x^i)$ или значение $f(x^i)$ вместе с помехой) — в этом случае близость приближающей функции к $f(x)$ понимается в соответствующем смысле (например, в смысле совпадения знаков или в смысле сходимости по вероятности). Далее в этом докладе приводятся примеры точных постановок задач такого рода.

Общая схема постановки и решения таких задач имеет следующий вид. Пусть $\varphi_i(x)$ $(i = 1, 2, \ldots)$ — некоторая полная (не обязательно ортонормированная) система функций, заданных на X.

Тогда восстанавливаемая функция $f(x)$ представима разложением [1])

$$f(x)=\sum_{i=1}^{\infty} c_i\varphi_i(x). \tag{1}$$

Разумеется, коэффициенты c_i заранее неизвестны. Предполагается, что функция $f(x)$ —«достаточно гладкая», т. е. что коэффициенты c_i убывают достаточно быстро. Это предположение будет далее уточнено.

В рассматриваемых задачах точки $x^1, \ldots, x^n, \ldots$ появляются последовательно, шаг за шагом, и независимо в соответствии с неизвестной плотностью вероятности $P(x)$. Каждый такой шаг называется «показом». Цель состоит в построении рекуррентного процесса, последовательно приближающего в нужном смысле функцию $f(x)$ с ростом числа показов на подмножестве из X, где $P(x)>0$.

Для решения формулируемых далее задач используется один и тот же метод, названный авторами методом потенциальных функций [2]). Вводится в рассмотрение функция двух переменных («потенциальная функция»)

$$K(x, y)=\sum_{i=1}^{\infty} \lambda_i^2\varphi_i(x)\varphi_i(y), \tag{2}$$

где действительные числа λ_i выбраны так, что функция $K(x, y)$ ограничена. При n-м показе строится n-е приближение функции $f(x)$ с помощью рекуррентного соотношения [3])

$$f_n(x)=f_{n-1}(x)+r_nK(x, x^n); \quad f_0(x)\equiv 0. \tag{3}$$

Здесь r_n определяется информацией об $f(x^n)$, полученной при n-м показе, и значением $f_{n-1}(x^n)$. Применение рекуррентной процедуры (3) в различных задачах отличается лишь конкретным способом выбора r_n. В любом случае выбор r_n таков, чтобы $f_n(x^n)$ приближало $f(x^n)$ лучше, чем $f_{n-1}(x^n)$, т. е. чтобы добавление члена $r_nK(x, x^n)$ улучшало приближение в точке, выбранной при очередном показе. Разумеется, при этом приближение функции в других точках (в том числе и в показанных ранее) может ухудшаться, но, несмотря на это, для всех рассматриваемых задач приводятся теоремы, устанавливающие сходимость процесса и указывающие каждый раз, в каком смысле этот процесс сходится. Более того, в ряде случаев удается оценить быстроту сходимости. Доказательство сходимости процедуры опирается на предположение о «достаточной гладкости» восстанавливаемой функции, о котором выше шла речь. Мы можем теперь уточнить это предположение следующим образом:

$$\sum_{i=1}^{\infty}\left(\frac{c_i}{\lambda_i}\right)^2<\infty. \tag{4}$$

Особый интерес представляет случай, когда можно предположить, что в разложении (1) содержится лишь конечное число N членов [1]):

$$f(x)=\sum_{i=1}^{N} c_i\varphi_i(x). \tag{5}$$

В этом случае условие (4) заведомо выполнено, если $\lambda_i\neq 0$ $(i=1, 2, \ldots, N)$. Кроме того, если выполнено условие (5), то потенциальную функцию можно выбирать в виде

$$K(x, y)=\sum_{i=1}^{N} \lambda_i^2\varphi_i(x)\varphi_i(y). \tag{6}$$

Для получения оценок сходимости делаются дополнительные предположения, которые будут сформулированы далее.

При практическом использовании метода потенциальных функций нет необходимости задавать систему функций $\varphi_i(x)$, а можно непосредственно фиксировать вид $K(x, y)$, причем этот вид в значительной степени произволен. Подробнее об этом будет идти речь в § 6. Все теоремы в настоящем докладе приводятся без доказательств.

Доказательства теорем о сходимости предлагаемых процедур опубликованы в [1—4]. В работах [2, 3] предполагалось, что система

---

[1]) При небольших и очевидных видоизменениях предлагаемые ниже построения пригодны и в случае, когда $f(x)$ представима интегралом

$$f(x)=\int c_\omega\varphi_\omega(x)\,d\omega.$$

Для простоты изложения в докладе везде предполагается, что имеет место представление (1).

[2]) Как показал Я. З. Цыпкин [7], задачи, решаемые этим методом, близки по своему характеру к задачам теории стохастической аппроксимации и, в частности, некоторые из приводимых ниже процедур сводятся к процессу Роббинса — Монро. Взаимоотношение метода потенциальных функций и теории стохастической аппроксимации в настоящем докладе не обсуждается.

[3]) Для решения некоторых задач методом потенциальных функций (см. § 5) используется более общее, нежели (3), рекуррентное соотношение

$$f_n(x)=q_nf_{n-1}(x)+r_nK(x, x^n).$$

[1]) Если пространство X таково, что полная система содержит конечное число M функций $\varphi_i(x)$, то предположение (5) нетривиально лишь тогда, когда $N<M$. Описываемые далее процедуры практически эффективны, если можно предположить, что $N\ll M$.

функций $\varphi_i(x)$ ортонормирована, а ряд (1) не содержит всех функций полной системы. Это ограничение снимается в [5]. Доказательства теорем, касающихся оценок скорости сходимости алгоритмов, даны в работе [6].

2. Идентификация статического объекта

При восстановлении статических характеристик объекта возникает задача об экстраполяции функции $y = f(x)$ по ее значениям в случайно наблюдаемых точках. Ниже рассматриваются две постановки этой задачи, отличающиеся учетом помех.

Восстановление функции при отсутствии помех. В целочисленные моменты времени $1, \ldots, n, \ldots$ наблюдаются значения векторов «входа» $x^1, \ldots, x^n, \ldots$, появляющиеся случайно и независимо с плотностью вероятности $P(x)$, и скалярные значения «выхода» $y^1, \ldots, y^n, \ldots$, где $y^n = f(x^n)$. Процедуры восстановления $y = f(x)$ основаны на предположениях (1), (4) и определяются формулами (2) и (3), где в данном случае полагается

$$r_n = \gamma_n \operatorname{sign}[y^n - f_{n-1}(x^n)] \tag{7}$$

или

$$r_n = \frac{1}{\Lambda}[y^n - f_{n-1}(x^n)]. \tag{8}$$

Здесь $\gamma_1, \ldots, \gamma_n, \ldots$ — произвольная последовательность положительных чисел, такая, что ряд $\sum_{n=1}^{\infty}\gamma_n$ расходится, а ряд $\sum_{n=1}^{\infty}\gamma_n^2$ сходится:

$$\sum_{n=1}^{\infty}\gamma_n = \infty, \quad \sum_{n=1}^{\infty}\gamma_n^2 < \infty; \tag{9}$$

Λ — произвольная положительная константа, удовлетворяющая лишь условию

$$\Lambda > \frac{1}{2}\max_{x\in X} K(x, x). \tag{10}$$

При таком выборе γ_n и Λ имеет место следующая

Теорема 1. *Пусть выполнено условие* (4). *Тогда последовательность функций* $f(x)$, *определяемая рекуррентной процедурой* (2), (3), *такова* [1]), *что при* $n \to \infty$

$$\int_X |f(x) - f_n(x)|\, P(x)\, dx \xrightarrow{\text{п. н.}} 0, \tag{11}$$

[1]) Символ $\xrightarrow{\text{п. н.}}$ в тексте теоремы 1 и далее означает сходимость «почти наверное» (с вероятностью единица).

если числа r_n *выбираются в соответствии с формулой* (7), *и*

$$\int_X [f(x) - f_n(x)]^2 P(x)\, dx \xrightarrow{\text{п. н.}} 0, \tag{12}$$

если числа r_n *выбираются в соответствии с формулой* (8).

Восстановление функции при учете помех. В постановке задачи, рассматриваемой выше, предполагалось, что при каждом показе в случайно выбранной точке x^n известно точное значение $y^n = f(x^n)$. Теперь предположим, что результат измерений содержит меньшую информацию о значениях функции; именно при каждом показе известно не y^n, а $\tilde{y}^n = f(x^n) + \xi^n$, где ξ^n — случайная величина (помеха), удовлетворяющая следующим условиям: значения ξ^n при разных n независимы, условная плотность вероятности $W(\xi^n \mid x^n)$ не зависит от n и условное математическое ожидание величины ξ^n при условии x^n равно 0, а ее дисперсия ограничена при всех x^n одной и той же константой. Задача состоит в приближении функции $y = f(x)$, по-прежнему удовлетворяющей предположению (1). В этом случае коэффициенты r_n в (3) предлагается выбирать по формуле

$$r_n = \gamma_n[\tilde{y}^n - f_{n-1}(x^n)], \tag{13}$$

где последовательность γ_n удовлетворяет условию (9). Тогда имеет место

Теорема 2. *Пусть выполнено условие* (4). *Тогда последовательность функций* $f_n(x)$, *определяемых рекуррентной процедурой* (2), (3), *где* r_n *выбирается в соответствии с формулой* (13), *такова, что* [1])

$$\int_X [f_n(x) - f(x)]^2 P(x)\, dx \xrightarrow{p} 0. \tag{14}$$

Теоремы 1 и 2 показывают, что последовательность $f_n(x)$ приближает $f(x)$ при достаточно больших n в тех точках $x \in X$, где $P(x) > 0$.

Скорость сходимости. Переходя к оценкам скорости сходимости процедур этого параграфа, будем предполагать, что для восстанавливаемой функции $f(x)$ выполняется условие (5), а потенциальная функция задается в форме (6). Кроме того, будем предполагать, что для любой функции $F(x)$, представимой конечным рядом по системе функций $\varphi_i(x)$,

$$F(x) = \sum_{i=1}^{N}\mu_i\varphi_i(x), \tag{15}$$

[1]) Символ $\xrightarrow{p}$ в тексте теоремы 2 означает сходимость по вероятности.

где $\mu_1, \ldots, \mu_N$ не равны одновременно нулю, выполняется неравенство [1])

$$\int_X F^2(x) P(x)\, dx > 0. \tag{16}$$

Из (15) и (16) следует, что матрица с элементами

$$s_{ik} = \lambda_i \lambda_k \int \varphi_i(x) \varphi_k(x) P(x)\, dx, \quad \lambda_i > 0, \quad i, k = 1, \ldots, N, \tag{17}$$

положительно определена.

Прежде чем перейти к оценкам скорости сходимости процедур настоящего параграфа, сделаем замечание, касающееся самого факта сходимости. Если помимо условий теоремы 2 дополнительно предполагается, что выполнены условия (15) и (16), то утверждение теоремы 2 может быть усилено. Именно, в этом случае имеет место не только сходимость к нулю по вероятности интеграла в выражении (14), но и сходимость к нулю этого интеграла «почти наверное».

Возвращаясь теперь к вопросу об оценке скорости сходимости процедур, обозначим через r минимальное характеристическое число матрицы (17); это число положительно в силу положительной определенности матрицы (17).

Оценки скорости сходимости приведенных выше процедур устанавливаются следующими теоремами:

Теорема 3. *При использовании процедуры* (3), (6), (8) *математические ожидания интеграла*

$$\beta_i = \int_X [f(x) - f_i(x)]^2 P(x)\, dx$$

удовлетворяют неравенству

$$M\{\beta_{n+1}\} \leqslant c(1 - ra)^n,$$

где c и a — положительные константы и $1 - ra > 0$.

При оценке скорости сходимости двух других процедур этого параграфа существенную роль играет возможность выбора положительного числа λ, удовлетворяющего неравенствам

$$\left(\frac{\gamma_n}{\gamma_{n+1}}\right)^{\lambda} (1 - r\gamma_{n+1}) \leqslant 1, \quad \sum_{i=1}^{\infty} \gamma_i^{2-\lambda} < \infty. \tag{18}$$

Теорема 4. *Если для любого $r > 0$ существует $\lambda(r) > 0$, удовлетворяющее неравенствам* (18), *то найдется такое λ^*, что при использовании процедуры* (3), (6), (7) *математическое ожидание интеграла*

$$\beta_i = \int |f(x) - f_i(x)| P(x)\, dx$$

удовлетворяет неравенству

$$M\{\beta_n\} \leqslant c\gamma_n^{\lambda^*/2},$$

где $c > 0$ — константа.

Если, кроме того, $\lambda(r) \equiv \lambda_0$ и фактически от r не зависит, то можно принять $\lambda^* = \lambda_0$.

Теорема 5. *Если существует $\lambda > 0$, удовлетворяющее неравенствам* (18), *то при использовании процедуры* (3), (6), (13) *математическое ожидание интеграла*

$$\beta_i = \int [f(x) - f_i(x)]^2 P(x)\, dx$$

удовлетворяет неравенству

$$M\{\beta_n\} \leqslant c\gamma_n^{\lambda},$$

где $c > 0$ — константа.

В силу теоремы 3 скорость сходимости процедуры (3), (6), (8) такова же, как и у геометрической прогрессии; сходимость же в остальных двух случаях зависит от выбора γ_n. Так, например, если положить $\gamma_n = 1/n$, то число λ, удовлетворяющее (18), можно выбрать так:

$$\lambda = \min\{1, r\}.$$

Если $\gamma_n = 1/n^{1-\varepsilon}$ $(0 < \varepsilon < 1/2)$, то число λ, удовлетворяющее (18), можно принять равным $\lambda = (1 - 2\varepsilon)/(1 - \varepsilon)$, т. е. не зависящим от r; и в том и в другом случае выбора γ_n теоремы 4 и 5 обеспечивают степенную сходимость соответствующих процедур.

3. Обучение машины распознаванию образов (детерминистская постановка задачи)

В настоящем параграфе имеется в виду следующая задача об обучении машины распознаванию образов. На вход машины последовательно поступают объекты, принадлежащие одному из двух классов A или B, причем множества A и B не пересекаются. В процессе обучения о каждом объекте сообщается, к какому классу он относится. Требуется, чтобы после достаточно большого числа показов машина правильно опознавала новые объекты, не появлявшиеся в процессе обучения.

[1]) Если функции $\varphi_i(x)$ и $P(x)$ непрерывны, то условие (16) эквивалентно требованию линейной независимости системы функций $\varphi_i(x)$ на множестве, где $P(x) > 0$. В этом случае, в частности, неравенство (16) удовлетворяется, если система $\varphi_i(x)$ линейно независима на X и $P(x) > 0$ везде на X.

Понимаемая так задача обучения машины распознаванию образов сводится к следующей экстраполяционной задаче. Каждому показываемому объекту ставится в соответствие точка некоторого пространства X. В пространстве X существуют два непересекающихся множества A и B и, следовательно, существует «разделяющая» функция $f(x)$, принимающая на A и B значения противоположных знаков. В дальнейшем используется более сильное предположение относительно разделяющей функции, а именно:

$$f(x)\begin{cases} >\varepsilon & \text{при } x\in A, \\ <-\varepsilon & \text{при } x\in B, \end{cases} \tag{19}$$

где $\varepsilon > 0$.

В процессе обучения случайно и независимо с неизвестной заранее плотностью вероятности $P(x)$ появляются точки $x^1, \ldots, x^n$ и сообщаются знаки $f(x^n)$ в этих точках. Требуется построить приближение к функции $\operatorname{sign} f(x)$ при $x \in A \cup B$ [1]. Для построения последовательности $f_n(x)$ в данном случае вновь используется процедура (2), (3), причем r_n задается формулой [2])

$$r_n = \frac{1}{2}\left[\operatorname{sign} f(x^n) - \operatorname{sign} f_{n-1}(x^n)\right], \tag{20}$$

а приближением к $\operatorname{sign} f(x)$ является $\operatorname{sign} f_{n-1}(x)$.

Теорема 6. *Пусть выполнено условие* (4). *Тогда последовательность функций* $f_n(x)$, *определяемых рекуррентной процедурой* (2), (3), *где* r_n *выбирается в соответствии с* (20), *такова, что*

$$\int \left|\operatorname{sign} f_n(x) - \operatorname{sign} f(x)\right| P(x)\,dx \xrightarrow{\text{п. н.}} 0. \tag{21}$$

Из выражения (21) видно, что по крайней мере там, где $P(x) > 0$, мера множества точек, где знаки функций $f_n(x)$ и $f(x)$ не совпадают, стремится к нулю с ростом n. Поэтому если потребовать строгую положительность $P(x)$ на $A \cup B$, то использование процедуры (2), (3), (20) гарантирует при $n \to \infty$ сходимость получаемой последовательности функций к разделяющей. Хотя теорема 6 и устанавливает сходимость процедуры при $n \to \infty$, однако в рассматриваемой задаче о распознавании образов имеет место более сильный результат, а именно сходимость в известном смысле за конечное число показов. Этот факт устанавливает

Теорема 7. *Пусть выполнено условие* (4) *и, кроме того, статистика показа удовлетворяет следующему условию: если к* q*-му показу (каково бы ни было* q*) не произошло еще полного разделения множеств* A *и* B, *то во время* q*-го показа существует строго положительная вероятность появления такой точки* x^q, *чтобы произошло исправление ошибки. Тогда с вероятностью единица найдется такое число* k (*может быть, свое для каждой конкретной реализации процесса*), *что* $\operatorname{sign} f_k(x) = \operatorname{sign} f(x)$ *для* $x \in A \cup B$.

Заметим, что требование, предъявляемое в теореме 7 к статистике показа, является естественным ослаблением требования строгой положительности $P(x)$ на $A \cup B$, которое упоминалось выше. Хотя при детерминистской постановке задачи о распознавании образов пока не удалось установить оценки скорости сходимости процедуры (2), (3), (20) по числу показов, тем не менее может быть дана оценка числа исправлений ошибок, после которого происходит разделение миожеств. Такую оценку устанавливает теорема 8 [1]).

Теорема 8. *Пусть* S — *произвольная бесконечная последовательность точек пространства* X *из* $A \cup B$, *показываемых в процессе обучения. Тогда число исправлений ошибок не превосходит некоторого конечного числа* m, *не зависящего от выбора* S *и такого, что*

$$m \leqslant \frac{\sup\limits_{x\in A\cup B} K(x, x)}{\inf\limits_{x\in A\cup B} f^2(x)} \sum_{i=1}^{\infty}\left(\frac{c_i^2}{\lambda_i^2}\right). \tag{22}$$

Из теоремы 8 следует, что при любом выборе последовательности S показываемых точек обучение закончится за конечное число показов и начиная с некоторого k-го показа весь бесконечный «хвост» последовательности S будет правильно относиться к A и B по знаку функции $f_k(x)$, построенной к k-му показу.

Хотя теорема 7 доказывает сходимость процесса за конечное число шагов, при практическом использовании метода не ясно, когда следует закончить «процесс обучения», т. е. начиная с какого

[1]) Если существует одна разделяющая функция, то их существует и бесконечно много, так как на $A \cup B$ существен лишь их знак, а вне $A \cup B$ их значения вообще несущественны.

[2]) Коэффициент r_n, вычисляемый по формуле (20), может принимать лишь значения 0, $+1$ и -1. Если $r_n = 0$, то $\operatorname{sign} f(x^n) = \operatorname{sign} f_n(x^n)$, машина «не ошибается» в точке x^n и в силу (3) разделяющая функция $f_n(x)$ на n-м шаге не меняется. Если же $\operatorname{sign} f(x^n) \neq \operatorname{sign} f_n(x^n)$, то машина «ошибается» в точке x^n и на n-м шаге происходит изменение $f_n(x)$ в силу (3); при этом говорят, что произошло «исправление ошибки».

[1]) Эта теорема является непосредственным обобщением теоремы А. Новикова [8].

показа можно считать, что построенная функция разделяет множества A и B.

Однако можно ввести следующее условие окончания процесса обучения. Обучение считается законченным, если после S исправлений ошибок в течение последующих $L_S = L_0 + S$ показов нового исправления ошибки не происходит. Обозначим через $\Pr(p < \varepsilon)$ вероятность того, что после окончания процесса обучения вероятность p ошибки при классификации новых точек будет меньше ε. Тогда выбор константы L_0 определяется следующей теоремой:

Теорема 9. *Для любых положительных* ε, δ *обеспечивается выполнение неравенства*

$$\Pr(p < \varepsilon) > 1 - \delta,$$

если L_0 *удовлетворяет условию*

$$L_0 > \frac{\ln \varepsilon\delta}{\ln(1-\varepsilon)}. \qquad (23)$$

4. Обучение машины распознаванию образов (вероятностная постановка задачи)

В настоящем параграфе в отличие от предыдущего предполагается, что каждая точка $x \in X$ не может быть с достоверностью отнесена к классу A или B, но вместе с тем существуют вероятности («степени достоверности») $D_A(x)$ и $D_B(x) = 1 - D_A(x)$ того, что эта точка x принадлежит A или B соответственно. Степени достоверности $D_A(x)$ и $D_B(x)$ неизвестны заранее, но на каждом n-м шаге при появлении точки x^n она относится к множеству A или B в соответствии с этими вероятностями [1]. Задача состоит в восстановлении функций $D_A(x)$ и $D_B(x)$.

Эта задача в принципе может быть решена с помощью формулы Байеса. Именно, по статистике появляющихся точек могут быть восстановлены условные плотности вероятности $P(x \mid A)$ и $P(x \mid B)$ появления в X точек из A и B соответственно [2], а затем по формуле Байеса могут быть вычислены апостериорные вероятности $D_A(x)$ и $D_B(x)$. Недостаток такого пути состоит в том, что функции $P(x \mid A)$ и $P(x \mid B)$ могут быть значительно более сложными для

[1] Следует иметь в виду, что при повторном появлении точки, ранее отнесенной к A, она может быть отнесена и к B, но при многократном показе этой же точки она будет относиться к A или к B в соответствии с объективно существующими, но неизвестными заранее вероятностями $D_A(x)$ и $D_B(x)$.

[2] В [2] показано, что задача восстановления неизвестной заранее плотности вероятности может быть решена также с помощью метода потенциальных функций.

приближения, чем функции $D_A(x)$ и $D_B(x)$. Более того, функции $D_A(x)$ и $D_B(x)$ могут существовать и в том случае, когда функции $P(x \mid A)$ и $P(x \mid B)$ не существуют. Нас в этой работе интересует другой путь решения задачи, связанный с непосредственным восстановлением степеней достоверности $D_A(x)$ и $D_B(x)$, минуя предварительное восстановление вероятностей $P(x \mid A)$ и $P(x \mid B)$ и использование формулы Байеса.

Будем предполагать, что для $f(x) \equiv D_A(x)$ по-прежнему справедливо разложение (1) и условие (4).

Ниже предлагается два способа решения поставленной задачи.

Первый способ заключается в формальном сведении задачи к задаче об идентификации объекта при наличии помех. Действительно, рассмотрим случайную функцию $\psi(x)$, принимающую в точке $x \in X$ значение 1, если x отнесено к A, и значение 0, если x отнесено к B. Тогда функция $\psi(x)$ принимает в точке x значение 1 с вероятностью $f(x) = D_A(x)$ и значение 0 с вероятностью $1 - f(x)$. Математическое ожидание функции $\psi(x)$ равно $f(x)$, и при каждом показе значение случайной функции $\psi(x)$ может быть представлено в виде

$$\tilde{y}^n = \psi(x^n) = f(x^n) + \xi^n,$$

где математическое ожидание ξ^n равно нулю. Рассматривая ξ^n как «помеху» и используя рекуррентную процедуру (2), (3), (13), можно восстановить функцию $f(x) = D_A(x)$ в смысле теоремы 2. При этом сохраняется оценка скорости сходимости процедуры, устанавливаемая теоремой 5.

Другой способ восстановления $D_A(x)$, описанный в [2, 5], также может быть понят как частный случай процедуры (2), (3). Рассмотрим оператор «черта сверху», определив его так:

$$\overline{\Phi}(x) = \begin{cases} 0, & \text{если } -\infty < \Phi(x) < 0, \\ \Phi(x), & \text{если } 0 \leqslant \Phi(x) \leqslant 1, \\ 1, & \text{если } 1 < \Phi(x) < \infty. \end{cases} \qquad (24)$$

В отличие от предыдущих процедур в определении числа r_n используется случайный акт. Пусть к n-му шагу в соответствии с (3) построена функция $f_{n-1}(x)$ и появилась точка x^n. Тогда с вероятностью $\overline{f}_{n-1}(x^n)$ «делается предположение», что x^n относится к A, и с вероятностью $1 - \overline{f}_{n-1}(x^n)$ — что она относится к B. Это предположение сравнивается с тем, к какому множеству (A или B) отнесена точка x^n в действительности. Тогда возникает одна из четырех ситуаций AA, AB, BA, BB (первая буква указывает, к какому множеству отнесена точка в действительности, а вторая —

к какому множеству она отнесена в результате описанного в тексте случайного акта). Число r_n определяется формулой

$$r_n = \begin{cases} 0 & \text{в случае } AA \text{ и } BB, \\ \gamma_n & \text{в случае } AB, \\ -\gamma_n & \text{в случае } BA, \end{cases} \qquad (25)$$

где γ_n — последовательность положительных чисел, удовлетворяющих условию (9). При этом имеет место

Теорема 10. *Последовательность функций $f_n(x)$, определяемых рекуррентной процедурой* (2), (3), (25), *удовлетворяет условию*

$$\int_X [f_n(x) - D_A(x)]^2 P(x)\, dx \xrightarrow{\text{п. н.}} 0.$$

Переходя к оценке скорости сходимости процедуры (3), (25), введем ряд дополнительных предположений, которые несколько отличаются от предположений, лежавших в основе теорем 3—5. Помимо условия (5) о конечности разложения функции $f(x) \equiv D_A(x)$ (и соответственно конечности ряда (6) для потенциальной функции) будем теперь предполагать, что выполнено не только (16), но и более сильное неравенство

$$\int_\Omega F^2(x) P(x)\, dx > 0, \qquad (26)$$

где Ω — любое множество, такое, что

$$\int_\Omega P(x)\, dx > 0.$$

Здесь по-прежнему $F(x) = \sum_{i=1}^{N} \mu_i \varphi_i(x)$ и μ_i — любые действительные числа, не равные нулю одновременно. Кроме того, о функциях $D_A(x)$ и $P(x)$ предполагается, что

$$\int_\Gamma P(x)\, dx > 0, \qquad (27)$$

где Γ — множество точек x, для которых $0 < D_A(x) < 1$ (строгие неравенства!). Условие (27) означает лишь, что имеется отличная от нуля вероятность появления точек, в которых $D_A(x)$ не равно нулю или единице, а это как раз означает, что рассматриваемая задача является вероятностной и не сводится к детерминистской.

При этих предположениях имеет место

Теорема 11. *Пусть для любого $r > 0$ существует $\lambda(r) > 0$, удовлетворяющее неравенствам* (18) *начиная с некоторого n. Тогда для любого $\delta > 0$ существуют такие $\lambda^*(\delta) > 0$ и $c(\delta) > 0$, что при использовании процедуры* (3), (6), (25) *вероятность появления такой реализации, для которой при всех n выполнено неравенство*

$$\int [D_A(x) - \overline{f}_n(x)]^2 P(x)\, dx < c(\delta)\, \gamma_n^{\lambda^*(\delta)}, \qquad (28)$$

больше чем $1 - \delta$.

Если, кроме того, существует такое число $\lambda_0 > 0$, что при $\lambda = \lambda_0$ неравенства (18) *удовлетворяются при любых r, то в* (28) *можно принять $\lambda^*(\delta) \equiv \lambda_0$.*

5. Задача обучения машины распознаванию образов без учителя

В § 3 задача обучения машины распознаванию образов (с учителем) понималась как задача построения в некотором пространстве X поверхности, разделяющей множества точек, соответствующих объектам из разных образов. При этом в разделяемых множествах предполагалось, что они достаточно «разнесены» в пространстве и что их границы (а значит, и разделяющая их поверхность) достаточно гладки; не «вычурны». Роль оператора («учителя») в этой задаче заключалась в том, что он сообщал машине, к какому множеству относится каждая появляющаяся в процессе обучения точка. При построении разделяющей поверхности эта информация существенно использовалась. Если множества, подлежащие разделению, достаточно «далеки» друг от друга, то после того, как машине показано много объектов, соответствующие им точки образуют в пространстве X достаточно «далекие» скопления точек — «кучки». И если бы удалось построить поверхность, разделяющую эти «кучки», не используя информацию о том, к какому множеству принадлежит каждая из появившихся на входе машины точек, то оказалось бы принципиально возможным обучение машины распознаванию образов без участия учителя. Как показал М. И. Шлезингер [9], формализация постановки этой задачи может быть проведена путем формулировки критерия качества разделения, являющегося функционалом от разделяющей поверхности. Тогда решение задачи обучения машины распознаванию образов без учителя может быть сведено к поиску разделяющей поверхности, экстремизирующей заданный функционал. Тот или иной конкретный вид критерия будет при этом приводить к различным разделяющим поверхностям.

При формулировке критерия, который применяется в настоящей работе [1]), использовано понятие обобщенного расстояния $\rho(x, y)$ между двумя точками x и y, принадлежащими X,

$$\rho^2(x, y) = K(x, x) + K(y, y) - 2K(x, y), \qquad (29)$$

где $K(x, y)$ — потенциальная функция вида (6).

Из (6) и (29) следует, что

$$\rho^2(x, y) = \sum_{i=1}^{N} \lambda_i^2 [\varphi_i(x) - \varphi_i(y)]^2. \qquad (30)$$

Обобщенное расстояние $\rho(x, y)$ превращается в обычное евклидово расстояние между векторами в случае, когда X есть N-мерное пространство, а $K(x, y) = (x, y)$ — скалярное произведение векторов x и y.

Пусть теперь проведено некоторое разделение пространства X поверхностью $f(x) = 0$. Будем говорить, что точки, для которых $f(x) > 0$, принадлежат множеству A, а точки, для которых $f(x) < 0$, принадлежат множеству B. Функцию $f(x)$ будем и в этом случае называть разделяющей функцией. Введем функционал от функции $f(x)$

$$\Psi(f(x)) = P_A M\{\rho^2(x, y)|_{x,y\in A}\} + P_B M\{\rho^2(x, y)|_{x,y\in B}\}, \qquad (31)$$

где $M\{\rho^2(x, y)|_{x,y\in R}\}$ — средний квадрат расстояния между точками, принадлежащими некоторому множеству $R = A, B$, а P_A и P_B — вероятности появления точек из A и из B соответственно. Таким образом, (31) имеет смысл среднего квадрата расстояния между двумя точками, принадлежащими одному множеству. Примем в качестве разделяющей функции такую функцию, которая минимизирует функционал (31).

Раскрывая выражение (31) с учетом (29), получаем, что исходная задача может быть сформулирована как задача построения такой функции $f(x)$, которая максимизирует функционал

$$\Phi(f(x)) = \frac{1}{P_A} \int\limits_{x,y\in A}\int K(x, y) P(x) P(y)\, dx\, dy + \frac{1}{P_B} \int\limits_{x,y\in B}\int K(x, y) P(x) P(y)\, dx\, dy. \qquad (32)$$

Введем в рассмотрение так называемое спрямляющее пространство Z, координаты которого определяются соотношениями

$$z_i = \lambda_i \varphi_i(x), \qquad i = 1, \ldots, N. \qquad (33)$$

[1]) Этот критерий является некоторой модификацией предложенного в [9] критерия, позволяющей применить для решения задачи обучения машины распознаванию образов без учителя метод потенциальных функций.

В спрямляющем пространстве Z функционал (32) принимает вид

$$\Phi(f(x)) = \frac{1}{P_A}\left[\int\limits_A zP(z)\, dz\right]^2 + \frac{1}{P_B}\left[\int\limits_B zP(z)\, dz\right]^2 = \frac{M_A^1}{P_A} + \frac{M_B^1}{P_B}, \qquad (34)$$

где $M_A = \int\limits_A zP(z)\, dz$ и $M_B = \int\limits_B zP(z)\, dz$. Разделяющую функцию в спрямляющем пространстве вновь будем обозначать $f(x)$.

Ниже формулируется теорема 12, устанавливающая класс функций, в котором следует искать разделяющую функцию, доставляющую экстремум произвольному функционалу Φ в том случае, когда этот функционал является дифференцируемой функцией от ненормированных моментов распределения M_R^k по множествам A и B:

$$M_R^k = \int\limits_R z^k P(z)\, dz, \quad R = A, B, \quad k = 0, 1, \ldots, r, \qquad (35)$$

где z^k есть k-я степень вектора z. Заметим, что z^k является числом $(|z|^2)^{k/2}$, когда k четное, и вектором $-z(|z|^2)^{(k-1)/2}$, когда k нечетное. Очевидно, (34) является функционалом этого типа. Ниже под выражением $\partial\Phi/\partial M_R^k$ будем понимать число в случае четного k (т. е. в случае, когда M_R^k число) и вектор с компонентами $\partial\Phi/\partial M_{R_i}^k$ ($M_{R_i}^k$ есть i-я компонента вектора M_R^k) в случае нечетного k, а под выражением (u, v) — либо произведение чисел u и v, либо скалярное произведение векторов u и v.

Теорема 12. *Пусть функционал Φ является дифференцируемой функцией от ненормированных моментов* (35) *до r-й степени включительно, а плотность вероятности $P(z)$ является непрерывной функцией, обращающейся в нуль вне некоторого ограниченного множества Z. Тогда если экстремум функционала Φ дается некоторой разделяющей функцией, то этот же экстремум дается разделяющей функцией, являющейся полиномом r-й степени*

$$f(z) = \sum_{k=0}^{r} (c_k z^k), \qquad (36)$$

причем

$$c_k = \frac{\partial\Phi}{\partial M_A^k} - \frac{\partial\Phi}{\partial M_B^k}. \qquad (37)$$

Если условия (36) *и* (37) *выполнены, то функционал Φ принимает стационарное значение.*

Применительно к функционалу (34), где $k = 0, 1$, утверждение теоремы означает, что разделяющую функцию в спрямляющем пространстве можно искать среди функций

$$f(z) = (c, z) - a, \tag{38}$$

причем экстремум достигается при

$$c = \frac{\partial \Phi}{\partial M_A} - \frac{\partial \Phi}{\partial M_B} = 2\left(\frac{M_A}{P_A} - \frac{M_B}{P_B}\right),$$

$$a = -\left(\frac{\partial \Phi}{\partial P_A} - \frac{\partial \Phi}{\partial P_B}\right) = \frac{M_A^2}{P_A^2} - \frac{M_B^2}{P_B^2}.$$

Переходя теперь от спрямляющего пространства Z к исходному X, из (38) получаем, что разделяющую функцию следует искать в классе функций

$$f(x) = \sum_{i=1}^{N} c_i \varphi_i(x) - a. \tag{39}$$

Из (38) следует, что если рассматривать лишь обычное евклидово расстояние, то, используя функционал (31), можно получить разделение лишь линейно разделимых множеств. Использование обобщенного расстояния вида (29) позволяет получить разделение более сложных множеств. Это соображение оправдывает введение обобщенного расстояния.

Для решения поставленной задачи ниже предлагается рекуррентная процедура, основанная на методе потенциальных функций.

Пусть последовательно появляются точки $x_1, x_2, \ldots, x_n$. В процессе реализации процедуры в соответствии с появляющимися точками строятся функции $\Phi_A^n(x)$ и $\Phi_B^n(x)$ и числа a_A^n и a_B^n, которые используются для построения n-го приближения разделяющей функции

$$f_n(x) = \Phi_A^n(x) - \Phi_B^n(x) - (a_A^n - a_B^n). \tag{40}$$

Если на $(n+1)$-м шаге появилась точка x^{n+1}, то принимается, что x^{n+1} принадлежит A, если $f_n(x^{n+1}) > 0$, и что x^{n+1} принадлежит B, если $f_n(x^{n+1}) < 0$. Пусть к $(n+1)$-му шагу появлявшиеся ранее точки n_A раз относились к A и $n_B = n - n_A$ раз — к B. Тогда строится $(n+1)$-е приближение, т. е. функции $\Phi_A^{n+1}(x)$ и $\Phi_B^{n+1}(x)$ и числа a_A^{n+1} и a_B^{n+1}, по следующему правилу:

а) если $x^{n+1} \in A$, то

$$\begin{aligned} \Phi_A^{n+1}(x) &= \Phi_A^n(x) + \gamma_{n_A}[K(x, x^{n+1}) - \Phi_A^n(x)], \\ a_A^{n+1} &= a_A^n + \gamma_{n_A}[\Phi_A^n(x^{n+1}) - 2a_A^n], \\ \Phi_B^{n+1}(x) &= \Phi_B^n(x), \quad a_B^{n+1} = a_B^n; \end{aligned} \tag{41а}$$

б) если $x^{n+1} \in B$, то

$$\begin{aligned} \Phi_A^{n+1}(x) &= \Phi_A^n(x), \\ a_A^{n+1} &= a_A^n, \\ \Phi_B^{n+1}(x) &= \Phi_B^n + \gamma_{n_B}[K(x, x^{n+1}) - \Phi_B^n(x)], \\ a_B^{n+1} &= a_B^n + \gamma_{n_B}[\Phi_B^n(x^{n+1}) - 2a_B^n], \end{aligned} \tag{41б}$$

где γ_n — некоторая заранее выбираемая последовательность положительных чисел; ограничения на выбор последовательности γ_n будут указаны далее.

Таким образом, на каждом шаге меняются либо только Φ_A^n и a_A^n, либо только Φ_B^n и a_B^n в зависимости от знака $f^n(x^{n+1})$. Начальные значения величин, входящих в рекуррентные соотношения (41), определяются по первым двум показанным точкам:

$$\begin{aligned} \Phi_A^2(x) &= K(x, x^1), \\ a_A^2 &= \frac{K(x^1, x^1)}{2}, \\ \Phi_B^2(x) &= K(x, x^2), \\ a_B^2 &= \frac{K(x^2, x^2)}{2}. \end{aligned} \tag{42}$$

Остановимся теперь на ограничениях, накладываемых на выбор последовательности γ_n:

1) последовательность чисел $\gamma_n > 0$ монотонно не возрастает;

2) $\sum \gamma_n$ расходится;

3) существуют два таких числа $\alpha > 0$ и $\lambda > 0$ и такой номер n_0, что

$$(1 - 2\gamma_{n+1} + \alpha\gamma_{n+1}^{\alpha-\lambda})\left(\frac{\gamma_n}{\gamma_{n+1}}\right)^{\lambda} \leqslant 1, \quad n > n_0,$$

а ряд $\sum \gamma_n^{1+\lambda}$ сходится;

4) для любого числа $\beta > 0$ найдется такое $L_1(\beta)$, что если $\frac{n_1}{n_2} > \beta$, то

$$\frac{\gamma_{n_1}}{\gamma_{n_2}} < L_1(\beta),$$

5) для любого $L_2 > 0$ найдутся $N(L_2) > 0$ и $\varkappa(L_2) > 0$, такие, что

$$\sum_{k=[\varkappa n]}^{n} \gamma_k > L_2, \quad n > N(L_2),$$

где $[\varkappa n]$ — целая часть произведения $\varkappa n$.

Этим пяти условиям удовлетворяет, например, последовательность $\gamma_n = 1/n^{1-\varepsilon}$, где $0 \leqslant \varepsilon < 1/2$.

Таким образом, ограничения на выбор последовательности γ_n формально оказываются более жесткими, нежели в иных применениях метода потенциальных функций [1]).

Теорема 13. *Пусть $P(z)$ — непрерывная функция, обращающаяся в нуль вне некоторого ограниченного множества. Тогда в силу правил* (41), (42) *функционал* (31) *с вероятностью единица стремится к стационарному значению.*

6. Реализация метода потенциальных функций

В основе метода потенциальных функций лежит предположение (1), (4) о свойствах разложения восстанавливаемой функции $f(x)$ по некоторой известной системе $\varphi_i(x)$. Поскольку потенциальная функция в силу (2) определяется системой $\varphi_i(x)$, то при решении конкретных задач прежде всего возникает вопрос, какова же эта система функций.

Может показаться поэтому, что при практической реализации метода система функций $\varphi_i(x)$ должна быть жестко определена заранее для каждой конкретной задачи. На практике, однако, дело обстоит проще и систему функций $\varphi_i(x)$ можно выбрать в значительной степени произвольно. Это связано с тем, что подлежащие восстановлению функции $f(x)$ обычно оказываются достаточно «гладкими», и если произвольно выбираемая система $\varphi_i(x)$ не слишком «вычурна», то коэффициенты c_i в разложении (1) убывают, а условие (4) выполняется. Соответственно в значительной степени произвольна и потенциальная функция $K(x, y)$, тем более что и при фиксированной системе $\varphi_i(x)$ оставалась бы возможность изменения $K(x, y)$ за счет выбора λ_i в формуле (2). Поэтому возникает возможность непосредственно задать (как отмечено выше, более или менее произвольно) функцию $K(x, y)$, не задавая явно системы $\varphi_i(x)$. При этом надо позаботиться о том, чтобы заданная симметричная функция $K(x, y)$ могла бы быть представлена в виде (2) с действительными λ_i. При выборе функции $K(x, y)$ полезна следующая теорема [5]:

Теорема 14. *Пусть X представляет собою*: а) *ограниченную область m-мерного евклидова пространства E_m или* б) *конечное множество точек в E_m. Пусть далее функция $K(|z|)$, где*

[1]) Остается невыясненным, в какой мере эти ограничения порождены принципиальными особенностями данной задачи, а в какой — принятым авторами методом доказательства теоремы 13 о сходимости процедуры.

$|z| = \sqrt{z_1^2 + \ldots + z_m^2}$, *непрерывна в E_m и ее Фурье-преобразование*

$$\int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} K(|z|) \exp\left[-i \sum_{k=1}^{m} \omega_k z_k\right] dz_1 \ldots dz_m$$

положительно при любых $\omega_1, \ldots, \omega_m$. Тогда функция $K(|x - y|)$ двух переменных при $x, y \in X$ может быть представлена формулой (2), *причем система функций $\varphi_i(x)$ полна в $L^2(X)$* [1]).

Если пространство удовлетворяет условию теоремы и в X вводится евклидова метрика с расстоянием $\rho(x, y)$, то удобно потенциальную функцию выбирать в виде $K[\rho(x, y)]$ (см. [10]). Приняв, например,

$$K[\rho(x, y)] = e^{-\alpha^2 \rho^2(x, y)} \tag{43}$$

и имея в виду, что Фурье-преобразование этой функции имеет вид $\frac{\sqrt{2\pi}}{\alpha} \exp\left[-\frac{1}{\alpha^2} \sum \omega_k^2\right] > 0$, в силу теоремы 12 устанавливаем, что функция (43) может быть использована в качестве потенциальной функции.

Далее рассматриваются два способа практической реализации метода потенциальных функций —«машинный» и «схемный»,— отличающиеся способом запоминания приближений, последовательно получаемых на каждом шаге процедуры. При этом оказывается, что при «машинной» реализации естественно задаваться непосредственно потенциальной функцией $K(x, y)$ и вообще не интересоваться конкретным видом функций $\varphi_i(x)$. Наоборот, при «схемной» реализации можно не вводить явно потенциальную функцию, а оперировать лишь с функциями $\varphi_i(x)$.

Машинная реализация. При этом способе реализации начинают с выбора потенциальной функции $K(x, y)$ с учетом сделанных выше замечаний. Далее, к n-му шагу в памяти машины хранят лишь коды точек $x^1, \ldots, x^n$ и числа $r_1, \ldots, r_n$, так что значение функции $f_n(x)$ в любой точке $x = x^*$ может быть вычислено по формуле

$$f_n(x^*) = \sum_{q=1}^{n} r_q K(x^q, x^*). \tag{44}$$

Из (44) следует, что те точки x^q, которым соответствует $r_q = 0$, могут не запоминаться.

[1]) $L^2(X)$ — пространство функций, для которых существует интеграл $\int_X f^2(x)dx$ (в случае б) вместо интеграла рассматривается соответствующая сумма).

При появлении $(n+1)$-й точки x^{n+1} вычисляется коэффициент r_{n+1} (в соответствии с избранной в зависимости от решаемой задачи процедурой), и если $r_{n+1} \neq 0$, то в памяти машины задается код точки x^{n+1} и это значение r_{n+1}. После окончания «процесса обучения» при появлении новой точки x^* машина «выдает» на выход в качестве $f(x^*)$ величину, подсчитанную по формуле (44).

Схемная реализация. В том случае, когда может быть сделано предположение, что восстанавливаемая функция представима конечным рядом (5) по некоторой выбранной системе $\varphi_i(x)$, возможна иная, схемная реализация процедуры.

Действительно, из формул (3), (5) и (6) следует, что если коэффициенты c_i^n удовлетворяют рекуррентной процедуре

$$c_i^n = c_i^{n-1} + r_n \lambda_i \varphi_i(x^n) \qquad (i = 1, \ldots, N), \quad c_i^0 = 0, \qquad (45)$$

то

$$\sum_{i=1}^{N} c_i^n \varphi_i(x) \equiv f_n(x), \qquad (46)$$

т. е. не только восстанавливаемая функция $f(x)$, но и любое ее приближение $f_n(x)$ представимо конечным рядом (46) по системе функций $\varphi_i(x)$, а коэффициенты этих рядов определяются процедурой (45).

Использование процедуры (45) позволяет хранить в памяти не нарастающее число кодов точек x^n, предъявленных в процессе обучения, как это было при машинной реализации, а всегда одно и то же число N коэффициентов c_i^n. На каждом шаге по (45) вычисляются и запоминаются новые значения этих коэффициентов, а старые их значения забываются.

Рекуррентная процедура (45) может быть реализована схемой, в которой сигнал x^n поступает на N нелинейных преобразователей $\psi_i(x) = \lambda_i \varphi_i(x)$. Выход $\psi_i(x)$ каждого преобразователя умножается на величину c_i^{n-1}, накопленную к n-му шагу в накапливающем сумматоре. Величины $c_i^{n-1}\psi_i(x^n)$ складываются и образуют функцию $f_{n-1}(x^n)$, которая поступает на вход устройства, формирующего на каждом шаге величину r_n. На второй вход этого устройства поступает информация о $f(x^n)$ в соответствии с рассматриваемой процедурой [1]. В каждом i-м канале величина r_n умножается на $\psi_i(x^n)$, и в следующем такте содержимое накапливающего сумматора c_i^{n-1} изменяется на величину $r_n \psi_i(x^n)$, образуя c_i^n.

[1]) Например, значение $f(x^n)$ или это значение вместе с помехой в задачах восстановления функции, знак $f(x^n)$ в детерминистской постановке задачи о распознавании образов и т. д.

В [1] было показано, что рассмотренная схема, примененная к детерминистской постановке задачи распознавания образов, является обобщением перцептрона Розенблатта [11].

Из содержания этого параграфа видно, что класс «перцептронных схем» решает общую задачу о восстановлении функции в том смысле, как это было сформулировано во введении.

Институт автоматики и телемеханики,
Москва, СССР

ЛИТЕРАТУРА

[1] Айзерман М. А., Браверман Э. М., Розоноэр Л. И., Теоретические основы метода потенциальных функций в задаче об обучении автоматов разделению входных ситуаций на классы, *Автомат. и телемехан.*, **25**, № 6 (1964).

[2] Айзерман М. А., Браверман Э. М., Розоноэр Л. И., Вероятностная задача об обучении автоматов распознаванию классов и метод потенциальных функций, *Автомат. и телемехан.*, **25**, № 9 (1964).

[3] Айзерман М. А., Браверман Э. М., Розоноэр Л. И., Метод потенциальных функций в задаче о восстановлении характеристики функционального преобразователя по случайно наблюдаемым точкам, *Автомат. и телемехан.*, **25**, № 12 (1964).

[4] Браверман Э. М., Метод потенциальных функций в задаче обучения машины распознаванию образов без учителя, *Автомат. и телемехан.*, **27**, № 10 (1966).

[5] Браверман Э. М., О методе потенциальных функций, *Автомат. и телемехан.*, **26**, № 12 (1965).

[6] Браверман Э. М., Пятницкий Е. С., Оценки скорости сходимости алгоритмов, основанных на методе потенциальных функций, *Автомат. и телемехан.*, **27**, № 1 (1966).

[7] Цыпкин Я. З., О восстановлении характеристики функционального преобразователя по случайно наблюдаемым точкам, *Автомат. и телемехан.*, **26**, № 11 (1965).

[8] Novikoff A. B. J., On convergence proofs for perceptrons, Tech. Rep. Stanford Research Institute, Politechn. Inst. Brooklin, Apr. 1962.

[9] Шлезингер М. И., О самопроизвольном различении образов, сб. «Читающие автоматы», изд-во «Наукова думка», Киев, 1965.

[10] Башкиров О. А., Браверман Э. М., Мучник И. Б., Алгоритмы обучения машин распознаванию зрительных образов, основанные на использовании метода потенциальных функций, *Автомат. и телемехан.*, **25**, № 5 (1964).

[11] Rosenblatt F., Principles of neurodynamics. Perceptron and the theory of brain mechanisms, Washington, 1962. Русский перевод: Розенблатт Ф., Принципы нейродинамики. Перцептрон и теория механизмов мозга, изд-во «Мир», 1965.

Секция 8

Section 8

EMBEDDING SMOOTH MANIFOLDS

WILLIAM BROWDER[1])

We shall consider the problem of smoothly embedding one differentiable manifold in another. Our aim will be to reduce the problem to certain homotopy problems, in the spirit of, and using the techniques of [2], [9], [10]. The results will be in terms of finding bundles, maps, and spaces which will play the role, up to homotopy, of the normal bundle and the inclusion of its boundary into the complement of the submanifold. The results are extensions of some results announced in [3], and generalize results of Levine [8]. The proofs are quite simplified compared to [3] and give a simpler approach to some of Levine's results. The main restrictions on our technique are that the big manifold should be of dimension $\geqslant 5$ and the complement of the submanifold simply connected. Unlike [3], no restrictions of simple connectivity are made on the submanifold.

We shall consider closed manifolds, embedded in closed manifolds, and consider only existence theorems for embeddings. The techniques yield similar theorems for bounded manifolds, with boundary embedded in boundary and some isotopy theorems, but to keep the exposition as short and simple as possible, we defer these to another time.

The techniques may also be applied to piecewise linear manifolds using the block bundles of Rourke and Sanderson [11] and the piecewise linear surgery techniques introduced in [4].

1. Statements of theorems

Let M^n be an n-dimensional closed smooth manifold and consider triples $\mathscr{A} = (\xi^k, f, Y)$ where ξ^k is a linear k-plane bundle over M, Y is a space and $f\colon E_0 \to Y$ is a continuous map, where E_0 is the associated sphere bundle to ξ^k (or if one prefers, the associated $(R^k - 0)$ bundle). We call such a triple a *system of codimension k* over M. If M^n is embedded in a manifold W^{n+k} with normal bundle ν then the system $(\nu, i, W - M)$ where i is the inclusion, is called the *normal system* of the embedding.

Two systems (ξ^k, f, Y) and (η^k, g, Z) of codimension k over M will be called *equivalent* if there is a linear bundle equivalence

[1]) The author was supported in part by an NSF grant.

$b\colon \xi \to \eta$ and a homotopy equivalence $h\colon Y \to Z$ such that the diagram

$$\begin{array}{ccc} E_0(\xi) & \xrightarrow{f} & Y \\ {\scriptstyle b}\downarrow & & \downarrow{\scriptstyle h} \\ E_0(\eta) & \xrightarrow{g} & Z \end{array}$$

commutes.

If $\mathscr{S} = (\xi^k, f, Y)$ is a system of codimension k, we define the *suspension* of $\mathscr{S}$ to be the system $\Sigma\mathscr{S} = (\xi^k + \varepsilon^1, f', \Sigma Y)$ of codimension $(k + l)$, where $\xi^k + \varepsilon^1$ is the sum of ξ^k with a trivial line bundle, ΣY is the suspension of Y, and $f'\colon E_0(\xi + \varepsilon^1) \to \Sigma Y$ is the suspension of f along each fiber. Note that if $\mathscr{S}$ is the normal system to an embedding of M^n in S^{n+k}, and if S^{n+k} is contained as the equator in S^{n+k+1}, then $\Sigma\mathscr{S}$ is equivalent to the normal system of M^n in S^{n+k+1}.

We now are in a position to state the theorem for embeddings in the sphere:

Theorem 1. Let M^n be a closed smooth manifold, $\mathscr{S} = (\xi^k, f, Y)$ a system of codimension k, $k \geqslant 3$, $n + k \geqslant 5$, Y 1-connected. Then $\mathscr{S}$ is equivalent to the normal system of an embedding of M^n in S^{n+k} if and only if $\Sigma_\ell\mathscr{S}$ is equivalent to the normal system of an embedding of M^n in S^{n+k+1}.

If q is sufficiently large, $q > n + 1$, then there is a unique (up to isotopy) embedding of M^n in S^{n+q}, and hence a unique normal system $\mathscr{S}'$ of this codimension (up to equivalence), which we shall call the *stable normal system* of M. Hence, Theorem 1 implies that, $\mathscr{S}$ is a normal system for an embedding in S^{n+k} if and only if $\Sigma^{q-k}\mathscr{S}$ is equivalent to $\mathscr{S}'$. This question naturally subdivides into three:

(1) Is there a bundle equivalence $b\colon \xi^k + \varepsilon^{q-k} \to \nu^q$, where ν^q = stable normal bundle of M in S^{n+q}?

(2) Is there a homotopy equivalence $h\colon \Sigma^{q-k}Y \to (S^{n+q} - M)$?

(3) Can b and h be chosen so that the diagram

$$\begin{array}{ccc} E_0(\xi^k + \varepsilon^{q-k}) & \xrightarrow{b} & E_0(\nu) \\ \downarrow{\scriptstyle f^q} & & \downarrow{\scriptstyle t} \\ \Sigma^{q-k}Y & \xrightarrow{h} & (S^{n+q} - M) \end{array}$$

commutes?

Of these questions, (1) is equivalent to the immersion problem, by Hirsch's Theorem [5], and has been studied intensively for particular manifolds. Many authors have considered (2) from the point of view of proving non-embedding resulst, but (3) has heretofore received less attention, except for the case of M^n a homotopy sphere (see [8] where Theorem 1 is proved in this case.)

Suppose that ξ^k is a vector bundle over M and that the total space $E(\xi)$ is embedded as an open subset in a space X. Then there is a natural collapsing map $C: X \to T(\xi)$, where $T(\xi) = E(\xi)/E_0(\xi)$ is the Thom complex of ξ. If ν^k is the normal bundle for an embedding of M^n in S^{n+k}, then the Tubular Neighborhood Theorem implies that $E(\nu)$ is embedded as an open subset of S^{n+k} and we call the homotopy class of the collapsing map $C: S^{n+k} \to T(\nu)$ a *normal invariant of the embedding*. Note that a bundle automorphism of ν induces a different embedding of $E(\nu)$ in S^{n+k} and a different map $C': S^{n+k} \to T(\nu)$.

If $k > n+1$, then we call the set C_M of all such elements in $\pi_{n+k}(T(\nu))$ the *set of normal invariants of* M. Note that $T(\xi + \varepsilon^1) = \Sigma T(\xi)$, the suspension of $T(\xi)$, and that if $M^n \subset S^{n+k}$ with normal bundle ν^k and $\alpha \in \pi_{n+k}(T(\nu))$ a normal invariant of the embedding, then the composite embedding $M^n \subset S^{n+k} \subset S^{n+k+1}$ has normal bundle $\nu^k + \varepsilon^1$ and normal invariant $\Sigma(\alpha) =$ the suspension of $\alpha \in \pi_{n+k+1}(\Sigma T(\nu)) = \pi_{n+k+1}(T(\nu + \varepsilon^1))$. Then Theorem 1 has the following corollary:

Corollary. Let M^n be a closed smooth manifold, ξ^k a k-plane bundle over M^n, $k \geqslant 2$, $n+k \geqslant 4$, $b: \xi^k + \varepsilon^{q-k} \to \nu^q$ a bundle equivalence with the normal bundle ν of M^n in S^{n+q}, $q > n+1$. Let $\alpha \in \pi_{n+k}(T(\xi))$ be such that $T(b)_*(\Sigma^{q-k}(\alpha)) \in C_M \subseteq \pi_{n+k}(T(\nu))$, the set of normal invariants of M. Then M embeds in S^{n+k+1} with normal bundle $\xi + \varepsilon^1$ and normal invariant $\Sigma(\alpha)$.

To prove the corollary one simply notes that $(\xi + \varepsilon^1, f, T(\xi) \bigcup_\alpha e^{n+k+1})$ is a system of codimension $k+1$ which suspends to the stable normal system of M, where f is the composite map $E_0(\xi + \varepsilon^1) \to E_0(\xi + \varepsilon^1)/\omega(M) = T(\xi) \to T(\xi) \bigcup_\alpha e^{n+k+1}$, (where ω is the canonical cross-section). An example due to J. Wagoner [12, § 6] shows that the corollary cannot be improved to get an embedding, in S^{n+k} (c.f. [8]). The corollary is related to a result of Levine [7] (c.f. [3]).

Theorem 1 has the virtue that it is rather simple to state, but unfortunately the general theorem is much more cumbersome.

Let M^n and W^{n+k} be smooth closed manifolds, let ν^q be the normal bundle of W^{n+k} in S^{n+k+q}, and ω^{k+q} the normal bundle of M^n in S^{n+k+q}, and denote by $C_W \subseteq \pi_{n+k+q}(T(\nu^q))$ and $C_M \subseteq \pi_{n+k+q}(T(\omega^{k+q}))$ the sets of normal invariants of W and M.

Theorem 2. Suppose W^{n+k} is 1-connected, $n+k \geqslant 5$ and let (ξ^k, f, Y) be a system over M^n such that Y is 1-connected, and $H_{n+k-1}(Y) = 0$. Suppose there is a homotopy equivalence $h: E(\xi) \bigcup_f Y \to W$ and a bundle equivalence $b: \xi^k + (h^*(\nu^q) \mid M) \to \omega^{k+q}$ such that $T(b)_* \eta_*(\alpha) \in C_M \subseteq \pi_{n+k+q}(T(\omega))$, where $\eta: T(h^*(\nu)) \to T(\xi^k + (h^*(\nu) \mid M))$ is the natural collapsing map, and $\alpha \in \pi_{n+k+q}(T(h^*(\nu)))$ is some element which goes into $C_W \subseteq \pi_{n+k+q}(T(\nu))$. Then M embeds in W with normal system (ξ, f, Y).

It would be possible to state the general theorem in an analogous way to Theorem 1, in terms of embedding in $W \times I$, but it seems more artificial and cumbersome in this case.

2. Proofs

First we outline how to deduce Theorem 1 from Theorem 2, and then we will outline the proof of Theorem 2.

Proof of Theorem 1 from Theorem 2. Suppose $\Sigma\mathscr{S}$ is equivalent the normal system of an embedding of M^n in S^{n+k+1}, so that $E(\xi + \varepsilon^1) \bigcup_f \Sigma Y$ is homotopy equivalent to S^{n+k+1}. A simple argument using the Mayer-Viettoris sequence then shows that $E(\xi) \bigcup_f Y$ is homotopy equivalent to S^{n+k}. The normal bundle of S^{n+k} in S^{n+k+1} is of course the trivial line bundle, and since $\Sigma\mathscr{S}$ is the normal system of M in S^{n+k+1} it follows that the normal invariant condition of Theorem 2 is satisfied, using $\alpha =$ identity map of S^{n+k+1}. Then we see that the conditions of Theorem 2 are satisfied and $\mathscr{S}$ is equivalent to a normal system of M^n in S^{n+k}.

Now we outline the proof of Theorem 2.

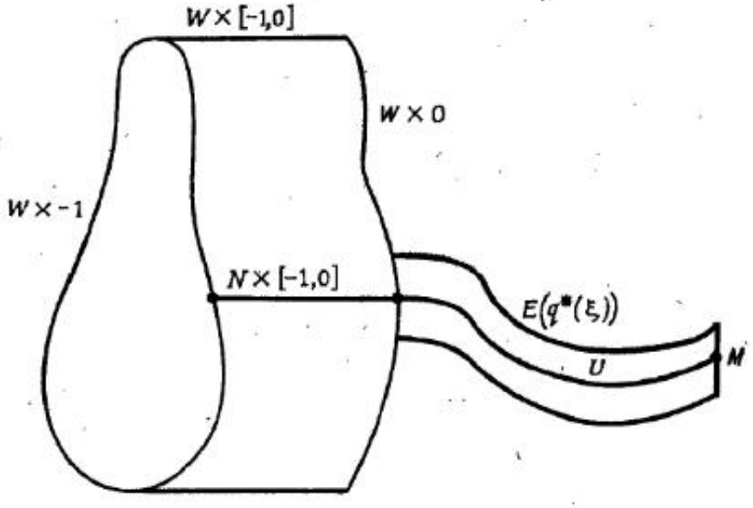
Proof of Theorem 2. Let $k = h^{-1}$ and make $k: W^{n+k} \to E(\xi) \bigcup_f Y$ t-regular on M^n to get a manifold $N^n \subset W^{n+k}$ with $k^{-1}(M) = N$, $k \mid N: N \to M$ of degree 1 and induces the normal bundle ξ' of N in W from ξ. The normal bundle of N^n in S^{n+k+q} is then $\xi' + (\nu \mid N)$, so that there is a bundle map $c: \xi' + (\nu \mid N) \to \xi + ((h^*\nu) \mid M)$ and then into ω^{k+q} by b, which covers the map $k \mid N: N \to M$. Set $a = bc$, $a: \xi' + (\nu \mid N) \to \omega$.

Now a normal invariant of $N^n \subset S^{n+k+q}$ is clearly given by $\bar{\eta}_*(\alpha)$, where $\alpha \in \pi_{n+k+q}(T(\nu))$ is the normal invariant of W and $\bar{\eta}: T(\nu) \to T(\xi' + (\nu \mid N))$ is the natural collapsing map. But clearly $T(a)_*\bar{\eta}_*(\alpha) = T(b)_*\bar{\eta}_*(\alpha) \in C_M$ by hypothesis. Hence

$$T(a) \circ \bar{\eta}: S^{n+k+q} \to T(\omega)$$

is homotopic to a map β such that $\beta^{-1}(M) = M$, $\beta \mid M =$ identity, and $(\beta \mid M)$ is covered by a bundle map of ξ to itself. Since $\xi + (h^*\nu \mid M)$ is bundle equivalent to ω, it follows that the same holds for the map into $T(\xi + (h^*\nu \mid M))$.

Making the homotopy t-regular on M^n we obtain a cobordism U^{n+1} of N with M, and a map q: $U^{n+1} \to M^n$ such that $q \mid N = k \mid N$ and $q \mid M =$ identity. Further $U^{n+1} \subset S^{n+k+q} \times I$ with normal bundle $q^*(\xi + (h^*\nu \mid M))$ and if B: $q^*(\xi + (h^*\nu \mid M)) \to \xi + (h^*\nu \mid M)$ is the bundle map covering q, $B \mid N = C$, where $C = C_1 + C_2$, C_1: $\xi' \to \xi$, C_2: $\nu \mid N \to (h^*\nu \mid M)$. Now C_1 extends to a bundle map of $q^*(\xi)$ into ξ, while C_2 extends to a bundle map of ν into $h^*(\nu)$, and also of $q^*(h^*\nu \mid M)$ into $h^*\nu \mid M$. Set $E = E(q^*(\xi)) =$ total space of $q^*(\xi)$. Then the normal bundle of E in $S^{n+k+q} \times I$ is $q^*(h^*\nu \mid M)$ so C_2 extends to a bundle map of the normal bundle of $E(q^*(\xi))$



into $h^*\nu \mid M$, and to a bundle map of the normal bundle ν of W^{n+k} in S^{n+k+q} to $h^*(\nu)$.

Now we take $W \times [-1, 0] \subset S^{n+k+q} \times [-1, 0]$ and $E \subset S^{n+k+q} \times [0, 1]$ and identify the tubular neighborhood of $N \subset W \times 0$ with $E(q^*(\xi)) \cap S^{n+k+q} \times 0$, where this is the total space of $(q \mid N)^*(\xi) \subset E$ (see figure).

Then $V = (W \times [-1, 0]) \cup E(q^*(\xi))$ (with the identification) defines a cobordism between W and a new manifold $W' = (W \times 0 - $ — (nbd. of N)) $\cup$ ($\partial E(q^*(\xi))$ — (nbd. of N)), with the obvious identification along the boundary of the tubular neighborhood of N. (We must round the corners in the usual way to get a differentiable manifold.) Then $V \subset S^{n+k+q} \times I$ with normal bundle the union of the normal bundles of $W \times [-1, 0]$ and $E(q^*(\xi))$. Hence the extensions of C_2 define a map of the normal bundle of V^{n+k+1} in $S^{n+k+q} \times I$ into $h^*(\nu)$, which covers the map K: $V \to E(\xi) \cup Y$, given by $k \circ$ projection on $W \times I$ and by C_1 on $E(q^*(\xi))$. Further $k^{-1}(M) = (N \times [-1, 0]) \cup U$, identifying $N \times 0$ with $N \subset \partial U$, which is diffeomorphic to U. Then $K^{-1}(M) \cap W' = M \subset U$ and $K \mid U = q$ so $K \mid M =$ identity.

Therefore, we may consider to have improved the situation by getting $M \subset W'$ will normal bundle ξ, but we now have W' instead of W, and it is easy to see that it may be a very different manifold. However, we have $W' = E(\xi) \cup X$, where $\partial X = E_0(\xi) =$ the associated sphere bundle to ξ and if $g = K \mid W'$, we have g: $(X, \partial X) \to (Y, E_0(\xi))$ (where we may assume f: $E_0(\xi) \to Y$ is an inclusion) and $g^*(h^*(\nu))$ is the normal bundle of X in S^{n+k+q}.

Lemma. Suppose $Z = A \cup B$, $A \cap B = C$ and
(1) Z satisfies Poincaré duality in dimension $m + 1$
(2) C satisfies Poincaré duality in dimension m and
(3) $H_m(A) = H_m(B) = 0$.
Then (A, C) and (B, C) are Poincaré pairs (satisfy relative Poincaré duality) in dimension $m + 1$.

This follows easily using the two Mayer-Viettoris sequences in homology and cohomology for Z, and the Five Lemma.

It follows that $(Y, E_0(\xi))$ is a Poincaré pair of dimension $n+k$, and therefore we may apply the general techniques of surgery to try to change the map g: $(X, \partial X) \to (Y, E_0(\xi))$ to a homotopy equivalence, by doing surgery in X. Further $g \mid \partial X$: $\partial X \to E_0(\xi)$ is a diffeomorphism so we may consider the problem of doing the surgery in interior X, leaving ∂X and $g \mid \partial X$ unchanged. Further Y is 1-connected and $n + k \geqslant 5$.

Now we summarize the fundamental result of the theory of surgery on simply connected manifolds (see [2], [6], [9], [10]).

The surgery problem

Let (A, B) be an m-dimensional Poincaré pair, with A 1-connected, $m \geqslant 5$. Let $(X, \partial X)$ be an m-dimensional smooth manifold with boundary and let g: $(X, \partial X) \to (A, B)$ be a map of degree 1, such that $g \mid \partial X$: $\partial X \to B$ is a homotopy equivalence. Note that this last condition is trivially satisfied if $\partial X = B = \varnothing$. Let ξ^q be a q-plane bundle over A and let b: $\nu \to \xi$ be a bundle map covering g, where $\nu =$ the normal bundle of X in S^{m+q}, q large.

A *cobordism* of g rel B will be a triple $(V^{m+1}, G, \bar{b})$, where V^{m+1} is an $(m + 1)$-manifold with $\partial V = X \cup (\partial X \times I) \cup X'$ where $X' \cap (\partial X \times I) = \partial X' = \partial X \times 1$, G: $(V, \partial X \times I) \to (A, B)$ with

$G \mid X = g$, $G \mid \partial X \times I = (g \mid \partial X) \circ$ (projection), and $\overline{b}$: $\omega \to \xi$ is a bundle map covering G, where ω = normal bundle of V in $S^{m+q} \times I$, $S^{m+q} \times 0 \cap V = X$ and $\overline{b} \mid \nu = b$ (where $\omega \mid X = \nu$). One asks the question: How much can one make X' resemble A by a cobordism rel B, for example, can one obtain from X by a cobordism rel B, a manifold X' homotopy equivalent to A?

Let $[a]$ = greatest integer $\leqslant a$, for a real number a.

Fundamental Theorem. In the surgery problem described above, one can get by a cobordism rel B to g': $(X', \partial X) \to (A, B)$ so that g' is $\left[\frac{m+1}{2}\right]$-connected. If m is odd, then g' is a homotopy equivalence. If m is even there is an obstruction σ defined, to making g' a homotopy equivalence, where $\sigma(g') \in \begin{cases} Z & \text{if } m = 4l \\ Z_2 & \text{if } m = 4l+2 \end{cases}$ and is defined "locally" as indicated below.

Suppose g: $(X, \partial X) \to (A, B)$ is k-connected, $m = 2k$, and we have an exact sequence

$$0 \to K_k(g) \to H_k(X) \xrightarrow{g_*} H_k(A) \to 0$$

where $K_k(g)$ = the kernel of g_*, $K_k \cong H_{k+1}(g)$. By the relative Hurewicz theorem $H_{k+1}(g) \cong \pi_{k+1}(g)$, and so K_k is the image of elements in $\ker g_{\#}$, $g_{\#}$: $\pi_k(X) \to \pi_k(A)$. Representing the elements by embedded $S^k \subset \operatorname{int} X$, using the Whitney embedding theorem, we may define quadratic forms Q: $K_k \to G$, $G = \begin{cases} Z & k \text{ even} \\ Z_2 & k \text{ odd} \end{cases}$ as follows:

If k is even Q is the quadratic form associated with the intersection bilinear form. If k is odd we define $Q(x) = 0$ if x is represented by an embedded sphere with a trivial normal bundle and $Q(x) = 1$, otherwise. This is quadratic, and the intersection form mod 2 is its associated bilinear form (see [6], for example).

Then σ = index Q if k is even, and σ = Arf invariant of Q if k is odd (see [6], [1]). It is clear that the value of σ is determined completely by the manifold in a *neighborhood* of a set of embedded spheres representing a basis of K_k. Hence we have the following corollary:

Corollary. Let g: $(X, \partial X) \to (A, B)$ and g': $(X', \partial X') \to (A', B')$ represent surgery problems as above. Suppose that $\dim X = \dim X' = m = 2k$ and g, g' are k-connected. Further suppose that $X \subset$ interior X', and s: $A \to A'$, such that $sg = g'i$, where i: $X \to X'$ is inclusion, and that $i_* \mid K_k(g)$: $K_k(g) \to K_k(g')$ is an isomorphism. Then $\sigma(g) = \sigma(g')$.

The "local" property of σ which yields the corollary is the key to the proof of Theorem 2, and if such a "local" obstruction could be defined in other circumstances, for example where X is not 1-connected, then one might get a similar embedding theorem with weaker hypotheses on W.

Now we return to the proof of Theorem 2. From the corollary it follows that the obstruction to doing surgery on interior X rel $E_0(\xi)$ to get a homotopy equivalence with $(Y, E_0(\xi))$ is the same as the obstruction to doing surgery on W' to get a homotopy equivalence with $E(\xi) \underset{f}{\cup} Y$. But W' was obtained by a cobordism of the type considered, from W which was homotopy equivalent to $E(\xi) \underset{f}{\cup} Y$. Hence the obstruction to the problem with W' and $E(\xi) \underset{f}{\cup} Y$ is zero, and hence by the corollary, the obstruction to doing surgery on X rel $E_0(\xi)$ is zero. Hence we get $(X', \partial X') \to (Y, E_0(\xi))$ by a homotopy equivalence, with $\partial X' \to E_0(\xi)$ a diffeomorphism. Then $W'' = E(\xi) \cup X'$ is mapped by a homotopy equivalence into $E(\xi) \underset{f}{\cup} Y$, and there is a cobordism as described above, of W'' with W. Therefore W and W'' are homotopy equivalent to $E(\xi) \underset{f}{\cup} Y$, their stable normal bundles are induced from $h^*(\nu)$ and they have the same normal invariant. Hence the theorem of Novikov (see [9] and [10]) applies to show that $W'' = W \# \Sigma$ where Σ is a homotopy sphere which bounds a π-manifold, $\#$ meaning connected sum. We note that Novikov's theorem is an easy consequence of the Fundamental Theorem, using the construction of Kervaire-Milnor, which by "plumbing" constructs a homotopy sphere bounding a π-manifold whose obstruction to doing surgery to get a disk is any desired value. (In the index case $m = 4l$, σ is always divisible by 8.) Hence, by adding the inverse of this homotopy sphere (away from $E(\xi) \subset W''$), we get back W, but with $M \subset E(\xi) \subset W$ and in fact the normal system equivalent to (ξ, f, Y).

This completes the proof of Theorem 2.

Princeton University,
Princeton, New Jersey, USA

Секция 14

Section 14

О МЕТОДАХ РЕШЕНИЯ НЕКОРРЕКТНО ПОСТАВЛЕННЫХ ЗАДАЧ

А. Н. ТИХОНОВ

Неоднократно высказывалась точка зрения, что корректность выражает физическую определенность задачи и является принципиальным условием как для применимости задачи к явлениям природы, так и для возможности получения приближенного решения. Подобная точка зрения наложила сомнение на целесообразность изучения решений некорректно поставленных задач.

Однако можно привести много классов некорректно поставленных задач как встречающихся при изучении явлений природы, так и представляющих основной аппарат математики. Сюда относятся:

1) определение равномерного приближения для производной $z = u'$ по приближенным данным в метрике C.

2) Определение суммы ряда Фурье в заданной точке по приближенным в l_2 значениям коэффициентов Фурье.

3) Равномерные приближения решений интегральных уравнений 1-го рода и многие задачи, к ним приводящие (аналитическое продолжение, операционное исчисление в действительной области) при возмущении входных данных в метрике L_2.

4) Линейные задачи на спектре при обычных дополнительных условиях, определяющих единственное решение. (Плохо обусловленные алгебраические системы, третья теорема Фредгольма.)

5) К этому кругу вопросов также относятся неустойчивые задачи оптимизации (неустойчивые задачи оптимального управления, линейного и динамического программирования).

Типичным классом некорректно поставленных задач, встречающихся при изучении явлений природы, являются «обратные задачи». При изучении объектов или явлений природы z, недоступных для непосредственного измерения, мы часто пользуемся изучением их физически детерминированных проявлений $u = Az$ (A — обычно вполне непрерывный оператор), так что определение z связано с «обратной задачей» (некорректно поставленной) определения z по u и заданному приближению. Все возрастающее значение имеет задача извлечения наиболее полной информации из результатов экспериментов с помощью математических методов (и автоматизация ее решения) вместо того, чтобы эту информацию получать при помощи усложнения экспериментальной техники.

В настоящее время определилось два подхода к решению некорректно поставленных задач.

Первый подход связан с предположением, что имеется дополнительная информация, ограничивающая класс возможных решений $\bar{Z}$ ($z \in \bar{Z}$) и позволяющая сделать заключение о компактности Z.

В этом случае решение обратной задачи $Az = u$ устойчиво. Изложение этого подхода к решению некорректно поставленных задач и сводка результатов даны в монографии М. М. Лаврентьева, где изучен также вопрос, когда входные данные задачи выходят из $\bar{U} = A\bar{Z}$. К этому направлению принадлежит понятие «квазирешения», введенное В. К. Ивановым. При таком подходе основным вопросом при решении конкретных задач является установление дополнительных ограничений на класс решений, делающий его компактным.

Второй подход трактует решение неустойчивых задач. При этом первым вопросом является определение того, как можно понимать приближенное решение неустойчивой задачи. При задании приближенных входных данных $\tilde{u}^\delta$ обычно задается их точность, т. е. величина δ возможного уклонения $\tilde{u}^\delta$ от точных значений $\tilde{u}$: $\rho_u(\tilde{u}^\delta, \tilde{u}) \leqslant \delta$. В качестве приближенного значения $\tilde{z}$ для некорректно поставленных задач нельзя брать точное решение задачи с входными данными $\tilde{u}$: $\tilde{z} = R(\tilde{u})$. Однако наличие дополнительного параметра δ позволяет искать приближенное решение при помощи параметрических операторов $\tilde{z} = R(\tilde{u}, \alpha)$, если значение параметра α согласовано с точностью δ задания $\tilde{u}^\delta$: $\alpha = \alpha(\delta)$.

Параметрический оператор $R(\tilde{u}, \alpha)$ называется регуляризирующим оператором для $R(u)$, если (1) $R(\tilde{u}, \alpha)$ определен для всех $\tilde{u} \in \tilde{U}$, где $\tilde{U}$ — класс возможных приближенных значений $\tilde{u}$, и если (2) из соотношения $z = R(\tilde{u})$ следует существование функции $\alpha(\delta)$, такой, что из $\rho_u(\tilde{u}^\delta, \tilde{u}) \leqslant \delta$ следует, что $\rho_z(\tilde{z}^\delta, \bar{z}) \leqslant \varepsilon(\delta)$ ($\varepsilon(\delta) \to 0$, $\delta \to 0$), где $\tilde{z} = R(\tilde{u}^\delta, \alpha(\delta))$.

Так, например, при вычислении производной обычное разностное отношение $R(\tilde{u}, \alpha) = 1/\alpha\,[\tilde{u}(x + \alpha) - \tilde{u}(x)]$ является регуляризирующим оператором при $\delta/\alpha(\delta) \to 0$. Таким образом, выбор в качестве приближенного решения $\tilde{z}^\delta$ представляет метод определения устойчивого приближения к $\bar{z}$, хотя задача и является неустойчивой.

Регуляризирующие операторы для широкого круга обратных задач могут быть получены при помощи стабилизации минимума уклонения $\rho_u(Az, u)$.

Пусть на множестве $\widetilde{Z}$ пространства возможных решений Z определен (стабилизирующий) функционал $\Omega[z]$, такой, что множество Z_C, определенное условием $\Omega[z] \leqslant C$ $(z \subset \widetilde{Z})$, компактно в Z. В этом случае значение регуляризирующего оператора $\widetilde{z}^\alpha = R(\widetilde{u}, \alpha)$ может быть определено как элемент $\widetilde{z}^\alpha$, реализирующий минимум функционала

$$M^\alpha[z, \widetilde{A}, \widetilde{u}] = \rho_u(\widetilde{A}z, \widetilde{u})^2 + \alpha\Omega \quad (z \in \widetilde{Z}).$$

Пусть уравнение $Az = \overline{u}$ имеет решение $\overline{z} \in \widetilde{Z}$, и пусть приближенные значения $\widetilde{u}^\delta$ и $\widetilde{A}^\delta$ таковы, что $\rho(\widetilde{u}^\delta, \overline{u}) \leqslant \delta$, $\rho(\widetilde{A}^\delta z, Az) \leqslant \eta(\delta) \times$ $\times f(\Omega(z))$ $(\eta(\delta) \to 0, \ \delta \to 0)$ и $f(\Omega)$ — возрастающая функция. Устанавливается зависимость $\alpha(\delta)$ $(\alpha(\delta) \to 0$ при $\delta \to 0)$ и существование $\delta_0(\varepsilon)$, такого, что $\rho(\widetilde{z}^{\alpha(\delta)}, \overline{z}) \leqslant \varepsilon$ при $\delta \leqslant \delta_0(\varepsilon)$.

Неустойчивые задачи оптимизации находятся в тесной связи с рассмотренными выше задачами. Пусть задана непрерывная функция $F(z)$, определенная в метрическом пространстве Z. Пусть существует единственный элемент z_0, реализующий минимум $F(z)$: $F(z_0) = F_0$. Будем говорить, что задача оптимизации $F(z)$ устойчива, если из $F(z_n) \underset{n\to\infty}{\longrightarrow} F_0$ следует $\rho(z_n, z_0) \to 0$, и что задача оптимизации неустойчива (или некорректно поставлена), если последовательность z_n может расходиться. Пусть $\Omega[z]$ — функционал, удовлетворяющий названным выше условиям и $z_0 \in \widetilde{Z}$. Заменяя функционал $F(z)$ на

$$M^\alpha(z, \widetilde{F}_\eta) = \widetilde{F}_\eta(z) + \alpha\Omega[z], \qquad |F(z) - \widetilde{F}_\eta(z)| \leqslant \eta(\Omega(z) + C),$$

мы получим устойчивую вариационную задачу, при помощи которой строятся последовательности $\widetilde{z}^{\alpha(\eta)}$, такие, что $\rho(\widetilde{z}^{\alpha(\eta)}, \overline{z}) \to 0$ при $\eta \to 0$.

Можно указать естественный класс задач оптимального управления, а также задач линейного и динамического программирования, являющихся неустойчивыми задачами, для которых приведенный выше метод позволяет находить устойчивые приближения.

Приведенные выше методы легко реализуются на электронных вычислительных машинах и представляют эффективный метод решения широкого класса задач.

Московский университет,
Москва, СССР

СОДЕРЖАНИЕ CONTENTS

ВВОДНАЯ ЧАСТЬ INTRODUCTION

ЧАСОВЫЕ ДОКЛАДЫ ONE-HOUR REPORTS

ПОЛУЧАСОВЫЕ ДОКЛАДЫ HALF-HOUR REPORTS

СЕКЦИЯ 1 SECTION 1

СЕКЦИЯ 2 SECTION 2

СЕКЦИЯ 3 SECTION 3

СЕКЦИЯ 4 SECTION 4

СЕКЦИЯ 5 SECTION 5

СЕКЦИЯ 6 SECTION 6

СЕКЦИЯ 7 SECTION 7

СЕКЦИЯ 8 SECTION 8

СЕКЦИЯ 9 SECTION 9

СЕКЦИЯ 10 SECTION 10

СЕКЦИЯ 11 SECTION 11

СЕКЦИЯ 12 SECTION 12

ДОКЛАДЫ, ПОЛУЧЕННЫЕ ПОСЛЕ 1 ЯНВАРЯ 1967 г.
REPORTS RECEIVED AFTER JANUARY 1, 1967

ТРУДЫ
МЕЖДУНАРОДНОГО КОНГРЕССА
МАТЕМАТИКОВ

Редактор *Д. Ф. Борисова, Г. М. Ильичева*
Художник *А. В. Шипов*
Художественный редактор *В. И. Шаповалов*
Технический редактор *В. П. Сизова*

Сдано в производство 18/VIII 1967 г.
Подписано к печати 26/IV 1968 г.
Бумага 60 × 90¹/₁₆ = 23 бум. л.
46 печ. л.
Уч.-изд. л. 46,5. Изд. № 1/4231
Цена 3 р. 43 к. Зак. 1220

ИЗДАТЕЛЬСТВО «МИР»
Москва, 1-й Рижский пер., 2

Московская типография № 16
Главполиграфпрома Комитета по печати
при Совете Министров СССР
Москва, Трехпрудный пер., 9.