Some Points of Analysis and Their History

Lars Gårding

Higher Education Press

American Mathematical Society

Some Points of Analysis and Their History

Lars Gårding

Higher Education Press
Beijing, China

American Mathematical Society
Providence,Rhode Island

(京)112号

ABSTRACT. The purpose of these lectures is to give historical background and leisurely accounts of some important results in analysis in this century. Most of them belong to the classical analysis and the theory of partial differential operators and are associated with Swedish mathematicians, but there is also the Tarski-Seidenberg theorem and Wiener's classical results in harmonic analysis.

The book is of interest to all specialists and students in analysis and partial differential equations.

图书在版编目(CIP)数据

分析学中的若干问题及其历史;英文/(瑞典)戈尔丁(Garding,L.)著.-北京:高等教育出版社,1999

ISBN 7-04-007369-2

Ⅰ.分… Ⅱ.①戈… Ⅲ.数学分析-英文 Ⅳ.O17

中国版本图书馆 CIP 数据核字(98)第 38005 号

书　名	Some Points of Analysis and Their History		
作　者	Lars Garding		

出版发行	高等教育出版社		
社　址	北京市东城区沙滩后街 55 号	邮政编码	100009
电　话	010—64054588	传　真	010—64014048
网　址	http://www.hep.edu.cn		
经　销	新华书店北京发行所		
印　刷	北京印刷一厂		
开　本	787×1092　1/16	版　次	1995 年 5 月第 1 版
印　张	6	印　次	1999 年 5 月第 1 次印刷
字　数	138 000	定　价	8.30 元

Author's Preface

The purpose of these lectures is to give historical background and leisurely accounts of some important results in analysis in this century.Most of the results in classical analysis and the theory of partial differential operators are associated with Swedish mathematicians, but we also include the Tarski-Seidenberg theorem and Wiener's classical results in harmonic analysis, which have demonstrated over time that simple things may lie behind problems that were once very famous and that engendered much work.

It goes without saying that the circle of problems treated here represent just a tiny fraction of the thousands of important results in analysis. Personal affinity rather than systematic selection has determined my sample.

The inspiration for the lectures was an invitation to join the centenary of Wuhan University in 1993.For various reasons I was not able to attend at that time , but I gave some of the lectures when I visited Nankai, Wuhan, Fudan, Jilin and Beijing universities a year later. I want to express my gratitude for the courtesy extended to me by all these universities.

I also thank Professor Li Ta-Tsien for arranging the printing of my lectures and their translation into Chinese. The present expanded version, including lectures on Picard's great theorem, Nevanlinna theory, and a personal essay on the impact of distributions in analysis, has been accepted by the American Mathematical Society and Higher Education Press, P.R.China.[1] Finally, I thank Jana Madjarova for careful proofreading, Natalya Pluzhnikov for expert editing, and Sven Spanne for helping me with a tricky font.

Lars Gårding
Lund, 1997

[1] The original title, "Some problems of analysis and their history", has now been changed to "Some points of analysis and their history".

Contents

CHAPTER 1

Picard's Great Theorem

Introduction

Charles Émile Picard (1856–1941) is famous for Picard's theorem. In its general form it says that a meromorphic function assumes all but two values in any neighborhood of an isolated essentially singular point. Note that a meromorphic function may have poles of finite order and that these are not essential singularities.

In all proofs the assumption that f avoids three different values in a neighborhood of an isolated essentially singular point leads to a contradiction. Picard's own proof was an unexpected fruit of the theory of elliptic functions. His tool was the so-called modular function, the result of half a century of intense study.

Picard's theorem was a radical improvement on Weierstrass's result that an analytic function comes arbitrarily close to any given value in any neighborhood of an isolated essentially singular point, but the theorem seemed mysterious for several reasons. The use of modular functions was out of proportion with the simple formulation of the theorem and gave no hint why precisely two exceptional values were the maximum. In the many later proofs of Picard's theorem the modular function was first eliminated by the use of various inequalities by Borel (1897) and Schottky (1904). Landau (1916, 1929) gave a terse account of Schottky's inequality in his classic *Neuere Ergebnisse der Funktionentheorie.* Rolf Nevanlinna's book (1929) was motivated by Picard's theorem and contains it as a special case of a general theory of the exceptional values of functions which are meromorphic outside a compact set, the point infinity excepted. Nevanlinna's theory will be sketched in Chapter 4. Finally, in (1935b) Lars Ahlfors gave a topological explanation why there are at most two exceptional values.

Tha aim of this paper is to present or at least sketch some of the proofs in this turn of events, starting with Picard's own proof.

Picard's proof

The task is to prove that a meromorphic function cannot avoid three values in any neighborhhod of an isolated essentially singular point. Here the exceptional values may be taken to be $0, 1, \infty$ so that it suffices to consider an analytic function at an isolated essentially singular point which does not take the values 0 and 1. In fact, if the values not taken are a, b, c, the function

$$g(z) = \frac{f(z) - a}{f(z) - b} : \frac{c - a}{c - b}$$

avoids the values $0, 1, \infty$, even if one of a, b, c already is infinity, and if f has an isolated essential singularity at z_0, so does g and conversely.

Picard proved two versions of his theorem. The first one (1879) says that an entire function whose range avoids two separate complex numbers is a constant. The main theorem was proved one year later (1880). In his papers he could just refer to known properties of modular functions. For completeness we shall now describe the one he used.

The modular function

The elliptic integral

$$u = \int_0^y \frac{1}{\sqrt{(1-t^2)(1-\kappa^2 t^2)}}dt$$

where the *module* κ^2 is not $0, 1, \infty$, defines Jacobi's elliptic function $y = s(\kappa, u)$, the *sinus amplitudinis.* For $0 < \kappa^2 < 1$ it has two canonically defined periods σ_1, σ_2 obtained by integration along certain closed cycles on the Riemann surface of the curve $y^2 = (1 - x^2)(1 - \kappa^2 x^2)$. All other periods are then linear combinations of these two periods, which can always be chosen so that the quotient $\omega = \sigma_1/\sigma_2$ has a positive imaginary part.

When the module $z = \kappa^2$ avoids the values $0, 1, \infty$, the quotient $\omega = \omega(z)$ is an analytic many-valued function of $z = \kappa^2$ with values in the upper half-plane. Under closed loops, $\omega(z)$ is subject to certain Möbius transformations which form a discrete group Γ. More precisely, the images under ω of the lower and upper half-planes form a tesselation of the upper half-plane by non-Euclidean triangles with all three corners on the real axis. The inverse of ω is a function from the upper half-plane to itself which is automorphic in the sense that it is invariant under Γ.

Picard's two papers

In his first paper Picard used only the fact that the quotient $\omega = \omega(z)$ is analytic and many-valued with values in the upper half-plane when z is not equal to $0, 1, \infty$. After a Möbius map which makes an entire function $f(z)$ avoid the points $0, 1$, the proof of Picard's first theorem is now obvious: $\omega(f(z))$ can be continued analytically everywhere in the complex plane, hence it is a single-valued entire function with range in the upper half-plane and must be constant so that f is constant.

Very soon afterwards Picard could prove also his second theorem by a variation of the same trick. Actually, Picard's proof is elementary modulo the existence of an automorphic function defined in the upper half-plane. It is difficult to read only because the author uses the theory of Möbius maps in a complicated way. This was before the present canonical theory of linear algebra, and it is hoped that the rendering of the proof below is more readable. For completeness the text includes a sketch of the classical construction of automorphic functions.

Möbius maps

A Möbius map is an invertible fractional linear map

$$z \to \frac{az+b}{cz+d}, \quad ad - bc = 1,$$

with complex coefficients. When A is the matrix $(a,b)/(c,d)$, it is convenient to write the right side above as

$$A[z] = \frac{az+b}{cz+d}.$$

It is immediate to verify that A is invertible and that

$$AB[z] = (AB)[z].$$

Reflection $z \to w$ in a circle with center α and radius r is given by the formula

$$\overline{(z-\alpha)}(w-\alpha) = r^2,$$

and it is known by elementary geometry that reflections map circles to circles. Our map may be written as $w = A[\bar{z}]$ with a certain invertible matrix A. Hence every reflection in circles (including straight lines) is an improper Möbius map, that is, a Möbius map preceeded or followed by a conjugation. The product of two such maps is a Möbius map. In fact, all proper and improper Möbius maps form a group, the full Möbius group M, generated by reflections. All its elements map circles into circles.

Tesselations of the upper half-plane. Automorphic functions

A triangle bounded by circular arcs which touch each other at the corners so that all corner angles vanish needs a simple name. Let us call it a vanishing triangle. Automorphic functions are closely connected with tesselations of the upper half-plane H by vanishing triangles with all corners on the real axis. Such a tesselation may start with a triangle K in H bounded by the lines $x = 0$ and $x = 1$ and the half-circle $|z - 1/2| = 1/2$. Reflections in the sides will then give three adjoining vanishing triangles whose sides like those of K meet the real axis under right angles and belong to H. Repeated reflections will then produce a tesselation of the upper half-plane.[1] At the same time they generate a subgroup G of the full Möbius group which maps the tesselation to itself. This group is discrete in the sense that if $A, B \in G$ and $z \in H$, then $A = B$ when Az and Bz are sufficiently close.

We can now construct automorphic functions simply by using Riemann's mapping theorem to map K conformally to the upper half-plane H by a function φ which maps the corners $0, 1, \infty$ to themselves. If a reflection R maps K to any of its neighbors, then, by Schwarz's reflection principle, $\varphi(z) = \varphi(R^{-1}z)$ extends φ to RK across their common boundary in such a way that RK is mapped to the lower half-plane. Continuing this process we have a function $I(z)$ defined in the upper half-plane and invariant under G. The corners of the tesselation are all mapped to one of the points $0, 1, \infty$.

The inverse function $J(z)$ is many-valued but has the crucial property of being singular only at the points $0, 1, \infty$. When z runs through a closed path γ from $z_0 \in H$ back to z_0 which avoids these points, then $J(z)$ assumes a new value $A[J(z_0)]$ for some $A \in G$. Since γ crosses the real line an even number of times, we are sure that A is a Möbius map.

[1]The reader is advised to draw a figure himself or to look up a corresponding figure in some standard treatise.

Normal forms of Möbius maps of H

In our version of Picard's proof of his theorem, we shall need to know the normal forms of Möbius maps $z \to A[z]$ of H to itself. Then the real axis is mapped to itself so that A may be assumed to be real. Further, since

$$\operatorname{Im} \frac{az+b}{cz+d} = (ad - bc)\frac{\operatorname{Im} z}{|cz+d|^2},$$

the determinant $ad - bc$ is positive. We normalize it to 1. Then the eigenvalues of A have the form $\lambda, 1/\lambda$ and, since their sum is real, they are either both real or both of absolute value 1.

The possible normal forms of A under similarity maps $A \to SAS^{-1}$ are as follows:

1. Two complex eigenvalues, S maps H to the unit circle and $SAS^{-1} = D$ is diagonal with non-real elements $e^{i\theta}, e^{-i\theta}$.
2. Two real eigenvalues $\lambda > 1$ and $1/\lambda$, S maps H to itself, SAS^{-1} is diagonal with elements $\lambda, 1/\lambda$.
3. Two eigenvalues equal to 1, A is not diagonalizable, but there is an S which maps H to itself such that SAS^{-1} is the matrix $(1,1)/(0,1)$.
4. A is the unit matrix.

Proof of Picard's theorem

Using the elementary statements of the previous section we can now prove the

THEOREM. *A function $f(z)$ which is analytic and single-valued when restricted to a neighborhood N of ∞ can avoid at most one value.*

In the proof we may assume that $f(z)$ is never 0 or 1 and use the inverse $J(z)$ of the automorphic function $I(z)$ defined above. The following lemma is taken as a matter of course by Picard.

LEMMA. *With f as above, the function $g(z) = J(f(z))$ is analytic in a connected neighborhood N of ∞ with values in H and, under a turn T in N in positive direction around the origin,*

$$Tg(z) = A[g(z)] \tag{1}$$

for some Möbius map $A \in G$.

PROOF. Since $f(z)$ is never $0, 1, \infty$, $g(z)$ can be continued analytically and indefinitely in N. Also, if T refers to a path $\gamma \subset N$ in positive direction from a point z_0 and back, it is clear that (1) holds with $z = z_0$ and some $A \in G$. A slight modification of γ will change A to some $A' \in G$ close to A, but since G is discrete, $A'z_0$ cannot come arbitrarily close to Az_0 unless $A' = A$. Hence A does not depend on the choice of γ. Similarly, it cannot depend on the choice of z_0. The last statement follows from Weierstrass's theorem.

In the rest of the proof we shall see that (1) leads to situations where the range of $f(z)$ for large z cannot be dense in the upper half-plane contradicting Weierstrass's theorem.

1. Suppose that A has complex eigenvalues $e^{i\theta}, e^{-i\theta}$ and let S be a diagonalizing matrix mapping H to the unit disk. We may assume that $0 < \theta < \pi$. Then

$$S[Tg(z)] = e^{2i\theta} S[g(z)]$$

so that

$$S[g(z)] = z^{\theta/\pi} h(z)$$

where $h(z)$ is single-valued and $|S[g(z)]| \leq 1$. This is possible only if $h(z) = O(1/z)$ and then the left side tends to zero as $z \to \infty$. But then $g(z) = J(f(z))$ has a limit in the upper half-plane as $z \to \infty$ so that f cannot be singular at infinity.

2. Suppose that A has real eigenvalues $\lambda > 1$ and $1/\lambda$ so that $SH \subset H$. Now $z^{\log\lambda/\pi i}$ changes by a factor of λ^2 under T and hence

$$S[T^n g(z)] = z^{n \log\lambda/\pi i} h(z)$$

where $h(z)$ is single-valued and the left side belongs to H. Here, since $\log \lambda > 0$, we can put $z = e^{m/\log\lambda}$ where $m > 0$ is a large integer and then the range of

$$e^{nr \log\lambda/\pi i} = e^{nm/\pi i}$$

is dense in the unit circle when n varies. Hence $g(z)$ is not in H and this is a contradiction.

3. We may suppose that $SA[w] = w + 1$ and that $SH \subset H$. Then

$$S[Tg(z)] = \frac{\log z}{2\pi i} + h(z) \tag{2}$$

for all n where $h(z)$ is analytic and single-valued for large arguments. Hence

$$e^{iS[Tg(z)]} = z^{1/2\pi} e^{ih(z)}$$

where the left side is bounded and its range dense in the unit disk for all regions $|z| >$ const. But then $e^{ih(z)}$ tends to zero at least as $1/z$ and this is a contradiction.

4. Suppose that $g(z)$ is single-valued. Then this function cannot have values in H unless it is regular at infinity, and this means that $g(z)$ tends to a limit as $z \to \infty$. But the range of $g(z) = J(f(z))$ is dense in the range of J in every neighborhood of ∞, whence a contradiction.

The proofs by Borel and Schottky

Picard's proof of Picard's theorem explores the absurd consequences of the assumption that there exists an entire function which avoids two separate values or the absurd consequences of the existence of a function which avoids three values in the neighborhood of an isolated essential singularity and is meromorphic outside. The ensuing proofs of Borel (1897) and Schottky (1904) avoid the theory of elliptic functions. Borel's proof, which is simple only in principle, only concerns Picard's first theorem about entire functions.

In the first edition of his classic *Leçons sur les fonctions entières* (1900) Emile Borel devoted a chapter to Picard's theorem and almost proved its analytic version by a very simple argument using the concept of growth of entire functions. A simple paraphrase of Borel's argument runs as follows.

As has been remarked before, it suffices to consider an entire function $f(z)$ which is never 0 or 1. Then $f(z) = e^{g(z)}$ for some entire function $g(z)$ never equal to an integral multiple of $2\pi i$ and hence also

$$f(z) = e^{-2\pi i g(z)}$$

where g does not take integral values, in particular not 0 or 1. Hence, if

$$M(f,r) = \max_{|z|=r} |f(z)|, \quad A(f,r) = \max_{|z|=r} \operatorname{Im} f(z),$$

we must have

$$M(f,r) \leq e^{2\pi A(g,r)}.$$

Now the value of f at a point w with $|w| = r' < r$ may be explicitly expressed by an integral of $\operatorname{Im} f$ over a circle $|z| = r$ plus a term $\operatorname{Re} f(0)$. Hence, for instance,

$$M(f, r/2) \leq \text{const}\, A(f,r) + |f(0)|$$

so that, since $M(f,r)$ tends to infinity with r,

$$M(g, r/2) = O(\log M(f,r)).$$

In particular, if $M(f,r) = O(e^{r^m})$ for some $m > 0$, then $M(g,r) = O(r^m)$. But then g is a polynomial and assumes all values, which is a contradiction. If $\exp^{(n)}$ denotes the function exp iterated n times, the same argument and induction show that an entire function f which does not take two values cannot have a bound

$$|f(z)| \leq \exp^n(O(|z|^m))$$

for any integer $n > 0$ and number $m > 0$. It follows that no entire function of reasonable growth can avoid two values. This is also the point where Borel stops in his lectures.

In the beginning of the century Schottky (1904) gave the first proof after Borel of Picard's theorem without using a modular function. It was followed by a flurry of papers by Landau, Hurwitz, Caratheodory and others. In the second edition (1929) of his book *Ergebnisse* Landau gives a number of properties of an analytic function in a disk which avoids the values 0 and 1. One of them, called the theorem by Schottky, says that a function f which is regular in the unit disk and does not assume the values 0 and 1 has a bound for $|z| < \theta < 1$ which only depends on θ and a bound of $|f(0)|$ away from 0 and infinity.

From this theorem, Picard's theorem can be deduced as follows. Assume that $F(z)$ is analytic for $0 < |z| < 1$, has an essential singularity at the origin and is never equal to 0 or 1. Put

$$F(e^t) = g(t)$$

so that $g(t)$ is defined when $\operatorname{Re} t < 0$ and is periodic with the period $2\pi i$. By Weierstrass's theorem there is a sequence of radii r_n tending to zero and points z_n

on the corresponding circles where $|F(z_n) - 2| < 1/2$. Putting $z = e^t$ we then have $|g(t_n) - 2| < 1/2$ where $\operatorname{Re} t_n = \log r_n$ tends to $-\infty$. Now the function

$$h(u) = g(t_n + 4\pi u)$$

is analytic when $|u| < 1$, it is never 0 or 1 and $|h(0) - 2| < 1/2$. Hence, by Schottky's theorem, h has an absolute bound when $|u| < 1/2$ and this suffices to cover an interval between t_n and $t_n + 2\pi i$. Hence the function $F(z)$ has a uniform bound on all circles $|z| = r_n$ and this is a contradiction.

The shortest proof of Schottky's theorem uses again the inverse $J(z)$ of the modular function $I(z)$ defined above. In fact, let $f(z)$ be a function analytic in the unit disk with a fixed $a = f(0)$, assume that f does not take the values 0 and 1 and consider the function

$$g(z) = J(f(z))$$

which maps the unit disk to the upper half-plane and so has the form

$$g(z) = \frac{ce^{i\theta} - \bar{c}z}{e^{i\theta} - z}, \quad \operatorname{Im} c > 0,$$

with

$$\operatorname{Im} g(z) = \frac{\operatorname{Im} c(1 - |z|^2)}{|1 - e^{i\theta} z|^2}.$$

Hence if $\operatorname{Im} g(z)$ tends to zero for some z in a closed disk $D : |z| \leq b < 1$, then $\operatorname{Im} c$ tends to zero and hence $\operatorname{Im} g(z)$ tends to zero uniformly for all $z \in D$. Similarly, if $g(z)$ tends to infinity for some $z \in D$, then c tends to infinity and hence $g(z)$ tends to infinity uniformly for all $z \in D$.

Now consider a family F of functions f analytic in the open unit disk for which $a = f(0)$ stays in a compact set not containing 0 and 1. With z restricted to the disk $D : |z| \leq b < 1$, suppose that $f(z)$ comes very close to $0, 1$ or is very large for some $f \in F$ and some $z \in D$. Then $\operatorname{Im} g(z)$ must come very close to the real axis or be very large and hence all of $g(D)$ has this property uniformly. But this contradicts the assumption about $f(0)$ and this proves Schottky's theorem. A modern proof where the topological content of the theorem is evident is available in Nevanlinna (1953).

Ahlfors's topological proof

Ahlfors's paper (1935b) which gave him a Fields medal is actually a topological proof of the essential part of Nevanlinna's theory. We shall now sketch the idea of the proof and how it can be used to prove Picard's theorem that a function $f(z)$ which is analytic outside a circle $|z| = r_0$ can avoid at most one value without being meromorphic at infinity.

Ahlfors's proof uses the Euler index, i.e. the number of corners minus the number of lines plus the number of triangles in a triangulation of a two-dimensional set. The Euler index is known to be a topological invariant. For a bounded set in the plane it equals $1 - q$ where q is the number of holes in the set.

The basis of the proof is a theorem by Hurwitz about covering maps $T : \bar{S} \to S$ of two-dimensional compact manifolds. If $T\bar{S}$ covers S N times, the theorem says that

$$\chi(\bar{S}) = N\chi(S) - \sum(\nu(P) - 1)$$

where χ is the Euler index, P runs through the points of S and $\nu(P)$ is the number of points of $\bar{S}$ over P. The proof[2] is immediate if we use all multiple points as corners in a triangulation of $\bar{S}$. It follows as a special case that

$$\chi(\bar{S}) \leq N\chi(S).$$

Imagine now f as a map from the region $S_0 : R < |z| < \infty$ to a region S consisting of the complex plane C minus q points. We then have $\chi(S) = 1 - q$. Also, $\chi(S_0) = 0$, and this is also the Euler index of the image $\bar{S} = f(S_0)$. Hence, if we apply Hurwitz's theorem to this non-compact situation, we get

$$0 \leq N(1 - q)$$

with some large N, perhaps infinity, and this means that $q \leq 1$. This reasoning is of course complete nonsense, but in (1935b) Ahlfors got precisely this last inequality as a special case. Roughly speaking he arrived at this result by taking restrictions of f to ring-shaped regions $r_0 \leq |z| \leq r$ with a large r, by taking boundaries into account and by replacing N by a quotient of spherical areas. A fuller but not complete account of Ahlfors's arguments is given at the end of the chapter on Nevanlinna theory.

Bibliography

AHLFORS L.

1935. a. *Über eine neue Methode in der Theorie der meromophen Funktionen*, Soc. Sci. Fenn. Comm. Phys.-Math. **8** (1935), no. 10, 1–14.

1935. b. *Zur Theorie der Überlagerungsflächen*, Acta Math. **65** (1935), 157–194.

BOREL E.

1897. *Sur les zéros des fonctions entières*, Acta Math. **20** (1957), 357–396.

1904. *Leçons sur les fonctions entiéres*, Paris, 1900.

LANDAU E.

1929. *Darstellung und Begründung einiger neuerer Ergebnisse der Funktionentheorie*, Berlin 1929.

NEVANLINNA R.

1929. *Le théorème de Picard-Borel et la théorie des fonctions méromorphes*, Coll. Borel, Paris, 1929.

1936. *Eindeutige Analytische Funktionen*, Springer, 1936, zw. Auflage, 1953.

PICARD E.

1879. *Sur une propriété des fonctions entiéres*, Oeuvres I, p. 19.

1880. *Mémoire des fonctions entières*, Oeuvres I, pp. 39–60.

SCHOTTKY F.

1904. *Über den Picard'schen Sats und die Borel'schen Ungleichungen*, Sitzungsber. Der Kgl. Preussischen Akad. d. Wiss. Berlin Jahrgang 1904, pp. 1244–1262.

[2]The reader is advised to think through this argument.

CHAPTER 2

On Holmgren's Uniqueness Theorem

Introduction

The basic result about analytic solutions of partial differential equations with analytic coefficients is the Cauchy-Kovalevskaya theorem. Although it applies in great generality for non-linear equations, let us state it for the first order linear operator

$$P(x, D) = A_1(x)D_1 + \cdots + A_n(x)D_n + B(x), \quad D_k = \partial/\partial x_k,$$

where $A_1(x), \ldots, A_n(x)$ are $m \times m$ matrices which are analytic in the real variables $x = (x_1, \ldots, x_n)$. Let $P_1 = \sum A_k D_k$ be the principal part of P. A hypersurface $S : s(x) = 0$ is said to be characteristic for P at a point x if $P_1(x, s_x)$ is not invertible. By the Cauchy-Kovalevskaya theorem, the boundary problem

$$Pu = v, \quad u = w \quad \text{when} \quad s(x) = 0$$

has a local analytic solution u at non-characteristic points of S when P, w, s are analytic. In a sense, this result parametrizes the solutions of $P(x, D)u = v$ close to a non-characteristic point of S.

On the turn of the century there was a growing interest in classes of functions which are not analytic, only sufficiently differentiable in some sense, for instance those with continuously differentiable derivatives up to some order. One mathematician working in this area was the Swede Erik Holmgren, later professor at Uppsala university. He had the bright idea to ask himself what happens in the linear Cauchy-Kovalevskaya theorem when the solution u is not analytic and only sufficiently differentiable. The answer is given by Holmgren's uniquenss theorem (1901): if the solution exists, it is unique.

This lecture gives the simple proof of Holmgren's theorem followed by an example that it fails for non-analytic coefficients and a simple, recent example by Métivier that it fails also for non-linear systems with analytic coefficients. In a final section, the theorem is extended to non-analytic operators of the above form which are elliptic and almost commute with their adjoints. It illustrates the use of weight functions first introduced by Torsten Carleman (1939) in a uniqueness proof for first order systems in two variables. His result was extended to several variables by Calderón (1958). We shall follow the treatment in Hörmander (1985 III) which is close to Carleman's. Hörmander's complete presentation of Calderón's theorem (1985 IV) is beyond the scope of a simple lecture.

At present, Holmgren's uniqueness theorem is just a convenient heading for various developments connected with the original result (see Hörmander (1994)).

Proof of Holmgren's uniqueness theorem

By an analytic change of the variables x and a linear change of the unknown function u we may assume that the coefficients of $P(x, D)$ are analytic at the origin, that S is given by $x_1 = x_2^2 + \cdots + x_n^2$ and that $A_1(x)$ is the unit matrix E. It suffices to prove that u vanishes close to the origin when $Pu = 0$, $u = 0$ on S. Let (u, v) be the Euclidean scalar product and let

$$P'(x, D) = -\sum D_k A_k'(x) + B'$$

be the adjoint of P so that $(Pu, v) - (u, P'v) = \sum D_k(A_k u, v)$. If we integrate over a region $K = K(c) : c > x_1 > x_2^2 + \cdots + x_n^2$ with upper boundary K_+ where $c = x_1$ and put $x' = (x_2, \ldots, x_n)$, we get

$$\int_{K_+} (u, v) dx' = \int_K (u, P'v) dx.$$

Here we can let v be an analytic solution of $P'v = 0$ with data on K_+. Since these quantities can be given arbitrarily, u must vanish on every K_+. Varying the size and position of $K(c)$ shows that $u = 0$ close to the origin.

No uniqueness

For non-analytic coefficients Holmgren's uniqueness theorem is no longer true and several counterexamples were constructed in the 1950's. A general construction is presented in Hörmander (1983), a simple example of which is the following: there is a C^∞ function $a(t, x)$ which vanishes for $t \leq 0$ and whose support contains the origin such that the equation

$$\partial_t u + a \partial_x u = 0$$

has a solution $u = f(t, x) \neq 0$ which vanishes for $t \leq 0$. It was remarked by Métivier (1993) that this permits construction of a non-linear *analytic* system of equations

$$\partial_t u + v \partial_x u = 0, \quad \partial_t v + \partial_y v = 0$$

for which Cauchy's problem with data on $t = 0$ has two different solutions u, v which coincide for $t - 0$. It suffices to put either

$$u = g(y) f(t - y, x), \quad v = a(t - y, x)$$

where $g \in C^\infty$ is supported in $y \geq 0$ and vanishes otherwise, or

$$u = 0, \quad v = a(t - y, x).$$

The two solutions are smooth and different, but they are equal when $t = 0$. Hence Holmgren's theorem cannot be true for non-linear analytic systems. Métivier also has a couple of other counterexamples.

Uniqueness for non-analytic coefficients

Holmgren's result is much more difficult to prove for operators with non-analytic coefficients. The first proof in this case is due to Carleman (1939). He treated linear first order operators in two variables of the form

$$P(x,D) = D_1 u + A_2(x)D_2 + B(x), \quad D_k = \partial/i\partial x_k,$$

(since we shall use the Fourier transform later, we now use the imaginary gradient D). It is important that the square matrix $A_2(x)$ may be uniformly diagonalized so that, by a change of variables, Carleman could assume that $A_2(x)$ is already diagonalized. This implies in particular that

$$(1) \qquad [P(x,D), P^*(x,D)] = O(|u(x)||Du(x)|), \quad P^*(x,D) = \bar{P}'(x,D),$$

provided the coefficients are uniformly Lipschitz continuous. We shall assume (1) also in the general case where

$$P(x,D) = \sum_1^n A_k(x)Dx_k + B(x)$$

and the coefficients are $m \times m$ Lipschitz continuous matrices. When the coefficients of $P(x,D) = P(D)$ are constant and (1) holds, the right side vanishes, so that $P(D)$ and $P^*(D)$ commute. When $n > 2$ it seems difficult to imitate Carleman by making preliminary changes of the coefficients in order to achieve (1).

Our proof of the uniqueness theorem below is parallel to the proof in Hörmander (1985 III) of a corresponding result in the scalar case where the analogue of (1) is automatic.

Outline of the theorem and a proof for constant coefficients

Let K be the region

$$K : c > x_1 > b(x_2^2 + \cdots + x_n^2), \quad 1 > b, c > 0.$$

We shall consider solutions u of the inequality

$$(*) \qquad |P(x,D)u| = O(|u|), \quad x \in K,$$

for small x_1. Here u with locally square integrable derivatives is assumed to vanish below the lower boundary $S : x_1 = b(x_2^2 + \cdots + x_n^2)$ of K. The coefficients $A_k(x), B(x)$ are supposed to be bounded and Lipschitz differentiable in $\bar{K}$. We shall find conditions under which such a u vanishes for small x_1. The main tool of the proof is the use of a function $h(x) = x_1 - x_1^2$ for small $x_1 > 0$ and a corresponding change of the unknown function,

$$v = e^{-h(x)}u, \quad u = e^{h(x)}v.$$

Then $D_j e^h v = e^h(D_j - ih_j)v$ where $h_j = \partial_j h$ and

$$\int_K e^{-2h}|P(x,D)u|^2 dx = \int_K |P(x, D - i\partial h)v|^2 dx.$$

Estimates referring to such norms were first called Carleman's estimates by Lars Hörmander.

Let us now note that the commutator

$$[D_j - ih_j(x), D_k + ih_k(x)]$$

vanishes unless $j = k = 1$, in which case it equals $h_{11}(x) = -2$. From this simple formula and (1) follows the main ingredient of our future proof, namely the identity

$$\int_K |P(x, D - \tau ih'(x))v(x)|^2 dx = \int_K |P^*(x, D + \tau ih'(x))v(x)|^2 dx + 2\tau \int_K |v(x)|^2 dx + \int_K O(|v(x)||Dv(x)|)dx \tag{2}$$

where $\tau > 0$. When P has constant coefficients, this holds without an error term and the desired uniqueness follows by inserting (*). In fact, then

$$2\tau \int_K e^{-2\tau h(x)}|u(x)|^2 dx \leq \int e^{-2\tau h(x)} O(|u(x)|^2) dx.$$

Hence, letting τ tend to infinity, it follows that $u = 0$ in K.

Permitted simplifying assumptions

In the general case, one can try to find properties of $P(x, D)$ besides (1) which make the first integral on the right in (2) so positive for large τ that the kind of argument just given goes through. Before proceeding further we shall now state some permitted simplifications which influence neither the assumption (*) nor the conclusion of the theorem.

1) $B(x) = 0$.

2) The solution u has compact support close to the origin.

3) $P(x, D)$ is replaced by

$$P_\varepsilon(x, D) = \varepsilon P(\varepsilon x, D_{\varepsilon x}) = \sum A_k(\varepsilon x) D_k, \quad \varepsilon > 0.$$

To see 2), replace u by κu where $\kappa(x) \in C_0^1$ equals 1 for small x. The effect of 3) is to make the error term of (1) small when P is replaced by P_ε. In the sequel we shall make tacit use of these assumptions.

Ellipticity assumption and the full proof

Besides (1) we now assume that $P(0, D)$ is elliptic at the origin, i.e. that

$$|P(0, \xi)a|^2 \geq C|\xi|^2|a|^2, \quad a \in C^n. \tag{1'}$$

With $N = h'(0) = (1, 0, \ldots, 0)$, this means in particular that

$$|P^*(0, \xi + i\tau N)a|^2 + \tau^2|a|^2 \geq C(|\xi|^2 + \tau^2)|a|^2$$

with another constant $C > 0$. In fact, the left side vanishes only when $a = 0$ and it is homogeneous of order two in a and (ξ, τ). By a Fourier transformation, some easy estimates and a passage to P_ε^*, this shows that

$$\int (|P_\varepsilon^*(x, D + i\tau h'(x))v(x)|^2 + \tau^2|v|^2) dx \geq C \int (|Dv|^2 + \tau^2|v|^2) dx$$