Printed at the Mathematical Centre, 49, 2e Boerhaavestraat, Amsterdam.

The Mathematical Centre, founded the 11-th of February 1946, is a non-profit institution aiming at the promotion of pure mathematics and its applications. It is sponsored by the Netherlands Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O).

MC SYLLABUS 41

# COLLOQUIUM NUMERICAL TREATMENT OF INTEGRAL EQUATIONS

H.J.J. te RIELE (ed.)

MATHEMATISCH CENTRUM AMSTERDAM 1979

1980 Mathematics subject classification: 65R20, 45L10, 78A30, 78A45, 92A15, 45Mxx, 76Bxx

ISBN 90 6196 189 0

# CONTENTS

CONTENTS v

PREFACE vii

PART I, GENERAL NUMERICAL METHODS

1. H.J.J. te Riele
   Introduction and global survey of numerical methods for integral equations. 1-25
2. H. Schippers
   Multi-grid techniques for the solution of Fredholm integral equations of the second kind. 27-46
3. H.J.J. te Riele
   Numerical methods for first kind Fredholm integral equations. 47-65
4. J.J.M. Cuppen
   Regularization methods and parameter estimation methods for the solution of Fredholm integral equations of the first kind. 67-85
5. R.W. de Vries
   Invariant imbedding methods for integral equations. 87-96

PART II, APPLICATIONS

6. G. De Mey
   Numerical application of integral equations in semiconductor physics. 97-114
7. O. Diekmann
   Integral equations and population dynamics. 115-149
8. P.J. Zandbergen
   Over de integraalvergelijking van de dragende vlak theorie. 151-169
9. Th.E. Labrujère
   Integral equation methods for multi-element airfoil analysis and design. 171-219
10. G. Mur
    Computational aspects of integral-equation methods for the scattering of time-harmonic waves by penetrable obstacles. 221-240
11. J.C.W. Berkhoff
    Diffraction of water waves. 241-258

ADDRESSES OF CONTRIBUTORS 259

# PREFACE

During the period of October 1978 - May 1979 the Department of Numerical Mathematics of the Mathematical Centre organized a Colloquium which was devoted entirely to the Numerical treatment of integral equations.

The purpose of this Colloquium was to give insight into the large amount of existing numerical methods for the solution of integral equations and related equations, like integro-differential equations. Moreover, much attention was paid to concrete practical problems, as an attempt to narrow the gap between, on the one side, the field of numerical mathematics which aims at the design, the analysis and the implementation of numerical methods for *classes* of equations and, on the other hand, the areas of application where often special problems arise, which require a special treatment, and which are not amenable for treatment by general purpose numerical methods.

This syllabus comprises the extended versions of eleven lectures presented at the monthly meetings of the Colloquium. We have divided them into two parts: I. General numerical methods (5 papers); and II. Applications (6 papers, in the fields of semiconductor physics, population dynamics, fluid mechanics (airfoil analysis and design, diffraction of water waves, subsonic lifting surface theory) and electromagnetic scattering). The papers of three other lectures have already appeared elsewhere and, therefore, have not been included in this Syllabus. They are: H.K. Kuiken, Heat or mass transfer from an open cavity, J. Engng. Math., 12 (1978), pp. 129-155; P.J. van der Houwen (speaker) and P.H.M. Wolkenfelt, On the stability of multistep formulas for Volterra equations of the second kind, Report NW 59/78, Math. Centre, 1978; P.J. van der Houwen, H.J.J. te Riele and P.H.M. Wolkenfelt (speaker), On the stability of multistep formulas for systems of Volterra integro-differential equations, Report NW 63/78, Math. Centre, 1978.

Finally, we would like to thank all those who have contributed to the realization of the Colloquium and of this Syllabus, especially the guest speakers J.C.W. Berkhoff, J.J.M. Cuppen, G. De Mey, O. Diekmann, H.K. Kuiken, Th.E. Labrujère, G. Mur, R.W. de Vries and P.J. Zandbergen.

H.J.J. te Riele

# PART I, GENERAL NUMERICAL METHODS

H.J.J. te Riele (ed.)

Colloquium Numerical treatment of integral equations

MC Syllabus 41

C O R R I G E N D U M

The title-pages 1 and 47, belonging to the first and the third paper, respectively, should be interchanged.

# NUMERICAL METHODS FOR FIRST KIND FREDHOLM INTEGRAL EQUATIONS

H.J.J. te RIELE

## 1. INTRODUCTION

Integral equations occur very frequently in mathematical models which describe some real world phenomena, and the mathematical literature of the last decade reflects a growing interest in the (numerical) solution of these equations. The mathematical *theory* of integral equations is well-developed (cf. the references given in the first chapter of DELVES & WALSH [20]), and one should not study the numerical treatment of integral equations before knowing the most important results from the theory. Only in the last years *books* begin to appear which are devoted to the *numerical solution* of integral equations.

We intend to give here an introductory survey of the most important numerical methods which have been developed in the past years for various classes of integral equations. Unless otherwise indicated, we assume existence and uniqueness of the solution of the integral equation. Much attention will be paid to an explanation of how the methods work, and to the numerical problems which can occur in the actual application of the methods. For simplicity, we shall confine ourselves mostly to *linear* integral equations. For a further study one is referred to the books of ATKINSON [4], BAKER [5], DELVES & WALSH [20], IVANOV [30], JASWON & SYMM [31] and TIHONOV & ARSENIN [55], and the references given there to the original papers. Three survey papers deserve to be mentioned here explicitly, viz. FOX & GOODWIN [22], NOBLE [44] and NOBLE [46], the first two of which can now safely be called *classical*.

## 2. CLASSIFICATION OF INTEGRAL EQUATIONS

Generally spoken, an integral equation is a functional equation where the unknown function $f(x)$ occurs under an integral sign. There is a wide variety of different kinds of integral equations, which makes a classification difficult. Nevertheless, one can safely say that the following four classes play an independent and dominant rôle in the theoretical and numerical literature on integral equations.

$$g(x) = \int_a^b K(x,y)f(y)dy, \qquad a \le x \le b, \tag{2.1}$$

called a *Fredholm integral equation of the first kind*;

$$f(x) = g(x) + \int_a^b K(x,y)f(y)dy, \qquad a \le x \le b, \tag{2.2a}$$

called a *Fredholm integral equation of the second kind*; closely related to this equation is the *eigenvalue problem*

$$\lambda f(x) = \int_a^b K(x,y)f(y)dy; \tag{2.2b}$$

$$g(x) = \int_0^x K(x,y)f(y)dy, \qquad 0 \le x \le b, \tag{2.3}$$

called a *Volterra integral equation of the first kind*;

$$f(x) = g(x) + \int_0^x K(x,y)f(y)dy, \qquad 0 \le x \le b, \tag{2.4}$$

called a *Volterra integral equation of the second kind*.

In these equations the "kernel" $K(x,y)$ and the "driving term" $g(x)$ are given functions. Both *scalar* equations and *systems* of equations occur; in the latter case f, g and K are vector functions. All four classes of equations are *linear* in f. *Nonlinear* equations arise when instead of $K(x,y)f(y)$ we have $K(x,y,f(y))$ under the integral sign, where the latter function is nonlinear in f.

Equations which arise frequently in practice are those with a *weakly singular* kernel, e.g., $K(x,y) = H(x,y)|x-y|^{-\alpha}$, $0 < \alpha < 1$, where $H(x,y)$ is continuous. Also $\alpha = 1$ can occur, but then $\int_a^b H(x,y)dy/(x-y)$ is to be interpreted as a *Cauchy principal value*, i.e., the limit, if it exists,

$$\lim_{\varepsilon \downarrow 0}\{\int_a^{x-\varepsilon} \frac{H(x,y)}{x-y}\,dy + \int_{x+\varepsilon}^b \frac{H(x,y)}{x-y}\,dy\}.$$

Another important class of integral equations are those where the unknown function depends on two or three variables.

There is a well-known close connection between integral equations of the second kind and differential equations. First, consider the initial-value problem $f''(x) + g(x)f(x) = h(x)$, $x \in [0,b]$, with $f(0) = f_0$ and $f'(0) = f_0'$. Integrating the differential equation twice, replacing the resulting double integral by the corresponding single integral, and using the given initial conditions, we obtain the Volterra integral equation of the second kind

$$f(x) + \int_0^x (x-y)g(y)f(y)dy = f_0 + xf_0' + \int_0^x (x-y)h(y)dy, \quad x \in [0,b].$$

From the character of numerical methods for solving initial value problems, and from the above relationship, it follows that methods for solving Volterra integral equations will generally be based on a *recursive* process.

Consider now the boundary value problem $f''(x) + g(x)f(x) = h(x)$, $x \in [a,b]$, with $f(a) = f_a$ and $f(b) = f_b$. Proceeding in a similar way as before, this problem can be rewritten as a Fredholm integral equation of the second kind, namely

$$f(x) + \int_a^b K(x,y)g(y)f(y)dy = f_a + \frac{f_b - f_a}{b-a}(x-a) + \int_a^b K(x,y)h(y)dy, \tag{2.5}$$

$$x \in [a,b];$$

here, $K(x,y)$ is a Green's function

$$K(x,y) = \begin{cases} (b-x)(a-y)/(b-a), & a \le y \le x \le b, \\ (b-y)(a-x)/(b-a), & a \le x \le y \le b. \end{cases}$$

Since the solution of (2.5) does not only depend on "backward" information (i.e. $y \le x$), but, for any given $x$, involves values of the kernel at each $y \in [a,b]$, the methods for solving Fredholm integral equations will involve the solution of a (generally large) system of algebraic equations.

There is also a relationship between Volterra integral equations of the second kind and so-called Volterra integro-differential equations. Differentiating (2.4) yields

$$f'(x) = g'(x) + K(x,x)f(x) + \int_0^x \frac{\partial K}{\partial x}(x,y)f(y)dy, \quad f(0) = g(0),$$

which is an example of a *Volterra integro-differential equation*

$$f'(x) = \Phi(x,f(x),z(x)), \quad f(0) = f_0, \tag{2.6}$$

where

$$z(x) = \int_0^x K(x,y,f(y))dy.$$

On the other hand, integrating (2.6), and defining $f_1(x) := f(x)$, $f_2(x) := z(x)$, we obtain the *system* of Volterra integral equations of the second kind

$$\begin{pmatrix} f_1(x) \\ f_2(x) \end{pmatrix} = \begin{pmatrix} f_0 \\ 0 \end{pmatrix} + \int_0^x \begin{pmatrix} \Phi(y,f_1(y),f_2(y)) \\ K(x,y,f_1(y)) \end{pmatrix} dy.$$

Hence, methods for solving (2.4) can be useful for solving (2.6), and vica versa.

## 3. TWO MAIN PRINCIPLES FOR NUMERICALLY SOLVING INTEGRAL EQUATIONS

In this section we shall describe two principles for numerically treating integral equations. In general, they apply to all classes of equations given in section 2, although in certain cases the results will be more successful than in other cases. In subsequent sections, where the four classes are treated separately, more attention will be paid to these different results.

The first, and very obvious, idea is to replace the integral by some suitably chosen quadrature formula, like the trapezoidal rule, the Simpson rule, a Gaussian rule, a Gregory rule, etc.. For *Fredholm-type* equations we need one single quadrature formula for the integral over the interval [a,b]. Writing down the resulting equation for different values of x, usually collocating with the abscissas of the quadrature formula, we obtain a system of algebraic equations, where the unknowns are the values of the unknown function f(x) in the abscissas of the quadrature formula. The system is linear or nonlinear according as the original integral equation is linear or nonlinear. The more accurate the quadrature formula, the more accurately will the solution of this system approximate f(x) in the abscissas. A suitably chosen interpolation formula then yields f(x) in other points, if necessary. For *Volterra-type* equations, one should select a *sequence* of quadrature formulas because of the variable upperbound in the integral. First write down the Volterra integral equation for a number of suitably chosen values of x (e.g., $x_i = ih$, $i = 0,1,\ldots,N$; $Nh = b$), and in all these equations replace the integral by a quadrature formula with abscissas which coincide with these values of x. This yields a system of equations (again linear or nonlinear) where the upper-triangle of the coefficient matrix is zero. Thus, solving Volterra equations by direct quadrature is generally cheaper than treating Fredholm equations in this way.

Direct quadrature may fail when the kernel of the integral equation is (weakly) singular. In this case one may remove the singularity by a so-called

*product integration formula*. The basic idea of product integration is the formula

$$\int_a^b K(x,y)f(y)dy = \sum_{j=1}^{n} \alpha_j(x)f(y_j) + E_\alpha,$$

where $\alpha_j(x)$ is chosen such that the remainder $E_\alpha$ vanishes when $f(y)$, *rather than the whole integrand,* is a polynomial of appropriate degree. By writing successively $f(y) = 1,y,y^2,\ldots,y^{n-1}$ in this formula, we obtain a system of n linear equations for the weights $\alpha_j$. However, this must be done for *each point* x, so we have to pay a heavy price for the removal of the singularity. Various forms of composite formulae have been suggested to decrease the amount of work, but we shall not enter into this subject here. Product integration can be viewed as a special case of direct quadrature, but it is also equivalent to a so-called collocation method, which can be derived from the second main idea.

This idea is to write the unknown function $f(x)$ as a linear combination of suitably chosen basis functions and substituting this into the original equation. Subtracting the right-hand side from the left-hand side yields a residue function which depends on x and on the (unknown) coefficients of the expansion of $f(x)$. Unless the true solution of our problem is a linear combination of the basis functions, one cannot choose the coefficients such that the residual function vanishes identically. Therefore, we must content ourselves with a choice of the coefficients which makes the residual function small, in some sense. One obvious way is to require the residual function to be zero in a number of suitably chosen points, which number equals the number of basis functions chosen initially. This yields a (linear or nonlinear) system of equations, and this method is known as the *collocation method*. If this number of points is less or greater than the number of basis functions the resulting system of equations is underdetermined, resp. overdetermined. Another method, the classical Galerkin method, is obtained, when we require the inner product of the residual function with each of the basis functions to vanish. This also yields a linear or nonlinear system of equations to solve. In the linear and symmetric case (i.e., $K(x,y) = K(y,x)$), this method reduces to the so-called Rayleigh-Ritz method, which can also be obtained from a variational formulation. When the inner products are not taken with the basis functions, but with other, suitably chosen, functions, one speaks of the method of moments. There is also a linear programming method, which

minimizes the infinity norm of the residual function over all possible choices of the coefficients.

All methods derived from the second idea, are called *expansion* methods. These methods may be used with profit in case the kernel has a (weak) singularity. Although it is not our intention here to compare the merits of direct quadrature methods with those of expansion methods, one rule of thumb may be emphasized: direct quadrature methods are much cheaper than expansion methods, and should always be tried first.

## 4. FREDHOLM INTEGRAL EQUATIONS OF THE FIRST KIND

In this section we shall consider the (linear) Fredholm integral equation of the first kind

$$\int_a^b K(x,y)f(y)dy = g(x), \qquad a \le x \le b, \tag{4.1}$$

or, in operator notation,

$$Kf = g.$$

This problem is known to be ill-posed in the sense that the solution f does not depend continuously on the (data) function g. We want to make some observations which illustrate the difficulties one can expect when trying to solve this problem numerically.

(i) Suppose the operator K has an eigenvalue zero with eigenfunction $\psi(x)$, so $\int_a^b K(x,y)\psi(y)dy = 0$, then it follows that if $f(x)$ is a solution of (4.1) than also $f(x) + \psi(x)$ is a solution.

(ii) Suppose the variables x and y in K can be separated, i.e., suppose that $K(x,y) = X(x)Y(y)$, then

$$\int_a^b K(x,y)f(y)dy = \gamma X(x),$$

where $\gamma = \int_a^b Y(y)f(y)dy$. Hence a necessary condition for the existence of a solution of (4.1) is that $g(x)$ is a multiple of $X(x)$. As a consequence, when not $g(x)$, but $g(x) + \varepsilon(x)$ is known, no longer a solution may exist.

(iii) According to the Riemann-Lebesgue theorem, if $K(x,y)$ is absolute integrable then

$$h_n(x) := \int_a^b K(x,y)\cos ny\, dy \to 0, \qquad a \le x \le b, \text{ as } n \to \infty.$$

Therefore, given any $\varepsilon > 0$, as small as you want, for n large enough we have $\max_{a\le x\le b} |h_n(x)| \le \varepsilon$. Now suppose that the equation $Kf = g$ has a unique solution f, and let g be given a small perturbation $\delta g = Ch_n(x)$, then the corresponding change in f is $\delta f = C \cos nx$. Clearly, by choosing n sufficiently large, the ratio $\|\delta f\|/\|\delta g\|$ can be made as large as we like.

Now we give a numerical example, borrowed from NOBLE ([46]). Let $a = 0$, $b = 1$, $f(x) \equiv 1$, and

$$K(x,y) = \frac{1}{\pi}\left\{\frac{d}{d^2+(x-y)^2} + \frac{d}{d^2+(x+y)^2}\right\}$$

and set g(x) to the resulting integral in (4.1). Solving the integral equation numerically on an 8-decimal machine by approximating the integral by Simpson's rule with n points and collocating at the quadrature points, yields

| d | n | f(0) |
|---|---|---|
| 1 | 4 | 1.278 |
| | 8 | 0.237 |
| | 16 | -117.521 |

| d | n | f(0) |
|---|---|---|
| 0.1 | 4 | 1.263 |
| | 8 | 1.479 |
| | 16 | 1.500 |

It is striking that the results get worse, as n increases, and also that the smoother the kernel, the worse the results are! An explanation can be given in terms of eigenfunctions and eigenvalues (for symmetric kernels; see NOBLE [46]) or singular functions and - values (for nonsymmetric kernels; see Chapter 13, Fredholm equations of the first kind, by G.F. MILLER, in DELVES & WALSH [20]). A general conclusion we can draw is that direct quadrature methods for Fredholm equations of the first kind can give very poor results, when the kernel is smooth. RICHTER [48] quantifies the relationship between the smoothness of the kernel, and the amenability of Fredholm equations of the first kind to standard numerical methods.

Before entering into some detail about methods for solving (4.1) we notice that one should always use as much information as can possibly be gained *a priori* about the solution of (4.1), e.g. its positivity.

Roughly spoken, there are four methods for solving (4.1). Good expositions of these four methods may be found in the two references NOBLE and MILLER given above. Here we will only give an indication of the main lines of thought.

## 4.1. Expansion methods

When the kernel function is symmetric in x and y one could try to find an expansion of f(x) in eigenfunctions of K. This was first investigated in 1964 by BAKER, FOX, MAYERS & WRIGHT [6]. Extensions to the general (non-symmetric) case have been considered by HANSON [26], VAINSTEIN [57, 58] and CRONE [14]. The principal difficulty is the choice of the number of terms in the expansion.

Other expansion methods, of the type discussed in section 3, can work out quite well, especially if one knows a set of basis functions which is particularly well suited to the problem, i.e., such that f(x) can be well represented by means of only a few terms. An interesting review paper on this subject is TURČIN, KOZLOV & MALKEVICH [56].

## 4.2. Regularization methods

Instead of solving $Kf = g$, one solves $(K+\alpha R)f = g$, where R is a so-called regularizing operator and $\alpha$ is a regularization parameter. The term $\alpha Rf$ has a stabilizing effect; another formulation of the regularization method is that instead of the *first* kind problem, one solves a *second* kind problem which is close to the first kind problem. This method was developed independently by TIHONOV [53] and PHILLIPS [47]. The method suggested by Phillips for determining R is to minimize

$$\| Kf_\alpha - g \|^2 + \alpha \| f_\alpha'' \|^2,$$

where the second term is chosen to help suppress oscillations. The difficulty with this method is how to choose the value of the regularization parameter. Much research has been done with respect to the regularization method since the publications of Phillips and Tihonov. We only mention here the survey paper of TIHONOV, IVANOV and LAVRENTIEV [54] and the book of TIHONOV and ARSENIN [55].

## 4.3. Iteration

The simplest iteration method is given by the scheme (see LANDWEBER [37])

$$f^{(r+1)} = f^{(r)} + \beta K^T (g - Kf^{(r)}), \qquad r = 0,1,\ldots, \quad f^{(0)} = 0,$$

with $0 < \beta < 2/\lambda_1^2$, where $\lambda_1$ is the largest singular value of K. If no errors are present, $f^{(r)}$ will tend to the solution of $Kf = g$ as r tends to infinity. In practice the initial iterates will be smooth, but spurious oscillations tend to appear as r increases. The difficulty with iteration methods is to know when to stop the iteration process.

SHAW [49] and STRAND [51] considered the iteration of the regularization method.

### 4.4. Probabilistic methods

In this class of methods, the problem of solving (4.1) is viewed from a probabilistic or statistical standpoint. Some a priori knowledge about the data of the problem is needed. Important papers on this subject have been written by TURČIN, KOZLOV & MALKEVICH [56], STRAND & WESTWATER [52], FRANKLIN [23], ANDERSSEN & BLOOMFIELD [3] and WAHBA [59]. In fact, these methods may be regarded as an extension of the regularization method, and the actual computation processes are somewhat similar.

### 4.5. Additional references

The following recent references deserve to be mentioned here: they are all concerned with theoretical or practical aspects of the four methods described above: ANDERSSEN [2], CULLUM [15, 16], CUPPEN [17], ELDÉN [21], GRAVES & PRENTER [25], FRANKLIN [24], KÖCKLER [36], JENNINGS [32], LEE & PRENTER [38] and NATTERER [43].

## 5. FREDHOLM INTEGRAL EQUATIONS OF THE SECOND KIND AND EIGENVALUE PROBLEMS

It is convenient to write Fredholm integral equations of the second kind in the form

$$\lambda f(x) = g(x) + \int_a^b K(x,y) f(y)\,dy, \qquad a \le x \le b, \tag{5.1}$$

where $\lambda$ is a constant. In operator notation:

$$(\lambda - K) f = g. \tag{5.2}$$

According to the theory, if K and g are continuous, then (5.1) has a unique solution, except when $\lambda$ is an eigenvalue of K, or zero. The following usual methods for solving (5.1) numerically are all based on one of the two ideas described in section 3. We do not discuss here so-called invariant imbedding methods (cf. KAGIWADA & KALABA [33]), which have as "prime distinctive feature the exact transformation of the functional equation into a Cauchy problem" ([34]). This Cauchy problem is then to be solved by methods for the numerical integration of large systems of ordinary differential equations. In their area of applicability invariant imbedding methods are often quite good, although the amount of computational work may be considerable.

## 5.1. Direct quadrature (Nyström's method)

The integral in (5.1) is replaced by a finite sum, using a standard quadrature formula, and assuming some appropriate smoothness condition for $K(x,y)f(y)$. This gives

$$\lambda f(x) = g(x) + \sum_{j=1}^{n} w_j K(x,x_j) f(x_j) + e(x),$$

where the $w_j$'s are the weights of the quadrature formula and $e(x)$ is the error. Neglecting $e(x)$ and collocating at the points $x = x_j$, $1 \le j \le n$, yields the set of simultaneous algebraic equations

$$\lambda f_i = g(x_i) + \sum_{j=1}^{n} w_j K(x_i,x_j) f_j, \qquad 1 \le i \le n, \tag{5.3}$$

where the $f_i$'s are (hopefully) approximations of $f(x_i)$. The function

$$\tilde{f}(x) = \frac{1}{\lambda}\{g(x) + \sum_{j=1}^{n} w_j K(x,x_j) f_j\},$$

which obviously satisfies $\tilde{f}(x_i) = f_i$, $1 \le i \le n$, is called the Nyström extension of the computed vector $(f_1,\ldots,f_n)^T$. This "interpolating" function is a linear combination of $g(x)$ and the functions $K(x,x_j)$, $1 \le j \le n$. Hence, in a certain sense, this direct quadrature method may be considered as an expansion method. Of course, the Nyström method can also be used, with appropriate modifications, for nonlinear equations of Fredholm of the second kind.

Many interesting remarks on smoothness, error control and related practical aspects of solving (5.3) may be found in ATKINSON [4], DELVES & WALSH

[20] and NOBLE [46]. Very recently, SCHIPPERS ([50]) has applied multigrid techniques (see BRANDT [9]) to the solution of (5.3).

## 5.2. Degenerate kernel methods

Degenerate kernel methods are "inspired" by the fact that if the kernel of (5.1) is *degenerate*, i.e. it can be written *exactly* as

$$K(x,y) = \sum_{j=1}^{m} X_j(x)Y_j(y),$$

then the solution of (5.1) is given by

$$f(x) = \frac{1}{\lambda} g(x) + \frac{1}{\lambda} \sum_{j=1}^{m} a_j X_j(x), \tag{5.4}$$

where the $a_j$'s are the solutions of a linear system of algebraic equations. Indeed, substituting (5.4) into (5.1) yields

$$a_j = \int_a^b Y_j(y) f(y)\,dy; \tag{5.5}$$

multiplying (5.4) by $Y_k(x)$ and integrating from a to b gives

$$\int_a^b f(x)Y_k(x)\,dx = \frac{1}{\lambda} \int_a^b g(x)Y_k(x)\,dx + \frac{1}{\lambda} \sum_{j=1}^{m} a_j \int_a^b X_j(x)Y_k(x)\,dx, \tag{5.6}$$

$$k = 1,2,\ldots,m.$$

Substituting (5.5) into the left side of (5.6) and introducing the abbreviations $b_k = \int_a^b g(x)Y_k(x)\,dx$ and $c_{kj} = \int_a^b X_j(x)Y_k(x)\,dx$, we obtain the linear equations

$$\lambda a_k = b_k + \sum_{j=1}^{m} c_{kj} a_j, \qquad k = 1,2,\ldots,m.$$

Suppose now that we want to solve a Fredholm equation (5.1) where the kernel is *non-degenerate*. Then an obvious approach is to approximate the kernel by a *degenerate* kernel, which can be done in many ways (see e.g. ATKINSON [4]). The resulting Fredholm equation is solved as outlined above. The major computational effort will go into the approximation of the m(m+1) integrals $b_k$ and $c_{kj}$, and it is clear that this degenerate kernel method is only feasible when m is not too large.

## 5.3. The collocation method

In the collocation method for solving (5.1), one assumes that f can be approximated by a series of the form

$$f(x) \approx f_m(x) = \sum_{j=1}^{m} a_j \phi_j(x), \tag{5.7}$$

where the $\phi_j$'s are suitably chosen basis functions and the $a_j$'s are still to be determined. Defining the residual function

$$r_m(x) := (\lambda - K) f_m - g,$$

and setting $r_m(x_i) = 0$, for $i = 1,2,\ldots,m$, yields the following system of linear algebraic equations

$$\lambda \sum_{j=1}^{m} a_j \phi_j(x_i) - \sum_{j=1}^{m} a_j \int_a^b K(x_i,y)\phi_j(y)dy - g(x_i) = 0, \quad 1 \le i \le m;$$

the major computational work, again, lies in the approximation of the so-called "moment integrals"

$$\int_a^b K(x_i,y)\phi_j(y)dy.$$

## 5.4. The method of weighted residuals and Galerkin's method

In these methods the $a_j$'s in (5.7) are determined by the conditions

$$\int_a^b r_m(x)\psi_i(x)dx = 0, \qquad 1 \le i \le m, \tag{5.8}$$

where the $\psi_i(x)$ are appropriate weight functions. In the Galerkin method, one chooses $\psi_i = \phi_i$, $1 \le i \le m$. The method of moments and the method of least squares are also special cases of conditions (5.8).

The choice of the basis functions, e.g. Chebyshev series, can be of great importance in practice.

## 5.5. The eigenvalue problem

All the methods given in sections 5.1 - 5.4 can be used for the

calculation of the eigenvalues and the eigenfunctions in the problem

$$\int_a^b K(x,y)f(y)dy = \lambda f(x), \qquad a \le x \le b.$$

Instead of a system of linear equations, we get an eigenvalue problem for a matrix. An extensive and thorough survey of these methods can be found in BAKER [5].

## 6. VOLTERRA INTEGRAL EQUATIONS OF THE FIRST KIND

Practical experience with the numerical solution of first kind integral equations shows that Volterra integral equations of the first kind

$$\int_0^x K(x,y)f(y)dy = g(x), \qquad 0 \le x \le b, \tag{6.1}$$

often give fewer difficulties than Fredholm equations of the first kind (eq. (4.1)). This may be explained as follows. From the theory of first kind integral equations one can derive that the situation is usually improved upon when the kernel $K(x,y)$ behaves badly. When in (6.1) $K(x,y) > 0$, one can expect better results than in case $K(x,y) = 0$ for certain values of $x,y$, since (6.1) can be viewed as a Fredholm equation of the first kind with kernel $K_+(x,y) = K(x,y)$ for $y \le x$, and $K_+(x,y) = 0$ for $y > x$. This kernel is discontinuous on the line $x = y$, when $K(x,y) > 0$. Another indication for numerical difficulties is the fact that non-uniqueness of the solution of (6.1) can occur, when $K(x,y) = 0$ for some $x$, $y$ in the domain of definition $0 \le y \le x \le b$, e.g., for $K(x,y) = 2x - 3y$ and $g(x) = x^2$, $f(x) = ax + 2$ is a solution of (6.1) for any real $a$.

Before discussing some methods for solving (6.1) we remark that if $K(x,x) \neq 0$, (6.1) can be converted by differentiation into the Volterra integral equation of the second kind

$$K(x,x)f(x) + \int_0^x \frac{\partial K}{\partial x}(x,y)f(y)dy = g'(x), \qquad 0 \le x \le b, \tag{6.2}$$

which is generally easier to solve than (6.1). However, practical difficulties may arise when $g(x)$ and $g'(x)$ are not known analytically. If $K(x,x) \equiv 0$, one may repeat this differentiation process, but this method seems to be only of practical relevance when $\frac{\partial^r}{\partial x^r} K(x,y)\Big|_{y=x} \neq 0$ for some small value of $r$.

## 6.1. Direct quadrature methods

These methods are based on rules of the form

$$\int_0^{rh} \phi(y)\,dy \approx \sum_{j=0}^{r} w_{rj}\phi(jh), \qquad 1 \le r \le N, \quad Nh = b. \tag{6.3}$$

Substituting this into (6.1) yields

$$\sum_{j=0}^{r} w_{rj}K(rh,jh)f_j = g(rh), \qquad 1 \le r \le N. \tag{6.4}$$

If $w_{rr} \neq 0$ $(1 \le r \le N)$ this is a system of N equations in N + 1 unknowns $f_0, f_1, \ldots, f_N$, and cannot be solved directly. An extra condition may be derived from (6.2), viz. $f_0 = g'(0)/K(0,0)$. An obvious choice for (6.3) is the trapezoidal rule, but it has been observed that instabilities can arise when h is too large. Also higher order Gregory quadrature formulas (of which the trapezoidal rule is the simplest case) produce instabilities in (6.4). Better results have been obtained by other one-step methods, Runge-Kutta methods and block methods. Also higher order linear multistep methods have been developed. For an extensive discussion and list of references, compare BRUNNER [12].

## 6.2. Product-integration methods

The simplest product-integration method is the so-called modified midpoint method, where the approximation

$$\int_{(j-1)h}^{jh} K(x,y)f(y)\,dy \approx f((j-\tfrac{1}{2})h) \int_{(j-1)h}^{jh} K(x,y)\,dy$$

is used. This yields the scheme

$$\sum_{j=0}^{r-1} f_{j+\frac{1}{2}} \int_{jh}^{(j+1)h} K(rh,y)\,dy = g(rh), \qquad 1 \le r \le N,$$

where $f_{j+\frac{1}{2}}$ is an approximation of $f((j+\frac{1}{2})h)$. Product-integration is particularly appropriate when K(x,y) is badly behaved (or weakly singular), whereas the solution f(x) is not. See ANDERSSEN & WHITE [1], WEISS & ANDERSSEN [60]

and LINZ [41].

### 6.3. Collocation methods

Very recently, BRUNNER [10, 11, 12] has developed a general theory of collocation methods for first kind Volterra integral equations, where the solution is projected into the space of piecewise polynomials of a given degree, with jump discontinuities on the set of knots. Many of the well-known direct methods (section 6.1) are shown to result from particular discretizations of the "moment" integrals occurring in this projection method, and a unified convergence analysis for these direct methods is therefore possible. Moreover, Brunner's approach yields the tools to handle the question of the connection between the location of the collocation points and the order of convergence of the method. No distinction has to be made between equations with regular and weakly singular kernels.

## 7. VOLTERRA INTEGRAL EQUATIONS OF THE SECOND KIND

All methods for Volterra equations of the first kind, described in section 6, are suitable for solving Volterra equations of the second kind

$$f(x) = g(x) + \int_0^x K(x,y)f(y)dy, \qquad 0 \le x \le b. \tag{7.1}$$

In the literature, relatively little attention has been paid to expansion methods. In section 7.2 we shall describe one special method of Bownds, which might be viewed as an expansion method.

### 7.1. Direct quadrature methods

The integral occurring in (7.1) is discretized by use of the formula (6.3). Writing down the resulting equation for $x = ih$, $0 \le i \le N$, we obtain the set of equations

$$\begin{aligned} f_r &= g(rh) + \sum_{j=0}^{r} w_{rj}K(rh,jh)f_j, \qquad 1 \le r \le N, \\ f_0 &= g(0). \end{aligned} \tag{7.2}$$

Various choices for the weights $w_{rj}$ can be made. In contrast to Volterra

equations of the first kind, Gregory quadrature formulas do not present instability problems here, thanks to the "stabilizing" term f(x) in (7.1) outside the integral.

Instead of the trapezoidal rule or higher order Gregory rules, one could also try the repeated Simpson rule. However, this asks for an *even* number of intervals, so it can only be used in (7.1) for *even* r. When r is odd the 3/8ths rule should be used at the lower end or at the upper end of the integration interval, in order to maintain the local truncation error of $O(h^5)$. When the 3/8ths rule is used at the *lower* end, the weight matrix reads

$$
\begin{array}{lllllllllll}
 & j=0 & 1 & 2 & & 2i-2 & 2i-1 & 2i & 2i+1 \\
w_{2i,j} = & \frac{1}{3}h & \frac{4}{3}h & \frac{2}{3}h & \cdots & \frac{2}{3}h & \frac{4}{3}h & \frac{1}{3}h & \\
w_{2i+1,j} = & \frac{3}{8}h & \frac{9}{8}h & \frac{9}{8}h & (\frac{3}{8}+\frac{1}{3})h \cdots & \frac{4}{3}h & \frac{2}{3}h & \frac{4}{3}h & \frac{1}{3}h,
\end{array}
\tag{7.3}
$$

and when the 3/8ths rule is used at the *upper* end, the weight matrix reads

$$
\begin{array}{lllllllll}
 & j=0 & 1 & 2 & 2i-2 & 2i-1 & 2i & 2i+1 \\
w_{2i,j} = & \frac{1}{3}h & \frac{4}{3}h & \frac{2}{3}h \cdots & \frac{2}{3}h & \frac{4}{3}h & \frac{1}{3}h & \\
w_{2i+1,j} = & \frac{1}{3}h & \frac{4}{3}h & \frac{2}{3}h \cdots & (\frac{1}{3}+\frac{3}{8})h & \frac{9}{8}h & \frac{9}{8}h & \frac{3}{8}h.
\end{array}
\tag{7.4}
$$

The work of KOBAYASI [35], LINZ [39], NOBLE [45] and BAKER & KEECH [7] has indicated that the use of (7.4) is to be preferred above (7.3), since (7.4) has a so-called *repetition factor* 1 and (7.3) a repetition factor >1 (viz. 2). The repetition factor p associated with the weights in (7.2) is defined as the smallest integer such that $w_{rj} = w_{r+p,j}$ for $u \le j \le r-v$, where u and v are fixed numbers, *independent* of r.

When using higher order quadrature formulas in (7.2) (in order to increase the accuracy) some starting values $f_1, f_2, \ldots, f_k$ need to be computed. DAY [18, 19] has developed a number of starting procedures, which are useful in certain cases, but which tend to become unstable when K(x,y) (or $\partial K/\partial f$ in the nonlinear case) assumes large values.

Also Runge-Kutta and block methods have been developed for second kind Volterra equations. For references, see BAKER [5].

VAN DER HOUWEN & TE RIELE [27] have recently developed a method for

solving (7.1) which is based on the use of a Gregory formula combined with a backward differentiation formula. This method has very attractive stability properties. Very recently, VAN DER HOUWEN & WOLKENFELT [29] have given a general stability analysis of multistep formulas for systems of Volterra integral equations of the second kind. This partly extends the work of BAKER & KEECH [7] on this subject.

## 7.2. An expansion method of Bownds

BOWNDS [8] has developed a method for solving (7.1) which is inspired by the degenerate kernel method for Fredholm equations of the second kind (cf. section 5.2). Suppose the kernel $K(x,y)$ can *exactly* be written as

$$K(x,y) = \sum_{j=1}^{m} X_j(x) Y_j(y), \tag{7.5}$$

then the solution of (7.1) can be written as

$$f(x) = g(x) + \sum_{j=1}^{m} a_j(x) X_j(x), \tag{7.6}$$

where the $a_j(x)$ are the solutions of the system of differential equations

$$\begin{cases} a_j'(x) = Y_j(x)\left[g(x) + \sum_{i=1}^{m} a_j(x) X_j(x)\right], & 1 \le j \le m, \\ a_j(0) = 0. \end{cases} \tag{7.7}$$

Indeed, substituting (7.5) into (7.1) and using the expansion (7.6) in the left hand side yields

$$a_j(x) = \int_0^x Y_j(y) f(y)\, dy,$$

which is equivalent with (7.7), when using (7.6).

If there is no expansion for $K(x,y)$ of the form (7.5), then Bownds suggests to *approximate* the kernel by such an expansion. Then he solves (7.7) by a fourth-order standard Runge-Kutta method.

## 7.3. Volterra integro-differential equations

The most important methods for solving Volterra integro-differential equations

(7.8) $\qquad f'(x) = \Phi(x,f(x),z(x)), \qquad f(0) = f_0,$

where $z(x) = \int_0^x K(x,y,f(y))dy$, are linear multistep methods of the form

$$\sum_{\ell=0}^{k} a_\ell f_{n+1-\ell} = h \sum_{\ell=0}^{k} b_\ell \Phi(x_{n+1-\ell}, f_{n+1-\ell}, z_{n+1-\ell}),$$

$$z_{n+1-\ell} = \sum_{j=0}^{n+1-\ell} w_{n+1-\ell_j} K(x_{n+1-\ell}, x_j, f_j), \qquad 0 \le \ell \le k.$$

Here $a_\ell$ and $b_\ell$ are the coefficients of some linear multistep method for the "differential" equation (7.8), and $z_{n+1-\ell}$ is a quadrature formula approximation of $z(x_{n+1-\ell})$. These methods have been studied by LINZ [40], BRUNNER & LAMBERT [13], MATTHYS [42] and WOLKENFELT [61]. Recently, VAN DER HOUWEN, TE RIELE & WOLKENFELT [28] have developed a general stability analysis for linear multistep methods for systems of Volterra integro-differential equations.

In DELVES & WALSH [20, ch. 14] Baker gives a survey of Volterra- and other types of integro-differential equations, and of methods for their numerical solution.

REFERENCES

[1] ANDERSSEN, A.S. & E.T. WHITE, *Improved numerical methods for Volterra integral equations of the first kind,* The Computer Journal, 14 (1971), pp. 442-443.

[2] ANDERSSEN, R.S., *Computing with noisy data with an application to Abel's equation,* pp. 61-79 in: *Error, approximation and accuracy,* Proc. Sem. Austr. Nat. Univ. Canberra, Univ. Queensland, St. Lucia, 1973.

[3] ANDERSSEN, R.S. & P. BLOOMFIELD, *Numerical differentiation procedures for non-exact data,* Numer. Math., 22 (1974), pp. 157-182.

[4] ATKINSON, K.E., *A survey of numerical methods for the solution of Fredholm integral equations of the second kind,* SIAM, 1976.

[5] BAKER, C.T.H., *The numerical treatment of integral equations,* Clarendon Press, Oxford, 1977.

[6] BAKER, C.T.H., L. FOX, K. WRIGHT & D.F. MAYERS, *Numerical solution of Fredholm integral equations of the first kind,* The Computer Journal, 7 (1964), pp. 141-148.

[7] BAKER, C.T.H. & M.S. KEECH, *Stability regions in the numerical treatment of Volterra integral equations,* SIAM J. Numer. Anal., 15 (1978), pp. 394-417.

[8] BOWNDS, J.M., *On an initial value method for quickly solving Volterra integral equations: a review,* J. Opt. Theory and Applics., 24 (1978), pp. 133-151.

[9] BRANDT, A., *Multi-level adaptive solutions to boundary-value problems,* Math. Comp., 31 (1977), pp. 333-390.

[10] BRUNNER, H., *Discretization of Volterra integral equations of the first kind,* Math. Comp., 31 (1977), pp. 708-716.

[11] BRUNNER, H., *Discretization of Volterra integral equations of the first kind (II),* Numer. Math., 30 (1978), pp.117-136.

[12] BRUNNER, H., *The approximate solution of integral equations by projecting methods based on collocation,* Report Mathematics and Computation No. 1/78, Dept. of Mathematics, Univ. of Trondheim, Trondheim, 1978.

[13] BRUNNER, H. & J.D. LAMBERT, *Stability of numerical methods for Volterra integro-differential equations,* Computing, 12 (1974), pp. 75-89.

[14] CRONE, L., *The singular value decomposition of matrices and cheap numerical filtering of systems of linear equations,* J. Franklin Inst., 294 (1972), pp. 133-136.

[15] CULLUM, J., *Ill-posed problems, regularization and singular value decomposition,* Report RC6465 (#27919), IBM Thomas J. Watson Research Center, Yorktown Heights, 1977.

[16] CULLUM, J., *The choice of smoothing norm in regularization - a key to effectiveness,* Report RC6511 (#28126), IBM Thomas J. Watson Research Center, Yorktown Heights, 1977.

[17] CUPPEN, J.J.M., *Regularization of ill-posed problems* (in Dutch), Eindhoven University of Technology, febr. 1978.

[18] DAY, J.T., *On the numerical solution of Volterra integral equations,* BIT, 8 (1968), pp. 134-137.

[19] DAY, J.T., *A starting method for solving nonlinear Volterra integral equations,* Math. Comp., 21 (1967), pp. 179-188.

[20] DELVES, L.M. & J.E. WALSH (eds.), *Numerical solution of integral equations,* Proc. of Liverpool-Manchester summer school, Clarendon Press, Oxford, 1974.

[21] ELDÉN, L., *Algorithms for the regularization of ill-conditioned least squares problems,* BIT, 17 (1977), pp. 134-145.

[22] FOX, L. & E.T. GOODWIN, *The numerical solution of nonsingular linear integral equations,* Phil. Trans. Royal Soc. A., 245 (1953), pp. 501-534.

[23] FRANKLIN, J.N., *On Tihonov's method for ill-posed problems,* Math. Comp., 28 (1974), pp. 889-907.

[24] FRANKLIN, J.N., *Minimum principles for ill-posed problems,* SIAM J. Math. Anal., 9 (1978), pp. 638-650.

[25] GRAVES, J. & P.M. PRENTER, *Numerical iterative filters applied to first kind Fredholm integral equations,* Numer. Math., 30 (1978), pp. 281-299.

[26] HANSON, R.J., *A numerical method for solving Fredholm integral equations of the first kind using singular values,* SIAM J. Numer. Anal., 8 (1971), pp. 616-622.

[27] HOUWEN, P.J. VAN DER & H.J.J. TE RIELE, *Backward differentiation formulas for Volterra integral equations of the second kind,*
*I. Convergence and stability,* Report NW 48/77, Sept. 1977,
*II. Numerical experiments,* Report NW 57/78, May 1978, Mathematical Centre, Amsterdam.

[28] HOUWEN, P.J. VAN DER, H.J.J. TE RIELE & P.H.M. WOLKENFELT, *On the stability of multistep formulas for systems of Volterra integro-differential equations,* Report NW 63/78, Oct. 1978, Mathematical Centre, Amsterdam.

[29] HOUWEN, P.J. VAN DER & P.H.M. WOLKENFELT, *On the stability of multistep formulas for Volterra integral equations of the second kind,* Report NW 59/78, July 1978, Mathematical Centre, Amsterdam.

[30] IVANOV, V.V., *The theory of approximate methods and their application to the numerical solution of singular integral equations,* Noordhoff, Leiden, 1976.

[31] JASWON, M.A. & G.T. SYMM, *Integral equation methods in potential theory and elastostatics,* Acad. Press, 1977.

[32] JENNINGS, L.S., *Regularization techniques for first kind integral equations,* pp. 81-94 in: *Error, approximation and accuracy,* Proc. Sem. Austr. Nat. Univ. Canberra, Univ. Queensland, St. Lucia, 1973.

[33] KAGIWADA, H.H. & R. KALABA, *Integral equations via imbedding methods,* Addison-Wesley, 1974.

[34] KAGIWADA, H.H. & R. KALABA, *Imbedding methods for integral equations with applications,* J. Opt. Theory and Applics., 24 (1978), pp. 29-57.

[35] KOBAYASI, M., *On the numerical solution of the Volterra integral equations of the first kind by trapezoidal rule,* Rep. Stat. Appl. Res. JUSE, 14 (1967), pp. 1-14.

[36] KÖCKLER, N., *Parameterwahl und Fehlerabschätzung bei der regularisierten Lösung von inkorrekt gestellten Problemen,* Doctor's diss., Mainz, 1974.

[37] LANDWEBER, L., *An iteration formula for Fredholm integral equations of the first kind,* Amer. J. Math., 73 (1951), pp. 615-624.

[38] LEE, J.W. & P.M. PRENTER, *An analysis of the numerical solution of Fredholm integral equations of the first kind,* Numer. Math., 30 (1978), pp. 1-23.

[39] LINZ, P., *The numerical solution of Volterra integral equations by finite difference methods,* MRC. Tech. Summ. Rept. #825, Univ. of Wisconsin, Madison, Wisconsin, 1967.

[40] LINZ, P., *Linear multistep methods for Volterra integro-differential equations,* J. ACM, 16 (1969), pp. 295-301.

[41] LINZ, P., *Product integration methods for Volterra integral equations of the first kind,* BIT, 11 (1971), pp. 413-421.

[42] MATTHYS, J., *A-stable linear multistep methods for Volterra integro-differential equations,* Numer. Math., 27 (1976), pp. 85-94.

[43] NATTERER, F., *Regularisierung schlecht gestellter Probleme durch Projektionsverfahren,* Numer. Math., 28 (1977), pp. 329-341.

[44] NOBLE, B., *The numerical solution of nonlinear integral equations and related topics,* pp. 215-318 in: P.M. Anselone (ed.), *Nonlinear integral equations,* The Univ. of Wisconsin Press, 1964.

[45] NOBLE, B., *Instability when solving Volterra integral equations of the second kind by multistep methods,* pp. 23-29 in: J. Ll. Morris (ed.), *Conference on the Numerical solution of differential equations,* LNM #109, Springer-Verlag, 1969.

[46] NOBLE, B., *The numerical solution of integral equations,* pp. 915-966 in: D. Jacobs (ed.), *The State of the art in numerical analysis,* Acad. Press, 1977.

[47] PHILLIPS, D.L., *A technique for the numerical solution of certain integral equations of the first kind,* J. ACM, 9 (1962), pp. 84-97.

[48] RICHTER, G.R., *Numerical solution of integral equations of the first kind with nonsmooth kernels,* SIAM J. Numer. Anal., 15 (1978), pp. 511-522.

[49] SHAW, C.B. JR., *Improvement of the resolution of an instrument by numerical solution of an integral equation,* J. Math. Anal. Applics, 37 (1972), pp. 83-112.

[50] SCHIPPERS, H., *Multi-grid techniques for the solution of Fredholm integral equations of the second kind,* this colloquium.

[51] STRAND, O.N., *Theory and methods related to the singular-function expansion and Landweber's iteration for integral equations of the first kind,* SIAM J. Numer. Anal., 4 (1974), pp. 798-825.

[52] STRAND, O.N. & E.R. WESTWATER, *Statistical estimation of the numerical solution of a Fredholm integral equation of the first kind,* J. ACM, 15 (1968), pp. 100-114.

[53] TIHONOV, A.N., *On the solution of incorrectly posed problems and the method of regularization,* Soviet Math., 4 (1963), pp. 1035-1038.

[54] TIHONOV, A.N., V.K. IVANOV & M.M. LAVRENTIEV, *Improperly posed problems,* Amer. Math. Soc. Transl., Series 2, 105 (1976), pp. 313-332.

[55] TIHONOV, A.N. & V.Y. ARSENIN, *Solution of ill-posed problems,* John Wiley and Sons, 1977.

[56] TURČIN, V.F., V.P. KOZLOV & M.S. MALKEVICH, *The use of mathematical-statistics methods in the solution of incorrectly posed problems,* Soviet Phys. Usp., 13 (1971), pp. 681-702.

[57] VAINSTEIN, L.A., *Filtering of noise in a numerical solution of integral equations of the first kind,* Soviet Phys. Dokl., 17 (1972), pp. 519-521.

[58] VAINSTEIN, L.A., *Numerical solution of integral equations of the first kind using a priori information on the function to be determined,* Soviet Phys. Dokl., 17 (1972), pp. 532-534.

[59] WAHBA, G., *Practical approximate solutions to linear operator equations when the data are noisy,* SIAM J. Numer. Anal., 14 (1977), pp. 651-667.

[60] WEISS, R. & R.S. ANDERSSEN, *A product integration method for a class of singular first kind integral equations,* Numer. Math., 18 (1972), pp. 442-456.

[61] WOLKENFELT, P.H.M., *Backward differentiation formulas for Volterra integro-differential equations,* Report NW 53/77, Nov. 1977, Mathematical Centre, Amsterdam.

# MULTI-GRID TECHNIQUES FOR THE SOLUTION OF FREDHOLM INTEGRAL EQUATIONS OF THE SECOND KIND

H. SCHIPPERS

## 1. INTRODUCTION

In this paper we deal with iterative methods for the solution of Fredholm equations of the second kind:

$$\lambda f(x) - \int_a^b K(x,y)f(y)dy = g(x), \qquad a \le x \le b, \tag{1.1}$$

where $\lambda$ is a constant ($\lambda \neq 0$).

Our methods are essentially multi-grid methods that work with a sequence of grids of increasing refinement. The solution of different but related problems on these grids interact with each other to obtain an approximation to the solution of the original problem. Recently the multi-grid technique is strongly propagandized by BRANDT [5], who applies it to boundary value problems. On each grid in a multi-grid process a system of equations is solved by a relaxation scheme (for example: Jacobi relaxation). However, it is well-known that convergence of such a relaxation scheme is only fast as long as the residuals have strong fluctuations (i.e. high-frequency errors) compared with the scale of the grid. As these fluctuations are smoothed out, convergence slows down. At this stage the relaxation sweeps should be stopped, since the low-frequency errors in the residual can be liquidated on coarser grids.

Another point of view is to consider our methods as an extension of the iterative schemes given by ATKINSON [2,3]. These schemes are based on the Nyström method for which Atkinson has given an abstract error analysis using the collectively compact operator approximation theory of ANSELONE [1]. Within this framework he was able to analyse the convergence of his iterative methods.

Recently, STETTER [8] exposed the common structural principle of some techniques for error estimation and iterative improvement of discretization methods. Stetter describes the basic principle of these techniques, which he calles the *Defect Correction* principle. He shows that many iterative techniques can be interpreted as special cases of defect correction. As examples he gives a.o. iterative improvement of linear algebraic equations, the modified Newton method and deferred correction.

In section 2 we shall give a brief description of the Defect Correction principle of Stetter. In section 3 we discuss the Nyström method and we shall see that the iterative methods of Atkinson can be formulated

within the framework of the Defect Correction principle. Using the multigrid philosophy of BRANDT [5] we give an extension of the methods of Atkinson in section 4. We also present an Algol-68 algorithm and some numerical results. In section 5 we draw some conclusions and formulate some questions.

## 2. THE DEFECT CORRECTION PRINCIPLE

The Defect Correction formulation appears to be an expedient tool for describing iterative methods as those of ATKINSON [2,3] and BRANDT [5]. This fact motivates our short treatment here.

STETTER [8] considers two normed linear spaces X and Y and a continuous mapping $A: X \to Y$. Assume that A is bijective between a domain $F \subset X$ and a domain $G \subset Y$. With $g \in G$ it is the aim to find an approximation to the unique solution $f^* \in F$ of

$$Af = g. \tag{2.1}$$

This is done by means of an approximate inverse $\tilde{B}: G \to F$; $\tilde{B}$ shall be bijective and continuous. Stetter considers $\tilde{B}$ to be the solution operator of an approximated equation (2.1):

$$\tilde{A}f = g. \tag{2.2}$$

Let $f_0 = \tilde{B}g$ be the first approximation to f. Further we define the *defect* $d_0 = Af_0$. The basic principle of defect correction is to form a sequence $f_i$ so that the defect $d_i$ approaches g. Such a sequence is given by the following iterative process:

$$f_{i+1} = (I-\tilde{B}A)f_i + \tilde{B}g, \tag{2.3}$$

where I is the identity operator. It is clear that $f_i$ converges to the fixed point $f^*$ if $(I-\tilde{B}A)$ is a contraction in F.

A simple example is the modified Newton method in Banach spaces. For this let $A: F \to G$ be Frechet-differentiable and take

$$\tilde{A}f = A\bar{f} + A'(\bar{f})(f-\bar{f}), \qquad \forall f \in F,$$

with $\bar{f} \in F$ fixed, so that

$$\tilde{B}y = \bar{f} - [A'(\bar{f})]^{-1} (A\bar{f}-y), \qquad \forall y \in G.$$

Then (2.3) yields:

$$f_{i+1} = f_i - [A'(\bar{f})]^{-1} (f_i-g).$$

We get an extension of the Defect Correction principle if the approximate inverse $\tilde{B}$ depends on the iteration number i. An example of such a process is the original Newton method (i.e. the last equation where $\bar{f}$ is replaced by $f_i$).

Stetter introduces another version, called "Iterative Updating Defect Correction" (IUDEC), which arises after discretization of an infinite-dimensional problem. Instead of (2.1) we solve the discretized problem

$$A^h f^h = g^h, \tag{2.4}$$

where $A^h$ is a continuous mapping from $X^h$ into $Y^h$. The superscript h is related to the mesh-width and $X^h$ and $Y^h$ are finite-dimensional linear spaces. In solving integral (or differential) equations by a Nyström method (resp. finite differences), $X^h$ and $Y^h$ denote spaces of grid functions on a grid $\mathfrak{G}^h$. Application of defect correction yields

$$f^h_{i+1} = (I-\tilde{B}^h A^h) f^h_i + \tilde{B}^h g^h \tag{2.5}$$

where $\tilde{B}^h$ is an approximate inverse of $A^h$. The principle of iterative updating defect correction is to update $A^h$ after a number of iteration steps and therefore we need a sequence $A^h_i: X^h \to Y^h$. This sequence is so selected that a next $A^h_i$ is a closer approximation (a not less accurate approximation) to the operator A. Equation (2.5) now takes the form

$$f^h_{i+1} = (I-\tilde{B}^h A^h_i)\, f^h_i + \tilde{B}^h g^h. \tag{2.6}$$

In general the IUDEC-principle is defined by:

a. a sequence of operators $A^h_i$ (i.e. approximations to A);
b. an approximate inverse $B^h$ of $A^h_i$ (an extension of the idea of Stetters IUDEC-principle is to consider a sequence $B^h_i$ of approximate inverses);

c. a right-hand side $g^h$.

The aim of updating defect correction is to find a sequence $f_i^h$ so that the defect $d_i = A_i^h f_i^h$ approximates g with the additional condition that $A_i^h \to A$.

Summarizing we can conclude that defect correction is a very general technique, but the remaining problem is still difficult: How to choose the sequences $\{A_i^h\}$ and $\{\tilde{B}_i^h\}$ so that $f_i^h$ converges to $f^h$ in an optimal way? ($f^h$ will approximate $f^*$ if $d_i \to g$ and $A_i^h \to A$). In the following sections we make choices for these sequences.

## 3. ITERATIVE METHODS OF ATKINSON

In this section we discuss the iterative schemes given in ATKINSON [2,3] for the numerical solution of equation (1.1). We shall assume that $g \in C[a,b]$ and the kernel $K(x,y)$ is so that the associated integral operator K is compact from $C[a,b]$ into $C[a,b]$. Further we assume that $\lambda$ is not an eigenvalue of K and $\lambda \neq 0$. According to the Fredholm Alternative theorem these assumptions imply that (1.1) has a unique solution.

ATKINSON [2] discusses iterative schemes which are based on the Nyström method. The integral in equation (1.1) is replaced by a finite sum

$$\int_a^b K(x,y)f(y)dy = \sum_{j=1}^{N} k_j(x)f(x_j) + e(x)$$

where $\{x_j\}$ are the nodal points in $[a,b]$. The weight-functions $k_j(x)$ are supposed to be continuous. They can be obtained in different ways. If $K(x,y)$ and $f(x)$ are smooth, then usually $k_j(x) = w_j K(x,x_j)$ where the $w_j$'s are the weights of a quadrature formula.

Neglecting $e(x)$ we get the following problem

$$\lambda f^h(x) - \sum_{j=1}^{N} k_j(x) f^h(x_j) = g(x). \tag{3.1a}$$

In operator notation:

$$(\lambda I - K^h)\, f^h = g, \tag{3.1b}$$

where $f^h \in C[a,b]$ and $K^h$ is the finite sum operator from $C[a,b]$ into $C[a,b]$.

Collocation at the points $x = x_j$, $1 \leq j \leq N$, yields a system of linear

equations

$$(3.2a) \qquad \lambda f_i^h - \sum_{j=1}^{N} k_j(x_i) f_j^h = g_i, \qquad 1 \le i \le N,$$

where $f_i^h = f^h(x_i)$. We denote the vector $(f_1^h,\ldots,f_N^h)^T$ with $\bar{f}^h$. Let us introduce the restriction operator $P^h$: $C[a,b] \to \mathbb{R}^N$ and the finite sum operator $\bar{K}^h$: $\mathbb{R}^N \to C[a,b]$, defined by

$$\bar{f}_i \equiv (P^h f)_i = f(x_i)$$

and

$$\bar{K}^h \bar{f} = \sum_{j=1}^{N} k_j(x) \bar{f}_j .$$

We can now write (3.2a) in operator notation as

$$(3.2b) \qquad (\lambda I - P^h \bar{K}^h)\, f^{-h} = \bar{g} \equiv P^h g.$$

This equation follows from (3.1b) by realizing that $K^h f^h = \bar{K}^h P^h f^h = \bar{K}^h \bar{f}^h$. If we assume that $(\lambda I - P^h \bar{K}^h)^{-1}$ exists, equation (3.2) yields

$$(3.3) \qquad \bar{f}^h = (\lambda I - P^h \bar{K}^h)^{-1} \bar{g}.$$

Moreover, we get the (continuous) solution $f^h$ of equation (3.1) by means of the "natural" interpolation formula of Nyström:

$$(3.4) \qquad f^h = \frac{1}{\lambda}\{g + \bar{K}^h \bar{f}^h\} \equiv Q^h \bar{f}^h$$

where $Q^h$ denotes the natural interpolation operator: $\mathbb{R}^N \to C[a,b]$.

Resuming we have the following scheme:

$$\begin{array}{ccc} f^h & \xrightarrow{\;(\lambda I - K^h)\;} & g \\ {\scriptstyle Q^h}\uparrow & & \downarrow{\scriptstyle P^h} \\ \bar{f}^h & \xleftarrow{\;(\lambda I - P^h \bar{K}^h)^{-1}\;} & \bar{g} \end{array}$$

It will be clear that the function $f^h \in C[a,b]$, which can be computed by (3.3) - (3.4), satisfies equation (3.1). Therefore from the above scheme one can read

$$(\lambda I - K^h)^{-1} = Q^h (\lambda I - P^h \bar{K}^h)^{-1} P^h .$$

Depending on the number of nodal points in equation (3.2), the linear system may become too large to be solved directly. For its solution ATKINSON [2] discusses two iterative schemes, which we shall describe below as special cases of the defect correction principle. Let us introduce two grids $\mathfrak{G}^{h(0)}$ and $\mathfrak{G}^{h(1)}$ with $N_0$ and $N_1$ nodal points, respectively ($N_1 > N_0$). Further, we introduce short notations for grid $\mathfrak{G}^{h(p)}$ ($N_p$ nodal points):

$$\mathfrak{G}^{h(p)} = \mathfrak{G}_p, \quad f^{h(p)} = f_p, \quad K^{h(p)} = K_p .$$

We assume that $N_0$ is small, so that $(\lambda I - P_0 \bar{K}_0)\bar{u} = \bar{r}$ can be solved directly with

$$A^h = (\lambda I - K_1) \quad \text{and} \quad \tilde{B}^h = (\lambda I - K_0)^{-1} .$$

Equation (2.5) yields

$$f_{1,i+1} = (I - (\lambda I - K_0)^{-1}(\lambda I - K_1)) f_{1,i} + (\lambda I - K_0)^{-1} g ,$$

which can be rewritten as

$$(\lambda I - K_0) f_{1,i+1} = g + (K_1 - K_0) f_{1,i} , \tag{3.5}$$

which is exactly the iterative method 1 of ATKINSON [2].
Equation (3.5) can also be written as

$$\begin{cases} (\lambda I - K_0) v_{1,i} = g - (\lambda I - K_1) f_{1,i} \equiv r_{1,i} \\ \\ v_{1,i} = f_{1,i+1} - f_{1,i} . \end{cases} \tag{3.5.1}$$

The right-hand side of (3.5.1), i.e. the residual $r_{1,i}$, which is computed by means of interpolation of a coarse grid solution, usually will oscillate on the fine grid $\mathfrak{G}_1$. This has been realized first by BRAKHAGE [4], who added a Jacobi-relaxation sweep to smooth the residual in (3.5.1). This leads to the second iterative method in ATKINSON [2], which is defined by

(3.6.a) $$f_{1,i+\frac{1}{2}} = \frac{1}{\lambda}\{g + K_1 f_{1,i}\}$$

(3.6.b) $$(\lambda I - K_0) f_{1,i+1} = g + (K_1 - K_0) f_{1,i+\frac{1}{2}}$$

In fact this method, including the Jacobi-relaxation sweep (3.6.a), is an example of the generalized defect correction principle (2.6), where

$$A_i^h = A_{i+\frac{1}{2}}^h = (\lambda I - K_1),$$

$$\tilde{B}_i^h = (\lambda I - K_0)^{-1} \quad \text{and} \quad \tilde{B}_{i+\frac{1}{2}}^h = \frac{1}{\lambda} I.$$

In this simple case, where the number of relaxation sweeps is fixed, the generalized principle (2.6) can also be considered as the principle (2.5) where

$$\tilde{B}^h = (\lambda I - K_0)^{-1}[I + \frac{1}{\lambda}(K_1 - K_0)].$$

Using the collectively compact operator theory of ANSELONE [1], Atkinson analyses the convergence of both methods (3.5) and (3.6). Without proof we give the following theorem from ATKINSON [2]. The hypotheses on K and $\{K_p \mid p \geq 0\}$ are as follows:

A1. K and $K_p$, $p \geq 0$, are linear operators on the Banach space X into X.

A2. $K_p f \to Kf$ as $h(p) \to 0$, for all $f \in X$. ($h(p)$ represents the mesh-width on grid $\mathbb{G}_p$).

A3. $\{K_p\}$ is a collectively compact family.

The following operator norm is used: $\|K_p\| = \max_{a \leq x \leq b} \sum_{j=1}^{N_p} |k_j(x)|$.

THEOREM 1. *Under the assumptions* A1-A3 *it can be shown that for* $p \geq 0$:

(3.7) $$\|K - K_p\| \geq \|K\|,$$

(3.8) $$\|(K - K_p)K\| \to 0 \quad \textit{as} \quad h(p) \to 0,$$

(3.9) $$\|(K - K_p)K_p\| \to 0 \quad \textit{as} \quad h(p) \to 0.$$

*If* $(\lambda I - K)^{-1}$ *exists, then for all sufficiently small* $h(p)$, *say* $h(p) \leq H(\lambda)$, $(\lambda I - K_p)^{-1}$ *exists and is uniformly bounded by* $c(\lambda)$. *Moreover*:

(3.10) $\|f-f_p\| \leq c(\lambda)\|Kf-K_pf\|, \quad h(p) \leq H(\lambda).$ □

In view of the following theorems we define:

$$d = \sup_{p\geq 0} \|K_p\|,$$

$$a_p = \sup_{q\geq p} \sup_{\ell\geq 0} \|(K-K_q)K_\ell\|.$$

It follows from (3.8) and (3.9), that $a_p \to 0$ as $h(p) \to 0$. Using theorem 1 we continue with the convergence analysis of method 1 of Atkinson. From (3.5) we get for the error

(3.11) $$f_1 - f_{1,i+1} = (\lambda I-K_0)^{-1}(K_1-K_0)(f_1-f_{1,i}).$$

The sequence $\{f_{1,i}\}$ will converge to $f_1$ if

$$\|(\lambda I-K_0)^{-1}(K_1-K_0)\| < 1.$$

Generally, this condition is not satisfied because of (3.7). But by applying (3.5) twice, a satisfactory result is obtained.

THEOREM 2 (ATKINSON [2]):

(3.12) $$\|[(\lambda I-K_0)^{-1}(K_1-K_0)]^2\| \leq A(\lambda) \equiv \frac{2(a_0+a_1)}{|\lambda|} c(\lambda)[1+dc(\lambda)].$$ □

It is clear that $A(\lambda)$ will be less than one for $h(0)$ sufficiently small. From (3.12) we get

$$\|f_1-f_{1,i+2}\| \leq A(\lambda)\|f_1-f_{1,i}\|, \quad i \geq 0,$$

so that convergence is obtained for $h(0)$ small enough. We remark that the errors can have an irregular behaviour as i increases. This imperfection is not present in method 2 of Atkinson.

The convergence analysis of method 2 is based on substituting (3.6.a) into (3.6.b), so that we get for the error

(3.13) $$f_1-f_{1,i+1} = \frac{1}{\lambda}(\lambda I-K_0)^{-1}(K_1-K_0)K_1(f_1-f_{1,i}).$$

THEOREM 3 (ATKINSON [2]):

$$(3.14)\qquad \|\frac{1}{\lambda}(\lambda I-K_0)^{-1}(K_1-K_0)K_1\| \le B(\lambda) \equiv \frac{1}{\lambda}c(\lambda)(a_0+a_1). \quad \square$$

It is again straightforward that $B(\lambda) < 1$ for $h(0)$ sufficiently small.

Atkinson gives in [3] a Fortran program which is based on method 2. It is divided into two stages:

A. The aim of this stage is to find a grid $\mathbb{G}_0$ for which $h(0)$ is small enough, i.e. so that the iteration (3.6) is successful.

B. In this stage the operator $K_1$ is found. In fact, this operator is determined iteratively as follows. For $p = 1,2,3,\ldots$, let $h(p) = \frac{1}{2}h(p-1)$. An operator $K_p$ is found on a grid $\mathbb{G}_p$ and a solution $f_p$ of $(\lambda I-K_p)f_p = g$ is computed by means of (3.6), where $K_1$ is replaced by $K_p$. Since $h(p) = h(0) * 2^{-p}$, we know that $\lim_{p\to\infty} h(p) = 0$ and therefore $K_p f$ will approximate $Kf$ as $p \to \infty$. This process is continued until $\|f_p-f_{p-1}\|$ is small enough.

We remark that stage B of the computational process can be considered as an application of the iterative updating defect correction principle with an approximate inverse

$$\tilde{B}_p = (\lambda I-K_0)^{-1}[I+\frac{1}{\lambda}(K_p-K_0)],$$

where $K_0$ has been fixed.

In the following section we extend method 2 of Atkinson by introducing the approximate inverse

$$(3.15)\qquad \tilde{B}_p = (\lambda I-K_{p-1})^{-1}[I+\frac{1}{\lambda}(K_p-K_{p-1})], \qquad p > 0,$$

with

$$\tilde{B}_0 = (\lambda I-K_0)^{-1}.$$

## 4. MULTI-GRID METHODS

The basic idea of multi-grid methods is to work with a sequence of grids of increasing refinement. These grids are simultaneously used to obtain an approximation of the solution of the original problem. Recently several authors were using this technique in relation with finite difference

or finite element equations arising from boundary-value problems. Here we mention the work of BRANDT [5], NICOLAIDES [6,7] and WESSELING [9,10]. In this paper we apply multi-grid methods to Fredholm integral equations of the second kind. We realize that the methods can be used in many ways. Here we focus our attention on the ideas of BRANDT [5] and we show their relation with the methods of the previous sections.

On a grid $\mathbb{G}_p$ a multi-grid method solves iteratively a system of algebraic equations by means of some relaxation algorithm and prescribed interactions with solutions on coarser grids. Let equation (3.1) be the problem to be solved. Using the notation of section 3 and defining $A_p = (\lambda I - K_p)$ we have to solve on grid $\mathbb{G}_p$

$$A_p f_p = g \equiv g_p. \tag{4.1}$$

Our multi-grid method to solve this problem on grid $\mathbb{G}_p$ is of a recursive type. Recursion takes place with respect to p. For each grid $\mathbb{G}_p$ the method is iterative and it consists of repeating the following two steps:

a. Perform a (fixed) number of relaxation sweeps, $\sigma$, on grid $\mathbb{G}_p$ and denote the result by $f_{p,i+\frac{1}{2}}$.
b. Solve on grid $\mathbb{G}_{p-1}$ the equation

$$A_{p-1} f_{p,i+1} = g_p + (A_{p-1} - A_p) f_{p,i+\frac{1}{2}} \equiv g_{p-1}. \tag{4.2}$$

This last equation is solved by the same multi-grid method, except when it has to be solved on grid $\mathbb{G}_0$. In that case it is solved directly, i.e. the linear system $P_0 A_0 f_{1,i+1} = P_0 g_0$ is solved by some direct method.

We terminate this multi-grid process on grid $\mathbb{G}_p$ if the residual is small enough, i.e.

$$\|r_p\| \equiv \|g_p + A_p f^I\| < \varepsilon_p, \tag{4.3}$$

where $f^I$ is the iteratively determined approximation for the solution $f_p$.

In the above method we can still choose the relaxation algorithm and the number of relaxation sweeps. In BRANDT [5] the relaxation algorithm depends on the problem to be solved and the number of sweeps depends on the absence of high-frequency errors in the residual. The heuristic arguments to relate the number of relaxation sweeps to the residual function are

given in the introduction. Using local mode analysis Brandt was able to formulate a suitable criterion for finite difference equations. However, for integral equations this approach needs some modification, which requires further research.

Following section 3 we propose to choose the Jacobi-relaxation algorithm. Fixing the number of sweeps on zero or one we get multi-grid variants of methods 1 and 2 of Atkinson, respectively. Performing one relaxation sweep we get the following algorithm for solving (4.1):

$$(4.4.a) \qquad f_{p,i+\frac{1}{2}} = \frac{1}{\lambda} g_p + (I - \frac{1}{\lambda} A_p) f_{p,i}$$

$$(4.4.b) \qquad A_{p-1} f_{p,i+1} = g_p + (A_{p-1} - A_p) f_{p,i+\frac{1}{2}}.$$

This iterative method can be interpreted as a special case of defect correction with the approximate inverse $\tilde{B}_p$ of $A_p$ given by (3.15). If we restrict (4.4) to two computational grids $\mathbb{G}_0$ and $\mathbb{G}_1$ it is exactly (3.6), i.e. method 2 in ATKINSON [2]. The convergence of (3.6) is given in theorem 3, which can be generalized to the convergence of (4.4). For stage B of Atkinson's program (discussed at the end of section 3) the convergence factor $B(\lambda)$ of (3.14) is given by $\frac{1}{\lambda} c(\lambda)(a_0+a_p)$, while for (4.4) it is given by $\frac{1}{\lambda} c(\lambda)(a_{p-1}+a_p)$. Because of $a_p \leq a_{p-1}$ it is to be expected that, once we have found two grids $\mathbb{G}_0$ and $\mathbb{G}_1$ for which (3.6) converges, (4.4) will do for any pair of grids $\mathbb{G}_{p-1}$ and $\mathbb{G}_p$, $p \geq 1$.

If our multi-grid process converges, we get a continuous approximation $f^I \in C[a,b]$ to the solution of (4.1). However, in a practical implementation we compute the discrete solution on the finest grid. In the following Algol 68-comparable program we describe the presented multi-grid process where the fixed number of consecutive relaxation sweeps is $\sigma$. In the program below the procedure call P(k,z) yields the restriction of the gridfunction z on $\mathbb{G}_{k+1}$ to the grid $\mathbb{G}_k$ and the procedure call Q(k,f,z) yields the natural interpolation of the gridfunction z on $\mathbb{G}_{k-1}$ to the grid $\mathbb{G}_k$, i.e. $Q(k,f,z) = \frac{1}{\lambda}(f+K_{k-1}z)$. Thus the procedures P and Q are analogous to the operators $P^h$ and $Q^h$ in section 3.

```
proc solve on grid 𝔾 = (int k, vec f_k, g_k, real ε_k) void:
if k = 0
then f_k := A_0^-1 * g_k
else for i while vec r_k = g_k - A_k * f_k;
                 maximum norm (r_k) > ε_k
```

```
    do if i mod (σ+1) = 0
       then vec f_{k-1} = P(k-1,f_k);
            vec g_{k-1} = P(k-1,r_k) + A_{k-1} * f_{k-1};
            solve on grid G(k-1,f_{k-1},g_{k-1},ε_{k-1});
            f_k := f_k + Q(k,r_k,f_{k-1}-P(k-1,f_k))
       else f_k := f_k + (1/λ) r_k
       fi
    od
fi;
```

We conclude this section with a numerical example which has also been considered by ATKINSON [2,3] and BRAKHAGE [4]:

$$\lambda f(s) - \int_0^{2\pi} K_\rho(s,t) f(t)\,dt = g(s)$$

where

$$K_\rho(s,t) = \frac{1}{2\pi}(1-\rho^2)[1-2\rho\cos(s-t)+\rho^2]^{-1}.$$

This integral equation is a classical reformulation of the Dirichlet problem for $\Delta u \equiv 0$ on an ellips. The right-hand side g(s) is chosen so that f(s) = = cos s. The parameters $\lambda$ and $\rho$ are chosen to be 0.9 and 0.8, respectively, so that $\|(\lambda I - K_\rho)^{-1}\|$ is large. This means that large linear systems are necessary to obtain a reasonably accurate solution $f^h(s)$ with respect to the true solution f(s). Further, we remark that $\lambda_{max}$, the largest eigenvalue of $K_\rho$, is equal to one. Hence, if we would apply Jacobi-relaxation only, this process would diverge because its convergence factor is $\frac{1}{\lambda}\lambda_{max} > 1$.

The operator $\bar{K}^h$ is defined using the repeated trapezoidal rule. We are considering grids $G_p$ in which the nodal points $\{x_j\}$ are equidistantly distributed. Further, we let h(p) = ½h(p-1). Hence, the weights $\{w_j\}$ on grid $G_p$ are given by $\{\frac{1}{2}h(p), h(p), \ldots, h(p), \frac{1}{2}h(p)\}$.

In view of the question of how many relaxation sweeps we have to perform, so that the sequence $\{f_{p,i}\}$ converges to the true solution f in an optimal way, we have set the maximum number of sweeps $\sigma$ to 0,1,2,3, respectively. For the above choice of $\lambda$ and $\rho$ the initial grid $G_0$ needs 33 nodal points, which is the same number as in [3]. The multi-grid process is terminated as an estimated error $\|f - f_{p,i}\| < tol$. (The construction of this

estimate requires further research.)

For tol = $10^{-5}$ we illustrate the behaviour of the residual $r_{p,i}$ of our multi-grid algorithm in the tables 1.1-1.4. We also mention the number of kernel evaluations (= KEVAL) and the final error $\|f-f^I\|$.

| grid $G_p$, p = | $\|r_{p,i}\|$ |
|---|---|
| 0 | 0.0 |
| 1 | $3.1\ 10^{-3}$ |
| 0 | 0.0 |
| 1 | $1.5\ 10^{-3}$ |
| 0 | 0.0 |
| 1 | $8.2\ 10^{-5}$ |
| 0 | 0.0 |
| 1 | $2.3\ 10^{-5}$ |
| 0 | 0.0 |
| 1 | $1.7\ 10^{-6}$ |
| 2 | $3.6\ 10^{-6}$ |
| ---- | ----- |
| KEVAL | 112,192 |
| $\|f-f^I\|$ | $2.3\ 10^{-5}$ |

Table 1.1 ($\sigma$=0, tol=$10^{-5}$), No relaxation sweeps.

| grid $G_p$, p = | $\|r_{p,i}\|$ |
|---|---|
| 0 | 0.0 |
| 1 | $3.1\ 10^{-3}$ |
| 1 | $1.3\ 10^{-3}$ |
| 0 | 0.0 |
| 1 | $3.9\ 10^{-5}$ |
| 1 | $1.7\ 10^{-5}$ |
| 0 | 0.0 |
| 1 | $5.2\ 10^{-7}$ |
| 2 | $2.6\ 10^{-6}$ |
| ---- | ----- |
| KEVAL | 100,864 |
| $\|f-f^I\|$ | $1.4\ 10^{-5}$ |

Table 1.2 ($\sigma$=1, tol=$10^{-5}$), One relaxation sweep.

| grid $\mathbb{G}_p$, p = | $\|r_{p,i}\|$ |
|---|---|
| 0 | 0.0 |
| 1 | $3.1\ 10^{-3}$ |
| 1 | $1.3\ 10^{-3}$ |
| 1 | $1.1\ 10^{-3}$ |
| 0 | 0.0 |
| 1 | $3.2\ 10^{-5}$ |
| 1 | $1.3\ 10^{-5}$ |
| 1 | $1.2\ 10^{-5}$ |
| 0 | 0.0 |
| 1 | $3.3\ 10^{-7}$ |
| 2 | $2.6\ 10^{-6}$ |
| ---- | ---- |
| KEVAL | 109,440 |
| $\|f-f^I\|$ | $1.3\ 10^{-5}$ |

Table 1.3 ($\sigma$=2, tol=$10^{-5}$), Two relaxation sweeps.

| grid $\mathbb{G}_p$, p = | $\|r_{p,i}\|$ |
|---|---|
| 0 | 0.0 |
| 1 | $3.1\ 10^{-3}$ |
| 1 | $1.3\ 10^{-3}$ |
| 1 | $1.1\ 10^{-3}$ |
| 1 | $1.0\ 10^{-3}$ |
| 0 | 0.0 |
| 1 | $2.8\ 10^{-5}$ |
| 1 | $1.2\ 10^{-5}$ |
| 1 | $1.0\ 10^{-5}$ |
| 1 | $9.0\ 10^{-6}$ |
| 0 | 0.0 |
| 1 | $2.5\ 10^{-7}$ |
| 2 | $2.5\ 10^{-6}$ |
| ---- | ---- |
| KEVAL | 118,016 |
| $\|f-f^I\|$ | $1.3\ 10^{-5}$ |

Table 1.4 ($\sigma$=3, tol=$10^{-5}$), Three relaxation sweeps.

From these tables we see that for this example one relaxation sweep gives the minimal number of kernel evaluations. However, ATKINSON [3] gives a numerical example where $\lambda$ is close to zero, for which method 1 (no relaxation sweep) converges, whereas method 2 (one relaxation sweep) diverges. Obviously, the optimal number of relaxation sweeps depends on the problem to be solved.

In table 2 we show the behaviour of $\|r_{p,i}\|$ for $tol^* = 10^{-8}$ and one relaxation sweep. In table 3 we give the results for tol = $10^{-5}$, one relaxation sweep, while $\bar{K}^h$ is now defined by the Simpson rule.

| grid $G_p$, p = | $\|r_{p,i}\|$ |
|---|---|
| 0 | 0.0 |
| 1 | $3.1\ 10^{-3}$ |
| 1 | $1.3\ 10^{-3}$ |
| 0 | 0.0 |
| 1 | $3.9\ 10^{-5}$ |
| 1 | $1.7\ 10^{-5}$ |
| 0 | 0.0 |
| 1 | $5.2\ 10^{-7}$ |
| 1 | $2.3\ 10^{-7}$ |
| 0 | 0.0 |
| 1 | $6.9\ 10^{-9}$ |
| 1 | $3.1\ 10^{-9}$ |
| 2 | $2.4\ 10^{-6}$ |
| 3 | $9.9\ 10^{-7}$ |
| 3 | $8.7\ 10^{-7}$ |
| 2 | $7.7\ 10^{-7}$ |
| 1 | $6.8\ 10^{-7}$ |
| 0 | 0.0 |
| 1 | $1.9\ 10^{-8}$ |
| 2 | $8.0\ 10^{-9}$ |
| 3 | $7.0\ 10^{-9}$ |
| 4 | $6.1\ 10^{-9}$ |
| KEVAL | 881,280 |
| $\|f-f^I\|$ | $5.6\ 10^{-8}$ |

Table 2 ($\sigma$=1, tol=$10^{-8}$), $\bar{K}^h$ defined by the trapezoidal rule,
One relaxation sweep.

| grid $G_p$ p = | $\|r_{p,i}\|$ |
|---|---|
| 0 | 0.0 |
| 1 | $3.2\ 10^{-2}$ |
| 1 | $7.2\ 10^{-3}$ |
| 0 | 0.0 |
| 1 | $2.2\ 10^{-3}$ |
| 1 | $6.0\ 10^{-4}$ |
| 0 | 0.0 |
| 1 | $1.8\ 10^{-4}$ |
| 1 | $4.7\ 10^{-5}$ |
| 0 | 0.0 |
| 1 | $1.4\ 10^{-5}$ |
| 1 | $3.8\ 10^{-6}$ |
| 2 | $7.0\ 10^{-4}$ |
| 2 | $1.4\ 10^{-4}$ |
| 3 | $1.2\ 10^{-4}$ |
| 3 | $1.1\ 10^{-4}$ |
| 2 | $9.4\ 10^{-5}$ |
| 1 | $8.3\ 10^{-5}$ |
| 0 | 0.0 |
| 1 | $2.4\ 10^{-5}$ |
| 1 | $5.4\ 10^{-6}$ |
| 2 | $5.1\ 10^{-6}$ |
| 3 | $4.1\ 10^{-6}$ |
| 4 | $3.6\ 10^{-6}$ |
| KEVAL | 918,784 |
| $\|f-f^I\|$ | $3.3\ 10^{-5}$ |

Table 3 ($\sigma$=1, tol=$10^{-5}$), $\bar{K}^h$ defined by Simpson's rule,

One relaxation sweep.

From the results given in the tables 1.1 and 1.2 we expect that similar results can be produced by the iterative schemes given in ATKINSON [2]. In fact, in the initial stages only two computational grids $\mathbb{G}_0$ and $\mathbb{G}_1$ are used and finally the solution is interpolated to grid $\mathbb{G}_2$. However, in table 2 ($tol=10^{-8}$) we see that a multi-grid sweep has been performed on grid $\mathbb{G}_3$. In the near future we shall investigate the efficiency of the program of ATKINSON [3] for this case.

The use of a quadrature rule of higher order (i.e. Simpson's rule) for definition of the operator $\bar{K}^h$ gives worse results for this problem as can be concluded from comparing the tables 1.2 and 3. In fact, Simpson's rule generates a larger convergence factor $B(\lambda)$ in (3.14) than the trapezoidal rule does.

Summarizing we conclude that our multi-grid algorithm gives good results for the considered problem. In general, we expect that the final solution can be found in about $0(N^2)$ kernel evaluations, where N is the number of unknowns on the finest grid. However, further research is required to make firm statements about computational complexity.

Finally we remark that our multi-grid method can also be used (and is used) for the numerical solution of non-linear integral equations. In that case, only on grid $\mathbb{G}_0$ Newton iteration is used to solve the non-linear system of equations resulting from applying the Nyström method.

## 5. DISCUSSION

In this paper we have introduced a new iterative method for the numerical solution of Fredholm equations of the second kind. This method is essentially a multi-grid method. It can be interpreted as an extension of the iterative schemes discussed in ATKINSON [2,3]. Both these schemes and our multi-grid method can be formulated within the general framework of the defect correction principle as introduced by STETTER [8]. In section 3 and 4 we discussed iterative processes of type (2.6):

$$f^h_{i+1} = (I-\tilde{B}_{p,i}A_p)f^h_i + \tilde{B}_{p,i}g.$$

We made the following choices for the approximate inverses $\tilde{B}_{p,i}$ of $A_p$:

a) $\tilde{B}_p = \frac{1}{\lambda} I$, the Jacobi relaxation sweep, and

b) $\tilde{B}_p = A^{-1}_{p-1}$, the correction from a coarser grid.

From the numerical results in section 4 we expect that this method can be more efficient than the methods discussed in ATKINSON [2,3]. However, from the general formulation (2.6) we are led to the following questions:

- how should we alternate between relaxation sweeps and coarse grid corrections in order to (approximately) minimize the amount of computational work,
- are there better choices for $\tilde{B}_p$, for which the solution $f^h$ can be obtained with even less computational work.

From our numerical example we have also seen a clear dependence between the computational work and the definition of the operator $K_p$ ($A_p = \lambda I - K_p$). This obviously raises the question of how to define the operator $A_p$ (or $A_{p,i}$).

All these questions are combined in the one question formulated at the end of section 2: How to choose the sequences $\{A_{p,i}\}$ and $\{B_{p,i}\}$ in order to find efficiently a $f_i^h$ which approximates in an optimal way the solution f?

ACKNOWLEDGEMENT

The author wishes to thank dr. P.W. Hemker and dr. H.J.J. te Riele of the Mathematical Centre (Amsterdam) for the many helpful discussions.

REFERENCES

[1] ANSELONE, P.M., *Collectively compact operator approximation theory*, Englewood Cliffs, New Jersey, Prentice-Hall, 1971.

[2] ATKINSON, K.E., *Iterative variants of the Nyström method for the numerical solution of integral equations*, Numerische Mathematik, 22 (1973), pp. 17-31.

[3] ATKINSON, K.E., *A survey of numerical methods for the solution of Fredholm integral equations of the second kind*, SIAM, 1976.

[4] BRAKHAGE, H., *Uber die numerische behandlung von integralgleichungen nach der Quadratur-formelmethode*, Numerische Mathematik, 2 (1960), pp. 183-196.

[5] BRANDT, A., *Multi-level adaptive solutions to boundary-value problems*, Mathematics of computation, 31 (1977), pp. 333-390.

[6] NICOLAIDES, R.A., *On multiple grid and related techniques for solving discrete elliptic systems,* T. Comp. Phys. 19, (1975), pp. 418-431.

[7] NICOLAIDES, R.A., *On the $\ell^2$ convergence of an algorithm for solving finite element equations,* Mathematics of Computation, 31 (1977), pp. 892-906.

[8] STETTER, H.J., *The defect correction principle and discretization methods,* Numerische Mathematik, 29, (1978), pp. 425-443.

[9] WESSELING, P., *Numerical solution of the stationary Navier-Stokes equations by means of a multiple grid method and Newton iteration,* Report NA-18, Delft University of Technology (1977).

[10] WESSELING, P., *A convergence proof for a multiple grid method,* Report NA-21, Delft University of Technology (1978).

# INTRODUCTION AND GLOBAL SURVEY OF NUMERICAL METHODS FOR INTEGRAL EQUATIONS

H.J.J. te RIELE

# 1. INTRODUCTION

A linear Fredholm integral equation of the first kind is given by

$$(1.1) \qquad \int_0^1 K(x,y)f(y)dy = g(x), \qquad 0 \le x \le 1,$$

where K (the kernel function) and g (the data function) are given functions, defined on $[0,1] \times [0,1]$ and $[0,1]$, respectively, and the function f is sought on $[0,1]$. We shall assume f and g to be elements of certain spaces F and G, respectively. Defining the linear operator $K: F \to G$ by $(Kf)(x) := \int_0^1 K(x,y)f(y)dy$, we write (1.1) in operator notation as

$$(1.2) \qquad Kf = g, \qquad g \in G \text{ given}, \ f \in F \text{ sought}.$$

In general, numerical solution of (1.1) is a difficult task, because this problem belongs to the class of so-called *ill-posed* problems. The problem (1.2) is ill-posed (in the sense of Hadamard) if at least one of the following assertions is false (F and G are assumed to be complete metric spaces):

(i) for every $g \in G$ there exists a solution $f \in F$;

(ii) its solution is unique;

(iii) its solution depends continuously on the data g.

Note that a problem is ill-posed with respect to two given metric spaces. It may be well-posed (i.e., when assertions (i)-(iii) are true) in other metrics.

Usually, the ill-posedness of (1.1) is shown by the following argument. According to a theorem of Riemann-Lebesgue (see e.g. NATANSON [17, p. 281]), if K(x,y) is integrable in the sense of Lebesgue (as a function of y for any $x \in [0,1]$), then

$$\lim_{m\to\infty} \int_0^1 K(x,y)\sin my\, dy = 0, \qquad \text{for any } x \in [0,1].$$

Replacing f(y) in (1.1) by $f(y)+c\sin my$ yields a perturbation of the right-hand side g(x), which can be made arbitrarily small by choosing m large enough. This perturbation can be caused by any finite perturbation of f (with maximum value $|c|$), so that the solution f of (1.1) does not depend continuously on the data g; hence assertion (iii) for (1.1) is false.

In practical situations the data function g(x) is very often the output of some measuring process, and then it is only approximately known on a discrete set of points $x_i \in [0,1]$. Consequently, instead of (1.2) it is more

realistic to consider the problem

$$Kf = \tilde{g}, \tag{1.3}$$

where only $\tilde{g}$ and $\varepsilon$ are known such that $\|g-\tilde{g}\| \leq \varepsilon$. This can cause $\tilde{g}$ to lie outside the range of the operator $K$, so that there exists no solution of (1.3).

Another possibility is that the kernel is degenerate (also called finitely decomposable), i.e., it can be written as

$$K(x,y) = \sum_{i=1}^{n} K_{1i}(x)K_{2i}(y), \tag{1.4}$$

where n is as small as possible, i.e., both $\{K_{1i}(x)\}_{i=1}^{n}$ and $\{K_{2i}(x)\}_{i=1}^{n}$ are a set of linear independent functions. Substituting (1.4) into (1.1) shows that a necessary condition for the existence of a solution of (1.1) is that g(x) is a linear combination of the functions $K_{11}(x),\ldots,K_{1n}(x)$. Obviously, a small error in g can violate this condition.

Also assertion (ii) can readily be false for (1.2), i.e., when there exists a function $\phi$ such that $K\phi = 0$. It follows that when f is a solution of (1.2), then also $f + c\phi$ is a solution, since $K(f+c\phi) = Kf + cK\phi = Kf = g$.

So far, we have only spoken of properties of the *continuous* problem (1.1). It is evident that, whenever one of the assertions (i), (ii) or (iii) is false for (1.2), this will cause serious troubles when one tries to compute a numerical solution of a discretized version of (1.1) which is "close" to it, in some sense. Usually, equation (1.1) is discretized as follows. The integral in (1.1) is replaced by a quadrature formula with weights $w_1,\ldots,w_n$, and quadrature points $y_1,\ldots,y_n$. Writing down the resulting equation for $x = x_1,\ldots,x_m$ yields the matrix-vector equation

$$AW\vec{f} = \vec{g}, \tag{1.5}$$

where $W = \mathrm{diag}(w_1,\ldots,w_n)$, A is an m×n-matrix with elements $A_{ij} = K(x_i,y_j)$, $\vec{g}$ is an m-vector whose i-th component represents $g(x_i)$, and $\vec{f}$ is an unknown n-vector whose j-th component (hopefully) represents an approximation of $f(y_j)$. However, direct solution of the system (1.5) generally yields a vector $\vec{f}$ whose components will oscillate very wildly around the corresponding values of the solution of (1.1).

Fredholm integral equations of the first kind occur in mathematical models of many physical problems (cf. HILGERS [11] and NEDELKOV [18]). One often has a situation where f corresponds to some actual physical phenomenon that one wishes to measure, whereas one can only measure g which is the result of the influence of some (smoothing) device which acts on f and which is characterized by the kernel K(x,y).

An example is the problem of the reconstruction of subsurface structures from gravity measurements made on the surface of earth. Other examples occur in spectroscopy, X-ray scattering, chemical analysis, calculations of atmospheric temperature profiles from satellite measurements, optimal control, and so on.

Classical mathematical problems which can be formulated as Fredholm integral equations of the first kind are: the problem of harmonic continuation, numerical inversion of the Laplace transform, the backwards heat equation, and numerical differentiation.

When solving boundary value problems in partial differential equations, it can sometimes be convenient to convert them into a Fredholm integral equation of the first kind. For an interesting paper on this topic we refer to NOBLE [19].

In this paper we shall discuss a number of common numerical methods for solving Fredholm integral equations of the first kind, with smooth and with non-smooth kernels. In Section 2 we quote some results from the Hilbert-Schmidt theory on (1.1), in terms of singular functions and singular values. This theory is extensively described in SMITHIES [24]. It provides the apparatus for the analysis of the numerical methods. In Section 3 we shall discuss some methods for Fredholm equations of the first kind with *non-smooth* kernel. This type of equations is generally less difficult to solve numerically than equations with a smooth kernel. In Section 4 we describe and discuss numerical methods for equations (1.1) with a smooth kernel. In his contribution to this colloquium, J.J.M. Cuppen discusses regularization methods, where he estimates the regularization parameter with a simplified version of a method recently proposed by N. Köckler. Therefore, in this contribution to the colloquium we pay relatively little attention to regularization methods (and relatively much attention to truncated series and iterative methods).

## 2. SINGULAR FUNCTION ANALYSIS

Equation (1.1) can be formally studied in terms of singular values and singular functions. We assume from now on that in (1.2), $F = G = L^2[0,1]$.

We define the adjoint operator $K^*$: $L^2[0,1] \to L^2[0,1]$ of $K$ by

$$(K^*g)(y) = \int_0^1 K(x,y)g(x)\,dx. \tag{2.1}$$

With the usual $L^2$-inner product, if $f \in L^2[0,1]$, then

$$\begin{aligned}(Kf,g) &= \int_0^1\int_0^1 K(x,y)f(y)g(x)\,dy\,dx \\ &= \int_0^1 f(y)\{\int_0^1 K(x,y)g(x)\,dx\}dy = (f,K^*g).\end{aligned} \tag{2.2}$$

Now we suppose that the kernel $K(x,y)$ of (1.1) has the following expansion in terms of the singular functions $u_i(x)$, $v_i(y)$ and singular values $\kappa_i$:

$$K(x,y) = \sum_{i=1}^{\infty} \kappa_i u_i(x) v_i(y).$$

The sequences $\{u_i(x)\}$ and $\{v_i(y)\}$ are then orthonormal systems with the properties

$$Kv_i = \kappa_i u_i \quad \text{and} \quad K^*u_i = \kappa_i v_i. \tag{2.3}$$

Hence, $KK^*u_i = \kappa_i^2 u_i$ and $K^*Kv_i = \kappa_i^2 v_i$, so that $u_i$ and $v_i$ are eigenfunctions of the symmetric operators $KK^*$ and $K^*K$, respectively, both with eigenvalues $\kappa_i^2$. The $\kappa_i^2$ are positive, and, without loss of generality, we may assume $\kappa_i > 0$. Furthermore, we assume that the $\kappa_i$ are ordered so that $\kappa_1 \geq \kappa_2 \geq \ldots \geq \kappa_i \geq \kappa_{i+1} \geq \ldots$ . Now one readily verifies that, if $g(x)$ can be expanded in the form

$$\begin{aligned}g(x) &= \sum_{i=1}^{\infty} g_i u_i(x), \qquad \text{with} \\ g_i &= (g,u_i) = \int_0^1 g(x)u_i(x)\,dx,\end{aligned} \tag{2.4}$$

then, at least formally, a solution of (1.1) is given by

$$f(y) = \sum_{i=1}^{\infty} \frac{g_i}{\kappa_i} v_i(y). \tag{2.5}$$

Indeed, we have by (2.5), (2.3) and (2.4)

$$Kf = \sum_i \frac{g_i}{\kappa_i} Kv_i = \sum_i g_i u_i = g.$$

Now we can state the following results:

(i) (PICARD [21]) (1.1) has an $L^2$-solution if and only if
   (a) $\Sigma_{i=1}^{\infty} |g_i|^2/\kappa_i^2 < \infty$ and
   (b) whenever u is such that $K^*u = 0$, then $(g,u) = 0$;

(ii) the solution of (1.1) is unique if and only if the equation $Kv = 0$ has only the trivial solution $v = 0$.

The ill-posed character of (1.1) can also be shown from this singular function analysis as follows (cf. MILLER [16]). If the solution of (1.1) exists and is unique, then according to the Hilbert-Schmidt theory the sequence of singular values is infinite, and $\lim_{i\to\infty} \kappa_i = 0$. Perturbing g(x) by changing the coefficient $g_i$ in (2.5) by an amount $\delta g_i$ causes a perturbation of f in (2.5) of $\delta f = \delta g_i v_i/\kappa_i$. Hence the ratio $\|\delta f\|/\|\delta g\| = 1/\kappa_i$ can be made arbitrarily large by choosing i large enough. From (2.5) we can now draw the following conclusion: the problems in solving (1.1) numerically are fully determined by the rate with which the singular values $\kappa_i$ approach zero, as i tends to infinity. The faster this sequence tends to zero, the slower will the series (2.5) converge. Some insight can be obtained from the case of the one-dimensional difference kernel $K(x,y) = x - y$. The asymptotic behaviour of the corresponding $\kappa_i$ can be studied by Fourier methods, and it then turns out (cf. NOBLE [19]) that the *smoother* the kernel, the faster the $\kappa_i$ tend to zero as i increases. When solving first kind integral equations numerically, it can therefore be of much help, to have a (weakly) singular kernel!

## 3. SINGULAR FREDHOLM EQUATIONS OF THE FIRST KIND

Singular Fredholm equations of the first kind occur very frequently in potential theory and in elastostatics. The book of JASWON and SYMM [14] provides an interesting survey of this matter. In this colloquium G. De Mey has discussed some applications in semiconductor physics.

As we have seen in Section 2, a (weak) singularity of the kernel gives us hope to be able to solve the integral equation numerically, without too much trouble. One often uses product integration, assuming the unknown function to be piecewise constant between two grid points. A difficulty which can occur now is the (numerical) evaluation of the moment integrals (over the singular kernel). Generally, *no* numerical problems are caused here by the ill-posedness associated with first kind equations.

As an example, we mention the following problem discussed by NOBLE [19]. We are given a flat conducting rectangular lamina in free space, provided with an electrostatic charge. Suppose that the total charge density at a point $(\xi,\eta)$ on the lamina, lying in $z = 0$, $-a \le x \le a$, $-b \le y \le b$, is $f(\xi,\eta)$. If the plate is assumed to be at unit potential, one can deduce for $f(\xi,\eta)$ the two-dimensional Fredholm integral equation (with singular kernel):

$$\int_{-a}^{a}\int_{-b}^{b} \frac{f(\xi,\eta)\,d\eta\,d\xi}{[(x-\xi)^2+(y-\eta)^2]^{\frac{1}{2}}} = 1, \qquad \begin{cases} -a \le x \le a \\ -b \le y \le b \end{cases}.$$

Special attention is paid to the choice of the subdivisions for x and y, since on physical grounds it is known that the charge distribution $f(x,y)$ tends to infinity as $(x,y)$ tend to the edge of the rectangular domain $[-a,a] \times [-b,b]$. Because of the symmetry of the problem, one needs only consider the domain $[0,a] \times [0,b]$. Noble chooses $x = a_i$, $1 \le i \le m$, where $a_i$ is determined by the formula

$$\frac{a_i}{a} = \frac{(m-i+1)^k+\ldots+(m-1)^k+m^k}{1^k+2^k+\ldots+(m-1)^k+m^k}.$$

Here k is a constant which, when increased, narrows the intervals near $x = a$. The interval $[0,b]$ is subdivided in a similar way. The numerical results are satisfactory.

Recently, RICHTER [23] has quantified the relationship between the smoothness of the kernel and the amenability of the problem (1.1) to standard (direct discretization) methods. Richter uses a *smoothing* index $\ell$ which essentially estimates the number of derivatives which are added by the integral operator $K$ to $L^2$-functions to which it is applied. The more severe the singularity in the kernel function, the smaller the value of $\ell$ will be. Richter suggests that when $\ell = 1$ or 2 one has a good chance, to find an acceptable solution of (1.1) via a direct product integration method.

Other important references are HSIAO and MACCAMY [12] and HSIAO and WENDLAND [13], where elliptic boundary value problems are solved via integral equations of the first kind with (among others) logarithmic singularities in the kernel.

## 4. METHODS FOR SOLVING (1.1) WITH A SMOOTH KERNEL

All numerical methods for solving (1.1) can be interpreted in terms of weighting the higher terms of the series (2.5). The problem is that with increasing i the singular functions $v_i$ become increasingly rapidly oscillating, and when g is contaminated with error (from a measuring process or from the discretization of the continuous problem), one can no longer distinguish between real fluctuations in f and fluctuations caused by the error in g. Therefore it seems reasonable to drop some terms, or weight them by a filter, and this has been the subject of many investigations.

We shall describe here four methods, viz. regularization methods, expansion methods, iterative methods and probabilistic methods. It is important to notice that with neither of these four methods problem (1.1) can be solved in a fully satisfactory way. Because of the ill-posedness of the problem, one should always try to use as much information about its solution (known on physical grounds) as possible.

### 4.1. Regularization methods.

Regularization methods for the numerical solution of (1.1) were first considered by TIHONOV [27 - 30] and PHILLIPS [20], independently from each other, and later on by TWOMEY [35], RIBIÈRE [22], TIHONOV and GLASKO [32, 33], among many others. The recent book of TIHONOV and ARSENIN [31] gives a good impression of the extensive Russian literature on this subject.

The regularization method consists in replacing the ill-posed problem (1.3) by the well-posed problem:

(4.1) minimize the functional

$$\Phi_\alpha(f) = \|Kf-\tilde{g}\|^2 + \alpha\|Cf\|^2$$

over all f in a compact set satisfying $\|Kf-\tilde{g}\| \leq \varepsilon$.

Here, $\alpha$ is some fixed positive number, the so-called regularization

parameter, and C is some linear operator, e.g., Cf = f, or Cf = f', or Cf = f", or if some initial approximation $\hat{f}$ of f is known: Cf = f - $\hat{f}$. Under certain mild conditions problem (4.1) has a unique solution, denoted by $f_\alpha$. Moreover, $f_\alpha$ will converge to the solution of (1.1) as $\alpha \to 0$, provided that $\varepsilon^2 \to 0$ no less rapidly than $\alpha$. The problem is how to choose $\alpha$. If we let $\alpha$ increase, then the residue $\|Kf_\alpha - \tilde{g}\|$ will increase, whereas $\|Cf_\alpha\|$ will decrease, and conversely. The linear operator C is often chosen in such a way that it will help to suppress wild oscillations in functions f for which $\|Kf - \tilde{g}\| \leq \varepsilon$. Of course, this effect should not be too strong so that *all* oscillations in f are damped out! Using a variational argument, the solution of (4.1) can be shown to be the solution of

$$(K^*K + \alpha C^*C)f = K^*\tilde{g}. \tag{4.2}$$

In case C = I, the identity operator, one easily obtains the singular function expansion for $f_\alpha$:

$$f_\alpha(y) = \sum_{i=1}^{\infty} \frac{\kappa_i \tilde{g}_i}{\kappa_i^2 + \alpha} v_i(y), \tag{4.3}$$

provided that $\tilde{g}(x)$ admits the expansion

$$\tilde{g}(x) = \sum_{i=1}^{\infty} \tilde{g}_i u_i(x), \qquad \tilde{g}_i = (\tilde{g}, u_i). \tag{4.4}$$

Comparing (4.3) with (2.5) we see that the effect of regularization is the insertion of a *filter* factor $\kappa_i^2/(\kappa_i^2 + \alpha)$. This factor is close to unity as long as $\kappa_i$ is large compared with $\alpha$ (small i), but tends to zero as $\kappa_i \to 0$ ($i \to \infty$). Other choices of C can lead to other filter factors.

More about regularization methods and a method where the singular function expansion is truncated can be found in the contribution of J.J.M. Cuppen to this colloquium.

## 4.2. Expansion methods

One obvious filtering method is to calculate successive terms of the singular function expansion of f, and terminate at a suitable point. BAKER, FOX, MAYERS and WRIGHT [3] were the first who proposed such a method for symmetric kernels K(x,y). They immediately discretized (1.1) with m = n,

yielding a square symmetric matrix A in (1.5), and they expressed their method in terms of eigenvalues and eigenvectors. In the case of a non-symmetric kernel K(x,y) (yielding a non-symmetric matrix A), one can use the concepts of singular values and singular vectors. We shall sketch here the method of Baker c.s. in some more detail, because of its fundamental importance. Suppose that $\lambda_i$ is an eigenvalue and $\vec{f}^{(i)}$ a corresponding eigenvector of the matrix AW in (1.5). Then if the vector $\vec{g}$ can be expressed as $\vec{g} = \Sigma_{i=1}^{n} a_i \vec{f}^{(i)}$ then the solution of (1.5) is given by

$$\vec{f} = \sum_{i=1}^{n} \frac{a_i}{\lambda_i} \vec{f}^{(i)}. \tag{4.5}$$

The supposition now, as in the continuous case, is that to the smallest eigenvalues correspond high frequency components of $\vec{g}$, which are caused by the error in $\vec{g}$ and which therefore should be neglected. The problem now, of course, is where to "chop" the series (4.5). The strategy suggested by Baker c.s. is based on numerical experiments with problems with a known solution. Let $\vec{f}^{(s)} = \Sigma_{i=1}^{s} (a_i/\lambda_i)\vec{f}^{(i)}$, $1 \le s \le n$. They suggest to compute successively $\vec{f}^{(1)}, \vec{f}^{(2)}, \ldots$ and to examine, each time a new eigenvector and - value has been computed, how accurately the function g can be fitted by a linear combination of the computed eigenvectors. If sufficient accuracy has been obtained, one is suggested to compute these eigenvectors more accurately, using smaller intervals.

Generalizations of this work to nonsymmetric kernels have been investigated by HANSON [8], CRONE [4] and VARAH [36]. Varah gives an error estimate which, when minimized, predicts the number of terms to be used in the singular vector expansion of the solution of (1.5). In the three examples with known solution which he treats in detail (harmonic continuation, numerical inversion of the Laplace transform and the backwards heat equation), the *theoretically* optimal number of terms agreed with the number of terms prescribed by the error estimate.

## 4.3. Iterative methods

LANDWEBER [15] was the first who considered an iterative method for solving (1.1), viz., the scheme

$$f^{(k+1)} = f^{(k)} + K^*(g - Kf^{(k)}), \qquad k = 0,1,\ldots, \; f^{(0)} = 0. \tag{4.6}$$

It is not difficult to show that, in terms of singular values and functions, if $g \in U = \text{span}(u_1, u_2, \ldots)$, then

$$f^{(k)}(y) = \sum_{i=1}^{\infty} \{1 - (1 - \kappa_i^2)^k\} \frac{g_i}{\kappa_i} v_i(y), \tag{4.7}$$

and that this series converges to the solution of (1.1) provided that $0 < \kappa_1^2 < 2$. Notice the effect of the filtering factor $\{1 - (1-\kappa_i^2)^k\}$ with increasing k.

The iteration scheme (4.6) has been generalized in various ways. Important studies have been published by STRAND [25], and, very recently, by GRAVES and PRENTER [7].

Strand considers the scheme

$$f^{(k+1)} = f^{(k)} + DK^*(g - Kf^{(k)}), \qquad k = 0,1,\ldots, \tag{4.8}$$

where D is some linear operator. It will appear that D, when properly chosen, will act as a filter on the terms $g_i v_i/\kappa_i$ in the expansion of the solution f of (1.1).

From (4.8) we see that D acts on elements of the form $K^*u$, for $u \in L^2[0,1]$. Now one can prove that $R(K^*) \subset \text{span}(v_1, v_2, \ldots)$, so that D can be defined entirely by specifying for each singular function $v_i$ the image under D. Strand chooses

$$Dv_i = p_i v_i, \qquad i = 1,2,\ldots, \tag{4.9}$$

assuming that $0 < p_i \kappa_i^2 < 2$, for all i, and that the sequence $\{p_i\}_{i=1}^{\infty}$ is bounded. Now with the choice (4.9) one easily proves that

$$f^{(k)}(y) = \sum_{i=1}^{\infty} \{1 - (1 - p_i \kappa_i^2)^k\} \frac{g_i}{\kappa_i} v_i(y). \tag{4.10}$$

With this definition of D, Strand proves that $f^{(k)}$ tends to the least squares solution of minimum norm for (1.1), provided that $P_U g \in R(K)$, and $f^{(0)} \in \text{span}(v_1, v_2, \ldots)$. Here, by $P_U g$ we mean the projection of g on U. He also proves that $\|f^{(k)}\| \to \infty$ if $P_U g \notin R(K)$. For this reason, numerical difficulties may arise when g is contaminated with error.

In order to determine a practical choice of D, Strand reasons as follows. Defining the Heaviside function

$$H_a(x) = \begin{cases} 1, & x \geq a \\ 0, & x < a, \end{cases}$$

we have

$$\sum_{i=1}^{\infty} \frac{g_i}{\kappa_i} H_a(\kappa_i^2) v_i = \sum_{\kappa_i^2 > a} \frac{g_i}{\kappa_i} v_i, \tag{4.11}$$

which is simply a truncated series expansion of the solution of (1.1). He now requires that the factor $\{1 - (1 - p_i\kappa_i^2)^k\}$ in (4.10) approximates the factor $H_a(\kappa_i^2)$ in (4.11). The simplest way (recalling (4.9)) is to choose $D = F(K^*K)$, where $F(\lambda)$ is a polynomial, and $0 < \lambda F(\lambda) < 2$, for $0 < \lambda \leq 1$. Then, since $\{v_i\}_{i=1}^{\infty}$ is an eigensystem of the operator $K^*K$ with eigenvalues $\{\kappa_i^2\}_{i=1}^{\infty}$, this implies for (4.9) that

$$p_i = F(\kappa_i^2). \tag{4.12}$$

The curve defined by

$$R_k(\lambda) = 1 - (1 - \lambda F(\lambda))^k, \qquad 0 < \lambda \leq 1,$$

is called the response curve after k iterations associated with $D = F(K^*K)$.

From STRAND [25] we reproduce here the response curve of the original Landweber iteration scheme, for which $F(\lambda) \equiv 1$, so that $R_k(\lambda) = 1 - (1-\lambda)^k$, and a response curve determined by minimizing the quantity

$$\int_0^1 (1 - P(\lambda))^2 d\lambda,$$

where $P(\lambda) = \lambda F(\lambda) = a_1\lambda + \ldots + a_7\lambda^7$ (fig. 1 resp. fig. 3).

Strand also analyses the filter effect of iterating the regularization method. If in (4.1) we suppose that $Lf = f - \hat{f}$, where $\hat{f}$ is some a priori estimate of f (possibly 0 if nothing better is known), then the resulting linear equations read

$$(K^*K + \alpha I)f = K^*g + \alpha\hat{f}.$$

This suggests the iteration process

$$(K^*K + \alpha I) f^{(k+1)} = K^*g + \alpha f^{(k)}, \qquad k = 0,1,\ldots$$

and one easily deduces that

$$f^{(k)}(y) = \sum_{i=1}^{\infty} \{1 - (\frac{\alpha}{\kappa_i^2+\alpha})^k\} \frac{g_i}{\kappa_i} v_i(y).$$

One immediately sees that this filter is obtained by choosing $D = (K^*K + \alpha I)^{-1}$ in (4.8). For this D it follows that $F(\lambda) = (\alpha+\lambda)^{-1}$ and the resulting response curve is $R_k(\lambda) = 1 - \{\alpha(\alpha+\lambda)^{-1}\}^k$, which is also reproduced here from STRAND [25] (fig. 2).

Response curves $R_k(\lambda)$

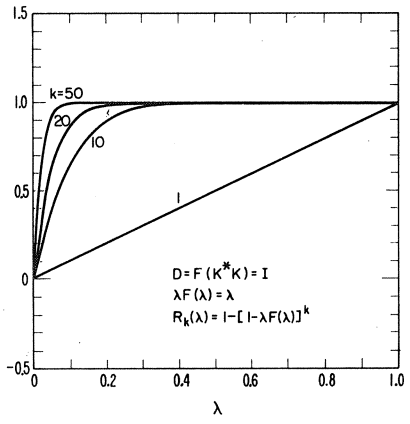



Fig. 1. The original Landweber iteration

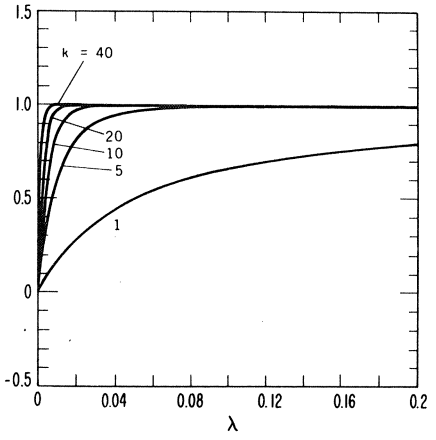
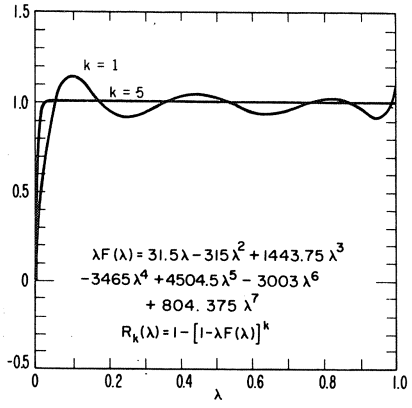



Fig. 2. The iterated Tihonov regularization process, $\alpha = 0.05$



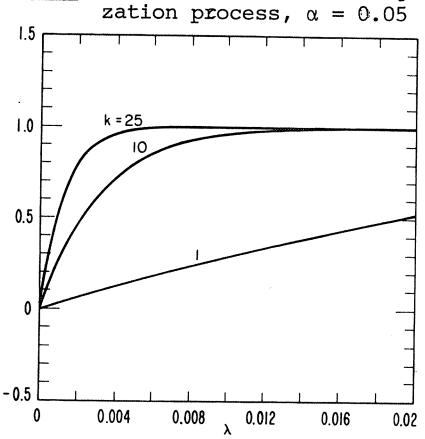




Fig. 3. A seventh degree polynomial of Strand

## 4.4. Statistical and probabilistic methods

These methods may be regarded as an extension of the regularization methods, and the actual computational processes are somewhat similar. Details and references may be found in papers of TURCHIN, KOZLOV and MALKEVICH [34], STRAND and WESTWATER [26] and WAHBA [37]. Some a priori information about the unknown function f is needed, together with statistical information (mean, covariance) of the data function g. ANDERSSEN and BLOOMFIELD [1, 2] have applied two statistical procedures (regression analysis and time series analysis) to the problem of numerical differentiation of noisy data. Anderssen and Bloomfield show that a regularization procedure of CULLUM [5] for numerical differentiation of noisy data is equivalent to a certain spectrum analysis procedure. This enables them to derive a fast Fourier transform implementation for regularization, where an acceptable value of the regularization parameter is estimated directly from a time series formulation. This procedure involves $O(n \log n)$ operations (where n is the number of data points) instead of $O(n^3)$ operations for the original regularization procedure.

## 4.5. A hybrid method of Hanson and Phillips

HANSON and PHILLIPS [9] have given a method based on a combination of the concepts of regularization, expansion and iteration. We shall describe it in some detail. The solution of (1.1) is found in the space of continuous piecewise linear polynomials based on a given partition $y_1 < y_2 < \ldots < y_n$ of the interval [0,1]. Representing f as an (unknown) linear combination of basis functions in this space, substituting this into the original equation, and writing down the resulting equation for the points $x = x_1,\ldots,x_m \in [0,1]$ yields an m×n-system of linear algebraic equations. The moment integrals in the matrix of this system are evaluated using two-point Gauss quadrature. The linear system is next stabilized, by requiring that its solution minimizes a functional which measures both the residual and the size of the solution. Special attention is paid to the choice of the points $y_1,\ldots,y_n$ and $x_1,\ldots,x_m$. After the computation of an approximate solution $\tilde{f}$ based on $n_1$ equally spaces knots $y_1,\ldots,y_{n_1}$, the total variation V in $\hat{f}$ is determined and a new set of $n_2 > 1.5\, n_1$ knots is chosen so that the variation in $\hat{f}$ is constant $(=V/(n_2-1))$ in each subinterval. Finally, a new knot is added at the midpoint of any subinterval whose length exceeds twice that of

the adjacent subintervals. This process of knot refinement can be repeated as often as one wishes. The points $x_1,\ldots,x_m$ are chosen according to the same strategy. Hanson and Phillips have also incorporated the possibility of constraining the solution. This is very important in the light of the requirement, when solving ill-posed problems, to include as much information about the unknown function as possible. They allow for the following constraints: (i) $f(y) \geq 0$; (ii) $f(y)$ is non-increasing (non-decreasing); (iii) $f(y)$ is convex upward (or concave downward); (iv) $f(y) \leq \beta$ = constant; (v) $\int_0^1 f(y)dy = 1$. The choice of f in the space of continuous piecewise linear polynomials facilitates the implementation of these constraints to a great extent. The numerical results reported are impressive, although the amount of computational work must have been considerable.

HANSON & PHILLIPS [10] have also given a generalization of their methods to two-dimensional equations.

REFERENCES

[1] ANDERSSEN, R.S. & P. BLOOMFIELD, *Numerical differentiation procedures for non-exact data,* Numer. Math., 22 (1974), pp. 157-182.

[2] ANDERSSEN, R.S. & P. BLOOMFIELD, *A time series approach to numerical differentiation,* Technometrics, 16 (1974), pp. 69-75.

[3] BAKER, C.T.H., L. FOX, D.F. MAYERS & K. WRIGHT, *Numerical solution of Fredholm integral equations of first kind,* The Computer J., 7 (1964), pp. 141-148.

[4] CRONE, L., *The singular value decomposition of matrices and cheap numerical filtering of systems of linear equations,* J. Franklin Inst., 294 (1972), pp. 133-136.

[5] CULLUM, J., *Numerical differentiation and regularization,* SIAM J. Numer. Anal., 8 (1971), pp. 254-265.

[6] DELVES, L.M. & J. WALSH (eds.), *Numerical solution of integral equations,* Clarendon Press, 1974, Oxford.

[7] GRAVES, J. & P.M. PRENTER, *Numerical iterative filters applied to first kind Fredholm integral equations,* Numer. Math., 30 (1978), pp. 281-299.

[8] HANSON, R.J., *A numerical method for solving Fredholm integral equations of the first kind using singular values,* SIAM J. Numer. Anal., 8 (1971), pp. 616-622.

[9] HANSON, R.J. & J.L. PHILLIPS, *An adaptive numerical method for solving linear Fredholm equations of the first kind,* Numer. Math., 24 (1975), pp. 291-307.

[10] HANSON, R.J. & J.L. PHILLIPS, *Numerical solution of two-dimensional integral equations using linear elements,* SIAM J. Numer. Anal., 15 (1978), pp. 113-121.

[11] HILGERS, J.W., *Non-iterative methods for solving operator equations of the first kind,* MRC Tech. Summ. Rept. #1413, Univ. of Wisconsin, Madison, Wisconsin, 1974.

[12] HSIAO, G. & R.C. MACCAMY, *Solution of boundary value problems by integral equations of the first kind,* SIAM Review, 15 (1973), pp. 687-705.

[13] HSIAO, G.C. & W.L. WENDLAND, *A finite element method for some integral equations of the first kind,* J. Math. Anal. Appl., 58 (1977), pp. 449-481.

[14] JASWON, M.A. & G.T. SYMM, *Integral equation methods in potential theory and elastostatics,* Acad. Press, 1977.

[15] LANDWEBER, L., *An iteration formula for Fredholm integral equations of the first kind,* Amer. J. Math., 73 (1951), pp. 615-624.

[16] MILLER, G.F., *Fredholm equations of the first kind,* Chapter 13 in [6], pp. 175-188.

[17] NATANSON, I.P., *Theorie der Funktionen einer reellen Veränderlichen,* Akademie-Verlag, Berlin, 1956.

[18] NEDELKOV, I.P., *Improper problems in computational physics,* Comput. Phys. Comm., 4 (1972), pp. 157-164.

[19] NOBLE, B., *Some applications of the numerical solution of integral equations to boundary value problems,* pp. 137-154 in: John Ll. Morris (ed.), *Conference on applications of numerical analysis,* (Dundee, Scotland, March 23-26, 1971), Springer, 1971.

[20] PHILLIPS, D.L., *A technique for the numerical solution of certain integral equations of the first kind,* J. ACM, 9 (1962), pp. 84-96.

[21] PICARD, E., *Sur un théorème générale relatif aux équations intégrales de première espèce et sur quelques problèmes de physique mathématique,* R.C. Mat. Palermo, 29 (1910), pp. 615-619.

[22] RIBIÈRE, G., *Régularisation d'opérateurs,* Rev. Franç Inf. Rech. Opér., 1 (1967), pp. 57-79.

[23] RICHTER, G.R., *Numerical solution of integral equations of the first kind with nonsmooth kernels,* SIAM J. Numer. Anal., 15 (1978), pp. 511-522.

[24] SMITHIES, F., *Integral equations,* Cambridge, 1958.

[25] STRAND, O.N., *Theory and methods related to the singular function expansion and Landweber's iteration for integral equations of the first kind,* SIAM J. Numer. Anal., 11 (1974), pp. 798-825.

[26] STRAND, O.N. & R. WESTWATER, *Statistical estimation of the numerical solution of a Fredholm integral equation of the first kind,* J. ACM, 15 (1968), pp. 100-114.

[27] TIHONOV, A.N., *Solution of incorrectly formulated problems and the regularization method,* Soviet Math. Dokl., 4 (1963), pp. 1035-1038.

[28] TIHONOV, A.N., *Regularization of incorrectly posed problems,* Soviet Math. Dokl., 4 (1963), pp. 1624-1627.

[29] TIHONOV, A.N., *Solution of nonlinear integral equations of the first kind,* Soviet Math. Dokl., 5 (1964), pp. 835-838.

[30] TIHONOV, A.N., *Nonlinear equations of the first kind,* Soviet math. Dokl., 6 (1965), 559-562.

[31] TIHONOV, A.N. & V.JA. ARSENIN, *Solutions of ill-posed problems,* Transl. from the Russian, Scripta series in mathematics, John Wiley and Sons, New York etc., 1977.

[32] TIHONOV, A.N. & V.B. GLASKO, *The approximate solution of Fredholm integral equations of the first kind,* USSR Comp. Math. and Math. Phys., 4, No. 3 (1964), pp. 236-247.

[33] TIHONOV, A.N. & V.B. GLASKO, *Use of the regularization method in non-linear problems,* 5, No. 3 (1965), pp. 93-107.

[34] TURCHIN, V.F., V.P. KOZLOV & M.S. MALKEVICH, *Use of the methods of mathematical statistics to solve ill-posed problems,* Soviet Phys. Usp., 13 (1971), pp. 681-702.

[35] TWOMEY, S., *On the numerical solution of Fredholm integral equations of the first kind by the inversion of the linear system produced by quadrature,* J. ACM, 10 (1963), pp. 97-101.

[36] VARAH, J.M., *On the numerical solution of ill-conditioned linear systems with applications to ill-posed problems,* SIAM J. Numer. Anal., 10 (1973), pp. 257-267.

[37] WAHBA, G., *Practical approximate solutions to linear operator-equations when the data are noisy,* SIAM J. Numer. Anal., 14 (1977), pp. 651-667.

# REGULARIZATION METHODS AND PARAMETER ESTIMATION METHODS FOR THE SOLUTION OF FREDHOLM INTEGRAL EQUATIONS OF THE FIRST KIND [+)]

J.J.M. CUPPEN

[+)] This paper will be submitted for publication elsewhere.

## 1. INTRODUCTION

In this paper we shall consider the numerical solution of linear Fredholm integral equations of the first kind on the interval [0,1]:

$$(1.1) \qquad \int_0^1 k(s,t)f(t)dt = g(s), \qquad s \in [0,1]$$

with a given, smooth kernel k and a given right hand side g or a perturbed right hand side $\tilde{g}$. We shall consider regularised solutions to this problem obtained by three methods:

1) Tychonov's method
2) Filtered least squares minimal C-norm approximation
3) A method based on the regularising effect of coarse discretisation in a form connected with method 2.

Regularisation methods usually depend on a parameter governing the level of regularisation. A number of methods to estimate the optimal choice of this regularisation parameter are discussed. A discussion of numerical experiments on a number of the test problems used in the literature is given. We shall assume the reader's familiarity with the ill-posedness of the problem, with the concept of regularisation and with Tychonov's method of regularisation [17,15,13].

## 2. DISCRETISATION OF THE INTEGRAL EQUATION, TYCHONOV'S METHOD

In order to be able to speak the language of linear algebra rather than the language of functional analysis, we will first discretise the integral equation. It should be emphasised, though, that neither the methods of regularisation nor the methods of parameter estimation are restricted in their application to the particular way of discretising the equation as is given below (viz. collocation + cubic splines).

Consider the integral equation (1.1) in least squares form, where f is to minimise the $L_2$-norm of the residue:

$$(2.1) \qquad \left\| \int_0^1 k(s,t)f(t)dt - g(s) \right\|_2 = \left( \int_0^1 \left( \int_0^1 k(s,t)f(t)dt - g(s) \right)^2 ds \right)^{\frac{1}{2}}.$$

To the outer integral we apply a quadrature rule with abscissae $s_i$ and

weights $w_i^2$ $(i = 0,1,\ldots,m)$ and get

$$(2.2) \qquad \sum_{i=0}^{m} \left(w_i \int_0^1 k(s_i,t)f(t)dt - w_i g(s_i)\right)^2 \quad \text{minimal.}$$

The discretisation of the second integral could be performed in the same way, but I prefer to approximate the solution f rather than the integral. Then the calculations can be performed in terms of an approximating function and also the result will be an approximating function rather than a set of approximating function values on some discretisation grid. Moreover, we can introduce here in a natural way some smoothness in the approximating function by choosing a smooth approximation space. We will use cubic splines on an equidistant grid $t_j = j/n$, $j = 0,1,\ldots,n$. For convenience we add the points $t_{-1} = -1/n$ and $t_{n+1} = 1+1/n$ and put

$$(2.3a) \qquad f(t) = \sum_{j=-1}^{n+1} f_j B_j(t) = \sum_{j=-1}^{n+1} f_j B(n(t-t_j)), \qquad t \in [0,1],$$

where

$$(2.3b) \qquad B(t) = \begin{cases} 0 & t \geq 2, \\ (2-t)^3/6 & t \in [1,2], \\ 2/3 - t^2 + t^3/2 & t \in [0,1], \\ B(-t) & t \leq 0. \end{cases}$$

With (2.3) the solution of (2.2) becomes the least squares solution of

$$(2.4) \qquad Kf = g,$$

$$K_{ij} = w_i \int_0^1 k(s_i,t)B_j(t)dt, \qquad \begin{array}{l} i = 0,1,\ldots,m; \\ j = -1,0,\ldots,n+1. \end{array}$$

$$g_i = w_i g(s_i),$$

In our notation we will neither distinguish between a cubic spline f(t) and its representation vector $(f_j)_{j=-1,\ldots,n+1}$, nor between the right hand side g(x) and its discretisation $(g_i)_{i=0,\ldots,m}$.

The norms taken of "spline-vectors" $(f_j)$ shall be Z-norms:

$$\|f\|_Z = \|Zf\|_2$$

where Z is a matrix such that

$$\|(f_j)\|_Z^2 = (f_j)^T Z^T Z (f_j) = \int_0^1 (f(t))^2 dt = \|f(t)\|_2^2,$$

so

$$(Z^T Z)_{i,j} = \int_0^1 B_i(t) B_j(t) dt.$$

The matrix Z is well-conditioned and can easily be computed as an upper-triangular bandmatrix with 3 codiagonals.

Because of the ill-posedness of the integral equation (1.1) the system (2.4) is extremely ill-conditioned. In the regularisation method of Tychonov, problem (2.4) is replaced by:

(2.5) $$\text{minimise } \|K(f)-g\|_2^2 + \alpha \|Cf\|_2^2$$

for some $\alpha > 0$. If the solution of the integral equation is known to be smooth, this method can effectively overcome the disastrous amplification of noise, discretisation and round-off errors occurring in the solution of (2.4). The method stands or falls, however, with the choice of $\alpha$, which will be considered below.

In (2.5) we assume C to be a matrix such that $\|Cf\|_2 = \|f''\|_2$, the $L_2$-norm of the second derivative of the cubic spline, following Ribière's advice [13] to regularise with the second derivative (other choices are possible). C is positive semi-definite with a nullspace of dimension 2, and can be expressed as an uppertriangular bandmatrix with 3 codiagonals. Since k is a smooth kernel, we can assume K to be nonsingular on the null-space of C which corresponds with the constant and linear functions.

We calculate C as the Choleski decomposition of the matrix $C^T C$ which is given by

$$(C^T C)_{i,j} = \int_0^1 B_i''(t) B_j''(t) dt.$$

This integral can be evaluated analytically and depends only on the difference between i and j if i nor j is close to its boundary values. The integral is zero for $|i-j| > 3$. If one does not use cubic splines but another expansion

$$f(t) = \sum_j f_j \phi_j(t),$$

for example Chebyshev polynomials, we have

$$(Z^TZ)_{ij} = \int_0^1 \phi_i(t)\phi_j(t)dt,$$

$$(C^TC)_{ij} = \int_0^1 \phi_i''(t)\phi_j''(t)dt.$$

In general, $Z^TZ$ and $C^TC$ will be full matrices and consequently Z and C will be full triangular matrices. However, if the $\phi_i$ are orthonormal then $Z^TZ$ and Z reduce to the identity matrix.

The solution to (2.5) with g replaced by a perturbation $\tilde{g}$ is given by:

$$\tilde{f}_\alpha = (K^TK + \alpha C^TC)^{-1}K^T\tilde{g} \qquad (\alpha > 0). \tag{2.6}$$

Now suppose there exists a cubic-spline solution f to (2.4) satisfying $Kf = g$. We would like to choose $\alpha$ in (2.5) and (2.6) such that $\tilde{f}_\alpha$ approximates f as close as possible. Using (2.6) we have:

$$\begin{aligned}\|f-\tilde{f}_\alpha\|_Z &= \|Z(K^TK+\alpha C^TC)^{-1}K^T(Kf-\tilde{g}) + \alpha Z(K^TK+\alpha C^TC)^{-1}C^TCf\|_2 \\ &\le \|Z(K^TK+\alpha C^TC)^{-1}K^T\|_E\|Kf-\tilde{g}\|_2 + \alpha\|Z(K^TK+\alpha C^TC)^{-1}C^T\|_E\|Cf\|_2.\end{aligned} \tag{2.7}$$

So if we have constants $\delta$ and $\tau$ with

$$\begin{aligned}&\|Kf-\tilde{g}\| = \|g-\tilde{g}\| \le \delta \\ &\|Cf\| \le \tau,\end{aligned} \tag{2.8}$$

then

$$\|f-\tilde{f}_\alpha\|_Z \le \delta\,\|Z(K^TK+\alpha C^TC)^{-1}K^T\|_E + \alpha\tau\|Z(K^TK+\alpha C^TC)^{-1}C^T\|_E. \tag{2.9}$$

Since the right hand side of (2.9) can be computed for arbitrary $\alpha > 0$ we could use (2.9) for the selection of a semi-optimal $\alpha$ which minimizes the right-hand side of (2.9) instead of its left-hand side. This method was advocated by KÖCKLER [6], but a corresponding rule of thumb was already given by RIBIÈRE [13]. Ribière advised to choose $\alpha$ in such a way that the error in $\tilde{f}_\alpha$ caused by the perturbation of the equation by noise etc. (the first term of the right-hand side of (2.9)) would be of the same order of magnitude as the error in $\tilde{f}_\alpha$ caused by the regularisation itself (the second term). Köckler showed how to calculate (2.9) efficiently for a large number of $\alpha$'s and we

shall consider a modified version of his algorithm below. In Köcklers experiments [6] the estimate turned out to be very pessimistic, *but of the right shape*. Moreover, both the estimate and the actual error behaved like a convex function with a very flat minimum, admitting a straightforward and effective application of the idea described above. Note, however, that the resulting $\alpha$ does not depend on $\delta$ and $\tau$ themselves, but on their ratio $\delta/\tau$. If, in practice, this number is not approximately known, an inspection of the realised residue $\tilde{\delta} = \|K\tilde{f}_\alpha - \tilde{g}\|_2$ and smoothness $\tilde{\tau} = \|C\tilde{f}_\alpha\|_2$ might help.

## 3. REDUCTION OF $(K^TK+\alpha C^TC)^{-1}$ TO A SIMPLE FORM

Recall that K is a $(m+1)\times(n+3)$ matrix and C a $(n+3)\times(n+3)$ matrix. Let r be rank(C), $r \le n+3$, and suppose that K is non-singular on $N(C)$. Let $(M_1|N)$ be a non-singular matrix with

$$(M_1|N)^TC^TC(M_1|N) = \left(\begin{array}{c|c} I_r & 0 \\ \hline 0 & 0 \end{array}\right); \tag{3.1}$$

the columns of N span $N(C)$ and $(M_1|N)$ is a generalised inverse of C. In the setting of section 2 it is easy to compute $M_1$ and N. Now consider

$$(M_1|N)^TK^TK(M_1|N) = \left(\begin{array}{c|c} M_1^TK^TKM_1 & M_1^TK^TKN \\ \hline N^TK^TKM_1 & N^TK^TKN \end{array}\right). \tag{3.2}$$

If we replace $M_1$ by

$$M = (I-N(KN)^+K)M_1, \tag{3.3}$$

then (3.1) holds for M too, since $CN = 0$, so $CM = CM_1$ and we have

$$(M|N)^TC^TC(M|N) = \left(\begin{array}{c|c} I_r & 0 \\ \hline 0 & 0 \end{array}\right)$$

$$(M|N)^TK^TK(M|N) = \left(\begin{array}{c|c} M^TK^TKM & M^TK^TKN \\ \hline N^TK^TKM & N^TK^TKN \end{array}\right) = \left(\begin{array}{c|c} M^TK^TKM & 0 \\ \hline 0 & N^TK^TKN \end{array}\right), \tag{3.4}$$

since

$$N^TK^TKM = N^TK^T(I-KN(KN)^+)KM_1 = 0.$$

Using the singular value decomposition

$$KN = E\Lambda_2F^T = \sum_{i=r+1}^{n+3} e_i\lambda_i f_i^T \tag{3.5}$$

of KN, KM in (3.3) can be easily calculated as

$$KM = (I-EE^T)KM_1. \tag{3.6}$$

With the singular value decomposition

$$KM = P\Lambda_1 Q^T = \sum_{i=1}^{r} p_i \lambda_i q_i^T \tag{3.7}$$

it follows

$$(K^TK+\alpha C^TC)^{-1} =$$

$$(M|N)\left(\left(\begin{array}{c|c} M^TK^TKM & 0 \\ \hline 0 & N^TK^TKN \end{array}\right) + \alpha\left(\begin{array}{c|c} I_r & 0 \\ \hline 0 & 0 \end{array}\right)\right)^{-1}(M|N)^T =$$

$$(M|N)\left(\begin{array}{c} Q \\ \hline F \end{array}\right)\left(\begin{array}{c|c} \Lambda_1^2+\alpha I & \\ \hline & \Lambda_2^2 \end{array}\right)^{-1}\left(\begin{array}{c|c} Q^T & \\ \hline & F^T \end{array}\right)(M|N)^T. \tag{3.8}$$

With (3.8) the estimate (2.9) transforms to

$$\begin{aligned} \|f-\tilde{f}_\alpha\|_Z &\le \delta\|Z(MQ|NF)\left(\begin{array}{c|c} (\Lambda_1^2+\alpha I)^{-1}\Lambda_1 & \\ \hline & \Lambda_2^{-1} \end{array}\right)\|_E + \\ &+ \alpha\tau\|ZMQ(\Lambda_1^2+\alpha I)^{-1}Q^TM^TC^T\|_E \\ &= \delta(\|ZMQ(\Lambda_1^2+\alpha I)^{-1}\Lambda_1\|_E^2 + \|ZNF\Lambda_2^{-1}\|_E^2)^{\frac{1}{2}} + \\ &+ \alpha\tau\|ZMQ(\Lambda_1^2+\alpha I)^{-1}\|_E \\ &= \delta\left(\sum_{i=1}^{r}\Delta_i\left(\frac{\lambda_i}{\lambda_i^2+\alpha}\right)^2+s\right)^{\frac{1}{2}} + \tau\left(\sum_{i=1}^{r}\Delta_i\left(\frac{1}{1+\frac{\lambda_i^2}{\alpha}}\right)^2\right)^{\frac{1}{2}}, \end{aligned} \tag{3.9}$$

where $\Delta_i$ is the $i^{th}$ diagonal element of $Q^TM^TZ^TZMQ$ and $s = \|ZNF\Lambda_2^{-1}\|$. This clearly shows that (2.9) can be computed with $O(n)$ operations per $\alpha$, provided the preliminary work of forming M, N, the SVD's (3.5) and (3.7), s and the $\Delta_i$ is done once and for all in advance (Compare [6], [7], [9] and [19]).

## 4. FILTERED LEAST SQUARES MINIMAL C-NORM APPROXIMATION

If L is the generalized inverse to C as formed in section 3, L = (M|N), nonsingular, then

(4.1) $\quad f = L(KL)^{+}g$

is the least squares solution to (2.4) with minimal C-norm $\|f\|_C = \|Cf\|_2$. This follows from (2.5), (2.6) and (3.8) by taking the limit $\alpha \to 0$. With the singular value decomposition of KL:

(4.2) $\quad KL = U\Sigma V^T = \sum_i u_i \sigma_i v_i^T, \qquad \sigma_1 \geq \sigma_2 \geq \ldots \geq 0,$

(4.1) transforms to

(4.3) $\quad f = LV\Sigma^{+}U^Tg = \sum_i Lv_i \sigma_i^{+} u_i^T g.$

In these formulas the $u_i$ and $v_i$ denote the columns of U and V, the singular vectors of KL, $\Sigma$ is a diagonal matrix containing the singular value $\sigma_i$, and $\Sigma^{+} = \mathrm{diag}(\sigma_1^{+},\sigma_2^{+},\ldots)$ where $\sigma_i^{+} = 1/\sigma_i$ if $\sigma_i > 0$ and $\sigma_i^{+} = 0$ if $\sigma_i = 0$. The ill-posedness of problem (1.1) is reflected in the fact that the $\sigma_i$ become very small with increasing i. A filtered least squares minimal C-norm approximation is given by

(4.4) $\quad f_q = LV\Sigma^{q+}U^Tg = \Sigma_{i=1}^{q} Lv_i \sigma_i^{+} u_i^T g$

where $\Sigma^{q+}$ is defined by $\Sigma^{q+} = \mathrm{diag}(\sigma_1^{+},\sigma_2^{+},\ldots,\sigma_q^{+},0,\ldots,0)$. In our case C has a nullspace of dimension 2 and the corresponding singular functions do not contribute to $\|Cf\|$. Therefore, in practice we modify the method slightly and always include the corresponding components in our approximation.

Filtered least squares solutions were considered by, among others, VARAH [18] and HANSON [4]. The regularising effect of the method is based on the observation that for smooth solutions f the solution and the right hand side are well-represented by the smooth singular functions (vectors) belonging to the larger singular values, whereas the smaller singular values are responsible for the magnification of errors in g in (4.1) or (2.4). (Compare also [1]).

An error estimate for the obtained approximation, suitable for the selection of a semi-optimal q, will illustrate the phenomena described above. Let $\tilde{g}$ be a perturbation of the right hand side g and let f be given by (4.3). Suppose we have $\delta$ and $\tau$ with

$$\|Kf-\tilde{g}\|_2 \leq \delta$$

(4.5)

$$\|Cf\|_2 \leq \tau,$$

then for $\tilde{f}_q = LV\Sigma^{q+}U^T\tilde{g}$ it follows

$$\begin{aligned}\|f-\tilde{f}_q\|_Z &= \|ZLV\Sigma^{+}U^{T}Kf - ZLV\Sigma^{q+}U^T\tilde{g}\|_2 \\ &\leq \delta\|ZLV\Sigma^{q+}U^T\|_E + \|ZLV(\Sigma^{+}-\Sigma^{q+})U^{T}KL\|_E\|Cf\|_2 \\ &\leq \delta\|ZLV\Sigma^{q+}\|_E + \tau\|ZLV\left(\begin{array}{c|c}\overset{q\times q}{0} & 0\\ \hline 0 & I\end{array}\right)\|_E \\ &= \delta\Big(\sum_{i=1}^{q}\Big(\frac{\Delta_i}{\sigma_i^2}\Big)\Big)^{\frac{1}{2}} + \tau\Big(\sum_{i=q+1}^{n+3}\Delta_i\Big)^{\frac{1}{2}}, \end{aligned} \tag{4.6}$$

where $\Delta_i$ is the $i^{th}$ diagonal element of $V^TL^TZ^TZLV$.

Remarks in connection with (4.6).

1. The error in the computed solution $\tilde{f}_q$ as dependent on q, behaves in much the same way as that in the Tychonov approximation $\tilde{f}_\alpha$ as dependent on $\alpha$. For large q (small $\alpha$) the magnified perturbation, discretisation and round-off errors dominate, whereas for small q (large $\alpha$) the error introduced by the regularisation dominates.
2. In case L is the discretisation of the inverse of a differential operator, L has, like K, a smoothing character. It is to be expected that the first columns of V are the smooth ones and the latter columns become increasingly oscillatory. Therefore the $\Delta_i$, being nothing else than $\|Lv_i\|_Z^2$, will decrease with increasing i, to the benefit of the effectiveness of both the regularisation method itself and the parameter estimation with (4.6). The smoother L is, i.e., the higher the order of C as a differential operator is, the more pronounced this effect will be. A well-chosen C of high order (with $\|Cf\|$ small) modifies the singular functions of KL in such a way that f and g fit better in the first few of them than in the first few singular functions of the plain K.
3. The error estimate (4.6) is present in KÖCKLER [6]. Also VARAH [18] uses an error bound of the form of (4.6) to select a semi-optimal q. In his context L = I and he derives:

$$\|f-\tilde{f}_q\| \le \delta \|V\Sigma^{q+}\| + \|V(\Sigma^{+}-\Sigma^{q+})U^T g\|$$

$$\le \delta/\sigma_q + \left(\sum_{i\ge q+1} \left(\frac{\beta_i}{\sigma_i}\right)^2\right)^{\frac{1}{2}},$$

where $\beta_i = u_i^T g$. The exact $\beta_i$ decrease faster than the $\sigma_i$ in case of a smooth true solution. They are, however, not available during computation and the approximations $\tilde{\beta}_i = u_i^T \tilde{g}$ will not have this property in the presence of non-smooth noise. HANSON [4] mentioned the possible usefulness of the coordinate transformation introduced by C and L. LEE and PRENTER [8] give theoretical backgrounds and convergence results for filtered least squares minimal norm approximations.

## 5. REGULARISATION THROUGH (COARSE) DISCRETISATION

NATTERER [10,11] investigates the regularising effect of a relatively coarse discretisation of the integral equation (1.1). The idea is the following. Let us think of the discretisation as governed by the dimension of the cubic spline space, in which we choose our approximate solution fitting to the equation in some sense. For not too large dimensions this space represents smooth functions very well, but wildly oscillating functions not at all. It is clear that such a cubic spline space does represent our smooth solution very well, but not the troublemakers, the wild singular functions belonging to the small singular values of the kernel. Therefore, solving with a low-dimensional solution space means regularising the ill-posed problem. The attractiveness of this method is, of course, that one can work with low order systems, thereby reducing the work to be done enormously. In our approach the set up of the discretisation (2.4) takes, for example, the computation of order m.n element integrals. For a further discussion the reader is referred to [2], [10] and [11].

Now, when we take a closer look at the filter method of section 4 we see that this method does nothing else than restrict the solution space to the subspace spanned by the first q singular functions of the discretized kernel. It is clear that in general this restriction to dimension q of a cubic spline space of dimension n, say, represents the first q singular functions of the continuous kernel better than a cubic spline space of dimension q, and a similar effect is to be expected in connection with the absentness of the oscillatory singular functions of higher order. Therefore, we combine the advantages of regularisation through discretisation, being

- little work because of the low order matrices

with those of the filtration method, being

- easy variation of the amount of smoothing,
- availability of the estimate of the optimal amount of regularization using (4.6),
- effectiveness,

and we get the following practical method:

I Set up the equations with an n which is a little larger than the expected dimension of the cubic spline space necessary to adequately represent those smooth singular functions of our kernel which compose the desired solution;

II Apply the method of section 4 to this problem to arrive at an estimated optimal dimension for the restricted solution space;

III Repeat with a larger n is the outcome of II (too close to n) shows that the necessary singular functions might not all be adequately represented in the discretisation.

## 6. NUMERICAL EXPERIMENTS

The methods of sections 2, 4 and 5 were combined in one program, and experiments were carried out with most examples to be found in the literature which are of the form

$$\int_{t_1}^{t_2} k(s,t)f(t)dt = g(s), \qquad s \in [s_1,s_2].$$

These were transformed to $[t_1,t_2] = [s_1,s_2] = [0,1]$. A number of these examples will be discussed below. To each of them tables are given containing, besides the results, the transformed kernel, the exact solution and the right hand side. The parameters which govern the experiments are:

a) the discretization parameters m and n (see section 2),

b) the perturbation parameters a, $\omega_1$ and $\omega_2$; in all experiments but one the right hand side g(s) is replaced by $\tilde{g}(s) = g(s).(1+a(\sin(\omega_1 s)+\cos(\omega_2 s)))$, an arbitrary enough, reproducable perturbation. Unless mentioned otherwise we have $\omega_1 = \omega_2 = 200$.

Using a cubic spline interpolation (s"(0) = s"(1) = 0) of the reference solution f, and the discretized right hand side $\tilde{g}$, the smoothness $\tau = \|f"\|$ and the residue $\delta = \|Kf-\tilde{g}\|$ were calculated. Note that $\delta$ represents both

perturbation and discretisation errors. Regularized solutions $\tilde{f}_r$ ($r = \alpha,q$) were calculated for a number of given values of $\alpha$ and for a number of values of q ranging from 2 to q max (given). The error bounds (2.9) and (4.6), divided by $\|f\|$, residues $\tilde{\delta}_r = \|K\tilde{f}_r - \tilde{g}\|$, smoothnesses $\tilde{\tau}_r = \|C\tilde{f}_r\|$ and relative errors $\|\tilde{f}_r - f\|/\|f\|$ were calculated and printed. For selected values of $\alpha$ and q these results are given in tables below.

Underlined in these tables are:

1) the (or a) value of the regularisation parameter r which is semi-optimal with respect to the estimate with $\tau$ and $\delta$ replaced by $\tilde{\tau}_r$ and $\tilde{\delta}_r$ (the "modified estimate"),
2) the minimal value of the error bound (2.9) or (4.6),
3) the maximal $\tilde{\delta}_r$ not larger than $\delta$ (regularising as much as possible under this restriction).
4) the maximal $\tilde{\tau}_r$ not larger than $\tau$ (regularising as little as possible under this restriction),
5) the minimal relative error $\|f - \tilde{f}_r\|/\|f\|$.

Actually, 3) and 4) provide two more well-known estimation methods for the optimal regularisation level. We will refer to them as "selection on $\delta$" and "selection on $\tau$" respectively. The calculations were performed on a CDC 73/173 system, all integrals were calculated using 16-point Gaussian quadrature.

In our first example (tables 1 and 2) the kernel is not very smooth and has singular values $\sigma_i = i^{-2}\pi^{-2}$ which do approach zero for i to infinity, but not very fast. The example nevertheless illustrates the regularisation and estimation processes very well.

$k(s,t) = \begin{cases} (1-s)t & 0 \le t \le s \le 1 \\ (1-t)s & 0 \le s \le t \le 1, \end{cases}$ $f(t) = t - 2t^3 + t^4$, $g(s) = \frac{s}{30}(3 - 5s^2 + 3s^4 - s^5)$

Tychonov's method

$a = 5_{10}{}^{-3}$ m=n=16, $\delta = 1.04_{10}{}^{-4}$, $\tau = 2.19$, $\|f\| = 2.22_{10}{}^{-1}$

| $\alpha$ | Bound on Rel. Err. | $\tilde{\delta}_\alpha$ | $\tilde{\tau}_\alpha$ | Rel. Err. |
|---|---|---|---|---|
| 1 | $2.61_{10}{}^{+5}$ | $8.65_{10}{}^{-4}$ | $3.06_{10}{}^{-6}$ | $4.80_{10}{}^{-1}$ |
| $1_{10}{}^{-7}$ | $3.79_{10}{}^{-1}$ | $3.36_{10}{}^{-4}$ | 1.31 | $1.91_{10}{}^{-1}$ |
| $10_{10}{}^{-8}$ | $2.88_{10}{}^{-1}$ | $\underline{1.02_{10}{}^{-4}}$ | $\underline{2.00}$ | $5.32_{10}{}^{-2}$ |
| $\underline{1_{10}{}^{-9}}$ | $\underline{2.87_{10}{}^{-1}}$ | $7.74_{10}{}^{-5}$ | 2.31 | $\underline{4.52_{10}{}^{-2}}$ |
| $1_{10}{}^{-20}$ | $2.20_{10}{}^{+2}$ | $5.56_{10}{}^{-10}$ | $2.89_{10}{}^{+2}$ | $3.90_{10}{}^{-1}$ |

Table 1, example 1, TRICOMI, STRAND [16], GRAVES and PRENTER [3].

Let us first consider the results of Tychonov's method as given in table 1. The following observations can be made:

i) too much ($\alpha$=1) or too little regularisation gives inferior results;

ii) the estimate of the regularisation level based on (2.9) gives an optimal result and the results for selection on $\delta$ and $\tau$ are comparable.

$k(s,t) = \begin{cases} (1-s)t & 0 \le t \le s \le 1 \\ (1-t)s & 0 \le s \le t \le 1, \end{cases}$ $f(t) = t-2t^3+t^4$, $g(s) = \frac{s}{30}(3-5s^2+3s^4-s^5)$

Filter method

$a=5_{10}{}^{-3}$ m=n=4, $\delta=6.03_{10}{}^{-5}$, $\tau=2.19$, $\|f\|=2.22_{10}{}^{-1}$

| q | Bound on Rel. Err. | $\tilde{\delta}_q$ | $\tilde{\tau}_q$ | Rel. Err. |
|---|---|---|---|---|
| 6 | $1.19_{10}{}^{+9}$ | $3.07_{10}{}^{-2}$ | $2.55_{10}{}^{+13}$ | $2.81_{10}{}^{+11}$ |
| 5 | $6.42_{10}{}^{-1}$ | $8.80_{10}{}^{-14}$ | 2.19 | $3.04_{10}{}^{-2}$ |
| 4 | $6.36_{10}{}^{-1}$ | $1.32_{10}{}^{-5}$ | 2.15 | $4.17_{10}{}^{-2}$ |
| 3 | $6.37_{10}{}^{-1}$ | $2.33_{10}{}^{-5}$ | 2.12 | $3.35_{10}{}^{-2}$ |
| 2 | $9.68_{10}{}^{-1}$ | $8.87_{10}{}^{-4}$ | $3.68_{10}{}^{-7}$ | $4.80_{10}{}^{-1}$ |

Table 2, example 1, TRICOMI, STRAND [16], GRAVES and PRENTER [3].

The results for the filter method are given in table 2. We observe that

i) the optimal result (q=5) is, with this low order discretisation, even a little better than with Tychonov's regularisation method. In all examples the optimal relative errors obtained with the two methods were comparable;

ii) $\tau$-selection delivers the optimal q, as does the modified estimate;

iii) the results of estimate (4.6) and $\delta$-selection are comparable to the optimal result.

| $k(s,t) = e^{st}$, $f(t) = e^{t}$, $g(s) = (e^{s+1}-1)/(s+1)$ | | | | |
|---|---|---|---|---|
| Tychonov's method | | | | |
| $a=0$, | $m=n=16$, | $\delta=3.39_{10}-5$, | $\tau=1.74$, | $\|f\|=1.79$ |
| $\alpha$ | Bound on Rel. Err. | $\tilde{\delta}_\alpha$ | $\tilde{\tau}_\alpha$ | Rel. Err. |
| $1_{10}-9$ | $2.79_{10}-2$ | $8.61_{10}-6$ | 1.37 | $3.92_{10}-3$ |
| $1_{10}-14$ | $3.48_{10}-1$ | $4.01_{10}-9$ | 1.69 | $1.88_{10}-3$ |
| $a=10^{-4}$, | $m=n=16$, | $\delta=2.10_{10}-4$, | $\tau=1.74$, | $\|f\|=1.79$ |
| $\alpha$ | Bound on Rel. Err. | $\tilde{\delta}_\alpha$ | $\tilde{\tau}_\alpha$ | Rel. Err. |
| $1_{10}-5$ | $5.58_{10}-2$ | $3.06_{10}-4$ | $3.62_{10}-3$ | $3.54_{10}-2$ |
| $1_{10}-7$ | $5.25_{10}-2$ | $2.74_{10}-4$ | $2.90_{10}-1$ | $2.82_{10}-2$ |
| $1_{10}-9$ | $5.79_{10}-2$ | $2.05_{10}-4$ | 1.38 | $3.70_{10}-3$ |
| $1_{10}-10$ | $6.71_{10}-2$ | $2.05_{10}-4$ | 1.46 | $3.07_{10}-3$ |
| Filter method | | | | |
| $a=0$, | $m=n=8$, | $\delta=2.09_{10}-3$, | $\tau=1.61$, | $\|f\|=1.79$ |
| $q$ | Bound on Rel. Err. | $\tilde{\delta}_q$ | $\tilde{\tau}_q$ | Rel. Err. |
| 4 | $2.97_{10}+1$ | $6.95_{10}-9$ | 1.46 | $2.94_{10}-3$ |
| 3 | $5.96_{10}-1$ | $3.75_{10}-7$ | 1.42 | $3.82_{10}-3$ |
| 2 | $9.51_{10}-2$ | $2.25_{10}-4$ | $2.79_{10}-6$ | $3.55_{10}-3$ |
| $a=10^{-4}$, | $m=n=8$, | $\delta=2.13_{10}-3$, | $\tau=1.61$, | $\|f\|=1.79$ |
| $q$ | Bound on Rel. Err. | $\tilde{\delta}_q$ | $\tilde{\delta}_q$ | Rel. Err. |
| 4 | $3.02_{10}+1$ | $5.32_{10}-5$ | $1.11_{10}+2$ | $9.90_{10}-1$ |
| 3 | $6.07_{10}-1$ | $1.43_{10}-4$ | 1.34 | $4.11_{10}-3$ |
| 2 | $9.54_{10}-2$ | $2.56_{10}-4$ | $2.79_{10}-6$ | $3.54_{10}-2$ |

Table 3, example 2, BAKER et al. [1].

Our second example (table 3) is taken from BAKER et al. [1]. Here we have a smooth kernel ($e^{st}$) and a smooth exact solution ($e^{t}$). Therefore it is no surprise that both regularisation methods effectively regularise the ill-posed problem. Among the estimation methods $\tau$-selection is clearly

superior here, giving optimal results in all cases. The results of the other estimation methods are, nevertheless, still acceptable.

| $k(s,t) = e^{-8st}$, $f(t) = 2(2t+1)e^{-2t}$, $g(s) = ((4s+2)-(12s+4)e^{-8s+1})/(4s+1)^2$ | | | | |
|---|---|---|---|---|
| $a=0$, | $m=n=8$, | $\delta=3.14_{10}-4$, | $\tau=2.23$ | $\|f\|=1.51$ |
| Tychonov's method | | | | |
| $\alpha$ | Bound on Rel. Err. | $\tilde{\delta}_\alpha$ | $\tilde{\tau}_\alpha$ | Rel. Err. |
| $1_{10}-7$ | $4.26_{10}-2$ | $1.20_{10}-4$ | $7.02_{10}-1$ | $1.55_{10}-2$ |
| $1_{10}-14$ | 4.58 | $3.43_{10}-8$ | 2.09 | $1.09_{10}-3$ |
| Filter method | | | | |
| q | Bound on Rel. Err. | $\tilde{\delta}_q$ | $\tilde{\tau}_q$ | Rel. Err. |
| 6 | 5.06 | $7.72_{10}-9$ | 2.15 | $1.28_{10}-3$ |
| 3 | $4.34_{10}-2$ | $1.13_{10}-4$ | 1.72 | $1.72_{10}-2$ |
| $a=5_{10}-3$, | $m=n=16$, | $\delta=2.61_{10}-3$, | $\tau=2.47$, | $\|f\|=1.51$ |
| Tychonov's method | | | | |
| $\alpha$ | Bound on Rel. Err. | $\tilde{\delta}_\alpha$ | $\tilde{\tau}_\alpha$ | Rel. Err. |
| 1 | $8.41_{10}+2$ | $2.54_{10}-3$ | $1.76_{10}-5$ | $2.04_{10}-2$ |
| $1_{10}-4$ | $1.35_{10}-1$ | $2.54_{10}-3$ | $4.30_{10}-3$ | $2.03_{10}-2$ |
| $1_{10}-6$ | $9.03_{10}-2$ | $2.52_{10}-3$ | $1.79_{10}-1$ | $1.65_{10}-2$ |
| $1_{10}-7$ | $1.02_{10}-1$ | $2.49_{10}-3$ | $7.68_{10}-1$ | $1.31_{10}-2$ |
| $a=5_{10}-3$, | $m=n=8$, | $\delta=3.86_{10}-3$, | $\tau=2.23$, | $\|f\|=1.51$ |
| Filter method | | | | |
| q | Bound on Rel. Err. | $\tilde{\delta}_q$ | $\tilde{\tau}_q$ | Rel. Err. |
| 3 | $1.36_{10}-1$ | $3.12_{10}-3$ | 1.50 | $4.23_{10}-2$ |
| 2 | $1.03_{10}-1$ | $3.66_{10}-3$ | $2.73_{10}-6$ | $1.89_{10}-2$ |

Table 4, example 3, RIBIÈRE [13].

The third example comes from RIBIÈRE [13]. Again the kernel ($e^{-8st}$) and the solution ($2(2t+1)e^{-2t}$) are very smooth. In the unperturbed case we see

τ-selection to be optimal and the other estimates acceptable (acceptable because without regularisation the solution would be completely destroyed by oscillations). In the perturbed case the filter method provides one of the rare cases where the other estimates behave (a little) better than τ-selection.

$k(s,t) = e^{-10(s-t)^2}$, $f(t) = t(1-t)$, $g(s) = \int_0^1 k(s,t)f(t)dt$

Tyhonov's method, $\tilde{g}(s) = g(s)+a(\sin(\omega_1 s)+\cos(\omega_1 s))$ (exception)
$a=10^{-4}$, $\omega_1=10$, $\omega_2=500$, m=n=16, $\delta=9.39_{10}-5$, $\tau=1.96$, $\|f\|=1.83_{10}-1$

| α | Bound on Rel. Err. | $\tilde{\delta}_\alpha$ | $\tilde{\tau}_\alpha$ | Rel. Err. |
|---|---|---|---|---|
| $1_{10}-7$ | $1.46_{10}-1$ | $1.30_{10}-4$ | 1.62 | $3.29_{10}-2$ |
| $2_{10}-8$ | $1.33_{10}-1$ | $8.71_{10}-5$ | 1.69 | $1.84_{10}-2$ |
| $1_{10}-8$ | $1.30_{10}-1$ | $6.71_{10}-5$ | 1.75 | $1.24_{10}-2$ |
| $5_{10}-9$ | $1.30_{10}-1$ | $5.30_{10}-5$ | 1.82 | $1.07_{10}-2$ |
| $1_{10}-9$ | $1.43_{10}-1$ | $4.21_{10}-5$ | 1.92 | $1.54_{10}-2$ |
| $1_{10}-16$ | $4.08_{10}+1$ | $3.32_{10}-5$ | $8.37_{10}+2$ | 6.68 |

Filter method, $\tilde{g}(s) = g(s)+a(\sin(\omega_1 s)+\cos(\omega_2 s))$ (exception)
$a=10^{-4}$, $\omega_1=10$, $\omega_2=500$, m=8,n=4, $\delta=6.09_{10}-4$, $\tau=1.85$, $\|f\|=1.83_{10}-1$

| q | Bound on Rel. Err. | $\tilde{\delta}_q$ | $\tilde{\tau}_q$ | Rel. Err. |
|---|---|---|---|---|
| 6 | $1.05_{10}+1$ | $1.59_{10}-5$ | 4.66 | $6.03_{10}-2$ |
| 5 | $8.40_{10}-1$ | $3.80_{10}-5$ | 1.96 | $1.68_{10}-2$ |
| 4 | $5.23_{10}-1$ | $1.47_{10}-4$ | 1.62 | $3.83_{10}-2$ |
| 3 | $5.49_{10}-1$ | $1.48_{10}-4$ | 1.62 | $3.83_{10}-2$ |
| 2 | $9.16_{10}-1$ | $1.03_{10}-2$ | $2.31_{10}-7$ | $4.18_{10}-1$ |

Table 5, example 4, NATTERER [10].

The fourth and last example, from NATTERER [10], provides a typical example of a deconvolution problem as one could have in correcting for the profile of a measuring instrument. Table 5 gives the results for a special perturbation of the right hand side which is analogous to the one used by Natterer. As in the other examples, these results show the effectiveness and practical applicability of both the regularisation methods and the estimates for the optimal regularisation level.

## 7. CONCLUSION

Tychonov's regularisation method and Filtered least squares minimal C-norm approximation both effectively regularise the ill-posed problem of solving a Fredholm integral equation of the first kind. Köckler's error estimate, as well as "selection on $\tau$" provide those methods with good criteria for selecting the regularisation level. They might also help to partly automate the regularisation.

Natterers idea to regularise through coarse discretisation can be combined with the filtration method resulting in a relatively cheap but not less effective regularisation method.

## REFERENCES

[1] BAKER, C.T.H., L. FOX, D.F. MAYERS & K. WRIGHT, *Numerical solution of Fredholm integral equations of the first kind,* The Computer Journal, 7 (1964) 141-148.

[2] CUPPEN, J.J.M., *Regularisation of Ill-posed Problems* (in Dutch), Masters thesis, Eindhoven University of Technology, 1978.

[3] GRAVES, J. & P.M. PRENTER, *Numerical Iterative filters applied to first kind Fredholm Integral equations*, Numer. Math. 30 (1978) 281-300.

[4] HANSON, R.J., *A Numerical method for solving Fredholm integral equations of the first kind using singular values,* SIAM J. Numer. Anal. 8 (1971) 616-622.

[5] HILGERS, J.W., *On the equivalence of Regularization and certain Reproducing kernel Hilbert Space Approximations for solving first kind problems,* SIAM J. Numer. Anal. 13 (1976) 172-184.

[6] KÖCKLER, N., *Parameterwahl und Fehlerabschätzung bei der regularisierten Lösung von inkorrekt gestellten Problemen,* thesis, Mainz 1974.

[7] LAWSON, C.L. & R.J. HANSON, *Solving Least Squares Problems,* Prentice Hall, Englewood Cliffs, 1974.

[8] LEE, J.W. & P.M. PRENTER, *An analysis of the numerical solution of Fredholm integral equations of the first kind,* Numer. Math. 30 (1978) 1-23.

[9] LOAN, C.F. VAN, *Generalizing the Singular Value decomposition,* SIAM J. Numer. Anal. 13 (1976) 76-83.

[10] NATTERER, F., *The finite element method for ill-posed problems,* RAIRO An. Numer. 11 (1977) 271-278.

[11] NATTERER, F., *Regularisierung schlecht gestellter Probleme durch Projektionsverfahren,* Numer. Math. 28 (1977) 329-341.

[12] PHILLIPS, D.L., *A technique for the numerical solution of certain integral equations of the first kind.* J. ACM, 9 (1962) 84-97.

[13] RIBIÈRE, G., *Régularisation d' operateurs.* R.I.R.O. 1 (1967) 57-79.

[14] RIELE, H.J.J. TE, *Regularisatiemethoden voor integraalvergelijkingen van de eerste soort,* hoofdstuk 7 uit: Colloquium Numerieke Programmatuur deel 2, H.J.J. Te Riele (red.), Mathematisch Centrum Amsterdam, 1977.

[15] RIELE, H.J.J. TE, *Numerical methods for first kind Fredholm integral equations,* this colloquium.

[16] STRAND, O.N., *Theory and methods related to the singular function expansion and Landwebers iteration for integral equations of the first kind,* SIAM J. Numer. Anal. 4 (1974) 798-825.

[17] TYCHONOV, A.N., *Regularisation of incorrectly posed problems,* Sov. Math. Dokl. 4 (1963) 1624-1627.

[18] VARAH, J.M., *On the numerical solution of ill-conditioned linear systems with applications to ill-posed problems,* SIAM J. Numer. Anal. 10 (1973) 257-267.

[19] VOEVODIN, V.V., *The method of regularisation,* USSR Comp. Math. and Math. Phys. 9 (1969) 228-232.

# INVARIANT IMBEDDING METHODS FOR INTEGRAL EQUATIONS

R.W. de VRIES

## 0. INTRODUCTION

The method of invariant imbedding for the solution of boundary value problems originated in 1957, with a series of papers by R. Bellman, R. Kalaba and G.M. Wing. Subsequently R. Kalaba and H. Kagiwada developed an invariant imbedding method for the solution of integral equations. In this paper a brief introduction into the theory of invariant imbedding methods for integral equations is given and some numerical experiments are discussed.

## 1. CLASSIFICATION OF LINEAR INTEGRAL EQUATIONS

Linear integral equations can be classified as Volterra or Fredholm integral equations of the first or second kind.

An equation $\phi(x) = f(x) + \lambda \int_0^x k(x,r)\phi(r)dr$ is called a Volterra equation of the second kind, while an equation $\int_0^x k(x,r)\phi(r)dr = f(x)$ is a Volterra equation of the first kind. Equations of the type $\phi(x) = f(x) + \lambda \int_a^b k(x,r)\phi(r)dr$ and $\int_a^b k(x,r)\phi(r)dr = f(x)$, with a and b constants are called Fredholm equations of the second and first kind respectively.

In general the theory of equations of the second kind is simpler then the theory of equations of the first kind. Fortunately, equations of the second kind are far more frequent in mathematical physics and so we confine ourselves to these equations. In this paper attention is given especially to Fredholm equations of the second kind.

## 2. A CLASSIC SOLUTION METHOD

The equation $\phi(x) = f(x) + \lambda \int_a^b k(x,r)\phi(r)dr$, can be solved by putting formally:

$$\phi(x) = f(x) + \lambda\phi_1(x) + \lambda^2\phi_2(x) + \ldots .$$

The function $\phi_i(x)$ can be found by defining the iterated kernel:

$$k_n(x,t) = \int_a^b \int_a^b \ldots \int_a^b k_{n-1}(x,t_{n-1}) \ldots k(t_1,t)dt_1 \ldots dt_{n-1},$$

or

$$k_n(x,t) = \int_a^b k_{n-1}(x,t_{n-1})k(t_{n-1},t)\,dt_{n-1}$$

and

$$k_1(x,t) = k(x,t) \qquad \text{with } n = 2,3,\ldots .$$

If $|k_n(x,t)|$ is bounded for $n = 1,2,\ldots$ and $|\lambda|$ is small enough then the series:

$$R(x,t;\lambda) = k_1(x,t) + \lambda k_2(x,t) + \lambda^2 k_3(x,t) + \ldots$$

converges uniformly. See [3,4].

The function $\phi_i(x)$ is given by $\phi_i(x) = \int_a^b k_i(x,y)f(y)dy$, $i = 1,2,\ldots$ $\ldots,n$ and hence the solution of the integral equation is given by the so-called resolvent kernel $R(x,t;\lambda)$. The notion of the resolvent kernel is of importance in the theory of invariant imbedding. The resolvent kernel $R(x,t;\lambda)$ satisfies the following integral equations.

$$\begin{aligned} R(x,t;\lambda) &= k(x,t) + \int_a^b k(x,u)R(u,t;\lambda)\,du \\ R(x,t;\lambda) &= k(x,t) + \lambda \int_a^b k(u,x)R(x,u;\lambda)\,du. \end{aligned} \tag{2.1}$$

## 3. SOLVING A LINEAR FREDHOLM INTEGRAL EQUATION BY MEANS OF INVARIANT IMBEDDING

Consider the equation

$$u(t) = g(t) + \int_0^a k(t,y')u(y')dy'. \tag{3.1}$$

If $a = 0$ the solution of this equation is easily found to be $u(t) = g(t)$. As we are interested in the solution for $a \neq 0$, we consider the solution $u(t)$ as a function of the interval length, and instead of equation (3.1) we solve a set of "simpler" equations:

$$u(t;x) = g(t) + \int_0^x k(t,y')u(y';x)dy' \tag{3.2}$$

for x ranging from 0 to a.

The main idea of the imbedding method is now clear: if the solution of the equation (3.2) is known for $x = 0$, find a set of "imbedding" equations which determine the solution for $0 \le x \le a$.

The set of imbedding equations can be derived as follows [1,2].

Introduce dimensionless variables $\tau = t/x$, $\eta = y/x$ and $\eta' = y'/x$. For (3.2) we get:

$$u(\tau x,x) = g(\tau x) + x \int_0^1 k(\tau x,\eta' x)u(\eta' x,x)d\eta'. \tag{3.3}$$

If $R(\tau,\eta,x)$ is the resolvent kernel of (3.3) then according to (2.1), $R(\tau,\eta,x)$ satisfies the integral equation

$$R(\tau,\eta,x) = k(\tau x,\eta x) + x \int_0^1 k(\tau x,\eta' x)R(\eta',\eta,x)d\eta'. \tag{3.4}$$

Differentiate (3.4) with respect to x; then, in obvious notation:

$$\begin{aligned} R_x(\tau,\eta,x) = {} & k_x(\tau x,\eta x) + \int_0^1 k(\tau x,\eta' x)R(\eta',\eta,x)d\eta' + \\ & + x \int_0^1 k_x(\tau x,\eta' x)R(\eta',\eta,x)d\eta' + \\ & + x \int_0^1 k(\tau x,\eta' x)R_x(\eta',\eta,x)d\eta'. \end{aligned} \tag{3.5}$$

If

$$\begin{aligned} Q(\tau,\eta,x) = {} & k_x(\tau x,\eta x) + \int_0^1 k(\tau x,\eta' x)R(\eta',\eta,x)d\eta' + \\ & + x \int_0^1 k_x(\tau x,\eta' x)R(\eta',\eta,x)d\eta' \end{aligned}$$

then:

$$R_x = Q(\tau,\eta,x) + x \int_0^1 k(\tau x,\eta' x)R_x(\eta',\eta,x)d\eta'. \tag{3.6}$$

Equation (3.6) is a linear integral equation for $R_x$, with $Q(\tau,\eta,x)$ a forcing function.

The resolvent kernel of (3.6) is $R(\tau,\eta,x)$. This can be seen as follows. If $R(\tau,\eta,x)$ is a resolvent kernel of (3.6) then:

$$R_x(\tau,\eta,x) = Q(\tau,\eta,x) + x\int_0^1 Q(\eta',\eta,x)R(\tau,\eta',x)d\eta. \tag{3.7}$$

Substitution of (3.7) into (3.6) gives:

$$Q(\tau,\eta,x) + x\int_0^1 Q(\eta',\eta,x)R(\tau,\eta',x)d\eta' =$$

$$= Q(\tau,\eta,x) + x\int_0^1 k(\tau x,\eta' x)[Q(\eta',\eta,x) +$$

$$+ x\int_0^1 Q(\eta'',\eta,x)R(\eta',\eta'',x)dy'']d\eta'$$

or

$$\int_0^1 Q(\eta',\eta,x)R(\tau,\eta',x)dy' = \int_0^1 k(\tau x,\eta' x)Q(\eta',\eta,x)d\eta' +$$

$$+ x\int_0^1 k(\tau x,\eta' x)\int_0^1 Q(\eta'',\eta,x)R(\eta',\eta'',x)d\eta''d\eta.$$

Changing the order of integration in the second term of the right hand side yields:

$$\int_0^1 Q(\eta',\eta,x)R(\tau,\eta',x)d\eta' = \int_0^1 k(\tau x,\eta' x)Q(\eta',\eta,x)d\eta' +$$

$$+ x\int_0^1 Q(\eta'',\eta,x)\int_0^1 k(\tau x,\eta' x)R(\eta',\eta'',x)d\eta'd\eta''$$

or

$$R(\tau,\eta',x) = k(\tau x,\eta' x) + x\int_0^1 k(\tau x,\eta'' x)R(\eta'',\eta',x)d\eta''$$

and this is just the integral equation for the resolvent kernel $R(\tau,\eta,x)$.

If we solve the initial value problem:

$$R_x(\tau,\eta,x) = Q(\tau,\eta,x) + x\int_0^1 Q(\eta',\eta,x)R(\tau,\eta',x)d\eta'$$

(3.8)

$$R(\tau,\eta,0) = 0 \qquad 0 \le x \le 1$$

the solution u(t) of (3.1) is found by integration of the resolvent kernel. The integral equation (3.1) is equivalent with the initial value problem (3.8). A validation of this procedure can be found in [1].

## 4. NONLINEAR INTEGRAL EQUATION

Consider the general nonlinear integral equation

$$u(t,x) = g(t) + \int_0^x k(t,y,u(y,x)dy. \tag{4.1}$$

For this kind of equation no general resolvent kernel exists. The procedure is as follows. If we differentiate (4.1) with respect to x we get a linear Fredholm equation for $u_x$. For this equation there exists a resolvent kernel and (4.1) is again equivalent with a coupled system of initial value problems. The procedure is described in detail in [1] and here we give only the results.

Introduce again dimensionless variables $\tau$ and $\eta$ and the function $v(\tau,x) = u(\tau x,x)$. The imbedding equations are in this case:

$$v_x(\tau,x) = G(\tau,x); \quad v(\tau,0) = g(0);$$

$$R_x(\tau,\xi,x) = Q(\tau,\xi,x) + x\int_0^1 R(\tau,\eta,x)Q(\eta,\xi,x)d\eta;$$

$$R(\tau,\xi,0) = \frac{\partial}{\partial v}k(0,0,g(0)).$$

$G(\tau,x)$ and $Q(\tau,x)$ are functions which contain integrals over the first and second derivatives of the kernel k. This procedure applies only to smooth kernels which are twice differentiable with respect to all arguments. We shall call this procedure: "interval imbedding". A second procedure for the solution of nonlinear integral equations is the method of parameter imbedding. In this case we introduce a parameter $\lambda$ and consider the solution

of the equation:

$$u(t) = g(t) + \lambda \int_0^1 k(t,y,u(y))dy \tag{4.2}$$

as a function of $\lambda$.

Differentiation with respect to $\lambda$ gives again a linear Fredholm equation for $u_\lambda$. For equation (4.2) the following derivatives must exist: $\frac{\partial k}{\partial u}$ and $\frac{\partial^2 k}{\partial u^2}$. For a detailed account the reader is referred to [1,2].

## 5. NUMERICAL EXPERIMENTS

Two types of nonlinear integral equations are considered. As an example of an equation with a smooth kernel we take the following equation:

$$\phi(t) = 0.8t^2 + \int_0^1 t^2\phi^2(y)dy. \tag{5.1}$$

As a second example we consider:

$$\phi(x) = -\int_0^1 T(x,y)[F(y) - \alpha^2 \sin \phi(y)]dy \tag{5.2}$$

with $T(x,y)$ the "triangular (non-smooth) kernel":

$$T(x,y) = \begin{cases} x(1-y) & 0 \le x \le y \\ y(1-x) & y \le x \le 1. \end{cases}$$

For the computations we use the DEC 10 system of Twente University. The computer programs are written in Algol 60. For equation (5.1) we use interval imbedding as well as parameter imbedding and compare execution times. In both methods the right hand side of the imbedding equations contain integrals which are approximated by a 5-points Gaussian integration formula [6]. Hence we have to solve a coupled system of $5^2 + 5 = 30$ initial value problems. This system of initial value problems is solved by an extrapolation method given by Bulirsch and Stoer [5]. The routine "Diffsys", which is used here, gives a very efficient algorithm for the solution of systems of ordinary differential equations.

The results for $\phi(t)$, equation (5.1), are:

| t = Gaussian point | Interval imbedding | Parameter imbedding | Exact solution |
|---|---|---|---|
| $0.4691008\ 10^{-1}$ | $0.22005\ 10^{-2}$ | $0.22010\ 10^{-2}$ | $0.22006\ 10^{-2}$ |
| 0.230765345 | $0.53253\ 10^{-1}$ | $0.53253\ 10^{-1}$ | $0.53253\ 10^{-1}$ |
| 0.5 | 0.25 | 0.25 | 0.25 |
| 0.76923465 | 0.59172 | 0.59172 | 0.59172 |
| 0.953089925 | 0.90838 | 0.90838 | 0.90838 |

The execution time for interval imbedding is 38 sec., for parameter imbedding it is 7 sec..

Equation (5.2) can be transformed into an equation of Hammerstein type which has exactly one solution for $\alpha^2 < \pi^2$, see e.g. [4]. This problem is equivalent to the problem of forced oscillations of finite amplitude, which are governed by the following differential equation:

$$\frac{d^2\phi}{dt^2} + \alpha^2 \sin \phi(t) = F(t) \tag{5.3}$$

with boundary conditions $\phi(0) = \phi(1) = 0$.

For our computations $F(t) = \sin \pi t$ and $\alpha^2 = 3.1$. We solve (5.2) by using the following four methods:

(a) Interval imbedding and trapezoidal integration in the intervals [0,x] and [x,1].

(b) Parameter imbedding and Gaussian integration.

(c) Parameter imbedding and Simpson integration.

(d) Parameter imbedding and trapezoidal integration.

The obtained results are:

(1) with 10-points trapezoidal integration: $\phi(0.5) = ?$ (execution time > 300 sec.).

(2) with 5-points Gaussian integration: $\phi(0.5) = -0.15224$, execution time 7 sec. and 7-points: $\phi(0.5) = -014996$ in 14 sec..

(3) with 8-points Simpson integration: $\phi(0.5) = -0.14787$, execution time 32 sec..

(4) with 10-points trapezoidal integration: $\phi(0.5) = -0.14809$, execution time 87 sec..

To compare these results, we solved equation (5.3) by a multiple shooting method as given in [7,8]. If this method converges, it converges quadratically. As an accuracy test, the sum of the squares of the residuals is used. For this sum we found: $1.6.10^{-7}$, $1.1.10^{-7}$, $5.2.10^{-8}$, $2.0.10^{-19}$,

$3.2.10^{-30}$. In respect to these results we expect that the solution $\phi(0.5) = -0.14753$ found by the multiple shooting method is the most accurate one. The execution time is 8 sec..

## 6. CONCLUSIONS

For integral equations with a smooth kernel, invariant imbedding, and especially parameter imbedding, is an efficient method. These methods apply to linear as well as nonlinear equations. For singular kernels, invariant imbedding seems to be highly sensitive to the method used for the integrations. To improve the imbedding method for singular integral equations, integration routines should be conceived which take into account the singularity of the kernel.

## REFERENCES

[1] KAGIWADA, H.H. & R. KALABA, *Integral equations via imbedding methods,* Addison-Wesley, 1974.

[2] KAGIWADA, H.H. & R. KALABA, *Imbedding methods for integral equations with applications,* J. Opt. Theory and Applics., 24 (1978).

[3] SMIRNOW, W.J., *Lehrgang der höheren Mathematik IV,* VEB, Deutscher Verlag der Wissenschaften, 1962.

[4] TRICOMI, F.G., *Integral Equations,* Interscience Publishers, 1970.

[5] BULIRSCH, R. & J. STOER, *Numerical Treatment of Ordinary Differential Equations by Extrapolation Methods,* Numer. Math. Bd 8, 1966.

[6] ABRAMOWICH, M. & I.A. STEGUN, *Handbook of Mathematical Functions,* Dover Publications, 1965.

[7] STOER, J., *Einführung in die Numerische Mathematik I,* Springer-Verlag, 1972.

[8] STOER, J. & R. BULIRSCH, *Einführung in die Numerische Mathematik II,* Springer-Verlag, 1973.

# PART II, APPLICATIONS

# NUMERICAL APPLICATION OF INTEGRAL EQUATIONS IN SEMICONDUCTOR PHYSICS

G. De MEY

# 1. POTENTIAL DISTRIBUTION IN A HALL GENERATOR

## 1.1. Introduction

A Hall generator is a thin semiconductor layer placed in a magnetic field $\bar{B}$, directed perpendicular to the plane of the sample (fig. 1).

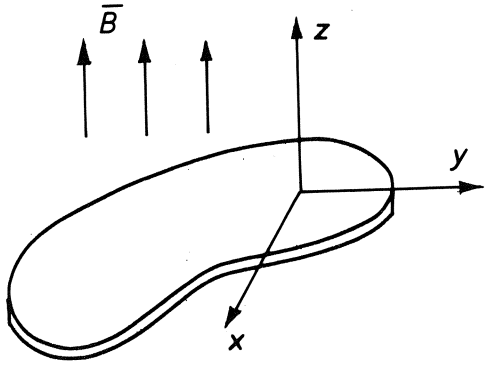



Figure 1.

The thickness being typical 1μ, whereas the dimensions in the other directions attain several mm. The potential problem can therefore be approximated by a two-dimensional equation in x and y.

The Hall generator is usually provided with four contacts (fig. 2).

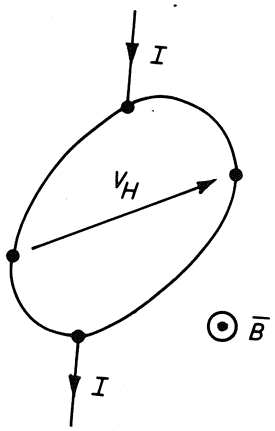



Figure 2.

Inserting a current I through two opposite contacts, a voltage proportional to the magnetic field $\bar{B}$ will be generated at the other contacts. This device can be used to measure the magnetic field strength [1]. If the magnetic field is known, this device is used intensively to determine important semiconductor parameters such as the Hall mobility and magnetoresistance [2].

Van der Pauw has shown that the relation between the Hall voltage and the supplied current I is given by [3]:

$$(1) \qquad V_H = \mu_H B \frac{I}{d\sigma} ,$$

where: $\mu_H$: mobility
d: thickness
$\sigma$: conductivity.

However, (1) is only valid if the four contacts are point shaped. On the other hand, point shaped contacts offer many technological difficulties [4], so that contacts with finite dimensions should be used. The Hall voltage is then no longer given by (1). The correction can only be found by calculating the potential distribution in the Hall generator [5], which will be done now by an integral equation technique.

## 1.2. Basic equations

In a Hall generator, the following relation exists between the current density $\bar{J}$ and the electric field $\bar{E}$:

$$(2) \qquad \bar{J} = \sigma[\bar{E} - \mu_H(\bar{E} \times \bar{B})].$$

The relation (2) means that the vectors $\bar{J}$ and $\bar{E}$ include a constant angle $\theta_H = \text{arctg}(\mu_H B)$, called the Hall angle. On the other hand, one has the vector equations for $\bar{J}$ and $\bar{E}$:

$$(3) \qquad \nabla \cdot \bar{J} = 0 \qquad \text{and} \qquad \nabla \times \bar{E} = 0.$$

Due to (3), the electric field $\bar{E}$ can be derived from a scalar potential function $\phi$:

$$(4) \qquad \bar{E} = -\nabla\phi.$$

From (2), (3) and (4) it can be easily shown that the potential $\phi$ will satisfy the Laplace equation:

$$\nabla^2\phi = 0. \tag{5}$$

The equation (5) does not change when the value of the Hall angle varies.

In order to explain the boundary conditions, one is referred to the particular geometry shown in fig. 3.

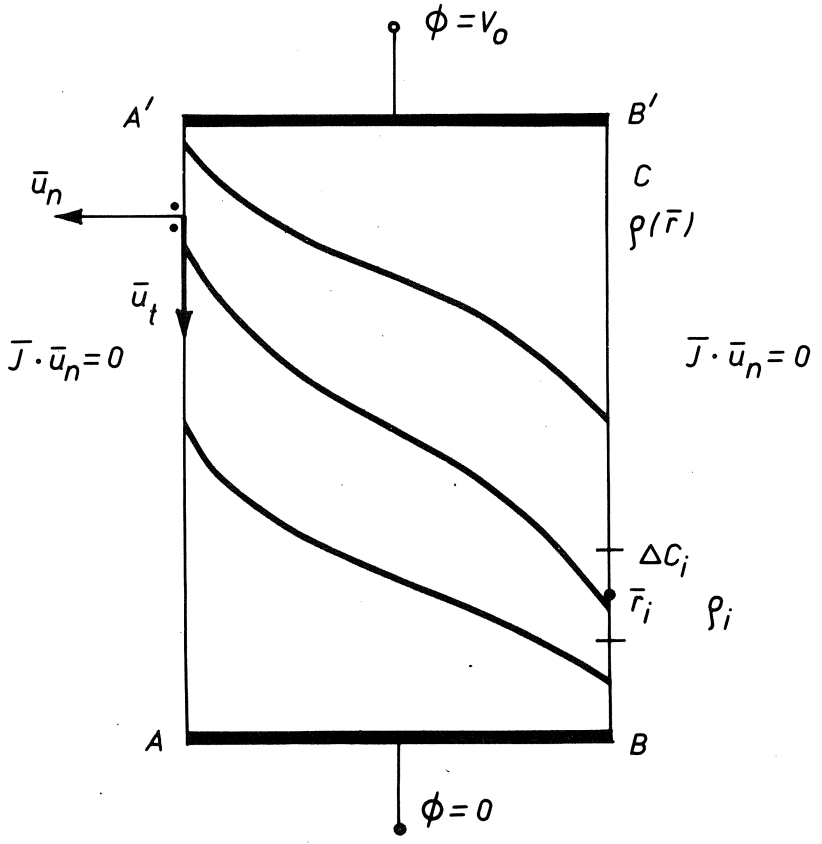



Figure 3.

This rectangular Hall plate is provided with two metallic contacts AB and A'B' supplying a current through the device. At the contacts, the potential is known which gives rise to the boundary condition:

$$\phi = 0, \qquad \bar{r} \in AB;$$

(6)

$$\phi = V_0, \qquad \bar{r} \in A'B'.$$

At the free boundaries AA' and BB', the current density $\bar{J}$ should be tangential as no current can flow out of the semiconductor without any metallic contact. Thus:

$$\bar{J}\cdot\bar{u}_n = 0, \qquad \bar{r} \in AA', BB'. \tag{7}$$

By using (2) and (4), the condition (7) is transformed into:

$$\nabla\phi\cdot\bar{u}_n - \mu_H B\nabla\phi\cdot\bar{u}_t = 0, \qquad \bar{r} \in AA', BB'. \tag{8}$$

For $\mu_H B = 0$, one gets the classic potential problem where either $\phi$ or the normal derivative $\nabla\phi\cdot\bar{u}_n$ is given along the boundary.

Due to the tangential component in (8), several classical methods for solving the Laplace equation such as separation of variables cannot be used even for the simple rectangular geometry shown in fig. 3. Only the conformal transformation method has been used successfully [6], [7], [8]. However, in order to treat arbitrary geometries, a numerical method based on an integral equation technique has been set up [9].

## 1.3. Integral equation

In order to establish an integral equation, one has to know the Green's function of the Laplace equation. This function is a solution of:

$$\nabla^2 G(\bar{r} \mid \bar{r}') = \delta(\bar{r} - \bar{r}'). \tag{9}$$

(9) has several solutions, but for our purpose we may retain the radial symmetric one depending on the distance $|\bar{r} - \bar{r}'|$:

$$G(\bar{r} \mid \bar{r}') = \frac{1}{2\pi} \ell n|\bar{r} - \bar{r}'|. \tag{10}$$

The solution of (5) is now written as:

$$\phi(\bar{r}) = \oint_C \rho(\bar{r}')G(\bar{r} \mid \bar{r}')dc', \tag{11}$$

where $\rho(\bar{r}')$ is still an unknown function defined along the boundary C of the Hall generator. Imposing the boundary conditions (6) and (8) on the proposed solution (11), one gets:

$$\frac{1}{2\pi}\oint_C \rho(\bar{r}')\ell n|\bar{r}-\bar{r}'|dc' = \begin{cases} V_0, & \bar{r} \in A'B', \\ 0, & \bar{r} \in AB; \end{cases}$$

(12)

$$-\frac{\rho(\bar{r})}{2} + \oint_C \rho(\bar{r}')\,\frac{(\bar{r}-\bar{r}')\cdot\bar{u}_n - \mu_H B(\bar{r}-\bar{r}')\cdot\bar{u}_t}{2\pi|\bar{r}-\bar{r}'|^2}\,dc' = 0, \qquad \bar{r} \in AA', BB'.$$

(12) is an integral equation in the unknown function $\rho$. It is evident that the two-dimensional Laplace equation and the boundary conditions have been replaced by an integral equation only involving the (one-dimensional) boundary C. Once the source function $\rho$ has been determined from (12), the potential $\phi$ can be calculated in any point $\bar{r}$ by (11).

The term $-\rho(\bar{r})/2$ occurring in (12) is called a fundamental discontinuity [10]. It is caused by the logarithmic behaviour of the Green's function.

## 1.4. Numerical solution

In order to solve the integral equation (12) numerically, the boundary C is divided into n parts $\Delta C_i$. The unknown function $\rho(\bar{r})$ is replaced by n unknown constants $\rho_i$, concentrated at the centre points $\bar{r}_i$ of each interval (fig. 3). The integral (11) can then be replaced by:

$$\phi(\bar{r}) = \sum_{j=1}^{n} \rho_j G(\bar{r} \mid \bar{r}_j)|\Delta C_j|, \tag{13}$$

where $|\Delta C_j|$ denotes the length of the j-th interval. When $\bar{r}$ is located on the boundary, a divergence occurs if $\bar{r}$ coincides with $\bar{r}_j$ as the Green's function $G = \infty$. It is then necessary to consider $\rho$ being approximated by a constant function in the interval $\Delta C_j$, giving a contribution of the following form:

$$\rho_j \int_{\Delta C_j} G(\bar{r}_j \mid \bar{r}')dc' = \frac{\rho_j}{\pi}\int_0^{|\Delta C_j|/2} \ell nx\,dx = \frac{\rho_j}{2\pi}|\Delta C_j|\left(\ell n\frac{|\Delta C_j|}{2} - 1\right). \tag{14}$$

The integral equation (12) becomes then:

$$\frac{\rho_i}{2\pi}|\Delta C_i|\left(\ell n\frac{|\Delta C_i|}{2}-1\right)+\sum_{\substack{j=1\\ j\neq i}}^{n}\rho_j\frac{1}{2\pi}\ell n|\bar{r}_i-\bar{r}_j||\Delta C_j|$$

$$(15)\qquad =\begin{cases}V_0', & \bar{r}_i\in A'B',\\ 0, & \bar{r}_i\in AB;\end{cases}$$

$$-\frac{\rho_i}{2}+\sum_{\substack{j=1\\ j\neq i}}^{n}\frac{(\bar{r}_i-\bar{r}_j)\cdot\bar{u}_n-\mu_H B(\bar{r}_i-\bar{r}_j)\cdot\bar{u}_t}{2\pi|\bar{r}_i-\bar{r}_j|^2}|\Delta C_j|\rho_j=0,\qquad \bar{r}_i\in AA',BB'.$$

(15) is a linear algebraic set which can be easily solved numerically.

Higher order approximations for the unknown function $\rho$ have also been investigated to solve the integral equation such as polynomials and spline functions [11], [12], [13]. However, this does not lead to more accurate results. It is believed that this fact is caused at the corners A, B, B' and A' where the boundary and hence the kernel suddenly change. Another experimental fact, which is not fully explained yet, is the behaviour of the relative error as a function of the number of unknowns n. With a good approximation a 1/n behaviour is found in practically all cases, even when higher approximations for $\rho$ were used. This fact should be a consequence of the logarithmic behaviour of the kernel. However, a rigorous proof of this assumption cannot be given at this moment.

## 1.5. Conclusional remarks

For a review of integral equation methods in potential, biharmonic and elastostatic problems, one is referred to the recent book "Integral Equation Methods", written by E. Jaswon and G. Symm [14].

Whereas some "classical" methods are not suited for the potential equation in a Hall generator, the integral equation offers no particular difficulties. The technique is easily applied to arbitrary geometries with more than two contacts. In order to investigate the influence of the geometry on the Hall voltage (section 1.1) a computer program has been written to calculate the potential distribution in a Hall plate with an arbitrary polygonal shape [15].

It should be notified that the integral equation (12) is a Fredholm equation of the first kind (not second kind) as for $\bar{r}\in AB$ or A'B' the

term before the integral disappears. Due to the change of the boundary conditions, the kernel is no longer symmetrical notwithstanding the fact that the Green's function does not vary by interchanging $\bar{r}$ and $\bar{r}'$.

## 2. MINORITY CARRIER CONCENTRATION IN SEMICONDUCTOR DEVICES

### 2.1. Introduction

The Hall effect studied in chapter 1 can be seen as a large scale problem, as Hall generators have dimensions of typical 1 mm and even more. For the study of pn junction diodes e.g., another approximation of the fundamental semiconductor equations should be done, as the phenomena occur in a much smaller scale. In this case the minority carrier concentration becomes the fundamental quantity. The potential distribution can be found analytically by the so called abrupt depletion approximation. For our purpose, the potential only appears in the boundary condition, and will not be considered further on. One will obtain a diffusion-type equation, which can also be transformed into an integral equation.

Several examples can be treated: more dimensional problems or time dependent phenomena occurring in semiconductor devices such as pn junctions, solar cells and transistors. We shall give a two-dimensional problem in a diode and a time dependent one in a solar cell [16], [17], [18].

### 2.2. Basic equations

A semiconductor device usually contains a junction of a p-type and a n-type material (fig. 4). In each domain, a so called depletion layer will be formed at the junction interface. Within the scope of this paper, it is sufficient to know that the boundary condition for the phenomena in the n-layer are expressed along the line A'B' and not at the junction A"B". Without loss of generality, we will restrict us to the n-layer, where the holes are the minority carriers with a concentration $p(\bar{r},t)$.

The current density $\bar{J}$ is related to the concentration p by:

$$\bar{J} = -qD\nabla p, \tag{16}$$

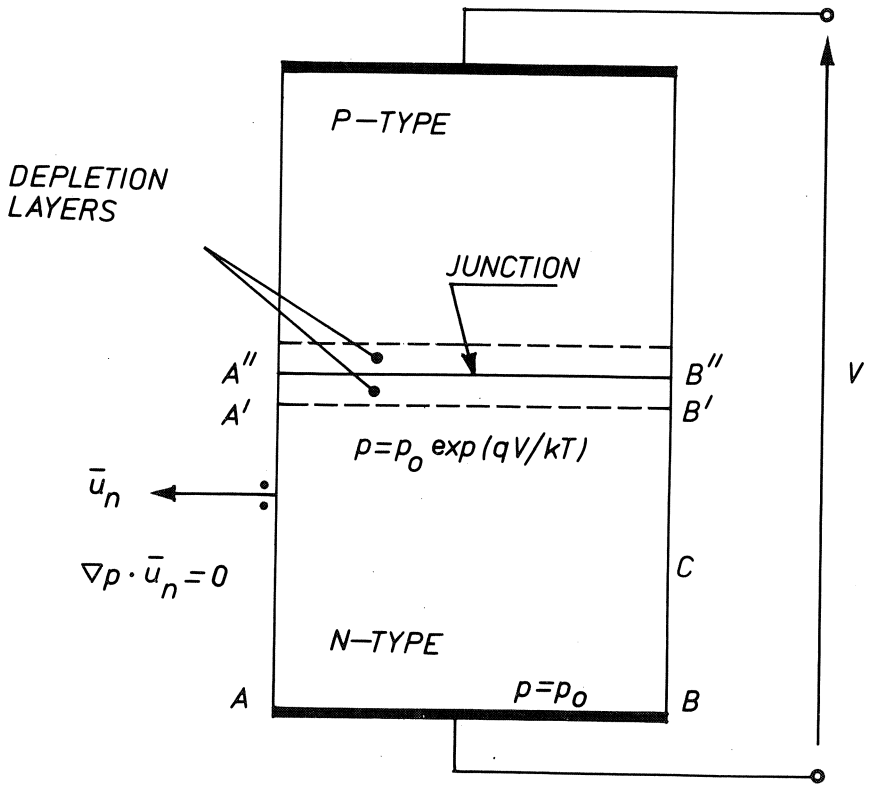



Figure 4.

where D is the diffusion constant and q = 1.6 $10^{-19}$ Cℓb the elementary charge. In a semiconductor, charge carriers can be generated by illumination or can disappear by recombination processes. These phenomena are described by the equation:

$$\nabla \cdot \bar{J} = - \frac{p-p_0}{\tau} + Q(\bar{r},t) - \frac{\partial p}{\partial t} , \tag{17}$$

where $p_0$ is the equilibrium concentration, $\tau$ the recombination time constant and G the number of carriers generated per unit time and volume. Putting (16) into (17) one obtains the equation:

$$\nabla^2 p - \frac{p-p_0}{L^2} = - \frac{Q(\bar{r},t)}{D} + \frac{\partial p}{\partial t} , \tag{18}$$

where $L = (D\tau)^{\frac{1}{2}}$ is the diffusion length.

If a voltage V is applied across the diode shown on fig. 4, the boundary conditions are:

$$\begin{aligned} p &= p_0 e^{qV/kT}, \qquad \bar{r} \in A'B', \\ p &= p_0, \qquad \bar{r} \in AB, \end{aligned} \tag{19}$$

where k is Boltzmann constant and T the absolute temperature. At a free boundary the current density $\bar{J}$ must be tangential, and due to (16) this gives result to:

$$\nabla p \cdot \bar{u}_n = 0, \qquad \bar{r} \in AA', BB'. \tag{20}$$

Therefore, a semiconductor device problem is described by the equation (18) and the boundary conditions (19) - (20), according to the theory of the abrupt depletion approximation.

Other boundary conditions than (20) can occur if more sophisticated phenomena such as surface recombination are taken into account. However, for the sake of clarity, we shall restrict us to the simple situation of condition (20).

## 2.3. Static characteristics of a pn junction diode

Let us consider a pn junction diode as shown on fig. 4. As we are only interested in static characteristics, the time dependence in (17) can be dropped. As diodes are normally encapsulated no light can penetrate into the semiconductor so that the generation term $G(\bar{r},t)$ in (17) equals zero. We get the equation:

$$\nabla^2 p - \frac{p-p_0}{L^2} = 0. \tag{21}$$

The boundary conditions (19) and (20) remain unchanged.

In order to construct an integral equation, the same procedure followed in section 1.3 will be followed. First of all, we need a Green's function. Rewriting (21) in polar coordinates, dropping the angular dependence and putting $\delta(\bar{r}-\bar{r}')$ in the right hand member, one gets the solution:

$$G(\bar{r} \mid \bar{r}') = -\frac{1}{2\pi} K_0\left(\frac{|\bar{r}-\bar{r}'|}{L}\right), \tag{22}$$

where $K_0$ is the zero-th order modified Bessel function of the second kind.

The unknown concentration $p(\bar{r})$ is now written as (cf. formula (11)):

$$p(\bar{r}) = p_0 + \oint_C \rho(\bar{r}')G(\bar{r} \mid \bar{r}')dc', \tag{23}$$

where $\rho(\bar{r})$ is an unknown source function defined along the boundary C. Imposing the boundary conditions (19) and (20) on the proposed solution (23), one obtains:

$$\begin{aligned} & p_0 - \frac{1}{2\pi}\oint_C \rho(\bar{r}')K_0\left(\frac{|\bar{r}-\bar{r}'|}{L}\right)dc' = \begin{cases} p_0 e^{qV/kT}, & \bar{r} \in A'B', \\ p_0, & \bar{r} \in AB; \end{cases} \\ & -\frac{\rho(\bar{r})}{2} + \oint_C \rho(\bar{r}')K_1\left(\frac{|\bar{r}-\bar{r}'|}{L}\right)\frac{(\bar{r}-\bar{r}')\cdot\bar{u}_n}{2\pi L|\bar{r}-\bar{r}'|}dc' = 0, \qquad \bar{r} \in AA', BB'. \end{aligned} \tag{24}$$

(24) is an integral equation in the unknown function $\rho$. It should be noted that (24) is quite similar to the equation (12) found in section 1.3 for a potential problem. Only the Green's function has to be adjusted if another equation is considered.

For the numerical solution, the same procedure outlined in section 1.4 can be used. From the experimental results, similar results have been observed so that most conclusions still hold.

Once p is found, the current density $\bar{J}$ is calculated by (16). By integrating $\bar{J}$ along A'B' the total current I supplied at the contacts can be calculated for different values of the applied voltage V so that the characteristics of the device are known.

## 2.4. Transient behaviour of a solar cell

A solar cell is a pn junction diode illuminated at one side. The thickness of the p-layer is sufficiently small so that the light intensity is still high at the junction interface (fig. 5). In both layers depletion regions will be formed, and we shall restrict us again to the study of minority carriers in the n-layer. It should be noted that the same method can still be applied if both layers are taken into account. The structure is assumed one-dimensional so that the transient behaviour of the minority

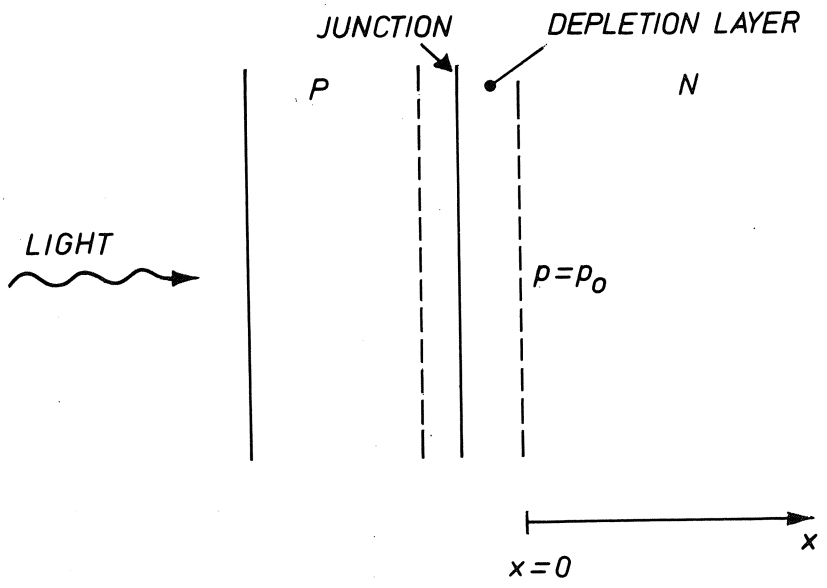



Figure 5.

carriers will be expressed by a function of the distance x and time t: p(x,t).

We want to calculate p(x,t) if the solar cell is illuminated at t = 0 by a constant light flux. As the light intensity decreases exponentially in a material and the generation of charge carriers is proportional to the decay of this light intensity, the equation (17) turns out to be:

$$\frac{\partial p}{\partial t} - D\frac{\partial^2 p}{\partial x^2} + \frac{p-p_0}{\tau} = Ae^{-\alpha x}u(t), \tag{25}$$

where A is a proportionality constant, $\alpha$ the light absorption coefficient and u(t) the unit step function. If the cell is operated under short circuit condition (zero applied voltage) the boundary condition read:

$$p = p_0 \qquad \text{at } x = 0, \text{ for all } t. \tag{26}$$

For $t < 0$, the cell was not illuminated and a stationary situation exists. It is then easily verified that $p = p_0$ for $t < 0$. Hence, the initial condition is:

(27) $\quad p = p_0 \qquad$ at $t = 0$, for all $x$.

In order to transform the equation (25) into an integral equation, the Green's function should be known. It is a solution of:

$$\frac{\partial G}{\partial t} - D\frac{\partial^2 G}{\partial x^2} + \frac{G}{\tau} = \delta(x)\delta(t). \tag{28}$$

In time dependent problems, one should take care that the causality principle is respected. Otherwise solutions can be obtained without any physical significance. This means that the Green's function must be zero for negative times. One finds:

$$G(x,t) = \begin{cases} \dfrac{e^{-t/\tau}}{2(\pi Dt)^{\frac{1}{2}}} e^{-x^2/4Dt}, & t \geq 0, \\ 0, & t < 0. \end{cases} \tag{29}$$

Neglecting the term $e^{-t/\tau}$ in (29), one obtains a Gaussian distribution for which the mean square deviation $\sigma \sim \sqrt{t}$, which is a well-known solution of the diffusion equation.

In order to construct an integral equation, the solution of (25) is written as:

$$p(x,t) = p_0 + p_1(x,t) + \int_0^t \rho(t')G(x,t-t')dt'. \tag{30}$$

$p_1(x,t)$ is a particular solution of (25), whereas the integral appearing in (30) only satisfies the equation (25) without right hand member. A particular solution is found by integration:

$$\begin{aligned} p_1(x,t) &= A \int_0^t dt' \int_{-\infty}^{+\infty} e^{-\alpha x'} G(x-x',t-t')dt' \\ &= Ae^{-\alpha x} \frac{\tau}{1-\alpha^2 L^2}\left[1 - e^{-(1-\alpha^2 L^2)t/\tau}\right]. \end{aligned} \tag{31}$$

For $t = 0$, one has $p_1(x,0) = 0$ and the integral in (30) vanishes too. This means that the initial condition (27) is fulfilled automatically. Only the boundary condition (26) should be imposed on the proposed solution (30), yielding:

$$(32) \qquad \int_0^t \rho(t')G(0,t-t')dt' = -p_1(0,t).$$

(32) is a Volterra integral equation inthe unknown function $\rho(t)$. For details regarding the numerical solution one is referred to the literature [19].

For the particular problem explained here, an analytic solution has been obatined by the Laplace transformation. The numerical results were found to be in good agreement with the exact values.

If the equation (25) would have been solved in a two-dimensional area with boundary C, a mixed Volterra-Fredholm type integral equation is obtained. The unknown source function $\rho(\bar{r},t)$ is then also defined along the boundary C.

## 2.5. Conclusional remarks

It has been shown that more-dimensional and time-dependent problems in semiconductor junctions can be analysed by integral equation techniques. The method is similar to the integral equation for potential problems outlined in the first section. Hence, the same conclusions can be drawn here.

# 3. CONCLUSION

Some semiconductor problems described by linear partial differential equations have been investigated by equivalent integral equations. This procedure has several advantages such as the reduction of the dimension, the economy in computation time and programming complexity. On the other hand, several problems such as the relationship error vs. number of unknowns, the non-increasing accuracy when higher order approximations are used, remain still unexplained. Several authors developed optimized methods to solve integral equations numerically [20], [21], [22], but they never applied their methods to equations with a kernel showing a singularity as (10), (22) or (29). So, more research is needed in order to understand the integral equation method and to develop optimal numerical solution methods.

The examples mentioned in this paper only give a limited view on the possibilities of integral equation methods. Similar techniques have also been used for the thermal diffusion problem [23], [24], eigenvalues of the Helmholtz equation [25], elastostatic problems [14] and electromagnetic field calculations in moving media. The construction of an integral equation can also be performed by other methods than those presented in this

contribution (cf. [26], [27]).

REFERENCES

[1] J. Haeusler und H. Lippmann:
*"Hallgeneratoren mit kleinen Linearisierungsfehler"*
Solid State Electronics, 1968, vol. 11, p. 173-182.

[2] E.H. Putley:
*"The Hall effect and semiconductor physics"*
Dover, New York, 1968, p. 42-55.

[3] L.J. Van der Pauw:
*"A method of measuring specific resistivity and Hall effect of discs of arbitrary shape"*
Philips Research Reports, 1958, vol. 13, p. 1-9.

[4] G. De Mey und J. Van Campenhout:
*"Hall Beweglichkeitsmessungen auf dünnen Halbleiterschichten"*
Archiv für Technisches Messen, 1975, vol. 4, p. 55-56.

[5] G. De Mey:
*"Influence of sample geometry on Hall mobility measurements"*
Archiv für Elektronik und Uebertragungstechnik, 1973, vol. 27, p. 309-313.

[6] J. Haeusler:
*"Exakte Lösungen von Potentialproblemen Beim Halleffekt durch konforme Abbildung"*
Solid State Electronics, 1966, vol. 9, p. 417-441.

[7] J. Haeusler:
*"Randpotentiale von Hallgeneratoren"*
Archiv für Elektrotechnik, 1971, vol. 54, p. 77-81.

[8] G. De Mey:
*"Field calculations in Hall samples"*
Solid State Electronics, 1973, vol. 16, p. 955-957.

[9] G. De Mey:
*"Integral equation for the potential distribution in a Hall generator"*
Electronics Letters, 1973, vol. 9, p. 264-266.

[10] J. Van Bladel:
*"Electromagnetic fields"*
McGraw - Hill, New York, 1964, p. 58.

[11] S. De Wolf and G. De Mey:
*"Numerical methods for solving integral equations of potential problems"*
Information Processing Letters, 1975, vol. 3, p. 121-124.

[12] K. Stevens and G. De Mey:
*"Higher order approximations for integral equations of potential problems"*
International Journal of Electronics (to be published).

[13] N. Papamichael (private communication).

[14] M. Jaswon and G. Symm:
*"Integral equation methods in potential theory and elastostatics"*
Academic Press, London, 1977.

[15] G. De Mey:
*"Program CALPOT1, a program to solve the Laplace equation in arbitrary geometries"*
Research Report LEM7801, Laboratory of Electronics, Ghent State University, Ghent, Belgium.

[16] P. De Visschere and G. De Mey:
*"Integral equation approach to the abrupt depletion approximation in semiconductor components"*
Electronics Letters, 1977, vol. 13, p. 104-106.

[17] G. De Mey, B. Jacobs and F. Fransen:
*"Influence of junction roughness on solar cell characteristics"*
Electronics Letters, 1977, vol. 13, p. 657-658.

[18] G. De Mey:
*"An integral equation method to calculate the transient behaviour of a photovoltaic solar cell"*
Solid State Electronics, 1978, vol. 21, p. 595-596.

[19] G. De Mey:
*"An integral equation method for the numerical calculation of ion drift and diffusion in evaporated dielectrics"*
Computing, 1976, vol. 17, p. 169-176.

[20] W. Squire:
*"Numerical solution of linear Volterra equation of the first kind"*
Research Report, GK 3799, West Virginia University, Department of Aerospace Engineering, 1969.

[21] S. Ljung and L. Ljung:
*"Fast numerical solution of integral equations with stationary kernels"*
Report LiTH, ISY - I - 0200, Linköping University Sweden, Department of Electrical Engineering, 1978.

[22] A. Haeghemans, Louvain University, Belgium
(private communication).

[23] R. Shaw:
*"An integral equation approach to diffusion"*
International Journal of Heat and Mass Transfer, 1974, vol. 17, p. 693-699.

[24] G. De Mey:
*"An integral equation approach to AC diffusion"*
International Journal of Heat and Mass Transfer, 1976, vol. 19, p. 702-704.

[25] G. De Mey:
*"Calculation of eigenvalues of the Helmholtz equation by an integral equation"*
Numerical Methods in Engineering, 1976, vol. 10, p. 59-66.

[26] G. De Mey:
*"Integral equations for potential problems with the source function not located on the boundary"*
Computers and Structures, 1978, vol. 8, p. 113-115.

[27] R. Hostens and G. De Mey:
*"An integral equation method for the numerical solution of the Laplace equation without Green's function"*
Computer Physics Communications (to be published).

# INTEGRAL EQUATIONS AND POPULATION DYNAMICS

O. DIEKMANN

## 0. INTRODUCTION

Mathematical models from population dynamics are a rich source of interesting integral equations. In this chapter I shall try to support this assertion through the presentation of some examples. The choice of the examples reflects my own interests and the survey is by no means complete.

In general, the models that will be discussed are not meant as a detailed description of reality. The phenomena that are to be modelled are much to complex to be "catched" in simple models. In fact these simple models are intended as theoretical experiments: one tries to expose the consequences of a set of assumptions in order to develop a feeling for the relation between interaction mechanisms and dynamical behaviour. And one hopes that in the long run a combination of the outcome of many such experiments leads to a better understanding of the phenomena which are observed in nature.

Consequently, the main interest is in obtaining a clear *qualitative* picture of the behaviour of solutions of the equations. This means that an analytical treatment of the equations, which are most of the time nonlinear, is preferable to a numerical treatment. So in this chapter the emphasis will be on analytical aspects. Yet *numerical experiments* can contribute quite a bit to the obtainment of qualitative insight and, in passing, I shall mention some problems which in my opinion are well suited for a numerical analysis.

The presentation will be rather informal. I intend to enlighten the most important ideas and results without worrying about precise mathematical formulations. However, all the time I shall give references where such formulations can be found.

The chapter is divided into three sections devoted to the topics population growth, selection and migration in population genetics, and epidemic models. I shall restrict my attention to deterministic models. For a general introduction to the mathematical theory of populations I refer to HOPPENSTEADT [31], LEVIN [38], MAY [43] and MAYNARD SMITH [45,46]. The reader interested in the historical papers of the pioneers will find reading SCUDO & ZIEGLER [60] and SMITH & KEYFITZ [64] a pleasant activity. Some older books on the subject are LOTKA [40], KOSTITZIN [36] and VOLTERRA [65].

## 1. POPULATION GROWTH

1.1. One of the earliest models of population growth, due to Malthus (± 1790), states that

$$\frac{dN}{dt} = bN - dN, \tag{1.1}$$

where N is the total population size, which is a function of the time t. The interpretation is as follows: the change in N is supposed to be equal to the difference of the effect of birth and death which are both assumed to be proportional to N itself (with respective constants of proportionality b and d). Clearly this differential equation has the solution

$$N(t) = N(0)e^{(b-d)t}. \tag{1.2}$$

Hence generically (i.e., for $b \neq d$) either the population evolves exponentially towards extinction or it grows exponentially beyond any fixed level. The parameter b-d is called the intrinsic rate of natural increase or the Malthusian parameter.

1.2. In ± 1840 Verhulst introduced density dependence into this model in order to explain the self-limiting growth which is often observed in nature. If one or both of the coefficients b and d are allowed to depend linearly on N, one can deduce from (1.1) the equation

$$\frac{dN}{dt} = (\beta - \delta N)N. \tag{1.3}$$

Now the equation is nonlinear but still it can be solved explicitly (note that the transformation $x = N^{-1}$ leads to a linear equation for x):

$$N(t) = \beta\delta^{-1}(1+Ce^{-\beta t})^{-1}, \tag{1.4}$$

where $C = \beta(\delta N(0))^{-1} - 1$.

For $0 < N(0) < \beta\delta^{-1}$ the solution has the form shown in Figure 1 and it is called the logistic curve.

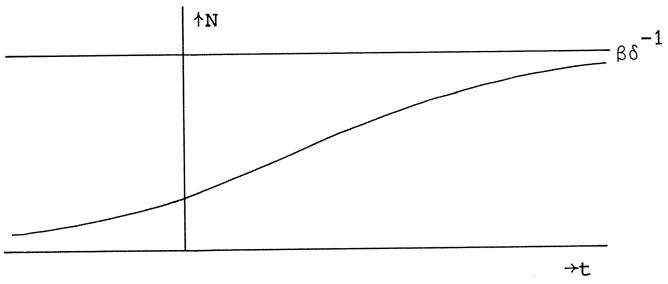



Figure 1

1.3. If one empirically observes the growth of a single bacterial species, then often the curve has the form shown in Figure 2, which clearly is quite

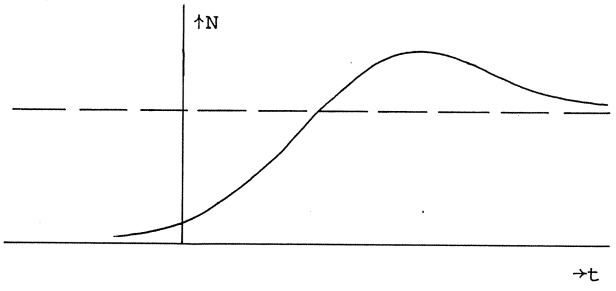



Figure 2

different from that of Figure 1. In order to explain this fact, V. VOLTERRA [60,p.47-56] introduced the following model for a population which produces toxic metabolic substances:

$$(1.5)\qquad \begin{cases} \frac{dN}{dt}(t) = [a - bN(t) - \int_0^t A(t-\tau)N(\tau)d\tau]\, N(t), & t > 0 \\ N(0) = N_0. \end{cases}$$

Here $A(\tau)$ describes the poisoning of the cultural medium at time $t = \tau$ due to the metabolic substances secreted by one bacterium at time $t = 0$. It is important to note that (1.5) corresponds to an initial value problem and

that for special solutions, like equilibria and periodic solutions, one should consider the autonomous (i.e., translation invariant) equation

$$(1.6) \qquad \frac{dN}{dt}(t) = [a - bN(t) - \int_{-\infty}^{t} A(t-\tau)N(\tau)d\tau]\, N(t), \qquad -\infty < t < \infty.$$

In the special case that

$$A(\tau) = A_0 e^{-\lambda\tau}$$

(1.5) is equivalent to a two dimensional system of ordinary differential equations (see CUSHING [9] or MACDONALD [41] for a more general treatment of this kind of equivalence) and one can verify that solutions exhibit the behaviour depicted in Figure 2, for suitable values of the parameters, by using phase plane analysis. For general A (say of class $L_1$) one has to resort to numerical analysis.

The equations (1.5) and (1.6) have stimulated much research. Typical questions that have been analyzed are the following:

- Under which conditions (on the parameters a and b and the function A) is it true that for any $N_0 > 0$ the solution N(t) of (1.5) satisfies

$$\lim_{t\to\infty} N(t) = N^*,$$

  where

$$N^* = a(b + \int_0^{\infty} A(\tau)d\tau)^{-1}$$

  is the positive constant solution of (1.6)?
- Under which conditions does the result above hold for all $N_0$ sufficiently near to $N^*$?
- Under which conditions does equation (1.6) have positive periodic solutions?

We refer to MILLER [49] and CUSHING [9] for a detailed account of these matters.

1.4. So far the models do not take into account the internal structure of the population in terms of age, weight, sex etc.. Clearly, for many populations (for instance the human) an important characteristic of the individual is its age, because fertility depends on age in a rather prominent way.

For this reason much effort has been put into the construction of a mathematical theory of age dependent population growth (A.J. LOTKA, ± 1920, being one of the main architects).

Let $n(t,a)$ be the age distribution of the population at time t. In other words,

$$\int_{a_1}^{a_2} n(t,a)da$$

gives the number of individuals with age between $a_1$ and $a_2$ at time t. First of all we assume that

$$n(t+h,a+h) - n(t,a) = -d(a)n(t,a)h + o(h)$$

for small h, or, by taking the limit $h \downarrow 0$,

$$\frac{\partial n}{\partial t} + \frac{\partial n}{\partial a} = -d(a)n. \tag{1.7}$$

Here the nonnegative function d is the age specific death rate which is assumed to be known. The equation (1.7) takes into account the process of aging and the probability of dying. The second assumption describes the birth process

$$n(t,0) = \int_0^\infty n(t,a)b(a)da\ . \tag{1.8}$$

Here $n(t,0)$ are the new-borns at time t and b is the age specific fertility which is assumed to be known. Moreover, in order to obtain a well-posed initial value problem, we suppose that the age distribution at time $t = 0$ is known:

$$n(0,a) = \phi(a). \tag{1.9}$$

At first sight the problem (1.7) - (1.9) seems to be out of place in a colloquium about integral equations. However, the first order partial differential equation (1.7) can be integrated along the characteristics $a = t +$ constant and then the boundary condition (1.8) leads to a well-known integral equation as we shall show.

The solution of (1.7) which satisfies the initial condition (1.9) is

given by

$$n(t,a) = \phi(a-t)\exp\Big(-\int_0^t d(\alpha+a-t)\,d\alpha\Big), \qquad t < a$$

$$n(t,a) = n(t-a,0)\exp\Big(-\int_0^a d(\alpha)\,d\alpha\Big), \qquad t > a. \tag{1.10}$$

Using the notation

$$B(t) := n(t,0),$$

$$K(a) := b(a)\,\exp\Big(-\int_0^a d(\alpha)\,d\alpha\Big),$$

$$f(t) := \int_0^\infty \phi(a)\exp\Big(-\int_0^t d(\alpha+a)\,d\alpha\Big)b(a+t)\,da,$$

(note that K(a) is the product of the fertility at age a and the chance that an arbitrary individual will reach the age a) and substituting (1.10) into (1.8) we obtain

$$B(t) = \int_0^t B(t-a)K(a)\,da + f(t). \tag{1.11}$$

This linear Volterra integral equation of convolution type is called the *renewal equation* (for obvious reasons) and it is well studied both because it arises from the model we are discussing right now and because it has many applications in the theory of industrial replacement. For a comprehensive discussion of (1.11) we refer to BELLMAN & COOKE [2], FELLER [23,24], HOPPENSTEADT [31], KEYFITZ [35] and LOPEZ [39].

Under suitable assumptions on K and f one can solve (1.11) by means of Laplace transformation. If we denote by $\bar{g}$ the Laplace transform of a function g, i.e.

$$\bar{g}(s) = \int_0^\infty e^{-st} g(t)\,dt,$$

then Laplace transformation of (1.11) leads to

$$\bar{B} = \bar{B}\bar{K} + \bar{f}$$

and hence to

$$\bar{B} = \frac{\bar{f}}{1-\bar{K}} .$$

Applying the inverse transformation we obtain

$$B(t) = \frac{1}{2\pi i} \int_L \frac{\bar{f}(s)}{1-\bar{K}(s)} e^{st}\, ds, \tag{1.12}$$

where L is a suitable line in the complex plane parallel to the imaginary axis.

Most of all we are interested in the asymptotic behaviour as $t \to \infty$. As is well-known this behaviour is determined by the poles of the integrand in (1.12) and hence by the roots of the characteristic equation

$$\bar{K}(s) = 1 , \tag{1.13}$$

which, in this context, is called *Lotka's equation*. One easily verifies that the nonnegativity of K implies that there is at most one real root, say $\sigma$, and that all other roots have real part less than $\sigma$. It follows that

$$B(t) \sim Ce^{\sigma t} \quad \text{as } t \to \infty,$$

where the constant C depends on f. Again $\sigma$ is called the intrinsic rate of natural increase or the Malthusian parameter.

Thus we have shown that the population will decay or grow exponentially according to the sign of $\sigma$, which is in fact equal to the sign of $\bar{K}(0) - 1$ (note that $\bar{K}$, as a function of a real variable, is monotone decreasing). The quantity

$$\bar{K}(0) = \int_0^\infty b(a) \exp\left(-\int_0^a d(\alpha)\, d\alpha\right) da$$

can be interpreted as the expected offspring of a new-born baby during his lifetime, and this interpretation makes our conclusion evident.

We have introduced the age structure in the model, but we did not yet draw any conclusion about the evolution of the age distribution in the course of time. Substitution of the asymptotic behaviour of B in the expression (1.10) for n(t,a) for $t > a$ leads to

$$n(t,a) \sim Ce^{\sigma t}\ (e^{-\sigma a}\exp(-\int_0^a d(\alpha)d\alpha)), \qquad t \to \infty.$$

Hence

$$\lim_{t\to\infty} \frac{n(t,a)e^{-\sigma t}}{C} = A(a)$$

where

$$A(a) := e^{-\sigma a}\exp(-\int_0^a d(\alpha)d\alpha). \tag{1.14}$$

So as time proceeds the age distribution gets a shape which is independent of the initial condition. For this reason the function A (which already, though in a somewhat different form, appears in a work of EULER [22] from 1760) is called the *stable age distribution*. Note that the function

$$n(t,a) = A(a)e^{-\sigma t}$$

is a solution of (1.7)-(1.8) which one can also obtain directly by applying separation of variables. The stable age distribution has been the subject of many demographical investigations and discussions (see COALE [6], KEYFITZ [35] and LOPEZ [39]).

1.5. As a next step it is natural to include both age - and density dependence in the model. If we assume that the death rate d and the fertility b depend on the total population size

$$N(t) := \int_0^\infty n(t,a)da,$$

we arrive at

$$\begin{cases} \frac{\partial n}{\partial t} + \frac{\partial n}{\partial a} = -d(a,N)n, \\ n(t,0) = \int_0^\infty n(t,a)b(a,N)da, \\ n(0,a) = \phi(a). \end{cases} \tag{1.15}$$

This problem has been studied by GURTIN & MACCAMY [28,29]. By integration along the characteristics they arrive at two coupled nonlinear Volterra

integral equations for B and N for which they prove the existence and uniqueness of solutions.

In an equilibrium situation N is constant and

$$(1.16) \qquad n = \bar{n}(a) = \frac{N \exp(-\int_0^a d(\alpha,N)d\alpha)}{\int_0^\infty \exp(-\int_0^\gamma d(\alpha,N)d\alpha)d\gamma} \quad .$$

Substitution of this expression into the boundary condition leads to a condition for N:

$$(1.17) \qquad \int_0^\infty b(a,N)\exp(-\int_0^a d(\alpha,N)d\alpha)da = 1.$$

Consequently, every root of (1.17) yields a steady state for which the corresponding age distribution is given by (1.16).

In [28] GURTIN & MACCAMY discuss the stability of equilibria. Linearization of the equation about the steady state followed by separation of variables leads to a linear eigenvalue problem. The eigenvalues turn out to be solutions of a transcendental equation. They show that the steady state is exponentially asymptotically stable if all roots of the transcendental equation lie in the left half plane.

It would be interesting to find conditions on b and d which guarantee that *every* solution of (1.15) approaches a steady state as $t \to \infty$, and to find complementary conditions which guarantee the existence of periodic solutions. In order to develop some feeling one can study special cases for which (1.15) is equivalent to a low dimensional system of ordinary differential equations. In [29] GURTIN & MACCAMY follow this approach. Another possibility would be to do some numerical experiments.

## 2. SELECTION AND MIGRATION

2.1. In the nucleus of the cells of many organisms one can recognize string-like structures called *chromosomes*. These are the carriers of genetic information. In fact there is a relation between certain properties of the individual and the composition of the chemicals it has at certain locations on its chromosomes. Such a location is called a *locus* and the corresponding collection of chemicals a *gene*. The variant forms of the gene are called *alleles*.

For many populations (for instance the human) chromosomes occur in

pairs and in the process of reproduction both parents contribute one chromosome of each pair to their offspring. In this section we shall consider such a diploid population. In order to make things as easy as possible we concentrate on one locus, assuming that there are only two alleles, A and a. Hence there are three possible *genotypes*: AA, Aa and aa (note that the chromosomes of one pair are indistinguishable from each other so that Aa and aA are identical). The types AA and aa are called *homozygotes* and Aa is called *heterozygote*.

Using Mendel's law of inheritance and various other assumptions we shall derive equations describing the propagation of the alleles through the successive generations. Again our approach will be to gradually make the models more complicated by allowing the population to have additional structure and again we shall concentrate on the asymptotic behaviour as time goes to infinity. For more background information we refer to CROW & KIMURA [8], HOPPENSTEADT [31] and NAGYLAKI [52], and to the general references given there (notably to the pioneering works of Haldane, Fisher and Wright).

## 2.2. Random mating and the Hardy-Weinberg equilibrium.

Let p, 2q and r denote the frequencies of respectively, the genotypes AA, Aa and aa in a certain generation (hence $p + 2q + r = 1$). Then, according to Mendel's laws of inheritance, the frequencies in the next generation can be read of from the following table:

| Type of mating | Frequency of this type of mating | Frequency of offspring of genotype | | |
|---|---|---|---|---|
| | | AA | Aa | aa |
| AA × AA | $p^2$ | $p^2$ | 0 | 0 |
| AA × Aa | $4pq$ | $2pq$ | $2pq$ | 0 |
| AA × aa | $2rp$ | 0 | $2rp$ | 0 |
| Aa × Aa | $4q^2$ | $q^2$ | $2q^2$ | $q^2$ |
| Aa × aa | $4qr$ | 0 | $2qr$ | $2qr$ |
| aa × aa | $r^2$ | 0 | 0 | $r^2$ |
| Sum | $(p+2q+r)^2 = 1$ | $(p+q)^2$ | $2(p+q)(q+r)$ | $(q+r)^2$ |

Here we take this law as an assumption but we remark that it can be derived as the deterministic limit, as the population size goes to infinity, of a stochastic model in which one assumes that the matings occur without regard to the genotype in question.

If we assume that the relevant characteristics of the life of an individual are independent of its genotype and that the generations are synchronized and non-overlapping, the dynamics are described by the difference equations

$$p_{n+1} = (p_n+q_n)^2,$$
$$q_{n+1} = (p_n+q_n)(q_n+r_n),$$
$$r_{n+1} = (q_n+r_n)^2,$$

where the subscript n counts the generations. Let $u_n$ denote the frequency of allele A in the n-th generation, i.e. $u_n := p_n + q_n$, then the equations can be written as

$$p_{n+1} = u_n^2, \quad q_{n+1} = u_n(1-u_n), \quad r_{n+1} = (1-u_n)^2.$$

Moreover,

$$u_{n+1} = p_{n+1} + q_{n+1} = u_n^2 + u_n(1-u_n) = u_n,$$

and consequently the system reaches a steady state in one step! The equilibrium state in which $p = u_0^2$, $q = u_0(1-u_0)$ and $r = (1-u_0)^2$ for some $u_0 \in [0,1]$ is called a Hardy-Weinberg equilibrium.

## 2.3. Selection

Now, in contrast with one of the assumptions above, we assume that the chance that a new-born individual reaches the age at which it becomes sexually mature depends on its genotype.

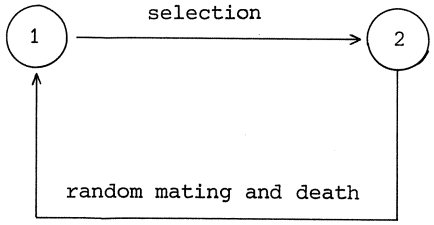



Figure 3: Life cycle

In particular we assume that this chance is $\alpha,\beta$ or $\gamma$ for the genotype AA, Aa or aa. Let $u_n$ denote the frequency of allele A in the n-th generation in stage 1, i.e. just after its birth. It is left as an exercise for the reader to verify that our assumptions lead to the difference equation

$$u_{n+1} = g(u_n) \tag{2.1}$$

where

$$g(u) = \frac{\alpha u^2+\beta u(1-u)}{\alpha u^2+2\beta u(1-u)+\gamma(1-u)^2}\,. \tag{2.2}$$

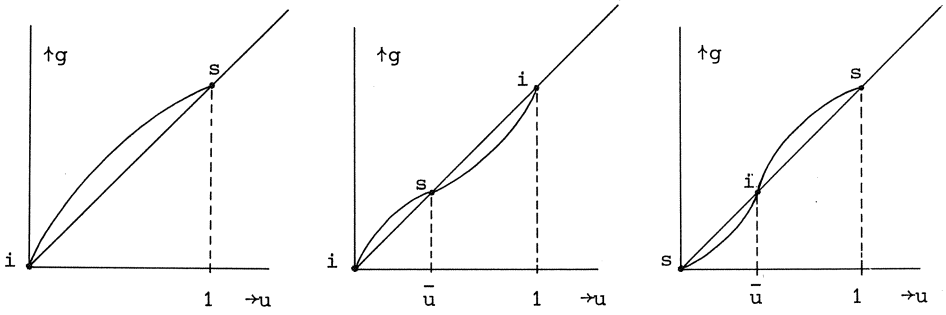
The relevant properties of the function g are:

$$g(0) = 0, \quad g(1) = 1,$$

$$g'(u) > 0 \quad \text{for} \quad 0 \le u \le 1,$$

(use the fact that $\alpha,\beta,\gamma \in (0,1]$). Hence if $u_0 \in [0,1]$ then $u_n \in [0,1]$ for all n as it should be!

If we choose the names A and a such that $\alpha \ge \gamma$, we can distinguish three qualitatively different cases:




I. $\alpha \geq \beta \geq \gamma$, $\alpha > \gamma$
heterozygote
intermediate

II. $\beta > \alpha \geq \gamma$
heterozygote
superior

III. $\alpha \geq \gamma > \beta$
heterozygote
inferior

s: stable i: unstable

Figure 4

In all cases (2.1) has the (trivial) equilibria u = 0 and u = 1 which correspond to the situation that one of the two alleles is absent in the population. In case I these are the only equilibria and u = 1 is globally asymptotically stable: for any $u_0 \in (0,1]$, $\lim_{n\to\infty} u_n = 1$. This means that the allele a is doomed to disappear. In case II there is a third equilibrium $\bar{u}$ which is globally asymptotically stable (note that the stability is determined completely by the sign of $g'(\bar{u}) - 1$). So in this case both alleles remain present for all time. In case III the equilibrium $\bar{u}$ is unstable and it depends on the initial condition which of the alleles will ultimately be the only one: if $u_0 \in (\bar{u},1]$ then $\lim_{n\to\infty} u_n = 1$, but if $u_0 \in [0,\bar{u})$ then $\lim_{n\to\infty} u_n = 0$.

## 2.4. Migration

Next, consider a population distributed over a certain territory (its "habitat"). One can easily imagine a situation where a selective advantage at some point becomes a selective disadvantage at others (for instance, think of the colour of the soil in relation to the protective colouring of some sort of animal, differences being due to the gene). According to SLATKIN [61] "Spatial variation in the intensity of natural selection can

play an important part in determining the genetic structure of natural populations". In this section we are going to study the effect of the joint action of selection and migration. However, we shall include one more stage in the life cycle of an individual, namely a process of regulation during which the total population density at each point is reduced to a fixed quantity, the carrying capacity. Although one can, to a large extent, justify this assumption by biological arguments, we admit that the desire to end with a tractable equation is an important motive too.

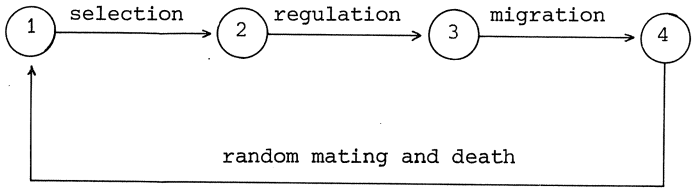



Figure 5: Life cycle

The symbols $u_n$, $\alpha,\beta$ and $\gamma$ have the same meaning as before, but now they are functions of the spatial variable x, which is supposed to vary in the domain $\Omega$. Let C(x) denote the carrying capacity at x and let b denote the number of descendants of one individual. We assume $b >> C(x)$.

The migration process is described through a given function $k = k(x,y)$ which has the interpretation that the chance that an individual which is at y during stage 3 shall be at x in stage 4 is k(x,y). On the other hand, let K(x,y) denote the chance that an individual which is at x during stage 4 descended from y in stage 3. If $\rho^3$ and $\rho^4$ denote the total population density in, respectively, stage 3 and 4 of one particular generation, then clearly

$$k(x,y)\rho^3(y) = K(x,y)\rho^4(x)$$

since both quantities describe what is at x and came from y. Moreover, by assumption,

$$\rho^4(x) = \int_\Omega k(x,y)\rho^3(y)\,dy.$$

Next we observe that the regulation process brings about that $\rho^3$, and hence $\rho^4$, are independent of the generation we are looking at. Consequently we can express K explicitly in term of k:

$$K(x,y) = -\frac{k(x,y)C(y)}{\int_\Omega k(x,\eta)C(\eta)\,d\eta} \quad . \tag{2.3}$$

We remark that

$$\int_\Omega K(x,y)\,dy = 1, \qquad \forall x \in \Omega, \tag{2.4}$$

as it should be, and that

$$\int_\Omega k(x,y)\,dx = 1, \qquad \forall y \in \Omega \tag{2.5}$$

if and only if no individuals are lost during the migration process (which might be unrealistic if we consider the wind borne dissemination of seeds of plants in an oasis).

Finally, as in (2.2) we define a function g, which now is a function of two variables,

$$g(x,u) = \frac{\alpha(x)u^2+\beta(x)u(1-u)}{\alpha(x)u^2+2\beta(x)u(1-u)+\gamma(x)(1-u)^2} \quad . \tag{2.6}$$

Skipping some details we derive the equation by following the frequency of the allele A through one cycle:

$$\underset{u_n(x)}{(1)} \longrightarrow \underset{g(x,u_n(x))}{(2)} \longrightarrow \underset{g(x,u_n(x))}{(3)} \longrightarrow \underset{\int_\Omega K(x,y)g(y,u_n(y))dy \;=:\; u_{n+1}(x)}{(4)} \longrightarrow \underset{u_{n+1}(x)}{(1)}$$

$$u_{n+1}(x) = \int_\Omega K(x,y)g(y,u_n(y))\,dy \quad . \tag{2.7}$$

We make several remarks:

1. The birth parameter b does not occur in the equation.
2. $g(x,0) = 0$ and $g(x,1) = 1$. Hence $u(x) \equiv 0$ and $u(x) \equiv 1$ are (trivial) equilibria.
3. A different order of the stages of the life cycle would have led to a different equation. These differences have biological interpretations. For instance, one can consider the situation where pollen is disseminated through insects or the wind and the situation where seeds are spread.
4. Under various assumptions one can "derive" from (2.7), through a certain limit procedure, the nonlinear diffusion equation

$$(2.8) \qquad \frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + f(x,u)$$

(see SLATKIN [61], NAGYLAKI [53], WEINBERGER [66] and FIFE [25]). This genetical interpretation has stimulated, beginning with the work of FISHER [27], much research in the field of nonlinear diffusion problems.

5. For a numerical analysis, equation (2.7) seems easier than equation (2.8).
6. Under suitable assumptions on K (i.e., on k and C) the right-hand side of (2.7) defines a mapping of $X = \{\phi \in C(\Omega) \mid 0 \leq \phi(x) \leq 1 \text{ for all } x \in \Omega\}$ into itself. Then for any initial condition $u_0 \in X$, the sequence $u_n$ is well-defined and $u_n$ belongs to X for all n.

There are lots of open problems in the qualitative analysis of (2.7). By looking at the corresponding literature for diffusion equations (see DIEKMANN & TEMME [18] and FIFE [25,26]) one can easily formulate conjectures and some of the proofs, in particular those depending on positivity and monotonicity arguments, can be imitated (see BOMINAAR [3]). Unfortunately the arguments based on the use of a Lyapunov functional break down.

But let us concentrate on some known results, in particular on the work of WEINBERGER [66]. He considers the case where $\alpha,\beta$ and $\gamma$ are constant, $\Omega = \mathbb{R}^n$ and $K(x,y) = A(x-y)$ for some radial symmetric, nonnegative, integrable function A, say of compact support. For the heterozygote intermediate case (see Figure 4.I) his most important result states that there exists a number $c_0$ such that:

(i) $\lim_{n\to\infty} \min_{|x|\leq nc} u_n(x) = 1$ for any $c < c_0$ and $u_0 \in X$, $u_0 \not\equiv 0$.

(ii) $\lim_{n\to\infty} \sup_{|x|\geq nc_0} u_n(x) = 0$ for any $u_0 \in X$ which has compact support.

In this situation one says that the equation exhibits the *hair-trigger effect*: no matter how little of the allele A is initially present in an arbitrarily small subset of $\mathbb{R}^n$, eventually the allele a is doomed to disappear everywhere. Moreover, the spread of A takes place with the *asymptotic speed* $c_0$.

In fact, $c_0$ is equal to the minimal speed for which travelling plane-wavefront solutions do exist. These are solutions of a fixed shape which, as time increases, are propagated with a fixed velocity in a fixed direction. So they are solutions of the form $u_n(x) = w(x.\nu+nc)$, where the unit vector $\nu$ defines the direction of the wave and the parameter c the speed. Substitution of this "ansatz" into (2.7) leads to a one-parameter family of convolution equations on the real line:

$$(2.9) \qquad w(\xi) = \int_{-\infty}^{\infty} g(w(\eta))A_c(\xi-\eta)d\eta$$

where

$$(2.10) \qquad A_c(\xi) := \int_{\mathbb{R}^{n-1}} A(\xi-c,x_2,\ldots,x_n)dx_2\ldots dx_n$$

(note that the equation does not depend on $\nu$ because of the radial symmetry of A).

The equation (2.9) has always, i.e. for all values of c, the constant solutions 0 and 1. We ask ourselves the question whether for some values of c the equation admits nonconstant solutions between 0 and 1? How can we characterize those values of c?

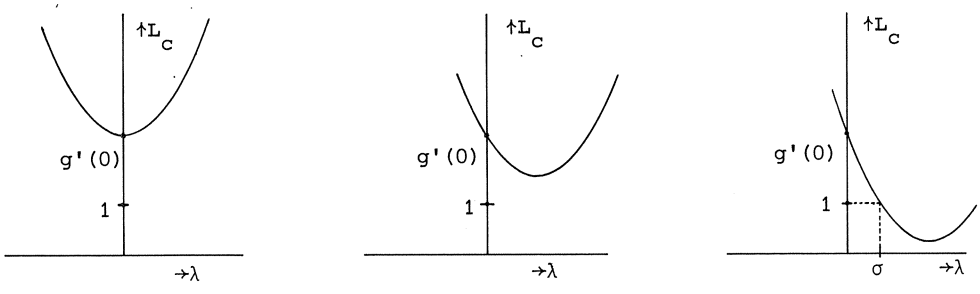
In order to answer this question one has to study the characteristic equation

$$(2.11) \qquad L_c(\lambda) = 1$$

where

$$(2.12) \qquad L_c(\lambda) = g'(0)\int_{-\infty}^{\infty} A_c(\eta)e^{-\lambda\eta}d\eta = g'(0)e^{-\lambda c}\int_{\mathbb{R}^n} e^{-\lambda x_1}A(x)dx.$$

This characteristic equation arises as the condition for the existence of an exponential solution of the equation obtained by linearizing (2.9) about zero, and one should compare its function with that of the characteristic polynomial in the case of an ordinary differential equation. The function $L_c(\lambda)$ is a monotone decreasing function of c and, on the real line, a convex function of $\lambda$.




$c = 0$ $\qquad$ $0 < c < c_0$ $\qquad$ $c > c_0$

Figure 6

Clearly the definition

$$(2.13) \qquad c_0 = \inf\{c > 0 \mid L_c(\lambda) = 1 \text{ for some } \lambda > 0\}$$

makes sense and $0 < c_0 < \infty$. In figure 6 we summarize the relevant knowledge about $L_c$.

It has been proved that for any $c \geq c_0$ there exists a nondecreasing solution w of (2.9) with $w(-\infty) = 0$ and $w(\infty) = 1$. For $c > c_0$, the basic idea of the proof in WEINBERGER [66] and DIEKMANN [12] is to use the information obtained from $L_c(\lambda)$ and the properties of g in the construction of two functions $\phi$ and $\psi$ such that $\phi \leq \psi$, $T\phi \geq \phi$, $T\psi \leq \psi$, where T denotes the formal integral operator that is associated with the right hand side of (2.9). Then the result follows from the monotonicity of T. For $c = c_0$, either one can follow the same procedure, but the construction is a little bit more complicated (see [66]), or one can resort on a limiting argument which shows that the set of speeds is closed (see BROWN & CARR [4]).

The characterization of the set of speeds is completed by the complementary result that for $0 \leq c < c_0$ there are no nonconstant solutions of (2.9) with $0 \leq w(\xi) \leq 1$. This is a corollary of Weinberger's result about the asymptotic speed, but here we sketch a different proof. Using Pitt's form of Wiener's general Tauberian theorem one can deduce that an arbitrary solution w of (2.9) with $0 \leq w(\xi) \leq 1$, has to decrease exponentially to zero as $\xi \to -\infty$. Furthermore, by using Laplace transformation, one can show that the exponent has to be a real root of the characteristic equation (2.11). Consequently, the nonexistence of such roots implies the nonexistence of this type of solution of (2.9) (see DIEKMANN & KAPER [17] for the details).

The advantage of this approach is that the same method is suitable for obtaining results about uniqueness modulo translation. For $c > c_0$, let $\sigma$ denote the smallest real root of (2.11). One can show that for every solution w of (2.9) between 0 and 1 there exists a constant C such that

$$w(\xi) \sim Ce^{\sigma\xi}, \qquad \xi \to -\infty.$$

If $w_1$ and $w_2$ are two such solutions we define a function v by

$$v(\xi) = (w_1(\xi - \sigma^{-1}\ln\frac{c_1}{c_2}) - w_2(\xi))e^{-\sigma\xi}.$$

From the fact that $w_1$ and $w_2$ satisfy (2.9) we deduce that

$$|v(\xi)| \leq g'(0) \int_{-\infty}^{\infty} A_c(\eta) e^{-\sigma\eta} |v(\xi-\eta)| d\eta.$$

Since moreover $v(-\infty) = v(\infty) = 0$, one can show that this inequality implies that $v(\xi) = 0$ for all $\xi$ (see [17]) and we have obtained that for fixed $c > c_0$, equation (2.9) admits, modulo translation, one and only one non-constant solution between 0 and 1. (We remark that BARBOUR [1] has treated the uniqueness problem by using probabilistic arguments.)

So we have seen that the global asymptotic stability of $u = 1$ in the heterozygote intermediate case carries over to the migration model in a strikingly strong sense, and we have found some new phenomena in the transient behaviour, that is in the manner in which solutions evolve towards the asymptotic state.

If the coefficients $\alpha, \beta$ and $\gamma$ do depend on $x$ one can, in certain situations, have monotonic (nonconstant) *stable* equilibria, which in this context are called clines (for instance, think of a gradually change, with respect to the distribution of genotypes, of the vegetation on a mountain-slope). This has been thoroughly investigated for the diffusion equation (2.8). (see NAGYLAKI [52,53,38], SLATKIN [61], PELETIER [57] and the references given there). For the integro-difference equation (2.7) one can obtain similar results (unpublished work of the author; also see DOWNHAM & SHAW [19,20,21]) but so far a detailed and careful working out is missing.

## 3. EPIDEMIC MODELS

3.1. In 1927, KERMACK & MCKENDRICK [33] published a by now classical paper about a deterministic model for the spread (as a function of time) of an infectious disease in a closed population. Strange enough many people seem unaware of the fact that the model, in its full generality, leads to a Volterra integral equation. Instead they speak and write only about a special case in which the model leads to a system of ordinary differential equations. For that reason we shall paraphrase part of their article.

Consider a population divided into two classes S and I. We shall denote the number of individuals belonging to a certain class with the same character and we shall consider these numbers as "smooth" real-valued functions of the time t (an assumption which clearly is unrealistic if we

consider a small population of, say, ten individuals).

The class S consists of those who are susceptible to a certain infectious disease and the class I of those who were infected. We provide the class I with an internal structure by keeping track of the class-age. This means that we distinguish the members of I according to the time elapsed since they were infected. So let $i(t,\tau)$ denote the density, at time t, of those members of I which have class-age $\tau$, then

$$I(t) = \int_0^\infty i(t,\tau)d\tau.$$

We assume that:

(i) the disease induces permanent immunity, so that the transition from I to S does not occur;

(ii) all changes are due to infection ($S(t)+I(t)$ = constant, the population is demographically closed);

(iii) a nonnegative function A is given such that

$$\dot{S}(t) = -S(t)\int_0^\infty i(t,\tau)A(\tau)d\tau. \tag{3.1}$$

With respect to (iii) we remark that the interaction between the classes S and I is of "mass action" type, but with the factor I replaced by a weighted average over the age-structured density. The function A is characteristic for the disease and a typical example is sketched in Figure 7

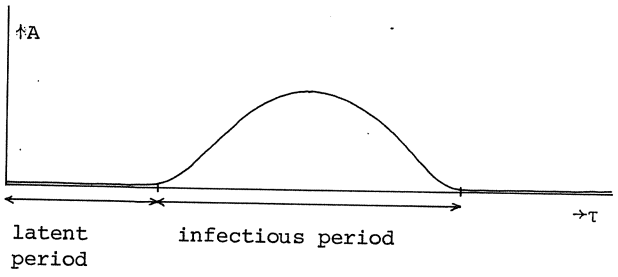



Figure 7

On account of these assumptions we have $\frac{\partial i}{\partial t} + \frac{\partial i}{\partial \tau} = 0$.

Hence

$$i(t,\tau) = i(t-\tau,0), \qquad t > \tau,$$

$$i(t,\tau) = i(0,\tau-t), \qquad t < \tau,$$

and

$$\dot{S}(t) = -\dot{I}(t) = -i(t,0).$$

In order to complete the model, we assume that all these relations hold for $t > 0$ and that on $t = 0$ we possess all the necessary information to derive a well-posed problem. This amount to putting

$$S(0) = S_0,$$

$$i(0,\tau) = i_0(\tau), \qquad 0 \le \tau < \infty,$$

Where $S_0$ is a given number and $i_0$ a given function. We rewrite (3.1):

$$\dot{S}(t) = -S(t)\left\{\int_0^t i(t-\tau,0)A(\tau)d\tau + \int_t^\infty i(0,\tau-t)A(\tau)d\tau\right\}$$

$$= S(t)\left\{\int_0^t \dot{S}(t-\tau)A(\tau)d\tau - \int_0^\infty i_0(\tau)A(t+\tau)d\tau\right\}.$$

Dividing the equation by $S(t)$ and integrating from 0 to t we obtain

$$\ln\frac{S(t)}{S_0} = \int_0^t (S(t-\tau)-S_0)A(\tau)d\tau - f(t),$$

where

$$f(t) = \int_0^t\int_0^\infty i_0(\tau)A(s+\tau)d\tau ds.$$

Thus, the function f is monotone nondecreasing. Introduction of a new dependent variable

$$u(t) = -\ln\frac{S(t)}{S_0}$$

leads to an equation in the standard form of nonlinear Volterra equations of convolution type (or, nonlinear renewal equations):

$$u(t) = S_0 \int_0^t g(u(t-\tau))A(\tau)d\tau + f(t), \tag{3.2}$$

where

$$g(u) = 1 - e^{-u}.$$

So, starting from assumptions about the interaction between the different groups, we have derived one equation for one dependent variable u, which measures the number of susceptibles. This is possible because in this model the number of newly infected equals the decrease of the number of susceptibles. From a mathematical point of view we do not have to keep books of the infected.

Under fairly weak conditions on A and f one can prove the existence and uniqueness of a solution of equation (3.2) (see MILLER [48]). The solution inherits nonnegativity from A and f and likewise monotonicity from f (in view of the biological interpretation these properties are necessary; in this way the monotonicity reflects our assumptions that the population is demographically closed and that the disease induces permanent immunity). Finally, assuming that

$$\gamma := \int_0^\infty A(\tau)d\tau < \infty$$

and

$$f(\infty) < \infty,$$

one can prove that the solution is bounded and consequently that $u(\infty)$ exists and satisfies the scalar equation

$$u(\infty) = \gamma S_0\, g(u(\infty)) + f(\infty). \tag{3.3}$$

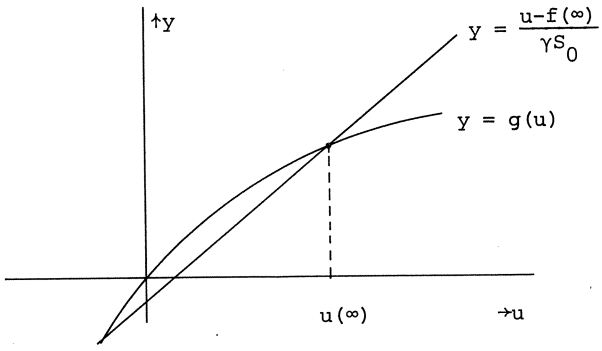



Figure 8

The drawing in Figure 8 should convince the reader that equation (3.3) has, for each positive $f(\infty)$, a unique positive solution $u(\infty)$. From the fact that $u(\infty) < \infty$, Kermack and McKendrick deduce the following conclusion

> "An epidemic, in general, comes to an end, before the susceptible population has been exhausted".

Subsequently they ask themselves what will be the consequence of the introduction of a very small number of infectives in a up to then uninfected population. Mathematically, this amounts to the question: what happens to $u(\infty)$ if $f(\infty)$ converges to zero?

If $f(\infty)$ approaches zero, the straight line $y = (\gamma S_0)^{-1}(u-f(\infty))$ in Figure 8 shifts upwards parallel to itself untill it passes through the origin. So we see that $u(\infty)$ is strictly increasing as a function of $f(\infty)$. Moreover this geometrical argument shows that the ratio between the slope $(\gamma S_0)^{-1}$ of the line and the slope 1 of the tangent line to the curve at zero determines whether $u(\infty)$ converges to zero or not. If $\gamma S_0 > 1$ the scalar equation

$$y = \gamma S_0 g(y)$$

has a positive root, which we shall call p, and

$$\lim_{f(\infty)\downarrow 0} u(\infty) = p.$$

If, on the other hand, $\gamma S_0 \leq 1$ then

$$\lim_{f(\infty)\downarrow 0} u(\infty) = 0,$$

and if $\gamma S_0 < 1$ we have the even stronger property

$$u(\infty) \leq (1-\gamma S_0)^{-1} f(\infty).$$

So we conclude that the qualitative behaviour of the solution depends crucially on the value of the parameter $\gamma S_0$. For this reason one says that the parameter has a *threshold value* one, and the remarks above form an informal mathematical formulation of Kermack and McKendrick's celebrated *Threshold Theorem*. Another formulation, with a biological flavour, is the following: if $\gamma S_0 > 1$, the fraction of the susceptible population which escapes from being infected is bounded from above (namely $\frac{S(\infty)}{S_0} < e^{-p}$) *independently* of the number of introduced infectives, whereas, on the contrary, for $\gamma S_0 < 1$ such a bound does not exist. For still other formulations we refer to DIEKMANN [11], HOPPENSTEADT [30,31], METZ [47] and REDDINGIUS [58].

There is a simple biological interpretation of the parameter $\gamma S_0$ which makes this result easy understandable. One can view $\gamma S_0$ as the number of those who will be infected by one freshly infected individual which is introduced in the population (see HOPPENSTEADT [30]). If this number exceeds one, an avalanche occurs, if it is less than one things blow over.

## 3.2. Spread in space and time

One can incorporate space dependence into the model by assuming that individuals, although they stay at their position, exert their infective influence in a certain neighbourhood (for instance, think of a cornfield where a disease spreads like a forest-fire). This leads to the equation

$$u(t,x) = u_0(t,x) + \int_0^t A(\tau) \int_{\mathbb{R}^n} g(u(t-\tau,x-\xi))V(\xi)\,d\xi\,d\tau, \tag{3.4}$$

where V is a given radial symmetric function (see [12]). For this equation results have been proved which are completely analogous to those we have discussed in connection with the heterozygote intermediate model: the equation exhibits the hair-trigger effect and there exists an asymptotic speed

of propagation which equals the minimal speed for travelling wavefronts (see [12,13,14] and, for a survey in the style of this chapter see [15]; also see the interesting review of MOLLISON [50]).

### 3.3. Limited immunity

One can as well generalize the model of section 3.1 in a different direction, namely by dropping the assumption of permanent immunity (see [34] for this and other generalizations). In [37] Lauwerier posed the question whether this could open the way to periodic solutions describing, to some extent, the phenomenon of periodic outbursts.

So let us assume that the support of A is contained in $[0,\tau_1]$ and that individuals retake their susceptibility exactly $\tau_2$ units of time after

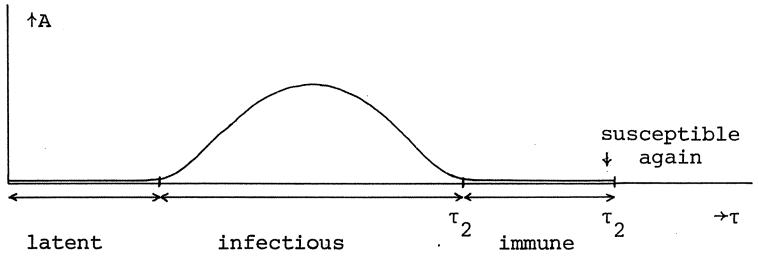



Figure 9

exposure (for some $\tau_1$ and $\tau_2$ with $\tau_2 > \tau_1 > 0$). As before we obtain

$$i(t,0) = S(t) \int_0^{\tau_1} i(t-\tau,0)A(\tau)d\tau .$$

The fact that the total population remains constant is expressed by the relation

$$S(t) + \int_0^{\tau_2} i(t-\tau,0)d\tau = C .$$

Elimination of S leads to an equation for i(t,0) alone, which after some straightforward changes of variables can be written as

$$(3.5) \qquad y(t) = \gamma\left(1 - \int_{t-1}^{t} y(\tau)d\tau\right) \int_{t-1}^{t} y(\tau)B(t-\tau)d\tau .$$

Here B is a nonnegative and integrable function such that

$$\int_{0}^{1} B(\tau)d\tau = 1$$

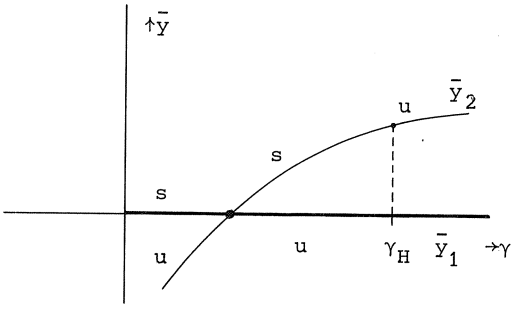
and such that supp $B \subset [0,1)$. The parameter $\gamma$ measures, as before, the total population size and the infective strength of the disease.

Let us first concentrate on steady state solutions $y(t) = \bar{y}$ of equation (3.5). These are found from the roots of the scalar equation

$$(3.6) \qquad \bar{y} = \gamma(1-\bar{y})\bar{y} ,$$

and consequently

$$\bar{y}_1 = 0, \qquad \bar{y}_2 = \frac{\gamma-1}{\gamma}$$




s: stable
u: unstable

Figure 10

If $\gamma < 1$, the state $y = 0$, in which the disease is absent, is globally exponentially stable (here stability has to be defined in terms of solutions of some appropriate initial value problem; see [16] for an extensive discussion of these matters). If $\gamma$ passes through one bifurcation and exchange of stability take place. The state $y = \bar{y}_2$, in which the disease is endemic, becomes positive (and hence biologically significant) and it takes over the stability from $y = 0$.

So if $\gamma$ increases beyond one the state $\bar{y}_2$ is initially stable. Does it

retain its stability?
Once again we are led to consider a characteristic equation

$$1 = \bar{B}(\lambda) + (1-\gamma)\,\frac{1-e^{-\lambda}}{\lambda}, \tag{3.7}$$

which arises as the condition for the existence of exponential solutions for the equation obtained by linearizing (3.5) about $\bar{y}_2$.

Numerical experiments and some analytical arguments (see MONTIJN [51]) indicate that for many (though not all) functions B a number $\gamma_H(B)$ exists such that as $\gamma$ passes through $\gamma_H(B)$ a pair of complex conjugated roots of equation (3.7) crosses the imaginary axis. Consequently $\bar{y}_2$ looses its stability and a periodic solution arises (for the general theory of this phenomenon, a so-called Hopf-bifurcation, we refer to MARSDEN & MCCRACKEN [42]; CUSHING [10] is a reference pertaining to the present context; also see [16]). Analytically one can obtain some information about the periodic solution for $\gamma$ in a small neighbourhood of $\gamma_H(B)$. Moreover some abstract results are known about the global behaviour of the bifurcating branch of periodic solutions (see [5,55,56]). However, for such a concrete example it could be interesting to follow the periodic solution numerically as a function of $\gamma$.

A numerical study of equation (3.5) aiming at an understanding of the global dynamical behaviour might be an exciting but difficult enterprise. Most of all through the investigation of difference equations (see MAY [44]) one has recently realized that innocent looking nonlinear equations can exhibit very exotic dynamical behaviour (see KAPLAN & MAROTTO [32]). I conjecture that equation (3.5) falls into this category. As an indication in this direction one should note that, as $\gamma$ tends to minus infinity, the roots of (3.7) tend to the roots of $1-e^{-\lambda} = 0$, and hence to the infinitely many points $\lambda = \pm 2n\pi i$, $n = 1,2,3,\ldots$, on the imaginary axis. By analogy with the situation described in ROSENBLAT & DAVIS [59] one is led to believe that equation (3.5) has, for large values of $\gamma$, infinitely many periodic solutions which bifurcate from infinity. It is our plan to study these problems within a general dynamical system framework for equations like (3.5).

It is important to observe that equation (3.5) is autonomous (that is, translation invariant). Recently much attention has been paid to epidemic models where seasonal variations are included in the equations as an explicit periodic time dependence. Although this theory leads to many

interesting results about nonlinear integral equations, we shall not comment on these matters here. Rather we refer to the pertinent literature, a representative choice being COOKE & KAPLAN [7], SMITH [62,63] and NUSSBAUM [54,55,56].

EPILOGUE. Not all of the questions presented above have "real" biological significance and most of the models are rather naive. However, it has been my intention to demonstrate that the interplay of building models and analyzing equations can be seen as a dynamic process which leads to the enrichment of both biological and mathematical theory.

REFERENCES

[1] BARBOUR, A.D., *The uniqueness of Atkinson and Reuter's epidemic waves*, Math. Proc. Camb. Phil. Soc. 82 (1977) 127-130.

[2] BELLMANN, R. & K. COOKE, *Differential-Difference Equations*, Academic Press, New York, 1963.

[3] BOMINAAR, S., Doctoral Thesis Univ. of Amsterdam (in Dutch), to appear.

[4] BROWN, K.J. & J. CARR, *Deterministic epidemic waves of critical velocity*, Math. Proc. Camb. Phil. Soc. 81 (1977) 431-443.

[5] CHOW, S-N. & J. MALLET-PARET, *The Fuller index and global Hopf bifurcation*, J. Diff. Equ. 29 (1978) 66-85.

[6] COALE, A.J., *The Growth and Structure of Human Populations, a Mathematical Investigation*, Princeton University Press, Princeton, 1972.

[7] COOKE, K.L. & J.L. KAPLAN, *A periodicity threshold theorem for epidemics and population growth*, Math. Biosc. 31 (1976) 87-104.

[8] CROW, J.F. & M. KIMURA, *An Introduction to Population Genetics Theory*, Harper & Row, New York, 1970.

[9] CUSHING, J.M., *Integrodifferential Equations and Delay Models in Population Dynamics*, Lect. Notes in Biomath. 20, Springer, Berlin, 1977.

[10] CUSHING, J.M., *Bifurcation of periodic solutions of Volterra integral equations*, To appear in: Proc. Helsinki Symposium on Integral Equations.

[11] DIEKMANN, O., *Limiting behaviour in an epidemic model*, Nonl. Anal., Th., Meth. & Appl. 1 (1977) 459-470.

[12] DIEKMANN, O., *Thresholds and travelling waves for the geographical spread of infection*, J. Math. Biol. 6 (1978) 109-130.

[13] DIEKMANN, O., *On a nonlinear integral equation arising in mathematical epidemiology*, In: Differential Equations and Applications (W. Eckhaus, E.M. de Jager, eds.), 133-140, North-Holland, Amsterdam, 1978.

[14] DIEKMANN, O., *Run for your life. A note on the asymptotic speed of propagation of an epidemic*. To appear in J. Diff. Equ.

[15] DIEKMANN, O., *A nonlinear integral equation describing the geographical spread of infection: the hair-trigger effect, travelling waves and the asymptotic speed of propagation*, Report III, 139, Istituto per le Applicazioni del Calcolo "Mauro Picone", Roma, 1978.

[16] DIEKMANN, O., *Volterra integral equations and semigroups of operators*, in preparation.

[17] DIEKMANN, O. & H.G. KAPER, *On the bounded solutions of a nonlinear convolution equation*, Nonl. Anal., Th., Meth. & Appl. 2 (1978) 721-737.

[18] DIEKMANN, O. & N.M. TEMME (eds.), *Nonlinear Diffusion Problems*, MC Syllabus 28, Math. Centrum, Amsterdam, 1976.

[19] DOWNHAM, D.Y. & S.M.M. SHAW, *An integral equation approach for models of clines*, Adv. Appl. Prob. 8 (1976) 625-628.

[20] DOWNHAM, D.Y. & S.M.M. SHAW, *Sufficiency conditions for the stability of an equilibrium*, Adv. Appl. Prob. 8 (1976) 4-7.

[21] DOWNHAM, D.Y. & S.M.M. SHAW, *Models of clines*, Adv. Appl. Prob. 6 (1974) 7-10.

[22] EULER, L., *Recherches générales sur la mortalité et la multiplication*, Mémoires de l'Académie Royale des Sciences et Belles Lettres XVI (1760) 144-164.

[23] FELLER, W., *On the integral equation of renewal theory*, Ann. Math. Statist. 12 (1941) 243-267. (reprinted in [64]).

[24] FELLER, W., *An Introduction to Probability Theory and Its Applications*, Vol. II, Wiley, New York, 1966.

[25] FIFE, P.C., *Mathematical Aspects of Reacting and Diffusing Systems*, Lect. Notes in Biomath. 28, Springer, Berlin, 1979.

[26] FIFE, P.C., *Asymptotic states for equations of reaction and diffusion*, Bull. Amer. Math. Soc. 84 (1978) 693-726.

[27] FISHER, R.A., *The wave of advance of advantageous genes*, Ann. of Eugenics 7 (1937) 355-369.

[28] GURTIN, M.E. & R.C. MACCAMY, *Nonlinear age-dependent population dynamics*, Arch. Rat. Mech. Anal. 54 (1974) 281-300.

[29] GURTIN, M.E. & R.C. MACCAMY, *Population dynamics with age dependence*, In: Nonlinear analysis and mechanics: Heriot-Watt Symposium, Vol. III (R.J. Knops, ed.), Research Notes in Math. 30 , 1-35, Pitman, London, 1979.

[30] HOPPENSTEADT, F., *Thresholds for deterministic epidemics*, In: Mathematical Problems in Biology, Victoria Conference 1973 (P. van den Driessche, ed.) Lect. Notes in Biomath. 2, 96-101, Springer, Berlin, 1974.

[31] HOPPENSTEADT, F., *Mathematical Theories of Populations: Demographics, Genetics and Epidemics*, SIAM Regional Conference Series in Applied Math. 20, SIAM, Philadelphia, 1975.

[32] KAPLAN, J.L. & F.R. MAROTTO, *Chaotic behavior in dynamical systems*, In: Nonlinear Systems and Applications (V. Lakshmikantham, ed.) 199-210, Academic Press, New York, 1977.

[33] KERMACK, W.O. & A.G. MCKENDRICK, *A contribution to the mathematical theory of epidemics*, Proc. Roy. Soc. A 115 (1927) 700-721.

[34] KERMACK, W.O. & A.G. MCKENDRICK, *Contributions to the mathematical theory of epidemics*. II - *The problem of endemicity*. III - *Further studies on the problem of endemicity*, Proc. Roy Soc. A 138 (1932) 55-83, 141 (1933) 94-122.

[35] KEYFITZ, N., *Introduction to the Mathematics of Population*, Addison-Wesley, Reading, 1968.

[36] KOSTITZIN, V.A., *Mathematical Biology*, Harrap, London, 1939.

[37] LAUWERIER, H.A., *Mathematische modellen voor epidemische processen*, Handout at the 12th Dutch Mathematical Congress, 1976.

[38] LEVIN, S.A., ed., *Studies in Mathematical Biology*, Part II: *Populations and communities*, Studies in Math. 16, Math. Ass. of America, 1978.

[39] LOPEZ, A., *Problems in Stable Population Theory*, Office of Population Research, Princeton, 1961.

[40] LOTKA, A.J., *Elements of Mathematical Biology*, Dover, New York, 1956.

[41] MACDONALD, N., *Time Lags in Biological Models*, Lect. Notes in Biomath. 27, Springer, Berlin, 1978.

[42] MARSDEN, J.E. & M. MACCRACKEN, *The Hopf Bifurcation and its Applications*, Springer, Berlin, 1976.

[43] MAY, R.M., *Stability and Complexity in Model Ecosystems*, Princeton Univ. Press, Princeton, 1973.

[44] MAY, R.M., *Simple mathematical models with very complicated dynamics*, Nature 261 (1976) 459-467.

[45] MAYNARD SMITH, J., *Mathematical Ideas in Biology*, Cambridge Univ. Press, Cambridge, 1968.

[46] MAYNARD SMITH, J., *Models in Ecology*, Cambridge Univ. Press, Cambridge, 1974.

[47] METZ, J.A.J., *The epidemic in a closed population with all susceptibles equally vulnerable; some results for large susceptible populations and small initial infections*, Acta Biotheoretica 27 (1978) 75-123.

[48] MILLER, R.K., *Nonlinear Volterra Integral Equations*, Benjamin, Menlo Park, 1971.

[49] MILLER, R.K., *On Volterra's population equation*, SIAM J. Appl. Math. 14 (1966) 446-452.

[50] MOLLISON, D., *Spatial contact models for ecological and epidemic spread*, J. Roy. Statist. Soc. B 39 (1977) 283-326.

[51] MONTIJN, R., Doctoral Thesis Univ. of Amsterdam (in Dutch).

[52] NAGYLAKI, T., *Selection in One- and Two-Locus Systems*, Lect. Notes in Biomath. 15, Springer, Berlin, 1977.

[53] NAGYLAKI, T., *Conditions for the existence of clines*, Genetics 80 (1975) 595-615.

[54] NUSSBAUM, R.D., *A periodicity threshold theorem for some nonlinear integral equations*, SIAM J. Math. Anal. 9 (1978) 356-376.

[55] NUSSBAUM, R.D., *Periodic solutions of some nonlinear integral equations*, In: Dynamical Systems (A.R. Bednarek & L. Cesari, eds.) 221-249, Academic Press, New York, 1977.

[56] NUSSBAUM, R.D., *Periodic solutions of some integral equations from the theory of epidemics*, In: Nonlinear Systems and Applications (V. Lakshmikantham, ed.) 235-255, Academic Press, New York, 1977.

[57] PELETIER, L.A., *The mathematical theory of clines*, Proceedings of the Bicentennial Congress of the Wiskundig Genootschap, Part II (P.C. Baayen, D. van Dulst, J. Oosterhoff, eds.) 295-308, MC Tract 101, Math. Centrum, Amsterdam, 1979.

[58] REDDINGIUS, J., *Notes on the mathematical theory of epidemics*, Acta Biotheoretica 20 (1971) 125-157.

[59] ROSENBLAT, S. & S.H. DAVIS, *Bifurcation from infinity*, Univ. of Melbourne Math. Research Report 6, 1978.

[60] SCUDO, F.M. & J.R. ZIEGLER, *The Golden Age of Theoretical Ecology*: 1923-1940, Lect. Notes in Biomath. 22, Springer, Berlin, 1978.

[61] SLATKIN, M., *Gene flow and selection in a cline*, Genetics 75 (1973) 733-756.

[62] SMITH, H.L., *On periodic solutions of a delay integral equation modelling epidemics*, J. Math. Biol. 4 (1977) 69-80.

[63] SMITH, H.L., *Periodic solutions for a class of epidemic equations*, J. Math. Anal. Appl. 64 (1978) 467-479.

[64] SMITH, D. & N. KEYFITZ, *Mathematical Demography* (Selected Papers) Biomathematics Vol. 6, Springer, Berlin, 1977.

[65] VOLTERRA, V., *Leçons sur la Théorie Mathématique de la Lutte pour la Vie*, Gauthier-Villars, Paris, 1931.

[66] WEINBERGER, H.F., *Asymptotic behavior of a model in population genetics*, In: Nonlinear Partial Differential Equations and Applications (J.M. Chadam, ed.) 47-96, Lect. Notes in Math. 648, Springer, Berlin, 1978.

# OVER DE INTEGRAALVERGELIJKING VAN DE DRAGENDE VLAK THEORIE

P.J. ZANDBERGEN

## 1. INLEIDING

Gedurende meer dan dertig jaar heeft in de aerodynamica de beschouwing van de z.g. dragende vlak theorie een belangrijke rol gespeeld. Ook in ons land is er een redelijke ervaring aanwezig met de behandeling van deze vergelijking en alhoewel de verdere ontwikkelingen van de hier te beschrijven methodes vrijwel tot een eind zijn gekomen, leek het toch goed, in het kader van dit colloquium, aandacht te besteden zowel aan de vergelijking als aan zijn afleiding en zijn oplossing.

We zullen de afleiding presenteren voor een incompressibele vloeistof, die om een vleugelachtig oppervlak stroomt.

Voor $x \to -\infty$ nemen we aan dat er een uniforme snelheid U is in de richting van de positieve x-as. Verder zullen we ervan uitgaan dat de door het oppervlak veroorzaakte stoorsnelheden u,v,w in respectievelijk de x,y en z richting klein zijn t.o.v. U (zie figuur 1).

De algemene vergelijkingen van de vloeistofdynamica zijn

$$\frac{\partial \rho}{\partial t} + \operatorname{div}(\rho \underline{v}) = 0 \tag{1.1}$$

en

$$\rho \frac{d\underline{v}}{dt} = -\operatorname{grad} p \tag{1.2}$$

waarbij $\rho$ de dichtheid is, $\underline{v}$ de snelheidsvector en p de druk.

Als $\rho$ constant is en de stroming stationair geldt

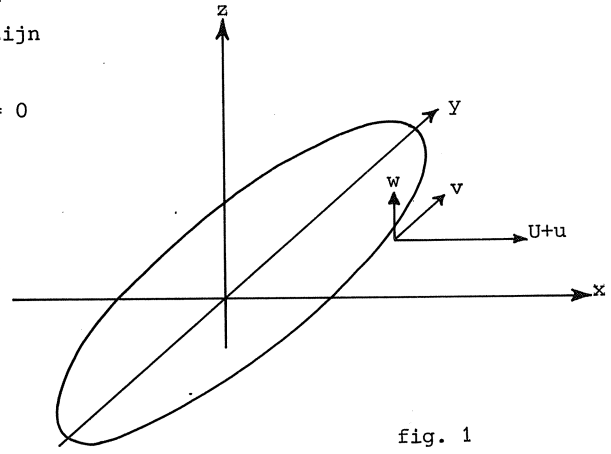



fig. 1

$$\operatorname{div} \underline{v} = 0, \tag{1.3}$$

$$(\underline{v}.\ \operatorname{grad})\ \underline{v} = -\frac{1}{\rho} \operatorname{grad} p. \tag{1.4}$$

Passen we op deze vergelijking de rot. operator toe, dan volgt

$\operatorname{rot}(\underline{v}.\operatorname{grad})\ \underline{v} = (\underline{v}.\operatorname{grad})\ \operatorname{rot}\ \underline{v} + \operatorname{div}\ \underline{v}\ \operatorname{rot}\ \underline{v} - (\operatorname{rot}\ \underline{v}.\operatorname{grad})\ \underline{v} = 0.$

Het is duidelijk dat $\operatorname{rot}\ \underline{v} = 0$ aan deze vergelijking voldoet. In dat geval vinden we dan

$$\underline{v} = \operatorname{grad} \phi \tag{1.5}$$

en op grond van de vergelijking (1.3) volgt er nu

$$(1.6) \qquad \text{div grad } \phi = \Delta\phi = \frac{\partial^2\phi}{\partial x^2} + \frac{\partial^2\phi}{\partial y^2} + \frac{\partial^2\phi}{\partial z^2} = 0.$$

Dit is de vergelijking voor de zogenaamde snelheidspotentiaal.
We kunnen ook een afleiding geven in termen van de zogenaamde versnellingspotentiaal $\psi$. Immers uit (1.2) volgt voor $\rho$ is constant

$$(1.7) \qquad \frac{d\underline{v}}{dt} = -\text{ grad } \frac{p}{\rho} \overset{\text{def}}{=} \text{grad } \psi.$$

Als we nu stellen dat voor $x \to -\infty$, $p \to p^\infty$ en $\psi \to 0$ dan volgt

$$(1.8) \qquad \psi = \frac{p^\infty - p}{\rho}.$$

Uit (1.4) volgt dan direkt dat

$$(1.9) \qquad \text{grad } (\frac{1}{2}\,\underline{v}.\underline{v} + \frac{p}{\rho}) = 0.$$

Dit betekent dus dat de uitdrukking tussen haakjes constant is in het gehele stromingsveld en er moet dus gelden

$$\frac{1}{2}\,\underline{v}.\underline{v} + \frac{p}{\rho} = \frac{1}{2}\,U^2 + \frac{p^\infty}{\rho}.$$

Voeren we in deze vergelijking nu in dat u,v en w klein zijn t.o.v. U dan volgt

$$\psi = Uu = U(\frac{\partial\phi}{\partial x} - U).$$

Als we in deze vergelijking nu de zogenaamde stoorsnelheidspotentiaal $\bar{\phi} = \phi - Ux$ invoeren die eveneens aan de Laplace vergelijking voldoet, dan krijgen we als resultaat

$$(1.10) \qquad \psi = U\,\frac{\partial\bar{\phi}}{\partial x}.$$

Het is nu direkt duidelijk dat ook $\psi$ aan de Laplace vergelijking moet voldoen, dus er geldt

$$(1.11) \qquad \Delta\psi = 0.$$

Deze vergelijking hebben we afgeleid, omdat het bij een draagvlak vooral gaat om de draagkracht die het resultaat is van de verschillen in druk tussen onder- en bovenkant van het draagvlak.

Zoals uit (1.8) volgt is $\psi$ hiervoor een geschikte grootheid omdat de druk daarmee rechtstreeks bepaald wordt.

Geschikte differentiaalvergelijkingen zijn nu verkregen. Nu moeten nog de randvoorwaarden worden afgeleid. Het gegeven oppervlak $F(x,y,z) = 0$ zal in het algemeen ondoordringbaar zijn voor de stroming, dat wil zeggen: de normaalsnelheid zal nul moeten zijn. Dit betekent dus als $\underline{n}$ een normaal is op het oppervlak dat

$$\underline{v}.\underline{n} = 0 \tag{1.12}$$

ofwel $(U + u)F_x + v F_y + w F_z = 0$.
Als nu de normaalrichting weinig van de z-richting afwijkt dan geldt $\frac{F_x}{F_z}$ en $\frac{F_y}{F_z}$ klein t.o.v. de eenheid. Als we dan bovendien gebruik maken van het feit dat u, v en w klein zijn t.o.v. U dan volgt

$$w \approx - U \frac{F_x}{F_z} . \tag{1.13}$$

Indien bovendien nog geldt dat de spanwijdte en de koorde van de vleugel groot zijn t.o.v. de afwijkingen van het vlak z = 0, dan is het gebruikelijk de randvoorwaarde (1.13) op te leggen op het vlak z = 0 en wel op het geprojekteerde vleugeloppervlak. In deze beschouwingen is dus dan een eventuele dikte van het draagvlak verdwenen, als we de randvoorwaarde voor de onderkant en de bovenkant van het draagvlak dezelfde nemen.

Om dus nu ons probleem op te lossen, zal dus een oplossing gezocht worden van verg. (1.6) of (1.11) met als randvoorwaarde verg. (1.13). Het is vanuit de potentiaaltheorie bekend, dat dergelijke problemen als integraalvergelijkingen kunnen worden geschreven.

## 2. DE AFLEIDING VAN DE TE GEBRUIKEN INTEGRAALVERGELIJKING

Gebruikmakend van de divergentiestelling van Gauss kan worden aangetoond dat voor een glad of bijna glad oppervlak geldt dat de oplossing $\bar{\phi}$ van $\Delta\bar{\phi} = 0$ kan worden geschreven als (zie ref (1) en ref (2))

$$(2.1) \qquad \bar{\phi}(x) = -\frac{1}{4\pi}\int\int_{F=0} \frac{1}{|x-\xi|}\frac{\partial\bar{\phi}}{\partial n}\,ds + \frac{1}{4\pi}\int\int_{F=0} \bar{\phi}\,\frac{\partial}{\partial n}\frac{1}{|x-\xi|}\,ds.$$

Hierin is x een punt van het uitwendige gebied rondom F terwijl $\xi$ een punt is van het oppervlak F = 0, n is de naar buiten gerichte normaal. Het is duidelijk dat de eerste integraal de potentiaal is tengevolge van een belegging van het oppervlak met bronnen met een sterkte $\frac{\partial\bar{\phi}}{\partial n}$. De tweede integraal is de potentiaal tengevolge van een belegging met dipolen met sterkte $\bar{\phi}$.

Het is nu mogelijk ons potentiaalprobleem op een geschikte manier te formuleren door gebruik te maken van het feit dat elke oplossing van de potentiaalvergelijking $\Delta\bar{\bar{\phi}} = 0$ voor het inwendige probleem moet voldoen aan

$$(2.2) \qquad 0 = \frac{1}{4\pi}\int\int_{F=0} \frac{1}{|x-\xi|}\frac{\partial\bar{\bar{\phi}}}{\partial n}\,ds - \frac{1}{4\pi}\int\int_{F=0} \bar{\bar{\phi}}\,\frac{\partial}{\partial n}\frac{1}{|x-\xi|}\,ds$$

voor een uitwendig punt x.
Door de vergelijkingen (2.1) en (2.2) te combineren ontstaat

$$(2.3) \qquad \bar{\phi}(x) = -\frac{1}{4\pi}\int\int_{F=0} \frac{1}{|x-\xi|}\{\frac{\partial\bar{\phi}}{\partial n} - \frac{\partial\bar{\bar{\phi}}}{\partial n}\}\,ds + \frac{1}{4\pi}\int\int_{F=0} (\bar{\phi}-\bar{\bar{\phi}})\,\frac{\partial}{\partial n}\frac{1}{|x-\xi|}\,ds.$$

De potentiaal $\bar{\phi}$ is nu naar vrije verkiezing zowel als een bronverdeling over F = 0 ($\bar{\phi} = \bar{\bar{\phi}}$ op F = 0) of als een dipoolverdeling weer te geven ($\frac{\partial\bar{\phi}}{\partial n} = \frac{\partial\bar{\bar{\phi}}}{\partial n}$).

In ons geval is het belangrijk om zich te realiseren, dat het in feite om een discontinuïteit in de druk, dat wil dus zeggen, om een discontinuïteit in $\psi$ gaat. Het ligt daarom zeer voor de hand te zoeken naar een oplossing van ons probleem in termen van een dipoolverdeling $\sigma$ over het oppervlak F = 0 voor de versnellingspotentiaal $\psi$

$$(2.4) \qquad \Psi = +\frac{1}{4\pi}\int\int_{F=0} \sigma\,\frac{\partial}{\partial n}\frac{1}{|x-\xi|}\,ds.$$

Gebruikmakend van vergelijking (1.10) en (1.13) volgt nu formeel

$$(2.5) \qquad w = \frac{1}{4\pi U}\frac{\partial}{\partial z}\int_{-\infty}^{x}\{\int\int_{F=0} \sigma\,\frac{\partial}{\partial n}\frac{1}{|x-\xi|}\,ds\}\,dx.$$

Het zou ons te ver voeren om hier precies aan te geven onder wat voor omstandigheden hiermee een eenduidige oplossing van ons probleem gewaarborgd is. Het is in ieder geval zeker dat voor voldoend gladde oppervlakken de funktie $\sigma$ eenduidig is te bepalen, en daarmee uiteraard ook $\bar{\phi}$ (althans op

een constante na). Helaas zou ons dit echter tot ongewenste oplossingen voeren, omdat volgens de paradox van D'Alembert een rotatievrije stroming slechts een oplossing kan opleveren, die de totale kracht die door de vloeistof op een lichaam wordt uitegeoefend nul maakt. Het was ons echter juist om een draagkracht te doen.

Hier rijzen dus grote problemen. De oplossing wordt verkregen door toe te laten dat er zogenaamde werveldraden in de vloeistof aanwezig zijn. Deze kunnen worden beschouwd als lijnen waar rot $\underline{v}$ singulier is, terwijl rot $\underline{v}$ overal elders nul is. In de draagvlaktheorie lopen deze draden af naar $x = +\infty$ waardoor de oplossing daar niet langer regulier is. Op deze wijze is de oplossing echter in het geheel niet meer bepaald. Deze bepaaldheid keert terug door gebruik te maken van het experimenteel waargenomen feit, dat de afstroming aan de vleugelachterrand glad is, wat wil zeggen dat daar moet gelden, dat er tussen onder- en bovenvlak van de vleugel geen drukverschil is, dus $\Delta\psi = 0$.

Het is aan te tonen dat ook in dit geval vergelijking (2.4) nog steeds geldt, daar hij in feite equivalent is met een belegging met elementaire "hoefijzerwervels" over het draagvlak, die afdruipen naar $x \to +\infty$.

We zullen nu vergelijking (2.5) wat nader beschouwen voor het geval het oppervlak $F = 0$ overeenkomt met het geprojekteerde vleugeloppervlak $z = 0$. In dit geval is $\frac{\partial}{\partial n} = \frac{\partial}{\partial z}$, terwijl er ter plaatse van $x = \xi$ een discontinuïteit in $\psi$ optreedt die gelijk is aan de lokale $\sigma$ waarde; terwijl deze discontinuïteit in $\psi$ juist het verschil $\Delta p/\rho$ is over de onderkant en de bovenkant van de vleugel.

Er ontstaat nu

$$(2.6) \qquad w = \frac{1}{4\pi U\rho}\frac{\partial}{\partial z}\int_{-\infty}^{x}\{\iint_{F=0}\Delta p\,\frac{z}{\{(x-\xi)^2+(y-\eta)^2+z^2\}^{3/2}}\,d\xi\,d\eta\}\,dx$$

$$(2.7) \qquad = -\frac{1}{4\pi U\rho}\iint_{F=0}\Delta p\,\frac{\partial}{\partial z}\,\frac{z}{(y-\eta)^2+z^2}\left(1+\frac{x-\xi}{\sqrt{(x-\xi)^2+(y-\eta)^2+z^2}}\right)d\xi\,d\eta.$$

Differentiëren we dit stomweg uit en stellen we dan $z = 0$, dan is het resultaat

$$(2.8) \qquad w = -\frac{1}{4\pi U\rho}\iint_{F=0}\frac{\Delta p}{(y-\eta)^2}\left\{1+\frac{x-\xi}{\sqrt{(x-\xi)^2+(y-\eta)^2}}\right\}d\xi\,d\eta.$$

Dit is de vorm waarin de vergelijking van de zogenaamde dragende vlak theorie vaak wordt aangetroffen. Het zal duidelijk zijn, dat de vergelijking in deze vorm in hoge mate singulier is. In dit geval is er een zin aan te geven door gebruik te maken van de hoofdwaarde van Hadaward.

$$(2.9)\qquad \fint_{-s}^{+s} \frac{f(\eta)}{(y-\eta)^2}\,d\eta = \lim_{\varepsilon\to 0}\left\{\int_{-s}^{y-\varepsilon} \frac{f(\eta)}{(y-\eta)^2}\,d\eta + \int_{y+\varepsilon}^{s} \frac{f(\eta)}{(y-\eta)^2}\,d\eta - \frac{2f(y)}{\varepsilon}\right\}.$$

We zullen nu pogen aan te geven hoe een dergelijke vergelijking als (2.8) kan worden opgelost. We moeten daarbij bedenken, dat oorspronkelijk door de aerodynamici een veel eenvoudiger vorm van deze vergelijking is gebruikt, de zgn. dragende lijn theorie, terwijl natuurlijk uit de theorie van de stroming om vleugelprofielen al wel bekend was wat men kon verwachten voor de drukverdeling in de koorderichting (x). Reeds voor de introductie van digitale rekenmachines is er daarom al getracht een handzame (letterlijk) oplossingstechniek voor vergelijking (2.8) aan te geven.
In het volgende hoofdstuk zullen we wat dieper op deze problemen ingaan.

## 3. OPLOSSINGSMETHODE VAN MULTHOPP

Een van de eerste en ook meest succesvolle methodes om een oplossing van het dragende vlak probleem te verkrijgen is afkomstig van MULTHOP [3]. Als de vleugel gegeven is als in fig. 2 en we voeren nu in

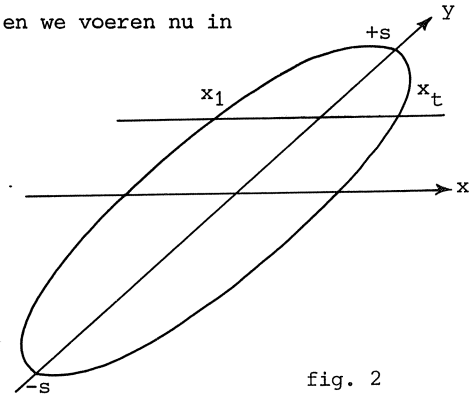


fig. 2

$$(3.1)\qquad \Delta c_p = -\frac{\Delta p}{\frac{1}{2}\rho U^2}\,,\quad \alpha = -\frac{w}{U},$$

$$(3.2)\qquad \xi = \frac{x}{s},\ \eta = \frac{y}{s},\ \ell(\eta') = \frac{x_t - x_1}{s},$$

en bovendien $X' = \dfrac{\xi-\xi_1}{\ell(\eta')}$,

dan is de integraalvergelijking te schrijven als

$$(3.3)\qquad \alpha(\xi,\eta) = -\frac{1}{2\pi} \fint_{-1}^{+1}\int_{0}^{1} \frac{\Delta c_p}{(\eta-\eta')^2}\,\frac{\ell(\eta')}{4s}\cdot$$

$$\cdot\left[1 + \frac{X - X'}{\sqrt{(X-X')^2 + \dfrac{s^2}{\ell(\eta')^2}\,(\eta-\eta')^2}}\right] dX'd\eta'.$$

De essentiële stap is nu om te stellen dat

$$(3.4) \qquad \Delta c_p = \frac{4s}{\ell(\eta')} \sum_{r=0}^{R} a_r(\eta')\, h_r\,(X').$$

Het voordeel van deze formule zal duidelijk zijn.

Door een geschikte keuze van het stelsel basisfunkties $h_r(X')$ kan een drukverdeling in koorde richting worden verkregen die zo goed mogelijk aansluit bij wat bekend is van twee-dimensionale drukverdelingen. De funkties $a_r(\eta')$ beschrijven dan het verloop in de spanwijdte richting. Voeren we dit nu in, in vergelijking (3.3), dan ontstaat

$$(3.5) \qquad \alpha(\xi,\eta) = -\frac{1}{2\pi} \sum_{r=0}^{R} \int_{-1}^{+1} \frac{a_r(\eta')H_r(\xi,\eta;\eta')}{(\eta-\eta')^2}\, d\eta'$$

met

$$(3.6) \qquad H_r(\xi,\eta;\eta') = \int_0^1 h_r\,(X')\left[1 + \frac{X - X'}{\sqrt{(X-X')^2 + \frac{s^2}{\ell(\eta')^2}\,(\eta-\eta')^2}}\right]$$

Op deze wijze hebben we dus nu het probleem teruggebracht tot het oplossen van een systeem van (R+1) onbekende funkties $a_r(\eta')$.
Multhopp voelt zich bij dit probleem onmiddellijk thuis, omdat het veel gemeen heeft met het welbekende probleem uit de dragende lijn theorie

$$(3.7) \qquad a(\eta) = -\frac{1}{2\pi} \int_{-1}^{+1} \frac{\gamma(\eta')\, d\eta'}{(\eta-\eta')^2}$$

waarvoor hij al eerder een succesvolle methode had ontworpen.
Deze methode maakt gebruik van het feit dat zowel voor $a_r$ als voor $\gamma$ geldt dat beide funkties zich gedragen als

$$(3.8) \qquad a_r \sim \sqrt{\varepsilon}$$

als $\varepsilon$ de afstand tot de vleugeltip is.

Voor $\gamma$ wordt dan een trigonometrische interpolatieformule gebruikt door in te voeren

$$(3.9) \qquad \eta = \cos\theta$$

en

(3.10) $\qquad \gamma = \sum \beta_\lambda \sin \lambda\theta$

waardoor automatisch aan (3.8) is voldaan.
Stel, dat we het interval $0 \leq \theta \leq \pi$ equidistant verdelen in (m+1) stukjes, dan geldt

$$\theta_n = \frac{n\pi}{m+1} \qquad n = 1(+1)m;$$

m wordt altijd even genomen.

Als nu in $\theta_n$, $\gamma$ gegeven is door $\gamma_n$, dan is uiteraard de vraag wat $\beta_\lambda$ is. Deze kan als volgt beantwoord worden.
Vergelijking (3.10) is te schrijven als volgt

$$\gamma = \sum_{n=1}^{n=m} \gamma_n g_n(\theta). \tag{3.11}$$

Als we nu stellen dat $g_n(\theta) = 1$ voor $\theta = \theta_n$ en $g_n(\theta_p) = 0$ voor $p \neq n$ dan is eenvoudig in te zien dat geldt

$$g_n(\theta) = -\frac{\sin(m+1)\theta \sin\theta_n}{(m+1)\cos(m+1)\theta_n \cdot (\cos\theta - \cos\theta_n)}. \tag{3.12}$$

En hieruit volgt nu

$$g_n(\theta) = \sum_{1}^{\infty} \delta_\lambda \sin\lambda\theta$$

(3.13)

waarbij $\qquad \delta_\lambda = \frac{2}{\pi} \int_0^\pi g_n(\theta) \sin\lambda\theta.$

Uitvoeren van deze integratie met gebruik van de volgende integraal

$$\frac{1}{\pi} \int_0^\pi \frac{\cos p\theta \, d\theta}{\cos\theta - \cos\theta_n} = \frac{\sin p\theta_n}{\sin\theta_n} \tag{3.14}^{*)}$$

---

*) Op de eenheidscirkel geldt $z = e^{i\theta}$, dus integraal is te schrijven als $\frac{1}{\pi i} \oint z^p/((z - e^{i\theta}n)(z - e^{-i\theta}n))\, dz = \frac{\pi i}{\pi i} \cdot$ Res (polen liggen op de integratieweg) $= e^{ip\theta n}/(e^{i\theta n} - e^{-i\theta n}) + e^{-ip\theta n}/(e^{-i\theta n} - e^{i\theta n}) =$
$= \sin p\theta_n / \sin\theta_n$.

geeft

$$\delta_\lambda = \frac{1}{m+1} \sin \lambda\theta_n \quad \text{voor } \lambda < m+1,$$
$$= 0 \qquad\qquad \text{voor } \lambda \geq m+1,$$

zodat uiteindelijk geldt

$$\gamma = \frac{2}{m+1} \sum_{n=1}^{n=m} \gamma_n \sum_{\lambda=1}^{m} \sin \lambda\theta_n \sin \lambda\theta. \tag{3.15}$$

Multhopp gebruikt nu deze formule voor oplossen van vergelijking (3.5) door een voorstelling zoals gegeven in vergelijking (3.14) toe te passen op de combinatie $a_r H_r$, zij het dan dat er eerst een correctie is gemaakt voor een logarithmische singulariteit die in de integrand van (3.5) optreedt en waar we later op terug zullen komen.

Met deze voorstellingswijze en door formeel over de singulariteit heen te integreren lukt het nu Multhopp een betrekkelijk simpel systeem van uitdrukkingen te verkrijgen. Als we aannemen dat het vrij eenvoudig is, voldoend nauwkeurige antwoorden voor $H_r(\xi,\eta;\eta')$ te produceren, dan is hiermee de zaak afgedaan. En eerlijk gezegd was de zaak hiermee vele jaren lang ook afgedaan omdat men in wezen alleen in geïntegreerde uitdrukkingen geïnteresseerd was, zoals b.v.

$$\int_{-1}^{+1} \int_{0}^{1} \Delta c_p \, dx' \, d\eta.$$

Multhopp volstond er daarom mee twee verschillende basisfuncties in de X-richting te kiezen en in het algemeen niet meer dan 7 basispunten $\theta_n$ in de spanwijdte richting.

Door nu in de X-richting 2 collocatiepunten te kiezen bij elk spanwijdte station, ontstaat in totaal een aantal lineaire vergelijkingen in de onbekende grootheden $a_r(\eta_n)$; door de daar ter plaatse voorgeschreven grootheid $\alpha$ is het rechterlid van deze vergelijkingen bepaald.

Toen in de loop der jaren de mogelijkheden van de digitale computer in principe het oplossen met veel meer onbekenden toeliet, terwijl ook de technici grotere nauwkeurigheid eisten, viel de Multhopp-methode door de mand. Er was geen sprake van convergentie in de resultaten bij het laten toenemen van met name R, dus het aantal collocatiepunten in de koorde richting. Maar vooral ook in de drukverdeling waren de verschillen groot. Een nader onderzoek was noodzakelijk.

Als we de Multhopp methode nu nader bekijken, dan vallen direkt een aantal merkwaardigheden op.
In de eerste plaats is het zo dat volgens verg. (3.5) en (3.7) de funktie $a_r H_r$ gezamenlijk door een Fourierreeks worden voorgesteld terwijl het toch eigenlijk om $a_r$ gaat. Bij nader onderzoek blijkt dat $H_r$ zich helemaal niet zo fijn gedraagt voor met name hogere waardes van R, wat dan betekent dat er van de nauwkeurigheid van integratie niet zoveel overblijft met een beperkt aantal punten in spanwijdte richting. Maar het is zeer goed mogelijk dat $a_r$ in het geheel niet zoveel Fouriercomponenten nodig heeft, om goed te worden gerepresenteerd.

Het ligt dus zeer voor de hand een scheiding te maken in de representatie van $a_r$ en die van $H_r$. Het voordeel daarvan is dat de integratienauwkeurigheid kan worden opgevoerd zonder dat het aantal vergelijkingen excessief toeneemt.

Een tweede punt om wat nader te bekijken lijkt de integratieprocedure volgens verg. (3.7) en (3.14). Een betere methode dan welke door Multhopp gesuggereerd, lijkt te zijn de singulariteit zodanig af te splitsen dat er een regulier deel overblijft, waarop dan wel en met aan te geven nauwkeurigheid een vergelijking als (3.15) kan worden toegepast.

Een dergelijke methode is iets meer dan 10 jaar geleden ontwikkeld (ref. 4) op het NLR en in het volgende hoofdstuk zullen we de voornaamste punten wat nader beschouwen.

## 4. DE OPLOSSINGSMETHODE VAN HET NLR

We gaan weer uit van verg. (3.5) en (3.6). Nader onderzoek van het gedrag van de funktie $H_r$ voor $\eta' \to \eta$ leert, dat $H_r$ een term bevat van de orde $(\eta-\eta')^2 \ln(\eta-\eta')$. Dit betekent dus dat de integrand van (3.5) een logaritmische singulariteit bevat. We voeren daarom in

$$F_r = H_r(\xi,\eta;\eta') + \frac{s^2}{\ell(\eta)^2}(\eta-\eta')^2 \ln|\eta-\eta'| \cdot \frac{dh_r}{dx'}(\eta=\eta'). \tag{4.1}$$

Bovendien nemen we aan dat we de funkties $a_r$ kunnen voorstellen door trigonometrische interpolatie, dus volgens verg. (3.15)

$$a_r(\eta') = \frac{2}{m+1} \sum_{n=1}^{m} a_r(\eta_n) \sum_{\mu=1}^{m} \sin \mu\theta' \sin \mu\theta_n. \tag{4.2}$$

Vullen we dit in in verg. (3.5) dan ontstaat

$$\alpha(\xi,\eta) = -\frac{1}{2\pi}\sum_{r=0}^{R}\frac{2}{m+1}\sum_{n=1}^{m}a_r(\eta_n)\sum_{\mu=1}^{m}\sin\mu\theta_n\int_0^{\pi}\frac{F_r(\xi,\eta;\theta')\sin\mu\theta'\sin\theta'}{(\cos\theta-\cos\theta')^2}\,d\theta' +$$

$$+\frac{1}{2\pi}\sum_{r=0}^{R}\frac{2}{m+1}\frac{s^2}{l(\eta)^2}\frac{dh_r}{dX'}(\theta=\theta')\sum_{n=1}^{m}a_r(\eta_n)\sum_{\mu=1}^{m}\sin\mu\theta_n\int_0^{\pi}\ln|\cos\theta-\cos\theta'|$$

(4.3) $\qquad \sin\mu\theta'\sin\theta'\,d\theta'.$

Het ligt nu natuurlijk erg voor de hand de singuliere integrand te regulariseren door het invoeren van een funktie

(4.4) $\qquad \bar{H}_r(\xi,\eta;\theta') =$

$$= \frac{\{F_r(\xi,\eta;\theta')-F_r(\xi,\eta;\theta)-(\cos\theta-\cos\theta')\frac{\partial F_r}{\partial n'}(\theta=\theta')\}\sin\theta'}{(\cos\theta-\cos\theta')^2}.$$

Een kleine opmerking is hier op zijn plaats. Het zou misschien meer voor de hand hebben gelegen om de sin θ' buiten deze definitie te hebben gelaten. Dat de sin θ' is toegevoegd heeft als reden dat de nu resulterende integralen alle analytisch kunnen worden opgelost, wat uiteraard een groot voordeel betekent. Om dat te bereiken wordt nu voor de funktie $\bar{H}_r$ ook een trigoniometrische voorstelling gekozen, echter met aanzienlijk meer basispunten dan die voor $a_r$, of om het anders te zeggen, onafhankelijk daarvan.

We kiezen voor het aantal basispunten a(m+1) - 1; het getal a beheerst daarmede dus de integratienauwkeurigheid; we kunnen dan schrijven:

$$\bar{H}_r(\xi_p,\theta_\nu;\theta') = \frac{2}{a(m+1)}\sum_{\lambda=1}^{a(m+1)-1}\bar{H}_r(\xi_p,\theta_\nu;\theta_\lambda)\sum_{\mu=1}^{a(m+1)-1}\sin\mu\theta'\sin\mu\theta_\lambda$$

(4.5)

$$= \sum_{\mu=1}^{a(m+1)-1}\frac{1}{a(m+1)}\bar{H}_r(\xi_p,\theta_\nu;\theta_\lambda)(-1)^\lambda\frac{\sin\theta_\lambda\sin a(m+1)\,\theta'}{\cos\theta_\lambda-\cos\theta'}$$

volgens verg. (3.12).

Het totale systeem kan nu worden geschreven als

$$(4.6)\qquad \alpha(\xi_p,\eta_\nu) = -\frac{1}{2}\sum_{r=0}^{R}\sum_{n=1}^{m} a_r(\eta_n)\Big\{\sum_{\lambda=1}^{a(m+1)-1}\frac{1}{a(m+1)}\,\bar{H}_r(\xi_p,\theta_\nu;\theta_\lambda)\gamma_{n\lambda} +$$

$$+ F_r(\xi_p,\theta_\nu;\theta_\nu)\,\frac{\sin\theta_n}{\sin\theta_\nu}\,\varepsilon_{\nu n} + \frac{\partial F_r}{\partial\eta'}(\theta'=\theta_\nu)\,\zeta_{\nu n} - \frac{\beta^2 s^2}{\ell(\eta_\nu)^2}\,\frac{dh_r}{dX'}(\theta'=\theta_\nu)\,s_{\nu n}\Big\}$$

waarbij dan geldt

$$(4.7)a\qquad \gamma_{n\lambda} = \frac{2(-1)^\lambda \sin\theta_\lambda}{\pi(m+1)}\sum_{\mu=1}^{m}\sin\mu\theta_n\int_0^\pi \frac{\sin a(m+1)\,\theta'\sin\mu\theta'}{\cos\theta_\lambda-\cos\theta'}\,d\theta',$$

$$(4.7)b\qquad \varepsilon_{\nu n} = \frac{2\sin\theta_\nu}{\pi(m+1)\sin\theta_n}\sum_{\mu=1}^{m}\sin\mu\theta_n\int_0^\pi \frac{\sin\theta'\sin\mu\theta'}{(\cos\theta_\nu-\cos\theta')}\,d\theta',$$

$$(4.7)c\qquad \zeta_{\nu n} = \frac{2}{\pi(m+1)}\sum_{\mu=1}^{m}\sin\mu\theta_n\int_0^\pi \frac{\sin\theta'\sin\mu\theta'}{\cos\theta_\nu-\cos\theta'}\,d\theta',$$

$$(4.7)d\qquad s_{\nu n} = \frac{2}{\pi(m+1)}\sum_{\mu=1}^{m}\sin\mu\theta_n\int_0^\pi \ln|\cos\theta_\nu-\cos\theta'|\sin\mu\theta'\sin\theta'\,d\theta'.$$

We zullen niet al deze integralen uitwerken, maar het blijkt dat ze allemaal analytisch zijn uit te werken. Als voorbeeld zullen we de eerste beschouwen.

$$\int_0^\pi \frac{\sin a(m+1)\,\theta'\sin\mu\theta'}{\cos\theta_\lambda-\cos\theta'}\,d\theta' = \frac{1}{2}\int_0^\pi \frac{\cos\{a(m+1)-\mu\}\theta' - \cos\{a(m+1)+\mu\}\theta'}{\cos\theta_\lambda-\cos\theta'}\,d\theta'.$$

Dit is volgens formule (3.14) te schrijven als

$$-\frac{1}{2}\pi\,\frac{\sin\{a(m+1)-\mu\}\theta_\lambda - \sin\{a(m+1)+\mu\}\theta_\lambda}{\sin\theta_\lambda} = \pi(-1)^\lambda\,\frac{\sin\mu\theta_\lambda}{\sin\theta_\lambda}$$

ofwel

$$\gamma_{n\lambda} = \frac{2}{m+1}\sum_{\mu=1}^{m}\sin\mu\theta_n\sin\mu\theta_\lambda.$$

Maken we nu gebruik van verg. (3.12), (3.13) en (3.15) dan volgt hieruit

$$(4.8)a\qquad \gamma_{n\lambda} = \frac{(-1)^n}{m+1}\,\frac{\sin\frac{\lambda}{a}\pi\sin\theta_n}{\cos\theta_n-\cos\theta_\lambda}\qquad \text{voor } \theta_n \neq \theta_\lambda,$$

$$(4.8)b\qquad \gamma_{n\lambda} = 1\qquad \text{voor } \theta_n = \theta_\lambda.$$

Op dezelfde wijze zijn voor de overige grootheden uit verg. (4.7) verrassend eenvoudige uitdrukkingen af te leiden. Hiervoor verwijzen we naar ref. 4. Het is duidelijk dat voor het verkrijgen van een goed resultaat ook de grootheden $\frac{\partial F_r}{\partial \eta'}$ en $\bar{H}_r(\xi,\theta_\nu;\theta_\nu)$ berekend moeten worden. Hoewel dat op zichzelf best een interessant stukje analytisch geconstrueer is, moeten we dat hier achterwege laten en verwijzen naar de bijlage van ref. 4. Vooral de uitdrukking voor $\bar{H}_r(\xi,\theta_\nu;\theta_\nu)$ is bepaald niet eenvoudig.
De ook benodigde uitdrukking voor $F_r(\xi_p,\theta_\nu;\theta_\nu)$ is gelukkig wel zeer eenvoudig. Uit verg. (3.6) volgt rechtstreeks

$$(4.9) \qquad F_r(\xi_p,\theta_\nu;\theta_\nu) = 2\int_0^{X_p} h_r\,(X')\,dX'.$$

Maar deze uitdrukking geeft ons aanleiding op te merken, dat we tot dusverre nog niets hebben opgemerkt over de toch ook zeer gewichtige keuze van de funkties $h_r$. Deze funkties zullen zo moeten zijn dat ze karakteristieken van de drukverdeling in koorderichting goed weergeven.

Uit de theorie voor een oneindig lange vlakke plaat is bekend dat deze drukverdeling een wortelsingulariteit aan de voorzijde bevat, terwijl aan de achterkant de druk naar nul gaat. Tegenwoordig is het algemeen gebruikelijk om hiervoor te kiezen

$$(4.10) \qquad h_r(X') = \frac{1}{\pi}\,\frac{T_r(1-2X') + T_{r+1}\,(1-2X')}{\sqrt{X'(1-X')}}\,.$$

Door nu te stellen

$$(4.11) \qquad X' = \frac{1 - \cos\psi}{2}$$

gaat deze formule over in

$$(4.12) \qquad h_r(\psi) = \frac{2}{\pi}\,\frac{\cos\frac{2r+1}{2}\psi}{\sin\frac{\psi}{2}}$$

en uit deze voorstelling is direkt duidelijk, dat $h_r$ aan de gestelde eisen voldoet. Ook volgt nu onmiddellijk bijv.

$$(4.13)\qquad F_r(\xi_p,\theta_\nu;\theta_\nu) = 2\int_0^{X_p} h_r(X')\,dX' = \frac{4}{\pi}\int_0^{\phi_p} \cos\frac{2r+1}{2}\psi\,\cos\frac{\psi}{2}\,d\psi$$

$$= \frac{2}{\pi}\{\frac{1}{r}\sin r\phi + \frac{1}{r+1}\sin(r+1)\phi\}.$$

De algemene uitdrukking voor $H_r$ wordt

$$(4.14)\qquad H_r(\xi,\eta;\eta') = \frac{1}{2}\int_0^{\pi}[\cos r\psi + \cos(r+1)\psi]\,\frac{\beta_1+\cos\psi}{\sqrt{(\beta_1+\cos\psi)^2+\beta_2^2}}\,d\psi$$

met $\beta_1 = 2X-1$ en $\beta_2 = (\eta'-\eta)\frac{s}{\ell(\eta')}$ .

Het is uiteraard duidelijk dat vooral de nauwkeurigheid waarmee de funktie $H_r$ wordt bepaald van grote invloed is op de nauwkeurigheid van het totale resultaat. Uit de formule (4.14) blijkt al wel dat er problemen te verwachten zijn voor r relatief groot en $\beta_2$ klein. Trouwens de gehele problematiek rondom vergelijking (4.14) is nogal ingewikkeld. Zo is het bijv. niet moeilijk in te zien dat we met een zogenaamde elliptische integraal te maken hebben. Maar dat helpt ons niet veel want er zijn geen programma's om dat ding in voldoende nauwkeurigheid te berekenen. Daarom kan men beter voor een rechtstreekse aanpak kiezen. Nu is in het verleden nogal eens gekozen voor een integratieroutine met zelfregelende stapgrootte t.b.v. een gegeven nauwkeurigheid. Maar het bleek al snel dat dit een kostbare grap was, omdat dan de integraal onnodig met een nog twee maal zo fijne verdeling werd uitgerekend en dat kost tijd en geld.

Op het NLR is daarom destijds gekozen voor een weliswaar niet elegante maar rechtstreekse benadering. De integraal is berekend met de trapeziumregel, die voor een periodieke funktie ook tegelijkertijd optimaal is. Het aantal integratiepunten (L+1) is gekozen in afhankelijkheid van de parameters $\beta_1$ en $\beta_2$ en zodanig, dat een bepaalde nauwkeurigheid is gegarandeerd voor r = 0 tot r = 10. Op deze wijze kan snel worden gerekend met een vaststaande nauwkeurigheid. Alleen voor $|\beta_1| < 1$ en $|\beta_2| < 0{,}02$ moet gebruik gemaakt worden van een andere procedure omdat dan $f(\psi)$ bijna discontinu is en dus onnauwkeurigheid ontstaat door het wegvallen van cijfers door aftrekking. Door nu het grootste deel analytisch te integreren en de kleine rest weer numeriek lukt het ook hier de nauwkeurigheid te handhaven.

Op deze wijze is nu alles beschikbaar om vergelijking (4.6) om te zetten in een lineair stelsel vergelijkingen. Daartoe kiezen we een aantal collocatiepunten, nl. voor elk spanwijdte station $\theta_n$, (R+1) punten in koorde richting. We verkrijgen dan juist m(R+1) vergelijkingen voor de m(R+1) onbekenden $a_r(\eta_n)$ waarmee dan de oplossing bepaald is. Opgemerkt moet worden, dat we in het algemeen met een volle matrix te doen hebben.

Als nu voldaan is aan de basisveronderstellingen van de methode, nl. dat de rand van de vleugelplanform twee keer continu differentieerbaar is, dus een continue kromtestraal heeft, dan kan op deze wijze de oplossing van het dragende vlak probleem met een gewenste mate van nauwkeurigheid worden verkregen.

Zo is het nu mogelijk om voor een bepaalde gekozen waarde van m en R een oplossing met een voorgeschreven nauwkeurigheid te berekenen door a te variëren. Vervolgens is het mogelijk om de convergentie van de resultaten met betrekking tot m en daarna met betrekking tot R te bekijken. Al deze exercities leiden tot een convergent resultaat. Iets wat bij Multhopp ver te zoeken was.

Als een voorbeeld van het verschil in resultaat dat verkegen kan worden, volgen hier de waardes van $a_1$ voor $\eta = 0{,}5406$ voor een rechthoekige vleugel met $l(\eta) = \frac{1}{4}$ voor de oorspronkelijke Multhopp methode en voor de NLR methode met a = 1 en a = 8 voor verschillende waarden van R

| R = 1 | R = 2 | R = 3 | R = 4 | |
|---|---|---|---|---|
| 0.0118 | 0.0078 | -0.0015 | -0.0088 | Multhopp |
| 0.0055 | 0.0018 | -0.0002 | -0.0013 | NLR a = 1 |
| 0.0064 | 0.0064 | 0.0064 | 0.0064 | NLR a = 8 |

Het is duidelijk dat hier R = 1 al voldoende is om een goed resultaat te verkrijgen maar dat de onvoldoende integratienauwkeurigheid van de Multhopp methode, zowel als de NLR methode voor a = 1 dit niet kunnen opleveren. Overigens is dus Multhopp niet hetzelfde als NLR a = 1.

In de loop van de afgelopen 10 jaar zijn er uitvoerige numerieke studies gemaakt op basis van een beschikbaar programma. Zie ref. 5, 6 en 7.

AFSLUITENDE OPMERKINGEN

De hierboven beschreven methode voor de oplossing van de dragende vlak vergelijking is natuurlijk uit te breiden tot het oplossen van verwante problemen van een trillende vleugel met lage frequentie of voor scheeps-schroeven. Beide uitbreidingen zijn op basis van de hier beschreven theorie gegeven. Die voor scheepsschroeven bijv. in ref. 8.

Het is uiteraard ook mogelijk geheel andere methodes te benutten om de vergelijking op te lossen. Zo zou men het probleem ook kunnen oplossen door de vleugel in elementen te verdelen en op ieder element $\Delta c_p$ in geschikte vorm te geven, of zoals dat ook wel gebeurd is, werken met een verdeling van hoefijzerwervels over elk element.

Het is hier echter niet de plaats daar dieper op in te gaan, alhoewel er zeker aanleiding zou kunnen bestaan in het licht van recente ontwikkelingen het probleem opnieuw te bekijken. Maar de technische noodzaak daarvoor is eigenlijk niet aanwezig.

REFERENTIES

[1] KELLOGG, O.D., *Foundations of Potential Theory,* Dover Publishing Cy.

[2] LAMB, H., *Hydrodynamics,* Dover Publishing Cy.

[3] MULTHOPP, H., *Methods for calculating the lift distribution of wings* (subsonic lifting surface theory), Report no. Aero 2353 R.A.E. 1950.

[4] ZANDBERGEN, P.J., TH.E. LABRUJERE & J.G. WOUTERS, *A new approach to the numerical solution of the equation of subsonic lifting surface theory,* NLR report TR G.49 1967.

[5] LABRUJERE, TH.E., *Computer Application of subsonic lifting surface theory,* NLR TR 70088 U 1970.

[6] GARNER, H.C., B.L. HEWITT & TH.E. LABRUJERE, *Comparison of three methods for the evaluation of subsonic lifting surface theory,* NLR TN G.65 1968.

[7] LABRUJERE, TH.E. & P.J. ZANDBERGEN, *On the application of a new version of lifting surface theory to non slender and kinked wings,* Journal of Engineering Mathematics, Vol. 7, no.1 1973.

[8] VIS, F., *De berekening van de drukverdeling over de bladen van een scheepsschroef*, Doktoraalverslag TH Twente, aug 1975.

# INTEGRAL EQUATION METHODS FOR MULTI-ELEMENT AIRFOIL ANALYSIS AND DESIGN

Th.E. LABRUJÈRE

CONTENTS


1 INTRODUCTION 173

2 FORMULATION OF THE FLOW PROBLEMS 175

3 DERIVATION OF THE INTEGRAL EQUATIONS 176

4 THE ANALYSIS PROBLEM 181
4.1 The independent variable 181
4.2 The order of approximation 182
4.3 Solution of the integral equations 182
4.3.1 General principles 182
4.3.2 The integral equation resulting from the the boundary condition $\phi_n = 0$ 183
4.3.3 The integral equations resulting from the boundary condition $\psi = 0$ 184
4.3.4 The integral equations resulting from the boundary condition $\bar{\phi} = 0$ 185
4.4 Numerical results 186
4.5 Conclusions 188

5 THE DESIGN PROBLEM 189
5.1 Basic idea 189
5.2 Numerical approximations 192
5.3 Solution to the minimization problem 192
5.4 Numerical results 193
5.5 Conclusions 195

6 FINAL REMARKS 195

7 REFERENCES 196

2 tables
20 figures

( 48 pages in total)

LIST OF SYMBOLS 219

## 1 INTRODUCTION

In the past 10 - 15 years theoretical aerodynamic research based on the application of calculation methods has become increasingly important for aircraft developments. A great deal of these methods is based on the assumption of potential flow. This assumption allows the application of integral equation methods, which because of their dimension reducing character are especially attractive for the solution of large scale problems.

In aerodynamic wing design two types of flowproblems may be considered. The first type is the "analysis" or "direct" problem which involves the determination of the velocity distribution along the surface of a body in a given onset flow. The second type is the "design" or "inverse" problem, which involves the determination of the geometry of a body such that it generates in a given onset flow an apriori specified velocity distribution along its surface.

The analysis problem amounts to the Neumann problem of prescribed normal derivative of the potential along the surface of the body, which may be solved by considering a linear integral equation.

The design problem amounts to the solution of two coupled non-linear integral equations, one of which is associated with the Neumann problem. The other one is associated with the Dirichlet problem of prescribed potential.

In this paper the two-dimensional analysis problem for a single airfoil and the two-dimensional design problem for a wing-flap configuration are considered.

In general, if calculation methods for the solution of such problems are being developed for engineering purposes, a few problems appear that have little to do with the mathematical formulation of the solution, but that have a great influence on the ultimate practical applicability. The prospective user will have insight into the physics of the problems to be solved such that he will be able to correlate to a certain extent geometrical and aerodynamical properties. This knowledge can be used at the selection of the parameters controling the numerical procedure. Therefore these parameters should have a direct physical meaning. Then the user will be able to use the method without having a thorough under-

standing of the numerical details. A second problem is the fulfilment of the requirement to use the computer in an economic way, such that a given accuracy of the solution is obtained with computing costs as low as possible. A third problem is the wish to extend the applicability of the calculation method to three-dimensional problems. This implies that the algorithms to be applied should be chosen such that they allow an extension as straightforwardly as possible.

The solution of the analysis problem has already been formulated long ago (see e.g. Ref. 1) and for its determination computerprograms were developed (see e.g. Refs. 2, 3). Ever since, many researchers have been and still are working on this problem; their goal being to arrive at a calculation method that for their particular problem is optimally suited in terms of manpower and computereconimics.

At NLR a computerprogram has been in use for some ten years for the calculation of the velocity distribution on complete aircraft configurations (see Ref. 5). Since the development of this program new improved techniques have become available, so that when it was decided to attack the design problem, also the solution of the analysis problem was reviewed again.

At the solution of the design problem an extra difficulty may arise, i.e. the fact that an arbitrarily prescribed velocity distribution can lead to physically impossible shapes with e.g. locally negative thickness. However, engineering arguments available to the designer will often allow to prescribe the velocity distribution in a qualitatively or quantitatively approximated sense only. A plausible use of this fact can be made by exchanging the implied amount of freedom for control over the geometry to be determined. Another necessity for applying constraints on the geometry may arise from constructive or manufacturing requirements.

Up to now, there are only a few calculation methods available that lead to a more or less satisfactory solution of the design problem (Refs. 8, 9, 10, 11). The main drawback of the methods of References 8, 9, 10 is the fact that they offer little possibilities to incorporate constraints with respect to the geometry. The method of Reference 11 has the disadvantage of a not very flexible contour representation.

In the following a review of alternatives for the solution of the

analysis problem is given before describing the procedure that has been developed for the solution of the design problem. These methods are based on simulation of the flow by means of a doublet distribution along the contour of the airfoil. The unknown doubletstrength and the unknown contour are determined by solving the linear and non-linear integral equations associated with the respective problems. The design problem is solved with application of a constraint to the geometry. All matters are illustrated by means of numerical results.

This work has been performed by NLR under contract for the Netherlands Agency for Aerospace Programs (NIVR).

## 2 FORMULATION OF THE FLOW PROBLEMS

The two-dimensional incompressible inviscid flow around an airfoil is considered. The problems to be solved are the determination of the velocity distribution along the contour of a given airfoil (analysis problem) and the determination of the shape of an airfoil which generates approximately a specified velocity distribution (design problem).

In both cases there exists a velocity potential $\phi$ which satisfies the Laplace equation (see Fig. 1)

$$\phi_{xx} + \phi_{yy} = 0 \qquad \text{in the flow domain G} \qquad (2.1)$$

and which is subject to the conditions

$$\left.\begin{aligned} \phi_x &= V_\infty \cos\alpha \\ \phi_y &= V_\infty \sin\alpha \end{aligned}\right\} \qquad \text{at infinity} \qquad (2.2)$$

The solution to the analysis problem is determined by the boundary condition

$$\phi_n = 0 \qquad \text{at the } \underline{\text{given}} \text{ airfoil contour C} \qquad (2.3)$$

The solution to the design problem is determined by the boundary conditions

$$\left.\begin{aligned} \phi_n &= 0 \\ \phi_t &\approx V_t \end{aligned}\right\} \quad \text{at the } \underline{\text{unknown}} \text{ airfoil contour C} \qquad (2.4)$$

where $V_t$ is the specified velocity.

In both cases physically meaningful solutions are obtained only if also the socalled Kutta condition of smooth flow is fulfilled at the trailing edges of the airfoil elements.

The geometrical constraint for the design problem may be formulated as:

$$y\ (x) \approx \bar{y}\ (x) \qquad (2.5)$$

which expresses the requirement that the airfoil contour to be determined y (x) should deviate not too much from a given contour $\bar{y}$ (x).

## 3 DERIVATION OF THE INTEGRAL EQUATIONS

From general potential theory it follows that the velocity potential $\phi$ may be written

$$\phi = \phi_\infty + \phi_d \qquad (3.1)$$

where

$$\phi_\infty = V_\infty \cos \alpha\ x + V_\infty \sin \alpha\ y \qquad (3.2)$$

is the velocity potential due to the uniform flow, and where $\phi_d$ is the perturbation velocity potential due to the presence of the airfoil.

If there is a circulation around the airfoil i.e. in the case of lift, the perturbation potential exhibits a discontinuity of constant magnitude along a line extending from a point on the airfoil to infinity. Such a perturbation potential can be represented as the potential due to a doublet distribution $\mu$ of varying strength along the contour of the airfoil and a doublet distribution of constant strength $\mu_w$ along the line of discontinuity (see Fig. 2).

As the velocity due to the doublet distribution approaches zero at infinity the boundary condition for $\phi$ at infinity is fulfilled automatically.

As far as the boundary condition of zero normal velocity at the airfoil contour is concerned, the following remarks may be made. It can be shown (see Ref. 1) that if a double layer is approached from opposite sides, the normal derivatives of its potential have equal limits. Hence at the airfoil contour:

$$\phi_n^+ = \phi_n^- = 0 \tag{3.3}$$

where the + sign refers to the outer side and the - sign to the inner side.

Now, from Green's theorem (Ref. 1) it follows that $\phi$ is a constant throughout the flow region if the normal derivative of $\phi$ vanishes at the boundary. Also the reverse is true, which implies that $\phi_n = 0$ if $\phi$ is prescribed to be constant along the boundary. In other words the Neumann boundary condition for the flow around the airfoil may be replaced by a Dirichlet boundary condition

$$\phi^- = C \tag{3.4}$$

for the flow inside the airfoil.
The value of the constant is irrelevant and may thus be taken equal to zero.

A third alternative for the Neumann boundary condition appears if it is noted that the condition of zero normal velocity is derived from the fact that the airfoil contour should be a stream line. Hence the Neumann condition can also be replaced by the requirement that the streamfunction has a constant value along the airfoil contour.

Another wellknown property of a double layer is the fact that it gives rise to a discontinuity in the potential across the layer such that

$$\phi^+ - \phi^- = -2\,\pi\mu \tag{3.5}$$

Setting $\phi^- = 0$ leads to the observation that

$$\phi_t^+ = -2\,\pi\,\frac{d\mu}{dt}\,, \tag{3.6}$$

so that the boundary condition of prescribed tangential velocity at the airfoil contour may be replaced by the condition

$$\frac{d\mu}{dt} = -\frac{1}{2\pi} V_t \tag{3.7}$$

It can be shown (see e.g. Ref. 6) that the Kutta condition is fulfilled implicitly if the line of discontinuity in potential is chosen to emanate from the trailing edge, if the constant doublet strength along this line is chosen according to

$$\mu_w = \mu_u - \mu_L \tag{3.8a}$$

where $\mu_u$ and $\mu_L$ denote the doublet strength at the trailing edge at upper and lower side respectively (see fig. 2) and if furthermore

$$\left(\frac{d\mu}{ds}\right)_u + \left(\frac{d\mu}{ds}\right)_L = 0 \tag{3.8b}$$

Now consider the potential due to a doublet distribution as indicated above. At any point outside the airfoil contour C, the potential is given by:

$$\phi_d^o = -\int_0^S \mu(s) \frac{\partial}{\partial s}\left\{\tan^{-1}\frac{y-\eta}{x-\xi}\right\} ds - \mu_w \int_S^\infty \frac{\partial}{\partial s}\left\{\tan^{-1}\frac{y-\eta}{x-\xi}\right\} ds$$

$$= -\int_0^S \mu(s) \frac{\partial}{\partial s}\left\{\tan^{-1}\frac{y-\eta}{x-\xi}\right\} ds + \mu_w \tan^{-1}\frac{y-y_T}{x-x_T} \tag{3.9}$$

If C is approached from the exterior an appropriate limiting procedure shows that

$$\phi_d^+ = -\pi\mu + \phi_d^o \tag{3.10a}$$

And if C is approached from the interior

$$\phi_d^- = \pi\mu + \phi_d^o \tag{3.10b}$$

Furthermore it can be shown that

$$\frac{\partial \phi_d}{\partial n}^+ = \frac{\partial \phi_d}{\partial n}^- = \frac{\partial \phi_d}{\partial n}^o \tag{3.11}$$

The appearance of the integrand of eq. (3.9) suggests a possible benefit from partial integration before evaluating the integral. In that case the expression for $\phi_d^o$ becomes:

$$\phi_d^o = \int_0^S \frac{d\mu}{ds} \tan^{-1} \frac{y-\eta}{x-\xi} \, ds + \pi\mu_L \tag{3.12}$$

The stream function due to the doublet distribution is given by the expression:

$$\psi_d = -\int_0^S \mu(s) \frac{\partial}{\partial s} \left\{ \ell n \, r \right\} ds + \mu_w \, \ell u \, r_T \tag{3.13a}$$

or

$$\psi_d = \int_0^S \frac{d\mu}{ds} \, \ell n \, r \, ds \tag{3.13b}$$

where $r = \{(x-\xi)^2 + (y-\eta)^2\}^{\frac{1}{2}}$ ; $r_T = \{(x-x_T)^2 + (y-y_T)^2\}^{\frac{1}{2}}$.

Now, the application of the boundary conditions in their alternative forms in combination with the alternative expressions for the velocity potential or the stream function leads to a number of integral equations. From theoretical point of view these integral equations are fully equivalent as far as they refer to the same problem. Some combinations lead to equations of the first kind, the solution of which can be troublesome. Other combinations lead to equations of the second kind. It has appeared that the determination of an accurate solution to these equations is not trivial either.

In an attempt to choose that formulation which leads to a calculation method which is optimal with respect to computing time, range of application and flexibility as far as extensions to more complex and 3D problems are concerned, five alternative formulations for the solution of the analysis problem have been considered. Several other formulations could be devised. However, the associated integral equations will have more complicated kernels or have other computing time increasing features.

In relation to the boundary condition $\phi_n = 0$ the following integral equation may be derived:

$$\int_0^S \frac{d\mu}{ds} \frac{\partial}{\partial n}\left\{\tan^{-1}\frac{y-\eta}{x-\xi}\right\}ds = -V_\infty \cos\alpha \frac{\partial x}{\partial n} - V_\infty \sin\alpha \frac{\partial y}{\partial n} \qquad (3.14)$$

Application of the boundary condition $\phi = 0$ gives rise to the integral equations:

$$\int_0^S \mu \frac{\partial}{\partial s}\{\ell n\, r\}ds - \mu_w \ell n\, r_T = V_\infty (x \sin\alpha - y \cos\alpha) \qquad (3.15a)$$

or

$$-\int_0^S \frac{d\mu}{ds} \ell n\, r\, ds = V_\infty (x \sin\alpha - y \cos\alpha) \qquad (3.15b)$$

From the boundary condition $\bar{\phi} = 0$ the following integral equations result:

$$\pi\mu - \int_0^S \mu \frac{\partial}{\partial s}\left\{\tan^{-1}\frac{y-\eta}{x-\xi}\right\}ds + \mu_w \tan^{-1}\frac{y-y_T}{x-x_T} = -V_\infty (x \cos\alpha + y \sin\alpha) \qquad (3.16a)$$

or

$$\pi(\mu+\mu_L) + \int_0^S \frac{d\mu}{ds} \tan^{-1}\frac{y-\eta}{y-\xi}\, ds = -V_\infty (x \cos\alpha + y \sin\alpha) \qquad (3.16b)$$

For the solution of the Dirichlet problem for the outer flow the following integral equation may be considered

$$-\pi \frac{d\mu}{dt} + \int_0^S \frac{d\mu}{ds} \frac{\partial}{\partial t}\left\{\tan^{-1}\frac{y-\eta}{x-\xi}\right\}ds = -V_\infty \left(\frac{\partial x}{\partial t}\cos\alpha + \frac{\partial y}{\partial t}\sin\alpha\right) + V_t \qquad (3.17)$$

The design problem as formulated before resolves itself into the solution of one of the equations (3.14) to (3.16) in combination with eq. (3.7) or eq. (3.17) under the constraint given by eq. (2.5). These integral equations are linear with respect to the unknown doublet strength but nonlinear with respect to the function representing the airfoil contour.

As has been remarked before the solution of the analysis problem of determining the velocity distribution along a given airfoil in uniform flow will be considered first.

# 4 THE ANALYSIS PROBLEM

## 4.1 The independent variable

Within the expressions for the integral equations as derived before the arclength measured along the airfoil contour appears as the independent variable. In a way the arclength may be the "natural" choice for the independent variable, but it is not at all obvious that this would also be the optimal choice from the numerical point of view. On the contrary, in this respect far better choices may be made. As the choice of x or y as independent variable will certainly be no improvement, it will be clear that a better choice will involve the application of a transformation.

Such a transformation is used in the elegant collocation method of reference 12 for the solution of integral equation (3.16a). This method is based on cubic spline representation of both airfoil contour and doublet strength. A transformation is applied in order to be able to apply a uniform mesh in the new independent variable. This may lead to accurate results using only a limited number of knots for the splines. A drawback of the method is however the violation of the main side condition for the development of a practical method as mentioned in the introduction, i.e. the wish to avoid the application of parameters which control the numerical accuracy but which have no direct physical meaning. This will be illustrated by means of a few results for a socalled supercritical shock-free airfoil.

Such an airfoil derives its special properties, which are of great importance for to-day aircraft design, from a rather peculiar behaviour of the curvature (see fig. 3). As a direct consequence of this behaviour the determination of an appropriate transformation to a new independent variable is rather cumbersome. The application of transformation 1 of figure 4 as proposed in reference 12, which involves the choice of a parameter B by the user of the method, leads to a rather unsatisfactory result when a relatively small number of knots for the contour spline is used. This is illustrated in figure 5a where in the nose region the deviation from the correct solution is depicted as a function of the x-coordinate. From the results obtained by means of a second additional transformation it becomes clear that a small number of knots will be

sufficient if a more appropriate transformation is applied (see Fig. 5b). The correct solution to which these results have been compared was obtained by applying the latter transformation and 160 knots for the spline. With these results it is illustrated that the application of a transformation to a new independent variable may indeed lead to a relatively short computing time but in general only at the cost of determining by trial and error an appropriate transformation. At the NLR this conclusion and the observation that the derivation of an appropriate transformation in three dimensions would be a highly complicated matter has led to the decision to forget about transformations and to return to the arclength as the independent variable.

4.2 The order of approximation

The method which is in use at the NLR at present, is based on first order approximations and has as a consequence a number of drawbacks which could be avoided when higher order approximations would be applied. As a first attempt to improve upon this situation a third order method was considered. It appeared however that the complexity of the evaluation of the integrals then leads to an increase in computing time which is not recovered by a decrease associated with a smaller number of parameters required for the representation of the contour and the doublet distribution. This observation together with the experiences reported in the literature (see e.g. Refs. 7, 8) have led to the decision to focus upon second order approximations.

4.3 Solution of the integral equations

4.3.1 General principles

The integral equations are solved by means of collocation methods. The airfoil contour is assumed to be given by a number n + 1 of discrete points, counting the trailing edge point at the lower side as the first point and the trailing edge point at the upper side as the last point. The second and consecutive points up to the nth point are used as collocation points (see Fig. 6) i.e. the points where the boundary condition is applied. Furthermore the independent variable s is replaced by the approximation $\tau$ of the arclength which is obtained by measuring the arclength along the straight line segments that connect the given

contour points.

The integrals are evaluated by summation of the subintegrals over the elements in which the airfoil contour is divided. The end points of these elements are the points $\{ \tau = t_i \mid t_1 = \tau_1;\ t_i = \frac{1}{2}(\tau_i + \tau_{i-1});$ $t_n = \tau_{n+1} \}$ midway between the given contour points $\tau = \tau_i$. On each element the coordinates of the airfoil contour and the doublet strength are approximated by a three terms Taylor series expansion:

$$f(\tau) = f(\tau_i) + \dot{f}(\tau_i)(\tau - \tau_i) + \tfrac{1}{2}\ddot{f}_i(\tau - \tau_i)^2 \qquad (4.1)$$

where the dots refer to differentiation to $\tau$.
The constants in this expression are determined from the different representations that have been considered. All representations will be of interpolating type, which means that the function considered will assume specified values at the knots of the representation. In all cases the knots for the contour representation are taken at $\tau = \tau_i$.

4.3.2 The integral equation resulting from the boundary condition $\phi_n = 0$

The integral equation reads according to eq. (3.14):

$$\int_0^S \frac{d\mu}{ds}\frac{\partial}{\partial n}\left\{\tan^{-1}\frac{y-\eta}{x-\xi}\right\}ds = -V_\infty \cos\alpha\frac{\partial x}{\partial n} - V_\infty \sin\alpha\frac{\partial y}{\partial n}$$

Here only the derivative of the doublet strength appears, so that it is obvious to consider a new function:

$$\gamma = -\frac{d\mu}{ds} \quad * \qquad (4.2)$$

To be consistent with the other approaches this function should vary linearly over the contour elements. Two different representations of this function have been considered. In the first representation the knots of the $\gamma$ distribution are taken at $\tau = \tau_i$ so that at the (i-1) element

$$\gamma(\tau) = \gamma_i + \dot{\gamma}(\tau_i)(\tau - \tau_i) \qquad (4.3)$$

where $\dot{\gamma}(\tau_i)$ is expressed in terms of $\gamma_{i-1}$, $\gamma_i$ and $\gamma_{i+1}$ by means

* The minus sign is introduced to be consistent with the usual definition of vorticity.

of divided differences.
In this way the $\gamma$ distribution is determined by n+1 unknows. As there are only n-1 collocation points which give rise to n-1 algebraic equations, two extra equations are needed. These equations are obtained by setting

$$\gamma_1 = \gamma_{n+1} = 0, \tag{4.4}$$

which is in accordance with the Kutta condition.
This approach is indicated in table 1 as case Ia.

In the second representation the knots of $\gamma$ are taken at $\tau = t_i$ so that at the (i-1)th element

$$\gamma(\tau) = \bar{\gamma}_{i-1} \frac{\tau - t_i}{t_{i-1} - t_i} + \bar{\gamma}_i \frac{\tau - t_{i-1}}{t_i - t_{i-1}}, \tag{4.5}$$

where $\bar{\gamma}_i$ refers to the value of $\gamma$ at $\tau = t_i$.
In this way the $\gamma$ distribution is determined by n unknows. The one extra equation that is needed here is obtained by setting

$$\bar{\gamma}_1 + \bar{\gamma}_n = 0 \tag{4.6}$$

This approach will be referred to as case Ib.
The second representation of $\gamma$ is such that this function is continuous across the edges of the contour elements, whereas this continuity is not ensured in the first representation. In both cases the derivatives needed for the contour approximation are taken from divided differences.

4.3.3 The integral equations resulting from the boundary condition $\psi = 0$

The first integral equation (3.15a) to be considered reads:

$$\int_0^S \mu \frac{\partial}{\partial s}\{\ell n\} r \, ds - \mu_w \ell n \, r_T = V_\infty (x \sin \alpha - y \cos \alpha).$$

The knots for the doublet distribution are taken at $\tau = \tau_i$, which gives rise to n+1 unknows $\mu_i$. The derivatives of $\mu$ are obtained by means of divided differences. This approach is referred to as case IIa.

The second integral equation to be considered (3.15b) reads after substitution of eq. (4.2):

$$\int_0^S \gamma \ln r \, ds = V_\infty \, (x \sin\alpha - y \cos\alpha). \qquad (4.7)$$

For $\gamma$ the representation of eq. (4.3) is used.
This approach is referred to as case IIb.

In both cases the contour representation is the same as in cases Ia,b. The two extra equations that are needed to complete the system of algebraic equations in both cases are obtained by extrapolating quadratically the first and last unknown from the three neighboring points.

4.3.4 The integral equations resulting from the boundary condition $\phi^- = 0$

Both integral equations that have been considered

$$\pi\mu - \int_0^S \mu \frac{\partial}{\partial s}\left\{\tan^{-1}\frac{y-\eta}{x-\xi}\right\} ds + \mu_w \tan^{-1}\frac{y-y_T}{x-x_r} = -V_\infty (x\cos\alpha + y\sin\alpha) \qquad (3.16a)$$

and

$$\pi(\mu + \mu_L) + \int_0^S \frac{d\mu}{ds} \tan^{-1}\frac{y-\eta}{x-\xi} ds = -V_\infty (x\cos\alpha + y\sin\alpha) \qquad (3.16b)$$

have been treated in the same way.
For each equation two different representations of the doublet distribution have been considered, together with analogous representations for the airfoil contour. In both representations the knots are taken at $\tau = \tau_i$.

In the first representation the derivatives for the local expansions are obtained by means of divided differences. In the second representation a quadratic spline fit is used for both the doublet distribution and the airfoil contour in order to determine the derivatives. In both cases n+1 unknows have to be determined. The two extra equations are obtained by setting the first derivative of $\mu$ at the trailing edge points at upper and lower side equal to the first derivative at the neighboring points. The difference between the two representations is that there exist discontinuities in function and derivatives across the edges of the contour elements at the first

representation, while at the second representation both the function and its first derivative are continuous. In the following these approaches are referred to as methods IIIa, b and IIIc, d respectively.

## 4.4 Numerical results

It is rather difficult to establish a suitable norm for judging the practical usefulness of a particular method. However, with respect to the accuracy of the results a few general rules may be maintained:

- As a matter of course, the method should procedure increasingly better approximations of the true solution if the number of knots is increased which implies that the method should be able to produce an approximation of known exact solutions up to a given accuracy.
- Ever for a relatively small number of knots the method should procedure reliable results for airfoils which have fairly extreme geometrical qualities such as the appearance of large curvatures or a very small thickness compared to the chordlength.
- The accuracy of the solution should be of the same order of magnitude all over the contour with a possible exception of the region near the trailing edge.*

From the investigations carried out at the NLR it has appeared that a fairly severe test case is formed by a thin cambered airfoil belonging to the class of socalled Karman-Treftz airfoils (see e.g. Ref. 13). These are obtained by conformal mapping from a circle

$$z = z_o + ae^{i\theta} \tag{4.8}$$

using the transformation

$$\frac{\zeta - pc}{\zeta + pc} = \left(\frac{z - c}{z + c}\right)^p \tag{4.9}$$

with

$$z_o = -a\,(\varepsilon - i\gamma)$$

---

* The real flow close to the trailing edge will be largely influenced by viscous effects, so that the user of the methods considered will not be very interested in an accurate calculation of the local potential flow as long as it does not affect the solution on the remainder of the airfoil.

$$a = \frac{(1-\gamma^2)^{\frac{p}{2}} - \varepsilon^p}{p\,(\sqrt{1-\gamma^2} - \varepsilon)(1-\gamma^2)^{\frac{p}{2}}} \qquad (4.10)$$

$$c = a\,(\sqrt{1-\gamma^2} - \varepsilon)$$

The testcase considered has been obtained by setting

$$\varepsilon = .05 \quad \gamma = .1 \quad p = 1.99$$

and taking an angle of attack of $6^o$.

The methods described in section 4.3 and listed in Table 1 have been applied to this testcase; the results are compared to the exact solution in figures 7 to 15. In these figures the pressure coefficient defined by

$$C_p = 1 - (\phi_t^+)^2$$

is depicted as a function of x/c where c is the chordlength of the airfoil. The tangential velocity $\phi_t^+$ is calculated from

$$\phi_t^+ = -\pi \frac{d\mu}{dt} + \int_0^S \frac{d\mu}{ds} \frac{\partial}{\partial t}\left\{\tan^{-1}\frac{y-\eta}{x-\xi}\right\} ds + \bar{V}_\infty \cdot \bar{t} \qquad (4.11)$$

which expression is obtained by formal differentiation of eq. (3.10a) after substitution of eq. (3.12).

From figure 7 may be concluded that the application of method Ia leads to a very poor result. Apparently this is partly due to the discontinuities in the $\gamma$-representation as may be concluded from the results of method Ib given in figure 8. Here an approximation of the exact solution is obtained which shows more or less the right behaviour except in the region near the trailing edge. The reason for this defect becomes clear if it is realised that the analysis problem as it has been formulated, deteriorates into an underdetermined problem if the thickness of the airfoil tends to zero. So, it may be expected that the small thickness of the airfoil especially near the trailing edge causes the system of algebraic equations to be nearly linearly dependent. As appears from figure 9 this problem may be solved by reducing the order of the system by deflating the matrix. Then a result is obtained which shows a reasonable agreement with the exact solution.

When method IIa is applied, the approximated pressure distribution has a reasonable accuracy except in the noseregion where appreciable deviations from the true solution appear (see fig. 10). The alternative method IIb leads to a much better overall result as can be seen in figure 11.

Figure 12 shows that method IIIa leads to a rather poor result with a very large inaccuracy near the nose, while the result of method IIIc is also poor but then especially near the trailing edge.

It has been attemped to improve the results of the methods Ia, IIa and IIIa, c by increasing the number of specified points(n). In case IIIa this established a much better accuracy as may be seen in figure 14. With the other methods no appreciable improvement was obtained.

When quadratic spline approximations of both the airfoil contour and the doublet distribution are used, the alternative integral equations (3.16a) and (3.16b) become equivalent in the discretized form also, because of the continuity of both function and first derivative across the contour elements. This observation is confirmed by the results of methods IIIb and IIId which lead to exactly the same solution. Figure 15 shows that this solution has a reasonable overall accuracy.

## 4.5 Conclusisons

From the numerical results reviewed in the former section it may be concluded that there are four methods which can be considered as good alternatives when the overall accuracy of the solution for a given number of points is taken as a norm. These methods are Ib with deflation, IIb and IIIb or IIId.

Now some remarks made in section 1 may be recalled. Firstly, the method should be economic, secondly the method should be extensible to three dimensions.

The methods as described here involve approximately the same amount of computing time when using the same number of knots and as such may be considered equivalent. However, refinements with respect to the numerical evaluations may lead to a saving of computing time.

Possibly such an improvement may be obtained by developing a more economic method for the solution of the ultimate system of algebraic

equations. When large systems are considered iterative methods are very attractive if the number of iterations is much smaller than the number of equations. But iterative methods can only be applied successfully if the matrix of the system of equations has a suitable structure. Now, apart from the deflation in method Ib, which already needs extra computing time, the matrices resulting from application of the methods Ib and IIb are structured such that the success of iterative methods is questionable. The matrix resulting from method IIIb, d however has a more promising structure.

A second objection against methods Ib and IIb is the fact that the formulation leads to the determination of a function $\gamma$. In three dimensions this function would be replaced by a vetorial quantity. This involves the determination of two separate functions which doubles the number of unknows as compared with the doublet distribution itself. Though this does not lead to insurmountable difficulties, the doublet distribution will be preferred if the alternatives are more or less equivalent otherwise.

These observations lead to the inevitable conclusion that method IIIb should be followed for the solution of the analysis problem. Having arrived at this conclusion a further verification was made by applying method IIIb for the determination of the pressure distribution on the quasi-elliptical airfoil of figure 3. In table 2 the results for 40 knots are compared with a reference solution obtained with the method of reference 12 using 160 knots and with the result of the latter method using 40 knots. It can be observed that the present method leads to a slightly better overall accuracy than the method of reference 12 using the same number of knots. This has to be due to the fact that the present method involves the choice of knots based on physical insight of the user, while within the method of reference 12 the location of the knots is determined by the applied transformation.

## 5 THE DESIGN PROBLEM

### 5.1 Basic idea

The problem to be solved is to determine the shape of an airfoil

such that it will procedure a specified velocity distribution. It is assumed that the velocity distribution on all the airfoil elements will be given as a function of the x-coordinate, together with a guess of the true shape of the airfoil contour. One point of the airfoil contour will be fixed. In addition it will be required that the deviation between the true shape and the initial guess will be small. The latter requirement serves merely as an example of constraint on the geometry. As will be shown later on with the aid of some numerical examples there will be probably no need for such a constraint, from numerical point of view, if the prescribed velocity distribution is realistic in the sense that it can be generated by an airfoil.

At NLR the investigations on improvement of the calculation method for the analysis problem and the search for a useful method for the solution of the design problem took place more or less simultaneously. Thus, the conclusion of the former section was not reached yet when a first attempt was made to solve the design problem (Ref. 14). This is the reason why the method to be described here differs at a number of points from the method that would have been applied if the conclusion had been reached at an earlier stage. These points are the following:

- for simulating the flow a vorticity distribution is used in stead of a doublet distribution which implies that the perturbation velocity potential is given by eq. (3.12) after substitution of eq. (4.2).
- the Neumann boundary condition is replaced by $\phi_t^- = 0$ instead of by $\phi^- = 0$.
- the airfoil contour as well as the vorticity distribution are represented by cubic Hermite polynomials instead of quadratic splines.

According to eq. (4.11) after substitution of eq. (4.2) the tangential velocity at the outer side of the airfoil contour is given by:

$$\phi_t^+ = \pi\gamma(t) - \int_0^S \gamma(s) \frac{\partial}{\partial t}\left\{\tan^{-1}\frac{y-\eta}{x-\xi}\right\} ds + \bar{V}_\infty \cdot \bar{t} \qquad (5.1)$$

while the velocity at the inner side is given by:

$$\phi_t^- = -\pi\gamma(t) - \int_0^S \gamma(s)\,\frac{\partial}{\partial t}\left\{\tan^{-1}\frac{y-\eta}{x-\xi}\right\} ds + \bar{V}_\infty \cdot \bar{t} \tag{5.2}$$

Applying the conditions

$$\phi_t^+ \approx V_t \tag{5.3}$$

and

$$\phi_t^- = 0 \tag{5.4}$$

leads then to a set of integral equations which are linear with respect to the unknown vorticity $\gamma$ and non-linear with respect to the unknown airfoilcontour given by $y(x)$. These equations have to be solved under the constraint

$$y(x) \approx \bar{y}\,(x) \tag{5.5}$$

Now, minimizing the least squares functional

$$F = \int_0^S (F_t^2 + F_n^2 + F_c^2)\,dt \tag{5.6}$$

where

$$F_t = W_t\,(\phi_t^+ - V_t) \tag{5.7}$$

$$F_n = W_n\,\phi_t^- \tag{5.8}$$

$$F_c = W_c\,(y - \bar{y}\,) \tag{5.9}$$

the functions $y$ and $\gamma$ and the angle of attack $\alpha$ will be determined such that the equations (5.3) and (5.4) will be fulfilled in a least squares sense.

The fact that eq. (5.4) is fulfilled in a least squares sense instead of by collocation simplifies the optimization procedure and introduces some extra flexibility.

The weighting functions W may be used to counter balance the different requirements.

5.2 Numerical approximations

At each contour element the airfoil co-ordinates and the vorticity are represented by a cubic Hermite polynominal:

$$f(\tau) = \sum_j \{f_j h1_j(\tau) + f'_j h2_j(\tau) + f_{j+1} h3_j(\tau) + f'_{j+1} h4_j(\tau)\} \quad (5.10)$$

Here $f_j$ and $f_{j+1}$ are the values of f and $f'_j$, $f'_{j+1}$ are the derivatives of f with respect to $\tau$ at the knots j and j+1 respectively.
The cubic functions h are defined by:

$$h1_j = (1 + 2u)(1 - u)^2 \qquad h3_j = (3 - 2u)\, u^2$$
$$h2_j = (\tau_{j+1}-\tau_j)u(1-u)^2 \qquad h4_j = (\tau_j-\tau_{j+1})(1-u)\, u^2 \quad (5.11)$$

with

$$u = \frac{\tau - \tau_j}{\tau_{j+1} - \tau_j} \quad (5.12)$$

As the contour is unknown a priori and updating the parameter choice during the minimization process would be too costly, $\tau$ is chosen to be the arclength measured along the initial guess of the airfoil contour.

Because any airfoil contour can be derived from an initial guess by changing the function $y(\tau)$ only, the values of y and y' and $\gamma$ , $\gamma'$ at the knots and the angle of attack $\alpha$ are the parameters with respect to which the functional F will have to be minimized.

5.3 Solution to the minimization problem

The minimization problem is solved by means of a method that has been applied to several different problems at NLR and that has proven to be very efficient (see Ref. 15). It is based on a method proposed by Fletcher (Ref. 16). The generalized inverse of the matrix of derivatives of F to the minimization parameters is determined following a method according to Lawson and Hanson (Ref. 17).

The function F of eq. (5.6) has essentially the form:

$$F = \sum_{k=1}^{N} \{f_k(\bar{z})\}^2$$

where $\bar{z}$ is the vector of the parameters $y_j$, $y'_j$, $\gamma_j$, $\gamma'_j$ and $\alpha$ and where

the $f_k$ form together a vector $\bar{f}$.

This function is minimized by applying the follwing computational algorithm:

1) given $\bar{z} = \bar{z}^{(i)}$ compute the vector of residuals $\bar{f}^{(i)}$ and the matrix of derivatives $J^{(i)}$ ; determine a searchdirection by computing $\bar{s}^{(i)} = -J^{(i)+} \bar{f}^{(i)}$ where $J^{(i)+}$ is the generalized inverse of $J^{(i)}$

2) set $\bar{z}^{(i+1)} = \bar{z}^{(i)} + \lambda \bar{s}^{(i)}$ and determine $\lambda > 0$ such that $F(\bar{z}^{(i+1)}) < F(\bar{z}^{(i)})$

3) if preset conditions with respect to the variations of the parameters z or with respect to the residuals $f_k$ are fulfilled convergence is considered to be attained; then terminate the iteration process, otherwise repeat from 1).

## 5.4 Numerical results

The possibilities which the method outlined above offers for the solution of the design problem were explored to some extent by application to several cases. Here the solutions to five example cases are presented.

Example 1

The objective of the first example is to show that the method is capable of reconstructing a known geometric shape from the corresponding pressure distribution. Starting with an ellipse as initial guess of the geometry to be determined and with the exact pressure distribution of a circular cylinder in uniform flow as target, the circular shape is reconstructed straightforwardly as shown in figure 16. In this case the weighting functions were chosen as $W_t \equiv W_n \equiv 1$ and $W_c \equiv 0$, showing that there is no need to apply the geometric constraint of eq. (5.5) if the prescribed pressure distribution is known to be realistic in the sense mentioned before.

Example 2

The second example demonstrates compatibility between the design method described above and the analysis method based on eq. (5.4). By means of this method the velocity distribution on an airfoil-flap configuration has been calculated. Prescribing the latter distribution as "target" velocity distribution and starting with a combination of two

symmetric profiles as initial guess of the airfoilshape, the application of the design method results in the reconstruction of the airfoil-flap configuration and the velocity distribution in three iteration steps as shown in figure 17. During the iteration process the mutual position of the airfoil elements is determined by fixing the trailing edges, i.e. the trailing edges of the initial airfoil have the same co-ordinates as those of the airfoil, resulting from the iteration process.

Example 3

The next example shows how the present method could be used during a design process. It is supposed that a designer wants to alter the velocity distribution of figure 17 into one with a higher velocity at the upper side of the main profile and arooftop like distribution on the flap. Starting with the airfoil of figure 17 as initial guess and fixing one point on the flap in the neighbourhood of the main profile and the trailing edge of the main profile, the iteration process results in a rather slight change of the airfoil (fig. 18a) which produces apparently not completely the desired velocity distribution (fig. 18b).

Example 4

This example shows a possible application of the geometrical constraint of eq. (5.5). The velocity distribution of example 3 is chosen as target again, but now it is tried to maintain the shape of the main profile and the mutual position of the elements and to reshape the flap only. Figure 19 shows the results for three different values of $W_c$ on the main profile. It can be seen that the increase of the weight factor $W_c$ for the main profile leads eventually to maintaining the shape of the forward airfoil element. However this involves increasing deviations between real and target velocity distribution.

Example 5

In the last example the geometry of figure 17 is chosen as a start and the velocity distribution of example 3 is chosen as target again and the shape of the main profile is fixed by means of the weightfactor $W_c$. But now the position of the flap is fixed in x-direction only, so that it may rotate and translate in y-direction. The velocity that can be realised in this way has a strong resemblance to the velocity distribution resulting in the former example (with $W_c = 100$) except in

the region of the rooftop target distribution. However, the resulting shape and position of the flap are very different from those of the former example (see fig. 20).

### 5.5 Conclusions

It can be concluded from the results that the approach described here holds very promising aspects for the process of wing design. The results indicate that there is probably no need to apply a geometrical constraint if realistic velocity distributions are prescribed. In the event of uncertainty about the possibility to generate a realistic shape, the geometrical constraint as introduced here can be applied. Some results of application of this constraint for maintaining part of the airfoil shape have been shown.

In the present method the optimization parameters have been limited to comprise the ordinates of the airfoil and the angle of attack. In principle, however, any other parameters such as flap angle or mutual position of the airfoil elements could be included. The geometrical constraint as applied in the present method is rather simple, but the basic idea of the method allows certainly the application of more complicated constraints such as positive thickness and range of angle of attack.

Finally it should be remarked that the numerical representations applied in the present method are essential only as far as they ensure a good approximation of the functional. They are not essential to the optimization process and as such may be replaced by any other appropriate representation.

## 6 FINAL REMARKS

A review has been given of investigations carried out at NLR aiming at a well-founded choice for the apporach to be followed for the development of an improved integral equation method for the solution of both the analysis and the design problem. It has been concluded that the approach should be based on a doublet distribution for simulation of the flow and quadratic spline approximations for both this

distribution and the airfoil contour. The integral equations to be solved should be those associated with the Dirichlet problems for the inner and outer region. The optimization procedure as described seems to be very well suited for the solution of the design problem.

## 7 REFERENCES


1 Kellogg, O.D. — Foundations of potential theory
Fred. Ungar Publ. Cy. New York (1929)

2 Smith, A.M.O.
Pierce, J. — Exact solution of the Neumann problem.
Calculation of non-circulatory plane and axially symmetric flows about or within arbitrary boundaries.
Douglas Aircraft Cy. Rep. No Es 26988 (1958)

3 Rubbert, P.E. et al — Ageneral method for determining the aerodynamic characteristics of fan-in-wing configurations
The Boeing Co Rep. D 6-15047-1 (1967)

4 Baurdoux, H.I.
Boerstoel, J.W. — Symmetrical transonic potential flows around quasi-elliptical airfoil sections
NLR TR 69007U (1968)

5 Labrujère, Th. E.
Loeve, W.
Slooff, J.W. — An approximate method for the calculation of the pressure distribution on wing-body combinations at sub-critical speeds
AGARD CP No. 71 paper 11 (1970)
NLR MP 70014U (1970)

6 Ebihara, M. — A method for calculation of lifting potential flow problems. Part I. Theoretical basis
NAL TR-240T (1971)

7 Hess, J.L. — Consistent velocity and potential expansions for higher order surface singularity methods
Douglas Aircraft Cy Rep. No. MDC J6911 (1975)

8 Johnston, F.T.
Rubbert, P.E.
Advanced paneltype influence coefficient methods applied to subsonic flow.
AIAA paper 75-50 (1975)

9 Bristow, D.R.
A new surface singularity method for multi-element airfoil analysis and design.
AIAA paper 76-20 (1976)

10 Morchoisne, Y.
Methode de calcul inverse en bidimensionelle compressible.
A.A.A.F. 12ème Colloque d'Aérodynamic Appliquée (1975)

11 Van der Plaats,G.H.
Hicks, R.M.
Murman, E.M.
Application of numerical optimization techniques to airfoil design
NASA SP-347, part II (1975)

12 Botta, E.F.F.
Calculation of potential flow around bodies
Thesis, Math. Inst. University of Groningen (1978)

13 Durand, W.F.
Aerodynamic Theory
Vol. II Div. E. Springer, Berlin (1935)

14 Labrujère, Th. E.
Multi-elemnt airfoil design by optimization
NLR MP 78023 (1978)

15 Boerstoel, J.W.
Huizing, G.H.
An approximate conformal mapping method using harmonic polynomials
NLR MP 71019U (1971)

16 Fletcher, R.
Generalized inverse methods for the best least squares solution of systems of non-linear equations
Comp. J. 10 pp. 392 - 399 (1968)

17 Lawson, C.L.
Hanson, R.J.
Solving least squares problems
Prentice Hall Inc. (1974)

TABLE 1

Alternative methods for solving the analysis problem

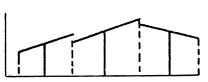
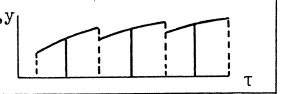
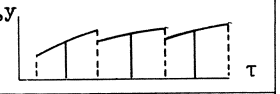
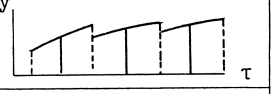
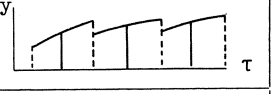
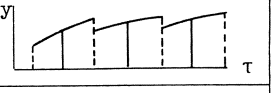
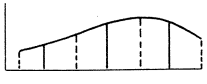
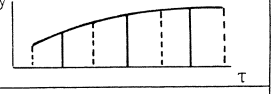
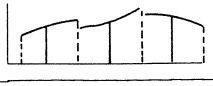
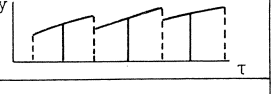
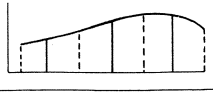
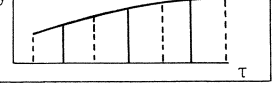
| Boundary condition | Method | Integral equation | Representations: Unknown function | Representations: Contour |
|---|---|---|---|---|
| $\phi_n = 0$ | Ia | $\int_0^T \gamma \frac{\partial}{\partial n}\left\{\tan^{-1}\frac{y-\eta}{x-\xi}\right\}\frac{ds}{d\tau}\,d\tau = \bar{V}_\infty.\bar{n}$ |   |   |
| | Ib | |   |   |
| $\psi = 0$ | IIa | $\int_0^T \mu \frac{\partial}{\partial\tau}\{\ln r\}d\tau - \mu_w \ln r_T = \psi_\infty$ |   |   |
| | IIb | $\int_0^T \gamma \ln r \frac{ds}{d\tau}\,d\tau = \psi_\infty$ |   |   |
| $\phi^- = 0$ | IIIa | $\pi\mu - \int_0^T \mu \frac{\partial}{\partial\tau}\left\{\tan^{-1}\frac{y-\eta}{x-\xi}\right\}d\tau + \mu_w \tan^{-1}\frac{y-y_T}{x-x_T} = -\phi_\infty$ |   |   |
| | IIIb | |   |   |
| | IIIc | $\pi(\mu+\mu_L) + \int_0^T \frac{d\mu}{d\tau}\tan^{-1}\frac{y-\eta}{x-\xi}\,d\tau = -\phi_\infty$ |   |   |
| | IIId | |   |   |

TABLE 2

Comparison of results obtained by the present method IIIb and results obtained by the method of Ref.12 using 40 knots

| x/c | $Cp_r$ (Ref.12;n=160) | $Cp_c$ (IIIb) | $Cp_c-Cp_r$ | $Cp_c$ (Ref.12;n=40) | $Cp_c-Cp_r$ |
|---|---|---|---|---|---|
| .0043 | .541 | .542 | .001 | .541 | .000 |
| .0100 | .027 | .027 | .000 | .026 | -.001 |
| .0159 | -.374 | -.372 | .002 | -.373 | .001 |
| .0224 | -.654 | -.650 | .004 | -.654 | .000 |
| .0317 | -.809 | -.809 | .000 | -.809 | .000 |
| .0384 | -.840 | -.840 | .000 | -.841 | -.001 |
| .0419 | -.847 | -.846 | .001 | -.849 | -.002 |
| .0451 | -.842 | -.842 | .000 | -.842 | .000 |
| .0478 | -.825 | -.822 | .003 | -.819 | .006 |
| .0521 | -.758 | -.758 | .000 | -.762 | -.004 |
| .0650 | -.629 | -.632 | -.003 | -.630 | -.001 |
| .0899 | -.541 | -.541 | .000 | -.541 | .000 |
| .1173 | -.498 | -.499 | -.001 | -.499 | -.001 |
| .1689 | -.459 | -.457 | .002 | -.459 | .000 |
| .3577 | -.375 | -.375 | .000 | -.376 | -.001 |
| .6524 | -.242 | -.238 | .004 | -.250 | -.008 |
| .8220 | -.085 | -.085 | .000 | -.077 | .008 |
| .8970 | .055 | .066 | .011 | .029 | -.026 |
| .9509 | .218 | .214 | .004 | .229 | .011 |

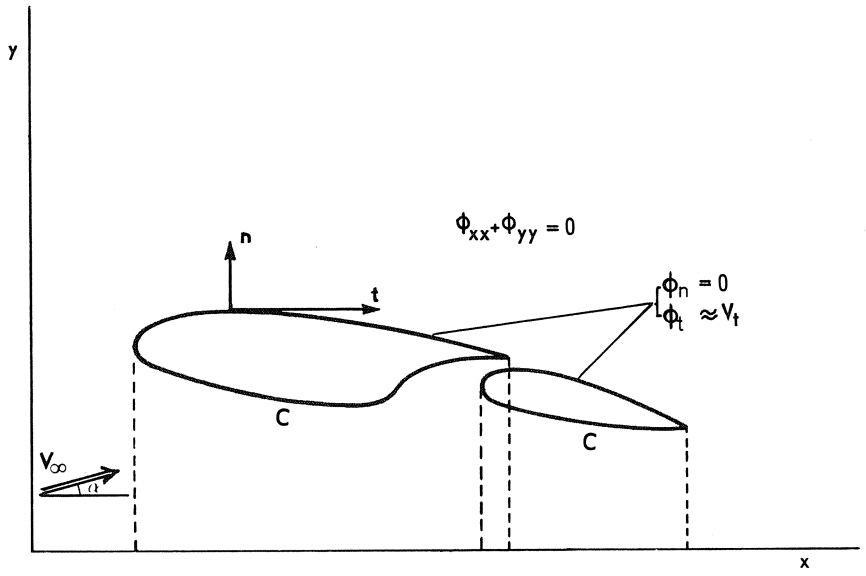



Fig. 1 The flow problem

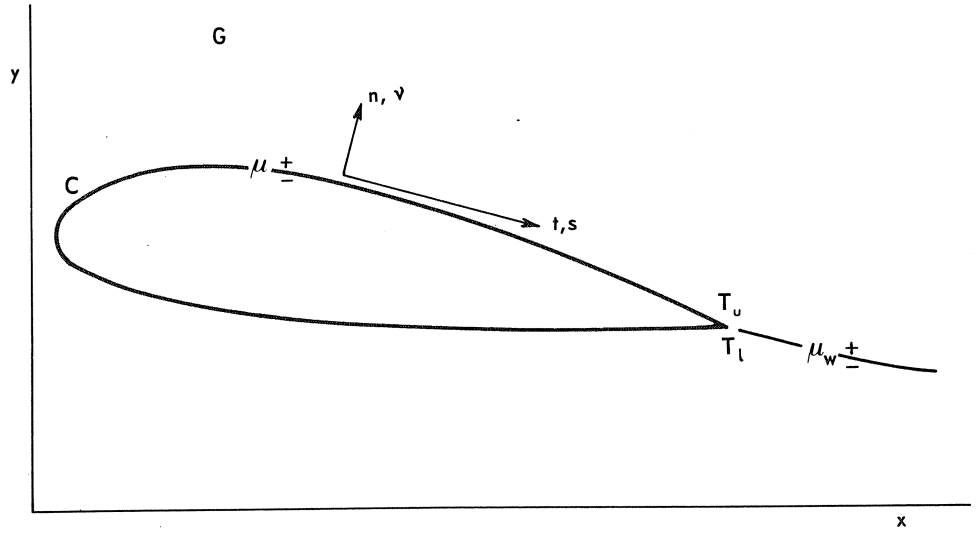



Fig. 2 The doublet distribution

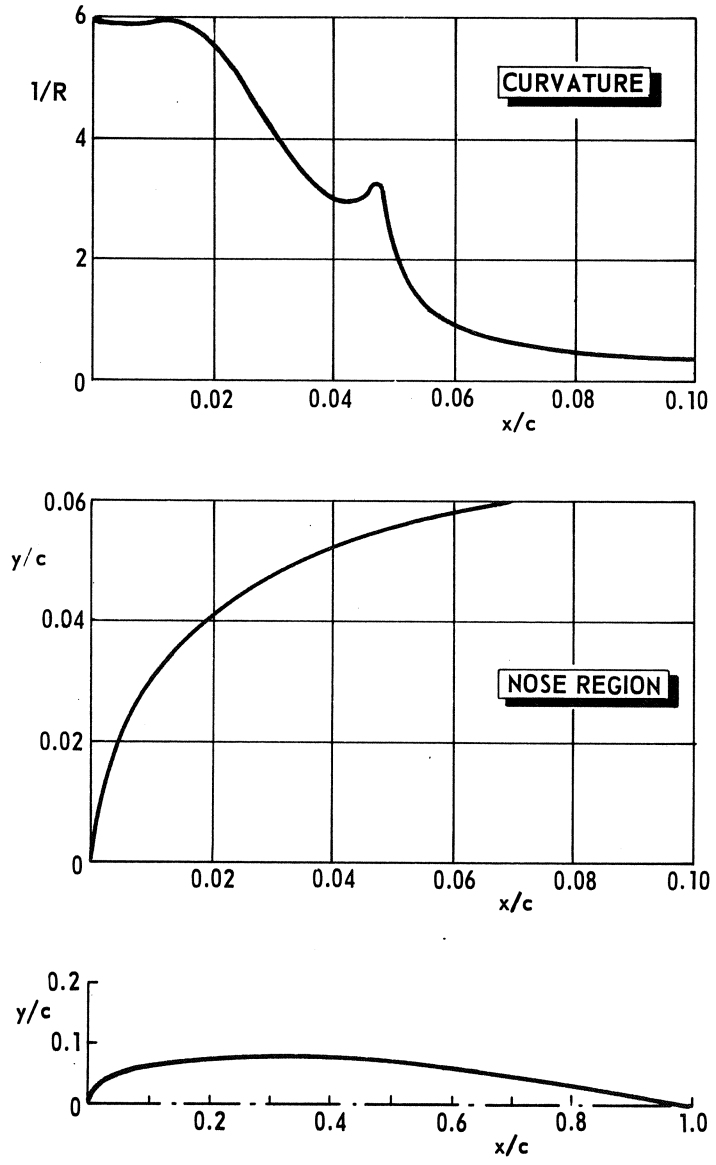
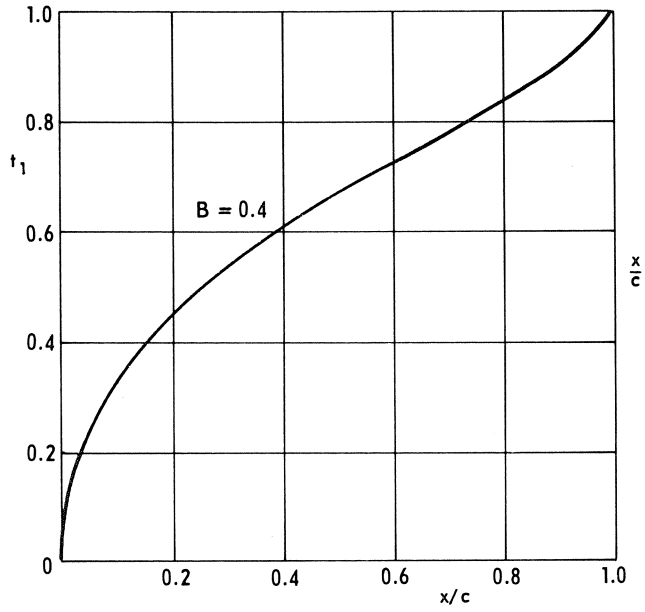
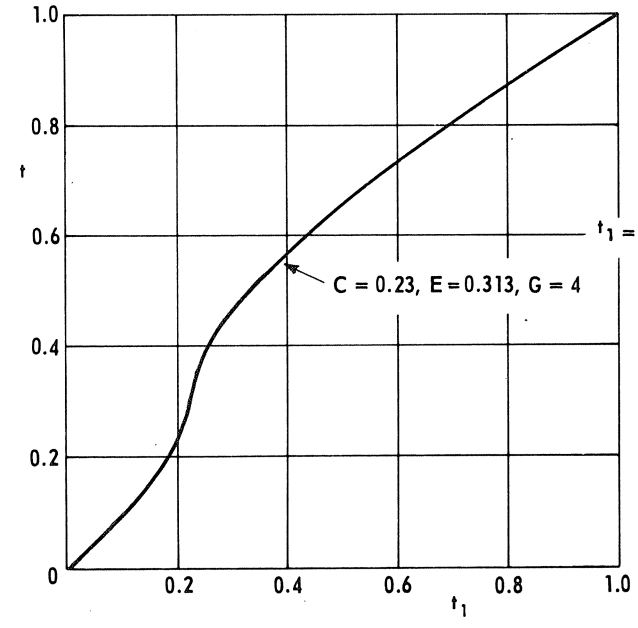



Fig. 3 Quasi-elliptical airfoil 0.1025-0.6750-1.3750 (Ref.4)




TRANSFORMATION 1

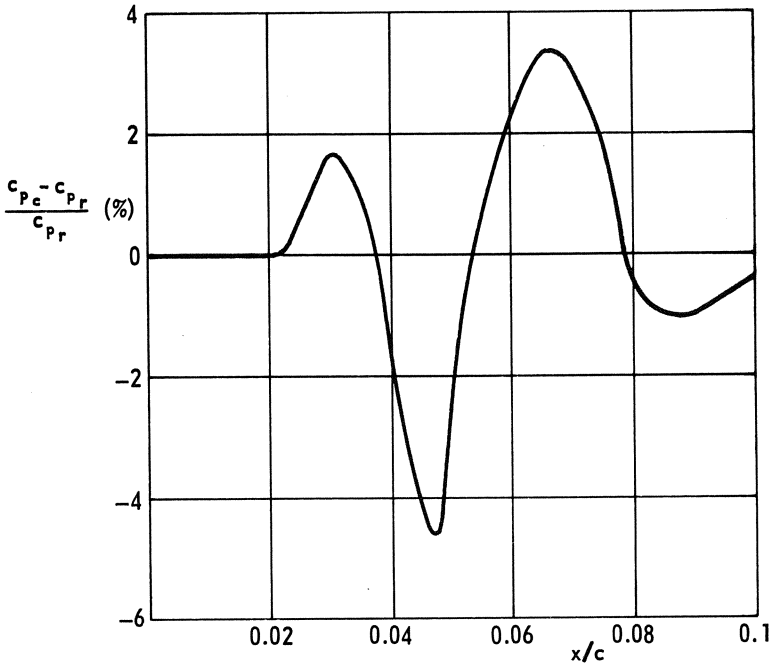
$$\frac{x}{c} = 1 - \left\{1 - Bt_1^2 + (B-1)\,t_1^4\right\}^2$$



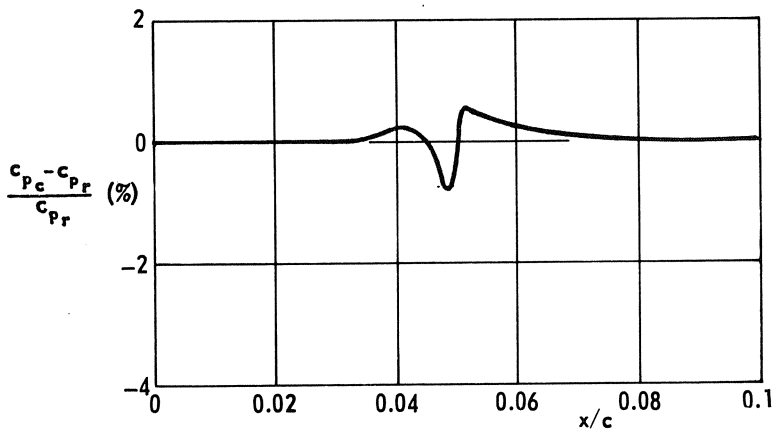

TRANSFORMATION 2

$$t_1 = C + D\,(t-E) + F\tan^{-1}\left\{G\,(t-E)\right\}$$

Fig. 4 Transformation to a new independent variable




a) TRANSFORMATION 1




b) TRANSFORMATION 2

Fig. 5 Accuracy of the solution obtained with the method of Ref. 12 based on 40 knots for the spline representing the contour

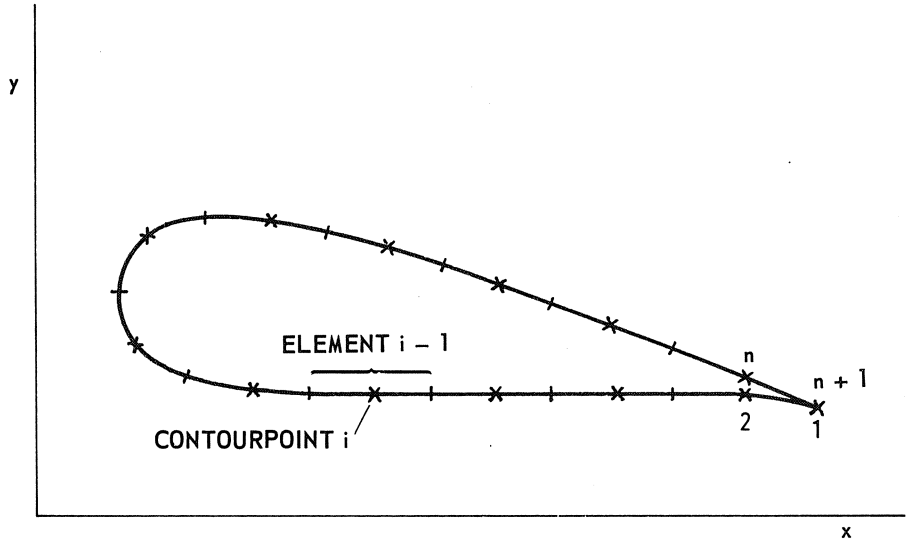



Fig. 6 Location of elements on airfoil contour

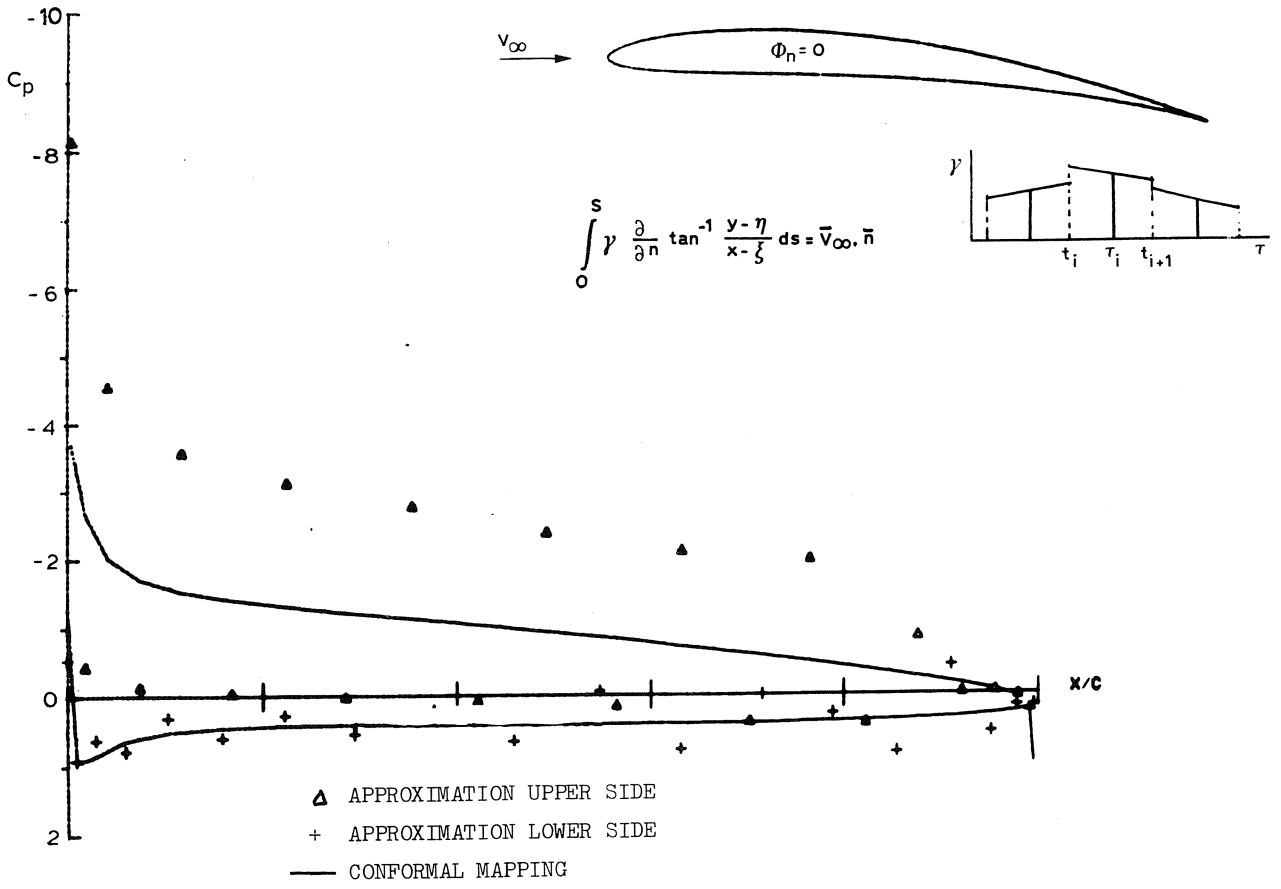



Fig. 7 Pressure distribution resulting from application of method Ia (n = 40)

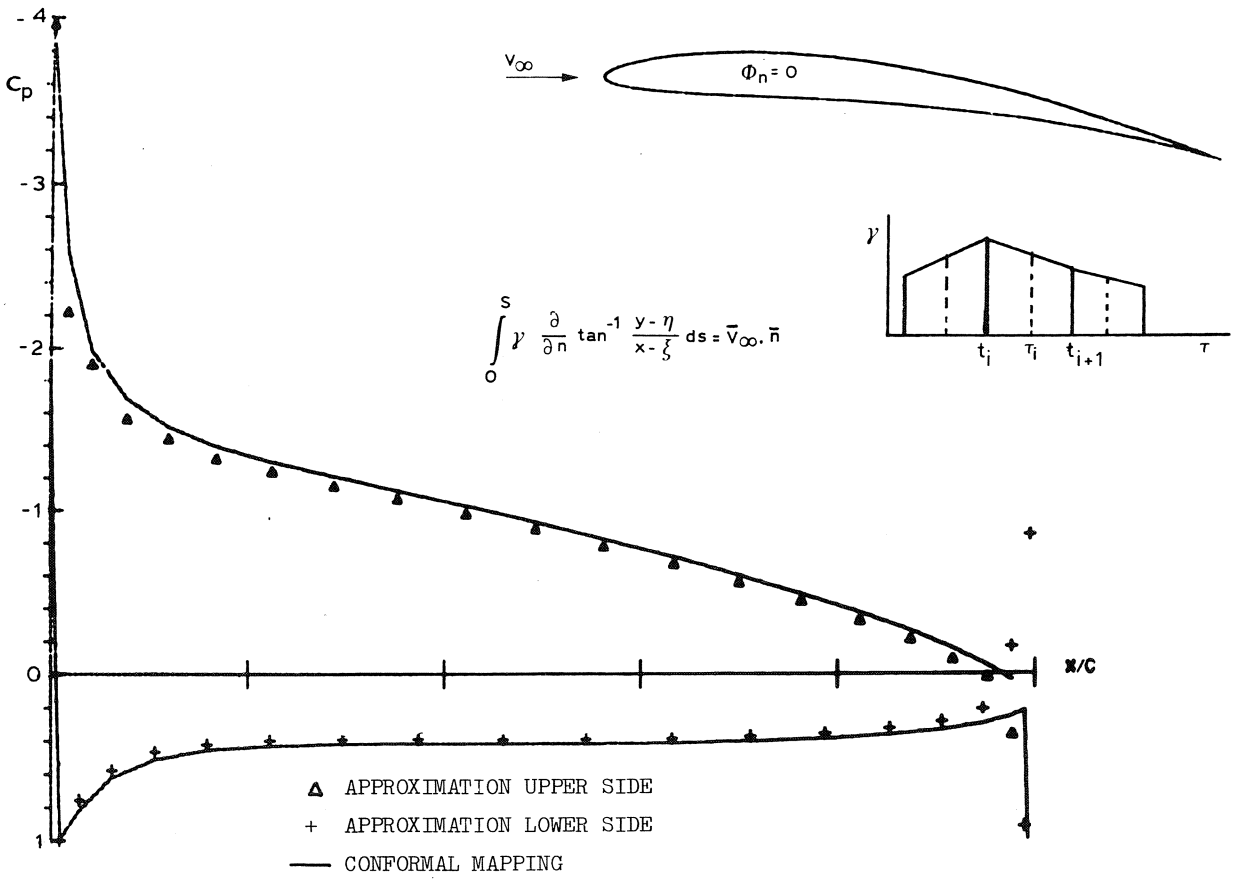



Fig. 8 Pressure distribution resulting from application of method Ib (n = 40)

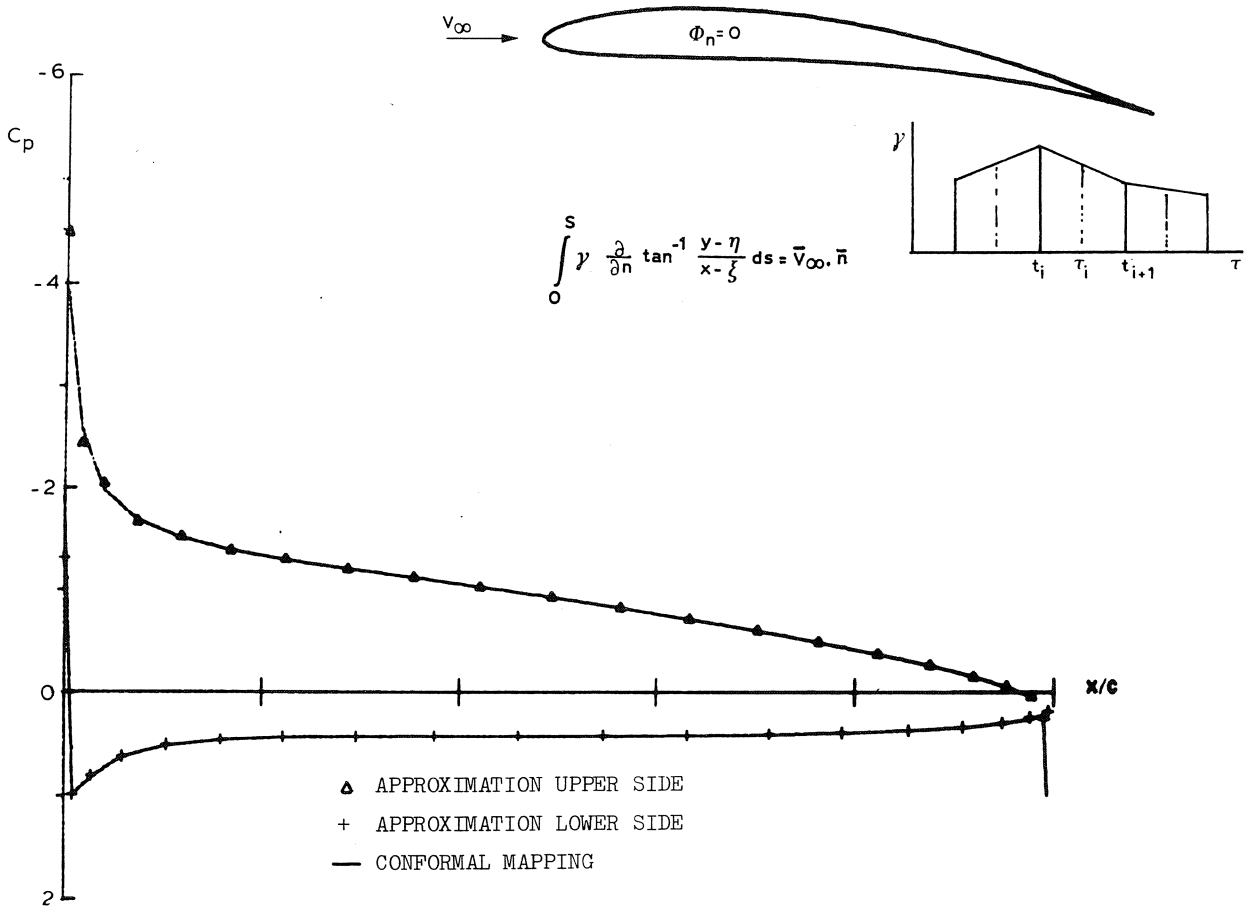



Fig. 9 Pressure distribution resulting from application of method Ib with deflation (n = 40)

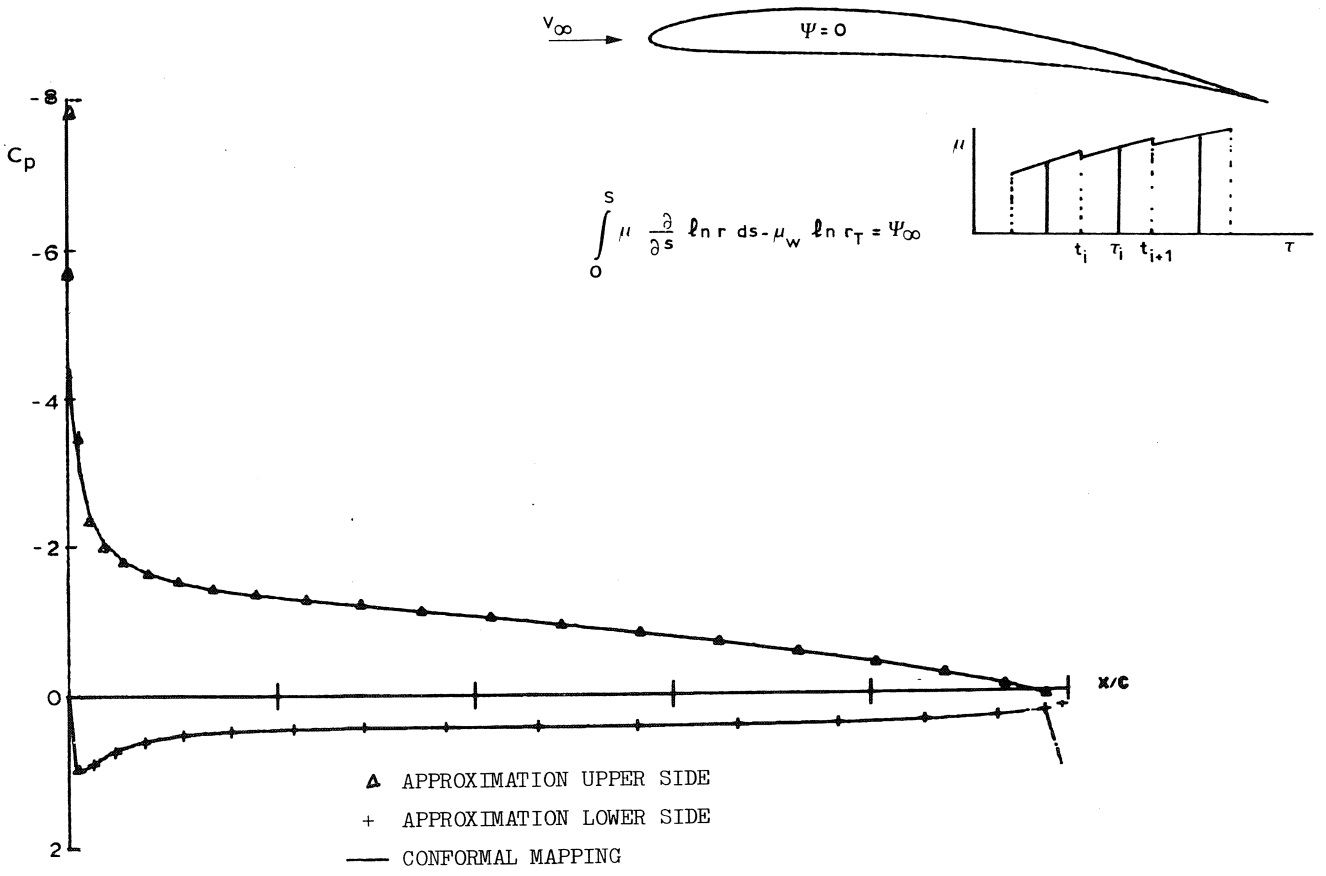



Fig. 10 Pressure distribution resulting from application of method IIa (n = 40)

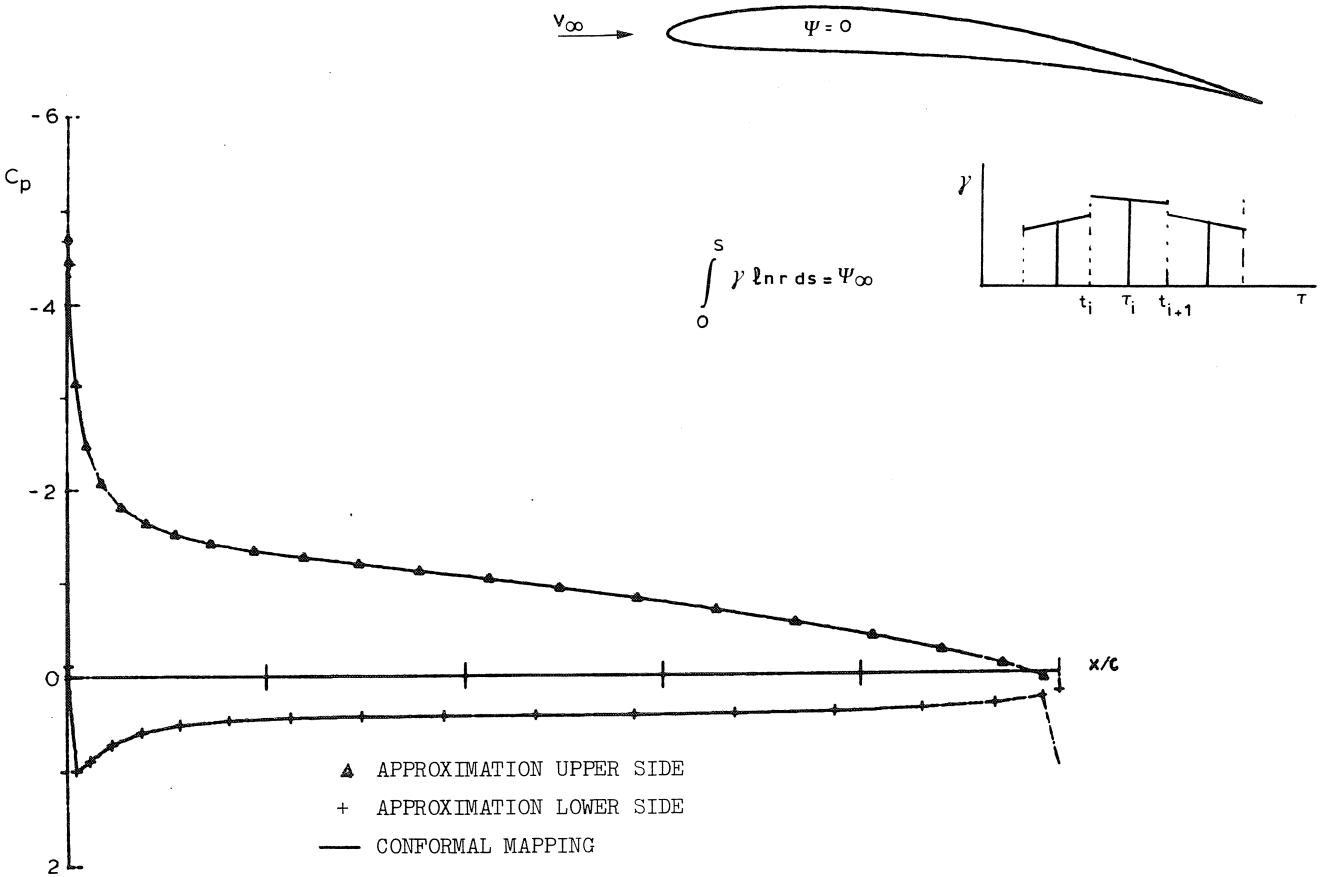



Fig. 11 Pressure distribution resulting from application of method IIb (n = 40)

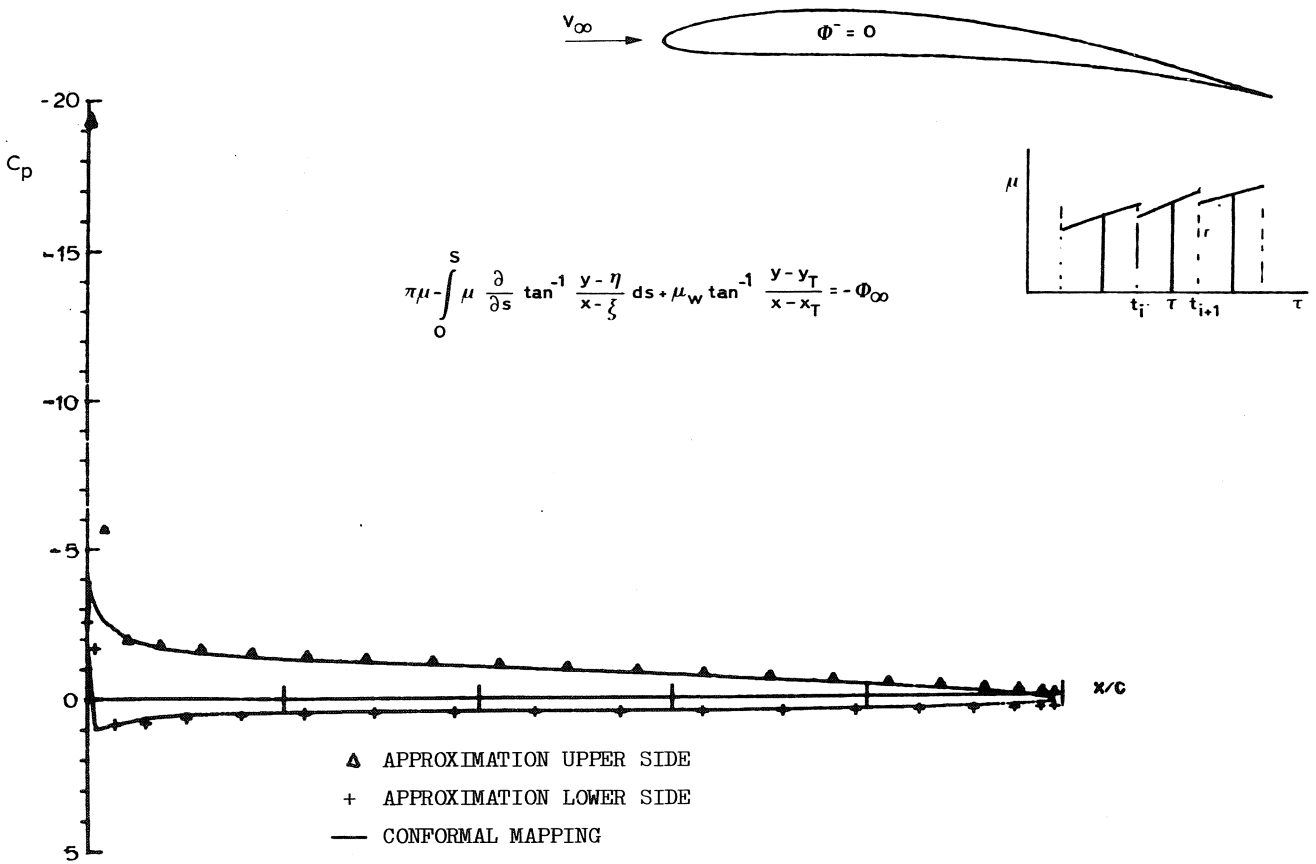



**Fig. 12 Pressure distribution resulting from application of method IIIa (n = 40)**

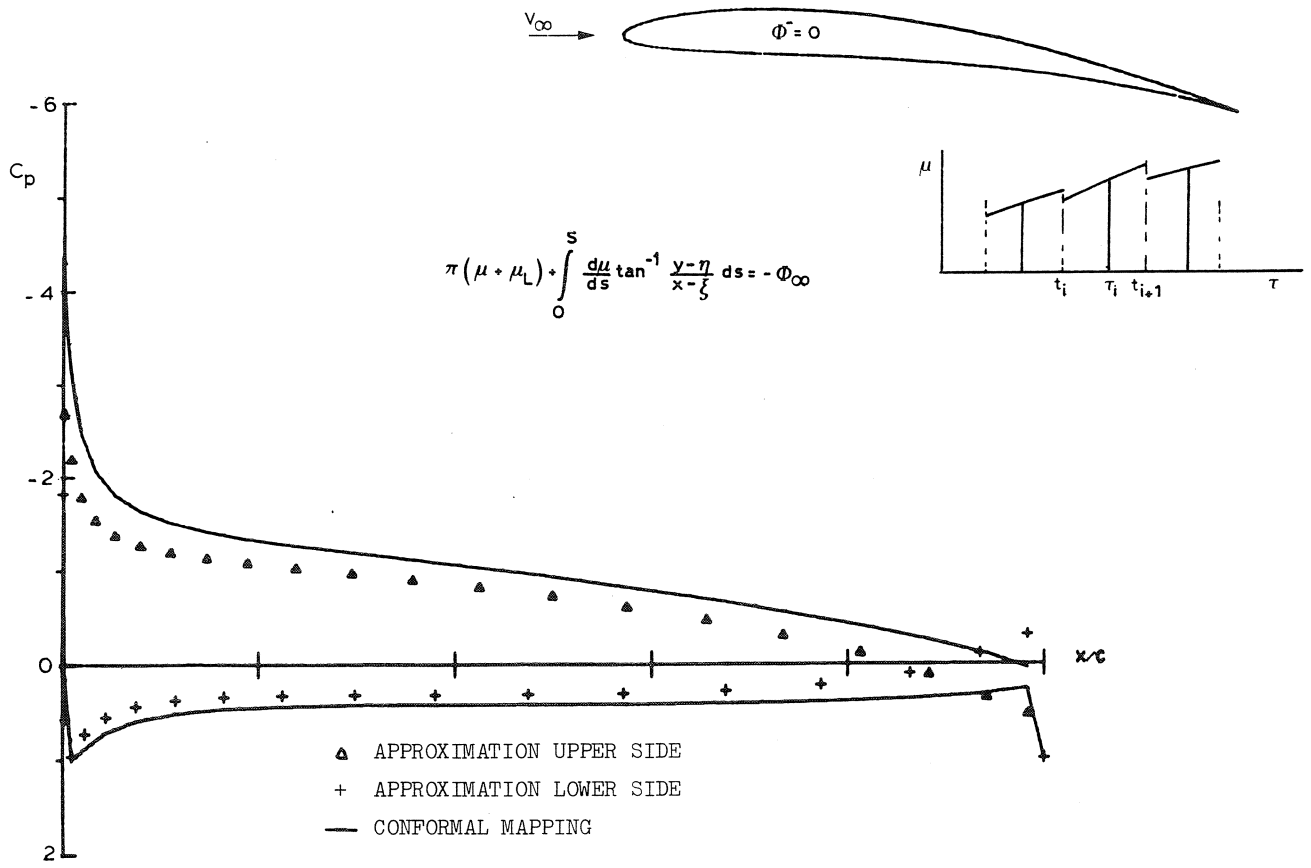



Fig. 13 Pressure distribution resulting from application of method IIIc (n = 40)

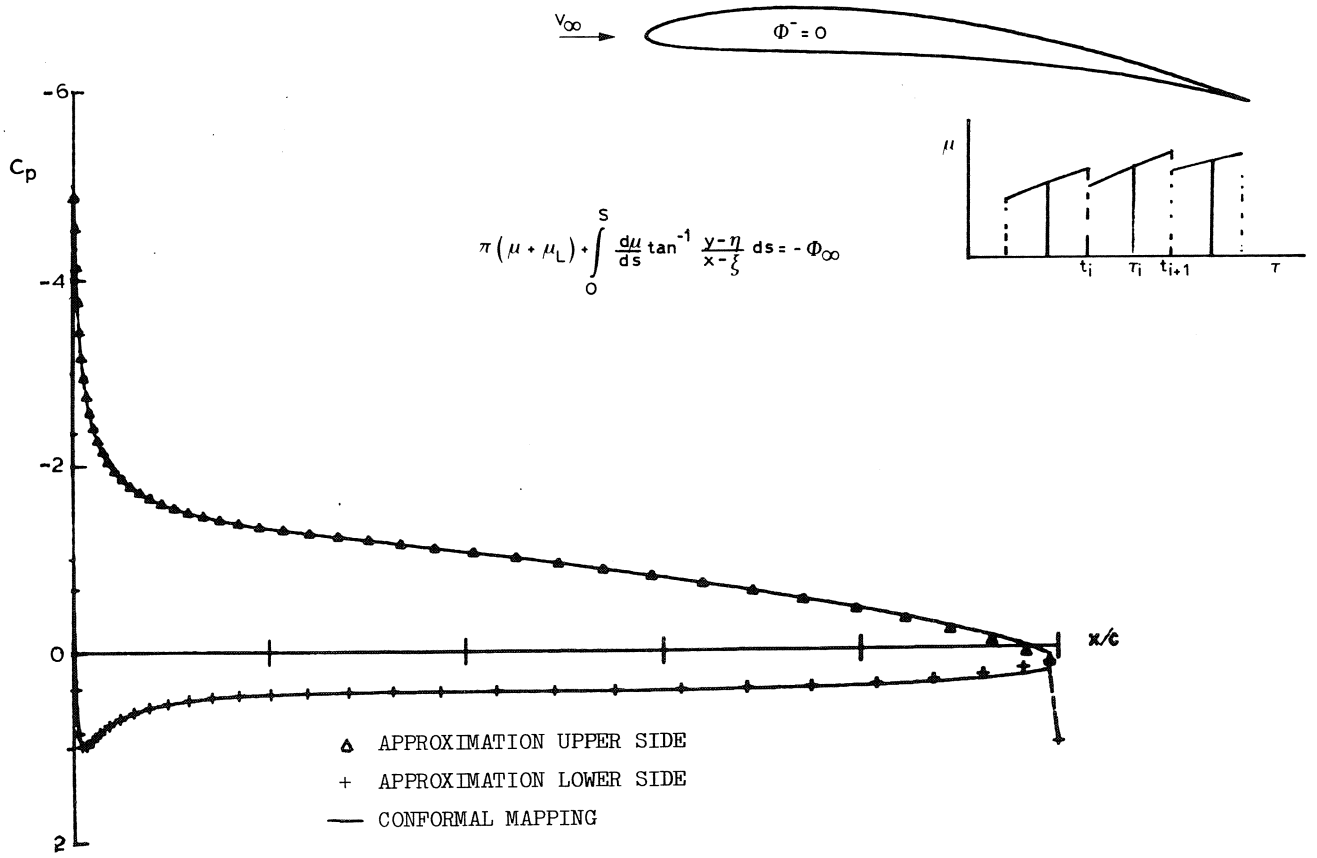



Fig. 14 Pressure distribution resulting from application of method IIIc (n = 80)

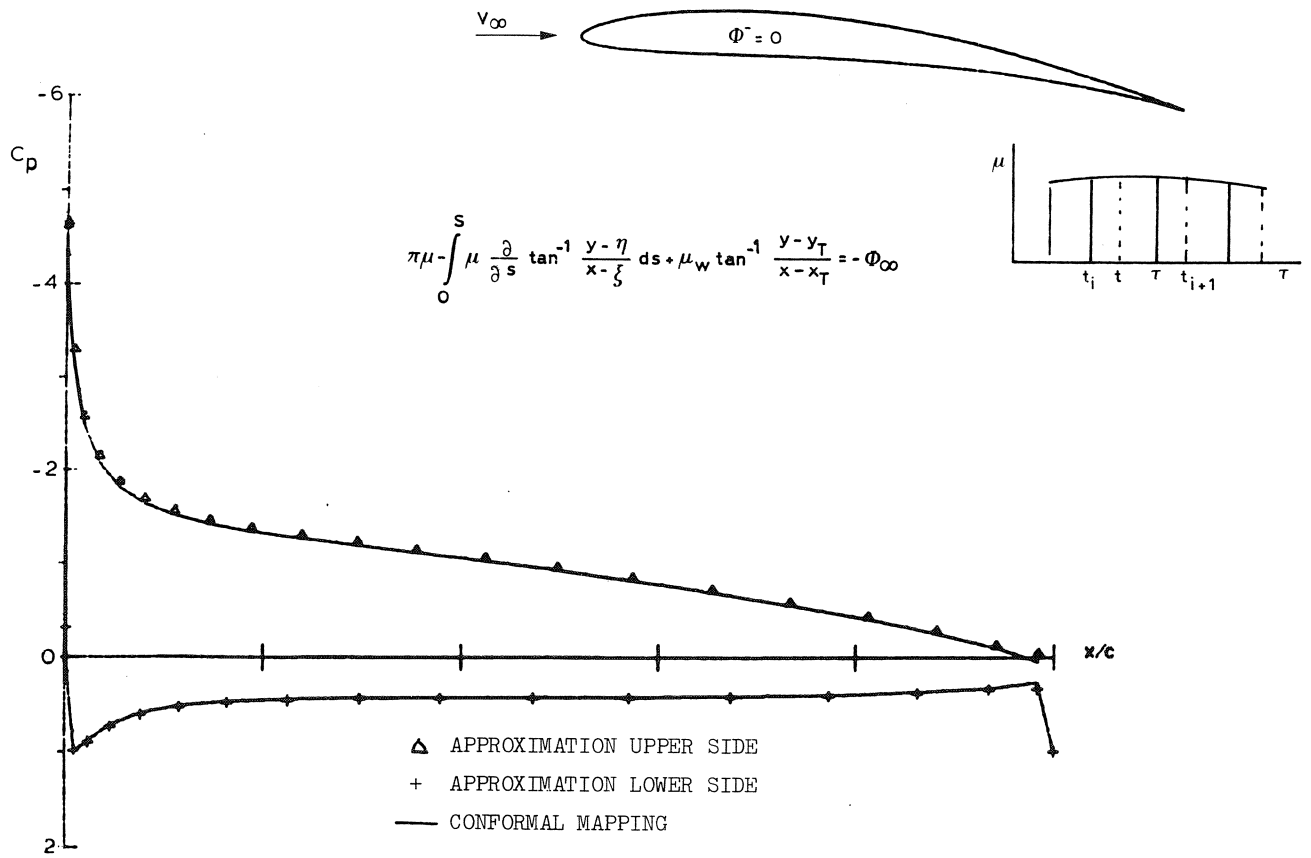
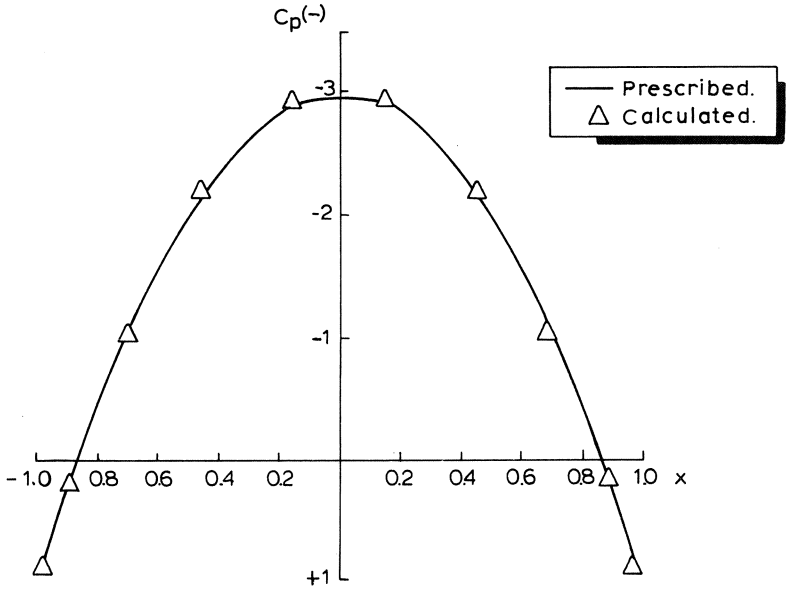
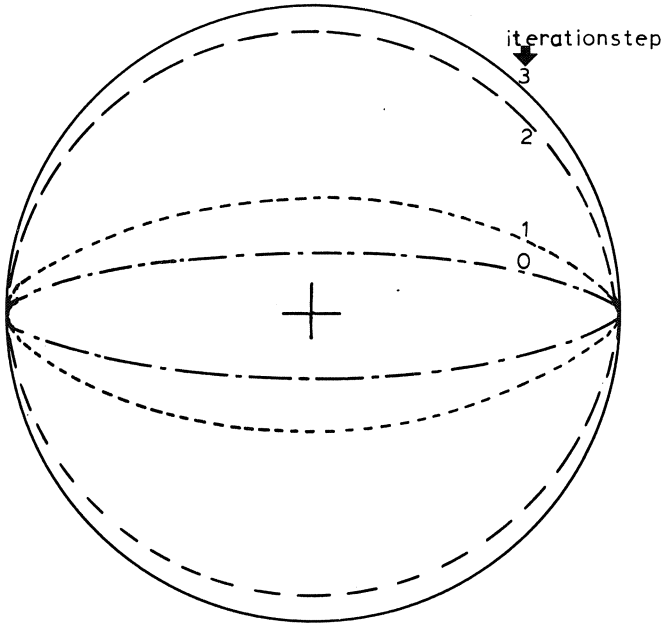



Fig. 15 Pressure distribution resulting from application of methods IIIb, d (n = 40)

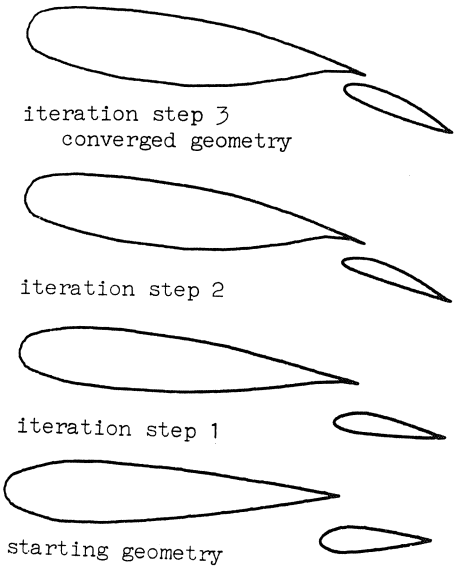



a. Pressure distribution.

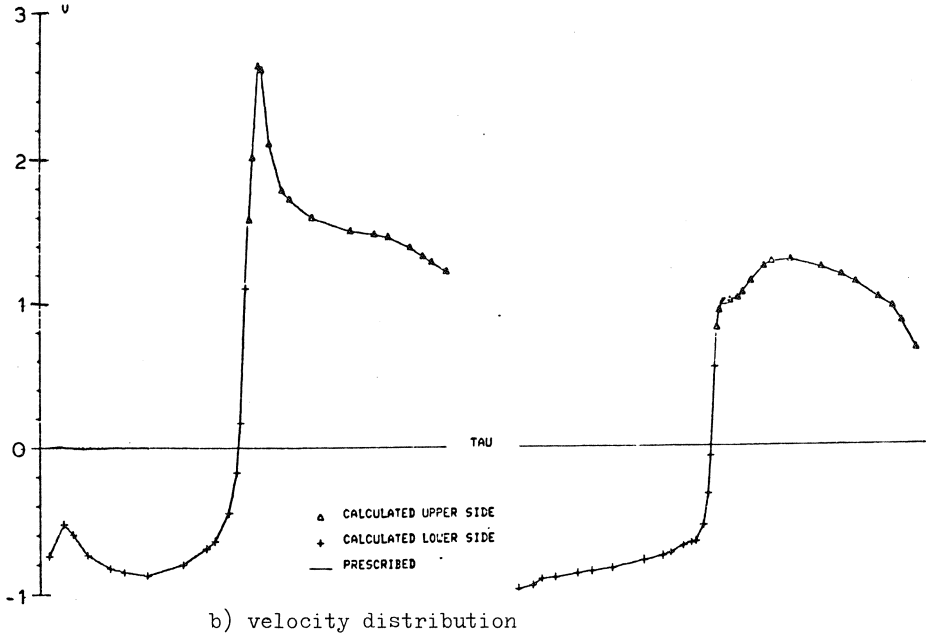



b. Contour.

**Fig. 16 Reconstruction of a circular cylinder**

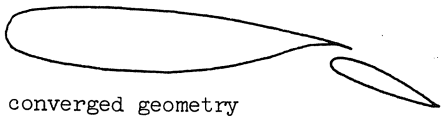



a) airfoil shapes

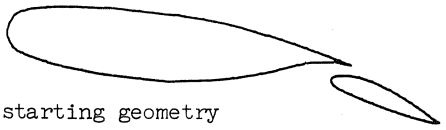



b) velocity distribution

**Fig. 17 Reconstruction of an airfoil-flap configuration**

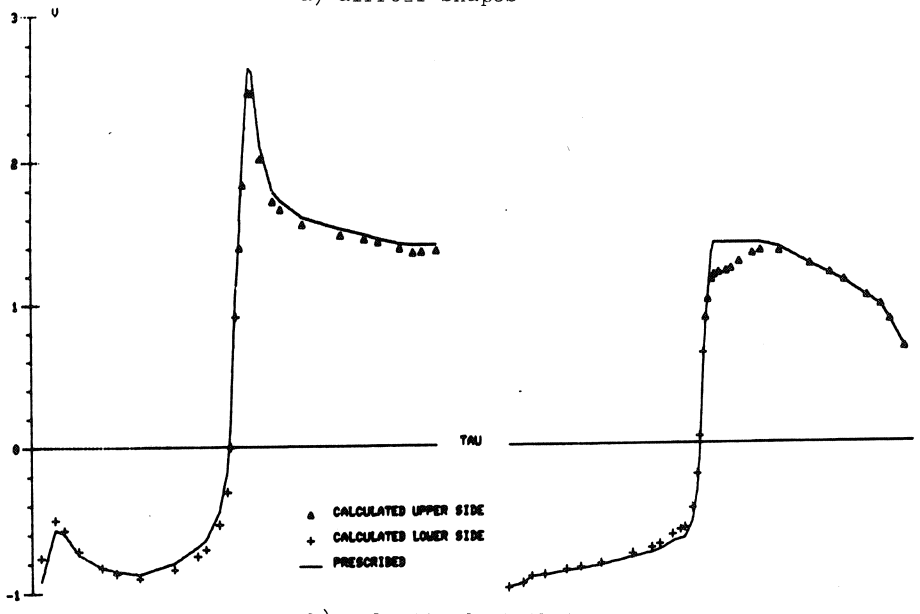


converged geometry



starting geometry

a) airfoil shapes




b) velocity distribution

**Fig. 18 Design problem solution with fixed mutual position of main profile and flap**

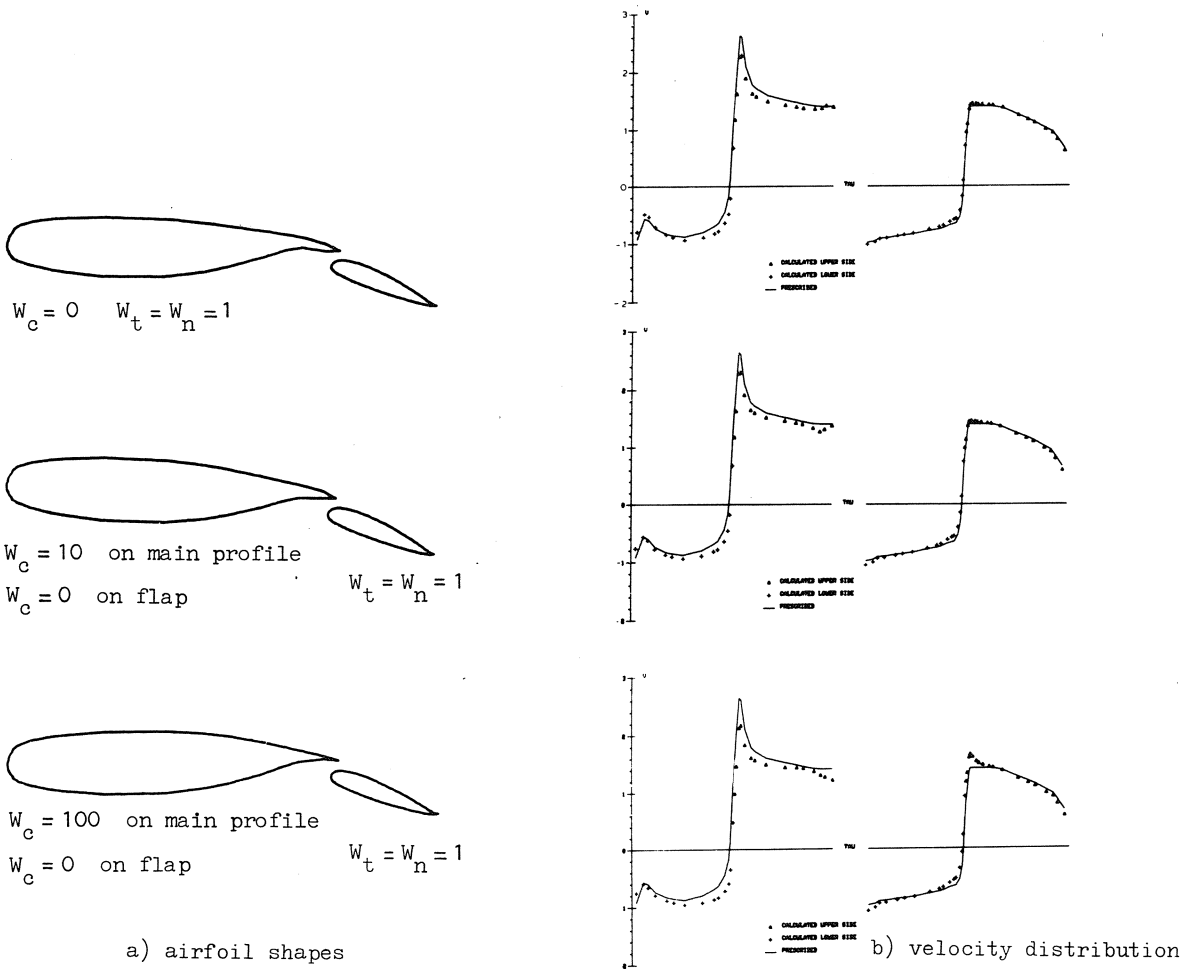
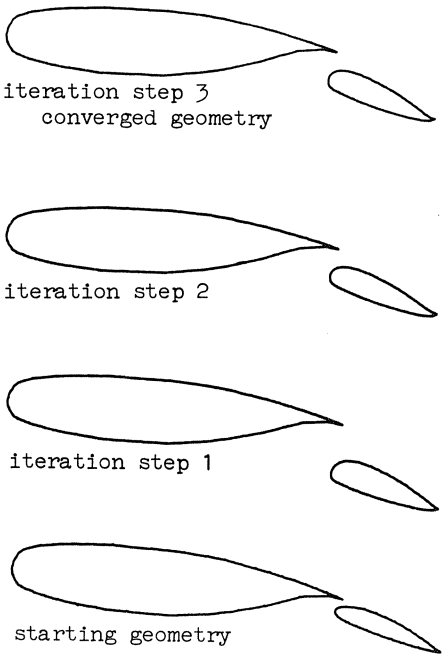
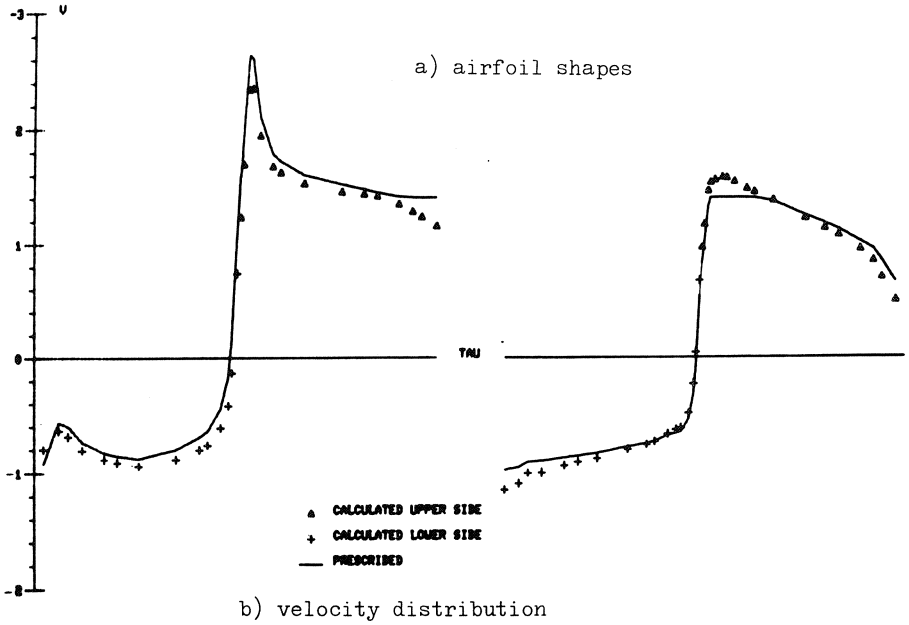



**Fig. 19 Design problem solution with a geometrical constraint on the main airfoil**




a) airfoil shapes




b) velocity distribution

**Fig. 20 Design problem solution with a fixed shape of the main airfoil and free shape and vertical position of the flap**

LIST OF SYMBOLS

| | |
|---|---|
| $n$ | normal to airfoil contour, taken positive when pointing outwards |
| $r$ | distance between two points |
| $s$, $t$ | arclength along airfoil contour, taken positive in the direction from trailing edge to leading edge at the lower side and in the direction from leading edge to trailing edge at the upper side |
| $x$, $y$ | Cartesian co-ordinate system |
| $C$ | airfoil contour |
| $Cp$ | pressure coefficient |
| $F$ $F_c$,$F_n$,$F_t$ | least squares error functions |
| $G$, | flow domain |
| $S$ | circumference of the airfoil contour |
| $V$ | velocity |
| $V_t$ | tangential velocity |
| $W_c^t$,$W_n$,$W_t$ | weighting functions |
| $\alpha$ | angle of attack |
| $\gamma$ | vorticity |
| $\xi$ ,$\eta$ | xy co-ordinates of an integration point |
| $\mu$ | doublet strength |
| $\phi$ | velocity potential |
| $\psi$ | stream function |
| $\tau$ | approximation of the arclength measured along the straight line segments connecting the given points |

Subscripts

| | |
|---|---|
| $d$ | refers to the doublet distribution |
| $n$ | denotes differentiation in normal direction |
| $t$ | denotes differentiation in tangential direction |
| $x$ | denotes differentiation in x-direction |
| $y$ | denotes differentiation in y-direction |
| $\infty$ | refers to infinity |
| $L$ | refers to the trailing edge at the lower side |
| $T$ | refers to the trailing edge |
| $U$ | refers to the trailing edge at the upper side |
| $W$ | refers to the line of discontinuity |

Superscripts

| | |
|---|---|
| $+$ | refers to the outer side of the airfoil contour |
| $-$ | refers to the inner side of the airfoil contour |

# COMPUTATIONAL ASPECTS OF INTEGRAL-EQUATION METHODS FOR THE SCATTERING OF TIME-HARMONIC WAVES BY PENETRABLE OBSTACLES

G. MUR

## 1. INTRODUCTION

In this paper three different integral-equation methods for computing the electromagnetic scattering properties of a transparent cylindrical obstacle are discussed and compared with each other. Each of the methods is formulated for a dielectric obstacle in a waveguide of rectangular cross-section. First a number of numerical difficulties that are inherent to the numerical implementation of the equations are discussed. Subsequently, some scattering problems are solved and a comparison of the methods is made with regard to their range of applicability and the amount of computation time that is required for arriving at a desired accuracy. In Section 6 we discuss the computation of fields in cylindrical obstacles in free space.

A rectangular waveguide with a cylindrical obstacle is shown in Fig. 1. A cartesian coordinate system is introduced with its y-axis

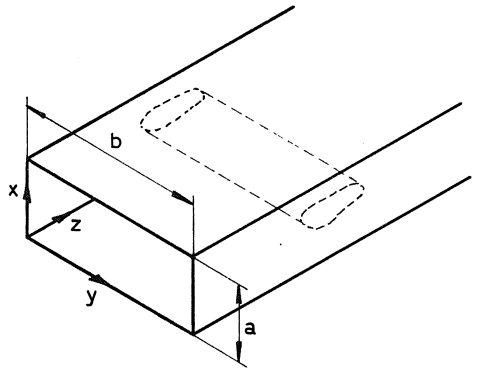



Fig. 1. Rectangular waveguide with cylindrical obstacle.

parallel to the direction in which the obstacle is uniform; the z-axis is chosen parallel to the axis of the waveguide. The waveguide walls are assumed to be perfectly conducting. The medium in the guide is taken to be homogeneous, linear, isotropic and lossless with (real) permittivity $\varepsilon_1$. The medium of the obstacle is taken to be linear, isotropic and uniform in the y-direction. It may be inhomogeneous (in the x,z-plane) as well as lossy; its complex permittivity is $\varepsilon_2 = \varepsilon_2(x,z)$. The permeability is taken to be the vacuum value $\mu_0 (=4\pi 10^{-7}$ H/m) everywhere. All fields are taken to vary sinusoidally in time with angular frequency $\omega$. The complex time factor $\exp(j\omega t)$ is omitted throughout. Depending on the method of solution to be

discussed, we divide, in two ways, the longitudinal cross-section of the

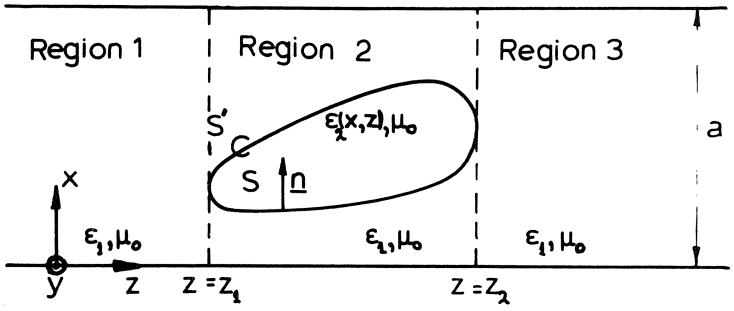



Fig. 2. Longitudinal cross-section of the configuration.

configuration, shown in Fig. 2, into different regions. The first way distinguishes the Region S inside the obstacle and the Region S' outside it. In the second one, we divide the waveguide into Region 1, which is uniform and occupies $-\infty < z < z_1$, Region 2 which contains the obstacle and occupies $z_1 < z < z_2$ and Region 3 which again is uniform and occupies $z_2 < z < \infty$. The choice of $z_1$ and $z_2$ is such that $z_2 - z_1$ is minimized. For simplicity, we only consider the E-polarized field that has a single-component electric field $\underline{E} = E_y \underline{i}_y$ and a magnetic field in the x,z-plane $\underline{H} = H_x \underline{i}_x + H_z \underline{i}_z$. Elimination of the magnetic-field components $H_x$ and $H_z$ from the electromagnetic field equations leads to the equation for the only nonvanishing component of the electric field

$$(\partial_x^2 + \partial_z^2 + k^2(x,z))E_y(x,z) = 0 \tag{1}$$

with

$$k(x,z) = \begin{cases} k_1 = \omega(\varepsilon_1 \mu_0)^{\frac{1}{2}} & \text{when } (x,z) \in S', \\ k_2(x,z) = \omega(\varepsilon_2(x,z)\mu_0)^{\frac{1}{2}} & \text{when } (x,z) \in S, \end{cases} \tag{2}$$

with $\mathrm{Re}(k) \geq 0$. In the uniform regions, the field can be expressed in terms of waveguide modes that satisfy the boundary conditions $E_y = 0$ on the waveguide walls $x = 0$ and $x = a$,

(3) $$\phi_m^+(x,z) = \sin(m\pi x/a)u_m(z), \quad \phi_m^-(x,z) = \sin(m\pi x/a)u_m^{-1}(z),$$

$m = 1,2,\ldots,$ with

(4) $$u_m(z) = \exp(-\Gamma_m z), \quad u_m^{-1}(z) = \exp(\Gamma_m z)$$

where

(5) $$\Gamma_m = ((m\pi/a)^2 - k_1^2)^{\frac{1}{2}}, \quad \mathrm{Re}(\Gamma_m) \geq 0, \quad \mathrm{Im}(\Gamma_m) \geq 0,$$

$m = 1,2,\ldots$ . As incident field $E_y^i$ we take the dominant mode travelling in the positive z-direction

(6) $$E_y^i = \phi_1^+.$$

The scattered field $E_y^s$ is now defined in the entire waveguide as

(7) $$E_y^s = E_y - E_y^i.$$

In Region 1, the scattered field can be written as

(8) $$E_y^s = \sum_{m=1}^{\infty} R_m \phi_m^-,$$

where $R_m$ denotes the reflection factor for the m-th mode. Using (6), (7) and (8) the total field in Region 1 is obtained as

(9) $$E_y = \phi_1^+ + \sum_{m=1}^{\infty} R_m \phi_m^-.$$

In Region 3, the transmitted field can be written as

(10) $$E_y = \sum_{m=1}^{\infty} T_m \phi_m^+,$$

where $T_m$ denotes the transmission factor for the m-th mode. In the next three sections, we shall successively present three different integral-equation methods for the computation of the reflection and transmission factors.

## 2. FREDHOLM INTEGRAL-EQUATION METHOD, BOUNDARY TYPE

### 2.1. Derivation of the Integral Equations

The boundary-type Fredholm integral equation employs as its main tool the two-dimensional Green's theorem. Since in this method Green's functions of the empty waveguide as well as of the obstacle are involved, which can only be determined easily in homogeneous regions, the applicability of this method is restricted to the case $\varepsilon_2$ = constant (i.e. a homogeneous obstacle). Using Green's theorem and the continuity of $E_y$ and its normal derivative $\partial_{\underline{n}} E_y$ on C [2], we arrive at

$$\oint_C (G_1 \partial_{\underline{n}} E_y - E_y \partial_{\underline{n}} G_1)ds + E_y^i(x_p,z_p) = \frac{1}{2} E_y(x_p,z_p),$$

(11) when $(x_p,z_p) \in C$,

$$\oint_C (G_2 \partial_{\underline{n}} E_y - E_y \partial_{\underline{n}} G_2)ds = -\frac{1}{2} E_y(x_p,z_p), \qquad \text{when } (x_p,z_p) \in C,$$

where

$$G_1 = \sum_{m=1}^{\infty} (a\Gamma_m)^{-1} \sin(m\pi x/a)\sin(m\pi x_p/a)\exp(-\Gamma_m|z-z_p|) \tag{12}$$

is the Green's function that satisfies the boundary conditions at the waveguide walls [1] and

$$G_2 = -\frac{1}{4} Y_0(k_2 R); \quad R = ((x-x_p)^2 + (y-y_p)^2)^{\frac{1}{2}} \geq 0 \tag{13}$$

is a free space Green's function for the interior of the obstacle. We note that the integrals should be evaluated as Cauchy principal value integrals, excluding the point $(x,z) = (x_p,z_p)$ for which $\partial_{\underline{n}} G$ has no value. Equation (11) constitutes a system of two coupled Fredholm integral equations of the second kind from which the unknown functions $E_y$ and $\partial_{\underline{n}} E_y$ on C are to be obtained. After solving the integral equations, the reflection and transmission factors can easily be computed [2].

### 2.2. The Numerical Solution

In order to obtain a numerical solution of (11), the integral equations

are transformed into a system of algebraic equations. To this aim, both $E_y$ and $\partial_{\underline{n}} E_y$ are expanded in terms of a series of $N_F$ pulse functions [3]. Pulse functions are functions that have the value unity over a subsection $\Delta C_j$ ($j = 1,\ldots,N_F$) of the interval of integration C ($\Sigma_j \Delta C_j = C$) and the value zero outside this subsection. Applying point matching [3] to enforce the equality sign in (11), we arrive at

$$\sum_{j=1}^{N_F} b_j \int_{\Delta C_j} G_1(x,z,x_i,z_i)ds - \sum_{j=1}^{N_F} a_j \int_{\Delta C_j} \partial_{\underline{n}} G_1(x,z,x_i,z_i)ds + E_y^i(x_i,z_i) = \frac{1}{2} a_i, \qquad 1 \le i \le N_F,$$

(14)

$$\sum_{j=1}^{N_F} b_j \int_{\Delta C_j} G_2(x,z,x_i,z_i)ds - \sum_{j=1}^{N_F} a_j \int_{\Delta C_j} \partial_{\underline{n}} G_2(x,z,x_i,z_i)ds = -\frac{1}{2} a_i, \qquad 1 \le i \le N_F,$$

where $a_j$ and $b_j$, $j = 1,\ldots,N_F$, denote the amplitudes of the pulse functions in the expansions of $E_y$ and $\partial_{\underline{n}} E_y$ respectively. In the case i = j, the integrals over $\partial_{\underline{n}} G$ are to be evaluated as Cauchy principal value integrals. As to the choice of the pulse functions we note that they are, usually, chosen to have equal width. However, in order to obtain a more accurate approximation, to the unknown field, the width of the pulse functions should be chosen smaller in those regions where $E_y$ and/or $\partial_{\underline{n}} E_y$ vary rapidly (for instance near the edges of the obstacle). The main numerical effort lies in the evaluation of the integrals in (14). We shall now make a series of remarks concerning the evaluation of these integrals. Since $\partial_{\underline{n}} G_1$ and $\partial_{\underline{n}} G_2$ are, usually, smooth functions on the intervals of integration, the integration of these functions can be carried out using a simple trapezoidal rule both for the case $i \neq j$ and for the case i = j when the integrals are principal value integrals. $G_1$ and $G_2$ are, usually, smooth functions of the integration variable and again a simple quadrature rule yields accurate results. When i = j, however, the integrand has a logarithmic singularity. In this case, the relevant integrals should be evaluated with care, preferably analytically. Integrals containing $G_1$ (see eqn(12)) require the evaluation of a series that converges slowly especially when $z_1 = z_2$. To save computation time, the number of terms taken into account has to be reduced with the aid

of a technique for accelerating the convergence [4]. Acceleration of convergence can be obtained in an alternative way by substituting (12) in (14), interchanging the summation and the integration and carrying out the resulting integrations analytically at least for those domains where this can be done. Finally we note that when $(x_i,z_i)$ and $(x_j,z_j)$ are located at opposite sides near a sharp edge of the obstacle, all integrands in (14) will vary considerably over the interval of integration and consequently a more accurate quadrature rule should be invoked.

## 3. FREDHOLM INTEGRAL-EQUATION METHOD, DOMAIN TYPE

### 3.1. Derivation of the Integral Equations

For the domain type of integral equation, the obstacle is replaced by an equivalent polarization current distribution. Using this approach, the scattered field can be written as an integral over the fields that are radiated by this current distribution [2]. Thus we obtain

$$E_y(x_p,z_p) = E_y^i(x_p,z_p) + \iint_S (k^2(x,z) - k_1^2)E_y(x,z)G_1(x,z,x_p,z_p)dA, \tag{15}$$

where $G_1$ denotes the Green's function for the uniform waveguide (12). Equation (15), when $(x_p,z_p) \in S$, constitutes a Fredholm integral equation of the second kind for the unknown function $E_y$. It is noted that, since no Green's function for the Region S is used, the medium of the obstacle may be inhomogeneous.

### 3.2. The Numerical Solution

In order to obtain a numerical solution of (15), the unknown function $E_y$ is expanded inside S in terms of two-dimensional pulse functions [5]. These pulse functions are two-dimensional functions in the x,z-plane, the value of which equals unity over a subdomain $\Delta S_j$, $j = 1,\ldots,M$ ($\Sigma_j \Delta S_j = S$) of the cross-section S of the obstacle and which equals zero outside it. Usually these subsections are chosen to be square. Again applying point-matching we arrive at a system of equations that can be written as

$$a_i = E_y^i(x_i,z_i) + \sum_{j=1}^{M} a_j \iint_{\Delta S_j} (k^2(x,z)-k_1^2)G_1(x,z,x_i,z_i)dA, \tag{16}$$

where $a_j$, $j = 1,\ldots,M$, denotes the amplitude of the j-th pulse function in the expansion of $E_y$. We note that, although $E_y$ is assumed to be constant on a subsection, $k^2(x,z)$ may be discontinuous on it. The main numerical effort lies in the evaluation of the integrals in (16). As in Section 2.2, a method for accelerating the convergence of the series representation (12) of $G_1$ has to be applied. In the case $k^2(x,z) - k_1^2$ = constant on $\Delta S_j$, the integrals in (16) for each term of $G_1$ can be carried out analytically, which yields a considerable acceleration of convergence especially when i = j. When the obstacle is inhomogeneous, however, the integrals must be evaluated numerically, for instance using a trapezoidal rule. In the inhomogeneous case a special algorithm has to be written for the case i = j in order to cope with the logarithmic singularity of $G_1$.

## 4. VOLTERRA INTEGRAL-EQUATION METHOD

### 4.1. The Derivation of the System of Equations

The Volterra integral equation method starts with the Fredholm integral equation (15) that was derived in the previous section. This equation is now transformed into a system of coupled integral equations of the Volterra type. Since this transformation is not as well-known as the derivation of the Fredholm equations, it is described in more detail. We start the transformation by writing

$$E_y(x,z) = \sum_{n=1}^{\infty} \psi_n(z)\sin(n\pi x/a), \qquad 0 \le x \le a;\ -\infty < z < \infty. \tag{17}$$

Using (3), we now substitute (12) and (17) in (15), multiply the right- and the left-hand side of (15) by $\sin(m\pi x_p/a)$, $m = 1,2,\ldots$, and integrate the resulting equation over the interval $0 \le x_p \le a$. As a result we obtain

$$\psi_m(z_p) = u_m(z_p)\delta_{m,1} + \int_{z=z_1}^{z_2} \exp(-\Gamma_m|z-z_p|) \sum_{n=1}^{\infty} V_{mn}(z)\psi_n(z)\,dz \tag{18}$$

for $m = 1,2,\ldots,-\infty < z_p < \infty$, with

$$V_{mn}(z) = (a\Gamma_m)^{-1} \int_{x=0}^{a} (k^2(x,z)-k_1^2)\sin(m\pi x/a)\sin(n\pi x/a)\,dx, \tag{19}$$

$$-\infty < z < \infty.$$

Using (3) and splitting up the interval of integration, (18) can be rewritten as

$$\psi_m(z_p) = u_m(z_p)\delta_{m,1} + u_m(z_p) \int_{z=z_1}^{z_p} u_m^{-1}(z) \sum_{n=1}^{\infty} V_{mn}(z)\psi_n(z)dz$$

(20)

$$+ u_m^{-1}(z_p) \int_{z=z_p}^{z_2} u_m(z) \sum_{n=1}^{\infty} V_{mn}(z)\psi_n(z)dz,$$

$m = 1,2,\ldots$ . Using matrices, (20) can more succinctly be written as

$$\underline{\psi}(z_p) = \underline{\underline{U}}(z_p)(\underline{\delta}^{(1)} + \int_{z=z_1}^{z_p} \underline{\underline{U}}^{-1}(z)\underline{\underline{V}}(z)\underline{\psi}(z)dz)$$

(21)

$$+ \underline{\underline{U}}^{-1}(z_p) \int_{z=z_p}^{z_2} \underline{\underline{U}}(z)\underline{\underline{V}}(z)\underline{\psi}(z)dz,$$

where we have introduced the column vectors $\underline{\psi}(z) = (\psi_1(z),\psi_2(z),\ldots)^T$ and $\underline{\delta}^{(1)} = (1,0,0,\ldots)^T$ and the square matrices $\underline{\underline{U}}(z) = \mathrm{diag}(u_1(z),u_2(z),\ldots)$, and $\underline{\underline{V}}(z) = (V_{mn}(z))$. Equation (21) can be rewritten as [6]

$$\underline{\psi}(z_p) = \underline{\underline{U}}(z_p)\underline{T} + \underline{\underline{U}}^{-1}(z_p) \int_{z=z_p}^{z_2} \underline{\underline{U}}(z)\underline{\underline{V}}(z)\underline{\psi}(z)dz$$

(22)

$$- \underline{\underline{U}}(z_p) \int_{z=z_p}^{z_2} \underline{\underline{U}}^{-1}(z)\underline{\underline{V}}(z)\underline{\psi}(z)dz,$$

with

$$\underline{T} = \underline{\delta}^{(1)} + \int_{z=z_1}^{z_2} \underline{\underline{U}}^{-1}(z)\underline{\underline{V}}(z)\underline{\psi}(z)dz. \qquad (23)$$

We note that $\underline{T}$ is an as yet unknown constant column vector. In order to obtain the solution of (22), we introduce the matrix $\underline{\underline{\Psi}}(z) = (\Psi_{mn}(z))$, which is defined as the solution of

$$\underline{\underline{\Psi}}(z_p) = \underline{\underline{U}}(z_p)\Big(\underline{\underline{1}} - \int_{z=z_p}^{z_2} \underline{\underline{U}}^{-1}(z)\underline{\underline{V}}(z)\underline{\underline{\Psi}}(z)dz$$

(24)

$$+ \underline{\underline{U}}^{-1}(z_p) \int_{z=z_p}^{z_2} \underline{\underline{U}}(z)\underline{\underline{V}}(z)\underline{\underline{\Psi}}(z)dz,$$

where $\underline{\underline{1}}$ denotes the unit matrix. Equation (24) constitutes a system of coupled Volterra integral equations of the second kind. Once the system (24) is solved, the solution of (22) is obtained as

$$\underline{\psi}(z_p) = \underline{\underline{\Psi}}(z_p)\underline{T}, \tag{25}$$

where

$$\underline{T} = \Big(\underline{\underline{1}} - \int_{z=z_1}^{z_2} \underline{\underline{U}}^{-1}(z)\underline{\underline{V}}(z)\underline{\underline{\Psi}}(z)dz\Big)^{-1}\underline{\delta}^{(1)}. \tag{26}$$

Using (17) and (3) we can now identify $\underline{T}$ as the column matrix containing the transmission factors $T_m$ that have been defined in (10). The column matrix $\underline{R}$ containing the reflection factors is subsequently obtained as

$$\underline{R} = \Big(\int_{z=z_1}^{z_2} \underline{\underline{U}}(z)\underline{\underline{V}}(z)\underline{\underline{\Psi}}(z)dz\Big)\underline{T}. \tag{27}$$

4.2. The Numerical Solution of the System of Equations

For the numerical solution of (24), the expansion (17) of $E_y$ is truncated after $M_v$ terms where $M_v$ is chosen such that a sufficiently accurate solution is obtained. As a consequence of this truncation, the vectors in the above analysis contain $M_v$ elements and the square matrices have the dimension $M_v * M_v$. As to the computation of $\underline{\underline{V}}$, we note that, using the product formulas for goniometric functions, it can be reduced to the computation of 2M+1 integrals of the form

$$W_m(z) = \int_{x=0}^{a} (k^2(x,z)-k_1^2)\cos(m\pi x/a)dx, \qquad m = 0,1,\ldots,2M_v. \tag{28}$$

These integrals are computed either numerically or analytically depending on the x-dependence of $k^2(x,z)$. We shall now describe the numerical solution of (24). First it is noted that from (24) we easily obtain

$$\partial_{z_p} \underline{\underline{\Psi}}(z_p) = -\underline{\underline{\Gamma}}\,\underline{\underline{U}}(z_p)\Big(\underline{\underline{1}} - \int_{z=z_p}^{z_2} \underline{\underline{U}}^{-1}(z)\underline{\underline{V}}(z)\underline{\underline{\Psi}}(z)\,dz\Big)$$

$$+\underline{\underline{\Gamma}}\,\underline{\underline{U}}^{-1}(z_p) \int_{z=z_p}^{z_2} \underline{\underline{U}}(z)\underline{\underline{V}}(z)\underline{\underline{\Psi}}(z)\,dz, \tag{29}$$

where $\underline{\underline{\Gamma}} = \mathrm{diag}(\Gamma_1,\Gamma_2,\ldots,\Gamma_{M_v})$. Thus, by comparing (24) and (29) we see that $\partial_{z_p}\underline{\underline{\Psi}}(z_p)$ can easily be computed from the same system of integrals that is used for the computation of $\underline{\underline{\Psi}}(z_p)$. The evaluation of these integrals over the interval $(z_1,z_2)$ is carried out in a stepwise manner [2,7], with N steps of size $h = (z_2 - z_1)/N$, from $z = z_2$ to $z = z_1$. The integration over the n-th step (n = 1,2,...,N) runs as follows. First we notice that the values of the integrals and therefore of $\underline{\underline{\Psi}}$ and $\partial_{z_p}\underline{\underline{\Psi}}$ are known at the upper boundary of the interval ($z_p = z_2 - (n-1)h$). We now make an intermediate approximation for $\underline{\underline{\Psi}}(z_2 - nh)$ and $\partial_{z_p}\underline{\underline{\Psi}}(z - nh)$ at the lower boundary of the n-th interval, for example, by using a linear extrapolation. Using this approximation, we compute the integrals over the interval under consideration with the aid of the Euler-McLaurin summation formula [8]. The value of these integrals being known, we determine a new approximation for $\underline{\underline{\Psi}}$ and $\partial_{z_p}\underline{\underline{\Psi}}$ at $z_p = z_2 - nh$, and update the numerical approximation of these functions. This iterative process is continued until a sufficiently accurate result is obtained. In practice, only one or two steps in the iteration process are required.

## 5. NUMERICAL RESULTS AND CONCLUSIONS FOR THE WAVEGUIDE SCATTERING PROBLEM

In this section, the integral-equation methods are applied to the computation of the scattering by the obstacle depicted in Fig. 3. The angular frequency $\omega$ is chosen such that $\omega = 1.5\omega_c$ where $\omega_c/2\pi$ denotes the cut-off frequency of the dominant mode of the waveguide. As a consequence, there is only a single propagating mode in the uniform waveguide. Computer programs have been written for the Fredholm integral-equation method (boundary type) of Section 2 and for the Volterra integral-equation method of

Section 4. The domain type Fredholm integral equation method of Section 3 was discarded because, by the time this research was carried out, it was expected to compare unfavourably with the boundary type method. This expectation was based on the assumption, which appears to be generally accepted (see for instance [9]), that reducing the number of dimensions of the problem will reduce the computational effort that is required to solve it. In Section 7 we shall comment on this assumption. In Figs. 4 and 5, the modulus of the reflection factor $R_1$ and of the transmission factor $T_1$,

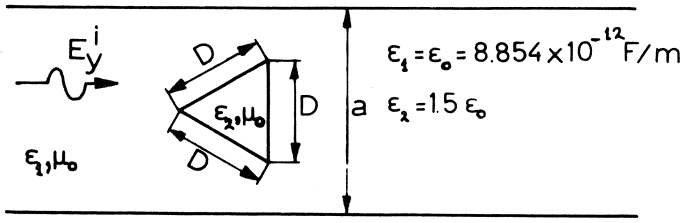



Fig. 3. Longitudinal section of the configuration for which numerical results are presented.

respectively, are depicted as a function of D/a (see Fig. 3). In these figures, we have also inserted results that have been obtained using the Rayleigh-Gans approximation $E_y = E_y^i$ in the right-hand side of (15). The results obtained using the two integral-equation methods differ by less

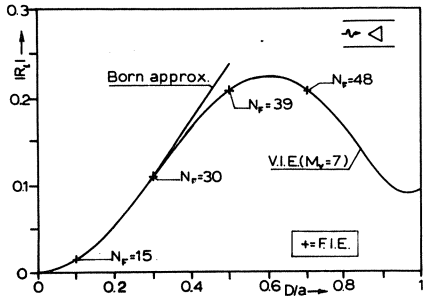



Fig. 4. Reflection factor $|R_1|$.

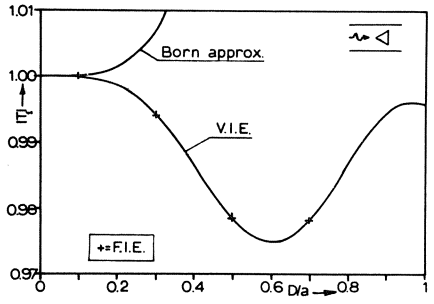



Fig. 5. Transmission factor $|T_1|$.

than 1 per cent. In Table 1, the methods of solution are compared with regard to the computation time and storage requirements for the computation of the scattering by an obstacle with D/a = 0.5. We observe that the Fredholm integral-equation method (F.I.E.) turns out to be very expensive.

| Method | Computation time (seconds) | Storage req. (kbyte) | $\|R_1\|$ |
|---|---|---|---|
| F.I.E. ($N_F = 36$) | 200 | 448 | 0.20700 |
| V.I.E. ($M_V = 7$) | 3 | 160 | 0.20682 |

Table 1. Comparison of results for obstacle with D/a = 0.5.

A second advantage of the Volterra integral-equation (V.I.E.) method is, that for a specific problem only a subroutine for computing the integral in (28) needs to be written, the remaining part of the programs being unchanged. For obstacles that show strong local variations in their permittivity, a large number of terms in (17) has to be taken into account in order to arrive at the desired accuracy and this, even with the V.I.E. method, leads to considerable computation times. For such obstacles the domain-type Fredholm integral-equation method turns out to be the most attractive. An illustration of this is given in the next sections. Another limitation of the V.I.E. method appears to be that it cannot easily be generalized for

application to configurations in which the Green's function has a more complicated structure.

## 6. FIELDS IN CYLINDRICAL OBSTACLES IN FREE SPACE

In this section Fredholm integral-equation methods of the types that are discussed in Sections 2 and 3 will be formulated for the computation of fields scattered by cylindrical obstacles in free space. In particular, we shall compute the current distribution, due to an incident field, in an infinitely long conductor having a conductivity $\sigma$ and a permittivity $\varepsilon = \varepsilon_0 = 8.8544*10^{-12}$ F/m [10]. The incident field is assumed to be E-polarized, i.e. the electric fieldvector $E_y \underline{i}_y$ is parallel to the axis of the cylinder. We note that for dielectric cylinders, the analysis runs along the same lines as the analysis for conducting cylinders [5,11]. The conductor is assumed to have an arbitrary cross-section S with boundary C.

### 6.1. Boundary Type Integral Equation

Proceeding in the same way as in Section 2, we arrive at two coupled integral equations of the Fredholm type

$$\int_C (G_1 \partial_{\underline{n}} E_y - E_y \partial_{\underline{n}} G_1)ds + E_y^i(x_p,z_p) = \frac{1}{2} E_y(x_p,z_p),$$

(30) when $(x_p,z_p) \in C$,

$$\int_C (G_2 \partial_{\underline{n}} E_y - E_y \partial_{\underline{n}} G_2)ds = -\frac{1}{2} E_y(x_p,z_p),$$

with the free space Green's function

$$G_1 = -(j/4)H_0^{(2)}(k_1R) \tag{31}$$

where

$$k_1 = \omega(\varepsilon_0\mu_0)^{\frac{1}{2}}, \tag{32}$$

and with

$$G_2 = -\frac{1}{4} Y_0(k_2R) \tag{33}$$

where

$$k_2 = (\omega^2\varepsilon_0\mu_0 - j\omega\sigma\mu_0)^{\frac{1}{2}}, \qquad \mathrm{Re}(k_2) \geq 0. \tag{34}$$

The solution of (30) runs along the same lines as the solution of (11). The integrations over the logarithmic singularities of $G_1$ and $G_2$ are carried out analytically, using the series representation of these functions [12].

## 6.2. Domain Type Integral Equation

When, as in Section 3.1, we replace the obstacle by a polarization current distribution, we obtain

$$E_y(x_p,z_p) = E_y^i(x_p,z_p) + \iint_S (k_2^2(x,z)-k_1^2)E_y(x,z)G_1(x,z,x_p,z_p)dA, \tag{35}$$

which constitutes a two-dimensional Fredholm integral equation of the second kind and where $G_1$ denotes the free space Green's function (31). As in Section 3.2, this integral equation can be solved using two-dimensional pulse functions for the expansion of the unknown function $E_y$ inside S followed by a point matching to derive a system of algebraic equations from which the field distribution can be solved [5,10]. We shall discard this possibility and describe an alternative way that appears to be very efficient. First, we observe that when the cross-section S of the conductor does not have a "large" length to width ratio, i.e. when S has a form that resembles a circle or a square (which will often be the case), the field distribution inside S is similar to the field distribution in a circular obstacle. To a high degree of accuracy the field distribution in the circular cylinder is proportional to $J_0(k_2r)$, where r denotes the distance of the point of observation to the center of the circular cross-section. In the non-circular case, we can write [10]

$$E_y(x,z) = f(x,z)J_0(k_2\rho), \qquad (x,z) \in S, \tag{36}$$

where $\rho$ denotes the distance from the point of observation (x,z) to the "center" of the obstacle. If the above assumption regarding the cross-section applies, f(x,z) will vary slowly with x and z, as compared with the variation of $E_y$ with x and z. Now f(x,z) is expanded in terms of a series of suitably chosen functions. We write

$$f(x,z) = \sum_{ij} a_{ij}\cos(i\pi x/a)\cos(j\pi z/b), \tag{37}$$

where a and b denote the width and the height of the smallest rectangle enclosing S. It appears that the expansion (37) converges rapidly [10]. We note that the present method only applies to conducting obstacles since for dielectric obstacles f(x,z) would be singular for those values of x and z for which $J_0(k_2\rho)$ is zero. For conducting obstacles $k_2$ is complex and $J_0(k_2\rho)$ will equal be zero nowhere for real $\rho$. In order to obtain a system of algebraic equations, we again apply point-matching. The main numerical effort now lies in the evaluation of the integrals in the matrix of coefficients of the system of algebraic equations. These integrals have the form

$$I_{ij} = \iint_S (k_2^2(x,z)-k_1^2)\cos(i\pi x/a)\cos(j\pi z/b) \tag{38}$$
$$J_0(k_2\rho)G_1(x,z,x_p,z_p)dA.$$

Integrals of this type can only be evaluated numerically. We note that $G_1$ has a logarithmic singularity at $(x,z) = (x_p,z_p)$. In order to obtain accurate results, the method of integration in the neighbourhood of this point should take into account this singularity. This can be done by assuming that, near the singularity, all factors in the integrand, except $G_1$ are locally constant. Now the remaining integral over this interval can be evaluated easily and added to the integral over the remaining part of S.

## 7. NUMERICAL RESULTS FOR THE CONDUCTING OBSTACLE

The methods that have been described in Section 6, are applied to the computation of the current distribution in conductors of various cross-sections. In Fig. 6 we have, for a number of frequencies of the incident field, plotted the current density on one of the diagonals of the square cross-section of a copper conductor ($\sigma = 57*10^6$ S/m). Results of the type that are given in Fig. 6 have been obtained in three different ways, viz. (a) using eq. (30), (b) using eq. (35), and solving this equation using pulse functions and (c) using eq. (35), and solving it using the expansion of eqns. (36) and (37). It appears that the results differ by less than 1 per cent. Comparing the three methods as regards the computation time it

appears that both methods based on the domain-type method (35) are much faster than the boundary-type method (30). This shows that reducing the number of dimensions of the problem not necessarily reduces the computational effort required to solve it. As to the comparison of the two domain-type

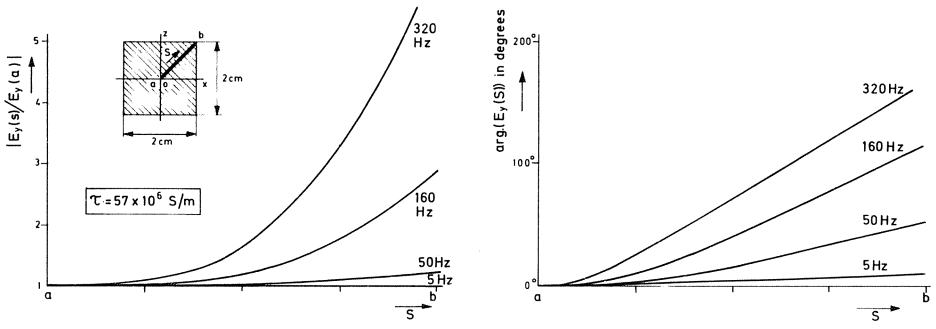



Fig. 6. Normalized amplitude and phase of current density on a - b in a copper conductor of square cross-section.

methods, we note that, for cross-sections having a large length to width ratio, the pulse function expansion appears to be the most attractive, whereas for cross-sections that more or less resemble a circle or a square the expansion of (36) and (37) is the most attractive one. As to the complexity of the computer programs, we note that the programs that are based on (35), are much simpler and much easier adaptable to different cross-sections than those based on (30).

## 8. CONCLUSIONS

In Table 2 we summarize the results of the comparison made between the three different integral-equation methods for computing the electromagnetic scattering properties of the transparent cylindrical obstacles we have investigated. When reading this table, the reader should bear in mind that the conclusions presented in it only hold for cases where the relevant methods are applicable. Furthermore it is noted that results regarding computation time and storage requirements depend on both the size and the shape of the cross-section of the cylinder. This dependence differs from one type of integral equation to the other. For cylinders having a larger cross-section than those investigated in this paper, the boundary type integral

equation might prove to be the most attractive.

| Type of integral equation | Boundary type Fredholm | Domain type Fredholm | Volterra |
|---|---|---|---|
| Applicable to inhomogeneous cylinders | No | Yes | Yes |
| Applicable to cylinders with a high contrast | Yes | Yes | No |
| Applicable to problems involving more complicated Green's functions | Yes | Yes | No |
| Adaptability of computer programs to different obstacles | Poor | Good | Very good |
| Complexity of the computer programs | High | Low | High |
| Computing time | Long | Short | Short |
| Storage requirements | High | Low | Very low |

Table 2. Comparison of integral-equation methods.

REFERENCES

[1] COLLIN, R.E., *Field Theory of Guided Waves,* New York, McGraw-Hill Book Company, Inc., 1960.

[2] MUR, G., D. QUAK & G.J. van DIJK, *Computational aspects of the scattering of electromagnetic waves by a dielectric obstacle in a waveguide of rectangular cross-section,* Conference paper, International Conference on Numerical Methods in Electrical and Magnetic Field Problems, S. Margherita Ligure (Italy), June 1-4, 1976.

[3] HARRINGTON, R.F., *Field Computation by Moment Methods,* New York, The Macmillan Company, 1968, p.5.

[4] JONG, G. de, *Scattering by a perfectly conducting cylindrical obstacle,* Int. J. Electronics, Vol. 32 (1972), pp. 153-167.

[5] RICHMOND, J.H., *Scattering by a dielectric cylinder of arbitrary cross-section shape,* IEEE Trans. on Antennas and Prop., Vol. AP-13, (1965), pp. 334-341.

[6] KOURI, D.J., *On the noniterative solution of integral equations for scattering of electromagnetic waves,* J. Math. Phys., Vol. 14, (1973), p. 1116.

[7] MUR, G. & A.J.A. NICIA, *Calculation of reflection and transmission coefficients in one-dimensional wave propagation problems,* J. Appl. Phys., Vol. 47, (1976), pp. 5218-5221.

[8] ISAACSON, E. & H.B. KELLER, *Analysis of Numerical Methods,* New York, Wiley, 1966, p. 340.

[9] MITTRA, R. (ed.), *Computer techniques for electromagnetics,* Oxford, Pergamon Press, 1973, Chapter 4, *Integral equation solutions of three-dimensional scattering problems,* by A.J. Poggio and E.K. Miller, p. 159.

[10] DEGENS, A.J., *The computation of skin effect and eddy currents in conducting cylinders using integral equation methods,* Master's Thesis, Laboratory of Electromagnetic Research, Dept. of Electrical Engineering, Delft University of Technology, Report No. 1978-5, in Dutch.

[11] TABBARA, W., *Light scattering by a single cylindrical lens: numerical results from a rigorous theory,* J. Opt. Soc. Am., Vol. 63, Jan. 1973, pp. 17-24.

[12] ABRAMOWITZ, M. & A. STEGUN, *Handbook of Mathematical Functions,* New York, Dover Publications Inc., 1968, p. 360.

# DIFFRACTION OF WATER WAVES

J.C.W. BERKHOFF

## 1. INTRODUCTION

Diffraction of water waves is the phenomenon that the propagation of a progressive field of surface water waves is disturbed by the presence of obstacles. Related to the diffraction problem is the problem of finding the wave field induced by the movement of oscillating objects.

The possibility for computing a diffracted wave field will be of great value for the engineers which have to design for instance off shore constructions and harbour lay-outs.

The diffraction problem is in general a time dependent three dimensional problem (in space). In practical application the time dependency will be restricted to harmonic oscillations and the number of space coordinates will be reduced to two if possible.

Mathematically the problem is to solve a three- or two-dimensional Laplace or Helmholtz equation on a domain which is bounded by boundaries of arbitrary shape, with more or less complicated boundary conditions. These kind of problems can be solved by the integral equation approach.

## 2. BASIC EQUATIONS

The diffraction formulation is based on the linear theory for water waves [1]. In that theory it is assumed that the fluid is ideal, the motion is irrotational so a potential function can be introduced, and the deviation of the free water surface from its equilibrium state is small (linear waves).

The three dimensional wave potential function $\Phi$ and the free surface function $\eta$ must satisfy the following equation (see definition sketch fig. 1)

$$\frac{\partial^2\Phi}{\partial x^2} + \frac{\partial^2\Phi}{\partial y^2} + \frac{\partial^2\Phi}{\partial z^2} = 0,$$

$$\frac{\partial\eta}{\partial t} - \frac{\partial\Phi}{\partial z} = 0 \qquad \text{at } z = 0,$$

$$\eta + \frac{1}{g}\frac{\partial\Phi}{\partial t} = 0 \qquad \text{at } z = 0,$$

$$\frac{\partial\Phi}{\partial z} = 0 \qquad \text{at } z = -h,$$

with:

$\Phi$ = three-dimensional time dependent wave potential function.

x,y,z = space coordinates

t = time

$\eta$ = free surface elevation function

h = constant waterdepth;

the two equations at z = 0 can be combined to one by eliminating the function $\eta$.

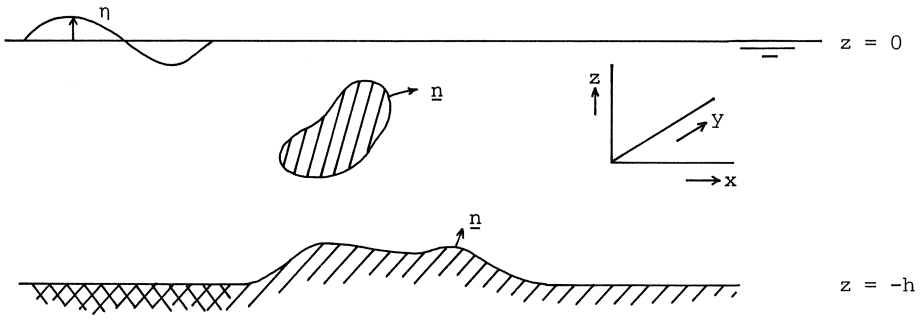



FIGURE 1

Definition Sketch

The solution is restricted to be simple harmonic in time so it is possible to write:

$$\Phi(x,y,z,t) = \phi_1(x,y,z)\cos \omega t + \phi_2(x,y,z)\sin \omega t$$

in which

$\omega$ = angular frequency

or with the aid of the complex function $\phi = \phi_1 + i\phi_2$, $\Phi = Re(\phi e^{-i\omega t})$. This function $\phi$ must also satisfy the equations (1) and (4). The free surface condition becomes:

$$(6) \qquad \frac{\partial \phi}{\partial z} - \frac{\omega^2}{g}\phi = 0 \qquad \text{at } z = 0.$$

Of course there are also needed boundary conditions in the horizontal plane and at the surface of obstacles, if present. When obstacles are present the most common boundary condition is the condition of normal velocity component at the surface

$$\frac{\partial \phi}{\partial n} = 0$$

in which $\partial/\partial n$ is the normal derivative.

In the horizontal directions there is the condition of a known incoming wave field and the condition that reflected or diffracted waves are only outgoing waves (radiation condition).

## 3. PRACTICAL PROBLEMS

### 3.1. Wave Forces on Three-Dimensional Structures of Arbitrary Shape.

The full three-dimensional problem must be solved when one is interested in the forces and moments exerted by waves on a structure of arbitrary shape.

The problem is treated in the case of a fixed structure (see Figure 2). For free floating bodies the problem is of course more complicated. The incoming wave field is given and will be indicated by the potential function $\tilde{\phi}(x,y,z)$ which is a solution of the Laplace equation and the boundary conditions at the free surface (5) and at the constant bottom (4).

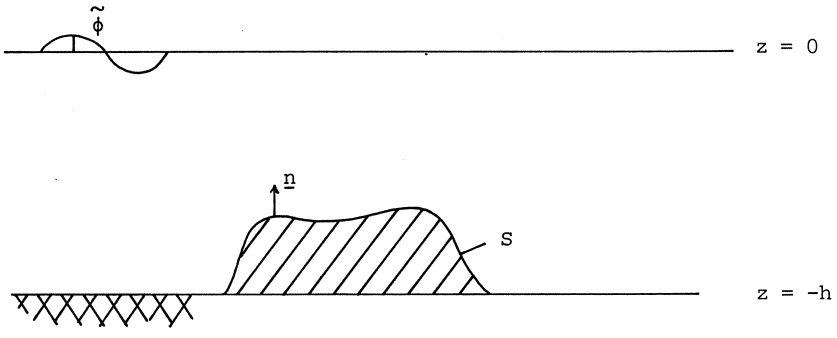



FIGURE 2

Three-Dimensional Diffraction Around
a Fixed Structure on the Bottom

The total potential $\phi$ can be written as the superposition of the potential $\tilde{\phi}$ of the incoming wave and a potential $\phi_d$ caused by the presence of the obstacle. This potential function $\phi_d$ must also be a solution of the Laplace equation

$$\Delta\phi_d = 0$$

and must fulfill the boundary conditions:

$$\frac{\partial\phi_d}{\partial z} - \frac{\omega^2}{g}\phi_d = 0 \qquad \text{at } z = 0,$$

$$\frac{\partial\phi_d}{\partial z} = 0 \qquad \text{at } z = h,$$

$$\frac{\partial\phi_d}{\partial n} = -\frac{\partial\tilde{\phi}}{\partial n} \qquad \text{at the surface S of the obstacle,}$$

and the radiation condition at infinity.

Once the diffracted wave potential $\phi_d$ has been found, the total wave potential $\phi$ can be computed by $\phi = \tilde{\phi}+\phi_d$, the pressure p at the surface S by

$$p = -\rho g z - \rho\frac{\partial\Phi}{\partial t}$$

and the wave forces by integration over the surface of the structure.

## 3.2. Diffraction around Cylindrical Obstacles

In the case that the obstacle is cylindrical and extending from the horizontal bottom to the free surface it is possible to reduce the three-dimensional problem to a two-dimensional in the horizontal x - y plane by separation of variables = $Z(z)\phi(x,y)$.

The dependency of the Z-direction appears to be as

$$Z(z) = \cosh\{k(h+z)\}$$

in which k is the real positive root of the equation

$$\omega^2 = g\, k \tanh(kh).$$

The two-dimensional potential function $\phi(x,y)$ must be now a solution of the Helmholtz equation

$$\frac{\partial^2\phi}{\partial x^2} + \frac{\partial^2\phi}{\partial y^2} + k^2\phi = 0$$

with boundary conditions at the contour C, which is the cross-section of the obstacle with the x - y plane, and at infinity, (see Figure 3).

At the contour C a reflection condition can be required and at infinity the condition that reflected and diffracted waves are outgoing radiating waves.

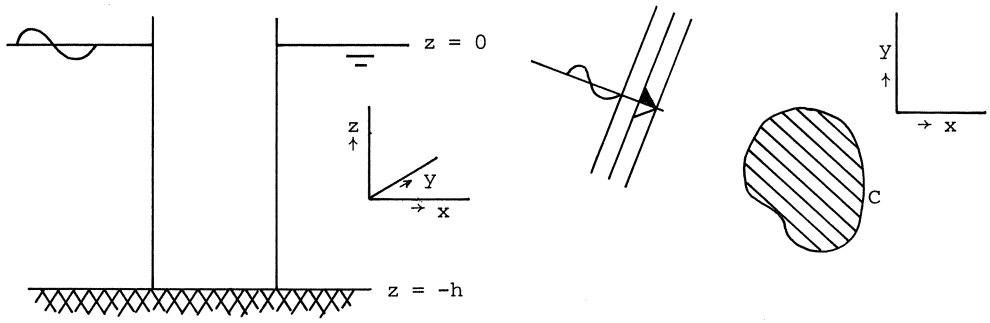



FIGURE 3

Diffraction around Cylindrical Obstacles

## 3.3. Wave Penetration into a Harbour of Arbitrary Shape

A modification of the problem mentioned in par. 3.2 is the wave penetration problem. If the interest is to compute the wave pattern in a harbour with one or more basins (see Figure 4) one has to solve the Helmholtz equation in each basin taking into account the reflection condition at fixed boundaries and continuity condition at the common open boundaries between the basins.

Also the radiation condition at the sea must be fulfilled.

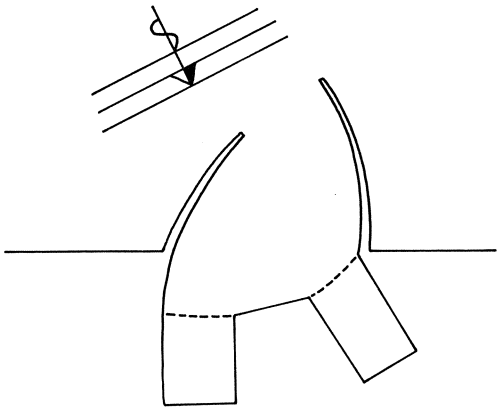


FIGURE 4

Wave Penetration
into a Harbour

3.4. Surface Waves Induced by an Oscillating Body

A two-dimensional example of this problem is the generation of waves in a flume by means of a forced oscillation of some object of arbitrary shape (see Figure 5).

In this case a two-dimensional Laplace equation must be solved in the x - z plane:

$$\frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial z^2} = 0,$$

with the boundary conditions

$$\frac{\partial \phi}{\partial z} - \frac{\omega^2}{g} \phi = 0 \qquad \text{at } z = 0,$$

$$\frac{\partial \phi}{\partial z} = 0 \qquad \text{at } z = -h,$$

$$\frac{\partial \phi}{\partial n} = f \qquad \text{at the contour of the object,}$$

and the radiation condition at infinity.

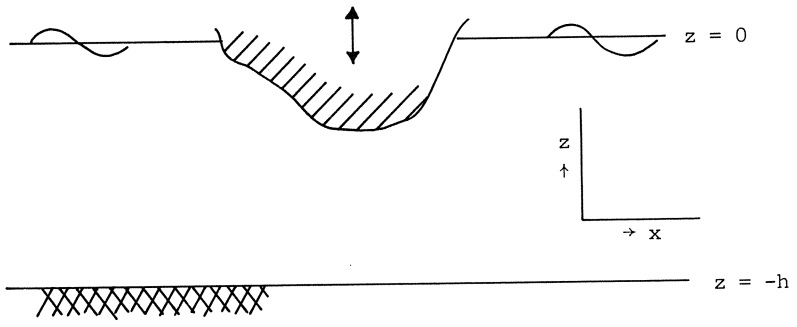



FIGURE 5

Oscillating Body
Wave Generation

## 4. Integral Equations

The problems described in par. 3 can be formulated in the form of an integral equation by using the principle of superposition of source solutions. In the three-dimensional case it can be derived that the solution in a point P outside the obstacle can be written as:

$$\phi(P) = \tilde{\phi}(P) + \iint_S \alpha(M) G(P;M) dO,$$

in which the point M is the integration point along the surface S. The function G must have the following properties:

(i) G as a function of the coordinates of the point P is a solution of the Laplace equation including the boundary conditions at the surface z = 0 and the bottom z = -h.

(ii) G also satisfies the radiation condition at infinity.

(iii) G is regular analytic in the whole domain except at the point M.

(iv) The singularity in G is of the form 1/R with R the distance between the two points P and M, so the function G+1/2πR is regular analytic in the whole domain outside the structure.

In the full three-dimensional problem the Green's function G is given by [2]:

$$G(P;M) = -\frac{1}{2\pi R} - \frac{1}{2\pi R'}$$
$$- \text{P.V.}\ \frac{1}{\pi}\int_0^\infty \frac{(k+\nu)e^{-kh}\cosh\{k(z+h)\}\cosh\{k(\zeta+h)\}}{\cosh(kh)\{k\tanh(kh)-\nu\}} J_0(kr)\,dk$$
$$- \frac{2(k^2-\nu^2)}{h(k^2-\nu^2)+\nu} i\cosh\{k(z+h)\}\cosh\{k(\zeta+h)\}J_0(kr),$$

$$P = (x,y,z), \quad M = (\xi,\eta,\zeta), \quad \nu = \frac{\omega^2}{g},$$

$$R = \sqrt{(x-\xi)^2+(y-\eta)^2+(z-\zeta)^2}, \quad R' = \sqrt{(x-\xi)^2+(y-\eta)^2+(z+\zeta+2h)^2},$$

$J_0(\ )$ = Bessel function,

P.V. = Cauchy principal value.

The function $\alpha$ along the surface of the body is related to the strength of the source distribution. This function can be computed with the aid of the boundary condition at the surface of the body.

Considering $\partial\phi/\partial n$ and taking the limit that P coincides with M, the following integral equation for the function $\alpha$ can be derived:

$$\left(\frac{\partial\phi}{\partial n}\right)_P = \alpha(P) + \iint_S \alpha(M)\,\frac{\partial G}{\partial n_P}\,dO_M; \qquad P \text{ and } M \text{ on } S,$$

in which the left-hand side is a known function using the boundary condition at the surface S.

In the case of the two-dimensional x - y diffraction problem the influence function is given by:

$$G(P;M) = \frac{1}{2i} H_0'(kr),$$

$$r = \sqrt{(x-\xi)^2+(y-\eta)^2}, \quad H_0'(\ ) = \text{Hankel function}.$$

The integral equation for the source strength distribution $\mu$ along the contour C is then:

$$\left(\frac{\partial\phi}{\partial n}\right)_P = \mu(P) + \int_C \mu(M)\frac{\partial G}{\partial n_P}\,ds_M; \qquad P \text{ and } M \text{ on } C.$$

Again the boundary condition at the contour C gives the information to compute the left-hand side of this equation. For wave penetration studies partial reflection must be taken into account. This can be done with the aid of a mixed boundary condition of the type:

$$\frac{\partial \phi}{\partial n} + ka\phi = 0$$

in which the coefficient a is related to the reflection coefficient. For the Laplace problem in the two-dimensional vertical x - z plane the Green's function sounds:

$$G(P;M) = \frac{1}{\pi} \ell n(\frac{r}{h}) + \frac{1}{\pi} \ell n(\frac{r'}{h}) +$$

$$- P.V. \frac{2}{\pi} \int_0^\infty \left[ \frac{(k+\nu) e^{-kh} \cosh\{k(z+h)\} \cosh\{k(\zeta+h)\}}{k \cosh(kh)\{k \tanh(kh) - \nu\}} \cos\{k(x-\xi)\} + \frac{e^{-kh}}{k} \right] dk$$

$$- \frac{2(k^2-\nu^2)}{h(k^2-\nu^2)+\nu} i \cosh\{k(z+h)\} \cosh\{k(\zeta+h)\} \frac{\cos\{k(x-\xi)\}}{k} ,$$

$$P = (x,z), \quad M = (\xi,\zeta),$$

$$r = \sqrt{(x-\xi)^2+(z-\zeta)^2}, \quad r' = \sqrt{(x-\xi)^2+(z+\zeta+2h)^2};$$

the integral equation for the source strength distribution function is given by:

$$(\frac{\partial \phi}{\partial n})_P = \alpha(P) + \int_C \alpha(M) \frac{\partial G}{\partial n_P} ds_M ; \qquad P \text{ and } M \text{ on } C.$$

## 5. NUMERICAL APPROACH

Because of the arbitrariness of the shape of the structures it is in general not possible to solve the integral equations analytically.

A numerical approach will be necessary. There are several reasons why the numerical approach is not so easy as it seems to be. The shape of the body or contour must be represented in a fairly good way by computing points, surface elements and normal vectors. The influence functions for the Laplace problems are very complex due to the free surface boundary condition. The integration interval for computing these influence functions is infinitely long and in this interval the integrand has a singular point so the Cauchy

principal value must be computed. The influence functions are also oscillating functions with respect to the horizontal space coordinates. Because of these difficulties the numerical method of solution has been taken simple and straightforward.

The surface of the body or the contour of the body or harbour in the two-dimensional case will be divided into finite surface elements. The shape of these elements will be in general plane (panels or facets in the three-dimensional problem and line-segments in the two-dimensional problem).

The unknown source strength distribution function will be approximated piecewise in each element by some shape function with a number of parameters.

Each element has one or more computing points and the integral equation together with continuity conditions between the approximating functions in the elements can give enough equations to compute the set of unknown parameters in the shape functions.

This approach results in the creation of a large set of linear equations for the element parameters.

For practical applications the most simple approach is used, taking a constant as the shape function in each element and only one collocation point in the center of the element in which the integral equation must be satisfied.

Even with this simple numerical approach a lot of work must be done for evaluating the matrix coefficients in the case of three- and two-dimensional Laplace problems with a free surface condition.

The described method of solution has much in common with the finite element method and therefore in literature this method is called the boundary element method.

The boundary element method can be used in combination with a finite element method to describe the solution in a far eventually infinite field taking into account some analytical properties (hybrid elements, infinite elements).

## 6. SOME RESULTS

As an example of a three-dimensional diffraction computation the problem of the computation of wave forces in a gravity type drilling platform is given.

The shape of the platform is given in Figure 6.

The incident waves are propagating in the negative y-direction. The computed wave force vector (Fx,Fz) during one wave period is given in Figure 7.

A comparison between a computed and measured wave pattern in a harbour (hydralic model) is given in the Figures 8, 9 .

The diffraction pattern around a number of cylindrical structures can be seen in Figure 10.

The potential belonging to an oscillating strip is given in Figure 11.

## 7. REFERENCES

[1] JOHN, F. (1950), *On the motion of floating bodies II, Simple harmonic motions,* Comm. on Pure and Appl. Math., 3, 45-100.

[2] WEHAUSEN, J.V. & E.V. LAITONE (1960), *Surface waves,* Encyclopedia of Physics, 9, 446-778.

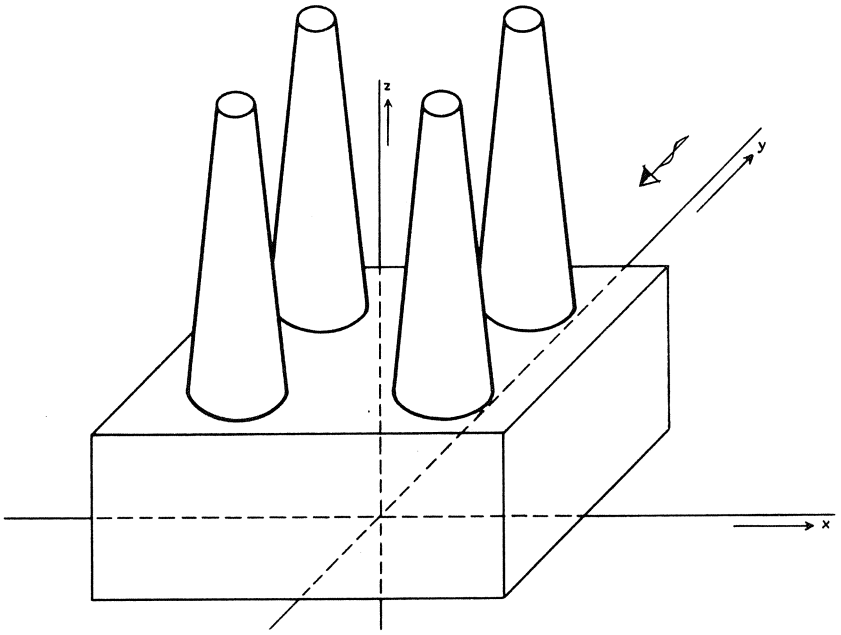



Figure 6

Shape of the platform

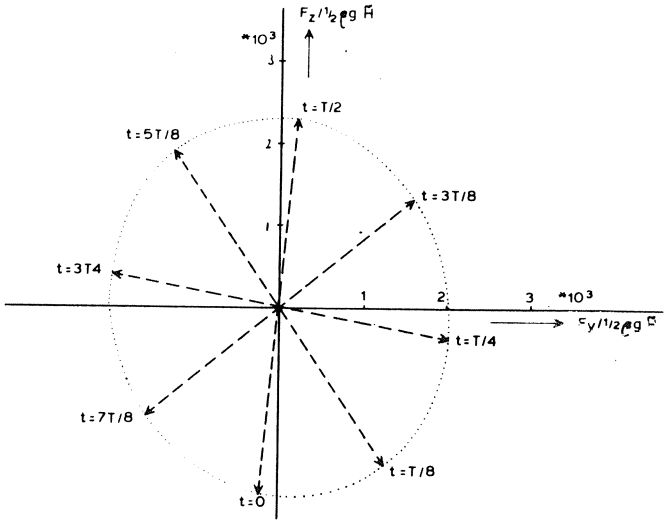
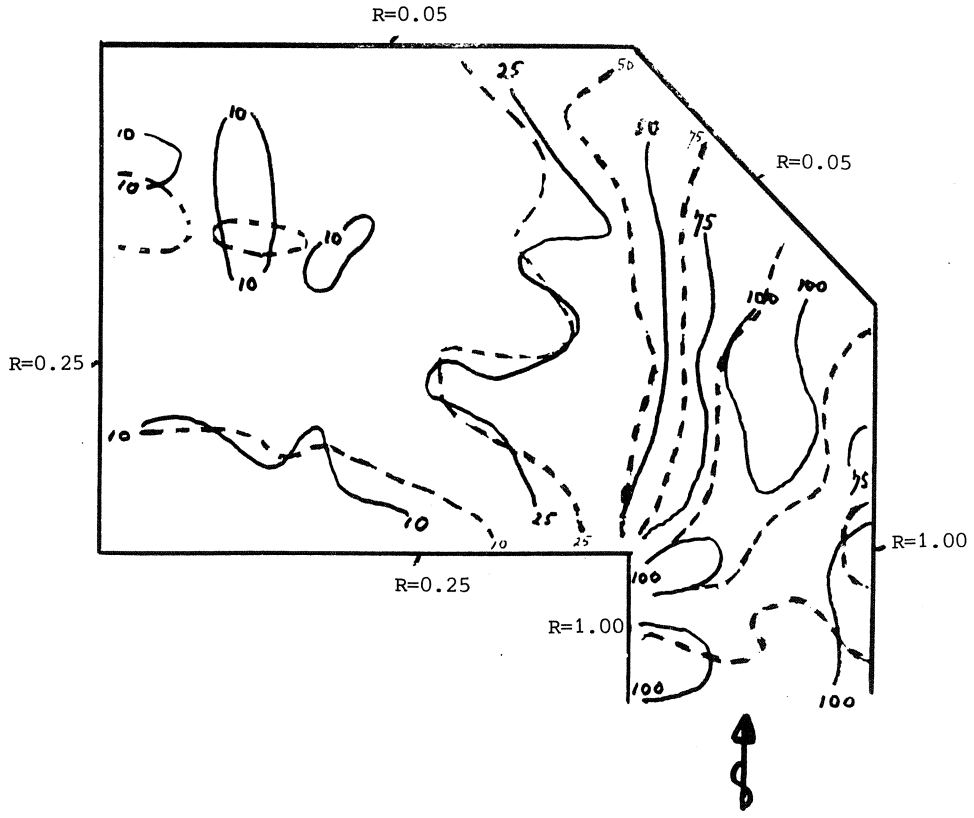



Figure 7

Wave force during one waveperiod




h = 0.30 m

T = 1.4 sec  R = REFL. COEFFICIENT

--- MEASUREMENT

___ COMPUTATION

Figure 8

Measured and computed wave pattern

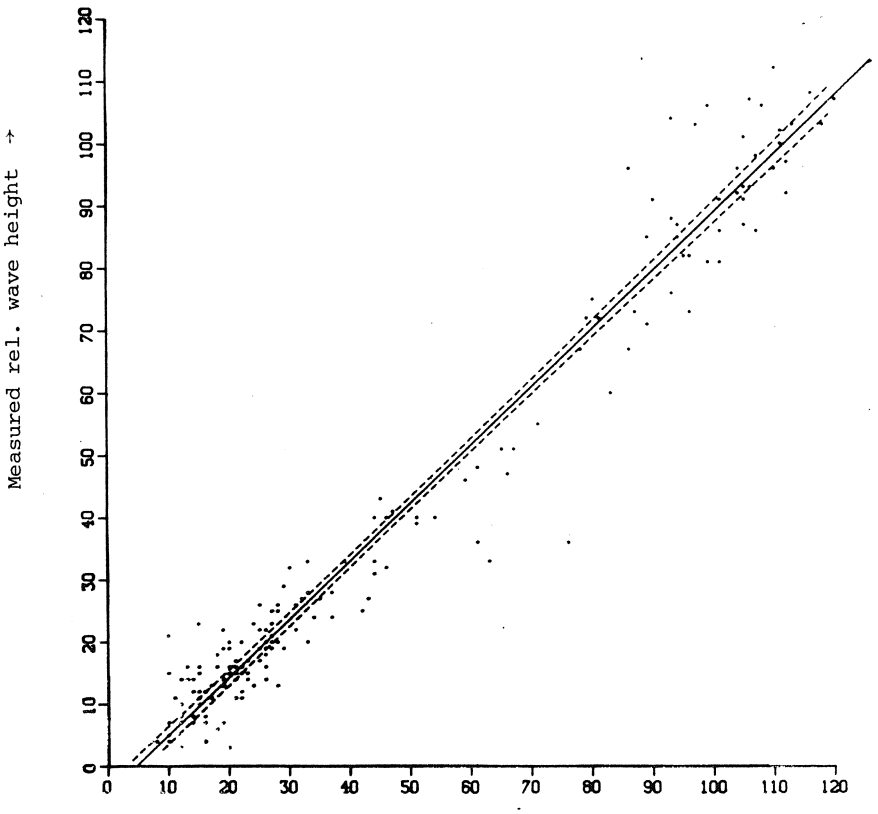



Figure 9
Comparison between measured and computed wave heights

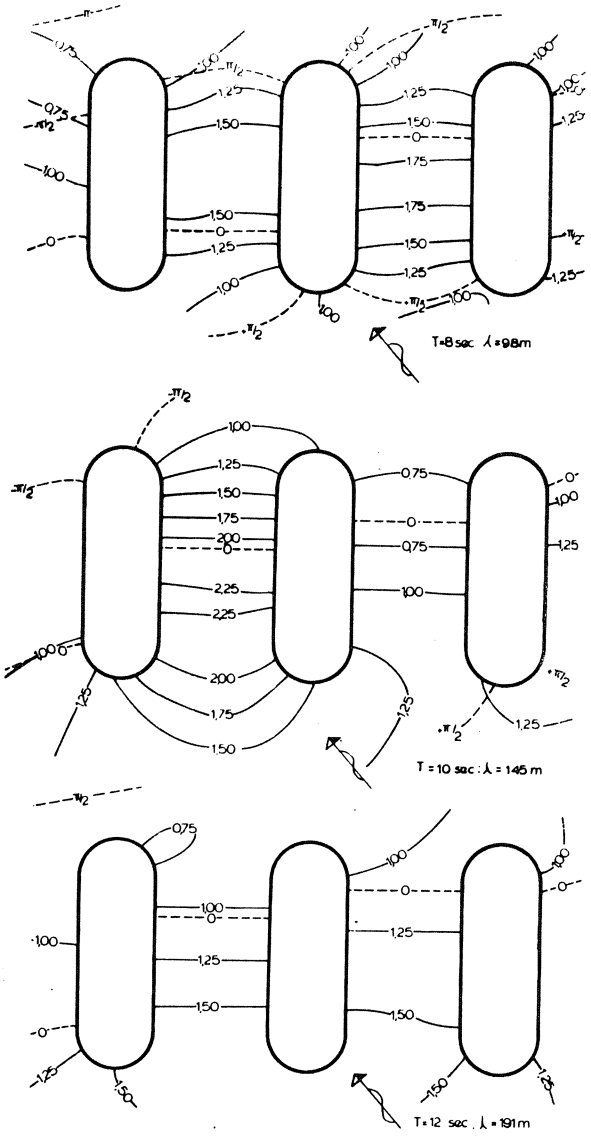



Figure 10

Diffracted wave pattern around three cylindrical structures

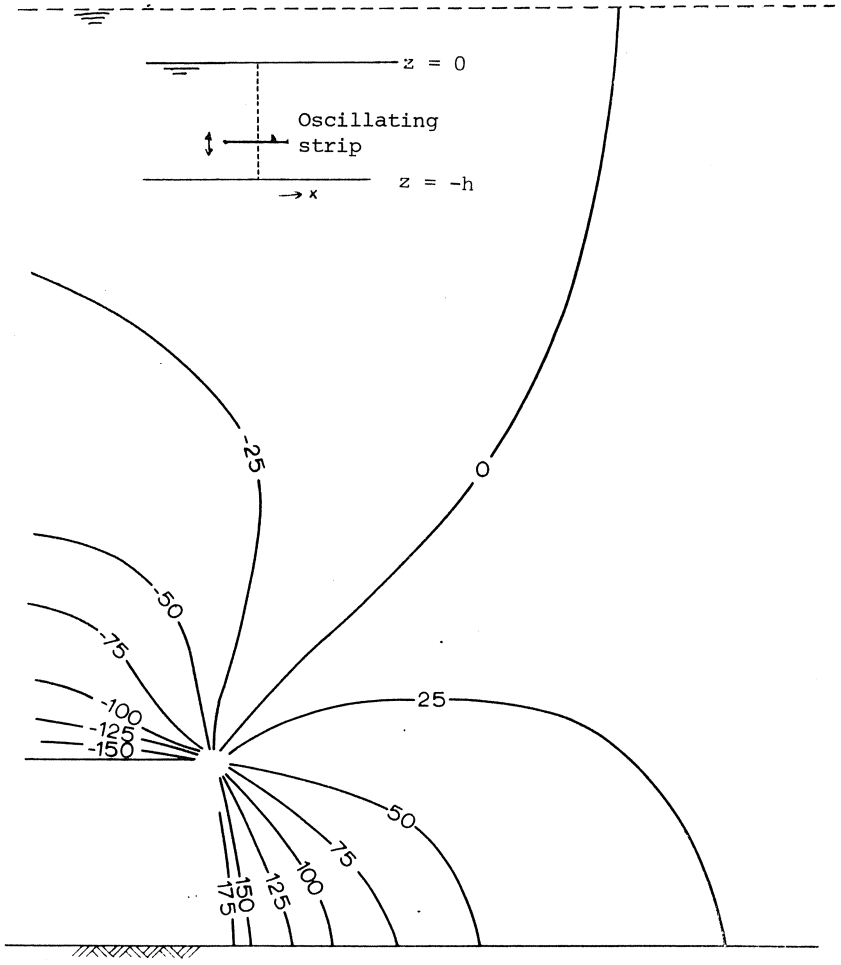



Figure 11

Potential lines belonging to an oscillating strip

# ADDRESSES OF CONTRIBUTORS

J.C.W. Berkhoff
Waterloopkundig Laboratorium, Laboratorium De Voorst
Postbus 152, 8300 AD Emmeloord

J.J.M. Cuppen
Instituut voor Toepassingen van de Wiskunde
Roetersstraat 15, 1018 WB Amsterdam

G. De Mey
Laboratorium voor Electronica, Rijksuniversiteit Gent
Sint Pietersnieuwstraat 41, 9000 Gent, België

O. Diekmann
Stichting Mathematisch Centrum, afdeling Toegepaste Wiskunde
2e Boerhaavestraat 49, 1091 AL Amsterdam

P.J. van der Houwen
Stichting Mathematisch Cemtrum, afdeling Numerieke Wiskunde
2e Boerhaavestraat 49, 1091 AL Amsterdam

H.K. Kuiken
Philips Natuurkundig Laboratorium
Eindhoven

Th. E. Labrujère
National Lucht - en Ruimtevaartlaboratorium
Postbus 153, 8300 AD Emmeloord

G. Mur
Afdeling der Electrotechniek, Technische Hogeschool Delft
Postbus 5031, 2600 GA Delft

H.J.J. te Riele
Stichting Mathematisch Centrum, afdeling Numerieke Wiskunde
2e Boerhaavestraat 49, 1091 AL Amsterdam

H. Schippers
Stichting Mathematisch Centrum, afdeling Numerieke Wiskunde
2e Boerhaavestraat 49, 1091 AL Amsterdam

R.W. de Vries
Onderafdeling der Toegepaste Wiskunde, Technische Hogeschool Twente
Postbus 217, 7500 AE Enschede

P.H.M. Wolkenfelt
Stichting Mathematisch Centrum, afdeling Numerieke Wiskunde
2e Boerhaavestraat 49, 1091 AL Amsterdam

P.J. Zandbergen
Onderafdeling der Toegepaste Wiskunde, Technische Hogeschool Twente
Postbus 217, 7500 AE Enschede

## UITGAVEN IN DE SERIE MC SYLLABUS

Onderstaande uitgaven zijn verkrijgbaar bij het Mathematisch Centrum, 2e Boerhaavestraat 49 te Amsterdam-1005, tel. 020-947272.

---

MCS 1.1 F. GÖBEL & J. VAN DE LUNE, *Leergang Besliskunde, deel 1: Wiskundige basiskennis*, 1965. ISBN 90 6196 014 2.

MCS 1.2 J. HEMELRIJK & J. KRIENS, *Leergang Besliskunde, deel 2: Kansberekening*, 1965. ISBN 90 6196 015 0.

MCS 1.3 J. HEMELRIJK & J. KRIENS, *Leergang Besliskunde, deel 3: Statistiek*, 1966. ISBN 90 6196 016 9.

MCS 1.4 G. DE LEVE & W. MOLENAAR, *Leergang Besliskunde, deel 4: Markovketens en wachttijden*, 1966. ISBN 90 6196 017 7.

MCS 1.5 J. KRIENS & G. DE LEVE, *Leergang Besliskunde, deel 5: Inleiding tot de mathematische besliskunde*, 1966. ISBN 90 6196 018 5.

MCS 1.6a B. DORHOUT & J. KRIENS, *Leergang Besliskunde, deel 6a: Wiskundige programmering 1*, 1968. ISBN 90 6196 032 0.

MCS 1.6b B. DORHOUT, J. KRIENS & J.TH. VAN LIESHOUT, *Leergang Besliskunde, deel 6b: Wiskundige programmering 2*, 1977. ISBN 90 6196 150 5.

MCS 1.7a G. DE LEVE, *Leergang Besliskunde, deel 7a: Dynamische programmering 1*, 1968. ISBN 90 6196 033 9.

MCS 1.7b G. DE LEVE & H.C. TIJMS, *Leergang Besliskunde, deel 7b: Dynamische programmering 2*, 1970. ISBN 90 6196 055 X.

MCS 1.7c G. DE LEVE & H.C. TIJMS, *Leergang Besliskunde, deel 7c: Dynamische programmering 3*, 1971. ISBN 90 6196 066 5.

MCS 1.8 J. KRIENS, F. GÖBEL & W. MOLENAAR, *Leergang Besliskunde, deel 8: Minimaxmethode, netwerkplanning, simulatie*, 1968. ISBN 90 6196 034 7.

MCS 2.1 G.J.R. FÖRCH, P.J. VAN DER HOUWEN & R.P. VAN DE RIET, *Colloquium Stabiliteit van differentieschema's, deel 1*, 1967. ISBN 90 6196 023 1.

MCS 2.2 L. DEKKER, T.J. DEKKER, P.J. VAN DER HOUWEN & M.N. SPIJKER, *Colloquium Stabiliteit van differentieschema's, deel 2*, 1968. ISBN 90 6196 035 5.

MCS 3.1 H.A. LAUWERIER, *Randwaardeproblemen, deel 1*, 1967. ISBN 90 6196 024 X.

MCS 3.2 H.A. LAUWERIER, *Randwaardeproblemen, deel 2*, 1968. ISBN 90 6196 036 3.

MCS 3.3 H.A. LAUWERIER, *Randwaardeproblemen, deel 3*, 1968. ISBN 90 6196 043 6.

MCS 4 H.A. LAUWERIER, *Representaties van groepen*, 1968. ISBN 90 6196 037 1.

MCS 5 J.H. VAN LINT, J.J. SEIDEL & P.C. BAAYEN, *Colloquium Discrete wiskunde*, 1968.
ISBN 90 6196 044 4.

MCS 6 K.K. KOKSMA, *Cursus ALGOL 60*, 1969. ISBN 90 6196 045 2.

MCS 7.1 *Colloquium Moderne rekenmachines, deel 1*, 1969. ISBN 90 6196 046 0.

MCS 7.2 *Colloquium Moderne rekenmachines, deel 2*, 1969. ISBN 90 6196 047 9.

MCS 8 H. BAVINCK & J. GRASMAN, *Relaxatietrillingen*, 1969.
ISBN 90 6196 056 8.

MCS 9.1 T.M.T. COOLEN, G.J.R. FÖRCH, E.M. DE JAGER & H.G.J. PIJLS, *Elliptische differentiaalvergelijkingen, deel 1*, 1970.
ISBN 90 6196 048 7.

MCS 9.2 W.P. VAN DEN BRINK, T.M.T. COOLEN, B. DIJKHUIS, P.P.N. DE GROEN, P.J. VAN DER HOUWEN, E.M. DE JAGER, N.M. TEMME & R.J. DE VOGELAERE, *Colloquium Elliptische differentiaalvergelijkingen, deel 2*, 1970.
ISBN 90 6196 049 5.

MCS 10 J. FABIUS & W.R. VAN ZWET, *Grondbegrippen van de waarschijnlijkheidsrekening*, 1970. ISBN 90 6196 057 6.

MCS 11 H. BART, M.A. KAASHOEK, H.G.J. PIJLS, W.J. DE SCHIPPER & J. DE VRIES, *Colloquium Halfalgebra's en positieve operatoren*, 1971.
ISBN 90 6196 067 3.

MCS 12 T.J. DEKKER, *Numerieke algebra*, 1971. ISBN 90 6196 068 1.

MCS 13 F.E.J. KRUSEMAN ARETZ, *Programmeren voor rekenautomaten; De MC ALGOL 60 vertaler voor de EL X8*, 1971. ISBN 90 6196 069 X.

MCS 14 H. BAVINCK, W. GAUTSCHI & G.M. WILLEMS, *Colloquium Approximatietheorie*, 1971. ISBN 90 6196 070 3.

MCS 15.1 T.J. DEKKER, P.W. HEMKER & P.J. VAN DER HOUWEN, *Colloquium Stijve differentiaalvergelijkingen, deel 1*, 1972. ISBN 90 6196 078 9.

MCS 15.2 P.A. BEENTJES, K. DEKKER, H.C. HEMKER, S.P.N. VAN KAMPEN & G.M. WILLEMS, *Colloquium Stijve differentiaalvergelijkingen, deel 2*, 1973. ISBN 90 6196 079 7.

MCS 15.3 P.A. BEENTJES, K. DEKKER, P.W. HEMKER & M. VAN VELDHUIZEN, *Colloquium Stijve differentiaalvergelijkingen, deel 3*, 1975.
ISBN 90 6196 118 1.

MCS 16.1 L. GEURTS, *Cursus Programmeren, deel 1: De elementen van het programmeren*, 1973. ISBN 90 6196 080 0.

MCS 16.2 L. GEURTS, *Cursus Programmeren, deel 2: De programmeertaal ALGOL 60*, 1973. ISBN 90 6196 087 8.

MCS 17.1 P.S. STOBBE, *Lineaire algebra, deel 1*, 1974. ISBN 90 6196 090 8.

MCS 17.2 P.S. STOBBE, *Lineaire algebra, deel 2*, 1974. ISBN 90 6196 091 6.

MCS 17.3 N.M. TEMME, *Lineaire algebra , deel 3*, 1976. ISBN 90 6196 123 8.

MCS 18 F. VAN DER BLIJ, H. FREUDENTHAL, J.J. DE IONGH, J.J. SEIDEL & A. VAN WIJNGAARDEN, *Een kwart eeuw wiskunde 1946-1971, Syllabus van de Vakantiecursus 1971*, 1974. ISBN 90 6196 092 4.

MCS 19 A. HORDIJK, R. POTHARST & J.Th. RUNNENBURG, *Optimaal stoppen van Markovketens*, 1974. ISBN 90 6196 093 2.

MCS 20 T.M.T. COOLEN, P.W. HEMKER, P.J. VAN DER HOUWEN & E. SLAGT, *ALGOL 60 procedures voor begin- en randwaardeproblemen*, 1976. ISBN 90 6196 094 0.

MCS 21 J.W. DE BAKKER (red.), *Colloquium Programmacorrectheid*, 1975. ISBN 90 6196 103 3.

MCS 22 R. HELMERS, F.H. RUYMGAART, M.C.A. VAN ZUYLEN & J. OOSTERHOFF, *Asymptotische methoden in de toetsingstheorie; Toepassingen van naburigheid*, 1976. ISBN 90 6196 104 1.

MCS 23.1 J.W. DE ROEVER (red.), *Colloquium Onderwerpen uit de biomathematica, deel 1*, 1976. ISBN 90 6196 105 X.

MCS 23.2 J.W. DE ROEVER (red.), *Colloquium Onderwerpen uit de biomathematica, deel 2*, 1976. ISBN 90 6196 115 7.

MCS 24.1 P.J. VAN DER HOUWEN, *Numerieke integratie van differentiaalvergelijkingen, deel 1: Eenstapsmethoden*, 1974. ISBN 90 6196 106 8.

MCS 25 *Colloquium Structuur van programmeertalen*, 1976. ISBN 90 6196 116 5.

MCS 26.1 N.M. TEMME (ed.), *Nonlinear analysis, volume 1*, 1976. ISBN 90 6196 117 3.

MCS 26.2 N.M. TEMME (ed.), *Nonlinear analysis, volume 2*, 1976. ISBN 90 6196 121 1.

MCS 27 M. BAKKER, P.W. HEMKER, P.J. VAN DER HOUWEN, S.J. POLAK & M. VAN VELDHUIZEN, *Colloquium Discretiseringsmethoden*, 1976. ISBN 90 6196 124 6.

MCS 28 O. DIEKMANN, N.M. TEMME (EDS), *Nonlinear Diffusion Problems*, 1976. ISBN 90 6196 126 2.

MCS 29.1 J.C.P. BUS (red.), *Colloquium Numerieke programmatuur, deel 1A, deel 1B*, 1976. ISBN 90 6196 128 9.

MCS 29.2 H.J.J. TE RIELE (red.), *Colloquium Numerieke programmatuur, deel 2*, 1976. ISBN 144 0.

* MCS 30 P. GROENEBOOM, R. HELMERS, J. OOSTERHOFF & R. POTHARST, *Efficiency begrippen in de statistiek*, . ISBN 90 6196 149 1.

MCS 31 J.H. VAN LINT (red.), *Inleiding in de coderingstheorie*, 1976. ISBN 90 6196 136 X.

MCS 32 L. GEURTS (red.), *Colloquium Bedrijfssystemen*, 1976. ISBN 90 6196 137 8.

MCS 33 P.J. VAN DER HOUWEN, *Differentieschema's voor de berekening van waterstanden in zeeën en rivieren*, 1977. ISBN 90 6196 138 6.

MCS 34 J. HEMELRIJK, *Oriënterende cursus mathematische statistiek*, ISBN 90 6196 139 4.

MCS 35 P.J.W. TEN HAGEN (red.), *Colloquium Computer Graphics*, 1977. ISBN 90 6196 142 4.

MCS 36 J.M. AARTS, J. DE VRIES, *Colloquium Topologische Dynamische Systemen*, 1977. ISBN 90 6196 143 2.

MCS 37 J.C. van Vliet (red.), *Colloquium Capita Datastructuren*, 1978. ISBN 90 6196 159 9.

MCS 38.1 T.H. KOORNWINDER (ED.), *Representations of locally compact groups with applications*, 1979. ISBN 90 6196 161 0.

MCS 38.2 T.H. KOORNWINDER (ED.), *Representations of locally compact groups with applications*, 1979. ISBN 90 6196 181 5.

MCS 39 O.J. VRIEZE & G.L. WAANROOIJ, *Colloquium Stochastische spelen, 1978*. ISBN 90 6196 167 X.

MCS 40 J. VAN TIEL, *Convexe Analyse*, 1979. ISBN 90 6196 187 4.

MCS 41 H.J.J. TE RIELE (ED.), *Colloquium Numerical Treatment of Integral Equations*, 1979. ISBN 90 6196 189 0.

De met een * gemerkte uitgaven moeten nog verschijnen.